Robust Vehicle Localization for Spherical Camera Models: Solution, Framework, and Verification

Yunfeng Lu, Yanmei Jiao, Dibin Zhou, Xiumei Li, *Member, IEEE*, Rong Xiong, *Senior Member, IEEE*, and Yue Wang, *Member, IEEE*

***Abstract*— Vehicle visual localization uses vision sensors to capture environmental information, enabling precise localization of autonomous vehicles within their surroundings. However, current visual localization methods generally have some shortcomings: on one hand, they are limited by the camera's field of view, on the other hand, their robustness is often inadequate under challenging conditions such as lighting changes, long-term scene changes, or occlusions. To address these issues, we formulate a general spherical camera model for both fisheye and panoramic cameras and propose a minimal solution for pose estimation using this model based on vehicle motion characteristic. The minimal solution cannot filter outliers, so a robust estimation framework is necessary. For outlier-rejection, we introduce two frameworks: a probabilistic optimal RANSAC and a globally optimal graph-based framework. We conduct a probabilistic analysis of the RANSAC to demonstrate its enhanced robustness given by the proposed minimal solution. To achieve robustness to extreme outliers (higher than 90%), we decouple the rotation and translation space through the minimal solution to construct maximum consensus graph for the two sub-problems. We then employs a maximum clique search algorithm to find the optimal solutions, achieving deterministic convergence while maintaining real-time performance. Extensive experiments with synthetic data, real-world fisheye images, and 360° panoramic images validate the robustness and efficiency of our proposed algorithms.**

***Index Terms*— Vehicle localization, visual localization, fisheye cameras, panoramic cameras, maximum consensus.**

I. Introduction

THE past decade has witnessed the increasingly significant role of visual localization in autonomous vehicle driving [1], [2], [3]. Visual localization involves vehicles leveraging visual perception information to determine their precise position and orientation within a pre-constructed map. The success of localization is crucial to the effectiveness and even safety of navigation and trajectory planning [4], especially when driving in complex road scenarios. However, it is widely recognized that visual systems are vulnerable in environments with dynamic occlusions, low textures, and long-term changes, particularly for monocular systems. Therefore, robust visual localization remains a key focus of the community.

This paper aims to address these challenges in visual localization through the following two aspects. On one hand, we can expand the camera's field of view (FOV). A larger FOV introduces more visual observations [5], significantly reducing the proportion of observations affected by dynamic occlusions, while increasing the likelihood of observing richly textured areas. On the other hand, we need to explore more robust outlier-rejection methods to deal with outliers caused by long-term changes.

A. Wide FOV

Fisheye and panoramic cameras with a larger FOV have attracted the attention of researchers. A plethora of studies targeting panoramic perception tasks, such as depth estimation [7] and semantic segmentation [8], have emerged. However, camera pose estimation still primarily relies on pinhole cameras. Consequently, there is significant research value in integrating panoramic observations into downstream tasks to reduce the number of cameras required in a vehicle system.

Specifically, the problem of map-based visual localization can be formulated as camera pose estimation based on 2D-3D observations, also known as the Perspective-n-Point (PnP) problem. PnP problems based on the pinhole camera model have been extensively studied [9], [10], [11]. Although it is possible to rectify fisheye or panoramic images into perspective images, this approach obviously loses the advantage of the inherent wide FOV observations of these models. Moreover, when the camera's FOV exceeds 180 degrees, feature distortion becomes non-linear, making accurate rectification difficult. Therefore, conventional PnP methods based on the pinhole model cannot make use of wide FOVs. Although MLPnP [12] has been proposed for fisheye camera pose estimation using a generic camera model, it cannot handle 360° panoramic cameras. Hence, a PnP formulation that is universally applicable to fisheye and panoramic camera models is still required.

B. Outlier-Rejection

RANSAC is a popular outlier-rejection framework that achieves efficient robust estimation by randomly sampling feature sets and voting for inliers [13]. However, RANSAC can only handle cases where the outlier proportion is moderate. When the outlier proportion is high, RANSAC struggles to achieve successful estimation due to its probabilistic nature

Received 25 July 2024; revised 5 December 2024, 21 January 2025, and 31 March 2025; accepted 2 April 2025. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China through the China-Croatia Bilateral Science and Technology Cooperation Project under Grant LQ24F030009. The Associate Editor for this article was G. Mao.
(Corresponding author: Yanmei Jiao.)

Yunfeng Lu, Yanmei Jiao, Dibin Zhou, and Xiumei Li are with the School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 313000, China (e-mail: ymjiao@hznu.edu.cn).

Rong Xiong and Yue Wang are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou 310027, China.

Digital Object Identifier 10.1109/TITS.2025.3558257

[14]. Then a guaranteed optimal robust framework for achieving robust localization under high outlier rate conditions is required. However, existing deterministic convergence frameworks based on Branch-and-Bound (BnB) [15] or Mixed Integer Programming (MIP) [16] have high computational complexity, making it difficult to scale to realistic problem instances.

To address above problems, this paper proposes new methods from both solution and framework perspectives. Specifically, we derive a minimal solution for vehicle localization based on a unified spherical camera model, which is applicable to any fisheye or panoramic camera model and requires only two feature matches for camera pose estimation. Based on this minimal solution, we first propose a sampling-based probabilistically optimal real-time robust estimation framework and demonstrates a significant improvement in estimation success rate through probabilistic analysis. It has been verified in the experiments that the proposed method can increase the tolerance of RANSAC-like methods to outlier proportions from 50% to 80%. Additionally, we discover from the derivation of the minimal solution that the estimation of rotation and translation can be decoupled. Then we formulate a graph-based maximum consensus set problem to efficiently solve rotation and translation sub-problems, thereby achieving efficient deterministic convergence solving under extreme outlier proportions of up to 95%. In summary, the contributions of this paper are listed as follows.

- We derive a minimal solution based on a unified spherical camera model to solve the camera pose estimation problem using only 2 correspondences for both fisheye and panoramic cameras.
- We embed the minimal solution into a statistically optimal framework and prove the enhancement of robustness through probabilistic analysis, enabling real-time estimation with moderate outlier ratios.
- We propose a guaranteed optimal framework by decoupling the solution space and formulating a graph-based maximum consensus set problem, facilitating efficient deterministic convergence under extreme outlier ratios.
- We conduct extensive simulation and real-world datasets experiments, verifying the effectiveness of our proposed method in challenging environments.

II. Related Work

A. PnP Solutions

1) Pinhole PnP: In the domain of robotics and vehicle navigation, visual localization via query images against 3D maps is a prevalent technique, with the estimation of camera pose using a PnP solver being a crucial step. PnP algorithms estimate camera pose from n 2D-3D point correspondences. Notable variants include EPnP [17], which represents 3D points as a linear combination of four virtual control points to solve for the pose; UPnP [18], which formulates the problem as an unconstrained nonlinear optimization, estimating both camera parameters and pose without prior calibration; and RPnP [19], which transforms the problem into solving a quartic polynomial, yielding candidate pose solutions. EProPnP [20] is an advanced object pose estimation network that learns 2D-3D correspondences by backpropagating through the probability density of the pose. oDLT [21] cleverly integrates measurement uncertainty into the DLT framework, constructing a sophisticated weighted least-squares system to directly solve camera pose that approaches the maximum likelihood estimate. These methods are all designed for pinhole cameras, which differ in computational efficiency, robustness, and accuracy.

2) Wide FOV PnP: Visual localization can be constrained by limited fields of view, therefore, expanding the field of view is a direct means to enhance robustness. Images with wide fields of view provide ample textures, which are fundamental for robust pose estimation. ORB-SLAM3 [22], integrating the Kannale-Brandt camera model [23], supports fisheye image input, employing MLPnP [12] as the camera relocalization algorithm. MLPnP explicitly considers the impact of observation noise. It assumes that the observed values of 2D image points follow a Gaussian distribution and finds the optimal camera pose by maximizing the likelihood function. SOS [24] proposes an innovative real-time direct monocular visual odometry for omnidirectional cameras. For panoramic cameras, current methods often address the 2D-2D relative pose estimation problem rather than the PnP problem. 360-8PA [25] redesigns the normalization algorithm in the classical eight-point method to estimate the essential matrix for spherical projections. 360vo [26] introduces a direct visual odometry method for 360 cameras, estimating camera poses by optimizing photometric error, but this approach has a high computational cost. Therefore, there is still a need for a PnP solution suitable for fisheye and panoramic camera models.

B. Robust Estimation for Visual Localization

1) RANSAC: Direct visual odometry directly utilizes image pixel intensities to estimate camera motion. However, PnP methods typically rely on the extraction and matching of feature points, such as SIFT [27] and ORB [28]. Deep learning is advancing rapidly and is being widely applied in autonomous driving and mobile robots. A control framework based on deep reinforcement learning [29] is proving effective for exploration tasks of autonomous mobile robots in unknown environments. With its powerful image feature extraction capabilities, deep learning is enhancing the environmental perception of autonomous vehicles. SuperPoint [30] is an end-to-end trained neural network capable of both feature point detection and descriptor computation. LOFTR [31], using a Transformer architecture, captures both global and local information during the feature point matching process, excelling in scenarios with occlusions and significant viewpoint changes. Despite their advancements, these methods inevitably generate some erroneous matches and structured noise. Robust estimation is crucial in visual localization to handle noise and outliers. Recently, a multitude of RANSAC variants [32], [33] have emerged. BANSAC [34] employs a dynamic Bayesian network to adaptively update the inlier scores of data points at each RANSAC iteration, enabling more efficient sampling and termination criteria. However, due to RANSAC's random

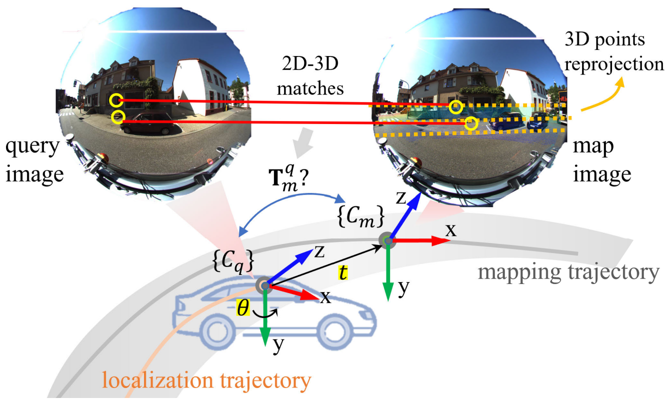


Fig. 1. Illustration of the proposed vehicle visual localization problem definition. Based on the local planar motion characteristics of the vehicle, the pitch and roll of the query camera can be approximated by those of the map camera. Therefore, vehicle visual localization can be formulated as solving for the yaw angle and translation of the query camera relative to the map camera based on 2D-3D observations. The example images are from KITTI-360 dataset [6].

sampling nature, a high proportion of outliers often necessitates more iterations to test different sample sets, increasing computational time and resource consumption, and potentially failing to find the correct camera pose.

2) Global Methods: Another class of robust estimation methods is global methods [35], which possess global convergence properties [36] and ensure that the estimated camera pose has the highest reliability within the given data. Liu et al. proposed a globally optimal algorithm based on BnB [37], leveraging line features to achieve accurate pose estimation and tracking in highly dynamic scenarios. The key to the BnB algorithm is to find the upper and lower bounds of the objective function, so each new algorithm often introduces new bounds. However, searching in a six-dimensional space remains very time-consuming. Additionally, methods based on MIP aim to eliminate outliers by introducing robust objective functions, setting distance thresholds, and incorporating outlier detection [38]. Nevertheless, existing methods have limitations in real-time performance. When apply to actual localization problems, the computation time increases exponentially with the number of feature points, making it challenging to meet real-time localization requirements.

Therefore, there is still a need for an efficient globally optimal method to solve localization problems in practical applications. We propose a novel globally optimal framework that decouples the problem into rotation and translation subproblems. By introducing a corresponding consensus function and using graph search methods to solve for the maximum consensus set, we significantly improve the efficiency of global search.

III. Overview

The process of visual localization typically adheres to a structured protocol. Initially, a current image is captured and identified as the query image. This image is then matched with a similar image retrieved from a map database, serving as the reference image. As depicted in Fig. 1, subsequent steps involve performing feature matching between the query and reference images to compute the pose of the query image. This pose, characterized by six degrees of freedom (DoF) comprising three for rotation and three for translation, is then transformed into the vehicle's pose using the extrinsic parameters that link the camera to the vehicle.

An overview of our proposed pose estimation method is shown in Fig. 2. In Sec. IV-A.1, we first reduce the 6DoF problem to 4DoF. Then, we propose a unified spherical camera model suitable for fisheye or panoramic camera models in Sec. IV-B. We use it to convert 2D-3D correspondences into projection vectors. Then, in Sec. IV-C and IV-D, based on this camera model, we propose a minimal solution for the camera pose estimation problem that requires only two 2D-3D correspondences. In Sec. V, we propose two outlier-rejection frameworks. First, we propose a sampling-based probabilistic optimal real-time robust estimation framework based on the idea of RANSAC. Then, we decompose the minimal solution into rotation and translation subproblems and construct a graph-based maximum consensus problem to achieve efficient deterministic convergence in solving.

IV. Minimal Solution

A. Problem Formulation

1) 4DoF Formulation: Given that vehicles predominantly travel on locally planar surfaces, it is plausible to assume that the pitch and roll angles of the query vehicle's pose closely approximate those of the reference pose. Precision in these angles can be further refined using data from an inertial measurement unit. Consequently, this adjustment enables us to formulate the transformation from the map camera coordinate system $\mathcal{C}_m$ to the vehicle camera coordinate system $\mathcal{C}_q$ with just four degrees of freedom, which is succinctly described by the transformation matrix $\mathbf{T}_m^q = [\mathbf{R}|\mathbf{t}]$, to be more specific:

$$\mathbf{R} = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}. \quad (1)$$

The details of reducing the 6DoF pose to 4DoF can be obtained from previous work [39].

B. Unprojection

The camera model encapsulates the mathematical correlation between the coordinates of 3D points in spatial space Ω and their corresponding projected 2D pixel coordinates within the image plane Ψ. Conventionally, the coordinates of 3D map points are designated as $\mathbf{p} = [x, y, z]^\top \in \Omega \subset \mathbb{R}^3$, and the image coordinates are represented as $\mathbf{u} = [u, v]^\top \in \Psi \subset \mathbb{R}^2$. Furthermore, $\mathbf{\Pi} : \Omega \to \Psi$ symbolizes the projection function, which is different corresponding to different camera models:

$$\mathbf{u}_i = \mathbf{\Pi}(\mathbf{R}\mathbf{p}_i + \mathbf{t}, \mathbf{K}, \mathbf{d}), \quad (2)$$

where i is the index of 2D-3D correspondence, $\mathbf{K}$ is the camera intrinsic parameters and $\mathbf{d}$ is the distortion coefficients.

For different panoramic or fisheye cameras, their projection and distortion models vary (Fig. 3), which in turn leads to distinctions in the derivation of camera poses. To unify pose estimation algorithms across these cameras, we initially

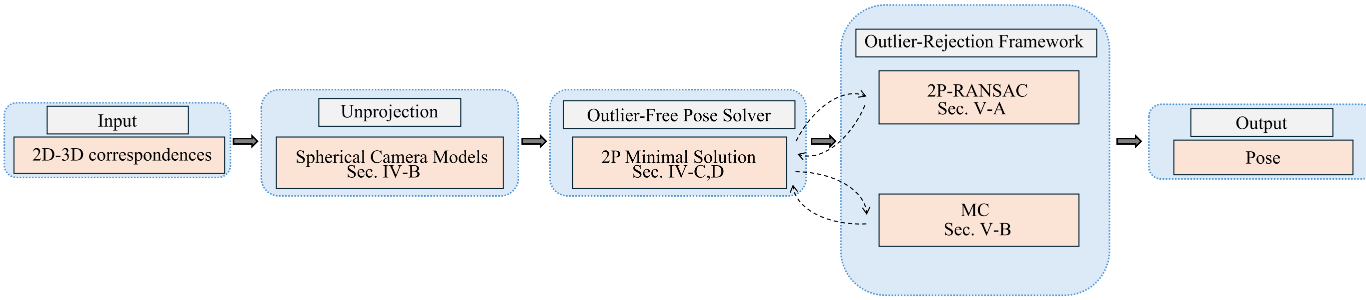
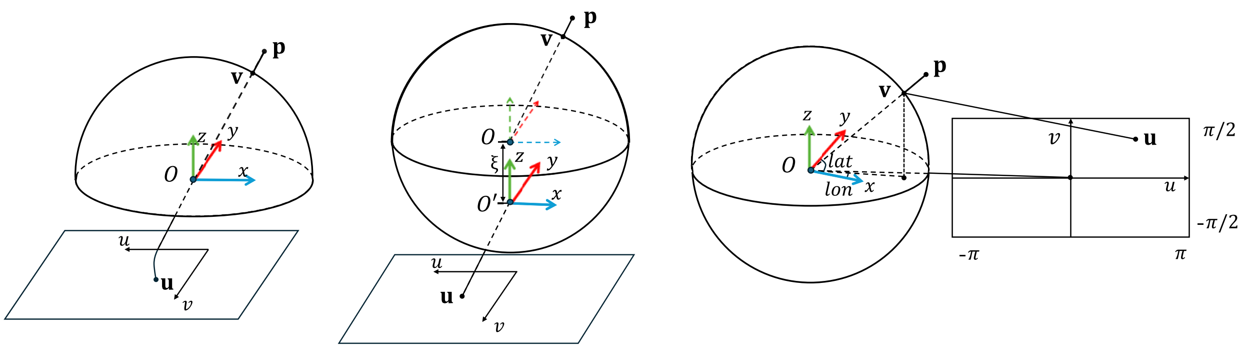


Fig. 2. Overview: The 2D-3D correspondences, which are the results of feature matching between the query image and the reference image, serve as the input to our pipeline. In the Minimal Solution, the 2D-3D correspondences are first unprojected into projection vectors through the camera model, and the problem's dimensionality is reduced to 4DoF. Then, our outlier-free solution computes the camera pose based on the relationships between these projection vectors. Finally, our outlier-rejection framework eliminates outliers based on the distribution of the 2D-3D correspondences to obtain the final camera pose.



(a) Kannale-Brand model (b) MEI model (c) Panoramic camera model

Fig. 3. Three different mappings between spherical coordinates and planar image coordinates.

transform 2D points on the image plane Ψ into projection vectors onto a spherical projection surface Θ emanating from the camera's optical center, which can be represented as $\tilde{\mathbf{v}} = [\tilde{x}_c, \tilde{y}_c, \tilde{z}_c]^\top \in \Theta \subset \mathbb{R}^3$. Furthermore, we can normalize this spherical surface to a unit sphere and obtain a normalized vector denoted as $\mathbf{v} = \tilde{\mathbf{v}}/\|\tilde{\mathbf{v}}\| = [x_c, y_c, z_c]^\top$. Subsequently, by leveraging the correspondence between these projection vectors and the matched 3D points, we proceed to derive the camera's pose. This representation based on projection vectors onto a unit spherical surface is universally applicable to any spherically modeled camera system. Following this, we outline three popular spherical camera models and detail the methodology for converting 2D image points in Ψ into such projection vectors in Θ.

1) Fisheye Camera Model of Kannale-Brandt [23]: The projection process of common fisheye camera models generally follows these steps. First, the 3D point coordinates in Ω are mapped to the normalized spherical coordinate system in Θ, and then projected onto the imaging plane where the lens distortion model is applied. Finally, they are mapped to the image pixel coordinates in Ψ. In the Kannale-Brandt model, we first use the camera's intrinsic parameters $\mathbf{K}$ to lift the pixel coordinates to the imaging plane:

$$\begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \quad (3)$$

Then the distortion model is represented as follows:

$$\theta_d = \sqrt{x_0{}^2 + y_0{}^2} = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + k_4\theta^8), \quad (4)$$

where $\mathbf{d} = [k_1, k_2, k_3, k_4]$ denotes the distortion parameters. We can solve for θ numerically, which is then used to un-project the pixel coordinates onto the spherical coordinates as:

$$r = \tan(\theta), \quad (5)$$

where r represents the norm of $\mathbf{u}$ after distortion correction on the imaging plane.

$$a = \frac{x_0}{\theta_d} r, \ b = \frac{y_0}{\theta_d} r, \quad (6)$$

$$\tilde{\mathbf{v}} = \begin{bmatrix} \tilde{x}_c \\ \tilde{y}_c \\ \tilde{z}_c \end{bmatrix} = \begin{bmatrix} a \\ b \\ 1 \end{bmatrix}. \quad (7)$$

2) Fisheye Camera Model of MEI [40]: The MEI model is a prevalent fisheye camera model. The core concept of this model lies in introducing an offset parameter ξ, thereby enabling the projection of horizontally incident rays onto the pixel plane. Similar to the Kannale-Brandt model, we first get the lifted coordinates x_0, y_0 as in (3). Then according to:

$$\begin{aligned} &x(1 + k_1(x^2 + y^2) + k_2(x^2 + y^2)^2) = x_0, \\ &y(1 + k_1(x^2 + y^2) + k_2(x^2 + y^2)^2) = y_0. \end{aligned} \quad (8)$$

we solve for the numerical approximate solution for x and y. We use MATLAB's fsolve function with the Levenberg-Marquardt algorithm, setting function tolerance to 1e-6 and maximum iterations to 100, to solve the nonlinear equations in (4) and (8) for obtaining the distortion parameter θ or undistorted coordinates x and y. And $\mathbf{d} = [k_1, k_2]$. The point

$[x, y, 1]^\top$ is the undistorted point in the z-normalized imaging plane, which is obtained by normalizing the z-coordinate of $\mathbf{v}$ after applying an offset of ξ along the z-axis. Specifically, we have:

$$\mathbf{v} = \left[\tilde{x}_c/r \ \tilde{y}_c/r \ \tilde{z}_c/r\right]^\top, \tag{9}$$

$$r = \|\tilde{\mathbf{v}}\| = \sqrt{\tilde{x}_c^2 + \tilde{y}_c^2 + \tilde{z}_c^2}. \tag{10}$$

Then the point after offset is:

$$\mathbf{v}' = \left[\tilde{x}_c/r \ \tilde{y}_c/r \ \tilde{z}_c/r + \xi\right]^\top. \tag{11}$$

After normalizing z-coordinate, we obtain:

$$\left[\tfrac{\tilde{x}_c/r}{\tilde{z}_c/r+\xi} \ \tfrac{\tilde{y}_c/r}{\tilde{z}_c/r+\xi} \ 1\right]^\top = \left[x \ y \ 1\right]^\top. \tag{12}$$

Denoting $\lambda = \tilde{z}_c/r + \xi$, and substituting (10) into the expression of λ, we have

$$\lambda = \frac{\xi + \sqrt{1 + (x^2 + y^2)(1 - \xi^2)}}{1 + x^2 + y^2}. \tag{13}$$

Then we can solve the normalized spherical coordinates as:

$$\mathbf{v} = \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} \tilde{x}_c/r \\ \tilde{y}_c/r \\ \tilde{z}_c/r \end{bmatrix} = \begin{bmatrix} \lambda x \\ \lambda y \\ \lambda - \xi \end{bmatrix}. \tag{14}$$

3) Panoramic Camera Model [41]: In panoramic cameras, camera coordinates are conventionally expressed using a spherical coordinate system. As shown in Fig. 3(c), within this framework, lon and lat represent the camera's longitude and latitude, respectively, adhering to the constraints $-\pi < lon < \pi$ and $-\pi/2 < lat < \pi/2$. These coordinates are subsequently projected onto a 2D image utilizing a cylindrical equidistant projection approach. The image height is denoted by H, while the width is represented by W. In an ideal scenario, the camera's intrinsic parameters $\mathbf{K}$, can be determined solely by specifying the image dimensions (H, W) as:

$$\mathbf{K} = \begin{bmatrix} W/2\pi & 0 & W/2 \\ 0 & -H/\pi & H/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

Then we lift the pixel coordinates to the imaging plane as:

$$\begin{bmatrix} lon \\ lat \\ 1 \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{16}$$

Then the un-projection process is expressed as follows:

$$\tilde{\mathbf{v}} = \begin{bmatrix} \tilde{x}_c \\ \tilde{y}_c \\ \tilde{z}_c \end{bmatrix} = \begin{bmatrix} \cos(lat)\cos(lon) \\ \cos(lat)\sin(lon) \\ \sin(lat) \end{bmatrix}. \tag{17}$$

C. Outlier-Free Solution

As described in the previous section, after unifying the 2D observations detected on the image plane across three camera models into projection vectors from optical center to points on the unit sphere, the constraint derived from the i-th 2D-3D correspondence in (2) is transformed into the collinearity constraint as:

$$s_i \mathbf{v}_i = \mathbf{R}\mathbf{p}_i + \mathbf{t}, \tag{18}$$

which can be expanded as:

$$s_i \begin{bmatrix} x_{c_i} \\ y_{c_i} \\ z_{c_i} \end{bmatrix} = \begin{bmatrix} x_i \cos(\theta) + z_i \sin(\theta) + t_x \\ y_i + t_y \\ -x_i \sin(\theta) + z_i \cos(\theta) + t_z \end{bmatrix}, \tag{19}$$

where s_i represents an unknown scale factor associated with depth. According to (19), we can represent unknown s_i as:

$$s_i = \frac{y_i + t_y}{y_{c_i}}. \tag{20}$$

Then substituting it into (19), we can get two independent constraints as:

$$\begin{aligned} y_{c_i} x_i \cos(\theta) + y_{c_i} z_i \sin(\theta) + y_{c_i} t_x - y_i x_{c_i} - x_{c_i} t_y &= 0, \\ y_{c_i} z_i \cos(\theta) - y_{c_i} x_i \sin(\theta) + y_{c_i} t_z - y_i z_{c_i} - z_{c_i} t_y &= 0. \end{aligned} \tag{21}$$

D. Closed-Form Solution With Gravity Alignment

By exploiting the local planar motion characteristics of vehicle movement, we can reduce the minimum number of feature matches required for camera pose estimation. As shown in Sec. IV-A.1, by leveraging local planar motion characteristics or integrating inertial measurements, we can always align the camera's orientation with gravity. At this point, the unknowns in the camera pose are reduced to 4DoF. According to (21), each 2D-3D correspondence provides two independent constraints, thus only two pairs of correspondences are needed to solve for the camera pose. By combining these four constraints, we can eliminate the unknown translation components, resulting in an equation that depends only on the rotational unknown, specifically the yaw angle:

$$d_1 \cos(\theta) + d_2 \sin(\theta) + d_3 = 0, \tag{22}$$

where

$$\begin{aligned} d_1 &= y_{c_i} y_{c_j} [x_i - x_j - \gamma(z_i - z_j)] \\ d_2 &= y_{c_i} y_{c_j} [z_i - z_j - \gamma(x_j - x_i)] \\ d_3 &= y_j y_{ci} (x_{c_j} - \gamma z_{c_j}) + y_i y_{c_j} (\gamma z_{c_i} - x_{c_i}) \\ \gamma &= \frac{x_{c_j} y_{c_i} - x_{c_i} y_{c_j}}{z_{c_j} y_{c_i} - z_{c_i} y_{c_j}}. \end{aligned} \tag{23}$$

And j is the index of another 2D-3D correspondence. Then, by reintroducing the solved value of $\hat{\theta}$ into the constraint equations we can obtain a system of linear equations with respect to $\mathbf{t}$:

$$\mathbf{H}\mathbf{t} = \mathbf{b}, \tag{24}$$

where

$$\mathbf{H} \triangleq \begin{bmatrix} y_{c_i} & -x_{c_i} & 0 \\ 0 & -z_{c_i} & y_{c_i} \\ y_{c_j} & -x_{c_j} & 0 \\ 0 & -z_{c_j} & y_{c_j} \end{bmatrix}, \tag{25}$$

$$\mathbf{b} \triangleq \begin{bmatrix} y_i x_{c_i} - y_{c_i} x_i \cos(\hat{\theta}) - y_{c_i} z_i \sin(\hat{\theta}) \\ y_i z_{c_i} - y_{c_i} z_i \cos(\hat{\theta}) + y_{c_i} x_i \sin(\hat{\theta}) \\ y_j x_{c_j} - y_{c_j} x_j \cos(\hat{\theta}) - y_{c_j} z_j \sin(\hat{\theta}) \\ y_j z_{c_j} - y_{c_j} z_j \cos(\hat{\theta}) + y_{c_j} x_j \sin(\hat{\theta}) \end{bmatrix}, \tag{26}$$

and employing a least squares optimization approach, we can determine the unknown translation vector as:

$$\hat{\mathbf{t}} = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{b}. \tag{27}$$

E. Benefits of Closed-Form Solution

As a pure solution, our proposed minimal solution 2P cannot filter outliers. It needs to be combined with an outlier-rejection strategy for robust performance. Details are provided in Sec. V. The outlier-rejection strategy should provide the solution with a sufficient number of correct 2D-3D correspondences. Likewise, a good solution can improve the efficiency and robustness of the outlier-rejection strategy. Our proposed method offers two key advantages in this regard: First, it requires only two pairs of 2D-3D correspondences, which significantly improves the efficiency of outlier filtering. Second, our method provides a closed-form solution. This allows us to decouple the pose estimation process and filter outliers using intermediate results. This decoupling leads to a significant improvement in both efficiency and robustness.

V. Robust Estimation Framework

The above solution can accurately estimate the pose only when all matches are correct, i.e., inliers. However, in practical applications, due to long-term changes in environmental appearance, it is challenging to ensure the reproducibility of descriptors during the feature matching process, leading to incorrect matches, or outliers. Therefore, the aforementioned solution needs to be embedded within a robust estimator to eliminate mismatches. Next, we will introduce two types of robust estimators, each offering different levels of optimality.

A. Statistically Optimal Estimator

RANSAC [13] is a popular robust statistical method utilized predominantly for outlier-rejection. This technique operates through iterative random sampling of subsets from the data to construct and fit models. The efficacy of each model is assessed based on its congruence with the observational data, primarily measured by the quantity of inliers, which are data points that fit well with the model under consideration. The algorithm proceeds by selecting the most optimal model, typically characterized by the maximal aggregation of inliers, thus ensuring the statistical integrity and robustness of the model's estimations against outliers.

1) 2P RANSAC: The proposed minimal solution can be directly embedded into the RANSAC framework as model estimation. To identify inliers, we calculate the re-projection error and determine whether it falls below a preset inlier threshold (2). After a fixed number of iterations, we select the pose corresponding to the largest inlier set as the initial pose. Subsequently, we use this initial pose to apply a 6DoF Gauss-Newton optimization to iteratively refine the pose. Subsequent refinement of the initial estimate is a common procedure, as performed in various studies such as [17], [42].

Additionally, the 6DoF optimization helps address errors that may occur in aligning with gravity, specifically addressing the noise issues in pitch and roll observations. The detailed algorithm steps are shown in Algorithm 1.

2) Probabilistic Analysis: Additionally, probabilistic analysis reveals that the proposed minimal solution plays a crucial role in enhancing the success rate of the outlier removal process and reducing computational complexity in RANSAC.

Algorithm 1 2P-RANSAC

Input: 2D-3D correspondences set $\mathcal{C}$
Output: The guaranteed optimal $\mathbf{R}_{opt}$, $\mathbf{t}_{opt}$
Initialize $bestInliers \leftarrow 0$
for $k = 1$ **to** n **do**
  Randomly select a minimal sample set $\mathcal{S}$ from the $\mathcal{C}$.
  Estimate the pose $(\mathbf{R}_k, \mathbf{t}_k)$ using $\mathcal{S}$ and (22) (27).
  for each correspondence $(\mathbf{v}_i, \mathbf{p}_i)$ from the $\mathcal{C}$ **do**
    Compute the reprojection error e_i using (2).
    Determine the inlier set $\mathcal{I}_k$ using e_i.
  end for
  if $|\mathcal{I}_k| > bestInliers$ **then**
    $bestInliers \leftarrow |\mathcal{I}_k|$
    $bestPose \leftarrow (\mathbf{R}_k, \mathbf{t}_k)$
  end if
end for
$\mathbf{R}_{opt}, \mathbf{t}_{opt} = NonlinearOptim(bestInliers, bestPose)$.

Specifically, the factors affecting the success rate of estimation during the RANSAC process have been analyzed, as illustrated in (28), where k is the iteration number, w is the inlier rate in data, n is the minimal number of elements for model estimation, and P is the success rate.

$$P = 1 - (1 - w^n)^k. \tag{28}$$

During the application of RANSAC, the two common termination strategies are achieving a specific success rate or setting a fixed number of iterations. We find that under a specified success rate, the computational complexity of the RANSAC estimator increases exponentially with the minimal number of data points required. When the number of iterations is fixed, reducing the minimum number of features, n, required for model estimation significantly increases the success rate of the estimation. Therefore, the proposed minimal solution can significantly enhance the robustness of vehicle localization under long-term changes.

B. Globally Optimal Estimator

The RANSAC algorithm is characterized by its robustness, operational efficiency, and simplicity in implementation. However, given its stochastic nature as a random sampling method, it cannot unequivocally assure the procurement of the global optimum solution, thereby potentially relegating the process to suboptimal local maxima. Therefore, under challenging conditions with few feature matches or a high outlier rate, estimations based on RANSAC will fail. To ensure global optimality and achieve robust localization even under extreme mismatch proportions, we further propose a deterministic convergence solution framework based on the decoupling of pose estimation.

Overview: First, we introduce outliers to further define the problem and define the *Maximum Consensus* objective. Then we define a consensus function to determine whether correspondences are in consensus. Based on our proposed minimal solution method, we can first solve the maximum consensus objective for rotation estimation. Then we propose a graph-based consensus function to find the maximum consensus set for rotation. Next, we propose a similar translation consensus function to further eliminate outliers in the results of the rotation maximum consensus set. Finally, we can obtain the globally optimal solution.

1) Consensus Maximization Formulation: In the problem of 2D-3D correspondence-based visual localization, 2D observations $\mathbf{u}_i \in \Psi$ correspond one-to-one with 3D map points $\mathbf{p}_i \in \Omega$. Following Sec. IV-B, after unifying the 2D observations onto the unit sphere Θ, the matching set can be defined as $\mathcal{C} = \{(\mathbf{v}_i, \mathbf{p}_i) : \mathbf{v}_i \in \Theta, \mathbf{p}_i \in \Omega\}_{i=1}^{N_\mathcal{C}}$, where $\mathbf{v}_i$ represents the projection vector corresponding to $\mathbf{u}_i$, and $N_\mathcal{C}$ is the total number of correspondences. In practical applications, due to the presence of noise in observations and faulty data association, $\mathcal{C}$ typically contains an unknown number of outliers. Considering these factors, the 2D-3D observation constraint in (18) should be updated to the following model to represent these realities:

$$s_i \mathbf{v}_i = \mathbf{R}\mathbf{p}_i + \mathbf{t} + \mathbf{o}_i + \mathbf{e}_i, \tag{29}$$

where $\mathbf{o}_i = \mathbf{0}$ iff $\mathcal{C}_i$ is an inlier and $\|\mathbf{e}_i\| \leq \tau$ is assumed to be bounded measurement noise. $\mathcal{C}_i$ represents i-th correspondence in set $\mathcal{C}$, and $\|\cdot\|$ denotes the Euclidean norm. To estimate the pose robustly from the input correspondences, we formulate a *Maximum Consensus* objective (MC-objective) which aims to find the pose that is consistent up to threshold τ with maximum number of correspondences:

$$\max_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N_\mathcal{C}} \mathbb{I}(\|s_i \mathbf{v}_i - \mathbf{R}\mathbf{p}_i - \mathbf{t}\| \leq \tau), \tag{30}$$

where $\mathbb{I}(\diamond) = 1$ when $\diamond$ is true, otherwise $\mathbb{I}(\diamond) = 0$. Then the resulting consensus set corresponding to an estimated pose is given by:

$$\mathcal{I}(\mathbf{R}, \mathbf{t}) = \{\mathcal{C}_i : \|s_i \mathbf{v}_i - \mathbf{R}\mathbf{p}_i - \mathbf{t}\| \leq \tau\}. \tag{31}$$

2) Decoupled Formulation: During the derivation of the 2P minimal solution in Sec. IV-C, we find that the yaw angle of rotation and the translation components in the pose to be estimated can be decoupled, thus reducing the original 4DoF pose estimation problem into two lower-dimensional sub-problems, detailed as follows.

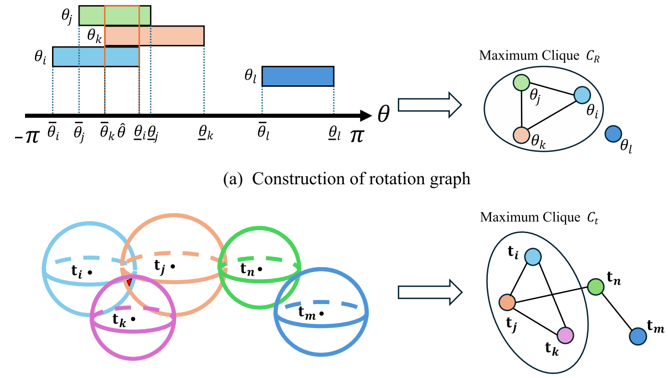
Translation invariant MC-objective: For each two pairs of correspondences $\mathcal{C}_i$ and $\mathcal{C}_j$ in set $\mathcal{C}$, by combining the collinearity constraints, we can eliminate the translation component, resulting in a constraint solely dependent on the yaw angle of the rotation matrix, as shown in (22). This can be expressed, considering noise and outliers, as follows:

$$f_{\mathbf{t},ij}(\theta) + o_{ij} + e_{ij} = 0, \tag{32}$$

where

$$f_{\mathbf{t},ij}(\theta) = d_{1,ij}\cos(\theta) + d_{2,ij}\sin(\theta) + d_{3,ij}. \tag{33}$$

and the coefficients d_* are calculated from (22) using $\mathcal{C}_i$ and $\mathcal{C}_j$. The new noise bound $\|e_{ij}\| \leq \tau_{ij}$ can be computed trough error propagation in process of (22). And o_{ij} is defined as an arbitrary value when at least one of $\mathcal{C}_i$ or $\mathcal{C}_j$ is an outlier, otherwise it is 0. For simplicity of the expression, we denote $k \triangleq ij$. Then we can formulate a translation invariant MC-objective as follows:



(a) Construction of rotation graph

(b) Construction of translation graph

Fig. 4. Visualization examples of the consensus function and the construction of rotation and translation graph. In (a), the overlapping intervals of rotation measurements i, j, and k indicate they belong to a consensus set. In (b), the overlapping regions of two translation measurement vectors indicate consensus, with the highlighted red area representing the maximum clique.

$$\max_{\theta} \sum_{k=1}^{K} \mathbb{I}(\|f_{\mathbf{t},k}(\theta)\| \leq \tau_k), \tag{34}$$

where $K = \binom{N_\mathcal{C}}{2}$. The translation invariant MC-objective indicates that the optimal yaw angle to be estimated should have the maximum number of consistent correspondences.

Rotation invariant MC-objective: By solving the translation invarinat MC problem, we can determine an optimal yaw angle $\hat{\theta}$, thereby obtaining an optimal estimate of the rotation matrix $\hat{\mathbf{R}}$. With the rotation component fixed as in (29), we can then derive a measurement solely related to translation:

$$f_{\mathbf{R},i}(\mathbf{t}) = s_i \mathbf{v}_i - \hat{\mathbf{R}}\mathbf{p}_i - \mathbf{t}. \tag{35}$$

Then the rotation invariant MC-objective is formulated as:

$$\max_{\mathbf{t}} \sum_{i=1}^{N_\mathcal{C}} \mathbb{I}(\|f_{\mathbf{R},i}(\mathbf{t})\| \leq \tau). \tag{36}$$

The rotation invariant MC-objective indicates that the optimal translation found in the maximum consensus set constructed with the optimal rotation should have the maximum number of inliers.

3) Consensus Function: To solve equations (34) and (36), we introduce consensus functions, leveraging it to construct a graph (as shown in Fig. 4) representation of the consensus structure among measurements.

Rotation consensus function: For the estimation of yaw angle, we have the uncertainty bound for the k-th candidate angle denoted as $\hat{\theta}_k \in [\overline{\theta}_k, \underline{\theta}_k]$ through error propagation. If the measurement intervals of θ_{k_1} and θ_{k_2} overlap, then correspondence set $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ belong to the same consensus set. As shown in Fig. 4(a), our rotational measurements are one-dimensional, and we can determine whether they belong to the same consensus set by the overlap of one-dimensional intervals. Therefore, we propose the following rotation consensus function:

$$\mathbb{I}_{\theta}(\theta_{k_1}, \theta_{k_2}) = \begin{cases} 1 & \text{if } \bar{\theta}_{k_1} \leq \underline{\theta}_{k_2} \wedge \bar{\theta}_{k_2} \leq \underline{\theta}_{k_1} \\ 0 & \text{otherwise} . \end{cases} \tag{37}$$

Then we construct a Rotation Graph, where each θ estimation is represented as a node. If the rotation consensus function of two θ estimations is 1, we consider these two nodes to be connected.

Translation consensus function: For translation estimation, we follow the similar idea to construct a translation graph. For each two correspondence $\mathcal{C}_i$ and $\mathcal{C}_j$, We first substitute the estimated optimal $\hat{\theta}$ into (24), and obtain an estimation of translation $\hat{\mathbf{t}}_k$ as in (27). We can also compute the uncertainty radius r_k of $\hat{\mathbf{t}}_k$ such that the ground truth $\mathbf{t}_k$ lies in the sphere centered at $\hat{\mathbf{t}}_k$ with radius r_k. Then, an estimate $\mathbf{t}$ exists for two measurements set $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ belonging to the same consensus set only if their spheres overlap. As shown in Fig. 4(b), our translation measurements are three-dimensional, and we determine whether they belong to the same consensus set by checking for overlap in three-dimensional space. Then we propose the following translation consensus function:

$$\mathbb{I}_{\mathbf{t}}(\hat{\mathbf{t}}_{k_1}, \hat{\mathbf{t}}_{k_2}) = \begin{cases} 1 & \text{if } \left\| \hat{\mathbf{t}}_{k_1} - \hat{\mathbf{t}}_{k_2} \right\| \leq r_{k_1} + r_{k_2} \\ 0 & \text{otherwise} . \end{cases} \tag{38}$$

We construct a Translation Graph, where each node represents an estimated translation $\hat{\mathbf{t}}_k$. The connection of each two nodes is guided by the translation consensus function.

4) Graph-Based Deterministic Estimation: Based on the decoupled rotation and translation consensus functions, we propose an efficient graph-based deterministic convergence robust solving framework, as shown in Algorithm 2. After constructing the rotation graph and translation graph, any existing maximum clique finding method can be applied to solve for the rotation and translation consistency sets. We use the Parallel-Maximum-Clique method [43]. After identifying the maximum consensus set, we apply the 6DoF nonlinear optimization using the final set of inliers for the correction of errors in the pitch and roll priors, thereby improving estimation accuracy.

Algorithm 2 Graph-Based Deterministic Estimation

Input: 2D-3D correspondences set $\mathcal{C}$.
Output: The guaranteed optimal $\mathbf{R}_{opt}$, $\mathbf{t}_{opt}$.
Initialize rotation graph $\mathcal{R}_{\mathcal{G}}$ and translation graph $\mathcal{T}_{\mathcal{G}}$.
for $k = 1 \ldots \binom{N_{\mathcal{C}}}{2}$ **do**
  Compute θ_k and $[\underline{\theta}_k, \bar{\theta}_k]$ with $\mathcal{C}_k$ by solving (22).
  Add node θ_k to $\mathcal{R}_{\mathcal{G}}$.
end for
for each two nodes k_1, k_2 in $\mathcal{R}_{\mathcal{G}}$ **do**
  if $\mathbb{I}_{\theta}(\theta_{k_1}, \theta_{k_2})$ **then**
    Add edge between nodes k_1, k_2.
  end if
end for
$\hat{\theta}$, $\mathcal{C}_R$ = $PMC(\mathcal{R}_{\mathcal{G}})$.
for $k = 1 \ldots \binom{N_{\mathcal{C}_R}}{2}$ **do**
  Compute $\hat{\mathbf{t}}_k$ and r_k with $\mathcal{C}_k$ and $\hat{\theta}$ by solving (27).
  Add node $\hat{\mathbf{t}}_k$ to $\mathcal{T}_{\mathcal{G}}$.
end for
for each two nodes k_1, k_2 in $\mathcal{T}_{\mathcal{G}}$ **do**
  if $\mathbb{I}_{\mathbf{t}}(\mathbf{t}_{k_1}, \mathbf{t}_{k_2})$ **then**
    Add edge between nodes k_1, k_2.
  end if
end for
$\hat{\mathbf{t}}$, $\mathcal{C}_t$ = $PMC(\mathcal{T}_{\mathcal{G}})$.
$\mathbf{R}_{opt}$, $\mathbf{t}_{opt}$ = $NonlinearOptim(\mathcal{C}_t, \mathbf{R}(\hat{\theta}), \hat{\mathbf{t}})$.

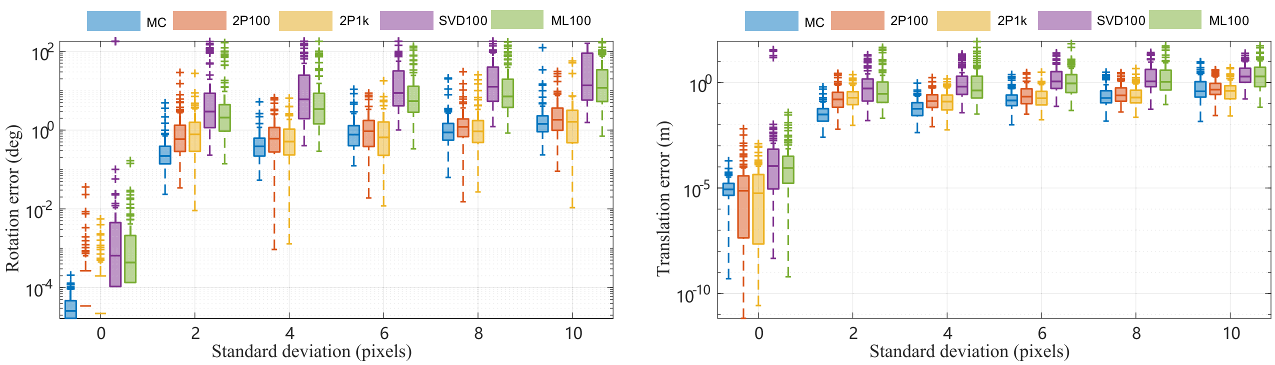
VI. Experiments

In this section, we evaluate the accuracy, robustness, and real-time performance of our proposed algorithm through simulations and real-world experiments. We use the SVD and MLPnP methods as comparison methods for our proposed minimal solution-based algorithm. The SVD is our implemented 6DoF pose estimation algorithm using the unified spherical camera model, while the MLPnP is a widely used fisheye image pose estimation algorithm provided by the toolbox [12]. To make MLPnP suitable for panoramic cameras, we input the unified spherical projection observations into MLPnP for pose estimation of panoramic images, following the method proposed in the Sec. IV-B.3. For robust estimation, we embed SVD and MLPnP into the RANSAC framework (denoted as SVD and ML) and compare them with the proposed 2P minimal solution-based RANSAC (denoted as 2P) and the graph-based deterministic convergence algorithm (denoted as MC). All methods and experiments are implemented in MATLAB, and all experiments are conducted on a portable computer with an AMD Ryzen 7 4800H CPU, clock speed of 2.90 GHz, and 16.0GB RAM, without using GPU acceleration.

A. Simulation Experiments

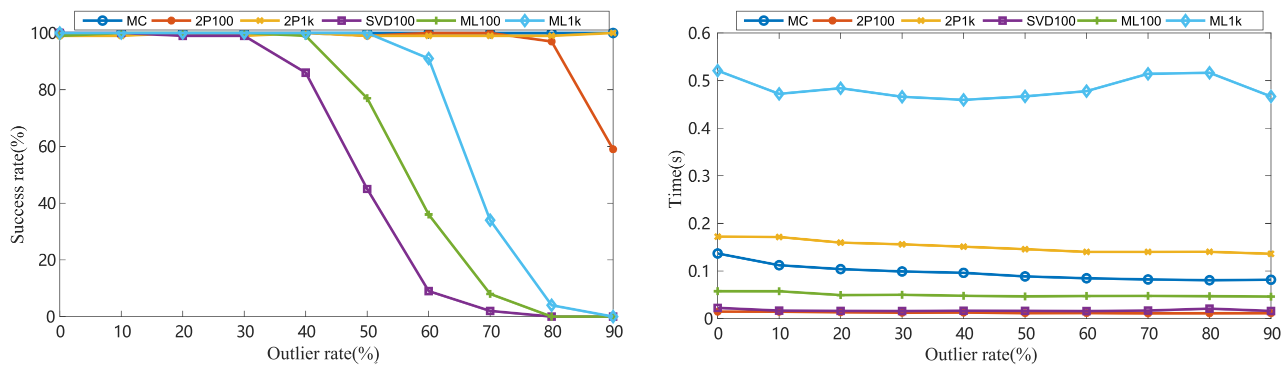
We generate a world cube of size $[-5, 5]^3$, within which we randomly generate 3D map points. Next, we randomly generate query camera poses and project the map points onto the image plane to produce the associated 2D image points. For the rotation in camera pose, we set the pitch and roll to 0 to satisfy the assumption of local planar motion, and then randomly generate the yaw angle within the interval $[-\pi, \pi]$. The translation values are randomly generated within the $[-5, 5]^3$ range. All random numbers in the experiment are generated following a standard normal distribution. To simulate spherical camera imaging, we use the intrinsic parameters of a real fisheye camera. The camera model is the MEI model (described in Sec. IV-B.1) with a resolution of 1432×1410, principal point at $(716, 705)$, distortion parameters $[0.017, 1.65, 0.0004, 0.0004]$, and offset parameter 2.21. For evaluation, we calculate the measurement error between the ground truth rotation $\mathbf{R}_{gt}$ and the estimated rotation $\mathbf{R}$ as $\arccos(\frac{trace(\mathbf{R}_{gt}^T \mathbf{R} - 1)}{2}) \frac{180}{\pi}$ degrees. The translation error between the ground truth translation $\mathbf{t}_{gt}$ and the estimated translation $\mathbf{t}$ is computed as $\left\| \mathbf{t}_{gt} - \mathbf{t} \right\|$ meters.

1) Accuracy: We add Gaussian noise with a mean of 0 and standard deviation ranging from 0 to 10 pixels to the 2D image plane, generating 100 test samples at each noise level, with



(a) Accuracy of rotation (b) Accuracy of translation

Fig. 5. Comparison of estimation accuracy under different noise standard deviations for 100 correspondences.



(a) Success rate (b) Time consumption

Fig. 6. Comparison of estimation success rate and time consumption under different outlier ratios for 100 correspondences.

each sample containing 100 pairs of 2D-3D correspondences. As shown in Fig. 5, we compare the rotation and translation errors of all methods. At almost all noise levels, our proposed MC framework demonstrates superior accuracy compared to the RANSAC framework. Specifically, at lower noise levels (less than 4 standard deviations), the superiority of the MC framework is more pronounced. In Fig. 5(b), in the noise-free experiments, the accuracy of all methods is on the order of 10^{-5}, which is close to the micrometer level. The accuracy advantage of our proposed MC method is not significant, but it demonstrates greater stability, as its median and interquartile range are closer. Additionally, a comparison of methods within the RANSAC framework shows that MLPnP slightly outperforms SVD, and our proposed 2P algorithm exhibits significantly better accuracy than both SVD and MLPnP. This improvement is attributed to our local planar motion assumption in Sec. IV-A.1, which utilizes prior information about pitch and roll to reduce errors.

2) Robustness: We initially set an outlier rate of 0-90%, generating 100 test samples at each outlier rate, with each sample containing 100 pairs of 2D-3D points. Based on the outlier rate, we generate the corresponding number of incorrect 2D-3D points as outliers. Each pair of outliers is created by randomly generating an incorrect pose for projection. The criteria for success are a rotation error of less than 1 degree and a translation error of less than 0.1 meters. Note that to highlight the probabilistic optimality of the RANSAC framework, we set different maximum iteration counts for the RANSAC-based methods for comparison. For instance, our proposed 2P RANSAC with 100 iterations is denoted as 2P100, and with 1000 iterations, it is denoted as 2P1k. The success rates of the various methods are shown in the Fig. 6.

At outlier rates of 0-90%, our proposed guaranteed optimal MC method maintains a 100% success rate, and 2P1k also achieves nearly 100% success. The success rate of 2P100 starts to decline at an 80% outlier ratio, while SVD100 and ML100 show a significant drop at 50%. This indicates that, with 100 iterations, our proposed 2P algorithm increases the tolerance to outliers from 50% to 80%. This observation is consistent with our probabilistic analysis of RANSAC (28). Even when increasing the iterations of MLPnP to 1k, its success rate still begins to drop significantly at a 60% outlier ratio.

From an efficiency perspective, although 2P1k's robustness is close to that of MC, it requires more computational time. ML1k's efficiency is significantly lower than other methods. The 2P100 method has the highest efficiency, making it a highly cost-effective option for practical applications. However, SVD and MLPnP require more computational resources compared to the 2P methods. Comparative experiments demonstrate that our proposed 2P minimal solution method offers the highest efficiency while significantly enhancing RANSAC's robustness to outliers. The proposed MC method provides the highest accuracy and robustness, achieving better

TABLE I
DATASET FEATURE COMPARISON

Feature	KITTI-360	Matterport360°	Newer College
Environment	Urban/Suburban Roads	Indoor Building Scenes	University Campus (Handheld Capture)
Sensor Type	Fisheye Cameras, IMU/GPS	Panoramic Camera	Fisheye Cameras, LiDAR, IMU
Image Resolution	1400×1400(pixels)	2048×1024(pixels)	720×540(pixels)
Camera Model	MEI Model	Cylindrical Equidistant Projection	Kannala-Brandt Model
Scene Characteristics	Varying lighting, dynamic objects, occlusions	Low-texture areas, uniform lighting, panoramic distortion	Fast motion, shaking, rapid lighting changes
Point Cloud Density	Moderate	Dense with rich local features	Sparse; textureless/repetitive regions
Localization Challenges	Lighting variations, occlusions, dynamics	Low texture, distortion, limited matches	Fast motion, severe occlusions

performance in less time than 2P1k when the outlier rate is high.

3) Extreme Outliers: To further compare the performance of MC and RANSAC, we conduct experiments under extreme outlier rates (90-95%). At these extreme outlier rates, the RANSAC framework exhibits numerous cases of localization failure, prompting us to increase the iteration counts for these methods and set the total number of 2D-3D correspondences to 200. Specifically, we perform 1k and 10k iterations for 2P, and 10k iterations for both SVD and MLPnP. In addition, we conduct a precision comparison experiment to evaluate the inlier precision rate of the algorithms. The results of all experiments are shown in Fig. 7. Our proposed MC, 2P1k, and 2P10k methods consistently demonstrate very high accuracy, with MC producing significantly fewer outlier points compare to other methods. Conversely, SVD10k and ML10k almost fail to localize successfully, highlighting the importance of specialized research in vehicle localization rather than general localization problems. Our MC method maintains a 100% precision rate even at 95% outlier rate, attributes to the global optimality of the global search approach. 2P10k also approaches 100%, with very few errors, while 2P1k shows a noticeable decline.

In terms of efficiency, MLPnP exhibits the lowest efficiency at 10k iterations, whereas 2P10k outperform MLPnP. Due to the requirement for more constraint equations in the SVD method, our implementation terminates unverified cases early under extreme outlier rates, leading to discrepancies in SVD's timing compared to previous experiments. The experiments demonstrate that MC exhibits superior robustness, while RANSAC (as seen in 2P10k) requires substantial computational resources to approach MC's performance. This further validates that the proposed MC method can achieve efficient and robust solutions even under very high mismatch ratios.

B. Real-World Experiments

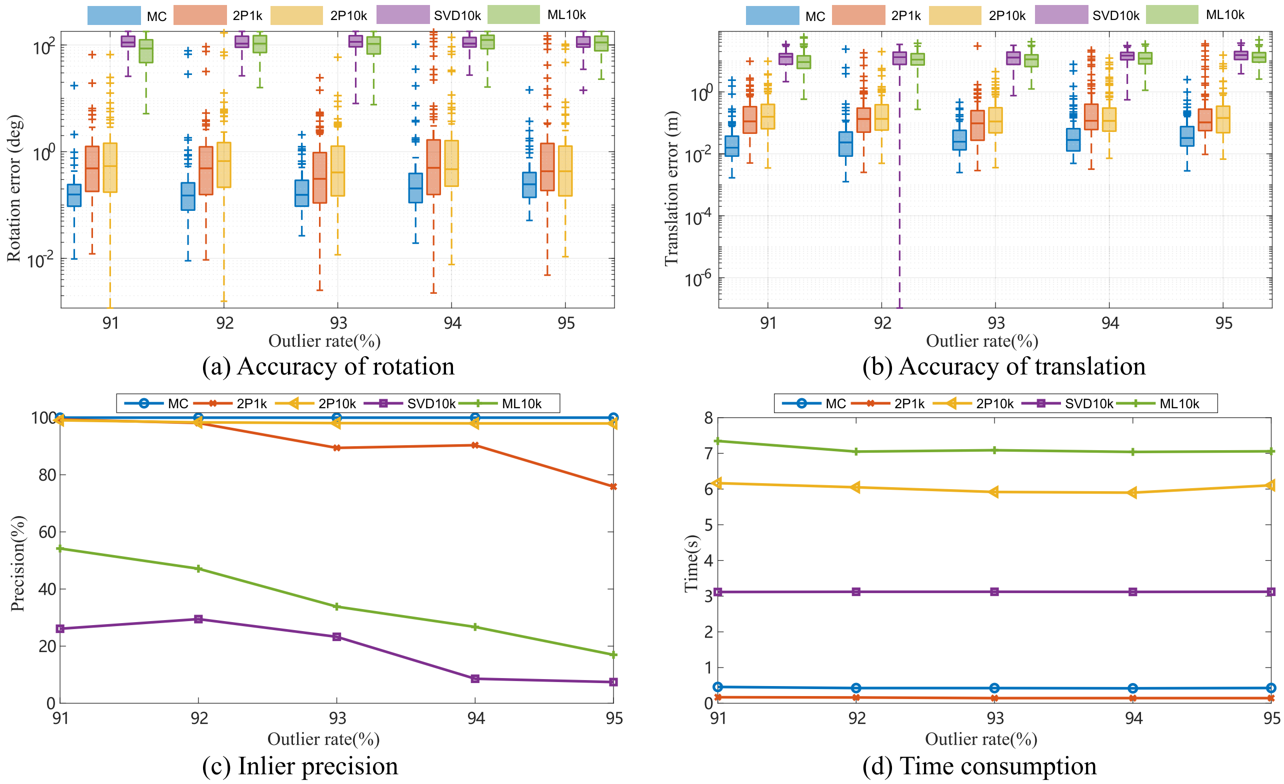
In this section, we use three real-world publicly available datasets that include indoor, outdoor, and road scenes. The image data in these datasets are collected using fisheye cameras and 360° panoramic cameras, covering the three camera models described in Sec. IV-B. The details of the three datasets are compared in Tab. I. In practical applications, the robustness of localization is more important than its accuracy, as localization failure can lead to more severe consequences. Therefore, in our real-world experiments, we primarily evaluate the robustness of the localization algorithms, namely the success rate. The comparison methods are the same as those in the simulation experiments. To control efficiency, we set 2P to iterate 100 and 1k times, and SVD and MLPnP to iterate 100 times.

TABLE II
PERFORMANCE COMPARISON OF KITTI-360

session	method	success rate $0.5/1/2(m)$ $1/2/5(°)$	time (s)	recall (%)	precision (%)	F1 (%)
seq1	MC	**7.3**/66.7/**92.3**	0.14	**100**	**44.84**	**60.74**
	2P100	6.7/66.6/87.5	**0.03**	**100**	41.97	58.55
	2P1k	6.4/**68.5**/88.2	0.16	**100**	42.17	58.67
	SVD100	7.0/65.5/87.9	0.15	**100**	41.84	57.68
	ML100	4.6/64.5/91.3	0.09	**100**	42.74	59.18
seq2	MC	**0.7**/**10.3**/38.8	0.02	**73.99**	**30.81**	**41.47**
	2P100	0.4/07.8/38.7	**0.01**	70.32	29.29	39.52
	2P1k	0.2/08.4/**40.6**	0.14	71.80	30.44	40.80
	SVD100	0.1/06.8/27.7	0.03	20.69	20.46	16.39
	ML100	0.4/08.8/39.8	0.08	55.37	20.77	27.97

1) KITTI-360: KITTI-360 is a large-scale dataset collected by a vehicle system, containing rich sensor data and ground truth annotations, recording multiple suburban scenes. The vehicle system is equipped with fisheye cameras and a laser scanner, providing a complete 360° field of view. Additionally, the system includes an IMU/GPS positioning system. The fisheye camera uses the MEI camera projection model in Sec. IV-B.2. The dataset is divided into 11 separate sequences, each corresponding to a continuous driving trajectory. We select three sequences with overlapping trajectories (sequences 4, 5 and 6) to test our localization algorithm, with sequence 5 serving as the reference image set and the other two as query image sets. We use image retrieval technique NetVLAD [44] to extract the reference images most similar to the query images. Feature extraction and matching between the query and reference images are performed using the SURF algorithm. Subsequently, we extract the laser data corresponding to the reference image frames and project the laser points onto the image using the extrinsic parameters based on the MEI camera projection model to obtain 3D information. In our proposed 2P minimal solution, we constrain the vehicle to local planar motion. We can directly obtain the pitch and roll angles of the query image from the reference image. By combining the pitch and roll transformations with the estimated yaw, we can construct the complete relative rotation matrix.

As shown in the Table. II, we compare the localization success rates under different thresholds and the time consumption of the algorithms. In sequence1, the MC method achieves the highest success rates at 0.5m and 2m thresholds, while 2P1k achieves the highest success rate at a 1m threshold. The



(a) Accuracy of rotation (b) Accuracy of translation (c) Inlier precision (d) Time consumption

Fig. 7. Comparison of Robustness and Efficiency under extreme outlier ratios for 200 correspondences.

robustness and efficiency of the algorithms are consistent with the results from the simulation experiments. In sequence2, similarly, our MC scheme demonstrates the best performance at most thresholds, with 2P1k achieving the highest success rate at a 2m threshold. The performance of ML100 is close to that of 2P100, and SVD100 performs worse than other methods. In terms of efficiency, since sequence2 mainly comes from suburban scenes, resulting in sparse image features, the MC method has lower time consumption. Overall, MC performs the best on this dataset, with 2P1k showing robustness close to MC but with lower efficiency. Meanwhile, 2P100 is an efficient compromise.

In addition, we calculate the median values of the recall, precision, and F1-score for inlier estimation. These metrics can comprehensively and quantitatively evaluate the effectiveness of the proposed algorithm for outlier rejection in feature matching. The threshold for determining inliers is set to 8 pixels, consistent with OpenCV's RANSAC library default value. As shown in Tab. II, it can be observed that MC consistently achieves the highest recall, precision, and F1-score across both sequences. In sequence 2, which has a higher outlier ratio, 2P also outperforms ML100. This result aligns with our findings from simulation experiments, further demonstrating the robustness advantage of our proposed method in high outlier scenarios.

2) Matterport360° Dataset: Since autonomous driving datasets, such as KITTI, typically obtain a 360° view using multiple fisheye or pinhole cameras rather than directly capturing a panoramic image, we select the Matterport360° [45] dataset to verify that the proposed method is also applicable under panoramic cameras. Matterport360° is an extension of Matterport3D [46] and includes 194,400 RGB-D images and 10,800 panoramic views from 90 building-scale scenes. Matterport data is collected using tripod-mounted cameras in home environments, still satisfying the local planar motion assumption of the proposed method. Matterport360° provides additional dense, high-resolution 360° depth maps on top of Matterport3D. It offers 9,684 samples from 90 different buildings. Following the protocol in [45], we use the test set distributed across 18 buildings and divide it into 7 test subsets. We use image retrieval technique NetVLAD [44] to obtain the reference images and perform feature extraction and matching using the SURF algorithm. The depth information of the images is provided by Matterport 360°. We use the panoramic camera model in Sec. IV-B.3 to lift the projection of the planar images, obtaining spherical 3D coordinates. Since the Matterport3D dataset does not provide ground truth camera poses, we cannot objectively measure the error. Therefore, we quantitatively estimate the point cloud registration accuracy between the reference image and the query image to evaluate the success rate of the estimation.

The success rates of different methods are shown in the Table. III. Note that in this dataset, the presence of large areas with low-texture regions, such as walls, and significant panoramic distortion result in a low number of feature matches and a high proportion of outliers. In such extreme high-mismatch rate scenarios, all RANSAC-based methods perform similarly. Within the RANSAC framework, our proposed 2P100 performs similarly to ML100, and 2P1k slightly outperforms both 2P100 and ML100. However, the

TABLE III
SUCCESS RATE COMPARISON OF MATTERPORT360

m	session00 0.4/0.6/0.8	session01 0.4 / 0.6/ 0.8	session02 0.4/0.6/0.8	session03 0.4/0.6/0.8	session04 0.4/0.6/0.8	session05 0.4/0.6/0.8	session06 0.4/0.6/0.8	ALL 0.4/0.6/0.8
MC	**21.3/31.9/44.7**	**09.3/13.7/19.5**	**13.0/40.6/53.6**	**16.5/34.1/44.0**	**02.6/06.5/09.1**	**08.0/26.7/36.0**	**21.6/37.8/48.6**	**13.1/27.3/36.5**
2P100	02.1/08.5/21.3	02.9/07.8/10.7	05.8/24.6/36.2	07.7/17.6/28.6	00.0/00.3/01.3	06.7/16.0/21.3	05.4/08.1/24.3	04.4/12.0/20.5
2P1k	02.1/08.5/23.4	02.9/07.8/09.8	05.8/26.1/37.7	07.7/17.6/29.7	00.0/01.3/01.3	06.7/17.3/22.7	05.4/08.1/24.3	04.4/12.4/21.3
SVD100	02.1/08.5/14.9	01.0/02.9/05.4	04.3/11.6/15.9	04.4/07.7/17.6	00.0/01.3/01.3	04.0/12.0/16.0	00.7/05.4/21.6	02.6/07.1/13.2
ML100	02.1/12.8/23.4	02.9/07.8/10.7	05.8/26.1/36.2	08.8/16.5/28.6	00.0/01.3/01.3	06.7/18.7/21.3	05.4/05.4/21.6	04.5/12.6/20.4

proposed optimality-guaranteed MC method achieves significantly higher success rates than all RANSAC methods, even showing a tenfold improvement (see session00). This demonstrates that in extremely challenging scenarios, our proposed globally optimal MC is significantly more robust than RANSAC-based approaches, whereas RANSAC with a low number of iterations can no longer guarantee a reliable localization success rate.

3) Newer College Dataset: Newer College [47] is a comprehensive multi-camera dataset covering a walking distance of 4.5 km, collected using a handheld device on the University of Oxford campus. This device includes multiple fisheye cameras, LiDAR, and an IMU. The fisheye cameras use the Kannala-Brandt model in Sec. IV-B.1. Although this dataset is collected using handheld devices and exhibits significant shaking, making it difficult to satisfy the local planar motion assumption, we utilize the IMU data to obtain prior information on pitch and roll angles. Therefore, the proposed method can still be adapted.

The dataset includes three scenes collected at different times of the year, featuring challenging aspects such as rapid motion, significant shaking, quick lighting changes, and textureless surfaces, and is categorized into different difficulty levels. We select the easy and hard levels of the quad scene and the medium and hard levels of the math scene for our experiments. We use COLMAP [48] to build a 3D map from the reference sequences. We employ NetVLAD [44] for reference image retrieval to find the reference images most similar to the query images. Feature extraction is performed using SuperPoint [30], and feature matching is conducted using the Nearest Neighbor method. For SuperPoint, we use the official PyTorch implementation and its provided pre-trained model, with a detection threshold of 0.015(applied to detector confidence scores) and an NMS radius of 4 pixels.

In this section, we incorporate EPnP [17] and two SOTA approaches, EProPnP [20] and oDLT [21]. For a fair comparison, we use RANSAC to remove outliers for them. Since they are only applicable to pinhole cameras, we unproject the fisheye images using the spherical camera model, and then we re-project them using the pinhole camera model. Ultimately, we eliminate the fisheye distortion in the images, but the rectification process typically compresses the edges of the image towards the center, which results in a reduced visible field of view. As shown in the Table. IV, the experimental results for the quad scene indicate that the easy level involves typical walking speeds, while the hard level includes aggressive movements, proximity to walls, and rapid lighting changes. It can be seen that under these slight attacks, our proposed MC solution significantly outperforms other methods. The robustness of 2P100, 2P1k, and EProPnP is similar, and all three are significantly better than oDLT. By comparing 2P100 and EPnP, we can see that the advantage of directly using the fisheye model to obtain an analytical solution is significant.

TABLE IV
SUCCESS RATE COMPARISON OF NEWER COLLEGE

m °	quad-easy 0.25/0.5/1.0 2.0/5.0/8.0	quad-hard 0.25/0.5/1.0 2.0/5.0/8.0	math-medium 0.25/0.5/1.0 2.0/5.0/8.0	math-hard 0.25/0.5/1.0 2.0/5.0/8.0
MC	**74.5/90.4/97.7**	**30.6/44.1/50.2**	**57.2/81.8**/87.1	**43.5/67.8/75.4**
2P100	70.8/87.0/95.3	27.3/40.3/48.0	55.2/80.8/**87.4**	39.7/66.3/74.7
2P1k	71.3/88.1/96.1	27.9/40.5/48.6	55.8/80.9/87.2	39.7/66.5/74.8
SVD100	47.8/68.6/82.1	18.4/31.9/38.8	38.0/67.5/78.7	24.1/51.4/62.5
ML100	55.9/79.0/92.5	22.3/37.1/44.7	38.5/74.4/85.0	23.6/58.1/71.2
EProPnP	37.3/66.2/81.5	14.3/27.8/38.1	28.5/59.5/74.9	16.3/40.4/54.3
R+EPro	63.1/86.5/96.6	27.7/42.4/49.4	55.8/81.5/87.1	40.6/66.8/73.6
R+EPnP	50.5/66.3/72.0	22.8/35.0/40.0	49.5/73.5/78.1	37.4/62.5/67.8
R+oDLT	53.0/79.8/93.3	22.0/37.4/46.6	49.9/77.9/85.1	36.8/64.6/72.9

Note: R+* represents using RANSAC to filter outliers with 100 iterations.

In the math scene, the medium level involves brisk walking with occasional changes in direction, while the hard level includes aggressive movements, fast walking, textureless surfaces, rapid device rotations, and shakes up to 5.5 rad/s. Under such intense attacks, our MC method still demonstrates superiority. The success rates of 2P100 and 2P1k are close, noticeably higher than those of SVD100 and ML100. The results indicate that the robustness of the 2P minimal solution is also evident in real data, outperforming SVD and ML methods. Similarly, the 2P and EProPnP each have their strengths and weaknesses under different thresholds. Even in scenarios with non-local planar motion, our 2P, as a 4DoF method, exhibits comparable performance to the advanced deep learning-based 6DoF method EProPnP. This result indicates that our dimensionality reduction approach is effective in most cases. oDLT's robustness is superior to MLPnP but inferior to our proposed method and EProPnP. Under intense attacks, the robustness of MC is also extremely significant.

To more comprehensively evaluate the performance of our method, Tab. V shows the median recall, precision and F1-score of different algorithms across different scenes in the Newer College dataset. From Tab. V, we can observe that our MC method achieves the best F1-score across all scenes, indicating that MC achieves the best overall performance balance on the Newer College dataset. Specifically, in quad-easy and math-medium, MC leads in all metrics compared to other methods, while 2P performs similarly to state-of-the-art R+EPro and R+oDLT. However, in the math-hard scene with less texture information and more complex motion, MC's performance advantage becomes more pronounced, while

TABLE V

OUTLIER REJECTION PERFORMANCE ON NEWER COLLEGE

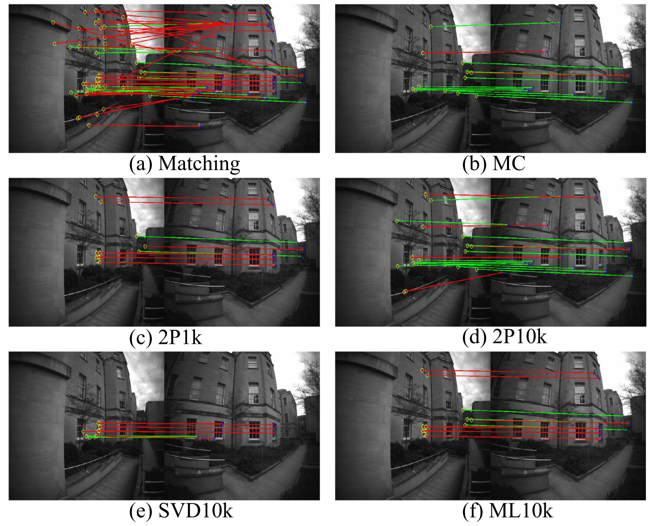
	quad-easy recall/precision/F1 % / % / %	quad-hard recall/precision/F1 % / % / %	math-medium recall/precision/F1 % / % / %	math-hard recall/precision/F1 % / % / %
MC	**88.19/65.42/74.74**	**84.84/49.53/62.68**	**78.05/49.36/60.00**	**78.95/40.40/53.01**
2P100	87.71/65.22/74.12	83.05/48.85/61.07	77.19/48.98/59.57	75.00/40.00/51.57
2P1k	87.22/65.41/73.98	82.62/48.98/60.74	76.21/48.98/59.08	75.00/40.32/51.61
SVD100	74.38/63.08/64.43	62.05/45.88/48.18	59.18/45.12/48.18	70.97/37.82/47.06
ML100	80.53/64.84/71.09	71.47/48.22/56.85	60.19/46.97/51.15	74.47/39.47/50.91
EProPnP	65.91/62.14/59.68	46.19/35.83/34.50	41.86/38.87/34.30	50.00/14.29/21.31
R+EPro	86.78/65.32/74.08	78.19/40.39/53.40	74.68/43.43/54.55	74.00/27.00/39.20
R+EPnP	83.09/57.77/67.41	76.29/39.36/49.94	75.00/48.28/58.33	66.67/19.40/28.40
R+oDLT	83.58/65.36/72.76	77.10/42.00/59.13	75.00/49.10/58.74	68.29/26.42/35.19

TABLE VI

CASE STUDY

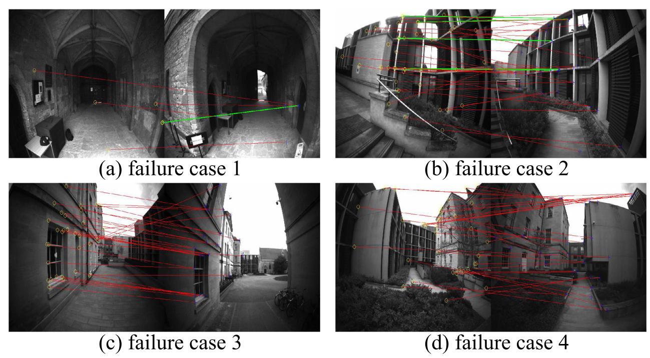
ExpID	gtinlier/match	methods	inlier(TP/TP+FP)	$\Delta\mathbf{t}$ m	$\Delta\mathbf{R}$ $^\circ$	time s
1	13/52	MC	10/12	**0.79**	**1.19**	**0.13**
		2P1k	2/11	8.93	8.71	0.21
		2P10k	12/18	1.33	3.19	5.75
		SVD10k	2/9	7.21	11.24	3.73
		ML10k	3/12	9.04	9.35	11.3
		EProPnP	7/11	6.16	9.45	-
		R+EPro	2/11	20.62	82.99	-
		R+EPnP	0/12	6.79	5.52	-
		R+oDLT	2/7	4.28	33.02	-
2	9/42	MC	8/14	**0.84**	**2.55**	**0.10**
		2P1k	8/14	**0.84**	**2.55**	0.17
		2P10k	8/14	**0.84**	**2.55**	4.7
		SVD10k	7/17	1.21	2.60	4.72
		ML10k	8/18	1.07	3.89	11.2
		EProPnP	1/1	3.19	32.89	-
		R+EPro	11/22	0.87	3.83	-
		R+EPnP	8/21	1.03	4.26	-
		R+oDLT	8/21	0.92	5.25	-
3	28/130	MC	18/23	**0.70**	**1.42**	**0.18**
		2P1k	19/24	0.8	1.82	0.2
		2P10k	19/24	0.76	1.50	5.45
		SVD10k	19/27	1.25	7.08	4.6
		ML10k	13/23	1.05	4.51	13.1
		EProPnP	1/4	11.45	41.22	-
		R+EPro	16/21	5.49	31.47	-
		R+EPnP	0/16	23.27	179.89	-
		R+oDLT	0/22	7.73	5.72	-

Note: For further comparison, we increased the number of iterations in the RANSAC framework, consistent with the extreme outlier rate simulation experiment. R+* iterates 10k times. EPropnp utilizes a GPU, whereas the RANSAC components within Epnp and oDLT are primarily implemented in C++. Therefore, we did not compare their time consumption.



(a) Matching (b) MC
(c) 2P1k (d) 2P10k
(e) SVD10k (f) ML10k

Fig. 8. Visualization of the original feature matching and inlier estimation results for case 1. The red line indicates outliers, while the green line indicates inliers.



(a) failure case 1 (b) failure case 2
(c) failure case 3 (d) failure case 4

Fig. 9. The examples of failure cases. The red line indicates outliers, while the green line indicates inliers.

R+EPro and R+oDLT show significantly reduced capability in comprehensive inlier identification and outlier filtering compared to 2P. This aligns with our conclusion that our proposed MC and 2P-RANSAC methods demonstrate stronger robustness under high outlier conditions.

As shown in the Table. VI, to further compare the performance of the methods under real-world extreme outliers, we detail some typical cases. Fig. 8 shows the inlier visualization in case1. Consistent with the results from the simulation experiments, MC achieves the highest precision rate under extreme outlier rates while maintaining excellent real-time performance. Additionally, EProPnP, et al. are not designed for fisheye cameras, so we use rectified images. Consequently, we have not included visualization of their inliers.

Moreover, as illustrated in Fig. 9, we present an analysis of several failure cases to reveal when localization becomes unreliable, so as to avoid erroneously updating the pose and thus causing potential hazards. However, the experimental comparison results demonstrate that the proposed method has already minimized the probability of localization failure compared to other methods. We observed three typical types of failure cases. The first type occurs when the absolute number of inliers is insufficient to provide the minimum number required by the algorithm. As shown in Fig. 9(a), this inevitably leads to localization failure. The second type arises when the outlier ratio is excessively high, as depicted in Fig. 9(b). In such scenarios, the largest consensus set formed by inliers may be smaller than the largest consensus set formed by outliers. The third type, shown in Fig. 9(c) and Fig. 9(d), involves the retrieval of an entirely incorrect reference image during place recognition.

Finally, as shown in the Fig. 10, we plot the localization trajectory for the math-medium scene. We map the localization error to a color bar, where the bluer the trajectory points, the smaller the error. The trajectory points generated by our proposed MC method are mostly blue, indicating minimal error. The 2P100 method also shows low errors, with its trajectory points being close to blue. The overall trajectory of MC is smoother compared to other methods. Although 2P100 is slightly inferior to MC, its trajectory remains very smooth. In contrast, the trajectory generated by the ML100 method displays a variety of colors, indicating larger errors, and is less smooth than the trajectories of the MC and 2P100 methods.

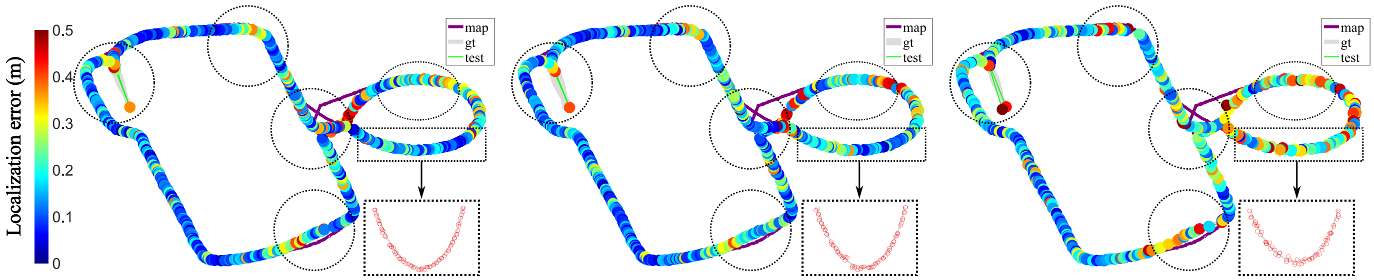


Fig. 10. Comparison of localization trajectories and errors among three methods.

These results demonstrate the excellent performance of our proposed method in real-world scenarios.

C. Application Discussion

The application of visual localization algorithms in autonomous driving navigation is of significant importance. In autonomous driving systems, visual localization plays a critical role in providing real-time position feedback. Any failure in localization can lead to significant positional deviations, jeopardizing the accuracy of downstream tasks such as trajectory planning and optimization [4], [49] and motion planning [50]. Incorrect motion planning can have catastrophic consequences, such as the vehicle steering into the wrong lane and causing accidents. Therefore, ensuring a high success rate for visual localization is paramount for safe driving. Our proposed method aims to improve the robustness and real-time performance of localization. In particular, the MC method further enhances global optimality. Localization systems are typically divided into two parts: the front end and the back end. The front end is responsible for feature extraction and matching, establishing correspondences between the query image and the map. The back end uses robust estimators to solve for the camera pose based on these correspondences. Therefore, the algorithms proposed in this paper are primarily designed for back-end algorithm development. Regarding the significance of the proposed method for practical vehicle applications, the following points should be clarified.

1) Modular Design: The proposed solution is an independent pose estimation module that is easy to integrate and can be compatible with any front-end components. It can significantly enhance performance in existing visual vehicle navigation systems.

2) Motion Adaptation: The 4DoF formulation in our methods aligns well with vehicle motion characteristics. They provide prior values for the pitch and roll angles without the need for measurements or calculations from other sensors.

3) Framework Selection: The two proposed robust solving frameworks exhibit different characteristics in practical applications. As shown in the experiments, the proposed 2P RANSAC can achieve robust localization with high efficiency under an outlier rate below 80%. On the other hand, the proposed deterministic convergence MC framework can achieve robust estimation at higher outlier rate, including extreme cases (80-95%). In practical applications, visual odometry operates at a high frequency, while visual localization is comparatively low-frequency. Visual localization is not required to operate at the same frequency as visual odometry, as its primary function is to provide periodic corrections to the accumulated drift. In our experiments, the MC method achieves a localization frequency of approximately 5Hz, which meets the requirements of typical visual localization applications. When combined with IMU interpolation, the odometry frequency generally reaches above 100Hz [51]. Additionally, the MC algorithm has a significant advantage over RANSAC in terms of stability, as its solutions are not influenced by random factors, ensuring deterministic convergence. This stability is particularly useful for reproducing error scenarios during engineering debugging.

Overall, when the vehicle operates in structured and relatively stable environments, such as urban roads or industrial logistics settings, 2P RANSAC can meet the needs for efficient and robust localization. In unstructured environments with long-term changes and poor feature matching conditions, the MC framework can ensure localization success in real-time estimation scenarios. In practical applications, the system's frontend can roughly estimate the outlier rate during the feature matching process. This allows for selecting the appropriate framework based on the matching conditions and localization success rate requirements.

In addition, the MC framework can be used as the initialization algorithm, while 2P RANSAC serves as the online estimation algorithm in a complete system. The initialization algorithm has higher requirements for success rates, allowing some computational time to ensure successful initialization. In contrast, online estimation demands higher computational real-time performance.

4) Other Applications: Additionally, besides vehicle applications, our algorithms can easily extend to robots, drones, and other applications with the inclusion of an IMU sensor (as validated in Newer College dataset). This is because the prior information on pitch and roll angles can be obtained from inertial measurements. The proposed methods also demonstrate its robustness in highly variable and complex indoor environments (as validated in Matterport360 dataset).

VII. Conclusion And Further Works

In this paper, addressing the limited field of view in pinhole cameras, we formulate a general spherical camera model to accommodate widely used fisheye and 360° panoramic cameras and propose two robust localization methods. Utilizing local planar motion, we introduce the 2P minimal solution algorithm. To cope with the impact of outliers caused by long-term changes, based on the 2P algorithm, we propose the 2P RANSAC framework and a graph-based globally optimal framework to tackle the issue of outlier interference. In summary, this research introduces and validates two robust visual localization methods, enhancing probabilistic RANSAC's outlier robustness from 50% to 80%, and demonstrating the strong robustness and efficiency of the proposed globally optimal framework MC, and supporting various applications, including robotics and drones. It significantly advances localization in complex environments, enriches theoretical research, and provides valuable practical references.

In future work, we plan to conduct data-driven analysis of environmental factors. For example, we can utilize deep learning techniques to detect challenging scenarios based on image complexity, noise levels, and feature point distribution. This analysis will allow us to differentiate localization difficulty before performing visual localization. Based on the deep learning analysis, the system can dynamically adjust the localization algorithm, RANSAC or MC.

REFERENCES

[1] J. Nie et al., "A training-free, lightweight global image descriptor for long-term visual place recognition toward autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1291–1302, Feb. 2024.

[2] T. Wen, K. Jiang, B. Wijaya, H. Li, M. Yang, and D. Yang, "TM^{3}Loc: Tightly-coupled monocular map matching for high precision vehicle localization," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20268–20281, Nov. 2022.

[3] Z. Rozsa, M. Golarits, and T. Sziranyi, "Immediate vehicle movement estimation and 3D reconstruction for mono cameras by utilizing epipolar geometry and direction prior," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23548–23558, Dec. 2022.

[4] R. Chai, A. Tsourdos, S. Chai, Y. Xia, A. Savvaris, and C. L. P. Chen, "Multiphase overtaking maneuver planning for autonomous ground vehicles via a desensitized trajectory optimization approach," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 74–87, Jan. 2023.

[5] Z. Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of large field-of-view cameras for visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 801–808.

[6] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.

[7] G. Pintore, E. Almansa, A. Sanchez, G. Vassena, and E. Gobbetti, "Deep panoramic depth prediction and completion for indoor scenes," *Comput. Vis. Media*, vol. 10, no. 5, pp. 903–922, Oct. 2024.

[8] J. Zhang, C. Ma, K. Yang, A. Roitberg, K. Peng, and R. Stiefelhagen, "Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9478–9491, Jul. 2022.

[9] R. Horaud, B. Conio, O. Leboulleux, and B. Lacolle, "An analytic solution for the perspective 4-point problem," *Comput. Vis., Graph., Image Process.*, vol. 47, no. 1, pp. 33–44, Jul. 1989.

[10] X. Xie and D. Zou, "Depth-based efficient PnP: A rapid and accurate method for camera pose estimation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9287–9294, Nov. 2024.

[11] V. Garro, F. Crosilla, and A. Fusiello, "Solving the PnP problem with anisotropic orthogonal Procrustes analysis," in *Proc. 2nd Int. Conf. 3D Imag., Modeling, Process., Vis. Transmiss.*, Oct. 2012, pp. 262–269.

[12] S. Urban, J. Leitloff, and S. Hinz, "MLPnP—A real-time maximum likelihood solution to the perspective-n-point problem," 2016, *arXiv:1607.08112*.

[13] 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.

[14] F. Camposeco, A. Cohen, M. Pollefeys, and T. Sattler, "Hybrid camera pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 136–144.

[15] T. Huang, Y. Liu, B. Yang, and Y.-H. Liu, "Efficient and globally optimal camera orientation estimation with line correspondences," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 10232–10239, Nov. 2024.

[16] T. Achterberg and R. Wunderling, "Mixed integer programming: Analyzing 12 years of progress," in *Facets of Combinatorial Optimization: Festschrift for Martin Grötschel*. Cham, Switzerland: Springer, 2013, pp. 449–481.

[17] 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, pp. 155–166, Feb. 2009.

[18] A. Penate-Sanchez, J. Andrade-Cetto, and F. Moreno-Noguer, "Exhaustive linearization for robust camera pose and focal length estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2387–2400, Oct. 2013.

[19] S. Li, C. Xu, and M. Xie, "A robust O(n) solution to the perspective-n-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1444–1450, Jul. 2012.

[20] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li, "EPro-PnP: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2781–2790.

[21] S. Henry and J. A. Christian, "Optimal DLT-based solutions for the perspective-n-point," 2024, *arXiv:2410.14164*.

[22] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[23] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1335–1340, Aug. 2006.

[24] C. Jaramillo, L. Yang, J. P. Muñoz, Y. Taguchi, and J. Xiao, "Visual odometry with a single-camera stereo omnidirectional system," *Mach. Vis. Appl.*, vol. 30, nos. 7–8, pp. 1145–1155, Oct. 2019.

[25] B. Solarte, C.-H. Wu, K.-W. Lu, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Robust 360-8PA: Redesigning the normalized 8-point algorithm for 360-FoV images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 11032–11038.

[26] H. Huang and S.-K. Yeung, "360VO: Visual odometry using a single 360 camera," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 5594–5600.

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

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[29] R. Chai, H. Niu, J. Carrasco, F. Arvin, H. Yin, and B. Lennox, "Design and experimental validation of deep reinforcement learning-based fast trajectory planning and control for mobile robot in unknown environment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5778–5792, Apr. 2022.

[30] 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.

[31] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8922–8931.

[32] P. Shi, S. Yan, Y. Xiao, X. Liu, Y. Zhang, and J. Li, "RANSAC back to SOTA: A two-stage consensus filtering for real-time 3D registration," *IEEE Robot. Autom. Lett.*, vol. 9, no. 12, pp. 11881–11888, Dec. 2024.

[33] T. Wei, Y. Patel, A. Shekhovtsov, J. Matas, and D. Barath, "Generalized differentiable RANSAC," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17649–17660.

[34] V. Piedade and P. Miraldo, "BANSAC: A dynamic Bayesian network for adaptive SAmple consensus," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3715–3724.

[35] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1455–1461, Jul. 2017.

[36] T.-J. Chin and D. Suter, *The Maximum Consensus Problem: Recent Algorithmic Advances*. Cham, Switzerland: Springer, 2022.

[37] Z. Liu, B. Guan, Y. Shang, Q. Yu, and L. Kneip, "Line-based 6-DoF object pose estimation and tracking with an event camera," *IEEE Trans. Image Process.*, vol. 33, pp. 4765–4780, 2024.

[38] Y. Khoo and A. Kapoor, "Non-iterative rigid 2D/3D point-set registration using semidefinite programming," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 2956–2970, Jul. 2016.

[39] 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.

[40] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3945–3950.

[41] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Sep. 2007.

[42] L. Ferraz, X. Binefa, and F. Moreno-Noguer, "Very fast solution to the PnP problem with algebraic outlier rejection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 501–508.

[43] R. A. Rossi, D. F. Gleich, and A. H. Gebremedhin, "Parallel maximum clique algorithms with applications to network analysis," *SIAM J. Sci. Comput.*, vol. 37, no. 5, pp. C589–C616, Jan. 2015.

[44] 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.

[45] M. Rey-Area, M. Yuan, and C. Richardt, "360MonoDepth: High-resolution 360deg monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3762–3772.

[46] A. Chang et al., "Matterport3D: Learning from RGB-D data in indoor environments," 2017, *arXiv:1709.06158*.

[47] L. Zhang, M. Camurri, D. Wisth, and M. Fallon, "Multi-camera LiDAR inertial extension to the newer college dataset," 2021, *arXiv:2112.08854*.

[48] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

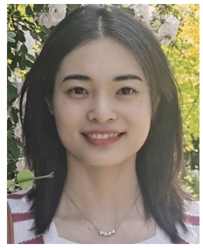
[49] R. Chai et al., "A two phases multiobjective trajectory optimization scheme for multi-UGVs in the sight of the first aid scenario," *IEEE Trans. Cybern.*, vol. 54, no. 9, pp. 5078–5091, Sep. 2024.

[50] R. Chai, Y. Guo, Z. Zuo, K. Chen, H.-S. Shin, and A. Tsourdos, "Cooperative motion planning and control for aerial-ground autonomous systems: Methods and applications," *Prog. Aerosp. Sci.*, vol. 146, Apr. 2024, Art. no. 101005.

[51] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," 2019, *arXiv:1901.03642*.

Yunfeng Lu received the B.S. degree in computer science from Zhejiang Gongshang University, China, in 2022. He is currently pursuing the M.S. degree with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China.

His current research interests include vision-based localization.



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.

Dibin Zhou received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2008.

He is currently with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou. His research interests include machine vision, deep learning, and medical image and engineering computing.

Xiumei Li (Member, IEEE) received the Ph.D. degree in information engineering from Nanyang Technological University, Singapore, in 2010.

She is currently a Professor with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China. Her current research interests include compressed sensing and its applications, time-frequency analysis and its applications, signal detection and estimation, and fast algorithms of signal processing methods.

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 in the Department of Control Science and Engineering, Zhejiang University. His current research interests include mobile robotics and robot perception.