Robust Depth-Aided Visual-Inertial-Wheel Odometry for Mobile Robots

Xinyang Zhao, Qinghua Li, Changhong Wang, *Senior Member, IEEE*, Hexuan Dou, and Bo Liu

***Abstract*—This article introduces visual-depth-inertial-wheel odometry (VDIWO), a robust approach for real-time localization of mobile robots in indoor and outdoor scenarios. Notably, VDIWO achieves accurate localization without relying on prior information. This approach integrates the RGB-D camera, inertial measurement unit, and odometer measurements in a tightly coupled optimization framework. First, we introduce the depth measurement model based on Gaussian mixed model to predict the depth uncertainty of feature points. Then, we propose a hybrid depth estimation method that utilizes both depth measurement fusion and multiview triangulation to estimate the depth of landmarks and simultaneously identify high-quality landmarks. Furthermore, we integrate visual reprojection with depth measurement constraints and odometer preintegration constraints into the tightly coupled optimization framework to further enhance pose estimation accuracy. We evaluate the performance of the VDIWO method using OpenLORIS datasets and real-world experiments. The results demonstrate the high accuracy and robustness of VDIWO for state estimation of mobile robots.**

***Index Terms*—Sensor fusion, simultaneous location and mapping (SLAM), visual-inertial-odometer odometry, wheeled robots.**

I. Introduction

SIMULTANEOUS localization and mapping (SLAM) is the approach to estimate the position and orientation of an agent. It holds a pivotal role in the field of robotics [1], [2], [3]. In recent years, visual SLAM (V-SLAM) and visual odometry (VO) have garnered significant attention. This can be attributed to their advantages, including low cost, compact size, and simplified hardware layout [4].

V-SLAM generally employs VO and visual loop closure techniques to estimate camera ego-motion and construct a precise global map [5]. VO techniques facilitate the estimation of camera pose and localization of landmarks. Simultaneously, visual place recognition provides potential candidates for the loop closure module when the robot revisits known regions [6]. By utilizing these candidates, visual loop closure effectively mitigates localization drift and maintains the consistent generation of maps [7]. Despite the good positioning accuracy of VO or V-SLAM, monocular SLAM methods cannot recover the metric scale of motion and are vulnerable in low-texture, dynamic scenes [8]. To overcome these limitations, researchers have proposed visual-inertial odometry (VIO) methods that leverage both camera and inertial measurement unit (IMU) measurements. Loosely coupled VIO systems, such as [9] and [10] are straightforward. However, they yield suboptimal state estimation due to disregarding correlations between different sensors. Tightly coupled VIO systems fuse measurements from the camera and IMU, simultaneously providing full state estimation. Tightly coupled methods can be divided into two categories [11]: filter-based method and optimization-based method. The former approaches propagate system states using IMU measurements and update them with visual measurements. One notable filter-based method is multistate constraint Kalman filter [12], which employs an efficient extended Kalman filter for state estimation. Based on this foundation, many improved methods [13], [14], [15], [16] have been proposed. Filter-based methods excel in real-time state estimation performance. However, linearization errors may degrade accuracy. Optimization-based methods [17], [18], [19], [20] estimate motion states using batch measurements within a nonlinear optimization framework, offering higher accuracy compared with filter-based methods. Noteworthy optimization-based methods include VINS-Mono [21], VINS-Fusion [22], VI-ORBSLAM [23], ORBSLAM3 [24], etc.

Deploying a monocular VIO system on a wheeled robot makes the system susceptible to additional unobservable directions that result from degenerate motions [25]. These degenerate motions encompass constant speed or acceleration along a straight line and approximate flat ground [26]. To tackle this issue, researchers have proposed integrating direct distance measurements into VIO systems [27]. VINS-RGBD [28], which extends the foundation of VINS-Mono [21] and incorporates the RGB-D camera for ground robots. Although VINS-RGBD effectively addresses scale drift concerns in indoor scenes, it still encounters substantial scale estimation errors in outdoor environments. In addition, this method overlooks the depth uncertainty model when fusing depth measurements, leading to reduced robustness in complex surroundings.

To improve the positioning performance of ground mobile robots, researchers have integrated wheel odometer measurements into VIO systems [29]. Lee et al. [27] proposed a visual-inertial-wheel odometry system. This system performs pose

Manuscript received 7 December 2022; revised 18 May 2023 and 18 August 2023; accepted 29 September 2023. Date of publication 26 October 2023; date of current version 13 April 2024. This work was supported by the Aeronautics Science Foundation under Grant 20175877011. *(Corresponding author: Changhong Wang.)*

The authors are with the Space Control And Inertial Technology Research Center, Harbin Institute of Technology, Harbin 150001, China (e-mail: xinyangzhao@hit.edu.cn; huahit@hit.edu.cn; cwang@hit.edu.cn; douhexuan@hit.edu.cn; 17b904042@stu.hit.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2023.3323731.

Digital Object Identifier 10.1109/TIE.2023.3323731

estimation, incorporates intrinsic and extrinsic calibration of the wheel encoder, and analyzes observability. Liu et al. [30] employed a bidirectional trajectory computation method to tackle the unobservability of extrinsic parameters before the first turning. Another approach, the special euclidean group in two dimensions (SE(2))-constrained VO-based ground vehicle SLAM system, achieves highly accurate 6-D pose estimation [31]. However, this system exhibits exceptional positioning performance exclusively on nearly flat ground, rendering it more suitable for indoor scenarios and unsuitable for complex outdoor road conditions. In addition, the challenge of constructing dense maps using a monocular camera persists, significantly limiting its application in robot navigation.

Motivated by the preceding discussions, this article presents a robust system named visual-depth-inertial-wheel odometry (VDIWO). The system aims to achieve accurate and robust motion estimation through the integration of RGB-D camera and wheel odometer measurements. By formulating a depth measurement model for the RGB-D camera, we propose a Gaussian mixed model (GMM)-based approach to estimate the depth uncertainty associated with varying range depth measurements. We adopt a hybrid depth estimation method to address the limitations of the depth camera's measurement range. This approach capitalizes on both depth measurement fusion and multiview triangulation, effectively estimating feature depth within and beyond the RGB-D camera's measurement range. Simultaneously, it identifies features of high quality and distinguishes outliers. In addition, we incorporate the vehicle's dynamic model to perform preintegration with wheel odometer measurements. Expanding upon conventional tightly coupled VIO optimization methods that rely on 3D–2D reprojection constraints and IMU preintegration constraints, we augment our system by integrating visual reprojection constraints with depth measurements and odometer preintegration constraints. Through the comprehensive fusion, our system achieves excellent positioning accuracy in both indoor and outdoor environments. The main contributions of this article can be summarized as follows.

1) We propose a novel optimization-based tightly coupled odometry that integrates measurements from the RGB-D camera and wheel odometer based on VIO. By incorporating hybrid visual reprojection constraints and wheel odometer pre-integration constraints, our system achieves exceptional accuracy and robustness in motion estimation in both indoor and outdoor environments.
2) To enhance the precision of depth estimation, we introduce the depth measurement model for the RGB-D camera and develop a depth uncertainty estimation method based on GMM. This method efficiently accounts for both spatial and range domains of depth measurements, thereby preserving edge information while minimizing noise.
3) We present a hybrid depth estimation method that utilizes depth measurement fusion and multiview triangulation. This method enables accurate depth estimation for landmarks, and simultaneously identifies high-quality landmarks for back-end optimization.

The rest of this article is organized as follows. Section II describes the main work of our system pipeline. Implementation details are presented in Section III. The experiments and evaluation of the results are shown in Section IV. Finally, Section V conludes this article.

II. System Overview

Fig. 1 illustrates the pipeline of the VDIWO system. The front-end module consists of four parallel threads dedicated to caching measurements from the RGB-D camera, IMU, and wheel odometer. Features are extracted and tracked from RGB images, then aligned with their corresponding depth images. Simultaneously, IMU and wheel odometer measurements are utilized to predict the pose of the incoming frame. Employing a probabilistic depth framework based on GMM, the system accurately quantifies the depth uncertainty associated with these features.

We introduce a hybrid depth estimation method that seamlessly utilizes depth measurement fusion and multiview triangulation, facilitating the estimation of landmarks in the sliding window and the identification of high-quality features. Finally, the visual reprojection constraints with depth measurements and odometer preintegration constraints are integrated into the visual-inertial optimization framework, resulting in a holistic solution that achieves robust state estimation.

III. Methodology

As demonstrated by Wu et al. [25], when implementing the monocular-based VIO system on ground vehicles constrained by nonholonomic motion, the observability of the system's scale factor is inevitably compromised. This limitation arises due to the vehicle's consistent linear acceleration or straight-line movement. To enhance the effectiveness of the VIO system for wheeled robots, we introduce the method that integrates two additional direct measurements: 1) depth measurements from the RGB-D camera and 2) odometer measurements.

A. Depth Measurement Preprocessing

This system utilizes the RealSense D435i camera, an active stereo depth camera with an infrared projector [32]. The infrared projector augments depth accuracy in low-texture environments by projecting infrared patterns. However, based on the active stereo and structured light, the employed depth camera inherently suffers from disparity quantization. In this regard, Ahn et al. [33] conducted a series of depth measurement experiments to establish a depth noise model for the D435 camera. Drawing from this work, we adopt a similar depth uncertainty model for estimating the axis noise of the D435i camera. The model is represented as

$$\sigma_z = a_0 + a_1 z + a_2 z^2 \tag{1}$$

where a_0, a_1, and a_2 correspond to the respective coefficients. This expression fits well the planar residuals in [33] but does not fully consider the lateral error and the depth measurement of its neighbors in a local window.

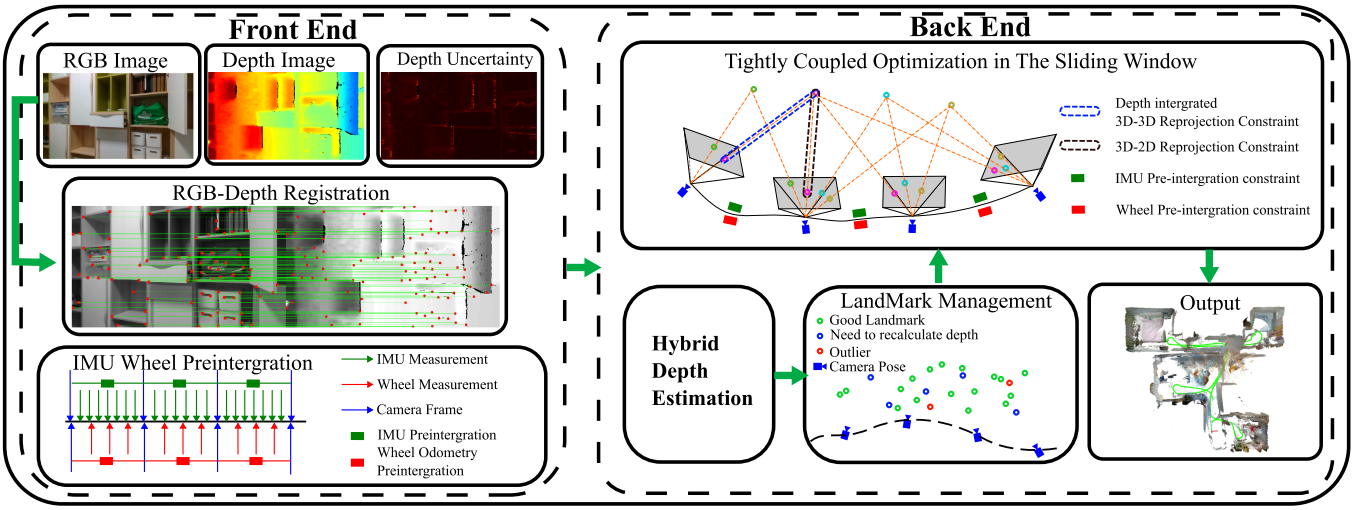


Fig. 1. Pipeline of the proposed VDIWO system, including the front-end and back-end modules.

To address this limitation, we develop a novel approach based on GMM [34]. Our method effectively smooths noise while preserving edges by constructing a kernel function that integrates both spatial and range domains.

Consider a feature point with corresponding pixel $p_k = [u, v, z]$ on the depth image. Here, u and v denote pixel coordinates, while z represents the depth measurement from the RGB-D camera. According to the analysis in [33], z conforms to a Gaussian distribution. To model this distribution, we introduce a random variable $\hat{z}$, representing a mixture of depth measurements within a local window $i :\in [u-1, u+1], j :\in [v-1, v+1]$. The weights assigned to the local window, denoted as W_{ij}, are based on the following Gaussian kernel:

$$W = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{2}$$

where W denotes the Gaussian kernel matrix used to capture the spatial correlation between neighboring pixels. To further extend the range correlation, we introduce an extension method and use range correlation factors to constrain the mixture weights, denoted as w_{ij}. This constraint can be mathematically expressed as

$$w_{ij} = \frac{W_{ij}}{\sum_{i,j} W_{ij}} \exp\left(-\frac{\zeta(z_{ij} - z)^2}{2\sigma_z^2}\right) \tag{3}$$

where σ_z represents the depth uncertainty of p_k, derived from (1). The variable z_{ij} corresponds to the depth measurement of the pixel (i, j) within the local window, while ζ represents the scale factor. The second part of (3) integrates the range correlation component into the weight computation, capturing the similarity among depth measurements.

In flat regions, the depth measurements of adjacent pixels have minimal discrepancies, causing range kernel weights to approach 1. In such cases, Gaussian kernel weights strongly influence the depth estimation. This results in less noisy depth estimation, capitalizing on the favorable attributes of the GMM. Conversely, in regions with edges, the local window contains depth measurements with significant variations. This leads to the range kernel weights approaching 0, which has a negligible effect on the ongoing depth estimation. As a result, the details of the depth measurements are preserved. The estimated depth $\hat{z}$ for p_k can be computed by

$$\hat{z} = \frac{1}{\sum_{i,j} w_{ij}} \sum_{i,j} w_{ij} z_{ij} \tag{4}$$

and the depth uncertainty $\hat{\sigma}_z^2$ for p_k can be expressed as

$$\hat{\sigma}_z^2 = \frac{1}{\sum_{i,j} w_{ij}} \sum_{i,j} w_{ij}(z_{ij}^2 + \sigma_{z_{ij}}^2) - \hat{z}^2 \tag{5}$$

where σ_{ij} represents the depth uncertainty of the pixel (i, j) within the local window, calculated using (1).

To evaluate the effectiveness of our proposed depth uncertainty estimation method, we gathered 900 depth images captured by an RGB-D camera within a static scene. For each pixel in these images, we computed the mean and standard deviation of their corresponding depth measurements. The standard deviation was employed as the ground truth for depth uncertainty. We compared the depth uncertainty predictions derived from different methods, including the simple depth measurement model prediction, GMM, bilateral filter (BF), and our approach, against the ground truth. The root mean square error (RMSE) was computed to quantify disparities between predictions and the ground truth. Fig. 2 shows that our proposed method provides the most accurate uncertainty estimation, particularly along object edges, with the smallest estimation error.

Throughout the depth and depth uncertainty estimation process, we omit pixels with depth measurements exceeding 3 m, those within 0.1 m, or those with missing depth values. Such pixels are assigned a depth value of "0" and excluded from the depth uncertainty estimation process.

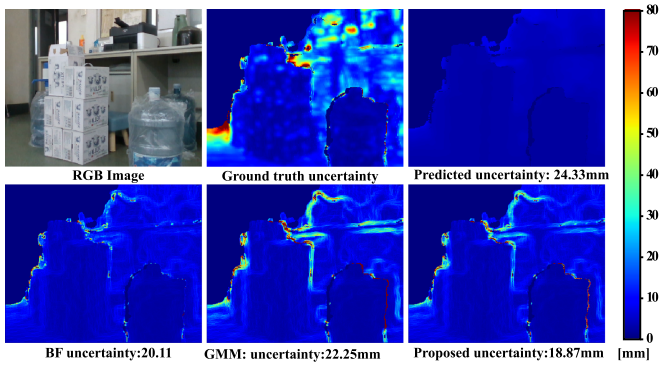


Fig. 2. RMSE of depth uncertainty estimation for four distinct methods: the simple depth measurement model prediction, GMM, BF, and the method proposed in Section III-A.

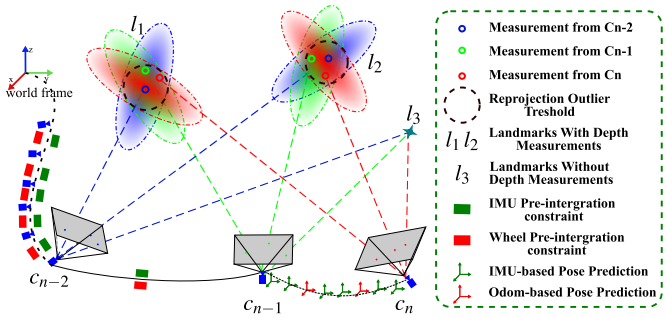


Fig. 3. Illustration of the hybrid depth fusion method. The ellipses represent the uncertainty associated with the same landmark observed from various positions, and the circular distribution symbolizes the acceptable reprojection error, indicated by the thick dashed lines.

B. Hybrid Depth Estimation

The robust initial value is crucial in nonlinear optimization to ensure efficient convergence toward the global optimum. When a new frame arrives, the system recalculates the depth of landmark, thus providing the improved initial value for subsequent nonlinear optimization. As depicted in Fig. 3, the sliding window has n frames. Specifically, the nth frame corresponds to the newly incoming frame, and its state is predicted using IMU and wheel odometer measurements based on the state of $(n-1)$th frame. Subsequently, the state of the nth frame serves as the initial value for the next optimization iteration. Meanwhile, the states of the remaining $n-1$ frames will be updated by optimizing within the sliding window.

On this basis, we employ the hybrid depth estimation approach that utilizes depth measurement fusion and multiview triangulation to jointly estimate the depth of landmarks within the sliding window. This approach illustrated in Fig. 3, the landmarks are categorized as $l^d = \{l_1, l_2\}$ and $l^{2D} = \{l_3\}$ based on their availability of depth measurements. Consider the landmark l_1 within the l^d category. Its 3-D coordinates in the camera frame C_n, corresponding to pixel $p = \{u, v\}$ on an RGB image with depth measurement d, can be given by

$$P_{c_n}^{l_1} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} d(u - c_x)/f_x \\ d(v - c_y)/f_y \\ d \end{bmatrix} \tag{6}$$

where c_x and c_y represent the camera's principal points, while f_x and f_y denote the focal length ratios relative to the image sensor size in the x and y dimensions. We can compute the covariance matrix $\Sigma_{P_{C_n}^{l_1}}$ of $P_{C_n}^{l_1}$ as follows:

$$\Sigma_{P_{C_n}^{l_1}} = \mathbf{J}_{P_{C_n}^{l_1}} \begin{bmatrix} \sigma_p^2 \mathbf{I}_2 & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & \sigma_d^2 \end{bmatrix} \mathbf{J}_{P_{C_n}^{l_1}}^{\mathrm{T}} \tag{7}$$

where σ_p^2 includes the variance values σ_u^2 and σ_v^2. These values remain consistent based on the provided camera parameters and feature detection algorithm. $\mathbf{J}_{P_{C_n}^{l_1}}$ denotes the Jacobian of (6) with respect to the pixel measurement $p = \{u, v, d\}^{\mathrm{T}}$. Utilizing $\mathbf{R}_{c_n}^{c_0}$, which indicates the rotation between the nth frame and the first frame c_0 within the sliding window, we can derive the covariance $\Sigma'_{P_{C_0}^{l_1}}$ of $P_{C_n}^{1}$ in the C_0 frame

$$\mathbf{\Sigma}'_{P_{C_0}^{l_1}} = \mathbf{R}_{c_n}^{c_0} \mathbf{\Sigma}_{P_{C_n}^{1}} \mathbf{R}_{c_n}^{c_0\,\mathrm{T}}. \tag{8}$$

The standard deviation σ'_n for the depth domains of $P_{C_n}^{l_1}$ in C_0 can be obtained by extracting the element from the third row and third column of $\mathbf{\Sigma}'_{P_{C_0}^{l_1}}$. Likewise, when dealing with N projected points of the same landmark within the sliding window, we can estimate the depth

$$\bar{d} = \frac{1}{\sum_{i=1}^{N} \sigma_i'^{-2}} \sum_{i=1}^{N} \frac{d'_i}{\sigma_i'^2} \tag{9}$$

where d'_i represents the depth measurement that has been transformed from C_i to the first frame C_0 within the sliding window.

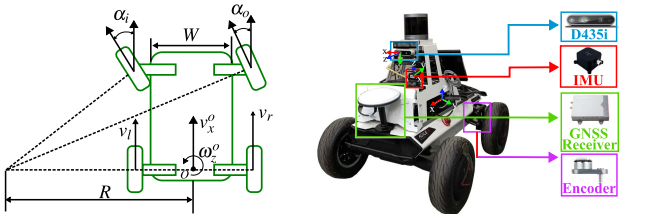
In practical applications, the system employs the reprojection error to serve as a criterion for selecting reliable 3-D features, as depicted by the dashed circle in Fig. 3. Only those 3-D features exhibiting a reprojection error within a predefined range are considered suitable for the depth estimation of landmarks in (9). Multiview observations are utilized to evaluate the accuracy of landmark depth estimation. Let $X_i^k = [X_i^k, Y_i^k, Z_i^k]^{\mathrm{T}}$ denote the 3-D coordinates of the ith landmark at the kth camera frame, and $\bar{X}_i^k = [\bar{X}_i^k, \bar{Y}_i^k, 1]^{\mathrm{T}}$ represent the normalized X_i^k. $T_{c_k}^w$ denotes the transformation matrix from C_k frame to the world coordinate system. By applying the multiview triangulation method, we can obtain

$$\begin{aligned} r_{k_1} &= -\bar{X}_i^k T_3 - T_1 \\ r_{k_2} &= \bar{Y}_i^k T_3 - T_2 \end{aligned} \tag{10}$$

where T_m, $m \in \{1, 2, 3\}$, denotes the mth row of the transformation matrix. The residuals of (10) concerning all observations to the landmark comprise the matrix $M = \{r_{0_1}, r_{0_2}, \ldots, r_{n_1}, r_{n_2}\}^{\mathrm{T}}$. The triangulation error of the depth fusion method can be computed by

$$\epsilon_i = \frac{\|M\bar{d}\bar{X}_0^k\|_2}{n} \tag{11}$$

where $\bar{X}_0^k$ denotes the normalized 3-D coordinates of the first frame within the sliding window that observed the landmark. The landmark is labeled as "good" when ϵ_i falls within the



(a) (b)

Fig. 4. (a) Kinematic model of the Ackermann drive wheel robot, W denotes the baseline of the robot. α_i and α_o are the steering angles of the inner and outer wheels with respect to the straight ahead direction, respectively. R is the steering radius. (b) Experimental platform: the Ackermann drive wheel robot and onboard sensors.

predefined threshold ϵ_{th}, and the number of tracking occurrences exceeds 4. Alternatively, the landmark is labeled as "incomplete" if these conditions are not met. Those labeled as "incomplete" will undergo the hybrid depth estimation process in the subsequent loop, while "good" landmarks will not. This strategy ensures both depth estimation accuracy and conservation of computational resources.

C. Wheel Odometer Preintegration and Noise Propagation

In this system, we utilize the preintegrated measurements from the IMU and wheel odometer to predict the transformation between consecutive keyframes in the sliding window. The preintegration of the IMU has been thoroughly examined in [21, Sec. IV-B]. In this section, we detail preintegration derivation for the wheel odometer. As depicted in Fig. 4(a), we illustrate the kinematic model of the Ackermann drive wheeled mobile robot as follows:

$$\begin{bmatrix} v_x^o \\ \omega_z^o \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{W} & \frac{1}{W} \end{bmatrix} \begin{bmatrix} v_l \\ v_r \end{bmatrix} \tag{12}$$

where W denotes the baseline of the robot, and v_l and v_r stand for the velocities of the left and right wheels, respectively. The notation $(\cdot)^o$ signifies the robot odometer frame, while O denotes the origin of the odometer coordinate system. The raw odometer measurements are $\tilde{\boldsymbol{v}}^o = \begin{bmatrix} \boldsymbol{v}_x & 0 & 0 \end{bmatrix}$ and $\tilde{\boldsymbol{\omega}}^o = \begin{bmatrix} 0 & 0 & \boldsymbol{\omega}_z^o \end{bmatrix}$, subject to zero-mean Gaussian noise characterized by $\boldsymbol{n}^v \sim \mathcal{N}(0, \delta_v^2)$, and $\boldsymbol{n}^\omega \sim \mathcal{N}(0, \delta_\omega^2)$.

Inspired by the IMU preintegration presented in [21], we adopt a similar approach to preintegrate the odometer measurements between consecutive image frames for predicting camera transformation. Given that the frequency of odometer measurements is significantly higher (approximately 50 Hz) in comparison to image frames, we define the discrete-time formulation of preintegration terms as follows:

$$\begin{aligned} \boldsymbol{\alpha}_{o_j}^{o_i} &= \sum_{k=i+1}^{j} \bar{\boldsymbol{v}}_k^o \Delta t = \sum_{k=i+1}^{j} \frac{1}{2} \left(\boldsymbol{R}_{k-1}^{o_i} \left(\tilde{\boldsymbol{v}}_{k-1}^o - \boldsymbol{n}_{k-1}^v \right) \right. \\ &\quad \left. + \boldsymbol{R}_k^{o_i} \left(\tilde{\boldsymbol{v}}_k^o - \boldsymbol{n}_k^v \right) \right) \Delta t \end{aligned}$$

$$\begin{aligned} \boldsymbol{\gamma}_{o_j}^{o_i} &= \prod_{k=i+1}^{j} \operatorname{Exp}\left(\bar{\boldsymbol{\omega}}_k^o \Delta t \right) \\ &= \prod_{k=i+1}^{j} \operatorname{Exp}\left(\frac{1}{2} \left(\tilde{\boldsymbol{\omega}}_{k-1}^o + \tilde{\boldsymbol{\omega}}_k^o - \boldsymbol{n}_{k-1}^\omega - \boldsymbol{n}_k^\omega \right) \Delta t \right) \end{aligned} \tag{13}$$

where $\boldsymbol{\alpha}_{o_j}^{o_i}$ and $\boldsymbol{\gamma}_{o_j}^{o_i}$ represent the preintegration of translation and rotation, respectively, from the ith frame to the jth frame in the odometer coordinate system. The above equation for odometer preintegration is based on the assumption of the ideal scenario where noise has been eliminated. To compute the covariance of the preintegrated odometer, we can analyze the noise propagation of (13) as follows: (14) shown at the bottom of the next page, where $\boldsymbol{J}_r^k = \boldsymbol{J}_r^k(\frac{1}{2}(\tilde{\boldsymbol{\omega}}_{k-1}^o + \tilde{\boldsymbol{\omega}}_k^o)\Delta t)$ is the right Jacobian of $\mathrm{SO}(3)$. The covariance of the preintegrated odometer measurement noise can be obtained iteratively using $\boldsymbol{F}$ and $\boldsymbol{G}$

$$\boldsymbol{\Sigma}_{\mathcal{O}_k^{O_i}} = \mathbf{F} \boldsymbol{\Sigma}_{\mathcal{O}_{k-1}^{O_i}} \mathbf{F}^{\mathrm{T}} + \mathbf{G} \boldsymbol{\Sigma}_\eta \mathbf{G}^{\mathrm{T}}. \tag{15}$$

The initial covariance is set as $\boldsymbol{\Sigma}_{\mathcal{O}_i^{O_i}} = \mathbf{0}_{6\times 6}$, and the noise covariance matrix is defined as $\boldsymbol{\Sigma}_\eta = \operatorname{diag}(\boldsymbol{\sigma}_v^2, \boldsymbol{\sigma}_\omega^2, \boldsymbol{\sigma}_v^2, \boldsymbol{\sigma}_\omega^2)$.

D. Tightly Coupled State Optimization

In this article, a tightly coupled nonlinear optimization approach is employed to achieve accurate pose estimation for a wheeled robot. We extend the conventional VIO optimization framework to enhance the localization performance of the wheeled robot. This extension involves the incorporation of visual reprojection constraints with depth measurements and odometer preintegration constraints.

We perform state variable optimization using bundle adjustment within the sliding window framework, aiming to minimize the following cost function:

$$\begin{aligned} \min_{\boldsymbol{x}} \Bigg\{ & \left\| \mathbf{r}_p - \mathbf{H}_p \boldsymbol{x} \right\|^2 + \sum_{i \in \beta} \left\| \mathbf{r}_{\text{imu}} \left(\hat{\mathbf{z}}_{b_{i+1}}^{b_i}, \boldsymbol{x} \right) \right\|_{\Sigma_{b_{i+1}}^{b_i}}^2 \\ & + \sum_{i \in \beta} \left\| \mathbf{r}_{\text{odom}} \left(\hat{\mathbf{z}}_{o_{i+1}}^{o_i}, \boldsymbol{x} \right) \right\|_{\Sigma_{o_{i+1}}^{o_i}}^2 \\ & + \sum_{\substack{j \in \text{obs}(l), l \in L \\ 0.1 < \bar{\mathbf{d}}_l^{c_j} <= 3}} \rho \left\| \mathbf{r}_{3d3d} \left(\hat{\mathbf{z}}_l^{c_j}, \bar{\mathbf{d}}_l^{c_j}, \boldsymbol{x} \right) \right\|_{\Sigma_{l_d}^{c_j}}^2 \\ & + \sum_{\substack{j \in \text{obs}(l), l \in L \\ \bar{\mathbf{d}}_l^{c_j} < 0.1; \bar{\mathbf{d}}_l^{c_j} > 3}} \rho \left\| \mathbf{r}_{3d2d} \left(\hat{\mathbf{z}}_l^{c_j}, \boldsymbol{x} \right) \right\|_{\Sigma_l^{c_i}}^2 \Bigg\} \end{aligned} \tag{16}$$

where $\boldsymbol{x}$ represents the estimated state variable, while the Mahalanobis norm $\| \cdot \|^2$, weighted by covariance Σ, is utilized. The prior information, represented by $[\mathbf{r}_p, \mathbf{H}_p]$, is obtained through marginalization. The Huber norm ρ is employed to mitigate the impact of mismatching. The parameter β

represents the frames in the sliding window, and obs(l) indicates the covisible keyframes where the lth landmark is observable. The residual components linked with IMU and odometer measurements are, respectively, denoted by $\mathbf{r}_{\text{imu}}$ and $\mathbf{r}_{\text{odom}}$. Enhancements to the 3D–2D reprojection error term are realized through the introduction of $\mathbf{r}_{3d3d}$, which is constrained by depth measurements, alongside the conventional $\mathbf{r}_{3d2d}$. Covariance matrices, such as $\Sigma_{b_{i+1}}^{b_i}$, $\Sigma_{o_{i+1}}^{o_i}$, $\Sigma_{l_d}^{c_j}$, and $\Sigma_l^{c_i}$ correspond to the residuals within the optimization framework.

The explicit definitions of prior and IMU terms have been presented in [21]. The odometer and visual terms are described as follows.

1) Odometer Residual: The residuals originating from the preintegrated odometer measurements establish constraint factors connecting consecutive keyframes b_i and b_{i+1}. The residual term can be mathematically defined as follows:

$$\mathbf{r}_{\text{odom}}\left(\hat{\mathbf{z}}_{o_{i+1}}^{o_i}, \boldsymbol{x}\right) = \begin{bmatrix} [\mathbf{R}_{b_i}^w \mathbf{R}_o^b]^{\mathrm{T}} (\mathbf{R}_{b_j}^w \mathbf{P}_o^b + \mathbf{P}_{b_j}^w - \mathbf{R}_{b_i}^w \mathbf{P}_o^b - \mathbf{P}_{b_i}^w) - \tilde{\boldsymbol{\alpha}}_{o_{i+1}}^{o_i} \\ \text{Log}(\tilde{\boldsymbol{\gamma}}_{o_{i+1}}^{o_i} [\mathbf{R}_{b_j}^w \mathbf{R}_o^b]^{\mathrm{T}} \mathbf{R}_{b_i}^w \mathbf{R}_o^b) \end{bmatrix} \tag{17}$$

where $\tilde{\boldsymbol{\alpha}}_{o_{i+1}}^{o_i}$ and $\tilde{\boldsymbol{\gamma}}_{o_{i+1}}^{o_i}$ represent the preintegration of translation and rotation. $\mathbf{R}_{b_i}^w$ and $\mathbf{P}_{b_i}^w$ are the rotation and translation of the body frame b_i relative to the world coordinate within the sliding window. The logarithm map $\text{Log}(\cdot)$ associates a rotation matrix $\mathbf{R} \in \mathbf{SO}(3)$ to the Lie algebra $\xi \in \mathfrak{so}(3)$.

2) Hybrid Visual Residual: For the lth 3-D landmark in the sliding window, initially observed as $\hat{\mathbf{z}}_l^{c_i}$ by the camera in the ith frame, the observation in the jth frame is denoted as $\hat{\mathbf{z}}_l^{c_j} = [\tilde{\mathbf{u}}_l^{c_j}, \tilde{\mathbf{v}}_l^{c_j}, \tilde{\mathbf{d}}_l^{c_j}]^{\mathrm{T}}$, where $\tilde{\mathbf{d}}_l^{c_j}$ denotes the corresponding depth measurement. To effectively integrate depth information, we employ the depth-constraint 3D–3D reprojection error term when the measured depth $\bar{\mathbf{d}}_l^{c_j}$ satisfies the condition

$$\eta_1 < \bar{\mathbf{d}}_l^{c_j} < \eta_2 \tag{18}$$

where η_1 and η_2 denote the minimum and maximum depth thresholds (set as $\eta_1 = 0.1$ m and $\eta_2 = 3$ m based on trial-and-error). The definition of the depth constraint 3D–3D reprojection error term is as follows:

$$\begin{aligned} \mathbf{r}_{3\,d3d}\left(\hat{\mathbf{z}}_l^{c_j}, \tilde{\mathbf{d}}_l^{c_j}, \boldsymbol{x}\right) &= \begin{bmatrix} \left[\mathbf{p}_l'^{c_j}\right]_1 / \left[\mathbf{p}_l'^{c_j}\right]_3 - \left[\tilde{\mathbf{p}}_l^{c_j}\right]_1 \\ \left[\mathbf{p}_l'^{c_j}\right]_2 / \left[\mathbf{p}_l'^{c_j}\right]_3 - \left[\tilde{\mathbf{p}}_l^{c_j}\right]_2 \\ 1/\left[\mathbf{p}_l'^{c_j}\right]_3 - 1/\tilde{\mathbf{d}}_l^{c_j} \end{bmatrix} \\ \tilde{\mathbf{p}}_l^{c_j} &= \pi^{-1}\left(\left[\hat{\mathbf{z}}_l^{c_j}\right]_{1,2}\right) \\ \mathbf{P}_l'^{c_j} &= (\mathbf{R}_c^b)^{\mathrm{T}} \left((\mathbf{R}_{b_j}^w)^{\mathrm{T}} \left(\mathbf{R}_{b_i}^w \left(\mathbf{R}_c^b \frac{1}{\boldsymbol{\lambda}_l} \pi^{-1}\left(\begin{bmatrix} \tilde{\mathbf{u}}_l^{c_i} \\ \tilde{\mathbf{v}}_l^{c_i} \end{bmatrix}\right) + \mathbf{p}_c^b\right) + \mathbf{p}_{b_i}^w\right) - \mathbf{p}_{b_j}^w\right) - \mathbf{p}_c^b \end{aligned} \tag{19}$$

where $[\cdot]_k$ represents kth row of the vector. π^{-1} stands for the backprojection function of camera, which is determined by the intrinsic calibration parameters. $\mathbf{p}_c^b$ and $\mathbf{R}_c^b$, respectively, indicate the translation and rotation from the camera's coordinate system to that of the IMU. These values are precalibrated and remain constant throughout the optimization process.

For observations without depth measurements or when the depth measurement $\bar{\mathbf{d}}_l^{c_j}$ fails to meet the criteria stated in (18), we employ the 3D–2D reprojection error term. This term is formulated as follows:

$$\mathbf{r}_{3d2d}\left(\hat{\mathbf{z}}_l^{c_j}, \boldsymbol{x}\right) = \begin{bmatrix} \left[\mathbf{p}_l'^{c_j}\right]_1 / \left[\mathbf{p}_l'^{c_j}\right]_3 - \left[\tilde{\mathbf{p}}_l^{c_j}\right]_1 \\ \left[\mathbf{p}_l'^{c_j}\right]_2 / \left[\mathbf{p}_l'^{c_j}\right]_3 - \left[\tilde{\mathbf{p}}_l^{c_j}\right]_2 \end{bmatrix}. \tag{20}$$

IV. Experiment and Discussion

In this section, the performance of our system is evaluated through experiments on the public OpenLORIS-Scene dataset [35] and real-world scenarios. Our approach is compared against five state-of-the-art SLAM methods: OpenVINS [16], ORBSLAM3 [Monocular (IMU) mode and RGB-D mode] [24], VINS-Fusion [22], VINS-RGBD [28], and RTABMAP (RGB-D/IMU/Wheel mode) [36].

A. Accuracy Evaluation Test on OpenLORIS-Scene Dataset

We evaluated the accuracy of the VDIWO method using the OpenLORIS-Scene dataset. This dataset was recorded by a wheeled robot equipped with an RGB-D camera, IMU, and wheel odometer. The ground truth trajectories were obtained

$$\begin{bmatrix} \delta\boldsymbol{\alpha}_k^{o_i} \\ \delta\boldsymbol{\gamma}_k^{o_i} \end{bmatrix} = \boldsymbol{F} \begin{bmatrix} \delta\boldsymbol{\alpha}_{k-1}^{o_i} \\ \delta\boldsymbol{\gamma}_{k-1}^{o_i} \end{bmatrix} + \boldsymbol{G} \begin{bmatrix} \boldsymbol{n}_{k-1}^v \\ \boldsymbol{n}_{k-1}^\omega \\ \boldsymbol{n}_k^v \\ \boldsymbol{n}_k^\omega \end{bmatrix}$$

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{I}_{3\times3} & -\frac{1}{2}\left(\Delta\tilde{\boldsymbol{R}}_{k-1}^{O_i}\left(\tilde{\boldsymbol{v}}_{k-1}^o\right)^\wedge + \Delta\tilde{\boldsymbol{R}}_k^{O_i}\left(\tilde{\boldsymbol{v}}_k^O\right)^\wedge \Delta\tilde{\boldsymbol{R}}_{O_i}^{k-1^T}\right)\Delta t \\ \boldsymbol{0}_{3\times3} & \tilde{\boldsymbol{R}}_{O_k}^{k-1^T} \end{bmatrix}$$

$$\boldsymbol{G} = \begin{bmatrix} \frac{\Delta t}{2}\Delta\tilde{\boldsymbol{R}}_{k-1}^{O_i} & -\frac{\Delta t^2}{4}\Delta\tilde{\boldsymbol{R}}_k^{O_i}\tilde{\boldsymbol{v}}_k^{O^\wedge}\boldsymbol{J}_r^k & \frac{\Delta t}{2}\Delta\tilde{\boldsymbol{R}}_k^{O_i} & -\frac{\Delta t^2}{4}\Delta\tilde{\boldsymbol{R}}_k^{O_i}\tilde{\boldsymbol{v}}_k^{O^\wedge}\boldsymbol{J}_r^k \\ \boldsymbol{0}_{3\times3} & \frac{1}{2}\boldsymbol{J}_r^k\Delta t & \boldsymbol{0}_{3\times3} & \frac{1}{2}\boldsymbol{J}_r^k\Delta t \end{bmatrix} \tag{14}$$

TABLE I
ABLATION EXPERIMENT RESULTS OF RMSE OF ATE ON OPENLORIS-SCENE DATASET (UNIT:M)

Datasets	WHEEL	WHEEL IMU	BASELINE	VDIO w/o TC	VDIO	VDIWO
Home1-1	0.983	0.679	0.692	0.588	0.455	**0.316**
Home1-2	1.005	0.552	6.710(drift)	0.779	0.772	**0.367**
Home1-3	0.494	0.474	0.572	0.339	0.261	**0.235**
Home1-4	0.671	0.391	0.885	0.617	0.455	**0.310**
Home1-5	0.235	0.227	0.208	0.160	0.158	**0.152**
Cafe 1	1.571	0.655	0.740	0.604	0.525	**0.407**
Cafe 2	2.030	0.952	1.767	0.573	0.533	**0.488**
Corridor 1	17.894	11.021	3.813	2.865	**2.642**	3.044
Average	3.110	1.871	1.239	0.816	0.725	**0.670**

The bold value indicates the best result.

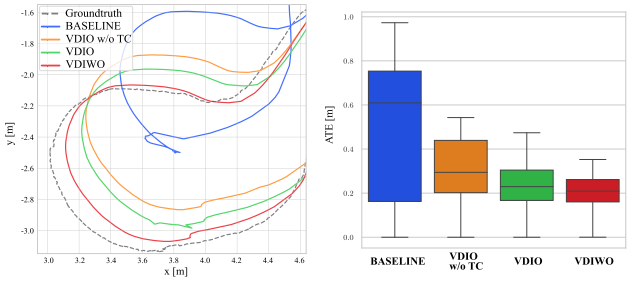


Fig. 5. Trajectories and boxplots comparison on Home 1-3 dataset. (a) Trajectory (b) Boxplot of ATE.

through the OptiTrack MCS system and Hokuyo UTM-30LX LiDAR sensors mounted on the robot.

1) Ablation Experiment: This section evaluates the efficacy of integrating hybrid depth estimation and wheel odometer into the VIO system. We compare the performance of six configurations: 1) WHEEL, using only wheel integration; 2) WHEEL+IMU, employing the extend Kalman filter-based fusion of wheel and IMU measurements; 3) BASELINE, representing the VDIWO method without depth and wheel measurements, including only the multiview triangulation-based landmark classification (TC) approach; 4) VDIO without TC, indicating the BASELINE method with hybrid depth estimation but without TC; 5) VDIO, signifying the BASELINE method with full hybrid depth estimation; 6) VDIWO method. Table I presents the results of experiments. The VDIWO method achieves an average RMSE of 0.67 m, a significant reduction of 45% compared with the BASELINE method. VDIO also demonstrates improved performance, though to a lesser extent than VDIWO. After removing TC, VDIO's average RMSE increased by 12.5%. Despite wheel and IMU integration advancements over the wheel-only method, it still falls short of the performance of the tightly coupled VDIWO method. Fig. 5(a) and (b) depict the estimated trajectories and the boxplot of absolute trajectory error (ATE) for the four configurations on the Home 1-3 datasets to further illustrate the performance improvement of VDIWO. It can be seen that the estimated trajectory of VDIWO is close to the ground truth and exhibits the smallest distribution of ATE.

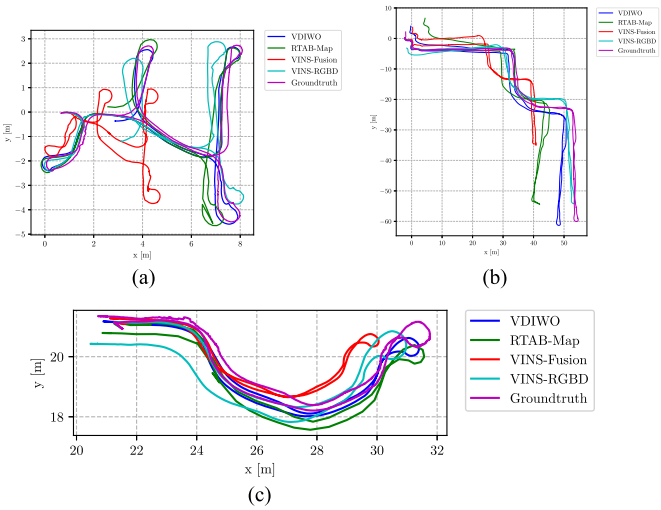


Fig. 6. Comparison of the estimated trajectories from VDIWO, VINS-Fusion, VINS-RGBD, and RTABMAP in sequences Home 1-1, Corridor 1-1, and Cafe 1-1. (a) Home 1-1. (b) Corridor 1-1. (c) Cafe 1-1.

2) Comparison With Other SLAM Systems: The experiments evaluated OpenVINS, ORBSLAM3, VINS-Fusion, VINS-RGBD, RTABMAP, and VDIWO in Home, Cafe, and Corridor scenes. To better evaluate the localization performance, all methods were executed without the loop closure module, except RTABMAP, which relies on loop closure. The ATE and relative trajectory error (RTE) [37] were computed for each sequence, and the results are presented in Table II. VDIWO has the best performance for both ATE and RTE in most sequences. While RTABMAP showed slightly superior ATE in Home 1-3, its RTE was worse. RTABMAP performs better in small range of scenarios, such as in Home 1-5. Interestingly, despite RTABMAP having a loop closure module, it failed in all OpenLORIS datasets, possibly due to the different starting and ending viewpoints. VINS-RGBD, utilizing RGB-D camera and IMU, performed well in Corridor 1-1 with only a slight increase in ATE compared with VDIWO. VINS-Fusion ran successfully in most sequences, while OpenVINS exhibited frailty in low-texture environments and failed to run in over half of the home and corridor sequences. Notably, when OpenVINS succeeded, its error was reduced by over half against VINS-Fusion, even surpassing VDIWO in the Cafe 2 sequence with lower ATE. In terms of ORBSLAM3, both monocular inertial and RGB-D modes cannot even complete initialization in Home and Corridor scenes. Although initialization succeeded in the Cafe scene, the ATE exceeded that of VDIWO. Fig. 6 depicted the estimated trajectories in Home 1-1, Cafe 1-1, and Corridor 1-1. The trajectory estimated by VDIWO aligned most closely with ground truth and exhibited the smallest relative translation and rotation error distribution, as demonstrated by the boxplots in Fig. 7.

Fig. 8 visualizes feature extraction and point cloud maps. Utilizing the Point Cloud Library, we combined estimated states with corresponding depth maps to visually depict trajectory estimation accuracy.

TABLE II
ATE AND RTE PERFORMANCE WITH DIFFERENT SLAM METHODS TESTED ON OPENLORIS-SCENE DATASET (UNIT:M)

Sequences	Length(m)	OpenVINS (Mono+IMU)		VINS-Fusion (Mono+IMU)		VINS-RGBD		ORBSLAM3-M1 (Mono+IMU)		ORBSLAM3-D2 (RGBD)		RTABMAP		VDIWO		Environment Description
		ATE.	RTE.	ATE.	RTE.	ATE.	RTE.	ATE.	RTE.	ATE.	RTE.	ATE.	RTE.	ATE.	RTE.	
Home 1-1	40.419	×		2.593	0.691	0.624	0.161	×		×		0.462	0.132	**0.316**	**0.092**	Illumination changing, Dynamic objects, Violent rotation, low texture
Home 1-2	31.321	×		2.190	0.325	0.784	0.182	×		×		0.403	0.163	**0.367**	**0.090**	
Home 1-3	23.830	×		×		0.441	0.179	×		×		**0.233**	0.130	0.235	**0.091**	
Home 1-4	21.057	1.044	0.245	1.957	0.175	0.696	0.207	×		×		0.429	0.186	**0.310**	**0.077**	
Home 1-5	5.898	0.399	0.708	0.702	1.133	0.252	0.208	×		×		**0.107**	**0.066**	0.152	0.077	
Cafe 1	30.973	0.471	0.153	0.967	0.275	0.793	0.170	2.827	1.002	0.742	0.167	0.545	0.165	**0.407**	**0.117**	Dynamic objects, Lots of People
Cafe 2	47.143	**0.415**	0.160	1.561	0.254	0.780	0.163	0.716	0.249	0.540	0.147	0.632	0.141	0.488	**0.111**	
Corridor 1	231.465	×		15.328	0.641	3.534	0.131	×		×		10.234	0.113	**3.044**	**0.075**	low texture

The bold value indicates the best result.

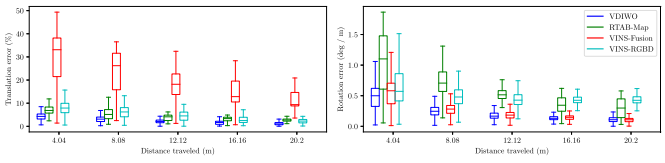


Fig. 7. Relative translation and rotation error of VDIWO, RTABMAP, VINS-Fusion, and VINS-RGBD for the Home 1-1 sequence.

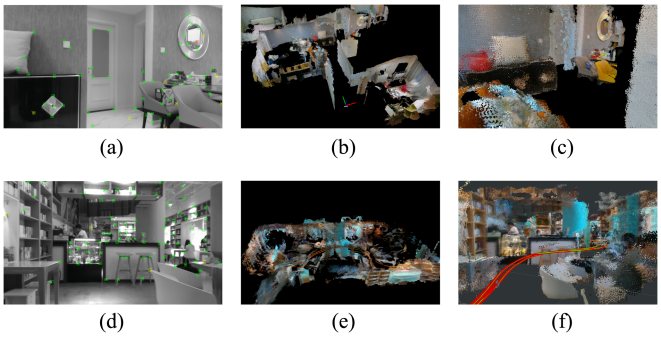


Fig. 8. (a)–(c) And (d)–(f) illustrate the visualization of feature extraction and point cloud map in Home 1-1 and Cafe 1-1, respectively. (b) And (e) are viewed from the top, and (c) and (f) are viewed from inside.

B. Real-World Experiments

This section focuses on evaluating the performance of VDIWO through real-world experiments conducted in various indoor and outdoor environments.

1) Experiment Platform: Fig. 4(b) illustrates the experimental platform, Agilex Ackermann drive wheel robot, employed in our experiments. This robot integrates a D435i camera, which captures color and aligned depth images at 30 Hz. Odometer measurements, at 50 Hz, are obtained via two Hall sensors within the brushless motors. An IMU sensor (CH110) outputs data at 400 Hz. The robot operates on the Ubuntu 20.04 system and utilizes an onboard Intel NUC with an i7-1165 CPU. For outdoor scenarios, the Sino M100 system (RTK+GPS+IMU) offers ground truth with exceptional positioning precision within 2 cm. In indoor scenarios, where GPS signals are unavailable, the accuracy of the proposed methods in terms of start-to-end drift estimation is evaluated using Apritag markers.

TABLE III
ATE PERFORMANCE WITH DIFFERENT SLAM METHODS ON REAL-WORLD EXPERIMENTS (UNIT:M)

Sequences (Length m)	WHEEL	WHEEL +IMU	VINS-Fusion	VINS-RGBD	RTAB MAP	VDIO	VDIWO
Corridor(82.13)	2.188	0.953	1.595	0.725	3.267	**0.564**	**0.418**
Room(67.71)	2.582	1.527	1.232	1.068	**0.163***	0.690	**0.438**
Garden(178.46)	4.754	2.910	6.274	7.666	3.421	**2.479**	**1.703**
Road(420.59)	8.219	**7.695**	28.537	39.817	11.108	19.675	**6.990**
Park(488.41)	41.624	11.892	44.303	84.763	**6.627***	23.981	**8.901**

The best result is highlighted in bold with **black**. The second-best result is highlighted in bold with **blue**. * point out that loop closing happened.

2) Indoor Experiment: Indoor experiments were conducted in corridor and room scenes, depicted in Fig. 9(a)–(b) and (f)–(g). The wheeled robot was guided to drive in these areas and returned to the start point. We evaluated start-to-end drift to assess the system's performance, with results presented in Table III. The Corridor sequence, featuring low-texture and repetitive scenes, served as a robustness test. VDIWO outperformed with a translation drift of 0.418 m, followed by VDIO at 0.564 m. Notably, RTABMAP failed in loop closure, leading to conspicuous drift due to sharp rotations at the end of the corridor. In the room sequence, RTABMAP achieved minor translation drift (0.163 m) due to successful loop closure, though VDIWO still performed better. VINS-Fusion slightly outperformed the wheel-only method, relying solely on monocular and IMU.

3) Outdoor Experiment: The efficacy of the proposed VDIWO method was also evaluated outdoors. Fig. 9(c)–(e) shows the garden, road, and park scenes, with corresponding trajectories in Fig. 9(h)–(j). In the garden dataset, the trajectory of VDIWO is closely aligned with the ground truth, with an APE of 1.703 m, slightly larger for VDIO (2.479 m). VINS-RGBD and VINS-Fusion exhibited significant scale drift, accentuated by rough terrains in the road and park sequences. This led to an increase in trajectory estimation errors across all methods. Successful loop closure enabled RTABMAP to perform best in the park scene (APE of 6.627 m), followed closely by VDIWO (8.901 m). Remarkably, WHEEL+IMU surpassed many visual-inertial methods outdoors, potentially due to lower quality features for state estimation of VIO.

Experiments demonstrate that VDIWO outperforms other methods in both indoor and outdoor scenarios, showcasing

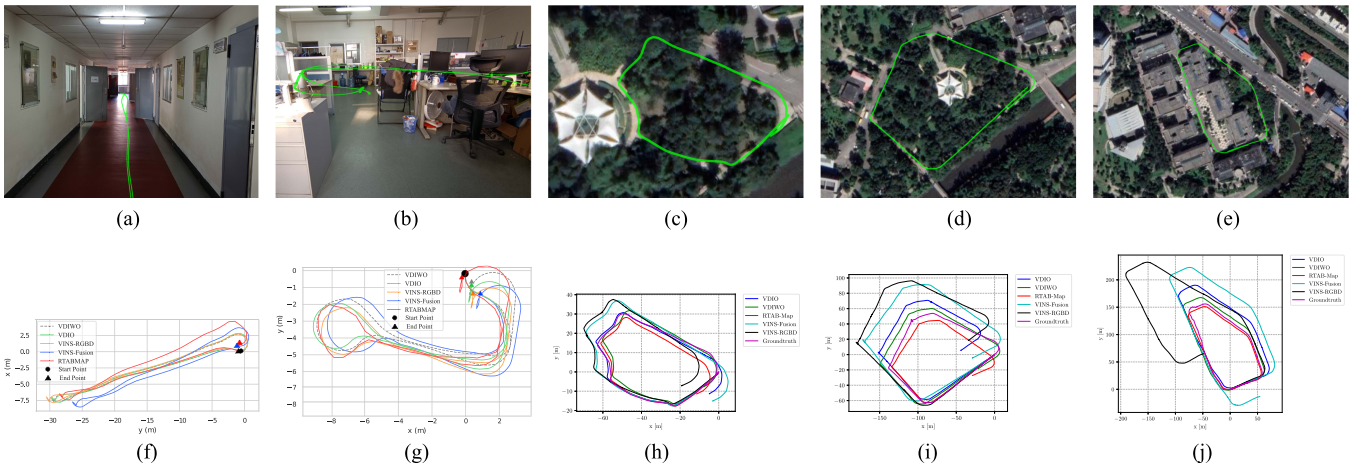


Fig. 9. (a)–(e) Illustrate the trajectories of VDIWO in the corridor, room, garden, road, and park scenes, respectively. (f)–(j) show the comparison of trajectories produced by VINS-Fusion, VINS-RGBD, RTABMAP, VDIO, VDIWO, and the ground truth.

superior accuracy and robustness. Notably, even in instances where RTABMAP loop closure succeeds, the performance of VDIWO remains closer, further proving its applicability for wheeled robots.

4) Potential Implementation for Nonplanar Scenes: The method proposed in this article focuses primarily on approximating planar environments. Several enhancements can be incorporated to extend its applicability to broader scenarios, including situations, such as parking buildings with slopes. First, the detection of ground slope changes can be achieved through IMU measurements. The wheel odometer preintegration constraints will be discarded when the system encounters significant slope changes. Second, the method proposed in [38] can be used to segment the ground into different planes according to the slope change using LiDAR, and subsequently, the parameters of these planes can be utilized to constrain the state of the system. Third, the polynomial parameterization method proposed in [39] can be used to approximate the 6-D motion manifold of ground robots to enhance the adaptability of the system for localization in complex environments. By integrating these strategies, the system can effectively cope with various nonplanar environments.

V. Conclusion

This article presented VDIWO for mobile robots, which estimated the robot state by integrating RGB-D camera, IMU, and odometer measurements in a tightly coupled scheme. By incorporating the depth measurement model, the hybrid depth estimation method could accurately estimate the depth of landmarks, and simultaneously identify high-quality landmarks. In the backend, the visual reprojection constraints with depth measurements and odometer preintegration constraints were integrated to further refine the pose estimation. Extensive evaluation of the OpenLORIS datasets demonstrated the excellent performance of VDIWO, whereas real-world experiments validate its enhanced robustness. Consequently, VDIWO exhibited significant potential for facilitating the autonomous navigation of wheeled robots in both indoor and outdoor environments.

In future work, we will incorporate LiDAR measurements into the VDIWO framework to obtain more accurate environmental information, thereby achieving a complete SLAM system with high performance, especially in complex scenarios.

References

[1] G. Huang, "Visual-inertial navigation: A concise review," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 9572–9582.

[2] T.-J. Lee, C.-H. Kim, and D.-I. D. Cho, "A monocular vision sensor-based efficient SLAM method for indoor service robots," *IEEE Trans. Ind. Electron.*, vol. 66, no. 1, pp. 318–328, Jan. 2019, doi: 10.1109/TIE.2018.2826471.

[3] Y. Zhou et al., "Toward autonomy of micro aerial vehicles in unknown and global positioning system denied environments," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7642–7651, Aug. 2021, doi: 10.1109/TIE.2020.3008378.

[4] S.-Y. Hwang and J.-B. Song, "Monocular vision-based slam in indoor environment using corner, lamp, and door features from upward-looking camera," *IEEE Trans. Ind. Electron.*, vol. 58, no. 10, pp. 4804–4812, Oct. 2011, doi: 10.1109/TIE.2011.2109333.

[5] C. Cadena et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016, doi: 10.1109/TRO.2016.2624754.

[6] S. Lowry et al., "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016, doi: 10.1109/TRO.2015.2496823.

[7] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19929–19953, Nov. 2022, doi: 10.1109/TITS.2022.3175656.

[8] Y. Liu et al., "Stereo visual-inertial odometry with multiple Kalman filters ensemble," *IEEE Trans. Ind. Electron.*, vol. 63, no. 10, pp. 6205–6216, Oct. 2016, doi: 10.1109/TIE.2016.2573765.

[9] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 957–964, doi: 10.1109/ICRA.2012.6225147.

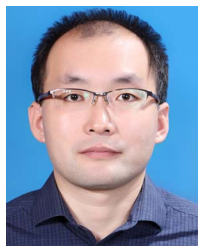
[10] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to MAV navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 3923–3929, doi: 10.1109/IROS.2013.6696917.

[11] J. Jiang, J. Yuan, X. Zhang, and X. Zhang, "DVIO: An optimization-based tightly coupled direct visual-inertial odometry," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11212–11222, Nov. 2021, doi: 10.1109/TIE.2020.3036243.

[12] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3565–3572, doi: 10.1109/ROBOT.2007.364024.
[13] H. Yu and A. I. Mourikis, "Vision-aided inertial navigation with line features and a rolling-shutter camera," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 892–899, doi: 10.1109/IROS.2015.7353477.
[14] M. K. Paul, K. Wu, J. A. Hesch, E. D. Nerurkar, and S. I. Roumeliotis, "A comparative analysis of tightly-coupled monocular, binocular, and stereo VINS," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 165–172, doi: 10.1109/ICRA.2017.7989022.
[15] G. Huang, K. Eckenhoff, and J. Leonard, "Optimal-state-constraint EKF for visual-inertial navigation," in *Robotics Research*. Cham, Switzerland: Springer, 2018, pp. 125–139.
[16] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 4666–4672, doi: 10.1109/ICRA40945.2020.9196524.
[17] D. Strelow and S. Singh, "Motion estimation from image and inertial measurements," *Int. J. Robot. Res.*, vol. 23, no. 12, pp. 1157–1195, Dec. 2004, doi: 10.1177/0278364904045593.
[18] V. Indelman, S. Williams, M. Kaess, and F. Dellaert, "Information fusion in navigation systems via factor graph based incremental smoothing," *Robot. Auton. Syst.*, vol. 61, no. 8, pp. 721–738, Aug. 2013, doi: 10.1016/j.robot.2013.05.001.
[19] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visualinertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, Mar. 2015, doi: 10.1177/0278364914554813.
[20] A. Patron-Perez, S. Lovegrove, and G. Sibley, "A spline-based trajectory representation for sensor fusion and rolling shutter cameras," *Int. J. Comput. Vis.*, vol. 113, no. 3, pp. 208–219, Jul. 2015, doi: 10.1007/s11263-015-0811-3.
[21] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018, doi: 10.1109/TRO.2018.2853729.
[22] T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3662–3669, doi: 10.1109/IROS.2018.8593603.
[23] R. Mur-Artal and J. D. Tardos, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017, doi: 10.1109/LRA.2017.2653359.
[24] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. Montiel, and J. D. Tardos, "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, doi: 10.1109/TRO.2021.3075644.
[25] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "VINS on wheels," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 5155–5162, doi: 10.1109/ICRA.2017.7989603.
[26] Y. Zhai and S. Zhang, "A novel LiDAR-IMU-odometer coupling framework for two-wheeled inverted pendulum (TWIP) robot localization and mapping with nonholonomic constraint factors," *Sensors*, vol. 22, no. 13, Jun. 2022, Art. no. 4778, doi: 10.3390/s22134778.
[27] W. Lee, K. Eckenhoff, Y. Yang, P. Geneva, and G. Huang, "Visual-inertial-wheel odometry with online calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4559–4566, doi: 10.1109/IROS45743.2020.9341161.
[28] Z. Shan, R. Li, and S. Schwertfeger, "RGBD-inertial trajectory estimation and mapping for ground robots," *Sensors*, vol. 19, no. 10, May 2019, Art. no. 2251, doi: 10.3390/s19102251.
[29] J. Liu, W. Gao, and Z. Hu, "Visual-inertial odometry tightly coupled with wheel encoder adopting robust initialization and online extrinsic calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 5391–5397, doi: 10.1109/IROS40897.2019.8967607.
[30] J. Liu, W. Gao, and Z. Hu, "Bidirectional trajectory computation for odometer-aided visual-inertial SLAM," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1670–1677, Apr. 2021, doi: 10.1109/LRA.2021.3059564.
[31] F. Zheng, H. Tang, and Y.-H. Liu, "Odometry-vision-based ground vehicle motion estimation with SE(2)-constrained SE(3) poses," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2652–2663, Jul. 2019, doi: 10.1109/TCYB.2018.2831900.
[32] A. Grunnet-Jepsen, J. N. Sweetser, P. Winer, A. Takagi, and J. Woodfill, "Projectors for Intel® RealSense™ depth cameras D4xx," Intel Support, Interl Corporation, Santa Clara, CA, USA, White Paper, 2018.
[33] M. S. Ahn, H. Chae, D. Noh, H. Nam, and D. Hong, "Analysis and noise modeling of the Intel RealSense D435 for mobile robots," in *Proc. 16th Int. Conf. Ubiquitous Robots*, 2019, pp. 707–711, doi: 10.1109/URAI.2019.8768489.
[34] I. Dryanovski, R. G. Valenti, and J. Xiao, "Fast visual odometry and mapping from RGB-D data," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 2305–2310, doi: 10.1109/ICRA.2013.6630889.
[35] X. Shi et al., "Are we ready for service robots? The OpenLORIS-Scene datasets for lifelong SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 3139–3145, doi: 10.1109/ICRA40945.2020.9196638.
[36] M. Labbe and F. Michaud, "RTAB-map as an open-source LiDAR and visual simultaneous localization and mapping library for large-scale and long-term online operation," *J. Field Robot.*, vol. 36, no. 2, pp. 416–446, 2019, doi: 10.1002/rob.21831.
[37] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 7244–7251, doi: 10.1109/IROS.2018.8593941.
[38] X. Wei, J. Lv, J. Sun, E. Dong, and S. Pu, "GCLO: Ground constrained LiDAR odometry with low-drifts for GPS-denied indoor environments," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 2229–2235, doi: 10.1109/ICRA46639.2022.9812336.
[39] M. Zhang, X. Zuo, Y. Chen, Y. Liu, and M. Li, "Pose estimation for ground robots: On manifold representation, integration, reparameterization, and optimization," *IEEE Trans. Robot.*, vol. 37, no. 4, pp. 1081–1099, Aug. 2021, doi: 10.1109/TRO.2020.3043970.

Xinyang Zhao received the bachelor's degree in measurement, control technology, and instrumentation from Harbin Engineering University, Harbin, China, in 2015, and the master's degree in control engineering from the Harbin Institute of Technology, Harbin, in 2017.

His current research interests include computer vision, simultaneous localization and mapping, robotics, and deep learning.



Qinghua Li received the B.S., M.S., and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2001, 2003, and 2008, respectively.

From 2011 to 2012, he was a Research Assistant with the Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI, USA. His current research interests include magnetic information navigation, inertial navigation, and motion control.

Changhong Wang (Senior Member, IEEE) received the B.S. degree in automation, the M.E. degree in control science and engineering, and the Ph.D. degree in navigation guidance and control from the Harbin Institute of Technology, Harbin, China, in 1983, 1986, and 1991, respectively.

His current research interests include intelligent control and intelligent systems, inertial technology and its testing equipment, robotics, precision servo systems, and network control.

Hexuan Dou received the BA Eng. degree in automation and the MA Eng. degree in control engineering from the Harbin Institute of Technology, Harbin, China, in 2013 and 2017, respectively, where he is currently working toward the Ph.D. degree in control science and engineering.

His research interests include visual SLAM and GNC for autonomous unmanned vehicles.

Bo Liu received the B.Eng. degree in control science and engineering in 2017 from the Harbin Institute of Technology, Harbin, China, where he is currently working toward the Ph.D. degree in control science and engineering.

His current research interests include smart UAV, heterogeneous swarm robotic system, multi-agents reinforcement learning, multi-agent formation control, and swarm intelligence.