Autonomous Semantic Mapping for SLAM Systems

Yong He[1], Chi Chen[1]*, Leyi Zhao[1], Yuhang Xu[1], Shangzhe Sun[1], Zongtian Hu[2], Ang Jin[1]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
[2]Institute of Artificial Intelligence, School of Computer Science, Wuhan University

Keywords: Semantic Mapping, Segmentation, SLAM, Multimodal, Real-time.

Abstract

Semantic mapping is crucial for intelligent obstacle avoidance and planning in SLAM systems. We proposed an autonomous semantic mapping approach that integrates multimodal semantic segmentation and SLAM techniques to construct a dense 3D semantic map in real time. Multimodal semantic segmentation based on camera images and LiDAR point clouds is performed in each frame, which assigns image segmentation labels to LiDAR points, generating per-frame 3D semantic information. These segmented frames are then incrementally fused within the SLAM framework to produce a globally consistent semantic map of the environment. The proposed approach is validated through real-world experiments conducted around the Star Lake Building at Wuhan University using the Luo-Jia Explorer system. The experimental results show that our method achieves real-time performance with an inference speed of up to 14Hz on an RTX 4070 GPU, effectively processing sensor data on 10Hz while maintaining high segmentation accuracy in both indoor and outdoor scenarios.

1. Introduction

Simultaneous Localization and Mapping (SLAM) is a widely used framework for autonomous systems to estimate their position while constructing a map of the environment (Esparza and Flores, 2022). SLAM and autonomous exploration play a vital role in empowering unmanned systems to autonomously perceive and comprehend their surroundings.

Semantic segmentation assigns a class label to each point in the input data (Zhuang et al., 2021), allowing systems to interpret and understand the 3D structure of their surroundings. By distinguishing different objects and surfaces, semantic segmentation enhances the accuracy and reliability of SLAM systems (Qian et al., 2020). However, due to the inherent characteristics of point clouds—such as sparsity, uneven density, and occlusions leading to incomplete contours—solely relying on LiDAR data for 3D semantic segmentation presents challenges, particularly in large-scale environments where edge points are prone to misclassification. Since images provide rich textural details while point clouds capture geometric structures, integrating both modalities in multimodal semantic segmentation improves both accuracy and robustness.

Semantic SLAM techniques combine semantic information to filter out dynamic objects, such as pedestrians and vehicles, reducing localization drift and improving long-term mapping consistency. Additionally, the enriched map representation enables more intelligent decision-making, such as identifying navigable areas, classifying obstacles, and supporting high-level reasoning in robotics applications. However, the semantic SLAM methods often require significant computational resources due to the high complexity of deep learning-based semantic segmentation and data fusion, posing challenges on resource-constrained devices.

We proposed a lightweight autonomous semantic mapping approach that combines multimodal semantic segmentation with SLAM to generate a dense and globally consistent 3D semantic map. In the multimodal semantic segmentation module, image semantic segmentation is first performed and then LiDAR points are projected onto the corresponding image to assign semantic labels. The image semantic segmentation networks utilized in the multimodal segmentation module are Mask2Former (Cheng et al., 2022) based on experimental evaluation. The per-frame 3D segmentation results are then integrated into the entire scene through SLAM, ensuring accurate and real-time semantic mapping.

To validate its effectiveness, experiments were conducted around the Star Lake Building at Wuhan University using the Luo-Jia Explorer system (Wu et al., 2024). The results demonstrate real-time performance, achieving an inference speed of up to 14Hz on an RTX 4070 GPU while processing data at 10Hz, maintaining high mapping accuracy in both indoor and outdoor environments.

2. Related Work

2.1 Multimodal Semantic Segmentation

2D semantic segmentation is primarily deep learning-based methods (Guo et al., 2023). Early techniques relied on Convolutional Neural Networks (CNNs) for feature extraction (Chen et al., 2017; Ronneberger et al., 2015). However, with the introduction of the Transformer architecture (Vaswani et al., 2017), many segmentation models have adopted Transformer-based networks, achieving state-of-the-art results. The Swin Transformer (Liu et al., 2021) is widely used as a backbone for feature extraction, and some studies have integrated it into segmentation framework to enhance feature representation, such as the dual-encoder Swin Transformer Embedding U-Net (He et al., 2022). Mask2Former (Cheng et al., 2022), a unified segmentation model that employs masked attention to focus on local features, has demonstrated strong performance across multiple datasets.

3D semantic segmentation methods can be broadly classified into single-modal and multimodal approaches. Single-modal methods rely solely on point clouds and are further divided into point-based (Charles et al., 2017; Qi et al., 2017) and voxel-based techniques (Poux and Billen, 2019; Zhou et al., 2020). For multimodal 3D segmentation, LiDAR-camera fusion techniques enhance segmentation accuracy by leveraging both geometric and texture information. Fusion strategies are categorized into early-fusion, deep-fusion, and late-fusion approaches (Huang et al., 2022). Early-fusion methods (Meyer et al., 2019) integrate LiDAR and image data at the input or feature level, preserving both texture and geometric details while maintaining computational efficiency. Deep-fusion techniques (Huang et al., 2020; Zhang et al., 2023) extract features independently from each modality before merging them at a later stage, while late-fusion approaches (Pang et al.,

2020) process each modality separately and combine predictions at the decision-making stage. Deep-fusion and late-fusion require independent feature extraction for each modality, resulting in high computational costs. In contrast, early-fusion methods extract features only from images, making them more suitable for real-time processing and commonly used in SLAM applications (Li et al., 2020; Liu and Miura, 2021).

Our proposed method aligns with early-fusion approaches, providing a lightweight multimodal segmentation solution compatible with various SLAM systems. This approach achieves real-time, high-precision segmentation without requiring expensive hardware, making it practical for real-world applications.

2.2 Semantic SLAM

Nowadays, single-robot SLAM has made significant progress (Cong et al., 2024), classical methods such as FAST-LIO (Xu and Zhang, 2021) and ORB-SLAM (Mur-Artal et al., 2015) have been able to carry out high-precision 3D mapping. Learning-based methods further improve the performance of SLAM methods. For example, DROID-SLAM (Teed and Deng, n.d.) utilizes the advantages of RNN in processing image sequence and time data, iteratively updates camera attitude and depth information, thus to significantly improve the performance in complex environments. However, the lack of semantic information for SLAM remains a challenge.

Many studies integrating semantic information into SLAM methods (Chen et al., 2022), with a typical target to remove dynamic objects. SuMa++ (Chen et al., 2019) utilizes semantic information to filter dynamic objects, performing remarkably well in dynamic environments. YOLO-SLAM (Wu et al., 2022) integrates a lightweight Darknet19-YOLOv3 network with depth-based geometric constraints, thus effectively identifies dynamic object feature points. Combining semantic and geometric info removes moving object interference, Blitz-SLAM (Fan et al., 2022) enables accurate localization and clean mapping in dynamic scenes. Semantic information is also used in Collaborative SLAM, the fully distributed Kimera-Multi (Tian et al., 2022) system, leveraging visual-inertial sensors, is capable of constructing precise metric-semantic 3D meshes. SlideSLAM (Liu et al., 2024) presents a decentralized, real-time metric-semantic SLAM system, utilizing sparse object-level semantic maps to enhance inter-robot loop closures and facilitate seamless cooperation.

Our method integrates 3D semantic Segmentation and Semantic SLAM, achieves real-time, lightweight performance, enabling efficient environment understanding as well as low-resource deployment.

3. System Overview

The overall framework is illustrated in Figure. 1. The process begins with multimodal semantic segmentation for frame t, which utilizes camera images and LiDAR point clouds captured by the Luo-Jia Explorer system simultaneously. 2D semantic segmentation is applied on images first. Then, within the multimodal fusion process, each LiDAR point is projected onto the corresponding image pixel based on sensor calibration parameters, and the pixel's semantic label is assigned to the point, producing per-frame 3D semantic information. The SLAM module further integrates LiDAR and IMU data to estimate the system's pose and build a globally consistent map. By incrementally fusing the per-frame segmentation results using SLAM, a dense 3D semantic map of the environment is ultimately obtained.

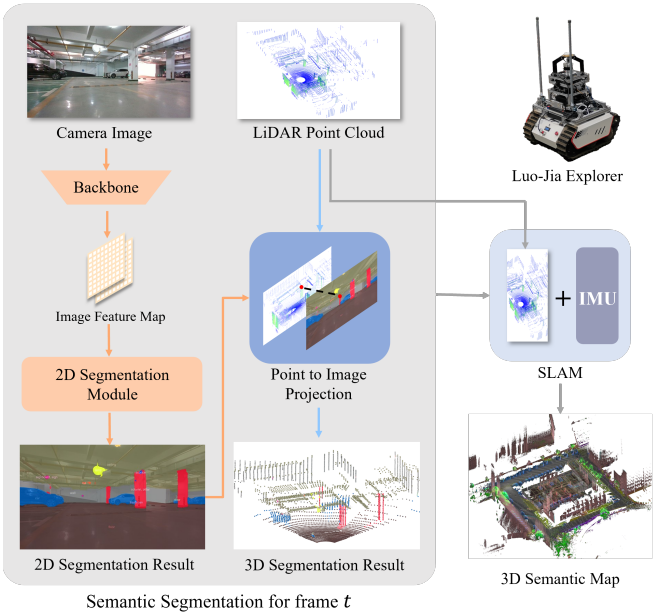

Figure 1. Overall framework of the semantic mapping approach.

3.1 Multimodal Semantic Segmentation

The multimodal semantic segmentation module employs both camera images and LiDAR point clouds, which are captured simultaneously by the Luo-Jia Explorer system. This process involves two main stages: 2D semantic segmentation and point-to-image projection, which together generate point cloud with semantic information.

The 2D semantic segmentation is driven by deep learning techniques that first extract features from the image using a backbone network, which produces image feature maps. A decoder then processes the feature map to generate the 2D segmentation results. Based on the performance and effectiveness, two 2D semantic segmentation models, Mask2Former and Tube-Link, were initially selected. The introduction of these two models is as follows.

Mask2Former is an enhanced version of MaskFormer (Cheng et al., 2021), replacing cross-attention with masked attention under the assumption that local features can represent most of the relevant information. Additionally, a pixel decoder generates multi-scale features to improve small object segmentation. Tube-Link (Li et al., n.d.) is a video segmentation model that partitions input video into sub-clips and applies contrastive learning with self-attention to correlate object queries and their corresponding masks across frames. Both models have been trained with different backbone networks, and the final choice will be made based on the experimental results, which are discussed in Section IV.

After obtaining the segmentation mask, each LiDAR point is mapped onto the corresponding image pixel to assign semantic labels and construct a per-frame 3D semantic representation. Assuming the current time step is t, the preprocessed point cloud is P_t, which is filtered beyond a predefined depth range, and each point is represented as (X, Y, Z). Using the rotation matrix R and the translation vector T calculated by the calibrated extrinsic parameters, the LiDAR coordinates are transformed into the camera coordinate system, resulting in the transformed coordinates (X_c, Y_c, Z_c). The coordinate conversion formula is shown in Equation (1).

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + T \quad (1)$$

These transformed coordinates are then projected onto a normalized plane to acquire the normalized coordinates (x_n, y_n), as shown in Equation (2). Distortion correction is applied using the distortion coefficients, and the corrected normalized coordinates (x_{nd} , y_{nd}) are obtained for image coordinates calculation.

$$x_n = \frac{X_c}{Z_c}, \; y_n = \frac{Y_c}{Z_c} \quad (2)$$

Using the provided camera intrinsic parameters, the camera matrix is formulated. This matrix includes four key parameters: f_x and f_y, which denote the focal lengths, and c_x and c_y, representing the optical center coordinates. Finally, the corrected normalized coordinates are mapped to the image coordinate system to determine the image coordinates (u, v), as expressed in Equation (3).

$$u = f_x \cdot x_{nd} + c_x, \; v = f_y \cdot y_{nd} + c_y \quad (3)$$

After projection, points within the camera's field of view are assigned segmentation labels based on their corresponding pixel coordinates in the mask. This results in a colored point cloud representing per-frame 3D semantic information.

Figure. 2 illustrates the point cloud segmentation result at a specific moment, while the full 360-degree point cloud is in the left and the segmented points within the camera's field of view highlighted on the full point cloud is in the right, overlaid with semantic colors. The yellow box marks the region where the segmented points are mapped on the original point cloud. It can be seen that the semantic segmentation is accurate.

Although single-frame segmentation is limited to the camera's field of view, SLAM integrates data collected along a closed-loop trajectory to construct a dense and globally consistent 3D semantic map of the environment (Chen et al., 2024).

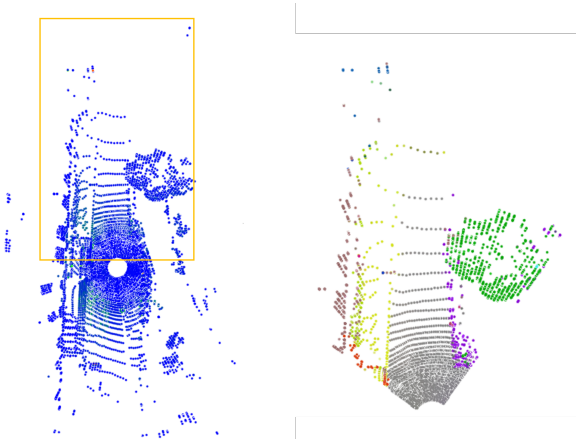

Figure 2. The original point cloud and the colored point cloud within camera's frustum.

3.2 SLAM

In this study, single-robot SLAM is utilized for semantic mapping, leveraging LiDAR and IMU data for accurate localization and map construction. The system employs LiDAR-based SLAM to generate a globally consistent point cloud representation, which forms the basis for 3D semantic mapping. To improve localization accuracy, IMU measurements are integrated to compensate for motion during data acquisition. The system fuses LiDAR and IMU data using an extended Kalman filter for state estimation, ensuring robust trajectory tracking and mapping.

As shown in Figure. 3, as a portion of the experimental scene, the SLAM trajectory in yellow and the corresponding registered point cloud are presented. The seamless alignment of per-frame point cloud demonstrates the accuracy of SLAM in both localization and map construction. After semantic labeling, the colored point cloud forms the 3D semantic map.

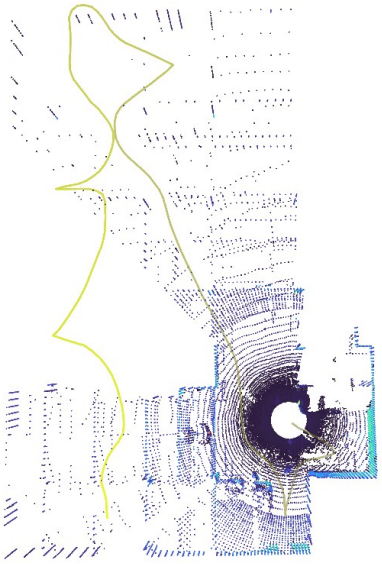

Figure 3. The SLAM trajectory and the corresponding registered point cloud.

4. Experimental Results

4.1 System Configuration and Data Collection

The Luo-Jia Explorer system consists of multiple unmanned ground vehicles (UGVs), among which UGV1 was deployed in the experiment. The vehicle is equipped with an Ouster OS1-128 LiDAR, an Intel RealSense D455 depth camera, and an IMU. The camera supports multiple resolutions, with 640×480 and 1280×720 being selected for this experiment. To facilitate efficient computation, the UGV is integrated with two GPUs, a GeForce RTX 2060 and an RTX 4070. Additionally, inference speed comparisons were conducted on external RTX 3090 and RTX 4060 GPUs to evaluate the performance of different segmentation models and backbone architectures.

To assess system performance, real-world data collection was performed around the Star Lake Building at Wuhan University, covering both an underground parking garage and surrounding outdoor areas. A key challenge encountered during data acquisition was the significant change in lighting conditions when transitioning from indoor to outdoor environments, requiring robust adaptation to illumination variations. The dataset was recorded at 10 Hz, capturing synchronized LiDAR point clouds and camera images, ultimately yielding 5,957 frames.

4.2 Multimodal Semantic Segmentation Results

The segmentation results produced by the two models are visualized in Figure. 4. Specifically, Figure. 4(b) and Figure. 4(c) depict results from the underground garage, while Figure. 4(e) and

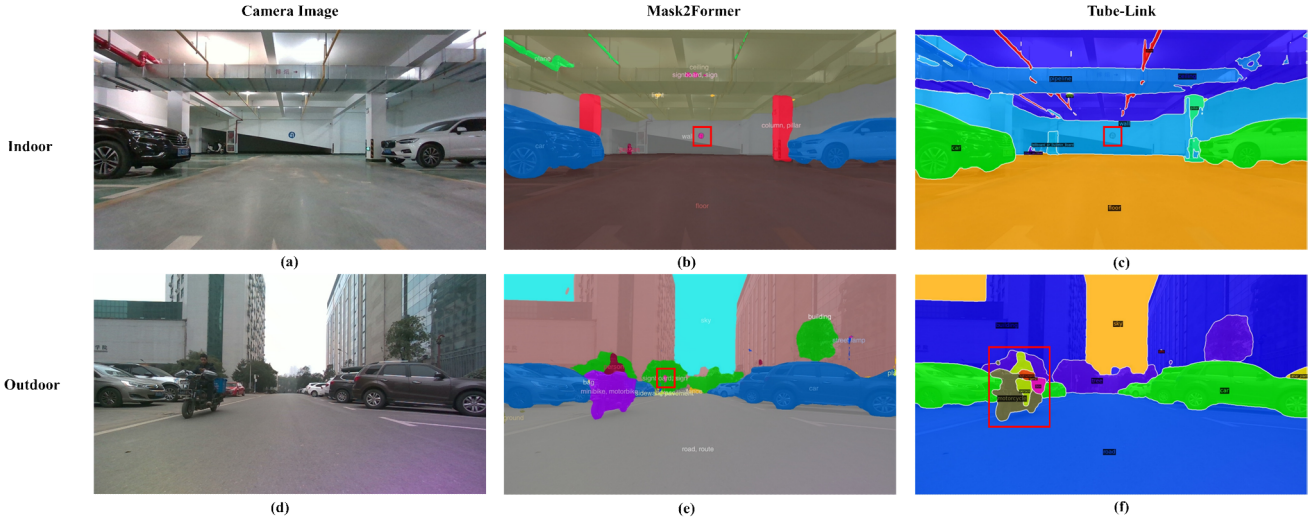

Figure 4. The camera images and segmentation results of Mask2Former and Tube-Link in indoor and outdoor scenes.

Figure. 4(f) illustrate the outdoor environment. In the indoor scenario, an object enclosed in the red box (identified as a signboard in Figure. 4(a)) is correctly segmented by Mask2Former, whereas Tube-Link fails to detect it. Similarly, in the outdoor environment, Figure. 4(e) shows that Mask2Former successfully recognizes a distant signboard, whereas Figure. 4(f) highlights Tube-Link's strength in identifying nearby objects, particularly a person riding a motorcycle, with well-defined segmentation contours. However, Tube-Link struggles with smaller objects.

To further analyze performance across consecutive frames, segmentation results were compiled into video sequences. Both models demonstrated instability in recognizing upper structures in the underground garage, frequently misclassifying categories. Additionally, a sudden category shift was observed in road segmentation in outdoor scenes. Despite these inconsistencies, the models achieved reliable performance across most object categories.

To evaluate the computational efficiency of the two semantic segmentation models, inference speeds under different configurations were recorded, as summarized in Table I. Since the projection algorithm executes within 0.01 seconds, the overall processing speed is predominantly influenced by the segmentation stage. Additionally, the results in Table I include the time required for visualization, which contributes to nearly one-sixth of the total processing time.

As indicated in Table I, when using the Swin-l backbone, Mask2Former significantly outperforms Tube-Link in processing 1280×720 resolution images. For a lower resolution of 640×480, Mask2Former with the Swin-t backbone achieves an inference speed of 14 Hz on RTX 4070, and this could potentially reach 17 Hz if visualization is excluded, while the inference speed will be lower on GPU such as the 3090 or 2090. Given its superior balance of accuracy and efficiency, Mask2Former with a Swin-t backbone was selected for deployment on the Luo-Jia Explorer system, supporting the recognition of 124 object categories.

Figure. 5 presents a 3D semantic segmentation result with the corresponding image segmentation mask for a specific frame. In Figure. 5(a), the colored point cloud within the camera's view frustum is depicted from multiple viewpoint. The front-facing perspective of the colored point cloud is provided in Figure. 5(b). A comparison with the segmentation mask in Figure. 5(c) reveals a high degree of projection accuracy, confirming the effectiveness of the method.

TABLE I. THE INFERENCE SPEED OF DIFFERENT MODELS AND DIFFERENT CONFIGURATIONS (WITH VISUALIZATION)

GPU	Mask2Former		Tube-Link
	640×480 (Swin-t)	*1280×720*	*1280×720*
2060	5Hz	-	-
3090	-	5Hz (Resnet-50) 4.5Hz (Swin-t) 2.5Hz (Swin-l)	8Hz (Resnet-50) - 0.6Hz (Swin-l)
4060	10Hz	5Hz (Swin-t)	-
4070	14Hz	-	-

By aggregating segmentation results across multiple frames, a complete 3D semantic map of the environment is constructed, as illustrated in Figure. 6. The zoomed-in visualization highlights various segmented objects, such as vehicles (blue), trees (green), and structural elements like pillars (red) in the underground garage, as well as indoor ceilings. The segmentation accuracy for these categories appears visually reliable. Additionally, the proposed mapping method enables real-time map construction. The semantic mapping approach provides valuable contributions to the reconstruction of both indoor and outdoor environments.

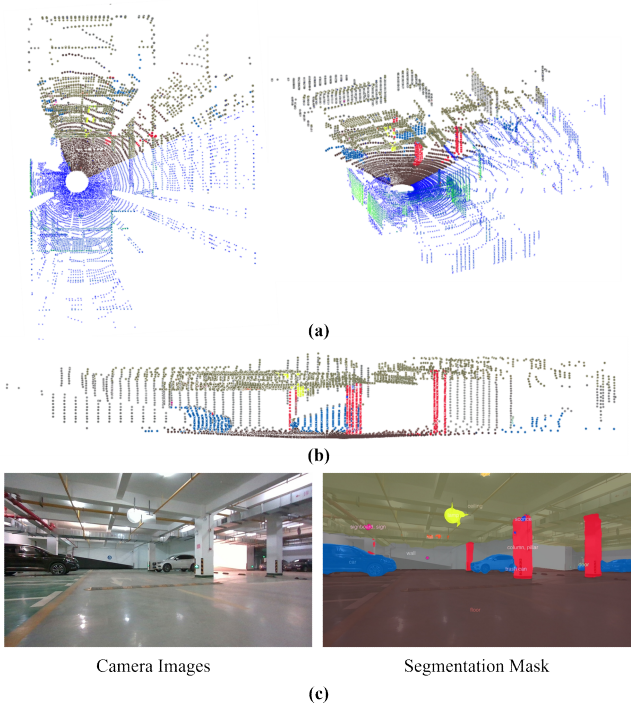

Figure 5. The projected colored point cloud displayed with the overall point cloud and the image segmentation mask.

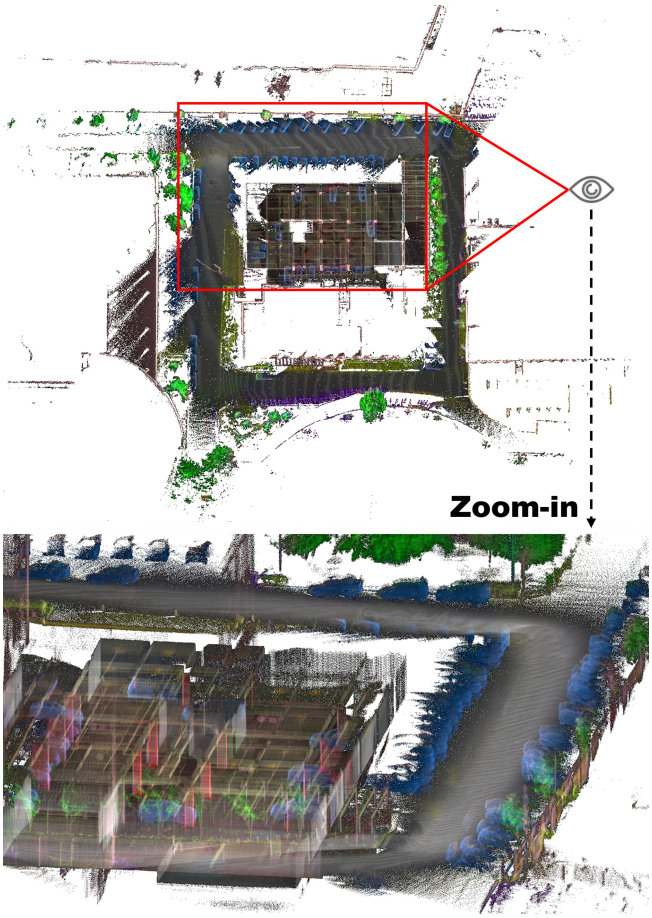

Figure 6. The 3D semantic segmentation map of the entire scene.

5. Conclusion

Semantic mapping plays a vital role in enhancing obstacle avoidance and path planning in autonomous SLAM systems. We proposed an autonomous semantic mapping approach that combines multimodal semantic segmentation with SLAM techniques to construct a dense and globally consistent 3D semantic map in real time. The experiments conducted around the Star Lake Building at Wuhan University using the Luo-Jia Explorer system validate the effectiveness of this approach. The results demonstrate that the method meets real-time performance requirements, achieving an inference speed of up to 14Hz on an RTX 4070 GPU while processing data at 10Hz. The generated 3D semantic maps maintain high segmentation accuracy across both indoor and outdoor environments. Future work will focus on accelerating the mapping process by optimizing SLAM efficiency and improving the real-time integration of semantic information. Additionally, enhancing the robustness of mapping in dynamic environments, such as handling moving objects and illumination changes, will be explored to improve the adaptability of the system in diverse real-world scenario.

References

Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 77–85. https://doi.org/10.1109/CVPR.2017.16

Chen, C., Jin, A., Wang, Z., Zheng, Y., Yang, B., Zhou, J., Xu, Y., Tu, Z., 2024. SGSR-Net: Structure Semantics Guided LiDAR Super-Resolution Network for Indoor LiDAR SLAM. IEEE Trans. Multimed. 26, 1842–1854. https://doi.org/10.1109/TMM.2023.3289752

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.

Chen, W., Shang, G., Ji, A., Zhou, C., Wang, X., Xu, C., Li, Z., Hu, K., 2022. An Overview on Visual SLAM: From Tradition to Semantic. Remote Sens. 14, 3010. https://doi.org/10.3390/rs14133010

Chen, X., Milioto, A., Palazzolo, E., Giguere, P., Behley, J., Stachniss, C., 2019. SuMa++: Efficient LiDAR-based Semantic SLAM, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Presented at the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Macau, China, pp. 4530–4537. https://doi.org/10.1109/iros40897.2019.8967704

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention Mask Transformer for Universal Image Segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation, in: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 17864–17875.

Cong, Y., Chen, C., Yang, B., Zhong, R., Sun, S., Xu, Y., Yan, Z., Zou, X., Tu, Z., 2024. OR-LIM: Observability-aware robust LiDAR-inertial-mapping under high dynamic sensor motion. ISPRS J. Photogramm. Remote Sens. 218, 610–627. https://doi.org/10.1016/j.isprsjprs.2024.09.036

Esparza, D., Flores, G., 2022. The STDyn-SLAM: A Stereo Vision and Semantic Segmentation Approach for VSLAM in Dynamic Outdoor Environments. IEEE Access 10, 18201–18209. https://doi.org/10.1109/ACCESS.2022.3149885

Fan, Y., Zhang, Q., Tang, Y., Liu, S., Han, H., 2022. Blitz-SLAM: A semantic SLAM in dynamic environments. Pattern Recognit. 121, 108225. https://doi.org/10.1016/j.patcog.2021.108225

Guo, Y., Nie, G., Gao, W., Liao, M., 2023. 2D Semantic Segmentation: Recent Developments and Future Directions. Future Internet 15, 205. https://doi.org/10.3390/fi15060205

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2022.3144165

Huang, K., Shi, B., Li, Xiang, Li, Xin, Huang, S., Li, Y., 2022. Multi-modal Sensor Fusion for Auto Driving Perception: A Survey.

Huang, T., Liu, Z., Chen, X., Bai, X., 2020. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. Comput. Vis. – ECCV 2020 35–52. https://doi.org/10.1007/978-3-030-58555-6_3

Poux, F., & Billen, R, 2019. Voxel-based 3D point cloud semantic segmentation: Unsupervised geometric and relationship featuring vs deep learning methods. ISPRS International Journal of Geo-Information, 8(5), 213.

Li, F., Chen, W., Xu, W., Huang, L., Li, D., Cai, S., Yang, M., Xiong, X., Liu, Y., Li, W., 2020. A Mobile Robot Visual SLAM System With Enhanced Semantics Segmentation. IEEE Access 8, 25442–25458. https://doi.org/10.1109/ACCESS.2020.2970238

Li, X., Yuan, H., Zhang, W., Cheng, G., Pang, J., Loy, C.C., n.d. Tube-Link: A Flexible Cross Tube Framework for Universal Video Segmentation.

Liu, X., Lei, J., Prabhu, A., Tao, Y., Spasojevic, I., Chaudhari, P., Atanasov, N., Kumar, V., 2024. SlideSLAM: Sparse, Lightweight, Decentralized Metric-Semantic SLAM for Multi-Robot Navigation. https://doi.org/10.48550/arXiv.2406.17249

Liu, Y., Miura, J., 2021. RDS-SLAM: Real-Time Dynamic SLAM Using Semantic Segmentation Methods. IEEE Access 9, 23772–23785. https://doi.org/10.1109/ACCESS.2021.3050617

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Meyer, G.P., Charland, J., Hegde, D., Laddha, A., Vallespi-Gonzalez, C., 2019. Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation. Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE. https://doi.org/10.1109/cvprw.2019.00162

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Trans. Robot. 31, 1147–1163. https://doi.org/10.1109/tro.2015.2463671

Pang, S., Morris, D., Radha, H., 2020. CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Qian, Z., Patath, K., Fu, J., Xiao, J., 2020. Semantic SLAM with Autonomous Object-Level Data Association.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Teed, Z., Deng, J., n.d. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras.

Tian, Y., Chang, Y., Herrera Arias, F., Nieto-Granda, C., How, J.P., Carlone, L., 2022. Kimera-Multi: Robust, Distributed, Dense Metric-Semantic SLAM for Multi-Robot Systems. IEEE Trans. Robot. 38, 2022–2038. https://doi.org/10.1109/TRO.2021.3137751

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz, Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Wu, W., Chen, C., Yang, B., 2024. LuoJia-Explorer: Unmanned Collaborative Localization and Mapping System, in: Qu, Y., Gu, M., Niu, Y., Fu, W. (Eds.), Proceedings of 3rd 2023 International Conference on Autonomous Unmanned Systems (3rd ICAUS 2023). Springer Nature, Singapore, pp. 66–75. https://doi.org/10.1007/978-981-97-1099-7_7

Wu, W., Guo, L., Gao, H., You, Z., Liu, Y., Chen, Z., 2022. YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint. Neural Comput. Appl. 34, 6011–6026. https://doi.org/10.1007/s00521-021-06764-3

Xu, W., Zhang, F., 2021. FAST-LIO: A Fast, Robust LiDAR-Inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter. IEEE Robot. Autom. Lett. 6, 3317–3324. https://doi.org/10.1109/lra.2021.3064227

Zhang, Zhiwei, Zhang, Zhizhong, Yu, Q., Yi, R., Xie, Y., Ma, L., 2023. LiDAR-Camera Panoptic Segmentation via Geometry-Consistent and Semantic-Aware Alignment, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Paris, France, pp. 3639–3648. https://doi.org/10.1109/ICCV51070.2023.00339

Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., Lin, D., 2020. Cylinder3D: An Effective 3D Framework for Driving-scene LiDAR Semantic Segmentation. https://doi.org/10.48550/arXiv.2008.01550

Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., Tan, M., 2021. Perception-Aware Multi-Sensor Fusion for 3D LiDAR Semantic Segmentation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 16260–16270. https://doi.org/10.1109/ICCV48922.2021.01597