Ground-VIO: Monocular Visual-Inertial Odometry with Online Calibration of Camera-Ground Geometric Parameters

Yuxuan Zhou, Xingxing Li, Shengyu Li, Xuanbin Wang, Zhiheng Shen

Abstract—Monocular visual-inertial odometry (VIO) is a low-cost solution to provide high-accuracy, low-drifting pose estimation. However, it has been meeting challenges in vehicular scenarios due to limited dynamics and lack of stable features. In this paper, we propose Ground-VIO, which utilizes ground features and the specific camera-ground geometry to enhance monocular VIO performance in realistic road environments. In the method, the camera-ground geometry is modeled with vehicle-centered parameters and integrated into an optimization-based VIO framework. These parameters could be calibrated online and simultaneously improve the odometry accuracy by providing stable scale-awareness. Besides, a specially designed visual front-end is developed to stably extract and track ground features via the inverse perspective mapping (IPM) technique. Both simulation tests and real-world experiments are conducted to verify the effectiveness of the proposed method. The results show that our implementation could dramatically improve monocular VIO accuracy in vehicular scenarios, achieving comparable or even better performance than state-of-art stereo VIO solutions. The system could also be used for the auto-calibration of IPM which is widely used in vehicle perception. A toolkit for ground feature processing, together with the experimental datasets, would be made open-source[1].

Index Terms—Visual-inertial odometry, autonomous vehicle navigation, camera-ground geometry, inverse perspective mapping.

I. INTRODUCTION

VISION-based solutions have been pivotal in the development of intelligent vehicle applications [1] [2]. The low-cost camera could provide high-resolution texture information of the environment, enabling high-level perception such as object detection [3] and scene parsing [4]. On the other hand, visual simultaneous localization and mapping (VSLAM) provides a feasible approach for accurate vehicle pose estimation, which could be used for navigation tasks [5] [6]. Such aspect of vision-based navigation is later enhanced with the introduction of inertial measurement unit (IMU), bringing about better stability and accuracy with very limited additional expenses [7]. Besides, IMU could resolve the scale ambiguity in monocular VSLAM, facilitating more practical use cases. The outstanding performance of visual-inertial odometry (VIO) and visual-inertial navigation system

The authors are with School of Geodesy and Geomatics, Wuhan University, China (e-mail: xxli@sgg.whu.edu.cn).
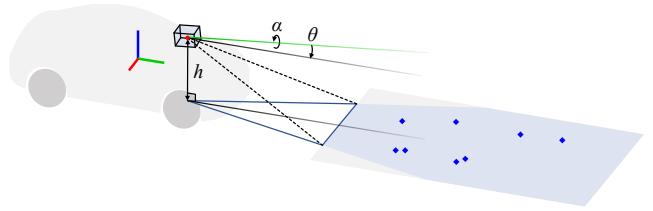
[1]https://github.com/GREAT-WHU/gv_tools

Fig. 1. Illustration of the camera-ground geometry. For a ground vehicle, the local ground plane could be expressed as a specific plane in the camera frame, which is parameterized as a height with a two-step rotation in this work (see Sect. III).

(VINS) has been demonstrated in unmanned aerial vehicle (UAV) applications [8] [9].

However, as a possible low-cost navigation solution, monocular VIO has been meeting challenges when applied to ground vehicles. Unlike UAVs, it is impractical for a ground vehicle to sufficiently excite the IMU during regular maneuvers. This would significantly affect the pose estimation performance of VIO due to lack of observability [10] [11], especially for the scale which is unresolvable in monocular vision. Although stereo camera setups can mitigate this issue to some extent [12], they entail additional expenses and computation cost. Other schemes turn to a higher integration level, introducing other sensors (e. g. wheel encoder, GNSS and LiDAR) to achieve better navigation performance [13]–[15].

It is noted that, the ground itself provides a natural and powerful constraint that could be utilized to enhance VSLAM/VIO performance. Considering the fact that the vehicle moves on the ground, the vehicle-mounted sensor and the local ground plane have a relatively fixed geometric relationship, depending mainly on the sensor installation and the vehicle size. For VSLAM/VIO, the local ground plane could be expressed as a specific plane in the camera frame, as shown in Fig. 1. We term this fixed relationship as the camera-ground geometry, which could be used to constrain the landmark depths, thu=s deriving metric-scale geometric information that is essential for high-accuracy pose estimation. However, few researches have given an in-depth insight into the application of camera-ground geometry in VIO.

In fact, such camera-ground geometry has been widely, sometimes implicitly, applied in vehicle perception. Typically, the well known inverse perspective mapping (IPM) technique [16] [17] is using the pre-calibrated camera-ground geometry to generate bird-eye view (BEV) images, or known as around-view monitoring (AVM) [18], thus to efficiently perceive the

arXiv:2306.08341v2 [cs.RO] 18 Jun 2023

surrounding road environment [19]–[21]. From this aspect, the online auto-calibration of the camera-ground geometry is also a meaningful issue and still remains unsolved.

In the proposed Ground-VIO, we introduce the online estimation of the camera-ground geometric parameters, termed as C-G parameters, into a monocular VIO, which could not only improve the odometry performance but also provide an approach for the auto-calibration of IPM. The contributions of this work are as follows:

- A vehicle-centered model is proposed to parameterize the camera-ground geometry, which is integrated into the monocular VIO for online calibration and to improve the navigation performance.
- A novel visual front-end is developed to precisely track features on the ground by making use of the camera-ground geometry and IPM.
- Both simulation tests and real-world experiments were conducted to validate different aspects of the system, including the estimation of C-G parameters, the odometry accuracy and the IPM calibration performance.
- We make the ground feature processing module and the test data sequences open-source.

The rest of the paper is organized as follows. In Section II, related works are discussed. In Section III, the system overview is presented. In Section IV, the core idea of camera-ground geometric model is explained. Section V presents the implementation details of Ground-VIO. The system performance is evaluated in Section VI and Section VII through simulation tests and real-world experiments respectively. The conclusion is finally given in Section VIII.

II. RELATED WORK

A. Visual-Inertial Odometry

The aspect of VIO/VINS has been extensively investigated in the past decades, applied in both UAV and ground vehicle applications. Generally, the implementations could be divided into filter-based and optimization-based methods. For filter-based methods, the representative framework is multi-state constraint kalman filter (MSCKF) [22], which maintains historical IMU poses in the state vector and uses common-view feature observations to construct geometric constraints among the poses. The variants of MSCKF have been developed to improve the framework by introducing observability constraint, extrinsic calibration [23], multi-IMU/camera configuration [24], landmark states [25] and so on. For optimization-based implementations, the mainstream methods use a factor graph which jointly optimizes IMU preintegration factors [26] and visual re-projection factors to estimate the navigation states and landmark positions [27]. These methods are expected to achieve better performance through iterations and relinearizations, in the expense of higher computation cost. The sliding window mechanism is a usual way to ensure real-time processing [8] [28], while some other implementations employ a local map to limit the problem size [29].

Despite the advantages of low cost and high accuracy in ideal conditions, VIO/VINS has been meeting challenges when employed for vehicular applications due to limited dynamics, fast motion, and lack of stable features. To improve the practicality of such methods, it is necessary to introduce other sensors or fully utilize the inherent constraint of a ground vehicle.

B. Vehicle Navigation Utilizing the Ground Constraint

It is a natural idea to utilize the ground constraint when designing a navigation system for ground vehicles. Most of these researches model the ground as a local plane (or manifold) and constrain the vehicle motion on it, which is sometimes implicitly comprised in a vehicle-centered non-holonomic constraint (NHC). As pointed out in [30], these implementations could be roughly divided into deterministic and stochastic SE(2) constraints. Deterministic SE(2) constraints strictly constrain the vehicle poses via parameterization or a deterministic model [31] [32], while stochastic SE(2) constraints are applied to SE(3) pose estimation with a time-variant and probabilistic constraint [33]–[36]. Generally, the former is more suitable for indoor or small-scale environments, while the latter shows superiority in outdoor environments with better resistance to outliers. Although some of these methods use a vision-based sensor setup, most of them don't directly associate the visual observations with the ground. Differently, in [30], stereo visual features are utilized to estimate the ground manifold representation, which is later used to constrain the vehicle pose.

Compared to the mentioned methods, our work focuses on the ground observed by the vehicle-mounted camera rather than the vehicle motion constrained on the ground. Actually, there is significant difference between "the vehicle maneuvers on the ground" and "the observed features are on the ground". For VSLAM/VIO, the latter statement could be used to constrain the landmark depths based on the relatively stable camera-ground geometry, thus to provide instantaneous scale-awareness to the system. This characteristic has been pointed out in [38]–[40], all of which use the observed ground structure to realize scale-aware VSLAM based on a monocular camera. However, the camera height should be pre-calibrated in these methods, which limits the usability. Literature [37] takes the camera-geometric geometry into the state vector of VIO, but discusses little about its mechanism. In this work, we would demonstrate that the camera-ground geometry could be calibrated online in a monocular VIO without other infrastructure and could greatly improve the odometry performance.

III. CAMERA-GROUND GEOMETRIC MODEL

In this section, the camera-ground geometric model utilized in the proposed system is firstly introduced.

For a camera mounted on a ground vehicle, it has a specific geometric relationship with the ground. Assuming the local ground is flat and the vehicle is a rigid body (temporarily ignoring the suspension system), the ground plane in the camera frame c is a specific plane that could be unambiguously determined by its normal vector and distance [38].

Here, for better convenience, we parameterize the local ground plane using the height h from the camera center to the ground and a two-step rotation which makes the X-Z plane of
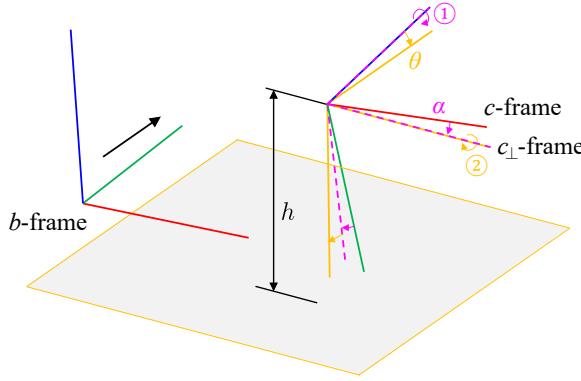
Fig. 2. Illustration of the C-G parameters. The b-frame is the vehicle body/IMU frame, c-frame is the camera frame, $c_\perp$-frame is the virtual camera frame whose X-Z plane is parallel with the local ground.

the camera frame parallel with the ground plane, as illustrated in Fig. 2. Specifically, we firstly rotate the real camera frame c around the Z-axis to make its X-axis parallel with the ground plane. Secondly, we rotate the obtained frame around the X-axis to get the expected virtual camera frame $c_\perp$, which could also be seen as the reference frame of IPM [20].

Thus, the ground plane in the camera frame could be expressed by the following one-row equation

$$\left(\mathbf{R}_{c_\perp}^c \left(\alpha, \theta \right)^\top \mathbf{p}_f^c \right)_y - h = 0 \tag{1}$$

with

$$\mathbf{R}_{c_\perp}^c \left(\alpha, \theta \right) = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{2}$$

where $(\cdot)_{x/y/z}$ denotes the first/second/third row of a three-row vector/matrix, α and θ are the magnitudes of the two-step rotation, corresponding to the roll and the pitch angles of the camera. The triplet (h, θ, α) is defined as the C-G parameters in this paper, which is similar to the parameterization in [37].

Such parameterization makes the estimation of camera-ground geometry straight-forward (i.e., one height and two angles), and it becomes easy to use IMU attitude to compensate the geometry (see Sect. IV-E). It is reasonable to expect the C-G parameters, which indicate the local ground plane, are statistically stable in common road environments without notable change of the sensor alignment or vehicle load. The proposed Ground-VIO fully utilizes this aspect, and several techniques would be given later to deal with complex conditions.

Given that a landmark f in the environment is observed by the camera, we could get

$$\mathbf{p}_f^c = \frac{\mathbf{u}_f}{\lambda_f}, \quad \mathbf{u}_f = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \pi_c^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3}$$

where $\begin{bmatrix} u & v \end{bmatrix}^\top$ is the pixel coordinates of f on the image, $\mathbf{u}_f = \begin{bmatrix} x & y & 1 \end{bmatrix}^\top$ is the normalized image coordinates, π_c is the camera projection matrix, λ_f is the inverse depth of f.
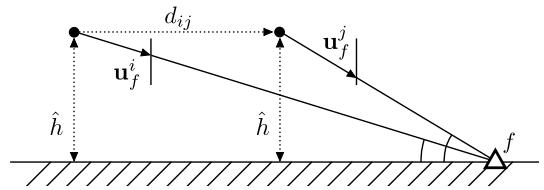

Fig. 3. A simple 1-D model to demonstrate how the camera-ground geometry works in VSLAM/VIO.

Equation (1) and (3) reveal that, with known camera-ground geometry, the metric-scale (inverse) depth of a ground feature on the image could be instantaneously recovered

$$\lambda_f = \frac{1}{h} \left(\mathbf{R}_{c_\perp}^c \left(\alpha, \theta \right)^\top \mathbf{u}_f \right)_y \tag{4}$$

which yields great significance for monocular VSLAM/VIO.

By combining (1) and (3), the camera-ground geometry can be applied in VSLAM/VIO to constrain the landmark depths. Here, a simplified 1-D model is used to qualitatively analyze how it makes sense in VSLAM/VIO. As shown in Fig. 3, the odometry system needs to estimate the traveled distance d_{ij} through the tracking of a landmark f. Assuming the C-G parameters are known and lead to a comprehensive camera height $\hat{h}$, the traveled distance d_{ij} could be obtained

$$d_{ij} = \hat{h} \cdot \left(1/(\mathbf{u}_f^i)_y - 1/(\mathbf{u}_f^j)_y \right) \tag{5}$$

where $\mathbf{u}_f^i$ and $\mathbf{u}_f^j$ are observations of f at epoch i and j.

Assuming the visual observations are noiseless, the estimation of d_{ij} only depends on the accuracy of $\hat{h}$

$$d_{ij} \propto \hat{h} \tag{6}$$

Given a typical case of vehicular scenario where the comprehensive camera height $\hat{h}$ is 2 m with 2 cm error, the estimation of d_{ij} would have 1% relative error. In other words, the tracking of just one ground feature could derive geometric information about the translation with 1% relative error, which is exceedingly meaningful for a monocular visual-inertial system.

It comes to the problems that, 1) how the camera-ground geometry could be integrated into the common VIO, and 2) how the C-G parameters could be obtained or estimated online. These problems would be explained in the following section.

IV. SYSTEM IMPLEMENTATION

In this section, the overview and implementation details of the proposed Ground-VIO will be presented.

A. System Overview

The overall structure of Ground-VIO is shown in Fig. 4. The basic structure of the system follows the classic pipeline of optimization-based VIO [8] but with additional camera-ground-related mechanisms.

Basically, the collected images and IMU data are processed for common VIO initialization and optimization routines.

On this basis, an additional front-end is designed for ground feature processing and works in parallel with the common
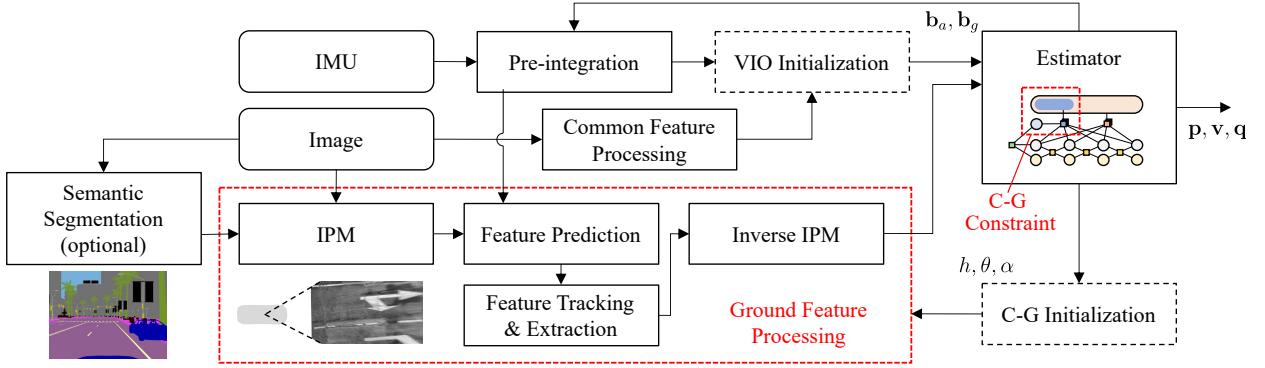
Fig. 4. Structure of the proposed system, adhering to the classic pipeline of optimization-based monocular VIO [8]. On this basis, a specially designed front-end for ground feature processing is added into the pipeline, which works in parallel with the common feature processor. The measurements from the IMU and the camera (including the ground features) are fused in the factor graph estimator. A module for the initialization of the C-G parameters is called periodically after the estimator until the parameters get initialized.

feature processor. The ground feature processor extracts and tracks features on the BEV images generated by IPM, which enables efficient and accurate tracking. A semantic segmentation module could be employed for ground segmentation but is not necessary, which would be discussed later.

In factor graph optimization, the ground features are treated as a subset of visual features with additional camera-ground geometric constraints. These constraints could significantly improve the VIO performance and enable the online estimation of C-G parameters.

Under the situation that the C-G parameters are completely unknown at the beginning, the C-G initialization module would be called every time after common factor graph optimization. Once initialized, the camera-ground-related mechanisms in feature processing and factor graph optimization would be switched on. The C-G parameters would then be continuously refined during VIO running.

B. Ground Feature Processing

In typical implementations of VIO, feature points in the images are continuously tracked to construct visual measurements. The proposed system follows the typical VIO routines [8] to detect and track environmental features in the camera view. To be specific, the feature detection method in [41] and KLT optical flow algorithm [42] are employed, and a fundamental matrix-based RANSAC is used to detect outliers.

As to the ground features, their unique distribution and fast motions on the perspective image make them hard to track. The near-to-far ground plane is highly "warped" on the image, and the near points move drastically despite their better observation geometry.

Instead of the common method, we develop a special module for more precise data association of the ground features. It is noted that, with the camera-ground geometry, the 3-D position of every pixel on the perspective image corresponding to the ground could be instantaneously obtained, referring to (4). From another perspective, we could efficiently generate a BEV image using IPM, and every pixel on the image is directly related to a 3-D position. The following mapping relationship

exists between the metric-scale 3-D point, the point on the perspective image and the point on the BEV image:

$$\mathbf{p}_f^c = \frac{1}{\lambda_f} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = h \cdot \mathbf{R}_{c_\perp}^c (\alpha, \theta) \begin{bmatrix} x_\perp \\ 1 \\ -y_\perp \end{bmatrix} \qquad (7)$$

where $\mathbf{u}_\perp = \begin{bmatrix} x_\perp & y_\perp & 1 \end{bmatrix}^\top$ is the normalized image coordinates of f on the BEV image, the inverse depth λ_f refers to (4). The generation of BEV images through IPM refers to [20].

The knowledge of camera-ground geometry makes the accurate prediction of ground feature tracking possible. Every time a new image comes, we could predict the position of an existing ground feature with the help of the IMU-predicted relative pose

$$\mathbf{p}_f^{c_{k+1}} = \hat{\mathbf{R}}_{c_k}^{c_{k+1}} \mathbf{p}_f^{c_k} + \hat{\mathbf{p}}_{c_k}^{c_{k+1}} \qquad (8)$$

where $\left(\hat{\mathbf{R}}_{c_k}^{c_{k+1}}, \hat{\mathbf{p}}_{c_k}^{c_{k+1}} \right)$ is the relative pose estimated by IMU integration. Combining (7) with (8), the prediction of ground features could be performed on either the perspective image or the BEV image, which could limit the search region of optical flow tracking to several pixels, thereby greatly improving the tracking performance.

In Ground-VIO, we choose to extract and track features on the BEV image, for the reason that the BEV image is less "warped" and has better tracking consistency. In fact, the KLT optical flow tracking doesn't guarantee scale and rotation invariance, with a failure case illustrated in Fig. 5. Fortunately, the IPM could recover the metric-scale geometry of ground features and eliminate most of the scaling effect during fast motion, thus contributing to better tracking precision. Fig. 6 illustrates the tracking of ground features on the BEV image with IMU-aided feature prediction. In addition, a homography matrix-based RANSAC method is used to efficiently detect outlier feature trackings [43]. In our implementation, we mainly focus on the rectangle area (15 m far and $\pm$ 3 m wide with 0.015 m spatial resolution) in front of the vehicle-mounted camera, which facilitates multi-frame tracking of the ground features during regular vehicle maneuvers and guarantees good tracking precision.
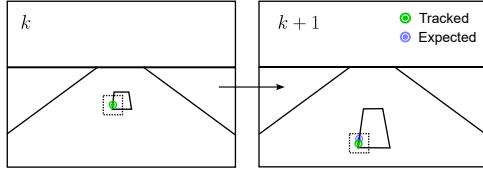
Fig. 5. Illustration of a failure case of optical flow tracking on the perspective image. The green scatter denotes the tracked feature, while the blue scatter is the expected accurate correspondence.
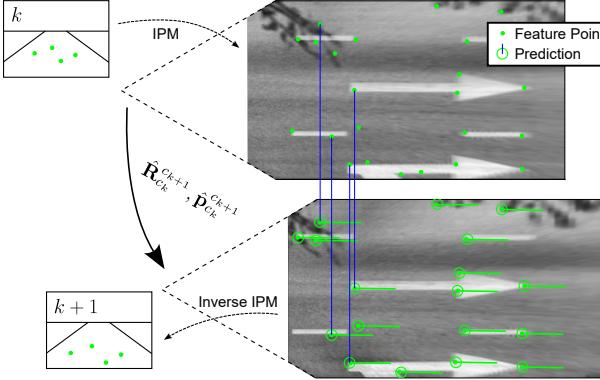

Fig. 6. Ground feature tracking on the BEV image.

After the feature processing on the BEV image, the obtained ground features are re-mapped to the perspective image through the inverse process of IPM, thus these features could be processed consistently with the common features. By doing so, the tracking on the BEV image helps improve the tracking precision without introducing systematic errors related to the C-G parameters. During the operation of VIO, the C-G parameters used for ground feature processing could be continuously updated.

Intuitively, to extract and track the ground features, it is needed to specifically identify the ground region in the image. This is not a hard task by applying deep learning-basd semantic segmentation [44] [45], which performs well in vehicular scenarios. Yet in the proposed method, the semantic segmentation is optional. The IPM processing itself could exclude most objects that are not on the ground surface, and the accurate feature prediction based on C-G parameters plus the RANSAC method could exclude outliers on the BEV image (e. g. vehicles, guardrails). Later in factor graph optimization, the influence of gross errors could be further mitigated through outlier detection methods. Therefore, although semantic segmentation could contribute to best performance of the system, it is not necessary.

In factor graph optimization, the extracted ground features are treated as a subset of visual features to construct the visual re-projection factors, while additional camera-ground constraints would be applied to them.

C. Optimization-based Visual-Inertial Odometry

Adhering to [8], we maintain a sliding-window factor graph to simultaneously estimate the navigation states, landmarks
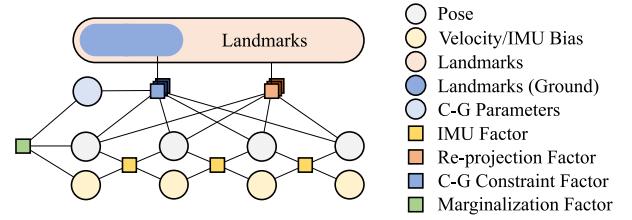

Fig. 7. The factor graph optimization of Ground-VIO.

and, additionally, the C-G parameters by optimizing different kinds of measurements.

The state vector of Ground-VIO is defined as follows

$$\mathcal{X} = (\mathbf{x}_0, \ \mathbf{x}_1, \ \cdots, \ \mathbf{x}_n, \ \mathbf{x}_\perp, \ \lambda_0, \ \lambda_1, \ \cdots, \ \lambda_m) \quad (9)$$

$$\mathbf{x}_k = \left(\mathbf{p}_{b_k}^w, \ \mathbf{q}_{b_k}^w, \ \mathbf{v}_{b_k}^w, \ \mathbf{b}_{a,b_k}, \ \mathbf{b}_{g,b_k} \right), \ k \in [0, \ n] \quad (10)$$

$$\mathbf{x}_\perp = (h, \ \theta, \ \alpha) \quad (11)$$

where $\mathbf{p}_{b_k}^w$, $\mathbf{q}_{b_k}^w$, $\mathbf{v}_{b_k}^w$ are the position, attitude and velocity of the k-th frame expressed in the world frame, $\mathbf{b}_{a,b_k}$ and $\mathbf{b}_{g,b_k}$ are the accelerometer bias vector and the gyroscope drift vector, $\lambda_0, \ \lambda_1, \ \cdots, \ \lambda_m$ are the inverse depths of the landmarks. Each landmark is anchored in the first observation frame within the sliding window.

The following factors are considered in the optimization:

1) IMU preintegration factor:

The IMU data between frames are preintegrated and used to construct the IMU preintegration factors. The residual could be expressed as

$$\mathbf{r}_{\text{IMU}} \left(\hat{\alpha}_{b_{k+1}}^{b_k}, \hat{\beta}_{b_{k+1}}^{b_k}, \hat{\gamma}_{b_{k+1}}^{b_k}, \mathbf{x}_k, \mathbf{x}_{k+1} \right) =$$
$$\begin{bmatrix} \mathbf{R}_{b_k}^{w \top} \left(\mathbf{p}_{b_{k+1}}^w - \mathbf{p}_{b_k}^w + \frac{1}{2}\mathbf{g}^w \Delta t_k^2 - \mathbf{v}_{b_k}^w \Delta t_k - \hat{\alpha}_{b_{k+1}}^{b_k} \right) \\ \mathbf{R}_{b_k}^{w \top} \left(\mathbf{v}_{b_{k+1}}^w + \mathbf{g}^w \Delta t_k - \mathbf{v}_{b_k}^w \right) - \hat{\beta}_{b_{k+1}}^{b_k} \\ 2 \left[\mathbf{q}_{b_k}^{w -1} \otimes \mathbf{q}_{b_{k+1}}^w \otimes \left(\hat{\gamma}_{b_{k+1}}^{b_k} \right)^{-1} \right]_{xyz} \\ \mathbf{b}_{a,b_{k+1}} - \mathbf{b}_{a,b_k} \\ \mathbf{b}_{g,b_{k+1}} - \mathbf{b}_{g,b_k} \end{bmatrix} \quad (12)$$

where k and $k+1$ are the epochs of adjacent frames, Δt_k is the time interval, $\hat{\alpha}_{b_{k+1}}^{b_k}$, $\hat{\beta}_{b_{k+1}}^{b_k}$, $\hat{\gamma}_{b_{k+1}}^{b_k}$ are the IMU preintegration terms [8].

The IMU measurements provide stable relative pose information based on the navigation state estimation, but they alone couldn't measure the absolute values of the translation and the velocity. When combined with visual measurements in VIO, metric-scale translation/velocity could be derived as long as the IMU is sufficiently excited [10].

2) Visual re-projection factor:

The visual features maintained in the sliding window, including the ground features, are used to construct the visual re-projection factors. The residual could be expressed as

$$\mathbf{r}_{\text{cam}} \left(\mathbf{u}_f^i, \mathbf{u}_f^j, \mathbf{x}_i, \mathbf{x}_j, \lambda_f \right) = \begin{bmatrix} (\mathbf{p}_f^{c_j})_x/(\mathbf{p}_f^{c_j})_z \\ (\mathbf{p}_f^{c_j})_y/(\mathbf{p}_f^{c_j})_z \end{bmatrix} - \mathbf{u}_f^j \quad (13)$$
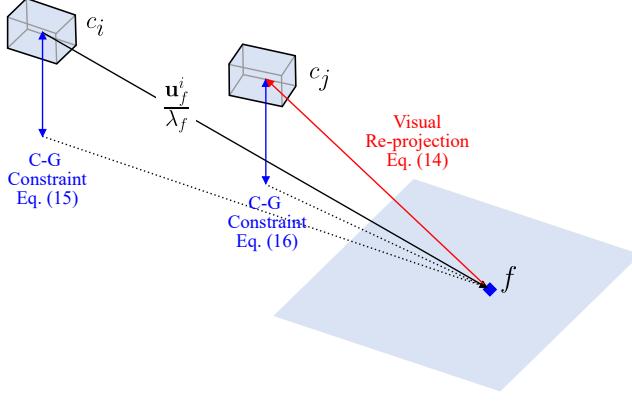
Fig. 8. Illustration of the visual re-projection factor and camera-ground constraint factors constructed upon a ground landmark f. Here, c_i is the anchor frame, while c_j is the target frame.

with

$$\mathbf{p}_f^{c_j} = \mathbf{R}_{b_j}^{w\top}\left(\mathbf{R}_{b_i}^{w}\left(\mathbf{R}_c^b\left(\frac{\mathbf{u}_f^i}{\lambda_f}\right) + \mathbf{p}_c^b\right) + \mathbf{p}_{b_i}^{w} - \mathbf{p}_{b_j}^{w}\right) \quad (14)$$

where $\mathbf{u}_f^i$ and $\mathbf{u}_f^j$ are visual observations of f at epoch i and j, $\left(\mathbf{R}_c^b,\ \mathbf{p}_c^b\right)$ are the IMU-camera extrinsic parameters.

The visual measurements are used to strongly constrain the vehicle poses and landmark positions through a bundle-adjustment (BA)-like model.

3) Camera-ground constraint factor:

Camera-ground constraints are applied to the ground features maintained in the sliding window, based on the model in Sect. III. In our implementation, there are two kinds of camera-ground constraint factors, depending on the anchor frame of the ground feature and the frame that the camera-ground constraint is applied, termed as the target frame.

If the anchor frame and the target frame are the same, the residual could be expressed as

$$r_{\text{C-G}}\left(\mathbf{u}_f^i, \lambda_f, \mathbf{x}_\perp\right) = h - \left(\mathbf{R}_{c_\perp}^c\left(\alpha,\theta\right)^\top \frac{\mathbf{u}_f^i}{\lambda_f}\right)_y \quad (15)$$

If the anchor frame and the target frame are different, the residual could be expressed as

$$r_{\text{C-G}}\left(\mathbf{u}_f^i, \lambda_f, \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_\perp\right) = h - \left(\mathbf{R}_{c_\perp}^c\left(\alpha,\theta\right)^\top \mathbf{p}_f^{c_j}\right)_y \quad (16)$$

where the i-th frame is the anchor frame, the j-th frame is the target frame, and $\mathbf{p}_f^{c_j}$ refers to (14).

By introducing the camera-ground geometric constraints into the estimator, the C-G parameters could be estimated and refined based on the information gained by VIO. Once the C-G parameters get converged, the constraints could reciprocally provide driftless, metric-scale geometric information to VIO. The mechanism of "make some parameters converge and use it to maintain the estimation performance" is similar to the IMU biases. Yet the converged C-G parameters are expected to have a more sustained influence, for: 1) the IMU biases are time-variant but the C-G parameters are relatively stable, 2) the C-G parameters are at the same order with pose estimation which don't need integration like IMU.

The optimization problem could then be expressed as minimizing above residuals and prior terms following

$$\min_{\mathcal{X}}\Big\{ \|\mathbf{r}_p - \mathbf{H}_p\mathcal{X}\|^2 +$$
$$\sum_{k\in[0,n]} \left\|\mathbf{r}_{\text{IMU}}\left(\hat{\alpha}_{b_{k+1}}^{b_k}, \hat{\beta}_{b_{k+1}}^{b_k}, \hat{\gamma}_{b_{k+1}}^{b_k}, \mathbf{x}_k, \mathbf{x}_{k+1}\right)\right\|_{\mathbf{P}_{\text{IMU}}}^2 +$$
$$\sum_{i<j\in[0,n],f\in\mathcal{F}} \rho_{\text{H}}(\left\|\mathbf{r}_{\text{cam}}\left(\mathbf{u}_f^i, \mathbf{u}_f^j, \mathbf{x}_i, \mathbf{x}_j, \lambda_f\right)\right\|_{\mathbf{P}_{\text{cam}}}^2) +$$
$$\sum_{i\in[0,n],f\in\mathcal{F}_\perp} \rho_{\text{C}}(\left\|r_{\text{C-G}}\left(\mathbf{u}_f^i, \lambda_f, \mathbf{x}_\perp\right)\right\|_{P_{\text{C-G}}}^2) +$$
$$\sum_{i<j\in[0,n],f\in\mathcal{F}_\perp} \rho_{\text{C}}(\left\|r_{\text{C-G}}\left(\mathbf{u}_f^i, \lambda_f, \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_\perp\right)\right\|_{P_{\text{C-G}}}^2)\Big\}$$
$$(17)$$

where $(\mathbf{r}_p, \mathbf{H}_p)$ is the prior information obtained from marginalization [8], $\mathcal{F}$ is the set of landmarks maintained in the sliding window, $\mathcal{F}_\perp$ is the set of ground landmarks, which is a subset of $\mathcal{F}$, $\rho_{\text{H}}(\cdot)$ and $\rho_{\text{C}}(\cdot)$ are Huber and Cauchy kernel functions [46] respectively, $\mathbf{P}_{\text{IMU}}, \mathbf{P}_{\text{cam}}, P_{\text{C-G}}$ are covariances/variances of the residuals. The ceres-solver [46] is employed to solve the optimization problem.

D. Initialization of Camera-Ground Parameters

If the C-G parameters are completely unknown at the beginning, the ground feature processing module couldn't work properly and it is hard to construct accurate camera-ground constraint factors. In this case, the system needs to online initialize the C-G parameters.

It is recognized that the monocular VIO has the capability to perceive metric-scale environmental structure with enough IMU excitation [10]. On this basis, it is completely possible to online initialize the C-G parameters without auxiliary information from other sensors. The specific procedure of the initialization is presented as follows.

After the VIO initialization, the common VIO routines start to work. So far, without the knowledge of C-G parameters, the ground features could only be tracked on the perspective image (without IPM processing). To achieve this, a conservative region of interest (ROI) on the image is used, which is determined by the IMU-camera extrinsics and a rough vehicle height. For the initialization of C-G parameters, the uncertainties (i. e., variances) of the ground landmarks in the sliding window are periodically checked. Once enough ground landmarks below the uncertainty threshold are obtained, observations of these landmarks are stacked together to estimate the C-G parameters, following

$$\left(\hat{h}, \hat{\theta}, \hat{\alpha}\right) = \underset{(h,\theta,\alpha)}{\arg\min}$$
$$\sum_{i\leq j\in[0,n],f\in F_\perp} \left\|h - \left[\mathbf{R}_{c_\perp}^c\left(\alpha,\theta\right)^\top \left(\mathbf{R}_{c_i}^{c_j}\frac{\mathbf{u}_f^{c_i}}{\lambda_f} + \mathbf{p}_{c_i}^{c_j}\right)\right]_y\right\|$$
$$(18)$$

After the initialization of C-G parameters, the ground feature processing module would be switched on for better tracking accuracy. At the same time, the camera-ground constraint

Fig. 9. Two typical cases of complex road conditions, (a) attitude vibration caused by road irregularity and vehicle dynamics. (b) change of the road slope.
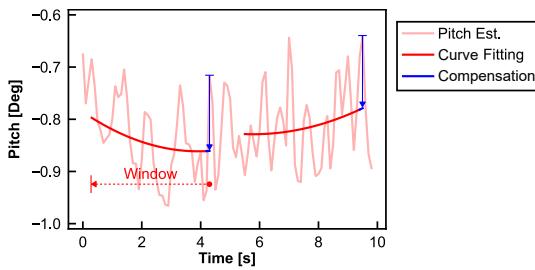

Fig. 10. An example of IMU-based pitch compensation when applying the camera-ground geometric constraint in realistic scenarios.

factors would be applied in the factor graph to enhance VIO, and the C-G parameters will be further refined.

E. Dealing with Complex Road Conditions

In real-world scenarios, the road conditions could be relatively complex and don't conform to the ideal camera-ground geometric model depicted in Sect. III. Such conditions could be mainly covered by the following two cases: 1) attitude vibration of the vehicle caused by road irregularity or vehicle dynamics, 2) change of the road slope. These two cases are illustrated in Fig. 9, which would lead to systematic errors and affect the system performance if not carefully considered.

In the proposed system, several tricks are employed to mitigate the effect of these problems. To deal with high-frequency attitude vibration of the vehicle (Fig. 9(a)), we use the local IMU attitude estimation to compensate the C-G parameters temporarily, as shown in Fig. 10. In our implementation, only the pitch component θ is compensated, which is more sensitive considering the ground region of interest ($\pm$ 3 m wide, 15 m far). To be specific, we use a 4-second window of historical pitch estimation to fit a quadratic curve and to calculate the pitch compensation of the current epoch, following

$$\theta_{comp} = \theta_{b_k}^w - \hat{\theta}_{b_k}^w \qquad (19)$$

where $\theta_{b_k}^w$ is the current IMU pitch estimation, $\hat{\theta}_{b_k}^w$ is the pitch predicted by curve fitting. And when applying the camera-ground constraint factors in this frame, the C-G parameter θ is compensated temporarily

$$\theta_k = \theta + \theta_{comp} \qquad (20)$$

where θ_k is the taken as the temporary C-G parameter at epoch k. By doing so, the noise of the camera-ground constraint caused by attitude vibration could be significantly mitigated.

To deal with the change of the road slope (Fig. 9(b)), firstly the ground feature processing could abandon some of the feature observations that don't conform to the planar ground assumption. Secondly, when constructing the camera-ground

Fig. 11. The vehicle actor and the 3-D environment applied in the CARLA simulator.

Fig. 12. Simulation trajectories and example images. The colorbar indicates the vehicle speed (m/s). Top: Seq. S-A, which uses the "Town10" world in CARLA simulator. The road texture is distinct and environmental features are rich in this world. Bottom: Seq. S-B, which uses the "Town06" world in CARLA simulator. Fewer distinct ground features, mainly the road markings, could be observed in this world.

factors, we use a relatively strict outlier removal strategy, with a cut-off threshold plus a Cauchy kernel function, in order to counter gross errors caused by drastic slope changes.

V. SIMULATION TESTS

Simulation tests are conducted to evaluate the system performance in relatively ideal conditions. The advantage of simulation is that the vehicle-sensor alignment and the environmental geometry are precisely known, which facilitates more in-depth analysis. The CARLA simulator [47], which provides exquisite 3-D scenes and realistic vehicle dynamics, is used to generate the vehicle poses and images. The IMU data is separately simulated based on B-spline fitting [25] of the 10 Hz ground truth poses, where custom biases and noises are added. The settings of the simulation are listed in TABLE I.

TABLE I
SENSOR SETTINGS OF THE SIMULATION TESTS.

Simulation settings	
Image resolution	1024×768
Field of view (FOV)	$60°$
Image frequency	10
IMU frequency	100
Velocity random walk (VRW)	$0.12\ (m/s/\sqrt{hr})$
Angle random walk (ARW)	$0.5\ (°/\sqrt{hr})$
Accelerometer bias[1]	$(1000, 1000, 1000)\ (mGal)$
Gyroscope drift[1]	$(100, 100, 100)\ (°/hr)$

[1] For simplicity, only constant biases are considered.

The vehicle trajectories and the captured images in the simulation tests are shown in Fig. 12. The simulation consists of two sequences, namely S-A and S-B, corresponding to urban and highway environments respectively. The vehicle
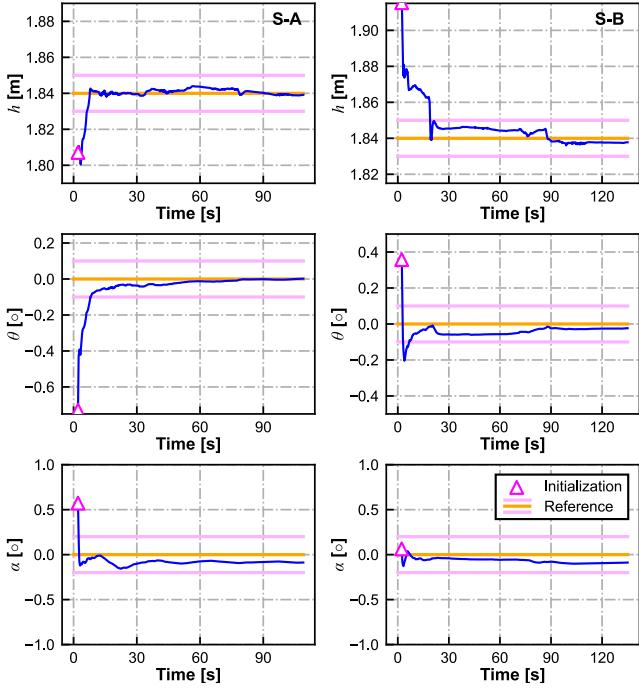
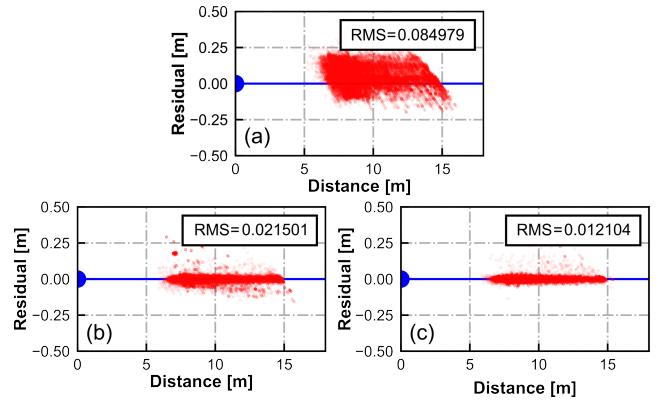Fig. 13. Convergence of the C-G parameters during online estimation (left: Seq. S-A, right: Seq. S-B).

Fig. 14. Residuals of camera-ground constraints (Eq. (15)) applying ground-truth C-G parameters in Seq. S-A. The residuals are calculated under different VIO schemes: (a) VINS-Fusion (Mono) without camera-ground constraints. (b) Ground-VIO without any compensation. (c) Ground-VIO with IMU pitch compensation. The X-axis indicates the perception distance of landmarks in the anchor frame. The root mean square (RMS) values are attached besides.
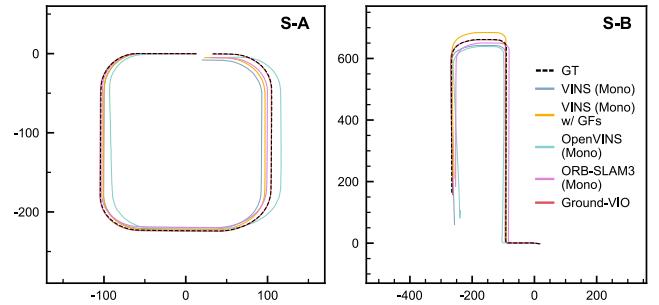

Fig. 15. Estimated vehicle trajectories of different VIO solutions in the simulation tests (left: Seq. S-A, right: Seq. S-B).

dynamics caused by the suspension system are considered, leading to up to $\pm$ 0.5° vibration of the vehicle attitude.

Different schemes of VIO are tested on the simulated data sequences, including: 1) VINS-Fusion (monocular), 2) VINS-Fusion (stereo), 3) VINS-Fusion (monocular) with ground features, 4) OpenVINS (monocular), 5) ORB-SLAM3 (monocular, with IMU) and 6) the proposed Ground-VIO. For VINS-Fusion (monocular) with ground features, the ground feature processing module is employed, in which ground-truth C-G parameters are applied for comparison. For Ground-VIO, the C-G parameters are unknown and would be estimated online.

For VINS-Fusion-based solutions and Ground-VIO, a 50-ms maximum optimization time limit (single thread, Intel i7-6700K) is set to guarantee real-time processing and provide a more equitable comparison. The maximum feature number of front-end common feature processing is 250, while the maximum number of ground features is set to 40. For an ideal analysis, the semantic images are used to determine the ground region in the simulation tests.

Firstly. the focus is put on the estimation of the C-G parameters. The convergence of the C-G parameters on Seq. S-A and Seq. S-B is shown in Fig. 13. In the two sequences, the initialization of the C-G parameters could be completed within 10 seconds with <0.1 m error of h and <1° error of θ and α, as long as enough geometric information is derived by the VIO system. After the initialization, the camera-ground geometric constraints are enabled in the factor graph, and the C-G parameters go on to be refined. It could be found that, with moderate vehicle dynamics and ideal planar grounds in the simulation, the C-G parameters could get converged in a very short time (10 secs for Seq. S-A and 20 secs for Seq. S-B) and achieve good accuracy (0.01 m for h, 0.1° for θ

and α). Relatively speaking, the convergence performance for Seq. S-A is better, which could be attributed to more available ground features. It is noted that the estimation accuracy of α (roll) is worse than θ (pitch), which is reasonable considering the region of interest (15 m far and $\pm$ 3 m wide) and the fact that only the pitch vibration is compensated (Sect. IV-E).

Secondly, we check the consistency between the landmark depths estimated by VIO and the ground-truth C-G parameters to analyze the model accuracy. To be specific, the residuals of single-frame camera-ground geometric constraints (15) are calculated using ground-truth C-G parameters and estimated landmark depths. As shown in Fig. 14, if the constraints are not applied (VINS-Fusion), the estimated landmark depths don't fit the camera-ground geometry well. The distribution of the residuals reflect the error of scale estimation, which could be over the level of $0.1/h \approx 5\%$. Once the camera-ground geometric constraints are taken into account (Ground-VIO), the residuals are consequently kept to around 0, which indicate an unbiased estimation of the scale. Furthermore, with the local compensation of the vehicle pitch, the noise level of the residuals is significantly lowered. This indicates better accuracy of the compensated model, as it compensates much of the model errors caused by vehicle dynamics.

Finally, the pose estimation accuracy of different solutions are investigated. The estimated vehicle trajectories are shown
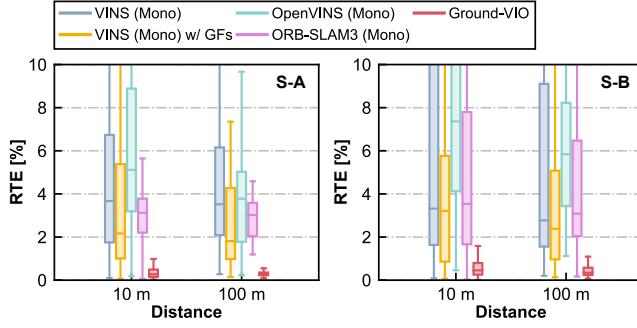
Fig. 16. Short-term relative translation errors (calculated with 10/100 m intervals) of different VIO solutions in the simulation tests (left: Seq. S-A, right: Seq. S-B).

in Fig. 15, and the distributions of the relative translation errors are shown in Fig. 16. Considering that different VIO solutions need different time (several seconds) to initialize the system, we align the estimated vehicle poses at $t = 10$ s when plotting the trajectories in Fig. 15. It could be found that, the attitude estimations of different solutions are comparable, yet almost all the monocular VIO solutions without C-G constraints show significant translation errors. As the scale observability is closely related to the dynamics in a monocular VIO, the long-time straight motions lead to inevitable scale drifting. The solution of VINS-Fusion (monocular) with ground features slightly improves the translation accuracy by introducing more stable features, but still suffers significant drift of the scale. For Ground-VIO, after the C-G parameters get converged in the beginning dynamic period (with acceleration and rotation), the camera-ground geometry could then provide unbiased information of the metric scale and helps VIO maintain accurate translation estimation. Consequently, the monocular Ground-VIO achieves superior translation estimation performance (relative error $< 0.5\%$) without introducing any other sensors, which is incredible considering the insufficient dynamics of a ground vehicle in the road environment.

The statistics of the navigation performance of different VIO solutions are listed in TABLE II. The relative translation and rotation errors are calculated by averaging all possible subsequences of length (100, ..., 800) meters, referring to [1]. The absolute trajectory error is calculated referring to [39].

TABLE II
POSE ESTIMATION ERRORS IN THE SIMULATION TESTS (S-A AND S-B, ONLINE CALIBRATION).

Method	Seq. S-A			Seq. S-B		
	t_{rel} (%)	r_{rel}[1]	t_{abs} (m)	t_{rel} (%)	r_{rel}[1]	t_{abs} (m)
VINS-Fusion (Mono)	3.81	0.14	7.78	5.05	0.21	22.25
VINS-Fusion (Mono) \w GFs	2.60	0.13	5.58	2.98	0.15	12.34
OpenVINS (Mono)	2.73	0.68	4.22	5.30	0.71	19.45
ORB-SLAM3 (Mono)	2.27	0.11	3.88	3.09	0.14	9.40
Ground-VIO	**0.24**	**0.10**	**0.36**	**0.38**	**0.11**	**1.43**

[1] Unit: $°/100\ m$.

Fig. 17. Appearance of the experimental vehicle.

VI. REAL-WORLD EXPERIMENTS

Real-world experiments were conducted on Oct. 12, 2022 to evaluate the performance of the proposed system under typical vehicular scenarios, including urban roads and highways. The appearance of the experimental vehicle is shown in Fig. 17. The experimental platform is equipped with two Flir BFS-PGE-31S4C cameras, a low-cost ADIS16470 MEMS IMU, a tactical grade XW-GI7660 IMU and a Septentrio AsteRx4 GNSS receiver. The data from the tactical grade IMU and the GNSS receiver (with base station availability) are post-processed to generate the reference trajectory. The specifications of the used IMUs are listed in TABLE III.

TABLE III
SPECIFICATIONS OF THE USED IMUS.

IMU	Noise Density		Bias Stability	
	Gyro. ($°/\sqrt{hr}$)	Accel. ($m/s/\sqrt{hr}$)	Gyro. ($°/hr$)	Accel. ($mGal$)
ADIS16470	0.34	0.037	8	1300
XW-GI7660	-	-	0.3	100

The reason to use self-collected datasets is for better representativeness of the evaluation, i. e., using low-cost visual-inertial sensor scheme under realistic vehicular scenarios, and especially focusing on feature-lacking highway scenarios with limited dynamics. For the visibility of our work, we will make the experimental data public available.

Notice that in the real-world experiments, we don't apply a semantic segmentation module but rely on the system itself to distinguish the ground features and resist possible outliers.

A. VIO with Unknown C-G Parameters

In this part, we evaluate the proposed system under the condition that the C-G parameters are completely unknown. In this case, the C-G parameters would be initialized online and continuously estimated during the data periods. Four 180-sec data sequences with moderate vehicle dynamics, namely Seq. R-A, R-B, R-C and R-D, are used for the evaluation, as shown in Fig. 18. Different solutions of VIO are tested on the data sequences. Compared to the simulation test, state-of-art VIO implementations with stereo camera setups are considered in this part to investigate the best achievable VIO performance in these real-world road environments.

The convergence of the C-G parameters is shown in Fig. 19. For the four sequences, the final estimation results of the C-G parameters show good consistency. Later in this part, the
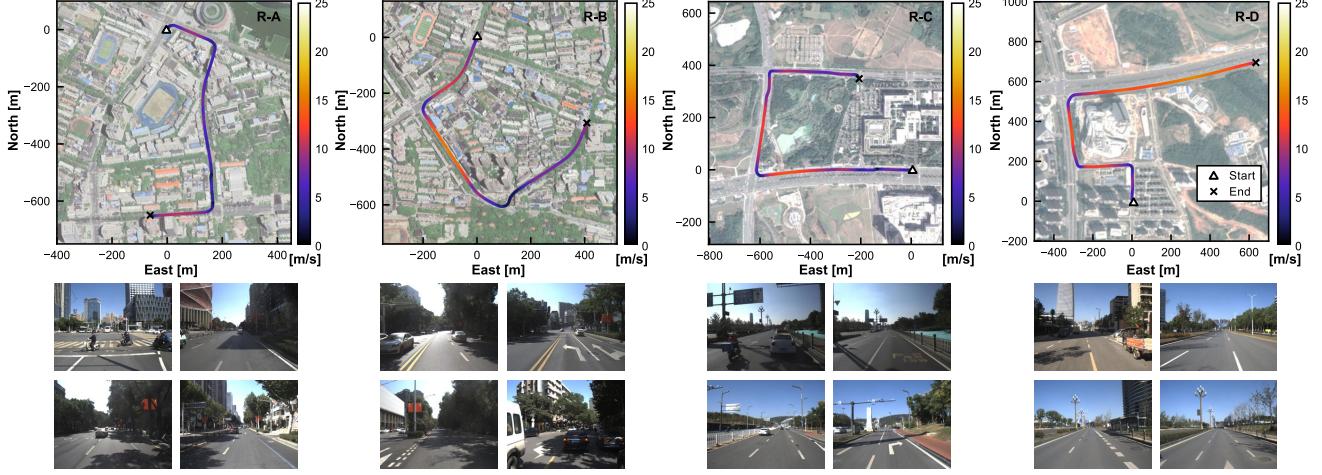
Fig. 18. Vehicle trajectories and example images in Seq. R-A, R-B, R-C and R-D. The colorbar indicates the vehicle speed (m/s). Seq. R-A and Seq. R-B are under the urban scenario, with narrow roads, abundant environmental textures and relatively low speed. Seq. R-C and Seq. R-B are mostly on the highway, with broad roads, fewer buildings and moderate speed.
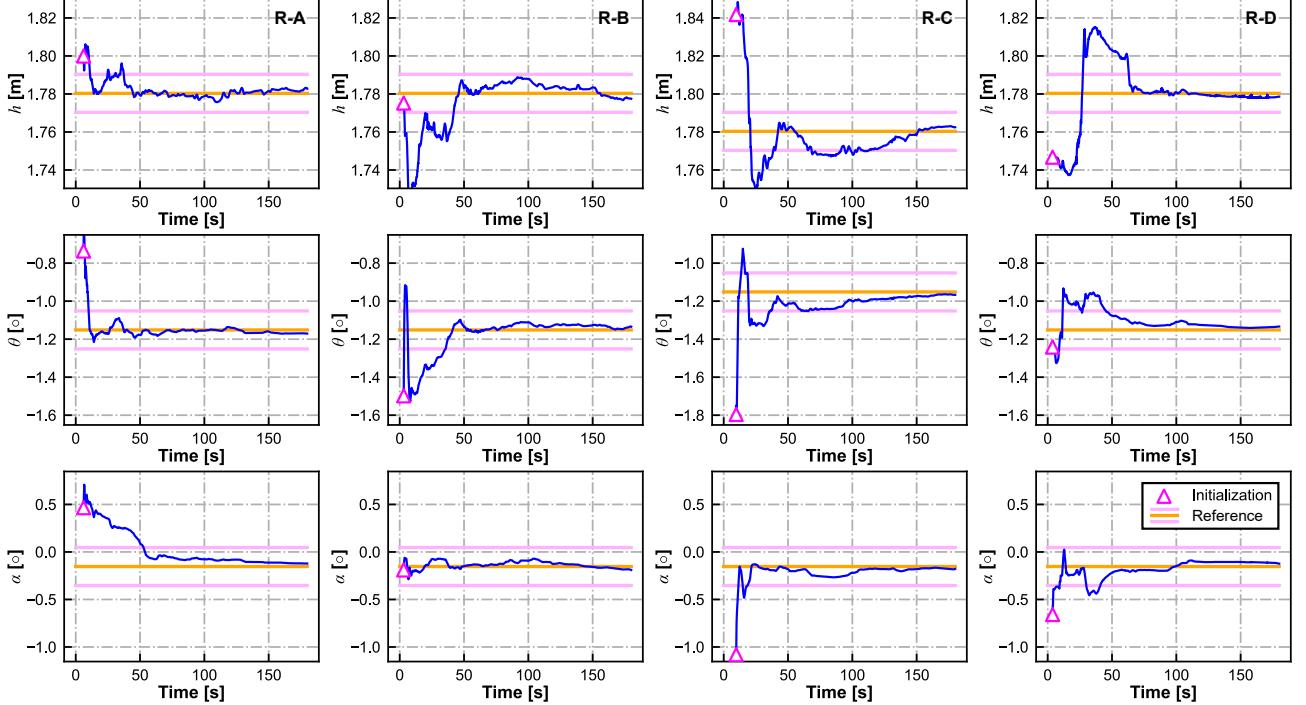

Fig. 19. Convergence of the C-G parameters during online estimation in Seq. R-A, R-B, R-C and R-D.

average value of the four sets of obtained C-G parameters is taken as the reference, which is (1.7803 m, -1.151°, -0.153°) for (h, θ, α). It is found from Fig. 19 that, the initialization of the parameters could be finished in a few seconds, and the initial accuracy is similar to the simulation tests (0.1 m for h, 1° for θ, α). Yet differently, the convergence of the C-G parameters is slower than the simulation tests. To be specific, around 30~60 seconds are needed to obtain ideal accuracy of the C-G parameters (0.01 m, 0.1°, 0.2° for h, θ, α). This could be attributed to more complex road conditions and smaller IMU excitation in the real-world experiments, which affect both the camera-ground geometric constraint and the monocular VIO itself. Roughly speaking, better observability

of the VIO system, sufficient ground features and smooth road surface could contribute to faster convergence of the C-G parameters.

The focus is then put on the pose estimation performance. Fig. 20 and Fig. 21 show the estimated vehicle trajectories and relative translation errors of different VIO schemes. The detailed statistics of the VIO performance are listed in TABLE IV. Similar to the simulation tests, the monocular VIOs (except Ground-VIO) obtain good attitude estimation, but show bad performance on the translation error due to the drift of the scale, which is significant during long-time straight motions. Comparatively, the stereo VIOs perform much better on the relative translation errors, but they still undergo significant
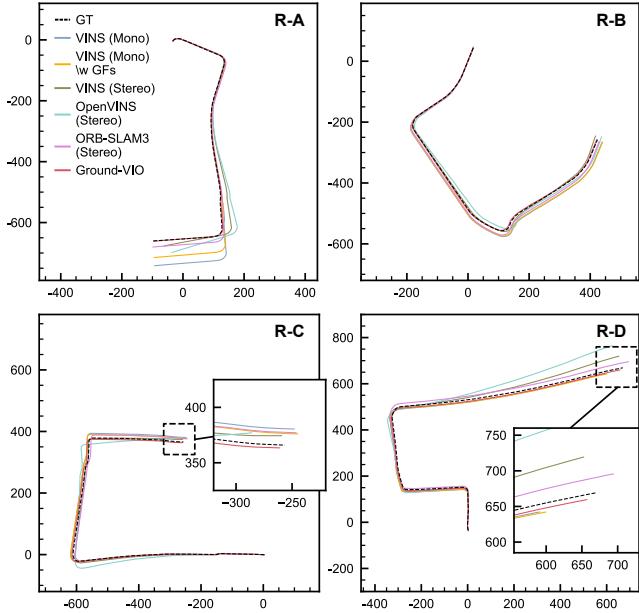
Fig. 20. Estimated vehicle trajectories of different VIO solutions in the simulation tests (left: Seq. R-A, right: Seq. R-B).
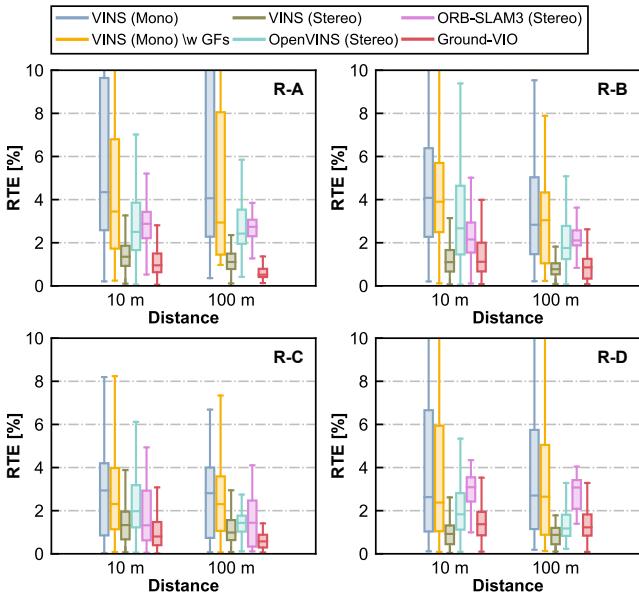

Fig. 21. Short-term relative translation errors (calculated with 10/100 m intervals) of different VIO solutions in the simulation tests (left: Seq. R-A, right: Seq. R-B).

pose errors as could be seen in Fig. 20. To be specific, the filter-based OpenVINS (stereo) undergoes relatively large heading errors on the four sequences, while the optimization-based VINS-Fusion (stereo) performs bad on Seq. R-A and R-D. The phenomenon that monocular VIO could outperform stereo VIO on attitude estimation could also be found in [28]. The ORB-SLAM3 (stereo, with IMU) scheme, although maintaining good heading estimation, undergoes non-negligible position drifting. The good attitude estimation performance could be attributed to the map-centered design of ORB-SLAM3, whose superiority is verified in [29]. However, the

road environment, with insufficient stable features and moving objects, has caused difficulty for it to achieve ideal translation estimation.

In contrast, the proposed Ground-VIO shows good translation estimation performance with the help of the camera-ground geometric constraints. Although the C-G parameters are unknown at the beginning, the vehicle dynamics are able to make them converge and continuously take effect in the remaining period. It is verified that, the camera-ground geometry, like stereo vision but in a different way, could help maintain precise and unbiased scale estimation in realistic vehicular scenarios. Generally, the Ground-VIO could achieve comparable relative translation error (0.5%∼1.0%) with state-of-art stereo VIO schemes, and the attitude estimation performance is even better. Thus, the Ground-VIO achieves the smallest position drift on almost all the four sequences.

In all, with moderate vehicle dynamics, the proposed Ground-VIO is able to online calibrate the C-G parameters and obtain good pose estimation accuracy simultaneously.

B. Pre-Calibrated VIO under Challenging Scenarios

It has been mentioned that, the online calibration of the C-G parameters relys on the vehicle dynamics, since it needs the observability of VIO to extract metric-scale environmental structure. Fortunately, with pre-calibration of the C-G parameters, the VIO performance could also be greatly improved even under dynamic-insufficient scenarios. Actually, the pre-calibration is not hard, since it could be automatically finished when moderate vehicle dynamics are available, as verified in Sect. VI-A.

In this section, two highway data sequences, namely Seq. R-E and Seq. R-F, are used to test the system performance with pre-calibrated C-G parameters. The vehicle trajectories and the representative images are shown in Fig. 18. These two data sequences are extremely challenging for VIO, with limited dynamics, insufficient environmental features and high vehicle speed. These conditions could cause difficulty in both feature tracking and the observability of the VIO system. For Ground-VIO, the pre-calibrated C-G parameters are obtained from the online estimation results in Sect. VI-A.

The estimated vehicle trajectories and relative translation error distributions are shown in Fig. 24 and Fig. 25. Despite our best efforts, some schemes can't work properly on the two sequences. To be specific, the OpenVINS (stereo) scheme can't successfully initialize on both sequences and the ORB-SLAM3 (stereo) scheme fails on Seq. R-F because of the difficulty in ORB feature matching, as the environmental textures are either weak or highly repetitive (e.g. building windows, guardrails).

The pose estimation results are presented in Fig. 24 and Fig. 25. As shown in Fig. 24, the monocular VIOs perform bad on the translation estimation on Seq. R-E, reaching a relative error over 10%. It is unexpected that state-of-art stereo VIO schemes are also unable to achieve good pose estimation on the sequences, despite the fact that the stereo vision could provide accurate scale information in principle. This could be mainly due to the lack of high-quality visual features in the highway environment, and the stereo matching even

TABLE IV
POSE ESTIMATION PERFORMANCE ON R-A, R-B, R-C AND R-D SEQUENCES (ONLINE CALIBRATION).

Method	Seq. R-A			Seq. R-B			Seq. R-C			Seq. R-D		
	t_{rel} (%)	r_{rel} (°/100 m)	t_{abs} (m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{abs} (m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{abs} (m)	t_{rel} (%)	r_{rel} (°/100 m)	t_{abs} (m)
VINS-Fusion (Mono)	10.4	0.17	33.9	3.06	0.13	9.15	2.17	**0.08**	7.12	2.96	0.09	18.0
VINS-Fusion (Mono) \w GFs	6.84	0.16	22.3	2.99	**0.11**	8.81	2.09	0.10	6.29	2.62	0.09	16.1
VINS-Fusion (Stereo)	2.03	0.80	4.44	1.15	0.40	3.74	0.88	0.19	3.22	**1.08**	0.26	8.01
OpenVINS (Stereo)	5.11	2.07	10.6	2.37	0.69	3.92	2.11	0.65	7.11	2.05	0.72	12.7
ORB-SLAM3 (Stereo)	2.45	0.22	6.74	2.19	0.24	5.80	1.36	0.15	4.52	2.58	0.10	12.0
Ground-VIO	**0.42**	**0.14**	**0.67**	**0.72**	**0.11**	**1.30**	**0.48**	0.09	**1.28**	**1.08**	**0.08**	**5.15**

Fig. 22. Vehicle trajectories and example images in Seq. R-E. The data sequence is on the highway with a relatively high speed. Nearby buildings, trees and traffic signs provide a limited number of visual features.
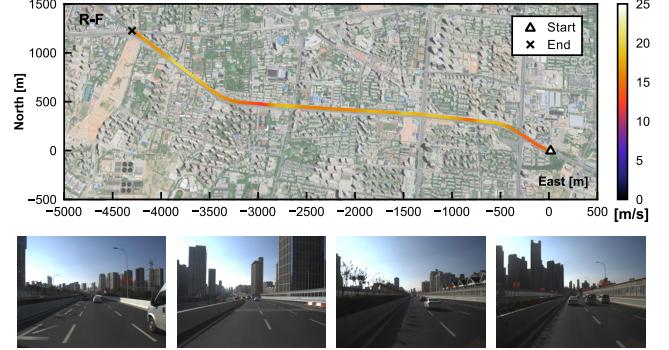

Fig. 23. Vehicle trajectories and example images in Seq. R-F. The data sequence is on the highway with moderate/high speed. The buildings in the view are faraway and the textures of the surrounding objects (buildings, guardrails and bushes) are mostly repetitive.

increases the risk of introducing gross errors. In contrast, with the pre-calibrated C-G parameters, the proposed Ground-VIO achieves an incredible 1% relative translation error (average). The camera-ground geometry not only provides unbiased scale information to the monocular VIO system, but also provides stable features via the specially designed ground feature processing module. In the highway environment, most visual features are faraway which can't provide accurate translation information. The ground feature processing module makes it possible to fully utilize the landmarks on the road (e. g. road markings, shadows) to mitigate the ill-conditioning, and what makes sense is that these landmarks depths are almost known. As a result, the estimation of vehicle velocity is effectively constrained and contributes to accurate pose estimation.

Similar results could be found in the more challenging Seq. R-F. As shown in Fig. 25, the Ground-VIO greatly outperforms state-of-art monocular and stereo VIO schemes, achieving 1% relative translation error (average).

The statistics of the pose estimation error on Seq. R-E and R-F are listed in TABLE V.

C. IPM Calibration Performance

To some extent, the online estimation of C-G parameters is equivalent to online calibrating IPM, which is widely used in vehicle perception. In this part, apart from the odometry performance, the effectiveness of the online IPM calibration is investigated qualitatively.

To be specific, the estimated C-G parameters in Sect. VI-A are used for IPM processing of an image sequence. Through
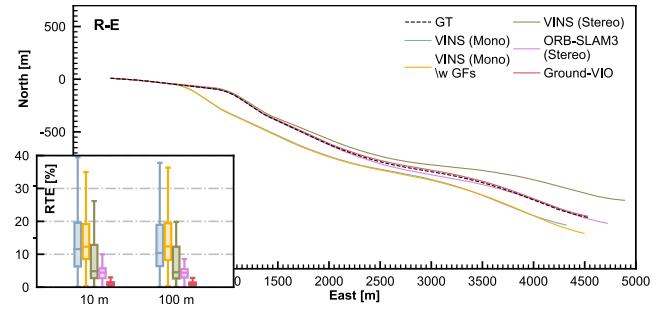

Fig. 24. Estimated vehicle trajectories and relative translation errors of different VIO solutions on Seq. R-E.
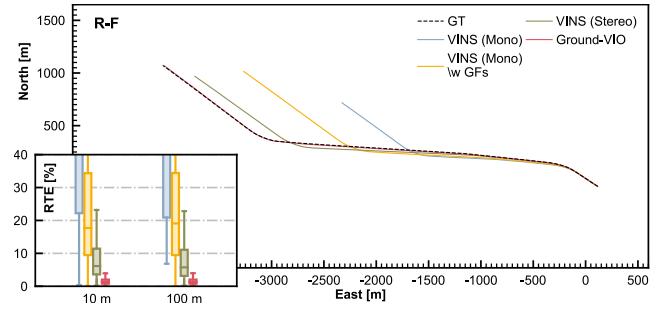

Fig. 25. Estimated vehicle trajectories and relative translation errors of different VIO solutions on Seq. R-F.

IPM, the images are transformed into metric-scale point clouds with color information. Based on the camera poses obtained by Ground-VIO, the point clouds could be merged together in

TABLE V
POSE ESTIMATION PERFORMANCE ON R-E AND R-F DATA SEQUENCES
(PRE-CALIBRATED).

Method	Seq. S-A			Seq. S-B		
	t_{rel} (%)	r_{rel}[1]	t_{abs} (m)	t_{rel} (%)	r_{rel}[1]	t_{abs} (m)
VINS-Fusion (Mono)	15.3	0.12	*	41.7	0.16	*
VINS-Fusion (Mono) \w GFs	17.2	0.12	*	22.4	0.11	*
VINS-Fusion (Stereo)	7.2	0.22	*	8.27	0.15	*
ORB-SLAM3 (Stereo)	4.25	0.13	67.6	-	-	-
Ground-VIO	**0.85**	**0.05**	**6.96**	**1.21**	**0.06**	**5.72**

[1] Unit: $°/100\ m$. * Greater than $100\ m$.

a reference frame. The consistency of the merged point could directly reflects the accuracy of IPM.

Fig. 26 shows the merged point clouds based on different C-G parameters. It could be seen that, with residual error on either attitude or height component of the C-G parameters, the merged point cloud is fuzzy and has stitching errors. This reflects the inconsistency of multiple point clouds, which further indicates that the generated IPM point clouds are not geometrically accurate. Roughly speaking, the longitudinal errors could reach over meter-level. In contrast, with calibrated C-G parameters, the merged point cloud is much more consistent and less fuzzy. Furthermore, once the IMU pitch compensation is applied, the details of the point cloud become clearer, which verifies the accuracy of the camera-ground geometric model. After the IPM calibration, a 10~15 meter effective perception range, with decimeter to centimeter level accuracy, could be expected.

In all, the proposed algorithm provides an approach to online calibrate the IPM parameters of vehicle-mounted cameras, which is based on only monocular visual-inertial data without the need of extra infrastructure.

VII. CONCLUSION

In this work, we presented Ground-VIO, which introduces the camera-ground geometry into monocular VIO to improve the odometry performance. The proposed works well with either unknown or pre-calibrated C-G parameters, achieving comparable or even better odometry accuracy than state-of-art stereo VIOs in vehicular scenarios. The method is expected to significantly improve the practicability of VIO applied in intelligent vehicle applications, which could work as an effective supplement to existing vehicle navigation schemes.

Besides, the proposed method provides an efficient way for online IPM calibration based on only monocular visual-inertial data. The auto-calibration doesn't need extra infrastructure and could handle long-term change of the sensor alignment, which is meaningful for better vehicle perception. It is our future interest to apply this technique in vision-based crowd-sourced mapping applications.
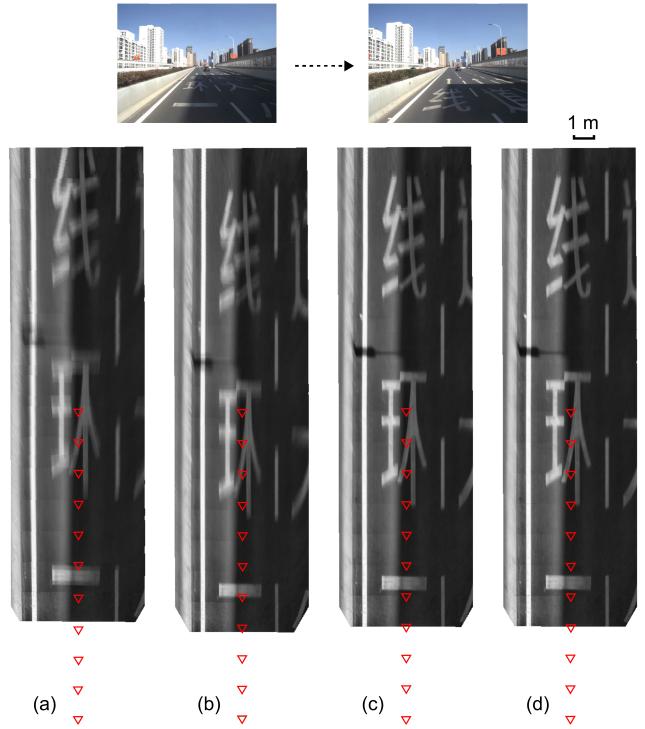
ACKNOWLEDGMENTS

Fig. 26. Merged IPM point clouds obtained from 11 successive images, based on different C-G parameters: (a) With $1°$ error of θ. (b) With 0.1 m error of h. (c) Calibrated C-G parameters. (d) Calibrated C-G parameters with IMU pitch compensation. The red triangles indicate the camera poses.

REFERENCES

[1] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3354-3361.

[2] J. Levinson et al., "Towards fully autonomous driving: Systems and algorithms," 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 2011, pp. 163-168.

[3] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler and R. Urtasun, "Monocular 3D Object Detection for Autonomous Driving," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2147-2156.

[4] K. Muhammad et al., "Vision-Based Semantic Segmentation in Scene Understanding for Autonomous Driving: Recent Achievements, Challenges, and Outlooks," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 22694-22715, Dec. 2022.

[5] C. Cadena et al., "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," in IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309-1332, Dec. 2016.

[6] H. Lategahn, A. Geiger and B. Kitt, "Visual SLAM for autonomous ground vehicles," 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 2011, pp. 1732-1737.

[7] G. Huang, "Visual-Inertial Navigation: A Concise Review," 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 2019, pp. 9572-9582.

[8] T. Qin, P. Li and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," in IEEE Transactions on Robotics, vol. 34, no. 4, pp. 1004-1020, Aug. 2018.

[9] K. Sun et al., "Robust Stereo Visual Inertial Odometry for Fast Autonomous Flight," IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 965-972, Apr. 2018.

[10] A. Martinelli, "Visual-inertial structure from motion: Observability and resolvability," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., pp. 4235-4242, Nov. 2013.

[11] J. Hernandez, K. Tsotsos and S. Soatto, "Observability identifiability and sensitivity of vision-aided inertial navigation," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 2319-2325, May 2015.

[12] Y. Yang and G. Huang, "Observability Analysis of Aided INS With Heterogeneous Features of Points, Lines, and Planes," in IEEE Transactions on Robotics, vol. 35, no. 6, pp. 1399-1418, Dec. 2019.

[13] W. Lee, Y. Yang and G. Huang, "Efficient Multi-sensor Aided Inertial Navigation with Online Calibration," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 5706-5712.

[14] J. H. Jung et al., "Monocular Visual-Inertial-Wheel Odometry Using Low-Grade IMU in Urban Areas," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 2, pp. 925-938, Feb. 2022.

[15] 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," Information Fusion, vol. 90, 2023, pp. 218-232.

[16] J. Jeong and A. Kim, "Adaptive Inverse Perspective Mapping for lane map generation with SLAM," 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Xi'an, China, 2016, pp. 38-41.

[17] J. Wang, T. Mei, B. Kong and H. Wei, "An approach of lane detection based on Inverse Perspective Mapping," 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, 2014, pp. 35-38.

[18] Y. -L. Chang, L. -Y. Hsu and O. T. . -C. Chen, "Auto-Calibration Around-View Monitoring System," 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, USA, 2013, pp. 1-5.

[19] J. Jeong, Y. Cho and A. Kim, "Road-SLAM : Road marking based SLAM with lane-level accuracy," 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 2017, pp. 1736-1473.

[20] Y. Zhou, X. Li, S. Li and X. Wang, "Visual Mapping and Localization System Based on Compact Instance-Level Road Markings With Spatial Uncertainty," in IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 10802-10809, Oct. 2022.

[21] N. Gosala and A. Valada, "Bird's-Eye-View Panoptic Segmentation Using Monocular Frontal View Images," in IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1968-1975, April 2022.

[22] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," Proceedings 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, 2007, pp. 3565-3572.

[23] M. Li and A. I. Mourikis, "Improving the accuracy of EKF-based visual-inertial odometry," 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 2012, pp. 828-835.

[24] K. Eckenhoff, P. Geneva and G. Huang, "MIMC-VINS: A Versatile and Resilient Multi-IMU Multi-Camera Visual-Inertial Navigation System," in IEEE Transactions on Robotics, vol. 37, no. 5, pp. 1360-1380, Oct. 2021.

[25] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang and G. Huang, "OpenVINS: A Research Platform for Visual-Inertial Estimation," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 4666-4672.

[26] C. Forster, L. Carlone, F. Dellaert and D. Scaramuzza, "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry," in IEEE Transactions on Robotics, vol. 33, no. 1, pp. 1-21, Feb. 2017.

[27] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization", International Journal of Robotics Research (IJRR), 2014.

[28] Tong Qin, Jie Pan, Shaozu Cao, Shaojie Shen, "A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors", arXiv:1901.03638 [cs.CV], Jan. 2019.

[29] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial, and Multimap SLAM," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, Dec. 2021.

[30] P. Zhou, Y. Liu, P. Gu, J. Liu and Z. Meng, "Visual Localization and Mapping Leveraging the Constraints of Local Ground Manifolds," in IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 4196-4203, April 2022.

[31] K. Konolige, G. Grisetti, R. Kümmerle, W. Burgard, B. Limketkai and R. Vincent, "Efficient Sparse Pose Adjustment for 2D mapping," 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 2010, pp. 22-29.

[32] D. Scaramuzza, F. Fraundorfer and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 2009, pp. 4293-4299.

[33] F. Zheng and Y. -H. Liu, "SE(2)-Constrained Visual Inertial Fusion for Ground Vehicles," in IEEE Sensors Journal, vol. 18, no. 23, pp. 9699-9707, 1 Dec.1, 2018.

[34] F. Zheng and Y. -H. Liu, "Visual-Odometric Localization and Mapping for Ground Vehicles Using SE(2)-XYZ Constraints," 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 2019, pp. 3556-3562.

[35] R. Kang, L. Xiong, M. Xu, J. Zhao and P. Zhang, "VINS-Vehicle: A Tightly-Coupled Vehicle Dynamics Extension to Visual-Inertial State Estimator," 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 2019, pp. 3593-3600.

[36] M. Zhang, X. Zuo, Y. Chen, Y. Liu and M. Li, "Pose Estimation for Ground Robots: On Manifold Representation, Integration, Reparameterization, and Optimization," in IEEE Transactions on Robotics, vol. 37, no. 4, pp. 1081-1099, Aug. 2021.

[37] M. Ouyang, Z. Cao, P. Guan, Z. Li, C. Zhou and J. Yu, "Visual-Gyroscope-Wheel Odometry With Ground Plane Constraint for Indoor Robots in Dynamic Environment," in IEEE Sensors Letters, vol. 5, no. 3, pp. 1-4, March 2021, Art no. 6000504.

[38] S. Song, M. Chandraker and C. C. Guest, "High Accuracy Monocular SFM and Scale Correction for Autonomous Driving," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 4, pp. 730-743, 1 April 2016.

[39] B. Lee, K. Daniilidis and D. D. Lee, "Online self-supervised monocular visual odometry for ground vehicles," 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 2015, pp. 5232-5238.

[40] R. Tian, Y. Zhang, D. Zhu, S. Liang, S. Coleman and D. Kerr, "Accurate and Robust Scale Recovery for Monocular Visual Odometry Based on Plane Geometry," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 5296-5302.

[41] J. Shi and C. Tomasi, "Good features to track," in Proc. IEEE Int. Conf. Pattern Recog., pp. 593-600, 1994.

[42] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proc. Int. Joint Conf. Artif. Intell., pp. 24-28, Aug. 1981.

[43] G. Bradski, "The OpenCV Library." [Online]. Available: https://docs.opencv.org/3.4/d9/d0c/group__calib3d.html

[44] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", in Proc. Eur. Conf. Comput. Vis., 2018, pp. 801-818

[45] X. Li et al., "Semantic flow for fast and accurate scene parsing," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 775-793.

[46] S. Agarwal and K. Mierle, "Ceres solver." [Online]. Available: http://ceres-solver.org

[47] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez and V. Koltun, "CARLA: An Open Urban Driving Simulator," Proceedings of the 1st Annual Conference on Robot Learning ser. Proceedings of Machine Learning Research, vol. 78, pp. 1-16, Nov. 2017.