VI-HSO: Hybrid Sparse Monocular Visual-Inertial Odometry

Wenzhe Yang, Yan Zhuang, *Senior Member*, *IEEE*, Dongting Luo, Xuetao Zhang, *Member*, *IEEE*, Wei Wang, *Senior Member*, *IEEE*, and Hong Zhang, *Fellow*, *IEEE*

***Abstract*— In this letter, we present VI-HSO, a hybrid sparse monocular visual-inertial odometry system based on two innovative techniques called adaptive interframe alignment (AIA) and dynamic inverse distance filter (DIDF). Although the sparse image alignment algorithm appears efficient for calculating frame-to-frame motion, it tends to fail in case of significant intensity changes and motion blur. To overcome these limitations, we propose an adaptive interframe alignment method that allows for an adaptive selection between the original Lucas-Kanade (LK) method and the inverse compositional method when constructing photometric errors, along with the addition of inertial information in the process. This approach enables the tracking phase to utilize the full image and inertial information. During intense motion, the inverse distance of the new candidate point often fails to converge, leading to either scale drift or tracking failure. We present a dynamic inverse distance filter that can adjust the convergence range to update candidate points' inverse distance. This adjustment is based on the convergence ratio of the inverse distance of keyframes, which enables more convergent map points aiding in robust tracking in regions lacking texture and during rapid rotation. We evaluate the performance of VI-HSO on public datasets and real-world experiments, and our system outperforms state-of-the-art algorithms. The code is published at https://github.com/luodongting/VI-HSO.**

***Index Terms*—Visual-Inertial Odometry, Monocular Vision, SLAM.**

I. Introduction

MONOCULAR visual-inertial odometry (VIO) and visual-inertial simultaneous localization and mapping have been widely used in various fields including automatic driving, mobile robots, and virtual reality in recent years [1]. These systems leverage the combination of cameras and an inertial measurement unit (IMU) to achieve sensor complementarity and enhance accuracy and robustness. In the case of monocular systems, the IMU provides precise short-term motion constraints and recovers the metric scale.

The odometry algorithms mainly use two parallel threads, the tracking and mapping pipeline [2]. In the tracking pipeline, the system is required to estimate the motion between two frames for feature matching and pose optimization. [3] uses the optical flow method in frame alignment and feature tracking, but the robustness of the system may reduce during intense motion. [4] relies on IMU to predict interframe motion but neglects image constraints, such as image intensity and covisibility relationship. The estimated value of motion relying solely on IMU may not be reliable in the case of large changes in image intensity. Combining optical flow methods and inertial methods can enhance the accuracy of interframe motion estimation. The challenges lie in constructing photometric errors reasonably and combining them with inertial errors.

The accuracy of the system depends on both the accuracy of interframe motion estimation in the tracking pipeline and the impact of local map point information in the mapping pipeline. In the mapping pipeline, an inverse distance filter is constructed for each candidate point detected from keyframes. [2] employs subsequent frames to update the filter. As soon as the filter uncertainty is small enough, candidate points are inserted into the map as 3D points for motion estimation in the tracking pipeline. [5] and [6] adopted a previous keyframe to accelerate the convergence speed of the filter. However, in regions lacking texture or suffering from motion blur, the Image quality in subsequent frames or a previous keyframe tends to deteriorate, presenting challenges in filter updating. This circumstance decreases the probability of candidate points converging into map points, resulting in a reduction of the system's robustness.

In this letter, we propose a hybrid sparse monocular visual-inertial odometry (VI-HSO) and evaluate it extensively on the EuRoC dataset [7], the TUM-VI dataset [8], and a self-collected real-world dataset. The results demonstrate that the proposed VI-HSO outperforms existing approaches in terms of both accuracy and robustness. The performance improvement is attributed to two major contributions:

1) Adaptive interframe alignment, a novel tracking strategy, is added as a module to the back of initialization replacing optical flow or pure inertial estimation. This strategy ensures accurate pose estimation and enhances the ability of the system to resist irregular changes in image intensity and rapid rotation.
2) We propose a dynamic inverse distance filter to expedite the convergence of the inverse distance of candidate points. The system dynamically adjusts the convergence region based on the convergence ratio, generating an adequate number of map points for tracking in texture-less regions and motion blur. This method significantly improves the robustness of the system.

This work was supported in part by the National Natural Science Foundation of China under Grants U22B2041, and in part by the Shenzhen Key Laboratory of Robotics and Computer Vision under Grant ZDSYS20220330160557001.

W. Yang, Y. Zhuang, D. Luo, X. Zhang and W. Wang are with the School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: ywz@mail.dlut.edu.cn, zhuang@dlut.edu.cn, luo_dt@vip. 163.com, zhangxuetao@dlut.edu.cn, wangwei@dlut.edu.cn).

H. Zhang is with the Department of electronic and electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: hzhang@sustech.edu.cn).

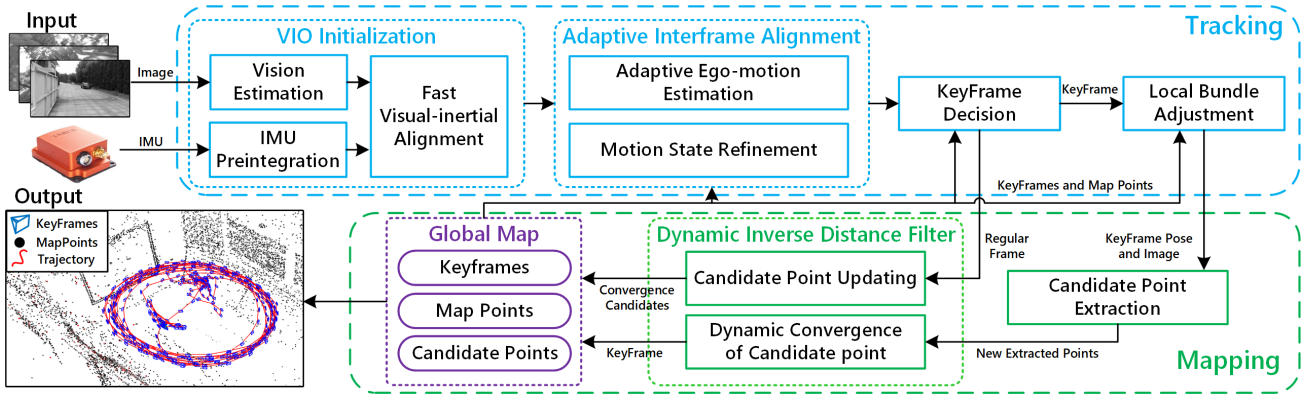


Fig. 1. The pipeline incorporates two parallel threads: the tracking thread in blue, and the mapping thread in green. The global map is represented in purple.

II. Related Works

Monocular vision-only methods are incapable of gaining scale information, thereby limiting their applicability in real-world scenarios. By combining the inertial measurement unit, monocular visual-inertial approaches can recover the metric scale accurately and improve the robustness to motion blur and low-texture [9]. There are two primary types of tightly coupled visual-inertial algorithms: filtering-based and optimization-based approaches [10].

Filtering-based approaches. MSCKF [11] is a classical filtering-based method, which reduces the dimension of the state vectors by using features to form geometric constraints between frames. ROVIO [12], [13] directly adopts pixel intensities of image patches instead of features and uses the intensity errors as the innovation term during the update step to link the multilevel patch features and EKF. The backend of the filtering-based approaches is lightweight, which improves the efficiency of the odometry. However, the backend cannot give the global optimum estimation to the system states. Due to the premature marginalization of Kalman filtering, accumulation errors are inevitably introduced to system [14].

Optimization-based approaches. OKVIS [15], [16] is the first keyframe-based visual-inertial odometry with bundle adjustment. VINS-Mono [3] is a versatile monocular visual-inertial estimator, tightly coupled extrinsic parameter calibration, and robust map-merging. VI-DSO [17] is a direct sparse visual-inertial odometry, which combines photometric information with inertial observation. In subsequent work DM-VIO [18], the system adopts a dynamic weight for visual residuals to address cases where the quality of the image is poor. BASALT [19] is a stereo-inertial odometry system with non-linear factors extracted from visual-inertial odometry. ORB-SLAM3 [4] is a visual-inertial and multi-map SLAM system, which allows various data association methods to reduce system drift. Optimization-based methods can have higher accuracy than filtering-based ones [20].

HSO [6] utilizes average gradients as an indicator to construct photometric residuals and updates the inverse distance filter with a previous frame. In comparison, our system adaptively constructs photometric residuals by considering both average gradients and motion, which are then combined with inertial residuals. Our method dynamically selects consecutive and covisibility frames based on the mean convergence ratio of candidate points to update the filter, resulting in improved accuracy and robustness.

III. Algorithm Overview

We propose a hybrid sparse monocular visual-inertial odometry based on HSO developed by our group [6]. Fig. 1 provides an overview of the proposed pipeline. The VI-HSO system is composed of two parallel threads: tracking and mapping.

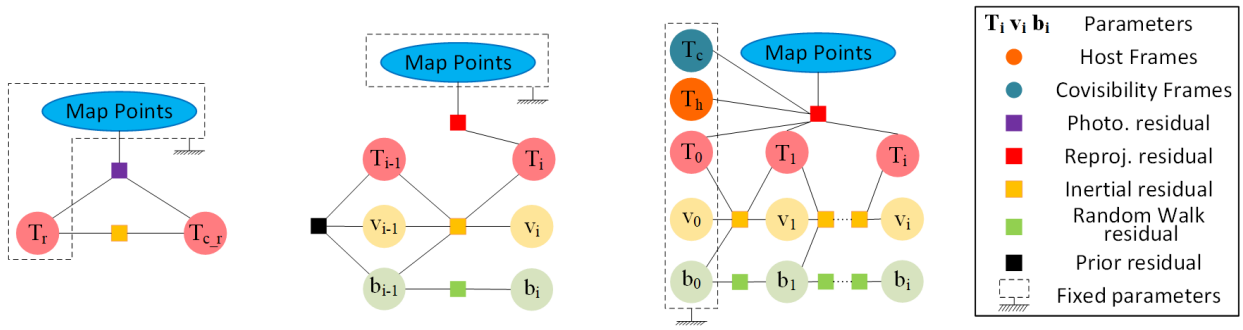
The tracking thread estimates and refines interframe states with the input images and inertial data. We use IMU preintegration and a few processed images to initialize the pose in the *VIO Initialization*. This part is explained in Section IV-*B*. We proposed an *Adaptive Interframe Alignment* method to estimate the relative pose of two frames from coarse to fine. An *Adaptive ego-motion estimation* module is applied to estimate the motion of adjacent frames. Then, the state of the current frame is precisely optimized by the *Motion State Refinement* module. The details of this part are explained in Section IV-*C*. A *Keyframe Decision and Local Bundle Adjustment* module is performed on the current keyframe, the consecutive keyframes, the covisibility keyframes, and the host frames (frames where map points are first observed). This part is explained in Section IV-*D*.

Map points are extracted and updated in the mapping thread. New candidate points are extracted in keyframes. A *Dynamic Inverse Distance Filter* is used to update the inverse distance probability model of candidate points. This part is introduced in Section IV-*E*.

IV. Methodology

A. Notations

We define matrices as bold upper-case letters, vectors as bold lowercase letters, and scalars as lowercase letters. $\{c\}$ is the camera frame. $\{b\}$ is the body frame, which is defined to be the same as the inertial frame. We consider $\{w\}$ as the world frame, which is coinciding with the origin of $\{b\}$ at the initial position. $\mathbf{T}_{cb} := \{\mathbf{R}_{cb}, \mathbf{t}_{cb}\} \in SE(3)$ represents the transformation from the IMU to the camera, where $\mathbf{R}_{cb} \in SO(3)$ is the rotation matrix and $\mathbf{t}_{cb} \in \mathbb{R}^3$ is the



(a) Adaptive Ego-motion Estimation (b) Motion State Refine (c) Local Bundle Adjustment

Fig. 2. Factor graph representation for different optimizations along the system.

translation vector. Poses are represented by $\mathbf{T}_i := \mathbf{T}_{wb_i}$, which is the transformation from IMU frame i to the world frame. ${}_w\mathrm{P} = (x, y, z)^T$ donates a 3D point in the world coordinate system, the left subscript w demonstrates the coordinate system where the map point is located in. $\mathbf{T}_{cb}\mathbf{T}_i^{-1}$ transforms the map point from the world frame to the camera frame, i.e., ${}_c\mathrm{P} = \mathbf{T}_{cb}\mathbf{T}_i^{-1}\,{}_w\mathrm{P}$. The state vector of the frame is defined as:

$$\mathbf{s}_i = \left\{\mathbf{T}_i, \mathbf{v}_i, \mathbf{b}_i^a, \mathbf{b}_i^g, \mathbf{m}_i^0, \mathbf{m}_i^1, \ldots \mathbf{m}_i^j\right\} \tag{1}$$

where $\mathbf{v}_i$ is the velocity in the world frame, $\mathbf{b}_i^a$ and $\mathbf{b}_i^g$ are the accelerometer and gyroscope biases respectively, and $\mathbf{m}_i^j$ are the map points observed in the current frame.

B. VIO Initialization

Monocular visual-inertial odometry is a highly nonlinear system that requires robust initialization to ensure accuracy and stability. The initialization period should be minimized to avoid introducing additional scale errors [18]. We design an initialization method to rapidly align visual and inertial information on the premise of accuracy, including scale, velocities, gravity direction, and IMU biases.

Vision Estimation. The sparse optical flow algorithm is adopted to track features including corner points and edge points. 1500-2000 features are detected in the first frame to ensure adequate points are projected to the subsequent frames. The initializer ensures sufficient motion by setting the minimum average moving distance of corresponding feature points in the pixel plane. To determine the appropriate transformation matrix $\mathbf{T}_i$ for vision initialization, we calculate and decompose the homography and essential matrices.

Fast Visual-inertial Alignment. In this step, we build up about 15 keyframes for alignment. The system inserts keyframes at a certain frequency (around 5-8 keyframes per second), similar to the approach used in [4]. Moreover, when the average parallax of tracked features between the current frame and the previous keyframe is greater than a certain threshold, we insert the current frame to prevent the missing feature points due to rotation. The initialization time is within 3 seconds.

IMU measurements are preintegrated between adjacent keyframes, i-1 and i [21], [22], and compute the inertial residual $\mathbf{r}^{\mathrm{I}}_{i-1,i}$. The optimization problem in alignment is:

$$\chi = \arg\min_{\chi}\left(\sum_{i=1}^{k}\left\|\mathbf{r}^{\mathrm{I}}_{i-1,i}\right\|^2_{\Sigma_I^{-1}} + \left\|\mathbf{b}^a\right\|^2_{\Sigma_{b^a}^{-1}} + \left\|\mathbf{b}^g\right\|^2_{\Sigma_{b^g}^{-1}}\right) \tag{2}$$

where $\chi = \{\mathbf{b}^a, \mathbf{b}^g, \mathbf{v}, \lambda, \mathbf{R}_{wg}\}$ is the initialization state vector, λ is the vision scale factor, and $\mathbf{R}_{wg}$ is the rotation matrix of the gravity vector from the body frame to the real-world frame. $\mathbf{b}^{\mathrm{a}}$ is accelerometer bias, and $\mathbf{b}^{\mathrm{g}}$ is gyroscope bias. Σ_I is the covariance matrix of IMU preintegration. Σ_b is added to make biases zero vectors during initialization. k is the number of all keyframes to be optimized. We employ velocities as explicit optimization variables to reduce non-uniform motion errors. After (2) is solved by the Levenberg-Marquardt (LM) method [23], states are updated.

C. Adaptive Interframe Alignment

In the tracking stage, obtaining a precise initial estimate value of ego-motion is crucial for achieving accurate motion state optimization [4]. To this end, we present an *Adaptive Interframe Alignment* method that tightly couples the inertial and visual data to estimate consecutive frame motion in the tracking thread. This method involves an *Adaptive ego-motion estimation* process that yields an accurate initial value, followed by *Motion State Refinement* that optimizes the pixel coordinates of features and the motion state of the current frame.

Adaptive Ego-motion Estimation. We minimize the photometric and inertial residuals between two frames in this step to achieve more precise ego-motion estimation, providing a solid initial value for further state refinement. To improve efficiency, we only optimize the transformation $\mathbf{T}_{c\_r}$ from the reference frame to the current frame. The factor graph of this problem is shown in Fig. 2(a):

$$\mathbf{T}_{c\_r} = \arg\min_{\mathbf{T}_{c\_r}}\left(\left\|\mathbf{r}^{\mathrm{I}}_{i-1,i}\right\|^2_{\Sigma_I^{-1}} + \mathbf{r}_i^{\mathrm{P}}\right) \tag{3}$$

where $\mathbf{r}_i^{\mathrm{P}}$ is photometric residual. In iterations, the Hessian matrix consists of photometric Hessian matrix $\mathbf{H}^{\mathrm{P}}$ and inertial Hessian matrix $\mathbf{H}^{\mathrm{I}}$, and the error vector is $\mathbf{b}_{total}$:

$$\mathbf{H}_{total} = \mathbf{H}^{\mathrm{P}} + \mathbf{H}^{\mathrm{I}} \text{ and } \mathbf{b}_{total} = \mathbf{b}^{\mathrm{P}} + \mathbf{b}^{\mathrm{I}} \tag{4}$$

Notice that the photometric residual and the inertial residual are defined in different coordinate systems. Therefore, compensate matrix $\mathbf{J}_c = -\mathrm{Adj}(\mathbf{T}_{cb}^{-1}\mathbf{T}_{c_i w})$ needs to be added to build the inertial Hessian matrix:

$$\mathbf{H}^{\mathrm{I}} = (\mathbf{J}^{\mathrm{I}}\mathbf{J}_c)^T \sum_I^{-1} (\mathbf{J}^{\mathrm{I}}\mathbf{J}_c) \text{ and } \mathbf{b}^{\mathrm{I}} = -(\mathbf{J}^{\mathrm{I}}\mathbf{J}_c)^T \sum_I^{-1} \mathbf{r}^{\mathrm{I}}_{i-1,i} \quad (5)$$

where Adj() is the adjoint matrix [24], $\mathbf{J}^{\mathrm{I}}$ represents the inertial Jacobian.

For the construction of photometric residuals, the original LK method [25] is computationally expensive due to the repeated gradient calculations in the current frame per iteration when computing photometric residuals. The inverse compositional method [5] reduces computational costs by computing the point gradient of the last frame only once, but it may fail to converge when gradient values on the last frame are unreliable. To address this, [6] use average gradients to evaluate the texture quality of the image and select the appropriate method for calculating photometric residuals to resist drastic changes in image intensity. However, relying solely on average gradients does not provide a complete comparison of image quality, especially in cases of rapid rotation. As a solution, we propose a state evaluation index *s* that considers both the overall texture of images and the motion state of two frames, which is given by:

$$s = \frac{1}{n}\sum_{j=0}^{n}\left(\left\|\nabla I_i(\mathbf{u}_j)\right\|^2 - \left\|\nabla I_{i-1}(\mathbf{u}_j)\right\|^2\right) + w\left\|\tau\left(\mathbf{R}_{i-1}^{\mathrm{T}}\mathbf{R}_i\right)\right\|^2 \quad (6)$$

where $\nabla I_i(\mathbf{u}_j)$ is the pixel gradient value at the position $\mathbf{u}_j$ in frame *i*, *n* represents the number of all pixels in the image. $\tau(\mathbf{R})$ represents the conversion from the rotation matrix to the Euler angle ΔE (mean of three-axis angles), and $\mathbf{R}_i$ is the attitude determination based on preintegration. $w = \nabla I_{\max} / \Delta E_{\max}$ is a scaling factor, which ensures the gradient and angle are in the same order of magnitude. The photometric residual of the interframe is then adaptively constructed based on *s*, which is defined as:

$$\mathbf{r}_i^{\mathrm{P}} = \begin{cases} \sum_{j=0}\left[I_i(\mathrm{W}(\mathbf{u}_j;\boldsymbol{\psi}+\triangle\boldsymbol{\psi})) - I_{i-1}(\mathbf{u}_j)\right]^2, & s > \theta \\ \sum_{j=0}\left[I_{i-1}(\mathrm{W}(\mathbf{u}_j;\triangle\boldsymbol{\psi})) - I_i(\mathrm{W}(\mathbf{u}_j;\boldsymbol{\psi}))\right]^2, & s \le \theta \end{cases} \quad (7)$$

where $I_i(\mathbf{u}_j)$ is the pixel value of the image at the position $\mathbf{u}_j$ in frame *i*, $\mathrm{W}(\mathbf{u};\boldsymbol{\psi})$ denotes the parameterized set of allowed warps [23], $\boldsymbol{\psi}$ is the vector of the estimated value, and $\triangle\boldsymbol{\psi}$ represents the increment. In our work, we regard the average value of *s* across three previous frames as the threshold θ, which is used to select the method for constructing photometric residual. The condition $s>\theta$ implies that either the gradient of I_i is significantly superior to that of I_{i-1}, or rapid motion has caused images to become blurred. In this case, the original method should be utilized to construct the photometric residual for enhancing robustness [23]. In the other case, the inverse compositional method is used to reduce the processing time while ensuring accuracy.

Motion State Refinement. Using the results of adaptive ego-motion estimation, 3D map points are projected into the image domain of the current frame to serve as the initial value of the 2D feature positions. Then, feature pixel coordinates $\mathbf{u}'_j$ are optimized individually by minimizing the photometric residual of the current image patch concerning the reference patch [6]:

$$\mathbf{u}'_j = \arg\min_{\mathbf{u}'_j}\left(\sum_{\mathbf{u}'_j \in N}\left\|I_{cur}(\mathbf{u}'_j) - e_{c,r}\,\mathrm{A}(I_{ref}(\mathbf{u}_j))\right\|^2\right) \quad (8)$$

where *N* is a set of pixels around $\mathbf{u}$, $e_{c,r}=e_c/e_r$ is the exposure ratio of two frames, And $\mathbf{A}(I)$ is a geometry affine warping to overcome the patch deformation from the reference frame to the current frame. A set of feature correspondences with subpixel accuracy are obtained since the patch of each feature is optimized by (8).

As shown in Fig. 2(b), refining all motion state variables of the current frame is treated as an optimization problem:

$$\mathbf{s}_i = \arg\min_{\mathbf{s}_i}\left(\left\|\mathbf{r}^{\mathrm{I}}_{i-1,i}\right\|^2_{\sum_I^{-1}} + \sum_{j=0}\left\|\mathbf{r}^{\mathrm{R}}_{ij}\right\|^2_{\sum_{ij}^{-1}} + \left\|\mathbf{r}^{\mathrm{B}}_{i-1,i}\right\|^2_{\sum_b^{-1}} + \left\|\mathbf{r}^{\mathrm{pri}}_{i-1}\right\|^2_{\sum_{pri}^{-1}}\right) \quad (9)$$

where $\sum_{pri}$ is the covariance matrix from the last frame optimization. The factor graph of optimization includes four residuals: the inertial residual, the reprojection residual, the bias residual, and the prior residual. $\mathbf{r}^{\mathrm{I}}_{i-1,i}$ is the interframe inertial residual. The reprojection residual $\mathbf{r}^{\mathrm{R}}_{ij}$ for point ${}_w\mathrm{P}_j$ observed in frame *i* is defined as:

$$\mathbf{r}^{\mathrm{R}}_{ij} = \mathbf{u}_{ij} - \Pi(\mathbf{T}_{cb}\mathbf{T}_i^{-1}\,{}_w\mathbf{P}_j) \quad (10)$$

where the image coordinate $\mathbf{u}_{ij}$ of the point *j* in the frame *i* can be obtained by projection using the camera geometric intrinsic $\mathbf{u}_{ij} = \Pi({}_c\mathrm{P}_j)$, and the visual covariance matrix is $\sum_{ij}$. The bias residual $\mathbf{r}^{\mathrm{B}}_{i-1,i}$ is defined as:

$$\mathbf{r}^{\mathrm{B}}_{i-1,i} = \left(\mathbf{b}^a_i, \mathbf{b}^g_i\right)^{\mathrm{T}} - \left(\mathbf{b}^a_{i-1}, \mathbf{b}^g_{i-1}\right)^{\mathrm{T}}. \quad (11)$$

The prior residual $\mathbf{r}^{\mathrm{pri}}_{i-1} = \{\mathbf{r}_{\mathbf{R}}, \mathbf{r}_{\mathbf{v}}, \mathbf{r}_{\mathbf{p}}, \mathbf{r}_{\mathbf{b}}\}$ is used to avoid erupting changes in states of the last frame [4]:

$$\begin{aligned} &\mathbf{r}_{\mathbf{R}} = \mathrm{Log}\left(\mathbf{R}_{i-1}^{\mathrm{T}}\tilde{\mathbf{R}}_{i-1}\right), \quad \mathbf{r}_{\mathbf{v}} = \tilde{\mathbf{v}}_{i-1} - \mathbf{v}_{i-1}, \\ &\mathbf{r}_{\mathbf{p}} = \tilde{\mathbf{p}}_{i-1} - \mathbf{p}_{i-1}, \quad \mathbf{r}_{\mathbf{b}} = \tilde{\mathbf{b}}_{i-1} - \mathbf{b}_{i-1}, \end{aligned} \quad (12)$$

where $(\tilde{\ })$ is the estimated state of the last frame, $\mathbf{b}_{i-1}$ is a bias vector $(\mathbf{b}^a_{i-1}, \mathbf{b}^g_{i-1})^{\mathrm{T}}$. After solving the optimization by the LM method, the motion state of the current frame is refined.

D. Keyframe Decision and Local Bundle Adjustment

The general keyframe decision criteria mainly include tracking quality, field of view changes, and time elapsed since the last keyframe. We suggest using the convergence ratio of inverse distance (*CRID*), discussed in Section IV-*E*, to improve the interaction between tracking and mapping. If the time interval between the current frame and the last keyframe reaches the mean convergence time (*MCT*), as in [6], and the *CRID* of the last keyframe is below 60%, the current frame is selected as a keyframe.

Local BA optimizes over a set of consecutive keyframes κ_s in a sliding window along with all map points ρ_s visible in these keyframes. Keyframes κ_c in the covisibility graph, not in κ_s, are also included in the optimization and remain fixed. In Section IV-*C*, the corresponding features of map points have subpixel accuracy. To reduce the number of parameters, we use inverse distance parameterization to

describe map points. In contrast to the 3D coordinate parameterization method, the inverse distance method links map points not only with the observation frames but also host keyframes κ_h in reprojection residual construction. With the host keyframes remaining fixed, the optimization problem is shown in Fig. 2(c):

$$\{\mathbf{s}_i\} = \arg\min_{\{\mathbf{s}_i\}} \left(\sum_{i\in\kappa_s} \left\| \mathbf{r}^{\mathrm{I}}_{i-1,i} \right\|^2_{\Sigma_I^{-1}} + \sum_{k\in\kappa_s\cup\kappa_c\cup\kappa_h} \sum_{j\in\chi_k} \left\| \mathbf{r}^{\mathrm{R}}_{kj} \right\|^2_{\Sigma_{kj}^{-1}} + \sum_{i\in\kappa_s} \left\| \mathbf{r}^{\mathrm{B}}_{i-1,i} \right\|^2_{\Sigma_b^{-1}} + \left\| \mathbf{r}^{\mathrm{pri}}_{0} \right\|^2_{\Sigma_{pri}^{-1}} \right) \tag{13}$$

where χ_k is the set of matches between the feature in keyframe k and a map point in ρ_s.

During the process of optimization, certain rules are followed: **i)** keyframes outside the window and the oldest keyframe in the window are kept fixed to prevent erupting changes, **ii)** 12 keyframes are optimized in the window, **iii)** the number of map points in the window is limited for efficiency, and **iv)** duplicate or outlier map points are culled to simplify the global map. Then, the new keyframe is sent to the mapping thread.

E. Dynamic Inverse Distance Filter

New candidate points are initialized at corners and along gradient edges in the mapping thread. The inverse distance estimation of a candidate point is modeled with a Gaussian distribution, updated by other frames. Previous work [2] uses subsequent frames to update the probability model of the inverse distance, but the convergence speed is slow. [5] and [6] improve convergence speed by updating filters with the last keyframe and a set of regular frames. However, these methods have two issues. Firstly, it will increase computational complexity during smooth motion and may cause tracking failures during rapid motion when a fixed number of frames are used to update the filter. Secondly, if only the motion's continuity relationship is used to update the filter while neglecting the covisibility relationship, it is difficult to find highly enough correlated patches to match candidate points in consecutive frames, particularly in regions with low texture or affected by motion blur.

To address these limitations, we propose the *Dynamic Convergence of Candidate Point* module, which jointly updates the filter using consecutive and covisibility frames and dynamically adjusts the ranges ($\boldsymbol{L_1}$ and $\boldsymbol{L_2}$) of selected frames, achieving more robust inverse distance estimation. $\boldsymbol{L_1}$ is the range of consecutive frames for the convergence of candidate points' inverse distance. We proposed the convergence ratio of inverse distance (*CRID*, the ratio at which all extracted corner and gradient edge points can converge successfully in a keyframe) as a new criterion to dynamically adjust the range $\boldsymbol{L_1}$ of the consecutive frames. We observed that the *CRID* decreases in the case of the nearby frames with violent motion or rapid rotation. As shown in Fig. 3, when the current keyframe k_i arrives, the mean convergence ratio of inverse distance M_{CRID} for a set of previous keyframes is:

$$M_{CRID} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{N_o + C}{N_a + C}\right) \tag{14}$$

where m is the number of keyframes (m=3 in our system), N_a is the number of candidate points, and N_o is the number of convergence map points in ith keyframe. Constant C is adopted to avoid erupting changes, which is set to the number of map points reprojected to keyframe i. The convergence range $\boldsymbol{L_1}$ includes N_f previous keyframes and the regular frames between these keyframes:

$$N_f = \frac{n_k}{\pi}\arccos\left(M_{CRID}\right) \tag{15}$$

where n_k is the number of keyframes in local BA. When M_{CRID} decreases, it indicates poor convergence and more frames are used to update the filter. In the other case, the filter reduces the number of update frames to improve efficiency.

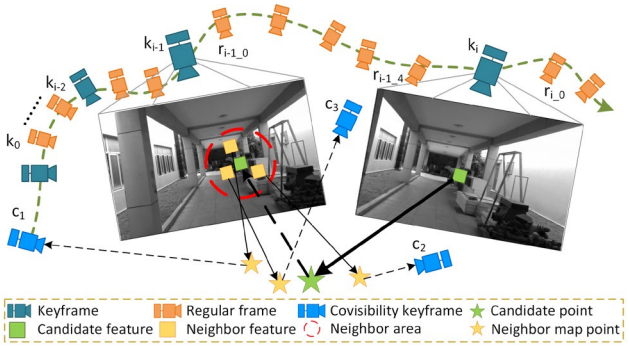


Fig. 3. Consecutive frames and covisibility keyframes in the dynamic inverse distance filter. Consecutive frame includes consecutive keyframes and consecutive regular frames.

$\boldsymbol{L_2}$ is the range of covisibility keyframes to converge the inverse distance of candidate points. Candidate points cannot find matching frames in $\boldsymbol{L_1}$ due to motion blur, but they are likely observed at the frame where the neighbor map points are located. These frames where neighbor map points are located are considered the covisibility frames for updating the filter. As shown in Fig. 3, we project candidate points onto the keyframes in $\boldsymbol{L_1}$ and utilize the resulting projection points as the search center to find neighbor features. The neighbor area is dynamically adjusted using a search radius, $d_r \times (1\text{-}M_{CRID})$, to obtain an appropriate number of corresponding neighbor map points, where d_r is the maximum distance between two feature points in the top-level image pyramid. When *MCRID* decreases, more neighbor map points will be discovered. The host keyframes of these neighbor map points are the range $\boldsymbol{L_2}$ of covisibility keyframes. Compared to the consecutive frames, these covisibility keyframes are reliable due to multiple optimizations, resulting in more accurate convergence results.

New candidate points can converge to the greatest extent possible by all frames in $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. Then, points are sent into the *Candidate Point Updating* module, which updates the probability model only through the newest frame. In challenging scenes, the *Dynamic Inverse Distance Filter* is crucial to the robustness of our system, especially when the tracking is unsatisfactory.

TABLE I
TRANSLATION ERROR [M] OF EACH SEQUENCE ON EUROC DATASET

Sequence	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203	**Avg**
MCSKF [11]	0.420	0.450	0.230	0.370	0.480	0.340	0.200	0.670	0.100	0.160	1.130	0.414
OKVIS [15]	0.330	0.370	0.250	0.270	0.390	0.094	0.140	0.210	0.090	0.170	0.230	0.231
ROVIO [12]	0.210	0.250	0.250	0.490	0.520	0.100	0.100	0.140	0.120	0.140	0.140	0.224
VINS-Mono [3]	0.150	0.150	0.220	0.320	0.030	0.079	0.110	0.180	0.080	0.160	0.270	0.159
VI-DSO [17]	0.062	0.044	0.117	0.132	0.121	0.059	0.067	0.096	0.040	0.062	0.174	0.089
BASALT [19]	0.070	0.060	0.070	0.130	0.110	0.040	0.050	0.100	0.040	0.050	-	0.072
DM-VIO [18]	0.065	0.044	0.097	0.102	0.096	0.048	0.045	0.069	0.029	0.050	0.114	0.069
ORB-SLAM3 [4]	0.062	0.037	0.046	0.075	0.057	0.049	0.015	0.037	0.042	0.021	0.027	0.043
VI-HSO (Ours)	0.021	0.033	0.041	0.047	0.062	0.035	0.019	0.023	0.025	0.020	0.023	0.032

* Red indicates the best result on a sequence, and blue represents the second-best result. "-" means the algorithm failed on this sequence.
* All other results are taken from the respective paper.

V. EXPERIMENTS

We evaluated our proposed method on three datasets: the EuRoC public dataset [7], the TUM-VI public dataset [8], and a self-collected real-world dataset. Additionally, we conducted ablation experiments on the proposed innovation points. We compared our method with other approaches by computing the root mean square error (RMSE) after aligning the estimated trajectory with the ground truth. To ensure reliability, we executed all sequences 10 times and used the median of the experiment results to accurately represent the behavior of the system. All experiments were performed in real-time mode on a desktop with an Intel Core i9-9900K at 3.6 GHz and 16 G RAM, using only CPU.

A. Visual-inertial Odometry Quantitative Comparison

1) Accuracy Evaluation on EuRoC Dataset. The EuRoC dataset includes 11 sequences collected in industrial machine hall and ordinary room scenes. Table I compares our method to state-of-the-art methods in visual-inertial odometry and our method clearly outperforms all other methods. Our method achieves more than double the accuracy of VINS-Mono and DM-VIO and 40% greater accuracy than ORB-SLAM3 on simple sequences, such as MH01, MH02, and V101. This improvement is attributable to our optimization module of Local BA, which leverages the observation information from nearby positions. By adding covisibility frames and host frames of map points in the sliding window to the optimization, we strengthen the association between the current and historical data even without loop closure. The V203 sequence presents a significant challenge for all methods due to motion blur and image intensity changes resulting from considerable motion. A similar issue leads to significant odometry errors on V103. However, our method employs adaptive interframe alignment, resulting in much smaller errors than other methods.

2) Accuracy Evaluation on TUM-VI Dataset. We also evaluated our method on the room and the corridor subsets of the TUM-VI dataset. On TUM-VI, the trajectory lengths of different sequences vary greatly, so the final average drift is computed with *(rmse·100/length)*%. The ground truth is only available in the room part of the sequence. Therefore, in Table II, we only evaluated the accuracy on room sequences and further evaluated the robustness on corridor and magistrale sequences. Compared to other methods, our proposed method shows the optimal result on most sequences and the suboptimal result on rest sequences, with a mean drift of 0.116. For the room sequences, which represent small environments, we reduce the average error to 0.0096m, below 1cm. The corridor sequences are more challenging, with rapid rotation, high acceleration motion, dramatic changes in image brightness, and low texture environments. Our method and DM-VIO exhibit more robust performance in these scenarios by leveraging photometric information. Compared to other methods, our method also exhibits good accuracy on magistrale sequences.

TABLE II
TRANSLATION ERROR [M] OF EACH SEQUENCE ON TUM-VI DATASET

Sequence	ROVIO	OKVIS	BASALT	VINS-Mono	DM-VIO	Ours
room1	0.16	0.06	0.09	0.07	0.03	0.008
room2	0.33	0.11	0.07	0.07	0.13	0.009
room3	0.15	0.07	0.13	0.11	0.09	0.01
room4	0.09	0.03	0.05	0.04	0.04	0.013
room5	0.12	0.07	0.13	0.2	0.06	0.008
room6	0.05	0.04	0.02	0.08	0.02	0.010
corridor1	0.47	0.33	0.34	0.63	0.19	0.1
corridor2	0.75	0.47	0.42	0.95	0.47	0.21
corridor3	0.85	0.57	0.35	1.56	0.24	0.35
corridor4	0.13	0.26	0.21	0.25	0.13	0.16
corridor5	2.09	0.39	0.37	0.77	0.16	0.22
magistrale1	4.52	3.49	1.2	2.19	2.35	0.75
magistrale2	13.43	2.73	1.11	3.11	2.24	0.98
magistrale3	14.8	1.22	0.74	0.4	1.69	2.39
magistrale4	39.73	0.77	1.58	5.12	1.02	2.76
magistrale5	3.47	1.62	0.6	0.85	0.73	0.71
magistrale6	X	3.91	3.23	2.29	1.19	1.85
Avg (%)	0.896	0.188	0.137	0.243	0.129	0.116

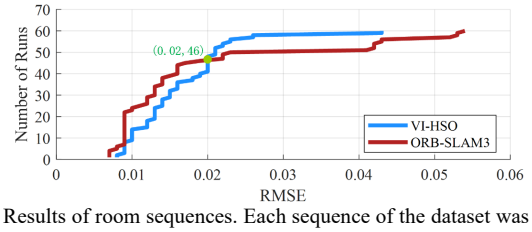
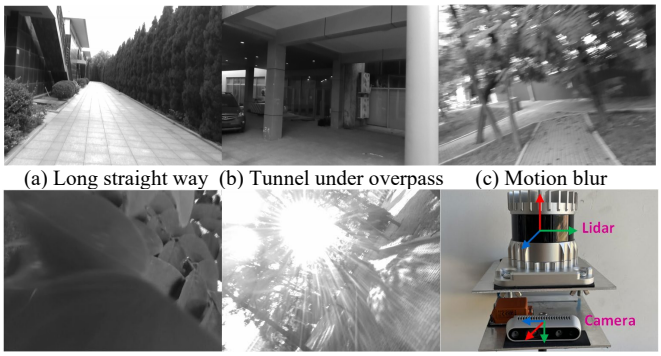


Fig. 4. Results of room sequences. Each sequence of the dataset was run ten times (60 experiments in total). The experimental results were arranged in ascending order of RMSE. The y-axis represents the number of runs. The x-axis is the RMSE of the corresponding run.

Fig. 4 shows the comparison test on our method and ORB-SLAM3. In the first 46 runs, the error of ORB-SLAM was slightly smaller than that of VI-HSO. As the number of runs increased, some significant errors occurred with ORB-SLAM3. In contrast, the error distribution of VI-HSO was more uniform and there was no sudden significant error. Therefore, ORB-SLAM3 performs well in terms of accuracy, while our method is superior in terms of robustness.

B. Robust Evaluation on Self-collected Real-world Dataset

To confirm the practicability and validity of our method in real-world applications, we collected a challenging dataset of outdoor settings on our university campus. We use a hand-held 3D-BOX device with a depth camera and a 32-line 3D lidar, as shown in Fig. 5. The camera and lidar have separate integrated IMU modules. The experimental setting involves buildings and groves, where GPS signals are mostly unavailable. Therefore, we approximated the results of FAST-LIO as the ground truth in GPS-denied environments.



(a) Long straight way (b) Tunnel under overpass (c) Motion blur

(d) Shelter (e) Strong sunlight (f) Hand-held 3D-BOX

Fig. 5. (a-e) shows the challenging environments where we collect datasets. (f) shows our Hand-held 3D-BOX device setup. The camera is an Intel RealSense D435i, and the lidar is an OUSTER OS1-32.

1) Robustness comparison on Building Sequence. We held the device by hand and walked around the laboratory building at a normal pace. The trajectory length of the building sequence is approximately 370m. There are overlapping parts in the path, and loop closure detection is available. The primary challenges peculiar to this sequence are: **i)** motion blur on some bend scenes, **ii)** scale drift caused by uniform motion, and **iii)** dramatic changes in image intensity. As shown in Fig. 6, our trajectory is closer to the ground truth trajectory than that of VINS-Mono and ORB-SLAM3. Although VINS-Mono successfully detected loop closures thanks to optimization in the sliding window, scale drift occurs in the corner. The irregular image intensity changes lead to large-scale drift of ORB-SLAM3 when facing the light source or entering and exiting a tunnel.

2) Robustness comparison on Grove Sequence. The grove sequence covers about 180m, during which we walked at a fast pace while shaking the handheld device vigorously. The varied view directions between the starting and ending points made it impossible to detect common regions. Moreover, the camera occasionally encountered obstructions from leaves or direct exposure to sunlight, increasing the sequence complexity. As shown in Fig. 7, our method has completed the challenging sequence with superior trajectory accuracy. In comparison, VINS-Mono displayed more drift in strong light environments and big corners. Unexpectedly, ORB-SLAM3 failed when the images were blocked by bushes. The blurred image causes the failure in matching the corresponding map points, and the rapid rotation exacerbates the difficulty of tracking, which is fatal for ORB-SLAM3 based on the feature method.

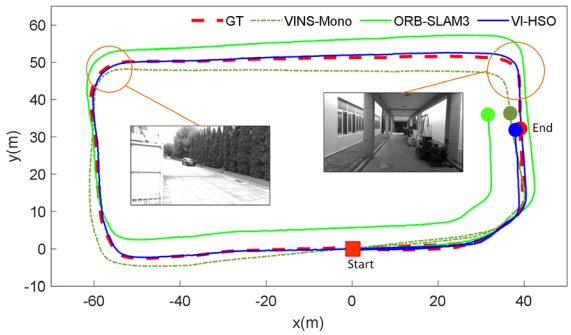


Fig. 6. Comparison results of the building sequence collected by our hand-held device. Yellow circle: the scene is prone to drift, including motion blur and dramatic image intensity changes.

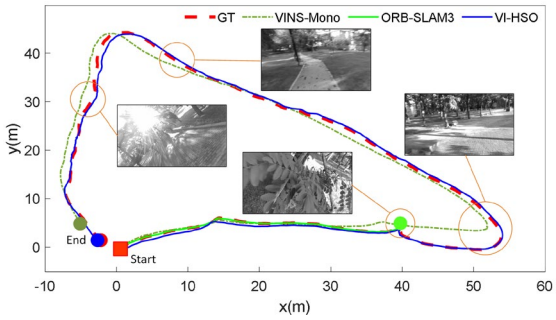


Fig. 7. Comparison results of the grove sequence collected by our hand-held device. Yellow circle: the scene is prone to drift, including image occlusion, violent shaking, rapid rotation, and intense illumination.

C. Ablation Studies

In this section, we perform ablation studies on AIA and DIDF. To evaluate their impact on the system, we individually disable these two modules. The experiments were conducted on the EuRoC dataset, with each sequence executed ten times, and the median value was taken as the experimental result. Table III demonstrates that the AIA and DIDF module can effectively improve the system accuracy.

TABLE III
TRANSLATION ERROR [M] WITH/WITHOUT AIA ON EUROC DATASET

AIA	DIDF	MH01	MH02	MH03	MH04	MH05	V101
-	-	0.053	0.072	0.095	0.11	0.119	0.037
-	√	0.040	0.049	0.057	0.052	0.087	0.036
√	-	0.028	0.039	0.062	0.064	0.077	0.036
√	√	0.021	0.033	0.041	0.047	0.062	0.035

AIA	DIDF	V102	V103	V201	V202	V203	**Avg**
-	-	0.024	0.036	0.035	0.045	0.126	0.068
-	√	0.021	0.028	0.027	0.022	0.056	0.043
√	-	0.022	0.042	0.027	0.021	0.081	0.045
√	√	0.019	0.023	0.025	0.020	0.023	0.032

D. Efficiency Analysis

As shown in Table IV, we present the mean and standard deviation of running time of main operations on EuRoC dataset. Our system mainly consists of two parallel threads: tracking and mapping. The most time-consuming part of the tracking thread is the Local Bundle Adjustment module (LBA), which only runs when the current frame is selected as the keyframe. For the system to operate stably, the LBA module needs to execute 4-6 times per second in our system. The other parts of the tracking thread take about 9ms. The mapping thread integrates the candidate point extraction module, the candidate point updating module, and the dynamic convergence of the candidate point module, and the running time of this thread is about 7ms.

TABLE IV
MEAN (μ) AND STANDARD DEVIATION (σ) OF RUNNING TIME [MS] OF MAIN OPERATIONS ON EUROC DATASET

Thread	Operation	μ ± σ
Tracking	Fast Visual-inertial Alignment	2.12±0.30
	Adaptive Ego-motion Estimation	3.32±0.76
	Motion State Refinement	3.54±1.06
	Keyframe Decision	0.03±0.01
	Local Bundle Adjustment (4-6 times/s)	36.18± 17.49
Mapping	Candidate Point Extraction	3.75±1.42
	Candidate Point Updating	1.31±0.84
	Dynamic Convergence of Candidate Point	2.65±1.08

VI. CONCLUSION

We present a hybrid sparse monocular visual-inertial odometry (VI-HSO). Specifically, we propose an adaptive interframe alignment method that facilitates the accurate estimation of motion between continuous frames, even in the presence of irregular image intensity changes. Moreover, we introduce a dynamic inverse distance filter that improves the convergence ratio of the inverse distance of new map points to mitigate the effect of texture-less regions and rapid rotation. Compared to other leading-edge solutions in the public datasets, our method has a superior performance in terms of accuracy. The results obtained from real-world experiments further confirm the robustness of our approach and its applicability in practical scenarios.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 15–22.

[3] 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, 2018.

[4] C. Campos, R. Elvira, J. J. G. Rodrguez, 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.*, pp. 1–17, 2021.

[5] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017.

[6] D. Luo, Y. Zhuang, and S. Wang, "Hybrid sparse monocular visual odometry with online photometric calibration," *Int. J. Robot. Res.*, vol. 41, no. 11, pp. 993-1021, 2022.

[7] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, 2016.

[8] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stueckler, and D. Cremers, "The TUM VI benchmark for evaluating visual-inertial odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018.

[9] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," *Int. J. Robot. Res.*, vol. 41, no. 7, pp. 667–689,2022.

[10] J. Gui, D. Gu, S. Wang, and H. Hu, "A review of visual inertial odometry from filtering and optimisation perspectives," *Advanced Robotics*, vol. 20, no. 20, pp. 1289-1301, 2015.

[11] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2007, pp. 3565–3572.

[12] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 298–304.

[13] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback," *Int. J. Robot. Res.*, vol. 36, no. 10, pp. 1053–1072, 2017.

[14] L. Jinyu, Y. Bangbang, and C. Danpeng. "Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 4, pp. 386-410, 2019.

[15] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial SLAM using nonlinear optimization," in *Proc. Robot.: Sci. Syst.*, pp. 314-334, 2013.

[16] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, " Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[17] L. von Stumberg, V. Usenko, and D. Cremers, "Direct sparse visual-inertial odometry using dynamic marginalization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018.

[18] L. von Stumberg, and D. Cremers. "DM-VIO: Delayed marginalization visual-inertial odometry," *IEEE Robot. Automat. Lett.*, vol.7, no. 2, pp. 1408-1415, 2022.

[19] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 422–429, 2020.

[20] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Visual SLAM: Why filter?" *Image and Vision Computing*, vol. 30, no. 2, pp. 65–77, 2012.

[21] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 61–76, 2012.

[22] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, 2017.

[23] G. Grisetti, R. Kümmerle, H. Strasdat. "g2o: A general framework for (hyper) graph optimization," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011.

[24] J. Solà, J. Deray and D. Atchuthan, "A micro lie theory for state estimation in robotics", *arXiv:1812.01537*, 2018, [online] Available: http://arxiv.org/abs/1812.01537.

[25] S. Baker and I. Matthews. "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221-255, 2004.