LF-VIO: A Visual-Inertial-Odometry Framework for Large Field-of-View Cameras with Negative Plane

Ze Wang1,4, Kailun Yang2, Hao Shi1, Peng Li1, Fei Gao3,4 and Kaiwei Wang1

Abstract—Visual-inertial-odometry has attracted extensive attention in the field of autonomous driving and robotics. The size of Field of View (FoV) plays an important role in Visual-Odometry (VO) and Visual-Inertial-Odometry (VIO), as a large FoV enables to perceive a wide range of surrounding scene elements and features. However, when the field of the camera reaches the negative half plane, one cannot simply use \([u, v, 1]^T\) to represent the image feature points anymore. To tackle this issue, we propose LF-VIO, a real-time VIO framework for cameras with extremely large FoV. We leverage a three-dimensional vector with unit length to represent feature points, and design a series of algorithms to overcome this challenge. To address the scarcity of panoramic visual odometry datasets with ground-truth location and pose, we present the PALVIO dataset, collected with a Panoramic Annular Lens (PAL) system with an entire FoV of \(360^\circ \times (40^\circ \sim 120^\circ)\) and an IMU sensor. With a comprehensive variety of experiments, the proposed LF-VIO is verified on both the established PALVIO benchmark and a public fisheye camera dataset with a FoV of \(360^\circ \times (0^\circ \sim 93.5^\circ)\). LF-VIO outperforms state-of-the-art visual-inertial-odometry methods. Our dataset and code are made publicly available at https://github.com/flysoaryun/LF-VIO

I. INTRODUCTION

With the rapid development of robotics and autonomous vehicles, Visual-Odometry (VO) and Visual-Inertial-Odometry (VIO) have been widely applied in navigation systems [1]–[4]. Meanwhile, cameras with a large Field of View (FoV) have been increasingly put into integration in such systems to enable a wide range of surrounding sensing of scene elements and features, which are often beneficial for upper-level vision perception and odometry tasks [5]–[12]. Some modern panoramic cameras [13]–[15] even have a negative plane field, that facilitates an ultra-wide surrounding understanding, where imaging points may appear on the negative plane \((z < 0)\), see Fig. 1.

Nowadays, there are many VIO frameworks [16]–[18] that support different kinds of camera models including pinhole-, fisheye-, and the omnidirectional camera model introduced by Scaramuzza et al. [19]. While some camera models, e.g., the model introduced by Scaramuzza et al., can support to work with cameras with a negative semi-planar imaging region, all of existing systems discard the points within the negative half plane during the subsequent processing, due to the usage of \([u, v, 1]^T\) to represent the location of image feature points. When the field of the camera reaches the negative half plane, one cannot simply use this representation anymore. This is a severe issue as such a negative semi-planar region may exist in various large-FoV systems like fisheye-, panoramic annular-, and catadioptric cameras, while discarding the features leaves abundant yet important features unused. Besides, \(u\) and \(v\) will increase rapidly near \(180^\circ\), which is disadvantageous to some following algorithms such as PnP and epipolar constraints for solving the rotation matrix \(R\) and the translation vector \(T\), leading to degraded tracking accuracy and even failures.

To tackle this challenge, we propose LF-VIO, a real-time VIO framework for cameras with very large FoV. In LF-VIO, we adapt a KLT sparse flow method [20] to extract feature points. We then propose to use a feature point vector with unit length to represent the features. And the RANSAC method is then used to pick out outliers. Specifically, in the process of initialization, the epipolar geometry is used to initialize if two frames have large enough parallax. After decomposing the essential matrix into rotation matrix and translation vector, the correct rotation matrix and translation vector are selected. Then, triangulation and EPnP alternation methods [21] are introduced to initialize the depth of feature points and every pose in the sliding window, and a tightly coupled optimization method is used to solve all the rotation and translation matrices given by the sliding window. After vision initialization, the IMU data and image data are aligned to recover the scale information. In our approach, we take the visual re-projection error, the IMU pre-integration error, and...
the marginalization error to solve the optimization problem.

To address the scarcity of panoramic visual odometry datasets with ground-truth location and pose, we introduce the \textit{PALVIO} dataset, which is collected via a Panoramic Annular Lens (PAL) system with an entire FoV of $360^\circ \times (40^\circ \sim 120^\circ)$, an IMU sensor, and a motion capture device (see Fig. 2). We conduct extensive quantitative experiments to verify our proposed LF-VIO framework on both the established PALVIO benchmark and a public fisheye camera dataset \cite{15} with a FoV of $360^\circ \times (0^\circ \sim 93.5^\circ)$. Our comprehensive variety of investigations on different FoVs confirms that the information of the negative plane are very important for a VIO system. The proposed method clearly outperforms state-of-the-art VIO frameworks. In addition, we show that our method is also beneficial when it is integrated with a LiDAR-Visual-Inertial Odometry \cite{15}. Furthermore, since feature extraction is performed on the raw panoramic image, bypassing the complex panorama unfolding and the step of cropping the unfolded panorama into 4–5 pinhole sub-images, LF-VIO has a fast speed, which is suitable for real-time mobile robotics applications.

In summary, we deliver the following contributions:

\begin{itemize}
  \item We propose \textit{LF-VIO}, a visual-inertial-odometry framework for cameras with a large field of view.
  \item We present a robust initialization method and algorithmic adjustments to support the feature points on the negative half plane.
  \item We create and release the \textit{PALVIO} dataset, which is collected with a Panoramic Annular Lens (PAL) system and an IMU sensor with the ground-truth location and pose obtained via a motion capture device.
  \item LF-VIO outperforms state-of-the-art visual-inertial-odometry methods on multiple wide-FoV datasets and experiment sequences with a negative plane.
\end{itemize}

\section{Related Work}

In this section, a brief review of representative works is presented on visual inertial odometry and panoramic SLAM (Simultaneous Localization And Mapping) frameworks.

\subsection{Visual inertial odometry}

Visual-Inertial-Odometry (VIO) is the process of a state estimator using vision and inertial measurement unit estimating the 3D pose (local position and orientation) and velocity relative to a local starting position. Qin \textit{et al.} \cite{17} proposed VINS, a robust and universal monocular vision inertial state estimator. Recently, the Semi-direct Visual-inertial Odometry (SVO2.0) \cite{16} has been released for monocular and multicamera systems. Campos \textit{et al.} \cite{18} proposed ORB-SLAM3, an accurate open-source library for vision, visual inertial, and multi-map SLAM. LiDAR-visual-inertial sensor fusion frameworks have also been developed in recent years such as LVI-SAM \cite{15} and R3LIVE \cite{22}. In addition to the feature-point-based methods, there are also direct methods for VO and VIO, which tend to be more sensitive to light than the feature point methods. Engel \textit{et al.} \cite{23} proposed Direct Sparse Odometry (DSO) using a fully direct probabilistic model and Wang \textit{et al.} \cite{24} combined it with inertial systems and stereo cameras. However, most of the works do not directly work well with panoramic camera systems which capture $360^\circ$ omnidirectional contents. Different from these previous works, we address panoramic visual inertial odometry for cameras with extremely large field of view and a negative plane.

\subsection{Panoramic SLAM frameworks}

Panoramic SLAM systems often leverage fisheye cameras, panoramic annular cameras, and multi-camera systems. Sumikura \textit{et al.} \cite{25} proposed OpenVSLAM, which is a VO framework supporting panoramic cameras. Wang \textit{et al.} \cite{26} proposed CubemapSLAM, a piecewise-pinhole monocular fisheye SLAM system. Seok and Lim \cite{4} introduced ROVINS, which uses four fisheye cameras. Chen \textit{et al.} \cite{1}, \cite{27} tackled VO with a panoramic annular camera. More recently, Wang \textit{et al.} \cite{28} proposed PAL-SLAM, which supports panoramic annular cameras. However, the FoV of their used panoramic camera is only $360^\circ \times (45^\circ \sim 85^\circ)$, which does not have a negative half plane. Moreover, based on different sensor types, many other odometry estimation methods emerge in the field, including fisheye-based \cite{29}, \cite{33}, multi-camera-based \cite{3}, \cite{34}–\cite{38}, and systems that combine omnidirectional imaging with other sensing units like GPS \cite{39} and LiDAR sensors \cite{40}, \cite{41}.

Some of the above algorithms have already completed the panoramic image unfolding step, and some algorithms use the direct method or the semi-direct method, which are greatly affected by the changes of illumination. While SVO2.0 \cite{16} and VINS \cite{17} support panoramic annular camera and fisheye camera, both of them convert the model into a pinhole camera model, which makes it impossible to correctly use the important negative half-plane information. Thereby, all the aforementioned VIO works do not support a panoramic annular camera or a catadioptric camera very well. Furthermore, most of the systems that support panoramic annular cameras are still VO frameworks \cite{1}, \cite{5}, \cite{27}, \cite{28}. To fill this gap, we propose a VIO framework to support all types of large-FoV cameras with a negative half plane.

\section{III. LF-VIO: Proposed Framework}

In this section, we describe in detail the proposed LF-VIO framework for large-FoV panoramic cameras with a negative
A. Tracking

In order to obtain the relationship of different image frames, we use a tracking method to establish correspondences between two frames. First, we extract Shi-Tomasi corners [42] and use the Lucas-Kanade method [20] to track these corners. If the corners are out of the border or the quality is lower than a threshold, we will eliminate these points. We then convert these points to a feature vector using a corresponding camera model.

The camera model represents the mathematical relationship between the pixel coordinates on the camera and the image plane. Due to the large FoV, the imaging projection process of a panoramic camera, e.g., the panoramic annular optical system [1], is significantly different from the pinhole camera model. It is necessary to introduce a model suitable for large-FoV imaging which modifies from existing models. For the imaging system, the mapping relationship between the camera coordinate system and the pixel coordinate system can be obtained by calibrating the intrinsic parameters, and the projection relationship between them can be described by the unified spherical model, which is a unit sphere with the camera as its origin. Each point of the pixel coordinate will correspond to a vector whose spherical center points to the unit sphere. For an omnidirectional system, we set the camera center $C$ on the Z-axis and take the quadric plane as the origin. For a space point $P(x, y, z)$ that is projected onto the unit sphere, the projection point $P_s$ is obtained:

$$ P_s = \frac{v}{\|v\|}, \quad (1) $$

$$ v = \begin{bmatrix} x_m \\ y_m \\ a_0 + a_1 \rho + a_2 \rho^2 + \ldots + a_n \rho^n \end{bmatrix}, \quad \rho = \sqrt{x_m^2 + y_m^2}, \quad (2) $$

where the coefficients $a_i$ ($i = 1, 2 \ldots N$) are the polynomial parameters and $x_m, y_m$ denotes the location of a pixel point. These coefficients are given via a panoramic camera calibration process with the OmniCalib calibration toolbox [19].

Since some panoramic annular cameras and fisheye cameras can see the scene behind the camera, that is, the negative half plane of the image ($z < 0$), we propose to use $[\alpha, \beta, \gamma]^T$ to represent the scene spot vector:

$$ \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = P_s = \pi_s^{-1} \begin{bmatrix} x_m \\ y_m \end{bmatrix}, \quad (3) $$

$$ \alpha^2 + \beta^2 + \gamma^2 = 1, \quad (4) $$

where $\pi_s$ means the mapping relationship from unit 3D coordinate point to pixel point. We can also use the MEI [43] or the Kannala-Brandt [44] model, which can also support negative plane to represent this relationship. On the PALVIO dataset, we choose the camera model from Scaramuzza et al. [19] for efficiency and distortion considerations. Afterwards, a series of algorithms need to be adjusted to adapt to work with a camera which has a negative plane.

In the original image, an optical flow method (KLT sparse flow) [20] is leveraged to track the feature points, and the polar geometry RANSAC method is used to eliminate the outlier points:

$$ x_2^T (\hat{T}^T R) x_1 = 0, \quad (5) $$

$$ x_1 = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix}, \quad (6) $$

where $x_1$ and $x_2$ are the unit vectors corresponding to their space points $P_1$ and $P_2$ in Fig. 3, and the notation $\hat{T}$ denotes the skew-symmetric cross product matrix of $T \in \mathbb{R}^3$.

B. Initialization

The initialization procedure, as shown in Fig. 4, includes the three following parts: $R_b^c$ estimation, Vision-only Structure from Motion (SfM), and Visual-inertial alignment. $R_b^c$ estimation: $R_b^c$ and $q_b^c$ means the rotation matrix and quaternion from the camera coordinate to the IMU coordinate. We first maintain several frames in a sliding window to keep a low computational complexity. From frame $k$ and frame $k + 1$, we can obtain the following equation:

$$ q_b^c \otimes q_{b_{k+1}} = q_{b_{k+1}}^c \otimes q_b^c, \quad (7) $$
where $q_{bk}^k$ is computed with IMU pre-integration and $q_{ck}^k$ is computed with epipolar constraint and essential matrix decomposition. From this equation, we can estimate $q_0^k$, and obtain $R_c^k$ according to the commonly used quaternion to rotation matrix algorithm.

**Vision-only SfM:** We first search for two frames which have a large parallax. Then, we use the epipolar constraint to acquire the essential matrix. Subsequently, four sets of solutions are obtained by decomposing the essential matrix. At this time, we use the method depicted in the following equation, to choose the right $R$ and $T$, so that more feature point pairs satisfy the requirement that the dot with the landmarks are always positive in Fig. 5:

$$\overrightarrow{O_1P_1} \cdot \overrightarrow{x_1} > 0 \& \& \overrightarrow{O_2P_2} \cdot \overrightarrow{x_2} > 0.$$ (8)

We then use the obtained $R$ and $T$ to triangulate landmarks \([x_i, y_i, z_i]\): \(\lambda \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = R \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + T.\) (9)

According to the landmarks, the EPnP method [21] is introduced to obtain the $R$ and $T$ between the initialization frames. After that, triangulation and EPnP methods are used alternately. By analogy, the three-dimensional spatial positions of every $R$ and $T$ and landmarks in the sliding window are acquired.

Finally, the re-projection error method is used to optimize the $R$ and $T$ of all frames in the sliding window:

$$\min_{R,T} \sum_{i=1}^{M} \sum_{j=1}^{N} \left\| \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix} - \frac{R P_j + T}{||R P_j + T||_2} \right\|^2,$$ (10)

where $M$ is the number of observations and $N$ is the number of landmarks.

**Visual-inertial alignment.** An illustration of the visual-inertial alignment is shown in Fig. 6. We match the up-to-scale visual structure with IMU pre-integration. Then, we calibrate the gyroscope bias, initialize velocity, gravity vector and metric scale, and finally align the gravity direction with the $Z$ axis. Because the feature vector is expressed by a unit vector, we innovatively design an initialization method and adjust a series of algorithms regarding epipolar constraint, triangulation, and re-projection error.
TABLE I

| Field of View | 40°~120° | 40°~110° | 40°~100° | 40°~90° | 40°~80° |
|---------------|----------|----------|----------|---------|---------|
| RPEt (%)      | 2.814    | 3.097    | 2.923    | 2.933   | 3.392   |
| RPEz (degree/m) | 0.397    | 0.393    | 0.402    | 0.433   | 0.631   |
| ATE (m)       | 0.093    | 0.182    | 0.112    | 0.124   | 0.171   |

| Field of View | 50°~120° | 60°~120° | 70°~120° | 80°~120° | 90°~120° |
|---------------|----------|----------|----------|---------|---------|
| RPEt (%)      | 2.585    | 2.712    | 3.584    | 3.525   | 5.223   |
| RPEz (degree/m) | 0.394    | 0.398    | 0.417    | 0.449   | 0.490   |
| ATE (m)       | 0.081    | 0.117    | 0.250    | 0.252   | 0.475   |

IV. EXPERIMENTS

A. Datasets

PALVIO dataset. To address the lack of panoramic visual odometry dataset with ground-truth location and pose, we collect and release our PALVIO dataset. PALVIO is collected with two panoramic annular cameras (see Fig. 2), an IMU sensor, and a RealSense D435 sensor with ground truth under motion capture in a computing platform with a quad-core Intel i7-8550U processor. The panoramic cameras capture monocular 1280×960 images at 30Hz, with a FoV of 360°×(40°~120°). The IMU provides the angular velocity and acceleration at 200Hz and the motion capture system provides the position and attitude at 10Hz. We collect 10 sequences in an area of of 8m×10m indoors (ID). For the vision data, the data captured by the stereo panoramic cameras and the RealSense camera are both made publicly available. In this work, we mainly use the data captured by the top panoramic camera for experiments.

LVI-SAM dataset. We use a public fisheye dataset LVI-SAM [15] to verify the generalizability of our approach. The data collection sensor suite includes: a Velodyne VLP-16 LiDAR sensor, an FLIR BFS-U3-04S2M-CS camera, a MicroStrain 3DM-GX5-25 IMU sensor, and Reach RS+ GPS. Jackal- and Handheld datasets are gathered by an unmanned ground vehicle. The FoV of their used fisheye camera is approximately 360°×(0°~93.5°). We use the LVI-SAM dataset together with their GPS measurements, which are treated as the ground truth for experiments.

B. Investigation on Different Field-of-View

To validate the importance of the information of the negative plane, we study the influence of using inputs with different FoVs, by only extracting features from the corresponding angle range. First, we gradually decrease the FoV from 40°~120° to 40°~80°, with a variation step of 10° (see Fig. 7), and perform the experiment using our LF-VIO framework on the PALVIO ID06 set, which is chosen as a representative sequence, where the results are shown in Table I. It can be seen that using the entire FoV of the panoramic camera (40°~120°) leads to the best performances in RPEt and ATE. When the FoV reduces to only the positive plane, i.e., in the situations of 40°~90° and 40°~80°, the performance degrades dramatically, verifying the importance of the information from the negative plane.

IV. EXPERIMENTS

A. Datasets

PALVIO dataset. To address the lack of panoramic visual odometry dataset with ground-truth location and pose, we collect and release our PALVIO dataset. PALVIO is collected with two panoramic annular cameras (see Fig. 2), an IMU sensor, and a RealSense D435 sensor with ground truth under motion capture in a computing platform with a quad-core Intel i7-8550U processor. The panoramic cameras capture monocular 1280×960 images at 30Hz, with a FoV of 360°×(40°~120°). The IMU provides the angular velocity and acceleration at 200Hz and the motion capture system provides the position and attitude at 10Hz. We collect 10 sequences in an area of of 8m×10m indoors (ID). For the vision data, the data captured by the stereo panoramic cameras and the RealSense camera are both made publicly available. In this work, we mainly use the data captured by the top panoramic camera for experiments.

LVI-SAM dataset. We use a public fisheye dataset LVI-SAM [15] to verify the generalizability of our approach. The data collection sensor suite includes: a Velodyne VLP-16 LiDAR sensor, an FLIR BFS-U3-04S2M-CS camera, a MicroStrain 3DM-GX5-25 IMU sensor, and Reach RS+ GPS. Jackal- and Handheld datasets are gathered by an unmanned ground vehicle. The FoV of their used fisheye camera is approximately 360°×(0°~93.5°). We use the LVI-SAM dataset together with their GPS measurements, which are treated as the ground truth for experiments.

B. Investigation on Different Field-of-View

To validate the importance of the information of the negative plane, we study the influence of using inputs with different FoVs, by only extracting features from the corresponding angle range. First, we gradually decrease the FoV from 40°~120° to 40°~80°, with a variation step of 10° (see Fig. 7), and perform the experiment using our LF-VIO framework on the PALVIO ID06 set, which is chosen as a representative sequence, where the results are shown in Table I. It can be seen that using the entire FoV of the panoramic camera (40°~120°) leads to the best performances in RPEt and ATE. When the FoV reduces to only the positive plane, i.e., in the situations of 40°~90° and 40°~80°, the performance degrades dramatically, verifying the importance of the information from the negative plane.

Then, we experiment by decreasing the FoV from 50°~120° to 90°~120°. In the worst case, when only the information from the negative plane is used (90°~120°), the VIO framework still works. But when incorporating slightly more information from the positive plane, i.e., in the cases of 80°~120° and 70°~120°, the performance recovers to the same level of that with pure positive plane. This again confirms the importance of information lying on the negative plane and the effectiveness of our method in properly harvesting and exploiting the negative-plane features, which are unused in previous works. Fig. 8 shows the top trajectories and error analyses based on the FoV variations, which illustrates the same trend. We note that the performance of the 50°~120° setting does not exceed that of 40°~120°, which is due to the lowest density (the number of vectors per unit area) in 40°~50°, so the use of information of this band may not improve the performance.

C. Comparison with the State-of-the-Art

With the data captured by the motion capture system as the ground truth, we provide the accuracy of our LF-VIO on the PALVIO benchmark, compared with state-of-the-art visual-inertial-odometry methods SVO2.0 [16] and VINS-Mono [17], in Table II, where the processing are conducted on a laptop with an R7-5800H processor. For fair comparisons, we use the camera model from Scaramuzza et al. in both SVO2.0 and VINS-Mono. We also report the results of LF-VIO using only information from the positive plane (40°~90°). We use Relative Pose Error in translation...
It can be seen that our method reaches the best precision in RPEt, RPEr, and ATE on most sequences, confirming the superior performance of LF-VIO. It is clearly seen that LF-VIO has a higher accuracy and robustness in situations which have relatively fast rotation and it also has a good performance in low-rotation scenarios. In summary, our approach is verified in almost all sequences, while the other two cannot always produce robust pose estimation. Further, we choose ID01, ID06, and ID10, as three typical and representative datasets for analysis in Fig. 9. In ID01 and ID10, the maximum Z-axis angular velocity reaches to 2.1 and 2.3 rad/s, and the images begin to become blurry. Thus, the accuracy of all the algorithms is less satisfactory. However, our method shows the relatively high accuracy in RPEt, RPEr, and ATE. In ID06, the maximum Z-axis angular velocity only reaches 1.5 rad and the ATE of all the compared algorithms are rather low. Moreover, in ID10, SVO2.0 has a poor performance under high angular velocity, due to unstable feature points and scale estimation in the back-end, but our system still maintains robust. Generally, our method has a higher accuracy and robustness in situations which have relatively fast rotation and it also has a good performance in low-rotation scenarios. In summary, our approach is verified...
Handheld  LF-LVI-SAM (Ours)  0.63857 0.00660 4.38248 0.84366 0.00393 0.00007 0.02180 0.00510 8.89108 0.08698 25.30987 11.57598  
LVI-SAM [15]  0.67193 0.01113 4.37260 0.87484 0.00397 0.00010 0.01947 0.00511 10.78413 0.09040 26.72794 13.52003  

Jackal  LF-LVI-SAM (Ours)  0.70924 0.00948 2.83097 0.79693 0.00265 0.00047 0.00761 0.00283 5.80506 0.80519 14.21643 6.28478  
LVI-SAM [15]  0.73586 0.01300 2.80597 0.82791 0.00274 0.00047 0.00781 0.00291 6.02435 0.65769 14.53112 6.53074  

TABLE III  
COMPARISON OF VIO METHODS ON THE LVI-SAM DATASET [15] (WITHOUT LOOP).  

D. Generalization to LiDAR-Visual-Inertial Odometry  

The proposed method supporting negative half plane can also be transplanted to and integrated with LiDAR-Visual-Inertial Systems, such as LVI-SAM [15]. We compare the accuracy of this system with and without our proposed method in Table III on the two sequences, i.e., Handheld- and Jackal datasets, provided by [15], where the results are obtained via a laptop with an R7-5800H processor. Their recording camera has a FoV of 360° (0° to 360°). 

In this experiment, we only change the visual part of LVI-SAM. We use a unit vector to represent feature points and adjust visual algorithms about it and the MEI camera model [43] following the original setup [15]. LVI-SAM [15] uses a mask that has many large field area, in order to avoid the problems coming with the feature points on the negative plane or those close to 90°. In our system, this challenge is solved, so we reduce the area of the mask to explore the best performance of our system. It can be seen that LF-LVI-SAM is consistently better than LVI-SAM in Mean and RMSE of RPE (%), RPEr (degree/m) and ATE (m). The RPE on the Handheld- and Jackal datasets are shown in Fig. 10 and Fig. 11. In summary, our method can not only improve the precision of VIO systems, but also be effective for LiDAR-VIO systems whose FoV reaches the negative half plane.
E. Speed Analysis
We report the efficiency of our approach in this subsection, which mainly composed of the front-end and the back-end. If the mapping relationship of the images is stored in memory, the image mapping process takes about 1~2ms. In the front-end of LF-VIO, this operation is not necessary, but the feature tracking and optical flow computation cost 15ms and 3ms respectively, and the total front-end costs about 40ms. In the back-end, the solver costs around 30ms mainly and the total costs about 60ms. Besides, the frond-end and back-end operate independently. In general, our LF-VIO system can reach at least 10Hz on the onboard computer with a quad-core Intel i7-8550U processor, which is reasonable for real-time mobile applications.

V. Conclusions
In this paper, we have proposed LF-VIO, a framework for large-FoV cameras with a negative plane, for performing real-time pose estimation. The proposed framework is composed of two parts: a robust initialization process and a tightly coupled optimization process. We create and release the PALVIO dataset with 10 sequences collected via a Panoramic Annular Lens (PAL) camera and an IMU sensor, for evaluating the performance of our proposed framework and fostering future research on panoramic visual inertial odometry. Our method improves the system robustness and accuracy, outperforming state-of-the-art VIO approaches. Our method generalizes well to other SLAM systems which involve large-FoV sensing and the feature point method, as proved when it is combined with LVI-SAM. Our implementation has been open sourced. In the future, we are interested in adding loop closure and combing the stereo panoramic information and LiDAR sensors in our system.

REFERENCES
[1] H. Chen et al., “PALVO: visual odometry based on panoramic annular lens,” OE, 2019.
[2] W. Hu, K. Wang, H. Chen, R. Cheng, and K. Yang, “An indoor positioning framework based on panoramic visual odometry for visually impaired people,” MST, 2019.
[3] H. Seok and J. Lim, “ROVO: Robust omnidirectional visual odometry for wide-baseline wide-FOV camera systems,” in ICRA, 2019.
[4] ——, “ROVINS: Robust omnidirectional visual inertial navigation system,” RA-L, 2020.
[5] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen, “DS-PASS: Detail-sensitive panoramic annular semantic segmentation through SwatNet for surrounding sensing,” in IV, 2020.
[6] M. Lin, Q. Cao, and H. Zhang, “PVO: Panoramic visual odometry,” in ICARM, 2018.
[7] B. Gao, D. Wang, B. Lian, and C. Tang, “LOVINS: Lightweight omnidirectional visual-inertial navigation system,” in ICSPPC, 2021.
[8] H. Chen, K. Yang, W. Hu, J. Bai, and K. Wang, “Semantic visual odometry based on panoramic annular imaging,” AOS, 2021.
[9] K. Yang, J. Zhang, S. Reif, X. Hu, and R. Stiefelhagen, “Capturing omni-range context for omnidirectional segmentation,” in CVPR, 2021.
[10] A. Jaus, K. Yang, and R. Stiefelhagen, “Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning,” in IV, 2021.
[11] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, “Omnisupervised omnidirectional semantic segmentation,” T-ITS, 2022.
[12] H. Shi et al., “PanoFlow: Learning optical flow for panoramic images,” arXiv preprint arXiv:2002.12385, 2020.
[13] D. Sun, X. Huang, and K. Yang, “A multimodal vision sensor for autonomous driving,” in SPIE, 2019.
[14] K. Yang, X. Hu, M. Bergasa, E. Romera, and K. Wang, “PASS: Panoramic annular semantic segmentation,” T-ITS, 2020.
[15] T. Shan, B. Englert, C. Ratti, and D. Rus, “LVI-SAM: Tightly-coupled LiDAR-visual-inertial odometry via smoothing and mapping,” in ICRA, 2021.
[16] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual-inertial odometry,” in ICRA, 2014.
[17] T. Qin, P. Li, and S. Shen, “VINS-Mono: A robust and versatile monocular visual-inertial state estimator,” T-RO, 2018.
[18] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM,” T-RO, 2021.
[19] D. Scaramuzza, A. Martinelli, and R. Siegwart, “A toolbox for easily calibrating omnidirectional cameras,” in IROS, 2006.
[20] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in IJCAI, 1981.
[21] V. Lepetit, F. Moreno-Noguer, and P. Fua, “EnPnP: An accurate O(n) solution to the pnp problem,” IJCV, 2009.
[22] J. Lin and F. Zhang, “RSLIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package,” arXiv preprint arXiv:2109.07982, 2021.
[23] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” TPAMI, 2018.
[24] R. Wang, M. Schwerer, and D. Cremers, “Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras,” in ICCV, 2017.
[25] S. Sumikura, M. Shibuya, and K. Sakurada, “OpenVSLAM: A versatile visual SLAM framework,” in MM, 2019.
[26] Y. Wang et al., “CubemapSLAM: A piecewise-pinhole monocular fisheye SLAM system,” in ACCV, 2018.
[27] H. Chen, W. Hu, K. Yang, J. Bai, and K. Wang, “Panoramic annular SLAM with loop closure and global optimization,” AO, 2021.
[28] D. Wang, J. Wang, Y. Tian, K. Hu, and M. Xu, “PAL-SLAM: a feature-based SLAM system for a panoramic annular lens,” OE, 2022.
[29] L. Heng and B. Choi, “Semi-direct visual odometry for a fisheye-stereo camera,” in IROS, 2016.
[30] P. Liu, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, “Direct visual odometry for a fisheye-stereo camera,” in IROS, 2017.
[31] H. Matsuki, L. Von Stumberg, V. Usenko, J. Stücker, and D. Cremers, “Omnidirectional DSO: Direct sparse odometry with fisheye cameras,” RA-L, 2018.
[32] S. Wang, J. Yue, Y. Dong, R. Shen, and X. Zhang, “Real-time omnidirectional visual SLAM with semi-dense mapping,” in IV, 2018.
[33] M. Ramezani, K. Khoshelham, and C. Fraser, “Pose estimation by omnidirectional visual-inertial odometry,” Robotics and Autonomous Systems, 2018.
[34] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, “Towards robust visual odometry with a multi-camera system,” in IROS, 2018.
[35] C. Jaramillo, L. Yang, J. P. Muñoz, Y. Taguchi, and J. Xiao, “Visual odometry with a single-camera stereo omnidirectional system,” in MVA, 2019.
[36] S. Yogamani et al., “WoodScape: A multi-task, multi-camera fisheye dataset for autonomous driving,” in ICCV, 2019.
[37] C. Won, H. Seok, Z. Cui, M. Pollefeys, and J. Lim, “OmniSLAM: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems,” in CVPR, 2020.
[38] V. R. Kumar et al., “OmniDet: Surround view cameras based multi-task visual perception network for autonomous driving,” RA-L, 2021.
[39] Y. Yu, W. Gao, C. Liu, S. Shen, and M. Liu, “A GPS-aided omnidirectional visual-inertial state estimator in ubiquitous environments,” in IROS, 2019.
[40] J. Kang, Y. Zhang, Z. Liu, A. Sit, and G. Sohn, “RPV-SLAM: Range-augmented panoramic visual SLAM for mobile mapping system with panoramic camera and tilted LiDAR,” in ICAR, 2021.
[41] Z. Xiang, Y. Ju, J. Li, and J. Su, “ViLiVo: Virtual LiDAR-visual odometry for an autonomous vehicle with a multi-camera system,” in IROS, 2019.
[42] J. Shi and C. Tomasi, “Good features to track,” in CVPR, 1994.
[43] C. Mei and P. Rives, “Single view point omnidirectional camera calibration from planar grids,” in IROS, 2007.
[44] J. Kannala and S. S. Brandt, “A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses,” TPAMI, 2006.
[45] S. Agarwal and K. Mierle, “Ceres solver;” http://ceres-solver.org.