IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

PO-GVINS: A Tightly Coupled GNSS-Visual-Inertial Navigation Framework Using Pose-Only Representation

Zhuo Xu, Feng Zhu, Zihang Zhang, Chang Jian, Jiarui Lv, Yuantai Zhang, and Xiaohong Zhang

***Abstract*—Accurate and reliable positioning is essential for perception, decision-making, and other high-level applications in autonomous driving, autonomous aerial vehicles, and intelligent robotics. Due to the inherent limitations of standalone sensors, integrating heterogeneous sensors with complementary capabilities is an effective approach to achieving this goal. The visual-inertial navigation system (VINS) fuses visual cameras and inertial measurement units (IMUs) to explore unknown environments. It requires a priori knowledge of 3D features and jointly estimates camera poses and feature positions, which inevitably introduces feature linearization errors. Meanwhile, the dimensionality of the system state increases with abundant textures, degrading real-time performance. To eliminate accumulated errors from VINS, frameworks further fuse measurements from the Global Navigation Satellite System (GNSS), but still suffer from similar limitations. To address the aforementioned issues, we propose a filtering-based, tightly coupled GNSS-visual-inertial positioning framework with a pose-only formulation applied to VINS, termed PO-GVINS. We first apply the PO formulation to our VINS (PO-VINS). GNSS raw measurements are subsequently incorporated, with integer ambiguities resolved, to achieve accurate and drift-free state estimation. Extensive experiments demonstrate that the proposed PO-VINS significantly outperforms the multi-state constraint Kalman filter (MSCKF) and achieves accuracy comparable to that of optimization-based VINS. By further incorporating GNSS measurements, PO-GVINS achieves accurate, drift-free state estimation, making it a robust solution for positioning in challenging environments.**

Received 16 June 2025; accepted 23 August 2025. Date of publication 5 September 2025; date of current version 11 September 2025. This article was recommended for publication by Associate Editor S. Santra and Editor A. Banerjee upon evaluation of the reviewers' comments. This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 42425003, in part by the National Key Research and Development Program of China under Grant 2022YFB3903802, and in part by the National Natural Science Foundation of China under Grant 42374031 and Grant 42388102. *(Corresponding author: Feng Zhu.)*

Zhuo Xu, Zihang Zhang, Chang Jian, Jiarui Lv, and Yuantai Zhang are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: zhuoxu@whu.edu.cn; zihangzhang@whu.edu.cn; 2024202140027@whu.edu.cn; lvjiauri@whu.edu.cn; officialtai@whu.edu.cn).

Feng Zhu is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China (e-mail: fzhu@whu.edu.cn).

Xiaohong Zhang is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with the Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430079, China (e-mail: xhzhang@sgg.whu.edu.cn).

Datasets used in this research are available at https://gitee.com/lv-jiarui/SmartPNT-MSF-Datasets.git.

Digital Object Identifier 10.1109/LRA.2025.3606792

***Index Terms*—Multi-sensor fusion navigation, pose-only formulation, tightly coupling, global navigation satellite system, visual-inertial system.**

I. Introduction

CONTINUOUS, reliable, and high-precision positioning is essential for emerging applications, such as Autonomous ariel vehicles (AAVs), autonomous cars, and other intelligent robots. However, due to inherent limitations of standalone sensors, multi-sensor fusion, which incorporates complementary information from heterogeneous sensors, is considered an effective means of addressing this issue.

Exploring unknown environments with a monocular camera has received lots of research attention, namely simultaneous localization and mapping (SLAM), due to its cheap, low-cost, and ample semantic information. However, it suffers from scale ambiguity, motion blur, and illumination change, leading to incapable use of estimates for robots. A feasible method is to incorporate inertial measurement unit (IMU), which provides rigorous dynamic model by observing motion of carriers and is independent of external infrastructures, enabling to recover the metric scale [1], [2], as well as to resist the degenerate cases of a single camera. The visual-inertial navigation system (VINS) has also received lots of research attention and been applied in spacecraft landing and descent [3], [4], and Autonomous ariel vehicle navigation [5]. However, there are several key challenges that VINS must address. First, it has been proven that VINS has four unobservable directions [1], where drifts will be accumulated along the four directions, leading to unacceptable estimations. Second, VINS relies on ample stationary textures from surroundings. On one hand, those rich features can make VINS more robust and can relieve accumulation of errors. On the other hand, however, the dimensionality of system state increases with 3D feature position modelled as parameters, degrading real-time performance. Third, VINS is a nonlinear system which needs linearization to apply optimization, inevitably introducing linearization error.

Complementarily, global navigation satellite system (GNSS) provides absolute and high-precision measurements from navigation satellites, which can theoretically achieve centimeter-level positioning in open-area environments. On one hand, fusing GNSS measurements can easily eliminate drift error caused by VINS and ensures drift-free positioning results. On the other

hand, the performance of GNSS rapidly degrades due to its vulnerable features under urban viaducts and other signal blocked areas, where VINS is capable to maintain acceptable positioning accuracy. Currently, frameworks of fusing GNSS, IMU, and monocular camera can be categorized into the loosely coupled and the tightly coupled. Loosely coupled methods generally regard GNSS and VINS as independent closed boxes and directly integrate their estimates. While this approach is time-efficient, it overlooks potential GNSS information, particularly in scenarios where the number of tracked satellites falls below four. This limitation often results in unnecessary accumulation of errors in VINS systems. Conversely, tightly coupled methods make full use of raw measurements from heterogeneous sensors, and thus it enables to achieve more accurate and consistent state estimation [6].

From the perspective of efficiency, in addition to constant dimension of state vector, i.e., camera poses in sliding window, the time cost is severely dependent on the number of correctly matched features in VINS. Dimensional explosion occurs due to additional parameters added to the state vector, i.e., position of each feature point [7]. To this end, researchers employed inverse depth expression in their fusion framework, reducing the dimension of each feature point from 3 to 1 [8]. Furthermore, MSCKF projects visual bearing measurements onto the null space of the feature Jacobian matrix, which explicitly eliminates feature points during state estimation [2]. However, the time complexity keeps increasing no matter using inverse depth representation or the null space projection.

Besides, as long as feature positions are modelled as parameters, linearization error of the feature is inevitably introduced. Although the null space projection used in MSCKF eliminates feature parameters, it operates *after* linearization, ignoring the higher order terms.

Recent research indicated that visual pose-only (PO) formulation expresses feature depth by a pair of camera poses, eliminating parameters of feature points *before* linearization. And it has been successfully applied to visual reconstruction, Inspired by this, this letter proposes a tightly integrated RTK-visual-inertial framework, where visual measurements are reformulated by PO model, namely PO-GVINS. In brief, the main contribution of this work includes:

a) We propose a filtering-based and tightly coupled framework to integrate raw measurements from GNSS-RTK (both pseudorange and carrier phase), IMU and a monocular camera with continuous integer ambiguity resolved.
b) The PO representation is applied in our visual observation model to avoid linearization errors introduced by 3D feature position when using null space projection in MSCKF and dimensional explosion. Thus, the PO formulation mitigates accumulative error, ensures an accurate estimation and improves the real-time performance.
c) The performance of proposed PO-GVINS is evaluated through dataset with challenging GNSS degenerated scenarios. Compared with the state-of-the-art open-source libraries, VINS-Fusion [9] and GVINS [10], the results show that PO formulation shows significant improvements and the proposed PO-GVINS achieves a more accurate and drift-free estimation. It is also shown that the PO formulation accelerates the VINS The dataset used in this research is available at: https://gitee.com/lv-jiarui/SmartPNT-MSF-Datasets.git.

II. Related Work

Researches in GNSS/IMU/camera have flourished in the past five years and numerous enlightening methods have been explored. Relevant literature is revisited from VINS and heterogeneous sensors fusion respectively.

A. Visual-Inertial Navigation System

Tightly coupled frameworks, where parameters of cameras and IMUs are joint estimated using raw measurements from each sensor, showing better accuracy [11]. A time efficient and profoundly influential framework adopts a sliding window filtering (SWF) to constrain multiple cameras and project visual bearing measurements onto the null space of the feature Jacobian matrix, rather than estimate them directly, namely MSCKF [2], [12]. Based on MSCKF, researchers further utilize mapped landmarks as well as tracked ones in spacecraft landing and descent [3], apply observability constraints to avoid wrong observability property for fast UAVs [13], [14]. Besides, other filtering methods, such as unscented Kalman filter (UKF), particle filter (PF) et al., are employed to handle the high degree of nonlinearity and non-Gaussian noise [15], [16], [17], [18], [19]. In [20], [21], invariant EKF (InEKF) has been introduced to preserve the observable structure of the navigation system. Besides, SchurVINS also utilizes stochastic clone to maintain a sliding window, but it utilizes Schur complement to build equivalent residual equation, and position of feature position can be recovered [22]. Yet the equivalent between Schur complement and null space projection has been demonstrated in [23], SchurVINS still suffers from ignoring higher order terms as stated before.

Another way to integrate IMU and camera measurements is to use graph optimization, where preintegration is utilized to process IMU measurements [24]. The earliest tightly coupled VIO system can be traced back to OKVIS [25]. VINS-mono is then proposed with loop-closure and provided ability to online estimate both intrinsic and extrinsic parameters [1]. Stereo camera is further supported in VINS-fusion [9]. Based on [26] and [27], ORBSLAM 3 is proposed which supports both monocular and stereo camera measurements fused with IMU preintegration measurements. Similarly, SVO originally supports visual and multiple cameras. And currently it has released supports of inertial sensors based on optimization [28]. In addition to indirect and semi-direct methods, [29] applied direct methods with stereo and inertial tightly integrated system. Sparse odometry (DSO)-VIO proposed a dynamic marginalization strategy in order to keep VINS consistent. [30]. And [31] proposes a delayed marginalization methods to inject IMU information into already marginalized states based on direct formulation.

Although there are works focusing on keeping system sparsity when marginalize old states [32], [33], reducing dimension of feature representation [8], and reusing previous calculations [34], both filtering-based and optimization-based approaches

still suffer from tradeoff between efficiency and accuracy. Recently, a visual pose-only (PO) representation has been proposed, which explicitly eliminates feature depth in visual measurement equation, and its equality with multiple geometry has been demonstrated [35], [36].

B. Integration of GNSS, IMU, and Visual Measurements

In order to eliminate accumulated drifts in VINS, GNSS is initially incorporated as a closed box, where position of antennas estimated by GNSS is integrated, i.e., loosely coupled [9], [37], [38], [39]. However, these loosely coupled methods cannot take full use of GNSS raw measurements, especially in those signal blocked scenarios [6]. On the contrary, tightly coupled manner provides the ability to fuse GNSS raw measurements even when the visible satellites are less than 4. [40] and [10] respectively utilize filtering-based and optimization-based framework to fuse raw measurements, but only pseudorange and doppler measurements are utilized, limiting accurate performance of GNSS. Recent researches further utilize high-precision phase measurements and ambiguity resolution techniques, enabling centimeter-level accuracy, where GNSS is respectively modelled as precise point positioning (PPP) [41], [42], [43], real-time kinematics (RTK) [6], [44], [45], and even the cutting-edge PPP-RTK technology [46], [47] with integer ambiguity resolved. However, PPP suffers a long-time convergence [6] and PPP-RTK is currently a developing technology, making RTK still the most well used and matured GNSS approach.

III. Visual-Inertial Odometry With Pose-Only Formulation

A. Pose-Only Formulation

Assume a feature point has been observed by n images, where $\boldsymbol{P}^w = (x^w, y^w, z^w)^T$ denotes feature position in world frame and $\boldsymbol{p}_i = (x_i, y_i, 1)^T$ denotes the normalized image coordinate of this feature expressed in i-th view ($i = 1, 2, \cdots, n$), satisfying

$$\boldsymbol{p}_i = \frac{1}{z_i^c}\boldsymbol{p}_i^c = \frac{1}{z_i^c}\boldsymbol{R}_i^T\left(\boldsymbol{P}^w - \boldsymbol{r}_i\right), i = 1, 2, \cdots, n \tag{1}$$

where $\boldsymbol{p}_i^c = (x_i^c, y_i^c, z_i^c)^T$ is the coordinate of this feature point in i-th frame and $(\boldsymbol{R}_i, \boldsymbol{r}_i)$ is the corresponding transformation from i-th camera frame to world frame.

Regarding two base frames i and j, the traditional two-view geometry can be expressed as

$$z_j^c\boldsymbol{p}_j = z_i^c\boldsymbol{R}_i^j\boldsymbol{p}_i + \boldsymbol{t}_{j,i} \tag{2}$$

where $\boldsymbol{R}_i^j = \boldsymbol{R}_j^T\boldsymbol{R}_i, \boldsymbol{t}_{j,i} = \boldsymbol{R}_j^T(\boldsymbol{r}_i - \boldsymbol{r}_j)$ denotes the local transformation.

To derive pose-only formulation, left multiply the antisymmetric matrix $\boldsymbol{p}_j^\wedge$ on the both side of (2):

$$-\boldsymbol{p}_j^\wedge\boldsymbol{t}_{j,i} = z_i^c\boldsymbol{p}_j^\wedge\boldsymbol{R}_i^j\boldsymbol{p}_i \tag{3}$$

Take the magnitude, then depth z_i^c can be expressed by a pair of frame poses:

$$z_i^c = \frac{\left\|\boldsymbol{p}_j^\wedge\boldsymbol{t}_{j,i}\right\|}{\theta_{i,j}} \triangleq d_i^{(i,j)}, \theta_{i,j} \triangleq \left\|\boldsymbol{p}_j^\wedge\boldsymbol{R}_i^j\boldsymbol{p}_i\right\| \tag{4}$$

Similarly, left multiplying $\left(\boldsymbol{R}_i^j\boldsymbol{p}_i\right)^\wedge$ on both side of (2), and then taking the magnitude, z_j^c can be also expressed by the same pair of poses:

$$z_j^c = \frac{\left\|\left(\boldsymbol{R}_i^j\boldsymbol{p}_i\right)^\wedge\boldsymbol{t}_{j,i}\right\|}{\left\|\left(\boldsymbol{R}_i^j\boldsymbol{p}_i\right)^\wedge\boldsymbol{p}_j\right\|} = \frac{\left\|\left(\boldsymbol{R}_i^j\boldsymbol{p}_i\right)^\wedge\boldsymbol{t}_{j,i}\right\|}{\theta_{i,j}^k} \triangleq d_j^{(i,j)} \tag{5}$$

Thus, (2) can be rewritten as:

$$d_j^{(i,j)}\boldsymbol{p}_j = d_i^{(i,j)}\boldsymbol{R}_i^j\boldsymbol{p}_i + \boldsymbol{t}_{j,i} \tag{6}$$

For any other l-th frame ($l \neq j$), we also have:

$$d_l^{(i,l)}\boldsymbol{p}_l = d_i^{(i,l)}\boldsymbol{R}_i^l\boldsymbol{p}_i + \boldsymbol{t}_{l,i} \tag{7}$$

According to definition, $d_i^{(i,l)} = d_i^{(i,j)} = z_i^c$, we substitute $d_i^{(i,l)}$ with $d_i^{(i,j)}$, deriving PO formulation of the feature point:

$$d_l^{(i,l)}\boldsymbol{p}_l = d_i^{(i,j)}\boldsymbol{R}_i^l\boldsymbol{p}_i + \boldsymbol{t}_{l,i} \tag{8}$$

Comparing with MSCKF's formulation, for example, the linearized formulation of (1) is:

$$\boldsymbol{e}^{(i)} = \boldsymbol{H}_x^{(i)}\delta\boldsymbol{x}_p^{(i)} + \boldsymbol{H}_f^{(i)}\delta\boldsymbol{x}_f^{(i)} + \boldsymbol{n}^{(i)} \tag{9}$$

where, $\boldsymbol{H}_x$ and $\boldsymbol{H}_f$ are Jacobis of camera poses and feature point respectively, and $\delta\boldsymbol{x}_p$, $\delta\boldsymbol{x}_f$ denote error state of camera poses and position of feature depth. $\boldsymbol{r}$ defines the residual and $\boldsymbol{n}$ denotes Gaussian noise. Then, by projecting $\boldsymbol{e}^{(i)}$ on the left null space of the matrix $\boldsymbol{H}_f^{(i)}$, i.e., MSCKF, $\delta\boldsymbol{x}_f^{(i)}$ is also eliminated, and thus a new residual can be derived [2]:

$$\boldsymbol{e}_o^{(i)} = \boldsymbol{H}_{ox}^{(i)}\delta\boldsymbol{x}_p^{(i)} + \boldsymbol{n}_o^{(i)} \tag{10}$$

It is important to highlight that, this null space projection only eliminates first order terms of the depth of feature point. Higher order terms are ignored during linearization. The pose-only formulation, however, is lossless when eliminating feature depth. Besides, camera poses are updated during iteration, thus the feature depth is also implicitly updated. While, it is difficult to recover feature depth in MSCKF, and even if the iteration strategy is applied, linearization point of 3D feature position maintains the original.

B. Propagation and Augmentation

IMU measurements are utilized to integrate and propagate, which is modelled as mechanization. The motion model can be described by

$$\begin{cases} \delta\dot{\boldsymbol{r}}_b = \delta\boldsymbol{v}_b + \boldsymbol{n}_r \\ \delta\dot{\boldsymbol{v}}_b = \boldsymbol{N}\delta\boldsymbol{r}_b - 2(\boldsymbol{\omega}_{ie}^w)^\wedge\delta\boldsymbol{v}_b + (\boldsymbol{f}^w)^\wedge\boldsymbol{\phi} + \boldsymbol{R}_b\delta\boldsymbol{b}_a + \boldsymbol{n}_v \\ \dot{\boldsymbol{\phi}} = -(\boldsymbol{\omega}_{ie}^w)^\wedge\boldsymbol{\phi} - \boldsymbol{R}_b\delta\boldsymbol{b}_g + \boldsymbol{n}_\phi \\ \dot{\boldsymbol{b}}_a = \boldsymbol{\eta}_a, \dot{\boldsymbol{b}}_g = \boldsymbol{\eta}_g \end{cases} \tag{11}$$

Where, $\delta\boldsymbol{x}$ denotes error state of a variable $\boldsymbol{x}$, which is defined by $\delta\boldsymbol{x} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ denotes an observation of true state $\boldsymbol{x}$. Specifically, $\boldsymbol{r}_b$, $\boldsymbol{v}_b$ denote position and velocity of body frame with respect to world frame. Subscript i stands for Earth-centered inertial (ECI) frame. $\boldsymbol{f}^w = \boldsymbol{R}_b\boldsymbol{f}^b$ denotes the accelerometers measurements. And $\boldsymbol{\omega}$ denotes angular velocity. $\boldsymbol{N}$ is the tensor of the gravitational gradients. Besides, $\boldsymbol{n}$ denotes the white Gaussian noise. Accelerometer and gyroscope biases, $\boldsymbol{b}_a$ and $\boldsymbol{b}_g$, are modelled as random walk process, $\boldsymbol{\eta}_a$ and $\boldsymbol{\eta}_g$ respectively. Additionally, error state of the rotation matrix can be obtained by $\boldsymbol{R}_b = (\boldsymbol{I} + \boldsymbol{\phi}^\wedge)\tilde{\boldsymbol{R}}_b$. In our implementation, world frame is chosen as Earth frame, and thus, $\boldsymbol{\omega}_{ie}^w = \boldsymbol{\omega}_{ie}^e$ is a constant denoting the rotation rate of the Earth. The initialization in e-frame of the whole system will be introduced in Section IV.

Then, (11) can be rewritten as matrix form

$$\delta\dot{\boldsymbol{x}}_{imu} = \boldsymbol{F}_t\delta\boldsymbol{x}_{imu} + \boldsymbol{G}_t\boldsymbol{w}. \tag{12}$$

where, IMU error state vector $\delta\boldsymbol{x}_{imu}$ and noise vector $\boldsymbol{w}$ have the following form

$$\begin{aligned}\delta\boldsymbol{x}_{imu} &= \begin{bmatrix}\delta\boldsymbol{r}_b^T & \delta\boldsymbol{v}_b^T & \boldsymbol{\phi}^T & \delta\boldsymbol{b}_a^T & \delta\boldsymbol{b}_g^T\end{bmatrix}^T,\\ \boldsymbol{w} &= \begin{bmatrix}\boldsymbol{n}_r^T & \boldsymbol{n}_v^T & \boldsymbol{n}_{\boldsymbol{\phi}}^T & \boldsymbol{n}_a^T & \boldsymbol{n}_g^T\end{bmatrix}^T\end{aligned} \tag{13}$$

When camera captures a new image at timestamp t, system error state is extended by stochastic clone [48]

$$\delta\boldsymbol{X}_t = \begin{bmatrix}\delta\boldsymbol{x}_{imu}\\ \delta\boldsymbol{x}_{t,imu}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I}_{15\times15}\\ \boldsymbol{F}_{imu}\end{bmatrix}\delta\boldsymbol{x}_{imu} \tag{14}$$

where, $\delta\boldsymbol{x}_{t,imu} = (\delta\boldsymbol{r}_{t,b}^T, \boldsymbol{\phi}_t^T)^T$ denotes the clone of IMU error state at t. $\boldsymbol{F}_{imu}$ is thus defined by

$$\boldsymbol{F}_{imu} = \begin{bmatrix}\boldsymbol{I}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times6}\\ \boldsymbol{0}_{3\times3} & \boldsymbol{0}_{3\times3} & \boldsymbol{I}_{3\times3} & \boldsymbol{0}_{3\times6}\end{bmatrix} \tag{15}$$

Then, prior constraints can be derived by covariance propagation law

$$\boldsymbol{D}_{t|t-1} = \begin{bmatrix}\boldsymbol{D}_{t-1} & (\boldsymbol{F}_{imu}\boldsymbol{D}_{t-1})^T\\ \boldsymbol{F}_{imu}\boldsymbol{D}_{t-1} & \boldsymbol{F}_{imu}\boldsymbol{D}_{t-1}\boldsymbol{F}_{imu}^T\end{bmatrix} \tag{16}$$

where, $\boldsymbol{D}_{t-1}$ denotes the covariance of the previous propagated system states. Thus, the state vector after augmentation is defined as

$$\boldsymbol{X}_{vins} = \begin{bmatrix}\delta\boldsymbol{x}_{imu}^T, \delta\boldsymbol{x}_{1,imu}^T, \cdots, \delta\boldsymbol{x}_{t,imu}^T, \cdots, \delta\boldsymbol{x}_{N,imu}^T\end{bmatrix}^T \tag{17}$$

where N denotes the window size.

C. Measurement Update

The PO residual $\boldsymbol{e}_l^{po}$ is defined as

$$\boldsymbol{e}_l^{po} = \boldsymbol{K}\frac{d_i^{(i,j)}\boldsymbol{R}_i^l\boldsymbol{p}_i + \boldsymbol{t}_{l,i}}{\boldsymbol{e}_3^T\left(d_i^{(i,j)}\boldsymbol{R}_i^l\boldsymbol{p}_i + \boldsymbol{t}_{l,i}\right)} - \tilde{\boldsymbol{u}}_l \tag{18}$$

where $\boldsymbol{e}_3^T = (0,0,1)$, $\boldsymbol{K}$, $\tilde{\boldsymbol{u}}_l$ define the camera intrinsic parameters and monocular on image plane respectively. (18) can be further simplified considering the definition of $d_i^{(i,j)}$:

$$\boldsymbol{e}_l^{po} = \boldsymbol{K}\frac{\left\|\boldsymbol{p}_j^\wedge\boldsymbol{t}_{j,i}\right\|\boldsymbol{R}_i^l\boldsymbol{p}_i + \theta_{i,j}\boldsymbol{t}_{l,i}}{\boldsymbol{e}_3^T\left(\left\|\boldsymbol{p}_j^\wedge\boldsymbol{t}_{j,i}\right\|\boldsymbol{R}_i^l\boldsymbol{p}_i + \theta_{i,j}\boldsymbol{t}_{l,i}\right)} - \tilde{\boldsymbol{u}}_l \triangleq \boldsymbol{K}\frac{\boldsymbol{Y}_l}{\boldsymbol{e}_3^T\boldsymbol{Y}_l} - \tilde{\boldsymbol{u}}_l \tag{19}$$

Note that our state vector contains IMU poses instead of the camera ones, the transformation from c-frame to b-frame can be obtained by extrinsic parameters:

$$\boldsymbol{r}_b = \boldsymbol{r}_c - \boldsymbol{R}_b^e\boldsymbol{l}_c, \boldsymbol{R}_b^e = \boldsymbol{R}_c^e\boldsymbol{R}_b^c \tag{20}$$

Where $\boldsymbol{l}_c$ denotes the camera position with respect to b-frame, and $\boldsymbol{R}_b^c$ rotates from b-frame to c-frame. Jacobian matrix can be derived according to the chain rule (21).

Thereafter, these measurements can be used to EKF update with constrained by $\boldsymbol{D}_{t|t-1}$.

As aforementioned, (i,j) is a pair of base frames. Considering $\theta_{\eta,\xi}$ which indicates a quality indicator [36], $\theta_{\eta,\xi}$ $(1 \le \eta,\xi \le n, \eta \ne \xi)$ is calculated for all candidates, and the maximum yields our base frames.

$$\begin{aligned}\boldsymbol{J}_l^{(i)} &= \begin{pmatrix}\frac{\partial\boldsymbol{e}_l^{po}}{\partial\delta\boldsymbol{r}_i} & \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{\phi}_i}\end{pmatrix} = \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{Y}_l}\begin{pmatrix}\frac{\partial\boldsymbol{Y}_l}{\partial\delta\boldsymbol{r}_i} & \frac{\partial\boldsymbol{Y}_l}{\partial\boldsymbol{\phi}_i}\end{pmatrix},\\ \boldsymbol{J}_l^{(j)} &= \begin{pmatrix}\frac{\partial\boldsymbol{e}_l^{po}}{\partial\delta\boldsymbol{r}_j} & \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{\phi}_j}\end{pmatrix} = \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{Y}_l}\begin{pmatrix}\frac{\partial\boldsymbol{Y}_l}{\partial\delta\boldsymbol{r}_j} & \frac{\partial\boldsymbol{Y}_l}{\partial\boldsymbol{\phi}_j}\end{pmatrix},\\ \boldsymbol{J}_l^{(l)} &= \begin{pmatrix}\frac{\partial\boldsymbol{e}_l^{po}}{\partial\delta\boldsymbol{r}_l} & \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{\phi}_l}\end{pmatrix} = \frac{\partial\boldsymbol{e}_l^{po}}{\partial\boldsymbol{Y}_l}\begin{pmatrix}\frac{\partial\boldsymbol{Y}_l}{\partial\delta\boldsymbol{r}_l} & \frac{\partial\boldsymbol{Y}_l}{\partial\boldsymbol{\phi}_l}\end{pmatrix}\end{aligned} \tag{21}$$

IV. Tightly Integration of GNSS, Visual and Inertial Raw Measurements

In order to eliminate accumulated errors introduced by VINS, measurements from GNSS, including both pseudorange and phase observations, are sequentially incorporated and then double-differenced integer ambiguity is resolved using LAMBDA, enabling high-precision positioning.

Considering signals from satellite s, its pseudorange and phase observations can be modelled as

$$\begin{cases}L_r^s = \rho_r^s + cdt_r - cdt^s + I_r^s + T_r^s + n_r^s\\ \phi_r^s = \rho_r^s + cdt_r - cdt^s + \lambda\cdot N_r^s - I_r^s + T_r^s + \varepsilon_r^s\end{cases} \tag{22}$$

Where, L_r^s, ϕ_r^s are pseudorange and phase measurements respectively. dt_r, dt^s denote clock error of receiver and satellite. I_r^s, T_r^s denote ionospheric and tropospheric delay. λ denotes the wavelength and N_r^i denotes the carrier phase ambiguity. e_r^s, ε_r^s denote white Gaussian noise of pseudorange and phase observations. ρ_r^s defines the distance between receiver and satellite, derived by (23), where $\boldsymbol{l}_a$ denotes the vector pointing from IMU center to GNSS antenna and $\hat{\boldsymbol{r}}_b$, $\hat{\boldsymbol{R}}_b$ denotes current estimates of IMU pose derived by VINS.

$$\rho_r^s = \left\|\hat{\boldsymbol{r}}_b - \boldsymbol{r}_s + \hat{\boldsymbol{R}}_b\boldsymbol{l}_a\right\|_2 \tag{23}$$

Regarding a stationary station u observing the same satellite s, the single-differenced observations can be derived by (24), which eliminates satellite clock error.

$$\begin{cases}\Delta L_{ru}^s = L_r^s - L_u^s\\ \quad = \rho_{ru}^s + dt_{ru} + I_{ru}^s + T_{ru}^s + n_{ru}^s\\ \Delta\phi_{ru}^s = \phi_{ru}^s - \phi_{ru}^s\\ \quad = \rho_{ru}^s + dt_{ru} + \lambda_f\cdot N_{ru}^s - I_{ru}^s + T_{ru}^s + \varepsilon_{ru}^s\end{cases} \tag{24}$$

Select a reference satellite k and then the double-differenced (DD) observations can be further derived by (25), which is also known as RTK formulation (short baseline scenario).

$$\begin{cases} \nabla\Delta L_{ru}^{sk} = \Delta L_{ru}^{s} - \Delta L_{ru}^{k} = \rho_{ru}^{sk} + n_{ru,f}^{sk} \\ \nabla\Delta \phi_{ru}^{sk} = \Delta \phi_{ru}^{s} - \Delta \phi_{ru}^{k} = \rho_{ru}^{sk} + \lambda \cdot N_{ru}^{sk} + \varepsilon_{ru,f}^{sk} \end{cases} \tag{25}$$

In addition to VINS state $\boldsymbol{X}_{vins}$, DD ambiguities are directly appended to the state vector and the priori can be derived by

$$\begin{aligned} \boldsymbol{X}_{gvins} &= \left[\boldsymbol{X}_{vins}^{T}, \boldsymbol{N}_{ru}\right]^{T} \\ \boldsymbol{D}_{gvins} &= \begin{bmatrix} \boldsymbol{D}_{t} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{amb} \end{bmatrix} \end{aligned} \tag{26}$$

where $\boldsymbol{N}_{ru}$ defines a vector containing all DD ambiguities at current timestamp, and $\boldsymbol{D}_{amb}$ denotes the prior covariance of DD ambiguities, which can be propagated from undifferenced ambiguities. $\boldsymbol{D}_t$ denotes the uncertainty of VINS related states.

Assume we have measurement set $\mathcal{Z}_t = \{\boldsymbol{z}_t^v, \boldsymbol{z}_t^g\}$ at timestamp t, where $\boldsymbol{z}_t^v$ and $\boldsymbol{z}_t^g$ denote visual and GNSS measurements respectively. Then $\boldsymbol{X}_{gvins}$ can be estimated by EKF.

$$\delta\hat{\boldsymbol{X}}_{gvins} = \delta\boldsymbol{x}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{l}_t - \boldsymbol{h}_t\delta\boldsymbol{x}_{t|t-1}\right) \tag{27}$$

Considering visual and GNSS measurements are dependent from each other, a sequential update strategy is utilized to separately update measurements of each sensor. We first update states with visual measurements. After obtaining the posteriori, GNSS measurements are then updated. It is important to note that, this update strategy is equivalent to (27), but makes program more flexible to handle hardware delay. For more details and meaning of each symbol, readers can refer to [49].

Specifically, in the case that the fusion system has received visual measurements and GNSS measurements have *t ms* delay. The fusion system will first carry out state augmentation, and then do visual measurement update. When GNSS measurements come, we will propagate system state to this timestamp by interpolating IMU measurement. Thus, GNSS measurements can be updated without creating new poses in sliding window.

Then, GNSS residual and the Jacobi can be easily derived from (25), and then an iterative EKF update is processed with IGG-III outlier removal strategy applied. After update, DD ambiguities are resolved through LAMBDA method [50]. Besides, in our implementation, a robust cascaded alignment is first conducted to initialize GNSS/IMU system [51], since GNSS provides absolute and high-precision positioning ability. Then, initial IMU pose, $\boldsymbol{r}_b$ and $\boldsymbol{R}_b$, can be determined. With known extrinsic parameters, camera poses can also be recovered with metric scale.

V. Experimental Results

For comparison, we implement Pose-Only VINS (PO-VINS), MSCKF, GNSS/IMU (GI), MSCKF (M)-GVINS, and the proposed PO-GVINS with our own platform. Besides, two state-of-the-art open-source libraries, VINS-Fusion and GVINS, are compared. In the subsequent analysis, the term VINS denotes the VINS-Fusion system operating without GNSS input. Given that the VINS module within GVINS closely resembles that of VINS-Fusion in both structure and methodology, a separate analysis is deemed unnecessary. EVO toolkit is utilized to evaluate [52].

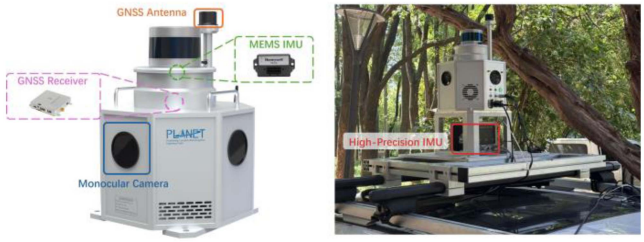

Fig. 1. Experimental hardware (left) and collection vehicle (right).

TABLE I
Sensor Specifications for the Device Used in Vehicle-Borne Experiments

Sensor Type & Item	Specification
High-precision IMU	**Novatel SPAN-ISA-100C**
Gyroscope bias	0.05 $^\circ/hr$
Gyroscope random walk	0.005 $^\circ/\sqrt{hr}$
Accelerometer bias	0.1 mg
Measurement frequency	200 Hz
MEMS-IMU	**HGuide I300**
Gyroscope bias	65 $^\circ/hr$
Gyroscope random walk	0.15 $^\circ/\sqrt{hr}$
Accelerometer bias	1.0 mg
Measurement frequency	200 Hz
Camera	**MER-131**
Shutter	Global shutter
Resolution	1280×1024 pixel
Measurement frequency	10 Hz
GNSS receiver	**Septentrio Mosaic-X5**
Antenna	Harxon HX-CH7609A
Measurement frequency	1 Hz

A. Dataset Description

As shown in Fig. 1, we evaluate the proposed PO-GVINS using data collected by a self-developed hardware, which includes a GNSS receiver together with its antenna, four monocular cameras (front-view camera is used in this research), and a MESE IMU. Besides, a high-precision IMU (ISA-100C) is utilized to obtain reference trajectory. All these sensors are hardware synchronized and extrinsic parameters relative to MEMS IMU are well calibrated. The base station, required by RTK algorithm, is equipped with a Trimble Alloy receiver, and its antenna is set at an open-sky view. The distance between rover and base station is no greater than 10km to meet the requirement of short baseline. In our test, reference trajectory is solved through the smoothed and combined solutions of multi-GNSS RTK/IMU(ISA-100C), using commercial software Inertial Explorer (IE) 8.9. MESE-IMU is used to evaluate our proposed PO-GVINS. The specifications of each important sensor are listed in Table I.

Four sequences are collected in Wuhan City, China, covering open-sky area, tree-line road, viaduct, illumination change and

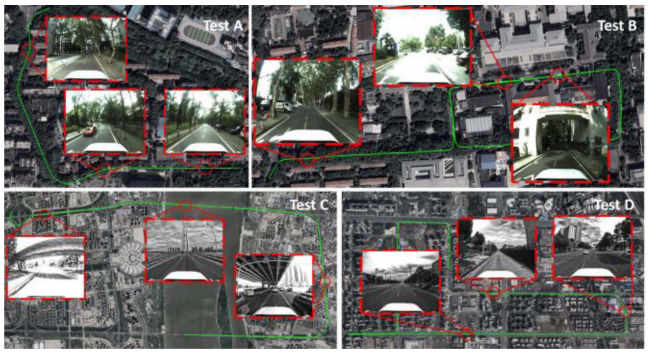

Fig. 2. Scenarios of collected sequences. Test A and B contain tree-lined road and other GNSS-denied scenarios with abundant textures. Test C contains viaducts, illumination change, and dynamic objects. Test D contains only open-sky areas. Visual textures in Both C and D are far from cameras.

TABLE II
TRAJECTORY LENGTH AND TIME DURATION OF EACH SEQUENCE

Sequence	Traj. Length (km)	Duration (min)
A	0.75	1.50
B	1.49	3.85
C	12.08	16.67
D	8.79	19.02

TABLE III
STATISTICS IN SEQUENCE A AND B

Methods	ATE (%)[1]		ARE (deg / m)[1]		RPE	
Sequence	A	B	A	B	A	B
MSCKF	5.35	2.25	**0.18%**	0.09%	0.74	0.16
VINS	37.0	15.6	1.63%	7.56%	1.45	0.91
PO-VINS	**3.12**	**1.58**	0.19%	**0.09%**	**0.27**	**0.16**
Methods	**ATE (m)**		**ARE (deg)**	
Sequence	**A**	**B**	**A**	**B**
GI	0.44	0.86	1.69	1.53
GVINS	20.62	-[2]	-	90.46
VINS-Fusion	44.37	26.53	-	-
M-GVINS	0.37	1.08	1.68	1.50
PO-GVINS	**0.35**	**0.64**	**1.58**	**1.42**

1 divided by trajectory length.
2 - denotes the value exceeds 100 or failure occurred.

other visual and GNSS complex environments, as shown in Fig. 2 and Table II.

B. Positioning Accuracy

First, we evaluate the proposed PO-VINS and PO-GVINS in Sequence A and B. The root mean square (RMS) of absolute translation error (ATE), absolute rotation error (ARE) and relative pose error (RPE) of each configuration are shown in Table III.

It is shown that PO-VINS outperforms MSCKF as well as VINS. Compared with MSCKF, the PO representation shows obvious improvements, especially on translation. Since we use the same feature selection methods, feature point linearization

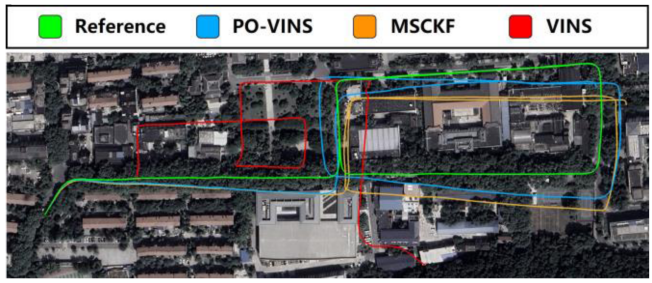

Fig. 3. Estimated trajectories of PO-VINS, MSCKF and VINS in Seq. B. Each trajectory starts at different positions due to different initialization strategies.

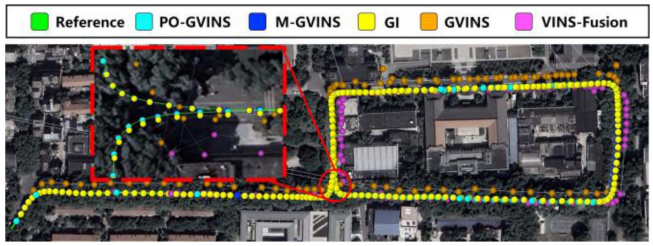

Fig. 4. Estimated trajectories of configurations with GNSS input in Seq. B.

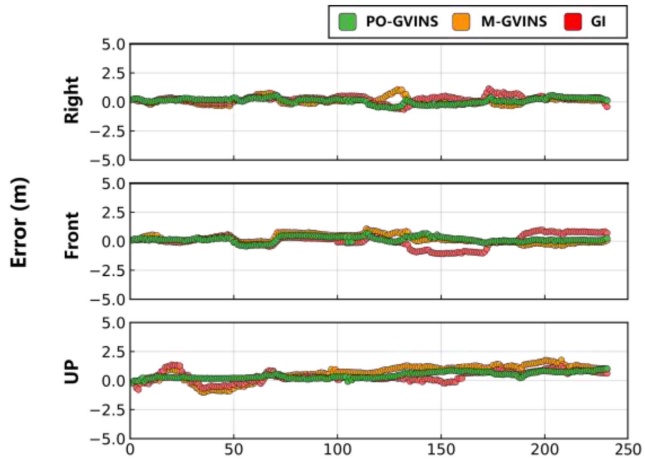

Fig. 5. PO-GVINS, M-GVINS, and GI Error time series of test B.

error introduced by MSCKF leads to significant increasing errors, achieving an average reduction on ATE and RPE by 35.7% and 31.8% respectively. As illustrated in Fig. 3, VINS initializes wrong scale at the beginning and thus the proposed PO-VINS shows better accuracy even compared with the optimization-based VINS.

Furthermore, Fig. 4 shows trajectories estimated with GNSS input. Besides, although GI can achieve drift-free estimation, it is vulnerable in GNSS challenging scenarios. In Seq. B, VINS-Fusion, integrating GNSS in loosely coupled manner, is severely affected by the wrongly estimated GNSS position. And thus, there are lots of failures and rotational angles are unable to converge. The GVINS, which tightly fuses pseudorange measurements, still suffers from outliers, leading to numerous wrong estimates. While, the PO-GVINS tightly

TABLE IV
STATISTICS IN SEQUENCE C AND D

Methods	ATE (m)		ARE (deg)	
Sequence	C	D	C	D
GI	**2.84**	0.03	1.44	1.53
GVINS	16.81	10.68	99.67	-
VINS-Fusion	72.47	13.30	-	-
M-GVINS	3.54	0.03	1.55	1.53
PO-GVINS	2.92	**0.03**	**1.44**	**1.53**

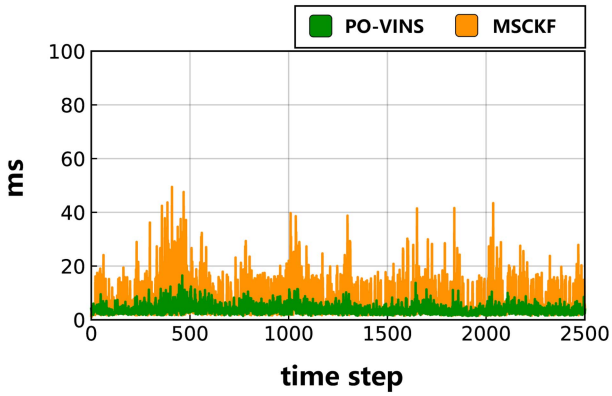

Fig. 6. Measurement update time cost of PO-VINS (green) and MSCKF (yellow), tested with Intel(R) Core(TM) i7-10510U CPU @ 1.80 GHz (8 CPUs) ~2.3 GHz in debug mode.

utilizes both pseudorange and carrier phase measurements and further employ ambiguity resolution methods, showing the best performance, as shown in Fig. 5. Specifically, compared with GI and M-GVINS, PO-GVINS achieves 23.3%, 23.1% reduction on ATE, and 8.8%, 5.2% reduction on ARE respectively.

Two long term experiments are analyzed further, as shown in Table IV. Even in open-sky scenarios (Seq. D), VINS-Fusion still fails, since it direct outputs the fusing results and the fusion estimates are ignored in VINS submodule. Besides, in this open-sky area, visual features are far from cameras, making it difficult for VINS to estimate. The positioning accuracy using only pseudorange measurements is typically several meters, depending on the scenario; therefore, it is reasonable that GVINS achieves approximately 15 meters in ATE. Due to the presence of distant features, the proposed PO-GVINS assigns greater weights to GNSS raw measurements and shows accuracy comparable to that of GI.

C. Time Complexity

Since Pose-only modelling eliminates parameters of feature depth, it can avoid dimensional explosion and ensure real-time applications. As illustrated in Fig. 6, PO-VINS enhances algorithmic efficiency by eliminating feature points before linearization. In comparison to MSCKF, it exhibits more stable computation time during the measurement update step.

VI. CONCLUSION

In this letter, a filtering-based GNSS, IMU, and monocular camera tightly coupled framework is proposed. Specifically, the pose-only formulation is utilized in our visual measurements processing, which avoids feature linearization errors compared with existing frameworks. Extensive experiments demonstrate the PO-VINS formulation outperforms MSCKF, and the PO-GVINS achieves accurate and drift-free estimation.

However, several issues remain to be addressed. In this work, base frames are selected based solely on maximum parallax. More informative selection strategies, such as those proposed in [53], should be considered. Additionally, mapping is not incorporated in this study. Given the significantly reduced computational cost and accurate pose estimation, accurate online maps could potentially be constructed. Furthermore, to address hardware delay, continuous-time state estimation methods may be explored.

REFERENCES

[1] 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.

[2] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. 2007 IEEE Int. Conf. Robot. Automat.*, Rome, Italy, Apr. 2007, pp. 3565–3572, doi: 10.1109/ROBOT.2007.364024.

[3] A. I. Mourikis, N. Trawny, S. I. Roumeliotis, A. E. Johnson, A. Ansar, and L. Matthies, "Vision-aided inertial navigation for spacecraft entry, descent, and landing," *IEEE Trans. Robot.*, vol. 25, no. 2, pp. 264–280, Apr. 2009, doi: 10.1109/TRO.2009.2012342.

[4] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, Sep. 2010, doi: 10.1002/rob.20360.

[5] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *Proc. 2018 IEEE Int. Conf. Robot. Automat.*, Brisbane, QLD, Australia, May 2018, pp. 2502–2509, doi: 10.1109/ICRA.2018.8460664.

[6] X. Wang, X. Li, H. Chang, S. Li, Z. Shen, and Y. Zhou, "GIVE: A tightly coupled RTK-inertial–visual state estimator for robust and precise positioning," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 1005615, doi: 10.1109/TIM.2023.3282296.

[7] Y. Ge, L. Zhang, Y. Wu, and D. Hu, "PIPO-SLAM: Lightweight visual-inertial SLAM with preintegration merging theory and pose-only descriptions of multiple view geometry," *IEEE Trans. Robot.*, vol. 40, pp. 2046–2059, 2024, doi: 10.1109/TRO.2024.3366815.

[8] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 932–945, Oct. 2008, doi: 10.1109/TRO.2008.2003276.

[9] T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems," in *Proc. 2018 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Madrid, Spain, Oct. 2018, pp. 3662–3669, doi: 10.1109/IROS.2018.8593603.

[10] S. Cao, X. Lu, and S. Shen, "GVINS: Tightly coupled GNSS-visual-inertial fusion for smooth and consistent state estimation," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2004–2021, Aug. 2022, doi: 10.1109/TRO.2021.3133730.

[11] G. Huang, "Visual-inertial navigation: A concise review," in *Proc. 2019 Int. Conf. Robot. Automat.*, Montreal, QC, Canada, May 2019, pp. 9572–9582, doi: 10.1109/ICRA.2019.8793604.

[12] M. Li and A. I. Mourikis, "Improving the accuracy of EKF-based visual-inertial odometry," in *Proc. 2012 IEEE Int. Conf. Robot. Automat.*, St Paul, MN, USA, May 2012, pp. 828–835, doi: 10.1109/ICRA.2012.6225229.

[13] K. Sun et al., "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 965–972, Apr. 2018, doi: 10.1109/LRA.2018.2793349.

[14] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "Observability-based rules for designing consistent EKF SLAM estimators," *Int. J. Robot. Res.*, vol. 29, no. 5, pp. 502–528, Apr. 2010, doi: 10.1177/0278364909353640.

[15] G. Bleser and D. Strickery, "Using the marginalised particle filter for real-time visual-inertial sensor fusion," in *Proc. 7th IEEE/ACM Int. Symp. Mixed Augmented Reality*, Cambridge, U.K., Sep. 2008, pp. 3–12, doi: 10.1109/ISMAR.2008.4637316.

[16] J. Georgy, A. Noureldin, and C. Goodall, "Vehicle navigator using a mixture particle filter for inertial sensors/odometer/map data/GPS integration," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 544–552, May 2012, doi: 10.1109/TCE.2012.6227459.

[17] H. Zhou, Z. Yao, C. Fan, S. Wang, and M. Lu, "Rao-blackwellised particle filtering for low-cost encoder/INS/GNSS integrated vehicle navigation with wheel slipping," *IET Radar, Sonar Navig.*, vol. 13, no. 11, pp. 1890–1898, Nov. 2019, doi: 10.1049/iet-rsn.2019.0108.

[18] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "A quadratic-complexity observability-constrained unscented Kalman filter for SLAM," *IEEE Trans. Robot.*, vol. 29, no. 5, pp. 1226–1243, Oct. 2013, doi: 10.1109/TRO.2013.2267991.

[19] T. Cantelobre, C. Chahbazian, A. Croux, and S. Bonnabel, "A real-time unscented Kalman filter on manifolds for challenging AUV navigation," in *Proc. 2020 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Las Vegas, NV, USA, Oct. 2020, pp. 2309–2316, doi: 10.1109/IROS45743.2020.9341216.

[20] M. Brossard, A. Barrau, and S. Bonnabel, "Exploiting symmetries to design EKFs with consistency properties for navigation and SLAM," *IEEE Sensors J.*, vol. 19, no. 4, pp. 1572–1579, Feb. 2019, doi: 10.1109/JSEN.2018.2882714.

[21] C. Liu, C. Jiang, and H. Wang, "InGVIO: A consistent invariant filter for fast and high-accuracy GNSS-visual-inertial odometry," *IEEE Robot. Automat. Lett.*, vol. 8, no. 3, pp. 1850–1857, Mar. 2023, doi: 10.1109/LRA.2023.3243520.

[22] Y. Fan, T. Zhao, and G. Wang, "SchurVINS: Schur complement-based lightweight visual inertial navigation system," in *Proc. 2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 17964–17973, doi: 10.1109/CVPR52733.2024.01701.

[23] Y. Yang, J. Maley, and G. Huang, "Null-space-based marginalization: Analysis and algorithm," in *Proc. 2017 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vancouver, BC, Canada, Sep. 2017, pp. 6749–6755, doi: 10.1109/IROS.2017.8206592.

[24] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," in *Proc. Robot.: Sci. Syst. (RSS)*, Rome, Italy, Jul. 2015, doi: 10.15607/RSS.2015.XI.006.

[25] 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, Mar. 2015, doi: 10.1177/0278364914554813.

[26] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[27] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.

[28] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, Apr. 2017, doi: 10.1109/TRO.2016.2623335.

[29] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *Proc. 2016 IEEE Int. Conf. Robot. Automat.*, May 2016, pp. 1885–1892, doi: 10.1109/ICRA.2016.7487335.

[30] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *Proc. 2018 IEEE Int. Conf. Robot. Automat.*, May 2018, pp. 2510–2517, doi: 10.1109/ICRA.2018.8462905.

[31] L. von Stumberg and D. Cremers, "DM-VIO: Delayed marginalization visual-inertial odometry," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1408–1415, Apr. 2022, doi: 10.1109/LRA.2021.3140129.

[32] K. Eckenhoff, L. Paull, and G. Huang, "Decoupled, consistent node removal and edge sparsification for graph-based SLAM," in *Proc. 2016 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Daejeon, South Korea, Oct. 2016, pp. 3275–3282, doi: 10.1109/IROS.2016.7759505.

[33] J. Hsiung, M. Hsiao, E. Westman, R. Valencia, and M. Kaess, "Information sparsification in visual-inertial odometry," in *Proc. 2018 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Madrid, Spain, Oct. 2018, pp. 1146–1153, doi: 10.1109/IROS.2018.8594007.

[34] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering," in *Proc. 2011 IEEE Int. Conf. Robot. Automat.*, Shanghai, China, May 2011, pp. 3281–3288, doi: 10.1109/ICRA.2011.5979641.

[35] Q. Cai, L. Zhang, Y. Wu, W. Yu, and D. Hu, "A pose-only solution to visual reconstruction and navigation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 73–86, Jan. 2023, doi: 10.1109/TPAMI.2021.3139681.

[36] Q. Cai, Y. Wu, L. Zhang, and P. Zhang, "Equivalent constraints for two-view geometry: Pose solution/pure rotation identification and 3D reconstruction," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 163–180, Feb. 2019, doi: 10.1007/s11263-018-1136-9.

[37] X. Niu, H. Tang, T. Zhang, J. Fan, and J. Liu, "IC-GVINS: A robust, real-time, INS-centric GNSS-visual-inertial navigation system," *IEEE Robot. Automat. Lett.*, vol. 8, no. 1, pp. 216–223, Jan. 2023, doi: 10.1109/LRA.2022.3224367.

[38] J. Hu, F. Zhu, D. Zhuo, Q. Xu, W. Liu, and X. Zhang, "Performance evaluation of stereo vision aided loosely coupled GNSS/SINS integration for land vehicle navigation in different urban environments," *IEEE Sensors J.*, vol. 23, no. 4, pp. 4129–4142, Feb. 2023, doi: 10.1109/JSEN.2023.3234216.

[39] T. Chu, N. Guo, S. Backén, and D. Akos, "Monocular camera/IMU/GNSS integration for ground vehicle navigation in challenging GNSS environments," *Sensors*, vol. 12, no. 3, pp. 3162–3185, Mar. 2012, doi: 10.3390/s120303162.

[40] G. Falco, M. Pini, and G. Marucco, "Loose and tight GNSS/INS integrations: Comparison of performance assessed in real urban scenarios," *Sensors*, vol. 17, no. 2, Jan. 2017, Art. no. 255, doi: 10.3390/s17020255.

[41] S. Li, X. Li, H. Wang, Y. Zhou, and Z. Shen, "Multi-GNSS PPP/INS/vision/LiDAR tightly integrated system for precise navigation in urban environments," *Inf. Fusion*, vol. 90, pp. 218–232, Feb. 2023, doi: 10.1016/j.inffus.2022.09.018.

[42] T. Li, L. Pei, Y. Xiang, W. Yu, and T.-K. Truong, "P^3-VINS: Tightly-coupled PPP/INS/visual SLAM based on optimization approach," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 7021–7027, Jul. 2022, doi: 10.1109/LRA.2022.3180441.

[43] B. Xu, S. Zhang, K. Kuang, and X. Li, "A unified cycle-slip, multipath estimation, detection and mitigation method for VIO-aided PPP in urban environments," *GPS Solutions*, vol. 27, no. 2, Apr. 2023, Art. no. 59, doi: 10.1007/s10291-023-01396-7.

[44] X. Li et al., "Continuous and precise positioning in urban environments by tightly coupled integration of GNSS, INS and vision," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 11458–11465, Oct. 2022, doi: 10.1109/LRA.2022.3201694.

[45] E. Dong et al., "Robust, high-precision GNSS carrier-phase positioning with visual-inertial fusion," Mar. 2023, *arXiv:2303.01291*.

[46] F. Wang and J. Geng, "GNSS PPP-RTK tightly coupled with low-cost visual-inertial odometry aiming at urban canyons," *J. Geodesy*, vol. 97, no. 7, Jul. 2023, Art. no. 66, doi: 10.1007/s00190-023-01749-7.

[47] X. Li et al., "Centimeter-accurate vehicle navigation in urban environments with a tightly integrated PPP-RTK/MEMS/vision system," *GPS Solutions*, vol. 26, no. 4, Oct. 2022, Art. no. 124, doi: 10.1007/s10291-022-01306-3.

[48] S. I. Roumeliotis and J. W. Burdick, "Stochastic cloning: A generalized framework for processing relative state measurements," in *Proc. 2002 IEEE Int. Conf. Robot. Automat.*, WA, DC, USA, 2002, vol. 2, pp. 1788–1795, doi: 10.1109/ROBOT.2002.1014801.

[49] F. Zhu, Z. Xu, X. Zhang, Y. Zhang, W. Chen, and X. Zhang, "On state estimation in multi-sensor fusion navigation: Optimization and filtering," 2024, *arXiv:2401.05836*.

[50] P. J. G. Teunissen, P. J. de Jonge, and C. C. J. M. Tiberius, "The least-squares ambiguity decorrelation adjustment: Its performance on short GPS baselines and short observation spans," *J. Geodesy*, vol. 71, no. 10, pp. 589–602, Sep. 1997, doi: 10.1007/s001900050127.

[51] W. Liu, R. Duan, and F. Zhu, "A robust cascaded strategy of in-motion alignment for inertial navigation systems," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 9, Sep. 2017, Article ID 1550147717732919, doi: 10.1177/1550147717732919.

[52] M. Grupp, "Evo: Python package for the evaluation of odometry and SLAM," 2017. [Online]. Available: https://github.com/MichaelGrupp/evo

[53] L. Wang, H. Tang, T. Zhang, Y. Wang, Q. Zhang, and X. Niu, "PO-KF: A pose-only representation-based Kalman filter for visual inertial odometry," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 14856–14875, May 2025, doi: 10.1109/JIOT.2025.3526811.