# 3D Model-Free Visual Localization System From Essential Matrix Under Local Planar Motion

Yanmei Jiao, Binxin Zhang, Peng Jiang, *Member, IEEE*, Chaoqun Wang, *Member, IEEE*, Haojian Lu, *Member, IEEE*, Rong Xiong, *Senior Member, IEEE*, and Yue Wang, *Member, IEEE*

***Abstract*— Visual localization plays a critical role in the functionality of low-cost autonomous mobile robots. Contemporary leading methods for precise visual localization are predominantly 3D scene-specific, necessitating extra computational and memory overhead to construct a 3D scene model in novel environments. An alternative approach of directly using a database of 2D images for visual localization offers more flexibility. However, such methods currently suffer from limited localization accuracy. In this paper, we propose an accurate and robust multiple checking-based 3D model-free visual localization system to address the aforementioned issues. To ensure high accuracy, our focus is on estimating the pose of a query image relative to the retrieved database images using 2D-2D feature matches. Theoretically, by incorporating the local planar motion constraint into both the estimation of the essential matrix and the triangulation stages, we reduce the minimum required feature matches for absolute pose estimation, thereby enhancing the robustness of outlier rejection. Additionally, we introduce a multiple-checking mechanism to ensure the correctness of the solution throughout the solving process. The efficacy of our approach is substantiated through both qualitative and quantitative assessments on simulated and two real-world datasets evidencing significant improvements in accuracy and robustness provided by our 3D model-free visual localization system.**

***Note to Practitioners*—The motivation of this article stems from the need to develop an accurate visual localization system with simplicity and flexibility of map construction and easy adaption to new environments. Such a system holds great practical value for a range of applications, including warehouse robots, service robots, and countless others. Existing visual localization systems that achieve high accuracy are dependent on a pre-built accurate 3D scene map, which pose challenges in terms of map construction and consume significant storage resources onboard, particularly for large scenes. And the aforementioned efforts need to be repeated when changing to a new scene. In this article, an accurate and robust 3D model-free visual localization system is proposed to handle this problem. The map construction is simplified to build a set of database images with associated camera poses, which is trivial as it amounts to adding posed images to a database. The core idea for achieving high accuracy and robustness is to model the local planar motion characteristic of general ground-moving robots into both essential matrix estimation and triangulation stages to obtain two minimal solutions. The proposed localization system simplifies the task of switching between different application scenarios for the robot, reducing additional workload and lowering the difficulty of use.**

***Index Terms*— Minimal solutions, essential matrix, local planar motion, visual localization.**

Manuscript received 11 January 2024; accepted 5 March 2024. This article was recommended for publication by Associate Editor W. Zhang and Editor P. Rocco upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62373322, in part by Zhejiang Provincial Natural Science Foundation of China under Grant LQ24F030009, and in part by "Ling-Yan" Research and Development Project of Zhejiang Province under Grant 2023C01176. *(Corresponding author: Yue Wang.)*

Yanmei Jiao, Binxin Zhang, and Peng Jiang are with the School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China.

Chaoqun Wang is with the School of Control Science and Engineering, Shandong University, Shandong 250100, China.

Haojian Lu, Rong Xiong, and Yue Wang are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou 310058, China (e-mail: ywang24@zju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2024.3374823.

Digital Object Identifier 10.1109/TASE.2024.3374823

## I. Introduction

GIVEN a query image, visual localization refers to the problem of estimating the pose, *i.e.*, position and orientation of the camera using visual information with respect to a environment. Visual localization is a core functionality for low-cost positioning in autonomous unmanned systems, which has important practical applications in various scenarios such as navigation, exploration, surveillance, and rescue [1], [2], [3], [4], [5], [6], [7].

The most accurate visual localization methods currently available are based on 3D information. The 3D scene information can be either explicit point cloud maps constructed by SfM (Structure-from-Motion) methods [8], or implicit scene encoded by neural networks [9], [10]. The general pipeline of 3D scene based visual localization approaches is to first obtain the 2D-3D feature matching between a query image and a pre-built 3D scene [11], [12], and then perform accurate localization based on absolute pose estimation algorithms [13], [14], [15]. These methods solve for high pose accuracy, but they are scene-specific, *i.e.*, they require extra time and computational resources to build a high precision 3D map when facing a new scene.

A more flexible way of visual localization is to localize directly based on a 2D scene database which is a set of reference images with associated camera poses. Currently, this line of method focuses on image retrieval or place recognition, which retrieves the most similar database image and approximates the pose of query image using the associated pose in database [16], [17]. However, the localization accuracy of this method is limited by the sampling interval when constructing the database, which is difficult to meet the requirements of tasks that require higher localization accuracy. To further improve the localization accuracy, the pose transformation between the query image and the database image can be

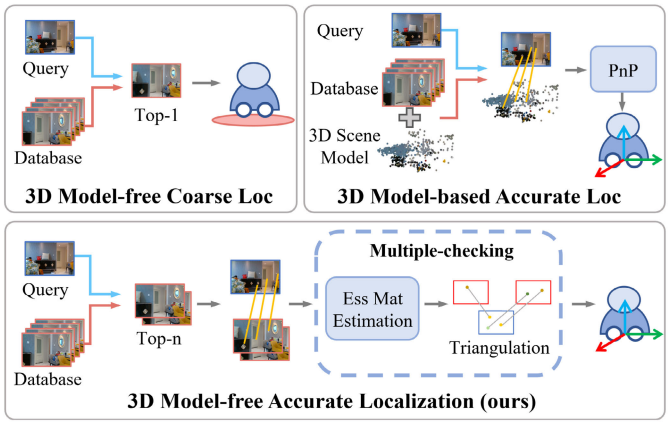



Fig. 1. Comparison of different problem settings of visual localization with or without a 3D scene model. The proposed 3D model-free visual localization system achieves accurate and robust localization directly based on a 2D image database, which (1) differentiates itself from a 3D model-based pipeline by eliminating the need for prior construction of a 3D environmental model, (2) achieves higher accuracy compared to typical 3D model-free method using the pose of retrieved database image for approximation, (3) leverages a multiple-checking procedure to ensure the correctness and robustness of essential matrix estimation and triangulation for absolute pose estimation.

estimated from 2D-2D feature matches, which is the focus of this paper.

There are two problems with the existing 2D-2D pose estimation methods for visual localization. First, these methods mostly perform relative pose estimation between two images based on essential matrix estimation, which provide the rotation matrix and the relative translation direction but not the absolute translation scale [18], [19], [20], [21]. Therefore, they cannot obtain the absolute pose of the camera with respect to the global coordinate system of the database, which is essential for complete visual localization. Second, most of them formulate the relative pose estimation as a 6DoF (degree-of-freedom) problem, which ignores the prior kinematic constraints in robotics. Mobile robots and ground vehicles usually operate on planar or at least locally planar surfaces such as floors in indoor environments and roads in outdoor settings. The planar motion constraint can simplify the 6DoF relative pose problem to a 3DoF one, which leads to less computation and higher accuracy.

In this paper, we propose two minimal solutions for absolute pose estimation from essential matrix and a multiple checking scheme for robust 3D model-free visual localization, addressing the above problems. The proposed minimal solutions are derived from simplified epipolar geometry that leverages local planar motion constraint. Specifically, based on the simplified epipolar constraint, only two feature point matches are required for essential matrix estimation. Then, a 2p2p minimal solution is naturally proposed, which estimates the absolute pose by triangulating the translation vector obtained from two sets of point correspondences provided by two reference images in the database. However, the 2p2p minimal solution still has redundancy in the constraints, and the number of feature matches required can be further reduced by considering the motion property into triangulation. Therefore, we propose a 2p1p minimal solution that only needs one additional feature match from the second reference image to uniquely determine the absolute translation scale and complete the absolute pose estimation. The proposed minimal solutions are embedded with a multiple checking scheme to perform a robust pose estimation. Note that reducing the number of necessary correspondences for pose estimation is meaningful, as it leads to higher success rate with fixed number of iterations in RANSAC, which means higher robustness of visual localization. In addition, by incorporating affine-invariant feature matching from our previous work [6], which has been reformulated with local planar motion property, a complete and robust 3D model-free visual localization system is present. Extensive simulation and real world experiments are conducted to demonstrate its effectiveness. In summary, the contributions of this paper are listed as follows:

- We derive two minimal solutions namely 2p2p and 2p1p for absolute pose estimation using 2D-2D matches and propose a multiple checking-based mechanism to perform outlier rejection in the early stage.
- We conduct a theoretical analysis of the minimal solutions from the aspects on lower bound of estimation accuracy, stability, robustness and sensitivity, which can guide practical applications.
- We propose a comprehensive 3D model-free visual localization system that combines the affine-invariant feature matching and multiple checking-based pose estimation, and the source code is available on github.[1]
- We conduct extensive experiments on synthetic data and real world datasets to validate the accuracy and robustness of the proposed system for visual localization of ground moving robots.

## II. Related Work

### A. 3D Model-Based Visual Localization

Generally, there are two lines of 3D model-based visual localization approaches. The first is direct approaches, which directly generate 2D-3D feature correspondences from a query image and a 3D scene model [22], [23], [24]. The typical methods match feature descriptors extracted from a 2D query image and 3D points in a SfM model and then produce accurate camera pose estimation from the matches. While achieving high accuracy, these methods struggle in scaling to large scenes due to memory consumption or arising ambiguities [25]. Recently, inspired by the success of deep learning in image understanding, several works adopt deep neural networks to directly learn 2D-3D feature matching function [26], [27] or regress camera pose by using 2D-3D matches as part of the loss function [28]. These methods utilize CNNs (convolutional neural network) to autonomously discover useful association information and can be performed end-to-end. However, their scalability to larger scenes is also limited and their generalization ability to new scenes has to be improved. The second is indirect approaches [29], [30], which follow a hierarchical localization pipeline by using image retrieval methods to retrieve a similar database image and match 2D-2D feature descriptors between

[1] https://github.com/slinkle/3D-Model-free-minimal-solutions

query image and database image. The 2D-3D matches are then obtained by retrieving the associated 3D points information of the database image stored in the scene model. Similar to direct methods, these methods are scene-specific which require extra computational resources and efforts to build a high precision 3D map when facing a new scene. While the proposed method in this paper aims to achieve accurate and robust visual localization without a pre-built 3D scene model.

### B. 3D Model-Free Visual Localization

Visual localization without 3D scene models refers to recovering the pose of a query image from a database of 2D images associated with pose information. It generally includes three steps: image retrieval, relative pose estimation, and absolute pose estimation. Image retrieval, also known as place recognition [16], [31], [32], [33], is the same as the first step of the indirect 3D model-based methods, which is based on image-level descriptors to extract top-$n$ similar reference images from the database.

For each retrieved reference image, the subsequent step involves computing the essential matrix through relative pose estimation algorithms. Relative pose estimation between two images is one of the classical geometric problems in computer vision. For the general 6DoF estimation problem, 8-point algorithm is the most common and standard method [18], [21]. And the minimal solver is 5-point algorithm which performs pose estimation using the least number of correspondences [19], [20]. By considering motion property in robotics, numerous efforts aim to reduce the number of features to solve the simplified relative pose estimation problems. As mobile robots and vehicles usually move on floors or roads, the problem can be reduced to 3DoF. Then 3-point algorithm formulated by a linear equation and the minimal solver based on iterative 2-point algorithm are proposed [34]. Then the non-iterative 2-point algorithm formulated by finding intersections of two ellipses is proposed [35] and be improved with less computation in [36]. With the assumption of planar and circular motion, the 1-point minimal solver is proposed using only a single correspondence [37]. However, the above algorithms only estimates the essential matrix between two images, which can be used to recover the rotation matrix and translation direction, but not the translation scale. The proposed minimal solutions in this paper aim to solve the absolute pose between the query image and the 2D image database, which can be directly used for visual localization.

The final step of absolute pose estimation usually involves triangulation from two translation directions obtained from two different reference images [18], [38]. However, the proposed method incorporates the local planar motion property not only into the essential matrix estimation, but also into the triangulation process in a creative way, thereby further reducing the number of required features, as shown in Fig. 4. To the best of our knowledge, this is the first minimal solution that models the local planar motion characteristics into triangulation. In addition, we propose a multiple checking procedure to ensure the robustness of the proposed visual localization method.

## III. Overview

We first clarify the definition of the visual localization problem studied in this paper. Given a set of reference images previously collected in a certain environment, these reference images are annotated with 6DoF poses expressed in a global reference coordinate system. The absolute localization problem addressed in this paper refers to obtaining the 6DoF pose of a robot relative to the global reference coordinate system when the robot re-enters this environment, based on a 2D query image acquired by the robot. The pipeline of the proposed 3D model-free visual localization system under local planar motion is depicted in Fig. 2. Initially, image retrieval is conducted to identify a set of reference images that correspond to the same location as the query image in the database. Subsequently, augmentation-based feature matching is performed to establish correspondences between the query image and each retrieved reference image (*c.f.* Sec. III-B). During this stage, the local planar motion property (*c.f.* Sec. III-A) is exploited to relax the affine distortion formulation, thereby achieving affine invariant feature matching. This approach significantly enhances the robustness of feature matching against significant viewpoint variations, resulting in a considerable increase in the number of inlier matches.

Using the established correspondences, the essential matrix is computed for each pair of query and reference images, encoding the relative pose transformation (*c.f.* Sec. IV-A). The relative poses that pass the cheirality test and Rcheck are then fed into the triangulation stage to estimate the absolute pose (*c.f.* Sec. IV-B and Sec. IV-C). Following the Pdcheck and consistency check, we perform 6DoF nonlinear optimization to refine the estimated localization result using identified inliers. To enhance the robustness, the local planar motion property is incorporated into both the essential matrix estimation and triangulation stages. This integration aims to reduce the minimum number of correspondences required for accurate relative and absolute pose estimation. By leveraging the inherent local planar motion property, both feature matching and pose estimation become more resilient to noise and outliers, thereby improving overall performance and robustness of the proposed 3D model-free visual localization system.

### A. Local Planar Motion Property

Ground mobile robots basically have an inherent motion property, *i.e.*, a local planar motion property. As shown in Fig. 3, there are only a $y$-axis rotation and 2D translation between query and reference camera views. Note that the coordinate assumption of $y$-axis pointing downwards is consistent with the general camera coordinate and can be easily satisfied by applying the appropriate rotation transformation to construct a virtual camera coordinate system of which the image plane is perpendicular to the ground. Some details are shown in Appendix-A. Then the camera pose transformation from query view $i$ to reference view $j$ can be represented as

$$\mathbf{R} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}, \quad \mathbf{t} = -\mathbf{R}\rho \begin{bmatrix} \sin\phi \\ 0 \\ \cos\phi \end{bmatrix}, \tag{1}$$

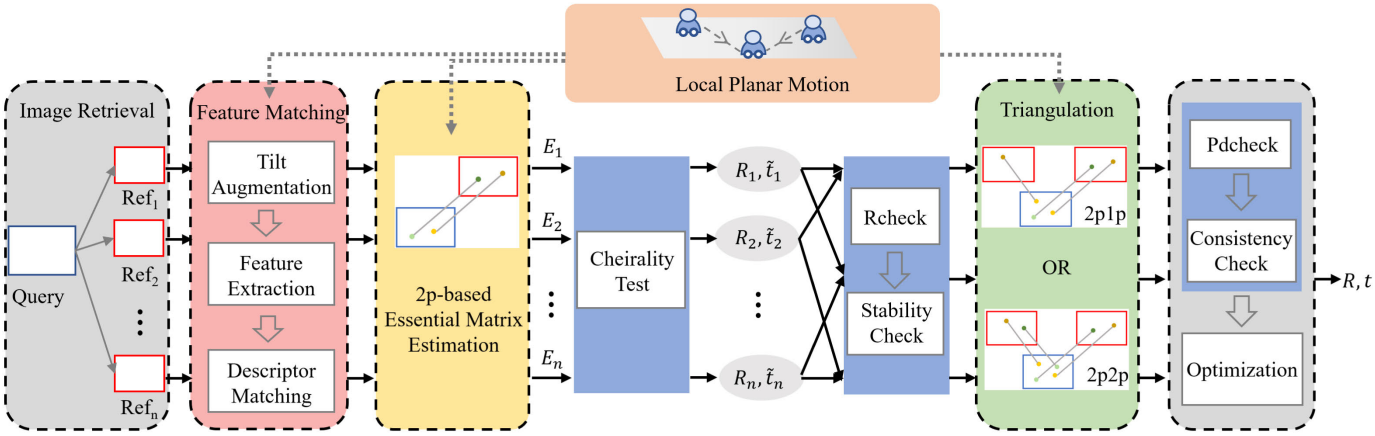



Fig. 2. The pipeline of the proposed 3D model-free visual localization system. The local planar motion property is introduced into feature matching, essential matrix estimation and triangulation stages, which improve overall performance and robustness of the proposed visual localization system. A multiple checking procedure is also introduced to enhance the correctness of the pose estimation.

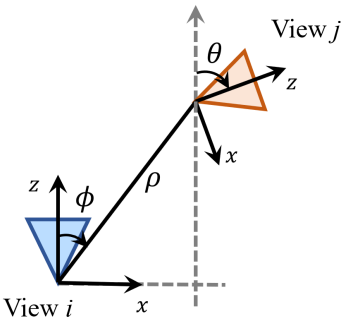



Fig. 3. The illustration of local planar motion between query and reference views. There are three unknowns: yaw angle $\theta$, translation direction $\phi$ and scale $\rho$ in absolute pose estimation problem.

where $\theta$ denotes the yaw angle around axis $y$, and $\rho$, $\phi$ present the scale and direction of 2D translation on the $xz$ polar coordinate. Then the essential matrix $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ under local planar motion is reformulated as:

$$\mathbf{E} = \rho \begin{bmatrix} 0 & \cos(\theta - \phi) & 0 \\ -\cos\phi & 0 & \sin\phi \\ 0 & \sin(\theta - \phi) & 0 \end{bmatrix}, \tag{2}$$

where $[\cdot]_\times$ denotes the skew symmetry matrix. Therefore, it is imperative to construct a dedicated visual localization system that reformulates feature matching and pose estimation based on the local planar motion property, rather than persisting with a general system designed for the 6 DoF problem.

### B. Augmentation-Based Feature Matching

For robust feature matching, we adopt TiltR2D2 [39] in our previous work to perform affine-invariant feature matching. The local planar motion property is formulated into projective deformation between query and reference images, leading to a simplified affine distortion model. Specifically, there are only a transition tilt which represents a longitudinal stretch of the image and a scale transformation between two views under local planar motion. Then we design a tilt augmentation strategy to achieve tilt-invariant which can be then integrated with the existing scale-invariant descriptors to achieve full affine invariance. For details about TiltR2D2, please refer to [39].

## IV. Minimal Solutions From Essential Matrix

For robust pose estimation, two minimal solutions from essential matrix are derived to perform robust 3D model-free visual localization. The key to reduce the number of feature correspondences for pose estimation is to leverage local planar motion property of wheeled mobile robots. Then we derive one minimal solution namely 2p2p which uses only 2 point correspondences for essential matrix estimation in Sec. IV-A and another 2 correspondences for absolute pose triangulation in Sec. IV-B. And the minimal solution namely 2p1p is derived in Sec. IV-C, which only requires one additional correspondence for absolute scale estimation.

### A. 2p-Based Essential Matrix Estimation

Given one 2D-2D feature correspondence between query view and reference view expressed as $p_i$ in query view and $p_j$ in reference view, the epipolar constraint is given as follows:

$$p_j^{\mathrm{T}} \mathbf{E} p_i = 0, \tag{3}$$

where $p_i = [\tilde{u}_i, \tilde{v}_i, 1]^{\mathrm{T}} = K^{-1}[u_i, v_i, 1]^{\mathrm{T}}$ and $p_j = [\tilde{u}_j, \tilde{v}_j, 1]^{\mathrm{T}} = K^{-1}[u_j, v_j, 1]^{\mathrm{T}}$ are the normalized homogeneous coordinates of a feature point in view $i$ and $j$, respectively. $K$ is the calibrated intrinsic matrix.

By substituting (2) into (3), the epipolar constraint under local planar motion can be reformulated as:

$$\tilde{v}_i \sin(\theta - \phi) + \tilde{v}_i \tilde{u}_j \cos(\theta - \phi) + \tilde{v}_j \sin\phi - \tilde{u}_i \tilde{v}_j \cos\phi = 0. \tag{4}$$

Given another 2D-2D feature correspondence denoted as $p_{i_2}$ in query view and $p_{j_2}$ in reference view, another constraint can be obtained as:

$$\begin{aligned} &\tilde{v}_{i_2} \sin(\theta - \phi) + \tilde{v}_{i_2} \tilde{u}_{j_2} \cos(\theta - \phi) \\ &\quad + \tilde{v}_{j_2} \sin\phi - \tilde{u}_{i_2} \tilde{v}_{j_2} \cos\phi = 0. \end{aligned} \tag{5}$$

For each two correspondences, the combination of (4)-(5) can be expressed as

$$\mathbf{A}\mathbf{x} = \mathbf{0}, \tag{6}$$

where $\mathbf{x} = [\sin(\theta-\phi), \cos(\theta-\phi), \sin\phi, \cos\phi]^{\mathrm{T}}$. To facilitate the description of the following derivation, we denote

$$\begin{cases} x_1 \triangleq \sin(\theta-\phi), & x_2 \triangleq \cos(\theta-\phi), \\ x_3 \triangleq \sin\phi, & x_4 \triangleq \cos\phi. \end{cases} \tag{7}$$

Then (6) can be reformulated as

$$a_i x_1 + b_i x_2 + c_i x_3 + d_i x_4 = 0, \quad i = 1, 2, \tag{8}$$

where coefficients $a_i$, $b_i$, $c_i$ and $d_i$ denote the problem coefficients in (4)-(5).

According to (8), $\mathbf{x}$ should lie in the null space of $\mathbf{A}$. Then we can obtain the general solution of $\mathbf{x}$ as

$$\mathbf{x} = \lambda_1 \mathbf{X}_1 + \lambda_2 \mathbf{X}_2, \tag{9}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are the null vectors of $\mathbf{A}$.

There are implicit constraints of trigonometric functions between the entries of $\mathbf{x}$ as follows:

$$\begin{cases} x_1^2 + x_2^2 = 1, \\ x_3^2 + x_4^2 = 1. \end{cases} \tag{10}$$

By substituting the general solution (9) into (10), the special solution of $\mathbf{x}$ can be solved. Once $\mathbf{x}$ has been obtained, the angles of $\phi$ and $\theta$ are

$$\begin{cases} \phi = \arctan 2(x_3, x_4), \\ \theta = \arctan 2(x_1, x_2) + \phi. \end{cases} \tag{11}$$

Then the essential matrix can be obtained from (2).

### B. 2p2p-Based Absolute Pose Estimation

For each query image, top-$n$ reference images can be picked by performing image retrieval method such as NetVLAD [16]. Therefore, there are total $n$ image pairs as shown in Fig. 2. For each image pair, we compute the essential matrix by selecting any 2 correspondences as derived in Sec. IV-A. Next, we extract the four relative poses $(\mathbf{R}, \tilde{t})$, $(\mathbf{R}, -\tilde{t})$, $(\mathbf{R}', \tilde{t})$,$(\mathbf{R}', -\tilde{t})$ from essential matrix, where $\tilde{t}$ is the normalized relative translation vector and $\|\tilde{t}\| = 1$, and $\mathbf{R}$ and $\mathbf{R}'$ are related by a 180° rotation around the baseline [18]. Traditionally, we perform a **cheirality test** using the feature correspondences to select the correct relative pose among the four candidates as in [18].

After extracting the relative pose of each image pair, we obtain the multiple directions between the query camera center and the top-$n$ reference camera centers $\tilde{t}_1, \ldots, \tilde{t}_n$. Then the absolute position of the query image can be uniquely determined by performing triangulation from each two directions. Thus the absolute pose of the query image can be solved which utilizes 2 point correspondences from 2 reference images, respectively. The solution is denoted as 2p2p. To deal with outliers in correspondences, we propose a robust estimation framework by performing a multiple checking procedure to speed up the outlier rejection and add multiple checks for filtering the optimal solution, as shown in Fig. 2.

*1) Rcheck:* We can perform a Rotation Check (Rcheck) before triangulation when selecting two relative poses $(\mathbf{R}, \tilde{t})$ and $(\mathbf{R}_2, \tilde{t}_2)$ as follows. According to the transformation matrix representation between the query view $i$ and the two selected reference views $j$ and $j_2$, we have

$$\begin{aligned} \mathbf{T}_{jj_2} &= \mathbf{T}_{ji}\mathbf{T}_{j_2 i}^{-1}. \\ i.e., \begin{bmatrix} \mathbf{R}_{12} & \mathbf{t}_{12} \\ \mathbf{0} & 1 \end{bmatrix} &= \begin{bmatrix} \mathbf{R} & \rho\tilde{t} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}_2 & -\rho_2 \mathbf{R}_2^{\mathrm{T}} \tilde{t}_2 \\ \mathbf{0} & 1 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{R}\mathbf{R}_2^{\mathrm{T}} & -\rho_2 \mathbf{R}\mathbf{R}_2^{\mathrm{T}} \tilde{t}_2 + \rho\tilde{t} \\ \mathbf{0} & 1 \end{bmatrix}. \end{aligned} \tag{12}$$

From the rotation part in (12), we can find that $\mathbf{R}_{12} = \mathbf{R}\mathbf{R}_2^{\mathrm{T}}$ must be satisfied for two inlier relative poses. Note that $\mathbf{T}_{jj_2}$ is the known transformation matrix between two reference views which can be retrieved from pre-built database. Subsequent triangulation is performed only if the selected relative position can pass the rotation check. Considering measurement noise and calibration error, we set a threshold $\tau_{th}$ to perform a relaxed Rcheck. Specifically, we compute the error between the two rotation matrix as $\tau = \arccos(0.5Tr(\mathbf{R}_{12}(\mathbf{R}\mathbf{R}_2^{\mathrm{T}})^{\mathrm{T}}) - 0.5)$ in degree. The Rcheck is considered passed if $\tau \leq \tau_{th}$.

*2) Triangulation and Pdcheck:* The absolute position of the query view can be solved through triangulation by solving the translation part in (12) as follows. Denoting $\mathbf{s} = [\rho, \rho_2]^{\mathrm{T}}$, the translation part in (12) can be expressed as

$$\mathbf{C}\mathbf{s} = \mathbf{b}, \tag{13}$$

where

$$\mathbf{C} = [\tilde{t}, -\mathbf{R}\mathbf{R}_2^{\mathrm{T}}\tilde{t}_2], \mathbf{b} = \mathbf{t}_{12}. \tag{14}$$

Then we find least-squared solution of $\mathbf{s}$ by

$$\mathbf{s} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{b}. \tag{15}$$

Note that the above solution procedure holds only when the $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ is invertible. If $Det(\mathbf{C}^{\mathrm{T}}\mathbf{C}) = 0$, the regular term needs to be added to make it invertible, as follows

$$\mathbf{s} = (\mathbf{C}^{\mathrm{T}}\mathbf{C} + \mathbf{I}_{2\times 2})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{b}, \tag{16}$$

where $\mathbf{I}_{m\times m}$ is $m \times m$ identity matrix. After triangulation, there is a Positive Depth Check (Pdcheck) which requires translation scales to be positive as $\rho > 0, \rho_2 > 0$.

*3) Consistency Check:* With the estimated scale, we can obtain the absolute pose of query view through the known pose of reference view in database. Then we can measure the consistency between the relative translation obtained from the estimated absolute pose and the retrieved relative translation from the essential matrix. Specifically, denoting the camera center of query view as $c_i$ and reference view as $c_j$, the relative translation between query view and reference view is $\hat{t} = \mathbf{R}_j^{\mathrm{T}}(c_i - c_j)$. And $\mathbf{R}_j$ is the rotation which rotates the vector in reference view to the global coordinate system. The direction of $\hat{t}$ should be consistent with the direction of $\tilde{t}$ recovered from the essential matrix. For quantification, we compute the angle $\alpha = \arccos(\tilde{t}^{\mathrm{T}}\hat{t}/\|\hat{t}\|)$. If the angle between the two predicted translation directions is less than a given threshold $\alpha_{th}$, the consistency check is considered passed.

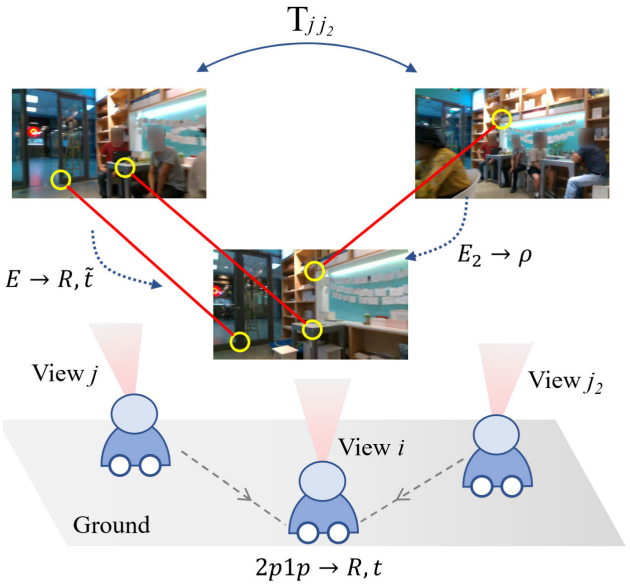



Fig. 4. Illustration of proposed 2p1p minimal solution for absolute pose estimation. By exploiting the local planar motion property, this solution requires only 2 points for essential matrix estimation and 1 point for triangulation, enhancing the robustness of the 3D model-free visual localization system.

*4) RANSAC and Nonlinear Optimization:* The proposed multiple-checking based minimal solution can be used for model estimation in RANSAC framework [40] for outlier rejection to achieve robustness pose estimation. Finally, we perform 6DoF nonlinear optimization using the identified 2D-2D inliers (and 2D-3D inliers if there is depth information in database) to optimize the absolute pose corresponding to the largest number of inliers in RANSAC. The nonlinear optimization can correct for noise in other DoF caused by occasional uneven ground.

### C. 2p1p-Based Absolute Pose Estimation

From the derivation in Sec. IV-A, we can find that essential matrix can be solved using 2 point correspondences between the query view and one reference view. Then only the absolute translation scale is unknown which needs to be solved with information provided by another reference view. As one point correspondence can provide one constraint based on epipolar constraint as in (3), the minimal solution for absolute pose estimation from essential matrix could be 2p1p, which means only one point correspondence from another reference view is required to solve the unknown scale, as shown in Fig. 4. The details in derivation of 2p1p-based minimal solution is shown as follows.

As in Sec. IV-A, we first solve the essential matrix using 2 point correspondences in reference view $j$. Then with the known transformation matrix $\mathbf{T}_{j_2 j}$ between reference view $j$ and $j_2$ retrieved from the database, we can express the $\mathbf{T}_{j_2 i}$ using the unknown absolute scale $\rho$ in $\mathbf{T}_{ji}$ as follows:

$$\begin{aligned}\mathbf{T}_{j_2 i} &= \mathbf{T}_{j_2 j}\mathbf{T}_{ji} \\ &= \begin{bmatrix}\mathbf{R}_{21} & \mathbf{t}_{21} \\ \mathbf{0} & 1\end{bmatrix}\begin{bmatrix}\mathbf{R} & \rho\tilde{t} \\ \mathbf{0} & 1\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{R}_{21}\mathbf{R} & \rho\mathbf{R}_{21}\tilde{t} + \mathbf{t}_{21} \\ \mathbf{0} & 1\end{bmatrix}. \end{aligned} \tag{17}$$

Then the essential matrix between query view $i$ and reference view $j_2$ can be expressed with the unknown scale $\rho$ as $\mathbf{E}_2 = [\mathbf{t}_{j_2 i}]_\times \mathbf{R}_{j_2 i} = [\rho\mathbf{R}_{21}\tilde{t} + \mathbf{t}_{21}]_\times \mathbf{R}_{21}\mathbf{R}$. Given one 2D-2D feature correspondence denoted as $p_{i_3}$ in query view and $p_{j_3}$ in reference view $j_2$, we can obtain one constraint about $\rho$ using the epipolar constraint as:

$$p_{j_3}^{\mathrm{T}} \mathbf{E}_2 p_{i_3} = 0, \tag{18}$$

*i.e.*,

$$p_{j_3}^{\mathrm{T}} [\rho\mathbf{R}_{21}\tilde{t} + \mathbf{t}_{21}]_\times \mathbf{R}_{21}\mathbf{R} p_{i_3} = 0. \tag{19}$$

By solving (19), we can uniquely determine the unknown scale $\rho$. We denote this minimal solution as 2p1p, which can be integrated into the RANSAC framework for robust pose estimation. Furthermore, we can apply the multiple checking procedure proposed in Sec. IV-B to enhance the robustness of our method.

## V. Theoretical Analysis

In the previous section, we derive two minimal solutions for absolute pose estimation of the camera based on 2D-2D feature matching: 2p2p and 2p1p. Subsequently, we will conduct a theoretical analysis of the proposed methods, focusing on estimation accuracy, solution stability, robustness to outliers, and sensitivity. Finally, discussion on the selection and combination of the solutions will be presented.

### A. Estimation Accuracy

We first analyze the lower bound of the estimation accuracy of the proposed minimal solutions. The estimation processes of the two minimal solutions can be delineated into two critical stages: essential matrix estimation and triangulation. We will next provide a detailed analysis of the lower bounds of estimation precision within these two stages.

*1) Essential Matrix Estimation:* The estimation process for the essential matrix is elaborated in detail in Sec. IV-A. The problem is defined as, given a pair of noisy 2D-2D feature correspondences as input, to solve for the essential matrix $\mathbf{E}$ via epipolar geometry. In the terminology of this paper, it can be expressed as determining the precision lower bound of the vector $\mathbf{x}$, which contains the critical unknowns of the essential matrix, under the presence of noise in the feature correspondences $p_i$, $p_j$, by utilizing constraint relation in (6). Assuming the presence of noise in feature matching causes perturbations $\delta\mathbf{A}$ in $\mathbf{A}$, we can analyze the impact of perturbations in $\mathbf{A}$ on the solution $\mathbf{x}$. Assuming the perturbed solution is $\mathbf{x} + \delta\mathbf{x}$, then we have:

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{0}, \tag{20}$$

$$(\mathbf{A} + \delta\mathbf{A})\delta\mathbf{x} = -\delta\mathbf{A}\mathbf{x}. \tag{21}$$

Since $\mathbf{A} + \delta\mathbf{A}$ may be singular, and

$$\mathbf{A} + \delta\mathbf{A} = \mathbf{A}(\mathbf{I}_{4\times 4} + \mathbf{A}^{-1}\delta\mathbf{A}), \tag{22}$$

then combining (21)-(22) we have:

$$\delta\mathbf{x} = -(\mathbf{I}_{4\times4} + \mathbf{A}^{-1}\delta\mathbf{A})^{-1}\mathbf{A}^{-1}\delta\mathbf{A}\mathbf{x}. \tag{23}$$

Then we have

$$\|\delta\mathbf{x}\| \leq \frac{\|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|\|\mathbf{x}\|}{1-\|\mathbf{A}^{-1}\delta\mathbf{A}\|}. \tag{24}$$

It is evident that the noise present in feature detection results in perturbations $\delta\mathbf{A}$ in $\mathbf{A}$, thereby causing errors $\delta\mathbf{x}$ in the obtained solution. Note that while the calculation of absolute error in the (24) requires the substitution of ground true values, we generally conduct error analysis in simulation experiments to guide practical applications. Additionally, in real-world scenarios, relative error is also an important factor, reflecting the degree to which the solution's error is affected by the input data's error. We can directly calculate the relative error as:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1-\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}. \tag{25}$$

In the following derivations, we denote the specific error upper bound obtained by calculating the right side of (24) is $\delta\mathbf{x}_{ub} = \begin{bmatrix}\sigma_1 & \sigma_2 & \sigma_3 & \sigma_4\end{bmatrix}^{\mathbf{T}}$. Then the error upper bound $\delta\mathbf{E}_{ub}$ in the obtained essential matrix can be expressed according to (2):

$$\delta\mathbf{E}_{ub} = \begin{bmatrix}0 & \sigma_2 & 0\\ \sigma_4 & 0 & \sigma_3\\ 0 & \sigma_1 & 0\end{bmatrix}. \tag{26}$$

*2) Pose Decomposition:* The first step in triangulation is to decompose $\mathbf{R}$ and $\mathbf{t}$ from the essential matrix. Recalling the decomposition process, it is achieved through Singular Value Decomposition (SVD), specifically involving:

$$\mathbf{E} = \mathbf{U}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{bmatrix}\mathbf{V}^{\mathrm{T}}, \tag{27}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and represent the SVD solution of $\mathbf{E}$. Then we have:

$$\mathbf{R} = \mathbf{U}\begin{bmatrix}0 & -1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 1\end{bmatrix}\mathbf{V}^{\mathrm{T}}, \tag{28}$$

and $\tilde{t}$ is the third column of $\mathbf{U}$. Calculating the specific error form of $\mathbf{R}$ and $\tilde{t}$ from the errors in $\mathbf{E}$ is complex. Here, we make some relaxations and compute a looser error lower bound for approximation. We assume each element in $\mathbf{U}$ and $\mathbf{V}$ is independent and has noise with a standard deviation of $\delta u$. Then by the error propagation formula [41], [42], comparing (27) and (28), we can infer that the absolute error of each element $r_{ij}$ in $\mathbf{R}$ is related to the absolute error of each element $e_{ij}$ in $\mathbf{E}$ as follows:

$$2\delta r_{ij} \simeq 3\delta e_{ij}. \tag{29}$$

The detailed derivation process can be found in Appendix-B. Then we have

$$\delta r_{ij} \leq 1.5\sigma, \tag{30}$$

where $\sigma = \max(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$. Then

$$\delta\mathbf{R}_{ub} = 1.5\sigma\mathbf{Ones}_{3\times3}, \tag{31}$$

where $\mathbf{Ones}_{m\times n}$ represents an $m$ by $n$ matrix of all ones. And $\delta\mathbf{R}_{ub}$ represent the upper bound of the estimation error of $\mathbf{R}$, which also represent the lower bound of the estimation accuracy.

For $\tilde{t}$, as it is the third column of $\mathbf{U}$, and we have made a relaxation assuming each element of $\mathbf{U}$ is independent with Gaussian noise of standard deviation $\epsilon$, the estimation error of $\tilde{t}$ can be represented by $\delta u$. Based on the error propagation formula for multiplication operations, combining the noise value in the essential matrix denoted in (26), we can calculate a loose upper bound of $\delta u$ as $\delta u_{ub} = 0.25\sigma$. The detailed derivation process can be found in Appendix-B. This implies that we have:

$$\delta\tilde{t}_{ub} = 0.25\sigma\mathbf{Ones}_{3\times1}. \tag{32}$$

Then according to (14), we can approximate the upper bound error of $\mathbf{C}$ (*c.f.* Appendix-B) as:

$$\delta\mathbf{C}_{ub} = \begin{bmatrix}0.25 & 18\\ 0.25 & 18\\ 0.25 & 18\end{bmatrix}\sigma. \tag{33}$$

Next, we will analyze the error in absolute scale estimation in $\mathbf{t}$ through triangulation.

*3) Triangulation:* The triangulation process constraints are presented in (13) for 2p2p and (19) for 2p1p. Note that (19) can also be expressed as $c\rho = b$, where $c$ and $b$ are scalars. Thus, the triangulation process for both 2p2p and 2p1p essentially involves solving a linear equation. We will analyze the error using (13) as an example.

We adopt Cramér-Rao Lower Bound (CRLB) [43], [44] to analyze the lower bound of the estimation accuracy of triangulation. To employ the CRLB, an likelihood function must be defined, illustrating the probability of observing the data $\mathbf{b}$ given the parameters $\mathbf{s}$. Without loss of generality, we assume $\mathbf{b}$ encompasses additive Gaussian noise, signified as $\mathbf{b} = (\mathbf{C}+\delta\mathbf{C})\mathbf{s}+\epsilon$, where $\epsilon$ represents noise with zero mean and standard variance $\sigma_{\mathbf{b}}$. As $\mathbf{b}$ is constituted by the pose of reference images in the map, $\sigma_{\mathbf{b}}$ represents the precision of map construction.

Then the likelihood function $L(\mathbf{s})$ is formalized as the probability density function of $\mathbf{b}$ as:

$$L(\mathbf{s};\mathbf{b}) = \frac{1}{\sqrt{(2\pi\sigma_{\mathbf{b}}^2)^3}}\exp\left(-\frac{1}{2\sigma_{\mathbf{b}}^2}\|(\mathbf{C}+\delta\mathbf{C})\mathbf{s}-\mathbf{b}\|^2\right). \tag{34}$$

Then we can calculate the Fisher information matrix as:

$$I(\mathbf{s}) = -E\left[\frac{\partial^2 \ln L(\mathbf{s};\mathbf{b})}{\partial\mathbf{s}^2}\right] = \frac{1}{\sigma_{\mathbf{b}}^2}(\mathbf{C}+\delta\mathbf{C})^{\mathrm{T}}(\mathbf{C}+\delta\mathbf{C}). \tag{35}$$

The CRLB of $\mathbf{s}$ estimation is the inverse of the Fisher information matrix's diagonal elements. Then we have

$$CRLB(\mathbf{s}) = \sigma_{\mathbf{b}}^2 diag(((\mathbf{C}+\delta\mathbf{C})^{\mathrm{T}}(\mathbf{C}+\delta\mathbf{C}))^{-1}). \tag{36}$$

Consequently, we have analyzed the lower bounds of estimation precision for the pose's rotation matrix, unit translation

vector, and absolute translation scale. It is evident that the precision of the first two is influenced by the error in keypoint detection during feature extraction, while the latter is also related to the precision of map construction.

### B. Solution Stability

We primarily analyze the numerical stability of the triangulation stage. According to (13), the condition number is defined as:

$$cond(\mathbf{C}) = \|\mathbf{C}^{-1}\| \cdot \|\mathbf{C}\|. \tag{37}$$

A larger value of the condition number implies that the solution is more sensitive to small changes in the input data.

From (14), we can see that $\mathbf{C}$ is composed of the direction vectors from the origin of the query camera coordinate system $i$ to the origin of coordinate system of reference camera $j$, and from the query origin to the origin of coordinate system $j_2$. Analyzing geometrically, the closer the distance between the origins of reference systems $j$ and $j_2$, or the more parallel the direction vectors from $i$ to $j$ and $i$ to $j_2$, the closer matrix $\mathbf{C}$ approaches singularity. This leads to a very large norm of the inverse $\mathbf{C}^{-1}$, thereby increasing the condition number $cond(\mathbf{C})$. A higher condition number indicates poorer stability of the solution, leading to high uncertainty and unreliable results. Therefore, we add an additional **Stability Check** prior to triangulation to exclude unstable solutions. In practical computation, this can be achieved by determining if the distance between the origins of the two reference frames or the directions from the query frame to the two reference frames used for triangulation is less than a set threshold.

### C. Robustness to Outliers

When outliers or mismatches exist in feature matching, we employ the RANSAC framework [40] to eliminate outliers and achieve correct estimation results. Based on the principles of RANSAC, we can analyze the robustness of the proposed minimal solution from a probabilistic perspective. Specifically, from (38), we can find that with the fixed number of RANSAC iterations $k$, the less minimal number of elements needed to solve the model ($n$), the higher success rate ($P$) RANSAC can achieve. And the assumption of fixed number of iterations is reasonable, as the max iterations is a common parameter to be set in RANSAC. Our proposed minimal solutions significantly reduce the value of $n$, thereby greatly enhancing the estimation's success rate. Additionally, as the proposed 2p1p minimal solution further reduces the value of $n$ based on the 2p2p minimal solution, it exhibits stronger robustness under extreme conditions, specifically when the proportion of mismatches is high and the absolute number of correct feature matches is also limited, which will be validated in subsequent experiments.

$$1 - P = (1 - w^n)^k. \tag{38}$$

### D. Sensitivity and Discussion

Given that the proposed minimal solutions leverage the local planar motion characteristic to reduce the dimensionality of the pose estimation problem, it is necessary to analyze the solution's sensitivity to the levelness of the ground. Intuitively, 2p2p and 2p1p differ in sensitivity to ground levelness. Compared to 2p2p, 2p1p introduces local planar motion constraints in the triangulation stage, requiring stricter ground levelness. Conversely, 2p2p can optimize for uneven ground errors during triangulation. This aspect has also been validated in subsequent sensitivity experiments. Overall, 2p1p exhibits stronger robustness against extreme outliers compared to 2p2p, while 2p2p demonstrates greater adaptability to uneven ground surfaces. The estimation accuracy of both is relatively similar. In practical applications, the choice between the minimal solutions should be guided by factors such as the levelness of the ground in the application environment, the significance of long-term changes, and the quality of feature matching, among others.

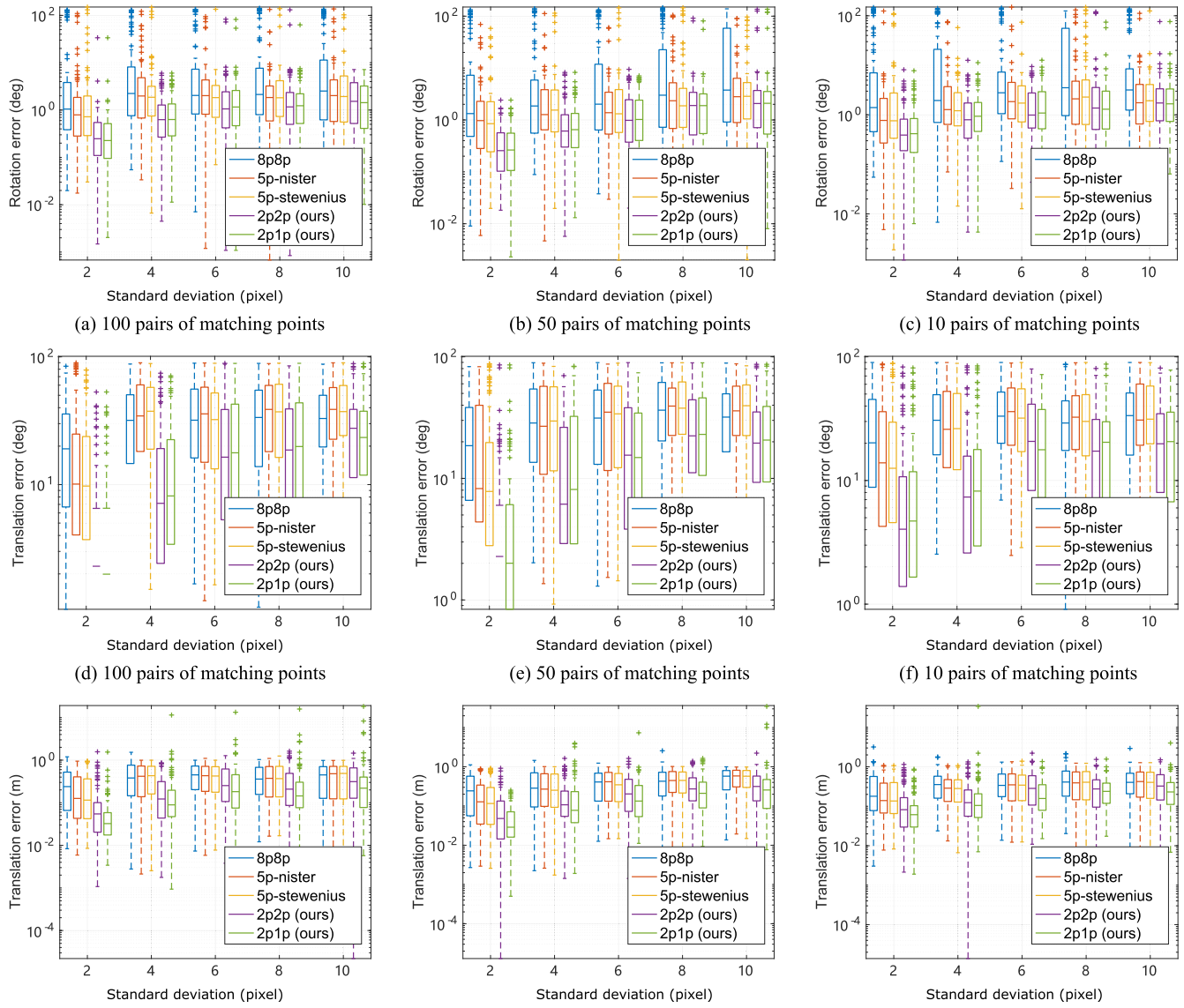
## VI. Experimental Results

To validate the performance of the proposed minimal solutions and the 3D model-free visual localization system, we conduct simulation experiments using synthetic data and real world experiments on two indoor datasets containing life-long changes. Specifically, we perform simulation experiments to test (i) the accuracy of the proposed minimal solutions, and (ii) the robustness of the proposed multiple-checking based framework against increasing outlier rate of correspondences, and (iii) the efficiency of the proposed minimal solutions by counting the computation time. In real world experiments, we first (iv) conduct ablation studies to show the effectiveness of each stage in the system, and (v) compare the success rate to show the robustness of the whole localization system in changing environments, and (vi) demonstrate the localization performance by comparing the whole trajectory, and finally (vii) show several special cases to illustrate the effectiveness of the proposed system. We also validate the sensitivity of the proposed solutions on both indoor and outdoor environments and conclude discussion on application implications to guide real world applications.

For comparison, we select the most classical 8-point solution [21] and two variations of 5-point minimal solution under 6DoF problem formulation [19], [20] to compute the essential matrix. we adopt the implementation of above algorithms in OpenGV [45]. To achieve absolute pose estimation, we perform the general triangulation for these solutions and the obtained methods are denoted as *8p8p*, *5p5p-stewenius* and *5p5p-nister*, respectively. We implement the proposed *2p2p* and *2p1p* minimal solutions and the multiple checking framework in Matlab. The $\tau_{th}$ and $\alpha_{th}$ in Rcheck and Consistency check are empirically set to 2 °. All experiments are performed on a desktop with CPU Intel i9-12900H 2.50GHz and 16G RAM.

### A. Simulation Experiments

To compare the theoretic performance of different absolute pose estimation algorithms, we generate test data in a synthetic world as in [46]. We first randomly generate 3D points in a cube $[-10, 10]^3$ as map points in the world. Then we project




(a) 100 pairs of matching points (b) 50 pairs of matching points (c) 10 pairs of matching points

(d) 100 pairs of matching points (e) 50 pairs of matching points (f) 10 pairs of matching points

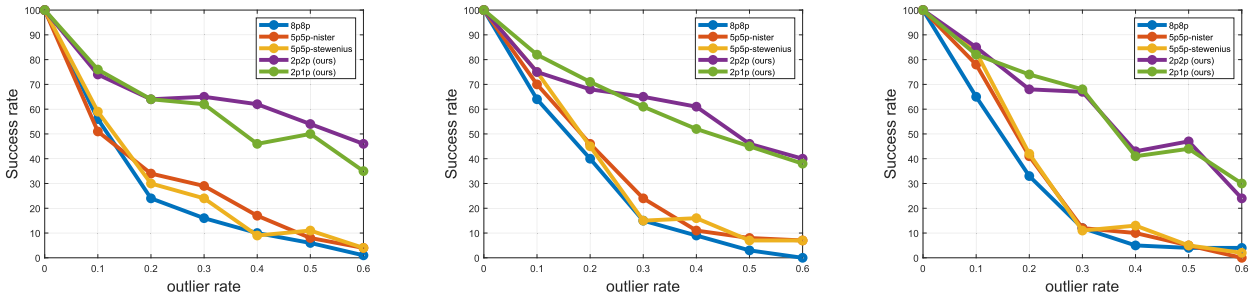
(g) 100 pairs of matching points (h) 50 pairs of matching points (i) 10 pairs of matching points

Fig. 5. Comparison of pose estimation accuracy with increasing 2D noise.

a number of co-visible points to query view and multiple reference views to obtain 2D-2D feature correspondences. All views are captured from a virtual camera with random poses sampled under local planar motion property. Specifically, the translation of the virtual camera is varying in range of $[-5, 5]$ along $x$-axis and $z$-axis, and the rotation is varying in range of $[-\pi, \pi]$ around $y$-axis. The focal length of the virtual camera is fixed as 800 with the resolution of $1280 \times 1080$ and the principal point as $(640, 540)$. For quantitative evaluation, we measure the error between the ground truth of the pose of query view $[\mathbf{R}_{gt}|\mathbf{t}_{gt}]$ and the estimated pose $[\mathbf{R}|\mathbf{t}]$. Similar in [47], we calculate the rotation error as $\arccos(0.5Tr(\mathbf{R}\mathbf{R}_{gt}^{T}) - 0.5)$ in degree, and the translation error as $|\mathbf{t} - \mathbf{t}_{gt}|$ in meter. And the relative translation vector error is computed as $\arccos(\tilde{t}^{\mathrm{T}}\mathbf{t}_{gt}/\|\mathbf{t}_{gt}\|)$, where $\tilde{t}$ is the normalized translation direction of $\mathbf{t}$.

*1) Accuracy:* We conduct accuracy evaluation experiment to verify the performance of the proposed pose estimation method under varying 2D noise levels and numbers of feature matches. Specifically, we add Gaussian nose with zero mean and increasing standard deviation from 2 to 10 pixels to the feature points of 2D-2D matching, forming eleven levels of noise. At the same time, we design three experiment groups with varying number of total feature matches: 100, 50, and 10. We generate 100 experimental tests for each noise level and perform 100 iterations of RANSAC for each method. The mean error of rotation estimation, translation direction and absolute translation estimation are shown in Fig. 5.

The results of rotation and translation direction estimation error in Fig. 5 (a)-(f) demonstrate the accuracy of different essential matrix estimation solutions. The proposed 2p-based method exhibits high accuracy under different levels of noise and number of feature matches, which is shown in results of 2p2p and 2p1p solutions. The performance of absolute pose estimation can be evaluated from the error of absolute translation estimation shown in Fig. 5 (g)-(i). The proposed solutions




(a) 100 pairs of matching points (b) 50 pairs of matching points (c) 20 pairs of matching points

Fig. 6. Comparison in robustness of pose estimation algorithms with increasing outlier rate.

outperform the compared methods, which is consistent with the results of essential matrix estimation. Particularly, even under high levels of noise and low numbers of correspondences, the proposed solutions are still able to achieve high accuracy in pose estimation. The accuracy evaluation experiment validates the effectiveness of leveraging local planar motion property into pose estimation. This is because less noise is introduced into pose estimation, leading to higher accuracy. In addition, the translation error of the proposed 2p1p minimal solution is smaller than the 2p2p solution, which further verifies the importance of formulating the motion property into triangulation.

*2) Robustness:* We conduct a robustness simulation experiment by adding increasing rates of outliers from 0% to 60% to the feature matches. Outliers are generated by projecting points using incorrect poses. The success rate of pose estimation is counted for each method to demonstrate robustness. There are three experimental groups according to the number of total feature matches: 100, 50, and 20. Note that the minimum number of feature matches is changed to 20, which is different from 10 in the accuracy experiment, as there are not enough inliers to perform pose estimation for the 8p8p method when the outlier rate increases to 60%. We consider the localization to be successful when the translation error is less than 0.1 m and the rotation error is less than 1 °. The comparison of the success rate of different methods is shown in Fig. 6.

Based on the experimental results, we can see that the performance of all methods gradually decreases as the outlier rate increases. However, the proposed methods (2p1p and 2p2p) outperform the comparison methods in all three experimental tests. The advantage of the proposed methods is more obvious with high outlier rate conditions, where all comparison methods failed to localize (e.g., outlier rate of 60%), while the proposed methods can still maintain a success rate of 30%-50%. And when the outlier rate is 60%, a comparison of the results between the proposed 2p1p and 2p2p methods under three different feature counts reveals that, as the number of features decreases, the success rate of the 2p1p method gradually surpasses that of the 2p2p method. This is because the 2p1p method requires fewer minimum feature matches for calculation, and its advantage becomes more pronounced when there is a high proportion of outliers with a low number of total features. Experimental results show that the proposed methods are more robust to outliers and can better handle the case of small number of feature matches, of which the situation is common in environments with lifelong changing variations in real-world robotic applications.

TABLE I
COMPUTATION TIME EVALUATION (ms)

| Points | 10 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|
| 2p2p | 1.394 | 1.673 | 1.777 | 2.354 | 5.638 |
| 2p1p | 0.532 | 0.557 | 0.613 | 1.046 | 1.983 |

*3) Efficiency:* The real-time performance is important for visual localization algorithms. We vary the number of point correspondences in the scene to demonstrate the efficiency of the proposed methods. In the test data, we add the Gaussian noise with zero mean and 5 pixel standard deviation and generate 10% outliers to simulate the real condition. We run the whole multiple checking procedure for 100 times to count the average execution time. The computation time evaluation results with increasing number of points are shown in Table. I. For the sake of fair comparison, the performance of compared methods implemented in C++ is not presented. It is important to note that our algorithms are currently implemented in MATLAB, which can be more efficient in C++. The results demonstrate that the proposed methods can achieve real-time performance, which is suitable for robotics tasks. In addition, as shown in Table. I, by introducing motion constraints to the triangulation stage, a significant reduction in computational complexity is achieved, resulting in a significantly faster computing speed for 2p1p compared to 2p2p.

### B. Experiments on OpenLORIS

In real-world experiments evaluation, we first perform thorough evaluations on a public indoor dataset: OpenLORIS-Scene dataset [48]. OpenLORIS-Scene is a multi-modal dataset containing RGB images, depth images and LiDAR data collected by a wheeled robot. The scenes in the dataset include common indoor environments such as offices, cafes, corridors and bedrooms. There are challenging lifelong variations including viewpoints, illuminations, dynamic objects changes and frequent human activities, which are generic and representative for localization algorithms validation.

*1) Experimental Settings:* We test the localization performance on 4 different scenes, including *home*, *corridor*, *office* and *cafe*. There are different number of sequences in each

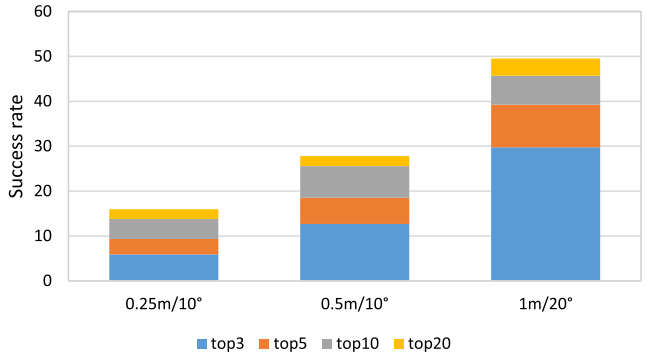



Fig. 7. The result of varying the number of reference images in the image retrieval stage on the localization performance.

TABLE II
ABLATION STUDIES: SUCCESS RATE COMPARISON ON *Home4*

| | method | 0.25m/10° | 0.5m/10° | 1.0m/20° |
|---|---|---|---|---|
| baseline | top5-R2D2-8p8p | 9.33 | 18.53 | 39.24 |
| +Tilt | top5-TiltR2D2-8p8p | 10.94 | 19.87 | 40.07 |
| improvement | | 1.61 | 1.34 | 0.83 |
| +2p2p | top5-R2D2-2p2p | 11.27 | 21.51 | 42.79 |
| improvement | | 1.94 | 2.98 | 3.55 |
| +2p1p | top5-R2D2-2p1p | 11.46 | 21.69 | 43.66 |
| improvement | | 2.13 | 3.16 | 4.42 |
| +Tilt-2p2p | top5-TiltR2D2-2p2p | 11.70 | 22.50 | 43.87 |
| improvement | | 2.37 | 3.97 | 4.63 |
| +Tilt-2p1p | top5-TiltR2D2-2p1p | 12.96 | 23.16 | 44.34 |
| improvement | | 3.63 | 4.63 | 5.1 |

[1] The improvement metrics in the table are all relative to the baseline.


Fig. 8. Localization examples extracted from *home4*, *corridor2* and *cafe1*.

scene which are collected in different time periods. Only RGB images in the dataset are used in the experiments for all methods. There are four steps for evaluation: (1) we perform NetVLAD algorithm [16] to retrieve reference images from database for each query image. (2) for feature matching between each retrieved image pairs, we use R2D2 algorithm [49] as a representative method in general motion, serving as a comparative method with the proposed TiltR2D2 in local planar motion. (3) using the known absolute poses of the retrieved reference images and 2D-2D feature matches, we estimate the absolute pose of the query image using different methods, respectively. (4) for quantitative evaluation, we measure the translation and rotation error between the ground truth pose and estimated poses and count the success rate under different error thresholds. 100 iterations are performed for each solution in RANSAC and nonlinear optimization is applied for further refinement.

*2) Ablation Study:* We conduct several experiments to validate the effectiveness of our design stages in the proposed 3D model-free localization system with success rate comparison in *home4* session. We first evaluate the impact of varying the number of reference images in the image retrieval stage on the localization performance. The result, as depicted in Fig. 7, demonstrates that an elevated number of reference images positively influences the localization success rate by facilitating a linear growth in the quantity of feature matches available. However, when the number of images exceeds 5, the performance improvement noticeably slows down. Considering that extracting additional features also increases the computational burden for subsequent pose estimation, we uniformly choose to extract the top 5 reference images in the subsequent experiments.

Then we quantitatively evaluate the impact of other stages in the localization pipeline, with the results shown in the Table. II. We select the combination of R2D2 features and the classic 8p8p pose estimation as our baseline for comparison. Based on this baseline, we evaluate the feature matching and pose estimation methods separately, expressing the performance changes of each method compared to the baseline in the "*improvement*" row of the table, highlighted in cyan. The results indicate that the proposed feature matching method brings performance improvement to any pose estimation method, including 8p8p, 2p2p, and 2p1p. Due to the presence of object variations and occlusions caused by humans in the *home4* scenario, the 2p1p method exhibits greater improvement compared to the 2p2p method. Furthermore, the proposed complete localization system under planar motion achieves optimal localization performance, as demonstrated by the "Tilt-2p1p" entry in the table, validating the effectiveness of the proposed localization system.

*3) Robustness Comparison:* The success rate comparison of different algorithms in all sessions of OpenLORIS-scene dataset is shown in Table. III. The results indicate that the proposed local planar motion aided localization system achieves the best performance in all scenes compared to the general 6DoF-based solvers, confirming the importance of leveraging motion property. In addition, comparing the results of Tilt-2p2p and Tilt-2p1p, we can find that the proposed Tilt-2p1p performs better in scenes where the localization success rate is less than 50% (such as *home4*, *corridor2*, *cafe1*). This is due to the fact that the challenges in these particular scenes are greater. For instance, frequent human activities in *home* and *cafe* scenes often result in dynamic occlusion and object movements, as shown in Fig. 8. Additionally, significant illumination changes occur in the corridor scene due to the presence of a large French window. The aforementioned challenges directly result in a lower number of matched features extracted from these scenes, along with a high rate of outliers. In such circumstances, the proposed Tilt-2p1p's advantage of requiring fewer features for localization becomes more pronounced, which is consistent with the aforementioned analysis in simulation experiments. In scenes where the localization success rate is generally higher than 50%, the advantage of the proposed Tilt-2p2p can be fully utilized thanks to the abundance of correct feature matches. Overall, the proposed methods outperform the comparative methods across the entire

TABLE III
COMPARISON OF SUCCESS RATES ON OPENLORIS-SCENE DATASET

| session | home2 | home4 | home5 | corridor2 | corridor3 | corridor4 |
|---|---|---|---|---|---|---|
| m | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 |
| degree | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 /20 | 10 / 10 / 20 |
| 8p8p | 18.49 / 28.74 / 44.39 | 9.33 / 18.53 / 39.24 | 56.61 / 66.32 / 77.33 | 2.57 / 5.36 / 11.62 | 0.48 / 1.62 / 4.77 | 3.34 / 5.61 / 10.80 |
| 5p5p-nister | 18.53 / 28.68 / 44.96 | 9.04 / 18.61 / 38.53 | 56.74 / 65.54 / 77.20 | 2.54 / 5.34 / 11.57 | 0.38 / 1.62 / 5.20 | 2.96 / 5.88 / 10.80 |
| 5p5p-stewenius | 18.06 / 28.48 / 45.16 | 9.19 / 18.51 / 38.20 | 55.96 / 64.38 / 76.04 | 2.19 / 5.31 / 11.34 | 0.29 / 1.38 / 5.00 | 3.65 / 6.03 / 11.22 |
| Tilt-2p2p (ours) | 22.18 / 32.82 / 48.03 | 11.70 / 22.50 / 43.87 | 59.82 / 69.06 / 79.72 | 2.82 / 6.10 / 12.88 | 0.95 / 3.19 / 6.48 | 4.29 / 7.25 / 15.14 |
| Tilt-2p1p (ours) | 22.68 / 31.52 / 47.10 | 12.96 / 23.16 / 44.34 | 60.08 / 68.55 / 79.33 | 3.62 / 7.79 / 13.60 | 0.91 / 3.43 / 6.86 | 4.34 / 6.56 / 13.34 |

| session | corridor5 | office3 | office4 | office5 | office7 | cafe1 |
|---|---|---|---|---|---|---|
| m | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 | 0.25 / 0.5 / 1.0 |
| degree | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 / 20 | 10 / 10 /20 | 10 / 10 / 20 |
| 8p8p | 9.04 / 16.66 / 31.02 | 31.74 / 31.74 / 40.17 | 51.90 / 59.03 / 66.05 | 62.24 / 76.09 / 89.43 | 50.22 / 67.67 / 81.76 | 6.91 / 12.29 / 22.09 |
| 5p5p-nister | 9.16 / 16.48 / 30.33 | 31.46 / 31.74 / 40.73 | 52.36 / 59.15 / 66.97 | 62.24 / 76.46 / 89.24 | 50.31 / 67.14 / 81.67 | 7.09 / 12.23 / 21.85 |
| 5p5p-stewenius | 9.04 / 16.36 / 30.61 | 31.74 / 32.30 / 41.01 | 52.36 / 59.61 / 67.09 | 62.37 / 76.46 / 88.86 | 49.69 / 67.49 / 81.94 | 7.09 / 12.17 / 21.91 |
| Tilt-2p2p (ours) | 11.85 / 21.29 / 34.47 | 32.50 / 33.89 / 43.33 | 56.21 / 60.80 / 68.74 | 66.46 / 79.23 / 90.56 | 54.51 / 69.76 / 83.17 | 7.49 / 13.93 / 23.95 |
| Tilt-2p1p (ours) | 12.27 / 20.90 / 33.90 | 32.22 / 34.44 / 43.06 | 52.41 / 59.89 / 69.31 | 69.04 / 79.55 / 89.93 | 54.34 / 70.73 / 82.91 | 8.08 / 14.40 / 24.18 |

[1] Considering the map coverage of the environment, the map sequences for each session are as follows: home1 and home3, corridor1, office1, office2 and office6, cafe2.

dataset, thus validating the effectiveness and robustness of the approach which combines local planar motion for localization.

*4) Error on Whole Trajectory:* In order to clearly compare the localization accuracy of different methods, we utilize a color bar to display the localization errors of the algorithms on the ground truth trajectory, as illustrated in Fig. 9. It is important to note that the localization is performed on each query image without sequential odometry assistance, and only cases with absolute translation error lower than 0.25m are presented on the trajectory. We select the *office5* sequence with the lowest localization difficulty to demonstrate the accuracy comparison. The bluer the color, the smaller the localization error, and higher quantity represents more successful localizations. The results demonstrate that the proposed algorithms achieve superior accuracy and robustness compared to the comparative methods across the entire trajectory, which is consistent with validation using synthetic data presented in Sec. VI-A.

*5) Case Study:* We take cases with relatively extreme outlier rates as examples to provide an intuitive demonstration of the advantage of the proposed motion property-aided algorithms against general 6DoF methods, as shown in Fig. 11. The lines crossing the query image on the left and the reference image on the right represent matched feature correspondences. Among them, the cyan lines indicate inliers that support the corresponding poses of the ground truth (top left) and the estimated poses of each method.

The first case displays a localization example with an outlier rate of 46%, where the proposed methods identify almost all inliers and achieve the highest localization accuracy. The second case presents a more challenging scenario where the scene is textureless, resulting in poor feature matching performance with only 6 inliers. As a result, only the proposed 2p1p method, which requires a minimal number of features for pose estimation, achieves high-precision localization. The 2p2p method successfully localizes the scene but with larger errors, leading to the inclusion of an outlier among the identified inliers. In contrast, all comparative methods fail to achieve successful localization as they consider a large number of spurious matches that are incorrectly identified as inliers.

### C. Experiments on Lobby

To further verify the effectiveness of the proposed methods in challenging localization scenarios, we perform validation on a self-collected Lobby dataset [39]. The dataset is collected by a mobile robot equipped with five pointgrey cameras and a 2D LiDAR as shown in Fig. 10 (a). More details about the hardwares can be found in [39]. The Lobby is a 11m $\times$ 11m square with two walls made up of large French windows, so the environment is subject to significant changes in external lighting during the day and artificial lighting at night. The static obstacles in the environment mainly include some tables, chairs and posters, while the dynamic obstacles are the moving pedestrians. The data in the Lobby dataset contains environmental changes from afternoon to evening.

There are total 50 images with a resolution of $1024 \times 540$ in the dataset and we select four images facing the four walls to form the database as shown in Fig. 10 (b) and (c), and the remaining 46 images as query images to perform evaluation. The ground truth of each image is obtained by aligning the dense synchronized 3D scans obtained by rotating the 2D LiDAR for 360 degree.

The localization performance is shown in Fig 12. The experimental results indicate that the proposed methods achieve significantly higher localization success rate than the comparative methods under all error thresholds. This is mainly due to the high difficulty of localization in the Lobby scene, with fewer feature matches and a higher proportion of outliers. As pointed out in the probabilistic analysis in Section V-C, our method significantly reduces the minimum number of features required for a single model estimation, thereby increasing the success rate of RANSAC estimation. The result is consistent with that obtained in the experiment conducted in the challenging scenes of OpenLORIS, such as *corridor*, which once again verifies the effectiveness of the proposed method in robust localization under long-term changing environments.

### D. Sensitivity Experiments

Our proposed localization system relies on the assumption of local planar motion. Consequently, we conduct sensitivity experiments to analyze the ground levelness in real-world

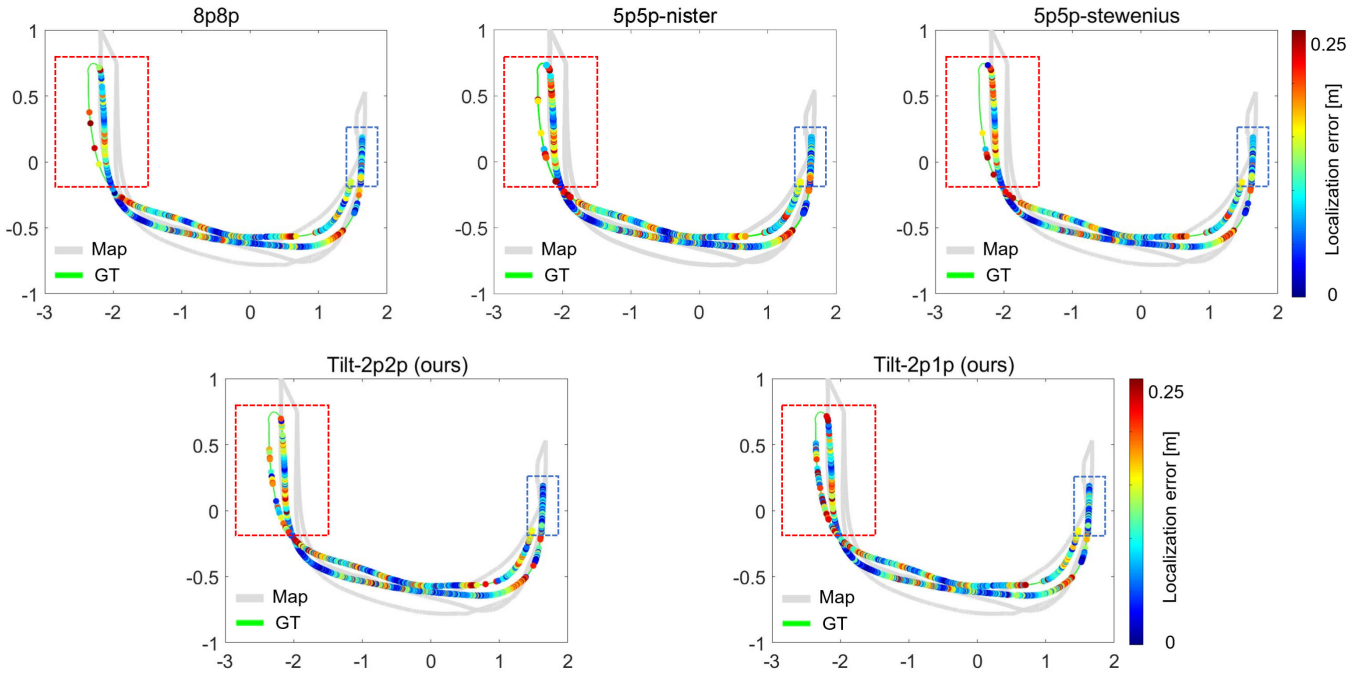



Fig. 9. Comparison of estimation errors on whole trajectory of *office5*. The color bar indicates different localization errors. The grey line represents the database trajectory and the green line indicates the ground truth of the query session. The lower the overlap between the gray and green lines, the lower the map coverage, indicating a higher difficulty in localization. As highlighted by the red dashed lines, the proposed methods demonstrate greater robustness in challenging localization scenarios. Additionally, the blue dashed box indicates that the proposed methods achieve higher localization accuracy in areas where map overlap is high.

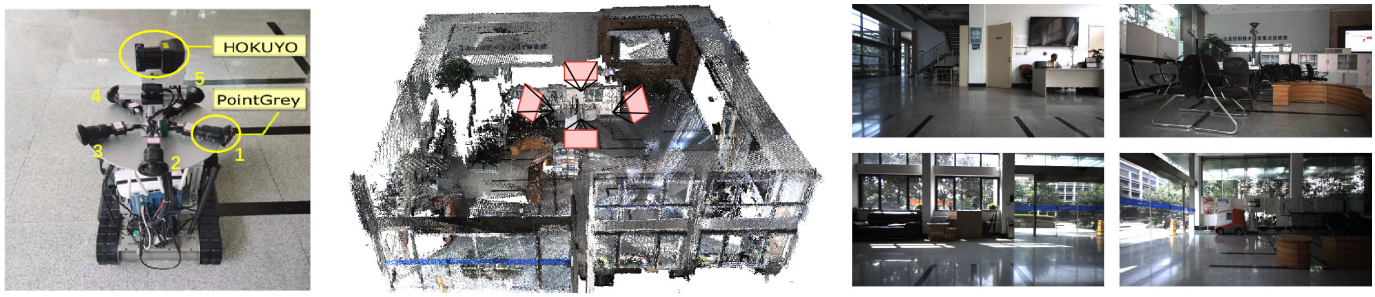
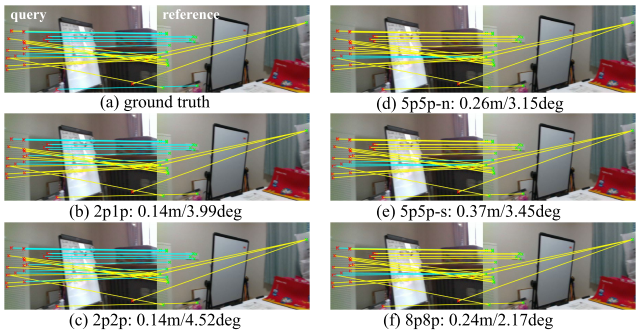


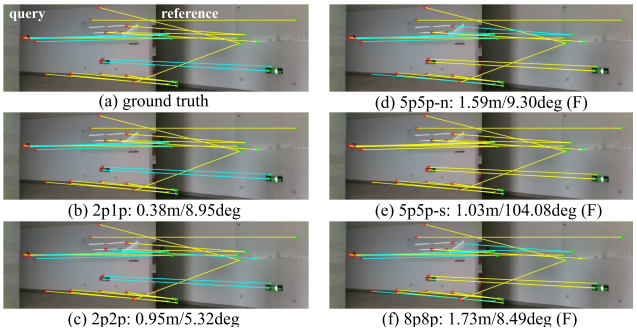

Fig. 10. (a) The platform used for Lobby dataset collection. (b) Illustration of the Lobby environment and camera positions to get the database images in (c).

environments and the tolerance of our method to varying degrees of ground unevenness. Considering that ground conditions generally differ between indoor and outdoor environments, we select the Lobby dataset for indoor scenarios and the Kaist dataset [50] collected from outdoor environments for validation and analysis. Additionally, we carry out experiments on the Kaist dataset to verify the system's robustness in positioning, with specific details provided in the Appendix-D. Here, we focus primarily on the sensitivity analysis. The experimental results are illustrated in the Fig. 13. The bar charts display histograms of the differences in pitch, roll angles, and $y$-direction translation between all queries and their corresponding references in the Lobby and Kaist datasets, respectively. These errors directly reflect the levelness of the ground. The results show that the levelness of the ground in indoor environments is significantly better than in outdoor urban settings, which aligns with our expectations.

We categorize the datasets into different query sets based on various error levels. The line charts in Fig. 13 represent the success rates of different methods on these respective query sets. Comparing the results from both datasets, it is evident that the proposed 2p1p method has lower tolerance to ground unevenness. It achieves the best localization success rate in scenarios with better ground levelness. However, its performance significantly declines on uneven grounds due to larger errors affecting both the essential matrix estimation and triangulation. In contrast, the proposed 2p2p method is less sensitive to ground levelness and maintains significant performance on the Kaist dataset, which has larger ground errors. This is consistent with our analysis in Sec. V. We also use the Pearson coefficient to quantitatively assess the success rate performance of the 8p8p and 2p2p methods. The Pearson correlation coefficient is calculated as the covariance of the two variables, adjusted for their means, divided by the product




(a) Case 1 with outlier rate of 46%.




(b) Case 2 with only 6 inliers.

Fig. 11. Comparison of identified inliers by different algorithms. The correspondences between the query and reference images are indicated by lines. The identified *inliers* that support the pose estimation by the corresponding algorithm are highlighted in cyan. The estimation errors of different methods are also indicated, with F marking a failure to localize.

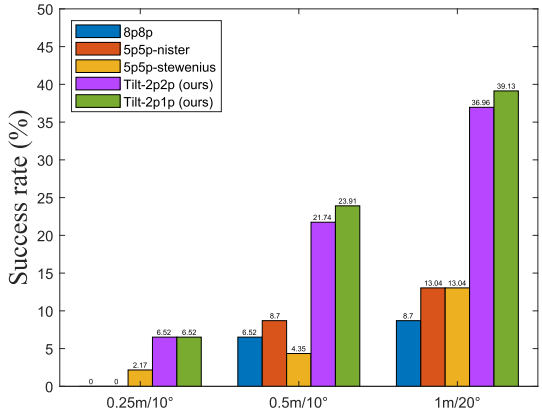



Fig. 12. Comparison of success rates on Lobby dataset.

of their standard deviations. In our analysis, the Pearson coefficient values of 8p8p and 2p2p are 96.88%. Since 8p8p is a 6DoF estimation method unaffected by ground unevenness, the high correlation between 2p2p and 8p8p suggests that there is no significant difference in the trend of the proposed method compared to 8p8p as ground unevenness errors increase.

The aforementioned analysis offers guidance for the selection and combination of the two proposed minimal solutions in practical applications. It is evident that their choice is closely related to environmental characteristics. In environments with good ground levelness, the more robust 2p1p method should be preferred to ensure the highest localization success rate. In scenarios where the ground has varying degrees of unevenness,

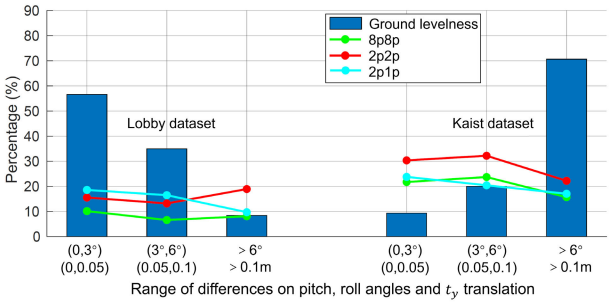



Fig. 13. The bar chart shows the histogram of errors on pitch, roll angles and $y$-direction translation on whole session of Lobby dataset and Kaist dataest. The line chart shows the success rate comparison in the corresponding query set among different range of ground levelness.

the less sensitive 2p2p method is recommended. In practical applications, the best combination can be implemented based on the changing characteristics of the environment.

### E. Discussion on Application Implications

While this paper introduces a complete localization system based on local planar motion without the need for a 3D model, it's important to note that the robust estimation backend, based on proposed minimal solutions, can be flexibly integrated with any feature matching front-end. This means no changes of plugin in front-end are required in the existing localization system to adopt the proposed minimal solutions to achieve absolute localization with a 3D model. Additionally, we wish to provide supplementary information on the system's real-time performance and accuracy trade-off. As detailed in [39], the tilt-enhanced feature matching front-end significantly improves localization performance by 67% with a 20% increase in computational time. For environments with minor perspective changes and rich textures, faster feature extraction methods, such as SuperPoint [12], may be paired with our minimal solution-based backend. This approach facilitates faster localization while maintaining accuracy. However, in environments with long-term dynamic changes, where localization becomes significantly more challenging, we recommend using the complete localization system proposed, accepting some time cost to ensure better localization performance.

## VII. Conclusion

In this paper, we propose a multiple checking-based 3D model-free robust visual localization framework and provide thorough validations through simulations and real-world experiments to demonstrate its effectiveness. The key insight of our approach is to utilize the local planar motion property both in feature matching and pose estimation, which is well-suited for general ground-moving robots. In addition, incorporating the motion property into the triangulation stage further reduces the minimal number of required features in pose estimation and enhances the robustness of the proposed system. The experiments reveal a significant improvement in the localization success rate of the proposed system, particularly in challenging localization scenarios. In the future, we plan to incorporate the minimal solution into the back-end of an end-to-end

learning-based visual localization pipeline and capitalize on the robustness of pose estimation to improve feature extraction in the front-end.

## APPENDIX A
## ASSUMPTION IN PLANAR MOTION PROPERTY

In Sec. III-A, there is a supposition regarding a coordinate system with the $y$-axis oriented downwards. Note that the coordinate system assumption aligns with the definitions in general camera imaging models, similar to [51] and [52]. Additionally, the assumption can be easily satisfied by applying the appropriate rotation transformation to construct a virtual camera coordinate system of which the image plane is perpendicular to the ground. This includes two main scenarios:

(1) As illustrated in Fig. 14 (a), there exists an axis in the given coordinate system (denoted as $i$) that aligns with the direction of gravity, though it is not the $y$-axis. In such cases, calculating the rotation matrix between these two coordinate systems is straightforward. Therefore, a virtual coordinate system, $i'$, with the y-axis pointing downwards can be established, as shown in Fig. 14 (c). All feature points in the coordinate system $i$ can be transformed into the virtual coordinate system $i'$ by applying the rotation matrix $R_{i'i}$. Subsequently, the minimal solution proposed can be applied to solve for the pose transformation between the virtual $i'$ and $j'$ coordinate systems. Finally, by applying the corresponding transformation matrix, the results can be re-expressed in the original coordinate system.

(2) As shown in Fig. 14 (b), the coordinate system is largely consistent with the assumptions in our paper, but due to installation constraints or errors, the y-axis of the camera coordinate system is not strictly vertical downwards. Then the virtual coordinate system can be reached by rotating the original camera coordinate system by certain angles around the $x$ and $z$ axes, and the corresponding rotation angle (i.e. pitch and roll) can be obtained from extrinsic calibration. Therefore, the assumption that the $y$-axis is perpendicular to the ground can be satisfied.

## APPENDIX B
## ACCURACY ANALYSIS IN POSE DECOMPOSITION

First, we perform an element-wise expansion of (27) and (28).

$$\mathbf{E} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{bmatrix} = \begin{bmatrix} u_{11}v_{11} + u_{12}v_{12} & u_{11}v_{21} + u_{12}v_{22} & u_{11}v_{31} + u_{12}v_{32} \\ u_{21}v_{11} + u_{22}v_{12} & u_{21}v_{21} + u_{22}v_{22} & u_{21}v_{31} + u_{22}v_{32} \\ u_{31}v_{11} + u_{32}v_{12} & u_{31}v_{21} + u_{32}v_{22} & u_{31}v_{31} + u_{32}v_{32} \end{bmatrix}, \tag{39}$$

$$\mathbf{R} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{bmatrix}, \tag{40}$$

$$\mathbf{R}_{11} = -u_{12}v_{11} + u_{11}v_{12} + u_{13}v_{13}. \tag{41}$$

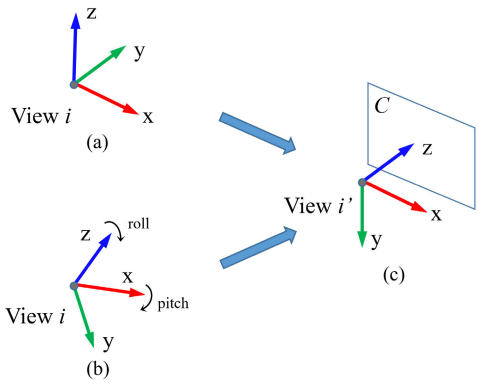



Fig. 14. (a)-(b) The original camera coordinate system. (c) The rotated virtual camera coordinate system.

Since we are not concerned with the specific subscript form of each element in $\mathbf{R}$, we expand just one element as an example for further analysis.

According to [41] and [42], the error propagation formula for addition and subtraction operations is as follows:

$$\delta(x_1^* \pm x_2^*) \simeq \delta(x_1) + \delta(x_2), \tag{42}$$

where $x_1^*$ and $x_2^*$ represent the values with errors, $\delta(\star)$ represents the absolute error of variable $\star$. Then, considering $u_* v_*$ as a single entity with absolute error denoted as $\sigma_{uv}$ and comparing the absolute error transmission forms of each element $e_{ij}$ in $\mathbf{R}$ and $r_{ij}$ in $\mathbf{R}$, we can deduce:

$$\delta e_{ij} \simeq 2\sigma_{uv}, \tag{43}$$

$$\delta r_{ij} \simeq 3\sigma_{uv}. \tag{44}$$

Then relations in (29) can be concluded.

To analyze the specific form of $\delta u$, we expand $\sigma_{uv}$ here. The error propagation formula for multiplication operations [41], [42] is as follows:

$$\delta(x_1^* x_2^*) \leq |x_1^*|\delta(x_2) + |x_2^*|\delta(x_1). \tag{45}$$

Then we have

$$\sigma_{uv} \leq |u|\delta u + |v|\delta u \leq 2\delta u, \tag{46}$$

as $u$ and $v$ are elements in orthogonal matrix. Note that this represents a relatively loose upper bound. Substituting (46) into (43), we obtain $4\delta u \leq \sigma$. Therefore, we have $\delta u_{ub} = 0.25\sigma$.

Similarly, let's analyze the upper bound of the error for $\mathbf{C}$. The first column of $\mathbf{C}$ is $\tilde{t}$, thus its error is the same as that in (32). Let's expand and analyze the second column of $\mathbf{C}$, expressed as $-\mathbf{R}\mathbf{R}_2^{\mathrm{T}}\tilde{t}_2$. We use the symbol $\diamond$ to represent any arbitrary element in the matrix obtained from the multiplication of the two matrices on the right side. Since error propagation is only relevant to arithmetic operations involving variables with errors, in the following analysis, we will overlook specific variable subscripts, merely focusing whether the variable is related to $r$ (rotation) or $t$ (translation) and the type of arithmetic operation involved. Therefore, we can simply the expanded form of $\diamond$ as:

$$\diamond = r_{*,1}\tilde{t}_{*,1} + r_{*,2}\tilde{t}_{*,2} + r_{*,3}\tilde{t}_{*,3}. \tag{47}$$

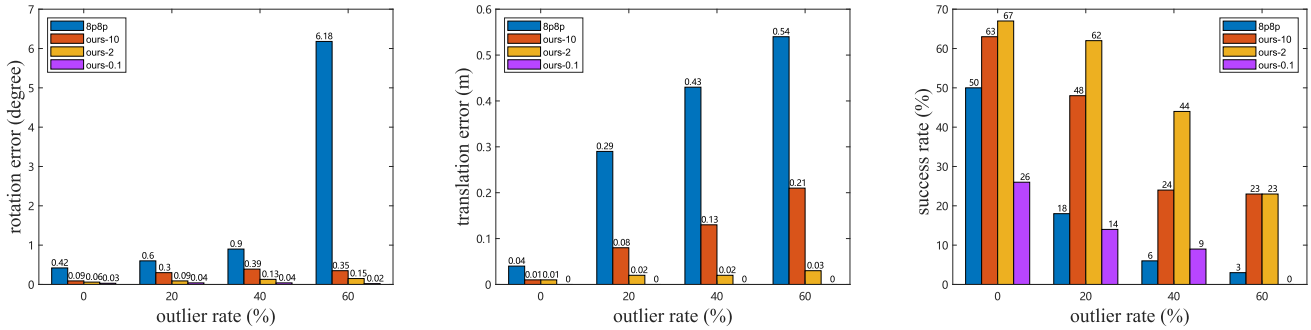



Fig. 15. Comparison in accuracy and robustness by selecting different thresholds.

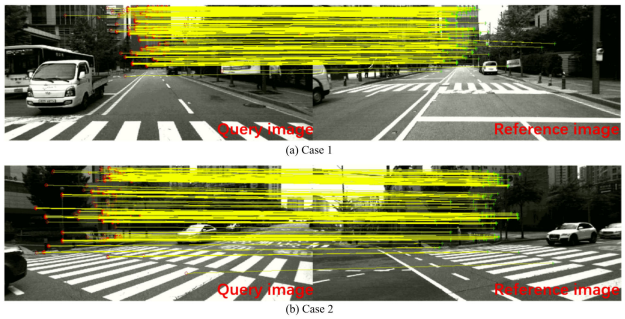



Fig. 16. Some examples in Kaist outdoor dataset. The yellow lines are inliers in feature matches.

TABLE IV
SUCCESS RATE COMPARISON ON KAIST DATASET

| session | Urban39 |
|---|---|
| m | 0.25 / 0.5 / 1.0 |
| degree | 10 / 10 / 20 |
| 8p8p | 2.87 / 16.09 / 45.79 |
| 5p5p-nister | 4.21 / 16.86 / 44.44 |
| 5p5p-stewenius | 3.26 / 15.90 / 45.98 |
| Tilt-2p2p (ours) | 6.90 / 23.95 / 58.81 |
| Tilt-2p1p (ours) | 5.36 / 18.01 / 47.89 |

And according to $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$, $e_{ij}$ can also be expanded into a form that is directly related to the components of $\mathbf{R}$ and $\tilde{t}$, as $e_{ij} = r_{*,4}\tilde{t}_{*,5} - r_{*,5}\tilde{t}_{*,5}$. Again considering $\sigma_{rt}$ as a single entity for error propagation analysis, we can derive $\delta\diamond \leq 1.5\sigma$. Then we can obtain the simplified expanded form of each element in the second column of $\mathbf{C}$ as $\dagger = -(r_{*,1}\diamond_1 + r_{*,2}\diamond_3 + r_{*,3}\diamond_3)$. And

$$\delta\dagger \leq 3(|r|\delta r + |\diamond|\delta\diamond) \leq 3(1.5\sigma + 3 \times 1.5\sigma) = 18\sigma. \quad (48)$$

## APPENDIX C
## THRESHOLD SELECTION

During the execution of Rcheck and Consistency Check, two parameter thresholds $\tau_{th}$ and $\alpha_{th}$ need to be set. We conduct parameter selection experiments to analyze the impact of parameter choices on method performance, thereby determining the optimal parameters. These two parameters are physically identical, both measuring the deviation angle between two direction vectors. Therefore, we consider setting these two parameters to the same value. Specifically, we select three different thresholds 0.1, 2, 10 for analysis of difference in localization accuracy and robustness. The experimental results are shown in Fig. 15.

To simulate real data, we introduce 10 pixels of Gaussian noise in 2D measurements. We also fix the total number of feature associations at 100 and gradually increase the proportion of outliers to test robustness. It's important to note that when calculating estimation accuracy, we only consider the cases where localization was successful. The results indicate that too tight a threshold, like 0.1, leads to the rejection of some noisy inliers, resulting in high accuracy but significantly reduced robustness. Conversely, too loose a threshold, like 10, can cause some outliers to be treated as inliers, thereby increasing estimation error. Overall, a 2-degree threshold seems to offer the best balance between accuracy and robustness.

## APPENDIX D
## EXPERIMENTS ON KAIST

We also validate the proposed system in the outdoor environment using the public urban dataset Kaist [50]. The Kaist dataset is collected with ground vehicles operating in highly diverse urban areas. As can be seen from the scenes shown in Fig. 16, the dataset is also consistent with the application of the proposed system, since the motion of the vehicle is also locally planar. Such a scenario is satisfied in most driverless applications in urban environments. In the dataset, there are two sequences, Urban38 and Urban39. The images in Urban38 sequence is used to form database and Urban39 is used to test the localization performance. The success rate comparison results are shown in Table. IV.

The experimental results demonstrate that the proposed system significantly outperforms comparative methods. We also observe that in this dataset, 2p2p clearly outperforms 2p1p. Integrating the results from both the Kaist and Lobby datasets, it is evident that 2p2p has a higher tolerance to ground levelness, hence its more pronounced advantage in the outdoor Kaist dataset. This aspect is discussed in more detail in the earlier sensitivity experiments. Overall, the proposed approach is also applicable to autonomous driving applications operating on local planar roads.

## REFERENCES

[1] X. Lin, J. Ruan, Y. Yang, L. He, Y. Guan, and H. Zhang, "Robust data association against detection deficiency for semantic SLAM," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 1, pp. 868–880, Jan. 2024.

[2] J. Cheng, C. Wang, and M. Q.-H. Meng, "Robust visual localization in dynamic environments based on sparse motion removal," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 658–669, Apr. 2020.

[3] J. Cheng, H. Zhang, and M. Q.-H. Meng, "Improving visual localization accuracy in dynamic environments based on dynamic region removal," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1585–1596, Jul. 2020.

[4] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "Indoor relocalization in challenging environments with dual-stream convolutional neural networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 2, pp. 651–662, Apr. 2018.

[5] F. Santoso, M. A. Garratt, and S. G. Anavatti, "Visual–inertial navigation systems for aerial robotics: Sensor fusion and technology," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 260–275, Jan. 2017.

[6] Y. Jiao, Y. Wang, X. Ding, M. Wang, and R. Xiong, "Deterministic optimality for robust vehicle localization using visual measurements," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5397–5410, Jun. 2022.

[7] Z. Zhang, Y. Song, S. Huang, R. Xiong, and Y. Wang, "Toward consistent and efficient map-based visual-inertial localization: Theory framework and filter design," *IEEE Trans. Robot.*, vol. 39, no. 4, pp. 2892–2911, Aug. 2023.

[8] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 4104–4113.

[9] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946.

[10] E. Brachmann et al., "DSAC—Differentiable RANSAC for camera localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2492–2500.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[13] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, "2-entity random sample consensus for robust visual localization: Framework, methods, and verifications," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4519–4528, May 2021.

[14] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.

[15] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EP$n$P: An accurate $O(n)$ solution to the P$n$P problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, Feb. 2009.

[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

[17] S. Lowry et al., "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[18] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[19] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–770, Jun. 2004.

[20] H. Stewénius, C. Engels, and D. Nistér, "Recent developments on direct relative orientation," *ISPRS J. Photogramm. Remote Sens.*, vol. 60, no. 4, pp. 284–294, Jun. 2006.

[21] M. A. Fischler and O. Firschein, *Readings in Computer Vision: Issues, Problem, Principles, and Paradigms*. Amsterdam, The Netherlands: Elsevier, 2014.

[22] L. Li et al., "Geo-localization with transformer-based 2D-3D match network," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4855–4862, Aug. 2023.

[23] S. Yan et al., "Long-term visual localization with mobile sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17245–17255.

[24] M. Geppert, P. Liu, Z. Cui, M. Pollefeys, and T. Sattler, "Efficient 2D-3D matching for multi-camera visual localization," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5972–5978.

[25] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3D point clouds," in *Large-Scale Visual Geo-Localization*. Cham, Switzerland: Springer, 2016, pp. 147–163.

[26] M. Li et al., "2D3D-MATR: 2D-3D matching transformer for detection-free registration between images and point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 14128–14138.

[27] E. Brachmann and C. Rother, "Learning less is more-6D camera localization via 3D surface regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4654–4662.

[28] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5974–5983.

[29] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12716–12725.

[30] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in *Proc. Annu. Conf. Robot Learn. (CoRL)*, 2018, pp. 456–465.

[31] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19516–19547, 2021.

[32] Y. Cai, J. Zhao, J. Cui, F. Zhang, T. Feng, and C. Ye, "Patch-NetVLAD+: Learned patch descriptor and weighted matching strategy for place recognition," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2022, pp. 1–8.

[33] Z. Li, C. D. W. Lee, B. X. L. Tung, Z. Huang, D. Rus, and M. H. Ang, "Hot-NetVLAD: Learning discriminatory key points for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 974–980, Feb. 2023.

[34] D. Ortín and J. M. M. Montiel, "Indoor robot motion based on monocular images," *Robotica*, vol. 19, no. 3, pp. 331–342, May 2001.

[35] C. C. Chou and C.-C. Wang, "2-point RANSAC for scene image matching under large viewpoint changes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 3646–3651.

[36] S. Hong, J. S. Lee, and T. Kuc, "Improved algorithm to estimate the rotation angle between two images by using the two-point correspondence pairs," *Electron. Lett.*, vol. 52, no. 5, pp. 355–357, Mar. 2016.

[37] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 4293–4299.

[38] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe, "To learn or not to learn: Visual localization from essential matrices," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3319–3326.

[39] Y. Jiao et al., "Leveraging local planar motion property for robust visual matching and localization," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7589–7596, Jul. 2022.

[40] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[41] T. Sauer, *Numerical Analysis*. Reading, MA, USA: Addison-Wesley, 2011.

[42] R. L. Burden, *Numerical Analysis*. Pacific Grove, CA, USA: Brooks/Cole Cengage Learning, 2011.

[43] H. Cramér, *Mathematical Methods of Statistics*, vol. 43. Princeton, NJ, USA: Princeton Univ. Press, 1999.

[44] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," in *Breakthroughs in Statistics: Foundations and Basic Theory*. New York, NY, USA: Springer, 1992, pp. 235–247.

[45] L. Kneip and P. Furgale, "OpenGV: A unified and generalized approach to real-time calibrated geometric vision," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1–8.

[46] Y. Jiao et al., "Robust localization for planar moving robot in changing environment: A perspective on density of correspondence and depth," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 4006–4012.

[47] T. Sattler et al., "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8601–8610.

[48] X. Shi et al., "Are we ready for service robots? The OpenLORIS-scene datasets for lifelong SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3139–3145.

[49] J. Revaud et al., "R2D2: Repeatable and reliable detector and descriptor," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 12405–12415.

[50] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *Int. J. Robot. Res.*, vol. 38, no. 6, pp. 642–657, May 2019.

[51] B. Guan, J. Zhao, Z. Li, F. Sun, and F. Fraundorfer, "Minimal solutions for relative pose with a single affine correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1926–1935.

[52] S. Choi and J.-H. Kim, "Fast and reliable minimal relative pose estimation under planar motion," *Image Vis. Comput.*, vol. 69, pp. 103–112, Jan. 2018.

**Yanmei Jiao** received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2022.

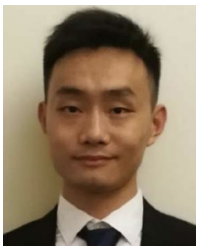
She is currently with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou. Her current research interests include mobile robotics and vision-based localization.

**Binxin Zhang** received the B.S. degree in engineering from Qianjiang College, Hangzhou Normal University, in 2022. He is currently pursuing the master's degree in software engineering with Hangzhou Normal University.

His current main research interests include visual SLAM.

**Peng Jiang** (Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2004.

He is currently a Professor with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou. His main research interests include wireless sensor networks, embedded systems and its application, intelligent instruments, and big data analysis.



**Chaoqun Wang** (Member, IEEE) received the B.E. degree in automation from Shandong University, Jinan, China, in 2014, and the Ph.D. degree in robot and artificial intelligence from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, in 2019.

During the Ph.D. study, he spent six months with The University of British Columbia, Vancouver, BC, Canada, as a Visiting Scholar. From 2019 to 2020, he was a Post-Doctoral Fellow with the Department of Electronic Engineering, The Chinese University of Hong Kong. He is currently a Professor with the School of Control Science and Engineering, Shandong University. His current research interests include autonomous vehicles, active and autonomous exploration, and path planning.

**Haojian Lu** (Member, IEEE) received the B.E. degree in mechatronic engineering from Beijing Institute of Technology, Beijing, China, in 2015, and the Ph.D. degree in robotics from the City University of Hong Kong, Kowloon, Hong Kong, in 2019.

He is currently a Professor with the State Key Laboratory of Industrial Control and Technology, and the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. His research interests include micro/nanorobotics, bioinspired robotics, medical robotics, micro aerial vehicles, and soft robotics.

**Rong Xiong** (Senior Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2009.

She is currently a Professor with the Department of Control Science and Engineering, Zhejiang University. Her current research interests include motion planning and SLAM.

**Yue Wang** (Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2016.

He is currently a Professor with the Department of Control Science and Engineering, Zhejiang University. His current research interests include mobile robotics and robot perception.