IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

RMSC-VIO: Robust Multi-Stereoscopic Visual-Inertial Odometry for Local Visually Challenging Scenarios

Tong Zhang, Jianyu Xu, Hao Shen, Rui Yang, and Tao Yang

***Abstract*—We present a Multi-Stereoscopic Visual-Inertial Odometry (VIO) system capable of integrating an arbitrary number of stereo cameras, exhibiting excellent robustness in the face of visually challenging scenarios. During system initialization, we introduce multi-view keyframes for simultaneous processing of multiple image inputs and propose an adaptive feature selection method to alleviate the computational burden of multi-camera systems. This method iteratively updates the state information of visual features, filtering out high-quality image feature points and effectively reducing unnecessary redundancy consumption. In the backend phase, we propose an adaptive tightly coupled optimization method, assigning corresponding optimization weights based on the quality of different image feature points, effectively enhancing localization precision. We validate the effectiveness and robustness of our system through a series of datasets, encompassing various visually challenging scenarios and practical flight experiments. Our approach achieves up to a 90% reduction in Absolute Trajectory Error (ATE) compared to state-of-the-art multi-camera VIO methods.**

***Index Terms*—SLAM, vision-based navigation, sensor fusion.**

I. Introduction

ACCURATE self-positioning is a foundation for robotics to achieve autonomy. Due to the lightweight, low power consumption, and complementary functionalities of cameras and Inertial Measurement Units (IMUs), methods based on VIO have been widely applied in robotics. Several notable algorithms [1], [2], [3], [4], [5] have achieved high accuracy and stable state estimation on publicly available datasets. However, since these algorithms rely solely on a single monocular or stereo camera, their ability to capture all orientations of a robot is limited. Consequently, in challenging visual environments characterized by inadequate or repetitive textures, abrupt variations in illumination, or local occlusions, these VIO algorithms are prone to failure due to the unavailability of visual information [6].

Facing complex visual challenges in the environment, many researchers have attempted to enhance the robustness of systems by introducing additional sensor modalities such as LiDAR, mmWave radar, and others [7], [8]. However, the integration of these additional sensor modalities presents challenges in synchronization, sensor fusion, calibration, and data processing, significantly increasing the complexity and computational requirements of the system. This makes real-time deployment challenging, particularly on UAVs with limited payload capacity.

In recent years, there has been a growing research interest in multi-camera VIO algorithms [9], [10], [11], [12], [13]. Multi-camera VIO can capture more visual information and offer redundant advantages, leading to enhanced robustness in challenging visual scenarios. However, multi-camera comes at the cost of additional computational burden. Furthermore, in challenging visual environments, the integration of poor-quality information directly impacts the accuracy of the system.

In this letter, we present a novel optimization-based **R**obust **M**ulti-**S**tereos**C**opic VIO algorithm: **RMSC**-VIO. Through innovative initialization strategies and tailored backend optimization, we harness high-quality visual features from multiple stereoscopic cameras, significantly improving the system's robustness in challenging visual scenarios. Additionally, our proposed Adaptive Feature Selection method (AFS) refines and selects high-quality feature points based on the state information of visual features in each image, reducing the computational burden on multi-camera systems and improving the quality of feature information for the back-end.

We conducted a comparative analysis of our proposed method, RMSC-VIO, against the state-of-the-art stereo method VINS-Fusion [4] and the multi-camera method MCVIO [13], using datasets consisting of various challenging scenarios. Our method demonstrated superior performance compared to VINS-Fusion, achieving a remarkable 60% to 80% reduction in Root Mean Square Error (RMSE) as measured by ATE. Furthermore, when compared to MCVIO, our approach exhibited exceptional effectiveness, achieving a substantial 60% to 90% reduction in ATE RMSE. To further validate the robustness and efficacy of our proposed approach, we conducted real-world experiments in both indoor and outdoor environments. The contributions of this letter can be summarized as follows:

Manuscript received 9 January 2024; accepted 8 March 2024. Date of publication 18 March 2024; date of current version 22 March 2024. This letter was recommended for publication by Associate Editor G. Costante and Editor P. Vasseur upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61603297 and in part by the Natural Science Foundation of Shaanxi Province under Grant 2023JC-YB-503. *(Corresponding author: Jianyu Xu.)*

Tong Zhang, Jianyu Xu, Hao Shen, and Tao Yang are with Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhangtong@nwpu.edu.cn; xujianyu@mail.nwpu.edu.cn; 2022-204583@mail.nwpu.edu.cn; yangtao@nwpu.edu.cn).

Rui Yang is with CIAD UMR7533, University Bourgogne Franche-Comte, UTBM, F-90010 Belfort, France (e-mail: rui.yang@utbm.fr).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3377008, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3377008

- We propose a Multi-Stereoscopic VIO system capable of integrating numerous stereo cameras, exhibiting excellent robustness in visually challenging scenarios.
- We propose an adaptive feature selection method that iteratively updates the state information of visual features, filtering high-quality image feature points and reducing the computational burden of multi-camera systems.
- We propose an adaptive tightly coupled optimization method that assigns optimization weights based on the quality of distinct image feature points, effectively enhancing the localization precision of the system.
- We conducted a comprehensive and extensive experimental evaluation in various challenging scenarios to validate the robustness and efficacy of our proposed method. Moreover, we have publicly released the dataset utilized in these challenging scenarios for future research and development purposes[1].

II. Related Work

A. Localization in Challenging Visual Environments

Estimating robot pose in challenging environments without GPS has been extensively researched. Although current VIO algorithms are widely used in robotics, they still face challenges in visually sparse or confusing scenes, rendering them highly unreliable. To enhance system robustness, researchers have explored the integration of information from other sensors. Khattak et al. [7] proposed a resilient multi-modal approach to pose estimation by fusing data from multiple sensors, including VIO, TIO, and LiDAR. This allows the system to continue operating even when a sensor fails. Similarly, Doer et al. [8] used Kalman filtering to fuse FMCW millimeter-wave radar with VIO and TIO, improving robustness when visual information is lost. While these methods effectively enhance system robustness in certain complex environments, they require the inclusion of additional sensor types, which increases system complexity and computational cost.

Alternately, some scholars have focused on improving traditional VIO algorithms to enhance system robustness. Chen et al. [14] incorporated a neural network into a conventional VIO to improve the system's ability to handle complex environments. Zhang et al. [15] proposed an inertial SLAM method based on point-line vision specifically designed for indoor environments under weak texture and dark conditions. However, these methods are limited by the narrow field of view of a single camera.

B. Multi-Camera VIO

Multiple cameras can significantly enhance visual constraints with a wider field of view, thereby providing redundancy for backend optimization. To overcome the challenges of visual positioning in complex scenarios, researchers have proposed the use of multiple cameras to improve system robustness. Initially, Schauwecker et al. [16] deployed multiple cameras on UAVs for pose estimation. They solved local poses separately using forward and downward-looking cameras and fused them to obtain pose estimates. Similarly, Yang et al. [17] utilized image features from two cameras to extend the Parallel Tracking And Mapping (PTAM) algorithm [18], enabling the autonomous navigation of drones. In a different approach, Miiller et al. [19] improved the robustness of visual-inertial navigation by fusing motion estimation from each camera's VO with IMU data via an Extended Kalman Filter (EKF). Eckenhoff et al. [20] proposed a general-purpose multi-camera visual-inertial navigation system based on the Multi-state Constraint Kalman Filter (MSCKF) framework. However, it is important to note that all these methods are filtering-based, and cumulative errors may occur over long-term operation, which can affect positioning performance.

Recent years have witnessed a tremendous increase in the processing power of minicomputers, leading to the emergence of optimization-based visual-inertial fusion localization methods as the mainstream approach. Houben et al. [21] proposed an extension of monocular ORB-SLAM for a system with multiple cameras and an IMU. However, their system lacks practical experimental verification. Yang et al. [10] formalized the problem of multi-camera SLAM with asynchronous shutters. Their framework groups input images into asynchronous multi-frames and extends feature-based SLAM to the asynchronous multi-view setting. Kuo et al. [9] proposed an adaptive SLAM system based on SVO [22] for multi-camera setups. They designed adaptive initialization schemes, entropy-based keyframe selection, and new map management. However, the accuracy of real experimental results remains nearly constant when using multiple cameras. Kaveti et al. [12] proposed a versatile multi-camera SLAM framework capable of accommodating various camera system configurations. Their work included a comprehensive examination of the impact of camera setups on SLAM performance, considering monocular, stereo, and multiple camera arrangements with and without overlapping Fields of View (FoVs). Abate et al. [23] extended Kimera to a multi-camera VI-SLAM system. Their extension enabled globally consistent trajectory estimation and the construction of dense 3D maps around the vehicle, facilitating obstacle avoidance and navigation. However, these methods did not consider the increased computational cost introduced by multi-camera systems.

To mitigate the issue of high computational consumption in multi-camera VIO systems, researchers have proposed various feature extraction and optimization algorithms. Jaekel et al. [24] introduced a feature point extraction algorithm called one-point RANSAC, which utilizes a fixed lag smoother to jointly optimize all poses and landmarks. However, their system lacks precise experimental verification, leaving uncertainty on its efficacy. He et al. [13] proposed a VIO algorithm that uses multiple non-overlapping monocular cameras based on VINS-Mono [1]. They employed GPUs to accelerate front-end feature processing and reduce the computational cost of multi-camera systems. Despite this optimization, their approach failed to achieve superior robustness and accuracy compared to conventional VIO algorithms. Zhang et al. [11] developed a multi-camera VIO system based on factor graph optimization using all available cameras simultaneously to estimate motion. Their research mainly focused on efficient feature tracking and selecting the best subset

[1]Available at http://github.com/884992491X/RMSC-VIO.

Algorithm 1: Adaptive Landmark Feature Selection.

Input: Multiview Keyframes $\mathcal{M}$.
Output: $\{\mathbf{R}_k^w, \mathbf{p}_k^w\}$: the pose of the k-th $\mathcal{M}$; Sup: the set of super feature points.

initialization:
  $k \leftarrow 1, i \in [1, n], \mathcal{E}_j^i \leftarrow 1, \mathcal{E}_{Sum} \leftarrow 0, Sum \leftarrow 0.$
foreach *multiview keyframes* $\mathcal{M}$ **do**
  $\{\mathbf{R}_k^w, \mathbf{p}_k^w\} \leftarrow PNP(\mathcal{F}_{t_k}^i)$
  foreach *stereo camera* c^i **do**
    for ${}^j\mathcal{F}_{t_k}^i \in \mathcal{F}_{t_k}^i$ **do**
      if ${}^j\mathcal{F}_{t_k}^i$ *is tracked continuously for more than 4 frames* **then**
        $\mathcal{E}_j^i \leftarrow Reprojection({}^j\mathcal{F}_{t_{start}}^i, {}^j\mathcal{F}_{t_k}^i)$
        $\mathcal{E}_{Sum} += \mathcal{E}_j^i$
        $Sum++$
      end
    end
    $\mathcal{E}_{ave}^i = \mathcal{E}_{Sum}/Sum$
    for ${}^j\mathcal{F}_{t_k}^i \in \mathcal{F}_{t_k}^i$ **do**
      if $\mathcal{E}_j^i < \mathcal{E}_{ave}^i$ **then**
        $Sup \leftarrow Sup \cup {}^j\mathcal{F}_{t_k}^i$
      end
    end
    $\mathcal{E}_{Sum} = 0$
    $Sum = 0$
  end
  $k++$
  $i \leftarrow \min(\mathcal{E}_{ave}^1, \mathcal{E}_{ave}^2, \cdots, \mathcal{E}_{ave}^n)$
end

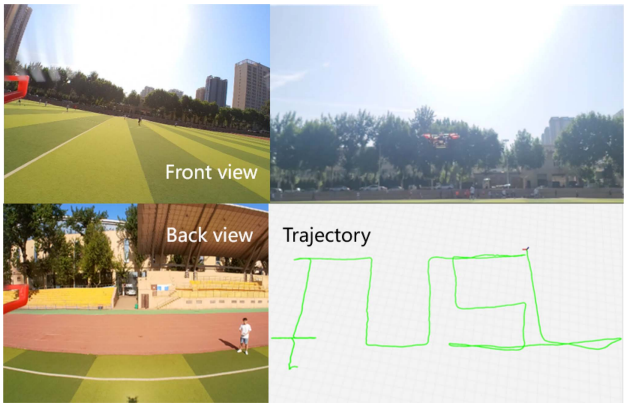


Fig. 1. Proposed algorithm was deployed on drones and validated in challenging visual scenes. The two images on the left show the front and back views of the drone, captured by two motion cameras. The graph in the lower right corner shows the trajectory of our proposed efficient multi-camera VIO system. More details of the experiment can be accessed https://youtu.be/_CWLo0V31og.

of features to ensure time-bounded computation. However, their feature selection method did not appear to improve the system's accuracy.

III. System Overview

The system architecture is illustrated in Fig. 2, providing a comprehensive overview of the proposed RMSC-VIO algorithm. The algorithm framework comprises three key components: front-end data preprocessing, system initialization, and back-end optimization. It is noteworthy that our approach leverages the features from multiple stereo cameras simultaneously, employing novel strategies in both initialization and back-end optimization to enhance the system's robustness and effectiveness in complex environmental conditions.

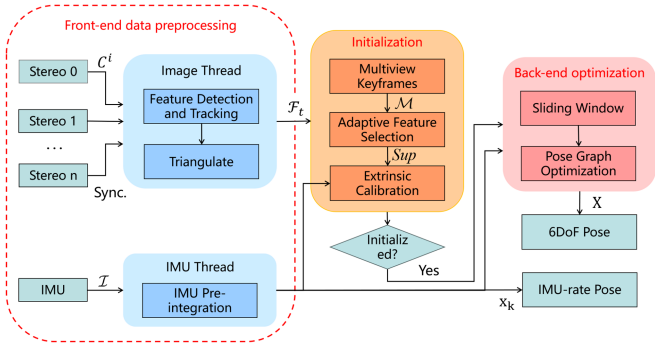


Fig. 2. Diagram illustrating the RMSC-VIO architecture, showcasing its three fundamental components: front-end data preprocessing, system initialization, and back-end optimization. The backend produces a pose with the same frequency as the IMU and a low-frequency 6DoF optimized pose.

A detailed description of the methodological improvements in front-end processing, initialization, and back-end optimization is provided in Section IV. Additionally, in Section V, we conduct an extensive set of experiments to analyze and evaluate the system's robustness and accuracy.

We define the notations and frame definitions used throughout the system overview as follows: The body frame of the robot is aligned with the IMU $\mathcal{I}$. We denote the world frame, body frame, and the i-th camera frame at time t as $(\cdot)^w$, $(\cdot)^b$, and $(\cdot)^{c_t^i}$, respectively. Rotation is represented using both rotation matrices $\mathbf{R}$ and quaternions $\mathbf{q}$. The operation of multiplying two quaternions is denoted by $\otimes$.

IV. Methodology

A. Front-End Feature Handling

In the multi-stereoscopic VIO system, we extract and match feature points from each stereo camera image to obtain environmental information from various robot orientations. For each stereo camera, we utilize the KLT optical flow algorithm [25] to track feature points in the previous frame of the left-eye image. Additionally, we employ the Shi-Tomasi algorithm [26] to extract new feature points, ensuring a minimum number of features is maintained. Subsequently, we perform KLT optical flow tracking on the left-eye image features towards the right-eye image. Finally, we use the stereo geometric constraints to triangulate the feature points and obtain their depth information. Through the aforementioned process, we generate a comprehensive set of feature points at the current time, denoted as $\mathcal{F}_t$ in Fig. 2.

B. Initialization

Multi-stereoscopic VIO systems provide rich feature information, including landmark features in various directions of

the robot. However, the computational complexity increases exponentially with the number of feature points. Additionally, the positioning accuracy is significantly influenced by the quality of each landmark feature. To address these issues, we introduce the concept of Multiview KeyFrames (MKF) during the initialization process and propose an Adaptive landmark Feature Selection method (AFS).

1) Multiview KeyFrames: We synchronize the timestamps of all stereo cameras and treat the images captured simultaneously from all cameras as a generalized image frame. Multiple-view keyframes are selected based on the computation of parallax to previous image frames and the quality of feature point tracking. If the average parallax of the tracked feature points exceeds a certain threshold or the number of tracked features drops below a certain threshold, the current frame is chosen as a multiple-view keyframe.

2) Adaptive Landmark Feature Selection: Multi-view keyframes include a large number of distinctive feature points from all images. To reduce the computational cost of processing feature points and improve their quality, we propose an adaptive landmark point selection method named AFS. When visual information for a specific orientation is unavailable, AFS selects alternative feature point sets that are accessible to initialize the pose of multiview keyframes. Conversely, in scenarios where all visual information is accessible, AFS strategically chooses a group of high-quality feature points for subsequent back-end optimization processing. Algorithm 1 shows the pseudo-code of the proposed AFS algorithm.

We first randomly select a set of feature points $\mathcal{F}_{t_k}^i$ from stereo camera c^i to initialize the pose of the first keyframe using the Perspective-n-Point (PnP) algorithm (Lines 3). After obtaining the initial pose, we proceed to filter the feature point set. For each stereo camera c^i, we iterate through its feature point set ${}^j\mathcal{F}_{t_k}^i$, checking whether the feature points appear in the current frame (Lines 5) and have been continuously tracked for more than 4 frames (Lines 6). If these conditions are met, we calculate the reprojection error $\mathcal{E}_j^i$ of these feature points between the current frame $\mathcal{M}_{t_k}$ and the frame $\mathcal{M}_{t_{start}}$ in which they were first observed (Lines 7). We designate $\mathcal{E}_j^i$ as the state information characterizing the visual features. The formulation is as follows:

$$\widehat{{}^j\mathbf{P}_k^i} = (\mathbf{R}_k^w)^{-1}((\mathbf{R}_{start}^w(\mathbf{R}_{c^i}^b\,{}^j\mathbf{P}_{start}^i + \mathbf{p}_{c^i}^b) + \mathbf{p}_{Start}^w) - \mathbf{p}_k^w)$$

$$\mathcal{E}_j^i = \left\| \pi_i \begin{pmatrix} \frac{{}^j\mathbf{P}_k^i(X)}{{}^j\mathbf{P}_k^i(Z)} \\ \frac{{}^j\mathbf{P}_k^i(Y)}{{}^j\mathbf{P}_k^i(Z)} \\ 1 \end{pmatrix} - \pi_i \begin{pmatrix} \frac{\widehat{{}^j\mathbf{P}_k^i(X)}}{\widehat{{}^j\mathbf{P}_k^i(Z)}} \\ \frac{\widehat{{}^j\mathbf{P}_k^i(Y)}}{\widehat{{}^j\mathbf{P}_k^i(Z)}} \\ 1 \end{pmatrix} \right\|^2 \tag{1}$$

where ${}^j\mathbf{P}_{start}^i$ denotes the three-dimensional coordinates of the feature point ${}^j\mathcal{F}_{t_{start}}^i$ when it is first observed in the i-th camera frame. The estimated value $\widehat{{}^j\mathbf{P}_k^i}$ and the observed value ${}^j\mathbf{P}_k^i$ represent the feature point ${}^j\mathcal{F}_{t_{start}}^i$ in the current frame's i-th camera coordinate.

Next, we calculate the average reprojection error $\mathcal{E}_{ave}^i$ for eligible feature points in each stereo camera (Lines 12). If the state information $\mathcal{E}_j^i$ of a feature point is less than the average $\mathcal{E}_{ave}^i$, it is designated as a super feature point Sup for backend optimization (Lines 14-16). Furthermore, we select the feature point set corresponding to the minimum average reprojection error to solve the pose of the next multi-view keyframe (Lines 22).

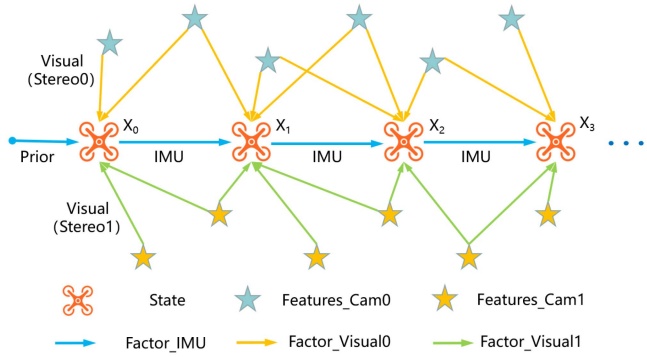


Fig. 3. Structure of the sliding-window factor graph, illustrating prior, visual, and pre-integrated IMU factors. Visual factors encompass feature points observed from various directions, with two directions depicted as examples.

Through iterative updates of multi-view keyframes, we ultimately obtain a higher-quality tracked feature point set Sup and more accurate initial poses $\{\mathbf{R}_k^w, \mathbf{p}_k^w\}$ for the multiview keyframes.

3) Calibration of Multi-Camera Extrinsics: Deploying multi-camera systems requires external parameter calibration to ensure accuracy and reliability. To simplify this process, we extend the online extrinsic calibration method proposed by Yang et al. [27] from single cameras to multi-camera systems. We formulate equation constraints by integrating IMU rotations and utilizing visual measurements between consecutive image frames. These constraints help estimate the rotational external parameters between individual cameras and the IMU. Once the rotation parameters for each camera are obtained, we integrate them into the backend to solve for the translation parameters and further optimize the system.

C. Tightly Coupled Multi-Stereoscopic VIO

After system initialization, we employ factor graphs to tightly couple optimization using data collected from various sensors. Instead of independently fusing visual information from each camera with IMU data and then merging multiple VIO outputs, we integrate the superior feature point sets Sup from all cameras into the backend for joint optimization. Our fusion method provides the following advantages:

- It avoids repetitive and complex VIO fusion computations, significantly reducing computational complexity and costs.
- It mitigates inconsistencies between multiple VIO results, particularly in complex environments, ensuring enhanced system robustness.
- It simultaneously integrates visual information from different perspectives into optimization, thereby improving the accuracy of system optimization.

The structure of the sliding window factor graph is depicted in Fig. 3. It comprises the state vector to be optimized, visual factors

from different stereo cameras, IMU pre-integration factors, and priori factors. The state vector to be optimized is defined as follows:

$$\begin{aligned}\mathbf{X} &= \left[\mathbf{x}_0, \ldots, \mathbf{x}_K, \mathbf{T}_{c^1}^b, \ldots, \mathbf{T}_{c^n}^b, \boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_n\right] \\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^w, \mathbf{v}_{b_k}^w, \mathbf{q}_{b_k}^w, \mathbf{b}_a, \mathbf{b}_g\right], k \in [0, K] \\ \mathbf{T}_{c^i}^b &= \left[\mathbf{p}_{c^i}^b, \mathbf{q}_{c^i}^b\right], i \in [1, n] \\ \boldsymbol{\rho}_i &= \left[\rho_{c_i,0}, \rho_{c_i,1}, \ldots, \rho_{c_i,f_{ci}}\right] \end{aligned} \tag{2}$$

where $\mathbf{x}_k$ represents the IMU state corresponding to the k-th multiview keyframes $\mathcal{M}_k$. It encompasses the position $\mathbf{p}_{b_k}^w$, velocity $\mathbf{v}_{b_k}^w$ and orientation $\mathbf{q}_{b_k}^w$ of the IMU in the world frame, as well as the acceleration bias $\mathbf{b}_a$ and gyroscope bias $\mathbf{b}_g$ in the IMU frame. In this context, K denotes the total number of multiview keyframes. Additionally, $\boldsymbol{\rho}_i$ signifies the inverse depth of each feature point in the i-th camera, while f_{ci} represents the total number of feature points for the i-th camera within the sliding window.

Given the assumption that all measurements are independent and uncertainties adhere to a Gaussian distribution, we transform the state estimation into a nonlinear least-squares problem, commonly referred to as bundle adjustment (BA). The formulation is as follows:

$$\begin{aligned}\min_{\mathbf{X}} \Bigg\{ &\left\|\mathbf{r}_p - \mathbf{H}_p\mathbf{X}\right\|^2 + \sum_{k\in\mathcal{B}} \left\|\mathbf{r}_{\mathcal{B}}\left(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathbf{X}\right)\right\|_{\mathrm{p}_{b_{k+1}}^{b_k}}^2 \\ &+ \sum_{i\in n}\sum_{(l,j)\in\mathbf{S}^i} \lambda_i \rho\left(\left\|\mathbf{r}_{\mathbf{S}^i}\left(\hat{\mathbf{z}}_l^{\mathbf{S}^{ij}}, \mathbf{X}\right)\right\|_{\mathrm{p}_l^{\mathbf{S}^{ij}}}^2\right)\Bigg\} \end{aligned} \tag{3}$$

where $\|\mathbf{r}_p - \mathbf{H}_p\mathbf{X}\|$, $\mathbf{r}_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathbf{X})$, and $\mathbf{r}_{\mathbf{S}^i}(\hat{\mathbf{z}}_l^{\mathbf{S}^{ij}}, \mathbf{X})$ represent the marginalized prior information, IMU residuals, and visual residuals, respectively. We adopt the method described in [1] for all the residual models, and the detailed definitions are presented in [1]. $\rho(x)$ is the robust kernel function used to suppress outliers, $\mathcal{B}$ denotes the set of all IMU measurements. Additionally, $\mathbf{S}^i$ refers to the set of super feature points from the i-th stereo camera selected previously.

In Section IV-B, we addressed the impact of unreliable visual features through the AFS method. In this section, we design and adaptively allocate visual residual weights λ_i based on the quality of different image feature points, effectively leveraging visual features from each direction. The definition of λ_i is formulated as follows:

$$\lambda_i = \left(\frac{N^i}{\sum_{i\in n} N^i}\right)^2 \tag{4}$$

where N^i is the total number of super feature points in the i-th camera.

D. Multi-Stereoscopic Loop Fusion

Multi-camera SLAM systems can offer a more comprehensive and extensive set of scene information since they are capable of observing the same scene from various angles and perspectives.

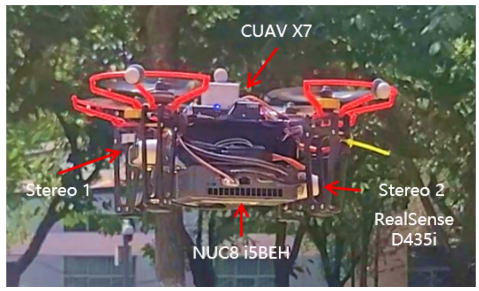


Fig. 4. Quadrotor drone employed in our experiment, equipped with two cameras featuring distinct orientations.

This capability enhances the effectiveness and robustness of loop closure detection.

We extend the loop closure detection scheme proposed in [1] to multi-view loop closure detection. For each newly acquired multi-view keyframe, we employ DBoW2 [28] to assess whether it revisits a previously observed area. Upon detecting a loop, we perform multi-view geometric validation to eliminate outliers in the loop closure. Subsequently, upon confirming a valid loop closure, we execute global bundle adjustment, leveraging all available information to optimize the entire trajectory, thereby significantly reducing drift in the majority of sequences.

V. Experimental Results

We assessed the performance of our algorithm across diverse challenging datasets indoor and outdoor, spanning environments ranging from underground parking lots to expansive open spaces. Our comprehensive evaluation involved a comparative analysis against state-of-the-art methods, considering both quantitative and qualitative metrics. Additionally, we conducted ablation studies to systematically demonstrate how our proposed contributions effectively reduce APE and computation time. Finally, to validate the practicality and effectiveness of our algorithm, we deployed it on quadrotors, offering real-world insights into its performance.

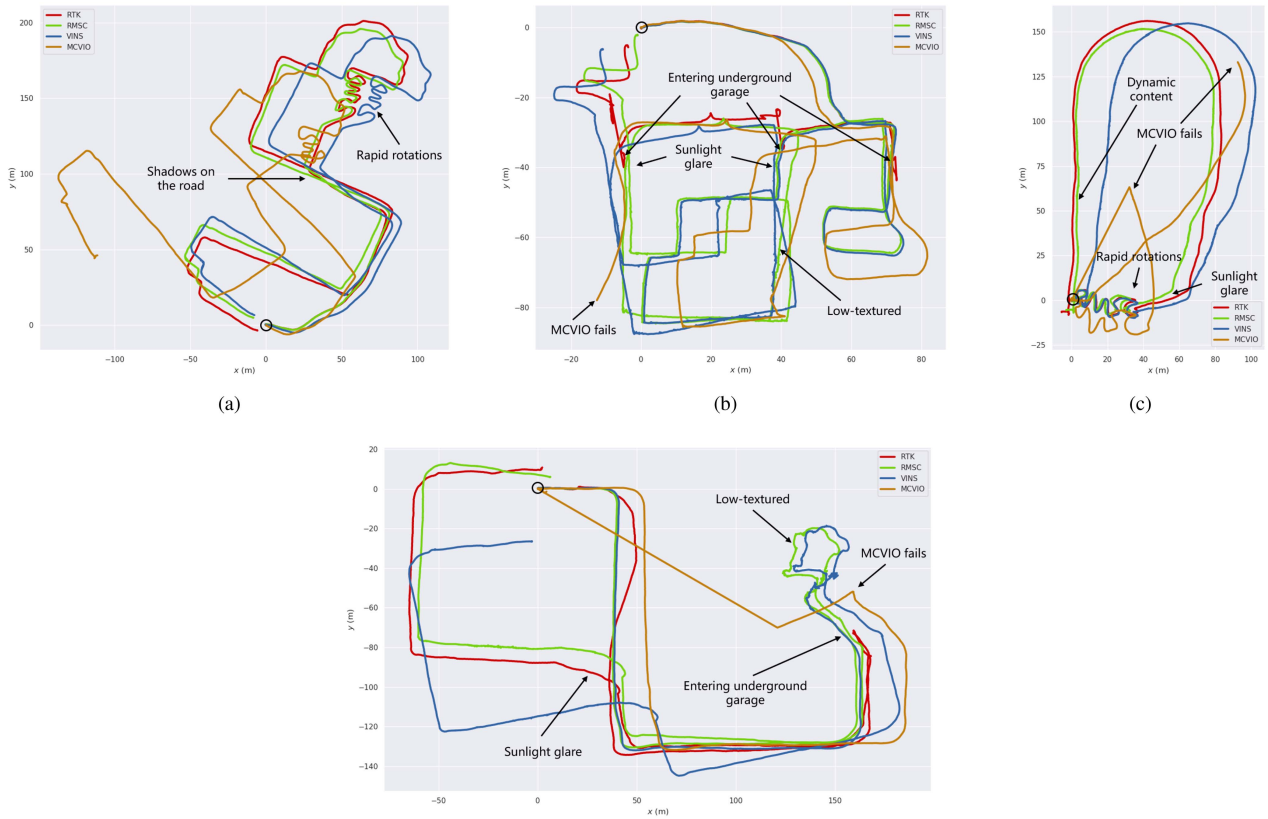
A. Hardware Setup and Datasets

The quadrotor drone utilized in our experiment is depicted in Fig. 4. To capture environmental information from various directions, we deployed two Intel RealSense D435i depth cameras. These cameras operated at a frequency of 15 Hz with a resolution of 640×480. An Intel NUC8i5BEH microprocessor was installed onboard to handle all data processing tasks and execute the motion pose estimation algorithms. Additionally, a CUAV X7 flight controller was employed to provide IMU measurements and execute ground station commands for UAV flight control.

We collected seven sequences, comprising three indoor and four outdoor scenes. These sequences included challenging visual scenarios (see in Fig. 5). For indoor sequences, we utilized NOKOV Motion Capture System to acquire ground truth poses with millimeter-level accuracy. For outdoor sequences, Real-Time Kinematic (RTK) technology was employed to obtain ground truth trajectories with centimeter-level precision.

(a) Indoor1 (b) Indoor2 (c) Indoor3 (d) Outdoor1 (e) Outdoor2 (f) Outdoor3 (g) Outdoor4

Fig. 5. Sample images showcase challenging scenarios such as featureless environments, occlusions, abrupt lighting changes, shadows on the road, low-textured underground garages, dynamic objects, and environments with sunlight glare.



(a) (b) (c)

(d)

Fig. 6. Top-down view of four dataset sequences comparing the estimated trajectory of VINS-Fusion, MCVIO, and RMSC-VIO against ground truth. The black circle indicates the start of the trajectory. Panels (a), (b), (c), and (d) correspond to the Outdoor1, Outdoor4, Outdoor3, and Outdoor2 dataset sequences, respectively.

TABLE I
PERFORMANCE COMPARISON BETWEEN DIFFERENT ALGORITHMS

Absolute Trajectory Error (ATE)				
Dataset	**VINS-Fusion** rmse(m)	**ORB-SLAM3** rmse(m)	**MCVIO** rmse(m)	**RMSC-VIO** rmse(m)
Indoor1	0.1444	**0.0298**	0.1399	0.0374
Indoor2	0.2756	Fail	0.1856	**0.0664**
Indoor3	0.3636	Fail	0.1528	**0.0593**
Outdoor1	15.2122	Fail	63.3857	**5.9833**
Outdoor3	13.2111	Fail	Fail	**3.5515**

B. Comparing With the State of the Art Algorithms

In this section, we conduct a comparative analysis of the performance of our algorithm with that of ORB-SLAM3 (a widely utilized sparse visual SLAM system), VINS-Fusion (a lightweight algorithm extensively employed for autonomous drone localization), and MCVIO (a recently open-sourced multi-camera visual-inertial odometry algorithm).

1) Qualitative Analysis: We ensured consistency in camera intrinsics and extrinsics, the quantity of extracted feature points, and other configuration parameters across all algorithms. The estimated trajectories of Outdoor sequences are illustrated in Fig. 6. Qualitative assessments of stability and accuracy for different algorithms can be made based on the trajectory performance.

In challenging visual scenarios, as depicted in Fig. 5 under conditions of shadows, rapid rotations, low texture, sunlight glare, light changing, and dynamic environments, the extraction of stable feature points from forward-looking images becomes significantly more challenging. VINS-Fusion, relying on IMU, can continue operating but suffers from increasing ATE and substantial trajectory drift in these situations.

TABLE II
ABLATION STUDIES: ADAPTIVE LANDMARK FEATURE SELECTION (AFS)

	Feature tracker time (ms)			**AFS** time (ms)	**Optimization** time (ms)			**Total time** time (ms)			**Feature cnt** /			**ATE** rmse (m)		
	N-AFS	**W-AFS**	**VINS**	**W-AFS**	**N-AFS**	**W-AFS**	**VINS**	**N-AFS**	**W-AFS**	**VINS**	**N-AFS**	**W-AFS**	**VINS**	**N-AFS**	**W-AFS**	**VINS**
Indoor1	21.04	21.11	14.79	0.07	53.11	40.66	33.02	74.14	61.85	47.81	223	131	89	0.071	0.037	0.101
Indoor2	21.40	20.78	14.80	0.08	54.19	39.36	33.46	75.59	60.22	48.27	229	127	92	0.086	0.066	0.276
Indoor3	20.24	20.20	14.52	0.07	54.69	38.79	33.80	74.93	59.07	48.32	238	120	88	0.085	0.059	0.364
Outdoor1	21.37	21.31	14.23	0.06	46.70	37.74	32.78	68.07	59.12	47.01	159	114	85	13.263	5.983	15.212
Outdoor2	21.44	21.51	14.87	0.06	40.37	37.04	31.51	61.80	58.62	46.38	132	105	77	/	/	/
Outdoor3	20.61	20.53	14.15	0.06	40.84	34.69	30.63	61.44	55.29	44.78	134	97	73	4.411	3.551	13.211
Outdoor4	20.33	20.35	14.94	0.06	37.48	32.97	28.15	57.81	53.39	43.09	110	89	60	/	/	/
Average	20.92	20.83	14.61	**0.065**	**46.77**	**37.33**	**31.91**	67.68	58.22	46.52	**175**	**112**	**81**	3.583	**1.939**	5.833

MCVIO, leveraging image information from different orientations, is anticipated to demonstrate superior performance. However, due to the limitations of monocular vision, scale inaccuracies may arise across various datasets. Moreover, using a primary camera for state estimation as a basis and then estimating the states of other cameras easily leads to unsatisfactory accuracy and system crashes, especially when the primary camera's image features become unavailable, as observed during rapid rotations or navigation through underground parking lots.

The proposed RMSC-VIO algorithm addresses scale inaccuracy by utilizing multiple stereoscopic cameras. It fuses feature point information from all stereo cameras, filters out high-quality feature point sets based on specific criteria through AFS, and feeds these inputs into the back-end multi-stereoscopic optimization, thereby improving the efficacy and robustness of the system. Our algorithm demonstrates excellent trajectory performance across all datasets.

2) Quantitative Analysis: Table I illustrates the Absolute Trajectory Error (ATE) for different algorithms across data sequences with complete ground truth. ATE serves as an intuitive metric, offering insights into the accuracy and global consistency of trajectories across various challenging scenarios. ORB-SLAM3 demonstrates exceptional precision in the Indoor1 sequence, featuring normal indoor visual characteristics. However, in sequences with challenging visual conditions, ORB-SLAM3 encounters difficulties due to the loss of feature points in the images, leading to performance issues. MCVIO exhibits enhanced accuracy compared to VINS-Fusion in indoor settings, but it encounters significant scale and drift issues in large outdoor scenes, surpassing errors observed in VINS-Fusion. Moreover, MCVIO is highly susceptible to failure in outdoor sequences. Therefore, in this study, we establish MCVIO as the baseline for indoor environments and VINS-Fusion as the baseline for outdoor environments.

Our algorithm demonstrates superior performance compared to VINS-Fusion, achieving a remarkable 60% to 80% reduction in RMSE as measured by ATE. Furthermore, when compared to MCVIO, our approach exhibits exceptional effectiveness, with a substantial 60% to 90% reduction in ATE RMSE. This remarkable performance can be attributed to several key factors. Firstly, the integration of environmental information from multiple fields of view allows for a more comprehensive understanding of the scene. Secondly, the utilization of the AFS method ensures the selection of high-quality feature points, thereby enhancing accuracy. Lastly, the adaptive tightly coupled multi-stereoscopic optimization method further improves the robustness and accuracy of the algorithm. Overall, our algorithm demonstrates outstanding performance and reliability when facing challenging visual conditions.

C. Ablation Study

We assessed the effectiveness of the AFS method by comparing its computational costs and impact on localization accuracy throughout the entire VIO process. Table II presents a summary of the performance of N-AFS (RMSC-VIO without AFS), W-AFS (RMSC-VIO with AFS), and VINS-Fusion across seven datasets. This evaluation aims to provide insights into the contribution of the AFS method to the overall multi-stereoscopic VIO process.

Compared to N-AFS, our proposed method (W-AFS) reduces the number of feature points incorporated in the optimization process by 60 through the selection of high-quality feature points. This reduction leads to an average time savings of 9 ms and an average RMSE reduction of 46%, resulting in improved localization accuracy while minimizing computational costs.

In comparison to VINS-Fusion, W-AFS increases the number of feature points added to the optimization process by 30, resulting in a significant enhancement of localization accuracy with an average RMSE reduction of 67%. The additional time required for backend optimization is only 6 ms, and the AFS method itself consumes a mere 0.06 ms. These results demonstrate the effectiveness of the AFS method in achieving high-quality localization accuracy with relatively low computational demands.

D. Demonstration on Quadrotor Drone

We conducted real-flight experiments by deploying the proposed algorithm on a quadcopter to demonstrate its practicality and effectiveness in local visually challenging scenarios. Our experimental settings are showcased in Fig. 1.

On the playground, the quadcopter autonomously completed the entire flight mission along the pre-designed trajectory, relying on algorithmic localization results, while encountering challenging scenarios such as rapid yawing, abrupt lighting changes, and dynamic crowds. Additionally, we simulated scenarios of local visual degradation by frequently occluding the

camera alternately. Throughout these practical experiments, the quadcopters exhibited excellent stability, thus confirming the algorithm's practicality and effectiveness in handling visually challenging scenarios.

VI. Conclusion

In this letter, we presented a Multi-Stereoscopic VIO system capable of seamlessly integrating an arbitrary number of stereo cameras, demonstrating exceptional robustness in visually challenging scenarios. During system initialization, we introduced multi-view keyframes and the AFS method, which filtered out high-quality image feature points and alleviated the computational burden of multi-camera systems. In the backend optimization, we employed an adaptive tightly coupled optimization method, assigning appropriate weights to different image feature points and significantly enhancing the system's localization precision. Quantitative and qualitative comparative experiments with state-of-the-art algorithms, as well as practical flight tests, validated the effectiveness and robustness of our system in challenging visual environments. Our future efforts will explore the integration of stereo and event cameras in multi-camera systems, further extending the applicability of our methodology.

References

[1] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[3] K. Sun et al., "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 965–972, Apr. 2018.

[4] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," 2019, *arXiv:1901.03638*.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[6] C. Cadena et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[7] S. Khattak, H. Nguyen, F. Mascarich, T. Dang, and K. Alexis, "Complementary multi–modal sensor fusion for resilient robot pose estimation in subterranean environments," in *Proc. IEEE Int. Conf. Unmanned Aircr. Syst.*, 2020, pp. 1024–1029.

[8] C. Doer and G. F. Trommer, "Radar visual inertial odometry and radar thermal inertial odometry: Robust navigation even in challenging visual conditions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 331–338.

[9] J. Kuo, M. Muglikar, Z. Zhang, and D. Scaramuzza, "Redesigning slam for arbitrary multi-camera systems," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 2116–2122.

[10] A. J. Yang, C. Cui, I. A. Bârsan, R. Urtasun, and S. Wang, "Asynchronous multi-view slam," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 5669–5676.

[11] L. Zhang, D. Wisth, M. Camurri, and M. Fallon, "Balancing the budget: Feature selection and tracking for multi-camera visual-inertial odometry," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 1182–1189, Apr. 2022.

[12] P. Kaveti, S. N. Vaidyanathan, A. T. Chelvan, and H. Singh, "Design and evaluation of a generic visual slam framework for multi camera systems," *IEEE Robot. Automat. Lett.*, vol. 8, no. 11, pp. 7368–7375, Nov. 2023.

[13] Y. He, H. Yu, W. Yang, and S. Scherer, "Towards robust visual-inertial odometry with multiple non-overlapping monocular cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 9452–9458.

[14] D. Chen, N. Wang, R. Xu, W. Xie, H. Bao, and G. Zhang, "RNIN-VIO: Robust neural inertial navigation aided visual-inertial odometry in challenging scenes," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2021, pp. 275–283.

[15] T. Zhang, C. Liu, J. Li, M. Pang, and M. Wang, "A new visual inertial simultaneous localization and mapping (SLAM) algorithm based on point and line features," *Drones*, vol. 6, no. 1, 2022, Art. no. 23.

[16] K. Schauwecker and A. Zell, "On-board dual-stereo-vision for autonomous quadrotor navigation," in *Proc. IEEE Int. Conf. Unmanned Aircr. Syst.*, 2013, pp. 333–342.

[17] S. Yang, S. A. Scherer, and A. Zell, "Visual SLAM for autonomous MAVs with dual cameras," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 5227–5232.

[18] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE 6th ACM Int. Symp. Mixed Augmented Reality*, 2007, pp. 225–234.

[19] M. Miiller et al., "Robust visual-inertial state estimation with multiple odometries and efficient mapping on an MAV with ultra-wide FOV stereo vision," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3701–3708.

[20] K. Eckenhoff, P. Geneva, J. Bloecker, and G. Huang, "Multi-camera visual-inertial navigation with online intrinsic and extrinsic calibration," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 3158–3164.

[21] S. Houben, J. Quenzel, N. Krombach, and S. Behnke, "Efficient multi-camera visual-inertial slam for micro aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1616–1622.

[22] 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.

[23] M. Abate et al., "Multi-camera visual-inertial simultaneous localization and mapping for autonomous valet parking," 2023, *arXiv:2304.13182*.

[24] J. Jaekel, J. G. Mangelson, S. Scherer, and M. Kaess, "A robust multi-stereo visual-inertial odometry pipeline," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4623–4630.

[25] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[26] J. Shi and Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.

[27] Z. Yang and S. Shen, "Monocular visual–inertial state estimation with online initialization and camera–IMU extrinsic calibration," *IEEE Trans. Automat. Sci. Eng.*, vol. 14, no. 1, pp. 39–51, Jan. 2017.

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