Indoor 3D Semantic Robot VSLAM Based on Mask Regional Convolutional Neural Network

CHONGBEN TAO, ZHEN GAO, JINLI YAN, CHUNGUANG LI, AND GUOZENG CUI
1Suzhou Smart City Research Institute, Suzhou University of Science and Technology, Suzhou 215009, China
2Faculty of Engineering, McMaster University, Hamilton, ON L8S 0A, Canada
3School of Computer Information and Engineering, Changzhou Institute of Technology, Changzhou 213002, China

Corresponding author: Zhen Gao (gaozhen@mcmaster.ca)

This work was supported in part by the National Natural Science Foundation of China under Grant 61801323, Grant 61703059, and Grant 61750110519, in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 17KJB120011 and Grant 17KJD520001, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170291, in part by the Postdoctoral Science Foundation Fund of Suzhou under Grant SYG201708 and Grant SS2019029, in part by the Construction System Science and Technology Fund of Jiangsu Province under Grant 2017ZD066, in part by the Jiangsu Collaborative Innovation Center for Cultural Creativity under Grant XYN1703, and in part by the Natural Science Foundation of Changzhou Institute of Technology under Grant YN1603.

ABSTRACT During the construction of indoor environmental semantic maps by robot Vision SLAM (VSLAM), there exist some problems such as low label classification accuracy and low precision under the situation of sparse feature points. In this case, this paper proposes an indoor three-dimensional semantic VSLAM algorithm based on Mask Regional Convolutional Neural Network (RCNN). Firstly, an Oriented FAST and a Rotated BRIEF (ORB) algorithms are used to extract image feature points. Secondly, a Random Sample Consensus (RANSAC) algorithm is employed to eliminate mismatched points and estimate camera position-pose changes. Then, a Mask RCNN algorithm is applied to make partial adjustments to its hyper parameter. A self-made data set is used to transfer learning, fulfilling real-time target detection and instance segmentation of a scene. A three-dimensional semantic map is constructed in combination with VSLAM algorithm. The semantic information in the environment not only improves the accuracy of VSLAM construction and positioning, but also reduces the impact of object movement on the construction by marking movable objects. Meanwhile, the VSLAM algorithm is used to calculate the positional constraints between objects and improve the accuracy of semantic understanding. Finally, by comparing with other methods, it demonstrates that this method is more correct and effective. It was also verified that the proposed method can accurately interpret the semantic information in environment for the construction of three-dimensional semantic maps.

INDEX TERMS VSLAM, deep learning, target detection, instance segmentation, semantic map.

I. INTRODUCTION

The process of utilizing the established map to synchronize and update its position by robots in unknown location and environmental map is called the Simultaneous Localization and Mapping (SLAM) problem. Visual Simultaneous Localization and Mapping (VSLAM) is a SLAM problem that uses a camera as the only sensor. It can be seen that after decades of development, VSLAM technology tends to be mature with acceptable robustness and usability. However, for the mainstream VSLAM algorithms, there are few methods to utilize spatial semantic information. The surrounding environment is only a dense or sparse spatial point, which cannot provide semantic information and cannot meet higher task requirements of the robot. Presently, significant progress in deep learning can provide solutions to this problem. However, these methods usually contain some shortcomings such as low label classification accuracy and low accuracy in feature point sparseness.

This paper plans to apply a target detection method to semantically label environmental information. From RCNN to Fast RCNN and Faster RCNN, the deep learning method represented by RCNN in target detection sharply improves the accuracy of target detection [1]. He et al. added the Fully Convolutional Networks (FCN) to generate the corresponding MASK branch based on the original Faster
RCNN algorithm. Mask RCNN method can be used to accomplish various tasks such as target classification, target detection, and semantic segmentation [2].

Considering the problems of low label classification accuracy and low precision under the situation of sparse feature points included in the current semantic VSLAM and the fact that the image sequence processed by the VSLAM algorithm often contains various types of objects, richer and more accurate information about itself and the environment can be achieved by combing the semantic recognition advantages of deep learning and the coordinate information provided by SLAM algorithm. This paper proposes a semantic VSLAM algorithm by combining ORB feature extraction algorithm and Mask RCNN. The algorithm uses the robustness and real-time ORB feature extraction for front-end processing, which greatly improves the speed and accuracy of key points acquisition. Meanwhile, based on the utilization of the Mask RCNN method for effective target detection, instance segmentation tasks and semantic association of map points, it ultimately achieves the construction of high-precision semantic map.

The main contributions of this work can be summarized as follows:

(1) Established a new data set for migration learning, 21 types of objects which are common in family and laboratory scenes are selected and data augmentation such as random stretching and shearing are conducted on the data so as to make 10,000 images;

(2) Fulfill real-time target detection and instance segmentation. This article implies ORB feature extraction with better robustness and instantaneity for the front-end processing, greatly improving the speed and accuracy when obtaining key points. Meanwhile this paper uses the Mask RCNN method, which is effective in target detection and segmentation task, to carry out semantic association between map points and finally build semantic maps with high accuracy;

(3) Reduce the influence of moving objects on construction by marking movable objects. Firstly, the unusable points in the map are marked to reduce the influence of moving objects on drawing construction. Secondly, the relationship between map points and semantic information is established. The relation of semantic information is formed by locating the coordinates of map points corresponding to the feature points in the target object frame;

(4) Achieve an accurate three-dimensional environmental semantic map. In this algorithm, the geometric information provided by SLAM algorithm and the semantic information provided by deep neural network are effectively combined to establish the mapping relationship between geometric space and semantic map. This algorithm improves the precision of slam mapping and positioning through semantic information, and enhances the accuracy of semantic understanding through the object location information obtained by SLAM.

II. RELATED WORK

Nowadays, intelligent robots are not only widely used in the industrial field, but also have great application prospects in various environments coexisting with humans [3]. With the rapid development of computer vision, VSLAM increasingly attracts attention due to its wide application scenarios, low costs, capacity of extracting semantic information and other advantages.

A. VSLAM

Mur-Artal et al. modified the calculation method of the rotation angle of brief descriptor to better stand the influence of light intensity variation based on the monocular VSLAM system, and ORB feature is used to extract rows simultaneously in different threads to improve the usability and accuracy of VSLAM [4]. Navarrete et al. proposed a GMMS-based 3D VSLAM point cloud compression and registration method [5]. Firstly, a planar-based 3D model was employed to select and group points. Then, a Gaussian mixture model was utilized for fast variants and expectation maximization algorithms with a set of Gaussian distribution replacing the points grouped in the previous step. Finally, a 3D point cloud map was obtained by decompressing the model. Fan et al. studied the VSLAM dynamic object elimination based on image fusion algorithm [6]. The camera motion was decomposed into two states, translation and rotation, and two constraints were proposed to locate the dynamic region. The dynamic regions in the image sequence were emptied while one of the image planes was selected as the projection plane. The information of other images was mapped to the plane to obtain a fused image with the result containing all the environmental information. Kang et al. proposed DF-SLAM, which used the deep local feature descriptor obtained by neural network to replace traditional manual features, and improved the availability and stability in VSLAM system [7].

B. SEMANTIC VSLAM BASED ON DEEP LEARNING

Recently, significant progress in deep learning can provide solutions to utilize spatial semantic information [8], [9]. Whelan et al. proposed a construction method for dense three-dimensional semantic map called Semantic Fusion based on convolutional neural network, which relied on Elastic Fusion SLAM algorithm to provide indoor RGB-D video inter-frame for position-pose estimation, utilizing convolutional neural network to predict pixel-level object category label which ultimately combined the Bayesian upgrade strategy with the conditional random field model. It realized the probability escalation of CNN prediction values from different perspectives, and finally generated a dense three-dimensional semantic map containing semantic information [10]. Engel et al. [11] proposed a monocular and semi-dense three-dimensional semantic mapping construction method based on CNN and large scale direct (LSD) SLAM, which used LSD-SLAM method instead of Elastic Fusion method [12] to estimate camera pose. Firstly, a key frame was selected to estimate the camera position-pose and
improve processing speed. Secondly, a CNN architecture was applied in DeepLab-v2 for pixel-level classification [13]. Finally, a Bayesian network was used for upgradation of pixel classification probability, map regularization combined with conditional random field (CRF), and noise smoothing on the generated semantic segmentation map. Mur-Artal and Tardós et al. [14] utilized a four-wire Stereo-ORB-SLAM [15] system to construct a dense 3D point cloud map. They also applied RGB-D image as the input complete convolutional neural network architecture to obtain pixel segmentation and to merge geometric information and semantics information for a semantic map. Liu et al. [16] studied the depth prediction map generated by deep neural network, and proposed a depth metric obtained by monocular VSLAM, which proved that the application of depth prediction in the absolute scale of estimated reconstruction could overcome defects of monocular VSLAM. Finally, a method for efficiently synthesizing semantic tags obtained from single frames in dense SLAM was put forward. The reconstruction of semantically coherent scene was obtained from a single perspective. Huang and Liu [17] proposed a novel multi-exposure image enhancement method, which can generate a multi-exposure image sequence. Rublee et al. [18] proposed a multi-feature extraction algorithm to extract two kinds of image features simultaneously, which can work effectively if the single feature algorithm fails to extract enough feature points in SLAM.

III. SYSTEM OVERVIEW

The proposed algorithm combines an ORB feature extraction algorithm and a Mask RCNN algorithm. The ORB feature extraction is used for front-end processing, which greatly improves the speed and accuracy of key points acquisition. Meanwhile, a Mask RCNN method is utilized for effective target detection, instance segmentation tasks and semantic association of map points to eventually construct a high-precision semantic map. The overall architecture of the proposed algorithm is shown in Figure. 1.

A. SLAM ALGORITHM BASED ON ORB FEATURE EXTRACTION

The ORB feature extraction algorithm as a matching method proposed by Rublee et al. is significantly faster than traditional SIFT and SURF operations [19]. The ORB algorithm is divided into two parts, feature point extraction and feature point description, which were evolved from Features from Accelerated Segment Test (FAST) algorithm and Binary Robust Independent Elementary Features (BRIEF) algorithm, respectively [20], [21].

FAST Keypoint Orientation (oFAST) is an algorithm which uses FAST to extract feature points and then defines a feature point direction to improve the rotation of the feature points without deformation. The process of oFAST algorithm is as follows:

Step 1: Extract coarse-grained feature point
Based on the coarse-grained extraction of the gray value of the image around the feature point and the pixel values around the candidate feature point, it has

\[ N = \sum_{x} \sum_{y} |I(x) - I(y)| > \varepsilon_d \]  

where \( I(x) \) is the gray level at any point on the circumference, \( I(y) \) is the gray level of the center, and \( \varepsilon_d \) is the threshold value of the gray value difference. If \( N \) is greater than the given threshold, \( p \) is considered to be a feature point. In order to speed up the extraction, as shown in Figure. 2, the gray values at the positions of 1, 9, 5, and 13 should be detected at the first. If the pixel value of 3 or more positions out of the four is larger or smaller than the gray value of the \( P \) point, then \( P \) is the feature point. Otherwise, \( P \) is not.

Step 2: Screen optimal feature point
The ID3 algorithm is used to train a decision tree, and 16 pixels on the circumference of the feature point are input into the decision tree to filter out the optimal feature points.

Step 3: Cull local dense feature point
Non-maximum suppression method is applied to remove locally dense feature points.

Step 4: Solve the feature point direction
The moment method is used to determine the direction of the FAST feature points. The moments are defined as follows:

\[ M_{ij} = \sum_{x} \sum_{y} x^i y^j I(x, y) \]
where \( I(x, y) \) is the grayscale expression of the image, and the angular direction is

\[
C_{ori} = \tan^{-1}\left(\frac{C_y}{C_x}\right)
\]

(3)

where \( c_x = \frac{M_{12}}{M_{22}} \), \( c_y = \frac{M_{02}}{M_{22}} \).

The direction angle of feature points is included in the feature points previously calculated by the oFast algorithm. Suppose that the original BRIEF algorithm selects \( n \) pairs of points in the feature point \( S \times S \) neighborhood, it gets

\[
D = \begin{pmatrix} x_1, x_2, \ldots, x_{2n} \\ y_1, y_2, \ldots, y_{2n} \end{pmatrix}
\]

(4)

The original BRIEF descriptor does not have rotation invariance. By applying the centroid direction \( \theta \) of the feature point to descriptor, the problem of rotation invariance can be solved to some extent. According to the rotation information \( \theta \) of the image block and the corresponding rotation matrix \( R_\theta \), a new matrix can be obtained as follows,

\[
D_0 = R_\theta D
\]

(5)

By this method, a rBRIEF descriptor with rotation invariance can be obtained as follows

\[
g_n(p, \theta) = f_n(p) \mid (x_i, y_i) \in R_\theta
\]

(6)

Since the rough extraction method is used in the ORB feature extraction, a large number of mismatch points are inevitable. Therefore, this paper proposes a feature point screening method based on RANSAC algorithm. The improved RANSAC algorithm is utilized to find the optimal parameter matrix so that the number of data points satisfying the matrix is maximized. The detailed steps are as follows:

Step 1: Set up the inner group

A subset of 4 samples is extracted as an inner group from the sample set.

Step 2: Establish an inner group model

This step is to solve transformation matrix \( M \) according to the sample subset, and establish the inner group model. The formula is as follows

\[
t \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}
\]

(7)

where \((u, v)\) represents the corner position of the target image, \((u', v')\) is the corner position of the scene image, and \( t \) is the scale parameter.

Step 3: Determine the threshold

This step is to calculate the error of the residual point and \( M \) projection in the sample set to determine whether it is an inner group. The formula is as follows

\[
\sum_{i=1}^{n} \left( u_i \left( h_{11}u_i + h_{12}v_i + h_{13} \right) \right)^2 + \left( v_i \left( h_{31}u_i + h_{32}v_i + h_{33} \right) \right)^2
\]

(8)

Step 4: Perform iterative calculation

This step is about judging whether it is an optimal set according to the current number of inner groups. Besides, an iterative calculation is performed.

After ORB feature extraction is performed on two frames of images, the corresponding relationship between adjacent frames is obtained through feature matching algorithm, and feature points are matched and filtered through RANSAC algorithm. Figure 3 shows the results after screening. It can be seen from this figure that this method can effectively filter the feature matching points to improve accuracy of feature matching.

In order to extract geometric information from twodimensional images, it is indispensable to calibrate the depth camera in 3D machine vision to obtain camera internal parameters. A classic Zhang calibration method [22] is applied in this paper to calibrate the Kinect camera, as shown in Figure 4, which requires the usage of depth cameras, observing calibration plate from different angles and collecting images for calibration. After correct calibration of the depth camera, the coordinate information of any point in the spatial 3D point cloud can be obtained.

In order to solve the problem of low resolution in building 3D feature map with ORB-SLAM algorithm, a key frame extraction method and a sub map method in spatial domain are proposed to construct 3D feature map. The idea is that adjacent image frames are continuous during the movement of a robot, namely position and pose of adjacent frames change little in space. Therefore, only when Kinect camera moves a certain size in an adjacent frame, then the frame can be regarded as a key frame, which point cloud is superimposed on the existing map. Meanwhile, the whole map is divided into several sub maps with certain key frames.
When the map scale reaches a certain degree, namely a certain number of key frames, these sub maps are released from the calculation memory and saved in a computer hard disk. When the whole mapping process is completed, multiple sub maps are integrated into a complete map to improve the speed of the algorithm. For example, four sub maps are created by key frames, as shown in Figure 5.

Kinect is used as the visual sensor, which is composed of color camera and depth sensor. The color camera can obtain the RGB value of the pixel points, and the depth sensor can measure the distance from the pixel points, which can be fused together to calculate the 3D point cloud. First, the key frame is extracted, and its point cloud is spliced into the existing map. Then, the large-scale map is decomposed into small map of a certain scale by using the sub map method, so as to reduce the consumption of memory and improve the overall speed of the algorithm.

B. ALGORITHM OF TARGET DETECTION AND INSTANCE SEGMENTATION BASED ON MASK RCNN

The task of target detection is applied to find out all the targets/objects which are interested in the image, and to determine their positions and sizes which is one of the core issues in the field of machine vision. The traditional target detection process often utilizes a sliding window for region selection, and then applies SIFT, HOG and other methods for feature extraction. Finally, SVM and Adaboost are adopted for category judgment. However, the traditional method has poor robustness, which is insufficient to reflect the illumination changes, backgrounds diversity etc. Besides, its regional selection is not targeted, the time complexity is high, and the window is redundant. Furthermore, in order to detect the target, entity segmentation also requires marking the category of each pixel.

During the construction of semantic map, the target detection algorithm is required to provide the original semantic information. The real-time performance of VSLAM algorithm has demanding requirements on the running speed of the algorithm. In this paper, a deep neural network is designed based on the Mask RCNN. Object categories at output layer are set as 21 types of indoor objects. More details are described in the experimental part. The proposed Mask RCNN network architecture is shown in Figure 6.

The proposed Mask RCNN process can be divided into two phases. Firstly, an offer regarding the area is generated where the object may be present based on the input image. Secondly, the category of object is predicted. The bounding box is refined, and a mask is generated at the pixel level of the object according to the proposal in the first phase. The Mask RCNN backbone network is a standard convolutional neural network. The bottom layer detects low-level features (edges and corners, etc.), and the advanced layers detect higher-level features. Through the forward propagation of the backbone network, the image is converted from a tensor of $1024 \times 1024 \times 3$ into a feature map of $32 \times 32 \times 2048$. This feature map will be used as input for the next phase. The Feature Pyramid Network (FPN) is an extension of the backbone network which can better characterize targets on multiple scales [23]. FPN improves the performance of the standard feature extraction pyramid by adding a second pyramid, which selects advanced features from the first one and pass them down to the bottom layer. The network is shown in Figure 7. Through this process, FPN allows features of each level combination of both high-level and low-level features.

In the first phase, the sliding window is implemented by the convolution process of the Region Proposal Network (RPN), which does not scan the image directly, but scans the backbone feature map. Extracted features can be efficiently reused in RPN and double counting can be avoided. Through
predictions of RPN, the anchor containing the target can be selected to finely adjust its position and size. In the second stage, ROI Align replaces ROI Pooling to locate the relevant regions and sample at different points of the feature map.

Bilinear interpolation is applied to improve the accuracy of the detection model. Finally, object classes are generated by bounding boxes and masks after completing pixel-level alignment.

The loss function of the network is composed of three parts, namely classification error, detection error and segmentation error. The expression is as follows

$$L = L_{cls} + L_{box} + L_{mask}$$

where \(L_{cls}\) and \(L_{box}\) are the same as in the Faster RCNN, and the mask branch has a \(Kn^2\) dimensional output for each ROI, representing \(K\) binary masks of resolution \(m \times n\). \(K\) is the number of categories. One sigmoid is implemented for each pixel. \(L_{mask}\) is defined as the average binary cross entropy loss. If the region is related to the ground truth value category \(k\), then only the \(k^{th}\) mask is included.

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \leq i, j \leq m} \left[ y_{ij} \log \hat{y}_{ij} + (1 - y_{ij}) \log \left(1 - \hat{y}_{ij}\right)\right]$$

where \(y_{ij}\) is the label of cell \((i, j)\) in the real mask with a size of \(m \times n\), and \(\hat{y}_{ij}\) is the predicted value of the same unit in the ground truth class \(k\) learning mask.

Table 1 compares the results of the proposed Mask RCNN algorithm with other instance segmentation algorithms. It can be found that the advantages of the proposed method are obvious. In this table, AP stands for Average Precision, which is an indicator of the accuracy of the target detection algorithm.

### IV. CONSTRUCTION OF SEMANTIC MAP

In the VSLAM algorithm, the image information directly corresponds to the map point. The image feature in the target frame detected by the target detection algorithm in this paper can correspond to the stored map point and the target object, thereby a mapping relationship between the image data and the target object is established. The algorithm flow of the semantic map construction is shown in Figure 8.

First of all, the VSLAM algorithm for front-end tracking module decides whether to use the current image frame as a key frame. The selection rules of key frame are as follows:

1. Being at least 20 frames away from the last relocation;
2. Local map thread is idle, or 20 frames have passed since the last time which a key frame was added;
3. The current frame tracks at least 50 points;
4. The number of map points that the current frame tracks to the reference frame in local map is less than 90%.

It also should ensure that there are significant visual changes between key frames.

Secondly, after the current frame is selected as the key frame, the key frame will be processed by two-part algorithm: the first part is the target detection algorithm, the second is the VSLAM back-end graph optimization algorithm. The system will deal with the obtained semantic information in two aspects. The unavailable points in the map is marked initially. For example, in order to reduce the impacts of moving objects on the construction. Feature points in the dynamic category target boxes detected as people, cat, and dog are marked as unavailable. Points in the range of threshold coordinates are discarded directly and are not used in the construction.

The connection between map points and semantic information is established. The association of semantic information is formed by finding the corresponding coordinates of map points to the feature points in the target object frame. From the jointly calibrated RGB image and depth image, pixels with coordinates \((u, v)\) and depth \(d\) can be obtained. The three-dimensional coordinates can be expressed as:

$$\begin{bmatrix} X, Y, Z \end{bmatrix}^T = \begin{bmatrix} \frac{(u - c_x)Z}{f_x}, & \frac{(v - c_y)Z}{f_y}, & \frac{d}{s} \end{bmatrix}^T$$

where \(f_x\) and \(f_y\) are the focal lengths of the camera in the \(x\) and \(y\) axes respectively; \(c_x\) and \(c_y\) are the translation quantity of the pixel coordinate system and the imaging plane respectively, which can be obtained by the internal parameter matrix of camera; \(s\) is the scaling factor of the depth map.
In the process of constructing the semantic map, the continuous frame information needs to be updated in real time. The real-time update algorithm is as follows. Note that the \textit{Fun} function is equivalent to taking the union and merging the output of the two frames:

\textbf{Algorithm 1} Procedures of Updating the Global Map

\begin{enumerate}
\item Initialize the global map: \(\text{Map} = y = x\);
\item Initialize the global map: \(\text{Map} = y = x\);
\item Get \(\text{Map}_{\text{local}}\) by projecting \(y\) into current perspective, preserve other targets beyond the scope of current perspective as \(z\);
\item Remove the boxes under the threshold by function \(\text{Map}_{\text{local}} = \text{Fun}({x, \text{Map}_{\text{local}}})\);
\item Put the updated local map into the global map by function \(y = \text{Fun}(z, \text{Map}_{\text{local}})\);
\end{enumerate}

The target detection algorithm and the 3D point cloud coordinate fusion algorithm are as follows:

Figure 9(a) shows key points extracted by depth camera, which is stored as a target point cloud as shown in Figure 9(b). The Mask-RCNN is used to detect targets in Figure 9(c). Then, according to the coordinate correspondence, image semantic information is given by assigning the corresponding semantic information to the target point cloud.

When the object is blocked, the blocked part would be ignored at first and then completed when the camera turned to the position where the previously blocked part became visible. Point clouds from two different perspectives were managed by surface list method, and then registered by ICP improved algorithm with curvature feature. Firstly, surface list method is applied to process point clouds from two different perspectives according to the two pictures of massive point cloud data. Secondly, calculate the approximate curvature of each point cloud according to the surface obtained, and then classify the curvature of one of the point cloud by size and each category is sampled randomly. According to the result of random sampling and the principle of curvature first, distance second, locate corresponding points in the other picture and finally find out the rotation and translation matrix of the first point cloud so as to complete the registration process.

The same object will not be considered as two objects of the same category from different viewing angles. When an object is identified as belonging to a specific category from a certain perspective, judgement will be made about whether it has been labeled as a category. If not, label the corresponding object on the semantic map as a category. Otherwise, no operation will be made.

The closed-loop detection algorithm based on Bag of Words is used to detect the closed-loop, and the large loop is obtained. At the same time, the matching relationship of the adjacent frames is used to detect the possible local loop between the adjacent frames in real time. By using the method of graph optimization to optimize these loops, we can get the accurate trajectory and point cloud map.

\textbf{V. EXPERIMENTAL EVALUATION}

\textbf{A. INTRODUCTION TO THE EXPERIMENTAL PLATFORM}

An independently developed mobile robot platform is used in experiments, as shown in Figure 10, which is equipped with a Kinect 2.0 depth camera and a laptop. The main body is consisted of a console, several brackets, two drive wheels and a chassis.

Robot Operating System (ROS) is used in the experimental platform as middleware, which is an open source meta-OS for robots that provides many functions similar to traditional operating systems [26]. Additionally, ROS supplies relative tools and libraries for acquiring, compiling, editing, and running among multiple computers to complete the distributed calculation.

The operational architecture of ROS is a processing architecture which uses communication modules to implement loosely coupled network connections among the modules. ROS performs several types of communication, including service-based synchronous Remote Procedure Call (RPC) communication, topic-based asynchronous data stream
communication, as well as data storage on the parameter server. As long as ROS is in a network with good communication quality and master node works normally, performance can be guaranteed. Figure 11 shows the working mode of ROS.

The experimental platform has some functions such as voice interaction, cloud control, and autonomous navigation, which is capable of performing tasks such as indoor inspection and service guidance.

Deep learning and training is conducted in the Ubuntu18.04. The processor model is Intel i5-7500 and the memory is 32 GB. For higher training and testing speed, a GTX 1080 graphic is used to accelerate training.

**B. EXPERIMENTAL VERIFICATION**

In order to prevent the interference of unrelated semantic information during map construction, this paper makes an adjustment to the Mask RCNN network structure. This experiment adopts a self-made data set to select 21 types of objects commonly found in laboratory and home scenes, and data augmentation is performed such as random stretching and shearing. Eventually, 10,000 images are produced, where 90% of these images serve as training sets and 10% as the verification set. The object types are shown in Table 2.

Since the target detection algorithm takes a long time without GPU acceleration, and it is still difficult to carry a high-performance GPU on the mobile robot, this research puts the target detection algorithm on the host side. The communication between the host side and the mobile robot is conducted through ROS. The mainstream TensorFlow is used as framework in target detection algorithm of server, which is equipped with GTX1080 graphics card to accelerate operation. After obtaining the target detection result of the key frame, the target point cloud is given the corresponding semantic information according to the coordinate correspondence. The quality of the construction can be seen from Figure.12. Figure.12(a) shows the real environment photo. The semantic information of dense point cloud is shown in Figure.12(b).

A migration learning method was used in experiments. Figure 13 shows a comparison of Loss-Iteration curves of three algorithms. The red line in the diagram represents the loss value of the proposed OMASK-RCNN method,
which is always lower than Fast-RCNN. Although the loss value of OMASK-RCNN is higher than Faster-RCNN between 6000 iterations and 17000 iterations, the curve of OMASK-RCNN gradually stabilizes below Faster-RCNN after 17000 iterations. It can be seen that the proposed OMASK-RCNN method is more accurate than Fast-RCNN and Faster-RCNN after 17000 iterations.

Figure 14 shows the Precision-Recall curve comparison diagram of three algorithms. The value of vertical coordinate represents precision, which means the detection accuracy of the detected target. The value of horizontal coordinate represents recall, which means the total number of correctly detected targets divided by the total number of targets. Obviously, when the area under the curve is larger, the performance of the algorithm is better, and then the detection is more accurate and complete. It can be seen from the diagram that the area under Precision-Recall curve of the proposed algorithm is obviously larger than other two methods. Besides, the simulation results show the proposed algorithm has better performance than the other two methods. All the analyses show that the proposed OMASK-RCNN method is more accurate and has better performance than the other two methods.

In order to show the real environment and the effect of semantic recognition, a certain key frame in the process of construction is processed and four pictures of Figure 15 are obtained. Figure 15(a) is the real environment photo of the key frame, Figure 15(b) is the picture of semantic segmentation generated after the construction. Figure 15(c) is the picture of instance segmentation generated on the basis of Figure 15(b), Figure 15(d) is the resulting effect picture of overall target detection.

In order to obtain more accurate experimental results, this paper also uses the data set provided by Computer Vision Group to conduct the test. RGB-D data set provides RGB
image and depth map of 30 Hz frame rate, whose resolution is 640 × 480. The running effect is shown in Figure 16.

The front end part of the algorithm is the pose estimation and simultaneous localization module of the VSLAM. After saving the key frame image, the target detection task is performed. All these key frame images, the corresponding map points, the semantic map information, and the position information of the target object in the key frame are stored. The data fusion operation is performed by transmitting the data to the server so as to achieve real-time operation of semantic maps in the robot. The resulting three-dimensional semantic map is shown in Figure 17.

The global precision of interframe alignment is used to evaluate the accuracy of map construction. By calculating the displacement error and rotation error between the mapping result of each pair of adjacent key frames of key frame sequence extracted from ORB and the real value according to formula, 213 data are obtained for each of them. Their mean value and standard deviation are calculated respectively, and the results were shown in TABLE 3.

Judging from the TABLE 3, the error of semantic map is slighter than that of non-semantic map with or without moving objects. It can be seen that semantic map effectively reduces the influence of object motion and enhances the effectiveness of the final mapping accuracy.

VI. CONCLUSION

The proposed VSLAM algorithm adopts method including ORB feature extraction and RANSAC to solve problems of poor real-time performance and low efficiency in the inter-frame registration process. By combining the target detection and instance segmentation algorithm based on Mask RCNN, a semantic map is constructed for the space where a mobile robot is working inside. The proposed algorithm integrates the geometric information provided by VSLAM algorithm with the semantic information provided by deep neural network to establish the mapping relationship between geometric space and semantic map, which boosts the accuracy of localization and mapping through semantic information and enhances the accuracy of semantic understanding by using location information of objects generated by VSLAM. The experimental results demonstrate that the proposed method can meet the requirement of reconstructing a 3D semantic map accurately.

TABLE 3. Global error statistics of interframe transform.

|                  | non-semantic maps without moving objects | semantic maps without moving objects | non-semantic maps with moving objects | semantic maps with moving objects |
|------------------|------------------------------------------|-------------------------------------|--------------------------------------|----------------------------------|
| displacement error (e ‖) | 0.0068 (+0.0029) | 0.0045 (+0.0029) | 0.0071 (+0.0029) | 0.0057 (+0.0029) |
| rotation error (e θ)   | 0.0138 (+0.0057) | 0.0127 (+0.0057) | 0.0145 (+0.0057) | 0.0134 (+0.0057) |

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961–2969.

[3] Z. Cheng and G. Wang, “Real-time RGB-D SLAM with points and lines,” in Proc. 2nd IEEE Adv. Inf. Manage., Commun., Electron. Autom., Control Conf. (IMCEC), May 2018, pp. 119–122.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: A versatile and accurate monocular SLAM system,” IEEE Trans. Robot., vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[5] J. Navarrete, D. Viejo, and M. Cazorla, “Compression and registration of 3D point clouds using GMMs,” Pattern Recognit. Lett., vol. 110, pp. 8–15, Jul. 2018.

[6] Y. Fan, H. Han, Y. Tang, and T. Zhi, “Dynamic objects elimination in SLAM based on image fusion,” Pattern Recognit. Lett., vol. 127, pp. 191–201, Nov. 2019, doi: 10.1016/j.patrec.2018.10.024.

[7] R. Kang, J. Shi, X. Li, Y. Liu, and X. Liu, “DF-SLAM: A deep-learning enhanced visual SLAM system based on deep local features,” 2019, arXiv:1901.07223 [Online]. Available: http://arxiv.org/abs/1901.07223

[8] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, “Semanticfusion: Dense 3D semantic,” in Proc. ICRA May/Jun. 2017, pp. 4628–4635.

[9] M. Wheelan, S. Leutenegger, R. Salas-Moreno, B. Glocker, and A. Davison, “Elasticfusion: Dense SLAM without a pose graph,” Robot. Sci. Syst., 2015.

[10] X. Li and R. Belaroussi, “Semi-dense 3D semantic mapping from monocular SLAM,” 2016, arXiv:1611.04144 [Online]. Available: http://arxiv.org/abs/1611.04144

[11] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 834–849.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2014, arXiv:1412.7062 [Online]. Available: http://arxiv.org/abs/1412.7062

[13] L. Li, Z. Liu, U. Ozginer, J. Lian, Y. Zhou, and Y. Zhao, “ Dense 3D semantic SLAM of traffic environment based on stereo vision,” in Proc. IEEE Intell. Vehicles Symp. (IV), Jun. 2018, pp. 965–970.
[14] R. Mur-Artal and J. D. Tardos, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[15] K. Tateno, F. Tombari, I. Laina, and N. Navab, “CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6243–6252.

[16] S. Liu and Y. Zhang, “Detail-preserving underexposed image enhancement via optimal weighted multi-exposure fusion,” *IEEE Trans. Consum. Electron.*, vol. 65, no. 3, pp. 303–311, Aug. 2019.

[17] J. Huang and S. Liu, “Robust simultaneous localization and mapping in low-light environment,” *Comput. Animation Virtual Worlds*, vol. 30, nos. 3–4, pp. 1–10, May 2019.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[19] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.

[20] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “Brief: Binary robust independent elementary features,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 778–792.

[21] Z. Zhang, “Flexible camera calibration by viewing a plane from unknown orientations,” in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 666–673.

[22] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[23] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.

[24] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, “Fully convolutional instance-aware semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2359–2367.

[25] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. (2009). *ROS: An Open-Source Robot Operating System.* [Online]. Available: http://www.willowgarage.com/sites/default/files/icraoss09-ROS.pdf

**CHONGBEN TAO** received the B.E. and Ph.D. degrees from Jiangnan University, Wuxi, China. He is currently an Associate Professor with the School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China. His current research interests include deep learning, vision SLAM, semantic mapping, and embedded systems.

**ZHEN GAO** is currently an Assistant Professor with the W Booth School of Engineering Practice and Technology, Faculty of Engineering, McMaster University, Hamilton, ON, Canada. His current research interests include correlation between human intelligence and AI, human–machine natural interaction, and intelligent mechatronics.

**JINLI YAN** received the B.E. degree from the Suzhou University of Science and Technology, Suzhou, China, where he is currently pursuing the master’s degree with the School of Electronic and Information Engineering. His current research interests include vision SLAM and semantic mapping.

**CHUNGUANG LI** received the B.E. degree from the Wuhan University of Technology, Wuhan, China, and the M.E. and Ph.D. degrees from Jiangnan University, Wuxi, China. He is currently an Associate Professor with the School of Computer and Information Engineering, Changzhou Institute of Technology, Changzhou, China. His current research interests include deep learning, multiple agents, flocking control, embedded systems, and gait planning for biped robot.

**GUOZENG CUI** received the B.Sc. degree in applied mathematics from the Shandong University of Technology, Zibo, China, in 2009, the M.Sc. degree in applied mathematics from Qufu Normal University, Qufu, China, in 2012, and the Ph.D. degree in control science and engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2016. He is currently a Lecturer with the School of Electronic and Information Engineering, Suzhou University of Science and Technology. His research interests include adaptive control, intelligent control, and multiagent systems.

---

**CHONGBEN TAO** received the B.E. and Ph.D. degrees from Jiangnan University, Wuxi, China. He is currently an Associate Professor with the School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, China. His current research interests include deep learning, vision SLAM, semantic mapping, and embedded systems.