2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

EAO-SLAM: Monocular Semi-Dense Object SLAM Based on Ensemble Data Association

Yanmin Wu[1], Yunzhou Zhang[1,2], Delong Zhu[3], Yonghui Feng[2], Sonya Coleman[4] and Dermot Kerr[4]

***Abstract*— Object-level data association and pose estimation play a fundamental role in semantic SLAM, which remain unsolved due to the lack of robust and accurate algorithms. In this work, we propose an ensemble data associate strategy for integrating the parametric and nonparametric statistic tests. By exploiting the nature of different statistics, our method can effectively aggregate the information of different measurements, and thus significantly improve the robustness and accuracy of data association. We then present an accurate object pose estimation framework, in which an outliers-robust centroid and scale estimation algorithm and an object pose initialization algorithm are developed to help improve the optimality of pose estimation results. Furthermore, we build a SLAM system that can generate semi-dense or lightweight object-oriented maps with a monocular camera. Extensive experiments are conducted on three publicly available datasets and a real scenario. The results show that our approach significantly outperforms state-of-the-art techniques in accuracy and robustness. The source code is available on `https://github.com/yanmin-wu/EAO-SLAM`.**

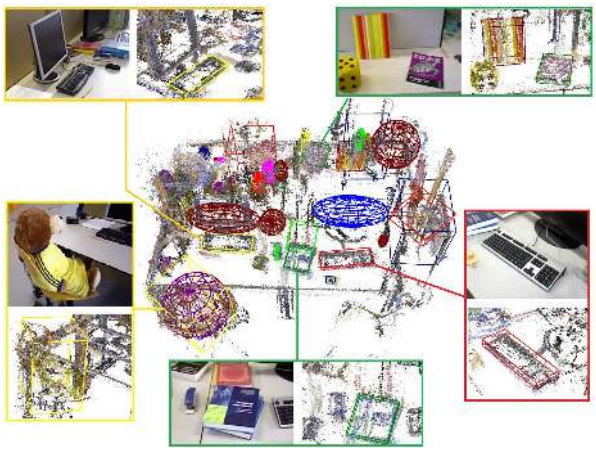

Fig. 1: A lightweight and object-oriented semantic map.

I. Introduction

Conventional visual SLAM systems have achieved significant success in robot localization and mapping tasks. More efforts in recent years are evolved in making SLAM serve for robot navigation, object manipulation, and environment representation. Semantic SLAM is a promising technique for enabling such applications and receives much attention from the community [1]. In addition to the conventional functions, semantic SLAM also focuses on a detailed expression of the environment, e.g., labeling map elements or objects of interests, to support different high-level applications.

Object SLAM is a typical application of semantic SLAM, and the goal is to estimate more robust and accurate camera poses by leveraging the semantic information of in-frame objects [2]–[4]. In this work, we further extend the content of object SLAM by enabling it to build lightweight and object-oriented maps, demonstrated in Fig. 1, in which the objects are represented by cubes or quadrics with their locations, orientations, and scales accurately registered.

The challenges of object SLAM mainly lie in two folds: 1) Existing data association methods [5]–[7] are not robust or accurate for tackling complex environments that contain multiple object instances. There are no practical solutions to systematically address this problem. 2) Object pose estimation is not accurate, especially for monocular object SLAM. Although some improvements are achieved in recent studies [8]–[10], they are typically dependent on strict assumptions, which are hard to fulfill in real-world applications.

In this paper, we propose the EAO-SLAM, a monocular object SLAM system, to address the data association and pose estimation problems. Firstly, we integrate the parametric and nonparametric statistic tests, and the traditional IoU-based method, to conduct model ensembling for data association. Compared with conventional methods, our approach sufficiently exploits the nature of different statistics, e.g., Gaussian and non-Gaussian measurements, hence exhibits significant advantages in association robustness. For object pose estimation, we propose a centroid and scale estimation algorithm and an object pose initialization approach based on the *isolation forest* (iForest). The proposed methods are robust to outliers and exhibit high accuracy, which significantly facilitates the joint pose optimization process.

The contributions of this paper are summarized as follows:

- We propose an ensemble data association strategy that can effectively aggregate different measurements of the objects to improve association accuracy.
- We propose an object pose estimation framework based

This work was supported by National Natural Science Foundation of China (No. 61973066,61471110) , Equipment Pre-research Fundation(61403120111), the Fundation of Key Laboratory of Aerospace System Simulation(6142002301), the Fundation of Key Laboratory of Equipment Reliability(61420030302), Natural Science Foundation of Liaoning (No.20180520040), and Fundamental Research Funds for the Central Universities(N172608005, N182608004).

[1]Yanmin Wu is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China.

[2]Yunzhou Zhang and Yonghui Feng are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (Corresponding author: Yunzhou Zhang, Email: `zhangyunzhou@mail.neu.edu.cn`).

[3]Delong Zhu is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China.

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

on iForest, which is robust to outliers and can accurately estimate the locations, poses, and scales of objects.
- Based on the proposed method, we implement the EAO-SALM to build lightweight and object-oriented maps.
- We conduct comprehensive experiments and verify the effectiveness of our proposed methods on publicly available datasets and the real scenario. The source code of this work is also released.

II. Related Work

A. Data Association

Data association is an indispensable ingredient for semantic SLAM, which is used to determine whether the object observed in the current frame is an existing object in the map. Bowman *et al.* [5] use a probabilistic method to model the data association process and leverage the EM algorithm to find correspondences between observed landmarks. Subsequent studies [7], [11] further extend the idea to associate dynamic objects or conduct semantic dense reconstruction. These methods can achieve high association accuracy, but can only process a limited number of object instances. Their efficiency also remains to be improved due to the expensive EM optimization process [12]. Object tracking is another commonly-used approach in data association. Li *et al.* [13] propose to project 3D cubes to the image plane and then leverage the Hungarian tracking algorithm to conduct association using the projected 2D bounding boxes. Tracking-based methods perform high runtime efficiency, but can easily generate incorrect priors in complex environments, yielding incorrect association results.

In recent studies, more data association approaches are developed based on maximum shared information. Liu *et al.* [14] propose random walk descriptors to represent the topological relationships between objects, and those with the maximum number of shared descriptors are regarded as the same instance. Instead, Yang *et al.* [8] propose to directly count the number of matched map points on the detected objects as association criteria, yielding a much efficient performance. Grinvald *et al.* [2] propose to measure the similarity between semantic labels and Ok *et al.* [3] propose to leverage the correlation of hue saturation histogram. The major drawback of these methods is that the designed features or descriptors are typically not general or robust enough and can easily cause incorrect associations.

Weng *et al.* [15] for the first time propose nonparametric statistical testing for semantic data association, which can address the problems in which the statistics do not follow a Gaussian distribution. Later on, Iqbal *et al.* [6] also verify the effectiveness of nonparametric data association. However, this method cannot address the statistics that follow Gaussian distributions effectively, hence cannot sufficiently exploit different measurements in SLAM. Based on this observation, we combine the parametric and nonparametric methods to perform model ensembling, which exhibits superior association performance in the complex scenarios with the presence of multiple categories of objects.

B. Object SLAM

Benefiting from deep learning techniques [16], [17], object detection is robustly integrated into the SLAM framework for labeling objects of interests in the map. The exploitation of in-frame objects significantly enlarges the application scopes of traditional SLAM. Some studies [15], [18], [19] treat objects as landmarks to estimate camera poses or for relocalization [13]. Some studies [20] leverage object size to constrain the scale of monocular SLAM, or remove dynamic objects to improve pose estimation accuracy [7], [21]. In recent years, the combination of object SLAM and grasping [22] has also attracted many interests, and facilitate the research on autonomous mobile manipulation.

Object models in semantic SLAM can be broadly divided into three categories: instance-level models, category-specific models, and general models. The instance-level models [9], [23] depend on a well-established database that records all the related objects. The prior information of objects provides important object-camera constraints for graph optimization. Since the models need to be known in advance, the application scenarios of such methods are limited. There are also some studies on category-specific models, which focus on describing category-level features. For example, Parkhiya *et al.* [10] and Joshi *et al.* [19] use the CNN network to estimate the viewpoint of objects and then project the 3D line segments onto image planes to align them. The general model adopts simple geometric elements, e.g., cubes [8], [13], quadrics [18] and cylinders [10], to represent objects, which are also the most commonly-used models.

In terms of the joint optimization of camera and object poses, Frost *et al.* [20] simply integrate object centroids as point clouds to the camera pose estimation process. Yang *et al.* [8] propose a joint camera-object-point optimization scheme to construct the pose and scale constraints for graph optimization. Nicholson *et al.* [18] propose to project the quadric onto the image plane and then calculates the scale error between the projected 2D rectangular and the detected bounding box. This work also adopts the joint optimization strategy, but with a novel initialization method, which can significantly improve the optimality of solutions.

III. System Overview

The proposed object SLAM framework is demonstrated in Fig. 2, which is developed based on ORB-SLAM2 [24], and additionally integrates a semantic thread that adopts YOLOv3 as the object detector. The ensemble data association is implemented in the tracking thread, which combines the information of bounding boxes, semantic labels, and point clouds. After that, the iForest is leveraged to eliminate outliers for finding an accurate initialization for the joint optimization process. The object pose and scale are then optimized together with the camera pose to build a lightweight and object-oriented map. In semi-dense mapping thread, the object map is combined with a semi-dense map generated by [25] to obtain the a semi-dense semantic map.

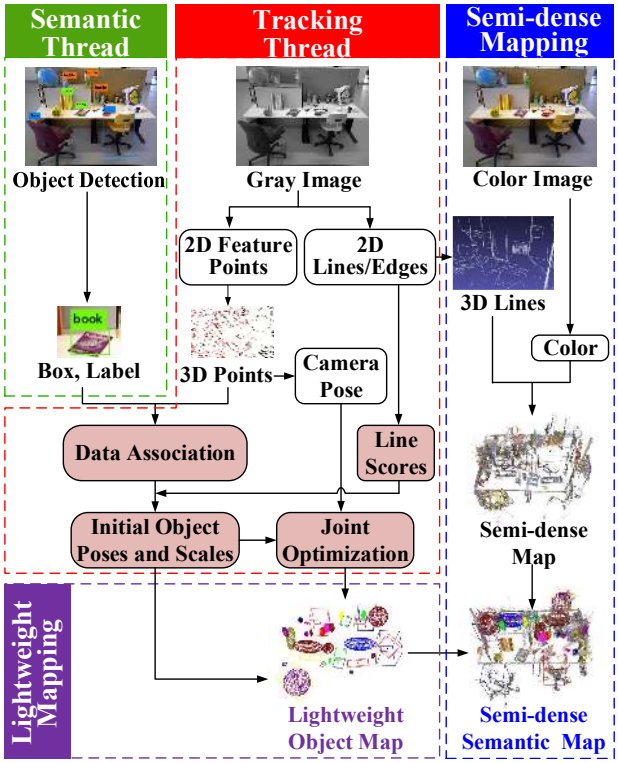

Fig. 2: The architecture of EAO-SLAM system. The main contributions of this work are highlighted with red colors.

IV. Ensemble Data Association

Throughout this section, the following notations are used:

- $P \in \mathbb{R}^{3\times|P|}, Q \in \mathbb{R}^{3\times|Q|}$ - the point clouds of objects.
- $\mathcal{R}$ - the rank (position) of a data point in a sorted list.
- $\mathbf{c} \in \mathbb{R}^{3\times1}$ - the currently observed object centroid.
- $C = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{|C|}] \in \mathbb{R}^{3\times|C|}$ - the history observations of the centroids of an object. C_1, C_2 are similar.
- $f(\cdot)$ - the probability function used for statistic test.
- $m(\cdot), \sigma(\cdot) \in \mathbb{R}^{3\times1}$ - the mean and variance functions.

A. Nonparametric Test

The nonparametric test is leveraged to process object point clouds (the red and green points in Fig. 3 (a)), which follows a non-Gaussian distribution according to our experimental studies (Section VI-A). Theoretically, if P and Q belong to the same object, they should follow the same distribution, i.e., $f_P = f_Q$. We use the *Wilcoxon Rank-Sum test* [26] to verify whether the null hypothesis holds.

We first concatenate the two point clouds $X = [P|Q] = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{|X|}] \in \mathbb{R}^{3\times(|P|+|Q|)}$, and then sort X in three dimensions respectively. Define $W_P \in \mathbb{R}^{3\times1}$ as follows,

$$W_P = \left\{ \sum_{k=1}^{|X|} \mathcal{R}(1\{\mathbf{x}_k \in P\}) - \frac{|P|(|P|+1)}{2} \right\}, \tag{1}$$

and W_Q is with the same formula. The Mann-Whitney statistics is $W = \min(W_P, W_Q)$, which is proved to follow a Gaussian distribution asymptotically [26]. Herein, we essentially construct a Gaussian statistics using the non-Gaussian point clouds. The mean and variance of W is calculated as follows:

$$m(W) = (|P||Q|)/2, \tag{2}$$

$$\sigma(W) = \frac{|P||Q|\Delta}{12} - \frac{|P||Q|(\sum_i \tau_i^3 - \sum_i \tau_i)}{12(|P|+|Q|)\Delta}, \tag{3}$$

where $\Delta = |P| + |Q| + 1$, and $\tau \in P \cap Q$.

To make the null hypothesis stand, W should meet the following constraints:

$$f(W) \geq f(r_r) = f(r_l) = \alpha/2, \tag{4}$$

where α is the significance level, $1-\alpha$ is thus the confidence level, and $[r_l, r_r] \approx [m - s\sqrt{\sigma}, m + s\sqrt{\sigma}]$ defines the confidence region. The scalar $s > 0$ is defined on a normalized Gaussian distribution $\mathcal{N}(s|0,1){=}\alpha$. In summary, if the Mann-Whitney statistics W of two point clouds P and Q satisfies Eq. (4), they come from the same object and the data association successes.

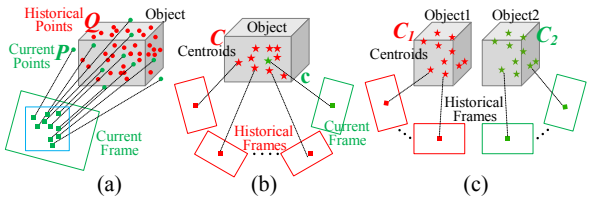

Fig. 3: Different Types of Statistics Used for data association.

B. Single-sample and Double-sample T-test

The single-sample t-test is used to process object centroids observed in different frames (the stars in Fig. 3 (b)), which typically follow a Gaussian distribution (Section VI-A).

Suppose the null hypothesis is that C and $\mathbf{c}$ are from the same object, and define t statistics as follows,

$$t = \frac{m(C) - \boldsymbol{c}}{\sigma(C)/\sqrt{|C|}} \sim t(|C|-1). \tag{5}$$

To make the null hypothesis stand, t should satisfy:

$$f(t) \geq f(t_{\alpha/2,v}) = \alpha/2 \tag{6}$$

where $t_{\alpha/2,v}$ is the upper $\alpha/2$ quantile of the t-distribution of v degrees of freedom, and $v = \sqrt{|C|} - 1$. If t statistics satisfies (6), $\mathbf{c}$ and C comes from the same object.

Due to the strict data association strategy above or the bad angle of views, some existing objects may be recognized as new ones. Hence, a double-sample t-test is leveraged to determine whether to merge the two objects by testing their historical centroids (the stars in Fig. 3 (c)).

Construct t-statistics for C_1 and C_2 as follows,

$$t = \frac{m(C_1) - m(C_2)}{\sigma_d} \sim t(|C_1| + |C_2| - 2) \tag{7}$$

$$\sigma_d = \sqrt{\frac{(|C_1|-1)\,\sigma_1^2 + (|C_2|-1)\,\sigma_2^2}{|C_1| + |C_2| - 2} \left(\frac{1}{|C_1|} + \frac{1}{|C_2|}\right)} \tag{8}$$

where σ_d is the pooled standard deviation of the two objects. Similarly, if t satisfies (6), $v = |C_1| + |C_2| - 2$, it means that C_1 and C_2 belongs to the same object, then we merge them.

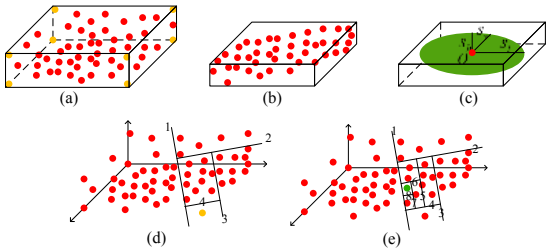

Fig. 4: Object representation and demonstration of iForest.

V. Object SLAM

Throughout this section, the following notations are used:

- $\boldsymbol{t} = [t_x, t_y, t_z]^T$ - the translation (location) of object frame in world frame.
- $\boldsymbol{\theta} = [\theta_r, \theta_y, \theta_p]^T$ - the rotation of object frame w.r.t. world frame. $R(\boldsymbol{\theta})$ is matrix representation.
- $T = \{R(\boldsymbol{\theta}), \boldsymbol{t}\}$ - the transformation of object frame w.r.t. world frame.
- $\boldsymbol{s} = [s_l, s_w, s_h]^T$ - half of the side length of a 3D bounding box, i.e., the scale of an object.
- $P_o, P_w \in \mathbb{R}^{3\times 8}$ - the coordinates of eight vertices of a cube in object and world frame, respectively.
- $Q_o, Q_w \in \mathbb{R}^{4\times 4}$ - the quadric parameterized by its semiaxis in object and world frame, respectively, where $Q_o = \text{diag}\left\{s_l^2, s_w^2, s_h^2, -1\right\}$.
- $\alpha(\cdot)$ - calculate the angle of line segments.
- K, T_c - the intrinsic and extrinsic parameters of camera.
- $\boldsymbol{p} \in \mathbb{R}^{3\times 1}$ - the coordinates of a point in world frame.

Object Representation: In this work, we leverage the cubes and quadrics to represent objects, rather than the complex instance-level or category-level model. For objects with regular shapes, such as books, keyboards, and chairs, we use cubes (encoded by its vertices P_o) to represent them. For non-regular objects without an explicit direction, such as balls, bottles, and cups, the quadric (encoded by its semiaxis Q_o) is used for representation. Here, P_o and Q_o are expressed in object frame and only depend on the scale $\boldsymbol{s}$. To register these elements to global map, we also need to estimate their translation $\boldsymbol{t}$ and orientation $\boldsymbol{\theta}$ w.r.t. global frame. The cubes and quadrics in global frame are expressed as follows:

$$P_w = R(\boldsymbol{\theta})P_o + \boldsymbol{t}, \tag{9}$$

$$Q_w = TQ_oT^T. \tag{10}$$

With the assumption that the objects are placed parallel with the ground, i.e., θ_r=θ_p=0, we only need to estimate $[\theta_y, \boldsymbol{t}, \boldsymbol{s}]$ for a cube and $[\boldsymbol{t}, \boldsymbol{s}]$ for a quadric.

Estimate $\boldsymbol{t}$ and $\boldsymbol{s}$: Suppose there is an object point cloud X in global frame, we follow conventions and denote its mean by $\boldsymbol{t}$, based on which, the scale can be calculated by $\boldsymbol{s} = (\max(X) - \min(X))/2$. The main challenge here is that X is typically with many outliers, which can introduce a large bias to the estimation of $\boldsymbol{t}$ and $\boldsymbol{s}$. One of our major contributions in this paper is the development of an outlier-robust centroid and scale estimation algorithm based on the iForest [27] to improve the estimation accuracy. The detailed procedure of our algorithm is presented in Alg. 1.

Algorithm 1 Centroid and Scale Estimation Based on iForest

Input: X - The point cloud of an object, t - The number of iTrees in iForest, ψ - The subsampling size for an iTree.
Output: $\mathcal{F}$ - The iForest, a set of iTrees, $\boldsymbol{t}$ - The origin of local frame, $\boldsymbol{s}$ - The initial scale of the object.

1: **procedure** PARAOBJECT(X, t, ψ)
2: $\quad\mathcal{F} \leftarrow$ BUILDFOREST(X, t, ψ)
3: $\quad$**for** point $\mathbf{x}$ in X **do**
4: $\quad\quad E(h) \leftarrow$ averageDepth($\mathbf{x}, \mathcal{F}$)
5: $\quad\quad s \leftarrow$ score($E(h), C$) ▷ Eq. (11) and (12)
6: $\quad\quad$**if** $s > 0.6$ **then** ▷ an empirical value
7: $\quad\quad\quad$remove($\mathbf{x}$) ▷ remove $\mathbf{x}$ from X
8: $\quad\quad$**end if**
9: $\quad$**end for**
10: $\quad\boldsymbol{t} \leftarrow$ meanValue(X)
11: $\quad\boldsymbol{s} \leftarrow$ (max(X) - min(X)) / 2
12: $\quad$**return** $\mathcal{F}, \boldsymbol{t}, \boldsymbol{s}$
13: **end procedure**
14: **procedure** BUILDFOREST(X, t, ψ)
15: $\quad\mathcal{F} \leftarrow \phi$
16: $\quad l \leftarrow$ ceiling($\log_2 \psi$) ▷ maximum times of iterations
17: $\quad$**for** $i = 1$ to t **do**
18: $\quad\quad X^{(i)} \leftarrow$ randomSample(X, ψ)
19: $\quad\quad\mathcal{F} \leftarrow \mathcal{F} \cup$ BUILDTREE($X^{(i)}, 0, l$)
20: $\quad$**end for**
21: **return** $\mathcal{F}$
22: **end procedure**
23: **procedure** BUILDTREE(X, e, l)
24: $\quad$**if** $e \geq l$ or $|X| \leq 1$ **then**
25: $\quad\quad$**return** exNode$\{|X|\}$ ▷ record the size of X
26: $\quad$**end if**
27: $\quad i \leftarrow$ randomDim(1, 3) ▷ get one dimension
28: $\quad q \leftarrow$ randomSpitPoint($X[i]$)
29: $\quad X_l, X_r \leftarrow$ split($X[i], q$)
30: $\quad L \leftarrow$BUILDTREE($X_l, e+1, l$) ▷ get child pointer
31: $\quad R \leftarrow$BUILDTREE($X_r, e+1, l$)
32: $\quad$**return** inNode$\{L, R, i, q\}$
33: **end procedure**

The key idea of the algorithm is to recursively separate the data space into a series of isolated data points, and then take the easily isolated ones as outliers. The philosophy is that, normal points is typically located more closely and thus need more steps to isolate, while the outliers usually scatter sparsely and can be easily isolated with less steps. As indicated by the algorithm, we first create t isolated trees (the iForest) using the point cloud of an object (lines 2 and 14-33), and then identify the outliers by counting the path length of each point $\mathbf{x} \in X$ (lines 3-9), in which the score function is defined as follows:

$$s(\mathbf{x}) = 2\exp\frac{-E(h(\mathbf{x}))}{C}, \tag{11}$$

$$C = 2H(|X|-1) - \frac{2(|X|-1)}{|X|}, \tag{12}$$

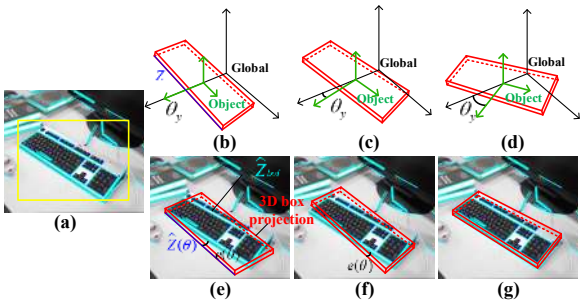

Fig. 5: Line alignment to estimate object direction.

where C is a normalization parameter, H is a weight coefficient, and $h(\mathbf{x})$ is the height of point $\mathbf{x}$ in isolated tree. As demonstrated in Fig. 4(d)-(e), the yellow point is isolated after four steps, thus its path length is 4, and the green point has a path length of 8. Therefore, the yellow point is more likely to be an outlier. In our implementation, points with a score greater than 0.6 are removed, and the remainings are used to calculate $\boldsymbol{t}$ and $\boldsymbol{s}$ (lines 10-12). Based on $\boldsymbol{s}$, we can initially construct the cubics and quadratics in the object frame, as shown in Fig. 4(a)-(c). $\boldsymbol{s}$ will be further optimized along with the object and camera poses later on.

Estimate θ_y: The estimation of θ_y is divided into two steps, namely to find a good initial value for θ_y first and then conduct numerical optimization based on the initial value. Since pose estimation is a non-linear process, a good initialization is very important to help improve the optimality of the estimation result. Conventional methods [13] usually neglect the initialization process, which typically yields inaccurate results.

The details of pose initialization algorithm is presented in Alg. 2. The inputs are obtained as follows: 1) LSD segments are extracted from t consecutive image and those falling in the bounding boxes are assigned to the corresponding objects (see Fig. 5a); 2) The initial pose of an object is assumed to be consistent with the global frame, i.e., θ_0=0 (see Fig. 5b). In the algorithm, we first uniformly sample thirty angles within $[-\pi/2, \pi/2]$ (line 2). For each sample, we then evaluate its score by calculating the accumulated angle errors between LSD segments Z_{lsd} and the projected 2D edges of 3D edges Z of the cube (lines 3-12). The error is defined as follows:

$$\begin{aligned} e(\boldsymbol{\theta}) &= ||\alpha(\hat{Z}(\boldsymbol{\theta})) - \alpha(\hat{Z}_{lsd})||^2, \\ \hat{Z}(\boldsymbol{\theta}) &= KT_c\left(R(\boldsymbol{\theta})Z + \boldsymbol{t}\right). \end{aligned} \tag{13}$$

A demonstration of the calculation of $e(\boldsymbol{\theta})$ is visualized in Fig. 5 (e)-(g). The score function is defined as follows:

$$\text{Score} = \frac{N_\text{p}}{N_\text{a}}(1 + 0.1(\xi - E(e))), \tag{14}$$

where N_a is the total number of line segments of the object in the current frame, N_p is the number of line segments that satisfy $e < \xi$, ξ is a manually defined error threshold (five degrees here), and $E(e)$ is the average error of these line segments with $e < \xi$. After evaluating all the samples, we choose the one that achieves the highest score as the initial yaw angle for the following optimization process (line 13).

Algorithm 2 Initialization for Object Pose Estimation

Input: $Z_1, Z_2, \ldots, Z_t$ - Line segments detected by LSD in t consecutive images, θ_0 - The initial guess of yaw angel.
Output: θ - The estimation result of yaw angel, e - The estimation errors.

1: $\mathcal{S}, \mathcal{E} \leftarrow \phi$
2: $\boldsymbol{\Theta} \leftarrow$ sampleAngles(θ_0, 30) ▷ see Fig. 5 (b)-(d)
3: **for** sample θ in $\boldsymbol{\Theta}$ **do**
4: $\quad s_\theta, e_\theta \leftarrow 0$
5: $\quad$ **for** Z in $\{Z_1, Z_2, \ldots, Z_t\}$ **do**
6: $\quad\quad s, e \leftarrow$ score(θ, Z) ▷ Eq. (13) and (14)
7: $\quad\quad s_\theta \leftarrow s_\theta + s$
8: $\quad\quad e_\theta \leftarrow e_\theta + e$
9: $\quad$ **end for**
10: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{s_\theta\}$
11: $\quad \mathcal{E} \leftarrow \mathcal{E} \cup \{e_\theta\}$
12: **end for**
13: $\theta^* \leftarrow$ argmax($\mathcal{S}$)
14: **return** θ^*, e_{θ^*}

Joint Optimization: After obtaining the initial $\boldsymbol{S}$ and θ_y, we then jointly optimize object and camera poses:

$$\{O, T_c\}^* = \underset{\{\theta_y, \boldsymbol{s}\}}{\text{argmin}} \sum (e(\boldsymbol{\theta}) + e(\boldsymbol{s})) + \underset{\{T_c\}}{\text{argmin}} \sum e(\boldsymbol{p}), \tag{15}$$

where the first term is the object pose error defined in Eq. (13) and the scale error $e(\boldsymbol{s})$ defined as the distance between the projected edges of a cube and their nearest parallel LSD segments. The second term $e(\boldsymbol{p})$ is the commonly-sued reprojection error in traditional SLAM framework.

VI. Experimental Results

A. Distributions of Different Statistics

For data association, the adopted statistics for statistical testing include the point clouds and their centroids of an object. To verify our hypothesis about the distributions of different statistics, we analyze a large amount of data and visualize their distributions in Fig. 6.

Fig. 6 (a) shows the distributions of the point clouds of 13 objects during the data association in the fr3_long_office sequence. It is obvious that such statistics do not follow a Gaussian distribution. We can be seen that the distributions are related to specific characteristics of the objects, and do not show consistent behaviors. Fig. 6 (b) shows the error distribution of object centroids in different frames, which typically follow the Gaussian distribution. This result verifies the reasonability of applying the nonparametric *Wilcoxon Rank-Sum test* for point clouds and t-test for object centroids.

B. Ensemble Data Association Experiments

We compare our method with the commonly-used Intersection over Union (IoU) method, nonparametric test (NP), and t-test. Fig. 7 shows the association results of these methods in TUM fr3_long_office sequence. It can be seen that some objects are not correctly associated in (a)-(c). Due to the lack of association information, existing objects are

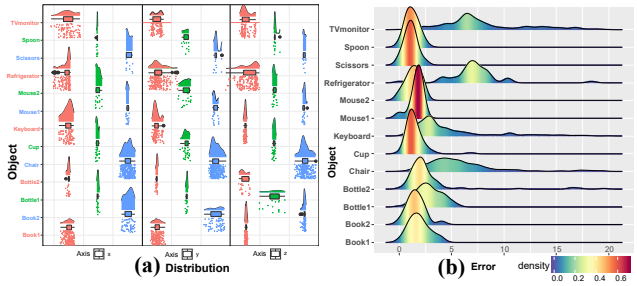

Fig. 6: Distributions of different statistics in data association. (a) position distribution of point clouds in three directions. (b) distance error distribution of centroids.

often misrecognized as new ones by these methods once the objects are occluded or disappear in some frames, resulting in many unassociated objects in the map. In contrast, our method is much more robust and can effectively address this problem (see Fig. 7(d)). The results of other sequences are shown in Table I, and we use the same evaluation metric as [6], which measures the number of objects that finally present in the map. The *GT* represents the ground-truth object number. As we can see, our method achieves a high success rate of association, and the number of objects in the map goes closer to GT, which significantly demonstrates the effectiveness of the proposed method.

We also compare our method with [6], and the results are shown in Table II. As is indicated, our method can significantly outperform [6]. Especially in the TUM dataset, the number of successfully associated objects by our method is almost twice than that by [6]. In Microsoft RGBD and Scenes V2, the advantage is not obvious since the number of objects is limited there. Reasons of the inaccurate association of [6] lie in two folds: 1) The method does not exploit different statistics and only used non-parametric statistics, thus resulting in many unassociated objects; 2) A clustering algorithm is leveraged to tackling the problem mentioned above, which removes most of the candidate objects.

TABLE I: DATA ASSOCIATION RESULTS

	IoU	IoU+NP	IoU+t-test	EAO	GT
Fr1_desk	62	47	41	**14**	16
Fr2_desk	83	64	52	**22**	25
Fr3_office	150	128	130	**42**	45
Fr3_teddy	32	17	21	**6**	7

C. Qualitative Assessment of Object Pose Estimation

We superimpose the cubes and quadrics of objects on semi-dense maps for qualitative evaluation. Fig. 8 is the 3D top view of a keyboard (Fig. 5(a)) where the cube characterizes its pose. Fig. 8(a) is the initial pose with large scale error; Fig. 8(b) is the result after using iForest; Fig. 8(c) is the final pose after our joint pose estimation. Fig. 9 presents the pose estimation results of the objects in 14 sequences of the three datasets, in which the objects are placed randomly and in different directions. As is shown, the proposed method achieves promising results with a monocular camera, which demonstrate the effectiveness of our pose

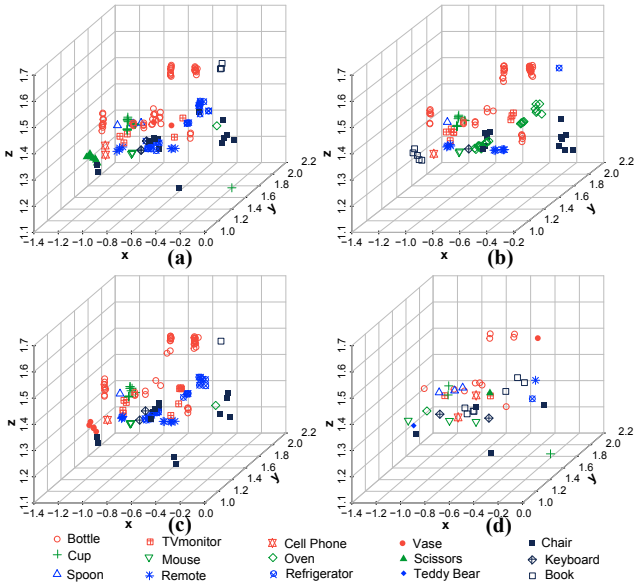

Fig. 7: Qualitative comparison of data association results. (a) IoU method. (b) IoU and nonparametric test. (c) IoU and t-test. (d) our ensemble method.

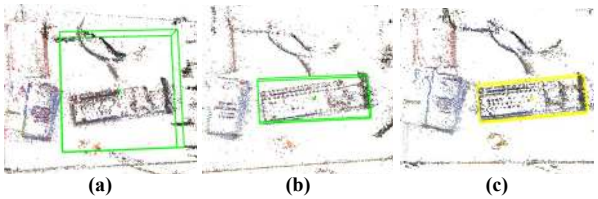

Fig. 8: Visualization of the pose estimation.

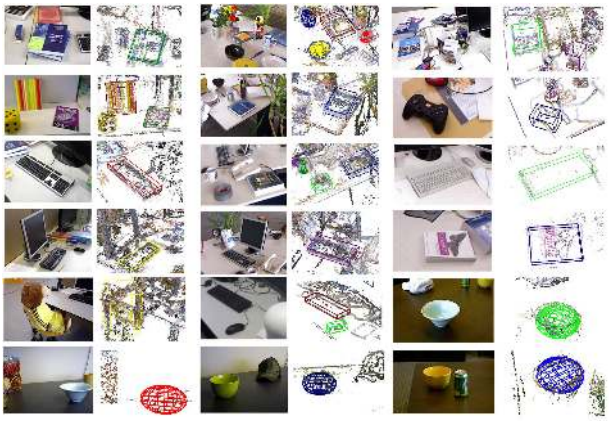

Fig. 9: Results of object pose estimation. Odd columns: original RGB images. Even column: estimated object poses.

estimation algorithm. Since the datasets are not specially designed for object pose estimation, there is no ground truth for quantitatively evaluate the methods. Here, we compare θ_y before initialization (BI), after initialization (AI), and after joint optimization (JO). As shown in Table III, the original direction of the object is parallel to the global frame, and there is a large angle error. After pose initialization, the error is decreased, and after the joint optimization, the error is further reduced, which verifies the effectiveness of our pose estimation algorithm.

D. Object-Oriented Map Building

Lastly, we build the object-oriented semantic maps based on the robust data association algorithm, the accurate object pose estimation algorithm, and a semi-dense mapping system. Fig. 10 shows two examples of TUM fr3_long_office and fr2_desk, where (d) and (e) show a semi-dense semantic map and an object-oriented map, build by EAO-SLAM. Compared with the sparse map of ORB-SLAM2, our maps can express the environment much better. Moreover, the object-oriented map shows the superior performance in environment understanding than the semi-dense map proposed in [25].

The mapping results of other sequences in TUM, Microsoft RGB-D, and Scenes V2 datasets are shown in Fig. 11. It can be seen that EAO-SLAM can process multiple classes of objects with different scales and orientations in complex environments. Inevitably, there are some inaccurate estimations. For instance, in the *fire* sequence, the chair is too large to be well observed by the fast moving camera, thus yielding an inaccurate estimation. We also conduct experiment in a real scenario, Fig. 12. It can be seen even the objects are occluded, they can be accurately estimated, which further verifies the robustness and accuracy of our system.

VII. Conclusion

In this paper, we present the EAO-SLAM system that aims to build semi-dense or lightweight object-oriented maps. The system is implemented based on a robust ensemble data association method and an accurate pose estimation framework. Extensive experiments show that our proposed algorithms and SLAM system can build accurate object-oriented maps with object poses and scales accurately registered. The methodologies presented in this work further push the limits of semantic SLAM and will facilitate related researches on robot navigation, mobile manipulation, and human-robot interaction.

References

[1] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," *arXiv preprint arXiv:2003.14338*, 2020.

[2] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric instance-aware semantic mapping and 3d object discovery," vol. 4, no. 3. IEEE, 2019, pp. 3037–3044.

[3] 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.

[4] T. Li, D. Zhu, and M. Q.-H. Meng, "A hybrid 3dof pose estimation method based on camera and lidar data," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2017, pp. 361–366.

[5] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[6] A. Iqbal and N. R. Gans, "Localization of classified objects in slam using nonparametric statistics and clustering," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 161–168.

[7] M. Strecke and J. Stuckler, "Em-fusion: Dynamic object-level slam with probabilistic data association," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5865–5874.

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

[9] 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.

[10] P. Parkhiya, R. Khawad, J. K. Murthy, B. Bhowmick, and K. M. Krishna, "Constructing category-specific models for monocular object-slam," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[11] S. Yang, Z.-F. Kuang, Y.-P. Cao, Y.-K. Lai, and S.-M. Hu, "Probabilistic projective association and semantic guided relocalization for dense reconstruction," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7130–7136.

[12] Z. Min, D. Zhu, H. Ren, and M. Q. . Meng, "Feature-guided nonrigid 3-d point set registration framework for image-guided liver surgery: From isotropic positional noise to anisotropic positional noise," *IEEE Transactions on Automation Science and Engineering*, pp. 1–13, 2020.

[13] J. Li, D. Meger, and G. Dudek, "Semantic mapping for view-invariant relocalization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7108–7115.

[14] Y. Liu, Y. Petillot, D. Lane, and S. Wang, "Global localization with object-level semantics and topology," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4909–4915.

[15] 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.

[16] D. Zhu, T. Li, D. Ho, C. Wang, and M. Q.-H. Meng, "Deep reinforcement learning supervised autonomous exploration in office environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7548–7555.

[17] D. Zhu, T. Li, D. Ho, T. Zhou, and M. Q. Meng, "A novel ocr-rcnn for elevator button recognition," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3626–3631.

[18] 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.

[19] N. Joshi, Y. Sharma, P. Parkhiya, R. Khawad, K. M. Krishna, and B. Bhowmick, "Integrating objects into monocular slam: Line based category specific models," *arXiv preprint arXiv:1905.04698*, 2019.

[20] D. Frost, V. Prisacariu, and D. Murray, "Recovering stable scale in monocular slam using object-supplemented bundle adjustment," *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 736–747, 2018.

[21] J. Peng, X. Shi, J. Wu, and Z. Xiong, "An object-oriented semantic slam system towards dynamic environments for mobile manipulation," in *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2019, pp. 199–204.

[22] A. K. Nellithimaru and G. A. Kantor, "Rols: Robust object-level slam for grape counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[23] S. Choudhary, L. Carlone, C. Nieto, J. Rogers, Z. Liu, H. I. Christensen, and F. Dellaert, "Multi robot object-based slam," in *International Symposium on Experimental Robotics*. Springer, 2016, pp. 729–741.

[24] 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.

[25] S. He, X. Qin, Z. Zhang, and M. Jagersand, "Incremental 3d line segment extraction from semi-dense slam," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1658–1663.

[26] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.

TABLE II: QUANTITATIVELY ANALYZED DATA ASSOCIATIONS

Seq	Tum				Microsoft RGBD					Scenes V2				
	fr1_desk	fr2_desk	fr3_long_office	fr3_teddy	Chess	Fire	Office	Pumpkin	Heads	01	07	10	13	14
[6]	-	11	15	2	5	4	10	4	-	5	-	6	3	4
Ours	**14**	**22**	**42**	**6**	**13**	**6**	**21**	**6**	**15**	**7**	**7**	**7**	**3**	**5**
GT	16	23	45	7	16	6	27	6	18	8	7	7	3	6

TABLE III: QUANTITATIVE ANALYSIS OF OBJECT ANGLE ERROR

Seq	fr3_long_office						fr1_desk				fr2_desk				Mean
Objects	book1	book2	book3	keyboard1	keyboard2	mouse	Book1	Book2	Tvmonitor1	Tvmonitor2	keyboard	Book1	Book2	mouse	
BI	19.2	11.4	16.2	10.3	7.4	11.3	33.5	15.2	32.7	22.5	8.9	15.5	16.9	8.7	16.4
AI	5.3	5.5	6.2	7.2	4.2	6.4	8.6	8.9	**6.0**	11.4	5.5	**3.8**	10.1	**7.5**	6.9
JO	**3.1**	**4.3**	**5.7**	**2.5**	**2.8**	**4.3**	**5.4**	**7.6**	8.7	**10.2**	**3.9**	5.1	**6.4**	7.9	**5.6**

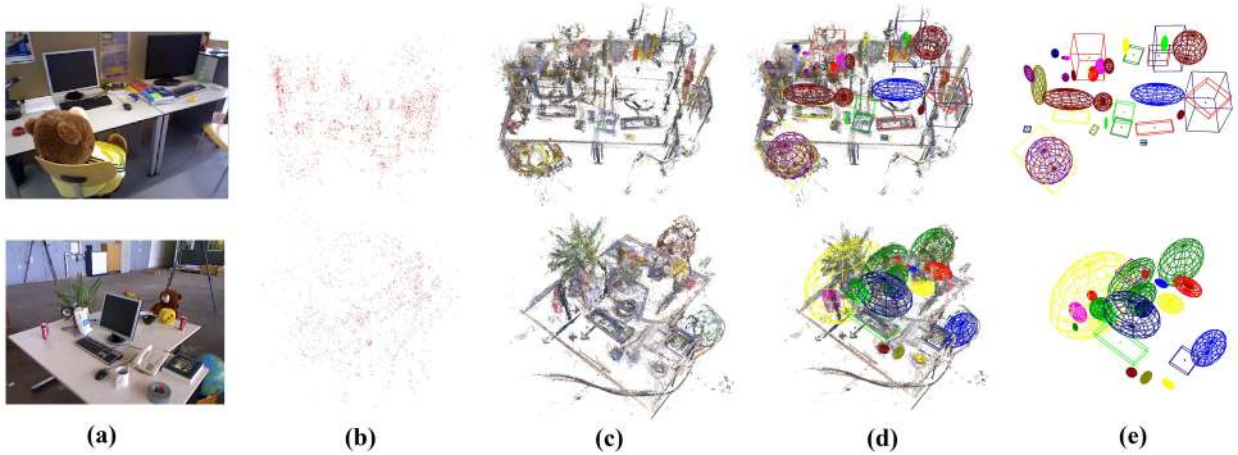

Fig. 10: Different map representations. (a) the RGB images. (b) the sparse map. (c) semi-dense map. (d) our semi-dense semantic map. (e) our lightweight and object-oriented map. (d) and (e) are build by the proposed EAO-SLAM.

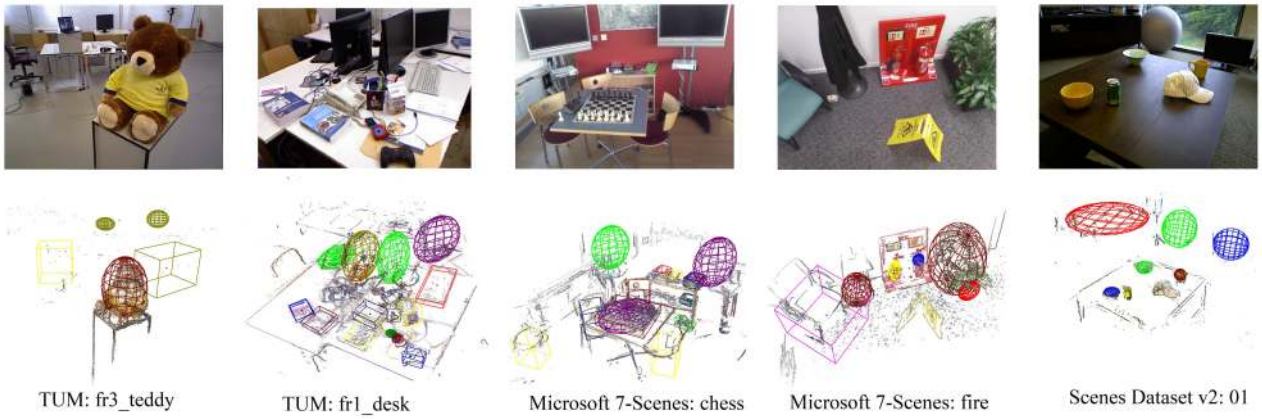

Fig. 11: Results of EAO-SLAM on the three datasets. Top: raw images. Bottom: simi-dense object-oriented map.

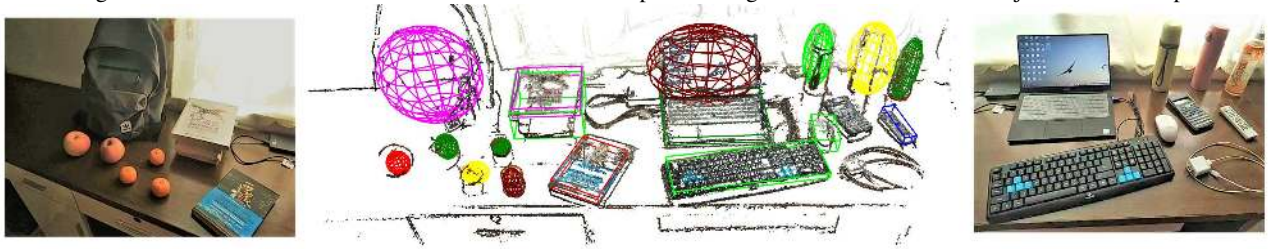

Fig. 12: Results of EAO-SLAM in a real scenario. Left and right: raw images. Middle: semi-dense object-oriented map.