2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Cross-Modal Monocular Localization in Prior LiDAR Maps Utilizing Semantic Consistency

Chi Zhang, Hengwang Zhao, Chunxiang Wang, Xuanlai Tang and Ming Yang

***Abstract*— Visual localization for mobile robots and intelligent vehicles in prior LiDAR maps can achieve high accuracy and low cost. However, algorithms for finding the cross-modal correspondences between images and LiDAR map points are not yet stable. In this paper, we propose a monocular visual localization system in prior LiDAR maps, which is based on the cross-modal registration to optimize the camera pose. To align the point clouds from vision and LiDAR map, a point-to-plane Iterative Closest Point algorithm utilizing semantic consistency is designed, and a decoupling optimization strategy is proposed to compute the affine transformation for the monocular scale ambiguity. Experiments on KITTI dataset show that utilizing the semantic consistency and geometric information of the map makes our system competitive with other methods. On the self-collected dataset, experiments on different light intensities demonstrate the robustness of the system in long-term localization tasks, and the ablation study demonstrates the effectiveness of the proposed algorithms.**

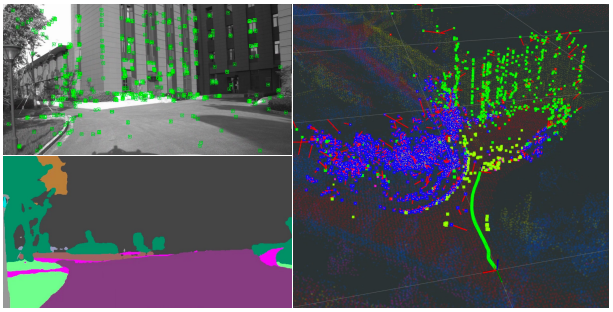

Fig. 1. This paper proposes a monocular localization system on LiDAR semantic map. The top left figure is the monocular image with features and the bottom left is the semantic segmentation. As the right figure shows, with the semantic visual 3D points (bright) and their corresponding LiDAR points in the map (translucent), accurate vehicle trajectory (green curve) can be computed.

I. Introduction

With the wide application of mobile robots and intelligent vehicles, a robust and accurate localization system shows great significance. LiDAR-based methods can achieve high precision, but the high cost of sensors hinders their promotion. Global Navigation Satellite System (GNSS) for localization is widely used, but its performance highly depends on the environment which gets invalid in indoor scenarios. The visual localization method gets concerned for the inexpensive camera sensors and the wide range of application scenarios [1]–[4]. To promote the application of visual localization, the difficulties lie in improving the accuracy and robustness. Specifically, the goal is to use camera sensors to obtain accurate and stable localization in prior maps, and the difficulty of map-based visual localization is to find the connection between camera images and map elements.

Using a camera to build the prior map is the traditional method. Hand-crafted visual descriptors [5]–[7] can be used as map elements in some light-stable scenes but will fail with illumination changes as they are sensitive to light. Although features from neural networks [8]–[10] make up for the above shortcomings, the model needs to be retrained as they are difficult to generalize to various scenarios, which limits its application. Semantic information exists in camera images, and some studies [11]–[14] show that explicitly marking some semantic elements in the map, such as lane lines, zebra crossings, etc., can improve the accuracy of visual localization. However, these works mainly focus on the semantic information of road markings, which limits their application scenarios.

Cross-modal localization has recently received attention, specifically, using LiDAR to build a prior map and then using a camera to localize in it. The LiDAR point cloud is not affected by light and can be used as a map element, which provides accurate geometric information and long-term stability. However, it is difficult to find the correspondences between the LiDAR point cloud and visual images. Some methods [15], [16] use structure from motion (SFM) to recover 3D visual point clouds from consecutive images, and then convert the visual localization task to a point cloud registration problem. But the traditional Iterative Closest Point (ICP) [17] algorithm is not satisfactory in this cross-modal registration task. The traditional ICP algorithm uses geometric information to achieve the matching process, which has a better performance in the point clouds from the same sensor. But the point clouds obtained by the camera and the LiDAR show a tremendous difference in accuracy and density, and a more robust registration algorithm is needed. Semantic information also exists in the LiDAR point cloud and can be used in the map. The recent work SemLoc [18] uses semantic LiDAR maps to introduce semantic elements into the optimization of SMF and achieves state-of-the-art accuracy in visual localization. However, their method lacks explicit identification of the semantic map elements, which makes some unmatched map points also participate in the

This work was supported by the National Natural Science Foundation of China (U22A20100/62173228). Chunxiang Wang and Ming Yang are the co-corresponding authors.

Chi Zhang, Hengwang Zhao, Chunxiang Wang and Ming Yang are with the Department of Automation, Shanghai Jiao Tong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China (phone: +86-21-34204533; email: Wangcx@sjtu.edu.cn, MingYANG@sjtu.edu.cn).

Xuanlai Tang is with the KEENON Robotics Co., Ltd..

optimization of localization. This shortcoming reduces the system's accuracy and limits the application.

Based on the demands above, a cross-modal monocular visual localization system on LiDAR semantic map is proposed in this paper. Aiming at the problem that the visual feature map is sensitive to illumination variation, the semantic LiDAR point cloud map is used to achieve long-term stability. The proposed system achieves accurate cross-modal visual localization based on the designed semantic point-to-plane ICP and decoupling optimization strategies, which robustly align visual point clouds with map LiDAR point clouds. Fig. 1 shows a monocular image with corresponding semantic segmentation, and the tracking trajectories can be obtained in the prior map. The main contributions of our work are as follows:

- A monocular visual localization framework for mobile robots and intelligent vehicles is proposed, which can provide the accurate 6 degrees of freedom (DoF) pose in the prior LiDAR semantic map.
- A novel registration method for cross-modal point clouds from camera and LiDAR is designed, which utilizes semantic consistency and a decoupling optimization strategy to improve the cross-modal registration accuracy.
- Experiments on KITTI dataset [19] and the self-collected dataset demonstrate the validity of our system, which achieves competitive performance in localization tasks.

II. Related Works

Map-based visual localization methods have attracted the attention of many researchers recently. The main idea of them is to find correspondences between images and elements in the map, which can eliminate accumulated errors in movement.

3D LiDAR point cloud map is a long-term map with rich geometric information, and approaches of using the camera to achieve cross-modal localization have developed in recent years. [15] reconstructs the visual point clouds from monocular consecutive images and turns the localization task into point clouds registration task, which is solved by the optimizing nonlinear least squares problem. Based on the registration idea, point-to-plane ICP is proposed for the visual point clouds in [16]. To overcome the geometric difference of cross-modal point clouds, the deep learning algorithm is introduced in [20] to extract a subset of the LiDAR map, and invariant geometric properties can be obtained. Besides the point-level registration, feature-level connections between LiDAR map and visual point clouds, like lines [21] and surfel [22], are used in camera localization. Although the LiDAR map is stable and precise, a robust and accurate visual matching algorithm for it is currently lacking.

In addition to registration between LiDAR map points and visual points, importing map points as constraints in the optimization of visual localization can also improve accuracy. The gaussian mixture model is used in [23] to introduce map points into the visual odometry, and the structure consistent can provide geometric constraints. In [24], [25], LiDAR map gets involved in pose graph optimization, which reduces odometry drift in their system. For a stereo camera system in [26], depth information can be calculated via binocular disparity, and reducing residuals between the LiDAR map and the constructed depth can improve localization accuracy. Deep learning method like [27] use a neural network to project map points to visual images and achieves high localization accuracy in datasets. However, the introduction of map elements lacks explicit identification, so mismatching in some scenarios is a common problem.

Compared to the LiDAR point cloud map, the semantic map contains more information, and is more suitable for intelligent vehicles. Modern structured scenes, with map markers such as lane lines and zebra crossings, can be used as map elements to assist localization. A traffic semantic map can be built in a crowd-sourced way in [13], and can be used for a long time. The registration idea is considered in [13], which uses the ICP algorithm to compute transformation between map and perceptual semantic segmentation of lane lines. Some researchers [11] annotate the location of lane lines in the road, and determines the camera pose through a coupled relation between map elements and images. TM^{3}Loc [28] uses semantic chamfer matching to combine semantic features in maps and images, and achieves high precision in outdoor large datasets. Compared to semantics in structured scenes, semantic information at the macro level such as road, building, and so on could broaden the application scenarios of vehicles. SemLoc [18] uses these semantics to build a prior map, and map elements provide data associations as additional constraints to optimize the camera poses. The semantics provides more information, but how to use semantics to achieve robust and accurate localization requires a further exploration.

Considering that semantic information is shared in pre-built semantic maps and images, we proposed a monocular visual localization system, which utilizes semantic consistency to find correspondences in cross-modal point clouds. To solve the scale ambiguity of monocular cameras and mismatching in cross-modal registration, we propose to optimize the camera pose and scale using a decoupling strategy.

III. Proposed Method

A. System Overview

With consecutive monocular images $\mathcal{F} \triangleq \{\mathcal{F}_1, \mathcal{F}_2, \dots \mathcal{F}_i\}$ and a prior LiDAR map $\mathcal{M}$ as the inputs of the system, our target is to obtain $T_{M_0 M_i}$, the 6DoF pose estimation of the camera in the map. The optimization idea of our system is based on registration. The monocular visual odometry can provide a rough camera pose without scale. After matching the visual point cloud with the map point cloud, the affine transformation matrix can be obtained through registration, which describes the relationship between the visual point cloud and the map point cloud. By applying this transformation to the rough pose in visual odometry, the precise camera pose can be obtained.

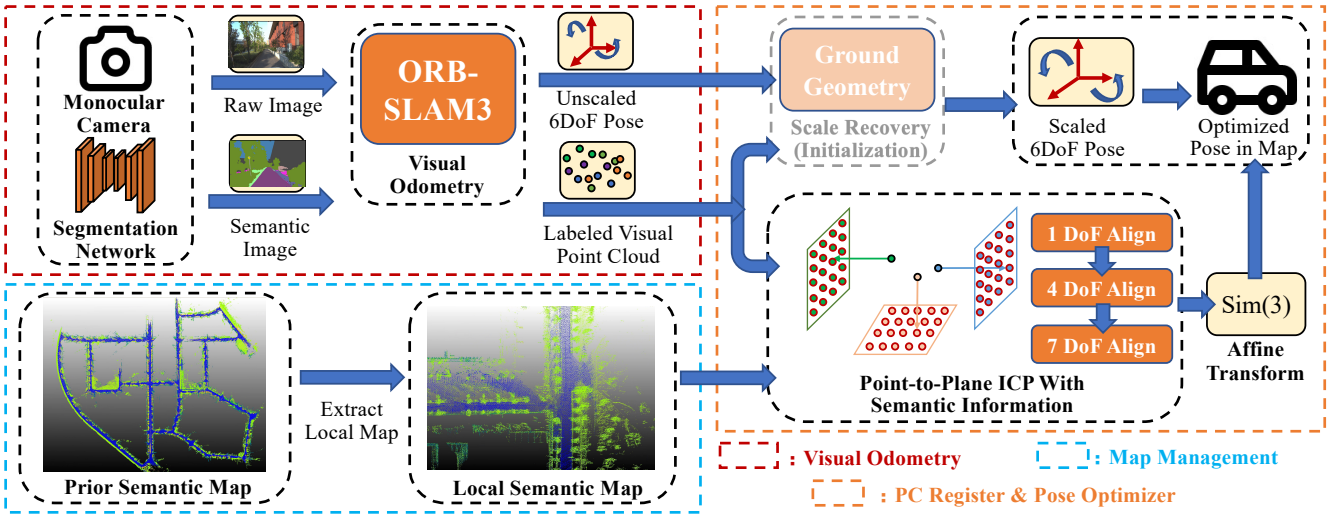

Fig. 2. System overview. The localization system includes three modules: visual odometry(red dotted box), map management(blue dotted box), and point clouds register & pose optimizer(orange dotted box). With the visual point cloud and map point cloud, the registration can be realized via the semantic consistency and our decoupling optimization strategy, and the camera pose of visual odometry can be optimized.

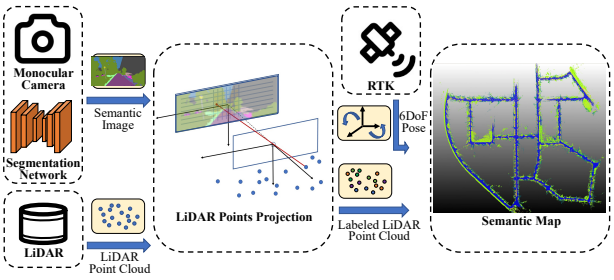

Fig. 3. The flow chart about how to build the prior semantic LiDAR map. With the extrinsic matrix of LiDAR and camera, combined with the camera model, the LiDAR points can be projected onto the image, and the category of each points can be annotated according to the result of semantic segmentation of the image. This simplifies the creation of semantic LiDAR maps

As shown in Fig. 2, the localization system includes three modules:

- Visual Odometry: based on the consecutive monocular images, ORB-SLAM3 [2] can provide unscaled 6DoF pose and visual 3D points. Semantic segmentation network [29] can label each pixel and corresponding 3D points.
- Map Management: as the vehicle moves on the map, the map management module will extract local map points based on the current location for the registration.
- Registration and Optimization: for the visual point cloud and local map point cloud, the registration module uses semantic consistency to align the cross-modal point clouds. With the 7DoF transformation matrix (6DoF for pose and 1DoF for scale), optimized vehicle pose in the map can be obtained.

Although semantic segmentation network can recognize more than ten categories, our system classifies them into abstract categories. The types of vegetations and the types of roads are not classified. Dynamic semantic objects such as vehicles and pedestrians are removed, and other static objects are uniformly marked as buildings. For any point p from visual odometry or LiDAR map, its corresponding label satisfies:

$$label\,(p) \in \{road, vegetation, building\}$$

B. Prior Map Construction and Management

Considering the high cost and time-consuming of manually annotated semantic point cloud maps, visual semantic segmentation is used to build the prior map. The mapping process is shown in Fig.3: with the extrinsic matrix of the camera and LiDAR, the camera model can be used to find the corresponding image pixel of the 3D point, and the semantic information of the point can be obtained by the network [29]. Use the pose provided by Real-Time Kinematic (RTK) to overlay multi-frame point clouds to create the map.

The map management module is based on the work in [16]. The normals of each point will be calculated after the loading of the map, which is an offline operation and will be used later. The local map extractor can update the local map based on the current camera location on the map. The map point cloud $\mathbf{P}_{M_i}$ of the local map for current frame $\mathcal{F}_i$ is presented as ${}^{M_0}P_{M_i}$ in homogeneous coordinates.

C. Visual Odometry and Scale Initialization

For image frame $\mathcal{F}_i$, visual odometry provides its 6DoF pose without scale $T_{C_0C_i} \in SE(3)$ in the visual coordinate system. Based on the visual local map, visual point cloud $\mathbf{P}_{C_i}$ for image frame $\mathcal{F}_i$ can be obtained, which is presented as ${}^{C_0}P_{C_i}$ in homogeneous coordinates.

Scale is unknown for pure monocular systems, while a rough scale relation between the computed odometry and the real odometry is necessary. In our system, scale is recovered in the initialization based on the ground point aggregation method. With the semantic information, visual

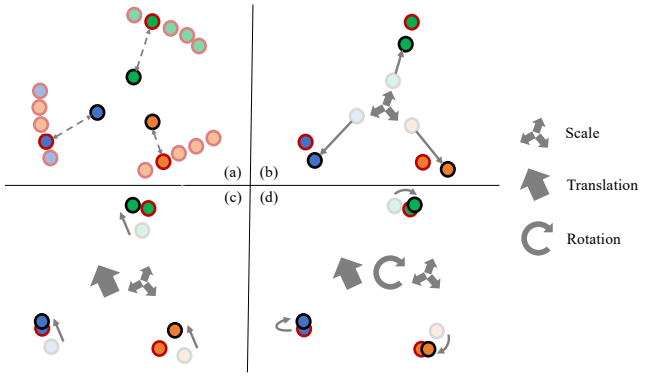

Fig. 4. Schematic diagram for point-to-plane ICP with semantic consistency. (a) shows how to use semantic consistency to find corresponding points, (b) to (d) shows the decoupling optimization strategy for the registration.

points belonging to the ground can be obtained, and the outliers can be filtered with a random sample consensus (RANSAC) plane fitting. The average height of these points $\overline{h}_c$, combined with the height of the camera in the real world h_r, can be used to calculate the scale: $s_0 = h_r/\overline{h}_c$

D. Preprocessing for Registration

Since the scale information of the monocular image is inaccurate, the affine transformation of the visual point cloud in the global coordinate system will easily lead to mismatches of registration. Therefore, we design a registration coordinate system $T_{M_0M_k}$ that changes periodically with the camera movement. It is necessary to transform the visual point cloud, map point cloud and camera pose into this coordinate system, which is conducive to reducing point cloud mismatches and error accumulation. For the visual point cloud $\mathbf{P}_{C_i}$ and map point cloud $\mathbf{P}_{M_i}$, they can be transformed to the registration coordinate system by:

$$\begin{cases} {}^{M_k}P_{C_i} = (T_{M_0M_k})^{-1} T_{M_0C_0} {}^{C_0}P_{C_i} \\ {}^{M_k}P_{M_i} = (T_{M_0M_k})^{-1} {}^{M_0}P_{M_i} \end{cases} \tag{1}$$

where $T_{M_0M_k}$ is the pose of frame $\mathcal{F}_k$ in the map, and k is updated periodically. $T_{M_0C_0}$ is the initial camera pose in the map and is considered to be known.

E. Point-to-Plane ICP with Semantic Information

For the registration task, the optimization target is to find an affine transformation matrix S_i between ${}^{M_k}P_{C_i}$ and ${}^{M_k}P_{M_i}$:

$$S_i = \begin{pmatrix} s_iR_i & t_i \\ 0 & 1 \end{pmatrix} \tag{2}$$

where $S_i \in Sim(3)$, $s_i \in \mathbb{R}$, $R_i \in SO(3)$, $t_i \in \mathbb{R}^3$.

The general ICP algorithm iterates the following two basic steps: find the correspondences between point clouds, then minimize the error function. The registration task of monocular visual point cloud and LiDAR point cloud faces great difficulties in these two steps: when finding correspondences, cross-modal point clouds have large differences in geometric features, so additional information is required as a constraint; when minimizing the error function, due to the uncertainty of the monocular scale, the optimization of nonlinear error function will fall into a local optimum and the optimization variable will be degraded. To solve the above problems, a semantic point-to-plane ICP with a decoupling optimization strategy is designed as follows, and the schematic diagram is shown in Fig. 4.

For each visual point $p_c \in {}^{M_k}P_{C_i}$, use the initial 7DoF transformation matrix S_i to find its candidate corresponding map points $P_m = \{p_{mj} | p_{mj} \in {}^{M_k}P_{M_i}, j = 1, 2, ..., K\}$. Based on the KD-Tree algorithm, nearest K neighbors at S_ip_c can be efficiently found. Semantic consistency is used for selecting the best corresponding map point p_m in the set P_m. Among the candidate points found by K-nearest neighbors, the map points that have the same semantics as the target point are closer at the semantic level. Based on the point-to-plane ICP [30], p_m should satisfies:

$$p_m = \arg\min_{p_m \in P_m} \{\lambda (S_ip_c - p_m)^T \Phi_{p_m} (S_ip_c - p_m)\} \tag{3}$$

$$\lambda = \begin{cases} 1, & label(p_c) = label(p_m) \\ +\infty, & others \end{cases}$$

where Φ_{p_m} is the orthogonal projection matrix and can be computed by using the normal vector at p_m and Gram–Schmidt process.

With the correspondences between ${}^{M_k}P_{C_i}$ and ${}^{M_k}P_{M_i}$, S_i can be computed with a decoupling optimization strategy, which includes three steps:

- for the initial S_i, fix R_i and t_i, optimize the scale s_i.
- inherit S_i from previous step, fix R_i, optimize the scale s_i and the translation t_i.
- inherit S_i from previous step, optimize the rotation R_i, the scale s_i and the translation t_i.

The nonlinear optimization is solved with g2o [31], and the final S_i is iteratively optimized by:

$$\arg\min_{S_i \in Sim(3)} \sum_{p_c \in {}^{M_k}P_{C_i}} \{(S_ip_c - p_m)^T \Phi_{p_m} (S_ip_c - p_m)\} \tag{4}$$

F. Derivation of Pose in Map

The above point cloud matching and pose optimization are completed in the registration coordinate system. To obtain the global pose of the vehicle on the map, the following calculations are required. The optimized pose $T_{M_0M_i}$ of $\mathcal{F}_i$ in map can be computed as:

$$T_{M_0M_i} = T_{M_0M_k} T_{M_kM_i} = T_{M_0M_k} (S_i \odot T_{M_kC_i}) \tag{5}$$

where $T_{M_kC_i} = (T_{M_0M_k})^{-1} T_{M_0C_0} T_{C_0C_i}$ and the $\odot$ defines the caculation between $S \in Sim(3)$ and $T \in SE(3)$:

$$\begin{aligned} S \odot T &= \begin{pmatrix} sR_S & t_S \\ 0 & 1 \end{pmatrix} \odot \begin{pmatrix} R_T & t_T \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} R_SR_T & sR_St_T + t_S \\ 0 & 1 \end{pmatrix} \in SE(3) \end{aligned} \tag{6}$$

TABLE I

ABSOLUTE TRAJECTORY ERROR(ATE) COMPARISON WITH DIFFERENT APPROACHES ON KITTI DATASET

KITTI Seq.	Stereo		Stereo+Map	Monocular		Monocular+Map			
	Stereo ORB SLAM [2]	**Stereo DSO** [3]	**Kento et.al** [16]	**Mono ORB SLAM-s** [2]	**Mono DSO-s** [4]	**Kento et.al** [16]	**Map ORB3** [2]	**SemLoc** [18]	**Ours**
00	1.1788	4.0934	0.7948	69.975	89.982	13.185	1.2282	1.4865	**0.5765**
05	0.7117	2.2656	2.0590	27.804	42.608	4.0380	1.9438	0.7221	**0.4783**
06	0.7171	2.7914	**0.4104**	42.322	113.57	-	1.8498	5.8789	0.7412
07	0.5031	2.8913	**0.4262**	13.067	13.108	-	1.2217	3.3182	0.6630
08	3.2105	3.0611	2.5680	40.576	83.141	-	3.4051	3.7445	**1.2784**
09	3.0213	4.0643	1.7039	44.566	38.841	-	2.4278	0.8810	**0.7168**
10	1.9542	0.6593	0.8555	4.9303	10.472	-	1.8145	1.0039	**0.5023**

Note: '-s' means scale correction. Bold fonts indicate the best results. The evaluation metric is average ATE, and the unit is meter.

IV. EXPERIMENTS

The proposed system is evaluated on KITTI Odometry dataset [19] and the self-collected dataset. Quantitative experiments with other localization methods are conducted on the KITTI. Besides, we use a vehicle to collect data on campus scenarios. In the first scene, three sequences were recorded at different times and with different illumination, which can evaluate the impact of light changes on the system in long-term localization tasks. Besides, the vehicle collected both structured and unstructured road data in the second scenario, which can evaluate the applicability and robustness of the system. The dataset contains images from the monocular camera and LiDAR point clouds from Panda40P, totaling over 7000 frames for localization. The ground truth of vehicle pose is from inertial and GNSS navigation systems with RTK.

The initial pose in the map is supposed to be known as a precondition for all experiments. As for semantic segmentation, we used the network in [29] with the pre-trained model. Average absolute trajectory error (ATE) is used as the evaluation metric for quantitative experiments. The unit of ATE is the meter and the lower value is better.

A. Quantitative Evaluations on KITTI Dataset

The performance of several influential visual SLAM systems, ORB-SLAM3 [2] and DSO [3], [4], are compared with our system. As for the map-based methods, Map ORB3 [2] uses image data to build the map with the visual feature ORB, and then uses the same image data to locate the map. Kento et.al [16] use LiDAR point cloud maps to achieve cross-modal localization with stereo or monocular images. SemLoc [18] is a semantic-map-based visual localization system, which performs state-of-the-art accuracy, and our system is compared with its monocular mode.

As shown in Table I, for the monocular odometry methods [2], [4], due to the scale drift, they perform a low localization accuracy after scale correction. For the stereo odometer methods [2], [3], limited by the accumulated error, the localization accuracy will decrease during long-distance motion. Our proposed method benefits from the prior map, and the absolute scale can be constrained in the monocular odometry, which enables our system to significantly reduce the trajectory error and perform competitive monocular accuracy with stereo odometry.

In the map-based approaches, Map ORB3 [2] shows impressive accuracy, but using ORB features as map elements is not robust to illumination changes. Besides, inaccurate visual features in maps lead to low precision. In the cross-modal solution provided by Kento et.al [16], the monocular mode failed in many sequences because of the scale ambiguity, and the stereo mode performs low accuracy due to the lack of algorithm design for cross-modal data. For the semantic-map-based methods SemLoc [18], the monocular scale degradation can easily happen due to the lack of explicit discrimination of map elements, which leads to low precision in some sequences. For the data association across modalities, we use semantics to match images and point clouds. For the problem of monocular scale uncertainty, our decoupling optimization strategy prioritizes the optimization of the scale value, thereby reducing the mismatch between point clouds. With accurate monocular scale information, a higher localization accuracy can be obtained.

B. Long-Term Localization under Different Light Intensities

Visual localization methods are sensitive to light intensities, which limits their usage in long-term localization task. In this subsection, we test the performance of our system at different light intensities. On the first day, we built the map and captured the image sequence SJTU01 when the illumination was medium. After 15 days in the same scene, we captured the image sequence SJTU02 in strong illumination at midday and the image sequence SJTU03 in weak illumination at dusk. The different light conditions are shown in Fig. 5. Compared with the scene when the map was built, the vegetation and roadblocks in the scene have changed after 15 days, but our map has not been updated accordingly, which is a challenge to localization. As a comparison, Map ORB3 [2] based on the visual map is also experimented with the above data.

The quantitative results are shown in Table II. Map ORB3 [2] uses visual features which do not have long-term stability in outdoor scenarios, and the localization failed after 15 days. As for our system, even if the vegetation and other elements in the scene have changed significantly in the past 15 days after the map building, the system can still complete the localization in the map, which reflects the robustness of

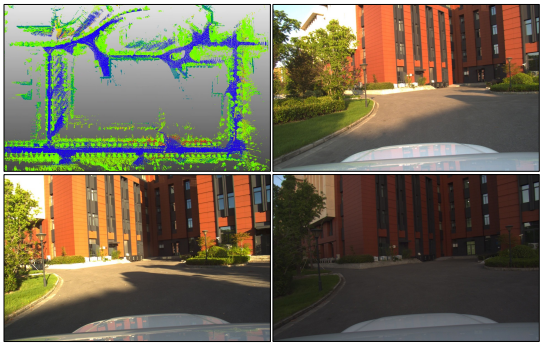

Fig. 5. Data overview for the long-term localization experiments. Top Left: prior LiDAR map for these evaluations. Top Right: medium illumination in SJTU01; Bottom Left: strong illumination in SJTU02; Bottom Right: weak illumination in SJTU03.

TABLE II

SYSTEM PERFORMANCE ON DATA WITH DIFFERENT CONDITIONS IN THE SAME SCENARIOS.

Seq.	Conditions			ATE	
	Length	Light	Past Days	ORB3 [2]	Ours
SJTU01	484.5m	Medium	0 day	5.130	0.818
SJTU02	364.7m	Strong	15 days	failed	0.971
SJTU03	470.2m	Weak	15 days	failed	0.979

our method and the reusability of the semantic LiDAR map. As for the experiments on the illumination, results show that semantics are effective in strong or weak illumination and accurate localization can still be obtained. This subsection proves that it is feasible to use semantics to connect visual images and LiDAR point cloud maps, and the semantic point-to-map ICP with a decoupling optimization strategy can make full use of semantic consistency to estimate the monocular camera pose. As for the run-time, the system can operate at about 10 Hz on the self-collected dataset.

C. Ablation Study on Campus Scenarios and KITTI

On campus dataset SJTU01 and SJTU04, ablation experiments are taken to examine the effect of each algorithm in our system. Fig. 6 shows the qualitative comparison of trajectories under different configurations, and corresponding quantitative results are in Table III. Visual odometry without a map shows serious trajectory drift, which is caused by scale drift and error accumulation. When a LiDAR map is introduced and the camera pose gets optimized with a basic point-to-plane ICP between map and visual points, the estimated locations get closer to the ground truth trajectory. However, due to the point cloud mismatch, the quantitative error is still large. Semantic map can reduce the occurrence of mismatch by finding correspondences with same semantics, which directly improves the accuracy of the SJTU01 and SJTU04 trajectories. After decoupling optimization is introduced to the semantic ICP process, more accurate scale estimation can further reduce the mismatch of point clouds, so that the system can perform more accurate monocular localization results.

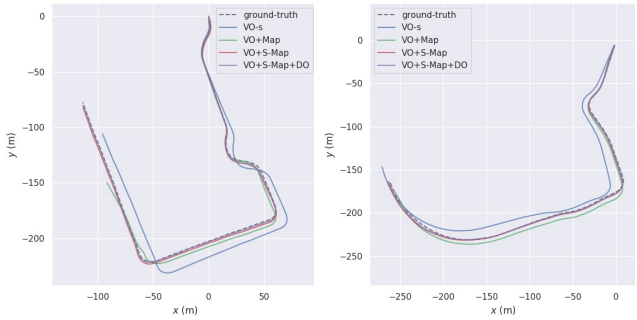

Fig. 6. Trajectory comparison of the ablation study. The trajectories (left: SJTU01, right: SJTU04) are qualitatively compared under different configurations. 'VO' means visual odometry; '-s' means scale correction; 'Map' or 'S-Map' means LiDAR map or semantic LiDAR map; 'DO' means decoupling optimization. The corresponding quantitative results are in Table III.

TABLE III

QUANTITATIVE ANALYSIS OF ABLATION STUDY.

Method	Sequences			
	SJTU01	SJTU04	KITTI06	KITTI07
SemLoc [18]	-	-	5.878	3.318
VO-s	4.656	2.136	45.77	11.10
VO+Map	8.215	4.676	25.50	failed
VO+S-Map	2.024	1.273	3.375	0.765
VO+S-Map+DO(Ours)	**0.818**	**1.166**	**0.741**	**0.663**

On the KITTI dataset, the KITTI06 and KITTI07 sequences are selected to compare our system with SemLoc [18]. In SemLoc [18], semantics are used as landmarks to take part in the pose optimization process, which does not explicitly handle the mismatches problem and gets sensitive to scale uncertainty in monocular mode. With the semantic map and traditional ICP algorithm, our visual localization already performs better accuracy than SemLoc [18], and the decoupling optimization can further improve the performance. That is because, our strategy for semantic information and the decoupling optimization process is explicitly designed to handle mismatch and scale uncertainty, and the qualitative evaluation of KITTI shows its robustness.

V. CONCLUSIONS

In this paper, a cross-modal monocular camera localization system is proposed, which utilizes semantic consistency and geometric information in the prior LiDAR map. In order to solve the cross-modal registration problem between the visual point cloud and LiDAR point cloud, a semantic point-to-plane ICP algorithm is designed with a decoupling optimization strategy to solve scale ambiguity in monocular odometry. In the experiments on KITTI dataset, our system achieves better performance than other state-of-the-art visual localization methods with semantic maps. On the self-collected dataset, the experiments about light intensity demonstrate the system ability to achieve accurate results in long-term localization tasks, and the ablation study demonstrates the effectiveness of the strategies in the system.

REFERENCES

[1] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "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.

[3] R. Wang, M. Schworer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3903–3911.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2010, pp. 778–792.

[6] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3304–3311.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[8] S. J. Lee, D. Kim, S. S. Hwang, and D. Lee, "Local to global: Efficient visual localization for a monocular camera," in *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2231–2240.

[9] D. Li, X. Shi, Q. Long, S. Liu, W. Yang, F. Wang, Q. Wei, and F. Qiao, "DXSLAM: A robust and efficient visual SLAM system with deep features," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4958–4965.

[10] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in neural information processing systems*, vol. 32, 2019.

[11] D. Wong, Y. Kawanishi, D. Deguchi, I. Ide, and H. Murase, "Monocular localization within sparse voxel maps," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 499–504.

[12] Z. Xiao, K. Jiang, S. Xie, T. Wen, C. Yu, and D. Yang, "Monocular vehicle self-localization method based on compact semantic map," in *Proceedings of International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3083–3090.

[13] T. Qin, Y. Zheng, T. Chen, Y. Chen, and Q. Su, "A light-weight semantic map for visual localization towards autonomous driving," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11 248–11 254.

[14] J. Hu, M. Yang, H. Xu, Y. He, and C. Wang, "Mapping and localization using semantic road marking with centimeter-level accuracy in indoor parking lots," in *Proceedings of International Conference on Intelligent Transportation Systems (ITSC)*, 2019, pp. 4068–4073.

[15] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1926–1931.

[16] K. Yabuuchi, D. R. Wong, T. Ishita, Y. Kitsukawa, and S. Kato, "Visual localization for autonomous driving using pre-built point cloud maps," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 913–919.

[17] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[18] S. Liang, Y. Zhang, R. Tian, D. Zhu, L. Yang, and Z. Cao, "Semloc: Accurate and robust visual localization with semantic and structural constraints from prior maps," in *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4135–4141.

[19] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[20] X. Ding, Y. Wang, R. Xiong, D. Li, L. Tang, H. Yin, and L. Zhao, "Persistent stereo visual localization on cross-modal invariant map," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4646–4658, 2019.

[21] H. Yu, W. Zhen, W. Yang, J. Zhang, and S. Scherer, "Monocular camera localization in prior lidar maps with 2d-3d line correspondences," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4588–4594.

[22] H. Ye, H. Huang, and M. Liu, "Monocular direct sparse localization in a prior 3d surfel map," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 8892–8898.

[23] H. Huang, H. Ye, Y. Sun, and M. Liu, "Gmmloc: Structure consistent visual localization with gaussian mixture models," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5043–5050, 2020.

[24] X. Zuo, P. Geneva, Y. Yang, W. Ye, Y. Liu, and G. Huang, "Visual-inertial localization with prior lidar map constraints," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3394–3401, 2019.

[25] X. Zuo, W. Ye, Y. Yang, R. Zheng, T. Vidal-Calleja, G. Huang, and Y. Liu, "Multimodal localization: Stereo over lidar map," *Journal of Field Robotics*, vol. 37, no. 6, pp. 1003–1026, 2020.

[26] Y. Kim, J. Jeong, and A. Kim, "Stereo camera localization in 3d lidar maps," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–9.

[27] D. Cattaneo, D. G. Sorrenti, and A. Valada, "Cmrnet++: Map and camera agnostic monocular visual localization in lidar maps," *arXiv preprint arXiv:2004.13795*, 2020.

[28] T. Wen, K. Jiang, B. Wijaya, H. Li, M. Yang, and D. Yang, "Tm^{3}loc: Tightly-coupled monocular map matching for high precision vehicle localization," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[29] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8856–8865.

[30] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp," in *Robotics: Science and Systems*, vol. 2, no. 4. Seattle, WA, 2009, p. 435.

[31] G. Grisetti, R. Kümmerle, H. Strasdat, and K. Konolige, "g2o: A general framework for (hyper) graph optimization," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 9–13.