

# Quasi-globally Optimal and Real-time Visual Compass in Manhattan Structured Environments

Pyojin Kim<sup>1</sup>, Haoang Li<sup>2</sup>, and Kyungdon Joo<sup>3,†</sup>

**Abstract**—We present a drift-free visual compass for estimating the three degrees of freedom (DoF) rotational motion of a camera by recognizing structural regularities in a Manhattan world (MW), which posits that the major structures conform to three orthogonal principal directions. Existing Manhattan frame estimation approaches are based on either data sampling or a parameter search, and fail to guarantee accuracy and efficiency simultaneously. To overcome these limitations, we propose a novel approach to hybridize these two strategies, achieving quasi-global optimality and high efficiency. We first compute the two DoF of the camera orientation by detecting and tracking a vertical dominant direction from a depth camera or an IMU, and then search for the optimal third DoF with the image lines through the proposed Manhattan Mine-and-Stab (MnS) approach. Once we find the initial rotation estimate of the camera, we refine the absolute camera orientation by minimizing the average orthogonal distance from the endpoints of the lines to the MW axes. We compare the proposed algorithm with other state-of-the-art approaches on a variety of real-world datasets including data from a drone flying in an urban environment, and demonstrate that the proposed method outperforms them in terms of accuracy, efficiency, and stability. The code is available on the project page: <https://github.com/PyojinKim/MWMS>

## I. INTRODUCTION

Estimating the 3-DoF rotational motion of autonomous agents is a fundamental problem for many applications in computer vision and robotics [1], [2], [3]. It is well-known that the rotational drift error and nonlinearity during the 3-DoF camera orientation estimation are the main sources of positioning inaccuracy [4] in visual odometry (VO) and simultaneous localization and mapping (SLAM). Thus, it is extremely important to obtain accurate and drift-free estimates of the rotational motion of autonomous agents for visual navigation in indoor or outdoor urban environments.

In particular, practical robotic platforms and computer vision applications (Pokémon GO, IKEA Place AR) rely heavily on proprietary visual-inertial odometry (VIO) and SLAM methods such as Apple ARKit and Google ARCore to obtain a reliable 3-DoF rotational motion of the camera. Although they can estimate the accurate roll and pitch angles (2-DoF) by utilizing the gravity direction [6] or the surface

<sup>1</sup>Pyojin Kim is with Department of Mechanical Systems Engineering, Sookmyung Women's University, Seoul, South Korea. {pjinkim}@sookmyung.ac.kr

<sup>2</sup>Haoang Li is with Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong, China. {haoang.li.chuk}@gmail.com

<sup>3</sup>Kyungdon Joo is with the Artificial Intelligence Graduate School and the Department of Computer Science and Engineering, UNIST, Ulsan, South Korea. kdjoo369@gmail.com, kyungdon@unist.ac.kr

<sup>†</sup>Corresponding author: Kyungdon Joo

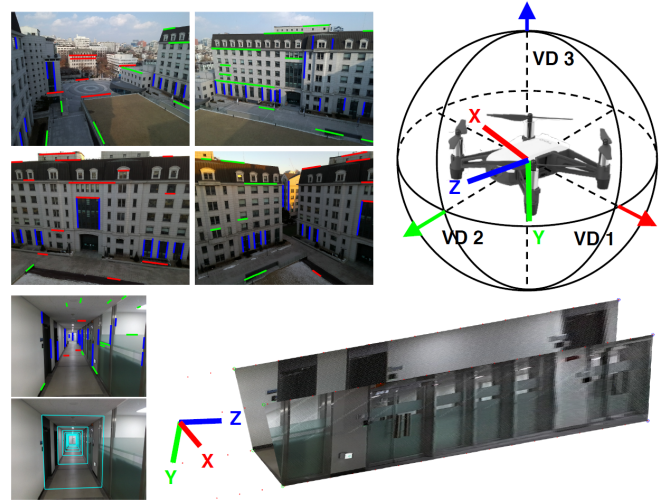


Fig. 1: In an urban area, the proposed method can serve as a drift-free visual compass for drones flying over a city by tracking the structural regularities – MF with respect to the camera frame (top). We conduct a single-view 3D reconstruction with 3D box priors [5] using the proposed method, which consistently reconstructs the interior 3D structures with a single RGB image (bottom).

normal to the ground plane [7], they still suffer from a drift error over time for the rotation about the vertical axis, the yaw angle as shown in Fig. 10.

Several recent studies [8], [9], [10] have focused on drift-free 3-DoF rotation estimation by utilizing the structural regularities in urban and indoor environments consisting of three mutually orthogonal dominant directions called a Manhattan world (MW) [11]. To compute the absolute 3-DoF camera orientation, the state-of-the-art approaches are based on data sampling [12], [13], [14], a parameter search [15], [16], or a combination of both [9]. The data sampling-based approaches hypothesize the Manhattan frame (MF) candidates using the sampled image lines and/or surface normals, providing high efficiency and stability. They fail to guarantee global optimality in terms of the number of inliers due to sampling randomness and uncertainty. The parameter search methods directly inspect the infinite MF hypothesized over the rotation search space, and continuously narrow down the search space. While they can achieve global optimality, their efficiency and effectiveness are limited in real-time robotic applications due to the high-dimensional search space and heavy computational load.

To address these issues, we propose a novel real-time visual compass that hybridizes the data sampling and efficient parameter search strategies to accurately and efficiently recognize the spatial regularities of orthogonal

structured environments (see Fig. 1). We first detect and track the normal vector to the ground plane (or gravity) from an RGB-D camera or an IMU as a vertical dominant direction (VDD) to compute the 2-DoF of the MF. We exploit the mine-and-stab (MnS) [10] approach to search for the optimal third DoF as a horizontal dominant direction (HDD) with the image lines. Thanks to our new MnS approach, which takes full advantage of the periodicity of the Manhattan structure, we can obtain an accurate and drift-free 3-DoF camera orientation in a “quasi-globally” optimal manner. Specifically, it guarantees the retrieval of the maximum number of inliers under the condition that the 2-DoF (VDD) is constrained. Furthermore, we refine the initial rotation estimate by minimizing the average orthogonal distance of the inlier lines parallel to the MW axes. Extensive evaluations show that our method can stably track the drift-free 3-DoF rotational motion of the camera in real-time in a variety of challenging indoor and outdoor environments. It can also be utilized as a visual compass in computer vision applications such as single-view 3D reconstruction with 3D box priors [5] as shown in Fig. 1. Our main contributions are as follows:

- We propose a novel visual compass to estimate accurate and drift-free rotational motion of the camera jointly from both the ground plane (or gravity) and lines by utilizing the structural regularities of the MW.
- We leverage the surface normal to the ground plane or gravity direction to efficiently compute the 2-DoF of the camera orientation, accelerating our rotation search by reducing the rotation search space.
- We propose an efficient MnS approach utilizing the periodicity of the Manhattan structure to search for the optimal third DoF of the camera orientation, achieving the quasi-global optimality in terms of maximizing the number of inlier lines.

In addition, we evaluate our visual compass on the ICL-NUIM [17] and York Urban [18] datasets, as well as on a new dataset from a low-cost drone traversing an outdoor urban area, showing accurate and drift-free rotation estimates.

## II. RELATED WORK

Research on the estimation of accurate rotational motion utilizing structural regularities such as the Manhattan [11], Atlanta [19] worlds has been actively studied in computer vision and robotics communities over the past decade. We can classify the existing approaches into two main categories with respect to the algorithms used: data sampling [8], [12], [13] and a parameter search [20], [15], [16], [10].

Data sampling-based approaches exploit RANSAC and its variants [21] given the lines from the RGB images or surface normals from the depth images. In [12], [13], the authors sample three image lines several times to hypothesize finite MF rotations in RANSAC, then retrieve the best MF hypothesis satisfying most inlier lines. Tardif et al. [22] utilize numerous MF hypotheses to define the image line descriptors and cluster lines concerning the MW using J-Linkage [21], a variant of RANSAC. Since these line-based sampling methods are sensitive and unstable in the presence

of spurious or noisy line segments, they are unsuitable for a robust orientation estimation of autonomous agents.

Recent studies [23], [24] have utilized the distribution of sampled surface normals to estimate dominant orthogonal directions in an MW from a depth sensor such as Intel RealSense and Microsoft Kinect. Although these surface normal-based sampling approaches demonstrate a stable and accurate rotation estimation [8], they require a dense surface normal distribution and at least two orthogonal planes must always be visible [25], which are unsuitable for use in outdoor robotic platforms such as flying drones in an open urban environment as shown in Fig. 1. Kim et al. [14] exploit the line and plane primitives together to estimate the drift-free camera orientation in an MW. However, this method still requires random sampling and is evaluated only in indoor datasets. Note that the above sampling-based approaches cannot guarantee global optimality in terms of the number of inliers due to sampling uncertainty.

The parameter search-based methods rely on a branch-and-bound (BnB) [15], [26] or the recently proposed mine-and-stab (MnS) [10]. Bazin et al. [15], [26] search for the optimal 3-DoF MF rotation satisfying the most image lines by iteratively narrowing down the rotation search space. Joo et al. [16] recently present a highly efficient BnB approach using the distribution of surface normals from a depth camera. While these BnB-based methods can guarantee global optimality in terms of maximizing the number of inliers, it takes more than three seconds per image to find the best solution. Since dense surface normals may be unavailable in practice due to the limited range of the depth camera, BnB-based methods lead to a relatively low applicability in various environments for indoor and outdoor robotic applications. Li et al. [10] recently propose an efficient parameter search method called the MnS algorithm; however, it is applied and tested only to determine the horizontal dominant directions of the Atlanta world.

Overall, existing approaches based on data sampling or a parameter search have failed to achieve high efficiency, stability, and accuracy simultaneously. Recent studies [27], [9] have tried hybridizing these two strategies, but they are still dependent on the computationally expensive BnB and unstable for use in robotic applications because they rely heavily on lines only. Our method overcomes these limitations thanks to the combination of data sampling and the newly proposed MnS approach, which takes full advantage of the periodicity of the Manhattan world.

## III. BACKGROUND ON 3D GEOMETRY

### A. Gaussian Sphere

The geometric interpretation of the image lines and surface normals is performed on a Gaussian sphere, which is a virtual unit sphere centered at the optical center of the camera. We project a line in an image onto a Gaussian sphere as a great circle (the intersection of the Gaussian sphere and the plane defined by the center of projection (COP) and the line, see Fig. 2). The great circle of each line can be expressed as a unit normal vector (gray dots). We transform

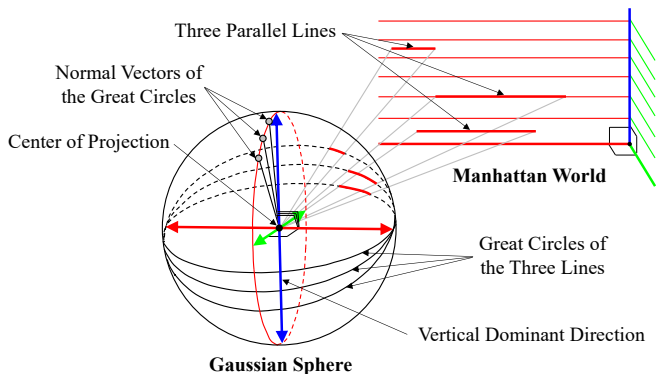


Fig. 2: Geometric relationships between the parallel lines and MW on the Gaussian sphere. We map the image lines onto the normal vectors of the great circles (gray dots). Each Manhattan direction and its corresponding line is drawn with the same color.

all image lines into the normal vectors of the great circles on a Gaussian sphere. The great circles from the parallel lines intersect at two antipodal points called a vanishing direction (VD) in the MW, which posits that every line and plane is perpendicular to one of the axes of a single coordinate system. We call this fixed coordinate system a Manhattan frame (MF). We summarize the abbreviations in Table I. The orthogonal surface normals of the MW planes exactly match the three orthogonal VDs defined by the parallel lines in an ideal MW, which are the basis of a MF. All normal vectors from the parallel lines pointing in the same direction should lie on the same great circle, and we utilize this geometric regularity to infer the MF in the proposed MnS approach.

### B. Rotational Motion with the Manhattan Frames

We represent the 3-DoF rotational motion of the camera and Manhattan frames as a  $3 \times 3$  rotation matrix  $R \in SO(3)$  in a Euclidean 3D space. We express the Manhattan frame (MF) with respect to the camera frame as:

$$R_{cM} = [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{r}_3] \in SO(3), \quad (1)$$

where each column  $\mathbf{r}_j$  denotes the  $x$ -,  $y$ -, and  $z$ -axes of the MF expressed in the camera frame. Our goal is to recognize the 3-DoF orientation of the MW for each  $k$ -th camera frame ( $R_{c_kM}$ ). Since the MW direction is fixed in a 3D space, we can track the 3-DoF rotational motion of the camera with respect to the fixed MF in a drift-free manner as follows:

$$R_{c_1c_k} = R_{c_1M}R_{c_kM}^{-1}, \quad (2)$$

where the notation follows a subscript cancellation rule. We assume that we can obtain the exact Manhattan frame directions from the initial camera frame.

## IV. PROPOSED METHOD

We propose a quasi-globally optimal and efficient rotational motion estimation method that requires a single vertical dominant direction (VDD) and at least a single image line following the MW directions for the horizontal dominant direction (HDD). The proposed method consists of three steps: 1) detection and tracking of the vertical dominant direction (2-DoF) from the depth images or IMU gravity

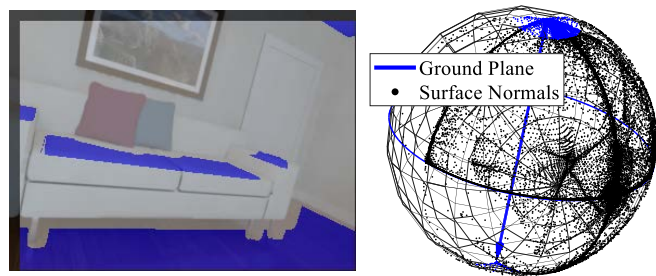


Fig. 3: Tracked VDD (blue axis) of the ground plane (blue) given the density distribution of the surface normals (black dots). We project the relevant surface normals inside a conic section of the VDD into the tangential plane to perform the mean shift [14].

direction, 2) a search for the optimal horizontal direction (1-DoF) with the lines from the RGB images, and 3) refining this initial rotation estimate with the inlier parallel lines. An overview of the proposed approach is shown in Fig. 4.

### A. Detection and Tracking of Dominant Direction

The proposed method requires only a single vertical dominant direction (VDD) such as the normal vector to the ground plane or gravity direction, which is a more practical condition for robotic applications compared to the conditions in [24], [23], which requires dense surface normals and the visibility of at least two orthogonal planes.

For structured environments with an RGB-D camera, we employ the dominant plane tracking approach in [14], in which we briefly summarize the pipeline (for full details, refer to [14]). We first detect a dominant plane<sup>1</sup> from a depth image using a three-point RANSAC. We track the detected dominant plane with a mean shift based on the tangent space Gaussian MF model, given the density distribution of the surface normals on the Gaussian sphere  $S^2$  [24] in Fig. 3.

We can alternatively detect and track the gravity direction as a VDD from an IMU where the use of the depth camera is infeasible due to the limited sensing range such as a drone flying over the city as shown in Fig. 4. We can estimate the gravity direction vector using Mahony and Madgwick filters [28] through raw IMU measurements, or obtain it directly from commercial platforms and libraries, e.g., Apple and Android devices. Any sensor, device, or algorithm that can detect and track the VDD can be used in combination with the proposed method.

### B. Manhattan Mine-and-Stab (MnS) Algorithm

We propose a new and efficient Manhattan MnS approach that effectively utilizes the periodicity of the MW to search for the optimal third DoF of the MF rotation, i.e., a single horizontal dominant direction (HDD) given the VDD from the previous section. In [10], the MnS is only used to find the horizontal dominant directions of the Atlanta world, and not for rotational motion tracking. Although the Atlanta MnS [10] can accurately find multiple VDs in the Atlanta world, it is unsuitable for real-time applications because it includes the computationally expensive BnB module.

<sup>1</sup>Without loss of generality, the dominant plane in any direction can be treated as a VDD in an indoor structured environment.

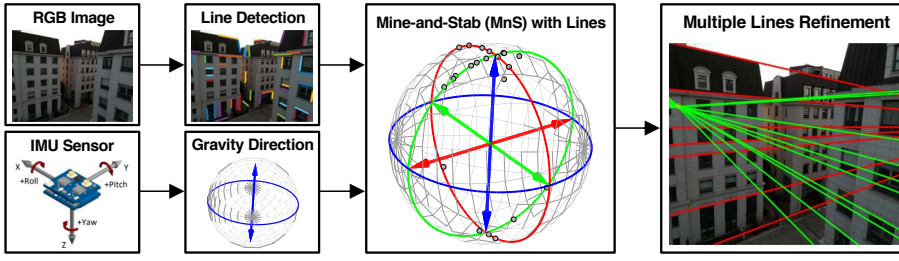


Fig. 4: We first detect/track the VDD (blue axis) to determine two orientation angles, employing the gravity direction vector from an IMU as a VDD. The proposed MnS method searches for the optimal 1-DoF horizontal direction (red and green axes) by effectively utilizing the periodicity of the MW, achieving quasi-global optimality. We refine the initial rotation estimate by minimizing the distance from the endpoints of the parallel and orthogonal lines.

TABLE I: List of Abbreviations

Abbreviation	Meaning
MW	Manhattan World
MF	Manhattan Frame
VD	Vanishing Direction
DD	Dominant Direction
VDD	Vertical DD
HDD	Horizontal DD
COP	Center of Projection
BnB	Branch-and-Bound
MnS	Mine-and-Stab

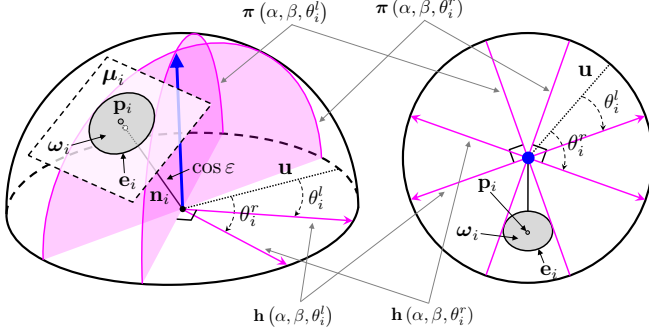


Fig. 5: 3D side view (left) and top orthographic view (right) of the Gaussian sphere with the tracked VDD (blue) and the projected  $i$ -th image line (gray dot). The sphere point  $\mathbf{p}_i$  (gray dot) lies on the horizontal dominant plane  $\pi$ . We expand  $\mathbf{p}_i$  into the spherical cap  $\omega_i$ , the candidate region to obtain the candidate interval  $[\theta_i^l, \theta_i^r]$ .

We briefly summarize the basic idea of the MnS (for full details, refer to [10]). In a one-dimensional space, all features treated as inliers within a specific range called the candidate interval are gathered as shown on the right side of Fig. 6. Given the candidate intervals mined above, our goal is to find an optimal probe that stabs as many candidate intervals as possible, i.e., maximizing the number of inliers.

1) *Parameterization and Candidate Region*: First, we parameterize the vertical dominant direction (VDD)  $\mathbf{v}$  with the azimuth  $\alpha \in [-\frac{\pi}{2}, \frac{\pi}{2}]$  and elevation  $\beta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ :

$$\mathbf{v}(\alpha, \beta) = [\cos \alpha \cos \beta, \sin \alpha \cos \beta, \sin \beta]^\top, \quad (3)$$

where  $\alpha$  and  $\beta$  denote the 2 DoFs of the MF rotation obtained in Section IV-A. We define a unit vector  $\mathbf{u} = [-\sin \alpha, \cos \alpha, 0]^\top$ , which is the basis reference vector representing the horizontal rotation orthogonal to the VDD. We parameterize the 1-DoF horizontal dominant direction (HDD)  $\mathbf{h}$  by rotating the basis vector  $\mathbf{u}$  around the VDD  $\mathbf{v}$  with an unknown-but-sought angle  $\theta \in [0, \frac{\pi}{2}]$  as follows:

$$\mathbf{h}(\alpha, \beta, \theta) = [[a_1, b_1] \mathbf{t}, [a_2, b_2] \mathbf{t}, [a_3, b_3] \mathbf{t}]^\top, \quad (4)$$

where  $\{a_i, b_i\}_{i=1}^3$  are a function of  $\alpha, \beta$ , and  $\mathbf{t}$  is a function of  $\theta$  (for full details and derivations, refer to [10] or the supplementary materials). Since we already know the VDD, we only need to find  $\theta$  using the distribution of the lines (gray dots) on the Gaussian sphere as shown in Fig. 5.

We parameterize the horizontal dominant plane  $\pi$  whose normal vector is the HDD  $\mathbf{h}(\alpha, \beta, \theta)$ , passing through the center of projection (COP)  $\mathbf{c}$ :

$$\pi(\alpha, \beta, \theta) : [x, y, z] \cdot \mathbf{h}(\alpha, \beta, \theta) = 0, \quad (5)$$

The normal vectors of the great circles from the parallel lines are orthogonal to the corresponding HDD  $\mathbf{h}$  as shown in Fig. 2. Thus, for a set of noise-free inlier image lines associated with the same HDD, the sphere point of the normal vector of the  $i$ -th image line on the Gaussian sphere  $\mathbf{p}_i$  should lie on the same horizontal dominant plane  $\pi$ . In practice, however, the sphere point  $\mathbf{p}_i$  cannot strictly lie on the horizontal dominant plane due to the inaccuracy and noise of the 2D line position in the image. To consider this error, we expand the sphere point  $\mathbf{p}_i$  into the spherical cap  $\omega_i$  on the Gaussian sphere, called the candidate region in Fig. 5. The geometric interpretation of the candidate interval  $[\theta_i^l, \theta_i^r]$  of the sphere point  $\mathbf{p}_i$  indicates that the horizontal dominant plane  $\pi$  intersects with the boundary of the candidate region, which is called the candidate region edge  $\mathbf{e}_i$ .

We define the 3D secant plane  $\mu_i$  to express the spherical cap  $\omega_i$  candidate region mathematically. The normal vector to the 3D secant plane  $\mu_i$  is  $\mathbf{n}_i$ , which is the normal vector of the great circle of the corresponding  $i$ -th image line. The distance between the 3D secant plane  $\mu_i$  and the sphere center (COP)  $\mathbf{c}$  is  $\cos \varepsilon$  where  $\varepsilon$  is a factor determining the size of the candidate region, defined by the user. We can find the edge  $\mathbf{e}_i$  of the candidate region  $\omega_i$  as the intersection of the secant plane  $\mu_i$  and the Gaussian sphere  $\mathbb{S}^2$  [10]. We redefine the inlier line if the candidate region  $\omega_i$  intersects with the horizontal dominant plane  $\pi$ . For each image line, we can compute and mine the candidate interval  $[\theta_i^l, \theta_i^r]$  based on the candidate region  $\omega_i$  as shown in Fig. 5. In the following, we introduce how we identify the horizontal inlier lines and how we leverage the periodicity of the MW in the proposed MnS to search for the unknown-but-sought angle  $\theta$  of the horizontal dominant plane  $\pi$ .

2) *Mining and Stabbing Candidate Intervals*: Given the candidate interval  $[\theta_i^l, \theta_i^r]$  of  $\omega_i$ , its range geometrically means the horizontal dominant plane  $\pi(\alpha, \beta, \theta)$  intersects with the candidate region edge  $\mathbf{e}_i$ . The quadratic equations have two distinct real solutions, which are the coordinates of two plane-edge intersections. We compute the real root of



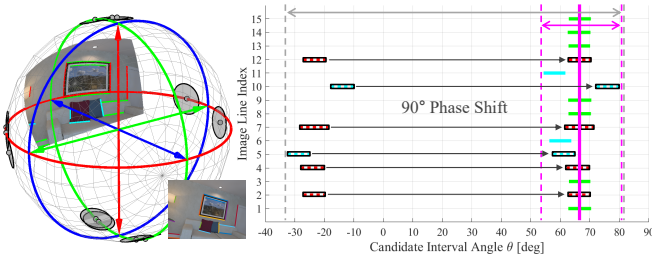


Fig. 6: Clustered image lines on the Gaussian sphere with the VDD – the blue axis (left) and corresponding candidate intervals  $[\theta^l, \theta^r]$  (right). By utilizing the  $90^\circ$  periodicity of the MW, we can reduce the search space of  $\theta$  from the gray to the magenta dotted lines. The proposed MnS can effectively find the optimal probe (magenta), which maximizes the number of stabbed intervals.

the polynomial using the SVD, and obtain the rotation range  $[\theta_i^l, \theta_i^r]$ , which correspond to the case in which the horizontal dominant plane is tangential to the candidate region edge as shown in Fig. 5. Our candidate interval computation leads to  $O(K)$  complexity where  $K$  is the number of lines. The proposed MnS utilizes the periodicity of the MW by shifting the  $90^\circ$  phase of the candidate interval values between  $-90^\circ$  and  $0^\circ$  as shown in Fig. 6. This accelerates our parameter search by reducing the search space of  $\theta$  from  $[-\frac{\pi}{2}, \frac{\pi}{2}]$  to  $[0, \frac{\pi}{2}]$ , resulting in “quasi-globally” optimal 3-DoF rotation estimate in real-time.

We mine  $K$  candidate intervals from the  $K$  image lines, and find the optimal probe, which stabs as many candidate intervals as possible, i.e., maximizing the number of horizontal inlier lines as shown in Fig. 6. We first sort all endpoints of the  $K$  candidate intervals in ascending order with the merge sort algorithm whose time complexity is  $O(K \log K)$ . We set the probe located at each endpoint, and sequentially scan the number of stabbed intervals in ascending order. We increase/decrease the number of stabbed intervals by one when the probe passes through a left/right endpoint, resulting in  $O(K)$  time complexity. Without loss of generality, we determine the optimal probe  $\theta^*$  passing through the median of the specific range with the maximum number of stabbed intervals as the representative (see Fig. 6). Given the optimal probe  $\theta^*$ , we can compute the optimal horizontal dominant direction  $\mathbf{h}(\alpha, \beta, \theta^*)$  using Eq. (4), achieving quasi-global optimality. The mining, endpoint sorting, and probe scanning of the  $K$  candidate intervals lead to a total complexity of  $O(K \log K)$ ; thus, the proposed MnS for the MW can run in polynomial time.

### C. Horizontal Inlier Lines based Nonlinear Optimization

The initial rotation estimate described in the previous section focuses on maximizing the number of inliers rather than minimizing the consistency error [13] of the inliers, resulting in a suboptimal 3-DoF MF rotation in terms of accuracy. To obtain a more accurate camera orientation, we further refine the initial rotation estimate by minimizing the average orthogonal distance using the inlier lines.

We define a cost function, which is a function of only the HDD  $\theta$  because the VDD is relatively accurate and smooth [14]. We express the 3-DoF rotational motion as the

axis-angle representation where the direction of the axis of rotation is the tracked vertical DD, and the magnitude of the rotation about the axis is the horizontal DD  $\theta$ . We can obtain an accurate, drift-free camera orientation by solving the following nonlinear optimization problem:

$$\theta^* = \arg \min_{\theta} \sum_{k=1}^2 \sum_{i=1}^{M_k} (d_{i,k}(\theta))^2, \quad (6)$$

where  $\{M_k\}_{k=1}^2$  is the number of parallel lines related to the  $k$ -th VD counted in the proposed MnS algorithm as inliers. In addition,  $d_{i,k}(\theta)$  denotes the distance of the  $i$ -th image line to the  $k$ -th VD in the image plane. We employ the Levenberg–Marquardt (LM) algorithm for solving Eq. (6). By additionally optimizing the horizontal DD  $\theta$  from the parallel and orthogonal inlier lines found in the proposed MnS, we can estimate a more accurate and consistent rotational motion compared to the initial rotation estimate.

## V. EVALUATION

We evaluate the proposed method on several datasets obtained from various sensor configurations in both indoor and outdoor structured environments:

- *York Urban Dataset* [18] (YUD) is composed of 102 calibrated RGB images captured in indoor and outdoor Manhattan environments. It is the de facto standard for evaluating rotational motion and line clustering because it provides a set of manually extracted lines and corresponding true MW labels.
- *ICL-NUIM Dataset* [17] is a synthetic RGB-D dataset captured in a living room and office with true 6-DoF camera poses. This dataset is useful for quantitatively evaluating the rotational motion because it contains a variety of image conditions and camera motions such as low texture and frequent on-the-spot rotations.
- *Tello Urban Dataset* is an author-collected dataset consisting of time-synchronized RGB images and gravity direction vectors from a DJI Tello drone through the ROS Tello driver, flying in an outdoor urban area.
- *iOS Logger Dataset* is an author-collected dataset containing an RGB image sequence, gravity direction vector, and camera poses of an Apple ARKit (VIO) from an iPhone 12 Pro Max with a custom iOS app.

We compare the proposed method against other state-of-the-art approaches including data sampling-based methods (OLRE [12] and SLRE [14]), parameter search-based methods (BnB [15]), and a combination of both (QBnB [9]). OLRE (or SLRE) retrieves the estimated MF rotation hypothesized by three sampled image lines (or a sampled single line and plane) from the RGB(D) images. BnB [15] searches for the optimal MF rotation over the rotation space with the image lines, achieving global optimality. QBnB [9] first computes the 2-DoF using two sampled image lines and then searches for the third DoF by BnB. All methods applied are implemented in MATLAB and tested on a desktop computer equipped with an Intel Core i7 (3.00 GHz) CPU and 16 GB of memory.

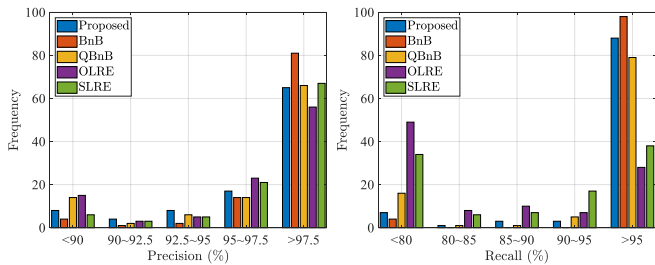


Fig. 7: Accuracy evaluation of the line clustering with ground-truth labels on 102 images of YUD [18]: precision (left) and recall (right).

TABLE II: Computational Time Analysis on York Urban Dataset

Module	Runtime
Candidate Interval Computation	1.716 ms
Optimal Probe Finding	0.300 ms
Nonlinear Optimization	3.938 ms
MF Direction Matching	0.025 ms

### A. York Urban Dataset

We evaluate the performance of the MF orientation estimation in terms of the precision and recall error metric, and the accuracy of the VDs in terms of the root mean square of the consistency error [13]. We compute the precision  $C/(C+W)$  and recall  $C/(C+M)$ , where  $C$ ,  $W$ , and  $M$  denote the number of correctly identified, wrongly identified, and missing inliers, respectively. The consistency error represents the orthogonal distance on the image plane in pixels from an endpoint of the image line  $l$  to a virtual line  $\hat{l}$  defined by the midpoint of  $l$  and an estimated VP.

The first row of Fig. 8 shows a representative evaluation result of the proposed method compared with other approaches in terms of precision, recall, consistency error, and run time. While the data sampling-based methods (OLRE and SLRE) show a fast computation time and good consistency error, they fail to guarantee precision and recall simultaneously. BnB can obtain all inlier lines, but it takes more than three seconds per image, which is unsuitable for real-time applications. Although QBnB can compute the MF rotations within 0.1 seconds, the existing parameter search-based approaches aim to maximize the number of inliers, and not minimize the consistency error. The proposed method can obtain the second-lowest consistency error similar to SLRE, and achieve 100% precision and recall accuracy, similar to that of BnB, simultaneously in real-time at 100 Hz.

We analyze the computational time of the key components of the proposed method given the VDD and image lines in Table II. First, it takes  $\sim 1.7$  ms to compute the candidate intervals  $[\theta_i^l, \theta_i^r]$  for the ten image lines. We sort all the endpoints of the ten candidate intervals, and find the optimal probe, taking  $\sim 0.3$  ms. Horizontal inlier line-based nonlinear optimization with the Levenberg–Marquardt (LM) takes  $\sim 4.0$  ms, and the other minor computations such as MF direction matching take  $\sim 0.03$  ms. Overall, the total computation time of the proposed method is approximately 6–14 ms per image depending on the number of lines.

We report the precision and recall accuracy of various

TABLE III: Evaluation Results on ICL-NUIM Dataset

Experiment	Proposed	OLRE [12]	SLRE [14]	OPRE [30]	Length (deg)
Living Room 0	<b>0.24</b>	×	0.31	×	705.10
Living Room 1	0.38	3.72	<b>0.38</b>	0.97	434.41
Living Room 2	<b>0.29</b>	4.21	0.34	0.49	475.36
Living Room 3	<b>0.29</b>	×	0.35	1.34	483.75
Office Room 0	0.28	6.71	0.37	<b>0.18</b>	704.96
Office Room 1	<b>0.17</b>	×	0.37	0.32	434.41
Office Room 2	<b>0.19</b>	10.91	0.38	0.33	475.35
Office Room 3	0.26	3.41	0.38	<b>0.21</b>	483.75

methods for all 102 images of the YUD in Fig. 7. OLRE and SLRE show a reasonable precision, but there are many missing inlier lines, resulting in a low recall accuracy. Most data sampling-based methods including OLRE and SLRE rely on RANSAC with randomness and uncertainty, which sometimes fail to estimate the MF rotations correctly. By contrast, our proposed method is less sensitive to noise than RANSAC, and achieves a high precision and recall accuracy similar to the parameter search-based BnB and QBnB approaches while working efficiently. The proposed method is the fastest and most stable among the top three algorithms with high accuracy.

### B. ICL-NUIM Dataset

We measure the mean value of the absolute rotation error (ARE) [30] in degrees, and report the quantitative evaluation results of various methods in Table III. The smallest rotation error for each dataset is indicated in bold. We exclude parameter search-based methods such as BnB and QBnB because they are unsuitable for this type of continuous 3-DoF rotational motion estimation. We add the result of OPRE [30], which is an orthogonal plane-based tracking.

Some of the methods such as OLRE and OPRE depending on multiple lines or planes sometimes fail to track the MF rotations (marked as  $\times$  in Table III) because multiple lines or orthogonal planes are not always visible throughout the entire image sequence. Particularly in ‘Living Room 0’, at one point the camera observes only a single line and plane with extremely low texture, which leads to a failure of other approaches. Theoretically, the proposed method only requires at least a single plane for a VDD and a single line for an HDD to estimate the MF rotations. The proposed method can keep tracking the absolute 3-DoF camera orientation stably and accurately for all image sequences as shown in Fig. 9. Empirically, the proposed method can reliably track the 3-DoF rotational motion if there are approximately five image lines satisfying the MW given the VDD.

The second row of Fig. 8 shows a representative result with precision, recall, consistency error, and run time. Our approach outperforms the parameter search-based methods, and achieves almost a similar absolute rotation error compared to data-sampling methods such as SLRE. The average ARE of the proposed method is 0.26 degrees, whereas OLRE, SLRE, and OPRE are 5.79, 0.36, and 0.55 degrees, respectively. The proposed method can stably track the absolute MF rotations even when the camera sees only a single planar surface with little texture by exploiting the theoretical minimal sampling, a single line and a single plane.


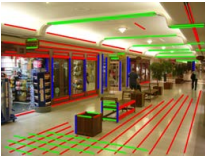





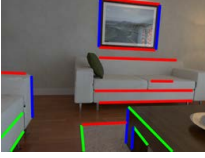
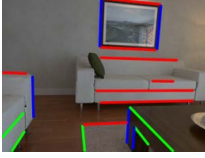
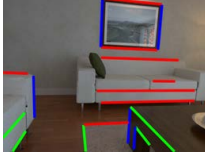
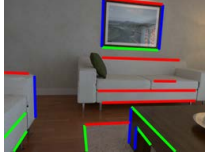
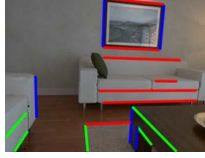

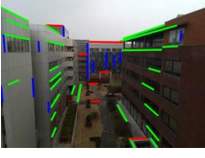
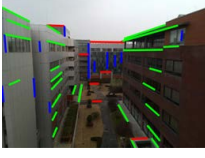
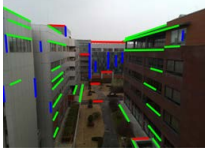
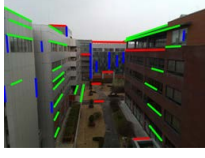
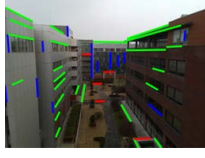
Lines	Proposed	BnB [15]	QBnB [9]	OLRE [12]	SLRE [14]
 York Urban [20] 92 Lines	 100%, 100% 0.48 pix., 0.013 sec.	 100%, 100% 1.68 pix., 6.412 sec.	 100%, 97.83% 3.18 pix., 0.094 sec.	 98.89%, 97.80% 1.88 pix., 0.008 sec.	 100%, 97.83% 0.27 pix., 0.009 sec.
 ICL-NUIM [19] 18 Lines	 100%, 100% 0.20 pix., 0.014 sec.	 100%, 100% 1.55 pix., 5.543 sec.	 100%, 100% 0.68 pix., 0.024 sec.	 94.44%, 100% 1.12 pix., 0.007 sec.	 100%, 94.44% 0.11 pix., 0.008 sec.
 DJI Tello Urban 44 Lines	 100%, 100% 0.85 pix., 0.012 sec.	 100%, 100% 5.07 pix., 2.372 sec.	 100%, 100% 1.87 pix., 0.043 sec.	 97.56%, 93.02% 2.91 pix., 0.008 sec.	 95.24%, 95.24% 0.33 pix., 0.006 sec.

Fig. 8: Representative evaluations on York Urban [18], ICL-NUIM [17], and our Tello Urban datasets. Each row represents a tested dataset, and each column denotes an evaluated algorithm. In the first row, we utilize the manually extracted lines. We extract the line segments with LSD [29] in the last two rows. The numbers below the images are the precision, recall, consistency error, and run time.

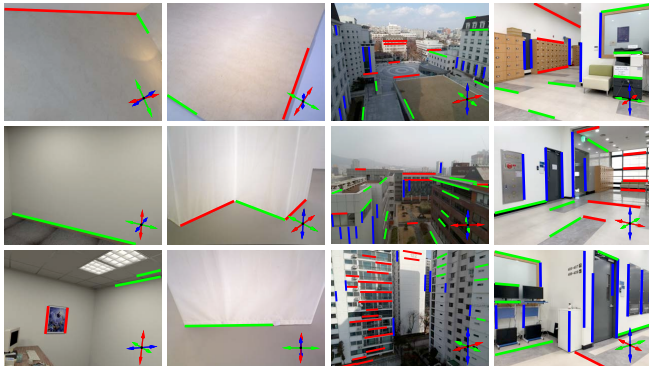


Fig. 9: Representative example images in various indoor and outdoor urban environments. Each column denotes the ICL-NUIM [17], TUM-RGBD [31], Tello urban, and iOS logger datasets. We cluster a set of image lines satisfying the inferred MW shown in the lower-right corner. The proposed method can stably track the MW orientations regardless of the amount of texture.

### C. Tello Urban Dataset

We evaluate the proposed method on real-world data from a DJI Tello drone flying in outdoor urban MW environments as shown in the third row of Fig. 8. We utilize the gravity direction from the IMU as a VDD, and search for the optimal third DoF of HDD with the image lines extracted using LSD [29]. Existing approaches relying on a depth camera cannot operate well on a drone flying over an open area due to the limited sensing range. The proposed method can stably track the absolute 3-DoF camera orientations in such a challenging outdoor flight environment as shown in Fig. 9, showing that it can operate as a drift-free visual-inertial compass for yaw angle correction.

TABLE IV: Computational Time Comparison on Tello Dataset

	Proposed	BnB	QBnB	OLRE	SLRE
Time (s)	0.011	5.461	0.264	0.008	0.009

We report the average run times of various methods in Table IV. The data sampling-based approaches (OLRE and SLRE) are efficient at the cost of sacrificing accuracy. Our method is significantly faster than BnB and is similar to the RANSAC-based approaches. The proposed method in MATLAB can run at 100~150 Hz, suggesting its potential when implemented in C/C++ for low-cost drones and various robotic applications with low computational power.

### D. iOS Logger Dataset

We evaluate the rotational accuracy of the proposed method with the Apple ARKit, one of the most accurate and reliable commercial VIO solutions. We set the camera orientation at the start and end points to be the same, and rotate the camera in place 16 times without any translational motion to quantitatively check the final rotation drift error (see Fig. 10). The rotation drift error of the ARKit gradually accumulates as the on-the-spot rotation continues, and finally a severe rotation drift of  $45^\circ$  is experienced. The translation estimation of the ARKit is also inaccurate on the right side of Fig. 10. The proposed method demonstrates extremely accurate and drift-free rotational motion estimation results of less than  $1^\circ$  even after 16 rotations while in place.

Please refer to the video clips and supplementary materials showing more details and additional 3-DoF complex roll, pitch, and yaw rotation experiments.



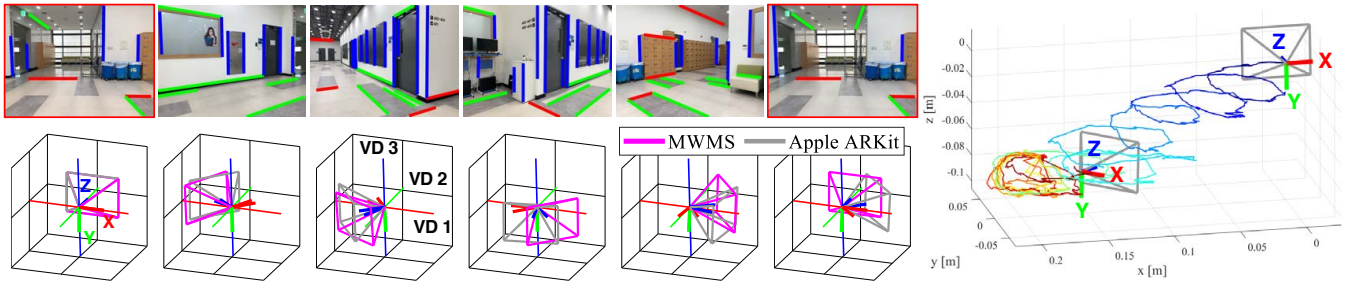


Fig. 10: Apple ARKit (gray), the proprietary visual-inertial odometry (VIO), shows a severe rotational drift error at the end (about  $45^\circ$ ) when rotating the camera in place 16 times, resulting in a degradation of the overall 6-DoF VIO motion estimation (right). The proposed Manhattan world max-stabbing (MWMS) can estimate the absolute 3-DoF camera orientation (magenta) with respect to the VDs by tracking the Manhattan patterns (left), ultimately demonstrating a 3-DoF rotational motion estimation error of less than  $1^\circ$ .

## VI. CONCLUSION

We propose a quasi-globally optimal and efficient MF rotation computation approach to estimate the absolute 3-DoF camera orientation with respect to the Manhattan world. We first detect and track the vertical dominant direction from an RGB-D camera or an IMU to compute the 2-DoF of the MF rotation, and then search for the optimal third DoF with the proposed Manhattan MnS, which effectively utilizes the periodicity of the Manhattan structure. Our sampling of the vertical dominant direction speeds up our parameter search in the horizontal direction by reducing the search space in 1-DoF. Our method is insensitive to noise and can achieve quasi-global optimality in real-time. Experiments show that the proposed method outperforms the state-of-the-art approaches in terms of accuracy, efficiency, and stability.

## ACKNOWLEDGEMENTS

Pyojin Kim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1061397). Kyungdon Joo was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01336, Artificial Intelligence Graduate School Program (UNIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1C1C1005723).

## REFERENCES

- [1] Y. Gao and A. L. Yuille, "Exploiting symmetry and/or manhattan properties for 3d object structure estimation from single and multiple images," in *CVPR*, 2017.
- [2] N. Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *ECCV*, 2012.
- [3] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3D SLAM: a survey on rotation estimation and its use in pose graph optimization," in *ICRA*, 2015.
- [4] P. Kim, B. Coltin, and H. J. Kim, "Linear rgb-d slam for planar environments," in *ECCV*, 2018.
- [5] Y. Li, J. Mao, B. Freeman, J. Tenenbaum, N. Snavely, and J. Wu, "Multi-plane program induction with 3d box priors," in *NeurIPS*, 2020.
- [6] R. C. Leishman, J. C. Macdonald, R. W. Beard, and T. W. McLain, "Quadrotors and accelerometers: State estimation with an improved dynamic model," *IEEE Control Systems Magazine*, 2014.
- [7] Y. Zhou, L. Kneip, and H. Li, "Real-time rotation estimation for dense depth sensors in piece-wise planar environments," in *IROS*, 2016.
- [8] J. Straub, N. Bhandari, J. J. Leonard, and J. W. Fisher, "Real-time Manhattan world rotation estimation in 3D," in *IROS*, 2015.
- [9] H. Li, J.-C. Bazin, and Y.-H. Liu, "Quasi-globally optimal and near/true real-time vanishing point estimation in manhattan world," *T-PAMI*, 2020.
- [10] H. Li, P. Kim, K. Joo, Z. Liu, and Y.-H. Liu, "Globally optimal and efficient vanishing point estimation in atlanta world," in *ECCV*, 2020.
- [11] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *ICCV*, 1999.
- [12] J.-C. Bazin and M. Pollefeys, "3-line RANSAC for orthogonal vanishing point detection," in *IROS*, 2012.
- [13] L. Zhang, H. Lu, X. Hu, and R. Koch, "Vanishing point estimation and line classification in a manhattan world with a unifying camera model," *IJCV*, 2016.
- [14] P. Kim, B. Coltin, and H. J. Kim, "Indoor rgb-d compass from a single line and plane," in *CVPR*, 2018.
- [15] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in manhattan world," in *CVPR*, 2012.
- [16] K. Joo, T.-H. Oh, J. Kim, and I. S. Kweon, "Robust and globally optimal manhattan frame estimation in near real time," *T-PAMI*, 2019.
- [17] A. Handa, T. Whelan, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *ICRA*, 2014.
- [18] P. Denis, J. H. Elder, and F. J. Estrada, "Efficient edge-based methods for estimating manhattan frames in urban imagery," in *ECCV*, 2008.
- [19] G. Schindler and F. Dellaert, "Atlanta world: An expectation maximization for simultaneous low-level edge grouping and camera calibration in complex man-made environments," in *CVPR*, 2004.
- [20] J.-C. Bazin, C. Demonceaux, P. Vasseur, and I. Kweon, "Rotation estimation and vanishing point extraction by omnidirectional vision in urban environment," *IJRR*, 2012.
- [21] R. Toldo and A. Fusiello, "Robust multiple structures estimation with j-linkage," in *ECCV*, 2008.
- [22] J.-P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *ICCV*, 2009.
- [23] J. Straub, O. Freifeld, G. Rosman, J. J. Leonard, and J. W. Fisher, "The Manhattan frame model—Manhattan world inference in the space of surface normals," *T-PAMI*, 2017.
- [24] Y. Zhou, L. Kneip, C. Rodriguez, and H. Li, "Divide and conquer: Efficient density-based tracking of 3D sensors in Manhattan worlds," in *ACCV*, 2016.
- [25] P. Kim, B. Coltin, and H. J. Kim, "Low-drift visual odometry in structured environments by decoupling rotational and translational motion," in *ICRA*, 2018.
- [26] J.-C. Bazin, Y. Seo, and M. Pollefeys, "Globally optimal consensus set maximization through rotation search," in *ACCV*, 2012.
- [27] H. Li, J. Zhao, J.-C. Bazin, W. Chen, Z. Liu, and Y.-H. Liu, "Quasi-globally optimal and efficient vanishing point estimation in manhattan world," in *ICCV*, 2019.
- [28] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol (UK)*, vol. 25, pp. 113–118, 2010.
- [29] R. G. Von Gioi, J.-M. Jakubowicz, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *T-PAMI*, 2008.
- [30] P. Kim, B. Coltin, and H. J. Kim, "Visual odometry with drift-free rotation estimation using indoor scene regularities," in *BMVC*, 2017.
- [31] J. Sturm, N. Engelhard, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *IROS*, 2012.