RESEARCH ON SEMANTIC SLAM SYSTEM TECHNOLOGY FOR DRIVERLESS VEHICLES

Xiaojing CHEN[1*], Libing ZHOU[1], Zhengqian YU[1], Jianjian WEI[1], Xueli JIANG[1], Baisong YE[1], Yexin ZHAO[1], Tianyu WANG[1], Guoqing WANG[1], Jun BIAN[1]

In response to the problem of a single sensor being unable to complete localization and map construction in large-scale scenarios, as well as dynamic obstacles reducing the accuracy of positioning and mapping, a framework called LIS_SLAM for simultaneous localization and map construction framework that combines image semantic segmentation and laser inertial odometer is proposed. First, the efficiency and performance of the image segmentation model are improved by replacing the backbone network and introducing an attention mechanism. Second, spatio-temporal synchronization between sensors is established, enabling semantic segmentation of single-frame point clouds. The framework also includes the establishment of a semantic SLAM system and the construction of a three-dimensional semantic map. Finally, the algorithm is verified in campus and urban environment roads. The experimental results show that LIS _ SLAM can achieve simultaneous localization and mapping in large-scale scenes.

Keywords: multi-sensor fusion; semantic segmentation; simultaneous localization and mapping; dynamic scenarios

1. Introduction

The rapid growth of China’s motor vehicle industry resulted in a sharply increased traffic volume, providing convenience to people but also leading to more frequent traffic accidents. Unmanned driving technology has emerged as a potential solution to address these challenges. SLAM technology, as a fundamental component of autonomous driving, enables vehicles to map their surroundings and accurately position themselves, thereby reducing accidents caused by human factors and propelling the automotive industry towards intelligence.

SLAM refers to a carrier equipped with sensors (such as camera, lidar, and IMU), which enable the perception of environmental information and the construction of an environment map without prior environmental information to achieve autonomous localization. There are two main types of SLAM: laser SLAM and visual SLAM. Visual SLAM can collect a wealth of characteristic environmental information at a low cost, but it is easy to be affected by light and

[1] Tiandi (Changzhou) Automation Co., China Coal Science and Engineering Group Changzhou Research Institute Co., Ltd. Changzhou, 213001, China

[*] corresponding author, e-mail: yjj20002022@163.com

prone to errors and drift. The advantage of laser SLAM is that it is not affected by light changes and can provide accurate depth information, but it is expensive and may have accuracy issues. In the case of fast movement and missing features, it may lead to low accuracy or even failure of the map. Based on the above analysis, it is difficult for a single sensor SLAM system to adapt the construction of three-dimensional maps in large-scale scenarios such as urban environments. To overcome these limitations, sensor fusion, such as using lidar and cameras, has become a future trend in SLAM development. Traditional SLAM assumes stationary objects in the environment to collect information for map construction and localization. At present, map construction in static environments has met the requirements of practical applications. However, most objects in the real environment are moving, and dynamic objects will affect the positioning accuracy of unmanned vehicles, resulting in errors in the constructed maps. In view of the above problems, this paper introduces advanced semantic information to realize the detection and elimination of dynamic objects, thereby improving the localization accuracy of semantic SLAM systems based on LIDAR in dynamic environments. To this end, a semantic SLAM system that integrates advanced semantic information is proposed to detect and eliminate dynamic objects. The integration of sensors such as LIDAR, camera, and IMU in this semantic SLAM system is important for building a high-precision and robust 3D environmental semantic map.

The original data collected by the lidar is preprocessed by distortion correction and ground segmentation, and the image information collected by the camera is semantically segmented using the Dv3p-RS algorithm. Spatio-temporal synchronization is performed on the point cloud data of lidar key frames to realize semantic segmentation. Then a surface element map is established to detect and eliminate dynamic obstacles, extract edge and plane features from the dynamically eliminated point cloud and reduce the time-consuming feature extraction. The semantic information is used to correct the mismatch of features, and improve the inertia of LIDAR. The efficiency and accuracy of the odometer can improve the overall localization accuracy of the algorithm. This, in turn, enables the calculation of the pose transformation relationship through inter-frame matching. After achieving the motion trajectory of the lidar a local semantic map is established. This local semantic map can be added to the global semantic map to create a 3D semantic point cloud map. Fig. 1 shows the semantic SLAM framework. The main work includes the following aspects:

(1) An improved image segmentation algorithm named Dv3p-RS is proposed, which improves the performance of the algorithm by replacing the backbone network and adding the attention module.

(2) A three-dimensional semantic slam framework is proposed. Through adding image semantic information to the LIDAR inertial SLAM system and eliminating dynamic obstacles based on the surface element model, autonomous

localization and mapping in large-scale scenes are realized.

The remaining part is organized as follows: Section 2 presents the related work; Section 3 introduces the Lidar inertial odometer, image semantic segmentation, single-frame point cloud segmentation, and dynamic obstacle removal; Section 4 carries out experimental verification, and Section 5 draws the conclusion.

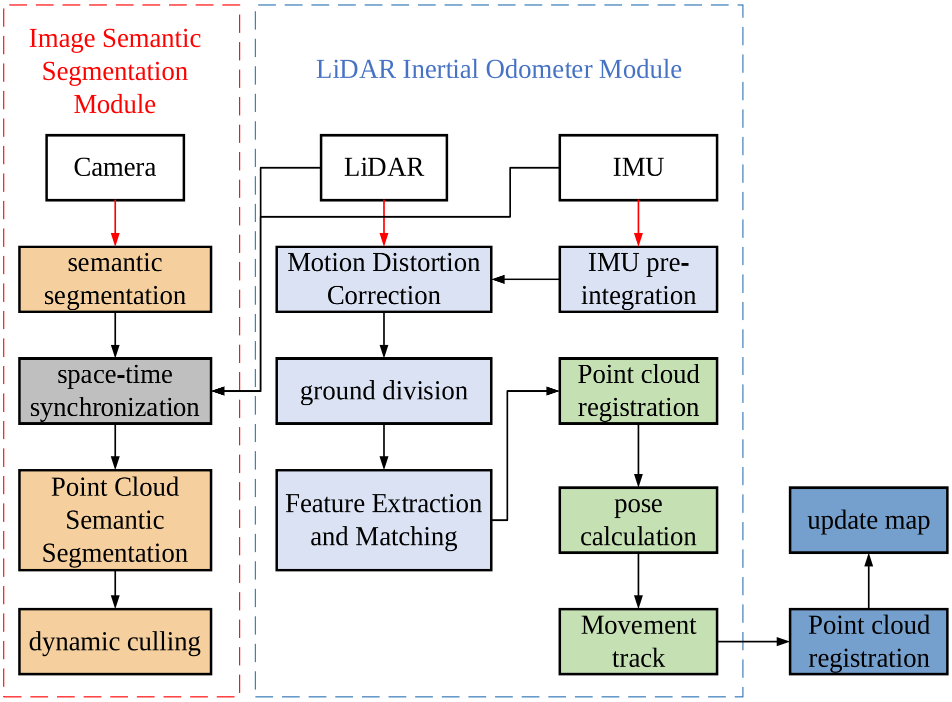


Image Semantic Segmentation Module
LiDAR Inertial Odometer Module
Camera
LiDAR
IMU
semantic segmentation
Motion Distortion Correction
IMU pre-integration
space-time synchronization
ground division
Point cloud registration
Point Cloud Semantic Segmentation
Feature Extraction and Matching
pose calculation
update map
dynamic culling
Movement track
Point cloud registration

Fig. 1. Semantic SLAM system framework

2. Previous Work

Lidar SLAM: The lidar is divided into two-dimensional lidar and three-dimensional lidar based on the number of lines. Two-dimensional laser SLAM can be classified into two types: filter-based method and optimization-based method. Thrun et al. [1] put forward a Fast SLAM based on particle filters, which combines Monte Carlo positioning with low-dimensional Kalman filtering to realize localization and map construction. Grisetti et al. [2] proposed Gmapping, which can effectively overcome the shortcomings of particle filter and use the Maximum Likelihood Estimation Method to improve the quality of sampling, while reducing the number of particles to alleviate the problem of memory explosion. The Cartographer proposed by Google uses correlation scanning matching for violent search at the rough level to avoid local extrema and uses gradient optimization for fine searches and find the optimal solution for linear interpolation. Additionally, it incorporates branch and bound method for loop-closing optimization, effectively eliminating cumulative error caused by frame-to-local sub-image matching [3]. According to the fusion of Lidar and IMU, 3D laser SLAM is divided into two categories: loose coupling and tight coupling. The laser load [4] and [5] belong to the loosely coupled methods. LIO mapping proposed by Ye et al. [6] and LIO-SAM proposed by Shan et al. [7] are both tightly coupled methods. LOAM proposes a novel feature extraction method, which divides feature points into plane points and edge points based on the smoothness of the plane. It narrows the range of feature extraction and proposes an accurate and fast matching method between frames and sub-images. The disadvantage is that there is no loopback detection, which will inevitably cause drift. LeGO-LOAM uses ground for feature point segmentation and point cloud clustering to eliminate noise points, improve the extraction accuracy of feature points, and introduce loop detection to improve the accuracy of localization and mapping. The LIO mapping algorithm proposes a rotation-constrained thinning algorithm, which optimizes all measurements but lacks real-time performance. The LIO-SAM algorithm constructs the odometry, pre-integration, GPS, and loop-closing factors, and uses the factor map to realize tight coupling and global optimization of lidar and IMU. Qi and Guan [8] proposed a real-time 3D positioning method for mechanical working surfaces based on laser SLAM to address the issue of difficulty in meeting the accuracy of the odometer for underground coal mine movement survey. This method uses inertial navigation to eliminate the motion distortion of Lidar and adopts the feature extraction method of principal component analysis. The LM method is used to solve the attitude transformation relationship and realize the attitude estimation.

Image Semantic Segmentation: Long et al. [9] proposed the fully convolutional network (FCN). This method replaces all fully connected layers in the traditional convolutional neural network (CNN) with convolution and restores the image dimension by upsampling. FCN cannot perform real-time reasoning and cannot utilize global context information. Chen [10] proposed the Deeplabv1 algorithm that combines deep learning convolutional neural network (DCNN) with conditional random field (CRF). It can effectively solve the problem of defect location and improve the accuracy of boundary segmentation. Chen et al. [11] proposed the Deeplabv2, which introduced the hollow pyramid pooling (ASPP) based on Deeplabv1. It can improve the segmentation of the network for different scales of targets, but still relies on fully connected conditional random fields [12]. Deeplabv3 introduced the Multi-Grid strategy and optimized ASPP structure, no longer relying on fully connected conditional random fields [13]. Deeplabv3 introduced a multi-grid strategy and optimizes the ASPP structure, no longer relying on fully connected conditional random fields [13].The Deeplabv3+ adopts an encoder-decoder structure. Using the Deeplabv3 network structure as the encoder, the decoder is introduced to obtain clearer segmentation boundaries [14]

Semantic SLAM: Vineet et al. [15] proposed a method based on the combination of hash and conditional random field models. This method evaluates the features extracted by image semantic segmentation through random forest, and uses a densely connected CRF model to reduce the computational burden and construct 3D semantic maps in real time. Combined with classical surface mapping methods, Chen et al. [16] used semantic information to improve the position and pose estimation accuracy of lidar. By using the methods such as flood filling and filtering, it can achieve semantic segmentation and denoising of point clouds. Bojko et al. [17] proposed a self-supervised dynamic elimination SLAM algorithm, which cannot only avoid the negative impact of the system caused by direct recognition, but also eliminate dynamic objects without prior information. Eslamian et al. [18] proposed a semantic map system based on Detectron2 and ORB-SLAM3 algorithm. In ORB-SLAM3, the depth information of feature points is obtained through camera movement, and dynamic points are eliminated by using the semantic information. The results are more accurate than the method using geometric information constraints, but this method is not suitable for outdoor and fast-moving scenes [19].

3. Methods

3.1. Lidar inertial odometer

First, to address the issue of motion distortion in Lidar, the pose transformation within one frame of lidar is obtained through IMU pre-integration, and the laser point coordinates are converted to the first laser point coordinate system to eliminate motion distortion. Second, when the laser radar collects information, a large amount of ground point cloud information will be obtained, which will reduce the operating efficiency of the algorithm. Through calculating the pitch angle of the distance image, the point cloud is divided into two parts: location and non-location, as shown in Fig. 2(a) and Fig. 2(b). Then, using the calculated plane smoothness, these feature points are divided into edge points and plane points, as shown in Fig. 2(c) and Fig. 2(d). Finally, the distance between two objects is obtained through point-to-line and point-to-surface feature matching methods, and then the pose transformation matrix is solved through levenberg-Marquart iteration [20].

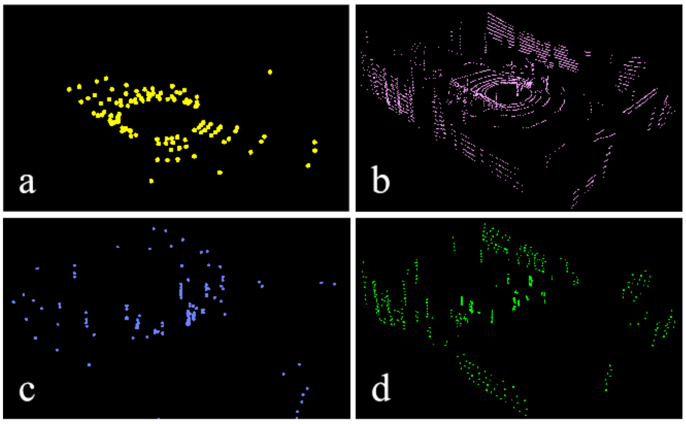


Fig. 2. Point cloud ground segmentation and feature extraction. (a) Location cloud (b) Non-location clouds; (c) Edge feature points; (d) Planar feature points

3.2. Image semantic segmentation

Taking the Deeplabv3+ algorithm as the basic framework of image segmentation, an improved Dv3p-RS algorithm is presented. It is optimized and improved from the following three aspects:

(1) Due to the Xception goal of the Deeplabv3+algorithm backbone network being to train a model that is easy to migrate, computationally efficient, and highly accurate, the research scenario here is an urban environment, aiming to process image information through a semantic segmentation algorithm. It is a laser radar point cloud that provides high-level semantic information. The ResNeXt network [21] has a higher efficiency under the same number of parameters. To this end, a lightweight ResNeXt is used as a feature extraction network to improve model efficiency.

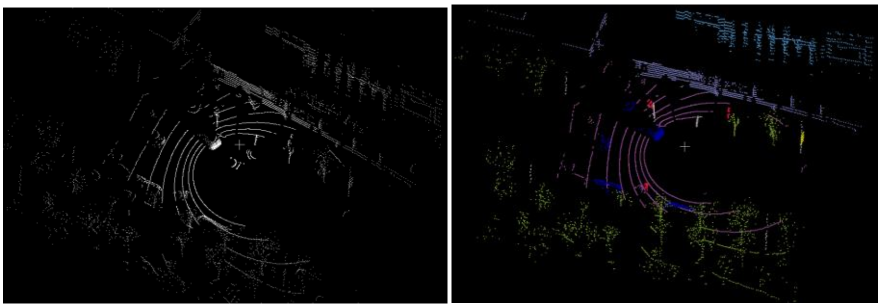
(2) After extracting deep and shallow feature maps from the backbone network, the SE attention module [22] is used to enhance channel characteristics, thereby improving the performance of the model.

(3) The Deeplabv3+ algorithm handles the problem of different dimensions between deep and the shallow feature maps through linear interpolation and up sampling.

However, unmanned vehicles may experience significant scale changes when collecting environmental image information, with many anomalies occurring between pixels. Linear relationship, deconvolution achieves high-precision upsampling through parameter learning. Here deconvolution instead of linear interpolation is used to ensure the accuracy of segmentation.

3.3. Single frame point cloud segmentation

The topic subscription mechanism of the time synchronizer under the robot operating system (ROS) is used to achieve soft synchronization of sensor time. The PTP network protocol synchronization method is used for clock source alignment, thereby realizing the hard synchronization of the sensor time, and obtaining the sensor time through external parameter calibration. Using a rotation matrix and a translation matrix, the coordinate values of each sensor are projected into the same coordinate system, thereby achieving realize spatio-temporal synchronization of sensors. Under spatio-temporal synchronization, the lidar and camera establish a mapping relationship between the point cloud and image, that is, the pixel coordinates corresponding to each point cloud are obtained. After semantic segmentation processing, the 2D images captured by the camera have consistent semantic labels for the same type of objects. The pixel coordinates correspond one-to-one with semantic labels to establish a mapping relationship between the three-dimensional point cloud and semantic tags. Fig. 3(a) is the original point cloud, and Fig. 3(b) is the semantic segmentation point cloud.



(a) Original point cloud; (b) Semantic segmentation point cloud

Fig. 3. Point cloud semantic segmentation

3.4. Dynamic obstacle removal

A surfel is a circular plane with directions and sizes in space. Compared with a point cloud, a surfel contains position information, normal vector information, and area information. The point cloud data with the same normal vectors is described by the same surfel, which cannot only reduce the data storage, but also provide richer geometric information for the data. To be suitable for dynamic environments, we extend the surface elements and define them as follows:

$$surfel' = \{p^s, \vec{n}^s, r^s, t_{ct}, t_{ud}, l_s\} \tag{1}$$

where, t_{ct} represents the creation time of the surfel; t_{ud} represents the update time of the surfel, and l_s is the probability value in logarithmic form, indicating the stability of the surfel. The 3D point cloud scanned by a single-frame lidar is projected into a 2D depth map, assuming a certain laser point$P^s(x, y, z)$.The calculation formula for the coordinate $P_D^s(u,v)$ of the depth map is as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}[1-\arctan(y,x)\cdot\pi^{-1}]\cdot w \\ [1-(\arcsin(z\cdot r^{-1})+f_{\max})f^{-1}]\cdot h \end{pmatrix} \tag{2}$$

$$f = f_{\max} + f_{\min} \tag{3}$$

where, $r = \|P\|_2$ represents the range;f represents the vertical field of view of the lidar; $f_{\max}$ and $f_{\min}$ are the maximum and minimum values of the vertical field of view, respectively;w and h are the width and height of the depth map, respectively. Suppose the center of surface element s is the point $P^{s'}(x',y',z')$, the coordinate $P_D^{s'}(u',v')$ in the depth map, the vectors formed by the adjacent point $P_D^{s'L}$ on the right side of $P_D^{s'}$ and the adjacent point $P_D^{s'U}$ on the upper side of $P_D^{s'}$ are $\vec{n}_L$ and $\vec{n}_U$ respectively.The normal vector of the panel is the outer product of $\vec{n}_L$ and $\vec{n}_U$, and the normal vector $\vec{n}^{s'}$ is calculated as follows:

$$\vec{n}^{s'} = \vec{n}_L \times \vec{n}_U = \left[P_D^{s'L}(u+1,v) - P_D^{s'}(u,v)\right] \times \left[P_D^{s'L}(u,v+1) - P_D^{s'}(u,v)\right] \tag{4}$$

The static objects and dynamic objects are distinguished by detecting the geometric consistency of surface elements. Assuming that there are two adjacent key frames $\mathcal{K}_t$ and $\mathcal{K}_{t+1}$, an object detected in keyframe $\mathcal{K}_t$ is at position $\mathcal{S}_t^L$ in the lidar coordinate system, and mapped in the global. The position in the coordinate system is $\mathcal{S}_t^W$, and the position of the object detected in key frame $\mathcal{K}_{t+1}$ is $\mathcal{S}_t^{L'}$ in the lidar coordinate system. If the position mapped in the global coordinate system is still $\mathcal{S}_t^W$, it indicates that the object is stationary; otherwise, theobjects in different positions are dynamic objects. When a dynamic object is detected, an additional penalty will be given for updating the bin stabilityl_s.The penalty function is as follows:

$$\begin{aligned} l_s^{(t)} = l_s^{(t-1)} &+ \operatorname{odds}\left(p_{\text{stable}} \exp\left(-\frac{\alpha^2}{\sigma_\alpha^2}\right)\exp\left(-\frac{d^2}{\sigma_d^2}\right)\right) \\ &- \operatorname{odds}(p_{\text{prior}}) - \operatorname{odds}(p_{\text{penalty}}) \end{aligned} \tag{5}$$

$$odds(p) = \log(p(1-p)^{-1}) \tag{6}$$

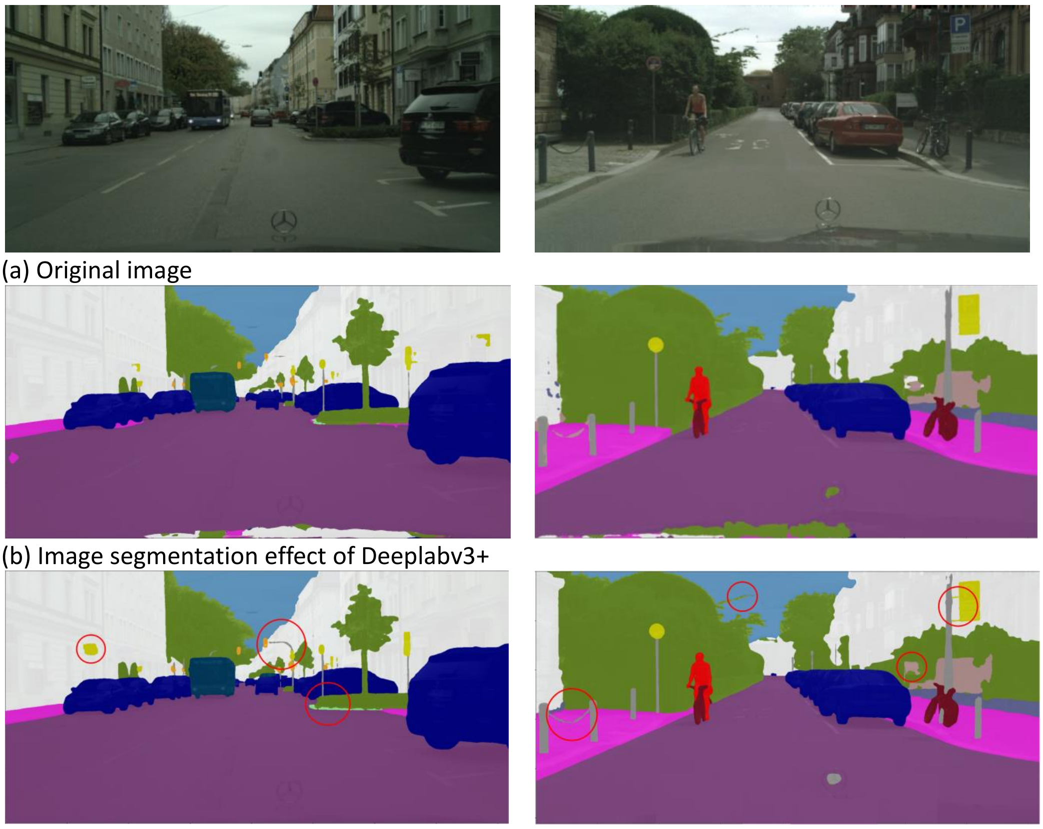
where, p_{stable} is the measured value of the bin;p_{penalty} is the prior probability of the bin; α is the angle between the bin normal vector $\vec{n}^s$ and the measurement normal vector, and d is the distance between the measurement normal vector and the bin. After multiple observations and updates, the stability of the bins belonging to moving objects will become very low. If it is lower than a certain threshold, it will be removed, while the stability of the bins of stationary objects will always be higher than the threshold. Through this method, static objects and dynamic objects can be distinguished, and dynamic objects can be eliminated.

4. Experimental Verification

4.1. Image segmentation network dataset test

The Cityscapes data set [23] is used to train the semantic segmentation network. This data set records image information of 50 urban roads in different seasons and climate environments, including 5000 finely labeled images and 2000 coarse-grained labeled images. It contains rich urban environment information, including 19 labels such as roads, buildings, vegetation and sky.

In the data set, the Deeplabv3+ and the Dv3p-RS proposed in this paper are used to perform image semantic segmentation, and the final segmentation results are shown in Fig. 4. Fig. 4(a) shows the original training image of the dataset; Fig. 4(b) shows the result of semantic segmentation using the Deeplabv3+ algorithm; Fig. 4(c) shows the result of semantic segmentation using the Dv3p-RS.



(a) Original image

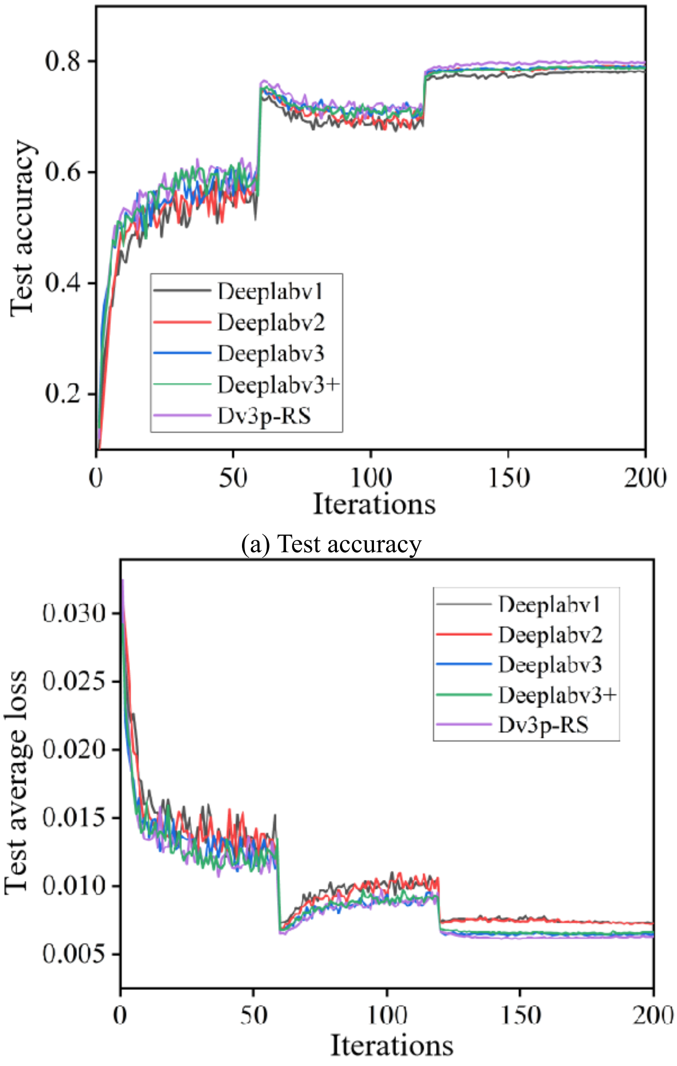
(b) Image segmentation effect of Deeplabv3+

(c) Image segmentation effect of Dv3p-RS

Fig. 4. Comparison of semantic segmentation results

The places marked with the red circle in the middle indicate they are different from the segmentation results of the Deeplabv3+ algorithm. According to the semantic segmentation results, the Dv3p-RS algorithm can efficiently refine the boundaries when segmenting objects such as pedestrians, vehicles, and traffic signs under urban roads. The segmentation network has higher precision performance.

The improvement of the Dv3p-RS algorithm compared with other mainstream semantic segmentation algorithms is quantitatively verified from the perspectives of test accuracy and average test loss, as shown in Fig. 5. Fig. 5 shows that Dv3p-RS has higher test accuracy and lower average loss than other algorithms, which again verifies that the Dv3p-RS algorithm has higher accuracy and efficiency.



(a) Test accuracy

(b) Test Average Loss

Fig. 5. Dataset test results

4.2. Urban road environment test

When collecting information in urban road environment, three original images are randomly selected from the image sequence collected by the camera, as shown in Fig. 6. The original image information is imported into the semantic segmentation network, and the image segmentation results are shown in Fig. 7. The semantic information of the segmented image is mapped to the laser point cloud under the same time stamp, and a local point cloud map with semantic labels is constructed, as shown in Fig. 8. The constructed local 3D semantic point cloud map is added to the global map to realize the construction of the global semantic map, as shown in Fig. 9.

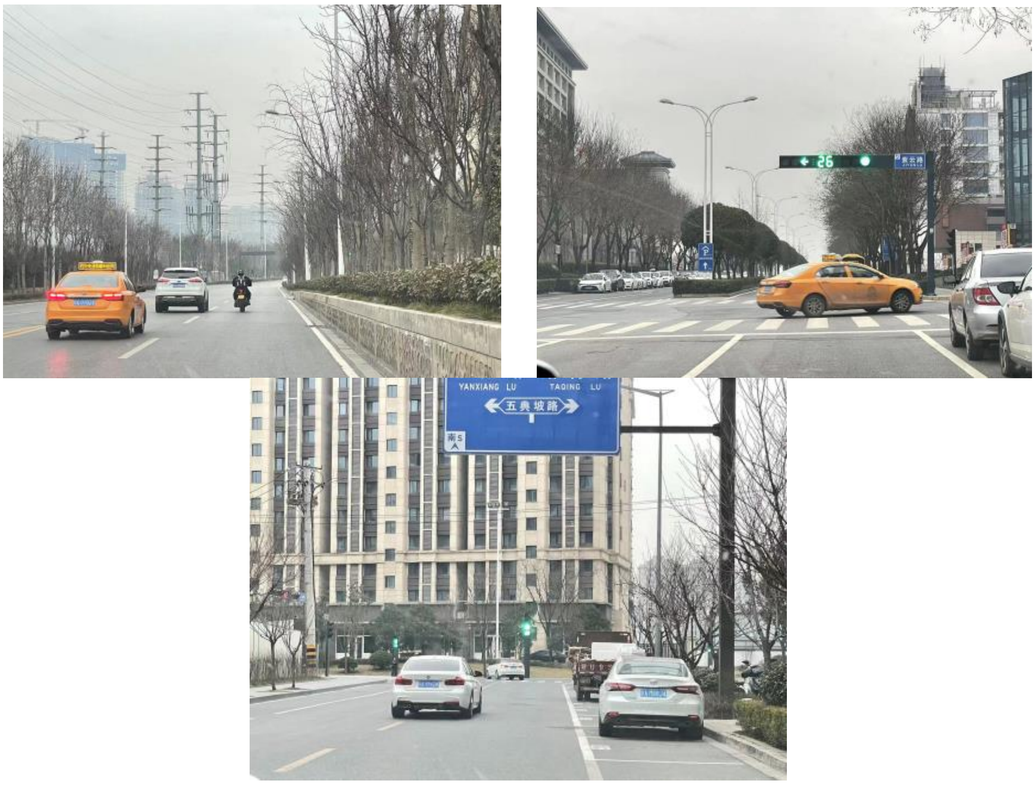


Fig. 6. Camera image

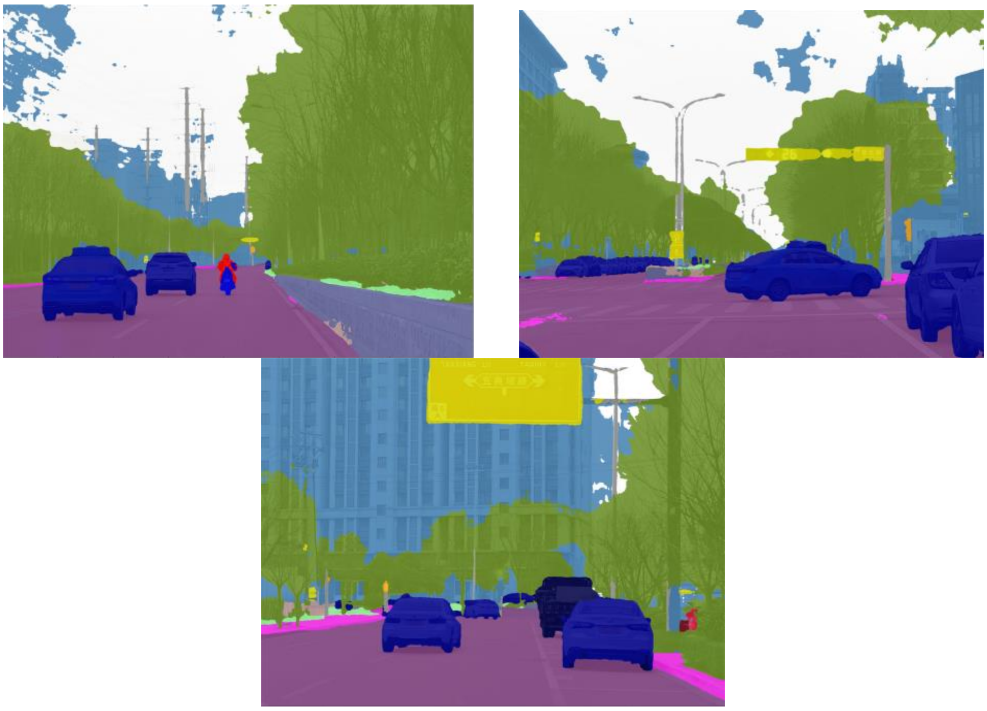


Fig. 7. Semantic segmentation image

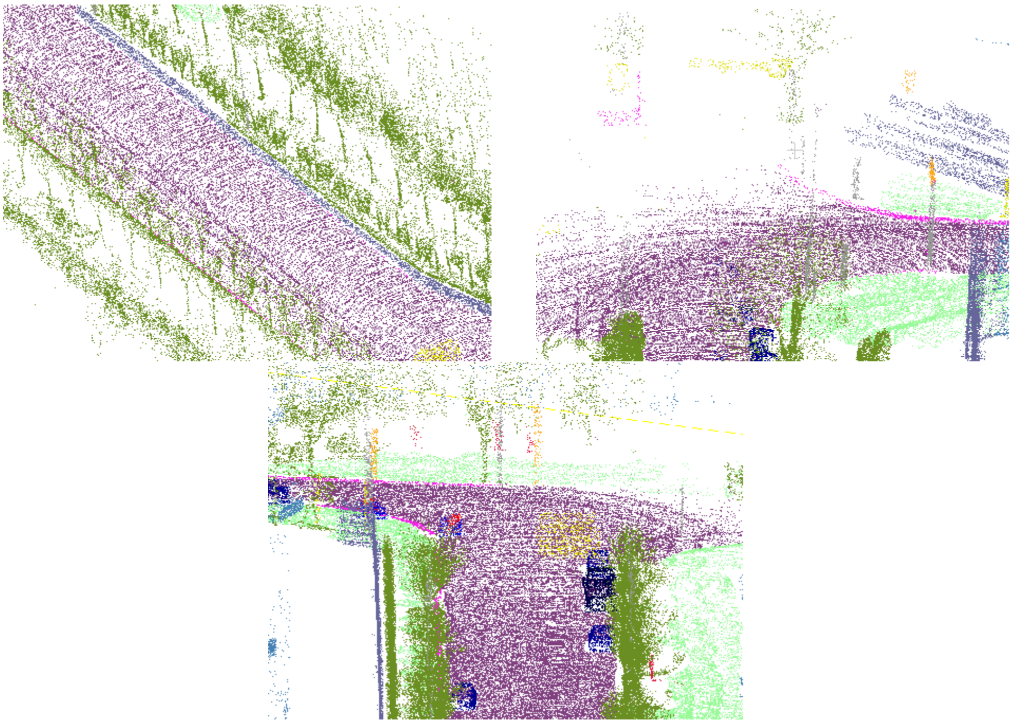


Fig. 8. Semantic point cloud map

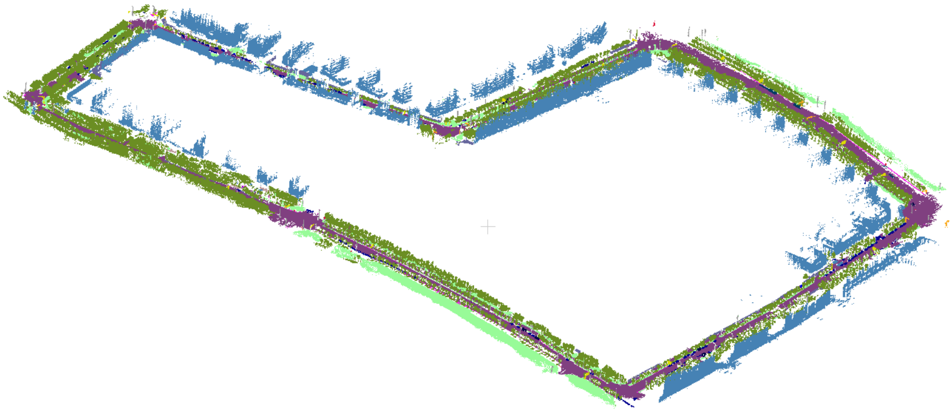
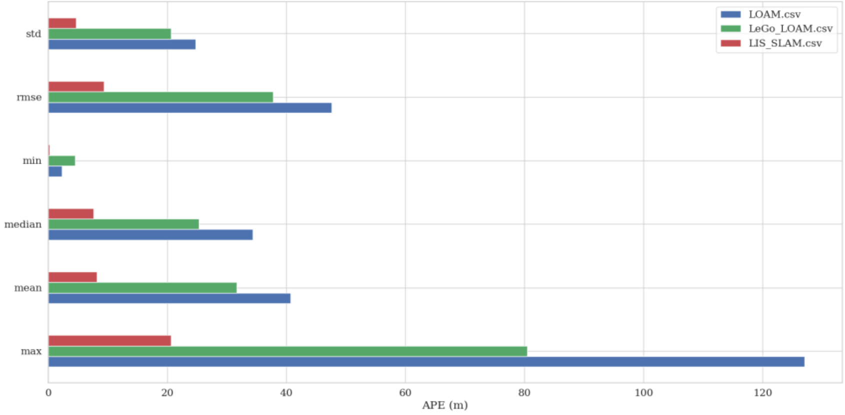
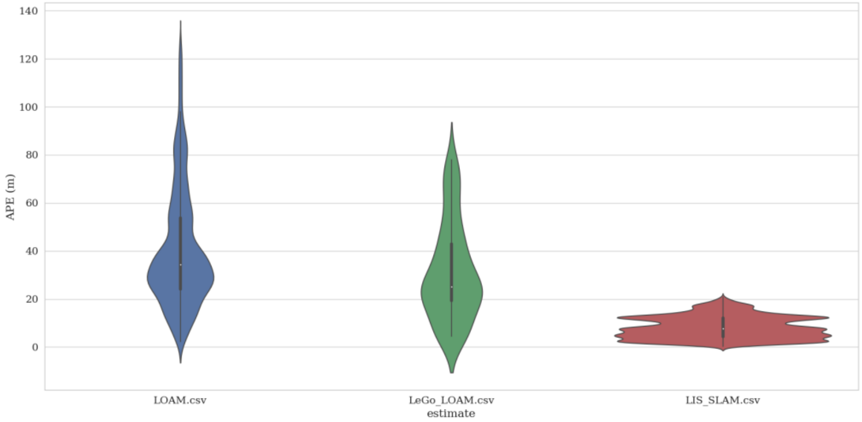


Fig. 9. Semantic map of urban road environment construction

In order to prove that removal of dynamic obstacles can improve the location precision of the system, the experimental verification is carried out in an urban road environment with denser dynamic objects, and compared with the representative algorithms LOAM and LeGo_LOAM of laser SLAM. We use the evaluation index APE (Absolute Position Error) to evaluate the system performance. APE is the calculation of the difference between the estimated position and the known reference position or ground truth. The lower APE indicates that the SLAM system can accurately estimate the absolute position of the camera or robot, while the higher APE indicates that the positioning error is large. The results of error comparison and distribution are shown in Fig. 10. The specific values of error results such as root mean square and standard deviation are shown in Table 1. The error rate of the LSI_SLAM algorithm is 0.42%, which is 1.25% and 0.95% lower than the error rates of the LOAM and the LeGo_LOAM, respectively. This indicates that the LSI_SLAM has higher positioning accuracy and robustness.



(a) APE error comparison chart



(b) APE error distribution map

Fig. 10. Autonomous positioning error of urban road environment

Table 1

Error results in urban road environment

Errors	LOAM	LeGo_LOAM	LIS_SLAM
Root mean square error	53.62	43.81	13.40
mean error	24.75	20.61	4.67
Error rate	1.67%	1.37%	0.42%

5. Conclusions

Combined image semantic segmentation technology with laser inertial ranging technology, a high precision SLAM system in a dynamic environment is presented. First, the laser odometer is constructed by fusing LiDAR and IMU data to estimate the attitude change and position displacement of the mobile robot using the scanned laser point cloud and the inertial measurement information from the IMU. Then the image semantic segmentation algorithm Dv3p-RS with high precision and efficiency is proposed. The single frame point cloud is spatio-temporal synchronized to complete semantic segmentation, and dynamic obstacles are removed through geometric consistency detection using the surface model. The algorithm is tested under an urban traffic environment and compared with LOAM and LeGo_LOAM. The experiment results show that the LIS_SLAM algorithm can achieve high precision in a dynamic environment. Although this study provides some clues for SLAM systems in dynamic environments, there are also some limitations. Specifically, the results of image segmentation will greatly affect the removal effect of dynamic obstacles and also increase additional operation time. Future work will focus on improving the performance of image segmentation models to achieve better rejection results and faster running speed.

Acknowledgement

This study was supported by Jiangsu Province Scientific and Technological Achievements Transformation Project (BA2022040); Science and Technology Innovation Venture Capital Project of Tiandi Technology Co., LTD. (2023-TD-ZD005).

REFERENCES

[1] *Thrun, S., Montemerlo, M., Koller, D., Wegbreit, B., Nieto, J., &Nebot, E.,* Fastslam: An efficient solution to the simultaneous localization and mapping problem with unknown data association. Journal of Machine Learning Research, vol. 4, no. 3, pp. 380-407, 2004.

[2] *Grisetti, G., Stachniss, C., &Burgard, W.,* Improved techniques for grid mapping with rao-blackwellized particle filters. IEEE Transactions on Robotics, vol. 23, no. 1, pp. 34-46, 2007. DOI: 10.1109/TRO.2006.889486

[3] *Montemerlo, M., Thrun, S., Koller, D., &Wegbreit, B.,* FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. IJCAI, vol. 3, no. 2003, pp. 1151-1156, 2003.

[4] *Zhang, J., & Singh, S.,* Low-drift and real-time lidarodometry and mapping. Autonomous Robots, vol. 41, no. 2, pp. 401-416, 2017. DOI:10.1007/s10514-016-9548-2

[5] *Shan, T., &Englot, B.,* Lego-loam: Lightweight and ground-optimized lidarodometry and mapping on variable terrain. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, IEEE, pp. 4758-4765, 2018. DOI: 10.1109/IROS.2018.8594299

[6] *Ye, H., Chen, Y., & Liu, M.,* Tightly coupled 3d lidar inertial odometry and mapping. 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, IEEE, pp. 3144-3150, 2019.DOI: 10.1109/ICRA.2019.8793511

[7] *Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., &Rus, D.,* Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), Las Vegas, NV, USA, IEEE, pp. 5135-5142, 2020. DOI: 10.1109/IROS45743.2020.9341176

[8] *Qi, Y, Guan, S.,* Real-time 3D mapping method of fully mechanized mining face based on laser SLAM. Journal of Mine Automation, vol. 48, no. 11, pp. 139-144, 2022. DOI: 10.13272/j.issn.1671-251x.2022060047

[9] *Long, J., Shelhamer, E., & Darrell, T.,* Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015. DOI: 10.1109/CVPR.2015.7298965

[10] *Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., &Yuille, A. L.,* Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062, 2014. https://doi.org/10.48550/arXiv.1412.7062

[11] *Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., &Yuille, A. L.,*Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834-848, 2017. DOI: 10.1109/TPAMI.2017.2699184

[12] *He, K., Zhang, X., Ren, S., & Sun, J.,* Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904-1916, 2015. Doi: 10.1109/TPAMI.2015.2389824

[13] *Chen, L. C., Papandreou, G., Schroff, F., & Adam, H.,* Rethinking atrous convolution for

semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. https://doi.org/10.48550/arXiv.1706.05587

[14] *Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H.,* Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), pp. 801-818, 2018. https://doi.org/10.48550/arXiv.1802.02611

[15] *Vineet, V., Miksik, O., Lidegaard, M., Nießner, M., Golodetz, S., Prisacariu, V. A., Torr, P. H.,* Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. 2015 IEEE international conference on robotics and automation (ICRA), Seattle, WA, USA, IEEE, pp. 75-82, 2015. DOI: 10.1109/ICRA.2015.7138983

[16] *Chen, X., Milioto, A., Palazzolo, E., Giguere, P., Behley, J., &Stachniss, C.,* Suma++: Efficient lidar-based semantic slam. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, IEEE, pp. 4530-4537, 2019. DOI: 10.1109/IROS40897.2019.8967704

[17] *Bojko, A., Dupont, R., Tamaazousti, M., &Borgne, H. L.,* Self-Improving SLAM in Dynamic Environments: Learning When to Mask. arXiv preprint arXiv:2210.08350, 2022. https://doi.org/10.48550/arXiv.2210.08350

[18] *Eslamian, A., Ahmadzadeh, M.R.,* Det-SLAM: A semantic visual SLAM for highly dynamic scenes using Detectron. arXiv preprint arXiv:2210.00278, 2022. https://doi.org/10.48550/arXiv.2210.00278

[19] *Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., &Tardós, J. D.,* Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, 2021. DOI: 10.1109/TRO.2021.3075644

[20] *Ranganathan A*., The levenberg-marquardt algorithm. Tutoral on LM algorithm, vol. 11, no. 1, pp. 101-110, 2004.

[21] *Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K.,* Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5987-5995, 2017. DOI: 10.1109/CVPR.2017.634

[22] *Hu, J., Shen, L., & Sun, G.,* Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[23] *Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Schiele, B.,* The cityscapes dataset. CVPR Workshop on the Future of Datasets in Vision, vol. 2, 2015.