Accurate and Robust Object-oriented SLAM with 3D Quadric Landmark Construction in Outdoor Environment

Rui Tian[1], Yunzhou Zhang[1*], Yonghui Feng[1], Linghao Yang[1],
Zhenzhong Cao[1], Sonya Coleman[2], Dermot Kerr[2]

Abstract—Object-oriented SLAM is a popular technology in autonomous driving and robotics. In this paper, we propose a stereo visual SLAM with a robust quadric landmark representation method. The system consists of four components, including deep learning detection, object-oriented data association, dual quadric landmark initialization and object-based pose optimization. State-of-the-art quadric-based SLAM algorithms always face observation related problems and are sensitive to observation noise, which limits their application in outdoor scenes. To solve this problem, we propose a quadric initialization method based on the decoupling of the quadric parameters method, which improves the robustness to observation noise. The sufficient object data association algorithm and the object-oriented optimization with multiple cues enables a highly accurate object pose estimation that is robust to local observations. Experimental results show that the proposed system is more robust to observation noise and significantly outperforms current state-of-the-art methods in outdoor environments. In addition, the proposed system demonstrates real-time performance.

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) is a fundamental technique in order for robots to perceive the environment. When compared with classic SLAM methods that use only the geometry of the scene, object-based SLAM has recently focused on creating maps with both geometry and high-level semantic objects within the environment [1]–[9]. This semantically-enriched information can help robots with target-oriented tasks like obstacle avoidance, robust relocalization and human-robot interaction. The improvement in the accuracy of semantic information acquisition, driven by deep learning networks [10]–[12], has led to the increasing introduction of object detection and semantic segmentation into visual SLAM systems to build semantically enriched maps and enhance the perception ability of robots.

Accurate object representation is a key issue in object-oriented SLAM research and 3D object models [13], cubic boxes [3]–[5] and ellipsoids [6]–[8] are among common methods utilized for object representation. Prior work like [4] and [5] use the cubic box to represent the object, where the pose
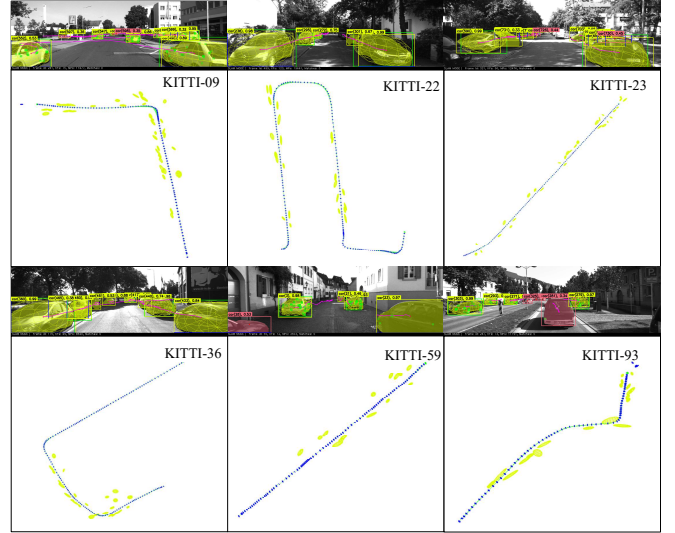
Fig. 1. Our proposed method uses 3D quadric landmarks to build the object-oriented map in outdoor environments. Yellow quadrics illustrate the accuracy of orientation and shape of the estimated ellipsoids when projected onto the current image frame. Object ID is marked, and the magenta lines show the center of the ellipsoids in previous frames when projected onto the image frame, indicating the accuracy of the object data association. The red bounding box is the object detected as dynamic and will not be constructed with ellipsoids. Finally, the object maps are also provided.

of the cubic box can be estimated by vanishing points and rotation sampling. Compared with the cubic box, the ellipsoid can also accurately represent the position, orientation and size of the object and has a more concise mathematical representation [6]. In projective geometry the quadric can be represented by a symmetric matrix [14] where the compact perspective projection model and the closed surfaces of ellipsoids are meaningful for object landmarks.

The accuracy and robustness of current quadric-based SLAM are not ideal, especially the quadric initialization process, which is limited by the parameter coupling of the direct linear solution method [6] or the necessity for point cloud fitting [7], [9]. QuadricSLAM [6] is a recently proposed object-oriented SLAM system that represents objects as quadrics; a dual quadric observation model based on the object detection is proposed. However, the closed-form constrained dual quadric parameterization and the lack of observation angles under the planar trajectory of the mobile vehicle make the initialization of the quadric difficult and sensitive to observation noise. In [9], multiple constraints combined with points, surface and quadrics are used in the optimization framework,

but the prior shape of the object is estimated based on deep learning which incurs a high computational complexity and is not robust. In [2], the texture plane and shape prior constraints are added to the quadric estimation which solves the problem of poor estimation performance when the observation angles change in road driving scenes. However, the assumption that the texture plane is parallel to the image plane during quadric initialization causes the estimation to be sensitive to noise.

In addition, in prior work such as [6], [15], data association methods have been proposed although they are typically not robust to outdoor scenes. Dynamic objects in outdoor scenes like moving cars and persons are a challenge for quadric estimation since false object associations will lead to false quadric initialization results.

To solve the aforementioned problems, we propose a robust and accurate quadric landmark initialization based on a method for decoupling of quadric parameters (DQP) and an object data association (ODA) algorithm in outdoor scenes. The robustness of DQP to observation noise is improved by independently estimating the quadric centroid translation and the yaw rotation constraint which is satisfied for autonomous vehicles in road planes in most cases. Then, an ellipsoid with improved accuracy can be obtained by a nonlinear optimizer combining the observation error, the texture plane error and the prior object size. In terms of data association, we propose a multiple-cues algorithm combined with the Hungarian assignment algorithm [16] which improves the robustness of object pose estimation.

We demonstrate the performance of the proposed system in both a simulation environment and using the KITTI Raw Data [17] datasets. The experimental results show that the proposed system is more robust to observation noise than other existing methods and improves the accuracy of the position, orientation and size of the object estimation in the outdoor environment.

The main contributions of this work are:

- To effectively overcome the observation noise, we propose an accurate and robust quadric landmark initialization method based on the DQP algorithm by decoupling of translation and rotation of quadric centroids.
- We proposed an ODA algorithm that combines the semantic inliers distribution, Kalman-based motion prediction, and ellipsoidal projection to achieve accurate object data association and object pose estimation.
- Based on the proposed algorithms, we implement real-time stereo visual SLAM with accurate and robust ellipsoids representing objects, aiming to build an object-oriented and semantically-enhanced map for outdoor navigation.

II. SYSTEM OVERVIEW

A. Mathematical Representation of a Quadric Model

For convenience of description, the notations used in this paper are as follows:

- $(\cdot)_w$ is the world coordinate, $(\cdot)_c$ is the camera coordinate, $(\cdot)_r$ is the reference camera coordinate of the object, and $(\cdot)_q$ is the quadric center frame.
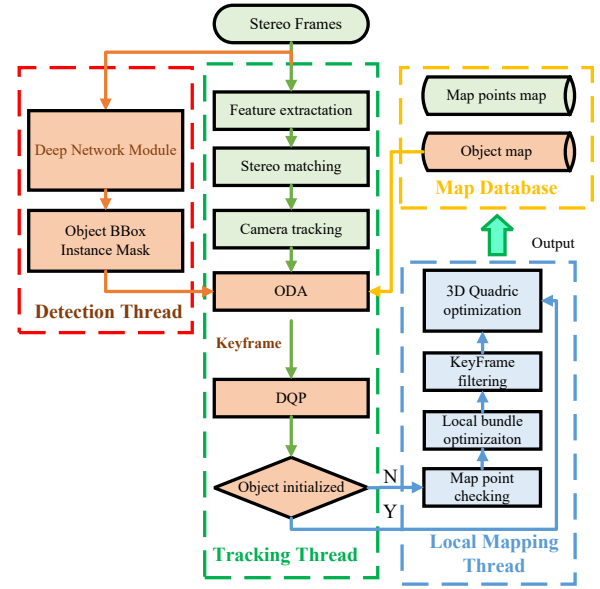

Fig. 2. Overview of our proposed system, there are two key modules: 1) The detection thread takes images and acquires semantic and detection results. 2) The tracking thread initializes quadrics with the DQP method and the ODA associates detected object quadrics with the mapped objects for further optimization. Finally, the object map is stored with ellipsoid representations.

- K - The intrinsic matrix of a pinhole camera model.
- $T_{cw} \in R^{4\times 4}$ - The transformation from world frame to camera frame, which is composed of a rotation $R_{cw} \in R^{3\times 3}$ and a translation $t_{cw} \in R^{3\times 1}$.
- $P = KT$ - The camera projection matrix that contains intrinsic and extrinsic camera parameters.
- $B = [x_1, y_1, x_2, y_2]^T$ - The 2D object detection bounding box (BBox).
- M is the segmentation instance mask, D is the detection instance, O is the object instance.
- D_j^i represent the detected object instance D_j that is assigned to the object O_i, $cls(D)$ and $cls(O)$ represent the class label of the detected instance and object instance respectively.
- $In(B, x)$ - The checking of image points x that are located in the B detection box.
- $Q \in R_{4\times 4}$ - The quadric matrix in 3D space and $Q^* \in R_{4\times 4}$ is denoted as the dual quadric matrix.
- $\Pi = [\pi_1, \pi_2, \pi_3, \pi_4]^T$ - The 3-D plane surface in homogeneous coordinate and all quadric plane fulfil $\Pi^T Q^* \Pi = 0$.
- $q = [a_x, a_y, a_z, t_x, t_y, t_z, \theta_x, \theta_y, \theta_z]^T$ - The 9-D vector representing the attributes of the quadric, including axial length, translation and rotation.

When a dual quadric is projected onto an image plane, it creates a dual conic, following the rule $C^* = PQ^*P^T$. For more specific properties of the quadric, please refer to [6].

B. System Architecture

The proposed system is shown in Fig.2. We implement our algorithms on the basis of ORB-SLAM3 [18], and a stereo camera is used to obtain a metric scale of the estimated

trajectory for the autonomous driving scene to avoid scale ambiguity caused by monocular SLAM [19]. However, we also highlight that our method can be used for monocular SLAM. There are two key modules, the visual SLAM module and the detection module. The visual SLAM module consists of parallel threads, including the tracking thread and the local mapping thread. Finally, the camera pose is estimated and a semantically-enhanced object map is also stored in the map database.

(1) **The detection thread** uses YOLOACT [10] to acquire semantic information from the left images of the stereo pair. The output results are object detection BBoxes and the instance segmentation masks.

(2) **The tracking thread** takes images and estimates the camera pose from consecutive frames. Meanwhile, the thread waits for the detection instances and associates them with the existing objects in the object map database or decides whether to create a new object using the ODA algorithm. In addition, if the current frame is a keyframe and an observation satisfies the quadric initialization condition, the DQP algorithm is used for robust and accurate quadric initialization.

(3) **The local mapping thread** optimizes the map points of keyframes with local bundle adjustment. In addition, when the objects are observed by newly inserted keyframes, the new observation can be added to the object optimizer for nonlinear optimization of the ellipsoidal representation of objects.

(4) **The map Database** stores the final maps, including the geometry information of map points and the object-oriented map with ellipsoids.

III. DECOUPLING OF QUADRIC PARAMETERS INITIALIZATION ALGORITHM

A. Decoupling of Quadric Central Translation

We present the mathematical analysis of the dual quadric parameters to illustrate the effect of the translation component on the estimation of rotation and shape. The dual form parameters of the ellipsoid can be decomposed by eigen-decomposition in the reference camera coordinates of the object:

$$
\begin{aligned}
Q_r^* &= T_{rq} Q_q^* T_{rq}^T \\
&= \begin{bmatrix} R_{rq} & t_{rq} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} D & 0 \\ 0^T & -1 \end{bmatrix} \begin{bmatrix} R_{rq}^T & 0 \\ t_{rq}^T & 1 \end{bmatrix} \\
&= \begin{bmatrix} R_{rq} D R_{rq}^T - t_{rq} t_{rq}^T & -t_{rq} \\ -t_{rq}^T & -1 \end{bmatrix} \\
&= \begin{bmatrix} Q_{r33}^* & -t_{rq} \\ -t_{rq}^T & -1 \end{bmatrix}
\end{aligned}
\tag{1}
$$

where $D \in R^{3 \times 3}$ is the diagonal matrix composed of the squares of the quadric axial lengths, and $t_{rq} \in R^{3 \times 1}$ is the quadric centroid translation in the reference camera coordinates. The parameters of the block matrix $Q_{r33}^* \in R^{3 \times 3}$ couple the rotation and translation of the quadric. Since the length of the quadric centroid translation is much larger than that of the rotation and axes, small errors in the estimation of the quadric centroid translation have a significant impact on the

accurate estimation of the dual quadric matrix, which is why QuadricSLAM [6] is sensitive to observation noise.

We can also see from Eq.1 that the translation parameters are independent in dual form parameters q_{14}^*, q_{24}^*, q_{34}^*. Therefore we estimate the translation component parameters $t_{rq} \in R^{3 \times 1}$ independently to eliminate the effect of coupling parameters, a key aspect of our approach. We triangulate the center of the 2D detection box x_{bc}^i and obtain the triangulation map point $\hat{t}_{rq}$, which is almost close with the quadric center t_{rq} in outdoor scenes. This assumption is proved by experiments in VI-A. Observations of two or more frames of detection centers form the overdetermined equation to solve t_{rq},

$$
\begin{bmatrix} P_c[1,*] T_{wr} - x_{bc}^i[1] \cdot P_c[3,*] T_{wr} \\ P_c[2,*] T_{wr} - x_{bc}^i[2] \cdot P_c[3,*] T_{wr} \end{bmatrix} t_{rq} = 0
\tag{2}
$$

where, $x_{bc}^i[i]$ is the i-th element of 2D detection center, $P_c[i,*]$ is the i-th row of the projection matrix P_c.

B. Decoupling of Quadric Rotation and Axial Length

The rotation and quadric axial lengths are considered after the quadric centroid translation has been estimated independently. We assume that the ellipsoid of the object, such as an autonomous vehicle or robot, is under the constraint of yaw rotation, while the pitch and roll are constant at zero. This is satisfied for autonomous vehicles on the road in outdoor scenes. Therefore, we can replace the rotation matrix in Eq.1 by:

$$
R_{rq} = \begin{bmatrix} cos\theta_y & 0 & sin\theta_y \\ 0 & 1 & 0 \\ -sin\theta_y & 0 & cos\theta_y \end{bmatrix}
$$

$$
Q_r^* = \begin{bmatrix} a_x^2 cos^2\theta_y + a_z^2 sin^2\theta_y - t_{rq}^{x2} & -t_{rq}^x t_{rq}^y & cos\theta_y sin\theta_y(a_z^2 - a_x^2) - t_{rq}^x t_{rq}^z & -t_{rq}^x \\ -t_{rq}^x t_{rq}^y & a_y^2 - t_{rq}^{y2} & -t_{rq}^y t_{rq}^z & -t_{rq}^y \\ cos\theta_y sin\theta_y(a_z^2 - a_x^2) - t_{rq}^x t_{rq}^z & -t_{rq}^x t_{rq}^y & a_x^2 sin^2\theta_y + a_z^2 cos^2\theta_y - t_{rq}^{z2} & -t_{rq}^z \\ -t_{rq}^x & -t_{rq}^y & -t_{rq}^z & -1 \end{bmatrix}
\tag{3}
$$

where, t_{rq}^x, t_{rq}^y, t_{rq}^z are elements of the quadric centroid translation vector.

We can simplify the linear form in [6] by using the landmark BBox observations B and the corresponding dual quadric planes Π by substituting the t_{rq}.

$$
M^T = \begin{bmatrix} \pi_1^2 \\ 2\pi_1\pi_3 \\ \pi_2^2 \\ \pi_3^2 \\ \pi_4^2 + 2\pi_1\pi_2 t_{rq}^x t_{rq}^y + 2\pi_1\pi_4 t_{rq}^x + 2\pi_2\pi_3 t_{rq}^y t_{rq}^z \\ + 2\pi_2\pi_4 t_{rq}^y + 2\pi_3\pi_4 t_{rq}^z \end{bmatrix}
\tag{4}
$$

$$
M \begin{bmatrix} q_{11}^* & q_{13}^* & q_{22}^* & q_{33}^* & q_{44}^* \end{bmatrix}^T = 0
\tag{5}
$$

The decoupled linear form of Eq.5 can be solved by singular value decomposition (SVD) [6], where q_{ij} is the remaining elements of the dual quadric to be estimated.

Finally, the 9-D vector q of the quadric with orientation, translation and axial lengths of the ellipsoid can be obtained by the estimated dual quadric matrix Q^*:

$$\theta_y = arctan(2Q_3/(Q_8 - Q_1))/2$$
$$a_x = \sqrt{|Q_1 - 2Q_3 + Q_8|}/2$$
$$a_y = \sqrt{Q_2}$$
$$a_z = \sqrt{|Q_1 + 2Q_3 + Q_8|}/2$$

(6)

where,

$$Q_1 = -\frac{q_{11}^*}{q_{44}^*} + t_{rq}^{x2} \quad Q_2 = -\frac{q_{22}^*}{q_{44}^*} + t_{rq}^{y2}$$
$$Q_3 = -\frac{q_{13}^*}{q_{44}^*} + t_{rq}^x t_{rq}^z \quad Q_8 = -\frac{q_{23}^*}{q_{44}^*} + t_{rq}^{z2}$$

(7)

IV. 3D OBJECT OBSERVATION CONSTRAINTS OPTIMIZATION

In the local mapping thread, we optimize the quadrics by using odometry factors and landmark factors combined with the observation of local keyframes. We define the set of detected objects as $\mathbb{D}$, and the set of mapped objects as $\mathbb{O}$. By minimizing the observation error between observed instances D_k and associated mapped instance O_i, q of the quadric can be optimized with the following constraint:

$$q = \arg\min_q(\sum H_b(f_b) + \sum H_a(f_a)) + \sum H_p(f_p) \quad (8)$$

The Huber kernel $H(.)$ is used to enhance the robustness of outlier observations, and the LM algorithm [20] is used to optimize the target cost function.

A. The 2D detection error

The 2D detection error is used to calculate the distance error between the 2D object BBox $B_{O_i}^k$ and the detected BBox B_{D_j} in the kth keyframe. Detection results near the edge of the image are ignored in order to eliminate the effect of occlusion.

$$f_b = e_b(B_{O_i}^k, B_{D_j})^T \Omega_b e_b(B_{O_i}^k, B^{D_j})$$
$$e_b(B_{O_i}^k, B_{D_j}) = B_{O_i}^k - B_{D_j}$$

(9)

B. Prior axial length error

The prior axial length error is calculated by the distance between the prior axial length a_{prior} and the object quadric axial length a_{O_i} with the same object class.

$$f_a = e_a(O_i)^T \Omega_a e_a(O_i)$$
$$e_a(O_i) = a_{prior}(cls(O_i)) - a_{O_i}$$

(10)

C. Texture plane error

Similar to the method proposed by [2], the texture plane error is obtained by the minimum distance between the fitted texture plane and the quadric landmark. The plane parameters of the texture plane is obtained by Delauney Triangulation of the object's map points with the normal vector n_{D_j} and plane distance Z_{D_j} of a texture plane Π_{D_j}. The texture plane distance error can be calculated as:

$$f_p = e_p(\Pi_{D_j}, \Pi_{O_i})^T \Omega_p e_p(\Pi_{D_j}, \Pi_{O_i})$$
$$e_p(\Pi_{Dj}, \Pi_{O_i}) = Z_{D_j} - Z_{O_i}$$

(11)

V. THE OBJECT DATA ASSOCIATION ALGORITHM

Multi-view geometry information is used for object landmark initialization, while the object detection results are obtained by the single-frame image. Therefore, it is necessary to correctly associate the detected instance of the same object within the map. We propose the ODA algorithm to integrate information for data association. The Hungarian algorithm [16] is used to complete the assignment with the minimum distance error. Three different distance metrics are used for affinity functions to obtain a_{ij}, which is the element of the cost matrix $\mathbb{A}$. The α, β and γ parameters are experimentally set to 0.8, 1, and 0.8 respectively.

$$a_{ij} = \alpha a_{ij}^p + \beta a_{ij}^g + \gamma a_{ij}^k \quad (12)$$

A. Semantic Inliers Points Distance

To overcome the overlap of the detection masks, we use Bi-directional Optical Flow (BODF) to track the keypoints within the detection mask M_j from the last keyframe and obtain the keypoints set $\{x_j^k\}$. We calculate the ratio of inliers corresponding to the same object class, where size{.} calculates the element numbers of the set:

$$a_{ij}^p = 1 - \frac{\text{size}\{In(M_{O_i}^k, X_j^k)\}}{\text{size}\{X_j^k\}} \quad (13)$$

B. Intersection of Union Distance

To calculate the intersection of union distance, we use the intersection ratio between the 2D quadric landmark projection BBox B_{Oi} of O_i and the 2D detection result B_{D_j} of the object instance D_j .

$$a_{ij}^g = 1 - \frac{B_{O_i} \cap B_{D_j}}{B_{O_i} \cup B_{D_j}} \quad (14)$$

C. Prior Object Size Distance

For each object instance O_i, the motion prediction method based on the Kalman filter [21] is adopted to predict the state of the detection in the image frame. The predicted 2D BBox of D_j is denoted as $B_{D_j}^a$, the prior object size distance a_{ij}^k is defined by:

$$a_{ij}^k = 1 - \frac{B_{O_i}^a \cap B_{D_j}^a}{B_{O_i}^a \cup B_{D_j}^a} \quad (15)$$

VI. EXPERIMENTS

The proposed system consists of two modules, including the SLAM module and the detection module. The overall system architecture is described in Fig.2. In order to evaluate the performance of our proposed method, we build an experimental simulation environment based on OpenGL to compare the robustness and accuracy against other state-of-the-art techniques. The KITTI Raw Data dataset [17] is adopted as the benchmark real-world dataset to demonstrate the effectiveness of our method in outdoor scenes. All the experiments are conducted using an Intel(R) Core(TM) i7-9750H CPU@2.6GHZ, 16G memory, and Nvidia GTX 1080 Ti.

We define the following criteria for evaluation:

(1) IoU_{2D}: The intersection ratio between the ground truth (GT) and the estimated quadric projection detection.

$$IoU_{2D} = \frac{B_{gt} \cap B_{pred}}{B_{gt} \cup B_{pred}} \quad (16)$$

(2) e_{trans}: The error of quadric centroid translation between the GT ellipsoid and the prediction estimation, indicating the accuracy of the ellipsoid position estimation.

$$e_{trans} = \|t_{wq}^{gt} - t_{wq}^{pred}\|_2 \quad (17)$$

(3) e_{axe}: The error of ellipsoid axial length between the GT ellipsoid and the predict estimation in the world coordinate, indicating the accuracy of the object shape estimation.

$$e_{axe} = \|a_{gt} - a_{pred}\|_2 \quad (18)$$

A. Quantitative Evaluation of Simulation

Simulation provides GT of object positions and it is easy to test the robustness of the methods with different types of disturbance. We create the synthetic dataset with OpenGL, five cameras are evenly deployed within $18°$ circular arcs to simulate the camera observation in the outdoor environment. An ellipsoid with varying shape and yaw rotation is deployed, the GT 2D object BBox and the position are provided. The yaw rotation of the ellipsoid is randomly sampled in the range of $\pm5°$ to simulate objects with rotation. To avoid the influence of random errors on the experimental results, for each type of noise, 10 ellipsoids are generated with Gaussian noise from 10 seeds resulting in a total of 100 trials.

To test the effect of different types of noise on the quadric initialization method, the relative camera poses are obtained by introducing zero-mean Gaussian noise with standard deviations in the range $5\% \sim 30\%$ to simulate the trajectory error. In addition, a detection BBox is simulated by adding the zero-mean Gaussian noise of $1\% \sim 6\%$ to the GT.

We compare methods of quadric initialization including (a) Nicholson *et al* [6] denoted as Q-SLAM, (b) Rubino *et al* [1] denoted as Conic-method, (c) the proposed method with only decoupling of the quadric central translation, denoted as Tri, and (d) the proposed initialization method denoted as Tri+Yaw.

Quantitative evaluation results of initialization methods with different types of noise are visualized in Fig.3. The plots show the trend of different evaluation criteria with the increase in noise. It can be seen from Fig.3, the results for all methods are consistent with the GT, demonstrating correctness of all methods with zero noise. The performance of all methods degrades when noise increases.

It is obvious that the Q-SLAM method is the most sensitive to noise among all the techniques. When either the translation noise reaches 15%, the rotation noise reaches 20%, or the detection BBox noise reaches 2%, Q-SLAM fails to construct ellipsoids.

Meanwhile, the Conic-method maintains relatively good results which show the robustness under the effect of translation and rotation noise. On the other hand, it can be seen that under the influence of detection BBox noise, the e_{trans} and

e_{axe} of the Conic-method also increase rapidly. When the detection BBox error exceeds 4%, the Conic-method fails to initialize the ellipsoids, which indicates that the Conic-method is also sensitive to detection noise. However, the performance of the proposed method is stable as it can be seen that with translation and rotation noise, the error remains stable with the maximum axial error of 0.45m and maximum translation error of 0.89m. These metrics are also influenced within a small range by the detection BBox error with the maximum axial error and translation error of 1.02m and 2.10m. These results show that our proposed method significantly improves the robustness of initialization with the minimal growth trend of noise. The visualization results of quadric initialization are shown in Fig.4, where the red ellipsoid is the GT, and green ellipsoid is the estimation. Our proposed method outperforms all compared methods.

B. Evaluation on KITTI Dataset

To evaluate the performance of the proposed method in outdoor environments, we select the KITTI Raw Data dataset [17] in particular the sequences -09, -22, -23, -36, -59, and -93, which were recorded in urban and residential areas with vehicles. The dataset provides GT for vehicles, including 3-DoF object size and 6-DoF object pose. With the extrinsic parameters of sensors, we can transform the object pose to camera coordinates.

Table I shows the success rate of initialization and ellipsoid construction by different methods using different sequences. Tables II, III and IV show the experimental results of successfully constructed ellipsoids under different evaluation criteria.

From Table I, we can see that our method constructs ellipsoids for 60.2% of the vehicles and reaches an increase of 62.6% (from 37.0% to 60.23%) and 99.2% (from 30.24% to 60.23%) in success rate compared with the Conic-method and the Q-SLAM, respectively, thus confirming the effectiveness of our initialization method. For the IoU_{2D} metric, larger values indicate better construction results. As can be seen from Table II, our method outperforms the other existing methods with respect to IoU_{2D} in sequence -09 and -36, with the overall best average of 73.03%. The compared methods give better results for individual sequences because they discard some detection results that fail to be initialized. For e_{trans} and e_{axe}, smaller values indicate better construction results. As can be seen from Table III and Table IV, our method outperforms the compared methods in all cases except for sequence-22, with the average ellipsoid central translation error of 2.127m, nearly 52.2% reduction in error. In addition, our average axial length error is 0.642 m, a 50.8% reduction in error, compared with 1.369 and 0.947 for the other techniques. These experimental results show the robustness and effectiveness of the proposed method for ellipsoid representations in outdoor scenes.

Finally, we show the constructed object maps in Fig.1. The yellow ellipsoids in the map represent static vehicles and the yellow quadrics illustrate the orientation and shape of the estimated ellipsoids when projected onto the image frame. The magenta lines show the center of the ellipsoids in previous
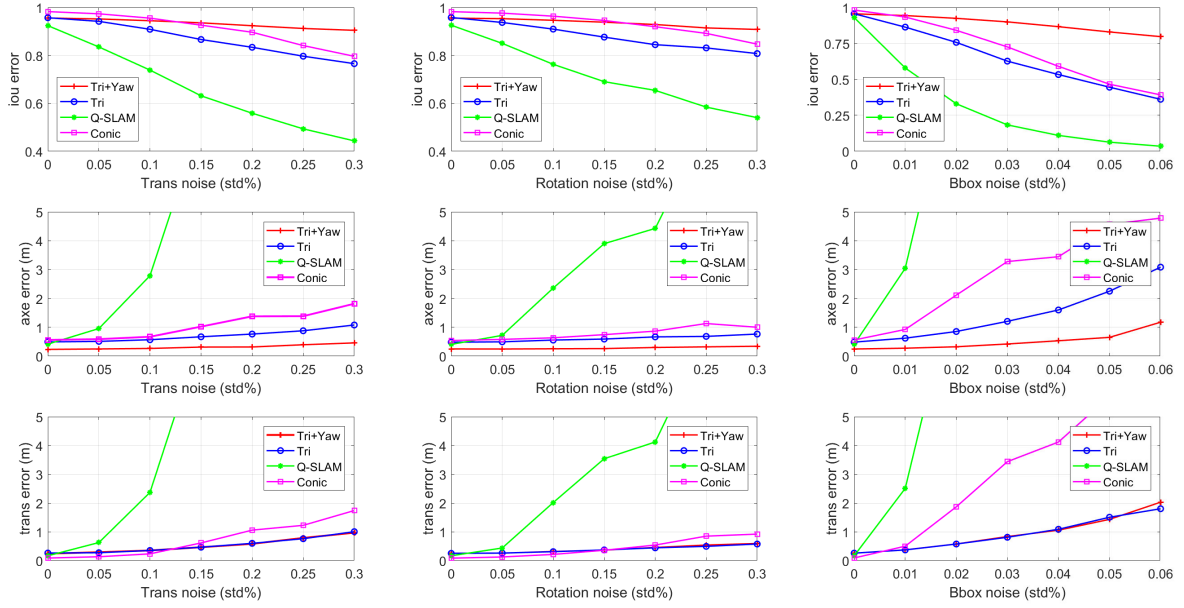
Fig. 3. The initialization performance of methods to different types of noise, the curves show the trend of different criteria with the increase of noise.

frames projected onto the current image frame, demonstrating the accuracy of the ODA algorithm. The red BBox represents the vehicles that are detected as dynamic objects and are not contained in the map.

TABLE I
SUCCESS RATE COMPARISON IN KITTI RAW DATA DATASET.

Sequence	Ours	Conic [1]	Q-SLAM [6]
09	**0.6912**	0.4468	0.2706
22	**0.6512**	0.3333	0.2923
23	**0.6230**	0.3949	0.2829
36	**0.6047**	0.4096	0.3514
59	**0.5625**	0.3556	0.2558
93	**0.4815**	0.2826	0.3617
Average	**0.6023**	0.3705	0.3024

TABLE II
IoU_{2D} COMPARISON IN KITTI RAW DATA DATASET.

Sequence	Ours	Conic [1]	Q-SLAM [6]
09	**0.7335**	0.7252	0.7031
22	0.7629	**0.7791**	0.7662
23	0.7509	**0.7529**	0.6959
36	**0.7604**	0.7558	0.7127
59	0.6508	**0.6878**	0.6500
93	0.7232	0.6751	**0.7433**
Average	**0.7303**	0.7293	0.7119

VII. CONCLUSION

In this work, a novel pipeline of real-time object-oriented stereo visual SLAM with 3D quadric landmarks is presented. A quadric initialization method based on the DQP algorithm is proposed to improve the robustness and success rate of ellipsoid construction. The data association is solved by the ODA algorithm which ensures highly accurate object pose estimation. Extensive experiments are conducted to show that the proposed system is accurate and robust to observation noise

TABLE III
TRANSLATION ERROR COMPARISON IN KITTI RAW DATA DATASET.

Sequence	Ours	Conic [1]	Q-SLAM [6]
09	**2.5456**	2.7819	3.6110
22	2.1769	**1.9552**	1.9651
23	**2.3341**	5.6605	8.7088
36	**1.8594**	2.7175	6.9814
59	**1.3276**	1.6883	1.3874
93	**2.5226**	5.4130	4.0654
Average	**2.1277**	3.3694	4.4532

TABLE IV
AXIAL LENGTH ERROR COMPARISON IN KITTI RAW DATA DATASET.

Sequence	Ours	Conic [1]	Q-SLAM [6]
09	**0.6271**	1.2618	1.1400
22	**0.5565**	0.7233	0.8356
23	**0.5494**	1.7886	0.6837
36	**0.7121**	1.2797	0.8799
59	**0.6706**	1.8467	1.2908
93	**0.7357**	1.3156	0.8574
Average	**0.6419**	1.3693	0.9479

and significantly outperforms other methods in an outdoor environment.

In further work, we will explore finding the semantic relationships between object ellipsoids, and using the semantic information of the object map to localize and perform re-localization.

REFERENCES

[1] C. Rubino, M. Crocco, and A. Del Bue, "3d object localisation from multi-view image detections," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1281–1294, 2017.

[2] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust object-based slam for high-speed autonomous navigation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 669–675.

[3] S. Yang and S. Scherer, "Monocular object and plane slam in structured environments," 2018.

[4] ——, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

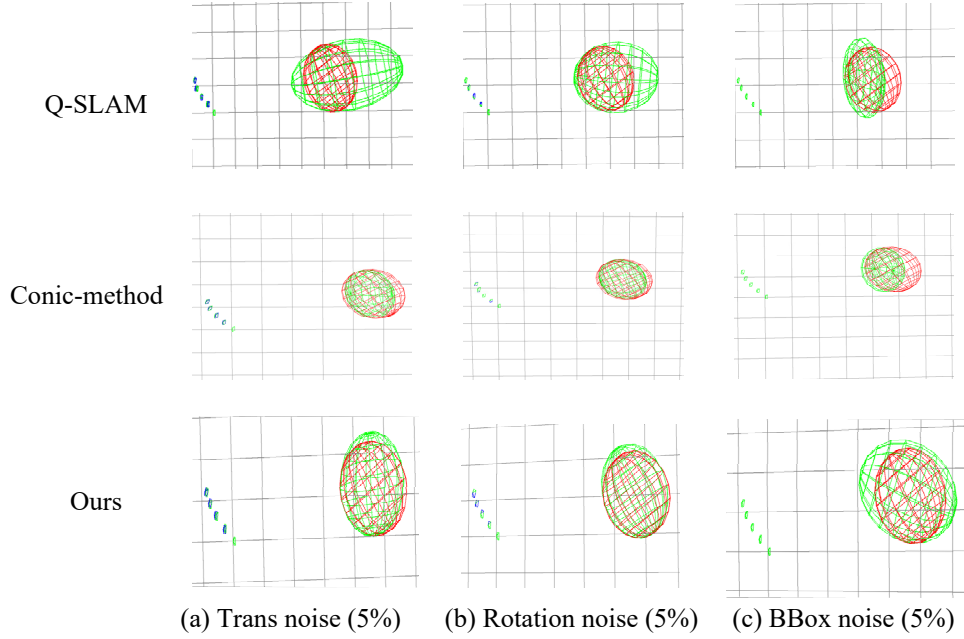| Q-SLAM | | |
Conic-method		
Ours		
(a) Trans noise (5%)	(b) Rotation noise (5%)	(c) BBox noise (5%)

Fig. 4. The visualization results of initialization performance of methods to different types of noise. Our method can initialize ellipsoid accurately and robustly.

[5] 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 *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4966–4973.

[6] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.

[7] Z. Liao, W. Wang, X. Qi, X. Zhang, and R. Wei, "Object-oriented slam using quadrics and symmetry properties for indoor environments," 2020.

[8] V. Gaudillière, G. Simon, and M. O. Berger, "Perspective-2-ellipsoid: Bridging the gap between object detections and 6-dof camera pose," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5189–5196, 2020.

[9] M. Hosseinzadeh, K. Li, Y. Latif, and I. Reid, "Real-time monocular object-model aware sparse slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7123–7129.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9157–9166.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[14] G. Cross and A. Zisserman, "Quadric reconstruction from dual-space geometry," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 25–31.

[15] Z. Qian, K. Patath, J. Fu, and J. Xiao, "Semantic slam with autonomous object-level data association," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021.

[16] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[18] C. Campos, R. Elvira, J. Rodríguez, J. Montiel, and J. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam," 2020.

[19] R. Tian, Y. Zhang, D. Zhu, S. Liang, S. Colman, and D. Kerr, "Accurate and robust scale recovery for monocular visual odometry based on plane geometry," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021.

[20] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *IEEE International Conference on Robotics Automation*, 2011.

[21] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016.