2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Robust Dual Quadric Initialization for Forward-Translating Camera Movements

Shujia Chen[†,1], Shuangfu Song[†,2], Junqiao Zhao[*,1], Tiantian Feng[2], Chen Ye[1], Lu Xiong[3], Deyi Li[4]

Abstract— **Herein, we present a novel approach for monocular dual quadric initialization that combines three-dimensional (3D) map points with two-dimensional (2D) object detection for forward-translating camera movements. The traditional approach using 2D detection bounding boxes in multiple views fails in straight vehicle motion scenarios as object observation is limited to few frames. Although single image initialization is possible when multiple constraints are introduced, such initialization is based on strong assumptions. In this work, we incorporate constraints from 3D map points with single-view 2D object detection to robustly initialize the dual quadric. Constraints from 3D map points are converted to planar constraints from their convex hull. Together with the projective planar constraints from bounding boxes, the proposed method can infer accurate dual quadric parameters. Further, comparison studies with the state of the art (SOTA) show that the proposed approach achieves the same accuracy of center localization but outperforms the existing methods in shape estimation and success ratio of initialization. The proposed method dose not rely on assumptions of dimension and pose of 3D objects; hence, it is more generic and accurate. Based on the KITTI raw dataset, the initialization success ratio is up to 97.7% with an average position error of 1.58 m, and 2D IoU of 80% when the number of map points per object accumulates to 60. When applied to the TUM RGB-D dataset, the proposed approach yields an initialization success ratio of 92.7% when the number of map points per object accumulates to 30, revealing a 16.2% increment compared with the SOTA using an RGB-D camera. Finally, we integrate the initialization method into a simultaneous localization and mapping system.**

Index Terms— **SLAM, Mapping, Dual Quadric Initialization, Forward-Translating Camera Movements**

I. INTRODUCTION

Object detection and localization in three-dimensional (3D) space are fundamental tasks in object-level simultaneous localization and mapping (SLAM). Through the localizing and mapping of 3D objects, a robot can build a lightweight high-level object map and robustly estimate the ego position with object-level observations.

Two primary approaches are available for 3D object localization. The first approach directly estimates object poses via end-to-end 3D regression using neural networks[1], [2]. This approach requires an expensive training process and it ignores an object's multiple frame observations. The second approach focuses on geometrical reasoning primarily based on 2D object detection results [3], [4]. [4] adopted cuboids with 9 degrees of freedom (DoF) as the geometrical representation of an object with a ground-object assumption; however, herein, we choose more ideal dual quadrics as object representation because quadrics[5] can be compactly parameterized and easily manipulated within the framework of projective geometry. In addition, dual quadrics can flexibly fit most of the 3D shape tightly, which means that strong shape assumptions of parameters such as e.g. the dimensions or the orientations need not be made.

QuadricSLAM proposed by [6] first incorporated dual quadrics into a SLAM system as landmark representation, wherein dual quadric initialization was implemented using 2D object detection (bounding boxes) from multiple views. However, this method is only suitable for scenarios with an orbiting camera path with several diverse viewpoints; otherwise, the 2D detections would be insufficient to constrain the exact shape of the dual quadric. To overcome this drawback, another approach using the texture on the object and semantic knowledge as extra constraints is proposed [7]. Though this method outperforms QuadricSLAM in forward-translating camera sequences, it relies on the extra assumption that the plane fitted by the triangulated features of the texture is parallel to the camera image plane. In addition, it assumes the center of the dual quadric to be projected onto the center of the 2D bounding box; the depth is the average scene depth; and its orientation prior is identical to the camera coordinate frame, with its shape prior resembling a Toyota Camry in terms of dimensions. Recently, [8] proposed a dual quadric SLAM system based on RGB-D camera for indoor environments. The method uses point cloud segmentation to extract object-supporting planes as extra constraints and assumes that the object's supporting planes are vertical to the direction of gravity.

Herein, we propose a monocular dual quadric initialization approach combining 2D object detection with 3D map points primarily focused on autonomous driving scenarios, wherein the camera mainly performs forward-translating. We accumulate the 3D map points for an object to a certain number and then construct planar constraints from their convex hull.

[1]Shujia Chen, Junqiao Zhao and Chen Ye are with the Department of Computer Science and Technology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 1830810@tongji.edu.cn; zhaojunqiao@tongji.edu.cn; yechen@tongji.edu.cn).

[2]Shuangfu Song and Tiantian Feng are with the School of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China (e-mail: songshuangfu@gmail.com; fengtiantian@tongji.edu.cn).

[3]Lu Xiong is with the Institute of Intelligent Vehicle, Tongji Unviersity, Shanghai, 201804 China (e-mail: xiong_lu@tongji.edu.cn).

[4]Deyi Li is with The 61st Research Institute of General Staff Department, Beijing, China (e-mail: lidy@cae.cn).

[†]Shujia Chen and Shuangfu Song contributed equally to this paper.

This work was supported by the National Natural Science Foundation of China (No. 41801335, No. U1764261, No. 41871370), the National Key Research and Development Program of China (No. 2018YFB0105103, No. 2018YFB0505400), and the Fundamental Research Funds for Central Universities (No. 22120180095).

Together with the planar constraints derived from the 2D object detection, our approach can provide a more robust solution to the least-square fitting problem than exclusively using multiple-view 2D object detection.

Comparing with [7], our approach is more accurate for forward-translating camera motion, particularly when the assumptions adopted by [7] do not hold. We also outperform [8] that uses depth information on the TUM RGB-D dataset.

II. RELATED WORKS

The dual quadric initialization required multiple constraints [5]. [9] derived a linear relationship between an ellipsoid and its perspective projection, i.e. the object contour from three views. [10] recovered a one-parameter family of quadrics from extracted object outlines in two views. However, multiple views must be used to determine the unique solution by adding additional epipolar constraints between matched points. Moreover, these approaches rely on extracted object contours, which are expensive.

Recent studies take advantage of low-cost object detection [11], [12] to initialize dual quadrics. These approaches are based on multiple-view detection bounding boxes. [13] presented a closed-form solution of dual quadrics from a set of 2D conics fitted at the object bounding boxes in multiple views. However, that solution is limited to synthesized virtual cameras that use orthographic projection. [3] extended the above approach for general perspective cameras. To resolve the problem of degenerate ellipsoids, they devised a non-linear optimization approach by forcing the dual quadric to lie on the subspace of ellipsoids. However, the above approaches require an ellipse-fitting step around each detected object, which leads to reliance on the conics' accurate shape estimation.

[6] directly estimated camera pose and dual quadric parameters from 2D bounding boxes in multiple views. They constructed planar constraints by back-projecting lines of bounding boxes. Then, they applied singular-value decomposition (SVD) to fit the dual quadric parameters. However, that approach requires many diverse viewpoints i.e. from an orbiting camera path. Consequently, this approach is unsuitable for autonomous driving scenarios with forward-translating motion. To overcome the above problem, [7] proposed ROSHAN by adding an additional planar constraint from image texture derived from 3D map points. For fast initialization, they relied on three strong assumptions as described in Section I with limited applicability.

Rather than using a monocular camera, [8] utilized an RGB-D camera to initialize the dual quadric by combining object detection and point cloud segmentation. However, majority RGB-D cameras are only suitable in the indoor environment, and this method relies on multiple-view bounding boxes from object detection when the point cloud is missing owing to illuminations or occlusions.

Typically, estimating the full 3D shape of objects from sparse views using purely classic geometric approaches is difficult. [14] adapted the Point Set Generation Net [15] trained on a CAD model repository to provide the accurate 3D shape of objects as point clouds from a single RGB image. Fitting the ellipsoid from the generated point cloud enriches the dual quadric inheritance coarse representation. However, this approach requires pretrained neural networks and lacks generality.

III. DUAL QUADRIC INITIALIZATION

A. Problem Formulation

The goal of this study is to model and represent static landmarks, e.g. static vehicles, as dual quadrics and infer their parameters in forward-translating camera movement.

Theoretically, a point quadric Q is a 4 by 4 symmetric matrix constrained by 3D points while a dual quadric Q^* is constrained by plane envelopes [5]. The relationship between them is $Q^* = Q^{-1}$, if Q is invertible. Accordingly, in a SLAM system, we intend to combine 3D points representing an object derived from triangulation and planar constraints of the object derived from 2D detections to initialize the dual quadric. This method does not rely on any assumption, and is therefore more generic than that presented in [7].

B. Constraints Derived from 2D Bounding boxes

Let us consider a set of image frames $F = 1...f$ (with $F > 3$) captured from straight-line vehicle motion under a single viewpoint. Using the object detector, each object i in an image frame f is identified by a 2D axis-aligned bounding box of a set of four lines l. The camera projection matrix $P = K[R|t]$ is calculated using the pose parameters $R|t$ of each frame f and the intrinsic K of the camera.

Thus, $\boldsymbol{\pi} = P^\top l$ represents the homogeneous vectors of planes by back-projecting these lines. These planar constraints will form a linear system $Aq = 0$ where A contains all coefficients derived from $\boldsymbol{\pi}$. The singular value decomposition (SVD) of Aq is $Aq = UDV^\top$. The last column of V can be considered the least-squares solution q that minimizes $||Aq||$.

Unfortunately, owing to the slight change in viewing angles during straight-line motion, similar planar constraints yield an ill-conditioned coefficient matrix A and unstable solution; the obtained dual quadric landmark may thus be imaginary or spawned behind the camera.

C. Constraints Derived from 3D Points

In a visual SLAM system, multi-view triangulation generates 3D points that represent an object. We tried to directly solve the below-listed equations by incorporating these 3D points.

$$\begin{cases} \boldsymbol{\pi}^\top Q^{-1} \boldsymbol{\pi} = 0 \\ \mathbf{x}^\top Q \mathbf{x} = 0 \end{cases} \tag{1}$$

Here, Q is the point quadric. $\mathbf{x}$ and $\boldsymbol{\pi}$ are both known 4 by 1 vectors representing 3D map points and planar constraints, respectively.

The vectorization vec operator [16] and Kronecker product $\otimes$[17] is employed to Equation 1:

$$vec(\boldsymbol{\pi}^\top Q^{-1} \boldsymbol{\pi}) = vec(\mathbf{x}^\top Q \mathbf{x}) \tag{2}$$

$$(\boldsymbol{\pi}^\top \otimes \boldsymbol{\pi}^\top) vec(Q^{-1}) = (\mathbf{x}^\top \otimes \mathbf{x}^\top) vec(Q). \tag{3}$$

So, we can rewrite Equation 2 as:

$$\begin{bmatrix} \mathbf{x}^\top \otimes \mathbf{x}^\top & -(\boldsymbol{\pi}^\top \otimes \boldsymbol{\pi}^\top) \\ \vdots & \vdots \\ \mathbf{x}^\top \otimes \mathbf{x}^\top & -(\boldsymbol{\pi}^\top \otimes \boldsymbol{\pi}^\top) \end{bmatrix} \begin{bmatrix} vec(Q) \\ vec(Q^{-1}) \end{bmatrix} = 0 \tag{4}$$

However, we find Equation 4 is difficult to solve directly even using numerical methods. Alternatively, we transform the 3D map points to their 3D convex hull [18], which is the smallest convex set containing the points. Then, we extract the surfaces of the convex hull as planar constraints $\boldsymbol{\pi}_c$. In this manner, the point constraints are converted to planar constraints:

$$\begin{cases} \boldsymbol{\pi}_b^\top Q^{-1} \boldsymbol{\pi}_b = 0 \\ \boldsymbol{\pi}_c^\top Q^{-1} \boldsymbol{\pi}_c = 0 \end{cases} \tag{5}$$

where $\boldsymbol{\pi}_b$ is the homogeneous vector of planes constructed from the 2D bounding box. $\boldsymbol{\pi}_c$ is the homogeneous vector of planes constructed from the 3D convex hull.

D. The Algorithm

The proposed initialization method can be integrated into visual-based SLAM systems, such as ORB-SLAM2 [19]. We adopt an object detector to detect the 2D bounding boxes of objects in each image. Then, the triangulated 3D map points within each bounding box are accumulated. If the number of map points exceeds a given threshold, the object will be marked as a candidate and its dual quadric representation will be solved.

1) Convex Hull Surface Extraction: Because the 2D bounding box cannot closely bound the outline of an object, extracted map points should be clustered. We adopt DBSCAN [20] to filter out the background map points and outliers; then, we construct the 3D convex hull based on the remaining map points located on the object.

Herein, we consider different strategies for extracting the convex hull surfaces. The first strategy is to utilize the entire surfaces of the convex hull and the other is to extract either visible or invisible surfaces from the convex hull. The former strategy is straightforward for considering the robust initialization of the dual quadric, because more constraints derived from surfaces are preferable. The second strategy of visible surface extraction is also intuitive because along the forward-translating camera motion, the backside of the object could not be observed. Therefore, most of the map points are located on the object's front side, that compose the visible surfaces. However, it's also interesting to compare it with the complementary strategy, i.e. invisible surface. All the three strategies are compared in Section V-A.

To extract the visible surfaces from a convex hull, the field of view (FOV) of the camera, the viewing vector from the camera to a surface vertex, and the surface normal vector should be considered. The visibility of a surface i is evaluated using the angle α_i between its normal vector (red arrow) and the vector V_i from the camera to any one of its vertices, as described in Figure 1 in 2D. When α_i is obtuse and V_i is within the FOV of the camera, the surface should be facing

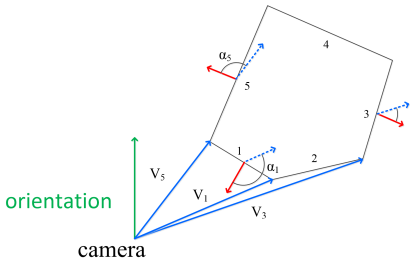


Fig. 1. The criterion for extracting the visible surface of a convex hull demonstrated in 2D, in which red arrows represent the normal vector of a surface, the green arrow represents camera orientation, and blue solid arrows represent the viewing vector, V_i, from the camera to vertices. α_i is the angle between V_i and the surface normal vector, which must be obtuse.

the camera and visible, as demonstrated by surface 1 and 5. It should be noted that visibility of an object's convex hull surfaces should be evaluated in all assoicated camera frames of the object rather than in a single view. The extracted visible surfaces are then introduced as planar constraints for the dual quadric initialization.

2) Solving the Combined Planar Constraints Equation: Similar to [6], [3], we utilize least-squares to solve the planar constraints equations and parametrize the dual quadric as ellipsoid with nine independent parameters. As the solution of the SVD is not constrained to an ellipsoid, we extract the dual quadrics rotation $R \in SO(3)$, translation $\mathbf{t} \in \mathbb{R}^3$, and shape $\mathbf{s} \in \mathbb{R}^3$ along the three semi-axes of the ellipsoid. The specific solution steps can be referred to [3].

IV. Experiment and Analysis

We evaluate the performance of our approach on the publicly available KITTI dataset [21] and TUM RGB-D dataset [22]. We choose six KITTI raw data sequences in the city and residential scenarios with the most number of ground truth object annotations[1]. Unlike the KITTI dataset, the dimension of generic objects in the TUM RGB-D dataset cannot be assumed and there is no ground truth for objects. Therefore, we choose three TUM sequences to demonstrate the generality of the proposed approach.

We choose the YOLOv3 detector [12] for 2D object detection, with a confidence of 0.6. On the KITTI dataset, the accumulated map point of an object is evaluated with the thresholds of 20, 40, and 60. While working on the TUM RGB-D dataset, the map point only needs to be accumulated to 30 owing to a closer observation of objects. In addition, the DBSCAN parameters are set as distance threshold = 0.1 or 0.5 and minimum number of points = 10 for both datasets.

A. Quantitative Evaluation Criteria

We evaluate the obtained results on the basis of the position error of dual quadrics, the number of imaginary solution, shape conformity in 2D intersection over union (IoU), and relative success ratio of initialization.

[1]The object annotations are provided by tracklets with object locations and types

1) Position Error: We measure the root mean square error (RMSE) between the center position of the dual quadric and the position provided in the ground truth as the position error.

2) The Number of Imaginary Solutions (Ino): We record the number of imaginary solutions, which are evaluated by the degeneratiy of projected ellipse using the criteria described in [23].

3) 2D IoU: If the ellipse solution is not imaginary, the center (x_0, y_0), semi-major axis a, semi-minor axis b, and rotation angle θ of the ellipse can be obtained from the ellipse's general equation. Then, the axis-aligned bounding box $bbox$ of the ellipse can be calculated, and the IoU is calculated between $bbox$ and the 2D bounding box from the ground truth. If the dataset does not provide the ground truth of objects, such as TUM, we propose to use the bounding boxes from the 2D object detection as an alternative.

4) Relative Success Ratio (RSR): A 2D IoU more than the threshold of 0.5 is considered a successful initialization herein. From static ground truth object annotations, we record the number of candidate objects Cno whose map points exceed the given threshold. Next, the RSR can be computed using Sno/Cno, where Sno represents the number of successfully initialized objects.

B. KITTI Results

For comparison, we re-implemented the approach of exclusively using 2D bounding boxes in multiple views according to [6] and the approach of [7]. As [6] assumes known data association, which implies that all observation frames of the object are used, we use all detection results of the object to construct planar constraints. For [7] implementation, we compute the ellipsoid depth from the average depth of the map points associated with the object for a more accurate initial estimate instead of using the rough scene depth. We set the initial orientation of the ellipsoid as identical to the camera coordinate frame. The ellipsoid's dimension are initialized as the prior car size (width = 1.6 m, length = 3.9 m, and height = 1.5 m) for the KITTI dataset [4]. If not specified, we use the entire surfaces of the convex hull and the 2D bounding box in the initialization.

1) The Impact of Different Map Point Accumulation Numbers: Table I, Table II, and Table III show quantitative results of the accumulated map point threshold of the evaluation object reaching 20, 40, and 60, respectively. The map point number can substantially influence the proposed approach in terms of localization accuracy. When the number of map points accumulates from 20 to 60, the average position error is reduced from 3.263 to 1.580 m, and the average 2D IoU and RSR increase from 0.685 to 0.80 and from 87.2% to 97.7%, respectively. and almost has no imaginary solution even if the number of map points is only accumulated 20. Compared with [7], the proposed method is considerably better in terms of shape estimation and success ratio of initialization even if the map points number is around 20, and it achieves the same accuracy of dual quadric center localization when the map points threshold is 60. In addition, we visualize an example to show the impact of different numbers of map points on dual quadric initialization in Figure 2 when the threshold reaches 20 and 40.

Considering the latency introduced by map points accumulation, we record the average number of frames required to accumulate about 20, 40, and 60 map points when the object was detected on the KITTI dataset in Table IV. The results show that the object can accumulate 40 map points and only require the observation of around four frames using sparse ORB features; thus, the proposed approach can guarantee real-time dual quadric initialization.

2) The Impact of Camera Movement: As shown in Figure 3, our approach performs well during camera turning. While [7] fails in this case because of its reliance on strong assumptions of the objects' dimension and the pose of the camera. As shown in Table V, [6] fails in forward-translating motion using all detection frames of the objects (over 11 frames in average).

C. TUM Results

Because assuming shape priors for generic objects is impossible, [7] is less applicable in this dataset. Additionally, [6] requires numerous observation views on the orbiting camera path to perform well, causing a serious initialization latency problem. Therefore, we compare the results obtained herein using the TUM RGB-D dataset with those presented in [8]. As shown in Table VI, we only need one observation view to accumulate around 30 map points, the results have no imaginary solution, and the average IoU is 0.802. The success ratio is 92.7%, revealing an increment of 16.2% compared with that obtained in [8] using RGB-D images.

V. Disscussions

A. The Effectiveness of Convex Hull Surfaces

We compared the results of using only visible convex hull surfaces (VC), using only invisible convex hull surfaces (IC), using visible convex hull surfaces and the bounding box as constraints (VCB) with those obtained using invisible surfaces and the bounding box (ICB) in Table VII. The comparison shows that the number of visible and invisible faces is basically the same, the results are not ideal either using only visible or using only invisible surfaces. However, the RSR significantly improves when they are combined with the bounding box respectively, especially for invisible surfaces; and combining visible surfaces and the bounding box causes more imaginary solutions. This counterintuitive result can be interpreted as the fact that the enveloping space fromed by ICB is convex, while the space fromed by VCB is concave. The former provides a better approximation of the dual quadric as an ellipsoid.

B. The Effectiveness of the Convex Hull Constraints and Native Clustered Map Points Constraints

It is possible to construct a quadric directly from the native clustered mapping points and invert it to find the dual quadric surface, then parameterize it into an ellipsoid. The comparison between AC and MP in Table VIII shows that constraints derived from convex hull outperform constraints

TABLE I

DUAL QUADRIC INITIALIZATION RESULTS ACCUMULATED ABOUT 20 MAP POINTS ON THE KITTI RAW SEQUENCE

sequencene	Position error		2D IoU		Gno[1]	Cno	Ino		RSR	
	[7]	Ours[2]	[7]	Ours			[7]	Ours	[7]	Ours
seq09	**1.891**	2.230	0.524	**0.710**	86	57	0	0	40.6%	**94.7%**
seq22	**1.967**	3.254	0.611	**0.732**	53	48	0	0	75.0%	**91.7%**
seq23	**1.899**	3.026	0.546	**0.628**	157	70	0	0	68.6%	**75.7%**
seq36	**3.040**	4.326	0.599	**0.699**	80	43	0	0	67.4%	**93.0%**
Seq59	**1.925**	2.420	0.522	**0.691**	54	29	0	0	44.8%	**86.2%**
Seq93	**2.440**	4.223	0.551	**0.649**	56	33	0	1	54.5%	**81.8%**
Mean	**2.194**	3.263	0.559	**0.685**	-	-	0	0.167	58.5%	**87.2%**

[1] The number of static objects with ground truth. [2] We use the entire surfaces of the convex hull and the 2D bounding box.

TABLE II

DUAL QUADRIC INITIALIZATION RESULTS ACCUMULATED ABOUT 40 MAP POINTS ON THE KITTI RAW SEQUENCE

sequencene	Position error		2D IoU		Gno	Cno	Ino		RSR	
	[7]	Ours	[7]	Ours			[7]	Ours	[7]	Ours
seq09	1.635	**1.575**	0.560	**0.787**	86	54	0	0	70.3%	**100%**
seq22	**1.819**	2.037	0.637	**0.807**	53	46	0	0	82.6%	**100%**
seq23	**1.625**	1.972	0.591	**0.733**	157	53	0	0	73.6%	**94.3%**
seq36	**2.455**	2.604	0.616	**0.745**	80	38	0	0	76.3%	**94.7%**
seq59	1.568	**1.433**	0.589	**0.807**	54	26	0	0	65.4%	**100%**
seq93	**1.811**	2.398	0.549	**0.721**	56	33	0	0	60.6%	**87.9%**
Mean	**1.819**	2.00	0.590	**0.767**	-	-	0	0	71.5%	**96.2%**

TABLE III

DUAL QUADRIC INITIALIZATION RESULTS ACCUMULATED ABOUT 60 MAP POINTS ON THE KITTI RAW SEQUENCE

sequencene	Position error		2D IoU		Gno	Cno	Ino		RSR	
	[7]	Ours	[7]	Ours			[7]	Ours	[7]	Ours
seq09	1.603	**1.231**	0.544	**0.807**	86	45	0	0	73.3%	**100%**
seq22	**1.624**	1.674	0.633	**0.807**	53	37	0	0	89.1%	**100%**
seq23	**1.285**	1.395	0.624	**0.789**	157	27	0	0	88.9%	**96.3%**
seq36	**2.080**	2.118	0.601	**0.782**	80	31	0	0	77.4%	**96.8%**
seq59	**1.078**	1.108	0.611	**0.846**	54	15	0	0	66.7%	**100%**
seq93	**1.650**	1.959	0.611	**0.766**	56	28	0	0	75%	**92.9%**
Mean	**1.553**	1.580	0.609	**0.800**	-	-	0	0	78.4%	**97.7%**

TABLE IV

THE NUMBER OF FRAMES REQUIRED FOR MAP POINT ACCUMULATION

sequencene	The number of frames		
	20	40	60
seq09	1.684	3.939	5.433
seq22	1.774	4.654	6.680
seq23	1.521	3.567	5.050
seq36	1.644	4.256	5.850
seq59	1.357	3.250	5.054
seq93	1.725	4.395	5.974
Mean	1.618	4.010	5.674

(a) The object in sequence93 when map points threshold is 20

(b) The object in sequence93 when map points threshold is 40

Fig. 2. Results for dual quadric initialization of the same object when the map point threshold is 20 and 40. The left and right images are representative frames, showing a close up of the views with the output of ground truth (red box) and the YOLOv3 detection (green box) and projections of the dual quadric estimated using [7] (yellow ellipse) and Ours (blue ellipses).

(a) Objects on sequence09

(b) Objects on sequence22

Fig. 3. Results for the dual quadric initialization when the camera turns. The left and right images are representative frames, showing a close-up of the views with the output of ground truth (red box) and the YOLOv3 detection (green box) together with the projections of the dual quadric estimated using [7] (yellow ellipse) and Ours (blue ellipses).

TABLE V

DUAL QUADRIC INITIALIZATION RESULTS OBTAINED BY EXCLUSIVELY USING MULTIPLE-VIEW BOUNDING BOXS ON THE KITTI RAW SEQUENCE

sequencene	OVno[1]	position error	2D IoU	Gno	Cno	Ino	RSR
seq09	9.754	3.334	0.204	86	47	1	10.6%
seq22	9.660	2.986	0.267	53	35	0	17.1%
seq23	7.868	6.558	0.159	157	49	9	12.2%
seq36	12.045	5.534	0.228	80	27	1	11.1%
seq59	12.676	4.648	0.153	54	23	1	8.7%
seq93	17.485	5.656	0.213	56	25	2	12%
Mean	11.581	4.786	0.204	-	-	2.33	11.95%

[1]The average number of observation views.

TABLE VI

DUAL QUADRIC INITIALIZATION RESULTS ON TUM RGB-D DATASETS

Dataset	AMPno[1]	OVno	2D IoU	Ino	Total Objects[2]	success ratio	
						Ours	[8]
fr2_dishes	32	1.232	0.849	0	5	**100%**	80%
fr2_desks	31	1.156	0.847	0	16	**93.8%**	81.3%
fr3_office	26	1.180	0.712	0	32	**84.4%**	78.1%
mean	29.6	1.189	0.802	0	-	**92.7%**	79.8%

[1]The number of average accumulated map points. [2]The total detected objects without ground truth.

derived from native clustered map points in a large margin. The reason might be that the convex hull provides a more reasonable approximation of the ellipsoid while map points represent the irregular shape of the object surface and are noisy.

C. The Effectiveness of Combined Planar Constraints

Table VIII also shows that the average position error decreases by 22.39% when a convex hull is exclusively used for planar constraints (AC) as compared to using combined planar constraints derived from convex hull and 2D detection (ACB). However, the average 2D IoU of AC is merely 0.389. This indicates that the map points are more precise for locating the object; however, they are unable to represent the entire object, resulting in a relatively small estimated shape. Thus, combining the convex hull with the 2D bounding box for planar constraints generates more reasonable dual quadric initialization.

D. Integrating Dual Quadric Initialization into a Monocular SLAM System

We integrate the proposed dual quadric initialization method into the monocular ORB-SLAM2 system. The successfully initialized dual quadrics participate both in the local and global bundle adjustment (BA) with the fixed estimated dimension. To refine the shape of the dual quadric, they will be regenerated on the basis of the later associated observation data and updated map points after BA.

Figure 4 shows the mapping results after final optimization on the KITTI odometry seqence00 and seqence07. Post optimization, the average 2D IoU slightly increases from 0.68 to 0.73. By filtering out objects whose 2D IoU falls below 0.5 during mapping, sequence00 retains 96.2% of the initialized 106 objects, whereas sequence07 retains 60.7% of the initialized 33 objects, indicating that multiple loop closures help in retaining the stability of object mapping.

VI. Conclusions

In summary, herein, we presented a novel robust dual quadric initialization approach by constructing combined planar constraints. One of the constraints is a 3D convex hull constructed from map points post clustering, and the other is back-projection of 2D object detections bounding box. On the KITTI raw dataset, the number of map points substantially impacts our method in terms of localization accuracy. Compared with the SOTA, we can achieve the same accuracy of center localization of the dual quadric when the map points accumulated to 60, with an average position error of 1.58 m. However, we improved the average 2D IoU and initialization success ratio by 20% and 24.6%, respectively. Moreover, because our approach does not rely on any assumptions or priors, it is more generic and suitable for common straight-line and turning in the field of autonomous driving. In addition, the performance of our approach on the TUM RGB-D dataset is better than that of the SOTA using RGB-D. Finally, we show the mapping results by integrating the proposed dual quadric initialization method into the monocular ORB-SLAM2 system. Future work would include the robust data association and optimization strategy for dual quadrics.

References

[1] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D Object Detection for Autonomous Driving," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 2147–2156.

[2] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 424–432.

[3] 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, pp. 1281–1294, 2018.

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

[5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[6] L. Nicholson, M. Milford, and N. Sunderhauf, "QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2019.

[7] K. Ok, K. Liu, K. Frey, J. P. How, and N. Roy, "Robust object-based SLAM for high-speed autonomous navigation," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, 2019, pp. 669–675.

[8] Z. Liao, W. Wang, X. Qi, and X. Zhang, "Rgb-d object slam using quadrics for indoor environments," *Sensors*, vol. 20, no. 18, p. 5150, 2020.

[9] S. D. Ma and L. Li, "Ellipsoid reconstruction from three perspective views," in *Proceedings of the 1996 International Conference on Pattern Recognition (ICPR '96) Volume I - Volume 7270*, ser. ICPR '96. USA: IEEE Computer Society, 1996, p. 344.

[10] G. Cross and A. Zisserman, "Quadric reconstruction from dual-space geometry," in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98. USA: IEEE Computer Society, 1998, p. 25.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[12] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv e-prints*, p. arXiv:1804.02767, Apr. 2018.

[13] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," 2016, pp. 4141–4149.

[14] M. Hosseinzadeh, K. Li, Y. Latif, and I. Reid, "Real-time monocular object-model aware sparse SLAM," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, 2019, pp. 7123–7129. [Online]. Available: http://arxiv.org/abs/1809.09149

[15] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017, pp. 2463–2471.

[16] H. V. Henderson and S. R. Searle, "Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics," *Canadian Journal of Statistics*, vol. 7, no. 1, pp. 65–81, 2010.

[17] J. W. Brewer, "Correction to "kronecker products and matrix calculus in system theory"," *IEEE Transactions on Circuits and Systems*, vol. 26, no. 5, p. 360, 1979.

[18] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *Acm Transactions on Mathematical Software*, vol. 22, no. 4, 1998.

[19] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on*

TABLE VII

DUAL QUADRIC INITIALIZATION RESULTS OF WHETHER USE VISIBLE CONVEX HULL SURFACES ON KITTI RAW SEQUENCE09

MPT[1]	the number of convex hull surfaces				position error				2D IoU				Ino				RSR			
	VC[2]	IC[3]	VCB[4]	ICB[5]	VC[2]	IC[3]	VCB[4]	ICB[5]	VC[2]	IC[3]	VCB[4]	ICB[5]	VC[2]	IC[3]	VCB[4]	ICB[5]	VC[2]	IC[3]	VCB[4]	ICB[5]
20	12	11	12	11	60.868	812.802	8.765	**5.516**	0.198	0.165	0.384	**0.720**	12	18	23	**1**	12.3%	8.8%	50.9%	**94.7%**
40	17	17	17	17	4.511	8.622	4.570	**3.333**	0.318	0.359	0.617	**0.774**	2	2	9	**0**	22.2%	22.2%	75.9%	**96.3%**
60	22	20	22	20	1.734	**1.206**	3.005	2.332	0.420	0.432	0.735	**0.800**	1	1	4	**0**	28.9%	31.1%	93.3%	**100%**
mean	17	16	17	16	22.371	274.21	5.447	**3.951**	0.312	0.318	0.579	**0.764**	5	7	12	**0.33**	21.3%	20.7%	73.4%	**97%**

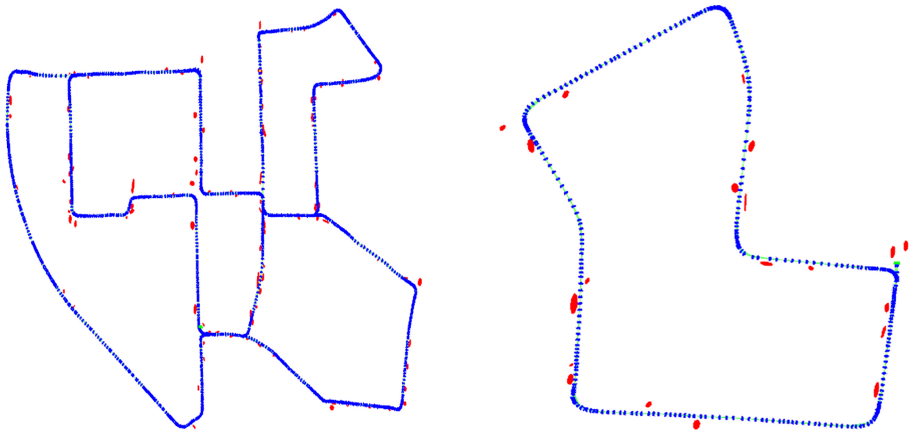
[1]The number of map points threshold. [2] Using visible convex hull surfaces for planar constraints. [3] Using invisible convex hull surfaces for planar constraints. [4] Using visible convex hull surfaces and the bounding box for planar constraints. [5] Using invisible convex hull surfaces and the bounding box for planar constraints.

TABLE VIII

DUAL QUADRIC INITIALIZATION RESULTS OF USING DIFFERENT CONSTRAINTS ON KITTI RAW SEQUENCE09

MPT[1]	the number of convex hull surfaces			position error			2D IoU			Ino			RSR		
	ACB	AC	MP[2]	ACB	AC	MP	ACB	AC	MP	ACB	AC	MP	ACB	AC	MP
20	23	23	-	2.230	**1.561**	3.467	**0.710**	0.291	0.245	0	0	0	**94.7%**	14.0%	7.0%
40	34	34	-	1.575	**1.288**	2.489	**0.787**	0.392	0.312	0	0	2	**100%**	24.1%	20.4%
60	42	42	-	1.231	**1.060**	3.559	**0.807**	0.485	0.298	0	0	1	**100%**	42.2%	20%
Mean	33	33	-	1.679	**1.303**	3.172	**0.768**	0.389	0.285	0	0	1	**98.2%**	26.8%	15.8%

[1] The number of map points threshold. [2]Using native map points after DBSCAN for point constraints.



(a) The mapping results on the KITTI odometry sequence00

(b) The mapping results on the KITTI odometry sequence07

Fig. 4. Mapping results in ORB-SLAM2 based on the proposed dual quadric initialization. Blue represents the estimated trajectory and red ellipsoids represent good objects with a 2D IoU of over 0.5 post global BA optimization.

Knowledge Discovery and Data Mining, ser. KDD'96. AAAI Press, 1996, p. 226–231.

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

[22] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.

[23] J. D. Lawrence, *A Catalog of Special Plane Curves*. Dover Publ, 1972.