VINS-Vehicle: A Tightly-Coupled Vehicle Dynamics Extention to Visual-Inertial State Estimator

Mingyu Xu, Rong Kang, Yuchen Kang, Peizhi Zhang Junqiao Zhao*, Lu Xiong, Zhuoping Yu

***Abstract*— In this paper, we propose VINS-Vehicle, a novel tightly-coupled vehicle dynamics extention to visual-inertial navigation system (VINS) framework. Degenerate motions, such as uniform linear motions or uniform cicular motions, which are most common for a ground vehicle, are not observable for a monocular VINS. Therefore, VINS can not be applied to vehicles, due to difficulties in initialization and low accuracy. To address this limitation, we extend VINS to tightly coupled with pre-integrated high-frequency motion information based on a two degree-of-freedom (DOF) vehicle dynamics model. By loosely aligning the structure from motion (SfM) results, pre-integrated IMU measurements and motion information, the system can be robustly initialized. A tightly-coupled, sliding window optimization method is proposed to obtain a accurate visual-inertial-dynamics odometry result. The experiments show that the system achieved significantly higher positioning accuracy compared with existing VINS methods. Moreover, the proposed method is robust in a texture-less underground parking lot and dynamic outdoor environments.**

I. Introduction

Visual odometry and simultaneously localization and mapping (SLAM) have drawn intensive attentions in the studies of autonomous driving because of their low-cost, small size and easy hardware layout [1]–[6]. However, the metric scale of the motion can not be recovered from a single camera, and low-level feature-based methods suffer from a lack of robustness when encountering texture-less areas or dynamic environments. The inertial measurement unit (IMU) was integrated to estimate scale and provide short-term motion constraints. Therefore, visual-inertial navigation system (VINS), where visual observations and IMU measurements are tightly coupled to estimate the 6-degree-of-freedom (DOF) poses of a platform navigating, has shown to achieve a high-accuracy and robust localization result even during fast motion or under strong illumination change [7]–[12].

However, if we deploy VINS directly on a platform that often performs degenerate motions, i.e., approximately planar and along arcs or straight lines at constant speed or acceleration, such as an autonomous vehicles or a wheeled robot, the system will have difficulties in initialization and a low localization accuracy.

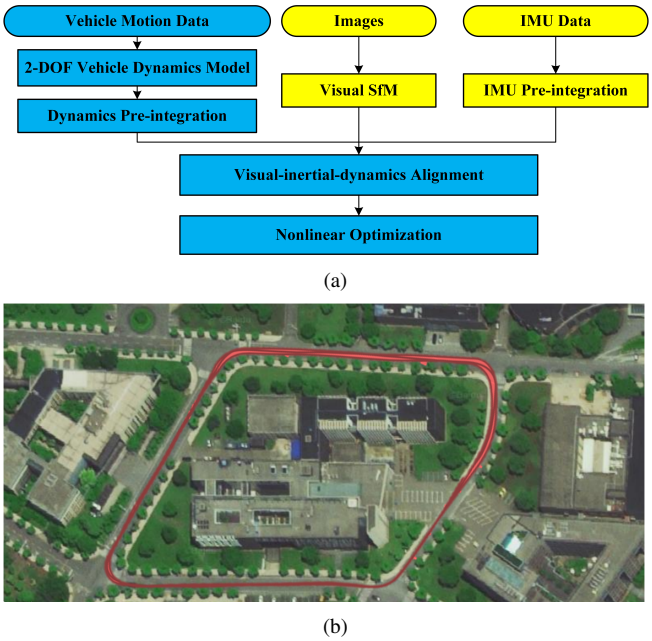


Fig. 1. (a) System overview of VINS-Vehicle. The blue blocks are our proposed extention to visual-inertial state estimator. (b) Results of outdoor experiments(red line) overlaid with Google map for visual comparison. Total trajectory length is 1.6 km(three loops).

Firstly give the reason for initialization difficulties and then for the low accuracy Firstly, due to the nonlinearity of the vision-inertial system and the lack of direct distance measurement, the initialization of the VINS system has always been a huge challenge. One the one hand, the monocular camera cannot estimate the metric scale and the vision algorithm is fragile during fast motion or in dynamic environment. One the other hand, during time-limited initilization period, the bias of gyroscope and accelerometer estimation are often difficult to converge due to the poor excitation of IMU. For the reasons given above, the initialization methods [11]–[15] that loosely couple inertial measurements with visual observations often fail. In this case, the initial attitude of the whole system, i.e., the velocity, gravity vector and bias of the gyroscope and accelerometer cannot to be fastly and accurately estimated, which will even lead to incorrect estimates. Besides, the ground vehicle's suspension system provides a stable platform on which the measurement of the IMU are restricted to only 3-DOF. And the metric scale becomes unobservable when a vehicle moves with a constant acceleration [16]. Nevertheless, the vehicle motion

This work is supported by the National Key Research and Development Program of China (No. 2018YFB0105103, No. 2017YFA0603104), the National Natural Science Foundation of China (No. U1764261, No. 41801335, No. 41871370), the Natural Science Foundation of Shanghai (No. kz170020173571, No. 16DZ1100701) and the Fundamental Research Funds for the Central Universities (No. 22120180095).

M. Xu, R. Kang, Y. Kang, P. Zhang, L. Xiong and Z. Yu are with the School of Automotive Studies, Tongji University, Shanghai.

J. Zhao are with the Department of Computer Science and Technology, School of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China zhaojunqiao@tongji.edu.cn

information from the wheel odometer [17] can make the metric scale observable.

In this paper, we propose a novel tightly-coupled vehicle dynamics extension to VINS, the VINS-Vehicle. With the high-frequency vehicle motion information obtained from the vehicle control unit (VCU) via the CAN bus, we employ vehicle dynamics model to estimate linear and angular velocity vector of the vehicle and pre-integrate the results. The sliding window optimization is adopted to estimate the initial state of the whole system and process tightly-coupled visual, IMU and vehicle dynamics residuals. complement the method in one sentence The coupled motion information helps estimate the accurate metric scale and the bias of accelerometers and gyroscopes. verify Furthermore, the introduction of motion information makes the system more robust for autonomous driving applications.

In summary, the main contributions of this paper are:

- A novel monocular visual-inertial-dynamics tightly-coupled framework for ground vehicles.
- A robust visual-inertial-dynamics initialization procedure which is able to bootstrap the sliding windows based optimization from unknown initial states.

II. Related Work

Research on monocular visual odometry/SLAM has made great progress in recent years. The representative works include PTAM [2], SVO [3], LSD-SLAM [4], ORB-SLAM [5] and DSO [6], which are based on indirect or direct method. Compared with the EKF-based method in earlier research, the optimization-based method has more advantages in accuracy. Common problems faced by all these pure monocular methods are the scale ambiguity of system and the poor robustness of algorithm.

To solve the problem of monocular system, motion estimation methods using camera and IMU have been proposed., The loosely coupled system [18], [19] is not robust and precise enough because the IMU information is not used in the visual odometry part. For this problem, tightly-coupled fusion approach that jointly optimizes raw camera and IMU measurements have been proposed. The representative tightly-coupled approaches include MSCKF [7], OKVIS [9] and VINS-Mono [12], VI-DSO [20]. The EKF-based MSCKF maintains several previous camera poses in the state vector, and uses visual constraints of the same feature across multiple camera views to update the state vector. VI-DSO is based on direct approach, which minimize photometric error while processing visual measurement. Indirect approaches are more frequently applied in engineering deployment due to its maturity and robustness. As the current typical tightly-coupled visual-inertial system, OKVIS and VINS-Mono are both based on indirect approach, which optimize over a bounded-size sliding window of recent states by marginalizing out past states and measurements. Compared with OKVIS, VINS-Mono costs less CPU resource and has better initialization module.

The initialization process is important to visual-inertial system. According to [21], through analyzing the relationship between SfM and inertial integration, the gravity vector and scale factor can be computed, which provided initial metric values for the state estimation filter. [11] presents an IMU initialization algorithm which based on the monocular ORB-SLAM, where An initial estimation of the scale, gravity direction, velocity and IMU biases are computed for the visual-inertial full BA. Similar with [11], initialization method [13] ignores acceleration bias in the initial step to ensure fast initialization because acceleration bias coupled with gravity usually lacks observability.

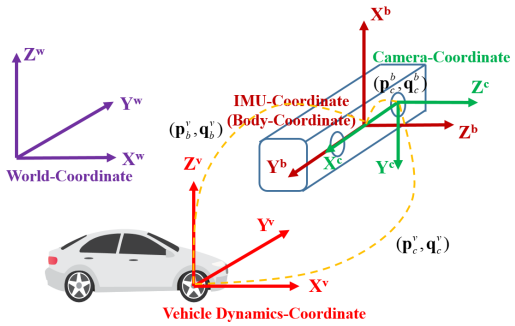


Fig. 2. The coordinates and extrinsic parameters between them.

Some studies have also considered the impact of sensor platforms on state estimation. Several works have been specifically designed for vehicles with motion constraints. The advantage is decreased computation time and improved accuracy. Some earlier works [22]–[24] focus on plane constrains of vehicle motion. [25], [26] introduced a one-point RANSAC outlier rejection based on the vehicle non-holonomic constraints to speed up egomotion estimation to 400 Hz. The Ackermann vehicle model is used in [27] to recover the scale of monocular visual odometry. However, this simple vehicle model did not consider the influence of the centroid side deflection angle. Similar to our approach, [17] incorporated wheel-encoder measurements into VINS and introduced mVINS that properly models the ground robot's almost-planar motion, but it didn't use vehicle dynamics model and tightly couple the vehicle motion information with other measurements.

III. VINS-Vehicle Framework

Throughout the paper we will use the following notation:We denote $(\cdot)^w$ as the world frame, where the gravity vector is aligned with z-axis. $(\cdot)^b$ represents the body frame, which is the same as the IMU frame. $(\cdot)^c$ is the camera frame. The vehicle dynamics model can be seen as an additional sensor. $(\cdot)^v$ is expressed as the vehicle dynamics frame, where the origin is at the center of the front axle of the vehicle. The defination of all the coordinates are shown in Fig. 2. We treat $(\cdot)_v^w$ as vehicle dynamics frame with respect to world frame. The noise measurement of a certain quantity is denoted by $\hat{(\cdot)}$. Rotation matrices $\mathbf{R}$ and Hamilton quaternions $\mathbf{q}$ are both used to represent rotation. $\mathbf{q}_v^w$ and $\mathbf{p}_v^w$ denote rotation and translation from the vehicle dynamics frame to the world frame. b_k is the kth body frame while

taking the kth image. c_k is the kth camera frame while taking the kth image. v_k is the kth vehicle dynamics frame while taking the kth image. $\otimes$ is the multiplication operation between two quaternions. At last, $\mathbf{g}^{c_0}$ represents the gravity vector in the first camera frame(visual reference frame) and $\mathbf{g}^w = [0, 0, g]^T$ is the gravity vector in the world frame.

A. 2-DOF Vehicle Dynamics Model

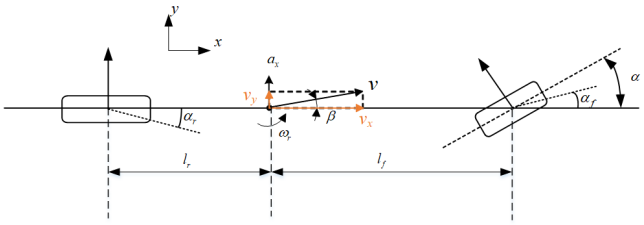


Fig. 3. 2-DOF vehicle dynamics model, α is the front wheel angle, α_f and α_r are front and rear wheel side angle. v is velocity of the center of mass, v_x and v_y are longitudinal and lateral components.l_f and l_r indicate respectively the distance from the center of mass to the front and rear axes.β is centroid side deflection angle and ω_r represents yaw velocity. The x axis is along the longitudinal direction and the y axis is along the lateral direction.

There are numerous degrees of freedom associated with vehicle dynamics. In this paper we choose the most basic vehicle dynamics model, a two-degree-of-freedom bicycle model [28], representing the lateral and yaw motions. The model is shown in Fig. 3.

The basic equation of motion for this model can be derived as follows:

$$\begin{aligned}(k_f + k_r)\beta + \frac{1}{v_x}(l_f k_f - l_r k_r)\omega_r - k_f\alpha &= m(\dot{v}_y + v_x\omega_r)\\ (l_f k_f - l_r k_r)\beta + \frac{1}{v_x}(l_f{}^2 k_f + l_r{}^2 k_r)\omega_r - l_f k_f\alpha &= I_z\dot{\omega}_r\end{aligned} \tag{1}$$

where I_z is moment of inertia around the z-axis. The meaning of the other letters in the formula has been illustrated in Fig. 3.

Consider the steady-state response of the vehicle, we substitue $\dot{v}_y = 0$ and $\dot{\omega}_r = 0$ to the equation (1), then centroid side deflection angle β and yaw velocity ω_r can be calculated through:

$$\begin{aligned}\beta_t &= \frac{1 + \frac{m}{2l}\frac{l_f}{l_r k_r}\hat{v}_t^2}{1 - K\hat{v}_t^2}\frac{l_r}{l}\alpha_t\\ \hat{\omega}_t &= \frac{1}{1 - K\hat{v}_t^2}\frac{\hat{v}_t}{l}\alpha_t\end{aligned} \tag{2}$$

where $K = m(l_f k_f - l_r k_r)/l^2 k_f k_r$, $\hat{v}_t$ is velocity measurement and $\hat{\omega}_t$ is yaw velocity measurement, α_t is front wheel angle, can be calculated by formula $\alpha_t{=}\delta_t/i$, where δ_t is steering wheel angle, i is transmission ratio from steering wheel to front wheel. The velocity $\hat{v}_t$ and steering wheel angle δ_t are both obtained from VCU via CAN bus.We introduce the plane hypothesis that exploits the fact that the vehicle moves on an approximately planar surface. The raw vehicle dynamics model measurements, $\hat{\boldsymbol{v}}_t$ and $\hat{\boldsymbol{\omega}}_t$, are given by:

$$\begin{aligned}\hat{\boldsymbol{v}}_t &= \begin{bmatrix}\hat{v}_t cos\beta_t & \hat{v}_t sin\beta_t & 0\end{bmatrix}^T\\ \hat{\boldsymbol{\omega}}_t &= \begin{bmatrix}0 & 0 & \hat{\omega}_t\end{bmatrix}^T\end{aligned} \tag{3}$$

The measurement noise, $\mathbf{n_v}$ and $\mathbf{n}_\omega$ are modeled as zero mean, white Gaussian noise, $\mathbf{n}_v{\sim}N(0, \sigma_v^2)$, $\mathbf{n}_\omega{\sim}N(0, \sigma_\omega^2)$.

B. Dynamics Pre-integration

The inspiration of dynamics pre-integration came from IMU pre-integration, which was first proposed in [29] and has became the basic module in subsequently proposed VINS. The frequency of the dynamics measurements is generally around 100hz, which is much higher than visual observations. Hence, we pre-integrate these measurements between selected keyframes like IMU pre-integration. In our work, we follow the continuous-time quaternion-based derivation as [12]. Assume that the state at time t is known, the continuous evolution of the dynamics orientation $\mathbf{q}_{v_{k+1}}^w$ and position $\mathbf{p}_{v_{k+1}}^w$ in the world frame can be obtained as follows:

$$\begin{aligned}\mathbf{p}_{v_{k+1}}^w &= \mathbf{p}_{v_k}^w + \int_{t\in[t_k,t_{k+1}]}[\mathbf{R}_t^w(\hat{\mathbf{v}}_t - \mathbf{n}_v)]dt\\ \mathbf{q}_{v_{k+1}}^w &= \mathbf{q}_{v_k}^w \otimes \int_{t\in[t_k,t_{k+1}]}\frac{1}{2}\boldsymbol{\Omega}(\hat{\boldsymbol{\omega}}_t - \mathbf{n}_w)\mathbf{q}_t^{v_k}dt\end{aligned} \tag{4}$$

where:

$$\boldsymbol{\Omega}(\boldsymbol{\omega}) \triangleq \begin{bmatrix}0 & -\boldsymbol{\omega}^T\\ \boldsymbol{\omega} & \boldsymbol{\omega}^\wedge\end{bmatrix} = \begin{bmatrix}0 & -\omega_x & -\omega_y & -\omega_z\\ \omega_x & 0 & \omega_z & -\omega_y\\ \omega_y & -\omega_z & 0 & \omega_x\\ \omega_z & \omega_y & -\omega_x & 0\end{bmatrix}$$

After we change the reference frame from world frame to local vehicle dynamics frame v_k, we are able to pre-integrate the parts which are only related to velocity $\hat{v}_t$ and angular velocity $\hat{\omega}_t$. In time interval $[t_k, t_{k+1}]$, the continuous-time form of pre-integration terms are:

$$\begin{aligned}\boldsymbol{\alpha}_{v_{k+1}}^{v_k} &= \int_{t\in[t_k,t_{k+1}]}[\mathbf{R}_t^{v_k}(\hat{\mathbf{v}}_t - \mathbf{n}_v)]dt\\ \boldsymbol{\gamma}_{v_{k+1}}^{v_k} &= \int_{t\in[t_k,t_{k+1}]}\frac{1}{2}\boldsymbol{\Omega}(\hat{\boldsymbol{\omega}}_t - \mathbf{n}_w)\boldsymbol{\gamma}_t^{v_k}dt\end{aligned} \tag{5}$$

For numerical calculation, considering that the value of the dynamics measurements is relatively stable between two consecutive visual frames, we use Euler integration for discretization. The dynamics integration expression above is the ideal state of ignoring noise. In order to get the information matrix and Jacobians of dynamics measurements residual in optimization part, we introduce the noise and derive discrete-time error-state kinematics:

$$\begin{aligned}&\begin{bmatrix}\delta\boldsymbol{\alpha}_{k+1}\\ \delta\boldsymbol{\theta}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & -\mathbf{R}_k\hat{\mathbf{v}}_k^\wedge\delta t\\ 0 & \mathbf{I} - \hat{\boldsymbol{\omega}}_k^\wedge\delta t\end{bmatrix}\begin{bmatrix}\delta\boldsymbol{\alpha}_k\\ \delta\boldsymbol{\theta}_k\end{bmatrix} +\\ &\begin{bmatrix}-\mathbf{R}_k\delta t & 0\\ 0 & -\delta t\end{bmatrix}\begin{bmatrix}\mathbf{n}_v\\ \mathbf{n}_w\end{bmatrix} = \mathbf{A}_t\delta\mathbf{z}_t^{v_k} + \mathbf{B}_t\mathbf{n}_t\end{aligned} \tag{6}$$

Through the equation (6) we can get the matrix $\mathbf{A_t}$ and $\mathbf{B_t}$ for the propagation of Jacobians and convariance. In practice

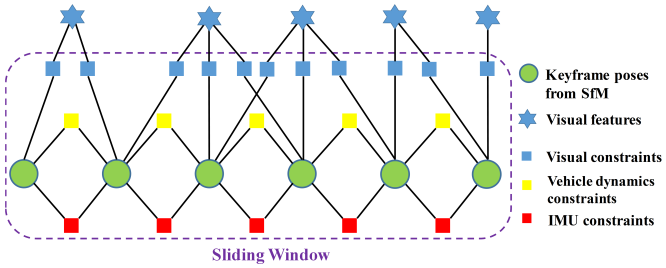


Fig. 4. Diagramm of the visual-inertial-dynamics alignment.

we found that the transmission ratio i (We use it to get the front wheel angle α_t in equation (2) and is not a fixed value, and will vary with the velocity of the car. Therefore, if the estimation of transmission ratio changes slightly, $\alpha^{v_k}_{v_{k+1}}$ and $\gamma^{v_k}_{v_{k+1}}$ will be adjusted by their first-order approximations with respect to the transmission ratio as:

$$\begin{aligned} \boldsymbol{\alpha}^{v_k}_{v_{k+1}} &\approx \hat{\boldsymbol{\alpha}}^{v_k}_{v_{k+1}} + \mathbf{J}^{\boldsymbol{\alpha}}_i \delta i \\ \boldsymbol{\gamma}^{v_k}_{v_{k+1}} &\approx \hat{\boldsymbol{\gamma}}^{v_k}_{v_{k+1}} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}\mathbf{J}^{\gamma}_i \delta i \end{bmatrix} \end{aligned} \tag{7}$$

where $\mathbf{J}^{\boldsymbol{\alpha}}_i = \frac{\partial \boldsymbol{\alpha}^{v_k}_{v_{k+1}}}{\partial i}$, $\mathbf{J}^{\gamma}_i = \frac{\partial \boldsymbol{\gamma}^{v_k}_{v_{k+1}}}{\partial i}$

According to equation (7), when the value of transmission ratio minorly changes, we can adjust it easily instead of re-integration.

C. Visual-inertial-dynamics Alignment Based Initialization

An illustration of visual-inertial-dynamics alignment is shown in Fig. 4. We match the visual SfM results, IMU pre-integration and dynamics pre-integration for estimating the initial value of metric scale, gyroscope and accelerometer bias, velocity as well as gravity vector.

Assume that we get the poses $(\bar{\mathbf{p}}^{c_0}_{b_k}, \mathbf{q}^{c_0}_{b_k})$ and $(\bar{\mathbf{p}}^{c_0}_{b_{k+1}}, \mathbf{q}^{c_0}_{b_{k+1}})$ of two consecutive frames from the visual-only SfM ($(\cdot)^{c_0}$ is the first camera and is set as the reference frame), and $\hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}}$, $\hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}}$, $\hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}}$ from IMU integration. $(\mathbf{p}^b_c, \mathbf{q}^b_c)$ are extrinsic parameters between camera and IMU, $(\mathbf{p}^v_c, \mathbf{q}^v_c)$ are extrinsic parameters between camera and vehicle dynamics, and extrinsic parameters between vehicle dynamics and IMU are $(\mathbf{p}^b_v, \mathbf{q}^b_v)$, which can all be measured and calibrated in advance. We will treat these known quantity as the input of our initialization module. The whole process is divided into several simpler subproblems:

1) Metric Scale Estimation: Dynamics and visual measurements are aligned to estimated the metric scale. The scale of camera trajectory computed by visual-only methods is arbitrary. Hence we need to introduce a scale factor s when the coordinate system is transformed between camera and vehicle dynamics:

$$s\bar{\mathbf{p}}^{c_0}_{v_k} = s\bar{\mathbf{p}}^{c_0}_{c_k} - \mathbf{R}^{c_0}_{v_k}\mathbf{p}^v_c \tag{8}$$

After that we combine the (4), (5) and (8) into the following linear equation:

$$\mathbf{R}^v_b \mathbf{R}^{b_k}_{c_0}(\bar{\mathbf{p}}^{c_0}_{c_{k+1}} - \bar{\mathbf{p}}^{c_0}_{c_k})s = \hat{\boldsymbol{\alpha}}^{v_k}_{v_{k+1}} + \mathbf{R}^v_b \mathbf{R}^{b_k}_{c_0}\mathbf{R}^{c_0}_{b_{k+1}}\mathbf{R}^b_v \mathbf{p}^v_c - \mathbf{p}^v_c \tag{9}$$

From (9) we can see $\mathbf{R}^{c_0}_{b_{k+1}}$, $\mathbf{R}^{c_0}_{b_k}$, $\bar{\mathbf{p}}^{c_0}_{c_k}$, $\bar{\mathbf{p}}^{c_0}_{c_{k+1}}$ from monocular visual SfM. To simplify expression, (9) can be written as:

$$\mathbf{F}^{v_k}_{v_{k+1}} s = \hat{\mathbf{z}}^{v_k}_{v_{k+1}} \tag{10}$$

Then we can get the scale factor s by solving this linear least square problem:

$$\min_s \sum_{k\in W} \left\| \hat{\mathbf{z}}^{v_k}_{v_{k+1}} - \mathbf{F}^{v_k}_{v_{k+1}} s \right\|^2 \tag{11}$$

where W indexes all frames in the window(assume we use sliding window based optimization, the same below).

2) Gyroscope Bias Estimation: We use dynamics and IMU measurements to calibrate the gyroscope bias $\mathbf{b}_g$. Since the gyroscope bias only affects the angular velocity measurements, we optimize it by minimizing the difference between gyroscope integration and vehicle dynamics angular velocity integration, for all pairs of consecutive keyframes:

$$\begin{aligned} &\min_{\delta \mathbf{b}_g} \sum_{k\in W} \left\| \mathbf{q}^b_v \otimes \hat{\boldsymbol{\gamma}}^{v_{k+1}}_{v_k} \otimes {\mathbf{q}^b_v}^{-1} \otimes \boldsymbol{\gamma}^{b_k}_{b_{k+1}} \right\|^2 \\ &\boldsymbol{\gamma}^{b_k}_{b_{k+1}} \approx \hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}\mathbf{J}^{\gamma}_{\mathbf{b}_g}\delta \mathbf{b}_g \end{bmatrix} \end{aligned} \tag{12}$$

We solve (12) by LDLT methods with a zero bias initial value. In such way we can get the initial estimation of gyroscope bias $\mathbf{b}_g$. After that we re-integrate the IMU measurements to obtain $\hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}}$, $\hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}}$, $\hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}}$ using the new gyroscope bias.

3) Velocity, Gravity Vector and Accelerometer Bias Estimation: After the matric scale and gyroscope bias are estimated, we move on to align the measurements of all the three sensors to estimate the velocity, gravity vector and accelerometer bias. Although we can already get the velocity from VCU via CAN bus, in fact, this velocity is calculated from the motor speed and essentially the speed of the front wheel. What we need to estimate is the velocity of body frame. In addition, the time synchronization is hard to realize for dynamics measurements and other equipments because of the triggering and transmission delays, which makes the existence of time offset. Based on these mentioned reasons, we add velocity into estimated variables. We do note that as proved in [13], [17], when a VINS platform has no rotation motion, it's difficult to calibrate the accelerometer bias in initialization procedure since acceleration is usually coupled with gravity under small rotation. The aggressive rotation movement in the beginning is infeasible especially for ground vehicles due to the dynamical constraints. However, this observability problem does not exist for our vehicle dynamics sensor. The dynamics measurements can be introduced to calibrate the accelerometer bias under nearly linear motions. The estimated variables in this step can be defined as:

$$\mathbf{X}_I = [\mathbf{v}^{b_0}, \mathbf{v}^{b_1}, \cdots \mathbf{v}^{b_n}, \mathbf{g}^{c_0}, \delta \mathbf{b}_a] \tag{13}$$

where $\mathbf{v}^{b_k}$ is velocity in the body frame while taking the k^{th} frame. $\mathbf{g}^{c_0}$ denotes the gravity vector in visual reference

frame c_0. Consider two consecutive frames b_k and b_{k+1}, we have the following equation:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k} =&\mathbf{R}_{c_0}^{b_k}[(\mathbf{p}_{b_{k+1}}^{c_0}-\mathbf{p}_{b_{k+1}}^{c_0})+\frac{1}{2}\mathbf{g}^{c_0}\Delta t_k^2-\mathbf{R}_{c_0}^{b_k}\mathbf{v}^{b_k}\Delta t_k]\\&-\mathbf{J}_{\mathbf{b}_a}^{\boldsymbol{\alpha}}\delta\mathbf{b}_a\\ \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} =&\mathbf{R}_{c_0}^{b_k}(\mathbf{R}_{b_{k+1}}^{c_0}\mathbf{v}^{b_{k+1}}+\mathbf{g}^{c_0}\Delta t_k-\mathbf{R}_{b_k}^{c_0}\mathbf{v}^{b_k})-\mathbf{J}_{\mathbf{b}_a}^{\boldsymbol{\beta}}\delta\mathbf{b}_a\end{aligned} \tag{14}$$

The relationship between the translation of vehicle dynamics frame and body frame can be described as:

$$\mathbf{p}_{b_{k+1}}^{c_0}-\mathbf{p}_{b_{k+1}}^{c_0}=\mathbf{R}_v^b\mathbf{R}_{c_0}^{b_k}\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}+(\mathbf{R}_{b_{k+1}}^{c_0}\mathbf{R}_v^b-\mathbf{R}_{b_k}^{c_0}\mathbf{R}_v^b)\mathbf{p}_b^v \tag{15}$$

After combining the equation (14) and (15), we can get the following linear measurement model:

$$\hat{\mathbf{z}}_{b_{k+1}}^{b_k}=\begin{bmatrix}\hat{\mathbf{z}}_1\\ \hat{\mathbf{z}}_2\end{bmatrix}=\mathbf{G}_{b_{k+1}}^{b_k}\mathbf{X}_I+\mathbf{n}_{b_{k+1}}^{b_k} \tag{16}$$

where

$$\begin{aligned}\hat{\mathbf{z}}_1=&\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}-\mathbf{R}_{c_0}^{b_k}[\mathbf{R}_v^b\mathbf{R}_{c_0}^{b_k}\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}+(\mathbf{R}_{b_{k+1}}^{c_0}\mathbf{R}_v^b-\mathbf{R}_{b_k}^{c_0}\mathbf{R}_v^b)\mathbf{p}_b^v]\\ \hat{\mathbf{z}}_2=&\hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k}\\ \mathbf{G}_{b_{k+1}}^{b_k}=&\begin{bmatrix}-\mathbf{I}\Delta t_k & 0 & \frac{1}{2}\mathbf{R}_{c_0}^{b_k}\Delta t_k^2 & -\mathbf{J}_{\mathbf{b}_a}^{\alpha}\\ -\mathbf{I} & \mathbf{R}_{c_0}^{b_k}\mathbf{R}_{b_{k+1}}^{c_0} & \mathbf{R}_{c_0}^{b_k}\Delta t_k & -\mathbf{J}_{\mathbf{b}_a}^{\beta}\end{bmatrix}\end{aligned} \tag{17}$$

We can see Δt_k is the time interval between two consecutive frames. The estimation problem is turned into a linear least-square problem:

$$\min_{\mathbf{X}_I}\sum_{k\in W}\left\|\hat{\mathbf{z}}_{b_{k+1}}^{b_k}-\mathbf{G}_{v_{k+1}}^{v_k}\mathbf{X}_I\right\|^2 \tag{18}$$

By solving this problem, we can estimate the initial value of variables in (13).

4) Plane Hypothesis Correction: We adopted the method proposed in [11], [13] to refine the gravity vector. So ,it's not necessary to introduce it here. After the gravity refinement we can get a rotation quaternion, which is used to adjust the direction of the gravity. In the world frame, the direction of gravity is perpendicular to the plane of motion. For subsequent dynamics measurements, we use this rotation quaternion to adjust $\hat{\boldsymbol{v}}_t$ and $\hat{\boldsymbol{\omega}}_t$ in (3).

We then rotate all the variables from the reference frame $(\cdot)_0^c$ to the world frame $(\cdot)^w$ and add the metric scale to the visual measurements from SfM. At this point, the whole initialization process is finished and all variables are used to bootstrap the sliding window based optimization backend.

D. Optimization with Vehicle Dynamics Residual

The current backend optimization part of tightly-coupled VINS mainly includes filter-based method and sliding windows based nonlinear method. We extend dynamics constrains to the sliding windows based nonlinear method, because it can not only reduce the computation complexity by removing old states but also maintain the information about the previous states of the system. The implementation to filter-based method is also feasible, but there are some differences in detail. The variables that we will optimize include:

$$\begin{aligned}\mathbf{X}&=[\mathbf{x}_{b0},\mathbf{x}_{b1},\cdots\mathbf{x}_{bn},\lambda_{c0},\lambda_{c1},\cdots\lambda_{cm},i_v]\\ \mathbf{x}_{bk}&=[\mathbf{p}_{b_k}^w,\mathbf{v}_{b_k}^w,\mathbf{q}_{b_k}^w,\mathbf{b}_a,\mathbf{b}_g],k\in[0,n]\end{aligned} \tag{19}$$

where the k_{th} IMU state $\mathbf{x}_{bk}$ includes position $\mathbf{p}_{b_k}^w$, velocity $\mathbf{v}_{b_k}^w$ and rotation $\mathbf{q}_{b_k}^w$ of body frame with respect to world frame. The visual observations are parameterized by inverse depth λ_{cl}. Condidering two consecutive vehicle dynamics frames v_k and v_{k+1} in the sliding window, the dynamics measurements residual can be written as:

$$\begin{aligned}\mathbf{r}_V(\hat{\mathbf{z}}_{v_{k+1}}^{v_k},\mathbf{X})=&\begin{bmatrix}\delta\boldsymbol{\alpha}_{v_{k+1}}^{v_k}\\ \delta\boldsymbol{\theta}_{v_{k+1}}^{v_k}\end{bmatrix}\\ =&\begin{bmatrix}\mathbf{R}_w^{v_k}(\mathbf{p}_{v_{k+1}}^w-\mathbf{p}_{v_k}^w)-\hat{\boldsymbol{\alpha}}_{v_{k+1}}^{v_k}\\ 2[\hat{\boldsymbol{\gamma}}_{v_{k+1}}^{v_k}{}^{-1}\otimes\mathbf{q}_{v_k}^{w}{}^{-1}\otimes\mathbf{q}_{v_{k+1}}^w]_{xyz}\end{bmatrix}\end{aligned} \tag{20}$$

Then we add the vehicle dynamics measurements residual to a visual-inertial bundle adjustment formulation and minimize the whole error function:

$$\begin{aligned}\mathbf{X}^*=\min_{\mathbf{X}}\{&\|\mathbf{r}_p-\mathbf{H}_p\mathbf{X}\|^2+\sum_{k\in W}\|\mathbf{r}_B\|_{\mathbf{P}_B}^2+\\&\sum_{k\in W}\left\|\mathbf{r}_V(\hat{\mathbf{z}}_{v_{k+1}}^{v_k},\mathbf{X})\right\|_{\mathbf{P}_V}^2+\sum_{(l,j)\in C}\rho(\|\mathbf{r}_C\|_{\mathbf{P}_C}^2)\}\end{aligned} \tag{21}$$

where $\mathbf{X}^*$ is the estimated state vector, $\|\cdot\|^2$ is the Mahalanobis norm weight by conariance $\mathbf{P}$. $\{\mathbf{r}_p,\mathbf{H}_p\}$ is the prior information from marginalization. $\mathbf{r_B}$, $\mathbf{r_V}$ and $\mathbf{r_C}$ are IMU, vehicle dynamical and visdual measurements respectively. $\rho(\cdot)$ is a robust huber norm. At last, we solve this maximum posterior estimation problem by Ceres Solver.

IV. Experimental Results

In this section, our aim is to validate the effect of the proposed vehicle dynamics extention to VINS. We test the accuracy and robustness of our method in underground and outdoor environments respectively. Note that the proposed extension is generic and not restricted to any particular VINS estimator. Compared to other VINS estimator, the state-of-art open-source solution VINS-Mono[1], is a robust and versatile SLAM system with not only tightly coupled VIO, but also relocalization and global optimization, and has been applied on drones. Hence we choose VINS-Mono to implement our extension.

Since the current public datasets does not contain both velocity and steering wheel angle, we use the datasets captured by the TiEV[2] autonomous driving platform at JiaDing campus of Tongji University. We use a mass-produced, low-cost sensor MYNT EYE Depth Camera D-Series[3] to collect stereo images(global shutter, 60 FPS, 1280×720) and synchronized IMU measurements(Bosch BMI088, 6 Axis). The other sensors we use in this experiment include Novatel span-KVH 1750(outdoor groundtruth), velodyne vlp-16 and USB-CAN(collect velocity and steering wheel angle from

[1]https://github.com/HKUST-Aerial-Robotics/VINS-Mono

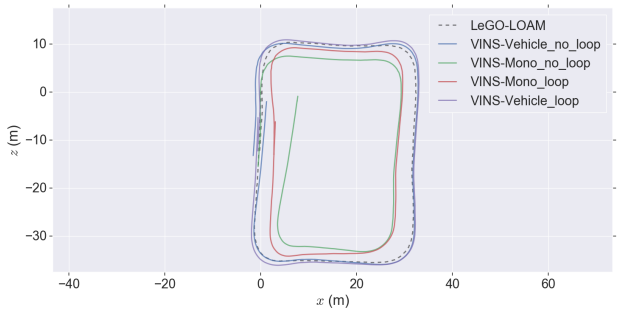
[2]cs1.tongji.edu.cn/tiev

[3]https://www.myntai.com/cn/mynteye/depth

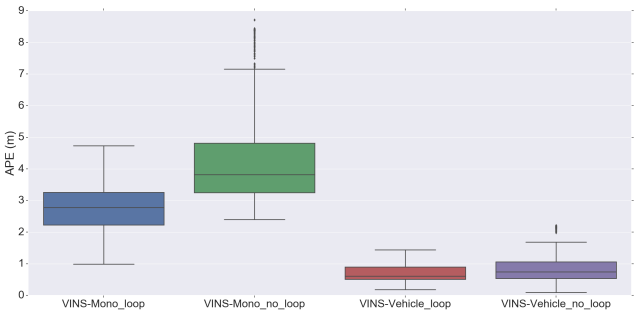
TABLE I

RMSE[M] IN PARKING LOT DATASETS

Datasets	Length(m)	Duration(s)	RMSE(m)			
			VINS-Mono	VINS-Vehicle	VINS-Mono_loop	VINS-Vehicle_loop
PL1	162.32	71.57	4.42	**0.98**	2.96	**0.74**
PL2	492.72	198.63	2.92	**2.62**	1.41	**1.26**
PL3	353.97	151.54	2.53	**2.30**	1.03	**0.68**
PL4	357.29	154.42	4.07	**3.09**	1.48	**0.94**

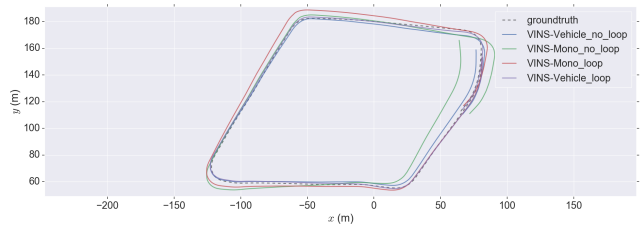


(a)

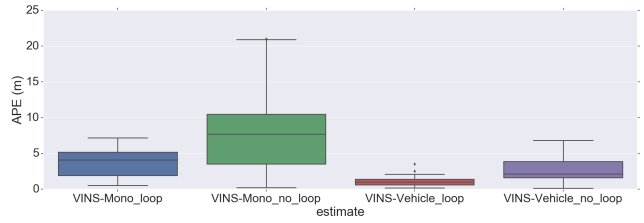


(b)

Fig. 5. Results of PL1 dataset. (a) Estimated trajectory in PL1 dataset, compared with VINS-Mono and LeGO-LOAM(groundtruth). (b) Absolute pose error in PL1 dataset.



(a)



(b)

Fig. 6. Results of Outdoor1 dataset. (a) Estimated trajectory in Outdoor1 dataset, compared with VINS-Mono and groundtruth. (b) Absolute pose error in Outdoor1 dataset.

CAN bus to PC). The proposed method runs in real-time for all experiments on a standard laptop (1.8GHz Quad-Core Intel Core i7-8550U processor, 8GB RAM). We adopt the evaluation scheme of evo[4], where the ground truth and estimated trajectories are aligned, and then the absolute pose error(APE) at each corresponding timestamp can be calculated. Besides, the statistical property of APE can also be analyzed by this tool.

A. *Underground Parking Lot Experiments*

We evaluate our proposed method in the different underground parking lots on Jiading Campus, Tongji University. They are illuminated by dim lights, and we collected four datasets there. Low-light and textureless environment brings great challenges to both state-of-art indirect and direct methods, such as OKVIS [9], VIORB[5] and DSO [6] with scale correct. The first two feature matching based VIO methods failed due to tracking lost and scale drift. Besides, the photometric error optimization based direct method failed because of bad illumination condition. Therefore, we compare the results to VINS-Mono which is the base to implement our proposed extension. In order to verify the effect of vehicle dynamics extension, we tested both VIO without loop-closure and SLAM system with loop-closure. There are no GPS in underground parking lot, so we take the results of LeGO-LOAM [30] as groundtruth.

The estimated trajectories in PL1(Parking Lot1) dataset are showed in Fig. 5(a). For VINS-Mono, noticeable position drifts occured with or without a loop-closure. In addition, on account of degenerate motions, significant scale drifts appears in VINS-Mono, which makes the circumference of the trajectory smaller than groundtruth. In the absolute pose error(APE) plot Fig. 5(a), VINS-Vehicle with a loop-closure outperforms others. The root-mean-square error (RMSE) of all underground parking lot datasets are shown in Tab. I, which is also evaluated by an APE. We can find that with the assistance of vehicle dynamics, VINS-Vehicle outperforms VINS-Mono whether with a loop-closure or not. Combining the aforementioned results of OKVIS and VIORB, it can be concluded that the method which involves vehicle dynamics is more robust and accurate than state-of-art VINS in textureless and low-light underground parking lot.

[4]https://github.com/MichaelGrupp/evo

[5]https://github.com/jingpang/LearnVIORB

TABLE II
RMSE[M] IN OUTDOOR DATASETS

Datasets	Length(m)	Duration(s)	Weather	Illumination	RMSE(m)			
					VINS-Mono	VINS-Vehicle	VINS-Mono_loop	VINS-Vehicle_loop
Outdoor1	594.96	112.91	sunny	good	8.95	**3.13**	4.14	**1.12**
Outdoor2	1650.56	319.50	cloudy	normal	35.67	**9.35**	7.61	**1.25**
Outdoor3	844.93	169.258	cloudy	normal	10.41	**7.44**	2.12	**1.49**
Outdoor4	785.11	156.91	sunny	strong	11.13	**7.92**	3.33	**1.78**
Outdoor5	663.65	124.06	drizzle	poor	21.27	**6.74**	12.36	**2.83**
Outdoor6	1239.53	243.83	drizzle	poor	42.94	**9.64**	14.26	**2.04**

B. Outdoor Experiments

In this experiment, the robustness and accuracy of our proposed method in large-scale dynamic outdoor environment are evaluated. In order to simulate the application scenario of autonomous driving, datasets are collected in different weather and illumination conditions(see Tab. II),and there are vehicles and pedestrians passing by. Besides, the range of velocity is controlled between 15km/h and 40km/h. Some scenes in datasets are less-feature because of the lack of tall buildings and trees. OKVIS and VIORB are also unable to work on any entire dataset due to the wrong feature matching and tracking failure. So we use the same methods for comparison as in underground experiments, and the groundtruth is from Novatel span-KVH 1750(RTK+GPS+IMU), which can achieve positioning accuracy within 5cm.

The estimated trajectories in Outdoor1 dataset are showed in Fig. 6(a). As same as underground experiments, for VINS-Mono, the accumulation of scale estimation drift due to degenerate motions results in obvious position error. From the absolute pose error plot in Fig. 6(b), we can see that even with the help of a loop-closure, our proposed method performs better than VINS-Mono. The RMSE of all the outdoor datasets are showed in Tab. II, which is also evaluated by APE. We can see that VINS-Vehicle outperforms VINS-Mono whether with a loop-closure or not on all the outdoor datasets. Moreover, the error percentages of VINS-Vehicle without a loop-closure on two long distance datasets, Outdoor2 and Outdoor6, are both less than 1%. With the help of loop-closure, the APE of our method is only 1 2m, which proves the accuracy under long-duration test. From the results on Outdoor4, Outdoor5 and Outdoor6 with too strong or poor illumination, it can be proved that our method is robust in different illumination conditions. Due to the introduction of vehicle dynamics, the proposed method performs much better than VINS-Mono in the case of poor visual estimation, such as raindrops and low-light interference in Outdoor5 and Outdoor6. In summary, it can be concluded that our proposed extension is able to improve the accuracy and robustness in outdoor dynamics environment under different weather and illumination conditions.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a vehicle dynamics extension for VINS. We use the vehicle dynamics model to pre-integrate vehicle motion information, and initialize the system with the results of visual SfM and IMU pre-integration, as well as involve vehicle dynamics constraints to the tight-coupled optimization backend. Extensive quantitative evaluation in reak world demonstrates that the introduction of vehicle dynamics enhances the accuracy and robustness of VINS, whether in underground parking lots or in outdoor dynamic environments. The potential of our proposed method in autonomous driving application is also proved.

In future work, we will establish more accurate vehicle dynamics model with more DOF and will consider the wheel slippage compensation.

REFERENCES

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1052–1067, 2007.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2007, pp. 1–10.

[3] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.

[4] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[7] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 3565–3572.

[8] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 298–304.

[9] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2014.

[10] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2015.

[11] R. Mur-Artal and J. D. Tardos, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2016.

[12] Q. Tong, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. PP, no. 99, pp. 1–17, 2017.

[13] T. Qin and S. Shen, "Robust initialization of monocular visual-inertial estimation on aerial robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4225–4232.
[14] J. Kaiser, A. Martinelli, F. Fontana, and D. Scaramuzza, "Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 18–25, 2016.
[15] J. Dominguez-Conti, J. Yin, Y. Alami, and J. Civera, "Visual-inertial slam initialization: A general linear formulation and a gravity-observing non-linear optimization," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2018.
[16] K. J. Wu and S. I. Roumeliotis, "Unobservable directions of vins under special motions," *university of of Minnesota, Dept. of Comp, Sci. & Eng., Tech. Rep.*, 2016.
[17] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "Vins on wheels," in *IEEE International Conference on Robotics and Automation*, 2017.
[18] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments," in *IEEE International Conference on Robotics and Automation*, 2013.
[19] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.
[20] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2510–2517.
[21] L. Kneip, S. Weiss, and R. Siegwart, "Deterministic initialization of metric state estimation filters for loosely-coupled monocular vision-inertial systems," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 2235–2241.
[22] B. Liang and N. Pears, "Visual navigation using planar homographies," in *IEEE International Conference on Robotics and Automation*, 2002.
[23] W. Hui, K. Yuan, Z. Wei, and Q. Zhou, "Visual odometry based on locally planar ground assumption," in *IEEE International Conference on Information Acquisition*, 2005.
[24] J. J. Guerrero, R. Martinez-Cantin, and C. Sagüés, "Visual map-less navigation based on homographies," *Journal of Robotic Systems*, vol. 22, no. 10, pp. 569–581, 2010.
[25] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *IEEE International Conference on Robotics and Automation*, 2009.
[26] D. Scaramuzza, "1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *International Journal of Computer Vision*, vol. 95, no. 1, pp. 74–85, 2011.
[27] W. Zong, L. Chen, C. Zhang, Z. Wang, and Q. Chen, "Vehicle model based visual-tag monocular orb-slam," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 1441–1446.
[28] D. E. Smith and J. M. Starkey, "Effects of model complexity on the performance of automated vehicle steering controllers: Model development, validation and comparison," *Vehicle System Dynamics*, vol. 24, no. 2, pp. 163–181, 1995.
[29] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.
[30] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.