Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.027467

ARTICLE

An Improved High Precision 3D Semantic Mapping of Indoor Scenes from RGB-D Images

Jing Xin[1,*], Kenan Du[1], Jiale Feng[1] and Mao Shan[2]

[1]Shaanxi Key Laboratory of Complex System Control and Intelligent Information Processing, Xi'an University of Technology, Xi'an, 710048, China

[2]Australian Centre for Field Robotics, The University of Sydney, Sydney, 2006, Australia

*Corresponding Author: Jing Xin. Email: xinj@xaut.edu.cn

Received: 31 October 2022 Accepted: 21 March 2023 Published: 03 August 2023

ABSTRACT

This paper proposes an improved high-precision 3D semantic mapping method for indoor scenes using RGB-D images. The current semantic mapping algorithms suffer from low semantic annotation accuracy and insufficient real-time performance. To address these issues, we first adopt the Elastic Fusion algorithm to select key frames from indoor environment image sequences captured by the Kinect sensor and construct the indoor environment space model. Then, an indoor RGB-D image semantic segmentation network is proposed, which uses multi-scale feature fusion to quickly and accurately obtain object labeling information at the pixel level of the spatial point cloud model. Finally, Bayesian updating is used to conduct incremental semantic label fusion on the established spatial point cloud model. We also employ dense conditional random fields (CRF) to optimize the 3D semantic map model, resulting in a high-precision spatial semantic map of indoor scenes. Experimental results show that the proposed semantic mapping system can process image sequences collected by RGB-D sensors in real-time and output accurate semantic segmentation results of indoor scene images and the current local spatial semantic map. Finally, it constructs a globally consistent high-precision indoor scenes 3D semantic map.

KEYWORDS

3D semantic map; online reconstruction; RGB-D images; semantic segmentation; indoor mobile robot

1 Introduction

Simultaneous localization and mapping (SLAM) is a crucial technology for autonomous robot navigation, aiming to reconstruct the map of the robot's surroundings while simultaneously determining the location of the robot itself relative to this map [1,2]. Visual SLAM (VSLAM) refers to only using visual information to complete the task of SLAM. The scenes map constructed by VSLAM, such as feature map and point cloud map, can be used not only for localization but also for obstacle avoidance and navigation. However, these map models lack a more detailed understanding of the functions and attributes of objects in the environment, which makes it difficult for robots to fully comprehend information from unknown scenes [3–6].

Environment understanding is mainly categorized into image classification, object detection and semantic segmentation. Compared with image classification and object detection, image semantic segmentation provides more detailed interpretability and is widely used in autonomous driving, human-computer interaction, medical imaging, and augmented reality [7,8]. Traditional image semantic segmentation methods classify objects based on manually selected features, which are often limited in expressing the target object, particularly in a complex environment and with changing illumination. In recent years, the continuous improvement and development of Deep Learning (DL) technology in the field of image perception have provided a new opportunity for improving the environment perception ability of intelligent robots. In particular, the outstanding achievements of Convolutional Neural Network (CNN) in image classification [8,9] have established a solid theoretical basis to help robots improve their environment perception ability. Many scholars have started to combine deep learning-based object detection, semantic segmentation and visual SLAM to establish an accurate mapping relationship between rich semantic information and spatial map points [10–14]. By using the semantic map, the robot system can not only get the spatial information of the environment and also obtain the semantic information of the objects in the environment. These advantages can enable robots to understand environmental information at the semantic level and imitate the way the human brain understands the environment, which is very crucial for achieving a higher level of intelligent operation [15,16].

Researchers in different fields also have varying definitions and understandings of "semantic map". According to the way of scene annotation, semantic map construction methods can be roughly divided into two categories: semantic map based on object detection boundary box annotation and semantic map based on pixel level semantic annotation.

1.1 Semantic Map Based on Object Detection Boundary Box Annotation

In this kind of mapping method, semantic annotation of objects is carried out by various object detection algorithms to obtain the relative position relationship between objects and robots in the scene. In 2017, Mur-Artal et al. [17] proposed an object-level semantic map construction method, which uses ORB-SLAM [18] to construct a 3D point cloud map of the surrounding environment. In the meantime, Single Shot MultiBox Detector (SSD) [19], a fast object detection network, is adapted to generate object classification bounding boxes on the key frame images filtered by SLAM. Finally, the point cloud corresponding to the current frame image is merged and the 3D semantic map is consequently updated by using the correlation of object information of the neighbors. However, in practical applications, the problem of object missing detection in the semantic map constructed by this algorithm is relatively serious, especially for the detection of small-scale objects. In addition to pure SSD, other object detection networks such as YOLO v2, YOLO v3, Mobile Net+SSD have also been used in semantic map construction [20–22]. In [22], dynamic and static feature point detection algorithm is added to semantic mapping to eliminate the dynamic objects, thus further realizing the semantic mapping in dynamic scenes. The semantic annotation method based on the object detection bounding box can provide the relative position relationship between the target object and the robot in the geometric model. However, the lack of accurate boundary information of the target object has caused a certain degree of interference for the lack of accurate boundary information of the target

object has caused a certain degree of interference for the robot to complete intelligent navigation, human-computer interaction and other complex tasks. Therefore, the construction of high-precision semantic map based on pixel-level semantic annotation has attracted the attention of many researchers.

1.2 Semantic Map Based on Pixel-Level Semantic Annotation

In mapping methods of this kind, the image semantic segmentation algorithm is adopted to provide the pixel-level category annotation for key frame images. Compared with object detection-based annotation approaches, this type of method provides more detailed scene interpretation, yet at the expense of higher computational complexity. Riazuelo et al. [23] developed a cloud-based robot semantic mapping system RoboEarth, which utilizes prior information to construct an environment map, and then uses knowledge-based reasoning to search for new objects. This method can obtain the sub-database related to the current task through semantic reasoning and improve the recognition rate by reducing the computational complexity and the false alarm rate. However, it can only build a sparse map model that does not reflect the detailed features in the environment. To address this limitation, McCormac et al. [24] proposed a dense semantic mapping method, which firstly uses ElasticFusion [25], a Surfel based SLAM mapping method to estimate the camera pose, and then uses DeconvNet [26] to predict object category label at pixel level, and finally integrate the image semantic segmentation results and the associated information of the point cloud map into a unified global dense semantic map by combining Bayesian updating and Conditional Random Fields (CRF). However, due to high count of parameters in DeconvNet, the time cost of obtaining semantic map model and the required computer resources are prohibitively high. In order to further improve the real-time performance of the system, Zhao et al. [27] proposed a pixel-voxel network, which can simultaneously identify and label the semantic categories of each point cloud in a 3D map. Also, a softmax weighted fusion stack is proposed, which can adaptively learn the different contributions of pixel network and voxel network and fuse according to their respective confidence, but the real-time performance of the system still needs to be improved. Chen et al. [28] designed a semantic mapping framework and proposed an improved Bayesian update model, which filters dynamic factors based on the prior knowledge generated by object categories and observations to improve the positioning accuracy in dynamic scenes. Subsequently, Bao et al. [29] proposed a dense semantic map construction system for the urban environment, which integrates semantic information obtained from deep neural networks into odometer and closed-loop detection to improve localization and mapping accuracy. Experimental results on data sets show that the system can build semantic maps for the urban environment, but the test results in real scenes have not been provided yet.

To sum up, the semantic map constructing algorithm based on objects can obtain a map model containing objects' type and location information in the real world. However, the semantic map object category is sparse and lacks the accurate positioning of object contour, leading to some error in the indoor navigation of intelligent robot. However, the semantic map construction algorithm based on pixels can obtain more detailed object information. Yet, this algorithm is of high complexity and is difficult to be applied in practical application, especially in complex indoor environments with severe occlusion and light illumination change. Although a series of research achievements have been made in the current study on environment perception based on deep learning, it still cannot achieve the goal of constructing a high-precision semantic map of the target scene in real time. Based on this, this paper proposes an improved high-precision 3D semantic mapping of indoor scenes from

RGB-D images, which uses image semantic segmentation as the annotation method to construct a high-precision semantic map of indoor scenes in real time. The ElasticFusion, a graph optimization based visual SLAM algorithm, is used to select the key frame sequence and construct the spatial model of indoor scenes; As for image semantic segmentation, VGG-16, which has powerful feature extraction capability, is used as the barebone network, and the segmentation performance of image semantic segmentation network is improved by means of image geometric structure information. Furthermore, an indoor RGB-D image semantic segmentation network with multi-scale features is proposed, which provides more detailed semantic annotation for the spatial point cloud model. The experimental results of semantic mapping in a real mobile robot demonstrate the ability of the proposed semantic mapping algorithm to construct a scene map with high-precision semantic information in real-time, helping robots to effectively perceive and understand the surrounding environment.

The main contributions of the proposed algorithm are as follows:

1) A novel indoor RGB-D image semantic segmentation network with multi-scale features fusion is proposed. The network integrates the visual color features and depth geometry features of RGB-D images to improve the accuracy of image semantic segmentation.

2) An improved method of high-precision 3D semantic mapping of indoor scenes from RGB-D images is proposed. The proposed mapping method combines the RGB-D image semantic segmentation algorithm with the traditional SLAM technology, and facilitates the robot system to construct a globally consistent high-precision indoor scenes 3D semantic map in real-time. This feature can also find its potential use in other applications such as autonomous driving.

The organization of this paper is as follows: Section 2 gives a brief introduction on the overall framework of the proposed algorithm, and the detailed introduction on the principle of the three main components of the system. The experimental results and performance analysis are presented in Section 3. Finally, conclusions and future works are given in Section 4.

2 Proposed Method

The block diagram of the proposed algorithm for 3D semantic mapping of indoor scenes from RGB-D images proposed in this paper is shown in Fig. 1. This system is mainly composed of RGB-D SLAM, indoor RGB-D image semantic segmentation network with multi-scale features fusion and semantic fusion module. Firstly, a popular RGB-D SLAM algorithm is used to select key frame images according to the image sequence captured by the RGB-D camera, and at the same time, inter-frame pose optimization and indoor scenes 3D model construction is carried out. Secondly, a indoor RGB-D image semantic segmentation network with multi-scale features fusion is proposed to segment the key frame image and predict the category at the pixel level. Then, a semantic fusion module is performed, i.e., the semantic annotation of the 3D map model is done incrementally using Bayesian updates, and the data correlation between multiple frames is used to optimize the consistency of the labeled semantic model category labels through conditional random fields (CRF), and finally a globally consistent 3D semantic map model is obtained. The implementation details of each module are as follows.

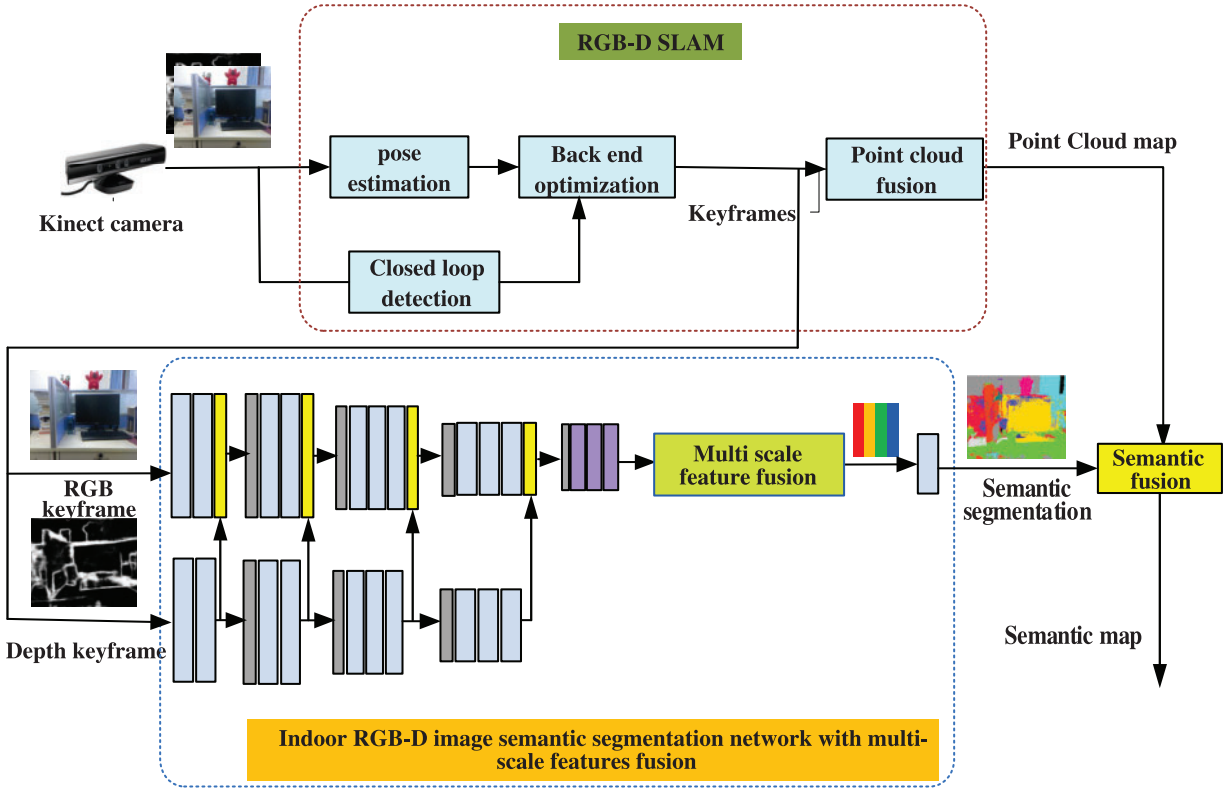


Figure 1: Overall block diagram of the proposed 3D semantic mapping algorithm of indoor scenes from RGB-D images

2.1 Indoor RGB-D Image Semantic Segmentation with Multi-Scale Feature Fusion

Indoor environment scenes are complex, and the semantic segmentation results obtained only using color features are easily affected by external factors such as illumination and occlusion, etc. Depth image contains geometric information about objects, which is helpful in improving the performance of the semantic segmentation network to a certain extent. Based on this, the image depth information is adopted to improve the accuracy of the semantic segmentation model in this paper. Depth images can often be captured by RGB-D sensors such as Kinect cameras. However, the depth values in the depth image are not all valid. There may also be wrong values. Therefore, this paper takes image depth information as auxiliary information for image semantic segmentation and proposes an indoor RGB-D image semantic segmentation network with multi-scale features fusion.

As shown in Fig. 2, the proposed semantic segmentation network is mainly composed of three parts: primary branch network, sub-branch network, and multi-scale feature fusion module. Among them, the inputs of the primary branch network and sub-branch network are RGB image and depth image, respectively. The depth features extracted from the sub-branches are fused with the color features extracted from the primary branches step by step, and then the features extracted from the multi-scale feature fusion module are extracted with different scales to obtain more accurate semantic segmentation results.

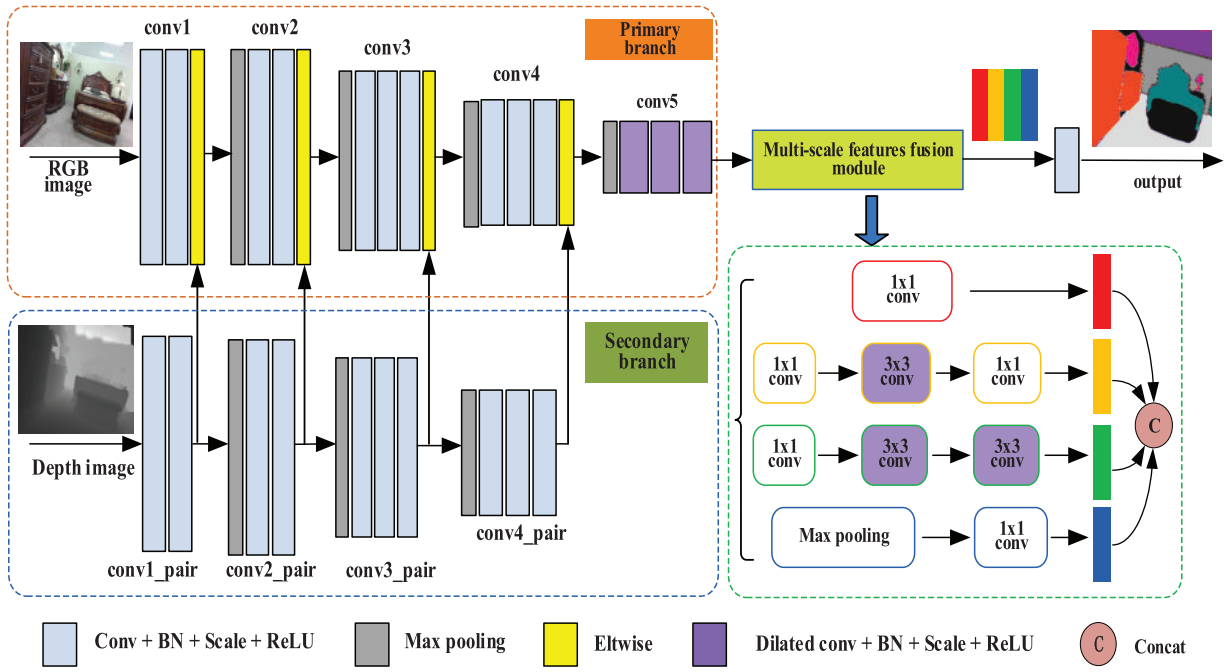


Figure 2: Some overall block diagram of indoor RGB-D image semantic segmentation network with multi-scale features fusion functions

In this paper, a simple and powerful VGG-16 network is adopted as the backbone network to extract image color features and geometric features, and the structure of the VGG-16 network is modified as follows:

(1) Pool5 pooling layer is removed to alleviate the loss of feature images resolution due to the down-sampling operation of the pooling layer;

(2) The fifth convolutional layer in the sub-branch is removed, and the fifth convolutional layer in the primary branch is modified into a dilated convolution with hole 2 to expand the receptive field of convolution operation without adding network parameters;

(3) In the meantime, the last two full connection layers of the VGG-16 network were removed in both branches to avoid losing the spatial information of the image.

2.1.1 Dilated Convolution

Obtaining dense and rich feature information is helpful for the efficient completion of segmentation tasks. In order to obtain richer feature information of the input object under the condition that the network parameters of each layers remain stable, we use dilated convolution to insert zeros between the filter templates in the process of CNN structure design and network construction, so that each convolution mapping feature map can contain a wide range of image information of the input object. For two-dimensional input signals, dilated convolution is similar to zero padding between sampled points of ordinary convolution, but the number of parameters and convolution operation remains unchanged because the number of multiply-add operations at each pixel position of the original image (feature map) has not changed.

Based on this, dilated convolution with the coefficient of dilation rate of 2 is introduced into all convolution layers of the last group of convolutions in the main primary branch network to eliminate the feature aliasing phenomenon generated after the final fusion of image geometric features and visual features. Inspired by the comparative experiment of dilated convolution performance with different sampling ratios in [30], the dilated convolution with a dilation rate of 12 is introduced into each $3 * 3$ convolution layer in the multi-scale feature fusion module described in the next section. Lastly, other convolution layers in the network adopt ordinary convolution operations.

2.1.2 Multi-Scale Features Fusion

In this paper, a multi-scale feature fusion module is constructed to extract the features of different scales from color features and geometric features after multiple fusion, and it obtains semantic and detailed information of objects at different levels by means of channel splicing. Since the computational cost of the $5 * 5$ convolution operation is 2.78 times that of the $3 * 3$ convolution operation, the $5 * 5$ convolution layer is decomposed into the series connection of two $3 * 3$ convolution layers, which increases the nonlinearity of the convolution feature map. In the multi-scale feature fusion module, simple and easy-to-operate conventional convolution and maximum pooling, such as $1 * 1$ and $3 * 3$, are applied to the same feature map to realize feature extraction of the same object at different scales. The specific design structure of the fusion module is shown in Fig. 3.

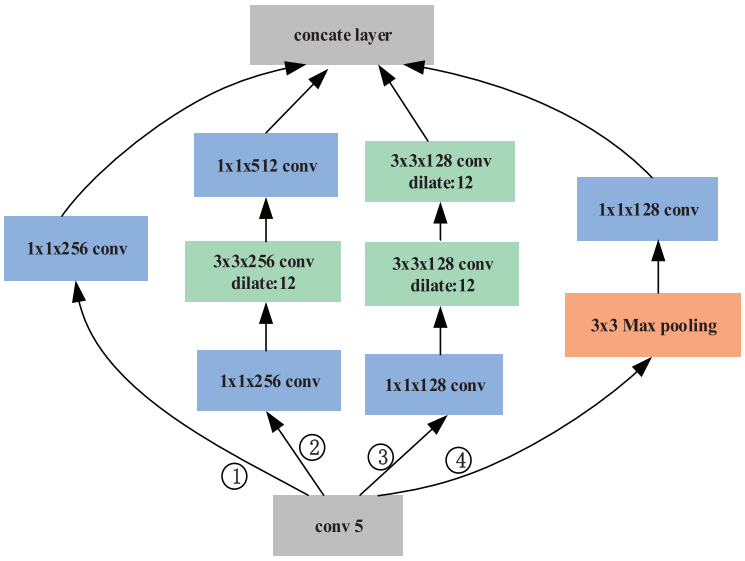


Figure 3: Block diagram of multi-scale feature fusion module

Considering that the geometric feature and color feature of the input image are finally fused, followed by extraction of features by the fifth convolution layer, the output 512-dimensional image features still lead to a large amount of computation. In addition, depthwise separable convolution [31] is further introduced here. As shown in Fig. 4, this convolution operation decomposes the traditional convolution process into two parts: Depthwise Convolution and Pointwise Convolution. Firstly, different convolution kernels are used for different input channels. Then the convolution feature maps of each channel are spliced into a new channel feature map group. Following that, the standard $1 * 1$ cross-channel convolution operation is performed. Assuming that the number of input channels is 192, the required number of output channels is 256, and the convolution kernel size is $3 * 3$, the

parameters of traditional convolution operation are $3 \times 3 \times 192 \times 256 = 442{,}368$, while the parameters of depthwise separable convolution are $3 \times 3 \times 192 + 192 \times 1 \times 1 \times 256 = 1728 + 49152 = 50{,}880$. Therefore, by using depthwise separable convolution, the number of parameters can be reduced to about 1/9 of that of traditional convolution, which greatly improves the running speed of the algorithm.

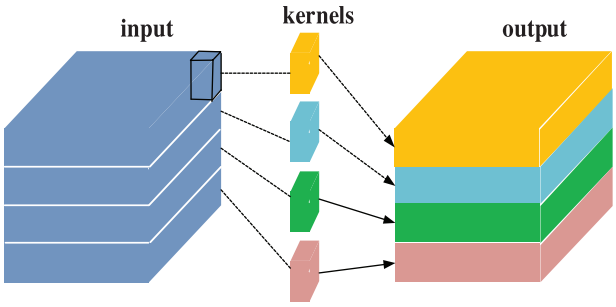


Figure 4: Schematic diagram of depthwise separable convolution

2.1.3 Weighted Loss

The weight calculation of each category in the data set is only for the training data set. Due to the serious uneven distribution of various categories of objects in the indoor image data set, the prediction accuracy of categories with large sample size is high or even over-fitted, while the prediction accuracy of categories with small sample size is unsatisfactory. In this paper, the weighted cross loss function [32] is used. Taking SUN RGB-D data set as an example, the weight of each category can be calculated by Eq. (1).

$$F_c = \frac{N_c}{P_c * W * H}$$

$$F_c = \frac{N_c}{P_c * W * H} \tag{1}$$

where, F_c represents the frequency of a certain type of pixel, N_c represents the total amount of such pixels in the training set, P_c represents the sample size of such pixels in the training set, and W, H are image sample sizes, so the weight W_c corresponding to each category can be obtained by the ratio of the median frequency M_F_c of all types of pixel to the frequency F_c of such pixel, as shown in the Eq. (2):

$$W_c = M\_F_c / F_c \tag{2}$$

The weights of 13 classes of objects in the SUN RGB-D data set are calculated by the above Eqs. (1) and (2), as shown in Table 1. The larger the weight is, the smaller the number of such pixels in the training set will be. The distribution of samples in SUNRGB-D training data set can be seen intuitively from the weights of samples shown in Table 1. Specifically, the weights of books and pictures are far greater than those of other types of objects, indicating that the sample sizes of these two types of objects are relatively small. On the other hand, the weights of beds, floors, sofas and walls are relatively small, indicating that these types of objects appear frequently.

Table 1: Weights of each category in SUN RGB-D data set

No.	Category	Weight	No.	Category	Weight
0	Background	0.5385	7	Object	1.9911
1	Bed	0.5425	8	Picture	4.3769
2	Book	4.8295	9	Sofa	0.6834
3	Ceiling	2.1971	10	Table	1.0519
4	Chair	0.8426	11	TV	2.6886
5	Floor	0.6415	12	Wall	0.5522
6	Furniture	0.9531	13	Window	1.3647

2.2 RGB-D SLAM

The traditional SLAM mapping method generally improves the accuracy of pose estimation. It builds the map model by constantly optimizing camera pose or feature points, which consumes a lot of computing resources and is difficult to achieve real-time mapping. Therefore, in this paper, the visual SLAM algorithm based on graph optimization—ElasticFusion [25] is adopted to construct the indoor environment spatial point cloud map in real time. The environmental data collected by the Kinect sensor is used to estimate the robot's pose. Then the selected key frame sequence is input into the indoor RGB-D image semantic segmentation network with multi-scale features as described in Section 2.1, so as to obtain the semantic segmentation result of a single frame key frame image.

2.3 Semantic Fusion and Model Optimization

In an unknown environment without prior information, when an RGB-D sensor is used to collect the image sequence of the surrounding environment, the 2D semantic segmentation results of a single frame image may lead to the inconsistency of semantic labels at the same point between two successive key frames. Therefore, we adopt the incremental fusion of semantic label to obtain the semantic label of point cloud map, and associates the semantic label in several consecutive key frames by recursive Bayesian updating. Assuming that the category of a point c in the 3D model at the current moment is c_t, and all the pixel measurement values related to this point can be expressed as $k_0^t = \{k_0, k_1, \ldots, k_t\}$, then the label probability of this point c can be expressed as Eq. (3).

$$p\left(c_t|k_0^t\right) = \frac{1}{Z_t} p\left(k_t|k_0^{t-1}, c_t\right) p\left(c_t|k_0^{t-1}\right) \tag{3}$$

where, $Z_t = p\left(k_t|k_0^{t-1}\right)$is the normalization factor. For $p\left(k_t|k_0^{t-1}, c_t\right)$, the first-order Markov hypothesis is adopted, that is, $p\left(c_t|k_0^{t-1}\right) \approx p\left(c_{t-1}|k_0^{t-1}\right)$.

According to the smoothness of the posterior hypothesis, $p\left(c_t|k_0^{t-1}\right) \approx p\left(c_{t-1}|k_0^{t-1}\right)$ then the above equation can be written as:

$$\begin{aligned} p\left(c_t|k_0^t\right) &= \frac{1}{Z_t} p\left(k_t|c_t\right) p\left(c_{t-1}|k_0^{t-1}\right) \\ &= \frac{1}{Z_t} \frac{p\left(c_t|k_t\right) p\left(k_t\right)}{p\left(c_t\right)} p\left(c_{t-1}|k_0^{t-1}\right) \end{aligned} \tag{4}$$

So, the current state $p\left(c_{t-1}|k_0^{t-1}\right)$ of the 3D point c can be obtained, and the state can be updated according to the 2D semantic segmentation results of subsequent frames. As the prior probability $p\left(k_t\right)$ is fixed, a new normalized factor Z_t' can be obtained by fusing it with Z_t, and the posterior probability is represented by $\hat{p}(c_t|k_t)$. Finally, the semantic label of 3D point cloud could be updated according to Eq. (5).

$$p(c_t|k_0^t) \leftarrow \frac{1}{\mathbf{Z}_t'}\frac{\hat{p}(c_t|k_t)}{p(c_t)}p(c_{t-1}|k_0^{t-1}) \tag{5}$$

After obtaining the 3D model with semantic labels, in order to smooth the noise, this paper adopts the same optimization method as in [24] to optimize the 3D semantic map model. That is, CRF is applied to optimize the label probability of 3D semantic map, and the spatial semantic map with the consistent label of the point cloud is built with the constraint of the unity of spatial coordinates and colors of the same object between consecutive key frames.

3 Experiment

In order to verify the effectiveness of the image semantic segmentation algorithm and semantic map online construction algorithm proposed in this paper, two sets of experiments were carried out, namely, image semantic segmentation in standard semantic segmentation data sets and semantic map online construction in a real scene.

3.1 Image Semantic Segmentation

The purpose of this experiment is to verify the effectiveness and superiority of the proposed indoor RGB-D image semantic segmentation network with multi-scale features fusion through two semantic segmentation data sets: SUN RGB-D and NYUDv2. In Ubuntu16.04 system, the Convolutional Architecture for Fast Feature Embedding (Caffe) deep learning framework is used to complete the construction of indoor RGB-D image semantic segmentation network with multi-scale features fusion. NVIDIA GeForce GTX 1070 graphics card is used for offline training of image semantic segmentation model, and CUDA8.0 was used to accelerate matrix operation in offline training process of the model.

In the process of model training, all RGB images, depth images and semantic label images in the training data set were scaled into $480 * 360$, and the original training sample size was expanded by 3 times through data augmentation by means of horizontal flip and color transformation. The weighted cross entropy was used as the loss function of the model, and the stochastic gradient descent (SGD) optimization algorithm was used to adjust the network parameters with a batch size of 4 samples. At the same time, the learning rate of network parameters is adjusted using Poly strategy, and the iterative formula of learning rate is as follows:

$$lr_{curr} = lr_{init} \times \left(1 - \frac{iter}{\max\_iter}\right)^{power} \tag{6}$$

where, the current learning rate lr_{curr} is equal to the product of the initial learning rate lr_{init} and the Poly factor, *iter* is the current iteration number, **max**_*iter* is the maximum iteration number of the model, and *power* is set as 0.9. Finally, all the training samples were fed into the network and trained for about 20 epochs. The offline training completes when the network loss was no longer drops. The model with high training accuracy and stable loss change is selected as the final prediction model.

3.1.1 SUN RGB-D Data Set

In order to verify the effectiveness of the semantic segmentation network of indoor RGB-D images constructed by combining multi-scale features, this paper selected 5050 test samples from SUNRGB-D standard semantic segmentation data set for experiments. The URL address of the SUN RGB-D DataSet is "http://cvgl.stanford.edu/data2/sun_rgbd.tgz". The experimental results are shown in Fig. 5. At the same time, an image semantic segmentation experiment without a deep information fusion branch is provided to analyze and verify the importance of image geometric information in indoor environment to the semantic segmentation network. As shown in Fig. 5, the image semantic segmentation network proposed in this paper can successfully fuse RGB information and depth information of images, and achieve good segmentation results in messy indoor scenes. Especially for objects with large sample size, such as tables, chairs and beds, the segmentation accuracy is higher. At the same time, from the results of semantic segmentation with RGB-D images and RGB images as input, it can be seen intuitively that the depth information of images can really play a certain auxiliary role in image segmentation.

Figure 5: (Continued)

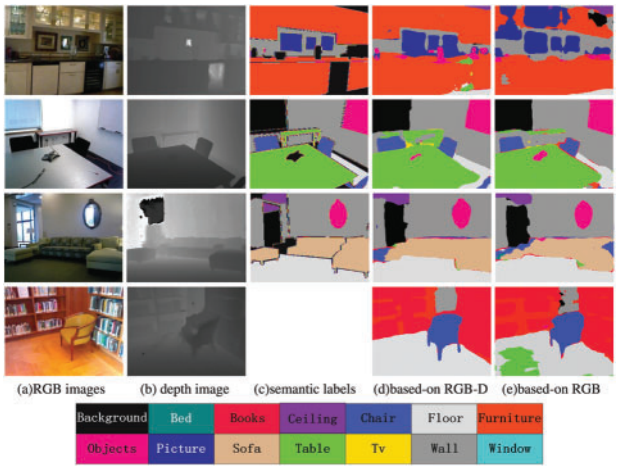


Figure 5: Semantic segmentation results of the SUN RGB-D test set

In order to evaluate the performance of the proposed algorithm quantitatively, we evaluate the performance of the constructed semantic segmentation network on the SUN RGB-D indoor image standard test set by using two semantic segmentation performance indicators, namely Mean Pixel Curacy (MPA) and Mean Intersection Over Union (MIOU). The results are shown in Table 2. After adding the depth information branch to the same semantic segmentation network model, the segmentation accuracy of the model is improved by 3.7% and 3.01% in MPA and MIOU, respectively, which indicates that the depth geometric information of the image plays an important role in assisting the image semantic segmentation task based on convolutional neural network.

Table 2: Comparison of model performance under two different inputs on SUN RGB-D data set

	MPA (%)	MIOU (%)
Based on RGB images	53.97	32.41
Based on RGB-D image	**57.67**	**35.42**
Performance gain	3.7	3.01

3.1.2 NYUDv2 Data Set

In this section, the image semantic segmentation performance of the proposed algorithm is compared with the four popular segmentation algorithms on NYUDv2 standard semantic segmentation data set. The URL address of the NYUDv2 DataSet is "https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html". As the SUN RGB-D indoor image training set contains a part of NYUDv2 data samples, the network model trained on SUNRGB-D data set is applied directly to NYUDv2 data set for evaluating the performance of the proposed algorithm. In NYUDv2 test samples, several scenes were selected for semantic segmentation experiments under different environments and different illumination conditions. Like the previous experiment, the image semantic segmentation experiment was also carried out under the condition of removing the branch of depth information fusion, and the importance of image geometric information in indoor environment to the semantic segmentation network was analyzed and verified. The experimental results are shown in Fig. 6. It can be seen from Fig. 6 that the proposed semantic segmentation model of indoor RGB-D

image trained by SUNRGB-D indoor image data set can also show good segmentation performance in NYUDv2 test samples. However, in this experiment, the model trained on SUNRGB-D indoor image data set, which contains some NYUDv2 data samples, is directly used for qualitative analysis, so the semantic segmentation visual effect of this model on NYUDv2 test samples is slightly inferior to that on SUNRGB-D data set, but it still performs well in the segmentation task of object categories with regular boundary contour and sufficient sample size, such as tables, chairs, beds, and pictures. In addition, from the semantic segmentation results with RGB-D and RGB images as input, it can be seen intuitively that the depth information of images can make up for the features effectively.

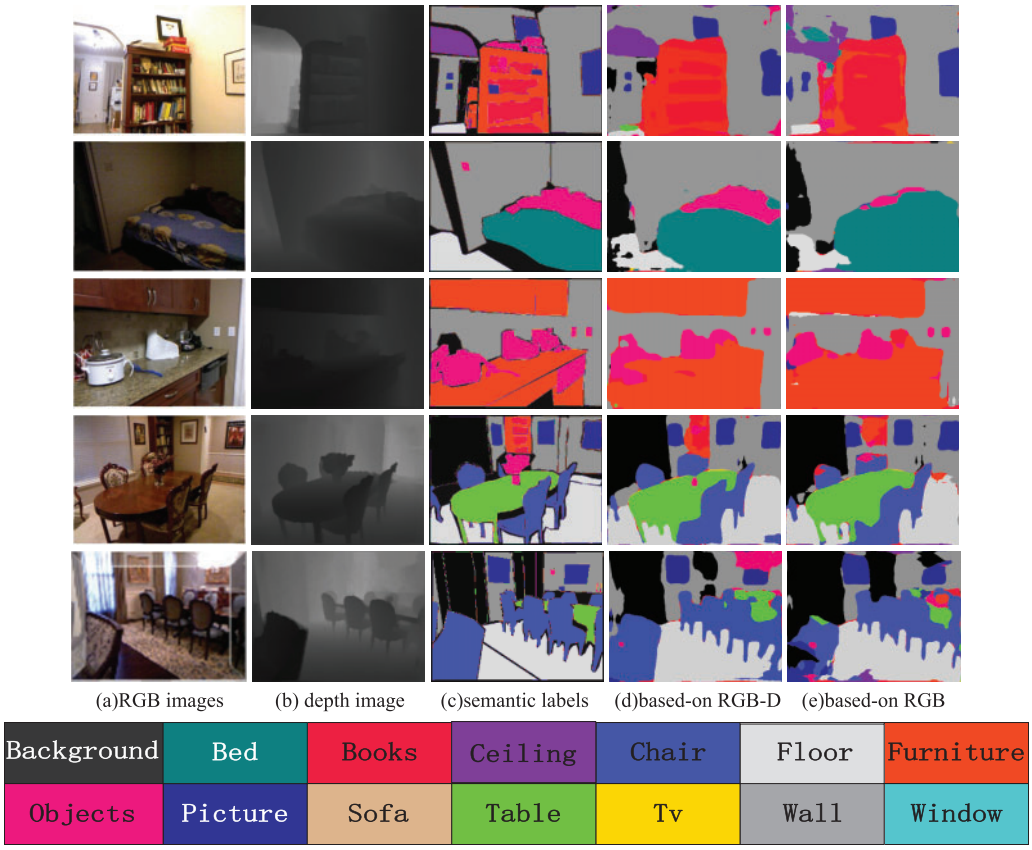


Figure 6: Semantic segmentation results of the NYUDv2 test set

The segmentation accuracy of the model under two types of input images (RGB image and RGB-D image) is shown in Table 3. It can be seen from Table 3 that the segmentation accuracy of the semantic segmentation network model is improved by 4.7% in MPA and 4.0% in MIOU after adding the depth information branch. From the image semantic segmentation results of the above two different image semantic segmentation data sets, it can be seen that incorporating image depth information can significantly improve the performance of the image semantic segmentation network.

Table 3: Comparison of model performance under two different inputs on NYUDv2 data set

	MPA (%)	MIOU (%)
Based on RGB images	54.8	35.4
Based on RGB-D image	**59.5**	**39.4**
Performance gain	+4.7	+4.0

As most of researches on semantic map construction are semantic annotation for 13 objects in NYUDv2 data set, in order to facilitate the comparative analysis, the proposed semantic segmentation network based on RGB-D image is compared with four other popular image semantic segmentation networks in the average accuracy of semantic segmentation categories for 13 class objects in this data set, and the quantitative estimation results are shown in Table 4.

Table 4: Average accuracy of 13 class objects on NYUDv2 data set

	RGBD [24]	RGBD-SF [24]	RGBD-SF-CRF [24]	RGBD [32]	Our RGB-D
Bed	62.5	61.7	62.0	**68.4**	62.2
Books	**60.5**	58.5	58.4	45.4	37.5
Ceiling	35.0	43.4	43.3	**83.4**	41.5
Chair	51.7	58.4	59.5	41.9	**72.9**
Floor	92.1	92.6	**92.7**	91.5	87.8
Furniture	54.5	63.7	**64.4**	37.1	59.0
Objects	**61.3**	59.1	58.3	8.6	49.5
Picture	**72.1**	66.4	65.8	35.8	59.0
Sofa	34.7	47.3	48.7	28.5	**53.8**
Table	26.1	34.0	34.3	27.7	**60.1**
TV	32.4	33.9	34.3	**38.4**	36.2
Wall	86.5	86.0	86.3	71.8	**87.7**
Window	53.5	60.5	62.3	46.1	**66.2**
MIOU	55.6	58.9	59.2	48.0	**59.5**

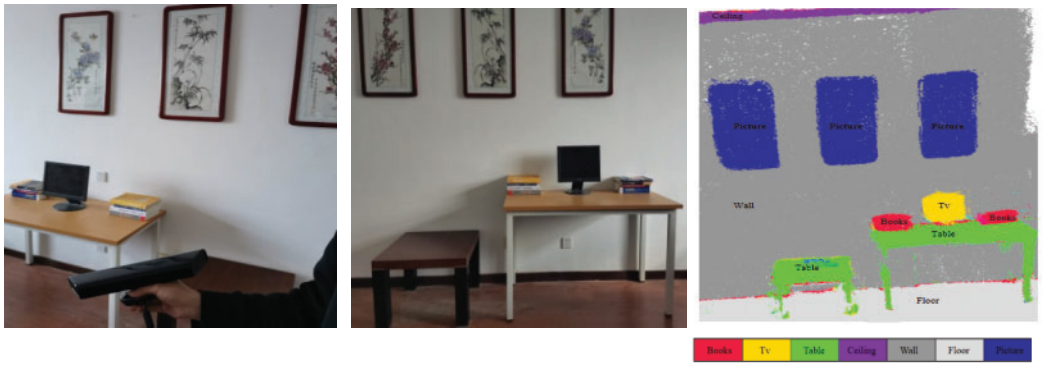
It can be seen from the quantitative estimation results shown in Table 4, the average accuracy of our RGB-D algorithm in 8 classes including ceiling, chair, and furniture is higher than that of the RGB-D method proposed in [24]. The average accuracy of the 7 types of objects, such as bed, chair, and sofa, was higher than that of the semantic fusion method RGBD-SF and the CRF optimized method RGBD-SF-CRF in [24]. Compared with the method proposed by Hermans et al. [32], our method achieves a better average accuracy of categories for 8 kinds of objects such as chairs, furniture, pictures, etc., only the segmentation accuracy of objects such as beds and books are slightly lower. Moreover, the average class accuracy of our RGB-D image segmentation algorithm is higher than that of the semantic segmentation algorithm proposed in [24] and Hermans et al. [32], validating the effectiveness and superiority of the proposed indoor RGB-D image semantic segmentation network.

3.2 Online Semantic Map Construction in Real Scene

In order to verify the effectiveness of the proposed semantic map construction method in real scenes, 3 online reconstruction experiments were carried out by using hand-held cameras and mobile robots equipped with cameras. The first group of 3D semantic map construction experiments was carried out by hand-held Kinect camera, and the second to third groups of 3D semantic map construction experiments were carried out by TurtleBot2 mobile robot equipped with a Kinect camera. The experimental results and analysis are as follows:

3.2.1 Online 3D Semantic Map Construction with Handheld Camera

In the experiment, the Online 3D semantic map construction of the laboratory scene was carried out with the handheld camera, and the experimental results are shown in Fig. 7. Figs. 7a and 7b are the original scenes RGB images of the laboratory from the third perspective and the first perspective, respectively. The objects in the scenes include a display, a desk, a book, a wall, a floor, a picture and a ceiling, and Fig. 7c is a constructed 3D semantic map of the laboratory environment with a hand-held camera. It can be seen from Fig. 7 that the online 3D semantic map construction algorithm proposed in this paper can construct a globally consistent 3D semantic map with accurate segmentation in the laboratory environment and can accurately label the semantics of various objects contained in the laboratory environment.



(a) The third perspective (b) The first perspective (c) Reconstructed 3D semantic map

Figure 7: Reconstruction result of 3D semantic map with handheld Kinect camera

3.2.2 Online 3D Semantic Map Construction with Mobile Robot Equipped with Kinect Camera

In the experiment, TurtleBot2 mobile robot equipped with Kinect camera (as shown in Fig. 8) was used to construct 3D semantic map in two laboratory scenes as shown in Figs. 9a and 9b. Among them, the objects included in the Lab scene 1 are TV, tables, books, walls, floors and pictures; Lab scene 2 includes TV, tables, books, chairs, walls, floors and objects (cartons).

Figure 8: Mobile robot equipped with Kinect camera

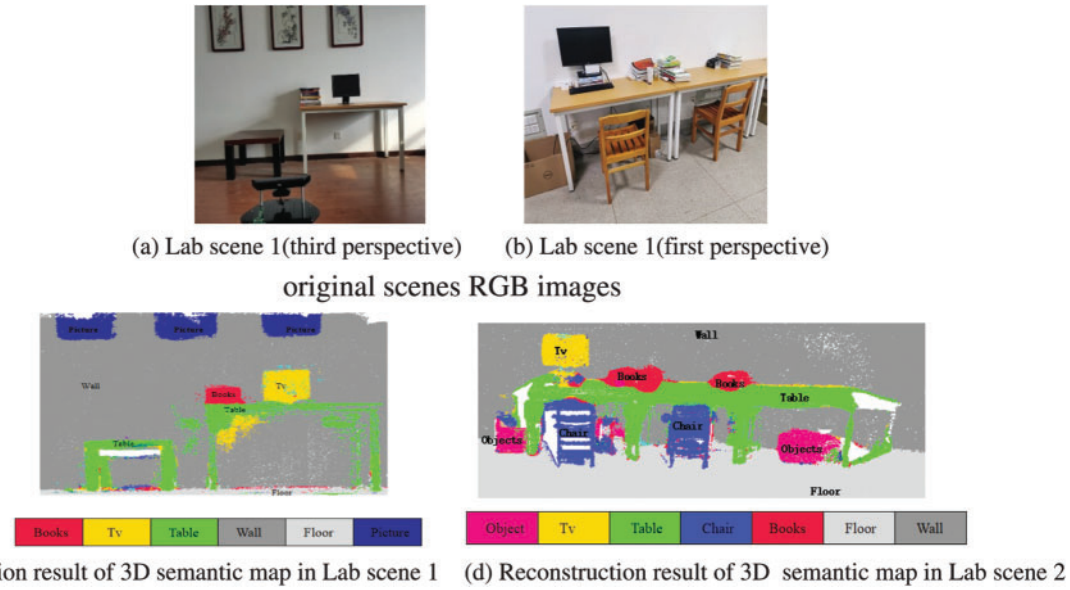
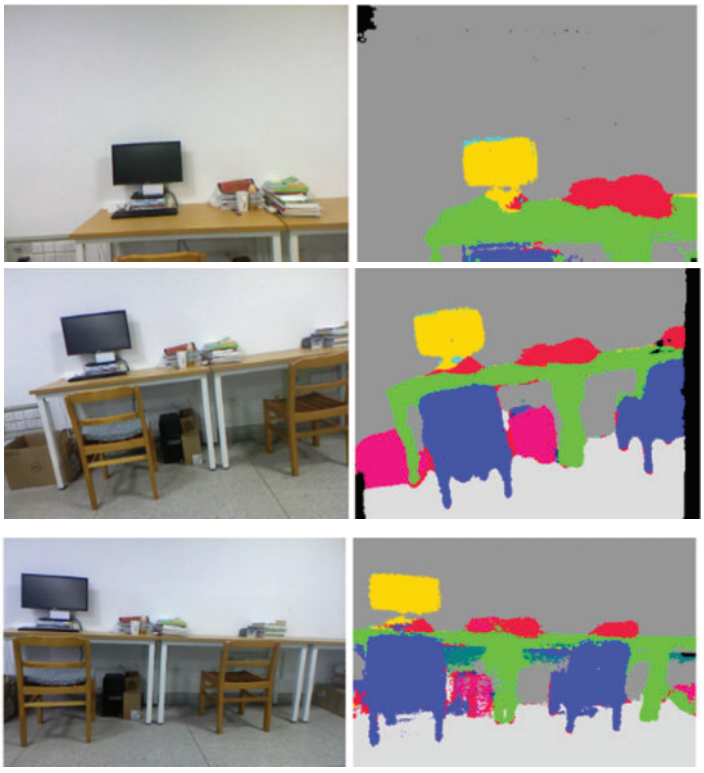


Figure 9: Reconstruction result of 3D semantic map with mobile robot equipped with Kinect camera

Figs. 9c and 9d respectively show the reconstructed 3D semantic maps of the above two laboratory scenes. It can be seen from the results in Fig. 9 that the semantic map construction algorithm proposed in this paper can construct accurate 3D semantic maps with mobile robot equipped with Kinect camera. In the process of constructing 3D semantic map of Lab scene 2 with TurtleBot2 mobile robot, the semantic segmentation results of some key frame images are shown in Fig. 10. In Fig. 10, the first

column is the key frame RGB image, and the second column is the semantic segmentation result of the key frame image. It can be seen form Fig. 10 that the image semantic segmentation network proposed in this paper can also achieve good segmentation results in real scenes when the original scene image captured by the mobile robot equipped with Kinect camera.



(a) keyframe images (b) semanticlabels

Figure 10: Semantic segmentation results of some key frame images with mobile robot equipped with Kinect camera

From the results of Figs. 9 and 10, it can be seen that the semantic map construction algorithm proposed in this paper can process the image sequence collected by the RGB-D sensors carried by the mobile robot in real time, and output the accurate semantic segmentation results of the single frame indoor scene image and the local spatial semantic map constructed at present, and finally construct a globally consistent high-precision indoor environment semantic map.

4 Conclusion

This paper proposes a novel indoor environment RGB-D image semantic segmentation network with multi-scale features fusion, which effectively utilizes the visual color features and depth geometric features of the image for more accurate indoor scene image semantic segmentation. Besides, an improved high-precision 3D semantic mapping of indoor scenes from RGB-D images for indoor scenes algorithm is proposed, combining the proposed image semantic segmentation algorithm with traditional SLAM technology. The effectiveness of the proposed approach is validated through experimental results of image semantic segmentation on public image semantic segmentation data sets and online construction of 3D semantic map in a real scene.

For future work, we will combine the proposed semantic map construction algorithm with a robot autonomous exploration algorithm to achieve autonomous semantic map construction. This will enable mobile robots to autonomously explore target scenes and construct three-dimensional semantic map, thereby improving the autonomy of the robots in the mapping process.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant U20A20225, 61833013, in part by Shaanxi Provincial Key Research and Development Program under Grant 2022-GY111.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Huang, W., Zhang, G., Han, X. (2020). Dense mapping from an accurate tracking SLAM. *IEEE/CAA Journal of Automatica Sinica, 7(6),* 1565–1574. https://doi.org/10.1109/JAS.2020.1003357
2. Niu, M. Y., Huang, Y. Q. (2022). An RGB-D SLAM algorithm based on dynamic coupling and spatial data association. *Robot, 44(3),* 333–342. https://doi.org/10.13973/j.cnki.robot.210151
3. Qi, S. H., Xu, H. G., Wan, Y. W., Fu, H. (2020). Construction of semantic mapping in dynamic environments. *Computer Science, 47(9),* 198–203. https://doi.org/10.11896/jsjkx.191000040
4. Fang, L. J., Liu, B., Wan, Y. C. (2020). Semantic SLAM based on deep learning in dynamic scenes. *Journal of Huazhong University of Science and Technology (Natural Science Edition), 48(1),* 121–126. https://doi.org/10.13245/j.hust.200122
5. Wu, H., Chi, J. X., Tian, G. H. (2019). Instance recognition and semantic mapping based on visual SLAM. *Journal of Huazhong University of Science and Technology (Natural Science Edition), 47(9),* 48–54. https://doi.org/10.13245/j.hust.190909
6. Qi, X. X. (2018). *Deep learning based semantic map construction in visual SLAM*, vol. 2. National University of Defense Technology, Changsha, China. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202101&filename=1020386688.nh
7. Huang, P., Deng, Q., Liang, C. (2020). Overview of image segmentation methods. *Journal of Wuhan University (Natural Science Edition), 66(6),* 519–531. https://doi.org/10.14188/j.167-8836.2019.0002
8. He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9),* 1904–1916. https://doi.org/10.1007/978-3-319-10578-9_23
9. Chang, S. Y. (2018). *Research on semantic mapping based on visual SLAM,* vol. 2018. Harbin Institute of Technology, China. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1018893761.nh
10. Hu, M. Y., Zhang, Y. Z., Qin, C., Liu, T. B. (2019). Semantic map construction based on deep convolutional neural network. *Robot, 41(4),* 452–463. https://doi.org/10.13973/j.cnki.robot.180406
11. Cheng, J., Wang, C., Mai, X., Min, Z., Meng, M. Q. H. (2020). Improving dense mapping for mobile robots in dynamic environments based on semantic information. *IEEE Sensors Journal, 21(10),* 11740–11747. https://doi.org/10.1109/JSEN.2020.3023696
12. Zhang, C., Liu, Z., Liu, G., Huang, D. (2019). Large-scale 3D semantic mapping using monocular vision. *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pp. 71–76, https://doi.org/10.1109/ICIVC47709.2019.8981035
13. Cui, X., Lu, C., Wang, J. (2020). 3D semantic map construction using improved ORB-SLAM2 for mobile robot in edge computing environment. *IEEE Access, 8,* 67179–67191.

14. Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J. et al. (2018). Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics, 14(9),* 4224–4231. https://doi.org/10.1109/TII.2018.2822828
15. He, W., Li, Z., Chen, C. L. (2017). A survey of human-centered intelligent robots: Issues and challenges. *IEEE/CAA Journal of Automatica Sinica, 4(4),* 602–609. https://doi.org/10.1109/JAS.2017.7510604
16. Sünderhauf, N., Pham, T. T., Latif, Y., Milford, M., Reid, I. D. (2017). Meaningful maps with object-oriented semantic mapping. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5079–5085. Vancouver, BC, Canada.
17. Mur-Artal, R., Tardós, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics, 33(5),* 1255–1262. https://doi.org.10.1109/TRO.2017.2705103
18. Jeong, J., Park, H., Kwak, N. (2017). Enhancement of SSD by concatenating feature maps for object detection. arXiv preprint arXiv:1705.09587.
19. Bai, Y. H. (2018). Research on semantic mapping based on slam algorithm and deep neural network. *Computer Applications and Software, 35(1),* 183–190. https://doi.org/10.3969/j.issn.1000-386x.2018.01.032
20. Dou, X. (2019). *Research on robot environment perception based on deep learning (M.S. Thesis)*. Harbin Engineering University, China. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201902&filename=1019141431.nh
21. Qi, S., Xu, H., Wan, Y., Fu, H. (2020). Construction of semantic mapping in dynamic environments. *The Journal of Computer Science, 47(9),* 198–203. https://doi.org/10.11896/jsjkx.191000040
22. Riazuelo, L., Tenorth, M., Marco, D. D., Salas, M., Gálvez-López, D. et al. (2015). RoboEarth semantic mapping: A cloud enabled knowledge-based approach. *IEEE Transactions on Automation Science and Engineering, 12(2),* 432–443.
23. McCormac, J., Handa, A., Davison, A., Leutenegger, S. (2017). SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4628–4635, https://doi.org/10.1109/ICRA.2017.7989538
24. Whelan, T., Salas-Moreno, R. F., Glocker, B., Davison, A. J., Leutenegger, S. (2016). ElasticFusion: Real-time dense SLAM and light source estimation. *The International Journal of Robotics Research, 35(14),* 1697–1716.
25. Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, vol. 2015, pp. 1520–1528. Santiago. https://doi.org/10.1109/ICCV.2015.178
26. Zhao, C., Sun, L., Purkait, P., Stolkin, R. (2018). Dense RGB-D semantic mapping with pixel-voxel neural network. *Sensors, 18(9),* 3099.
27. Cheng, J., Wang, C., Mai, X., Min, Z., Meng, M. Q. H. (2020). Improving dense mapping for mobile robots in dynamic environments based on semantic information. *IEEE Sensors Journal, 21(10),* 11740–11747.
28. Bao, Y., Pan, Y., Yang, Z., Huan, R. (2021). Utilization of semantic planes: Improved localization and dense semantic map for monocular SLAM in urban environment. *IEEE Robotics and Automation Letters, 6(3),* 6108–6115. https://doi.org/10.1109/LRA.2021.3091396
29. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. P., Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4),* 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
30. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. Honolulu, HI, USA.

31. Xie, S., Tu, Z. (2015). Holistically-nested edge detection. *International Journal of Computer Vision, 125,* 3–18.
32. Hermans, A., Floros, G., Leibe, B. (2014). Dense 3D semantic mapping of indoor scenes from RGB-D images. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2631–2638. Hong Kong, China.