PVI-DSO: Leveraging Planar Regularities for Direct Sparse Visual-Inertial Odometry

Bo Xu®, Xin Li®, Member, IEEE, Jingrong Wang, Chau Yuen®, Fellow, IEEE, and Jiancheng Li

Abstract—The monocular visual-inertial odometry (VIO) based on the direct method can leverage all available pixels in the image to simultaneously estimate the camera motion and reconstruct the denser map of the scene in real-time. However, the direct method is sensitive to photometric changes, which can be compensated by introducing geometric information in the environment. In this article, we propose a monocular direct sparse VIO, which exploits the planar regularities (PVI-DSO). Our system detects the planar regularities from the 3-D mesh built on the estimated map points. To improve the pose estimation accuracy with the geometric information, a tightly coupled coplanar constraint expression is used to express photometric error in the direct method. In addition, to improve the optimization efficiency, we elaborately derive the analytical Jacobian of the linearization form for the coplanar constraint. Finally, the inertial measurement error, coplanar point photometric error, noncoplanar photometric error, and prior error are added into the optimizer, which simultaneously improves the pose estimation accuracy and mesh itself. We verified the performance of the whole system on the simulation and real-world datasets. Extensive experiments have demonstrated that our system outperforms the state-of-the-art counterparts.

Index Terms—3-D mesh, direct sparse method, planar regularities, visual-inertial odometry (VIO).

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) are research hotspots in the field of robots, autonomous driving, augmented reality, etc. Camera and inertial measurement units (IMUs) are low-cost and effective sensors. Visual-inertial odometry (VIO) combines the complementary of the two sensors to improve the accuracy and robustness of the pose estimation. The existing VIO methods [1], [2], [3] generally form visual observations based on the feature point (indirect) method. However, in the weak texture environment, the lack of effective point features extracted by the indirect method makes the system fail to estimate the poses. Fortunately, the direct method [4], [5] can use all the available pixels of images to generate a more complete model, which effectively improves the performance of VIO in weak texture environments. However, the sensitivity to photometric changes makes it difficult to accurately estimate the depth of pixels.

It has been shown that the geometric features (e.g., lines and planes) in the environment can provide valuable information to VIO, and introducing additional constraints guides the optimization processes in the VIO system. Line features are already commonly used in SLAM systems [6], [7], [8], [9]. Compared with straight lines, plane features cannot be accurately recognized, which makes it difficult to introduce plane constraints into SLAM. The learning-based methods have made significant progress in plane detection in recent years [10], [11]. Nevertheless, these algorithms rely on GPUs, which consume relatively high amounts of power, making them impractical for computationally constrained systems. Building lightweight 3-D meshes based on the real-time map generated by VIO and extracting the scene structure from the meshes become a possible way of plane detection [12]. However, this method needs the system to produce denser maps efficiently. However, the indirect VIO only provides a sparse map of 3-D points [13], [14], which is hard to identify planar features. The dense mapping methods [15], [16] based on monocular vision and pixelwise reconstruction lead to high time complexity and low accuracy due to decoupling the trajectory estimation and mapping.

For the visual odometry (VO)/VIO based on the direct method, the accuracy of the pose estimation can be further improved after introducing the planar information. Wu and
Beltrame [17] fuse the coplanar constraints extracted from color image using PlaneNet [18] into DSO [4]. However, the CNN-based technique is time-consuming and can only be applied to global shutter color images. Concha and Civera [19] model the environment as high-gradient and low-gradient areas based on LSD-SLAM [20]. The plane features are segmented from low-gradient areas with superpixels [21] and estimated using singular value decomposition (SVD) for the points clustered in each superpixel. However, in texture-rich scenes, the assumption of extracting the plane information from low-gradient areas is not satisfied, resulting in wrong plane segmentation. We noted that the visual module in the direct method tracks the pixels with large enough intensity gradients. As shown in Fig. 1, sufficient visual observation makes the reconstructed map denser. Therefore, it is easy to extract plane regularities from 3-D mesh built on a real-time estimated denser map, which brings almost negligible computation burden. Furthermore, introducing geometric information that is less sensitive to the photometric changes into VIO can benefit both state estimation and mapping.

This article proposes a direct sparse VIO that leverages planar regularities, called PVI-DSO, which is an extension of DSO [4]. To improve the accuracy and robustness of the system, IMU measurements and coplanar information are integrated into the system. We detect the probable planar regularities from the 3-D mesh generated by the real-time estimated point cloud. Then inspired by the method in [22], a tightly coupled coplanar constraint expression is used to construct photometric error. Finally, we derive the analytic Jacobian for the linearized form of the coplanar constraint and present it in the Appendix. In summary, the main contributions of this work include the below.

1) We introduce the planar regularities which are detected through the 3-D mesh in the direct-method-based VIO to improve the accuracy of the system. The 3-D mesh segmentation which is performed on the real-time estimated denser map not only couples the state estimation and mapping but also brings marginal computation burden.

2) We adopt a tightly coupled coplanar constraint expression in the direct method to construct the photometric error. For the efficiency of the optimization, the analytical Jacobian in the linearization form for the coplanar constraint is derived in detail.

3) We design extensive experiments on the simulation data, the challenging EuRoC dataset [23], and TUM VI dataset [24]. The experimental results demonstrate that our system outperforms the state-of-the-art methods in pose estimation.

II. RELATED WORK

The most common features used in the SLAM/VIO algorithms are the points [1], [2], [3], [4], [25]. As a complement to the point features, the geometric information (line, plane) introduced in the system with point features can improve the accuracy of the pose estimation and mapping, which have received extensive attention in recent years.

A. Indirect Method With Geometric Regularities

Since the line features exist widely in the environment, it seems natural to fuse the line features into the framework based on points [6], [7], [26]. Furthermore, in the structural scenario, the lines with three perpendicular directions of the Manhattan world model can encode the global orientations of the local environments, which are used to improve the robustness and accuracy of pose estimation [8], [27], [28]. For the planar information, the main difficulty is how to accurately extract the planar regularities in the environment. Some works [29], [30], [31] extract plane features with the assistance of depth maps obtained by the RGBD camera. Rosinol et al. [12] propose a stereo VIO fusing plane information, called mesh-VIO. Mesh-VIO extracts the planar regularities from the 3-D meshes, which are generated by projecting 2-D Delaunay triangulation based on 2-D points to the corresponding 3-D points. However, sparse point clouds generated by indirect-based vision algorithms may result in an inaccurate 3-D mesh reconstruction. More recently, Li et al. [32] propose PVIO which leverages multiplane priors in the VIO. The system uses the 3-point RANSAC method to fit the plane among the estimated 3-D points with the RGB camera. Nevertheless, the RANSAC-based method is not stable when there are multipotential planes in the environment.

B. Direct Method With Geometric Regularities

The most prominent direct VO approach in recent years is direct sparse odometry (DSO). To improve the stability of DSO, the IMU measurements are fused into DSO, including: VI-DSO [5] with dynamic marginalization and DM-DSO [33] with delayed marginalization. Meanwhile, the structural information in the environment provides additional visual constraints, which can be used to reduce the drift of the estimator. For the line features used in the direct method,
some works [9], [34], [35] force the 3-D points in the map that satisfy the collinear constraints using straight lines, but not jointly optimizing the estimated poses. Zhou et al. [36] introduce the collinear constraints into the DSO more elegantly. The 3-D lines, points, and poses within a sliding window are jointly optimized. Cheng et al. [37] extract the Manhattan world regularity from the line features in the image and merge the structural information into the optimization framework of photometric error. Currently, there are few works on fusing planar features in the direct method. As mentioned previously, the CNN-based plane detection method [17] consumes a lot of computing resources. The superpixel segmentation method [19] leads to wrong segmentation of planes in texture-rich scenes.

### III. System Overview

The system proposed in this article is based on DSO [4]. We introduce the IMU measurements and geometric information in the environment to improve the accuracy and robustness of the system. As shown in Fig. 2, the proposed system contains two main modules running in parallel, the front end and the back end.

In the front end, we perform the coarse tracking to obtain the pose estimation and the photometric parameters’ estimation of the current frame, which serves as the initialization for the optimization in the back end. With the aid of IMU integration, the coarse tracking is executed for every frame, in which the direct image alignment based on the photometric bundle adjustment is used to estimate the state variables of the most recent frame. And then, we determine whether the current frame is the keyframe through the criterion described in [4]. If the current frame is a keyframe, it will be delivered to the back-end module. Otherwise, it is only used to track the candidate points to update the depth. To do so, we search along the epipolar line to find the correspondence with the minimum photometric error.

In the back end, the candidate points are tracked and refined with the latest keyframe to obtain more accurate depth. To activate the candidate points with low uncertainty, all the active points in the sliding window are projected onto the most recent keyframe. The candidate points (also projected into this keyframe) are activated as active points when their distance to any existing point is maximum [4]. Then, the planar regularities are extracted from the activated points in the point cloud. We construct the 3-D meshes through the depth map which is generated by the active points. The coplanar constraints and noncoplanar constraints are obtained through the segmentation of the 3-D meshes. Finally, the operations of the visual-inertial bundle adjustment are executed. We add the noncoplanar point residuals, coplanar point residuals, inertial residuals, and the corresponding prior residuals into the optimizer, which simultaneously improves the accuracy of the pose estimation and the mesh itself.

### IV. Notations and Preliminaries

In this section, coordinate transformations are defined, and the representations of point and plane features are given. Throughout the article, we denote vectors as bold lowercase letters $\mathbf{x}$, matrices as bold uppercase letters $\mathbf{H}$, scalars as lowercase letters $\lambda$, and functions as uppercase letters $E$.

#### A. Coordinate Transformation

The world coordinate system is defined as a fixed inertial framework in which the $z$-axis is aligned with the gravity direction. $\mathbf{T}_{w_i} \in \text{SE}(3)$ represents the transformation from the IMU frame to the world frame, where $\mathbf{R}_{w_i} \in \text{SO}(3)$ and $\mathbf{t}_{w_i} \in \mathbb{R}^3$ are the rotation and translation, respectively. The transformation from the camera frame to the IMU frame is defined as $\mathbf{T}_{ic} \in \text{SE}(3)$, and the transformation from the camera frame to the world frame $\mathbf{T}_{wc}$ can be calculated by:

$$\mathbf{T}_{wc} = \mathbf{T}_{w_i} \mathbf{T}_{ic}.$$  
Similarly, $\mathbf{R}_{ic}$ and $\mathbf{R}_{wc}$ represent the rotations from the camera frame to the IMU frame and from the camera frame to the world frame, respectively. $\mathbf{t}_{ic}$ and $\mathbf{t}_{wc}$ are the corresponding translations.

#### B. Point Representation

We use the inverse depth $d_p \in \mathbb{R}$ to parameterize the pixels from the host image frame in which it is extracted. With this parametric expression, we can project the pixels in the host image frame $I_h$ into the target image frame $I_t$, where it is observed. Assuming $\mathbf{p}_i$ is the pixel in $I_h$, the projection point $\mathbf{p}'_i$ in $I_t$ is given by

$$\mathbf{p}'_i = \Pi_t (\mathbf{R}_{c_i} \mathbf{R}_{c_h}^{-1} (\mathbf{p}_i, d_p) + \mathbf{t}_{c_i} \mathbf{t}_{c_h})$$

where $\mathbf{R}_{c_i} \mathbf{c_h}$ and $\mathbf{t}_{c_i} \mathbf{t}_{c_h}$ are the relative rotation and translation from image frame $I_h$ to $I_t$, respectively, $\Pi_t$ and $\Pi^{-1}_t$ are the projection and back projection of the camera, respectively.

#### C. Plane Representation

As shown in Fig. 3, a plane in the world frame can be represented by the Hessian normal form $\pi_w = [\mathbf{n}_w, d_w]^T$, where $\mathbf{n}_w = [n_x, n_y, n_z]$ is the normal of the plane, representing its orientation and $d_w$ is the distance from the origin of the world frame to the plane. The normal vector $\mathbf{n}_w$ has three parameters but only 2 degrees of freedom (DoFs) with $\|\mathbf{n}_w\|_2 = 1$.

To get a minimal parameterization of $\pi_w$ for optimization, we represent it as $q(\pi_w) = [\phi, \psi, d_w]$, where $\phi$ and $\psi$...
A. Coplanarity Detection

The planar regularities are detected from the 3-D mesh of the surrounding environment [12], [14]. The method first builds 2-D Delaunay triangulation with the feature points in the image, and then projects the triangular regularities to 3-D landmarks. The 3-D mesh is formed by organizing landmarks into 3-D patches. In the direct method VIO, as shown in Fig. 1(c), 2-D Delaunay triangulation is generated in the depth map. The reason for this is that the depth map is composed of active points whose depth has converged, ensuring the accuracy of 3-D mesh. Moreover, the depth map is anchored to the multi keyframes in the sliding window, avoiding the generation of 2-D Delaunay triangulation frame by frame, which limits memory usage. There might be invalid 3-D triangular patches generated from depth map. We use the method in [12] to filter out these outliers, which ensures that 3-D triangular patches should not be particularly sharp triangles, such as triangles with the aspect ratio high than 20 or an acute angle smaller than 5°.

The gravity direction of VIO can be used to improve the efficiency of plane detection. Thereby, we only detect the planes that are either horizontal (i.e., walls, the normal is perpendicular to the gravity direction) or vertical (i.e., floor, the normal is parallel to the gravity direction), which are commonly found in man-made environments. For horizontal plane detection, in most cases, the only horizontal plane we observed is the floor. Therefore, we only detect and optimize one horizontal plane at a time to ensure that it can be tracked for a long time. The specific strategy is: all the triangular patches that are parallel for the gravity direction are collected. Then we build a 1-D histogram of the average height of triangular patches. After statistics, a Gaussian filter is used to eliminate multiple local maximums as used in [12]. Finally, we extract the local maximum of the histogram and consider it to be the horizontal plane when it exceeds a certain threshold (σv = 20).

For vertical plane detection, a 2-D histogram is built, where one axis is the azimuth φ of the plane’s normal vector, and the other axis is the distance d from the origin to the plane. The histogram is divided into nφ × nσ bins with a bin size δφ × δd. We only count the triangular patches that the normal vector is perpendicular to the gravity direction. When the triangular patches fall into the corresponding bin, the number of bin will be increased by 1. After statistics, we take the candidates with more than 20 inliers.

B. Photometric Error of Noncoplanar Point

The direct method is based on the photometric invariance hypothesis to minimize the photometric error. In a similar way as [4], the photometric error for a noncoplanar point \( p_h \) with inverse depth \( d_{p_h} \) in the host image \( I_h \) observed by the target image \( I_t \) is defined as

\[
E_{p_h} = \sum_{p_i \in N_{p_h}} w_{p_i} \left\| I_t \left[ p_i \right] - b_i - \frac{t_e}{t_h e^{d_h}} \left( I_h \left[ p_h \right] - b_h \right) \right\|_y
\]

where \( I_h \) and \( t_i \) are the exposure times of the respective image \( I_h \) and \( I_t \), \( a_h, a_t, b_h, \) and \( b_t \) denote the affine illumination parameters.
transform parameters. \( \mathcal{N}_{\mathbf{p}_\mathbf{n}} \) represents a small set of pixels around the point \( \mathbf{p}_\mathbf{n} \). \( w_{\mathbf{p}_\mathbf{n}} \) stands for a gradient-dependent weighting, and \( \| \cdot \|_w \) indicates the Huber norm. \( \mathbf{p}_\mathbf{n} \) is the point projected into \( I_i \), which is obtained by (1). Therefore, the sum of the photometric error of the noncoplanar points observed by all the keyframes in the sliding window is defined as

\[
E_{\text{point}} = \sum_{i \in \mathcal{F}} \sum_{\mathbf{p}_\mathbf{n} \in \mathcal{C}_i} \sum_{j \in \text{obs}(\mathbf{p}_\mathbf{n})} E_{\mathbf{p}_\mathbf{n}} \tag{6}
\]

where \( \mathcal{F} \) is the keyframes in the sliding window, \( \mathcal{P}_l \) is the set of noncoplanar points in the host keyframe \( i \), and \( \text{obs} (\mathbf{p}_\mathbf{n}) \) is a set of observations of \( \mathbf{p}_\mathbf{n} \) in the other keyframes.

### C. Photometric Error of Coplanar Point

Assuming that the plane \( \pi_c \) detected from the 3-D mesh is expressed in the world frame, the photometric error of the coplanar points can be constructed with the plane constraint equation. Suppose the 3-D point corresponding to \( \mathbf{p}_\mathbf{r} \) in \( I_h \) lies on the plane \( \pi_c \) transformed into the host image frame, the point \( \mathbf{p}_\mathbf{r} \) satisfies the coplanar equation

\[
\mathbf{n}_c^T \Pi_{c}^{-1}(\mathbf{p}_\mathbf{r}, 1)/d_p + d_{\sigma_c} = 0 \tag{7}
\]

where \( \Pi_{c}^{-1}(\mathbf{p}_\mathbf{r}, 1) \) represents the point on the normalized image plane. \( \pi_c \) is the plane in the host image frame which can be obtained by (3). \( \mathbf{n}_c \) and \( d_{\sigma_c} \) are the normal vector and distance of \( \pi_c \), respectively. Substituting the inverse depth \( d_p \) of (7) which is regularized by plane constraints into (1), the projection point \( \mathbf{p}_\mathbf{r} \) of the planar points is given by

\[
\mathbf{p}_\mathbf{r} = \Pi_{c}(\mathbf{R}_c \mathbf{c}] \Pi_{c}^{-1}(\mathbf{p}_\mathbf{r}, -\mathbf{n}_c^T \Pi_{c}^{-1}(\mathbf{p}_\mathbf{r}, 1)/d_p) + \mathbf{t}_c \tag{8}
\]

The photometric error of the coplanar points is obtained by substituting \( \mathbf{p}_\mathbf{r} \) into (5), which is the same as that of the noncoplanar points. The sum of the photometric error of the coplanar points observed by all the keyframes in the sliding window can be written as

\[
E'_{\text{point}} = \sum_{i \in \mathcal{F}} \sum_{\mathbf{p}_\mathbf{n} \in \mathcal{C}_i} \sum_{j \in \text{obs}(\mathbf{p}_\mathbf{n})} E_{\mathbf{p}_\mathbf{n}} \tag{9}
\]

where \( \mathcal{C}_i \) is the set of coplanar points in the host keyframe \( i \), and \( \text{obs} (\mathbf{p}_\mathbf{n}) \) is a set of observations of \( \mathbf{p}_\mathbf{n} \) in the other keyframes.

For the optimization method, deriving the analytical Jacobian of estimated variables with the chain derivation method can significantly improve the efficiency of the optimizer. For a single photometric error, the Jacobian corresponding to the state variables can be decomposed as [4]

\[
J_{E_{\mathbf{p}_\mathbf{n}}} = [J_f \cdot J_{\text{geo}} \cdot J_{\text{photo}}] \tag{10}
\]

where \( J_f \) represents the image gradient. \( J_{\text{geo}} \) denotes the Jacobian of geometric parameters \( (T_{w_i}, T_{w_j}, \pi_w) \), and \( J_{\text{photo}} \) indicates the Jacobian of photometric parameters \( (a_{\mathbf{u}}, a_{\mathbf{v}}, b_{\mathbf{u}}, b_{\mathbf{v}}) \). Refer to the appendix for the specific forms of \( J_{E_{\mathbf{p}_\mathbf{n}}} \). It is worth noting that the coplanar constraints used in this article do not optimize the depth of optimization points simultaneously. As shown in Fig. 5, compared with the traditional point-to-plane constraints [12], [14], the efficiency of the algorithm can be improved significantly by reducing the dimension of the Hessian matrix in the optimizer.

### D. Inertial Error

To construct the inertial error with angular velocity and linear acceleration measurements, the preintegration method proposed in [38] is used to handle the high frequency of IMU measurements. This gives a pseudomeasured value between consecutive keyframes. Given the previous IMU state \( \mathbf{s}_{i} = [v_{w_i}, \mathbf{R}_{w_i}, \mathbf{b}_{w_i}, \mathbf{b}_{w_i}] \), which contains translation, rotation, velocity in the IMU frame, and gyroscope and accelerometer biases, the preintegration measurements provide us with the prediction state \( \mathbf{s}_{i+1} \) for the following state \( \mathbf{s}_{i} \) and a covariance matrix \( \mathbf{\Sigma}_{i} \). The resulting inertial error function penalizes deviations from the current estimated state to the predicted state

\[
E_{\text{inertial}} = (\mathbf{s}_{i} - \mathbf{s}_{i+1})^T \mathbf{\Sigma}_{i}^{-1} (\mathbf{s}_{i} - \mathbf{s}_{i+1}) \tag{11}
\]

The subtraction operation \( s_i \mp s_{i+1} \) is defined as \( \log(\mathbf{R}_i, \mathbf{R}_{i+1}) \) for rotation and a regular subtraction for vector values.

### E. Marginalization About Coplanar Constraints

To balance the efficiency and accuracy, the marginalization method is used in VO/VIO [4], [5], [33], [39]. When a new image frame is added to the sliding window, all the variables corresponding to the older frame are marginalized using the Schur complement [39]. The marginal keyframe is selected similar to the criteria in [4], which considers the luminance change in the images and the geometry distribution of the poses. Meanwhile, to maintain the consistency of the system, once the variable is connected to the marginalization factor, the first Jacobian estimation (FEJ) [40] is used to fix the variable's linearization point at the same value.

In the sliding window optimization, maintaining too many historical planes in the optimization will seriously affect the efficiency of the optimizer. The historical plane should also be “marginalized” to form the plane’s prior factor. For simplicity, the plane prior factor generated by marginalization is replaced by a plane-distance cost, which can be expressed by the distance from the prior plane \( \pi' = (\mathbf{\phi}', \mathbf{\psi}', d') \) to the currently optimized plane \( \pi = (\mathbf{\phi}, \mathbf{\psi}, d) \)

\[
E_{\pi'} = w_{\pi'} \left\| (\mathbf{\phi}', \mathbf{\psi}', d')^T - (\mathbf{\phi}, \mathbf{\psi}, d)^T \right\| _{\Sigma_{\pi}}^2 \tag{12}
\]
where $\Sigma_p$ is the covariance matrix, $w_n$ denotes the number of the coplanar constraints corresponding to the plane. If the plane is marginalized, a prior plane $\pi'$ is formed. When the plane is observed again, it is reactivated and optimized with prior constraint (12) in the sliding window.

**VI. EXPERIMENTS**

In this section, simulation experiments are first used to verify the effectiveness of our method, and then the evaluation on the real EuRoC MAV dataset [23] and TUM VI dataset [24] demonstrates the accuracy and efficiency. We provide a video\(^1\) to reflect the results of the experiments more intuitively. The related code about our analytical Jacobian of coplanar parametric representation is published on GitHub to facilitate communication.\(^2\) We run the system in the environment with Intel Core i7-9750H at 2.6 GHz, 32-GB memory.

**A. Quantitative Evaluation**

To evaluate the performance of the coplanar constraints in the VIO system, we reimplemented a VIO based on DSO [4], which is denoted as VI-DSO-RE to distinguish it from VI-DSO [5] without open source code, and the coplanar constraints are added into VI-DSO-RE, which is denoted as PVI-DSO. To verify the effectiveness of coplanar constraints in the direct method, we evaluate the performance of VI-DSO, VI-DSO-RE, and PVI-DSO and also compare PVI-DSO with the state-of-the-art VIOs fusing coplanar constraints: PVIO [32] with multiplane priors and mesh-VIO [12], which fuses the planar regularities generated by 3-D meshes.

The simulations and real experiments are conducted, in which we perform $\text{SE}(3)$ alignment against the ground truth to get the root mean square error (RMSE) of the translation and rotation error. The scale error is computed with $|1 - s|$, where $s$ is obtained from Sim(3) alignment. For the TUM VI dataset, since the trajectory lengths can vary greatly, the drifts in % are also computed with $(\text{RMSE} \cdot 100)/\text{length}$ as in [33]. Unless otherwise states all the methods are evaluated 10 times for the EuRoC MAV and TUM VI datasets on each sequence and the medium results are reported.

1) **Simulation Experiment**: Two simulation sequences with ideal coplanar environments are created to evaluate the efficiency concerning performance under different parametric representations. As shown in Fig. 6, we simulated camera motion with 10 Hz, forming an ellipse trajectory with sinusoidal vertical motion. The long-semi axis and short axis of the ellipse trajectory are 4 and 3 m, respectively. The number of landmarks in each plane is limited to 250, which consists of the vertical plane sequence and horizontal plane sequence. The Gaussian noise is added to the landmark observations $(\sigma_p = 1 \text{ pixel})$, plane parameters $(\sigma_n = 5^\circ, \sigma_d = 0.3 \text{ m})$, and camera poses $(\sigma_R = 5^\circ, \sigma_t = 0.3 \text{ m})$ to simulate the real observation environment. The comparison method includes: Point, PP (-L), and PP (-T), where Point denotes the point-based method, the implementation of which is obtained from open-sourced SLAM VINS [1]. Both PP (-L) and PP (-T) use coplanar constraints in the optimization module, but in different ways. PP(-L) fuses more constraints between point-to-plane with the loosely coupled method as in [12] and [14]. Whereas PP (-T) adopts the parametric representation used in this article to fuse plane constraints in a tightly coupled method. Considering it is difficult to simulate photometric error, Point, PP (-L), and PP (-T) all adopt the geometric reprojection error instead to verify the performance of different parametric representations in the simulation experiment.

As shown in Fig. 7, compared with the Point method, the translation error of PP (-L) and PP (-T) with coplanar constraints decreases on both the sequences, which demonstrates that the performance of the VIO can be improved with the structural information in the environment. The accuracy difference between PP (-L) and PP (-T) is marginal, whereas the effectiveness of PP (-T) is improved significantly, which is about 2.3 times and 2.4 times faster than Point and PP (-L), respectively. This is because PP (-T) reduces the dimension of the Hessian matrix in the optimization without estimating the state variables of coplanar points.

2) **EuRoC Dataset**: The EuRoC micro aerial vehicle (MAV) dataset consists of two scenes, the machine hall and the ordinary room, which contain different motion and visual scenes. We compared the positioning accuracy of VI-DSO, VI-DSO-RE, PVI-DSO, PVIO, and mesh-VIO. As shown in Table I, the average RMSE of VI-DSO, VI-DSO-RE, and PVI-DSO is 0.089, 0.094, and 0.083 m, respectively. VI-DSO adopts dynamic marginalization to maintain the consistency of the marginalization prior. VI-DSO-RE does not use any marginalization tricks, so the accuracy of VI-DSO is slightly higher than that of VI-DSO-RE. Furthermore, the accuracy of PVI-DSO leveraging planar regularities outperforms VI-DSO and VI-DSO-RE, which demonstrates that the prior structural

---

\(^1\)https://youtu.be/h-3fP6VP0_k
\(^2\)https://github.com/boxulibrary/PVI-DSO-SIM
information fused into VIO based on direct method suppresses the rapid divergence of the VIO, thereby reducing the cumulative error generated by long-time operation of VIO. By comparing the scale error of VI-DSO-RE and PVI-DSO, the average scale error of PVI-DSO decreases by 4%. For further analysis, we divided the trajectories of V21 and V22 into multiple segments with 10-s intervals to compute the scale error separately. As shown in Fig. 8, the scale error of PVI-DSO decreased in some segmented trajectories compared with VI-DSO-RE, which indicates that the consistency of the global scale estimation is improved after introducing the structural information. Finally, from the experimental results of PVI-DSO and two methods that introduce the planar information into VIO based on the indirect method, we can observe that our method is significantly better than PVIO and mesh-VIO. This is because VIO based on the direct method uses more visual information in the scene to estimate the state variables. Moreover, the denser map constructed by PVI-DSO contains rich structural information, which promotes the detection of the planar regularities in the map, so as to improve the positioning accuracy better after introducing the planar constraints.

3) TUM VI Dataset: We also evaluate our proposed system on the two sequences (corridor and magistrale) of the TUM VI dataset. The sequences contain a large number of images with motion blur and photometric changes, which is challenging for the direct method. In Fig. 9, we reveal the reconstructed map and the extracted coplanar points on the ground and walls running PVI-DSO. The drift of VIO can be effectively suppressed using the structural information in the scene. From Table II, it can be seen that compared with PVIO, no matter VI-DSO-RE or PVI-DSO, the translation error and rotation error decrease significantly. Furthermore, introducing the planar constraints into VI-DSO-RE, the average translation drift and rotation drift of PVI-DSO decrease by 13% and 23%, respectively, which verifies the effectiveness of our proposed method. It should be noted that there are sequences where the

| Sequence   | MH1 | MH2 | MH3 | MH4 | V11 | V12 | V13 | V21 | V22 | V23 |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| VI-DSO     | 0.062 | 0.044 | 0.117 | 0.132 | 0.121 | 0.059 | 0.067 | 0.096 | 0.040 | 0.062 | 0.174 |
| Scale Error(%) | 1.1  | 0.5  | 0.4  | 0.2  | 0.8  | 1.1  | 1.1  | 0.5  | 1.2  | 0.3  | 0.4  |
| VI-DSO-RE  | 0.081 | 0.046 | 0.054 | 0.137 | 0.205 | 0.051 | 0.108 | 0.095 | 0.060 | 0.057 | 0.138 |
| Scale Error(%) | 1.3  | 0.1  | 0.1  | 0.1  | 0.5  | 1.7  | 1.9  | 1.5  | 0.2  | 1.8  | 0.1  |
| PVI-DSO    | 0.073 | 0.046 | 0.055 | 0.114 | 0.175 | 0.056 | 0.100 | 0.087 | 0.050 | 0.044 | 0.111 |
| Scale Error(%) | 0.7  | 0.4  | 0.3  | 0.5  | 0.2  | 1.2  | 1.9  | 0.9  | 0.1  | 1.7  | 1.0  |
| mesh-VIO   | 0.145 | 0.130 | 0.212 | 0.217 | 0.226 | 0.057 | 0.074 | 0.167 | 0.081 | 0.103 | 0.272 |
| PVIO       | 0.163 | 0.111 | 0.119 | 0.353 | 0.225 | 0.082 | 0.113 | 0.201 | 0.063 | 0.157 | 0.280 |

1 results taken from [5].
2 results taken from [12].
3 results taken from [32].
accuracy of PVI-DSO decreases after introducing the planar information (e.g., corridor3-4, magistrale1-2). This is because the estimation of the plane parameters gradually converges during system operation. The visual observation noise may make the plane fail to converge to accurate position, and the inaccurate prior information can damage the performance of the system.

### B. Weight Determination of Photometric Error

Since the camera’s photometric calibration and distortion correction have significant influences on the standard deviations of the photometric error, it is difficult to determine the weight ratio of the photometric residuals and inertial residuals in the optimization. In this article, the parametric study method is used to obtain the optimal standard deviation of the photometric error. After setting different standard deviations, the cumulative error curves of 200 and 120 runs on the EuRoC and TUM VI datasets are counted. According to Fig. 10, it can be seen that the different weight ratios of photometric and inertial residual have a great impact on the positioning results of the system, and the experimental results on the TUM VI dataset are more sensitive to the different settings. At the same time, we can also observe that in the experiments, the best performance is obtained by setting the photometric error standard deviation 11 and 16 on the EuRoC and TUM VI datasets, respectively.

### C. Runtime Evaluation

To show the time-consuming details of the algorithm, the time consumption of different modules in the algorithm is carefully counted. We calculated the time consumption of VO based on the direct method, VI-DSO-RE introducing the inertial constraints in the VO, and PVI-DSO leveraging the planar regularities in the VI-DSO-RE. Considering the real-time performance of the system, we set the number of extracted pixels in the keyframe to 800, the size of the sliding window to 7, and the maximum number of iterations to 5.

The computation cost (in milliseconds) of different modules for running on EuRoC’s V11 sequence is shown in Table III. In pure visual VO, about two-thirds of the time is used for optimization, and the rest time is spent on the visual feature processing module in direct VO, including pixels’ tracking, candidates’ selection, points’ activation, pixels’ extraction, etc. After introducing the IMU measurements into VO, the total time consumption of VI-DSO-RE increases slightly. For PVI-DSO, plane detection and mesh construction have little burden on the system operation, which only cost 0.72 and 1.08 ms, respectively. Whereas compared with VI-DSO-RE, the time cost of the optimization and marginalization is decreased. The reason is that PVI-DSO does not need to optimize the coplanar points, which reduces the dimensions of the matrix to be solved. The statistics shows that in the sliding window, the average number of the state variables (plane number $\approx$ 2, point number $\approx$ 1182) of PVI-DSO is fewer than that (point number $\approx$ 1422) of VI-DSO-RE.

### VII. Conclusion

In this article, we present a direct sparse VIO that leverages planar regularities, which is called PVI-DSO. The denser
where \( \mathbf{f}_i \) is the point observation in the normalized plane of the host image frame. \((\mathbf{R}_{wi}, \mathbf{t}_{wi}), * \in \{h, t\}\) indicates the rotation and translation of the host image frame and target image frame. \((\mathbf{R}_{wi}, \mathbf{t}_{wi}), * \in \{h, t\}\) denotes the rotation and translation of the host IMU frame and target IMU frame. \(\mathbf{f}_i, \mathbf{f}_h, \mathbf{f}_*\) are calculated by \(\mathbf{f}_i = \mathbf{R}_i \mathbf{f}_i \) and \(\mathbf{f}_h = \mathbf{R}_h \mathbf{f}_*\), respectively. \(|.|_\times\) is the skew-symmetric operator. \(t_{cp}\) is obtained by \(t_{cp} = (\mathbf{R}_{wi}^T \mathbf{n}_w) \cdot \mathbf{f}_i\). And \(d_c\) is the distance of plane in the host image frame, which is obtained by (3).

The Jacobian of the target IMU frame \(\mathbf{T}_{wh}\) is written as

\[
\frac{\partial \mathbf{r}_k}{\partial \delta \mathbf{T}_{wh}} = J_{\text{f}} \cdot \frac{\partial \mathbf{p}}{\partial \mathbf{P}_c^{-1}(\mathbf{p}', \mathbf{d}_p)} \cdot \frac{\partial \mathbf{P}_c^{-1}(\mathbf{p}', \mathbf{d}_p)}{\partial \delta \mathbf{\xi}_{wh}}
\]

where \(\mathbf{J}_{\text{rot}} \in \mathbb{R}^{3 \times 3}\) is the Jacobian of the rotation part of the host IMU frame, and \(\mathbf{J}_{\text{trans}} \in \mathbb{R}^{3 \times 3}\) is the Jacobian of the translation part of the host IMU frame. They are given by

\[
\mathbf{J}_{\text{rot}} = - \frac{\mathbf{R}_{wi}^T \mathbf{R}_{wh} \left[ \mathbf{f}_t \mathbf{d}_t \right] T \mathbf{c}_p - \mathbf{d}_t \mathbf{f}_t \mathbf{f}_t^T \mathbf{R}_{wi}^T \mathbf{n}_w }{ t_{cp} \times t_{cp} } \quad \mathbf{J}_{\text{trans}} = - \frac{\mathbf{f}_t \cdot \mathbf{n}_w}{t_{cp}} + \mathbf{R}_{wi}^T. \tag{18}
\]

The Jacobian of the plane parameters is given by

\[
\frac{\partial \mathbf{p}}{\partial \delta \mathbf{\xi}_{wh}} = \mathbf{J}_{\text{rot}} \cdot \mathbf{J}_{\text{trans}}
\]

where \(\mathbf{J}_{\text{rot}} \in \mathbb{R}^{3 \times 3}\) is the Jacobian of the rotation part of the target IMU frame, and \(\mathbf{J}_{\text{trans}} \in \mathbb{R}^{3 \times 3}\) is the Jacobian of the translation part of the target IMU frame. They are given by

\[
\mathbf{J}_{\text{rot}} = \mathbf{R}_{wi}^T \mathbf{R}_{wc} \left[ \frac{- \mathbf{R}_{wc} \mathbf{f}_t \mathbf{d}_t}{t_{wc}} + \mathbf{t}_{wc} - \mathbf{t}_{wc} \right] T \mathbf{x}_c
\]

\[
\mathbf{J}_{\text{trans}} = - \mathbf{R}_{wi}^T
\]

(21)

The Jacobian of the plane parameters is given by

\[
\frac{\partial \mathbf{p}}{\partial \delta \mathbf{\xi}_{wh}} = \mathbf{J}_{\text{rot}} \cdot \mathbf{J}_{\text{trans}}
\]

where \(\mathbf{J}_{\text{rot}} \in \mathbb{R}^{3 \times 3}\) is the Jacobian of the rotation part of the host IMU frame. \((\partial \mathbf{P}_c^{-1}(\mathbf{p}', \mathbf{d}_p)) / (\partial \mathbf{\omega}_\pi)\) is calculated by

\[
\frac{\partial \mathbf{P}_c^{-1}(\mathbf{p}', \mathbf{d}_p)}{\partial \delta \mathbf{\omega}_\pi}
\]

where \(\mathbf{J}_{\text{rot}} \in \mathbb{R}^{3 \times 3}\) represents the Jacobian of the global parameters’ representation of the normal vector. \(\mathbf{J}_{n, l} \in \mathbb{R}^{3 \times 2}\) indicates the Jacobian of the global parameters with respect to the local parameters of the normal vector. \(\mathbf{J}_{d_c} \in \mathbb{R}^{3 \times 1}\) denotes the Jacobian of the distance. \(\mathbf{J}_{n, w} \) and \(\mathbf{J}_{d_c}\) are
written as
\[
\mathbf{J}_{\mathbf{n}_d} = -\frac{\mathbf{R}_{CR} \mathbf{f}_\mathbf{c} \left( \mathbf{t}_{w_\mathbf{c}} \times \mathbf{t}_{d} - \left( \mathbf{R}_{w_\mathbf{c}} \mathbf{f}_\mathbf{c} \right)^T \cdot \mathbf{d} \right)}{\mathbf{t}_{d} \times \mathbf{t}_{d}}
\]
\[
\mathbf{J}_{\mathbf{n}_p} = -\frac{\mathbf{R}_{CR} \mathbf{f}_\mathbf{c}}{\mathbf{t}_{d}}
\]
(24)
and \( \mathbf{J}_{\mathbf{n}_p} \) is given by
\[
\mathbf{J}_{\mathbf{n}_p} = \begin{bmatrix}
- \cos(\psi) \sin(\phi) & - \sin(\psi) \cos(\phi) & 0 \\
\cos(\psi) \cos(\phi) & - \sin(\psi) \sin(\phi) & 0 \\
0 & 0 & \cos(\psi)
\end{bmatrix}.
\]
(25)

ACKNOWLEDGMENT

The authors would like to thank Dr. Yijia He for the helpful discussion.

REFERENCES

[1] T. Qin, P. Li, and S. Shen, “VINS-Mono: A robust and versatile monocular visual-inertial state estimator,” IEEE Trans. Robot., vol. 34, no. 4, pp. 1004–1020, Aug. 2018.
[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” IEEE Trans. Robot., vol. 37, no. 6, pp. 1874–1890, Dec. 2021.
[3] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint Kalman filter for vision-aided inertial navigation,” in Proc. IEEE Int. Conf. Robot. Autom., Apr. 2007, pp. 3565–3572.
[4] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 611–625, Mar. 2018.
[5] L. Von Stumberg, V. Usenko, and D. Cremers, “Direct sparse visual-inertial odometry using dynamic marginalization,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2018, pp. 2510–2517.
[6] Y. He, J. Zhao, Y. Guo, W. He, and K. Yuan, “PL-VIO: Tightly-coupled monocular visual–inertial odometry using point and line features,” Sensors, vol. 18, no. 4, p. 1159, Apr. 2018.
[7] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, “PL-SLAM: Real-time monocular visual SLAM with points and lines,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2017, pp. 4503–4508.
[8] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu, “StructSLAM: Visual SLAM with building structure lines,” IEEE Trans. Veh. Technol., vol. 64, no. 4, pp. 1364–1375, Apr. 2015.
[9] L. Pei, Y. Ren, Y. Liu, M. Cheng, D. Frost, and V. A. Prisacariu, “Direct line guidance odometry,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2018, pp. 5137–5143.
[10] Z. Yu, J. Zheng, D. Lian, Z. Zhou, and S. Gao, “Single-image piecewise planar 3D reconstruction via associative embedding,” in Proc. IEEE/ACCV Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1029–1037.
[11] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, “PlaneRCNN: 3D plane detection and reconstruction from a single image,” in Proc. IEEE/ACCV Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4445–4454.
[12] A. Rosinol, T. Sattler, M. Pollefeys, and L. Carleone, “Incremental visual-inertial 3D mesh generation with structural regularities,” in Proc. Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 8220–8226.
[13] J. Wu, J. Xiong, and H. Guo, “Enforcing regularities between planes using key plane for monocular mesh-based VIO,” J. Intell. Robot. Syst., vol. 104, no. 1, p. 6, Jan. 2022.
[14] X. Li, Y. He, J. Lin, and X. Liu, “Leveraging planar regularities for point line visual-inertial odometry,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2020, pp. 5120–5127.
[15] R. A. Newcombe and A. J. Davison, “Live dense reconstruction with a single moving camera,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1498–1505.
[16] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, “DTAM: Dense tracking and mapping in real-time,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 2320–2327.
[39] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, “Visual-inertial mapping with non-linear factor recovery,” IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 422–429, Apr. 2020.
[40] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, “A first-estimates Jacobian EKF for improving SLAM consistency,” in Experimental Robotics. Cham, Switzerland: Springer, 2009, pp. 373–382.

Bo Xu received the B.E. degree from the School of Land Science and Technology, China University of Geosciences, Beijing, China, in 2018, and the M.E. degree from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics. His research interests include GNSS precise positioning, visual SLAM, visual inertial odometry (VIO), and multisensor fusion algorithm.

Xin Li (Member, IEEE) received the B.E. degree in measurement and control technology and instruments from Northeastern University, Qinhuangdao, China, in 2018, and the M.E. degree in electronics and communication engineering from Peking University, Beijing, China, in 2021. His research interests include SLAM, 3-D reconstruction, and scene understanding.

Jingrong Wang received the B.E. degree from the School of Land Science and Technology, China University of Geosciences, Beijing, China, in 2018, and the M.E. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the Ph.D. degree with the GNSS Research Center. His research interests include GNSS precise positioning, visual inertial odometry (VIO), and multisensor fusion algorithm.

Chau Yuen (Fellow, IEEE) received the B.Eng. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He was a Post-Doctoral Fellow with the Lucent Technologies Bell Laboratory, Murray Hill, NJ, USA, in 2005, and a Visiting Assistant Professor with The Hong Kong Polytechnic University, Hong Kong, in 2008. From 2006 to 2010, he was with the Institute for Infocomm Research, Singapore. From 2010 to 2023, he was with the Engineering Product Development Pillar, Singapore University of Technology and Design, Singapore. Since 2023, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His main research interests include advanced connectivity, the Internet of Things/IoT, and wireless networks.

Jiancheng Li received the Ph.D. degree from Wuhan University, Wuhan, China, in 1993. He is a Professor in Geodesy with Wuhan University. He was honored as an Academic Member of the Chinese Academy of Engineering, Beijing, China, in 2011. His main research interests include Earth gravity field and its engineering applications.