UAV-Based Collaborative Mapping Framework with Environmental Semantic Extraction

Cai, Haonan[1]; Zhong, Xuanke[1]; Zhou*, Baoding[1]

[1]College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China
– 2310474132@email.szu.edu.cn, 2210474161@email.szu.edu.cn, baodingzhou@szu.edu.cn

Keywords: UAV, SLAM, Collaborative Mapping, 3D LiDAR, Environmental Semantics

Abstract

The evolving low-altitude economy enables Unmanned Aerial Vehicles (UAVs) to gather diverse sensor data, including RGB images, 3D point clouds, and inertial measurements, offering untapped potential for environmental mapping. Traditional urban 3D modeling methods often face delays in updates and scalability issues. This paper introduces a novel UAV-based collaborative mapping framework that integrates heterogeneous data from multiple UAVs to efficiently reconstruct dynamic urban environments. The framework employs advanced visual recognition and optical character recognition (OCR) for semantic feature extraction, complemented by LiDAR inertial odometry for precise map construction. To address challenges posed by sparse LiDAR data in indoor settings, a temporal alignment mechanism is employed to generate synchronized keyframes, enhancing data coherence. Additionally, camera-LiDAR calibration combined with cross-modal registration and semantic-guided point cloud stitching boosts system robustness. Experimental results demonstrate that feature-guided point cloud registration, bolstered by semantic alignment, surpasses traditional methods, achieving efficient mapping and offering a scalable solution for urban 3D modeling, applicable in real-time urban planning and smart city development.

1. INTRODUCTION

Accurate and up-to-date 3D models of urban environments are crucial for various applications such as urban planning, smart city development and so on. However, the rapid pace of urbanization necessitates frequent updates to these models to maintain their relevance and accuracy(Goodchild, 2007). Traditional methods for constructing and updating 3D models face significant challenges related to data acquisition, processing time, and scalability. The continuous development of the low-altitude economy is transforming the capabilities of Unmanned Aerial Vehicles (UAVs), enabling them to gather diverse types of sensor data such as RGB images, 3D point clouds, and inertial measurements. This evolution presents significant opportunities for constructing dynamic and high-precision 3D models of urban environments.

Current methods for obtaining data for 3D models largely rely on satellite imagery, aerial remote sensing, or ground-based data collection techniques. While effective, these approaches are often prohibitively expensive, difficult to scale, and incapable of providing timely updates. Conventional model update methods typically involve full reconstruction techniques, which are time-consuming, failing to ensure the freshness and precision required to keep pace with the dynamic nature of urban environments (Nießner et al., 2013). The limitations of traditional methods highlight the need for more efficient, scalable, and timely solutions for 3D model construction and updating.

Leveraging collaborative mapping techniques offers a promising solution to these challenges. This approach can significantly reduce the cost and time associated with traditional data collection methods while enabling more efficient utilization of available resources. Current research on collaborative mapping predominantly targets applications in autonomous vehicles. These systems are typically designed to operate at ground level, which poses significant challenges when attempting to capture detailed information about structures at higher elevations, such as tall buildings or elevated infrastructure. This limitation restricts the comprehensiveness of the mapping outputs, especially in urban environments where vertical structures are prevalent and critical for accurate 3D modeling. Furthermore, the quality of collaborative mapping is closely related to communication between robots. This reliance on inter-robot communication becomes a critical bottleneck in complex or signal-compromised environments, such as urban canyons, where signals are significantly degraded due to multipath effects and signal blockages(Song et al., 2024a). In such settings, the integrity of data exchange and synchronization can be severely compromised, leading to interruptions in the mapping process or even complete failures.

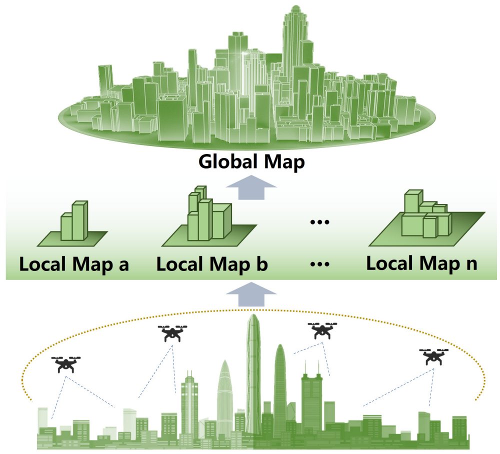


Figure 1. Overview of the proposed UAV-based collaborative mapping framework

This paper proposes a novel approach to collaborative mapping based on extracting environmental semantic landmarks with UAVs. Our approach enables large-scale data acquisition and dynamic updating of 3D models. By leveraging these diverse data streams, the proposed framework constructs high-precision 3D models and incorporates advanced algorithms to ensure the models remain accurate and up to date. This approach addresses the limitations of traditional 3D modeling methods and meets the growing demand for 3D models.

2. RELATED WORK

In recent years, the proliferation of UAVs has opened new avenues for urban 3D modeling by collecting diverse sensor data such as RGB images, 3D point clouds, and IMU measurements. However, these capabilities remain underutilized due to the limitations of traditional modeling techniques, which often suffer from update latency issues(Colomina and Molina, 2014). Our research addresses these challenges by introducing a UAV-based collaborative mapping framework that integrates heterogeneous data from multiple UAVs for efficient environmental reconstruction.

Significant advancements have been made in LiDAR-inertial odometry and mapping, which are crucial for 3D SLAM applications(Biber and Straßer, 2003). The LOAM framework is notable for its ability to decompose SLAM tasks into high-frequency odometry and low-frequency mapping threads, achieving high accuracy with minimal drift (Zhang et al., 2014, Klein and Murray, 2007). Despite these improvements, LOAM's performance on lightweight systems is limited by computational constraints. LeGO-LOAM addresses some of these issues by applying point cloud segmentation and advanced optimization techniques, though it remains challenged in environments without ground points (Shan and Englot, 2018). LIO-SAM further enhances accuracy and robustness by integrating IMU and GPS data through factor graph optimization (Shan et al., 2020). Recent innovations such as Fast-LIO and Fast-LIVO incorporate image data, optimizing state estimation and enabling more realistic 3D reconstructions (Xu and Zhang, 2021, Xu et al., 2022).Recent efforts in improving GNSS-visual-inertial state estimation include the R2-GVIO approach, which demonstrates enhanced robustness in urban settings (Song et al., 2024b). This serves as a foundation for further exploration into real-time processing techniques.

Collaborative mapping frameworks have also evolved to leverage data from multiple agents, improving mapping efficiency and coverage. C-SLAM is an approach that allows multiple robots to share and collaboratively optimize map data, thus reducing the computational burden on each robot. DDF-SLAM builds upon this by facilitating decentralized data fusion, enabling scalable map sharing with low communication overhead (Zhang et al., 2021). In a similar vein, CVI-SLAM fuses visual and inertial data to achieve consistency across shared maps, proving its effectiveness in handling large-scale urban environments (Li et al., 2019). In UAV-based contexts, MUCSLAM effectively enables collaborative exploration and mapping, enhancing both accuracy and computational efficiency (Zhao et al., 2019). Graph-based techniques, such as those used in DGR-SLAM, further enhance scalability and robustness in dynamic multi-agent environments (Lee et al., 2021). The integration of semantic information in frameworks like Semantic Collaborative SLAM (SCSLAM) enhances the interpretability of maps, benefiting applications such as autonomous navigation (McCormac et al., 2017).

Moreover, the integration of advanced visual recognition and optical character recognition (OCR) has improved semantic feature extraction from the environment. The YOLO series of models, particularly YOLOv7, have significantly advanced object detection by introducing dynamic label assignment and optimization techniques, enhancing both speed and accuracy (Redmon et al., 2016, Wang et al., 2023). YOLO-World extends these capabilities through vision-language modeling for effective zero-shot detection (Cheng et al., 2024). Complementary OCR technologies, such as Tesseract and PaddleOCR, provide robust multilingual text recognition, essential for enriching mapping data with semantic details.

These developments underscore the potential of integrating UAVs with multi-modal data to achieve efficient, robust, and scalable urban 3D modeling, addressing the challenges of dynamic and complex environments.

3. METHODS

3.1 System Overview

The proposed system leverages crowdsourced data from Unmanned Aerial Vehicles (UAVs) to construct and dynamically update high-precision real-world 3D models. UAVs, equipped with a combination of RGB cameras, LiDAR, and Inertial Measurement Units (IMUs), collect spatial data during their routine operations, such as urban monitoring or delivery tasks. This data is then processed to generate detailed 3D models of the urban environment, capturing both static structures and dynamic elements. The system integrates semantic feature extraction, local map generation, and map fusion to produce accurate models. Semantic landmarks, such as buildings, roads, and intersections, are extracted from the UAV imagery using advanced computer vision techniques. These landmarks serve as key reference points for aligning and merging local maps generated by individual UAVs. By utilizing these features, the system ensures precise map stitching, even in complex urban environments where traditional feature-based methods may struggle. Once the local maps are aligned, the system employs global optimization techniques to refine the merged map, minimizing errors and ensuring the consistency of the final 3D model. As new UAVs continue to gather data, the system allows for the dynamic updating of the model, integrating fresh information and maintaining its accuracy over time. This collaborative approach to map creation and updating enables the system to scale effectively, ensuring the continuous relevance of the 3D model in rapidly changing urban landscapes.The method architecture proposed in this article is shown in Figure 2.

3.2 Extraction of Environmental Semantic Landmarks

Accurate extraction of environmental semantic landmarks is crucial for enhancing the precision and robustness of 3D modeling in complex urban environments. This paper introduces an innovative approach that combines object detection and Optical Character Recognition (OCR) to extract semantic features from UAV-captured imagery, optimizing the mapping tasks.The process begins with preprocessing raw images captured by UAVs, applying transformations such as histogram equalization and Gaussian smoothing to enhance feature visibility under varying environmental conditions. These preprocessing steps improve image quality, making them suitable for detailed analysis.

For object detection, a sophisticated convolutional neural network is employed to identify and classify key features in the enhanced images. The network outputs bounding boxes and associated class probabilities, which are filtered to retain only

https://doi.org/10.5194/isprs-annals-X-1-W2-2025-1-2025

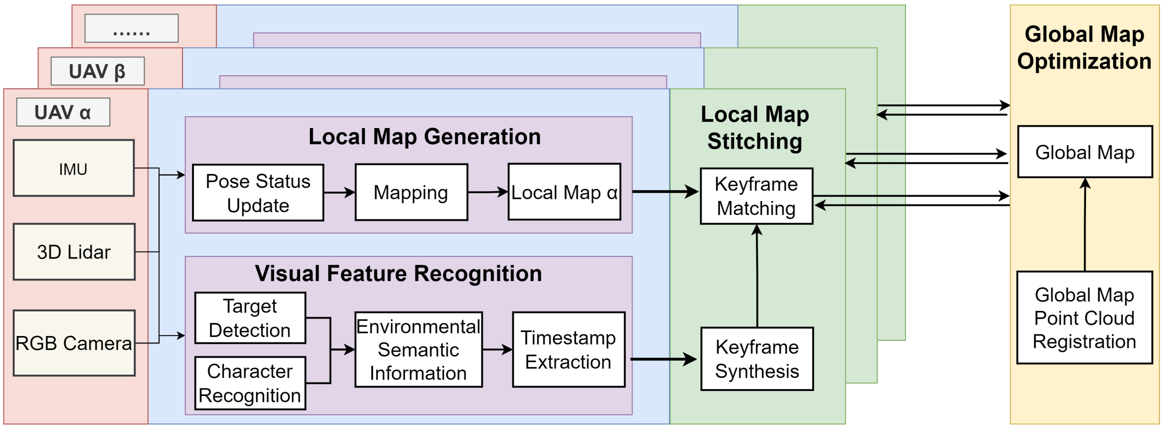


Figure 2. System overview

high-confidence detections, ensuring accuracy and reliability in feature extraction.

In tandem with object detection, textual information provides significant semantic context. OCR techniques are utilized to recognize and extract textual data from the processed images, capturing essential details such as street names, building numbers, and signage. This textual data enriches the semantic map, aiding navigation and localization tasks. The integration of detected objects and recognized text forms a comprehensive semantic map, combining geometric and semantic data to serve as critical inputs for subsequent processes like map stitching and alignment(Liu et al., 2016, Redmon et al., 2016).

This dual approach significantly enhances system robustness and reliability in dynamic and complex environments. By integrating object detection and OCR techniques, the method ensures comprehensive semantic extraction, facilitating the generation of scalable, accurate, and contextually relevant 3D models. This advancement addresses challenges posed by rapidly evolving urban landscapes, providing a powerful tool for real-time mapping and analysis.

Moreover, converting features into textual information reduces data storage requirements, enabling lightweight processing. This methodology not only offers a robust framework for extracting semantic landmarks but also supports precise urban modeling and applications such as autonomous navigation and smart city planning. The integration of object detection and text recognition provides a more comprehensive and contextually relevant solution to landmark extraction compared to traditional methods, ensuring robustness in rapidly changing and complex environments.

3.3 Local Map Generation

Local map generation is a crucial step in creating accurate and dynamic models of urban environments, relying on data collected from LiDAR sensors and Inertial Measurement Units (IMUs). This process is foundational for understanding and navigating complex spatial landscapes.

The process begins with the collection and synchronization of data from LiDAR and IMU sensors. Synchronization is essential as it provides a consistent framework for spatial and temporal measurements, which are necessary for precise state estimation and the construction of accurate maps. LiDAR sensors generate high-resolution point clouds that capture the geometric structure of the environment, detailing features such as buildings, streets, and natural elements with great accuracy. Meanwhile, IMUs offer dynamic motion data by measuring linear acceleration and angular velocity, which are critical for understanding the UAV's movement and orientation in three-dimensional space.

Figure 3. Extracting environmental semantic information through target recognition and OCR

A tightly-coupled sensor fusion approach is employed to continuously estimate the UAV's position, orientation, and velocity. This method integrates data from both LiDAR and IMU, optimizing the estimation process and addressing specific limitations associated with each sensor. For instance, LiDAR can struggle in areas without distinct features, leading to gaps in data, while IMUs can suffer from drift over time, which affects accuracy. By combining these data sources, the fusion approach enhances the reliability and precision of the mapping process.

The computational framework utilizes advanced optimization techniques to minimize errors in odometry measurements. LiDAR data is processed to generate dense point clouds, which are incrementally aligned using methods enhanced by inertial constraints. This alignment process is crucial for improving precision and ensuring that the data converges to form a coherent map. The resulting point clouds are then merged to create detailed local maps. These maps accurately represent the spatial configuration of the environment and are capable of accounting for dynamic changes, such as moving objects or environmental shifts. IMU data plays a pivotal role in maintaining ac-

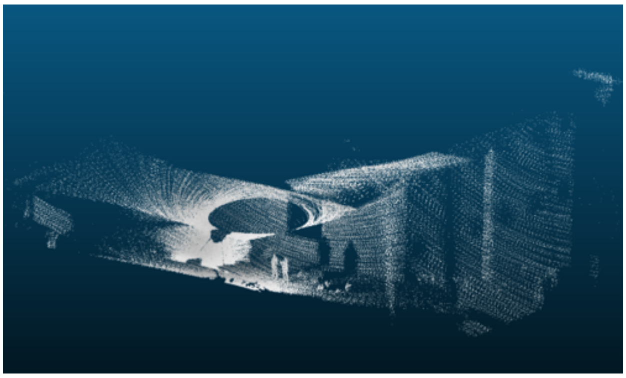


Figure 4. Keyframe extracted during the generation of local maps

curate pose estimation, especially in scenarios where LiDAR data might be compromised by occlusions or challenging conditions, like poor lighting or reflective surfaces. Extended Kalman filtering provides robust motion estimates by effectively combining the strengths of both LiDAR and IMU data, ensuring that the UAV accurately determines its position and orientation.

By integrating LiDAR and IMU data, this method can quickly and stably construct local maps, forming a basis for further processing and analysis. These local maps serve as essential building blocks for broader mapping applications, ensuring that the UAV can adapt to new data inputs and environmental conditions effectively. This robust mapping capability supports a wide range of applications, from urban planning and infrastructure management to environmental monitoring, providing a reliable tool for capturing the complexities of urban environments with high fidelity.

3.4 Local Map Stitching

The process of stitching local maps into a cohesive global representation is vital for building accurate and comprehensive 3D models. This integration leverages environmental semantic landmarks, which act as reliable anchors for aligning and merging individual map segments. These landmarks, such as identifiable buildings, road intersections, and other distinctive urban features, provide robust reference points that enhance both geometric and semantic accuracy in map stitching (McCormac et al., 2017).

Each local map $\mathcal{M}_i$ consists of detailed geometric features and semantic landmarks $\mathcal{L}_i = \{l_1, l_2, \ldots, l_n\}$. Each landmark l_k is identified by its spatial position $\mathbf{p}_k$ and associated semantic descriptor $\mathbf{d}_k$. These landmarks are detected using visual recognition and OCR techniques, where only those results that surpass a predefined confidence threshold are recorded. For each detected feature, the system logs the recognition output, confidence level, timestamp, and relative time from the video start.

During the local map stitching process, UAVs extract multiple keyframes from their recorded data.Keyframe registration is shown in the Figure5. Instead of matching every keyframe exhaustively, the process involves checking the semantic information associated with each keyframe. The keyframes extracted during the local map construction process are shown in Figure 4 Only keyframes with matching semantic information proceed to point cloud alignment using the Iterative Closest Point (ICP) algorithm. This step ensures that the alignment process is guided by robust semantic correlations, improving the accuracy and reliability of the map stitching (Li et al., 2019).

The stitching process employs a semantic-guided iterative closest point (ICP) algorithm designed to minimize the transformation error between overlapping map regions. The Iterative Closest Point (ICP) algorithm is a widely-used method for aligning three-dimensional point clouds, particularly in the context of computer vision and robotics. Its primary objective is to minimize the difference between two point clouds by iteratively refining the transformation—typically consisting of rotation and translation—that aligns them. The algorithm proceeds through several key steps: initially, it selects an arbitrary transformation as a starting point. Subsequently, it identifies the closest points between the source and target point clouds, calculates the optimal transformation to minimize the distance between these paired points, applies this transformation to the source cloud, and evaluates the alignment error. This process is repeated until the error converges to a minimal value or a maximum number of iterations is reached. This algorithm enhances traditional ICP methods by incorporating semantic information into the alignment process, which is particularly advantageous in environments where geometric features alone may be insufficient due to occlusions or sparse data (Cieslewski et al., 2018). The objective function for alignment is expressed as:

$$E(\mathbf{R}, \mathbf{t}) = \sum_{k=1}^{n} \omega_k \left\| \mathbf{R}\mathbf{p}_k^{\mathcal{M}_i} + \mathbf{t} - \mathbf{p}_k^{\mathcal{M}_j} \right\|^2, \tag{1}$$

where $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector applied to the landmarks of map $\mathcal{M}_i$. The weight factor ω_k is computed based on the semantic similarity between corresponding landmarks $\mathbf{d}_k^{\mathcal{M}_i}$ and $\mathbf{d}_k^{\mathcal{M}_j}$, using metrics like cosine similarity:

$$\omega_k = \frac{\mathbf{d}_k^{\mathcal{M}_i} \cdot \mathbf{d}_k^{\mathcal{M}_j}}{\|\mathbf{d}_k^{\mathcal{M}_i}\| \|\mathbf{d}_k^{\mathcal{M}_j}\|}. \tag{2}$$

This semantic weighting ensures that landmarks with higher semantic correlation exert greater influence on the alignment process, significantly enhancing robustness and precision (Zhang et al., 2021). The optimization involves iteratively adjusting $\mathbf{R}$ and $\mathbf{t}$ to minimize the error function $E(\mathbf{R}, \mathbf{t})$, using advanced optimization techniques such as Levenberg-Marquardt or gradient descent, which are well-suited for handling nonlinear problems. Once the optimal transformation parameters $(\mathbf{R}^*, \mathbf{t}^*)$ are determined, these are applied to merge the local maps into a unified global map $\mathcal{M}_{\text{global}}$. The fusion process updates the positions and semantic descriptors of landmarks, ensuring continuity and coherence across the global map:

$$\mathbf{p}_k^{\mathcal{M}_{\text{global}}} = \mathbf{R}^* \mathbf{p}_k^{\mathcal{M}_i} + \mathbf{t}^*, \tag{3}$$

$$\mathbf{d}_k^{\mathcal{M}_{\text{global}}} = \frac{\mathbf{d}_k^{\mathcal{M}_i} + \mathbf{d}_k^{\mathcal{M}_j}}{2}. \tag{4}$$

This systematic approach ensures both geometric consistency and semantic coherence, facilitating accurate representation within urban environments (Zhao et al., 2019).

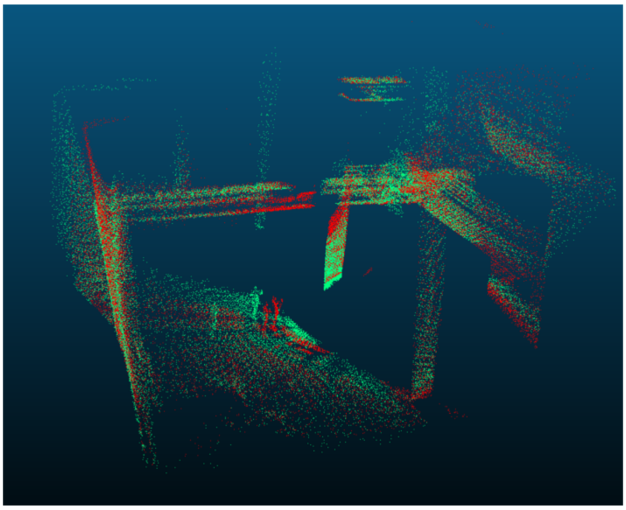


Figure 5. The red point cloud represents the keyframe extracted from one local map, and the green point cloud represents the keyframe extracted from another local map. The two keyframes are registered using ICP.

Moreover, adaptive strategies are incorporated to manage the dynamic elements and occlusions typical of urban landscapes. By continuously updating semantic descriptors and leveraging UAV data, the system dynamically adjusts to environmental changes, maintaining the integrity and accuracy of the 3D model. This semantic-driven approach significantly enhances the reliability of urban mapping, providing a robust solution for real-time 3D modeling (Lee et al., 2021).

The integration of semantic landmarks reduces susceptibility to noise and discrepancies that often affect purely geometric features. This capability is crucial for applications requiring high precision and real-time updates. Furthermore, the ability to dynamically adjust to changes in the environment ensures that the system remains robust and effective even in rapidly evolving urban landscapes (McCormac et al., 2017).

4. EXPERIMENT

4.1 Hardware Equipment

In the experimental setup of the UAV - based collaborative mapping framework, the Livox Mid360 LiDAR and the Walksnail Moonlight Kit camera are integrated, as depicted in Figure 7. The Livox Mid360 LiDAR is selected owing to its wide - ranging field of view and high - precision capabilities in generating 3D point clouds. It boasts a full 360 - degree field of view and a detection range reaching up to 100 meters. Operating at a frequency of 10 Hz, it facilitates data acquisition, which is indispensable for dynamic mapping tasks. Additionally, it is equipped with an integrated Inertial Measurement Unit (IMU), which furnishes essential motion data, including linear acceleration and angular velocity.The Walksnail Moonlight Kit camera is employed for its remarkable high - resolution imaging capabilities. It can capture 4K video, demonstrating superior performance under low - light conditions.The UAV platform is equipped with a robust onboard computer, the NVIDIA Jetson Orin NX, which effectively meets the computational demands of data processing. Its system architecture is designed to support the simultaneous processing and seamless integration of LiDAR, visual, and inertial data. Built on the Ampere architecture GPU, which is equipped with 1024 CUDA cores and

Figure 6. Overall display of UAV

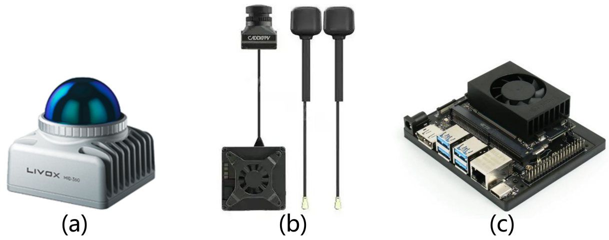


Figure 7. Hardware equipment used in the experiments: (a) Livox Mid360 LiDAR, which provides a wide field of view and high-precision 3D point cloud data; (b) Walksnail Moonlight Kit Camera, offering high-resolution imaging for capturing detailed visual data; (c) NVIDIA Jetson Orin NX, a powerful onboard computer for processing and integrating sensor data efficiently.

32 Tensor cores, it delivers a computing power of up to 100 TOPS. The inclusion of an Arm Cortex - A78AE 64 - bit octa - core CPU further enhances its overall processing efficiency.The overall situation of UAV is shown in the Figure 6.

4.2 Experimental Scenario

The experimental setup is designed around two distinct scenarios to thoroughly evaluate the proposed UAV-based collaborative mapping method.

The first scenario is located in an urban environment that features both indoor and outdoor elements, akin to an urban canyon. This area is characterized by numerous tall buildings and dense architectural structures, which simulate the complex environments UAVs are expected to navigate as the low-altitude economy continues to grow. Such environments are anticipated to become commonplace for UAV takeoff and landing operations.

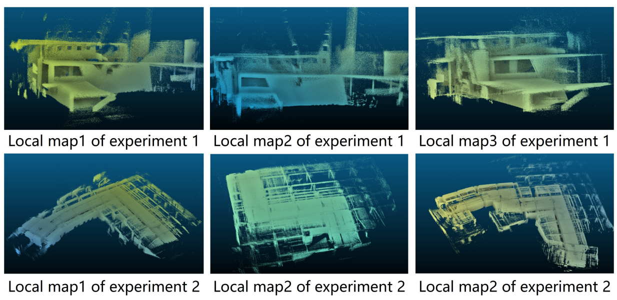
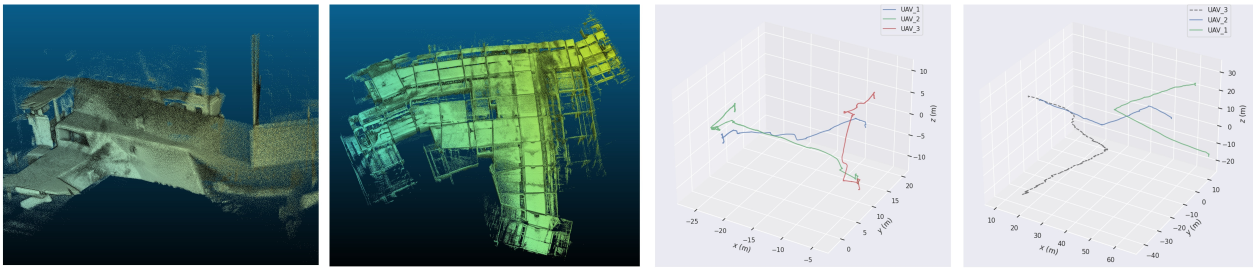


Figure 8. Local maps generated from two separate experiments

Table 1. Comparison of ground truth and measured values in SLAM mapping showing absolute and relative errors for two experiments

Experiment	Truth(m)	Measured Value(m)	Absolute Error(m)	Relative Error(%)
1	5.56	5.78	0.22	3.95
	7.58	7.45	0.13	1.71
	12.36	12.65	-0.29	2.34
	4.34	4.51	-0.17	3.91
	8.24	8.05	-0.19	2.31
	9.54	9.85	0.31	3.13
2	7.94	7.43	0.51	6.42
	15.89	14.76	0.63	3.96
	10.22	10.67	-0.45	4.40
	19.01	20.73	-1.02	5.36
	6.67	6.95	0.28	4.27
	6.27	6.08	-0.19	3.01



Global map of experiment 1 Global map of experiment 2
(a)
Trajectory of experiment 1 Trajectory of experiment 2
(b)

Figure 9. Figure (a) illustrates the global maps generated from the integration of data collected during two separate experiments. These maps highlight the effectiveness of our UAV-based collaborative mapping framework in accurately reconstructing detailed 3D models of the urban environment. Figure (b) depicts the flight trajectories followed by the UAVs in each experiment.

Testing in this scenario allows us to assess the method's ability to handle the intricate spatial dynamics and diverse structural features typical of urban settings.

The second scenario is characterized by an environment with significantly fewer structural features and reduced environmental semantic information compared to the urban setting. This scenario presents a greater challenge to the proposed method due to its lack of distinctive landmarks and diminished semantic context. By conducting experiments in such a setting, we can effectively evaluate the feasibility and robustness of the method under conditions that are less information-rich and more typical of constrained environments. This rigorous testing in a simplified environment helps demonstrate the method's adaptability and reliability when faced with minimal structural cues.

4.3 Experimental Results

The first experiment of this article was conducted in modern urban indoor and outdoor spaces, which have complex spatial structures and rich environmental semantic features. This experiment used three unmanned aerial vehicles for collaborative mapping. Three drones collected image data, laser point cloud data, and IMU information during their respective flights. The three trajectories in this experiment can effectively extract semantic information from the environment, find keyframes based on semantic information, and thus complete the stitching of three local maps.The local maps of the two experiments are shown in Figure 8. The global map and trajectories of Experiment 1 and Experiment 2 is shown in Figure 9.

In scenarios where aircraft struggle to receive accurate GPS satellite signals, and unlike ground-based unmanned vehicles, UAVs face challenges in performing error assessments based on loop closure detection. Therefore, we adopt a method leveraging specific optical markers to measure mapping accuracy. In one set of experiments, we deployed 10 sets of optical markers.To quantify the mapping error in laser SLAM, reflective tape is used as ground truth markers. This tape, characterized by high laser reflectivity, is strategically placed in the real-world environment, forming distinct features in the laser point cloud that can be precisely detected by the SLAM algorithm. The process involves measuring the actual distances between selected pairs of reflective tape using a high-precision laser rangefinder to establish ground truth baselines. These measurements are then compared with the Euclidean distances between the corresponding tape pairs in the SLAM-generated map. The mapping error is calculated using two metrics: the absolute error, defined as

$$\text{Absolute Error} = |D_{\text{map}} - D_{\text{truth}}|, \tag{5}$$

and the relative error, given by

$$\text{Relative Error} = \frac{|D_{\text{map}} - D_{\text{truth}}|}{D_{\text{truth}}} \times 100\%. \tag{6}$$

By computing these errors for multiple tape pairs, we obtain a quantitative assessment of the SLAM system's accuracy in indoor environments where GPS signals are unavailable. These calculations for each tape pair provide a quantitative assessment of the SLAM system's accuracy. This approach offers a practical and cost - effective means to validate mapping performance in indoor environments where GPS signals are unavailable.

5. CONCLUSION

This study presents a UAV-based collaborative mapping framework designed to efficiently integrate heterogeneous data from multiple UAVs for constructing and updating 3D models of urban environments. By leveraging LiDAR and IMU data alongside advanced visual and optical character recognition techniques, the framework addresses the limitations of traditional 3D modeling methods, particularly concerning update speed and scalability.

The integration of semantic feature extraction with LiDAR inertial odometry enhances the precision and robustness of mapping in complex urban settings. Our system also employs camera-LiDAR calibration and cross-modal registration to further improve mapping accuracy and reliability. Experimental results across diverse scenarios, including mixed urban indoor-outdoor and solely indoor environments, demonstrate the framework's adaptability and robustness. The feature-guided point cloud registration, supported by semantic alignment, consistently outperforms traditional methods, validating its effectiveness in dynamic urban modeling.

This research provides a scalable and efficient solution for real-time environmental mapping, applicable to urban planning, smart city development, and other domains benefiting from up-to-date 3D models.

References

Biber, P., Straßer, W., 2003. The normal distributions transform: A new approach to laser scan matching. *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, 2743–2748.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y., 2024. Yolo-world: Real-time open-vocabulary object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16911.

Cieslewski, T., Lynen, S., Bosse, M., Scaramuzza, D., 2018. Data-efficient decentralized visual slam. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2466–2473.

Colomina, I., Molina, P., 2014. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92, 79–97.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211–221.

Klein, G., Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, IEEE, 225–234.

Lee, H., Kim, S., Kim, A., 2021. DGR-SLAM: Distributed graph-based robust SLAM for multi-robot systems. *IEEE Robotics and Automation Letters*, 6(2), 1418–1425.

Li, W., Zhang, H., Zhao, C., 2019. Collaborative visual-inertial slam for multiple robots. *2019 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 6275–6281.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016. Ssd: Single shot multibox detector. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 21–37.

McCormac, J., Handa, A., Davison, A. J., Leutenegger, S., 2017. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 4628–4635.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6), 1–11.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Shan, T., Englot, B., 2018. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 4758–4765.

Shan, T., Englot, B., Meyers, D., Wang, W., Ratti, C., Rus, D., 2020. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 5135–5142.

Song, J., Li, W., Duan, C., Wang, L., Fan, Y., Zhu, X., 2024a. An optimization-based indoor–outdoor seamless positioning method integrating GNSS RTK, PS, and VIO. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 71(5), 2889–2893.

Song, J., Li, W., Duan, C., Zhu, X., 2024b. R2-GVIO: A robust, real-time GNSS-visual-inertial state estimator in urban challenging environments. *IEEE Internet of Things Journal*.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M., 2023. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475.

Xu, W., Cai, Y., He, D., Lin, J., Zhang, F., 2022. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4), 2053–2073.

Xu, W., Zhang, F., 2021. Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter. *IEEE Robotics and Automation Letters*, 6(2), 3317–3324.

Zhang, J., Kaess, M., Singh, S., 2021. Ddf-slam: Decentralized data fusion slam. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 1489–1496.

Zhang, J., Singh, S. et al., 2014. Loam: Lidar odometry and mapping in real-time. *Robotics: Science and systems*, 2number 9, Berkeley, CA, 1–9.

Zhao, B., Yuan, Z., Li, W., 2019. Multi-UAV cooperative SLAM for large-scale complex environments. *IEEE Access*, 7, 39816–39830.