Object-Aware SLAM Based on Efficient Quadric Initialization and Joint Data Association

ZhenZhong Cao[1], Yunzhou Zhang[1*], Rui Tian[1], Rong Ma[2], Xinggang Hu[1],
Sonya Coleman[3], Dermot Kerr[3]

Abstract— **Semantic simultaneous localization and mapping (SLAM) is a popular technology enabling indoor mobile robots to sufficiently perceive and interact with the environment. In this paper, we propose an object-aware semantic SLAM system, which consists of a quadric initialization method, an object-level data association method, and a multi-constraint optimization factor graph. To overcome the limitation of multi-view observations and the requirement of dense point clouds for objects, an efficient quadric initialization method based on object detection and surfel construction is proposed, which can efficiently initialize quadrics within fewer frames and with small viewing angles. The robust object-level joint data association method and the tightly coupled multi-constraint factor graph for quadrics optimization and joint bundle adjustment enable the accurate estimation of constructed quadrics and camera poses. Extensive experiments using public datasets show that the proposed system achieves competitive performance with respect to accuracy and robustness of object quadric estimation and camera localization compared with state-of-the-art methods.**

I. INTRODUCTION

Object representation is a key issue within object-level semantic SLAM, and appropriate representation can not only promote the robustness and accuracy of localization but also enhance the information of a semantic map oriented for human-robot interactions. There are many kinds of object representation methods, including preset models and general models, where prior object point clouds, cubes and quadrics are three kinds of common representation methods utilized for object-level semantic SLAM [1]–[9]. SLAM++ [1] presents an object-oriented 3D SLAM paradigm, which requires prior CAD models of objects. CubeSLAM [2] models objects as cubes, which is the first example of object-oriented SLAM. Compared with the cube methods, the quadric approach has a more compact mathematical form of a direct projection model which facilitates the update of model parameters, hence there is much research on quadric-based SLAM systems [3]–[9]. The approaches to quadric initialization can be broadly divided into two categories: methods based on object detection [3]–[6], [9], and methods based on point cloud fitting [7], [8].

QuadricSLAM [3] was the first approach to introduce a quadric representation of objects as landmarks and presented a novel method of initializing quadrics using 2D object detection results from multi-view frames. Subsequently, Hosseinzadeh *et al.* [4] used the idea of the quadric initialization in [3] but improved the optimization of quadrics by introducing plane landmarks and constructing tangent constraints between quadrics and planes. They further enhance this work in [5] and introduced a deep learning network [10] to estimate the point cloud distribution of the object to constrain the scale of the quadric. Ok *et al.* [6] addressed the problem of vehicle quadric construction, and reduced the difficulty of the quadric initialization with object detection, image texture, and prior semantic scale to jointly estimate the parameters of the quadric. However, the disadvantage of the above methods is that multiple frames of large viewing angles are required to initialize the quadrics.

The limitation of methods based on point cloud fitting are that a complete point cloud segmentation of the object is required, which can be difficult due to occlusions and noise. Liao *et al.* [7] introduced the hypothesis of symmetry, which is used to complement the point clouds of the object when fitting a more complete quadric. They also proposed a hybrid quadric estimation method based on point cloud segmentation and object detection, and extended the non-parametric data association method to the quadric for the first time [8]. Chen *et al.* [9] focused on scenes of outdoor forward translational motion and proposed a fast initialization method of quadrics based on sparse map points, which provided the inspiration for quadric initialization under small viewing angles.

In this paper, we propose an efficient quadric initialization (EQI) method based on object detection and surfel construction, which not only overcomes the limitation of multiple frames and large-view observations but also reduces the requirement of the object dense point cloud data when compared with state-of-the-art methods. With the extra constraints of the surfel construction, the EQI can construct quadrics within fewer frames and with small viewing angles. As for data association of object detection results in current frame and constructed quadrics in the map, we propose a robust object-level joint data association (JDA) method combining mixed information, which fully considers the factors of 2D image plane, 3D map projection, and statistic distributions. The multi-constraint factor graph including semantic and geometric constraints is presented for both quadrics optimization and joint bundle adjustment. Finally, combining the all above modules, a complete object-aware semantic SLAM system is formed.

[*] The corresponding author of this paper.

[1] ZhenZhong Cao, Yunzhou Zhang, Rui Tian, Xinggang Hu are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (Email: zhangyunzhou@mail.neu.edu.cn).

[2] Rong Ma is with Beijing Simulation Center, China (Email: mar_buaa@163.com).

[3] Sonya Coleman and Dermot Kerr are with School of Computing, Engineering and Intelligent Systems, Ulster University, N. Ireland, UK.

This work was supported by National Natural Science Foundation of China (No. 61973066), Major Science and Technology Projects of Liaoning Province(No. 2021JH1/10400049), Fundation of Key Laboratory of Equipment Reliability(No. WD2C20205500306) and Fundation of Key Laboratory of Aerospace System Simulation(No. 6142002200301).

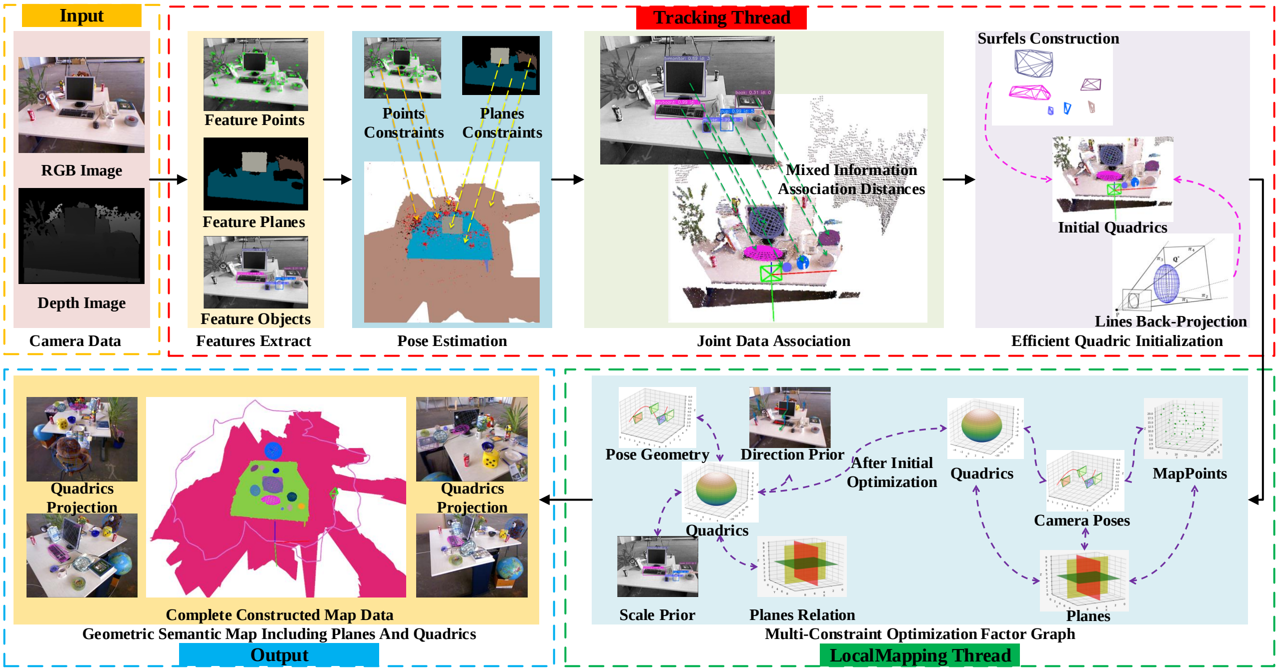


Fig. 1. Overview of the proposed system. There are two main parts in the system: 1) The Tracking thread extracts features, matches features, estimates current camera pose, associates detections with quadrics, and initializes the quadrics. 2) The LocalMapping thread optimizes camera poses, points, planes, and quadrics of the geometric semantic map.

We demonstrate the performance of the proposed system using public indoor datasets. Experimental results demonstrate that our system reaches competitive performance compared with the state-of-the-art methods with respect to both localization accuracy and mapping accuracy.

The main contributions of this work are as follows:

- We propose an efficient quadric initialization (EQI) method based on object detection and surfel construction which initializes quadrics using fewer frames with small viewing angles.
- We propose a robust object-level joint data association (JDA) method combining multi-dimensional information and statistic distributions.
- We propose a multi-constraint optimization factor graph for quadrics optimization and joint bundle adjustment.
- We implement a complete visual semantic SLAM system, aiming to build a novel object-oriented and semantically-enhanced map for indoor robot interaction.

II. System Overview

A. Mathematical Representation

The notations used in this paper are as follows:

- $T_{k,w} \in R^{4\times4}$ - The camera pose of image frame I_k in the global frame, which includes a rotation component $R_{k,w} \in R^{3\times3}$ and a translation component $t_{k,w} \in R^{3\times1}$. $T_{k,k-1} \in R^{4\times4}$ represents the relative camera pose between image frame I_k and image frame I_{k-1}.
- $Q_w \in R^{4\times4}$ - The quadric parameter matrix in the global frame, $Q_w^* \in R^{4\times4}$ is the dual form, which are both symmetric matrices, $q_w^* \in R^{10\times1}$ is the vector form of Q_w^*, $C_k^{Q_w^*} \in R^{3\times3}$ represents the dual conic projected from the quadric Q_w^* to image frame I_k.
- $\pi_w \in R^{4\times1}$ - The plane parameter vector in the global frame, which includes a normal vector $n \in R^{3\times1}$ and the distance to the origin $d \in R$, $\pi_k \in R^{4\times1}$ represents the plane parameter vector in frame I_k.
- $X_w \in R^{4\times1}$ - The 3D homogeneous coordinates in the global frame, $X_k \in R^{3\times1}$ represents the 3D homogeneous coordinates in frame I_k, $u_k \in R^{3\times1}$ represents the 2D homogeneous coordinates in frame I_k, $l_k \in R^{3\times1}$ represents line vector in frame I_k.
- D_k - The bounding box of object detection in frame I_k, $B_{D_k} \in R^{4\times4}$ represents the bounding box (BBox) of object detection D_k, $\text{cls}(D_k)$ represents the classification of D_k.
- O_w - The object of the map in the global frame.
- $K \in R^{3\times3}$ - The intrinsics of the pinhole camera model, $P_{k,w} = K[R_{k,w}|t_{t.w}] \in R^{3\times4}$ represents projection matrix from the global frame to frame I_k.

B. System Architecture

The framework of the system we proposed is illustrated in Fig.1, which is modelled on the RGB-D interface of ORB-SLAM2 [11]. The depth image is used to obtain depth information of the scene to avoid the monocular scale problem [12] and perform plane extraction. However, to highlight our work, we only present our added or modified parts instead of showing all details of ORB-SLAM2. The main changes in our approach are:

- **The Tracking thread** receives the RGB images and the depth images and then extracts point features using the ORB feature extractor, plane features are obtained using the plane segmentation algorithm [13] and object detection is performed with YOLOv5. Then the point features and plane features are matched with map points and map planes respectively. After using the associated relationship of points and planes to solve the initial pose, object-level JDA is used to determine the object detection results for the current frame and the quadrics that have been constructed in the map. The EQI is performed for objects that are associated but not yet initialized and those that are not associated.

- **The LocalMapping thread** performs the multi-constraint quadric optimization for the objects that have just been initialized or newly observed in the map. The optimized quadrics then participate in the tightly coupled joint optimization of poses and landmarks. Finally, it outputs a complete geometric semantic map.

III. Efficient Quadric Initialization

The reason why it is difficult to achieve the single frame initialization of the quadric based on object detection is that the parameters of the quadric are 10-dimensional, and the bounding box of an object has only four sides, which can only provide four tangent plane constraints. However, within a small viewing angle, even if there are enough constraints, the relationship between them is approximately linear, which will cause the solution to be limited by similar observations or even to be divergent, especially for the position and shape of the quadric, which is shown in Fig.2. Therefore, following the work in [9], we also use the map points associated with objects to find more constraints, but we adopt a more robust data association strategy, which is lacking in [9]. In this way, we can obtain sufficient tangent plane constraints for efficient quadric initialization.

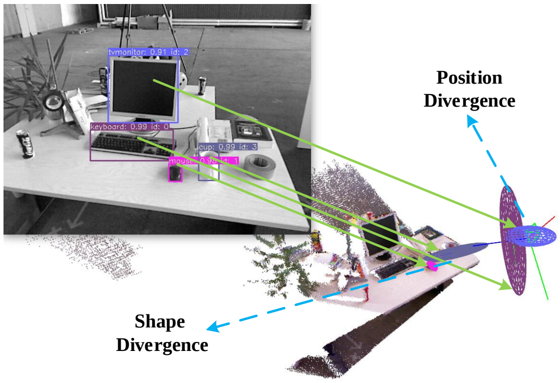


Fig. 2. The performance of quadric initialization that only uses the tangent planes provided by object detection within a small viewing angle.

A. Tangent Planes from Object Detection

Object detection can generate a bounding box for each object in the image, and the bounding box can be represented by points $U = \{u_k^1, u_k^2, u_k^3, u_k^4\}$ or lines $L = \{l_k^{12}, l_k^{23}, l_k^{34}, l_k^{41}\}$. The relationship between U and L is :

$$\begin{aligned} l_k^{12} &= u_k^1 \times u_k^2, l_k^{23} = u_k^2 \times u_k^3 \\ l_k^{34} &= u_k^3 \times u_k^4, l_k^{41} = u_k^4 \times u_k^1 \end{aligned} \tag{1}$$

Through the projection model of the camera, the line can be back-projected into a plane:

$$\pi_w = P_{k.w}^T l_k \tag{2}$$

Therefore, one bounding box can be converted to four tangent planes. With the movement of the camera, there will be many object detections corresponding to the same object. To make sufficient use of them, we use the JDA method to associate detections and objects, which will be introduced in detail in Section IV.

B. Tangent Planes from Surfel Construction

We have initially accumulated map points for the objects within the object detection boxes across frames. However, there are background map points that do not belong to these objects. As a result, we first use the depth distribution of points to filter initial outliers and then apply the isolated forest algorithm mentioned in EAO-SLAM [14] to cluster the map points and eliminate most of the remaining outliers. Finally, by referring to the convex polyhedron construction algorithm proposed in [15], we implement a surfel construction method that uses a 3D point set to obtain the tangent planes of the object surface. Fig. 3 shows the performance of surfel construction from a single frame.

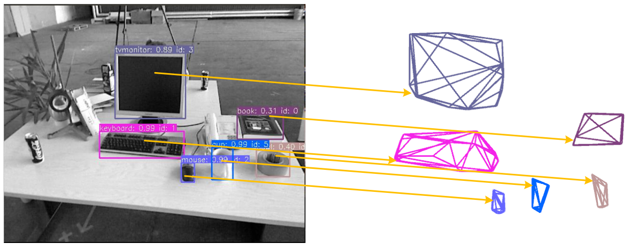


Fig. 3. Surfel construction performance from a single frame. The arrow associates the object with the corresponding surfels.

C. The Solution of Quadric Parameters

After obtaining the tangent planes $\{\pi_w^D\}$ and $\{\pi_w^X\}$ respectively, constructed by object detection and surfel construction, we define the operation $\rho(\cdot)$:

$$\begin{aligned} \rho(\pi) = \big(&\pi_1^2, 2\pi_1\pi_2, 2\pi_1\pi_3, 2\pi_1\pi_4, \pi_2^2, \\ &2\pi_2\pi_3, 2\pi_2\pi_4, \pi_3^2, 2\pi_3\pi_4, \pi_4^2\big) \end{aligned} \tag{3}$$

Using this, we can easily get the overall coefficient matrix:

$$A(\pi_w) = \begin{pmatrix} \rho(\pi_w^D) \\ \vdots \\ \rho(\pi_w^X) \end{pmatrix} \tag{4}$$

Then, we convert the tangent relationship $\pi_w^T Q_w^* \pi_w = 0$ into a linear expression:

$$A(\pi_w) q_w^* = 0 \tag{5}$$

$A(\pi_w) \in R^{n_{obs} \times 10}$, n_{obs} is the number of the effective tangent planes. Then Eq.(5) can be converted to a linear least squares problem:

$$(q_w^*)^{opt} = \arg\min_{q_w^*} \|A(\pi_w) q_w^*\|_2^2 \tag{6}$$

Using the Singular Value Decomposition (SVD) method, we can solve the above least squares problem. After obtaining q_w^*, we can get the parameter matrix of the dual quadric Q_w^*

according to the symmetry. Further, the parameter matrix of the original quadric Q_w can be calculated as:

$$Q_w = (Q^*_w |Q_w|^{-1})^{-1} = (Q^*_w)^{-1} \sqrt[3]{|Q^*_w|} \tag{7}$$

Due to the existence of noise, we can not completely trust the calculated Q_w. Therefore, after completing the calculation of the quadric parameters, we use the historical observation data accumulated in the JDA stage to evaluate Q_w. The specific process is to project the calculated quadric onto the image planes according to the historical observation poses and calculate the average 2D IoU with the associated object detections. If this value is greater than the threshold $\gamma = 0.5$ that we set in our experiment, we trust this initialization, otherwise we wait for more observations for quadric initialization of the object. In Alg. 1, we present the complete procedure of the EQI algorithm, which provides more details regarding the implementation.

IV. Object-level Joint Data Association

The accuracy of JDA not only affects the accuracy of the quadric initialization but also has an impact on the back-end optimization. To fully consider the mixed information in the scene, we design four kinds of association distance and assign different weights for them to construct the overall association distance matrix, where a_{ij} is the basic element. We set the weights k^q, k^w, k^e, and k^r to 0.2, 0.2, 0.2, and 0.4 respectively, in our experiment. Finally, we adopt the Hungarian algorithm [16] to solve this allocation problem:

$$a_{ij} = k^q a^q_{ij} + k^w a^w_{ij} + k^e a^e_{ij} + k^r a^r_{ij} \tag{8}$$

A. 2D Image IoU Association Distance

Since the JDA has been completed in I_{k-1}, the object detection results are all associated with constructed objects in the map. Therefore, we use the IoU between the i-th bounding box $B_{D^i_k}$ of I_k and the l-th bounding box $B_{D^l_{k-1}}$ of I_{k-1} that are associated with the j-th object in the map to indirectly compute the association distance:

$$a^q_{ij} = 1 - \frac{B_{D^i_k} \cap B_{D^l_{k-1}}}{B_{D^i_k} \cup B_{D^l_{k-1}}} \tag{9}$$

B. Object Projection IoU Association Distance

Some objects in the map have obtained the quadric parameters through the EQI algorithm. For these objects, we can use the projection model of the quadric to project them onto the image plane to obtain the quadratic curve, and then obtain the bounding box, so that we use the IoU between the projected bounding box $B_{O^j_w}$ and $B_{D^i_k}$ to calculate the association distance:

$$a^w_{ij} = 1 - \frac{B_{D^i_k} \cap B_{O^j_w}}{B_{D^i_k} \cup B_{O^j_w}} \tag{10}$$

Algorithm 1: Efficient Quadric Initialization (EQI)

```
Input: Object detections $\{D^i_k\}$ and 2D ORB features
       $\{u^j_k\}$ of current frame $I_k$, Objects $\{O^l_w\}$ of
       map
Output: Poses $\{T^m_w\}$ and shapes $\{S^m_w\}$ of new
        created quadrics $\{Q^m_w\}$
// first filter outliers of points
Points$\{\{X_w^{D^i_k}\}\} \leftarrow FilterByDepth(\{D^i_k\}, \{u^j_k\})$
Associations$\{D^i_k, O_w^{D^i_k}\} \leftarrow JDA(\{D^i_k\}, \{O^l_w\})$
for each object detection $D^i_k$ do
    Add $D^i_k$ and $\{X_w^{D^i_k}\}$ to {D} and {X} of $O_w^{D^i_k}$
    // second filter outliers of points
    $\{X\} \leftarrow FilterByIsolationForest(\{X\})$
    // count the number of object detections
    $DetsNum \leftarrow EffectiveDetection(\{D\})$
    // compute angle score among object detections
    $DetsScore \leftarrow ViewingAngleDiff(\{D\})$
    if $O_w^{D^i_k}$ is not initialized then
        // if detections are not adequate and suitable
        if $DetsNum < \alpha \vee DetsScore < \beta$ then
            // tangent planes from object detection
            $\{\pi^D_w\} \leftarrow LineBackProjection(\{D\})$
            // tangent planes from surfel construction
            $\{\pi^X_w\} \leftarrow SurfelsConstruction(\{X\})$
            // solve linear least square by SVD
            $Q^m_w \leftarrow SolveQuadric(\{\pi^D_w\}, \{\pi^X_w\})$
        end
        else
            $\{\pi^D_w\} \leftarrow LineBackProjection(\{D\})$
            $Q^m_w \leftarrow SolveQuadric(\{\pi^D_w\})$
        end
        // compute evaluation score for quadric
        $QuadricScore \leftarrow QuadricQualityCal(Q^m_w)$
        if $QuadricScore > \gamma$ then
            // discompose and parameterize quadric
            $T^m_w, S^m_w \leftarrow Discompose(Q^m_w)$
        end
    end
end
```

C. Map-point Projection Frequency Association Distance

During the implementation of EQI, we know that all objects in the map have obtained associated map points. Hence we project them onto the current frame and then calculate the number of map points for each object contained in each bounding box, and count the total number of map points associated with each bounding box. We define Num_{ij} as the number of map points of the j-th object projected onto the i-th bounding box. Finally, we compute the associated distance by calculating the frequency of map point observations:

$$a^e_{ij} = 1 - \frac{Num_{ij}}{\sum\limits_{n=1}^{N} Num_{in}} \tag{11}$$

D. Nonparametric Probability Association Distance

Referring to [17], we conclude that the posterior probability of the association between the object detection in the current frame and the landmark in the map is proportional to the product of the prior probability of the Dirichlet Process (DP), the likelihood probability of the object category, denoted as OC, and the likelihood probability of the landmark position denoted as LP. We continue this idea and calculate the nonparametric probability association distance as:

$$a_{ij}^{r} = 1 - DP(O_w^j) \cdot OC(D_k^i) \cdot LP(D_k^i, O_w^j) \tag{12}$$

V. Multi-Constraint Optimization Factor Graph

We propose a multi-constraint optimization factor graph to jointly optimize camera poses, quadrics, planes and points. The overall optimization function can be expressed as:

$$\begin{aligned}
&\{T_{k,w}, Q_w^*, \pi_w, X_w\}^{opt} = \\
&\mathop{\arg\min}_{\{T_{k,w},Q_w^*\}} \sum\sum f_{quadric}(T_{k,w}, Q_w^*) + \\
&\mathop{\arg\min}_{\{Q_w^*,\pi_w\}} \sum \left(f_{tangent}(\pi_w, Q_w^*) + f_{vert}(\pi_w, Q_w^*)\right) + \\
&\mathop{\arg\min}_{\{Q_w^*\}} \sum \left(f_{scale}(Q_w^*) + f_{orientation}(Q_w^*)\right) + \\
&\mathop{\arg\min}_{\{T_{k,w},X_w\}} \sum\sum f_{point}(T_{k,w}, X_w) + \\
&\mathop{\arg\min}_{\{T_{k,w},\pi_w\}} \sum\sum f_{plane}(T_{k,w}, \pi_w)
\end{aligned} \tag{13}$$

Equation (13) includes the following components:

- Pose-Quadric projection factor:

$$f_{quadric}(T_{k,w}, Q_w^*) = \left\| D_k^{Q_w^*} - \eta(T_{k,w}, Q_w^*) \right\|_{\Sigma_q}^2 \tag{14}$$

  where $D_k^{Q_w^*}$ represents the object detection associated with Q_w^* and $\eta(\cdot)$ calculates the projected bounding box of the quadric, discussed in [3].

- Plane-Quadric tangent factor:

$$f_{tangent}(\pi_w, Q_w^*) = \left\| \pi_w^T Q_w^* \pi_w \right\|_{\Sigma_t}^2 \tag{15}$$

  We use the distance from the center of the quadric to the plane in order to associate them.

- Plane-Quadric vertical factor:

$$f_{vert}(\pi_w, Q_w^*) = \left\| n(\pi_w)^T \phi(Q_w^*) \right\|_{\Sigma_v}^2 \tag{16}$$

  where $n(\cdot)$ calculates the normal vector of the plane, and $\phi(\cdot)$ calculates the principal axis direction vector of the quadric, discussed in [18].

- Quadric scale prior factor:

$$f_{scale}(Q_w^*) = \left\| Scale(Q_w^*) - SP(lable_{Q_w^*}) \right\|_{\Sigma_s}^2 \tag{17}$$

  where $Scale(\cdot)$ calculates the scale vector of the quadric and $SP(\cdot)$ represents the prior scale vector, which is set to prevent the optimization from falling into a local optimal solution rather than optimizing the scale to a fixed value.

- Quadric orientation prior factor:

$$f_{orientation}(Q_w^*) = \left\| 1 - |n(\pi_g)^T \phi(Q_w^*)| \right\|_{\Sigma_o}^2 \tag{18}$$

  where π_g represents the prior ground normal vector, which is used for the objects that do not have associated planes.

- Pose-Point projection factor:

$$f_{point}(T_{k,w}, X_w) = \left\| P_{k,w} X_w - u_k^{obs} \right\|_{\Sigma_{po}}^2 \tag{19}$$

- Pose-Plane projection factor:

$$f_{plane}(T_{k,w}, \pi_w) = \left\| T_{k,w}^{-T} \pi_w - \pi_k^{obs} \right\|_{\Sigma_{pl}}^2 \tag{20}$$

The designed factor graph can be solved by the existing nonlinear optimization library `g2o`.

VI. Experiments and Evaluation

We evaluate the performance of our proposed system using the public TUM RGB-D [19] dataset with respect to the quality of the quadric, the performance of data association, and the accuracy of camera localization. The confidence of the YOLOv5 detector is set to 0.5. All experiments are conducted using Intel(R) Core(TM) i7-9750H CPU@2.6GHz, 16G memory, and Nvidia GTX 2060 Super.

A. Quantitative Evaluation Criteria

We evaluate the quality of the quadrics from the following five aspects:

- **Number of Constructed Objects (NoCO)**: We record the number of objects that are successfully detected and constructed as quadrics.
- **Number of Not-constructed Objects (NoNO)**: We record the number of objects that are successfully detected but not constructed as quadrics.
- **Number of Frames (NoF)**: We record the number of frames that are used for quadric initialization for every object, and then compute the average number of frames for all objects.
- **2D IoU**: Due to the lack of ground truth for the objects in the TUM dataset, we propose to use 2D IoU to evaluate the accuracy of the quadric parameters of the object. We record the associated 2D object detection results for every object in the stages of JDA and then compute the average 2D IoU between the 2D object detection result and the projection of the quadric, and then compute the average 2D IoU for all objects.
- **Initial Success Rate (ISR)**: In the EQI, we set the thresholds $\alpha = 5$, $\beta = 0.7$, $\gamma = 0.5$. If the evaluation score of the quadric is greater than γ, we treat it as a successful initialization. As a result, we record the total initialization count and successful initialization count for every object and then compute the average initial success rate for all objects.

B. The Quality of Quadric

Fig.4 shows the qualitative performance of our proposed EQI method from one single frame, where the object detection results of the image and associated quadrics in the map have the same color. We can see that our method has the ability to use a single frame data to construct most or even all objects that have been detected in different scenarios. For

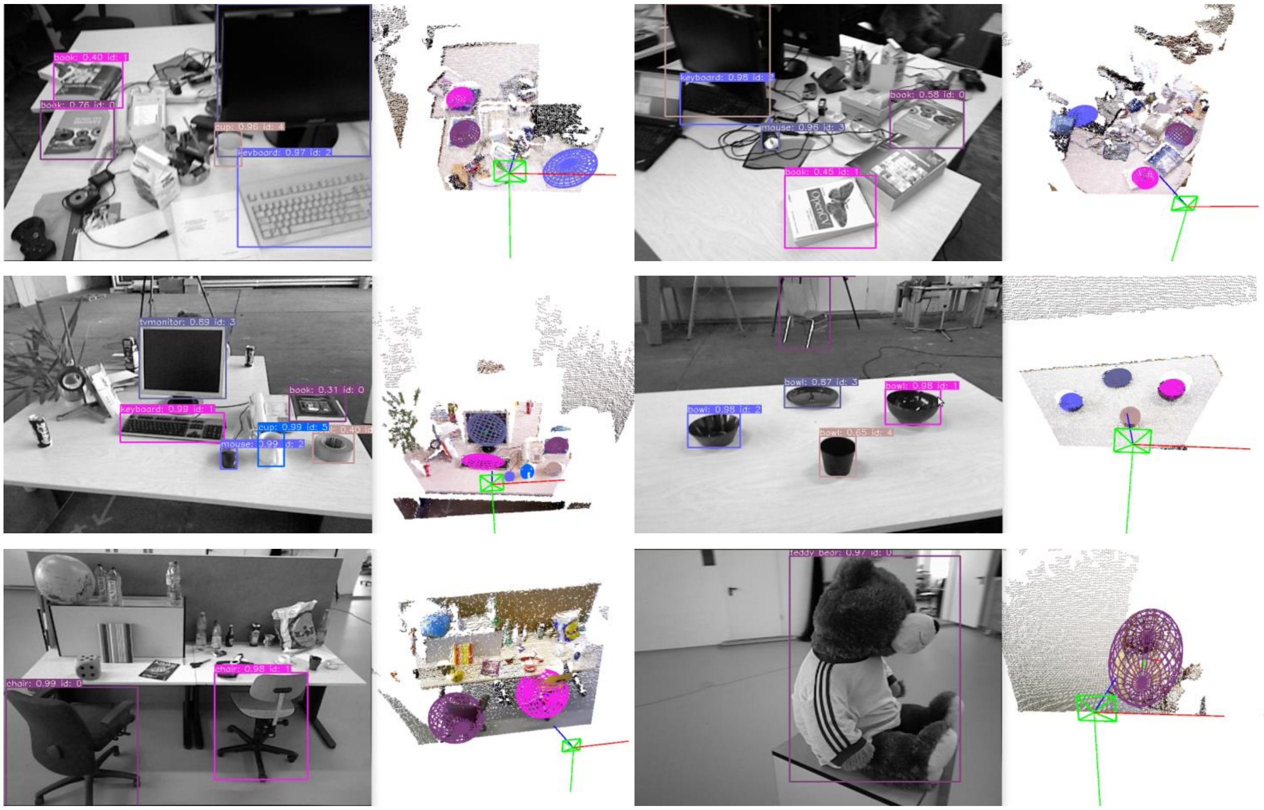


Fig. 4. The object detections of one single frame and the corresponding quadrics constructed by EQI on different sequences of TUM dataset. The left part of every scene presents the results of object detection, and the right part is the semi-dense point cloud map and the successfully initialized quadrics, which have the same color with the corresponding object detection results.

fair quantitative comparison, we designed two experiments: one where we disable the various constraint factors that are discussed in quadric optimization and only retain the pose-quadric projection factor keeping the approach consistent with Q-SLAM [3], and a second where we enable all constraint factors. The results are shown in Table I and Table II where TIN and TON respectively represent total image number and total object number of the sequence.

In Table I, we can see that the performance of our proposed method is better than Q-SLAM on most sequences, especially on NoF and ISR, which have an average improvement of 8.09 and 46% respectively. The comparison on both NoCO and NoNO indicates that our method can construct more effective objects. For 2D IoU, our method has an 8.21% improvement on average, which benefits from the fact that our proposed EQI algorithm can initialize quadrics for objects easily, hence we can have more observations to optimize them. However, Q-SLAM is better than our proposed method on the sequences fr1-xyz and fr3-teddy, because in these two sequences we have constructed more objects than Q-SLAM, but some of them cannot be further optimized due to the lack of observations, leading to lower average 2D IoU.

It can be seen from Table II, after enabling the whole constraint factors, denoted as Ours*, the average 2D IoU has a 3.78% improvement compared with Ours, and 12.59% improvement compared with Q-SLAM, which indicates that our proposed factor graph is effective for quadric optimization. However, the 2D IoU is slightly decreased in the sequences fr2-desk and fr3-desk. This is because the two sequences have a common feature: a large number of objects form a supporting relationship with planes, which will cause the plane-related constraint factors to account for a relatively large weight in the process of quadric optimization, and the projection constraint factor that directly promotes 2D IoU accounts for a small proportion of the error.

C. The Performance of JDA

We choose the fr2-desk sequence that has the most types of objects to verify our proposed JDA method. We define four kinds of data association (DA) methods by changing the weights of the four association distances proposed in JDA, including the 2D image plane DA ($k^q = 1.0$), the 3D map projection DA ($k^w = 0.5$, $k^e = 0.5$), the nonparametric PDA ($k^r = 1.0$) and our proposed JDA. All other weights are set to zero. We set 20 sets of different weights for object mapping, and then calculate the completeness and accuracy of the constructed map, and finally obtain the PR curve shown in Fig. 5, where we can see that the weights $k^q = 0.2$, $k^w = 0.2$, $k^e = 0.2$, and $k^r = 0.4$ are closest to the balance point. As a result, we use them as the weights of our JDA. Fig. 6 shows the object mapping performance comparison of four kinds of data association methods.

We can see that the 2D image plane DA has the worst performance. There is a large number of repeated quadrics for the same object, which is due to the omission of object

TABLE I

THE COMPARISON OF QUALITY OF QUADRICS USING TUM RGB-D SEQUENCES.

SEQUENCE	TIN	TON	NoCO		NoNO		NoF		2D IoU		ISR	
			Q-SLAM	Ours	Q-SLAM	Ours	Q-SLAM	Ours	Q-SLAM	Ours	Q-SLAM	Ours
fr1-desk	573	16	6.00	**10.00**	10.00	**6.00**	9.00	**2.40**	0.74	**0.80**	0.27	**0.75**
fr1-desk2	620	17	6.00	**11.00**	11.00	**6.00**	9.33	**1.82**	0.75	**0.81**	0.36	**0.71**
fr1-xyz	1352	9	2.00	**5.00**	7.00	**4.00**	21.00	**6.20**	**0.86**	0.83	0.06	**0.64**
fr2-room	792	22	5.00	**16.00**	17.00	**6.00**	4.00	**1.69**	0.45	**0.66**	0.63	**0.77**
fr2-desk	2893	18	11.00	**15.00**	7.00	**3.00**	6.55	**1.67**	0.79	**0.81**	0.45	**0.92**
fr2-dishes	2941	10	5.00	**7.00**	5.00	**3.00**	25.60	**6.00**	0.80	**0.84**	0.05	**0.63**
fr3-desk	2488	22	17.00	**19.00**	5.00	**3.00**	6.12	**3.16**	0.74	**0.77**	0.47	**0.67**
fr3-teddy	2327	5	1.00	3.00	4.00	2.00	7.00	**1.00**	**0.77**	0.73	0.13	**1.00**
mean	1748	14	6.63	**10.75**	8.25	**4.13**	11.08	**2.99**	0.74	**0.78**	0.30	**0.76**

TABLE II

THE COMPARISON OF 2D IOU USING TUM RGB-D SEQUENCES.

SEQUENCE	TIN	TON	2D IoU			Improvement(%)		
			Q-SLAM	Ours	Ours*	Ours/Q-SLAM	Ours*/Q-SLAM	Ours*/ours
fr1-desk	573	16	0.74	0.80	**0.83**	8.11	12.16	3.75
fr1-desk2	620	17	0.75	0.81	**0.85**	8.00	13.33	4.94
fr1-xyz	1352	22	**0.86**	0.83	0.84	-3.49	-2.33	1.21
fr-room	792	9	0.45	0.66	**0.73**	46.67	62.22	10.61
fr2-desk	2893	18	0.79	**0.81**	0.80	2.53	1.27	-1.23
fr2-dishes	2941	10	0.80	0.84	**0.85**	5.00	6.25	1.19
fr3-desk	2488	22	0.74	**0.77**	0.75	4.05	1.35	-2.60
fr3-teddy	2327	1	0.77	0.73	**0.82**	-5.19	6.49	12.33
mean	1748	14	0.74	0.78	**0.81**	8.21	12.59	3.78

TABLE III

THE COMPARISON OF ACCURACY OF LOCALIZATION USING TUM RGB-D SEQUENCES

SEQUENCE	ORB-SLAM2 [11]	Zhang et.al [20]	Q-SLAM [3]	Hosseinzadeh et.al [4]	Ours
fr1-desk	0.0146	0.0178	0.0632	0.0112	**0.0109**
fr1-desk2	0.0247	0.0265	0.0662	–	**0.0227**
fr1-xyz	0.0097	0.0099	–	0.0096	**0.0095**
fr2-desk	0.0083	0.0276	0.0568	0.0066	**0.0062**
fr3-desk	0.0109	0.0139	0.0765	**0.0087**	0.0088

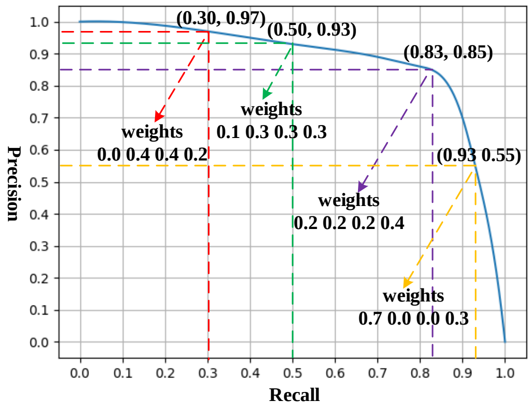
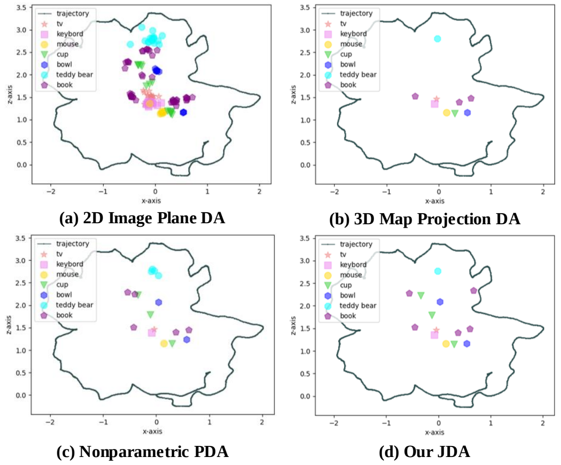


Fig. 5. PR curve of the mapping performance with different association distance weights. The weights of our JDA achieve a balance of completeness and accuracy for the object mapping.



(a) 2D Image Plane DA

(b) 3D Map Projection DA

(c) Nonparametric PDA

(d) Our JDA

Fig. 6. Object mapping performance comparison of 2D Image Plane DA, 3D Map Projection DA, Nonparametric PDA and our JDA. We respectively integrate these four data association methods into our system, and project the constructed map onto the XOZ plane to get the four 2D results.

detection results in false data associations. The performance of the 3D map projection DA on the object mapping is due to the fact that many objects have not been constructed, as the objects need to accumulate a certain amount of observation data before they are constructed. However, the 3D map projection DA is easily affected by occlusion and overlap of objects, which interrupts the accumulation of observation data or results in incorrect DA results.

In contrast, the nonparametric PDA proposed by [17] has achieved much better performance. Although there are still some cases where the same object is repeatedly constructed, this is rare, and the objects in the scene are constructed fairly completely. However, our JDA has achieved the best performance. In the case of having a complete construction for the objects in the scene, there is no obvious case of repeated quadrics for the same object.

D. The Accuracy of Localization

Table III shows the accuracy comparison of camera localization between our method and the baseline methods, including [11], [20], [3], [9], where the average translation error (ATE) is used as an evaluation metric. Since [11] and [20] are open-source, we use their code in the local environment to obtain the data. However, the code for [3] and [9] is not open-source, hence we directly use the experimental data of camera localization in their papers. Among the data, − indicates that there is no data for this sequence in the paper.

We can see that the performance of [3] is the worst because it only uses quadric-level landmarks to solve and optimize the poses. The combination of the poor constant noise estimation of the non-Gaussian bounding box measurement and the object occlusion will significantly and negatively affect the estimated trajectory, which leads to the accuracy of localization being much worse than that of the point-based method.

Our method outperforms all comparison methods on the sequences fr1-desk, fr1-desk2, fr1-xyz, and fr2-desk, especially compared with the method proposed by [9], which also uses the landmarks of planes and quadrics for pose optimization. The reason why our method is better on most sequences is that the proposed EQI method allows us to construct more quadrics, which means we can build more object-level constraints for pose optimization than [9]. However, on the sequence fr3-desk, [9] performs better, as [9] introduces Manhattan World constraints between plane variables in the optimization, which are strong and effective constraints, especially when there are many plane structures that meet the Manhattan world assumption, such as in the fr3-desk sequence.

VII. Conclusion

In this work, an object-aware semantic SLAM system is presented. The EQI algorithm which based on object detection and surfel construction is proposed to reduce the difficulty of initializing quadrics from a small viewing angle. The robust object-level data association is solved by the JDA method. As for back-end optimization, a multi-constraint factor graph is proposed to jointly optimize the camera poses and constructed landmarks. Extensive experiments are conducted to show that the proposed system achieved competitive or even better performance in indoor environments when compared with other state-of-the-art methods. Further work will focus on using the constructed geometric semantic map for loop detection and re-localization.

References

[1] 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.

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

[3] 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.

[4] M. Hosseinzadeh, Y. Latif, T. Pham, N. Suenderhauf, and I. Reid, “Structure aware slam using quadrics and planes,” in *Asian Conference on Computer Vision*. Springer, 2018, pp. 410–426.

[5] 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.

[6] 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.

[7] Z. Liao, W. Wang, X. Qi, X. Zhang, L. Xue, J. Jiao, and R. Wei, “Object-oriented slam using quadrics and symmetry properties for indoor environments,” *arXiv preprint arXiv:2004.05303*, 2020.

[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. Chen, S. Song, J. Zhao, T. Feng, C. Ye, L. Xiong, and D. Li, “Robust dual quadric initialization for forward-translating camera movements,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4712–4719, 2021.

[10] H. Fan, H. Su, and L. J. Guibas, “A point set generation network for 3d object reconstruction from a single image,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[11] 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.

[12] R. Tian, Y. Zhang, D. Zhu, S. Liang, S. Coleman, 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.

[13] C. Feng, Y. Taguchi, and V. R. Kamat, “Fast plane extraction in organized point clouds using agglomerative hierarchical clustering,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 6218–6225.

[14] 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.

[15] D. R. Chand and S. S. Kapur, “An algorithm for convex polytopes,” *Journal of the ACM (JACM)*, vol. 17, no. 1, pp. 78–86, 1970.

[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] B. Mu, S.-Y. Liu, L. Paull, J. Leonard, and J. P. How, “Slam with objects using a nonparametric pose graph,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4602–4609.

[18] N. Jablonsky, M. Milford, and N. Sünderhauf, “An orientation factor for object-oriented slam,” *arXiv preprint arXiv:1809.06977*, 2018.

[19] 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*. IEEE, 2012, pp. 573–580.

[20] X. Zhang, W. Wang, X. Qi, Z. Liao, and R. Wei, “Point-plane slam using supposed planes for indoor environments,” *Sensors*, vol. 19, no. 17, p. 3795, 2019.