2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Direct Sparse Stereo Visual-Inertial Global Odometry

Ziqiang Wang, Mei Li, Dingkun Zhou, Ziqiang Zheng

***Abstract*—Robust and accurate localization plays a key role in autonomous driving and robot applications. To utilize the complementary properties of different sensors, we present a novel tightly-coupled approach to combine the local (stereo cameras, IMU) and global sensors (magnetometer, GNSS). We jointly optimize all the model parameters through one active window. The visual part integrates constraints from static stereo into the photometric bundle adjustment pipeline of dynamic multi-view stereo. Accumulating IMU information between keyframes, magnetometer and GNSS measurements are all inserted into the active window as additional constrains among all the keyframes. Through these, our method can realize globally drift-free and locally accurate state estimation. We evaluate the effectiveness of our system on public datasets under with real-world experiments.**

I. Introduction

Robust and accurate localization is the basic requirement for the current existing self-driving systems. Since visual cameras are cheap and lightweight, they have drawn a huge attention of the robotics community. Over the past few decades, the pure visual odometry methods have achieved excellent progress. They can mainly be divided into two categories: 1) indirect and 2) direct methods. The former contain such representative works: MonoSLAM [1], PTAM [2] and ORB-SLAM [3]. The latter direct methods skip the pre-processing procedure and directly adopt the actual sensor values, which free from the need to extract and match feature points for reducing the additional computational consume. Compared with the indirect method, which is purely dependent on the pre-defined feature points, the direct methods (LSD-SLAM [4] and DSO [5]) can observe more various image information(including the edges and weak intensity variations). The redundant information resource can lead to more robustness under sparsely texture environments.

The monocular vision-only systems are incapable of recovering the metric scale due to lack of image depth information. By combining the inertial measurement unit (IMU), the monocular visual-inertial algorithms [6], [7] can achieve an accurate perception of the metric scale. One unavoidable dilemma of these methods is the metric scale factor convergence problem, which requires enough rotation or translation, otherwise the whole system cannot be well initialized and then fall into collapse. This is not user-friendly for autonomous driving or robotic applications. As a reason that the camera can obtain metric scale upon capturing first stereo photo, we can speed up the system initialization through one stereo camera. To utilize this intrinsic advantage, the stereo-inertial configuration of ORB-SLAM3 [8] provides more robust and accurate results than monocular, stereo and monocular inertial.

The authors are with the UISEE (Shanghai) Automotive Technologies LTD, 201800 Shanghai, China. Email: wang.ziqiang@qq.com {mei.li,dingkun.zhou,ziqiang.zheng}@uisee.com

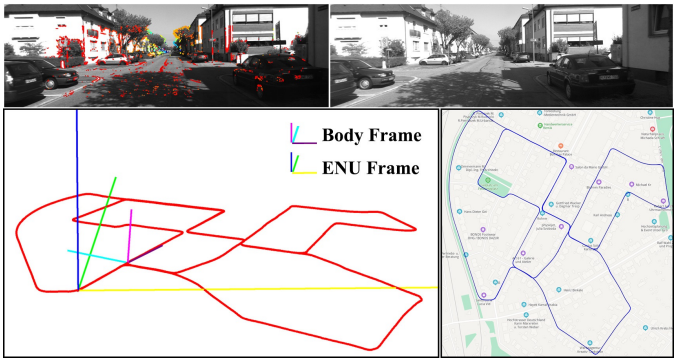

Fig. 1. KITTI dataset results of the proposed method. The top of this figure is a pair of stereo images, left image is coupled with sparse inverse depth map. The left bottom part is the estimated trajectory, while the right bottom part is the estimated global trajectory aligned with Google map.

Although these algorithms have conducted accurate state estimation within a local region, there are still several drawbacks in practice. One explicit disadvantage of local pose estimation is the poor reusability. It is extremely difficult to reuse the pose produced in local frames without fixed global coordinates. Besides, the local estimations are prone to cumulative drift in long distance travel. The loop closures [9] can alleviate the accumulated drift by increasing the computational complexity and memory requirements. Compared with local sensors, global sensors (such as GNSS and magnetometer) have a large superiority under the large-scale environment setting. They provide global measurements for fixed earth frames, which are drift-free as described in VINS-Fusion [10]. However, their measurements are usually not smooth, noisy, and cannot be directly used for precise localization. It is feasible to combine both the local and global sensors information to make good use of their complementary properties. In this paper, we propose a novel tightly-coupled direct sparse stereo visual-inertial global odometry termed as **DVIGO**. DVIGO is an optimization framework to fuse local estimations with global sensor measurements, while the results are shown in Fig. 1. To sum up, our main contributions are listed as follows:

- To the best of our knowledge, the proposed DVIGO is the first tightly-coupling multi-sensor fusion work based on direct sparse method, which achieves locally accurate and globally drift-free localization. We precisely fuse the measurements from the four complementary but asynchronous sensors (stereo camera, IMU, magnetometer and GNSS).
- We perform comprehensive quantitative evaluations on the KITTI and the EuRoC dataset. Comparisons to state-of-the-art methods like VINS-Fusion demonstrate the

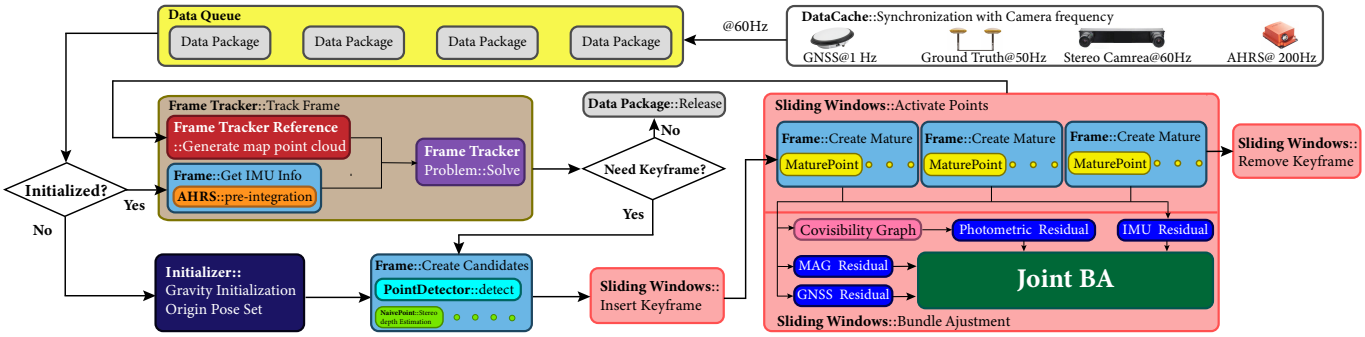

Fig. 2. The overview framework of our DVIGO.

superiority of the proposed method. In addition, we also conduct experiments under various real world scenarios (including both daytime and nighttime outdoor settings) to validate the effectiveness of our method.

II. Related Work

Recently, ORB-SLAM [3], [11]–[13] introduced an efficient visual SLAM solution based on ORB feature descriptor [14] and map reuse. Another category visual odometry use unprocessed intensities in the image to estimate the motion of the camera called direct methods. and were also applied to various works, in a dense [15], semi-dense [4], [16]–[18] and sparse [19]–[21]. Furthermore, The persistent map of DSM [22] handles reobservations, yielding the most accurate result on EuRoC [23] dataset through the direct manner.

IMU pre-integration was proposed by Lupton and Sukkarieh [24]. Forster *et al.* extended it to Lie group [25], and completed the combination with SVO [26]. Due to high computational efficiency, IMU pre-integration has been widely applied in VIO based on both direct and indirect optimization frameworks, including VI-ORB [8], [27], [28],VINS [6], [29], VI-LSD [30], VI-DSO [31]. Besides the above optimization-based frameworks, the filtering-based approaches contain multi-state constraint kalman filter (MSCKF) [32], [33], ROVIO [34] and OpenVINS [35].

Most visual-inertial global sensors fusion for navigation systems can also be divided into filtering-based and optimization-based approaches. Lee *et al.* [36] proposed a tightly-coupled MSCKF-based estimator to fuse the inertial, camera and asynchronous GNSS measurements. VINS-Fusion [10], Shin *et al.* [37], and GOMSF [38] adopted an optimization-based framework to fuse local VIO and global measurements.

III. Math

The proposed DVIGO aims to estimate poses of several consecutive IMU frames, velocities, biases, affine brightness parameters, camera intrinsics and landmarks' inverse depth within a sliding window. The state χ of our system at time i is described by:

$$\begin{aligned}\chi &= [\mathbf{x}_{i-N+1}, \mathbf{x}_{i-N+2}, \cdots, \mathbf{x}_i, d_{\mathbf{u}_0}, d_{\mathbf{u}_1} \cdots, d_{\mathbf{u}_M}, \mathbf{K}] \\ \mathbf{x}_i &= [{}_{\mathtt{WB}}\mathbf{R}_i, {}_{\mathtt{W}}\mathbf{p}_i, {}_{\mathtt{W}}\mathbf{v}_i, \mathbf{b}_i^g, \mathbf{b}_i^a, a_i^L, b_i^L, a_i^R, b_i^R],\end{aligned} \tag{1}$$

where N, M are keyframes and landmarks size of the sliding window. ${}_{\mathtt{WB}}\mathbf{T}_i = ({}_{\mathtt{WB}}\mathbf{R}_i, {}_{\mathtt{W}}\mathbf{p}_i) \in SE(3)$ indicate the rotation and translation from world frame (abbreviated as $\mathtt{W}$) to IMU body frame ($\mathtt{B}$). We identify the skew symmetric matrix with a vector $\mathbb{R}^3$ using hat operator ($\wedge$). Exponential map ($\mathtt{Exp}$) and logarithm ($\mathtt{Log}$) operator associate $\phi^\wedge \in \mathfrak{so}(3)$ to group $SO(3)$, whose right jacobian is $\mathtt{J}_r(\phi)$. Δt is the IMU sampling interval. The gyroscope and accelerometer measurements at time k (${}_{\mathtt{B}}\widetilde{\mathbf{a}}_k$ and ${}_{\mathtt{B}}\widetilde{\boldsymbol{\omega}}_k$) are affected by white noise and a slowly varying gyroscope and accelerometer bias ($\mathbf{b}_i^g, \mathbf{b}_i^a \in \mathbb{R}^3$). Homogeneous camera intrinsics is denoted as $\mathbf{K}$, while $d_{\mathbf{u}}$ is the inverse depth of a camera frame ($\mathtt{C}$) landmark $\mathbf{u}$. $a_i^L, b_i^L, a_i^R, b_i^R$ are affine brightness parameters left and right image for keyframe I_i.

IV. Methodology

Our DVIGO is rebuild on [39], DSM [22], DSO [5] and [25]. Fig. 2 demonstrates the overall framework. We summarize our method as three main procedures:

- **Data thread** caches all sensor information and synchronizes them according to the timestamp. Once GNSS data received, they will be interpolated into a data package, which contains the stereo images, several accelerometer, gyroscope and magnetometer observations.
- **Tracking thread** processes coarse pyramid tracking and computes the pose of the current frame with respect to the reference kayframe. It also determines whether the current frame belongs to the keyframe. When a new keyframe is created, all accelerometer and gyroscope data between two consecutive keyframes are pre-integrated following to [25].
- **Joint bundle adjustment thread** adds keyframes and landmarks to the sliding window, removes the redundant landmarks. For all keyframes in the current active window, a joint optimization of system state in (1), is performed to minimize the photometric, inertial, magnetic and GNSS error.

A. Local pose estimation

1) IMU Factors: There are m gyroscope, n accelerometer data ($m \neq n$) between two consecutive keyframes in Fig. 3. The pre-integrated relative rotation increment $\Delta\widetilde{\mathbf{R}}_{ij}$ is computed iteratively by rotational velocities. Considering the biases $\bar{\mathbf{b}}$ updated by a small amount $\delta\mathbf{b}$ during optimization, *i.e.*, $\mathbf{b} =$

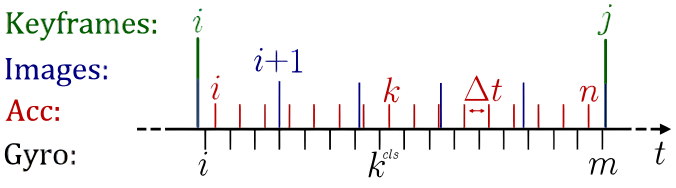

Fig. 3. There are different quantity samples of IMU. We take strategy that for accelerometer sample in time k, combining with time-closest k^{cls} gyro sensor sample.

$\bar{\mathbf{b}}+\delta\mathbf{b}$, we divide the pre-integrated rotation measurement into constant and perturbation counterparts:

$$\begin{aligned}
&\tfrac{\partial\Delta\bar{\mathbf{R}}_{im}}{\partial\mathbf{b}^g}=\sum_{k=i}^{m}\Delta\bar{\mathbf{R}}_{(k+1)m}^{T}\mathbf{J}_r((\widetilde{\omega}_k-\bar{\mathbf{b}}_i^g)\Delta t)\Delta t\\
&\Delta\widetilde{\mathbf{R}}_{im}(\bar{\mathbf{b}}_i^g)=\Delta\bar{\mathbf{R}}_{im}=\prod_{k=i}^{m}\mathtt{Exp}((\tilde{\omega}_k-\bar{\mathbf{b}}_i^g)\Delta t)\\
&\Delta\widetilde{\mathbf{R}}_{im}(\delta\mathbf{b}_i^g)=\Delta\widehat{\mathbf{R}}_{im}=\mathtt{Exp}(\tfrac{\partial\Delta\bar{\mathbf{R}}_{im}}{\partial\mathbf{b}^g}\delta\mathbf{b}_i^g),
\end{aligned}\tag{2}$$

we can obtain the pre-integrated rotation: $\Delta\widetilde{\mathbf{R}}_{ij}=\Delta\bar{\mathbf{R}}_{im}\Delta\widehat{\mathbf{R}}_{im}$. As a reason of that the pre-integrated relative velocity increment $\Delta\widetilde{\mathbf{v}}_{ij}$ is relevant to both gyro and accelerometer, we choose the time-closest $\Delta\bar{\mathbf{R}}_{ik}^{cls}$ and $\frac{\partial\Delta\bar{\mathbf{R}}_{ik}^{cls}}{\partial\mathbf{b}^g}$ for an accelerometer data and recursively calculate:

$$\begin{aligned}
&\Delta\bar{\mathbf{v}}_{in}=\sum_{k=i}^{n}\Delta\bar{\mathbf{R}}_{ik}^{cls}((\widetilde{\mathbf{a}}_k-\bar{\mathbf{b}}_i^a)\Delta t),\ \tfrac{\partial\Delta\bar{\mathbf{v}}_{in}}{\partial\mathbf{b}^a}=\sum_{k=i}^{n}-\Delta\bar{\mathbf{R}}_{ik}^{cls}\Delta t\\
&\tfrac{\partial\Delta\bar{\mathbf{v}}_{in}}{\partial\mathbf{b}^g}=\sum_{k=i}^{n}-\Delta\bar{\mathbf{R}}_{ik}^{cls}(\widetilde{\mathbf{a}}_k-\bar{\mathbf{b}}_i^a)^{\wedge}\tfrac{\partial\Delta\bar{\mathbf{R}}_{ik}^{cls}}{\partial\mathbf{b}^g}\Delta t\\
&\Delta\widehat{\mathbf{v}}_{in}=\tfrac{\partial\Delta\bar{\mathbf{v}}_{in}}{\partial\mathbf{b}^a}\delta\mathbf{b}_i^a+\tfrac{\partial\Delta\bar{\mathbf{v}}_{in}}{\partial\mathbf{b}^g}\delta\mathbf{b}_i^g,
\end{aligned}\tag{3}$$

we can get pre-integrated velocity $\Delta\widetilde{\mathbf{v}}_{ij}=\Delta\bar{\mathbf{v}}_{in}+\Delta\widehat{\mathbf{v}}_{in},$. After repeating the same process for $\Delta\widetilde{\mathbf{p}}_{ij}$, now it is easy to write the final residual errors: $\mathbf{r}_{I_{ij}}=[\mathbf{r}_{\Delta\mathbf{R}_{ij}}^T,\mathbf{r}_{\Delta\mathbf{v}_{ij}}^T,\mathbf{r}_{\Delta\mathbf{p}_{ij}}^T]\in\mathbb{R}^9$ as:

$$\begin{aligned}
&\mathbf{r}_{\Delta\mathbf{R}_{ij}}=\mathtt{Log}((\Delta\widetilde{\mathbf{R}}_{ij})^T\mathbf{R}_i^T\mathbf{R}_j)\\
&\mathbf{r}_{\Delta\mathbf{v}_{ij}}=\mathbf{R}_i^T(\mathbf{v}_j-\mathbf{v}_i-\mathbf{g}\Delta t_{ij})-\Delta\widetilde{\mathbf{v}}_{ij}\\
&\mathbf{r}_{\Delta\mathbf{p}_{ij}}=\mathbf{R}_i^T(\mathbf{p}_j-\mathbf{p}_i-\mathbf{v}_i\Delta t_{ij}-\tfrac{1}{2}\mathbf{g}\Delta t_{ij}^2)-\Delta\widetilde{\mathbf{p}}_{ij}.
\end{aligned}\tag{4}$$

We can get covariance Σ_{ij} recursively like previous calculations of asynchronized data. We model the bias through the "random walk". Covariance Σ_{bg} and Σ_{ba} are determined by discrete bias noises. We can formulate the bias residual error term as:

$$\mathbf{r}_{\mathbf{b}_{ij}}^g=\mathbf{b}_j^g-\mathbf{b}_i^g,\mathbf{r}_{\mathbf{b}_{ij}}^a=\mathbf{b}_j^a-\mathbf{b}_i^a.\tag{5}$$

2) Photometric Factors: For visual part of our system, we apply visual tracking strategy:

- We track the motion of the camera towards a reference keyframe and create a new keyframe according to DSO and whether the data package contains GNSS message.
- We estimate the inverse depth of selected points of the current reference keyframe from static and dynamic stereo cues.

Once creating a new keyframe, a sparse set of points selected from the image, which has sufficient image gradient, are termed as naive points. This keyframe will become the hostframe of these selected points. To make sure that the distribution of the selected points is sparse and even, the image is cropped to several small (16×12) blocks, and an adaptive threshold is adopted for each small block. Suppose a 2D image coordinate point $\mathbf{u}$ is selected in hostframe I_i and observed in corresponding stereo right image I_i^R. We search for the corresponding pixel on the epipolar line in I_i^R, obtaining the pixel with the highest NCC criterion similarity as the matching point. Then an inverse depth initialization $(d_{\mathbf{u}})$ of $\mathbf{u}$ can be calculated by using triangle calculating method. Furthermore, $\mathbf{u}$ can be projected to I_i^R as:

$$\mathbf{u}_s^{'}=d_{\mathbf{u}_s^{'}}\mathbf{K}({}_{\mathtt{RL}}\mathbf{T}((d_{\mathbf{u}})^{-1}\mathbf{K}^{-1}\mathbf{u})),\tag{6}$$

where $d_{\mathbf{u}_s^{'}}$ is inverse depth of $\mathbf{u}_s^{'}$, ${}_{\mathtt{RL}}\mathbf{T}$ is relative transformation between the left and right cameras.Then static one-view stereo residuals are defined as:

$$r_{\mathbf{u}}^s=I_i^R(\mathbf{u}_s^{'})-b_i^R-\tfrac{e^{a_i^R}}{e^{a_i^L}}(I_i(\mathbf{u})-b_i^L),\tag{7}$$

where a_i^L,b_i^L,a_i^R,b_i^R is the affine brightness parameters of frame I_i,I_i^R. Assuming that $\mathbf{u}$ can also be projected to another keyframe I_j, we use the IMU body frame pose $\mathbf{T}_i$ to map $\mathbf{u}$ from I_i to I_j as:

$$\mathbf{u}_d^{'}=d_{\mathbf{u}_d^{'}}\mathbf{K}({}_{\mathtt{BC}}\mathbf{T}^{-1}\mathbf{T}_j^{-1}\mathbf{T}_i((d_{\mathbf{u}})^{-1}{}_{\mathtt{BC}}\mathbf{T}\mathbf{K}^{-1}\mathbf{u})),\tag{8}$$

where $d_{\mathbf{u}_d^{'}}$ is inverse depth of $\mathbf{u}_d^{'}$, ${}_{\mathtt{BC}}\mathbf{T}$ is relative transformation between left camera and IMU. The dynamic multi-view residuals are defined as:

$$(r_{\mathbf{u}}^d)_{ij}=I_j(\mathbf{u}_d^{'})-b_j^L-\tfrac{e^{a_j^L}}{e^{a_i^L}}(I_i(\mathbf{u})-b_i^L),\tag{9}$$

where a_j^L,b_j^L is the affine brightness parameters of frame I_j.

B. Global Pose estimation

1) Magnetometer Factor: The magnetometer can help to determine the orientation in the world frame. Assuming that the magnetometer is calibrated offline without offset and bias, the magnetometer output ${}_{\mathtt{B}}\mathbf{m}_i$ at time i can provide partial information of the state rotation matrix as:

$${}_{\mathtt{B}}\mathbf{m}_i={}_{\mathtt{WB}}\mathbf{R}_i\ {}_{\mathtt{W}}\mathring{\mathbf{m}},\tag{10}$$

where ${}_{\mathtt{W}}\mathring{\mathbf{m}}$ is the approximately constant magnetic field of the earth at the position of the body expressed in the ENU (local East North Up) coordinates. Since the magnetic field is easily affected by the environment, we only adopt the normalized vector without length. The length is used to determine covariance Σ_m. We can easily formulate the residual error as:

$$\mathbf{r}_m={}_{\mathtt{B}}\widetilde{\mathbf{m}}_i-{}_{\mathtt{WB}}\mathbf{R}_i\ {}_{\mathtt{W}}\mathring{\mathbf{m}}.\tag{11}$$

2) GNSS Factor: In general, we can convert longitude, latitude and altitude to ENU $({}_{\mathtt{E}}\widetilde{\mathbf{p}}_i)$. We set the first ENU measurement as the origin point ${}_{\mathtt{E}}\widetilde{\mathbf{p}}_0$. At the same time, we can get origin rotation ${}_{\mathtt{EB}}\widetilde{\mathbf{R}}_0$ from ENU to Body frame according to magnetometer measurements. We also configure ${}_{\mathtt{EB}}\widetilde{\mathbf{T}}_0=({}_{\mathtt{EB}}\widetilde{\mathbf{R}}_0,{}_{\mathtt{E}}\widetilde{\mathbf{p}}_0)$ as world frame. Then world translation ${}_{\mathtt{W}}\widetilde{\mathbf{p}}_i$ can calculate from ${}_{\mathtt{WB}}\widetilde{\mathbf{T}}_i={}_{\mathtt{BE}}\widetilde{\mathbf{T}}_0\ {}_{\mathtt{EB}}\widetilde{\mathbf{T}}_i$. The number of satellites and GNSS quality when the measurements receive, determines the covariance Σ_g. Hence, we can construct the algebraic equation for the residual error as:

$$\mathbf{r}_g={}_{\mathtt{W}}\widetilde{\mathbf{p}}_i-{}_{\mathtt{W}}\mathbf{p}_i.\tag{12}$$

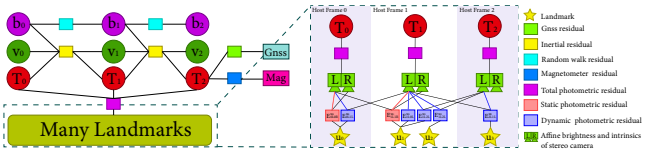

Fig. 4. Factor graph representation for optimization along the system. On the left, there are 3 keyframes poses relating to velocity, bias and many landmarks in sliding window. Detailed factor graph of landmarks are on the right. 4 landmarks are observed by 3 keyframes, depending on inverse depth of the landmark, their affine brightness correction factors and camera intrinsics.

TABLE I
APE OF THE ESTIMATED TRAJECTORIES ON KITTI DATASET FOR SEVERAL METHODS, (-) MEANS NO AVAILABLE DATA

Sequences	Length(m)	DVIGO wo. GNSS	VINS-Fusion wo. GNSS	DVIGO w. GNSS	VINS-Fusion w. GNSS
10_03_27	4268	**14.42**	14.72	**0.142**	0.28
10_03_34	5075	**4.673**	20.373	**0.221**	-
10_03_42	3714	**7.168**	8.472	**0.341**	0.63
09_30_16	397	**1.936**	2.448	**0.09**	0.12
09_30_18	2223	**4.253**	6.959	0.306	**0.24**
09_30_20	1239	**2.527**	4.159	**0.111**	0.27
09_30_27	695	**1.543**	3.37	0.444	**0.15**
09_30_28	5339	**6.286**	13.711	**1.805**	-
09_30_33	1717	**3.971**	6.951	**0.153**	0.27
09_30_34	919	**2.148**	4.684	0.236	**0.20**

C. Optimization

This optimization may be outlined as the factor-graph shown in Fig. 4. The total error is optimized iteratively using Gauss-Newton algorithm with IMU poses, velocity, biases, affine brightness and inverse depth parameters to be optimized as:

$$\begin{aligned}\min_{\chi}\Bigg(&\rho\left(||\mathbf{r}_g||^2_{\Sigma_g}\right)+\rho\left(||r_{\mathbf{m}}||^2_{\Sigma_m}\right)\\&+\sum_{i=0}^{N}||\mathbf{r}_{I_{i,i+1}}||^2_{\Sigma_i}+\sum_{i=0}^{N}||\mathbf{r}^g_{\mathbf{b}_{i,i+1}}||^2_{\Sigma_{bg}}+\sum_{i=0}^{N}||\mathbf{r}^a_{\mathbf{b}_{i,i+1}}||^2_{\Sigma_{ba}}\\&+\sum_{i\in F}\sum_{\mathbf{u}\in U_i}\left(\gamma\left(||r^s_{\mathbf{u}}||^2_{\Sigma_s}\right)+\sum_{j\in obs(\mathbf{u})}\gamma\left(||(r^d_{\mathbf{u}})_{ij}||^2_{\Sigma_d}\right)\right)\Bigg),\end{aligned}\tag{13}$$

where F is the set of the consecutive keyframes in the current window. The U_i is landmark set whose hostframe is $I_i \in F$, $obs(\mathbf{u})$ is the set of the keyframes in F that can observe $\mathbf{u}$. Σ_s and Σ_d are gradient-dependent weighting covariance matrices, which down-weight the high image gradients. $\rho(\cdot)$ is Cauchy norm, while $\gamma(\cdot)$ is Huber norm. To ensure the size of the active window is fixed, the old keyframes are removed by marginalizing a subset of variables.

V. RESULTS

A. KITTI Raw Data Benchmark

We evaluate our proposed system using KITTI Datasets [40]. The datasets are collected onboard a vehicle, contains stereo images (Point Grey Flea 2, 1382x512 monochrome, 10 FPS). Also, the ground truth states are provided by the Inertial Navigation System (OXTS RT3003). We use raw data of KITTI. Raw data IMU measurements are sampled at 100Hz. However, raw data IMU measurements have two main data problems. One is timestamps repetition. We manually modify data to correct timestamps. Another is IMU measurements missing. It can range from a few seconds to a few minutes. If there is not IMU data, we stop IMU pre-integrated, gyroscope, accelerometer bias update and compute velocity by position difference.

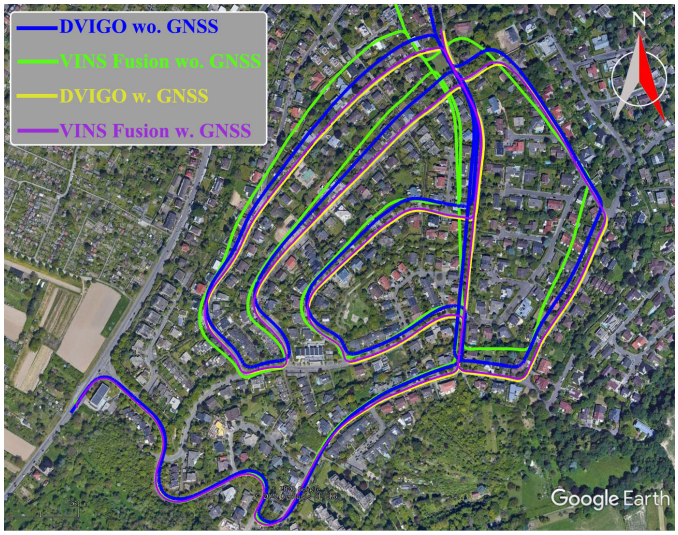

Fig. 5. Trajectories of one KITTI sequence (10_03_34) recovered from DVIGO with VIO only (blue), VINS-Fusion with KITTI odometry config(green), DVIGO with GPS fusion (yellow) and VINS-Fusion with KITTI GPS fusion config(purple).

We use the following IMU parameters: Gyroscope and accelerometer continuous-time noise density $\sigma^g = 0.02[rad/(s\sqrt{Hz}], \sigma^a = 0.06[m/(s^2\sqrt{Hz})]$, Gyroscope and accelerometer bias continous-time noise density $\sigma^{bg} = 0.007, [rad/(s^2\sqrt{Hz}], \sigma^{ba} = 0.019[m/(s^3\sqrt{Hz})]$. To be consistent with our self-developed sensor suite, we corrupt the ground truth and position measurements with Gaussian noise. We consider zero mean Gaussian noise with a standard deviation of 0.01m, and a sampling rate of 1 Hz.

In this experiment, we compare our results with VINS-Fusion [10], a state-of-the-art optimization based algorithm with re-localization capability. We evaluate APE (absolute pose error) results produced by DVIGO and VINS-Fusion. The results of absolute trajectory error for more sequences in KITTI datasets are shown in TABLE I. Firstly, We run the VINS-Fusion code with its KITTI odometry and loop closure configuration and obtain APE by tool proposed in [41]. In all ten sequences, the DVIGO outperforms VINS-Fusion. Secondly, we compute DVIGO with GNSS fusion results, which demonstrates that fusing GNSS measurements effectively increase the accuracy. The results of VINS-Fusion with GNSS are extracted directly from [10]. Trajectories of one KITTI sequence (10_03_34) recovered from VINS-Fusion and the proposed algorithm are shown in Fig. 5. Trajectories are aligned with Google Earth from the bird-eye view. Intuitively, GNSS corrects accumulated drifts in the long distance. DVIGO matches the road path more precisely and has less global drift than VINS-Fusion does, with or without GNSS fusion.

Trajectories of KITTI sequence (10_03_42) recovered from VINS-Fusion and the DVIGO without GNSS are shown on the left in Fig. 6. From the picture, we can see that the

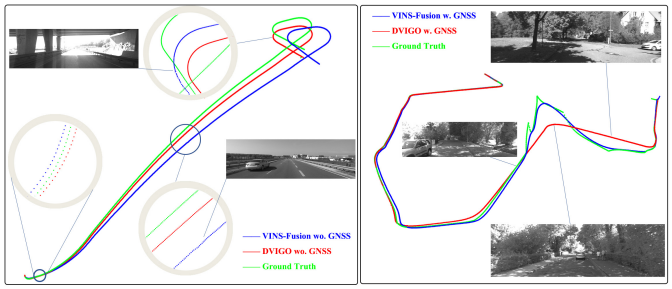

Fig. 6. Comparison between our methodology and VINS-Fusion in two challenging KITTI sequence. Left is on the highway, strong lighting changes and high-speed dynamic vehicles can impact visual odometry. Right is in the tall and dense trees area, leading to weak satellite signal.

estimated trajectory of VINS-Fusion and DVIGO have same drift near the start location. When high-speed dynamic cars show up and illumination changes drastic in camera view because of on the highway, estimated trajectory of DVIGO is much smoother than VINS-Fusion. Part trajectories of KITTI sequence (09_30_28) recovered from VINS-Fusion and the DVIGO with GNSS fusion are shown on the right in Fig. 6. In this sequence, the data acquisition vehicle has a long distance to drive in the tall dense trees area. Although GNSS navigation system (ground truth) returns good navigation quality(Real-Time Kinematic, fixed integers), there is a clear drift in altitude. DVIGO can handle this situation better than VINS-Fusion.

B. EuRoC Benchmark

We test the proposed odometry algorithm on outdoor machine hall sequences in EuRoC dataset [23], in which a firefly hex-rotor helicopter equipped with VI-sensor (an IMU @ 200Hz and dual cameras 752×480 pixels @ 20Hz) was used for data collection. Gyro and acc continuous-time noise density are $\sigma^g = 0.00016[rad/(s\sqrt{Hz}], \sigma^a = 0.002[m/(s^2\sqrt{Hz})]$, bias continous-time noise density(random walk) are $\sigma^{bg} = 0.000019[rad/(s^2\sqrt{Hz}], \sigma^{ba} = 0.003[m/(s^3\sqrt{Hz})]$.

We present all robot states estimation results in Fig. 7. There is no divergence on estimation of accelerometer bias. The gyroscope bias converges to a stable value. Attitude and velocity estimation match ground truth.

For the sequence MH_03_medium and MH_05_difficult, the estimated and ground truth trajectories is shown in Fig. 8. We also map APE error onto estimated trajectories. We show sparse inverse depth maps in Fig. 9. Strong motion blur and low illumination significant challenges for visual odometry estimation. Our method can still handle all the sequences.

TABLE II compares the performance of the DVIGO with VINS-Fusion both using stereo camera and IMU configurations. We compute APE using tool proposed in [41]. Our reported values are the median after five executions. As shown in the table, our method achieves more accurate result than VINS-Fusion.

C. Real-World Experiments

In this experiment, we use self-developed sensor suite which is equipped with multiple sensors. It contains stereo cameras

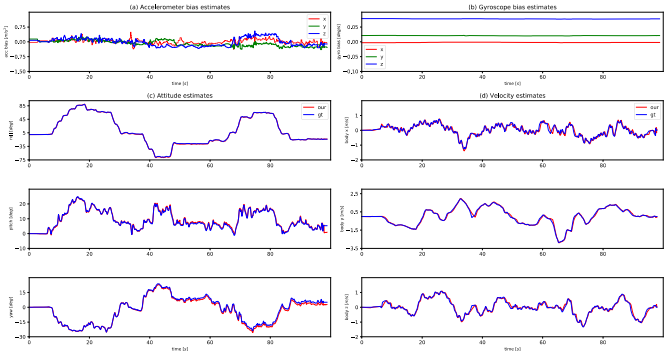

Fig. 7. EuRoC MH3 sequence(99 seconds)-The estimation results via the proposed method and ground truth.

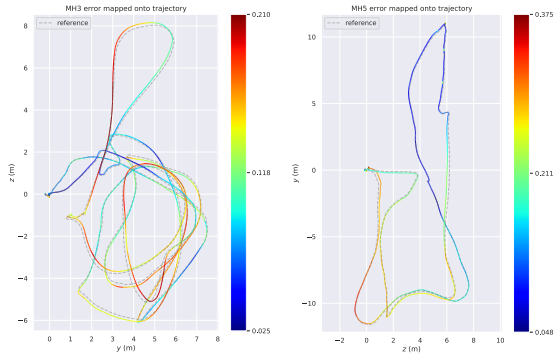

Fig. 8. APE error mapped onto trajectories in MH3(left) and MH5(right).

(MT9V032, global shutter gray scale camera, @60Hz, 752 × 480 resolution, 13.5cm baseline), Xsens MTi-300 consumer level AHRS which includes built-in IMU(@200 Hz), magnetometer(@100 Hz), Qianxun Cube (consumer level GNSS receiver, decimeter accuracy, @1 Hz). We also equip it with dual-antenna receiver-Trimble BD982, delivering accurate positions and precise heading. When BD982 works in single baseline RTK mode and uses observation space representation (OSR) GNSS correction services, it can achieve centimeter accuracy positioning results(@50Hz). We use Qianxun FindCM correction services. The sensor suite is shown in Fig. 11.

We further performed a road test using a bicycle equipped with our sensor suite, riding on different road in Jiading,

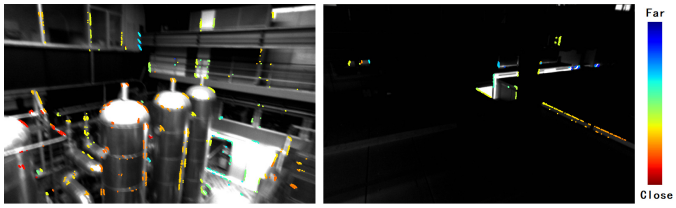

Fig. 9. Example sparse inverse depth maps from stereo camera and used for frame tracking in MH5. On the left is a motion blur scene, right is dark scene.

TABLE II
APE ON EUROC DATASET FOR DVIGO AND VINS-FUSION

Sequences	Length(m)	DVIGO	VINS-Fusion
MH_01_easy	80.6	**0.104**	0.161
MH_02_easy	73.5	**0.057**	0.113
MH_03_medium	130.9	**0.122**	0.182
MH_04_difficult	91.7	0.214	**0.209**
MH_05_difficult	97.6	**0.206**	0.332

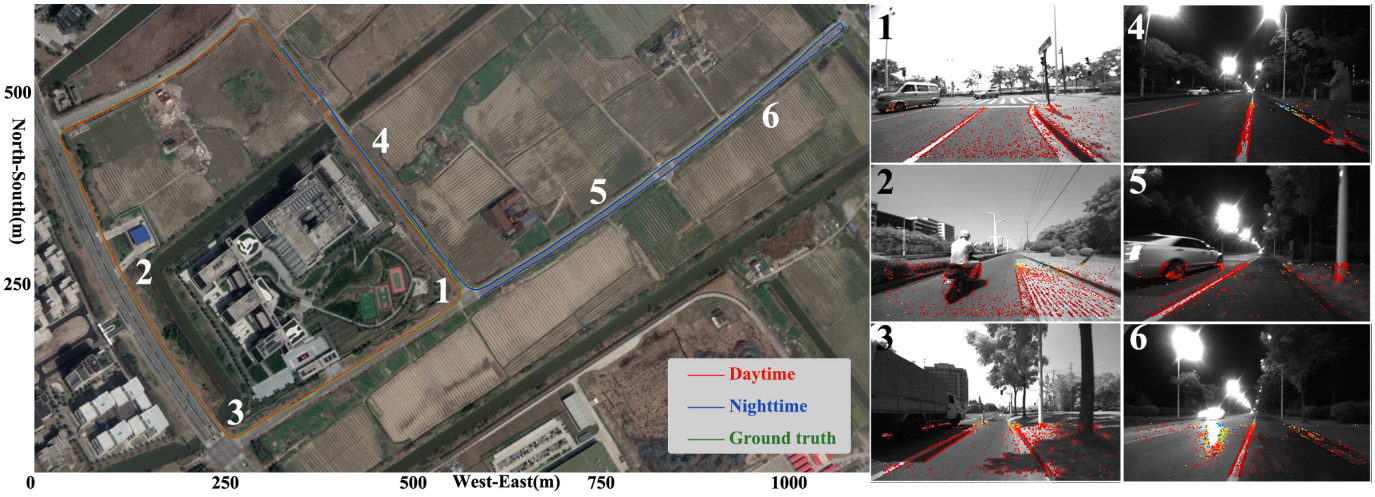

Fig. 10. Left is the trajectories of our large-scale daytime and nighttime outdoor tests recovered from DVIGO. Right is snapshots with sparse inverse depth during tests. It is important to note that the DVIGO is able to handle the (a) dark, (b) stroboscopic phenomenon of streetlight source and (c) dynamic vehicles and the pedestrians scenes of the real world.

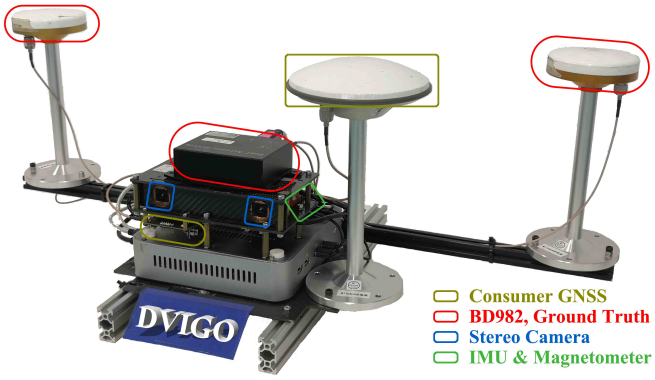

Fig. 11. The sensor suite used for the outdoor experiment.

TABLE III
APE OF THE DVIGO ON OUR OWN DATASET

Exp No.	Distance (m)	Avg vel (m/s)	Environment	No. vehicles and pedestrians	APE (m)
0	332	1.91	Day, V[1]	0	0.178
1	163	1.47	Day, F[2]	1	0.361
2	319	2.79	Night, C[3]	3	0.466
3	4204	3.61	Day, C,	19	0.659
4	1763	4.21	Day, C,	21	0.681
5	1779	5.69	Night, C,	34	0.675
6	3444	4.42	Day, C,	23	0.893

[1] Village potholes pavement. [2] Flat square. [3] City rough road.

Shanghai. Both daytime and nighttime data were collected for the road test. One daytime and one nighttime sequences DVIGO trajectories are shown in Fig. 10. On the left, we can see DVIGO trajectories match the ground truth and road well. On the right, we show example photos with sparse inverse depth. It is important to point out that the experiment is challenging primarily due to: (i) several traffic lights at which we must stop and wait for 20-30 seconds as shown, (ii) frequent stop/yield signs before which we must decelerate or stop, (iii) dynamic scenes including the running vehicles and the pedestrians in vicinity, and (iv) strong lens shake when riding on city rough road or village potholes pavement.

The complete test results with the 7 our own datasets are

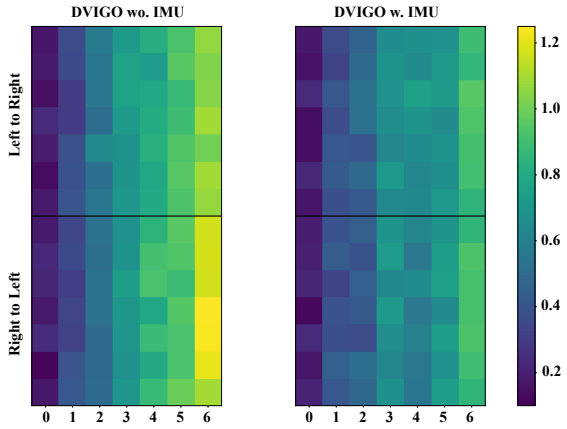

Fig. 12. All APE values for our dataset. "Right to Left" means that we use right image to track and make keyframes. We run each of the 7 sequences(horizontal axis), 7 times each (vertical axis).

listed in TABLE III. We also evaluate the impact of IMU constraint addition on the APE in Fig. 12. If the environment has a lot of dynamic pedestrians and vehicles, results with IMU are obviously more accurate than without IMU, otherwise, the two are roughly equal.

VI. Conclusion

We have presented a novel approach to direct sparse, tightly integrated visual-inertial global odometry. It combines a fully direct structure and motion approach – operating on per-pixel inverse depth instead of individual keypoint observations–with tight, minimization-based IMU integration. Our method can outperform VINS-Fusion approaches in terms of tracking accuracy. In future work, we are going to use a hybrid direct and indirect approach to build a system that is more robust.

VII. Appendix

A detailed derivation in an additional supplementary material https://github.com/ArmstrongWall/SupDVIGO. Experiments videos: https://www.youtube.com/playlist?list=PLnJ8pi4MhtBDpcEg6vZtTMooqEEQq33rI

REFERENCES

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007, pp. 225–234.

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

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

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[6] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[7] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6319–6326.

[8] 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 multi-map slam," *arXiv preprint arXiv:2007.11898*, 2020.

[9] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2198–2204.

[10] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv preprint arXiv:1901.03642*, 2019.

[11] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[12] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[13] R. Elvira, J. D. Tardós, and J. Montiel, "Orbslam-atlas: a robust and accurate multi-map system," *arXiv preprint arXiv:1908.11585*, 2019.

[14] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[15] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[16] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1449–1456.

[17] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1935–1942.

[18] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 141–148.

[19] R. Wang, M. Schworer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3903–3911.

[20] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.

[21] H. Matsuki, L. von Stumberg, V. Usenko, J. Stückler, and D. Cremers, "Omnidirectional dso: Direct sparse odometry with fisheye cameras," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3693–3700, 2018.

[22] J. Zubizarreta, I. Aguinaga, and J. M. M. Montiel, "Direct sparse mapping," *IEEE Transactions on Robotics*, 2020.

[23] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[24] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2011.

[25] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2016.

[26] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2016.

[27] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[28] C. Campos, J. M. Montiel, and J. D. Tardós, "Inertial-only optimization for visual-inertial initialization," *arXiv preprint arXiv:2003.05766*, 2020.

[29] T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems," pp. 3662–3669, 2018.

[30] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1885–1892.

[31] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2510–2517.

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

[33] M. Li and A. I. Mourikis, "High-precision, consistent ekf-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[34] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[35] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4666–4672.

[36] W. Lee, K. Eckenhoff, P. Geneva, and G. Huang, "Intermittent gps-aided vio: Online initialization and calibration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 5724–5731.

[37] S.-F. Ch'ng, A. Khosravian, A.-D. Doan, and T.-J. Chin, "Outlier-robust manifold pre-integration for ins/gps fusion," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7489–7496.

[38] R. Mascaro, L. Teixeira, T. Hinzmann, R. Siegwart, and M. Chli, "Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1421–1428.

[39] Z. Wang, C. Guo, L. Zhao, M. Li, and X. Qi, "Direct sparse visual-inertial odometry with stereo cameras," in IROS VINS Workshop, 2019. [Online]. Available: http://udel.edu/%7eghuang/iros19-vins-workshop/papers/07.pdf.

[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

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