IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 13, 2022, accepted February 20, 2022, date of publication February 24, 2022, date of current version March 2, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154086

Semantic SLAM Based on Improved DeepLabv3$^+$ in Dynamic Scenarios

ZHANGFANG HU, JIANG ZHAO, YUAN LUO, AND JUNXIONG OU

Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Jiang Zhao (2282781255@qq.com)

This work was supported in part by the National Natural Science Foundation for Young Scholars of China under Grant 61703067 and Grant 61803058, and in part by the National Natural Science Foundation of China under Grant 51775076 and Grant 61801061.

ABSTRACT Simultaneous Localization and Mapping (SLAM) plays an irreplaceable role in the field of artificial intelligence. The traditional visual SLAM algorithm is stable assuming a static environment, but has lower robustness and accuracy in dynamic scenes, which affects its localization accuracy. To address this problem, a semantic SLAM system is proposed that incorporates ORB-SLAM3, semantic segmentation thread and geometric thread, namely DeepLabv3$^+$_SLAM. The improved DeepLabv3$^+$ semantic segmentation network combines context information to segment potential a priori dynamic objects. Then, the geometry thread uses a multi-view geometry method to detect the motion state information of the dynamic object. Finally, a new ant colony strategy is proposed to find the group of all dynamic feature points through the optimal path, and avoids traversing all the feature points to reduce the dynamic object detection time and improve the real-time performance of the system. By conducting experiments on public data sets, the results show that the method proposed in this paper effectively improves the positioning accuracy of the system in a high-dynamic environment compared with similar algorithms, and the real-time performance of the system is improved.

INDEX TERMS DeepLabv3$^+$_SLAM, semantic, high-dynamic environment, new ant colony strategy.

I. INTRODUCTION

With the rapid development of robotics and computer science, autonomous mobile robots are widely used in various fields such as industry and agriculture. As one of the most advanced technologies in the field of robot motion, SLAM uses the sensor data from a robot for autonomous positioning and map construction. From the mutual dependence of robot autonomous localization and map construction, only accurate autonomous localization is necessary to build a correct map. A correct map can help the robot determine its position in the map accurately.

At present, most visual SLAM frameworks operate under the assumption of a static environment, such as ORB-SLAM [1], ORB-SLAM2 [2], ORB-SLAM3 [3], LSD-SLAM [4], RGB-D SLAM [5]. Among these frameworks, ORB-SLAM3 is considered to be the advanced method currently used in static scenes. ORB-SLAM3 is a system based on ORB-SLAM2 and ORB-SLAM-VI that can operate robustly in purely visual or visual inertial guidance systems, and is a complete and highly accurate generalized system. These algorithms can achieve satisfactory results in a static environment or an environment with a small number of dynamic objects. However, when the robot is operated in an environment with a large number of dynamic objects (e.g., people, vehicles), the performance of the visual SLAM algorithm will significantly decrease. This is a result of the visual features from dynamic objects in the environment, which affects the positional estimation of the robot and greatly decrease the positioning accuracy of the system. In recent years, with the development of deep learning technology, increasingly more excellent image algorithms have been applied to visual SLAM, which provide methods and ideas for improving the localization accuracy of the system.

In this article, we propose a multi-threaded parallel semantic SLAM system to solve the problem when facing dynamic objects. The system is mainly based on the ORB-SLAM3 algorithm framework and introduces semantic segmentation and multi-view geometry approaches to the original framework. In the semantic segmentation thread, the ResNest [6]

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenbao Liu.

classification network with higher accuracy is used to replace the original ResNet [7] in the DeepLabv3$^+$ [8] segmentation network, which helps segment object boundaries more accurately. The dilated convolution [9] with a smaller dilation rate is more effective in extracting low-resolution feature map information, so a new layer of dilated convolution is added to the Atrous Spatial Pyramid Pooling (ASPP) module of DeepLabv3$^+$, and the dilation rate size is adjusted. Simultaneously, to reduce the amount of network parameters and improve the efficiency and training speed of the network, we replace all the dilated convolutions with depthwise separable convolutions [10] and perform 2D decomposition. In the geometry thread, the method of multi-view geometry is used to determine the motion state of the object, and a new ant colony search strategy is proposed to avoid the multi-view geometry method having to analyze all the feature points using the distribution characteristics of feature points on the image. This improves the robustness and real-time performance of the system.

The rest of this paper is organized as follows. Section II briefly describes some achievements and shortcomings of various visual SLAMs in dynamic scenarios. Section III elucidates the architecture of our SLAM system. In Section IV, we conduct experiments on the TUM RGB-D dataset to verify the effectiveness and accuracy of the DeepLabv3$^+$_SLAM system. Finally, in Section V, we conclude and discuss the paper.

II. RELATED WORK

The main methods for obtaining the semantic information of objects include target detection and semantic segmentation. Target detection is the determination of the object bounding box, and semantic segmentation is the accurate classification of objects. Both target detection and semantic segmentation can be used to recognize dynamic objects in a scene. In comparison, semantic segmentation is better at recognizing the results of objects, because the contour of objects can be accurately segmented. However, the bounding box may contain pixels that do not belong to the object. After processing the abnormal objects using semantic segmentation, a static background model without any dynamic objects is established, thus improving the accuracy and robustness of the visual SLAM system in dynamic environments.

With the rise of neural networks, semantic segmentation has been gradually introduced into the SLAM semantic system. For instance, Yu *et al*. [11] proposed a DS-SLAM scheme, which combines the visual SLAM algorithm with the SegNet [12] network to filter the dynamic part using semantic information and motion feature points in dynamic scenes. This method improvs the accuracy of pose estimation, but the types of objects that can be recognized by the semantic segmentation network in this scheme are limited, which limits the scope of its application. Zhong *et al*. [13] combined ORB-SLAM2 and SSD [14] into a new coupling framework named Detect-SLAM, and proposed a method to propagate the motion probability of key points in real time to overcome the delay of target detection threads. Semantic information is used to eliminate the negative effects caused by moving objects in SLAM. This framework aims to improve the efficiency of target detection and sensitivity to the viewpoint transformation problem, and the real-time performance of the system must to be further optimized. Xiao *et al*. [15] proposed Dynamic-SLAM, constructed an SSD target detector based on a convolutional neural network, and proposed a missed detection compensation algorithm based on constant speed of adjacent frames to address the problem of low recall of SSD target detection network, which greatly improved the recall of detection. A selective tracking algorithm is also proposed to simply eliminate dynamic objects, which improves the robustness and accuracy of the system. Cui and Ma [16] proposed a semantic optical flow method, which combines the semantic information before motion, aids in the calculation of the epipolar geometry, filters out the true dynamic features, and keeps only the remaining static features fed into the tracking optimization module to achieve accurate estimation of the camera pose in a dynamic environment. Zhang *et al*. [17] proposed VDO-SLAM, a dynamic feature-based SLAM system, which utilizes image-based semantic information in the scene without prior knowledge of object pose or geometry to achieve localization, map building, and tracking of dynamic objects simultaneously. However, there are cases in which large errors occur due to problems with the algorithm or optimization function, and the real-time performance requires improvement. Chen *et al*. [18] proposed DM-SLAM, which combines the instance segmentation network Mask-R CNN with optical flow and epipolar geometry to constrain the outliers in the scene. Two different strategies for obtaining segmentation results of potential dynamic objects in the dynamic point detection segment are proposed. One method reprojects the feature points with depth information to the current frame, and uses the reprojection offset vector to distinguish dynamic points. The other method uses the epipolar geometric constraints. Long *et al*. [19] proposed PSPNet-SLAM, which integrates the semantic thread and geometric thread of the pyramid structure into ORB-SLAM2 through pyramid scene resolution SLAM, which uses a semantic thread combined with contextual information to segment dynamic objects. The best error compensation homography matrix is designed to improve the accuracy of dynamic point detection, but the ability of the network to process image frames affects the real-time performance of the system, and the ability to remove dynamic objects needs to be improved. Bescos *et al*. [20] proposed DynaSLAM, which processes monocular and RGB-D cameras differently. In the case of monocular, Mask R-CNN [21] is used to detect moving objects, and in RGB-D mode, the Mask R-CNN network and the multi-view geometric model are combined to detect moving objects. This method can detect multiple moving objects in the environment and repair the background occluded by dynamic objects. However, the system has difficulty operating in real-time, because the Mask R-CNN network is time and resource-consuming for images processing. Ai *et al*. [22]

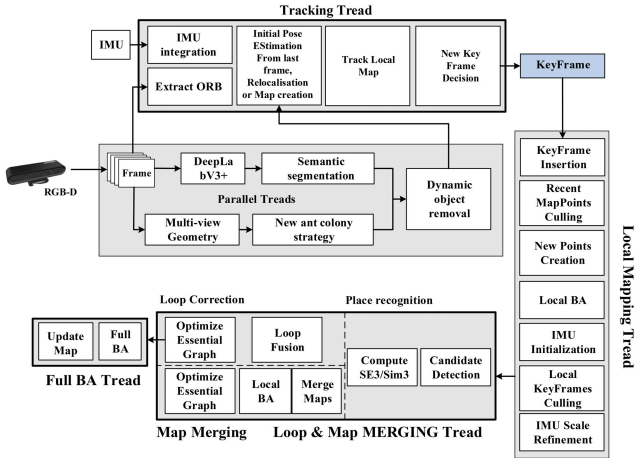

FIGURE 1. Structure of DeepLabv3+_SLAM.

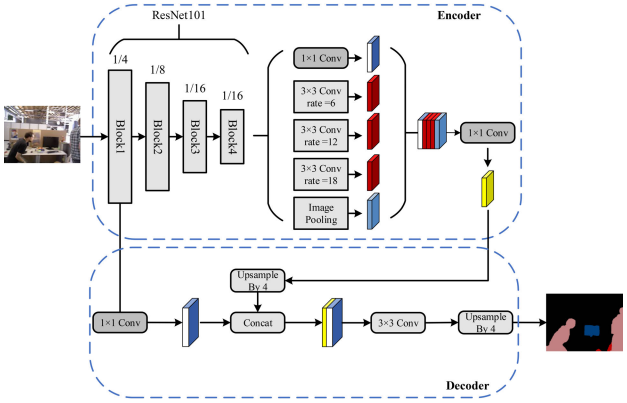

FIGURE 2. DeepLabV3+ network structure.

proposed the DDL-SLAM system, which improves the segmentation and background restoration abilities. By combining semantic segmentation and multi-view geometric algorithms to filter out dynamic objects in the scene, the static scene map can repair the background obscured by moving objects for restoration, thus improving the localization accuracy in highly dynamic environments. However, the real-time performance is remains insufficient.

Compared with the traditional ORB-SLAM3, although the various solutions proposed above how better performance when detecting the semantic information of objects, there is room for improvement and research on the correlation between objects in the semantic information, localization accuracy and the real-time performance of the system.

III. SYSTEM DESCRIPTION

The system proposed in this paper improves on the basis of ORB-SLAM3. The overall structure block diagram is shown in Fig. 1. In the improved framework, semantic threads and geometric threads are added. First, the RGB-D camera collects image data. Then, the data is passed into the tracking thread for pre-processing, and the DeepLabv3+ model subdivides all the a priori dynamic contents by pixels, while using the geometric thread module to distinguish the dynamic and static feature points in the image. Second, the segmentation results from the DeepLabv3+ model and the motion state information judged by the geometry module are combined and used to extract the contour regions of dynamic objects. Finally, feature points and spatial points of dynamic object regions are removed, and image frames with only static features are used for subsequent tracking and map building, thus improving the accuracy and robustness of the visual SLAM system in a highly dynamic environment.

A. SEMANTIC SEGMENTATION DeepLabv3+

In traditional semantic systems, convolutional neural networks such as fully convolutional neural network [23] (FCN), U-net based on codec architecture [24], SegNet and other algorithms are used in visual SLAM systems. However, each of these algorithms have problems, such as the lack of in ability to infer information from the context, the inability to handle the relationship between the scene and global information, or unable to effectively deal with the relationship between categories leading to the failure of label association, resulting in discontinuous predictions. DeepLabv3+ is the best segmentation model among a series of DeepLab [25]–[27] models proposed by Google, but the model is not superior in terms of processing speed and model capacity. The overall structure of DeepLabv3+ is shown in Fig. 2. This model introduces the idea of Encoder-Decoder based on Dilated FCN. The main function of Encoder is to gradually reduce the resolution of the feature map and provide high-level semantic information. The main body of Encoder is DCNN with dilation convolution, and the classification network used can be ResNet, Xception or another network, followed by the ASPP module, which introduces multi-scale information to capture rich contextual information by performing pooling operations at different resolutions.

Assuming H_k^r is expressed as a convolution operation with a convolution kernel size of k and a dilation rate of r, its output can be expressed as:

$$y = H_3^6(x) + H_3^{12}(x) + H_3^{18}(x), \tag{1}$$

The main function of the Decoder module is to further fuse the low-level features and high-level features to improve the accuracy of the segmentation boundaries and recover spatial information. The Decoder obtains a feature map with a resolution of 4 after bilinear upsampling 4 times from the feature map output by Encoder, and then splices and fuses this feature map with the feature map obtained after 1×1 convolution and dimensionality reduction in the backbone network. Finally, the module up-samples 4 times by 3×3 convolution to obtain the final predicted semantic segmentation map.

As the backbone network of DeepLabv3+, ResNet performs well. ResNet mainly uses a residual structure based on bottleneck design, which is generally used when the number of network layers is greater than 30, so that the network parameters can be significantly reduced and deeper networks

can be trained. The ResNet network has largely alleviated the problem of network degradation caused by the deepening of network layers to a great extent, so that the network can learn deeper image features. However, the size of its receptive field is fixed and single, which cannot be used to fuse multi-scale features, and does not take advantage of the interaction between cross-channel features. ResNest's proposal makes up for the shortcomings of ResNet.

ResNest is a modification of ResNet that combines the split attention of the feature map in a single network, and extends the attention mechanism of the channel dimension to the representation of the feature map group to form modularization, as shown in Fig. 3. Compared with ResNet or its variants, ResNest does not require additional calculations, and the result is a significant improvement compared to ResNet and its variants. Therefore, this article uses ResNest as the backbone network of DeepLabv3$^+$, so that the semantic thread in the SLAM system has better image segmentation performance.

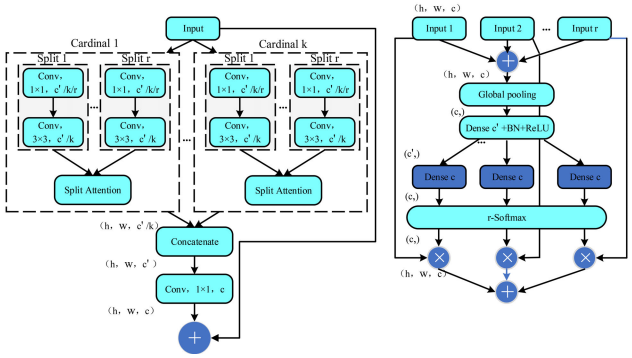

FIGURE 3. ResNest's split-attention block.

B. ASPP MODULE

In the Encoder session, the convolutional layers in the original ASPP module are 1 × 1 dilation convolution, 3 × 3 dilation convolution with a dilation rate of 6, 12, and 18, and a global average pooling layer. With the continuous extraction of image features by the backbone network, the resolution of the feature map will continue to decrease, and the dilation convolution with a larger dilation rate is not conducive to extracting feature map information with lower resolution. To address this problem, a new layer of dilation convolution is added to the original dilation convolution, and the dilation rate is adjusted to 4, 8, 12, and 16 to improve the extraction of low-resolution feature map information, the output of which can be expressed as:

$$y = H_3^4(x) + H_3^8(x) + H_3^{12}(x) + H_3^{16}(x), \tag{2}$$

ASPP stacks the dilation convolutions of different dilation rates in parallel to obtain multi-scale information gain. The one-dimensional mathematical expression of dilation rate is:

$$p = \sum_{s=1}^{S} x[i + r \times s]w[s], \tag{3}$$

where $x[i]$ means the input signal, $y[i]$ denotes the output signal, r is the step size of the sampling, $w[s]$ represents the size of the convolution kernel as a parameter at position s, and S means the size of the convolution kernel.

Comparing the depthwise separable convolution with the standard convolution, we found that depthwise separable convolution can largely reduce the excessive number of parameters in the training process. The number of parameters in the standard convolution is about three times the number of parameters in the depthwise separable convolution for the same input. Therefore, we replace all the dilation convolutions in ASPP with depthwise separable convolutions to improve the training performance and efficiency of the system with less impact on the segmentation accuracy.

The main function of ASPP is to extract multi-scale information from the feature map. However, the 3 × 3 convolution will learn redundant information, result in an increase in the number of system parameters that affects the speed of the system. In this paper, all 3 × 3 convolutions in ASPP are transformed into 3 × 1 and 1 × 3 convolutions using 2-dimensional decomposition without changing the dilation rate. This reduces the number of parameters compared with the original structure by about 1/3, effectively reducing the computation of this module, with faster training speed and the ability to extract important feature information.

The improved ASPP module is shown in Fig. 4. When the feature map generated by the backbone network is sent to ASPP for processing, the feature map is first subjected to a 1 × 1 convolution, the convolution with dilation rates of 4, 8, 12 and 16, and the global average pooling operation is performed. Then, the six feature maps obtained are spliced and fused in the channel dimension. Finally, the feature map containing high-level semantic features is obtained after 1 × 1 convolution and dimensionality reduction operation.

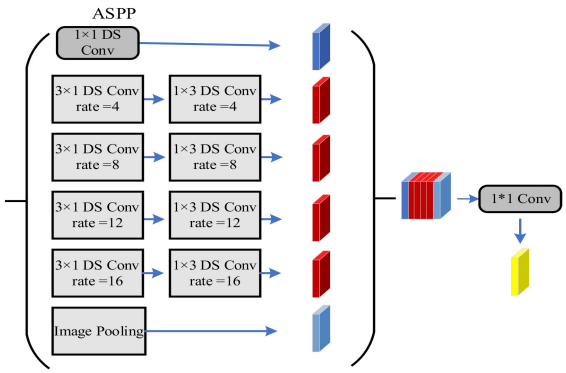

FIGURE 4. Modified ASPP framework.

C. DYNAMIC OBJECT DETECTION BASED ON MULTI-VIEW GEOMETRY

Semantic segmentation networks can only detect dynamic objects with a priori high probability, but in actual scenes, the SLAM system will often be disturbed by static objects. Books and chairs are examples of static objects. However,

when people move with books or chairs, they should be regarded as dynamic objects but are regarded as static objects to participate in the positioning and mapping. This results in a great impact on the SLAM system. Therefore, we use a dynamic object segmentation method based on multi-view geometry for processing. As shown in Fig. 5, the map point cloud is projected to the current frame, and the object is distinguished as a dynamic object or static object based on the viewpoint difference and size of the change in depth value. By calculating the viewing angle value v_{cf} of each key point in the current frame (cf) and the viewing angle value v_{hf} of the historical frame (hf), if the difference $\Delta v = |v_{cf} - v_{hf}|$ of the viewing angle value is greater than the set threshold, the key point is determined to be a dynamic point. At the same time, we also need to calculate the depth value d_{cf} of the key point in the current frame and the projection depth value d_{proj} of the historical frame in the current frame. If the difference between the depth values is $\Delta d = |d_{proj} - d_{cf}| = 0$, the key point is determined to be a static point. If Δd is greater than the set threshold d_{thresh}, the key point is considered dynamic.

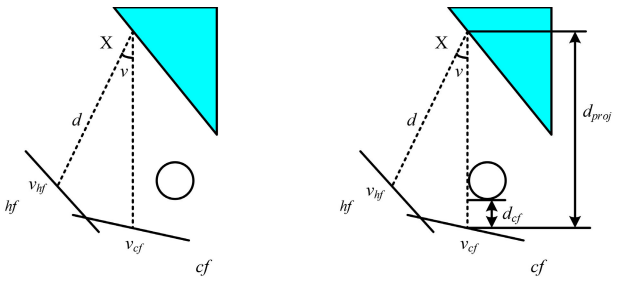

FIGURE 5. Multi-view geometry.

D. ANT COLONY STRATEGY

The ant colony algorithm [28] is a simulation optimization algorithm that simulates the foraging behavior of ants. Ants release pheromones related to the path length during movement. The path length is inversely proportional to the pheromone concentration, where the optimal path has the largest pheromone concentration. Ants choose their path according to pheromone concentration. The ant colony algorithm has two main processes: state transfer and pheromone update. Assuming that the probability of ant m moving from node i to node j is p_{ij}^m, its state transition rule is given by the following equation:

$$p_{ij}^m = \begin{cases} \dfrac{[\tau(i,j)]^\alpha \times [\eta(i,j)]^\beta}{\sum\limits_{S \in j_m(i)} [\tau(i,s)]^\alpha \times [\eta(i,s)]^\beta}, & j \in allowed_m \\ 0, & otherwise \end{cases} \tag{4}$$

where $\tau(i,j)$ denotes the pheromone concentration on the path from i to j, $\eta(i,j)$ is the corresponding heuristic information function, α is the information heuristic factor, β is the expected heuristic factor, and $allowed_m$ is the node not visited by the ant. The greater the value of α, the more likely the ant is to choose the path before moving, and the randomness of the search path is weakened. The smaller the value of α, the smaller the search range, and it is easy to fall into the local optimum. The larger the value of β, the easier it is for the ant colony to choose the local shortest path, and the convergence speed of the algorithm is accelerated. When the ant completes a path transfer, it will perform a pheromone update. The update rules are as follows:

$$\tau_{ij}(t+n) = (1-\rho) \times \tau_{ij}(t) + \Delta\tau_{ij}(t), \tag{5}$$

$$\Delta\tau_{ij}(t) = \sum_{m=1}^{M} \Delta\tau_{ij}^m(t), \tag{6}$$

where ρ is the information volatilization factor, $\rho \in [0, 1)$, $1-\rho$ denotes the residual factor, and $\Delta\tau_{ij}(t)$ is the pheromone increase from i to j at time t. When ρ is too small, there are too many pheromones that remain on each path, resulting in the continued search of invalid paths, which affects the efficiency of the algorithm. When ρ is too large, although invalid paths can be excluded from the search range, valid paths may also be excluded, affecting the search for the optimal solution.

E. NEW ANT COLONY STRATEGY

When the multi-view geometry method transforms the image of the historical frame into the current frame by projection, a large number of projected feature points will be obtained. A point is determined to be a static point or dynamic point by traversing all the projected feature points. However, there are thousands of feature points in the feature extraction, and if each feature must be determined to be static or dynamic, the real-time performance of SLAM will be limited. In this paper, based on the strategy of the ant colony algorithm, we propose a new ant colony strategy to find the group of all dynamic feature points through the optimal path, so as to avoid traversing all feature points, reduce the time-consumed by feature point extraction and improve the real-time performance of SLAM.

In the ant colony algorithm strategy, throughout the process from the origin to the destination, the ant colony avoids obstacles they encounter to find an optimal path to the destination. Based on this strategy, this paper sets a search path from the starting point to the destination, and searches feature points on the path in turn. Because the dynamic points or static points in the image are distributed in groups rather than chaotically scattered throughout the image, when a dynamic feature point is found, the search will be transferred to the group in which the feature point is located until all the feature points of the entire group are detected or the search exceeds the range of the group. The next dynamic feature point group will then be searched. When a static feature point is detected, the point and its group will not be processed, and the search will continue according to the path.

According to the distribution of feature points in the image, this paper designs a path l from the departure S to the destination T, as shown in Fig. 6. The search strategy is: the ant colony moves continuously from the feature point $m_i = 0$ on the path to the next point $m_i (i = 1, 2, \ldots, n)$ until reaching

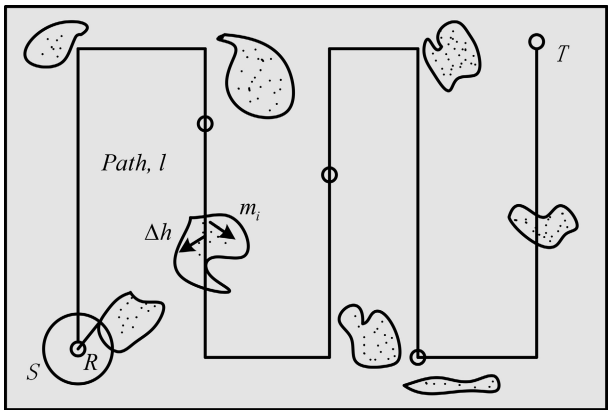

FIGURE 6. New ant colony strategy.

the destination target T. On the moving path, each feature point m will take itself as the origin, and search for feature points within a radius R. If a dynamic point is not found, the search will continue to move forward on path l. When a dynamic point is found, expand outward with the bandwidth Δh. If the next new dynamic point is found, continue to expand outward with Δh until no dynamic point is found in the expanded area, then return to path l and continue to search the next feature point m_i that matches the dynamic feature in turn until path l is completed.

IV. EXPERIMENT AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATA SET

In this section, to compare the performance of our semantic SLAM system and other excellent SLAM systems in dynamic environments, experiments are conducted on the data set TUM RGB-D. In addition, the proposed system is compared with the original ORB-SLAM3 to quantify its improvement in dynamic scenarios. All experiments were performed on a computer equipped with an Intel i7 CPU, RTX2080Ti GPU and 16 GB of memory.

The TUM dataset is an excellent dataset for evaluating camera positioning accuracy and provides an accurate ground truth for the sequences. The dataset contains 7 sequences recorded by an RGB-D camera at 30 fps with a resolution of 640 × 480. In this section, we use 5 sequences from the TUM dataset to evaluate the performance and demonstrate the effectiveness of DeepLabv3$^+$_SLAM in dynamic environments, namely fr3_s_static, fr3_w_static, fr3_w_rpy, fr3_w_xyz, fr3_w_halfsphere. Besides fr3_s_static which is a static sequence, the other sequences are dynamic sequences. The ''s'' in the sequence name means ''sitting'' and ''w'' means ''walking''. The word after the underscore indicates the state of the camera, for example, ''xyz'' indicates that the camera moves along the x-y-z axis.

To quantitatively evaluate the advantages of our algorithm, the overall performance of the system is evaluated using Absolute Trajectory Error (ATE), which indicates the global consistency of the trajectory, and Relative Pose Error (RPE), which measures translational and rotational drift. Root Mean Square Error (RMSE) can reflect the accuracy and robustness of the system better than the mean and median values, and the Standard Deviation Error (S.D.) can reflect the stability of the system. Therefore, in this paper, the RMSE value and S.D. value of ATE and RPE are obtained by processing each sequence separately to judge the positional accuracy and system stability.

B. EXPERIMENTAL RESULT

The ATE and RPE of ORB-SLAM3, DynaSLAM and DeepLabv3$^+$_SLAM algorithms were obtained by conducting experiments on 5 sequences. The results are shown in Tables 1-3.

As shown in the table, DeepLabv3$^+$_SLAM and DynaSLAM can significantly reduce the ATE and RPE of each sequence compared to ORB-SLAM3. In highly dynamic sequences, the method in this paper shows a significant improvement in ATE and PRE compared to DynaSLAM, and in terms of ATE, the improvement values of RMSE and S.D. reach 25.18 % and 31.88%, mainly because the proposed semantic segmentation network not only has better performance, but also considers the information correlation with geometric threads, so that the DeepLabv3$^+$_SLAM system can significantly improve its localization accuracy and robustness in high dynamic environments. In the low dynamic sequence fr3_s_static, the improvement of the method in this paper is not obvious compared with ORB-SLAM3. This is mainly because ORB-SLAM3 itself is designed for low dynamic environment and can handle low dynamic scenes well and achieve good results, so the room for improvement is limited.

Figs. 7-9 show the ATE and RPE of ORB-SLAM3, DynaSLAM and DeepLabv3$^+$_SLAM in the high dynamic sequence fr3_w_xyz. The black line represents the real trajectory of the camera and the blue line indicates the camera trajectory estimated by the SLAM algorithm. In the high dynamic environment, the motion trajectory estimated by the ORB-SLAM3 system has a large gap with the real trajectory, and even produces wrong trajectories in some regions. On the contrary, DynaSLAM and DeepLabv3$^+$_SLAM systems have high overlap between the estimated motion trajectories and the true trajectories because the dynamic objects in the scene are eliminated, and the motion trajectories estimated by DeepLabv3$^+$_SLAM are closer to the true trajectories than those estimated by DynaSLAM. This indicates that the method in this paper is more capable of handling highly dynamic scenes.

The purpose of this paper is to remove the feature points on dynamic targets and keep only the remaining static feature points. Therefore, to verify the effect of dynamic feature point rejection, this paper conducts experiments on the high dynamic sequence fr3_w_xyz. Fig. 10 shows the original image, the semantic segmentation image, and the image with unprocessed dynamic feature points from top to bottom, where the green dots represent the locations of ORB feature

TABLE 1. Absolute trajectory error results (ATE[m]).

sequences	ORB-SLAM3		DynaSLAM		DeepLabv3+_SLAM		Improvement (compare to ORB-SLAM3) percentage (%)		Improvement (compare to DynaSLAM) percentage (%)	
	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
fr3_s_static	0.0072	0.0044	0.0064	0.0044	0.0052	0.0035	27.78%	20.45%	18.75%	20.45%
fr3_w_static	0.3692	0.0528	0.0068	0.0034	0.0060	0.0028	98.37%	94.70%	11.76%	17.65%
fr3_w_xyz	0.6787	0.3258	0.0156	0.0079	0.0135	0.0065	98.01%	98.00%	13.46%	17.72%
fr3_w_rpy	0.8161	0.3024	0.0417	0.0276	0.0312	0.0188	96.18%	93.78%	25.18%	31.88%
fr3_w_halfphere	0.5939	0.3055	0.0302	0.0156	0.0276	0.0127	95.35%	95.84%	8.61%	18.59%

TABLE 2. Translation drift results (RPE[m/s]).

sequences	ORB-SLAM3		DynaSLAM		DeepLabv3+_SLAM		Improvement (compare to ORB-SLAM3) percentage (%)		Improvement (compare to DynaSLAM) percentage (%)	
	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
fr3_s_static	0.0079	0.0046	0.0081	0.0041	0.0068	0.0035	13.92%	23.91%	16.05%	14.63%
fr3_w_static	0.1752	0.1569	0.0093	0.0048	0.0086	0.0038	95.09%	97.58%	7.53%	20.83%
fr3_w_xyz	0.4105	0.2674	0.0206	0.0107	0.0178	0.0092	95.66%	96.56%	13.59%	14.02%
fr3_w_rpy	0.4145	0.3116	0.0592	0.0422	0.0406	0.0237	90.21%	92.39%	31.42%	43.84%
fr3_w_halfphere	0.3562	0.2734	0.0279	0.0132	0.0238	0.0108	93.32%	96.05%	14.70%	18.18%

TABLE 3. Rotational drift results (RPE[deg/s]).

sequences	ORB-SLAM3		DynaSLAM		DeepLabv3+_SLAM		Improvement (compare to ORB-SLAM3) percentage (%)		Improvement (compare to DynaSLAM) percentage (%)	
	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
fr3_s_static	0.2833	0.1223	0.2713	0.1179	0.2488	0.1039	12.18%	15.04%	8.29%	11.87%
fr3_w_static	3.2495	2.8642	0.2522	0.114	0.2342	0.1104	92.79%	96.15%	7.14%	3.92%
fr3_w_xyz	7.8974	5.4517	0.6229	0.3824	0.5785	0.3602	92.67%	93.39%	7.13%	5.81%
fr3_w_rpy	8.3134	6.0427	1.3207	0.9055	0.9274	0.5219	88.84%	91.36%	29.78%	42.36%
fr3_w_halfphere	6.3729	5.1487	0.7933	0.3865	0.7268	0.3447	88.60%	93.31%	8.38%	10.82%

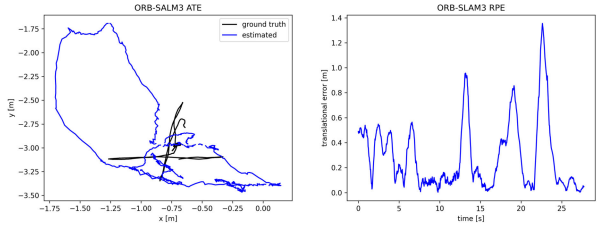

FIGURE 7. ATE and RPE of ORB-SLAM3 in fr3_w_xyz.

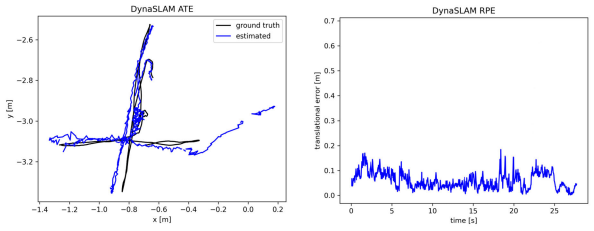

FIGURE 8. ATE and RPE of DynaSLAM in fr3_w_xyz.

points. As can be seen from the figure, the feature points falling on dynamic objects have been detected and removed by the method in this paper, while other feature points falling on static objects are retained. There are also feature points in some regions at the edges of the human body that are not well rejected, which is related to the accuracy of semantic segmentation.

In practical applications, real-time performance is an important metric for evaluating SLAM systems. Therefore, to evaluate the real-time performance, we let DeepLabv3+_SLAM and DynaSLAM run five sequences under the same hardware conditions and record the time consumed by the geometric threads, and the results are shown in Table 4. In terms of running time, since this paper introduces a new ant colony strategy in the geometric threads, which greatly reduces the time consumed by the geometric method to judge the object state information, the method in this paper has

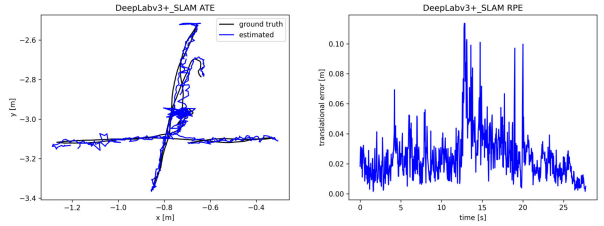

FIGURE 9. ATE and RPE of DeepLabV3$^+$_SLAM in fr3_w_xyz.

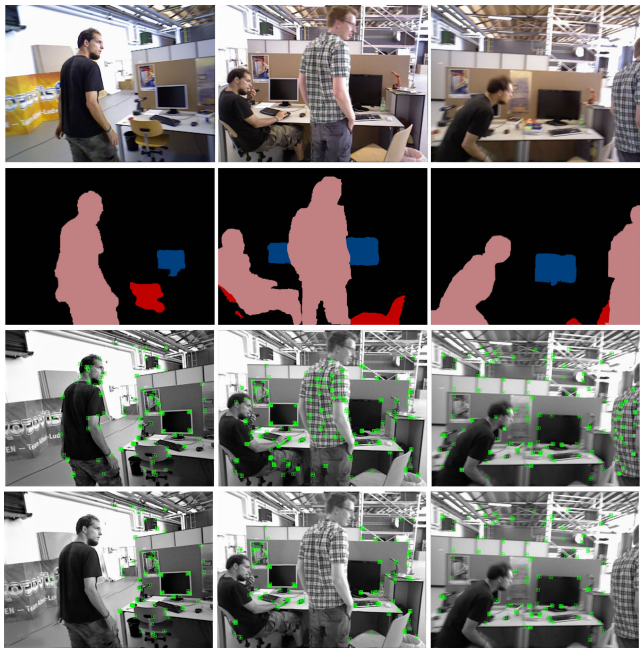

FIGURE 10. Experimental results on a highly dynamic sequence fr3_w_xyz from the TUM RGB dataset.

TABLE 4. Time evaluation.

Sequence	DynaSLAM [ms]	DeepLabV3$^+$_SLAM [ms]
fr3_s_static	156.42	146.21
fr3_w_static	227.32	136.28
fr3_w_xyz	323.48	172.17
fr3_w_rpy	209.84	149.89
fr3_w_halfphere	337.47	190.54

better real-time performance compared with DynaSLAM, thus improving the overall real-time performance of the SLAM system.

V. CONCLUSION

In order to eliminate the influence of dynamic objects on the positioning accuracy of the system, we propose the DeepLabv3$^+$_SLAM system. This system introduces semantic and geometric threads based on the original ORB-SLAM3. First, a priori dynamic information is obtained through semantic threads. Then, the dynamic feature points in the scene are detected in the geometry thread using a multi-view geometry approach, while a new ant colony strategy is proposed to selectively detect dynamic feature points using the distribution characteristics of the feature points in order to improve the real-time performance of the geometry thread. Finally, to verify the overall performance of the system in this paper, we conducted experiments and analyses on the TUM RGB-D dataset, and the results show that the localization accuracy and real-time performance of the system in this paper are improved in a highly dynamic environment compared with existing advanced SLAM frameworks.

Despite the progress in localization accuracy and real-time performance, there are still many deficiencies. On the one hand, the real-time performance of the system still needs to be improved, and the speed of geometric thread image frame processing needs to be improved. On the other hand, we still need to continuously optimize the semantic segmentation network to improve the accuracy of network segmentation, or select other excellent and lightweight networks to help the system more effectively eliminate the impact caused by dynamic objects.

REFERENCES

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

[2] R. Mur-Artal and J. D. Tardós, "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.

[3] C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," 2020, *arXiv:2007.11898*.

[4] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, Sep. 2014, pp. 834–849.

[5] G. Hu, S. Huang, L. Zhao, A. Alempijevic, and G. Dissanayake, "A robust RGB-D SLAM algorithm," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1714–1719.

[6] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[9] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[11] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1168–1174.

[12] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*.

[13] F. Zhong, S. Wang, Z. Zhang, C. Chen, and Y. Wang, "Detect-SLAM: Making object detection and SLAM mutually beneficial," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1001–1010.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," 2015, *arXiv:1512.02325*.

[15] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou, "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment," *Robot. Auton. Syst.*, vol. 117, pp. 1–16, Jul. 2019.

[16] L. Cui and C. Ma, "SOF-SLAM: A semantic visual SLAM for dynamic environments," *IEEE Access*, vol. 7, pp. 166528–166539, 2019.

[17] J. Zhang, M. Henein, R. Mahony, and V. Ila, "VDO-SLAM: A visual dynamic object-aware SLAM system," 2020, *arXiv:2005.11052*.
[18] J. Cheng, Z. Wang, H. Zhou, L. Li, and J. Yao, "DM-SLAM: A feature-based SLAM system for rigid dynamic scenes," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 4, pp. 1–18, 2020.
[19] X. Long, W. Zhang, and B. Zhao, "PSPNet-SLAM: A semantic SLAM detect dynamic object by pyramid scene parsing network," *IEEE Access*, vol. 8, pp. 214685–214695, 2020.
[20] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.
[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 2961–2969.
[22] Y. Ai, T. Rui, M. Lu, L. Fu, S. Liu, and S. Wang, "DDL-SLAM: A robust RGB-D SLAM in dynamic environments combined with deep learning," *IEEE Access*, vol. 8, pp. 162335–162342, 2020.
[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Nov. 2015, pp. 234–241.
[25] 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*.
[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[28] X. Dai, S. Long, Z. Zhang, and D. Gong, "Mobile robot path planning based on ant colony algorithm with A* heuristic method," *Frontiers Neurorobot.*, vol. 13, p. 15, Apr. 2019.

ZHANGFANG HU received the master's degree from the University of Electronic Science and Technology, Sichuan, China, in 1994. She was a Visiting Scholar at Zhejiang University, China. She is currently a Professor with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). Her research interests include photoelectric sensing and opto-electronic information processing.

JIANG ZHAO received the B.S. degree from Sichuan Agricultural University (SICAU), Sichuan, China, in 2018. He is currently pursuing the master's degree with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). His current research interests include mobile robot and semantic SLAM.

YUAN LUO received the M.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 1996, and the Ph.D. degree from Chongqing University, Chongqing, in 2003. She was a Visiting Scholar at the Université de Montréal, Canada, in 2006. She is currently a Professor with CQUPT. Her research interests include computer vision, photoelectric sensing, image processing, and mobile robots.

JUNXIONG OU received the B.S. degree from the Sichuan University of Science and Engineering (SUSE), Sichuan, China, in 2019. He is currently pursuing the master's degree with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). His current research interests include computer vision, deep learning, and V-SLAM.