A Learning-Free Method for Locomotion Mode Prediction by Terrain Reconstruction and Visual-Inertial Odometry

Shunyi Zhao, Zehuan Yu, *Graduate Student Member, IEEE*, Zhaoyang Wang, Hangxin Liu, *Member, IEEE*, Zhihao Zhou, *Member, IEEE*, Lecheng Ruan, and Qining Wang, *Senior Member, IEEE*

***Abstract*— This research introduces a novel, highly precise, and learning-free approach to locomotion mode prediction, a technique with potential for broad applications in the field of lower-limb wearable robotics. This study represents the pioneering effort to amalgamate 3D reconstruction and Visual-Inertial Odometry (VIO) into a locomotion mode prediction method, which yields robust prediction performance across diverse subjects and terrains, and resilience against various factors including camera view, walking direction, step size, and disturbances from moving obstacles without the need of parameter adjustments. The proposed Depth-enhanced Visual-Inertial Odometry (D-VIO) has been meticulously designed to operate within computational constraints of wearable configurations while demonstrating resilience against unpredictable human movements and sparse features. Evidence of its effectiveness, both in terms of accuracy and operational time consumption, is substantiated through tests conducted using open-source dataset and closed-loop evaluations. Comprehensive experiments were undertaken to validate its prediction accuracy across various test conditions such as subjects, scenarios, sensor mounting positions, camera views, step sizes, walking directions, and disturbances from moving obstacles. A comprehensive prediction accuracy rate of 99.00% confirms the efficacy, generality, and robustness of the proposed method.**

***Index Terms*— Locomotion mode prediction, terrain reconstruction, visual-inertial odometry, learning-free, wearable robots.**

I. Introduction

RECENT years have seen a surge in interest towards lower-limb wearable robotics, encompassing both prostheses and exoskeletons, owing to their significant impact on the restoration and enhancement of human locomotion. This burgeoning area of research has found extensive applications in healthcare and rehabilitation, as evidenced by numerous studies [1], [2], [3], [4]. It is critical to note that humans adopt distinct modes of locomotion contingent upon various types of terrains, prominently including level ground walking, and ascending or descending on ramps or stairs. It is therefore imperative for these wearable robots to anticipate the imminent locomotion mode and make appropriate adjustments to their control strategies accordingly.

The study of locomotion mode acquisition has been an extensive field of research. A collection of researchers have endeavored to discern locomotion modes by characterizing human movements. Specifically, these studies typically employ Inertial Measurement Units (IMUs) and other sensors affixed to designated positions on the human body to record corresponding postures and velocities, subsequently utilizing pattern recognition techniques for classification into distinct locomotion mode categories [5], [6], [7], [8], [9]. Notwithstanding, these techniques predominantly focus on *recognizing* the locomotion mode shortly after the gait cycle commences

Manuscript received 14 March 2023; revised 29 July 2023; accepted 25 September 2023. Date of publication 2 October 2023; date of current version 11 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 52005011, Grant 52375001, and Grant 91948302; and in part by the Beijing Municipal Science and Technology Project under Grant D181100000318002. *(Shunyi Zhao, Zehuan Yu, and Zhaoyang Wang contributed equally to this work.) (Corresponding authors: Lecheng Ruan; Zhihao Zhou.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of Peking University under Application No. 20180602.

Shunyi Zhao, Hangxin Liu, and Lecheng Ruan are with the National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI), Beijing 100080, China (e-mail: ruanlecheng@bigai.ai).

Zehuan Yu is with the National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI), Beijing 100080, China, and also with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

Zhaoyang Wang is with the Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University, Beijing 100871, China.

Zhihao Zhou is with the Institute for Artificial Intelligence, Peking University, Beijing 100871, China, and also with the Beijing Engineering Research Center of Intelligent Rehabilitation Engineering, Beijing 100871, China (e-mail: zhihaozhou@pku.edu.cn).

Qining Wang is with the Department of Advanced Manufacturing and Robotics, College of Engineering, and the Institute for Artificial Intelligence, Peking University, Beijing 100871, China, also with the School of Rehabilitation Sciences and Engineering, University of Health and Rehabilitation Sciences, Qingdao 266071, China, also with Peking University Third Hospital, Beijing 100191, China, and also with the National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI), Beijing 100080, China.

Digital Object Identifier 10.1109/TNSRE.2023.3321077

rather than *predicting* the mode in advance. Further, the general applicability of such methods presents a formidable challenge as the walking patterns of individuals traversing identical terrains can vary significantly [5]. Indeed, even a single individual may exhibit diverse walking patterns on the same terrain, including jogging, striding, and wandering [8], [10]. Furthermore, variations in terrain parameters, for instance, stair height or ramp slope, may also engender disparities in walking patterns and consequent degradation of accuracy. As such, these variables require a technically adept approach for effective management [8], [9].

Alternatively, a burgeoning body of research has concentrated on predicting locomotion mode by directly capturing the terrain ahead of a human using a camera. The resultant image is then typically subjected to the classification of locomotion modes employing Machine Learning (ML) techniques [11], [12], [13], [14], [15], [16]. These approaches often demonstrate greater robustness and universality, as well as advancements in prediction instance compared to the earlier discussed human movement characterization techniques. Nevertheless, akin to other ML tasks, these methods frequently necessitate considerable volumes of data and an intricate training process, and require sophisticated techniques for improvements [12]. Furthermore, these studies do not comprehensively incorporate walking information such as step size or walking direction relative to the viewed image. Indeed, variations in step size [16], camera pose [17], or walking direction [15] may result in the terrain captured in the camera view not reflecting the terrain of the next step. Moreover, to conserve computational power in wearable configurations, these methods typically rely on single image input, increasing their vulnerability to interference from moving objects and thus posing challenges for deployment in open environments [16].

To date, the realm of wearable robotics is yet to possess a highly accurate method for locomotion prediction that (1) effectively blends camera-view data and walking information to manage predictions in intricate environments with variations in step size and walking direction; (2) demonstrates robustness against a wide array of terrains, subjects, camera views, and locomotion disparities; (3) obviates the need for comprehensive data accumulation and sophisticated training processes typical of ML approaches; and (4) exhibits resilience amidst external disturbances. This paper proposes such a method. By employing 3D reconstruction, the terrain characterization is incrementally stabilized as more images are continuously incorporated, thereby mitigating the fluctuation in camera views and external disturbances from mobile obstacles, even in the absence of any learning process. Concurrently, by integrating Visual-Inertial Odometry (VIO), human walking information is explicitly considered, ensuring the maintenance of prediction accuracy across different subjects, regardless of alterations in walking directions and step size. Nevertheless, for successful implementation of locomotion mode prediction, the proposed VIO needs to (1) function efficiently within the computational power of wearable configurations; (2) display resilience under the randomness and abruptness inherent in human motions, especially when the camera is mounted on the head; and (3) uphold estimation precision in spite of sparse features, considering terrains frequently comprise vast planes. These prerequisites are comprehensively addressed within the framework of the proposed D-VIO.

The contributions of this paper are concluded as follows: (1) The proposition of a novel highly accurate learning-free method for locomotion mode prediction. This approach, for the first time, incorporates both terrain reconstruction and VIO, allowing for robust performance across diverse subjects and terrains. Remarkably, it maintains robustness despite variations in camera view, walking direction, step size, and disturbances from moving obstacles. (2) The development of the D-VIO algorithm, specifically tailored to meet the demands of locomotion mode prediction application. The algorithm is designed to be computationally efficient while exhibiting resiliency against random and abrupt human motions as well as sparse features. Its effectiveness on accuracy and time consumption is demonstrated through tests using an open-source dataset and closed-loop evaluations. (3) The undertaking of comprehensive experiments to assess the effectiveness of the proposed method for locomotion mode prediction. These comprehensive trials include different subjects, scenarios, sensor mounting types, camera views, walking step sizes and directions, along with the influence of moving objects as disturbances.

II. Method

The pipeline of the proposed method is described in Fig. 1. The terrain information is reconstructed from an environmental map, which is sourced from the Truncated Signed Distance Function (TSDF) algorithm [18]. By implementing the Ray Casting algorithm [19] for point cloud smoothing and localization, and introducing continuity and voting constraints during updates, our proposed method is capable of reconstructing the terrain despite various parameter variations and external disturbances. Walking information, such as the walking direction and step size, is explicitly considered within our proposed D-VIO. This approach demonstrates commendable performance even with sparse visual features and the intermittence and unpredictability characteristic of human motions.

A. Related Works

VIO has gained considerable attention as an effective technique for the estimation of kinematic states and the subsequent extraction of walking information for egocentric objects [20]. Certain methodologies, such as ORB-SLAM3 [21] and VIP-SLAM [22], strive for precise estimations contingent upon adequate computational capacity. In contrast, others, notably VINS-Mono [23] and OKVIS [24], prioritize computational efficiency. Among the lightweight VIOs, VINS-Mono distinguishes itself due to its high update frequency, robust initialization process, and facilitation of secondary development. This led to further advancements, namely VINS-RGBD [25] and VINS-Fusion [26], which seek to enhance the accuracy of VINS-Mono while maintaining marginal additions to computational load. However, for locomotion mode prediction applications, maintaining VIO accuracy presents a

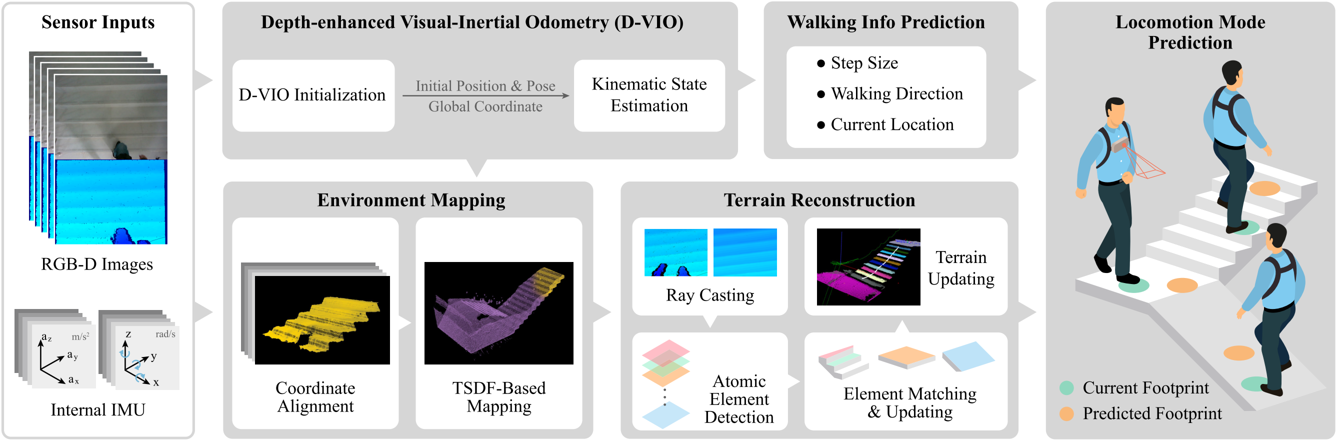

Fig. 1. The pipeline of the proposed locomotion mode prediction method.

significant challenge due to the randomness and abruptness of motion [27], coupled with the lack of environmental features, which need to be well addressed in our proposed method.

Extensive research has been conducted in the realm of real-time 3D reconstruction. One prevalent approach involves the application of learning-based implicit representation methods designed to establish relationships with 3D coordinates [28]. These techniques have exhibited advantages in settings encompassing large-scale scenarios when adequate computational power is accessible [29], [30]. Conversely, another category focuses on the integration of weighting and merging processes for single-shot depth information to bolster robustness against external disturbances such as moving objects, typically utilizing a Truncated Signed Distance Function (TSDF) [18]. This latter category generally presents lower computational demands and has demonstrated efficacy in local mapping applications, deeming it suitable for tasks involving locomotion mode prediction. Prominent reconstruction algorithms that adopt this methodology include ElasticFusion [31] and BundleFusion [32].

B. Notations

The sensor inputs for the D-VIO system consist of RGB images $\boldsymbol{I}_C$, depth images $\boldsymbol{I}_D$, and raw data from an Inertial Measurement Unit (IMU), including angular velocity $\boldsymbol{\omega}$ and linear acceleration $\boldsymbol{a}$, which are used to calculate the IMU preintegration measurement $\hat{\mathbf{z}}$. The D-VIO initializes a global coordinate G and an initial kinematic state $\mathcal{X}_0$, which consists of position ${}^G\mathbf{p}_\mathrm{I}$, velocity ${}^G\mathbf{v}_I$ and pose ${}^G\mathbf{q}_I$. The odometry then continuously estimates kinematic state $\mathcal{X}$. Using temporal estimated states $\boldsymbol{\mathcal{X}}$ and divided gait phases, the D-VIO calculates the walking direction $\bar{\mathbf{v}}^{nm}$ and step size δ_{nxt}. Additionally, the 2D footprint position $\mathbf{p}^{pd}$ is computed. The TSDF-based mapping algorithm aligns coordinates for each one-shot depth image and merges them into an integrated 3D map $\boldsymbol{S}$. Finally, the terrains are reconstructed into concise descriptions $\boldsymbol{T}$ by detecting atomic elements (planes $\boldsymbol{\eta}$) and updating information from $\boldsymbol{S}$. The locomotion mode is predicted directly by fusing the walking and terrain information.

TABLE I

FREQUENTLY USED TERMINOLOGIES

Term	Meaning	Term	Meaning
$\boldsymbol{I}_C, I_C$	color image	${}^G\mathbf{q}_I$	orientation
$\boldsymbol{I}_D, I_D$	depth image	$\hat{\mathbf{z}}$	IMU measurement
$\mathcal{X}$	human state	$\boldsymbol{S}$	3D map
$\mathcal{X}^e$	IMU biases	$\boldsymbol{\eta}, \eta$	plane
${}^G\mathbf{p}_I$	position	$\boldsymbol{\iota}$	voxel
${}^G\mathbf{v}_I$	velocity	$\boldsymbol{T}$	terrain description

We provide a list of frequently-used terminologies in TABLE I.

C. Visual Inertial Odometry (VIO)

This section proposes a D-VIO algorithm to estimate the kinematic state vector of human walking $\mathcal{X}$ in real time, in a gravity-aligned coordinate G using RGB-D images and IMU data. The proposed D-VIO algorithm is structured around VINS-Mono [23], due to its inherent benefits in lightweight computation, high update frequency, robust initialization process, and facilitation of secondary development. The introduction of supplementary constraints at both the initialization and estimation stages allows the proposed D-VIO to attain superior performance under conditions characterized by sparse visual features and erratic human motions - conditions that are typically encountered in locomotion mode prediction applications.

The VINS-Mono framework consists of two main components: a visual feature tracker and a kinematic state estimator. During initialization, the system is initialized with a sliding window $\mathcal{X}_W = \{\langle \mathcal{X}_i, \mathcal{X}_i^e \rangle\}_{i=1}^M$ consisting of M keyframes selected according to the visual difference, where $\mathcal{X}^e = \langle \mathbf{b}_a, \mathbf{b}_g \rangle$ denotes the accelerometer and gyroscope biases. The visual feature tracker uses the Kanade–Lucas–Tomasi (KLT) [33] optical flow algorithm to detect and track visual features, or landmarks, in the RGB image sequence. The state estimator solves the vision-only Structure-from-Motion (SfM) and visual-inertial alignment problem to estimate the kinematic state, which is constructed using visual projection constraints and solved using a nonlinear optimization method.

The visual projection factor $\mathbf{r}_{\mathcal{V}_{ij}}$ of the landmark ${}^{G}\mathbf{l}_j$ in frame i can be formulated as

$$\begin{aligned}\mathbf{r}_{\mathcal{V}_{ij}}&({}^{G}\mathbf{R}_{I_i}, {}^{G}\mathbf{p}_{I_i}, {}^{G}\mathbf{l}_j)\\ &= \frac{1}{Z}[{}^{I}\mathbf{R}_C^T({}^{G}\mathbf{R}_{I_i}^T({}^{G}\mathbf{l}_j -{}^{G}\mathbf{p}_{I_i}) -{}^{I}\mathbf{p}_C)]_{xy} - \mathbf{u}_m \end{aligned}\tag{1}$$

and the visual projection cost can be formulated as

$$\mathcal{C}_{\mathcal{V}} = \sum_{1\le i\le M}\sum_{1\le j\le|\mathcal{L}|}\rho(\|\mathbf{r}_{\mathcal{V}_{ij}}({}^{G}\mathbf{R}_{I_i}, {}^{G}\mathbf{p}_{I_i}, {}^{G}\mathbf{l}_j)\|^2_{\Sigma_{\mathcal{V}}}),\tag{2}$$

where Z denotes the depth of landmark, ${}^{G}\mathbf{l}_j \in \mathbb{R}^3$ denotes the position of landmark; ${}^{G}\mathbf{R}_{I_i}$ is the rotation matrix form of ${}^{G}\mathbf{q}_{I_i} \in \mathcal{X}$; ${}^{I}\mathbf{R}_C$ and ${}^{I}\mathbf{p}_C$ denotes the extrinsic parameters between the IMU and camera; $\mathbf{u}_m \in \mathbb{R}^2$ denotes the normalized plane position of the feature point; $\mathcal{L} = \{{}^{G}\mathbf{l}_i\}_{i=1}^n$ denotes the set of the landmarks; $\Sigma_{\mathcal{V}}$ represents the covariance matrix of visual re-projection measurement error; $\rho(\cdot)$ denotes the robust kernel function; $[\cdot]_{xy}$ denotes the operator that extracts the first two elements of a three-dimensional vector.

In our D-VIO system, we leverage depth measurements, which serve to eliminate outliers, expedite the optimization process, and enhance the robustness of the initialization process, to augment the VINS-Mono. The depth measurements enable the formulation of a depth-enhanced SfM problem, consisting of two types of factors: visual projection factors and depth measurement factors. The depth measurement factor $\mathbf{r}_{\mathcal{D}_{ij}}$ of the landmark ${}^{G}\mathbf{l}_j$ in frame i is formulated as

$$\begin{aligned}\mathbf{r}_{\mathcal{D}_{ij}}&({}^{G}\mathbf{R}_{I_i}, {}^{G}\mathbf{p}_{I_i}, {}^{G}\mathbf{l}_j)\\ &= [{}^{I}\mathbf{R}_C^T({}^{G}\mathbf{R}_{I_i}^T({}^{G}\mathbf{l}_j -{}^{G}\mathbf{p}_{I_i}) -{}^{I}\mathbf{p}_C)]_z - d_m \end{aligned}\tag{3}$$

and the corresponding cost is formulated as

$$\mathcal{C}_{\mathcal{D}} = \sum_{1\le i\le M}\sum_{1\le j\le|\mathcal{L}|}\rho(\|\mathbf{r}_{\mathcal{D}_{ij}}({}^{G}\mathbf{R}_{I_i}, {}^{G}\mathbf{p}_{I_i}, {}^{G}\mathbf{l}_j)\|^2_{\Sigma_{\mathcal{D}}}),\tag{4}$$

where $d_m \in \mathbb{R}$ denotes the measurement value of the feature depth; $[\cdot]_z$ denotes the operator that extracts the third element of a three-dimensional vector; $\Sigma_{\mathcal{D}}$ represents the covariance matrix of depth measurement error. Therefore, the nonlinear optimization problem is composed of two components shown as

$$\min_{\mathcal{X}_W} \mathcal{C}_{\mathcal{V}} + \mathcal{C}_{\mathcal{D}},\tag{5}$$

where $\mathcal{X}_W = \langle \mathcal{X}_0, \mathcal{X}_0^e \rangle$.

After the initialization stage, VINS-Mono solves the visual-inertial bundle adjustment in the sliding window to update the kinematic state $\mathcal{X}$ continuously. This process considers all temporal factors, including the prior factor, the IMU preintegration factor, and the visual projection factors. The prior factor $\mathcal{C}_{\mathcal{P}}$ is defined as

$$\mathcal{C}_{\mathcal{P}} = \|\mathbf{r}_{\mathcal{P}} - \mathbf{H}_{\mathcal{P}}\mathcal{X}\|^2,\tag{6}$$

where $\mathbf{r}_{\mathcal{P}}$ denotes the prior residual and $\mathbf{H}_{\mathcal{P}}$ denotes the prior hessian matrix, which is calculated using the marginalization factor including prior information from visual projection factors, depth measurement factors, and IMU preintegration factors. The IMU preintegration factor $\mathbf{r}_{\mathcal{B}}$ is formulated as

$$\begin{aligned}\mathbf{r}_{\mathcal{B}}&({}^{I_k}\hat{\mathbf{z}}_{I_{k+1}}, \mathcal{X}_k, \mathcal{X}_{k+1}, \mathcal{X}_k^e, \mathcal{X}_{k+1}^e)\\ &= \begin{bmatrix} {}^{I_k}\mathbf{R}_G\left({}^{G}\mathbf{p}_{I_{k+1}} -{}^{G}\mathbf{p}_{I_k} -{}^{G}\mathbf{v}_{I_k}\Delta t - \frac{1}{2}\mathbf{g}\Delta t^2\right) -{}^{I_k}\hat{\boldsymbol{\alpha}}_{I_{k+1}} \\ 2\cdot\left[\left({}^{I_k}\hat{\boldsymbol{\gamma}}_{I_{k+1}}\right)^{-1}\otimes\mathbf{q}\left({}^{G}\mathbf{R}_{I_k}\right)^{-1}\otimes\mathbf{q}\left({}^{G}\mathbf{R}_{I_{k+1}}\right)\right]_{xyz} \\ {}^{I_k}\mathbf{R}_G\left({}^{G}\mathbf{v}_{I_{k+1}} -{}^{G}\mathbf{v}_{I_k} - \mathbf{g}\Delta t\right) -{}^{I_k}\hat{\boldsymbol{\beta}}_{I_{k+1}} \\ \mathbf{b}_{a_{I_{k+1}}} - \mathbf{b}_{a_{I_k}} \\ \mathbf{b}_{g_{I_{k+1}}} - \mathbf{b}_{g_{I_k}} \end{bmatrix}\end{aligned}\tag{7}$$

and the visual projection factor $\mathbf{r}_{\mathcal{C}_{ij}}^{sw}$ is formulated as

$$\begin{aligned}\mathbf{r}_{\mathcal{C}_{ij}}^{sw}&(\mathcal{X}_{I_i}, \mathcal{X}_{I_j}, \xi_l, \mathbf{u}_{m_j})\\ &= [{}^{I}\mathbf{R}_C^T({}^{G}\mathbf{R}_{I_j}^T({}^{G}\mathbf{R}_{I_i}({}^{I}\mathbf{R}_C\frac{1}{\xi_l}\pi^{-1}(\mathbf{u}_{m_i}) +{}^{I}\mathbf{p}_C) +{}^{G}\mathbf{p}_{I_i})\\ &\quad -{}^{G}\mathbf{p}_{I_j}) -{}^{I}\mathbf{p}_C]_{xy} - \mathbf{u}_{m_j},\end{aligned}\tag{8}$$

where $\pi^{-1}(\cdot)$ is the back-projection function, ξ_l is the inverse depth of landmarks in the host frame which is the first frame that observes the landmark, $\mathbf{g}$ denotes the gravity vector, $\mathbf{q}(\cdot)$ denotes the equivalent quaternion of a rotation matrix, $\mathbf{u}$ is the observation of visual feature. In equation (7), ${}^{I_k}\hat{\mathbf{z}}_{I_{k+1}} = \left\{{}^{I_k}\hat{\boldsymbol{\alpha}}_{I_{k+1}}, {}^{I_k}\hat{\boldsymbol{\beta}}_{\mathbf{I_{k+1}}}, {}^{I_k}\hat{\boldsymbol{\gamma}}_{\mathbf{I_{k+1}}}\right\}$ is preintegration measurement representing 3D positions. In equation (8), frame i is the host frame of the landmark l, and frame j is a co-visible frame. The corresponding IMU preintegration cost $\mathcal{C}_{\mathcal{B}}$ is formulated as

$$\mathcal{C}_{\mathcal{B}} = \sum_{k\in\mathcal{B}}\|\mathbf{r}_{\mathcal{B}}({}^{I_{k+1}}\hat{\mathbf{z}}_{I_k}, \mathcal{X}_k, \mathcal{X}_{k+1}, \mathcal{X}_k^e, \mathcal{X}_{k+1}^e)\|^2_{\Sigma_{\mathcal{P}_{k+1}^k}}\tag{9}$$

and visual projection cost $\mathcal{C}_{\mathcal{V}}^{sw}$ is formulated as

$$\mathcal{C}_{\mathcal{V}}^{sw} = \sum_{(l,j)\in\mathcal{C}}\rho(\|\mathbf{r}_{\mathcal{C}_{ij}}^{sw}(\mathcal{X}_{I_i}, \mathcal{X}_{I_j}, \xi_l, \mathbf{u}_{m_j})\|^2_{\Sigma_{\mathcal{V}}}).\tag{10}$$

To enhance the estimation accuracy and adapt to the random motions, we construct depth measurement factors in the proposed D-VIO system. In the host frame, the host depth measurement factor is formulated as

$$\mathbf{r}_{\mathcal{D}_h}(\xi_l) = \frac{1}{\xi_l} - d_m\tag{11}$$

which is a residual between inverse depth ξ_l and host depth measurement d_m. Between the co-visible frame i and j, projection depth measurement factors are formulated as

$$\begin{aligned}\mathbf{r}_{\mathcal{D}_{ij}}^{sw}&(\mathcal{X}_{I_i}, \mathcal{X}_{I_j}, \xi_l, d_{m_j})\\ &= [{}^{I}\boldsymbol{R}_C^T({}^{G}\mathbf{R}_{I_j}^T({}^{G}\mathbf{R}_{I_i}({}^{I}\mathbf{R}_C\frac{1}{\xi_l}\pi^{-1}(\mathbf{u}_{m_i}) +{}^{I}\mathbf{p}_C)\\ &\quad +{}^{G}\mathbf{p}_{I_i}) -{}^{G}\mathbf{p}_{I_j}) -{}^{I}\mathbf{p}_C]_z - d_{m_j}.\end{aligned}\tag{12}$$

Their corresponding costs, $\mathcal{C}_{\mathcal{D}_h}$ is formulated as

$$\mathcal{C}_{\mathcal{D}_h} = \sum_{l\in\mathcal{L}}\rho(\|\mathbf{r}_{\mathcal{D}_h}(\xi_l)\|^2_{\Sigma_{\mathcal{D}}}),\tag{13}$$

and $\mathcal{C}^{sw}_{\mathcal{D}}$ is formulated as

$$\mathcal{C}^{sw}_{\mathcal{D}} = \sum_{(l,j)\in\mathcal{C}} \rho(\|\mathbf{r}^{sw}_{\mathcal{D}_{ij}}(\mathcal{X}_{I_i}, \mathcal{X}_{I_j}, \xi_l, d_{m_j})\|^2_{\Sigma_{\mathcal{D}}}). \tag{14}$$

Totally, the final nonlinear optimization problem is formulated as

$$\min_{\mathcal{X}_W} \mathcal{C}_{\mathcal{P}} + \mathcal{C}_{\mathcal{B}} + \mathcal{C}^{sw}_{\mathcal{V}} + \mathcal{C}^{sw}_{\mathcal{D}} + \mathcal{C}_{\mathcal{D}_h}. \tag{15}$$

After the sliding window optimization, all host depth measurement factors $\mathbf{r}_{\mathcal{D}_h}$ are dropped directly because they do not contribute to the subsequent optimization process. All projection depth measurement factors $\mathbf{r}^{sw}_{\mathcal{D}_{ij}}$ are used to construct the prior hessian matrix and prior residual in the marginalization process. The kinematic state $\mathcal{X}$ is then updated at a high frequency by combining the pre-integration result of IMU with the optimization result. Notice that global coordinate is set as default in the notations, thus omitted in the rest of this paper for simplicity. For example, ${}^{G}\mathbf{p}_I$ will be denoted as $\mathbf{p}$.

D. Walking Information Prediction

This section uses the estimated kinematic states $\boldsymbol{\mathcal{X}}$ to extract and estimate the walking information, *e.g.* step size, walking direction, and footprint of the subject. Force-sensitive insoles are adopted for footstep division assistance. Once the pressure signal is detected, the position and velocity are recorded. As a result, division positions $\{\bar{\mathbf{p}}^t_i\}^{n_{tp}}_{i=1}$ and corresponding velocities $\{\bar{\mathbf{v}}^t_i\}^{n_{tp}}_{i=1}$, which can represent the walking direction simultaneously, are cached as time series. The step size between two adjacent division positions is calculated using

$$\delta = \|\bar{\mathbf{p}}^t_i - \bar{\mathbf{p}}^t_{i-1}\|, \tag{16}$$

where $\bar{\mathbf{p}}^t_{i-1}$ and $\bar{\mathbf{p}}^t_i$ are two adjacent elements in $\{\bar{\mathbf{p}}^t_i\}^{n_{tp}}_{i=1}$. With the last two step sizes $\delta_{n_{tp}}$ and $\delta_{n_{tp}-1}$, the size of the next step δ_{nxt} can be calculated by a 1st-order calculation method, which is represented by

$$\delta_{nxt} = 2\delta_{n_{tp}} - \delta_{n_{tp}-1}. \tag{17}$$

Then, the footprint position can be predicted with δ_{nxt} and $\bar{\mathbf{v}}^t_{n_{tp}}$ as

$$\mathbf{p}^{pd} = \bar{\mathbf{p}}^t_{n_{tp}} + \bar{\mathbf{v}}^{nm}_{n_{tp}} \cdot \delta_{nxt}, \tag{18}$$

where the unit vector $\bar{\mathbf{v}}^{nm}_{n_{tp}} = \frac{\bar{\mathbf{v}}^t_{n_{tp}}}{|\bar{\mathbf{v}}^t_{n_{tp}}|}$ guarantees the scale-invariance of δ_{nxt}. The left and right feet are distinguished by adding specific offsets to $\mathbf{p}^{pd}$ respectively.

E. Environment Mapping

This section proposes an approach to integrate depth images $\boldsymbol{I}_D$ into a consistent 3D map to eliminate duplicate information, filter out disturbances caused by moving objects, and average sensor noise. We utilize a TSDF-based algorithm [18] to construct a 3D map $\boldsymbol{S}$. $\boldsymbol{S}$ comprises evenly distributed voxels, which have two attributes, a TSDF value λ and a weight W. Once a new depth image is received from the camera, the TSDF value and the weight of every voxel are updated accordingly. Finally, the surfaces of entities in the map, which contains the terrain information we need, are represented by the set of voxels with zero value.

The updating policy is composed of three steps: (1) Calculating the Signed Distance Function (SDF) value of each voxel as

$$\mathrm{SDF}(\iota_j) = \boldsymbol{d}_{\iota_j} - d_c, \tag{19}$$

where $\boldsymbol{\iota}_j$ denotes the voxel, $\boldsymbol{d}_{\iota_j}$ identifies the projection depth from the voxel to the camera optical center, d_c represents the voxel-corresponding depth value; (2) Updating the TSDF value from the SDF value, which involves (i) calculating the temporary TSDF value λ'_{ι_j} by truncating SDF value into $[-1, 1]$ as

$$\lambda'_{\iota_j} = \max(-1, \min(1, \frac{\mathrm{SDF}(\iota_j)}{t_\delta})), \tag{20}$$

where t_δ is a preset threshold, and (ii) calculating the TSDF value by weighting the existing TSDF value λ_{ι_j} and the temporary TSDF value with

$$\lambda_{\iota_j} = \frac{W_j \lambda_{\iota_j} + w_j \lambda'_{\iota_j}}{W_j + w_j}, \tag{21}$$

where w_j is a preset parameter; (3) The weight of each voxel is updated with

$$W_j = \min(W_j + w_j, W_{max}), \tag{22}$$

where W_{max} denotes the preset upper bound of weight.

F. Terrain Reconstruction

This section endeavors to progressively develop a succinct representation of terrain information, utilizing updates from the 3D map outlined in Section II-E. Firstly, in an effort to curtail superfluous computation, the Ray Casting algorithm [19] is implemented upon each update of the 3D map. This strategy ensures that the refinement of the terrain representation coincides solely with the acquisition of new information from the camera, which are subsequently presented as a smoothed depth image, denoted as I^s_D.

Secondly, planes are extracted from the smoothed depth image I^s_D, which will later serve as foundational elements of terrain representation. The extraction process employs the Agglomerative Hierarchical Clustering (AHC) algorithm [34], which comprises three main steps: (1) The depth image I^s_D is divided into cells of dimensions $H_p \times W_p$; (2) Each cell is fitted with a plane using Principal Component Analysis (PCA), excluding cells exhibiting apparent discontinuity; (3) Cells are then clustered based on the similarities of the fitted planes, characterized by attributes including the position of the plane center, the direction of the plane normal vector, and the plane fitting error. Upon completion of these steps, each cluster of cells forms a newly detected plane, represented as $\eta^i := \langle \mathbf{c}^i, \mathbf{n}^i, \mathrm{a}^i, \mathrm{s}^i, \mathrm{w}^i \rangle$ for the ith cluster. Here, $\mathbf{c}^i$ specifies the plane center, $\mathbf{n}^i$ represents the plane normal vector, a^i indicates the plane area, s^i identifies the plane ID, and w^i refers to the score value, which will be expounded upon in the ensuing step.

Thirdly, the newly detected planes are used to update the existing plane set for terrain representation, *i.e.* adding a new plane, merging multiple planes or deleting an existing plane, by evaluating the score value of each existing plane. Before calculating the score value of each function, we first pair each newly detected plane with every existing plane and calculate the Intersection over Union (IoU) as

$$IoU_{ij} = \frac{\eta_n^i \cap \eta_e^j}{\eta_n^i \cup \eta_e^j}. \tag{23}$$

The process of adjusting the scores of existing planes involves three distinct conditions: (1) In the event of a newly detected plane exhibiting a high IoU with an existing plane, the newly detected plane is merged into the existing one, and the score value of the latter is incremented by 1; (2) If a newly detected plane does not possess any pair with an existing plane, the newly detected plane is added to the set of existing planes with an initialized score value of 1; (3) If an existing plane in the scope of I_D^s does not possess a paired newly detected plane, its score value is decremented by 1. Any existing planes that exhibit a score value of zero are subsequently removed from the set of existing planes.

Finally, we reconstruct representations of terrain information with the set of existing planes. Ramps and level grounds are extracted directly based on slope θ and area a, while stairs are constructed iteratively from at least two planes with two preset thresholds, the distance of centers $d_{ij} = |\mathbf{c}_i - \mathbf{c}_j|$ and the vertical distance $h_{ij} = |z_i - z_j|$. Terrains consist of links between planes and are organized as hash tables, enabling easy projection of predicted footprints to grounds, and direct application of any updates to the planes. The succinct representation of terrain information is denoted by $\boldsymbol{T} = \{T := \langle \tau, \mathrm{s}, \boldsymbol{\Gamma} \rangle\}$, where τ denotes the type, s indicates the terrain ID, and $\boldsymbol{\Gamma}$ refers to the corresponding planes.

III. Experiments Setup

A. Experiment Preparation

Our proposed method is implemented in the experiment with Intel Realsense D455 as the sensor module, which contains an RGB-D camera and an internal IMU. The sensor module is attached to the human body with two configurations: mounted on a helmet or mounted on a chest bag. The gait phase is captured by a pair of force-sensitive insoles. All sensors are connected to a PC and the multi-thread communication is realized within the structure of Robot Operating System (ROS), as shown in Fig. 2. By conducting optimization on memory reading and writing, the prediction time consumption can be constrained within 34ms.

A total number of 15 subjects are recruited for the experiments, with a large variation of heights: group I (three subjects (165.10 cm, 165.50 cm, 164.80cm, average of 165.13 cm), group II (three subjects (170.80 cm, 172.00 cm, 171.10cm, average of 171.30 cm), group III (three subject (175.80 cm, 175.20cm, 174.10cm, average of 175.03 cm), group IV (three subjects (181.50 cm, 180.80 cm, 183.10 cm, average of 181.80 cm), and group V (three subjects (188.00 cm, 186.90cm, 188.20cm, average of 187.70 cm). All subjects signed the informed consent before experiments, and the experiments have been approved by the Local Ethics Committee of Peking University. It is important to underscore that while participants were categorized into five groups, each group underwent the identical set of experiments. Consequently, each experimental condition involved 15 participants, yielding statistically significant data. The rationale for this grouping was solely to ensure a diversity of participant heights, thereby testing the robustness of our proposed algorithm across different camera positions and step sizes.

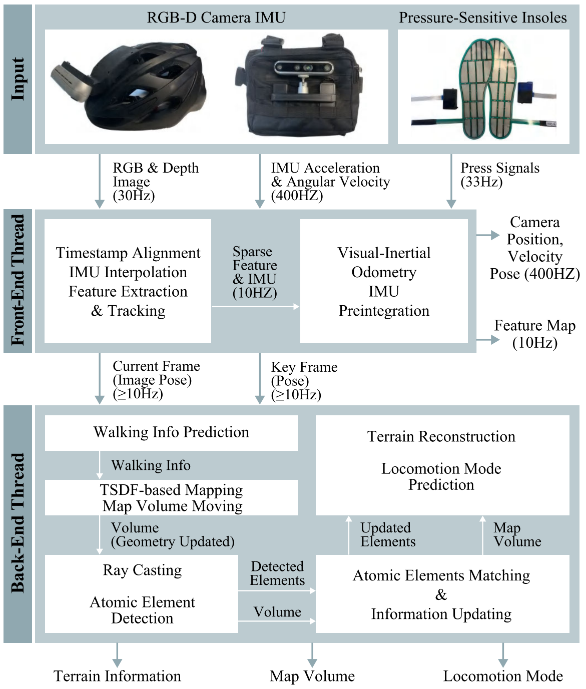

Fig. 2. The hardware and software implementation. ROS structure is applied to realize multi-thread communications among various sensors and processors.

B. Experiment Protocol

1) Generality Test: One major advantage of our proposed method is generality, which means that it can conduct accurate predictions for different subjects, terrain parameters, walking patterns, and even mounting positions with the same algorithm and only one set of fixed parameters. No tuning is required when transferring to different scenarios. Therefore, this generality test experiment aims to verify the robustness of our method to condition variations on human and the environment. Four types of condition variations are considered in this experiment.

a) Terrain type variation: This experiment selects three ramp slopes (11.8°, 15.8° and 17.8°), and three stair heights (15 cm, 20 cm, and 25 cm).

b) Subject height variation: The variation in subject heights can cause differences in sensor position and step size. As mentioned in the previous subsection, we recruit five groups of subjects with heights ranging from 164.80 cm to 188.20 cm.

c) Sensor position variation: This experiment selects two most popular sensor mounting positions: (1) on the head or (2) in front of the chest. Different mounting positions will greatly affect the sensing region. The head-mounting configuration also introduces additional disturbances to sensing from the frequent random movements of head.

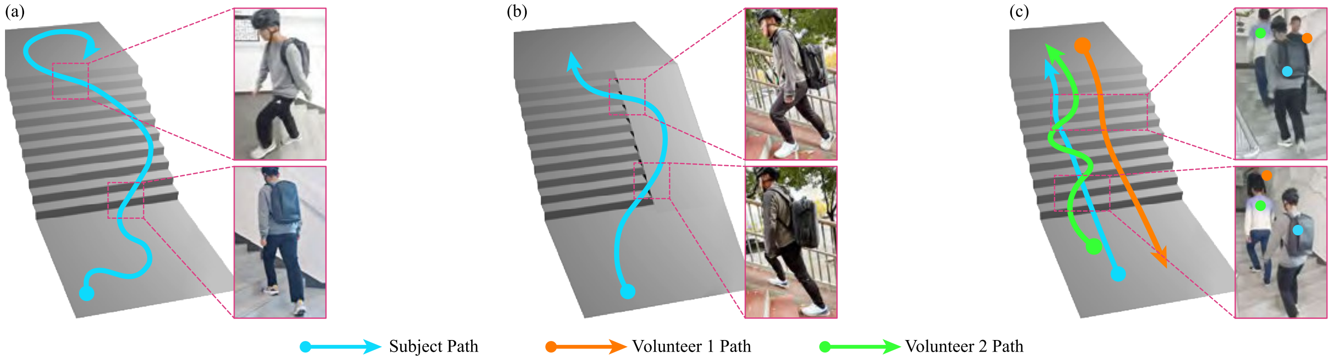

Fig. 3. Experiment protocols. (a) Unconventional Camera View: The subject walks upwards when frequently changing the camera view in a large angle so that the terrain is only partially captured. (b) Multi-direction and Outdoor Walking: The subject walks upwards while switching between two adjacent parallel terrains in the outdoor environment. (c) External Disturbance: The subject is walking on the stairs while two volunteers walking around the subject in different directions to partially or even largely block the camera view.

TABLE II
THE MEAN PREDICTION ACCURACY (%) OF EACH CONDITION COMBINATION IN THE GENERALITY TEST EXPERIMENTS

Group		Stair(15cm)		Stair(20cm)		Stair(25cm)		Ramp(11.8°)		Ramp(15.8°)		Ramp(17.8°)	
		Head	Chest	Head	Chest	Head	Chest	Head	Chest	Head	Chest	Head	Chest
I	Wander	100.00	100.00	100.00	98.15	100.00	98.21	100.00	97.73	98.44	98.46	98.63	98.67
	Walk	97.83	98.00	97.73	100.00	97.73	100.00	100.00	100.00	98.25	100.00	100.00	100.00
	Jog	97.96	100.00	97.87	100.00	97.96	100.00	96.97	100.00	100.00	98.18	98.11	98.28
II	Wander	100.00	100.00	100.00	100.00	100.00	98.08	98.63	100.00	98.08	100.00	98.68	98.68
	Walk	100.00	100.00	100.00	100.00	100.00	97.14	98.55	100.00	100.00	98.11	98.33	100.00
	Jog	100.00	97.74	97.14	98.31	97.72	97.78	100.00	98.59	100.00	100.00	100.00	98.18
III	Wander	99.06	98.17	97.75	98.06	98.25	100.00	100.00	97.18	100.00	98.51	96.77	98.46
	Walk	100.00	98.99	98.94	98.97	100.00	100.00	100.00	100.00	98.46	98.33	98.18	98.36
	Jog	98.72	100.00	100.00	100.00	100.00	97.87	97.72	100.00	98.15	100.00	100.00	98.18
IV	Wander	98.15	98.02	98.15	99.07	98.04	96.36	99.23	99.38	98.59	98.67	98.46	100.00
	Walk	100.00	99.07	100.00	98.99	100.00	100.00	99.10	99.17	98.36	100.00	98.44	98.46
	Jog	100.00	98.88	98.86	100.00	100.00	100.00	97.30	99.12	100.00	97.96	100.00	98.31
V	Wander	99.07	98.25	100.00	99.02	100.00	98.11	97.66	100.00	100.00	100.00	97.50	98.39
	Walk	100.00	98.11	98.86	98.55	100.00	98.11	98.55	100.00	98.31	100.00	100.00	98.21
	Jog	98.90	97.78	100.00	98.86	97.56	97.50	98.15	98.78	100.00	98.04	97.50	97.78

d) Walking pattern variation: Each subject is asked to walk in three different patterns: wandering, regular walking, and jogging. This can cause fluctuations in walking speed and also step size.

Five different locomotion modes are predicted in our experiments, including Level Ground (LG), Stair Ascending (SA), Stair Descending (SD), Ramp Ascending (RA), and Ramp Descending (RD). Each group of subjects are asked to walk in the LG → SA → LG → SD → LG procedure for each stair height, and LG → RA → LG → RD → LG for each ramp slope. Each terrain type is combined with every sensor position and walking pattern to form one condition combination. Each experiment is repeated three times to ensure repeatability. Therefore, a total number of 180 condition combinations and thus 540 experiments are conducted in this generality test.

2) Unconventional Camera View: Conventionally, the terrain information should largely occupy the central area of the camera view, and the view angle should be near-front. However, in many practical applications, the desired terrain may be only partially observed at a near-side or large-angle view, and may only occupy a corner of the camera view. This experiment is to show that, by applying terrain reconstruction techniques, our method is able to make a stable prediction even for these aforementioned unconventional camera views. The experiment protocol is shown in 3(a), where the subject is asked to walk with a large and changing view angle to the terrain, and make an abrupt change of walking direction in the middle of the path. A considerable percentage of the camera view during the walking is occupied by the wall on the sides. Each subject is requested to repeat the process three times.

3) Multi-Direction and Outdoor Walking: Another advantage of the proposed method is that VIO can detect the walking directions in the 3D space, so that accurate prediction can be made even if the subject changes walking directions and thus selects different terrain types. The importance of multi-directional walking and selection of terrain has been well addressed in [15]. In addition, it is also important to verify the effectiveness of the vision-based method in outdoor environments. The experiment protocol is shown in Fig. 3(b), where the subject walks outdoors while frequently switching between two parallel adjacent terrains, which contains switching among LG, SA and RA. Each subject is requested to repeat the process three times.

4) External Disturbances: One advantage of using terrain *reconstruction* instead of *detection* is that the terrain recognition results from reconstruction are very robust against various external disturbances, of which moving objects are representative. This set of experiments aims to test the robustness of our method in the existence of external moving obstacles that largely disturb the image input in a certain period of time. The experiment protocol is shown in Fig. 3(c). The subject walks along the LG → SA → LG path. Two volunteers walk around the subject in different directions to partially or even largely block the sensor view. Each subject is requested to repeat the process three times.

5) Odometry Performance: The efficacy of the proposed D-VIO is assessed through an open-source dataset focused on estimation accuracy and temporal efficiency, and a closed-loop test targeting estimation drift. As previously noted, the VINS-Mono framework was chosen due to its computational efficiency, high update frequency, robust initialization process, and provision for secondary development. Consequently, the objectives of this performance evaluation are two-fold: firstly, to conduct an ablation study demonstrating the enhancements made from the original VINS-Mono algorithm; and secondly, to engage in comparative analysis with other prominent VIO systems based on the VINS-Mono framework.

a) Open-source dataset test: The famous open-source dataset, EuRoC MAV [35], is applied. The proposed D-VIO is compared with VINS-Mono, VINS-RGBD, and VINS-Fusion on accuracy (measured by the Root Mean Square Error (RMSE) of Absolute Pose Error (APE) of the estimated trajectory) and time consumption. Depth information can be extracted from stereo images from the dataset.

b) Closed-loop test: Each subject is asked to complete a looping trajectory from the identical starting and ending points with the proposed hardware in this paper. The loop drifts of D-VIO, VINS-Mono and VINS-RGBD from the same trajectory are calculated for comparison. It should be noted that VINS-Fusion is not included in this test due to the inability of the RGB-D camera used in the experiment to provide stereo inputs.

IV. Experiment Results

A. Generality Test

1) Prediction Results: The experiment results of the 180 condition combinations in the generality test are demonstrated in TABLE II. The average accuracy of all experiments reaches 99.00% ± 0.95%. The prediction accuracy for each type of terrain, each motion pattern, and each sensor mounting position is concluded in TABLE III. It can be witnessed that the proposed method reaches high accuracy under all six types of terrains, three walking patterns, and two mounting positions. On the other hand, it can be shown in Fig. 4(a) that the proposed method reaches high accuracy in all five locomotion modes. Both show that our method is general to condition variations, such as terrain types, motion patterns, sensor positions, and subject heights.

2) Condition Variations: The condition variations are also measured explicitly in the experiments. The average sensor mounting height of the five subject groups are 1.64 m, 1.68 m, 1.75 m, 1.79 m, and 1.84 m for the head-mounting configuration, and 1.11 m, 1.11 m, 1.18 m, 1.24 m, 1.28 m for the chest-mounting configuration. The pose of the sensor fluctuates greatly during walking, especially for the head-mounting configuration, which leads to a large difference in camera views. As shown in Fig. 4 (b), the camera pitch angle has an average pitch of 0.14 rad, but ranges from −0.66 rad to 1.18 rad with a standard deviation of 0.26 rad. The variation of walking pattern and subject height leads to a variation of subject velocity, as shown in Fig. 4 (c). The subject velocity in the generality test experiments has an average of 0.78 m/s, but ranges from 0.05 m/s to 2.12 m/s with a standard deviation of 0.27 m/s.

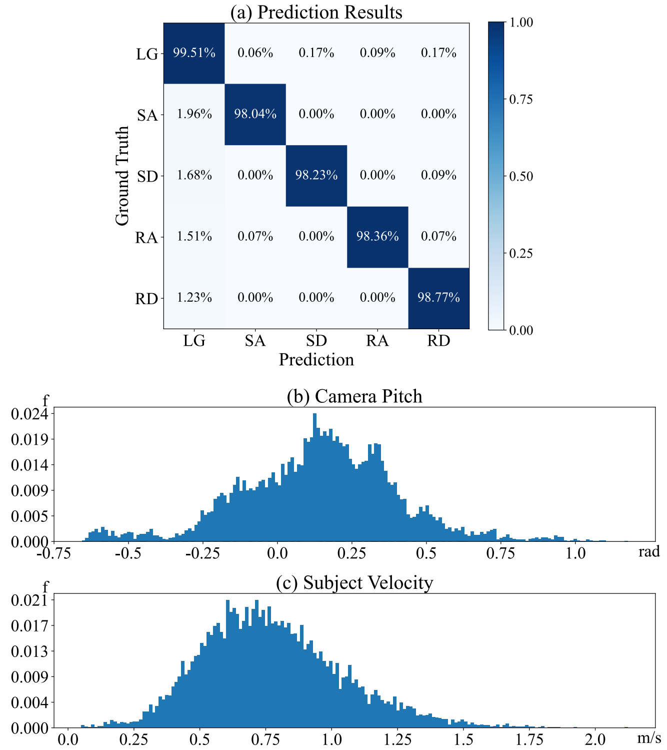

Fig. 4. (a) Confusion matrix of prediction accuracy in the generality test; (b) Histogram of camera pitch variation in the generality test; (c) Histogram of subject velocity variation in the generality test.

B. Unconventional Camera View

The average prediction accuracy for all subjects in the unconventional camera view experiment in Fig. 3(a) is 99.07% ± 1.41%, which reaches the accuracy of the conventional camera views shown in Section IV-A. This demonstrates that our method is able to fully handle the disturbance introduced by the large and abrupt changes of camera views.

C. Multi-Direction and Outdoor Walking

The average prediction accuracy for all subjects in the multi-direction and outdoor walking experiment in Fig. 3(b) is 98.18% ± 1.76%, which is of the same level with indoor, one-direction walking experiments. This verifies that our method is robust to direction change and the outdoor environment.

TABLE III
THE MEAN PREDICTION ACCURACY (%) FROM DIFFERENT CONDITION VARIATIONS

Terrain	Accuracy		Motion Pattern	Accuracy	Mounting Position	Accuracy
15 cm Stair	99.09 ± 0.88	99.03 ± 1.21	Wander	98.88 ± 1.02	Head	99.03 ± 1.04
20 cm Stair	99.11 ± 0.89					
25 cm Stair	98.82 ± 1.24		Walk	99.29 ± 0.87		
11.8° Ramp	99.06 ± 1.00	98.98 ± 1.14			Chest	98.90 ± 1.04
15.8° Ramp	99.09 ± 1.11		Jog	98.79 ± 1.17		
17.8° Ramp	98.63 ± 0.95					

TABLE IV
ODOMETRY TESTS ON THE EUROC MAV DATASET

	Odometry Sequence	Ours	VINS-Mono	VINS-RGBD	VINS-Fusion
RMSE of the APE	MH 01	**0.05**	0.10	0.09	0.12
	MH 02	**0.06**	0.18	0.16	0.23
	MH 03	**0.09**	0.24	0.20	0.28
	MH 04	**0.19**	0.30	0.29	0.22
	MH 05	**0.17**	0.31	0.32	0.19
	V1 01	**0.05**	0.10	**0.05**	0.10
	V1 02	**0.06**	0.08	**0.06**	0.11
	V1 03	**0.09**	0.18	0.15	0.10
	V2 01	**0.07**	0.08	**0.07**	0.09
	V2 02	**0.08**	0.10	0.09	0.09
Average Time Consumption (ms)		**31**	63	56	71

D. External Disturbances

The average prediction accuracy for all subjects in the external disturbance experiment in Fig. 3(c) is 98.59% ± 1.91%, which shows the same level of accuracy with other experiments. This shows that our terrain reconstruction technique can circumvent the negative influence of moving objects that block the camera views.

E. Odometry Performance

1) Open-Source Dataset Test: The accuracy and time consumption of D-VIO, VINS-Mono, VINS-RGBD and VINS-Fusion are shown in TABLE IV. It can be witnessed that the proposed D-VIO presents comparable or slightly superior performance than other VINS series in all test sequences, while significantly reducing the time consumption.

2) Closed-Loop Test: The loop drifts of D-VIO, VINS-Mono and VINS-RGBD are shown in TABLE V. The drift of VINS-Mono grows rapidly, as the locomotion mode prediction application has sparse visual features and human motion randomness. The proposed D-VIO shows comparative yet slightly better performance in contrast with the VINS-RGBD in the test case.

3) Time Consumption: As mentioned in TABLE IV, the proposed D-VIO demonstrates an approximate reduction of 50% on time consumption with other VINS-series VIOs for the EuRoC MAV Dataset.

4) Ablation Study: The ablation study is conducted on the open-source dataset test and closed-loop test, on both estimation accuracy and time consumption. Compared with the original VINS-Mono, the proposed D-VIO significantly improves the estimation accuracy and reduces the time consumption by half. The closed-loop test also shows that VINS-Mono faces challenges on locomotion mode prediction applications with sparse features and random motions.

V. DISCUSSION

Locomotion mode prediction has been a fundamental task for the assistance applications of lower-limb wearable robots, such as prostheses and exoskeletons. The basic idea is that the locomotion assistance strategy differs for different locomotion modes, which typically include ground walking, and ascending or descending on ramps and stairs. Compared with IMU-based algorithms [5], [6], [7], [8], [9] that recognize the locomotion mode by identifying and classifying the walking pattern, the vision-based algorithms [11], [12], [13], [14], [15], [16] release the burden of multiple-sensor deployment, advance the time instance to complete each prediction, and improves the robustness under individual variations on walking pattern for the same locomotion mode. With advanced ML techniques, the State-of-the-Art (SOTA) vision-based algorithms reach high accuracy under variations of terrain appearance.

This paper presents a learning-free method for vision-based locomotion mode prediction, by employing terrain reconstruction techniques. The walking information of subjects is explicitly measured and considered in the prediction with the introduction of VIO. Compared with other algorithms based on ML techniques, this method releases the workload of data collection through real-world experiments [7] or data augmentation [12], while reaching the same level of prediction accuracy (99.00% ± 0.97%) with SOTA ML-based techniques, under comprehensive variations of terrain appearance, walking pattern, subject heights and camera deployment, as shown in Section IV-A and also TABLE III.

Although the significance of walking information, such as step size and walking direction, has been emphasized in previous works, only limited ML-based works have taken it into consideration, either by introducing additional sensors [15], or setting constraints on motion [16]. Our method, by introducing VIO for the explicit consideration of walking information, shows commendable robustness under variations of step size and walking direction, as shown in Section IV-B and Section IV-C. The terrain reconstruction technique makes the proposed method robust to single image inconsistencies. As a result, the proposed method avoids the calibration process of camera position and pose [16]. Actually, the camera for our method can be mounted with different heights and poses, even on the head where the mounting position and pose remain fluctuating during the prediction, without any requirements for the modification or tuning of parameters, as verified

TABLE V
LOOP DRIFTS IN THE CLOSED-LOOP TEST

Odometry	Drift-x (m)	Drift-y (m)	Drift-z (m)
Ours	**-0.216**	**0.023**	**0.136**
VINS-Mono	>10	>10	>10
VINS-RGBD	-0.485	0.148	0.226

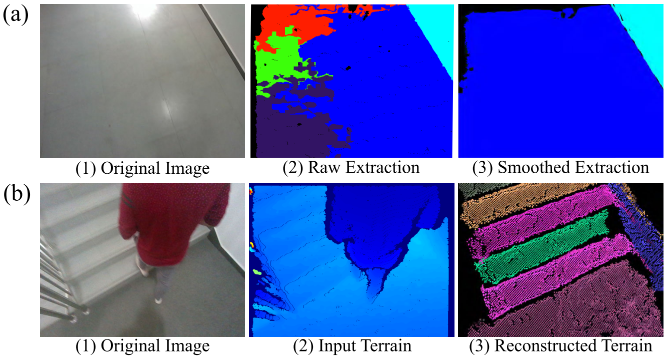

Fig. 5. The proposed terrain reconstruction method can (a) improve the completeness of the terrain surface under various lighting and environmental conditions. (b) provide robust terrain representation under external disturbance, such as a moving obstacle.

experimentally in TABLE III. This also indicates that our method can remain functional with external disturbance to camera images, such as moving obstacles, e.g. other people passing by and blocking the terrain in a certain period of time. This is demonstrated in the experiment in Section III-B4.

Our proposed method for the first time introduces VIO and terrain reconstruction in the field of wearable robots for locomotion mode prediction. However, the pipeline and internal algorithms need to be specially designed to achieve high performance for this specific scenario. As for VIO, the implementation on wearable configuration requires lightweight computation demand, robustness under random and abrupt human motions, and stability in accuracy under sparse features. Therefore, although researchers continue proposing novel algorithms with increasing performance, [22], [32], [36], this study adopts the VINS-Mono framework [23] for its comprehensive advantages to leverage accuracy and computation demand. In this paper, we present the D-VIO algorithm, which, by adding additional depth constraints on both initialization and estimation stages, greatly improves the accuracy and robustness under our specific scenario with random motion and sparse features. An ablation study is conducted to show our advantages from VINS-Mono, as shown in Section IV-E. The proposed D-VIO is also tested with the famous open-source dataset EuRoC MAV and also with closed-loop experiment on accuracy and time consumption, in contrast with other prestigious VIOs under the VINS-Mono framework, such as VINS-RGBD and VINS-Fusion. The results in TABLE IV and TABLE V show that the proposed VIO achieves better performance in accuracy and time in our application of locomotion mode prediction. As for terrain reconstruction, two major concerns are the stable and succinct representation of terrain information, and the robustness under external disturbance. By conducting Ray Casting and IoU-based plane updating algorithms on the TSDF mapping, the proposed method is able to extract smooth and relatively complete terrain surface under various lighting and environment conditions, as shown in Fig. 5(a), and also remain robust under external disturbance such as obstacles, as shown in Fig. 5(b).

One interesting finding of our experiment results is that, although high accuracy is maintained in all designed condition variations, the corner cases occur mostly in the wandering and jogging motion patterns, as indicated in TABLE III. The reason is that the subjects sometimes become bewildered when forced to walk in a pattern they are not familiar with, especially near the transition region between different terrains. In these cases, spikes in velocity occur and the walking information becomes unpredictable. This shall be largely improved in real applications, when people walk in a more natural and thus continuous manner.

VI. CONCLUSION

This paper proposes a novel locomotion mode prediction method, which incorporates terrain reconstruction and VIO techniques to build robust descriptions of the terrain and walking information, so that the data generation and training process in machine learning-based algorithms are omitted. The specially designed D-VIO and terrain reconstruction algorithm allow for high performance in our specific scenario, with the lack of geometric features for terrain images, and randomness and abruptness in human motions. The proposed method is able to maintain high accuracy (99.00% on average) with variations of subject heights, terrain geometric parameters, walking speeds and patterns, and also the mounting position and pose of the camera. Results also show that the accuracy can be maintained in the outdoor environment, multi-direction walking, or in the presence of moving objects to disturb the image inputs. Future works may include extracting gait phase information from the VIO results [37], converting the locomotion modes to parameterized continuous description and designing corresponding controllers for better performance in assistance, developing lightweight detection and representation technologies to facilitate the computation and storage level of embedded systems, and establishing an open-source dataset with a diversified group of subjects conducting locomotion on different terrains, under different walking patterns, with different camera types, positions and poses, so that a fair comparison can be conducted among different locomotion mode prediction algorithms.

ACKNOWLEDGMENT

The authors would like to thank Zhen Chen, Hao Liang, and Zihang Zhao for the technical supports, also would like to thank Dr. Boren Li for the insightful discussion on problem formulation, and also would like to thank G. Fu, H. Chen, H. Wang, H. Zhong, L. Yan, M. Wu, Q. Tang, X. Di, X. Guo, X. Wu, Z. Zhang, M. Jiang, R. Yang, Z. Wang, Q. Feng, and Z. Zhao for their participation in the experiments.

REFERENCES

[1] M. K. Shepherd, A. M. Simon, J. Zisk, and L. J. Hargrove, "Patient-preferred prosthetic ankle-foot alignment for ramps and level-ground walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 52–59, 2021.

[2] P. Slade, M. J. Kochenderfer, S. L. Delp, and S. H. Collins, "Personalizing exoskeleton assistance while walking in the real world," *Nature*, vol. 610, no. 7931, pp. 277–282, Oct. 2022.

[3] A. Alili et al., "A novel framework to facilitate user preferred tuning for a robotic knee prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 895–903, 2023.

[4] Q. Wang, K. Yuan, J. Zhu, and L. Wang, "Walk the walk: A lightweight active transtibial prosthesis," *IEEE Robot. Autom. Mag.*, vol. 22, no. 4, pp. 80–89, Dec. 2015.

[5] F. Gao, G. Liu, F. Liang, and W.-H. Liao, "IMU-based locomotion mode identification for transtibial prostheses, orthoses, and exoskeletons," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1334–1343, Jun. 2020.

[6] B.-Y. Su, J. Wang, S.-Q. Liu, M. Sheng, J. Jiang, and K. Xiang, "A CNN-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 1032–1042, May 2019.

[7] X. Liu and Q. Wang, "Real-time locomotion mode recognition and assistive torque control for unilateral knee exoskeleton on different terrains," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 6, pp. 2722–2732, Dec. 2020.

[8] J. Camargo, W. Flanagan, N. Csomay-Shanklin, B. Kanwar, and A. Young, "A machine learning strategy for locomotion classification and parameter estimation using fusion of wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1569–1578, May 2021.

[9] I. Kang, D. D. Molinaro, G. Choi, J. Camargo, and A. J. Young, "Subject-independent continuous locomotion mode classification for robotic hip exoskeleton applications," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 10, pp. 3234–3242, Oct. 2022.

[10] J. Camargo, A. Ramanathan, W. Flanagan, and A. Young, "A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions," *J. Biomech.*, vol. 119, Apr. 2021, Art. no. 110320.

[11] K. Zhang et al., "Environmental features recognition for lower limb prostheses toward predictive walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 465–476, Mar. 2019.

[12] C. Chen, K. Zhang, Y. Leng, X. Chen, and C. Fu, "Unsupervised sim-to-real adaptation for environmental recognition in assistive walking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1350–1360, 2022.

[13] N. E. Krausz and L. J. Hargrove, "Sensor fusion of vision, kinetics, and kinematics for forward prediction during walking with a transfemoral prosthesis," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 3, pp. 813–824, Aug. 2021.

[14] Y. Qian et al., "Predictive locomotion mode recognition and accurate gait phase estimation for hip exoskeleton on various terrains," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6439–6446, Jul. 2022.

[15] M. Li, B. Zhong, E. Lobaton, and H. Huang, "Fusion of human gaze and machine vision for predicting intended locomotion mode," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1103–1112, 2022.

[16] A. H. A. Al-Dabbagh and R. Ronsse, "Depth vision-based terrain detection algorithm during human locomotion," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 4, pp. 1010–1021, Nov. 2022.

[17] A. H. A. Al-dabbagh and R. Ronsse, "A review of terrain detection systems for applications in locomotion assistance," *Robot. Auto. Syst.*, vol. 133, Nov. 2020, Art. no. 103628.

[18] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

[19] H. Ray, H. Pfister, D. Silver, and T. A. Cook, "Ray casting architectures for volume visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 5, no. 3, pp. 210–223, Jul./Sep. 1999.

[20] 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.*, Apr. 2007, pp. 3565–3572.

[21] 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.

[22] D. Chen et al., "VIP-SLAM: An efficient tightly-coupled RGB-D visual inertial planar SLAM," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 5615–5621.

[23] 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.

[24] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[25] Z. Shan, R. Li, and S. Schwertfeger, "RGBD-inertial trajectory estimation and mapping for ground robots," *Sensors*, vol. 19, no. 10, p. 2251, May 2019.

[26] T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 3662–3669.

[27] J. Zhang, C. Zhu, L. Zheng, and K. Xu, "ROSEFusion: Random optimization for online dense reconstruction under fast camera motion," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–17, Aug. 2021.

[28] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[29] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "IMAP: Implicit mapping and positioning in real-time," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6229–6238.

[30] Z. Zhu et al., "NICE-SLAM: Neural implicit scalable encoding for SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12786–12796.

[31] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1697–1716, Dec. 2016.

[32] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 3, pp. 1–18, Jun. 2017.

[33] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2, 1981, pp. 674–679.

[34] C. Feng, Y. Taguchi, and V. R. Kamat, "Fast plane extraction in organized point clouds using agglomerative hierarchical clustering," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 6218–6225.

[35] M. Burri et al., "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, Sep. 2016.

[36] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[37] J. Luo, Y. Zhao, L. Ruan, S. Mao, and C. Fu, "Estimation of CoM and CoP trajectories during human walking based on a wearable visual odometry device," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 1, pp. 396–409, Jan. 2022.