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

Robust Improvement in 3D Object Landmark Inference for Semantic Mapping

Xubin Lin[1], Yirui Yang[1], Li He[1], Weinan Chen[2], Yisheng Guan[1], Hong Zhang[2]*

***Abstract*— Recent works on semantic Simultaneous Localization and Mapping (SLAM) utilizing object landmarks have shown superiority in terms of robustness and accuracy in tracking and localization. 3D object landmarks represented by a cubic or quadric surface are inferred from 2D object bounding boxes which are typically captured from multiple views by an object detector. Nevertheless, bounding box noises and small camera baseline may lead to an inaccurate 3D object landmark inference. Inspired by the dual quadric enveloping property, in this work, we introduce the horizontal support assumption to constrain rotation w.r.t. roll and pitch for a quadric representation. As the result, we reduce the number of quadric parameters and narrow down the solution space, and ultimately produce a relatively accurate inference. Extensive experimental evaluations under both simulated and real scenarios are conducted in this paper. Quantitative results demonstrate that our approach outperforms the state-of-the-art.**

***Index Terms*— Semantic Mapping, SLAM, Object Reconstruction, Dual Quadric**

I. INTRODUCTION

A classical visual simultaneous localization and mapping (VSLAM) system estimates on-board sensor poses online leveraging geometric primitives such as points, lines and planes extracted from an image stream, and simultaneously estimates primitives' spatial information. Such primitive-based algorithms are relatively mature, spawning a number of VSLAM systems like ORB-SLAM [1], LSD-SLAM [2] and SVO [3], etc. In a consequence of heavy dependence on pixel intensity gradient distribution, primitives extraction and matching method are sensitive to external changes. Although primitive-based systems provide promising local tracking, they often fail in the case of complicated data associations such as large scale change of viewpoint, significant illumination changes in long-term mapping and localization. Moreover, primitive-based map representation limits high-level tasks such as object manipulation and task-oriented interaction.

1 are with the Biomimetic and Intelligent Robotics Lab (BIRL), Guangdong University of Technology, Guangzhou, China, 510006.

2 are with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China.

* H. Zhang is the corresponding author (hzhong@ualberta.ca).

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61673125 and 61703115, in part by the Leading Talents of Guangdong Province Program under Grant No. 2016LJ06G498 and 2019QN01X761, in part by Guangdong Province Special Fund for Modern Agricultural Industry Common Key Technology R&D Innovation Team under Grant No. 2019KJ129.

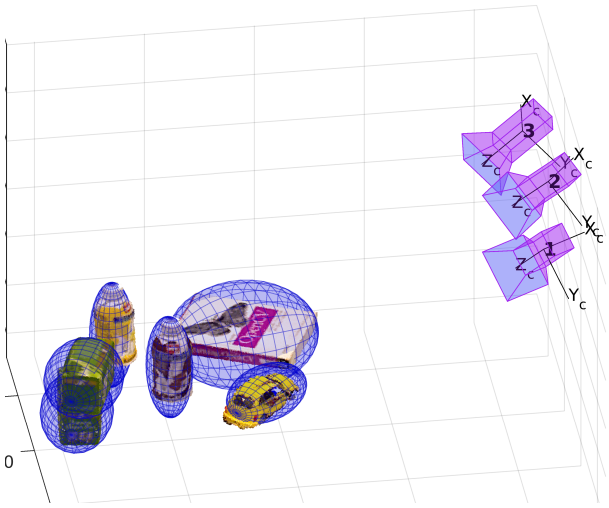

Fig. 1: An object in a semantic map is represented by an enveloping ellipsoid inferred from object bounding boxes in multiple views. Such a representation decreases the amount of stored data compared to dense point representation while maintaining essential information like semantic label, dimensions and pose.

On the other hand, the recent success of deep learning has boosted semantic processing including object detection, exhibiting its superiority of robustness and invariance against illumination and viewpoint changes. Rather than representing object elaborately with both dense points and texture, 3D enveloping surface inferred from 2D object bounding boxes in multiple views is widely adopted to represent object [4, 5], due to its potential to tremendously reduce the amount of data while maintaining essential information such as semantic label, dimensions and pose. In particular, quadric based representation, by which an object is modeled by a sphere or an ellipsoid, has been proposed intensively in recent years [6, 7], as shown in Fig. 1. Among existing works, the reconstructed quadric is refined via consecutive images, and a decent initial quadric estimate is crucial due to the non-convex property of such an optimization problem. However, bounding box noise and the small camera baseline may lead to a bad quadric initialization, and consequently result in the degradation of a SLAM system. To overcome such a problem, we introduce the horizontal support assumption and propose a new solution to obtain the quadric

presentation parameters based on such an assumption, in which the roll and pitch of a quadric is constrained. Extensive experimental evaluations both in simulated and real scenarios are conducted in this paper, and our method outperforms the baseline approaches under the conditions of varying noise and small camera baseline, showing the adaptability of our assumption.

The contributions of our work are as follows:

- We introduce a horizontal support assumption and propose a new solution to obtain the parameters of an orientation-constrained ellipsoid.
- Our method eliminates ambiguity of ellipsoid reconstruction within two views, rather than three views used in previous works.
- Our method narrows down the solution space and guarantees a solution lying in the ellipsoid space.
- We obtain higher accuracy and more robust performance than the state-of-the-art.

II. RELATED WORK

A. Object Landmark Representation in SLAM

Geometric primitives such as point, line and plane, regardless used alone or in combination, are adopted to represent an environment among the most current VSLAM system [1, 8]. Either feature-based or optical-based [2, 3], current VSLAM systems focus on tracking accuracy and robustness, rather than the representation and maintenance of the map. Elaborate information including 3D positions, pixel intensities and visual feature descriptions are preserved in dense mapping systems like RGBD-SLAM [9] or ElasticFusion [10]. However, such a representation carries no semantic information, which consequently may affect robustness and accuracy within challenging circumstances. As Bowman *et al.* [11] and Doherty *et al.* [12] declared that joint optimization of semantic landmark inference and data association could benefit each other, obtaining a better tracking accuracy.

Semantic landmarks can be utilized and represented in various ways. SemanticFusion [13] reconstructs a map with dense surfel enriched with semantic labels. Representation in SemanticFusion uses substantially point cloud clusters associated with semantic information. Such a non-object-oriented representation has a limitation of further application like global localization [14]. In contrast, CubeSLAM [4] infers 3D cuboids from 2D object bounding boxes offered by a Convolutional Neural Network (CNN) based object detector such as YOLO [15] or Faster R-CNN [16]. Li *et al.* [14] also proposed a cuboid based representation mapping approach but tend to address the viewpoint-invariant localization problem within this map. Quadric based representations have superiority of mathematical completeness in projective geometry, whereas a cuboid based approach relies on optimization solver in most cases. Rubino *et al.* [17], Hosseinzadeh *et al.* [7] and QuadricSLAM [6] all envelope objects with a 3D quadric surface or an ellipsoid more specifically.

B. Dual Quadric Reconstruction

Geometric properties related to the quadric locus and its duality quadric envelop have been thoroughly revealed in algebraic projective geometry. Internal relationship between 3D quadric surfaces and 2D occluding contours (conics) from multiple views are investigated in the 1990s. Ma *et al.* described in [18] a linear formulation scheme for reconstructing a general ellipsoid given three contours from corresponding perspective image views, while Cross *et al.* [19] recovered quadric given three contours or two contours with pairwise additional matched points. Both of them transformed quadrics into the dual space in which a quadric is defined by envelop planes, expressed in a more concise way. One of the remarkable conclusions in [19] is proved to be the unambiguous reconstruction of a general quadric with the demand of at least three contours from three views; otherwise a 1D linear family of quadric is recovered. Shashua *et al.* [20] recovered a quadric from one outline and four matched points, whereas Wijewickrema *et al.* [21]utilize two contours from two views to inference a sphere. Most of these works recovered quadric directly from image outlines.

In contrast, Rubino *et al.* [17] infer quadric leveraging ellipses fitted from 2D object bounding boxes, as an extension of [19]. To avoid ellipse fitting, QuadricSLAM [6] back-projected each edge of a bounding box to an enveloping plane, thus determining an optimal dual quadric tangent to the planes. These two works both take extra views into consideration rather than using minimal views only. As Rubino *et al.* [17] make efforts to unified multiple view information into a linear scheme with a closed-form solution, QuadricSLAM [6] develops a similar linear scheme to initialize a quadric with three views and subsequently optimize its estimate under the SLAM paradigm as consecutive images are captured. Both of them have not taken robot motion into consideration, as pointed out by Ok *et al.* [22], a small camera baseline caused by robot motion may result in gross estimate performance within the traditional linear scheme. Ok *et al.* [22] therefore proposed a texture plane as an extra constraint to quadric, which is a plane parallel to the image plane fitted to the highest gradient texture on an object. As for initialization, to reduce time consumption, they estimate a coarse quadric along the camera ray oriented the same as the camera, expecting a quick convergence to offer a fine quadric subsequently.

Different from the existing works, we focus on the ellipsoid initialization problem in the SLAM paradigm, which stimulate us to pursue an accurate enveloping ellipsoid estimate with minimal views. Rather than reconstructing an ellipsoid at full degrees of freedom, we reconstruct an orientation-constrained one under our horizontal support assumption, aiming to offer a better initial estimate to the SLAM back-end. Extensive evaluations suggest that the method implemented based on our parameterization is more robust and accurate than the baseline approach against noise and small camera baseline effect.

III. METHODOLOGY

A. Dual Quadric Formulation

In homogeneous coordinates, a quadric is a surface in $\mathbb{P}^3$ defined by a general equation of the form $\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} = 0$, where $\mathbf{Q}$ is a symmetric 4×4 matrix. A quadric has 9 Degrees of Freedom (DoF) since 10 independent elements $\mathbf{q} = \{q_1, q_2, \ldots, q_{10}\}$ corresponding to $\mathbf{Q}$ are equivalent up to scale. Quadric surfaces, such as ellipsoid, hyperboloid and cone, can be distinguished from the signature of matrix $\mathbf{Q}$ [23]. The 2D projection of a quadric is a 5-DoF conic which is usually known as occluding contour or outline of the quadric, represented by a 3×3 symmetric matrix $\mathbf{C}$ with 6 independent elements $\mathbf{c} = \{c_1, c_2, \ldots, c_6\}$. Intuitively, quadric and its conic outline have a closed geometric relationship in multiple view geometry. Nevertheless, the mathematical expression is somewhat complicated with primal form of $\mathbf{Q}$ and $\mathbf{C}$, which can be simplified by transforming the problem to the dual space.

Dual quadric, also known as quadric envelope, is defined by a bunch of tangent planes $\boldsymbol{\pi}$ to $\mathbf{Q}$, satisfying the equation $\boldsymbol{\pi}^\mathsf{T}\mathbf{Q}^*\boldsymbol{\pi} = 0$, where $\mathbf{Q}^* = \mathbf{adjoint}(\mathbf{Q})$. In particular, for a non-degenerate quadric (e.g. ellipsoid), $\mathbf{Q}^* = \mathbf{Q}^{-1}$ holds true up to scale, and it is also valid for a dual ellipse $\mathbf{C}^*$.

In dual form, a quadric and its conic outline can be connected via

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

where $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ is the camera projection matrix, and the homogeneous scale factor λ indicates that Eq. 1 is defined up to scale. This equation shows that a conic outline can offer at most 5 constraints on the quadric. Because of epipolar constraints, the cones back-projected by two conic outlines corresponding to the same quadric share two common tangent planes. Consequently, two views can only offer 8 constraints while we need at least 9 constraints from Eq. 1, which is insufficient to determine a unique quadric [19]. Furthermore, such a general parameterization $\mathbf{Q}^*$ is not explicitly restricted within the category of quadric. Because of observation noise, the solution is not guaranteed to be a closed surface, such as a sphere or an ellipsoid.

B. Horizontal Support Assumption based Parameterization

Semantic SLAM demands object enveloping quadric to be initialized from minimal views and noisy bounding boxes. To this end, we seek for an novel parameterization which reduce the degrees of freedoms of ellipsoid and explicitly restrict the solution lying in the ellipsoid space. It is intuitive to perceive that most objects at various scales in a scene are usually placed on the ground or desk-like horizontal support plane. Such an observation excludes parameter reduction in the aspect of ellipsoid semi-axes as proposed by [24] and [21]. We, therefore, propose a horizontal support assumption, which suggests that the degrees of freedom of an enveloping quadric in terms of pitch and roll are less necessary, with the compensation of varying semi-axes. According to such an assumption, the freedom w.r.t. roll and pitch of the ellipsoid is fixed to aligned with horizontal coordinate, reducing the solution space dimension from 9 to 7. Furthermore, our assumption explicitly parameterizes an ellipsoid, compelling the solution to lie in the ellipsoid space and avoiding a non-closed surface solution.

To investigate the parameters' internal relationship, we decouple the dual quadric matrix $\mathbf{Q}^*$. Denote $\breve{\mathbf{Q}}$ as an ellipsoid taking x, y, z direction as its semi-axes direction with length $a, b, c \in \mathbb{R}^3$, respectively. We therefore have its dual form

$$\breve{\mathbf{Q}}^* = \begin{bmatrix} a^2 & 0 & 0 & 0 \\ 0 & b^2 & 0 & 0 \\ 0 & 0 & c^2 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}. \tag{2}$$

Subsequently, with assumed orientation $\mathbf{R}$ and center position $\mathbf{t}$, a Euclidean transformation matrix can be synthesized as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_3^\mathsf{T} & 1 \end{bmatrix} \in \mathbb{SE}(3). \tag{3}$$

According to the transformation rule in the dual space, a general ellipsoid at specific scale and pose can be obtained as follows:

$$\mathbf{Q}^* = \mathbf{T}\breve{\mathbf{Q}}^*\mathbf{T}^\mathsf{T} \tag{4}$$

With our assumption, the rotation matrix $\mathbf{R}$ collapses into a subgroup of $\mathbb{SO}(3)$, i.e. 3D rotation regarding with yaw of value θ, and the transformation matrix turns into:

$$\mathbf{T_z} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 & t_x \\ \sin(\theta) & \cos(\theta) & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

Therefore, a pitch-and-roll-constrained ellipsoid in dual form can be expressed as

$$\mathbf{Q_z}^* = \mathbf{T_z}\breve{\mathbf{Q}}^*\mathbf{T_z}^\mathsf{T} \tag{6}$$

Notice that the solution to Eq. 6 is guaranteed to be an ellipsoid, since $\mathbf{Q}_\mathbf{z}^*$ has the same signature of $(+,+,+,-)$ with $\breve{\mathbf{Q}}$.

Finally, the 10 independent elements encoding 7 variables in symmetric matrix $\mathbf{Q}_\mathbf{z}^*$ can be expressed as:

$$\mathbf{q_z} = \begin{bmatrix} a^2\cos^2\theta + b^2\sin^2\theta - t_x^2 \\ (a^2 - b^2)\sin\theta\cos\theta - t_x t_y \\ -t_x t_z \\ -t_x \\ a^2\sin^2\theta + b^2\cos^2\theta - t_y^2 \\ -t_y t_z \\ -t_y \\ c^2 - t_z^2 \\ -t_z \\ -1 \end{bmatrix} \tag{7}$$

C. Reconstruction of Ellipsoid

Ellipsoid reconstruction problem is equivalently to solve Eq. 1 given conic outline $\mathbf{C}^*$ and projection matrix $\mathbf{P}$, which shows the linear relationship between a dual quadric and its dual conic outline. As suggested in [17], the quadratic term

existing in the conic outline may amplify the diversity of Eq. 1, and result in vulnerability against perturbation from noises of bounding box and projection matrix. This problem can be alleviated by transforming the original dual conic into its canonical form with a homogeneous transformation $\mathbf{H}$, as follows:

$$\mathbf{C}^* = \mathbf{H}\breve{\mathbf{C}}^*\mathbf{H}^\mathsf{T} \tag{8}$$

where

$$\mathbf{H} = \begin{bmatrix} h & 0 & t_1 \\ 0 & h & t_2 \\ 0 & 0 & 1 \end{bmatrix}, \breve{\mathbf{C}}^* = \begin{bmatrix} c_{11}^* & c_{12}^* & 0 \\ c_{12}^* & c_{22}^* & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{9}$$

$[t_1, t_2]$ is the center position of $\mathbf{C}^*$, and $h = \sqrt{l_1^2 + l_2^2}$ where $l_1, l_2 \in \mathbb{R}$ are two semi-axes of $\mathbf{C}^*$. Under such a homogeneous transformation, along with the substitution with our parameterization described in Eq. 6, Eq. 1 becomes

$$\lambda\breve{\mathbf{C}}^* = \mathbf{H}^{-1}\mathbf{P}\mathbf{Q}_\mathbf{z}^*\mathbf{P}^\mathsf{T}\mathbf{H}^{-\mathsf{T}}. \tag{10}$$

With vectorization representation, Eq. 10 can be rewritten to

$$\lambda \begin{bmatrix} \breve{c}_1^* \\ \vdots \\ \breve{c}_6^* \end{bmatrix} = \mathbf{B} \begin{bmatrix} q_{z1}^* \\ \vdots \\ q_{z10}^* \end{bmatrix} \tag{11}$$

or more concisely as $\breve{\mathbf{c}}^* = \mathbf{B}\mathbf{q}^*$ where $\mathbf{B}$ is a 6×10 matrix which is the quadratic in the elements of $\mathbf{H}^{-1}P$, $\breve{\mathbf{c}}^*$ consists of 6 elements corresponding to the preconditioned $\breve{\mathbf{C}}^*$ whereas $\mathbf{q_z}^*$ represents the ellipsoid with our parameterization detailed in Eq. 7.

Finally, the ellipsoid reconstruction problem can be formulated by collecting all conics $\breve{\mathbf{c}}_i^*$ and $\mathbf{H}_i^{-1}P_i$ corresponding to different views as:

$$\mathbf{M}\mathbf{v} = \mathbf{0} \tag{12}$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{B_1} & \breve{\mathbf{c}}_1^* & 0 \\ \mathbf{B_2} & 0 & \breve{\mathbf{c}}_2^* \end{bmatrix} \mathbf{v} = \begin{bmatrix} \mathbf{q_z}^* \\ \lambda_1 \\ \lambda_2 \end{bmatrix}$$

With the benefits of our parameterization, reconstruction ambiguity can be eliminated within two views although extra views with wide camera baseline will, typically, offer a better result. Note that Equation (12) is nonlinear in terms of $\mathbf{e} = \{\theta, a, b, c, t_x, t_y, t_z\}$ embedded in $\mathbf{q}_\mathbf{z}^*$. Toward such an optimization problem, we can find the solution by minimizing A closed-form solution to Eq. 12 is not easy to obtain, so we attempt to minimize

$$\tilde{\mathbf{e}} = \underset{\mathbf{e}, \lambda_1, \lambda_2}{\arg\min} \quad \|\mathbf{M}\mathbf{v}\|_2^2 \tag{13}$$

by instead.

The estimated ellipsoid $\mathbf{Q}_\mathbf{z}^*$ can be recovered from $\tilde{\mathbf{e}}$ via parameterization from Eq. 7. To facilitate the convergence of Eq. 13, we triangulate the 3D point from two bounding boxes centers, and regard its 3D coordinate as the ellipsoid initial position t_x, t_y, t_z. Depth value of other pixels provide a coarse range of the ellipsoid volume, offering the bounds of a, b, c in our problem. As for the reference coordinates, the ellipsoids are all with respect to the canonical world coordinate. [25] offered an effective method of computing the scene major axes in one image view, and then the camera coordinates can be aligned to such canonical scene coordinates.

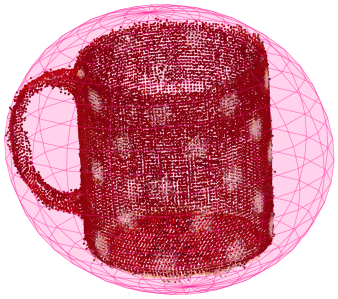

Fig. 2: Demonstration on the point cloud of the object (red mug with white spots) appeared in TUW dataset. Optimal enveloping ellipsoid toward such an object is estimated using point cloud and is regarded as ground truth.

IV. EXPERIMENTAL EVALUATION

We conduct three experiments under both simulated and real scenarios in a statistical manner to evaluate our method. Comprehensive performance was evaluated in two synthetic experiments on the exertion of different noise magnitudes and varying roll and pitch, which is different from our assumption. In addition, we have implemented two baseline approaches [17, 24] and compared with ours in a public dataset TUW [26], as demonstrated in Fig. 2. More details of this project can be found at `https://github.com/XubinLin/quadRecon`

As mentioned in Section III, our parameterization is able to reconstruct a unique ellipsoid with two views. However, to be consistent with the baseline approaches and present a more comprehensive evaluation, we implemented our methods using both three views and two views (denoted Ours-3views and Ours-2views, respectively). In both synthetic experiments, we assumed the extrinsic and the intrinsic matrices are the same as the first three views (or the first and third view in the 2-views implementation case) in sequence 7 in the TUW dataset. All the evaluations in this paper are measured by average precision (AP) of 3D Intersection over Union (IoU) between the reconstructed ellipsoid and the ground truth one, defined as:

$$AP = \frac{1}{N}\sum_{i=1}^{N}\frac{Q_{gt} \cap \tilde{Q}}{Q_{gt} \cup \tilde{Q}} \tag{14}$$

In Eq. 14, Q_{gt} is the ground truth ellipsoid whereas $\tilde{Q}$ is the ellipsoid reconstructed by all competing methods. For statistical purposes, the average precision is counted with N reconstruct samples.

A. Evaluation of Noise Effect

In this experiment, 100 ellipsoids at random scales are synthesized and placed randomly, and are projected onto

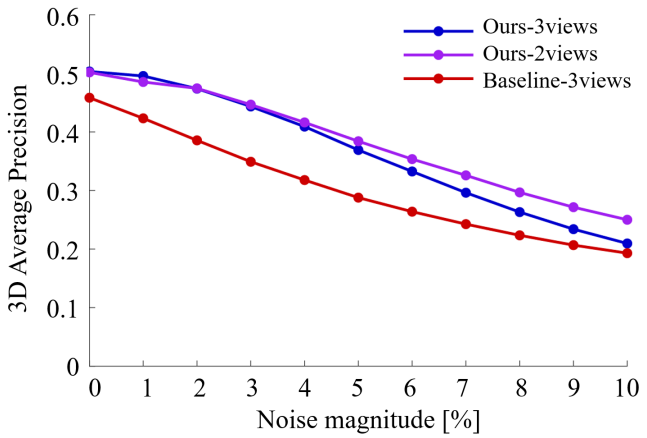

Fig. 3: Performance for noise magnitude from 0 to 10%.

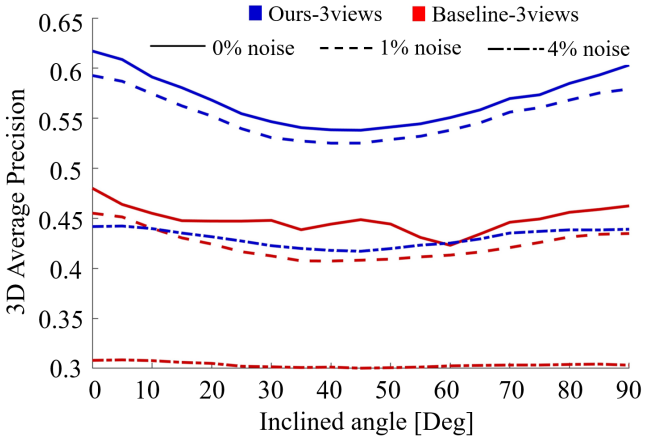

Fig. 4: Average precision curve on incline angle affection evaluations. Our precision is high when the incline angle is around $0°$ or $90°$,a case consistent with our horizontal support assumption.In addition, at the same incline angle and noise, our method outperforms the baseline in all tests.

each view, generating various ellipses. Since most popular Deep Learning object detectors generate axis-aligned bounding boxes, we envelop the ellipses with axis-aligned bounding boxes and push the generated bounding boxes to our methods. Typically, it is a tough task to detect the bounding box that perfectly envelops an object. Therefore, we add Gaussian noise to the coordinates of the top left and right bottom corners with zero means and standard deviation proportional to the diagonal length l of the bounding box. The standard deviation is set at $\sigma = l \times n$, where n is referred to noise magnitude ranging from 0 to 10% with interval 5%. At each specific noise magnitude, we take 250 samples from each bounding box in each view. Eventually, our 100 synthetic ellipsoids result in 100×250 bounding box samples at each noise magnitude, and the average precision is calculated using Eq. 14.

The results are depicted in Fig. 3, showing that noise has a negative influence on the average precision among all three competing methods, and our methods outperform the baseline approach [17] at each noise magnitude. As noise magnitude increasing, our 2-view-based implementation slightly outperforms our 3-view-based one. Since two views provide enough constraints on an orientation-constrained ellipsoid, additional noised view may depress the final accuracy, in particular in the case of strong noises. Notice that we assume there is no 2D rotation in our bounding box, meaning that the sampled ellipse used by all methods inevitably deviates from ground truth. Therefore, even with noise zero the average precision is not 1 in this case.

B. Validation of Assumption

We present this synthetic experiment to evaluate the adaptability of our horizontal support assumption by changing the incline angle of each synthetic ellipsoid. We synthesized 100 upright ellipsoids and placed them randomly. Afterward, each upright ellipsoid inclined gradually from 0 to $90°$ with interval $5°$ at a random direction. Axis-aligned bounding box in each view can be obtained by enveloping the outline corresponding to an ellipsoid at each inclined angle, and is regarded as the input to our 3-view-based implementation and the baseline method [17]. For each noise magnitude, we add Gaussian noise to bounding box corners and sample 250 boxes, and calculate the average precision at each incline angle. We repeat such experiments under noise magnitude $0\%, 1\%$ and 4%.

The result is shown in Fig. 4. Around incline angle $0°$ and $90°$, our method provides high precision whereas the precision of our method becomes low around $45°$, which is consistent with our assumption. In addition, our method outperforms the baseline approach even at incline angle $45°$, where our assumption is mostly violated, which implies the adaptability of our method.

C. TUW Dataset Evaluation

The TUW dataset [26] contains 15 annotated RGB-D sequences originally built for point cloud recognition, segmentation and registration. Each sequence offers 6 to 15 images with ground truth poses, along with point cloud of objects deployed in the scene. Same as [17], we discarded sequences with strong occlusions and retained 5 sequences for our evaluation. The point cloud of an object is enveloped with an ellipsoid which is regarded as the ground truth, as shown in Fig. 2. To investigate performance with minimal views, we take various combinations of three or two views on each sequence, and calculate the mean of average precision of each combination as the final score. In addition to the baseline method [17], our 3-view-based implementation (denoted

TABLE I: Average precision for the Sequences from TUW Dataset

	Seq1	Seq7	Seq8	Seq10	Seq11	Avg
Baseline	0.341	0.470	0.603	0.556	0.453	0.484
Ours-3views	0.440	**0.632**	**0.673**	0.649	**0.576**	**0.594**
Ours-2views	**0.472**	0.622	0.657	**0.652**	0.504	0.581
pQoR	0.06	0.08	0.16	0.13	0.04	0.10

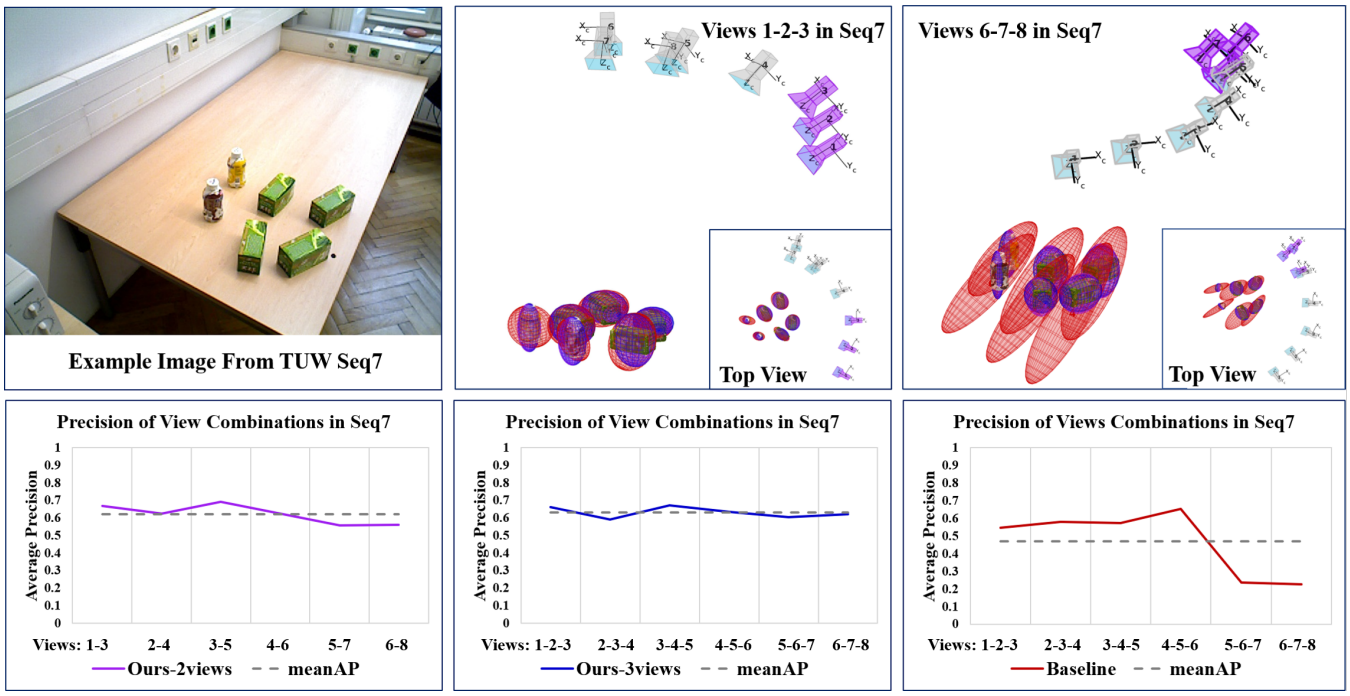

Fig. 5: Illustration of evaluation with regard to Seq. 7 in TUW dataset. The first row shows reconstruction results from Ours-2views (in magenta), Ours-3views (in blue) and the baseline approaches (in red), using different combinations of views. The second row depicted the average precision under different combinations of views, given by Ours-2views (left), Ours-3views (middle), and the baseline approach (right), suggesting that our implementations are relatively robust, especially in the case of views 6-7-8 which has a smaller camera baseline.

Ours-3views) and 2-view-based implementation (denoted Ours-2views), we also implemented the method proposed in [24] (denoted as pQoR). Average precision on each sequence is reported in TABLE I. Results suggest that the precision of our methods (both 3-view-based and 2-view-based) exceed other two competing methods in all conditions, whereas our 3-view-based implementation performs better than 2-view-based one. In the case of only two views available, our 2-view-based implementation outperforms pQoR, since pQoR has a strong assumption that one ellipsoid has to be strict prolate quadric of revolution. The pipeline of pQoR [24] demands to estimate the image of the revolution axis from the conic outline at first, which is inaccurate for such a bounding box, thus leading to a poor performance ultimately.

Concerning the influence of the camera baseline, we plot in Fig. 5 the average precision for each view combination. Fig. 5 shows that our implementations are relatively robust against camera baseline changes. In contrast, the baseline method [17] has significant degradation at small camera baseline, such as in the case of views-567 and views-678. This explains why the performance of the baseline method [17] is not as good as the original paper, in which all views are taken into consideration and thus a wide camera baseline is obtained.

V. CONCLUSIONS

This paper proposed a novel method of 3D object landmark inference for semantic mapping, specifically focusing on robustness improvement for the inference of ellipsoid representation. By introducing horizontal support assumption, we propose a new parameterization for an orientation-constrained ellipsoid, eliminating ambiguity of ellipsoid reconstruction within two views. Our method narrows down the solution space, and guarantees a solution lying in the ellipsoid space. Our method outperforms the state-of-the-art among simulated and real experimental evaluations, demonstrating the adaptability of our assumption and the robustness of our method.

REFERENCES

[1] 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.

[2] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[3] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2016.

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

[5] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese, "Semantic structure from motion with points, regions, and objects," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2703–2710.

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

[7] 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.

[8] R. Gomez-Ojeda, F.-A. Moreno, D. Zuñiga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, "Pl-slam: A stereo slam system through the combination of points and line segments," *IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 734–746, 2019.

[9] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *IEEE transactions on robotics*, vol. 30, no. 1, pp. 177–187, 2013.

[10] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "Elasticfusion: Real-time dense slam and light source estimation," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1697–1716, 2016.

[11] 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.

[12] K. Doherty, D. Fourie, and J. Leonard, "Multimodal semantic slam with probabilistic data association," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 2419–2425.

[13] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and automation (ICRA)*. IEEE, 2017, pp. 4628–4635.

[14] 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.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

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

[18] S. Ma and L. Li, "Ellipsoid reconstruction from three perspective views," in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 1. IEEE, 1996, pp. 344–348.

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

[20] A. Shashua and S. Toelg, "The quadric reference surface: Theory and applications," *International Journal of Computer Vision*, vol. 23, no. 2, pp. 185–198, 1997.

[21] S. N. Wijewickrema, A. P. Paplinski, and C. E. Esson, "Reconstruction of spheres using occluding contours from stereo images," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 151–154.

[22] 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.

[23] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[24] P. Gurdjos, V. Charvillat, G. Morin, and J. Guénard, "Multiple view reconstruction of a quadric of revolution from its occluding contours," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 1–12.

[25] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2136–2143.

[26] A. Aldoma, T. Fäulhammer, and M. Vincze, "Automation of "ground truth" annotation for multi-view rgb-d object instance recognition datasets," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 5016–5023.