Object SLAM With Robust Quadric Initialization and Mapping for Dynamic Outdoors

Object SLAM With Robust Quadric Initialization and Mapping for Dynamic Outdoors

Rui Tian, Yunzhou Zhang, *Member, IEEE*, Zhenzhong Cao, Jinpeng Zhang, Linghao Yang, Sonya Coleman, Dermot Kerr, and Kun Li

***Abstract*— Object SLAM is a popular approach for autonomous driving and robotics, but accurate object perception in outdoor environments remains a challenge. State-of-the-art object SLAM algorithms rely on assumptions and are sensitive to observation noise, limiting their application in real-world scenarios. To address these challenges, we propose a novel object SLAM system that utilizes a quadric initialization algorithm based on constrained quadric optimization, which does not rely on planar assumptions and is robust to partial observations. Additionally, we introduce an automatic object data association algorithm capable of detecting motion states while associating objects across frames. To further enhance the accuracy of the quadric mapping, an extra thread is used to refine the ellipsoid parameters within a local sliding window composed of keyframes. Our system utilizes a joint optimization framework that optimizes camera poses, object landmarks, and point clouds in the local mapping thread for further global optimization while maintaining a consistent map. Experimental results on the real-world KITTI dataset show that the proposed system is more robust and significantly outperforms current state-of-the-art methods in quadric initialization and mapping in outdoor scenarios. Moreover, our system achieves real-time performance, making it suitable for practical applications.**

***Index Terms*— Robotics, visual localization, quadric mapping, data association, object SLAM.**

I. Introduction

SIMULTANEOUS Localization and Mapping (SLAM) is a fundamental technique in order for autonomous vehicles to perceive environments. When compared with classic SLAM methods that use only the geometry of the scene [1], [2], object-based SLAM has recently focused on creating maps containing both geometry and high-level semantic objects within the environment [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. This semantic information promotes target-oriented tasks like obstacle avoidance [13], camera localization [10], [11], and robust relocalization [9], [12]. The improvement in the accuracy of semantic information acquisition, driven by deep learning networks [21], [22], [23], has led to an increased use of object detection and semantic segmentation in visual SLAM systems to build semantically enriched maps and enhance the perception ability.

Accurate object representation is a key issue in object-oriented SLAM research, commonly used object representation can be divided into the prior object model [18], [19], [20], and the generic object model [5], [6], [7], [8], [14], [15], [16], [17]. The prior object model methods rely on the prebuilt CAD model databases [18], [19] or point cloud models [20]. Since these models must be known in advance, the application scenarios of such methods are limited. The generic model methods adopt the cubic box or ellipsoid. In contrast to the cubic box, the ellipsoid can be compactly parameterized and easily manipulated within the framework of projective geometry, which has attracted attention in recent work due to the closed surface of the ellipsoid being meaningful for object landmarks [5], [6], [7], [8]. However, the accuracy and robustness of current quadric-based SLAM algorithms are not ideal, especially the quadric initialization process, which is limited by the parameter coupling of the direct linear solution method [5], the necessity of point cloud fitting [14], or the commonly used decoupled orientation methods rely on planar assumptions [6], [8], which limits their applications in real-world scenarios. Therefore, we propose a novel robust and accurate quadric initialization algorithm for outdoor scenes.

One of the core issues in object SLAM is object association since it establishes the correspondence between the same object in different frames and the global map, and these correspondences provide essential constraints for the optimization of object SLAM. Unlike data association methods using designed features [24], [25] or descriptors [26], object association methods can be divided into probabilistic methods [27], [28], [29] and assignment-based methods [30], [31]. However, the probabilistic EM algorithm is time-consuming and cannot handle large-scale scenarios, so its application scenarios are limited. Assignment-based methods are commonly utilized by using the constraints of overlaps between detection and projection results. However, the performance of these methods can be significantly degraded in the case of partial observations and occlusions, leading to false correspondences. Therefore,

Manuscript received 11 September 2022; revised 20 January 2023 and 6 April 2023; accepted 17 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61973066 and Grant 61471110, in part by the Major Science and Technology Projects of Liaoning Province under Grant 2021JH1/10400049, in part by the Foundation of Key Laboratory of Aerospace System Simulation under Grant 6142002200301, in part by the Foundation of Key Laboratory of Equipment Reliability under Grant WD2C20205500306, and in part by the Major Science and Technology Innovation Engineering Projects of Shandong Province under Grant 2019JZZY010128. The Associate Editor for this article was M. Yang. *(Corresponding author: Yunzhou Zhang.)*

Rui Tian, Yunzhou Zhang, Zhenzhong Cao, Jinpeng Zhang, and Linghao Yang are with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: zhangyunzhou@mail.neu.edu.cn).

Sonya Coleman and Dermot Kerr are with the School of Computing, Engineering and Intelligent Systems, Ulster University, BT52 1SA Coleraine, U.K.

Kun Li is with the DAMO Academy, Alibaba Group, Hangzhou 311121, China.

Digital Object Identifier 10.1109/TITS.2023.3281837

we propose a robust data association method considering occlusions and partial observations.

Dynamic objects such as moving vehicles are a challenge for object SLAM, moreover, incorrect correspondence of different motion states will lead to false object association and mapping results. Typical approach [1] treat dynamic features as outliers by using epipolar constraints. However, it suffers when the camera is rotating or moving in the same direction as the target. Previous work [32] proposed a method that detects dynamics based on the semantic segmentation. However, the potential static motion state is not considered, which degrades the performance of pose estimation. Recent quadric-based SLAM systems [3], [4], [5], [6], [7] still assume a static environment where the relative positions between objects do not change, degrades the performance of object association. Therefore, we propose a robust data association algorithm capable of detecting motion states while associating objects in consecutive frames.

In summary, the aim of this work is to propose a novel SLAM system for outdoors that builds maps with accurate quadric landmarks, representing the location, orientation, and shape of the objects. Challenges to be addressed include:

- Current quadric initialization methods are not robust under partial observations and planar assumptions, so there are no practical solutions for robust quadric initialization in outdoors;
- Object data association remains an unsolved problem in object SLAM, especially for dynamic outdoors with occlusions;
- An efficient system architecture for joint optimization of objects, camera poses, and point clouds has not been proposed.

A. The Main Contributions of This Work Are

- A novel quadric initialization algorithm based on the optimization of constrained quadric is proposed, which does not involve planar assumptions and effectively overcomes the observation noise.
- An additional thread is used to perform ellipsoid parameter refinement within a local sliding window of keyframes, further enhancing the accuracy of quadric mapping.
- An object association algorithm is proposed that can detect motion states while associating objects across frames, improving the robustness and accuracy of object association in dynamic scenes.
- A real-time object SLAM system is implemented, incorporating the proposed quadric initialization and mapping algorithms and the object association algorithm. The system aims to build an object-oriented and semantically-enhanced map for dynamic outdoors.

II. Related Work

A. Object Representation

Object representation methods are commonly categorized into the prior object model and the generic model. Prior object model methods establish observation constraints through prebuilt databases [18], [19]. However, the prebuilt models need to be known in advance, which limits their applicability to specific scenarios. On the other hand, the generic model methods can be divided into parametric and nonparametric approaches. Parametric approaches represent objects using regular 3D forms such as cubic boxes [16], [17] and dual quadrics [5], [6], [7], [8], [9], [10]. These methods tightly constrain the parameters of the 3D model by using the 2D bounding box corresponding to the object. Dual quadrics, in particular, have gained attention due to their compact mathematical parameterization and ease of manipulation within the framework of projective geometry [5]. Recent research has focused on using dual quadrics as object representations to improve system robustness and localization accuracy [5], [6], [7], [8], [9], [10]. In contrast, nonparametric approaches reconstruct and represent objects using a combination of geometric structures such as surfels [15], [27], [33], voluments [34] and clusters [28], [35]. These approaches describe objects in more detail, but at the cost of more memory and computation, which limits their applications.

B. Object Initialization

Unlike a cubic box, a dual quadric can be represented by a symmetric matrix with nine degrees of freedom [5]. In terms of quadric initialization, Nicholson et al. [5] proposed a method using 2D detection boxes of keyframes. However, the closed-constrained parameterization and limited viewing angles make it sensitive to observation noise. In [14], multiple constraints combined with points, surfaces and quadrics are used in the optimization framework, but the prior shape of the object is estimated by deep networks, which is computationally expensive. Cao et al. [7] fused object detection and surfel construction in the quadric initialization to overcome the limitations of multi-frame and large-view observations. However, the initialization method is designed for RGB-D systems in indoor scenes. Recent work [9] proposed a refinement-based quadric initialization method that initially reconstructs the object as a sphere. However, the method is not robust for outdoor scenes due to the frequent occurrence of false object associations. In [4], texture plane and shape prior constraints are added to the quadric initialization for outdoor scenes. However, the assumption that the texture plane is parallel to the image plane during initialization makes it sensitive to partial observations. Recently, the decoupling algorithms for quadric initialization based on the planar assumption have been proposed [6], [8], [16]. However, the planar assumption does not hold in real-world scenarios, especially when a vehicle is on a slope, leading to errors between the estimated vehicle rotation and the ground truth.

C. Object Association

Probabilistic methods model the statistical distribution and leverage the EM algorithm to find correspondences between observed landmarks [27], [28], [29]. However, they can only handle a limited number of object instances in complex environments. For assignment-based methods,

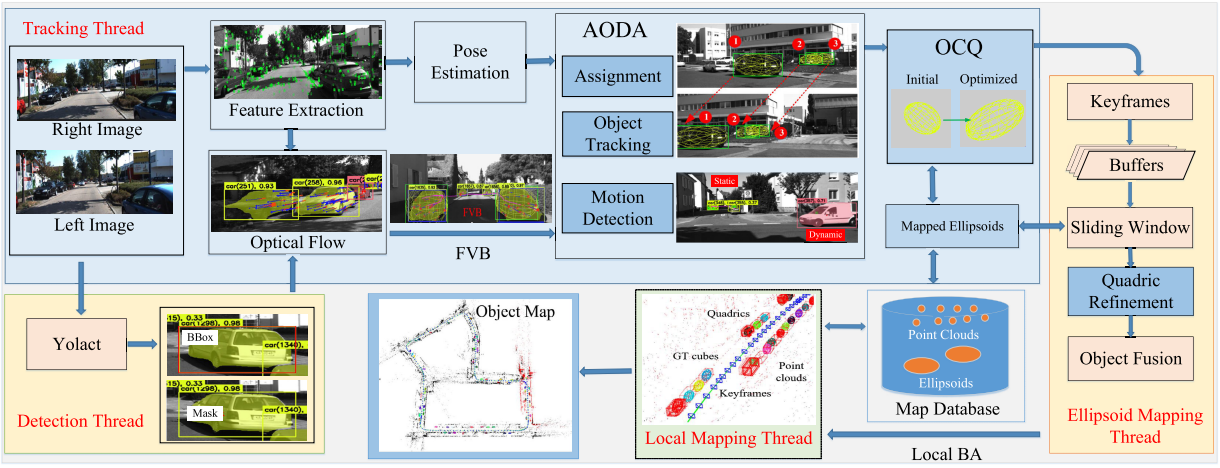


Fig. 1. The overview of the proposed system, including detection thread, tracking thread, ellipsoid mapping thread, and local mapping thread. The system combines detection results from the deep neural networks and constructs an object map of quadric representations.

Gawel et al. [31] proposed a random walk-based algorithm for assignments with semantic descriptors. Li et al. [36] established object correspondences under distinct viewpoints through graph matching. Yang and Scherer [16] proposed an efficient association method for counting the number of matched object features. However, these methods do not consider the effects of dynamics. Recently, Cao et al. [7] proposed a joint data association method that combines mixed information, but it was only applicable to indoor environments. In OA-SLAM [9], only the 2D IoU distance between the detection of consecutive frames and projection is used for object association, which works well for indoor scenes but is not suitable for dynamic outdoor scenes. Recent research has focused on dynamic objects in SLAM [3], [4], [5], [16], as they have a significant impact on the commonly used IoU-based association methods, which can lead to incorrect constraints and mapping results. Ballester et al. [37] considered potential dynamic objects by segmenting instances and photometric re-projection errors. Other work involved deep learning methods for dynamic instance segmentation [35], [38]. VDO-SLAM [39] tracked the dynamic objects and integrated dynamic information into a unified SLAM framework. However, these methods cannot be efficiently implemented due to bottlenecks imposed by the limited frequency of deep networks. Therefore, efficient and robust data association algorithms that can handle dynamic objects in real-time are urgently needed.

III. System Overview

A. Mathematical Definition

For convenience of description, the notations are as follows:

- $(\cdot)_w$ - world coordinate, $(\cdot)_c$ - camera coordinate, $(\cdot)_o$ - object coordinate.
- $D_i \triangleq \{M, B\}$ - Detection result, M and B are the semantic mask and detection bounding box (BBox).
- $In(M, x)$ - Check for features x located in M.
- $q = [a_x, a_y, a_z, t_x, t_y, t_z, \theta_x, \theta_y, \theta_z]^T \in \mathbb{R}^{9\times1}$ - the 9D vector representing the attributes of the ellipsoid, including semi-axial length, translation and rotation. The dual quadric is denoted by $\mathbf{Q}^* \in \mathbb{R}^{4\times4}$, with the projection dual conic denoted by $\mathbf{C}^*$,

$$\mathbf{C}^* = \mathbf{P}\mathbf{Q}^*\mathbf{P}^T, \tag{1}$$

where $\mathbf{P} = \mathbf{K}\mathbf{T}$ is the camera projection matrix contains intrinsic $\mathbf{K}$ and extrinsic camera parameters $\mathbf{T}$.

B. System Architecture

The overview of the proposed system is shown in Fig.1. We implement our algorithms based on the stereo configuration of ORB-SLAM3 [1] with additional modules of detection thread and ellipsoid mapping thread. The detection thread uses Yolact [22] to acquire semantic detections on the left image of the stereo pair, and output results are object BBoxes and semantic masks. The tracking thread takes images and estimates camera poses from consecutive frames using ORB features. The flow vector bound (FVB) algorithm combines optical flows to detect object motion, and further, objects are tracked with their motion states across frames through the automatic object data association (AODA) algorithm, i.e., objects are associated in 2D and reconstructed in 3D. The ellipsoids are initialized by the optimization of the constrained quadric (OCQ) algorithm and inserted into the map and are continuously refined in subsequent threads. The ellipsoid mapping thread performs ellipsoid parameters refinement in a sliding window of keyframes and runs in a parallel thread for system efficiency. The local mapping thread jointly optimizes quadric landmarks, camera poses, and map points in a nonlinear optimization of the bundle adjustment (BA) framework. The map database stores and updates map points, as well as optimized ellipsoids. The final perception results include a point clouds map and an object map for high-level applications such as localization and navigation.

IV. Optimization of Constrained Quadric

In outdoor scenes, object detection results face the problem of occlusion and failure, and State-of-the-art quadric initialization algorithms are sensitive to observation noise.

Meanwhile, the commonly used decoupling algorithms for quadric initialization rely on planar assumptions [6], [8], which limits their real-world applications. To solve these problems, we propose the OCQ algorithm for robust and accurate quadric initialization.

We present the mathematical analysis of dual quadric parameters to illustrate the OCQ algorithm. The parameters of a dual quadric can be decomposed by eigen-decomposition in a reference camera coordinate by Eq.(2),

$$\begin{aligned}\mathbf{Q}_w^* &= \mathbf{T}\mathbf{Q}_\mathbf{o}^*\mathbf{T}^T \\ &= \begin{bmatrix}\mathbf{R} & \mathbf{t}\\ \mathbf{0^T} & 1\end{bmatrix}\begin{bmatrix}\mathbf{A} & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}\mathbf{R}^T & \mathbf{0}\\ \mathbf{t}^T & 1\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{RAR}^T - \mathbf{tt}^T & -\mathbf{t}\\ -\mathbf{t}^T & -1\end{bmatrix} \triangleq \begin{bmatrix}\mathbf{Q}_{33}^* & -\mathbf{t}\\ -\mathbf{t}^T & -1\end{bmatrix},\end{aligned} \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{3\times3}$ is the diagonal matrix composed of the square of semi-axis of the ellipsoid, and $\mathbf{t} \in \mathbb{R}^{3\times1}$ is the center translation vector of the ellipsoid in the reference camera coordinate.

Note that the parameters of the block matrix $\mathbf{Q}_{33}^* \in \mathbb{R}^{3\times3}$ couple the rotation and translation of the ellipsoid. Since the length of the centroid translation is much larger than that of the rotation and axes, small errors in the estimation of the centroid translation have a significant impact on the accurate estimation of the dual quadric matrix, which is why the initialization method of [5] is sensitive to observation noise, which often occurs when the viewing angle is limited and the detection observation occluded in outdoors.

To avoid this problem, the initialization method of quadric parameters separation is proposed in our prior work [6]. The method estimates the translation vector independently to eliminate the effect of coupling parameters. Based on the assumption that autonomous vehicle is in the road plane, the yaw rotation is estimated and pitch and roll are constant to zero in reference camera coordinates. The constrained rotation matrix is given by Eq.(3),

$$\mathbf{R} = \begin{bmatrix} cos\theta_y & 0 & sin\theta_y \\ 0 & 1 & 0 \\ -sin\theta_y & 0 & cos\theta_y \end{bmatrix}. \tag{3}$$

Thus, the dimensions of quadric parameters q to be estimated are reduced from nine to seven, and the independent translation estimation improves the numerical stability of the initialization, as demonstrated in [6]. However, this assumption does not hold in real scenarios, especially when the vehicle is on a slope, causing the estimated vehicle rotation to be different from the ground truth, as illustrated in Fig.2. More experimental results of challenge cases are detailed in SectionVIII-D.

Herein, we propose the OCQ initialization in which the quadric is first initialized with equal axes (sphere) and then refined in the form of a constrained quadric as more detections are observed. The position of the ellipsoid is triangulated from centers of associated boxes, and axes are determined by the mean BBox size back-projected to the position of the center

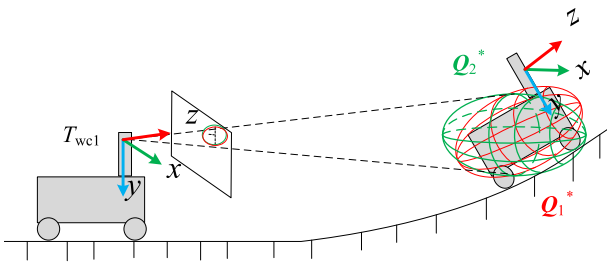


Fig. 2. Illustration of the planar assumption limitation. The rotation of estimated quadric (green) is different from the ground truth (red) due to the limitation of the rotation constraint with pitch and roll being zero.

of the ellipsoid, such that semi-axial length is obtained by,

$$a = \frac{1}{4n}\sum_{i=1}^{n} t_{iz}\left(\frac{w_i}{f_x} + \frac{h_i}{f_y}\right), \tag{4}$$

where t_{iz} is the depth of the ellipsoid center in the i-th camera coordinate, w_i and h_i are the pixel width and height of the 2D BBox, f_x and f_y are camera focal lengths and n is the number of observation associated to the object.

Observations of BBox centers form an overdetermined equation, and the RANSAC method is used to solve $\mathbf{t} \in \mathbb{R}^{3\times1}$. To ensure the accuracy of triangulation, the minimum angle and maximum distance of object observation for consecutive frames are set to 15 degrees and 50 m, respectively.

As the coarse initialization is associated with new observations, the ellipsoid parameters are refined with $q \in \mathbb{R}^{9\times1}$ and axes, positions and rotations are updated through nonlinear optimization. Note that the minimum eigenvalue of Q^* is constrained to -1 to avoid the degeneracy of quadrics. The nonlinear optimization is given by:

$$\begin{aligned} q &= \arg\min_q \{\sum_i \|e(B_i, q)\|^2_{\Sigma_o} + \|\mathbf{A} - \tilde{\mathbf{A}}\|^2_{\Sigma_a}\}, \\ e(B_i, q) &= B_i - \mathcal{B}(\mathbf{P}_i\mathbf{Q}^*\mathbf{P}_i^T), \end{aligned} \tag{5}$$

where $e(B_i, q)$ is the re-projection error between detection BBox and projection BBox. $\mathcal{B}$ is an operation for calculating the ellipse tangent bounding box [5]. $\mathbf{A}$ is the semi-axes of the ellipsoid. Note that the prior size error is used to constrain the ellipsoid shape, and the prior semi-axes is given by diagonal matrix $\tilde{\mathbf{A}} = diag(1, 1, 1/4)$.

The OCQ algorithm is a highly efficient method for quadric initialization, offering several advantages over existing approaches. Firstly, the spherical rotation does not affect the initialization result, which can be initialized as an identity matrix for efficiency, and the rotation can be further optimized with quadric parameters without relying on the planar assumption. Secondly, unlike existing methods such as [6] and [5], which only use keyframes, the OCQ algorithm can immediately associate the back-projection of the 3D initialization with potential matches in consecutive frames. This approach overcomes the limitations of keyframe-only methods, improving the robustness of initialization. Thirdly, the proposed OCQ algorithm shows significant improvement in

dynamic outdoors over the method proposed by Zins et al. [9]. The prior object error provides constraints for optimizing the constrained quadric, which prevents the solution from falling into a local minima. This, in turn, significantly improves the algorithm's robustness in outdoor environments, as demonstrated in Section VIII.

V. Ellipsoid Refinement in a Sliding Window

The aim of local ellipsoid refinement is to refine ellipsoid parameters with constraints provided by keyframes with accurate camera poses. As shown in the right module of Fig.1, an additional thread is used to refine the ellipsoid parameters within a local sliding window composed of keyframes.

For the ellipsoid stored in the local map, its re-projected ellipse can be obtained by taking the adjugate of the dual conic C^* in Eq.(1), which can be interpreted as 2D Gaussian distributions $\mathcal{N}(\mu, \Sigma)$, such that,

$$(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) = 1, \tag{6}$$

where $\mu = (c_x, c_y)^T$ is the center of ellipse and Σ is the covariance obtained by the dual conic decomposition.

Herein, motivated by the work of [11], we design the re-projection error by using the Bhattacharyya distance between two re-projected ellipses tangent to BBoxes of keyframes, which is given by:

$$\begin{aligned}\mathcal{E}(\mathcal{N}_1, \mathcal{N}_2) =& \frac{1}{8}(\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2) \\ &+ \frac{1}{2}\ln\left(\frac{\det(\Sigma_1+\Sigma_2)/2}{\sqrt{\det\Sigma_1 \det\Sigma_2}}\right).\end{aligned} \tag{7}$$

The ellipsoid parameters are optimized by minimizing re-projection errors by using observations from keyframes and associated ellipsoids stored in the local map database. Combined the center translation error, the optimization of dual quadric Q_j^* in the sling window is given by:

$$\begin{aligned}q_j = \arg\min_{q_j} \sum_{i,j} \{ &\| \mathcal{E}\left(\mathcal{N}_i, \mathbf{P}_i \mathbf{Q}_j^* \mathbf{P}_i^T\right) \|^2_{\Sigma_o} \\ &+ \| \pi(\mathbf{T}_i^{-1}\mathbf{t}_j) - u_i \|^2_{\Sigma_t} \},\end{aligned} \tag{8}$$

where $\mathcal{N}_i$ is the associated ellipse inscribed in the $i\text{-}th$ detection. $\pi(.)$ is the operation of re-projecting the ellipsoid center onto the frame, u_i is the center of the box corresponding to the keyframe and t_j is the position of the ellipsoid. The size of the sliding window buffer is n and $i \in \{1, 2, \ldots n\}$.

Note that the use of accurate pose estimations of keyframes and detection results in the sliding windows introduces more local constraints, which improves the robustness and accuracy of the quadric mapping to observation noise and partial observations. We adopt a keyframe insertion strategy similar to that proposed in [1], and to ensure the robustness of the quadric mapping, we additionally set the minimum angle of object observation for consecutive keyframes to 30 degrees.

VI. Automatic Object Data Association

Object association is of significance, since it links the correspondence between the image detection and mapped ellipsoids. The proposed AODA algorithm considers the complexity of outdoor scenes with dynamics and occlusions, and can be described as follows: (1) Object tracking; (2) Motion detection; (3) Object association by assignment.

A. Object Tracking

Object occlusions lead to false detection and hence incorrect object associations. In addition, outdoor moving objects have a significant impact on the commonly used IoU-based association methods [5], [9], [16]. To address these issues, we propose a Kalman Filter (KF) based multi-object tracking method with 2D BBox.

The displacement of a BBox with a linear constant velocity model is given by,

$$\boldsymbol{\chi}_i = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}], \tag{9}$$

where u, v are the box center and h and w are the pixel height and width of the box. $s = h \cdot w$ denotes the box scale. $r = h/w$ represents the aspect ratio of the box, which is considered as a constant. $\dot{u}$ and $\dot{v}$ are the pixel velocity at center. $\dot{s}$ is the change rate of the scale.

The state in Eq.(9) is updated by KF algorithm when the detection box of the current frame is associated with the box of the previous frame. The state is predicted without correction and a new object ID is assigned if no association is found, further details of the KF prediction can be found in [40]. To prevent incorrect associations in our implementation, the predicted BBox is compensated using the median-flow tracker [41] when the predicted and detected boxes are too far apart. The final BBox prediction $\widehat{B}_i$ is calculated with predicted state,

$$\widehat{B}_i = (\sqrt{\widehat{s}\cdot\widehat{r}}, \frac{\sqrt{\widehat{s}\cdot\widehat{r}}}{\widehat{r}}, \widehat{u}, \widehat{v}). \tag{10}$$

B. Motion Detection

For object motion detection, the FVB [42] is used to handle the case of the degenerate motion where when the camera is rotating or moving in the same direction as the target, which cannot be solved by the epipolar constraint correctly.

The object features within the semantic mask M are tracked by the Lucas Kanade (LK) optical flow [43], and normalized features of consecutive frames are denoted as x_i and x_{i+1}, respectively. The motion state of the object feature x_{i+1} is defined by,

$$\begin{aligned}S_{i+1} &= \begin{cases} 0, & d < 0 \\ 1, & otherwise \end{cases} \\ d &= (x_i + \frac{\mathbf{Kt}}{z} - x_{min})(x_i + \frac{\mathbf{Kt}}{z} - x_{max}) \\ x_{min} &= x_i + \frac{\mathbf{Kt}}{z_{max}},\ x_{max} = x_i + \frac{\mathbf{Kt}}{z_{min}},\end{aligned} \tag{11}$$

where S_{i+1} is the motion state, the binary values 0 and 1 denote the static and dynamic states of an object feature, respectively. $\mathbf{t}$ is the relative camera translation, and the depth z of the object feature is obtained by the stereo matching algorithm [1].

For our implementation, we set z_{max}=∞ and z_{min}=0.5 m as the upper and lower bounds on the possible depth of object points, respectively. We identify dynamic features as those exhibiting optical flow displacement beyond the predefined x_{min} and x_{max} boundaries. The initial detection of the motion state of the object feature is performed using the epipolar constraint. The ratio of dynamic LK features is calculated and used to update the motion state of the object through Bayesian updates, as described in [42].

C. Object Association by Assignment

The Hungarian algorithm [44] is used to complete the assignment of M detections with N ellipsoids through the cost matrix $\mathbb{A} \in \mathbb{R}^{M\times N}$, formed by elements a_{ij} with different weighted distances and are given by Eq.(12). In our implement, we set θ_i as 2, 1 and 1, respectively. We define B_k^t as the k-th detection BBox of t-th frame, and Q_j^* is the j-th ellipsoid in the map database, the aim of this process is to assign the current detection to the ellipsoid in the map, and associate BBox for KF prediction.

$$a_{ij} = \theta_1 a_{ij}^s + \theta_2 a_{ij}^d + \theta_3 a_{ij}^p. \tag{12}$$

1) Semantic Inliers Distance: To overcome the issue of object occlusions, we track the ratio of the LK feature set $\{x_j^{t-1}\}$ of the last frame within the feature set $\{x_i^t\}$ of the current detection mask M_i^t,

$$a_{ij}^s = \frac{\textbf{size}(In(M_i^t, x_j^{t-1}))}{\textbf{size}(\{x_i^t\})}. \tag{13}$$

2) Detection IoU Distance: We use the IoU between the re-projection BBox of j-th ellipsoid and i-th detection.

$$a_{ij}^d = \textbf{IoU}(B_i^t, \mathcal{B}(\mathbf{P}_i \mathbf{Q}_j^* \mathbf{P}_i^T)). \tag{14}$$

3) Prediction IoU Distance: We use the IoU between the predicted BBox of Eq.(10) and i-th detection.

$$a_{ij}^p = \textbf{IoU}(B_i^t, \widehat{B}_j^t). \tag{15}$$

The detection IoU distance of Eq.(14) provides long-term associations of objects, and the prediction IoU distance of Eq.(15) provides short-term associations of detection in consecutive frames. The assignment results are used for both initialization and KF correction. Note that the unsigned detection will be processed by the OCQ algorithm as a new quadric instance. If there is no association after more than 10 frames, the instance will be discarded.

VII. Joint Optimization in Local Mapping Thread

In the local mapping thread, we optimize ellipsoids, camera poses and static map points in a joint optimization framework. Using the observations omitting dynamics proposed by VI-B, the static map points P_k, feature observations u_k and the ellipsoid parameters q_j formulated the joint BA optimization problem by Eq.(16),

$$\begin{aligned}\mathbf{T}_i, P_k, q_j = \underset{\{\mathbf{T}_i, P_k, q_j\}}{\arg\min} \{ &\sum_{i,j} ||B_i - \mathcal{B}(\mathbf{P}_i \mathbf{Q}_j^* \mathbf{P}_i^T)||_{\Sigma_o}^2 \\ &+ \sum_{i,k} ||\pi(\mathbf{T}_i^{-1} P_k) - u_k||_{\Sigma_m}^2 \},\end{aligned} \tag{16}$$

where Σ_o and Σ_m are the covariance of BBox projection error and static map point projection error, respectively. Σ_m follows the similar strategy of multi-scale image pyramid as proposed by [1]. The Huber kernel and the Levenberg-Marquardt algorithm are use for optimization.

VIII. Experiments

In this section, we present the evaluation results of our proposed system and compare it with other state-of-the-art techniques, including [1], [3], [4], [5], [6], [9], [15], [16], and [32], on the KITTI dataset [45]. The KITTI tracking sequences were selected to validate multi-object tracking performance, while KITTI raw data sequences with the most ground truth (GT) object annotations were used to evaluate the quadric initialization and mapping performance, both quantitatively and qualitatively. Additionally, The camera localization accuracy was evaluated using KITTI odometry sequences with GT camera trajectories.

To ensure a fair comparison, we re-implemented the approaches of [5] and [9] with stereo configurations since the open-source codes of [5] and [9] were monocular versions. We set the 2D object detection confidence threshold for Yolact to 0.6, and the frame number of the sliding window in the ellipsoid mapping thread to 15. All experiments were conducted on a system with an Intel(R) Core(TM) i7-9700 CPU@3.00GHz, 16G memory, and Nvidia GTX 1080 Ti.

A. Evaluation Metrics

To evaluate the muti-object tracking, we follow the widely used CLEAR MOT metrics proposed in [46].

The Multi-Object Tracking Accuracy (MOTA) metric measures the overall tracking accuracy, m_t, fp_t, mme_t and g_t represent the number of misses, false positives, mismatches, and ground truth respectively.

$$\text{MOTA} = 1 - \frac{\sum_t m_t + fp_t + mme_t}{\sum_t g_t}. \tag{17}$$

The Multi-Object Tracking Precision (MOTP) metric measures the object localization precision, d_t^i represents the distance between a detection and its corresponding ground truth, and c_t is the number of all matches found.

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}. \tag{18}$$

We evaluate the constructed ellipsoids by the criteria of the translation error (TE), the axial length error (AE), the success ratio of initialization (SR), and the 2D IoU, which are defined as follows,

- TE: The average translation error between the position of a GT annotation and a constructed ellipsoid,

$$\text{TE} = \frac{1}{n} \sum_i \|\mathbf{t}_{gt} - \mathbf{t}_i\|^2. \tag{19}$$

- AE: The average axial length error between the size of a GT annotation and the ellipsoid,

$$\text{AE} = \frac{1}{n} \sum_i \|\mathbf{a}_{gt} - \mathbf{a}_i\|^2. \tag{20}$$

TABLE I
THE COMPARISON OF MULTI-OBJECT TRACKING PERFORMANCES WITH MOTA(%) AND MOTP(%) METRICS ON KITTI TRACKING BENCHMARK

Sequence	Cars	Ours		ODA [6]		Ours (w/o Semantics)		OA-SLAM [9]	
		MOTA(%)	MOTP (%)	MOTA(%)	MOTP(%)	MOTA(%)	MOTP(%)	MOTA(%)	MOTP (%)
0003	8	**46.38**	**74.85**	37.33	69.79	30.62	66.14	22.74	66.02
0004	40	**55.99**	**71.28**	42.45	67.03	20.06	69.97	20.39	66.19
0005	35	41.88	**67.14**	**49.80**	66.57	29.06	66.80	28.06	66.58
0010	21	**46.35**	**87.06**	36.70	80.95	21.36	71.43	23.89	86.33
0011	59	**32.13**	**81.39**	21.45	79.16	12.40	71.48	12.52	70.40
0018	20	**42.13**	**62.28**	35.30	59.87	22.40	60.79	21.30	61.34
0019	88	**45.96**	**74.02**	45.30	69.75	23.67	67.36	12.03	67.28
0020	133	35.10	80.19	**39.90**	**80.56**	28.39	76.61	30.62	78.16
Mean	50	**43.24**	**74.78**	38.53	71.71	23.50	68.82	21.42	70.29

Fig. 3. Visualization of multi-object tracking results of the AODA algorithm on KITTI tracking sequences, where 2D BBoxes are marked with Object IDs and confidence scores located at the top. The consistency of Object IDs across frames indicates the algorithm's accurate tracking performance.

- SR: The 2D IoU between projection and GT greater than 0.5 is defined as successfully initialized. The ratio is the number of successfully initialized ellipsoids divided by the number of GT annotations in the reference frame.
- 2D IoU: The intersection ratio of the 2D BBox projection of the successfully constructed ellipsoid and the GT annotation in the reference frame.

B. Multi-Object Tracking Evaluation

The KITTI tracking sequences contain images captured by cameras mounted on vehicles moving in dynamic environments. These frames are annotated with various object classes such as 'Car' and 'Pedestrian'. In our experiment, we evaluate the multi-object tracking performance using only the 'Car' object tracks. Specifically, we select 8 sequences that feature self-motion, as well as static and dynamic observed vehicles in the environments. The number of vehicles present in each sequence can be found in Table I.

Fig.3 shows the results of multi-object tracking with object ID and confidence score, where the red lines represent objects tracked in consecutive frames in highly dynamic scenes. The consistent ID results demonstrate the accuracy of object tracking with 2D BBoxes. Additionally, Fig.4 shows the motion detection results with 3D projected red cube GT annotations, where red and yellow 2D boxes denote dynamic and static detection results, respectively. The same object ID is maintained across frames, and the vehicle's motion state is accurately estimated. Static vehicles are successfully reconstructed as quadric landmarks, and the magenta line demonstrates the accuracy of object data association by showing the projection of the object center onto the current frame.

Table I presents the quantitative results of multi-object tracking with the MOTA and MOTP metrics, where larger values indicate better results. Our proposed AODA algorithm outperforms the compared methods in most cases. Compared to the methods of [6] and [9], the average MOTA metric increases by 12.22% and 101.87%, and the average MOTP metric increases by 4.28% and 6.38%. Moreover, our MOTA is larger than ODA's [6] for sequences with a large number of moving vehicles (e.g., sequence-0003, -0004, -0010, -0011, -0018, -0019), which demonstrates the effectiveness of motion detection and object tracking. In addition, we evaluate the proposed system using only 2D IoU distance without semantic inlier distance for object association (denoted as w/o Semantics). Compared with this method, our average MOTA and MOTP metrics improved by 84.03% and 8.65%, respectively. The improved performance is attributed to the robust tracking of semantic features in the object data association. In OA-SLAM [9], only 2D IoU distance between detection of consecutive frames and projected mapped ellipsoids are used for object association, which works well for indoor scenes but is not suitable for dynamic outdoor scenes. The relatively low MOTA metric of OA-SLAM [9] indicates the

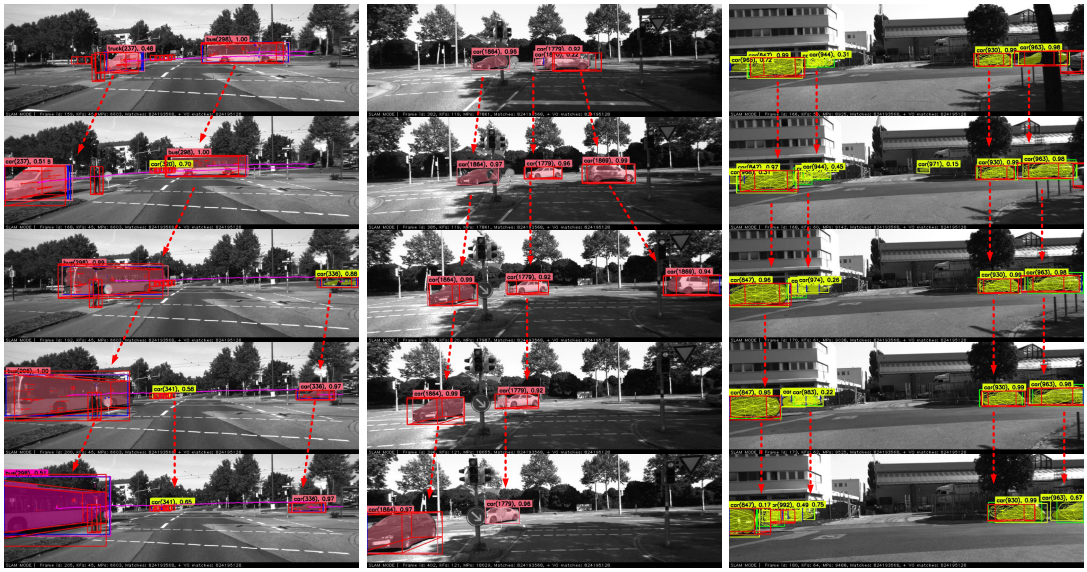


Fig. 4. Visualization of multi-object tracking results with ground truth annotations. Red and yellow boxes denote the detected dynamic and static objects. The left magenta lines show the projection of vehicles' centers. Static vehicles are reconstructed as ellipsoids and participate in the object association.

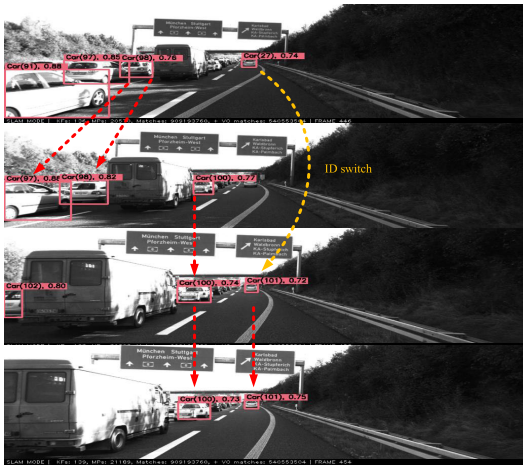


Fig. 5. Visualization of the ID switching of object tracking. The results show the effect of omitted distant detection results on the tracking process, resulting in the ID switching of vehicle 27 to 101, which usually occurs with moving objects.

instability of object tracking and association performance in dynamic outdoors, which also affects the accuracy of quadric mapping and is demonstrated with quantitative evaluation in Section VIII-E.

Note that in our proposed system, distant object detection is omitted to ensure quadric initialization accuracy, as distant feature depth estimation is usually inaccurate and leads to object ID switching, as shown in Fig.5. This strategy is a trade-off between the quadric mapping accuracy and the multi-object tracking accuracy, which degrades the overall performance of the multi-object tracking, especially for the MOTA metric. However, it does not affect static object initialization as object ID switching tends to occur during dynamic object tracking.

C. Quadric Initialization Evaluation

The KITTI raw data sequences consist of images taken by cameras mounted on vehicles moving in urban environments. These sequences provide object annotations for the vehicles that appear in each frame and the GT cubes can be obtained from these annotations. Fig.6 presents the results of quadric initialization on the KITTI raw data sequence-0009 using comparison initialization methods, including the Conic method of Rubino et al. [3], the QuadricSLAM method of Nicholson et al. [5], the ROSHAN method of Ok et al. [4], the SQP method of Tian et al. [6], the OA-SLAM method of Zins et al. [9] and our proposed OCQ algorithm.

Better convergence of the quadric projection and more enveloped vehicles indicate better initialization performance, as shown in Fig.6. The red cubes denote the GT annotations projected onto the current frame using the extrinsic parameters of the sensor calibration provided by [45], and yellow ellipses represent the initialized ellipsoids. The results show that the proposed algorithm (f) has a higher number of accurate quadric envelopes than the comparison algorithms (a)-(e), demonstrating the effectiveness of the proposed OCQ algorithm. In particular, the reconstruction results of the comparison algorithms are not ideal due to observation noise and partial observations in the scenes, resulting in significant scale and rotation errors, especially for the Conic method (a) and QuadricSLAM (b). Compared with the results of OA-SLAM (e), our proposed method (f) shows advantages in the scale and orientation of initialized quadrics. Compared with the SQP algorithm (d), our method (f) shows competitive results in quadric scale and orientation, while our OCQ method demonstrated higher initialization success rates and more accurate quadric mapping results due to the refinement of the ellipsoid parameters, as shown in Table II with quantitative results and analysis.

Table II provides a comparative analysis of the quadric initialization performance of the proposed OCQ algorithm against state-of-the-art methods, including [3], [4], [5], [6], and [9], based on SR and 2D IoU metrics, evaluated on the KITTI raw data sequences. Higher values indicate better initialization performance. The proposed algorithm outperforms the comparison methods in terms of SR, achieving the

(a) Conic method (b) QuadricSLAM (c) ROSHAN (d) SQP method (e) OA-SLAM (f) Ours

Fig. 6. Visualization of the quadric initialization results of comparison algorithms, the proposed algorithm (f) has a higher number of accurate quadric envelopes than the comparison algorithms (a) [3], (b) [5], (c) [4], (d) [6], (e) [9], demonstrating the effectiveness of the proposed OCQ algorithm. Red cube: GT annotation; Ellipse: projection of the quadric initialization; Bound box with ID: object detection result.

TABLE II

QUADRIC INITIALIZATION PERFORMANCE: THE COMPARISON OF THE SUCCESS RATIO (%) AND THE 2D IOU ON KITTI RAW DATA SEQUENCES

Sequence	Ours		OA-SLAM [9]		SQP [6]		ROSHAN [4]		Conic [3]		QuadricSLAM [5]	
	SR(%)	2D IoU	SR(%)	2D IoU	SR(%)	2D IoU	SR(%)	2D IoU	SR(%)	2D IoU	SR(%)	2D IoU
0009	**91.22**	**0.8052**	56.09	0.6417	69.12	0.7335	40.60	0.5240	44.68	0.7252	27.06	0.7031
0022	**87.44**	**0.8108**	41.82	0.5309	65.12	0.7629	75.02	0.6112	33.33	0.7791	29.23	0.7662
0023	**88.97**	**0.7993**	28.58	0.4881	62.30	0.7509	68.62	0.5462	39.49	0.7529	28.29	0.6959
0036	**76.02**	0.7582	22.27	0.6139	60.47	**0.7604**	67.42	0.5992	40.96	0.7558	35.14	0.7127
0059	**87.64**	**0.8098**	53.42	0.6886	56.25	0.6508	44.81	0.5223	35.56	0.6878	25.58	0.6500
0093	**81.51**	**0.8027**	36.58	0.5917	48.15	0.7232	54.51	0.5510	28.26	0.6751	36.17	0.7433
Mean	**85.47**	**0.7976**	39.79	0.5925	60.23	0.7303	58.50	0.5592	37.05	0.7293	30.24	0.7119

Fig. 7. Visualization of challenging cases with partial observations and occlusions on the KITTI datasets. The upper plots show the initialized quadric reconstruction results, while the lower plots show the corresponding camera inputs. The proposed system successfully initializes the observed vehicles with quadric representations despite the challenges posed by overlapping object detection.

highest value of 91.22% on the sequence-0009. The average SR metrics of the proposed method increase by 114.80%, 41.90%, 46.10%, 130.68%, and 182.26% over the comparison methods, respectively. The improved performance of the OCQ algorithm is attributed to the stable spherical initialization, which leverages the constraints from consecutive frames with reliable observations, as opposed to the comparison algorithms [4], [5], [6] that rely only on keyframes. In terms of 2D IoU, our method outperforms other algorithms except for sequence-0036 with the improvement of 34.62%, 9.21%, 42.63%, 9.36%, and 12.04%, respectively. The larger IoU of the SQP method [6] on sequence-0036 could be attributed to its strict keyframe selection strategy, which discards some detection results that failed to initialize. The failed

ellipsoid initialization estimations result in smaller SR and 2D IoU for OA-SLAM [9]. Compared with [9], the significant improvement of the proposed algorithm can be attributed to the scale constraint and the accurate object association, which ensures a more stable solution for the constrained ellipsoid during the non-linear optimization. These results demonstrate that the proposed OCQ algorithm achieves high accuracy and robustness of quadric initialization in comparison to state-of-the-art methods.

D. Partial Observation and Slope Cases

Fig.7 showcases the challenging scenarios of partial observation and occluded vehicles that are often encountered on the KITTI odometry sequences and KITTI raw data sequences. Ellipsoid reconstruction on these sequences is particularly challenging due to the narrow range of angles between vehicle and camera, which is almost limited to the azimuth plane. Furthermore, each vehicle appears in a limited subset of frames. As shown in Fig.7, the raw input frames are presented along with detection results, GT annotations, and quadric initialization projections. The yellow boxes denote detection results, the green boxes correspond to 2D BBoxes tangent to the projected quadric initialization, while the red cubes denote the GT annotations. Vehicles are parked on both sides of narrow roads, and occlusions are present. Some detection boxes are obscured or overlapped, and distant vehicles with limited viewing angles are starting to initialize and envelope the vehicle. Despite the challenges posed by overlapping object detection, the proposed system successfully initializes the observed vehicles with quadric representations.

In Fig.8, we compare the quadric initialization results of the proposed OCQ method with the SQP method [6] when the vehicle is driving downhill on the KITTI odometry sequence-10, and the observed vehicle is parked on the slope. The left plots show the quadric projection results of the proposed system and the right red dashed line marks the ground plane of the slope. It is evident that the pitch angle of the reconstructed ellipsoid under the rotation constraint of the SQP is aligned with the reference camera, resulting in rotation errors, as shown in Fig.8 (2). In contrast, our reconstructed ellipsoid's pitch angle remains parallel to the actual ground plane, as shown in Fig.8 (1). These results demonstrate that the proposed OCQ algorithm can overcome the problem of incorrect initialization when the planar assumption is not valid.

E. Quadric Mapping Evaluation

1) KITTI Odometry Sequences: Fig.9 shows the object map of the KITTI odometry sequence-07 with details. Vehicles in the frames have corresponding reconstructed ellipsoids with accurate estimated positions, rotations, and sizes. The static reconstruction results remain unaffected by these dynamics due to correct motion estimation and object association, as shown in regions (3) with the dynamic vehicle marked in red. In addition, the vehicle maintaining the same speed as the camera is also correctly detected and does not affect the quadric mapping, as shown in regions (5). The reduction of dynamic effects demonstrates the robustness of the proposed system.

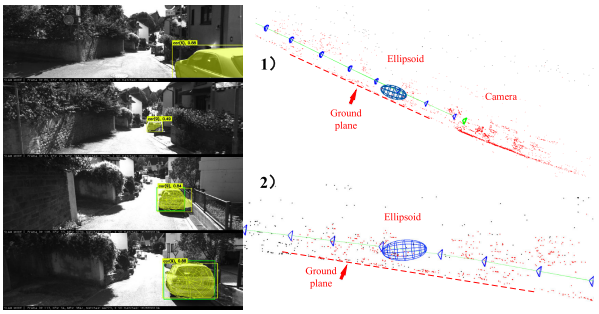


Fig. 8. Visualization of comparison mapping results when the vehicle is located on a slope. The left plots show the visualization results of our system. 1) The quadric initialization result of the proposed OCQ algorithm, where the pitch angle of the ellipsoid remains parallel to the actual ground. 2) The quadric initialization result of the SQP algorithm [6], where the pitch angle of the ellipsoid is aligned with the reference camera under rotation assumption, while the planar assumption does not hold, leading to mapping rotation errors.

Fig.10 shows the object map constructed by the proposed system and comparison algorithms, including [6], [9], and [16]. We use the offline reconstruction result of CubeSLAM [16] as baseline to show correct vehicle poses. The reconstructed ellipsoids of our system significantly outperform the SQP method [6] and OA-SLAM [9] in terms of scale, orientation, and quantity, for maps with quadric landmark representation. Note that ellipsoids reconstructed by OA-SLAM suffer from initialization failure and overlap. We contribute the improved performance of our system to the scale constraint of the OCQ algorithm and the correct object association of the AODA algorithm. These comparison results demonstrate the accuracy of the proposed system in quadric mapping.

2) KITTI Raw Data Sequences: Fig.11 shows visualization of object maps on KITTI raw data sequences. Red cubes are GT annotations for the quantitative evaluation of the reconstructed ellipsoids, and colored ellipsoids are reconstructed quadrics. Notably, some distant quadric landmarks are ignored with the GT annotations shown in the map, but this does not affect the accuracy of the camera localization. The results show that the reconstructed ellipsoids match the GT cubes in term of position, rotation and quantity, demonstrating the high accuracy and success rate of the proposed system.

Table III shows the results of TE and AE of the compared methods, including [3], [4], [5], [6], and [9]. The smaller values indicate better reconstructed results. It can be seen that our proposed method outperforms the comparison methods in all cases, with an average TE of 0.8184 m, reducing the errors by 136.58%, 159.98%, 168.11%, 311.70%, and 444.13%, respectively. In addition, our average AE is 0.5355 m, with an error reduction of 432.21%, 19.87%, 179.53%, 155.70%, and 77.01% for the compared algorithms, respectively. Note that the smaller AE of QuadricSLAM [5] compared to [4] and [3] is due to its smaller SR, which eliminates some of the results of the failed constructions. The average TE of OA-SLAM [9] is smaller than that of other methods [3], [4], [5], [6], which could be attributed to its initialization method of coarse sphere estimation, yielding more accurate

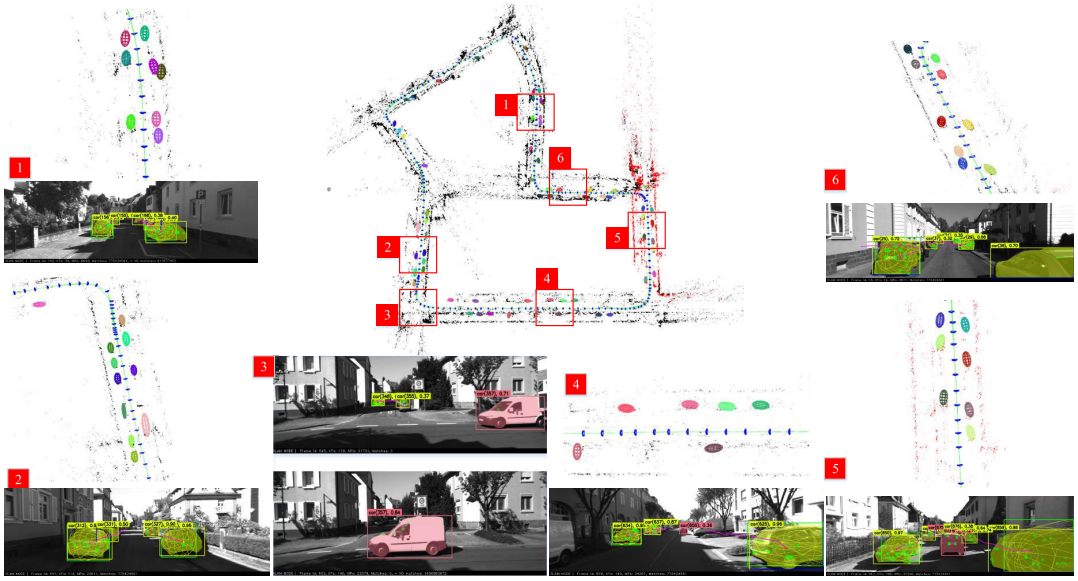
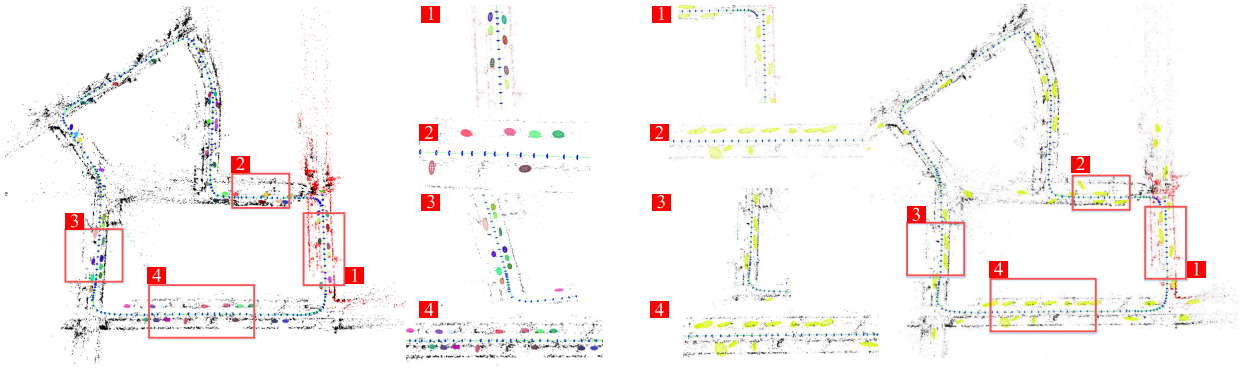
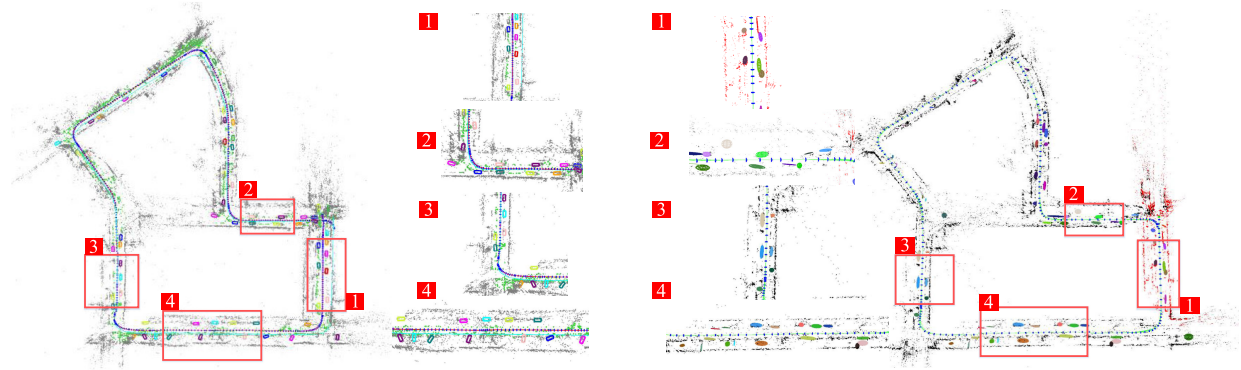


Fig. 9. Visualization of the object map on the KITTI odometry sequence with details. The static reconstruction results remain unaffected by these dynamics due to correct motion estimation and object association. This demonstrates the robustness of the proposed system.



(a) Ours

(b) SQP method [6]



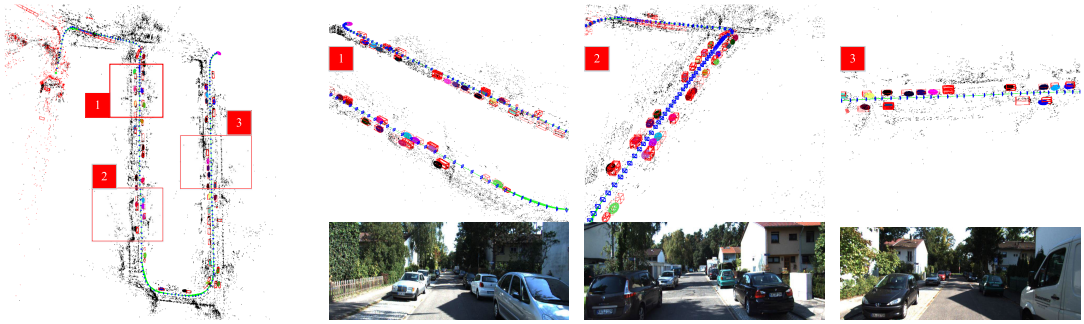
(c) CubeSLAM [16]

(d) OA-SLAM [9]

Fig. 10. Comparison of quadric mapping results. Regions (1)-(4) show local details of object mapping. Object map of [16] is a baseline to show correct vehicle poses. Our proposed system shows the accuracy of the quadric mapping in terms of scale, position and quantity, which demonstrates the advantages of the proposed system.

object centroids estimation. However, due to the lack of scale constraint, the reconstructed ellipsoid of OA-SLAM tends to fall into a local minima during optimization, resulting in an excessive scale error of the reconstructed ellipsoid, which leads to its larger AE. Compared to the SQP method [6], the improved TE metric of the proposed system benefits from

(a) KITTI raw data sequence-0009.



(b) KITTI raw data sequence-0022.

Fig. 11. Visualization of object maps with 3D GT annotations for quantitative evaluation of reconstructed ellipsoids in KITTI raw data sequences. The constructed color ellipsoid matches the GT cube in position, rotation, and quantity, demonstrating the quadric mapping accuracy of the proposed system.

the use of more local constraints in the sliding window with accurate pose estimation of keyframes.

We attribute our improvements in AE and TE metrics to the following factors. Firstly, the scale constraint of the OCQ algorithm ensures correct object shape during initialization optimization. Secondly, the AODA algorithm overcomes the effect of moving objects and ensures accurate object association. Thirdly, the sliding window-based optimization of the ellipsoid mapping thread includes more local constraints to refine ellipsoid parameters. These experimental results demonstrate high accuracy and robustness in the quadric mapping of our proposed system.

F. Localization Evaluation

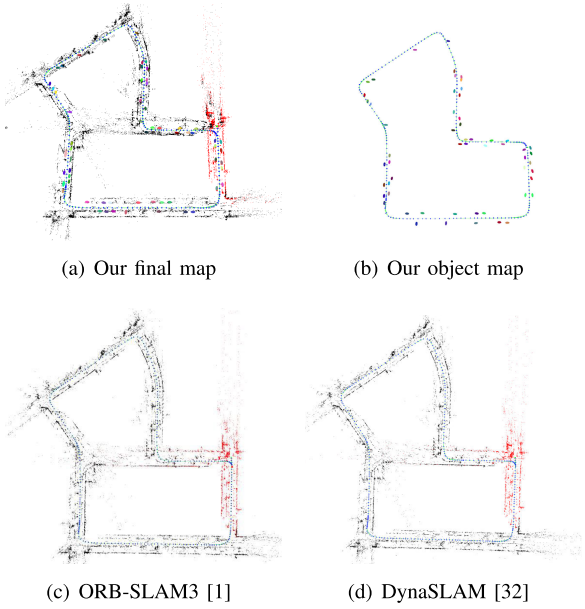
We evaluate the localization performance on KITTI odometry sequences in comparison to other state-of-the-art systems, including the object SLAM systems of [5], [6], [9], [15], and [16], the dynamic SLAM [32] and the baseline SLAM [1]. We use the criteria of the relative translation error RPE (%), the relative rotation error RRE (°/100 m) and absolute translation error ATE (m) on the KITTI odometry benchmark [45].

As illustrated in Table IV, the proposed system demonstrates a significant improvement in localization accuracy over CubeSLAM [16] and QuadricSLAM [5] in all the evaluated scenarios. In Sequence-01, which is a highway scene with moving vehicles, CubeSLAM and QuadricSLAM fail to initialize the objects, leading to tracking failure of the system. Particularly, QuadricSLAM fails to locate on the sequence-01, -04, -07, and -08, mainly due to the frequent occurrences of dynamic vehicles and quadric initialization failure. Compared to CubeSLAM and QuadricSLAM, the proposed system reduces the average ATE by 257.26% and 219.38%, respectively. These results demonstrate that for quadric initialization, the direct decomposition method of QuadricSLAM is numerically unstable, and the inaccurate object reconstruction without considering dynamics leads to degraded localization performance or tracking failure. Furthermore, compared to the baseline SLAM [1], we obtain better results with an average of RPE, RRE, and ATE metrics reduced by 5.71%, 4.54%, and 46.26%, respectively. Compared to DynaSLAM [32], our proposed system reduces the average RPE, RRE, and ATE by 7.14%, 4.5%, and 9.25%, respectively. These results indicate that integrating accurate object landmarks improves localization accuracy.

Compared to the recently proposed DSP-SLAM [15], which builds dense object maps, our proposed system reduces the average RPE by 8.45%, which we attribute to the elimination of dynamic effects. The advantages of our system are the concise quadric representation of the object map and the smaller memory footprint. In contrast, DSP-SLAM [15] utilizes dense mesh grid reconstruction and GPU with high computation. Moreover, compared to OA-SLAM [9], the average RPE, RRE, and ATE are reduced by 158.57%, 40.90%, and 129.07%, respectively. We attribute the improved localization performance of our proposed system to the accurate object association of the AODA algorithm and the robust quadric initialization of the OCQ algorithm, which ensure stable object observation for the ellipsoid refinement and accurate estimation of object landmarks. Finally, compared to our previous work [6], the proposed system reduces the RPE and ATE by 50.00% and 31.72%, respectively. These results indicate the effectiveness of accurate object landmark estimation in

TABLE III
QUADRIC MAPPING PERFORMANCE: THE COMPARISON OF THE TE (M) AND AE (M) ON KITTI RAW DATA SEQUENCES

Sequence	Ours		OA-SLAM [9]		SQP [6]		ROSHAN [4]		Conic [3]		QuadricSLAM [5]	
	TE(m)	AE(m)	TE(m)	AE(m)	TE(m)	AE(m)	TE(m)	AE(m)	TE(m)	AE(m)	TE(m)	AE(m)
0009	**0.4051**	**0.4096**	1.0021	1.4725	2.5456	0.6271	1.8912	1.2310	2.7819	1.2618	3.6110	1.1400
0022	**0.8196**	**0.4954**	1.9046	3.8609	2.1769	0.5565	1.9670	1.6244	1.9552	0.7233	1.9651	0.8356
0023	**0.7042**	**0.4711**	2.3930	1.6091	2.3341	0.5494	1.8992	1.2850	5.6605	1.7886	8.7088	0.6837
0036	**1.0496**	**0.5118**	1.9720	2.5901	1.8594	0.7121	3.0402	2.1181	2.7175	1.2797	6.9814	0.8799
0059	**0.8117**	**0.6395**	2.0024	3.8124	1.3276	0.6706	1.9256	1.0678	1.6883	1.8467	1.3874	1.2908
0093	**1.1204**	**0.6854**	2.3428	3.7551	2.5226	0.7357	2.4405	1.6552	5.4130	1.3156	4.0654	0.8574
Mean	**0.8184**	**0.5355**	1.9362	2.8500	2.1277	0.6419	2.1942	1.4969	3.3694	1.3693	4.4532	0.9479



(a) Our final map (b) Our object map

(c) ORB-SLAM3 [1] (d) DynaSLAM [32]

Fig. 12. In contrast to the mapping results of [1] and [32], which focus on constructing maps with sparse point clouds of the scene, our system provides object-level maps with enhanced semantic perception capabilities.

improving localization accuracy. Therefore, our proposed system improves object reconstruction by incorporating dynamic integration and accurate object association, while accurate object estimation further enhances camera localization.

Note that our proposed system shows competitive results in terms of average RPE, RRE, and ATE compared to the dynamic SLAM [32], with an improvement of the metrics by 7.14%, 4.57%, and 9.25%, respectively. The difference is that our system builds an object map as shown in Fig.12, providing semantic perception for vision-based localization. This is important for the perception and tracking of multiple objects as well as for autopilot navigation in our future work.

G. Discussion

Various factors affect the accuracy and robustness of camera localization, including the quadric initialization, ellipsoid landmark estimation, object data association, dynamic effects, and the multi-thread designed system architecture. We present a comparative analysis of the experimental results and investigate the effects of improving camera localization. The comparison of our proposed system with existing methods and the improvements in metrics are shown in Table V. Significant improvement values over 100% are highlighted.

We demonstrate that the data association algorithm used in OA-SLAM [9] is specifically designed for static environments, leading to poor performance in dynamic sequences and hindering the reconstruction of static objects, such as sequence-01, -03, -04, and -07. Through the analysis presented in Table V, we highlight the significant improvement in localization accuracy achieved by accurate quadric initialization and mapping, particularly in metrics such as SR, TE, and AE.

In direct comparison, our proposed system outperforms OA-SLAM with a 114.80% improvement in SR, a substantial 432.21% improvement in AE, and an enhanced ATE by 129.07%. Additionally, our system outperforms QuadricSLAM [5] with a 182.26% improvement in SR, a substantial 444.13% improvement in TE, and an improved ATE by 219.38%. The results emphasize the significance of accurate quadric initialization and mapping in enhancing localization performance. Furthermore, in assessing dynamic effects, we compare our proposed system with [1] and [32], confirming the positive impact of dynamic feature rejection on localization accuracy. While accurate object landmarks estimation significantly contributes to the accuracy of camera localization, it is crucial to address the negative impact of false object landmarks on localization accuracy. Notably, QuadricSLAM fails to localize in specific sequences, such as sequence-01, -04, -07, and -08.

In conclusion, the accurate object estimation is critical to enhancing camera localization accuracy. Our proposed system, incorporating OCQ and AODA algorithms, effectively improves quadric pose estimation, while accurate quadric initialization and mapping significantly enhance localization accuracy.

H. Real-Time Performance Evaluation

Table VI shows the average runing time of the frames processed by main modules, indicating our system can operate in real-time at a frequency of 10Hz. The detection thread takes on average 58.25 ms. The average time for the tracking thread is 96.08 ms, and the AODA and OCQ algorithms only take an average of 1.68 ms and 8.71 ms, respectively. The ORB extraction and LK flow take 59.67 ms and 17.35 ms, respectively. The optimization of the ellipsoid mapping thread and the local mapping thread take 164.20 ms and 372.76 ms, respectively. Since these operations are performed in parallel

TABLE IV
THE COMPARISON OF LOCALIZATION PERFORMANCES ON KITTI ODOMETRY SEQUENCES

Method	Metrics	00	01	02	03	04	05	06	07	08	09	10	Mean
Ours	RPE (%)	**0.68**	**1.18**	**0.72**	**0.68**	0.50	0.55	0.52	**0.48**	**1.01**	0.83	**0.56**	**0.70**
	RRE (°/100 m)	0.24	0.26	**0.21**	0.18	**0.09**	0.23	0.17	**0.24**	0.35	**0.20**	**0.22**	**0.22**
	ATE (m)	2.84	**7.02**	**5.72**	**0.39**	0.45	**0.63**	1.88	0.88	**2.57**	**1.62**	**1.01**	**2.27**
ORB-SLAM3 [1]	RPE (%)	0.70	1.38	0.76	0.71	0.79	0.40	0.51	0.50	1.07	**0.82**	0.58	0.74
	RRE (°/100 m)	0.25	**0.20**	0.23	**0.17**	0.26	0.16	**0.15**	0.28	0.31	0.25	0.28	0.23
	ATE (m)	3.86	13.23	6.07	0.62	1.89	0.83	1.26	0.98	3.16	3.27	1.40	3.32
DynaSLAM [32]	RPE (%)	0.74	1.57	0.80	0.69	**0.45**	**0.40**	**0.50**	0.52	1.05	0.93	0.67	0.75
	RRE (°/100 m)	0.26	0.22	0.24	0.18	0.09	0.16	0.17	0.29	0.32	0.29	0.32	0.23
	ATE (m)	**1.40**	9.41	6.75	0.65	**0.20**	0.80	**0.80**	**0.52**	3.54	1.94	1.35	2.48
Rui et.al [6]	RPE (%)	1.32	1.41	1.58	0.74	0.42	1.69	0.76	0.61	1.14	1.21	0.63	1.05
	RRE (°/100 m)	**0.22**	0.41	0.25	0.18	0.14	**0.16**	0.41	0.26	**0.27**	0.26	0.43	0.27
	ATE (m)	3.08	7.38	6.02	1.89	1.18	1.66	2.78	1.08	3.62	3.15	1.08	2.99
OA-SLAM [9]	RPE (%)	1.38	3.14	1.58	3.56	1.95	0.88	1.36	0.89	1.48	1.77	1.88	1.81
	RRE (°/100 m)	0.27	0.40	0.27	0.22	0.15	0.26	0.25	0.43	0.47	0.33	0.36	0.31
	ATE (m)	4.01	15.75	10.54	5.38	4.01	2.34	2.87	1.39	4.17	3.68	3.11	5.20
DSP-SLAM [15]	RPE (%)	0.71	1.45	0.75	0.73	0.47	0.57	0.57	0.51	1.02	0.87	0.65	0.75
	RRE (°/100 m)	0.24	0.30	**0.19**	0.19	0.11	0.23	0.22	0.29	0.32	0.26	0.31	0.24
	ATE (m)	-	-	-	-	-	-	-	-	-	-	-	-
CubeSLAM [16]	RPE (%)	2.40	*	4.25	2.87	1.12	1.64	3.20	1.63	2.79	3.16	4.34	2.49
	RRE (°/100 m)	-	*	-	-	-	-	-	-	-	-	-	-
	ATE (m)	13.94	*	26.21	3.79	1.10	4.85	7.04	2.67	10.70	10.52	8.39	8.11
QuadricSLAM [5]	RPE (%)	6.28	*	5.47	5.69	*	8.41	5.75	*	*	5.38	5.12	3.82
	RRE (°/100 m)	0.38	*	0.32	0.46	*	0.31	0.37	*	*	0.43	0.49	0.25
	ATE (m)	16.72	*	19.28	2.65	*	5.75	12.65	*	*	12.54	10.21	7.25

* Track failure; - Lack experimental data of the manuscript

TABLE V
THE CAMERA LOCALIZATION IMPROVEMENTS (CLI) OF OUR SYSTEM V.S. COMPARISON METHODS WITH EFFECT OF OBJECT ASSOCIATION, QUADRIC INITIALIZATION AND QUADRIC MAPPING

Method	Object Association		Quadric Initialization		Quadric Mapping		Camera Localization Improvements		
	MOTP*(%)	MOTA*(%)	SR*(%)	2D IoU*(%)	TE*(%)	AE*(%)	RPE*(%)	RRE*(%)	ATE*(%)
Rui et.al [6]	4.28 ↑	12.22 ↑	41.90 ↑	9.21 ↑	**159.98** ↑	19.87 ↑	50.00 ↑	22.73 ↑	31.72 ↑
OA-SLAM [9]	6.38 ↑	**101.87** ↑	**114.80** ↑	34.62 ↑	**136.58** ↑	**432.21** ↑	**158.57** ↑	40.90 ↑	**129.07** ↑
QuadricSLAM [5]	-	-	**182.26** ↑	12.04 ↑	**444.13** ↑	77.01 ↑	**445.71** ↑	13.63 ↑	**219.38** ↑
Dynaslam [32]	-	-	-	-	-	-	7.14 ↑	4.57 ↑	9.25 ↑
ORB-SLAM3 [1]	-	-	-	-	-	-	5.71 ↑	4.54 ↑	46.26 ↑

TABLE VI
RUN-TIME (MS) OF MAIN MODULES OF THE PROPOSED SYSTEM

Detection Thread	Segmentation	58.25
Tracking Thread	ORB Extraction	59.67
	LK flow	17.35
	AODA	1.68
	OCQ	8.71
	Total time	96.08
Ellipsoid Mapping	Optimization	114.38
	Total time	164.20
Local Mapping	Local BA	310.16
	Total time	372.76

threads, they do not affect the real-time performance of the rest of the system on the KITTI datasets captured at 10 Hz.

IX. CONCLUSION

In conclusion, this paper proposes a novel object SLAM system that combines a quadric initialization algorithm, an automatic data association algorithm, and a joint optimization framework to construct an accurate and robust object map in real-time for outdoor environments. In addition, the proposed system includes a multi-thread framework for ellipsoid parameter refinement, which significantly improves the efficiency of the system. Our proposed system outperforms the existing state-of-the-art object-based SLAM systems for object perception in outdoor scenarios as it considers partial observation, object occlusion, and dynamic objects. The novel quadric initialization algorithm and automatic data association algorithm enable accurate and robust quadric mapping, while the joint optimization framework ensures localization accuracy and real-time performance.

The proposed system has significant implications for autonomous driving and robotics, enabling more accurate and reliable object perception in complex outdoor environments. In our future work, the following issues will be important considerations: (1) Dynamic object representation for real-time tracking and joint dynamic pose estimation. (2) Relocalization and loop closure at the object level using quadric landmarks.

REFERENCES

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[3] C. Rubino, M. Crocco, and A. D. Bue, "3D object localisation from multi-view image detections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1281–1294, Jun. 2018.

[4] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust object-based SLAM for high-speed autonomous navigation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 669–675.

[5] L. Nicholson, M. Milford, and N. Sünderhauf, "QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM," *IEEE Robot. Autom. Lett.*, vol. 4, no. 1, pp. 1–8, Jan. 2019.

[6] R. Tian et al., "Accurate and robust object SLAM with 3D quadric landmark reconstruction in outdoors," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1534–1541, Apr. 2022.

[7] Z. Cao et al., "Object-aware SLAM based on efficient quadric initialization and joint data association," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9802–9809, Oct. 2022.

[8] Y. Meng and B. Zhou, "Ellipsoid SLAM with novel object initialization," in *Proc. IEEE 18th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2022, pp. 1333–1338.

[9] M. Zins, G. Simon, and M. Berger, "OA-SLAM: Leveraging objects for camera relocalization in visual SLAM," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2022, pp. 720–728.

[10] M. Zins, G. Simon, and M.-O. Berger, "Object-based visual camera pose estimation from ellipsoidal model and 3D-aware ellipse prediction," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 1107–1126, Apr. 2022.

[11] M. Zins, G. Simon, and M. Berger, "Level set-based camera pose estimation from multiple 2D/3D ellipse-ellipsoid correspondences," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 939–946.

[12] Y. Ming, X. Yang, and A. Calway, "Object-augmented RGB-D SLAM for wide-disparity relocalisation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 2203–2209.

[13] Y. Chen, S. Huang, and R. Fitch, "Active SLAM for mobile robots with area coverage and obstacle avoidance," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 3, pp. 1182–1192, Jun. 2020.

[14] M. Hosseinzadeh, K. Li, Y. Latif, and I. Reid, "Real-time monocular object-model aware sparse SLAM," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7123–7129.

[15] J. Wang, M. Rünz, and L. Agapito, "DSP-SLAM: Object oriented SLAM with deep shape priors," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 1362–1371.

[16] S. Yang and S. Scherer, "CubeSLAM: Monocular 3-D object SLAM," *IEEE Trans. Robot.*, vol. 35, no. 4, pp. 925–938, Aug. 2019.

[17] Y. Wu, Y. Zhang, D. Zhu, Y. Feng, S. Coleman, and D. Kerr, "EAO-SLAM: Monocular semi-dense object SLAM based on ensemble data association," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4966–4973.

[18] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1352–1359.

[19] Y. Dong et al., "A novel texture-less object oriented visual SLAM system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 36–49, Jan. 2021.

[20] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "CosyPose: Consistent multi-view multi-object 6D pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 574–591.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[22] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[23] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 17721–17732.

[24] Y. Li, N. Brasch, Y. Wang, N. Navab, and F. Tombari, "Structure-SLAM: Low-drift monocular SLAM in indoor environments," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6583–6590, Aug. 2020.

[25] Y. Li, R. Yunus, N. Brasch, N. Navab, and F. Tombari, "RGB-D SLAM with structural regularities," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 11581–11587.

[26] H. Huang, H. Ye, Y. Sun, and M. Liu, "Monocular visual odometry using learned repeatability and description," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8913–8919.

[27] M. Strecke and J. Stueckler, "EM-fusion: Dynamic object-level SLAM with probabilistic data association," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5864–5873.

[28] J. Huang, S. Yang, T. Mu, and S. Hu, "ClusterVO: Clustering moving instances and estimating visual odometry for self and surroundings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2165–2174.

[29] S. Yang, Z. Kuang, Y. Cao, Y. Lai, and S. Hu, "Probabilistic projective association and semantic guided relocalization for dense reconstruction," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7130–7136.

[30] Y. Liu, Y. Petillot, D. Lane, and S. Wang, "Global localization with object-level semantics and topology," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4909–4915.

[31] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1687–1694, Jul. 2018.

[32] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[33] X. Chen, A. Milioto, E. Palazzolo, P. Giguère, J. Behley, and C. Stachniss, "SuMa++: Efficient LiDAR-based semantic SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4530–4537.

[34] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "MID-fusion: Octree-based object-level multi-instance dynamic SLAM," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5231–5237.

[35] J. Huang, S. Yang, Z. Zhao, Y. Lai, and S. Hu, "ClusterSLAM: A SLAM backend for simultaneous rigid body clustering and motion estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5874–5883.

[36] J. Li, D. Meger, and G. Dudek, "Semantic mapping for view-invariant relocalization," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7108–7115.

[37] I. Ballester, A. Fontán, J. Civera, K. H. Strobl, and R. Triebel, "DOT: Dynamic object tracking for visual SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 11705–11711.

[38] M. Henein, J. Zhang, R. Mahony, and V. Ila, "Dynamic SLAM: The need for speed," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Oct. 2020, pp. 2123–2129.

[39] J. Zhang, M. Henein, R. Mahony, and V. Ila, "VDO-SLAM: A visual dynamic object-aware SLAM system," 2020, *arXiv:2005.11052*.

[40] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[41] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2756–2759.

[42] A. Kundu, K. M. Krishna, and J. Sivaswamy, "Moving object detection by multi-view geometric techniques from a single camera mounted robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 1–6.

[43] J.-Y. Bouguet et al., "Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm," *Intel Corp.*, vol. 5, nos. 1–10, p. 4, 2001.

[44] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[45] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[46] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Jan. 2008.

Rui Tian received the master's degree from the College of Information Science and Engineering, Northeastern University, Shenyang, China, where he is currently pursuing the Ph.D. degree. His research interests include visual SLAM, localization, and object SLAM.

Yunzhou Zhang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Northeastern University, Shenyang, China, in 2009. He is currently a Professor with the College of Information Science and Engineering, Northeastern University. His research interests include intelligent robots, computer vision, and sensor networks.

Zhenzhong Cao is currently pursuing the master's degree with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include visual SLAM and object SLAM.

Jinpeng Zhang is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include LiDAR SLAM and LiDAR place recognition.

Linghao Yang received the B.E. degree in automation from Northeastern University, Shenyang, China, where he is currently pursuing the master's degree with the College of Information Science and Engineering. His research interests include VSLAM, robot navigation, and localization.

Sonya Coleman is currently a Professor with the School of Computing and Intelligent System, Ulster University, Londonderry, U.K, and a Cognitive Robotics Team Leader with the Intelligent Systems Research Centre. Her research interests include robotics and computer vision.

Dermot Kerr is currently a Lecturer with the School of Computing, Engineering and Intelligent System, Ulster University Londonderry, U.K. His research interests include computational intelligence, biologically inspired image processing, and robotics.

Kun Li received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore. He is currently working with the DAMO Academy, Alibaba Group, Hangzhou, China. His research interests include SLAM, robotics, and localization.