中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)10-2839-20

论文引用格式:Zhang G F, Huang G, Xie W J, Chen D P, Wang N, Liu H M and Bao H J. 2024. A review of monocular visual-inertial SLAM. Journal of Image and Graphics, 29(10):2839-2858(章国锋, 黄赣, 谢卫健, 陈丹鹏, 王楠, 刘浩敏, 鲍虎军. 2024. 基于单目视觉惯性的同步定位与地图构建方法综述. 中国图象图形学报, 29(10):2839-2858)[DOI:10.11834/jig.230863]

基于单目视觉惯性的同步定位与地图构建方法综述

章国锋[1],黄赣[1],谢卫健[1,2],陈丹鹏[1,2],王楠[2],刘浩敏[2],鲍虎军[1*]

1. 浙江大学计算机辅助设计与图形系统全国重点实验室,杭州　310058; 2. 商汤研究院,杭州　311215

摘　要:单目视觉惯性同步定位与地图构建(visual-inertial simultaneous localization and mapping,VI-SLAM)技术因具有硬件成本低、无需对外部环境进行布置等优点,得到了广泛关注,在过去的十多年里取得了长足的进步,涌现出诸多优秀的方法和系统。由于实际场景的复杂性,不同方法难免有各自的局限性。虽然已经有一些工作对VI-SLAM方法进行了综述和评测,但大多只针对经典的VI-SLAM方法,已不能充分反映最新的VI-SLAM技术发展现状。本文首先对基于单目VI-SLAM方法的基本原理进行阐述,然后对单目VI-SLAM方法进行分类分析。为了综合全面地对比不同方法之间的优劣势,本文特别选取3个公开数据集对代表性的单目VI-SLAM方法从多个维度上进行定量评测,全面系统地分析了各类方法在实际场景尤其是增强现实应用场景中的性能。实验结果表明,基于优化或滤波和优化相结合的方法一般在跟踪精度和鲁棒性上比基于滤波的方法有优势,直接法/半直接法在全局快门拍摄的情况下精度较高,但容易受卷帘快门和光照变化的影响,尤其是大场景下误差累积较快;结合深度学习可以提高极端情况下的鲁棒性。最后,针对深度学习与V-SLAM/VI-SLAM结合、多传感器融合以及端云协同这3个研究热点,对SLAM的发展趋势进行讨论和展望。

关键词:视觉惯性同步定位与地图构建(VI-SLAM);增强现实(AR);视觉惯性数据集;多视图几何;多传感器融合

A review of monocular visual-inertial SLAM

Zhang Guofeng[1], Huang Gan[1], Xie Weijian[1,2], Chen Danpeng[1,2], Wang Nan[2], Liu Haomin[2], Bao Hujun[1*]

1. State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou　310058, China; 2. SenseTime Research, Hangzhou　311215, China

Abstract: Monocular visual-inertial simultaneous localization and mapping (VI-SLAM) is an important research topic in computer vision and robotics. It aims to estimate the pose (i. e., the position and orientation) of the device in real-time using a monocular camera with an inertial sensor while constructing the map of the environment. With the rapid development of various fields, such as augmented/virtual reality (AR/VR), robotics, and autonomous driving, monocular VI-SLAM has received widespread attention due to its advantages, including low hardware cost and no requirement for an external environment setup, among others. Over the past decade or so, monocular VI-SLAM has made significant progress and spawned many excellent methods and systems. However, because of the complexity of real-world scenarios, different methods have also shown distinct limitations. Although some works have reviewed and evaluated VI-SLAM methods, most of them only focus on classic methods, which cannot fully reflect the latest development status of VI-SLAM technology. Based on optimization type, VI-SLAM can be divided into filtering- and optimization-based methods. Filtering-based methods use filters to fuse observations from visual and inertial sensors, continuously updating the device's state information for

收稿日期:2023-12-08;**修回日期**:2024-07-11;**预印本日期**:2024-07-18

*通信作者:鲍虎军　baohujun@zju.edu.cn

基金项目:国家自然科学基金项目(61932003)

Supported by:National Natural Science Foundation of China (61932003)

localization and mapping. Additionally, depending on whether visual data association (or feature matching) is performed separately, existing methods can be divided into indirect methods (or feature-based methods) and direct methods. Furthermore, with the development and widespread application of deep learning technology, researchers have started to incorporate deep learning methods into VI-SLAM to enhance robustness in extreme conditions or perform dense reconstruction. This paper first elaborates on the basic principles of monocular VI-SLAM methods and then classifies them analytically into direct and filtering-, optimization-, feature-, and deep learning-based methods. However, most of the existing datasets and benchmarks are focused on applications like autonomous driving and drones, mainly evaluating pose accuracy. Relatively few datasets have been specifically designed for AR. For a more comprehensive comparison of the advantages and disadvantages of different methods, we select three public datasets to quantitatively evaluate representative monocular VI-SLAM methods from multiple dimensions: the widely used EuRoC dataset, the ZJU-Sensetime dataset suitable for mobile platform AR applications, and the low cost and scalable frarnework to build localization benchmark (LSFB) dataset aimed at large-scale AR scenarios. Then, we supplemented the ZJU-Sensetime dataset with a more challenging set of sequences called sequences C to enhance the variety of data types and evaluation dimensions. This extended dataset is designed to evaluate the robustness of algorithms under extreme conditions such as pure rotation, planar motion, lighting changes, and dynamic scenes. Specifically, sequences C comprise eight sequences, labeled C0-C7. In the C0 sequence, the handheld device moves around a room, performing multiple pure rotational motions. The C1 sequence involves the device mounted on a stabilized gimbal and moves freely. In the C2 sequence, the device moves in a planar motion, maintaining a constant height. The C3 sequence includes turning lights on and off during recording. In the C4 sequence, the device overlooks the floor while moving. The C5 sequence captures the exterior wall with significant parallax and minimal co-visibility, while the C6 sequence involves viewing a monitor during recording, with slight movement and changing screen content. Finally, the C7 sequence involves long-distance recording. On the EuRoC dataset, both filtering- and optimization-based VI-SLAM methods achieved good accuracy. Multi-state constraint Kalman filter (MSCKF), an early filtering-based system, showed lower accuracy and struggled with some sequences. Some methods such OpenVINS and RNIN-VIO enhanced accuracy by adding new features and deep learning-based algorithms, respectively. OKVIS, an early optimization-based system, completed all sequences but with lower accuracy. Other methods such as VINS-Mono, RD-VIO, and ORB-SLAM3 achieved significant optimizations, improving initialization, robustness, and overall accuracy. Direct methods such as DM-VIO and SVO-Pro, which we extended from DSO and SVO, respectively, showed significant improvements in accuracy through techniques like delayed marginalization and efficient use of texture information. Adaptive VIO, which is based on deep learning, achieved high accuracy by continuously updating through online learning, demonstrating adaptability to new scenarios. Furthermore, on the ZJU-Sensetime dataset, the comparison results of different methods are largely similar to those in EuRoC. The main difference is that the accuracy of the direct method DM-VIO significantly decreases when using a rolling shutter camera, whereas the semidirect method SVO-Pro has a slightly better performance. Feature-based methods do not show a significant drop in accuracy, but the smaller field of view (FoV) found in phone cameras reduces the robustness of ORB-SLAM3, Kimera, and MSCKF. Additionally, ORB-SLAM3 has high tracking accuracy but a lower completeness, while Kimera and MSCKF show increased tracking errors. HybVIO, RNIN-VIO, and RD-VIO have the highest accuracy, while HybVIO slightly outperforms the two others. The deep learning-based Adaptive VIO also shows a significant drop in accuracy and struggles to complete sequences B and C, indicating generalization and robustness issues in complex scenarios. On the LSFB dataset, the comparison results are consistent with those in small-scale datasets. The methods with the highest accuracy in small scenes, such as RNIN-VIO, HybVIO, and RD-VIO, continue to show high accuracy in large scenes. In particular, RNIN-VIO demonstrates even more significant accuracy advantages in large scenes. In large-scale scenes, many feature points are distant and lack parallax, leading to rapid accumulation of errors, especially in methods that are heavily rely on visual constraints. The neural inertial network-based RNIN-VIO can better maximize IMU observations, reducing dependence on visual data. The VINS-Mono also shows significant advantages in large scenes, as its sliding window optimization facilitating the early inclusion of small-parallax feature points, effectively controlling error accumulation. In contrast, ORB-SLAM3, which relies on local maps, requires sufficient parallax before adding feature points to the local map, which can lead to insufficient visual constraints in distant environments and ultimately cause error accumula-

tion and even tracking loss. The experimental results also show that optimization-based or combined filtering-optimization methods generally outperform filtering-based methods in terms of tracking accuracy and robustness. At the same time, direct/semidirect methods perform well when shooting with a global shutter camera, but are prone to error accumulation, especially in large scenes when affected by rolling shutter and light changes. Combining deep learning can improve robustness in extreme situations. Finally, the development trend of SLAM is discussed and prospected in this work based on three research hotspots: combining deep learning with V-SLAM/VI-SLAM, multisensor fusion, and end-cloud collaboration.

Key words: visual-inertial SLAM (VI-SLAM); augmented reality (AR); visual-inertial dataset; multiple-view geometry; multi-sensor fusion

0 引 言

随着增强/虚拟现实(augmented reality/virtual reality,AR/VR)、机器人和自动驾驶等领域的快速发展,同步定位与地图构建(simultaneous localization and mapping,SLAM)技术作为其中的关键技术,得到了人们的广泛关注。SLAM旨在通过使用传感器实时获取的数据来估计设备的位置和姿态,并同时构建周围环境的三维地图。在过去的几十年中,SLAM技术进展显著,并成功应用到无人机、机器人、自动驾驶和增强/虚拟现实等诸多领域。

因采用传感器的不同,SLAM方法可以分为视觉SLAM(visual SLAM,V-SLAM)、视觉惯性SLAM(visual-inertial SLAM,VI-SLAM)、激光SLAM等。其中,基于单目相机和惯性传感器(包括加速度计和陀螺仪等)的SLAM方法(简称单目VI-SLAM)因其具有成本和功耗低、鲁棒性高的优点,得到了广泛的应用。本文主要讨论单目VI-SLAM方法,并对代表性方法进行评测。

根据融合方法的不同,VI-SLAM可以分为基于滤波的方法和基于优化的方法。基于滤波的方法通过滤波器来融合视觉和惯性传感器的观测,不断地更新设备的状态信息来实现定位和建图。这些方法包括多状态约束卡尔曼滤波(multi-state constraint kalman filter, MSCKF)(Mourikis 和 Roumeliotis, 2007)及其扩展SR-ISWF(square-root inverse sliding window filter)(Wu等,2015)等。基于优化的方法(例如VINS-Mono(Qin等, 2018)、ORB-SLAM3(Campos等,2021)等)通过最大化似然函数等方式来实现对设备位姿的求解和地图构建。

另外,从视觉数据关联(或者特征匹配)是否需要单独进行,我们可以将现有方法分为非直接法(或者特征点法)和直接法两种。VINS-Mono、ORB-SLAM3这类方法都是先进行特征匹配,再进行融合优化,属于非直接法;而DSO(direct sparse odometry)系列,如VI-DSO(visual-inertial DSO)(Von Stumberg等,2018)、DM-VIO(delayed marginalization visual-inertial odometry)(Von Stumberg和Cremers,2022)等直接法则不提取特征点,而是直接通过比较像素颜色值来求解相机位姿。像SVO(semidirect visual odometry)系列,同时用到了特征点法和直接法,我们称之为半直接法,例如SVO Pro(Forster等,2017)就是其中一个代表性的工作。

随着深度学习技术的发展和广泛应用,研究人员开始把深度学习方法引入到VI-SLAM中提升在极端情况下的鲁棒性或进行稠密的在线重建,包括融合深度学习的惯导系统的RNIN-VIO(robust neural inertial navigation aided visual-inertial odometry)(Chen等,2021),以及融合基于深度网络单目估计的CodeVIO(Zuo等,2021)等。

随着V-SLAM/VI-SLAM技术的快速发展,一些综述性的文章(刘浩敏 等,2016;Cadena等,2016;Delmerico和Scaramuzza,2018;Li等,2019;Servières等,2021;曾庆化 等,2022;Wenzel等,2022)也陆续涌现出来,出现了一些相关的数据集或Benchmark,但这些数据集大部分是针对自动驾驶、无人机等应用场景,而且基本只评测位姿精度,相对来说针对AR的数据集比较少。PennCOSYVIO(Pfrommer等,2017)、ADVIO(an authentic dataset for visual-inertial odometry)(Cortés等,2018)等数据集虽然可以用来评测基于移动设备的VI-SLAM性能,但其真值的精度较低。ZJU-Sensetime数据集(Li等,2019)不但提供了一个面向AR的室内小场景视觉惯性数据集,而且提供了相应的评测标准,包括跟踪精度、初始化质量、跟踪鲁棒性、重定位时间等。在这之后,针对

AR应用的大尺度场景数据集也开始出现，例如LSFB数据集（Liu等，2020a）和LaMAR（benchmarking localization and mapping for augmented reality）（Sarlin等，2022）。此外，由于VI-SLAM技术发展很快，现有综述或评测大多只针对经典的VI-SLAM方法，已不能充分反映最新的VI-SLAM技术发展现状。

本文旨在综合全面地评估现有的代表性开源VI-SLAM算法（包括基于滤波、基于优化、特征点法、直接法以及结合深度学习的算法等），并深入分析这些方法的优缺点及其适用领域。为此，我们在ZJU-Sensetime数据集的基础上补充了新的数据以及评测标准，进一步完善其数据类型和评估维度，并采用LSFB数据集来评测VI-SLAM方法在大尺度场景下的跟踪定位精度。通过对这些方法的综合分析、比较和总结，并对未来该领域的研究方向进行展望，为SLAM研究者提供有关该领域的最新发展和未来趋势的综述参考。

1 基本原理

视觉SLAM（V-SLAM）可以根据多视图几何理论，从输入的视频流中恢复出相机在未知环境中的位姿，并同时构建环境的三维地图。视觉信息容易受到外界环境干扰，如光照变化、弱纹理场景、快速运动导致的图像模糊、动态物体等，从而导致V-SLAM的精度下降而且鲁棒性变差。此外，单目V-SLAM在没有先验信息的条件下无法估计实际的物理环境尺度。视觉惯性SLAM（VI-SLAM）在V-SLAM基础上，通过进一步融合惯性测量单元（inertial measurement unit，IMU）信息（加速度和陀螺仪信息），不但可以恢复出正确的尺度而且可以提升鲁棒性。

输入图像序列，V-SLAM可以恢复出每帧图像的相机姿态$\boldsymbol{C}_i$（包含相机的三维位置$\boldsymbol{p}_i$和旋转矩阵$\boldsymbol{R}_i$）以及场景的三维结构$\boldsymbol{X}_j$。将3D点$\boldsymbol{X}_j$投影至第i帧图像上得到投影点$\boldsymbol{x}_{ij}$，可以用下式计算

$$\boldsymbol{x}_{ij}=h(\boldsymbol{C}_i,\boldsymbol{X}_j)=\pi(\boldsymbol{R}_i^{\mathrm{T}}(\boldsymbol{X}_j-\boldsymbol{p}_i)) \tag{1}$$

式中，$h(\boldsymbol{C}_i,\boldsymbol{X}_j)$表示将三维点$\boldsymbol{X}_j$通过相机的位姿$\boldsymbol{C}_i$映射到图像平面上的过程，$\pi$是投影函数，通常假定内参矩阵已知，目的是将空间中3D点投影到相机平面中。

V-SLAM需要将不同图像中对应相同地图点的图像点匹配起来，从而形成同一个地图点在不同图像上的2D观测。对于m幅图像和n个3D地图点，可以通过最小化以下能量函数来同时求解每帧图像的相机位姿和地图点三维位置，即

$$\underset{\boldsymbol{C}_1\cdots\boldsymbol{C}_m,\boldsymbol{X}_1\cdots\boldsymbol{X}_n}{\arg\min}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|h(\boldsymbol{C}_i,\boldsymbol{X}_j)-\hat{\boldsymbol{X}}_{ij}\right\|_{\sum\boldsymbol{X}_{ij}}^{2} \tag{2}$$

式中，$\sum\boldsymbol{X}_{ij}$为图像观测点噪声，通常假设满足高斯分布。这个优化称为集束调整（bundle adjustment，BA）。通过BA得到最优的相机位姿和地图点，使得地图点在图像上的投影误差尽可能小。

V-SLAM需要准确而且足够多的匹配点，但是实际环境往往复杂多变，导致V-SLAM在实际应用中很容易遇到鲁棒性问题，而且单目V-SLAM难以准确恢复真实环境的尺度。IMU传感器不依赖外部环境，可以得到设备的自身加速度和角速度，但是其直接得到的数值与真实值存在一个偏置（随时间做随机游走的变化），一般需要结合其他信息来估计偏置，如果偏置估计不准，位姿估计的误差将迅速累积。IMU虽然不受外界环境影响，但没有其他信息的情况下一般只能较短时间内维持可靠的位姿估计，而基于视觉信息虽然可以维持长时间高精度的位姿估计，但容易受环境变化的影响。因为IMU和视觉信息有很强的互补性，所以可以将二者融合来提升定位建图的鲁棒性和精度，同时在单目情况下也能准确估计真实环境尺度。IMU传感器可以获得设备自身坐标系下的线性加速度$\boldsymbol{a}(t)$和角速度$\boldsymbol{\omega}(t)$，测量模型如下：

$$\begin{cases}\boldsymbol{\omega}(t)=\tilde{\boldsymbol{\omega}}(t)+\boldsymbol{b}_{\omega}+\boldsymbol{n}_{\omega}\\ \boldsymbol{a}(t)=\boldsymbol{R}_t^{\mathrm{T}}(\tilde{\boldsymbol{a}}_w(t)-\boldsymbol{g})+\boldsymbol{b}_a+\boldsymbol{n}_a\\ \dot{\boldsymbol{b}}_{\omega}=\boldsymbol{\eta}_{\omega}\\ \dot{\boldsymbol{b}}_a=\boldsymbol{\eta}_a\end{cases} \tag{3}$$

式中，$\tilde{\boldsymbol{\omega}}(t)$表示IMU坐标系下真正的角速度，$\tilde{\boldsymbol{a}}_w(t)$表示惯性世界系下真正加速度，$\boldsymbol{R}_t$是IMU在$t$时刻旋转矩阵。$\boldsymbol{\eta}_a$和$\boldsymbol{\eta}_{\omega}$是加速度计和陀螺仪的测量白噪声；$\boldsymbol{b}_a$和$\boldsymbol{b}_w$是加速度计和陀螺仪随机游走噪声，$\dot{\boldsymbol{b}}_{\omega}$和$\dot{\boldsymbol{b}}_a$表示导数，为随时间的变化率。将相邻两图像帧$(\boldsymbol{C}_i,\boldsymbol{C}_{i+1})$间的IMU数据标记为集合$\boldsymbol{Z}_i$。VI-SLAM通常优化的状态量包括每帧的相机位姿$\boldsymbol{C}_i$，速度$\boldsymbol{V}_i$，加速度偏置$\boldsymbol{b}_a$，陀螺仪偏置$\boldsymbol{b}_w$以及地图点$\boldsymbol{X}_j$。结合视

觉和IMU信息的VI-SLAM，一般可以通过求解如下目标函数来恢复相机的位姿和三维点：

$$\underset{\boldsymbol{C}_1,\ldots\boldsymbol{C}_m,\boldsymbol{X}_1,\ldots\boldsymbol{X}_n}{\arg\min}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|h(\boldsymbol{C}_i,\boldsymbol{X}_j)-\hat{\boldsymbol{X}}_{ij}\right\|_{\sum X_{ij}}^{2}+\sum_{i=1}^{m-1}\left\|f(\boldsymbol{C}_i,\boldsymbol{Z}_i)\ominus\boldsymbol{C}_{i+1}\right\|_{\sum Z_i}^{2}\right\}\tag{4}$$

式中，$f(\boldsymbol{C}_i,\boldsymbol{Z}_i)$项表示基于$\boldsymbol{C}_i$和$\boldsymbol{Z}_i$预测的$\boldsymbol{C}_{i+1}$。常用的做法是采用IMU预积分方法，在IMU偏置状态更新时能够比较方便地更新IMU预积分(Forster等，2017b)。$\ominus$表示计算两个位姿之间的误差，即相对变换。

除了通过最小化目标函数的优化方法，还可以使用滤波方法求解相机位姿和三维点。系统状态使用一个高斯概率模型表达，并通过滤波器不断更新。不同的状态设计和滤波方法将衍生出不同SLAM系统，滤波器常使用卡尔曼滤波或均方根滤波等方法。

2 代表性方法

将代表性的单目VI-SLAM方法按基于滤波的方法、基于优化的方法，直接法以及基于深度学习的方法这4类进行分别介绍，并在3个代表性的数据集上进行定量评测和对比分析。

2.1 滤波法

滤波方法通常是基于扩展卡尔曼滤波器理论，使用运动假设预测最新状态，相机观测进行概率更新，从而得到状态的最优估计。滤波方法是一种递归求解，只需一次预测和更新即可得到最优估计，相比于优化方法的多次重新构造问题进行迭代优化，效率更高。但是滤波方法需要较好的状态初值，否则滤波器可能无法正确收敛。因为VI-SLAM可以使用IMU预测较优的相机初值状态，所以滤波方法在VI-SLAM中广泛使用。基于滤波的VI-SLAM具备高效、低功耗等特性，在硬件受限或者对系统实时性要求较高的场景下得到了广泛的应用。典型代表方法有MSCKF(Mourikis和Roumeliotis，2007)、OpenVINS(Geneva等，2020)、RNIN-VIO(Chen等，2021)等。

MSCKF是早期提出的基于滤波的视觉惯性里程计(visual-inertial odometry，VIO)。MSCKF维护一个M帧的滑动窗口，估计状态包括M帧相机姿态和最新的IMU状态。为了高效求解问题，MSCKF并不求解3D点状态，而是利用相机姿态三角化3D点后，马上边缘化3D点，将投影约束直接变为相机位姿约束。这一方案使得VIO能够在保持较高计算效率的情况下获得较好精度。OpenVINS在MSCKF框架的基础上，加入SLAM的局部地图点、相机内外参标定和时间偏移校准等措施有效提升定位精度。由于视觉信息容易受外界环境影响，而没有视觉信息约束的情况下IMU的误差会迅速累积，因此长时间的视觉信息失效会导致VIO跟踪失败。因为神经惯性网络可以利用先验知识学习IMU运动模式，支持长时间下纯IMU的跟踪定位，为了提升极端情况下VIO的鲁棒性，RNIN-VIO(Chen等，2021)提出结合神经惯性网络到VIO中，实现了极端情况下的鲁棒跟踪定位。

2.2 优化法

与滤波方法基于马尔可夫假设不同，基于优化的方法每次都会利用历史观测来重新构造优化问题，并通过多次迭代来最小化观测误差以更新相机和地图点。因此基于优化的方法通常具有更高的精度和鲁棒性，但计算量也更大。基于优化的方法通常会维护一个大小固定的滑动窗口，每当新的帧加入到窗口中时，将窗口内维护的视觉观测和IMU观测一起构造优化问题，通过多轮迭代优化更新窗口状态。当窗口帧超过窗口大小时，通过一定的策略选出某一个历史帧进行边缘化，并将该历史帧以及相关的地图点从滑动窗口中移除。基于优化的视觉惯性方法的代表有OKVIS(Leutenegger等，2013)、VINS-Mono(Qin等，2018)、VINS-Fusion(Qin等，2019a，b)、RD-VIO(Li等，2024)、Kimera(Rosinol等，2020)、HybVIO(Seiskari等，2022)等。

OKVIS是早期基于滑动窗口的VIO系统，支持静止初始化，不包含回环检测和重定位模块。它的前端采用多尺度的Harris角点(Harris和Stephens，1988)结合BRISK描述子(Leutenegger等，2011)进行视觉跟踪。在窗口维护设计上，采用了N个关键帧和S个时序上连续的相邻帧的设计。VINS-Mono的窗口设计与OKVIS类似，前端则采用了Harris角点结合LK光流跟踪的方法进行视觉跟踪，节省了描述子的计算和匹配耗时，从而提高了视觉跟踪的效率。VINS-Mono设计了一套松耦合的初始化策略，首先对通过SfM(structure from motion)方法初始化相机运动轨迹，然后结合IMU进行在线VIO标定，从而实

现相机移动状态下的初始化。除此之外，VINS-Mono还添加了一个基于DBoW2（Gálvez-López和Tardos，2012）的回环检测线程，在检测到回环后进行4DoF的位姿图优化来消除误差累积。VINS-Fusion是VINS-Mono的拓展，相较于VINS-Mono支持更多传感器组合，包括单/双目相机+IMU，以及双目相机+GPS。ORB-SLAM3是第一个同时支持短、中、长期和多地图视觉关联的V-SLAM/VI-SLAM系统，其基本上延续了ORB-SLAM2的框架设计，但在每个线程（前端跟踪线程、局部建图线程、回路检测和全局优化线程）分别对IMU信息进行了适配，同时引入了多地图设计，在视觉长时间丢失、IMU积分不稳定的情况下自动生成新的子地图并重新初始化，然后通过地图合并实现地图的复用，增强系统的鲁棒性。

针对VIO跟踪中容易受外部运动物体干扰以及纯旋转等退化运动影响的问题，RD-VIO（Li等，2024）提出了一种新的面向增强现实的轻量级里程计。为了有效地辨识移动的关键点，RD-VIO引入了IMU运动先验信息，并通过自适应的IMU-PARSAC（IMU-prior-based adaptive RanSAC）算法来优化动态物体的剔除过程，该算法是基于纯视觉动态离群点剔除策略的扩展（Tan等，2013；潘小鹃 等，2023）。在处理纯旋转或静止等运动退化场景时，该系统首先检测传入图像帧的运动类型，会延迟地图点的三角化，以防止产生错误的深度估计。此外，RD-VIO还引入了子帧窗口的滑窗优化策略，通过有效提供位置约束来准确估计IMU的零偏。这些策略有效提升了跟踪定位的精度和鲁棒性。

Kimera是一个开源的用于实时度量语义的VI-SLAM系统，采用模块化的架构，包括4个模块（视觉惯性里程计、因子图优化器、3D网格生成和稠密3D语义重建），各个模块可以单独或组合运行。其中Kimera-VIO模块实现了基于关键帧的最大后验视觉惯性里程计，前端处理原始IMU和视觉数据，执行流形预积分和光流跟踪，而后端是系统的核心部分，使用因子图整合测量值以估计传感器状态（包括姿态、速度和传感器偏差）。在估计过程中，无结构的视觉模型使用DLT估计观测特征的3D位置，并从VIO状态中消除相应的3D点，此外通过去除退化点和异常值，提高了估计的鲁棒性，以及通过边缘化处理确保了系统的连贯性。

HybVIO创新性地将基于滤波的视觉惯性里程计（VIO）和基于优化的SLAM结合起来，以提高自主移动设备的自我定位和环境地图构建的准确性与可靠性。该研究不仅将概率惯性视觉里程计方法从单目扩展到双目，还引入了Orstein-Uhlembeck随机游走过程以改进IMU偏差建模，并通过改进的离群点检测、平稳性检测和特征轨迹选择机制，有效提高了系统的鲁棒性。特别地，HybVIO创新性地将基于扩展卡尔曼滤波的VIO与基于优化的SLAM相结合，实现了对物体运动的实时跟踪和一致性环境地图的构建。

2.3 直接法

传统基于特征点的视觉SLAM方法，都需要先提取和匹配特征点，再基于特征匹配的结果进行状态估计（即恢复相机位姿和特征点三维坐标），状态估计极大依赖于特征匹配的结果。在弱纹理的场景中，特征匹配十分困难，进而导致状态估计的鲁棒性也受到极大影响。这类方法的根本局限在于，特征匹配独立于状态估计，状态估计的结果未能辅助特征匹配消除匹配歧义。直接法通过将特征匹配和状态估计两个阶段合并来突破这一局限，直接通过最小化两帧图像间的光度误差进行状态估计，从而提升弱纹理情况下的鲁棒性。基于直接法的代表性视觉惯性方法有SVO Pro（Forster等，2017a）和DM-VIO（Von Stumberg和Cremers，2022）。

SVO Pro源于基于半直接法的纯视觉方法SVO（Forster等，2014）。通过最小化相邻两帧图像间光度误差以跟踪两帧相对运动。具体而言，将前一帧图像中梯度较大的像素，由其逆深度和相机相对运动投影至后一帧图像，最小化两帧图像的亮度差异。SVO由直接法跟踪相机运动后，仍需在投影位置附近搜索匹配点，并由BA通过最小化重投影误差以优化关键帧相机位姿和匹配点三维点坐标，故称为半直接法。SVO Pro在SVO的基础上，进一步采用OKVIS基于滑动窗口的视觉惯性优化方法以融合IMU。

DM-VIO（Von Stumberg和Cremers，2022）源于基于直接法的纯视觉方法DSO（Engel等，2018）。与半直接法的SVO不同，DSO直接由BA通过最小化光度误差以优化关键帧相机位姿和逆深度，同时优化每帧图像的光照参数以应对不同帧间的光照变化。DSO也采用了滑动窗口以控制BA的计算复杂度，对移出滑动窗口的变量进行边缘化。DM-VIO在DSO

的基础上融合IMU。与非直接法或半直接法相比，直接法融合IMU时对边缘化时刻的线性化误差更敏感。DM-VIO提出了一种延时边缘化技术来缓解线性化误差，系统维护了两个因子图，主因子图执行常规边缘化来确保边缘化的及时性，而延时因子图中状态边缘化的时间可以有所延时，其线性化点可以与其最新状态保持一致，从而减小线性化误差。

2.4 学习法

近年来，基于学习的方法在视觉里程计、惯性里程计和视觉惯性里程计领域得到了广泛研究，并取得了不错的成果。基于端到端学习的SLAM系统通常使用姿态和深度估计网络替代传统的跟踪或建图模块，并通过监督或者自监督方式进行训练。对于VO/V-SLAM，有人提出将深度学习与传统模块结合，DROID-SLAM（Teed 和 Deng，2021）和 DPVO（deep patch visual odometry）（Teed 等，2023）将迭代的视觉匹配更新与可微分的集束调整结合，在多个数据集上展示了卓越的性能。iSLAM（Fu 等，2024）结合基于学习的VO与图优化，并通过自监督学习进一步提升了性能。

与此同时，融合了IMU信息的VIO/VI-SLAM工作也相继出现，例如，VINet（Clark 等，2017）首次提出了端到端可训练的视觉惯性里程计方法，在中间特征表示级别执行数据融合。该方法不仅解决了相机和IMU手动同步和标定的需求，还整合了特定领域的信息，显著减少了漂移问题。DeepVIO（Han 等，2019）使用双目序列估计每个场景的深度和稠密点云，获得包括3D光流和6自由度姿态的3D几何约束作为监督信号，减少了相机和IMU标定的误差、数据不同步和丢失的影响。SelfVIO（Almalioglu 等，2022）利用网络编码并自适应地融合视觉和IMU信息，通过自监督学习姿态和深度估计。

Adaptive VIO（Pan 等，2024）将在线持续学习与传统的非线性优化相结合，通过两个网络来预测视觉匹配关系和IMU偏置。不同于端到端融合两种模态的特征直接预测位姿，Adaptive VIO将神经网络与视觉BA相结合，优化后的估计结果将反馈给视觉和IMU偏置网络，以自监督的方式优化网络。这种学习—优化相结合的框架和反馈机制使得系统能够进行在线持续学习。

2.5 对比分析

为了充分地对比不同方法之间的优劣势，我们选3个代表性的数据集（包括业界广泛使用的EuRoC数据集（Burri 等，2016）、适用于移动平台增强现实应用的ZJU-Sensetime数据集（Li 等，2019）和针对大尺度场景AR的LSFB数据集（Liu 等，2024）对代表性方法进行量化评测。现有的SLAM系统，如VINS-Mono（Qin 等，2018）和 ORB-SLAM3（Campos 等，2021）等，大多采用并行跟踪和建图架构，前端线程实时估计相机位姿，后端异步优化线程一般包含重定位、回路闭合和全局/局部优化模块，通过更多的时间来优化关键帧位姿和地图的精度，以及一旦丢失能及时恢复。考虑到AR等实时应用需要的是实时位姿，因此在我们的评测中，只考虑前端实时输出的位姿，不考虑后端离线优化后的位姿。所有的实验都在搭载Intel i7-12700F2. 1 GHz CPU和32 GB内存的主机上进行。操作系统为Ubuntu 20. 04，同时安装了ROS Noetic。每个系统在3个数据集的每个序列上都运行3次，结果取平均值，以减少偶然性带来的结果偏差。

EuRoC是一个基于六轴飞行器的数据集，包含了从3个场景（两个小房间和一个机房场景）捕获的11个序列。数据包括以20 Hz帧率采集的752 × 480像素分辨率的双目图像数据，以及以200Hz捕获的IMU传感器数据，其真值位姿是从VICON和Leica MS50设备获得（精度约为1 mm）。所有数据均以硬件同步的形式同步到统一的时钟下。该数据集用到的相机是全局快门相机，具有较高的图像质量。因为良好的硬件同步和高精度的真值位姿，使得EuRoC广泛地应用在各种VIO/VI-SLAM的精度评测中。

ZJU-Sensetime数据集是一个面向移动端AR的数据集，该数据集以iPhone X（iOS）和小米8（Android）这两个代表性的移动设备做为采集终端。图像数据以30 Hz的帧率采集，普通手机的相机均是卷帘曝光模式，因此图像质量逊色于EuRoC数据集。IMU数据分别从iOS和Android的接口中读取，帧率分别为100 Hz和400 Hz。真值位姿通过VICON设备获取。采集的场景是一个3 m × 5 m的小房间。

ZJU-Sensetime分为A、B两个子数据集。A数据包括直线、绕圈、旋转等常规运动，用于评估算法精度。B数据包括快速移动、快速旋转、短时间遮挡等挑战情况，用于评估算法鲁棒性。在此基础上，我们进一步补充了一批更有挑战性的子数据集C，用于

评估算法在纯旋转、平面运动、开关灯、动态场景等极端情况下的鲁棒性。具体来说，C数据集包含了C0-C7共8个序列，图1展示了8个序列的代表性图像，图2展示了各序列图像的真实轨迹。C0序列中，手持设备在房间内行走，伴随多次纯旋转运动。C1序列将设备固定在防抖云台上，进行自由运动。C2序列中，设备进行平面运动，高度保持不变。C3序列在录制过程中进行了灯光的开启和关闭。C4序列将设备俯瞰地板运动。C5序列环视外墙，具有较大视差和较小共视情况。C6序列在录制途中观看显示器，伴随轻微运动，画面发生变化。C7序列是长距离录制，其中在房间内记录了真实轨迹的开始和结束阶段。

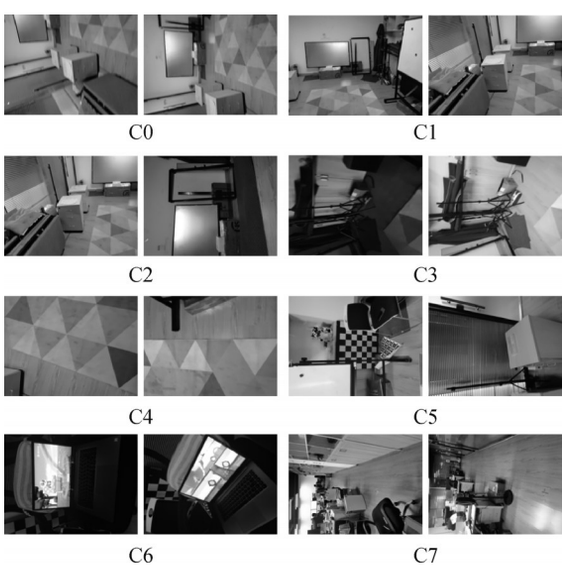

图1　ZJU-Sensetime补充C数据集的8个序列的代表图像

Fig. 1　The representative images of 8 sequences from ZJU-Sensetime C dataset

LSFB数据集是一个面向大尺度场景移动AR的的数据集，通过SfM技术重建场景高精地图，并通过与高精地图匹配后联合优化，恢复每段移动序列在高精地图中的运动轨迹作为位姿真值，真值精度可达厘米级。LSFB数据集包括在4个数百到上万平米的室内、室外大尺度场景中的常规运动数据，以及5个典型场景（如远景、长走廊、楼梯等）中的一系列典型运动数据。每个场景由华为手机、苹果手机和AR眼镜3种设备拍摄。本文选用华为手机拍摄的数据进行对比分析。

2.5.1　EuRoC数据集结果分析

表1对现有的VIO/VI-SLAM系统在EuRoC数据集上的绝对位置误差（absolute positional error，APE）（Li等，2019）进行了对比。由于ORB_SLAM3采用多子地图的模式，使用了实时跟踪最长的序列进行测评，并统计了其恢复的绝对位置的均方根误差和轨迹完整度（即括号内的数值）。表格中的第1列MH_01-05，V1_01-03，V2_01-03表示EuRoC数据集的11个序列，涵盖了3个场景。从表中可以看出，无论是基于滤波的方案（如OpenVINS、RNIN-VIO等）还是基于优化的方案（如VINS-Mono、ORB-SLAM3、DM-VIO等），都能取得不错的精度结果。

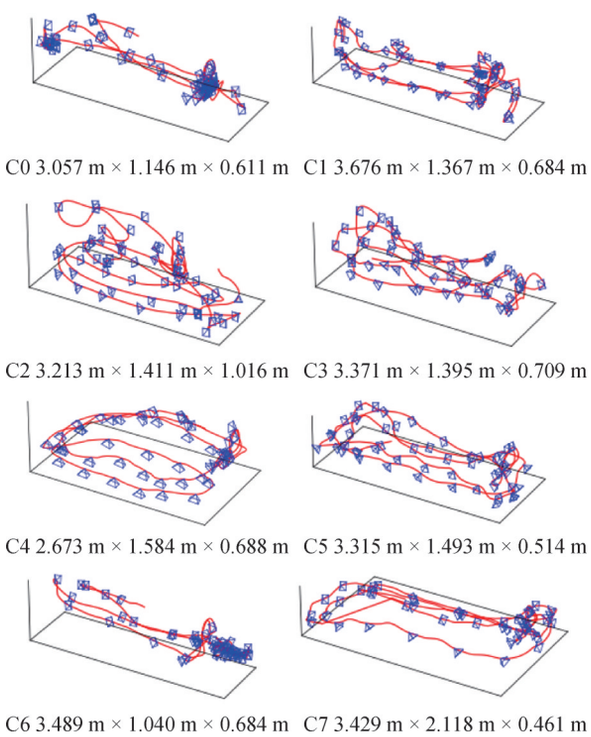

图2　ZJU-Sensetime补充C数据集的8个序列的图像真实轨迹

Fig. 2　The ground truth trajectories of ZJU-Sensetime C sequences

MSCKF作为最早的基于滤波的VIO系统，实现方案较为简单。由于原作者未开源其实现代码，本文实验使用了第三方实现的单目MSCKF-VIO[1]算法。可以看到，该方案的整体精度相对较低，且有部分序列不能完全跑完。后续的滤波方案在MSCKF的基础上做了较多的改进，如OpenVINS增加了对FEJ、SLAM特征、在线标定等的支持，精度有大幅提升。RNIN-VIO引入了基于深度学习的惯性导航算法，能够应对更加挑战性的场景，平均精度指标也有显著提升。

OKVIS是早期的采用优化的VIO系统，能够完整跑完所有的序列，但是平均精度也比较低。后续

表 1　EuRoC 数据集上的跟踪精度对比

Table 1　Tracking accuracy on the EuRoC dataset

	APE/m											
	VINS-Mono	ORB-SLAM3	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro	AdaVIO
MH_01	0.148	0.098 (85.01%)	0.260	0.109	0.848	0.184	0.655	0.344	0.093	0.069	0.162	**0.050**
MH_02	0.176	0.182 (76.22%)	0.216	0.115	1.594	0.061	0.463	0.183	0.129	**0.047**	0.145	0.055
MH_03	0.185	**0.067** (78.19%)	0.643	0.141	0.695	0.116	0.323	0.155	0.130	0.094	0.159	0.069
MH_04	0.356	0.169 (85.84%)	0.607	0.247	3.322	0.192	–	0.195	0.205	0.109	0.312	**0.092**
MH_05	0.300	**0.055** (72.16%)	0.550	0.267	1.182	0.297	0.995	0.538	0.218	0.125	0.214	0.124
V1_01	0.081	0.039 (100.0%)	0.152	0.060	0.353	0.064	0.143	0.064	0.047	0.061	0.091	**0.035**
V1_02	0.121	**0.022** (96.84%)	0.334	0.091	0.343	0.061	0.533	0.062	0.051	0.043	0.126	0.045
V1_03	0.164	**0.058** (94.42%)	0.403	0.168	0.256	0.116	–	0.074	0.079	–	0.101	0.073
V2_01	0.084	0.081 (100.0%)	0.125	0.058	0.481	0.049	0.236	0.120	0.049	**0.033**	0.115	0.052
V2_02	0.156	0.477 (100.0%)	0.247	0.100	0.393	0.079	–	0.064	0.075	**0.038**	0.124	0.086
V2_03	0.274	**0.085** (100.0%)	0.291	0.147	0.509	0.126	–	0.206	0.131	0.143	0.241	–
平均值	0.186	0.121 (89.88%)	0.348	0.136	0.907	0.122	0.478	0.182	0.110	0.076	0.163	**0.068**

注：加粗字体为每行最优值，“–”表示跟踪失败。

的算法如VINS-Mono、RD-VIO、ORB-SLAM3等，都基于优化方案做了大量改进。VINS-Mono对初始化和划窗优化策略做了大量的改进，且增加了回路闭合机制，使得算法可以在大多数场景下完成快速鲁棒的初始化和运动跟踪，同时也提升了算法精度。RD-VIO也是一种基于划窗优化的VIO方案，不同的是，RD-VIO增加了动态场景下算法对外点的过滤机制，使得系统在动态场景下仍然有较高的鲁棒性。ORB-SLAM3则借助了ORB-SLAM自身强大的地图管理机制，使得其在建图方面具备显著的精度优势，而这也会反过来提升VIO的精度。ORB-SLAM3由于采用了多子地图的模式，一旦当前地图跟踪丢失，会重新生成一个新的子地图，新子地图坐标系与上一个子地图没有关联。在我们的对比评测中，为了反映算法的真实精度，我们选取了ORB-SLAM3的实时位姿中最长的、未跟踪丢失的子地图序列用来进行评测，同时也统计了相应的轨迹的完整度作为参考。

DM-VIO和SVO-Pro是两个典型的直接法/半直接法的VIO方案，两者分别是DSO和SVO方法的拓展。DM-VIO采用了延迟边缘化的策略，通过利用更多帧的历史数据，来完成对VIO初始状态更加精准的估计，进而对后续的跟踪精度也有显著提升。另外DSO基于直接法的方案，充分利用了整幅图的纹理信息，在全局曝光的EuRoC数据集下，展现出非常好的精度。SVO-Pro则延续了SVO自身的高效率和鲁棒性，在EuRoC数据集上，算法的精度和效率都是适中的状态。基于深度学习的Adaptive VIO通过在线持续学习进行更新可以改善整个系统的性能，在该数据集上取得了很高的精度，显示出了对未见场景的潜在适应能力。

2.5.2　ZJU-Sensetime数据集结果分析

表2用ZJU-Sensetime-A数据对比了不同VIO/VI-SLAM方法的APE，作为不同方法的精度指标。在ZJU-Sensetime-A的常规运动数据下，不同方法的对比结果与EuRoC大部分类似。主要的区别在于，对于手机上卷帘快门的摄像头，基于直接法的DM-VIO的精度下降显著。基于半直接法的SVO-Pro虽然也有下降，但优于纯直接法的DM-VIO。其他基于特征点的方法，精度并没有明显下降。另一方面，手机摄像头与无人机搭载的鱼眼相机相比视场（field of view，FoV）更小，导致ORB-SLAM3、Kimera和MSCKF的鲁棒性明显下降。其中ORB-SLAM3虽然跟踪精度较高，但成功率明显下降。Kimera和MSCKF则表现为跟踪误差明显上升。其他方法的对比结果与EuRoC类似。所有系统中，HybVIO、RNIN-VIO和RD-VIO的精度最高，这3个系统的定位精度较为接近，其中HybVIO精度比其他两个系统略高一些。基于深度学习的Adaptive VIO在该序列中精度显著下降，并且难以完整地运行完序列B和C，这也说明了目前基于深度学习的VIO算法在

真实复杂场景尤其是极具挑战性的场景上存在泛化性和鲁棒性问题。因此，只对比了该算法在EuRoC数据集以及ZJU-Sensetime的简单序列A中的精度。

对于移动增强现实来说，快速而准确的初始化对于良好的用户体验非常重要。表3使用ZJU-Sensetime-A数据集对比了不同方法的初始化质量，选用Li等人(2019)提出的初始化质量指标，即

$$\varepsilon_{\text{init}} = t_{\text{init}}\left(\varepsilon_{\text{scale}} + 0.01\right)^{0.5} \tag{5}$$

式中，t_{init}为初始化时间，$\varepsilon_{\text{scale}}$为相对尺度误差。从表3也反映了初始化质量与表2的精度之间的联系。测试中，VINS-Mono、RD-VIO、HybVIO和RNIN-VIO体现出良好的初始化质量，而MSCKF通常需要更长的时间来收敛尺度，且在某些序列上初始化尺度误差不太稳定，导致不能进行完整跟踪。此外，基于直接法的DM-VIO和SVO-Pro由于过于依赖像素级的信息，初始化质量上明显要逊色于其他方法，尤其是在手机摄像头易受噪声和图像变化影响的情况下。

在鲁棒性方面，表4和表5分别用已有的ZJU-Sensetime-B数据集和本文工作新采集的ZJU-Sensetime-C数据集对比了不同方法的鲁棒性。选用Li等人(2019)提出的鲁棒性指标，即

$$\varepsilon_{\text{R}} = \left(\alpha_{\text{lost}} + 0.05\right)\left(\varepsilon_{\text{RL}} + 0.1\varepsilon_{\text{APE}}\right) \tag{6}$$

式中，α_{lost}为跟踪丢失的比例，ε_{RL}为重定位前后两段子序列轨迹与真值轨迹相似变换的差异，用于度量

表2　ZJU-Sensetime A序列跟踪精度对比

Table 2　Tracking accuracy on the ZJU-Sensetime A sequences

	APE/m											
	VINS-Mono	ORB-SLAM3	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro	AdaVIO
A0	0.140	0.129 (78.93%)	0.104	0.088	0.246	**0.050**	6.162	0.199	0.068	0.093	0.162	0.138
A1	0.106	–	0.106	0.067	0.493	**0.035**	–	0.721	0.063	–	–	0.137
A2	0.118	0.142 (83.18%)	0.094	**0.038**	0.193	**0.038**	–	0.233	0.049	0.162	0.071	0.150
A3	0.059	0.023 (80.86%)	0.038	**0.018**	1.320	0.022	–	0.035	0.019	0.031	0.060	0.028
A4	0.120	0.037 (81.22%)	0.086	0.032	0.118	0.019	–	0.082	**0.015**	0.090	0.318	0.104
A5	0.101	**0.032** (38.42%)	0.557	0.064	–	0.037	–	0.145	0.069	–	0.122	0.093
A6	0.068	0.032 (75.28%)	0.045	0.021	–	0.030	1.122	0.089	0.037	**0.014**	0.026	0.073
A7	0.071	**0.012** (87.70%)	0.063	0.028	0.291	0.018	–	0.153	0.030	0.041	0.018	0.040
平均值	0.095	0.058	0.137	0.045	0.443	**0.032**	3.642	0.207	0.044	0.211	0.115	0.095

注：加粗字体为每行最优值，“–”表示跟踪失败。

表3　ZJU-Sensetime A序列初始化质量指标$\varepsilon_{\text{init}}$对比

Table 3　Initialization quality $\varepsilon_{\text{init}}$ on the ZJU-Sensetime A sequence

	VINS-Mono	ORB-SLAM3	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro
A0	2.02	5.78	1.03	1.46	1.99	2.19	50.57	**1.03**	1.23	3.80	3.53
A1	**2.69**	8.09	4.03	3.55	7.78	3.50	30.25	4.03	2.88	39.61	–
A2	1.09	5.20	4.54	1.01	7.04	2.34	49.33	–	0.83	**0.45**	3.37
A3	0.97	2.82	0.35	0.71	6.83	**0.45**	–	0.50	1.22	1.06	4.53
A4	2.94	6.39	3.76	2.36	4.64	**0.27**	–	3.76	2.42	1.00	26.75
A5	1.15	4.86	8.59	2.70	–	**1.09**	–	30.59	3.42	–	12.08
A6	1.27	5.58	4.59	2.05	–	2.01	55.41	4.59	2.39	**1.07**	1.10
A7	1.57	4.66	1.61	1.70	–	1.79	–	1.61	**1.12**	–	1.59
平均值	1.71	5.42	3.56	1.94	5.66	**1.70**	46.39	6.59	1.94	7.83	7.56

注：加粗字体为每行最优值，“–”表示跟踪失败。

表 4　ZJU-Sensetime B 序列鲁棒性指标 ε_R 对比

Table 4　Tracking robustness ε_R on the ZJU-Sensetime B sequences

	VINS-Mono	ORB-SLAM3	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro
B0	0.87	–	1.60	0.91	–	0.54	–	0.40	**0.26**	–	1.21
B1	1.05	**0.26**	0.84	0.42	–	0.41	1.174	0.96	0.73	0.32	0.86
B2	4..21	–	3.89	–	–	–	–	–	2.92	**1.99**	–
B3	0.66	16.38	0.90	0.73	–	**0.30**	–	0.34	0.39	1.18	1.44
B4	1.05	–	1.35	1.09	–	0.41	–	**0.28**	0.39	–	–
B5	0.57	0.22	0.93	0.54	–	**0.20**	–	0.30	0.46	**0.20**	0.75
B6	0.72	–	1.12	0.63	5.48	**0.14**	–	0.38	0.30	0.99	1.06
B7	0.75	6.00	1.06	0.78	2.35	**0.24**	–	0.49	**0.24**	0.92	1.08
平均值	1.23	5.71	1.46	0.73	3.91	**0.32**	1.174	0.45	0.71	0.93	1.07

注：加粗字体为每行最优值，“–”表示跟踪失败。

表 5　ZJU-Sensetime C 序列鲁棒性指标 ε_R 对比

Table 5　Tracking robustness ε_R on the ZJU-Sensetime C sequences

	VINS-Mono	ORB-SLAM33	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro
C0	0.33	0.26	7.85	0.55	1.27	–	–	–	**0.21**	1.15	1.17
C1	0.58	–	1.33	0.44	0.98	1.13	8.80	**0.23**	0.32	0.28	1.21
C2	**0.54**	1.70	2.11	0.77	1.87	2.11	–	0.86	0.66	2.53	1.40
C3	0.43	–	1.23	0.42	5.34	0.98	–	0.47	**0.34**	0.92	0.60
C4	0.53	0.89	2.65	0.27	1.28	1.07	–	–	**0.18**	0.34	1.42
C5	0.52	–	0.98	0.48	1.34	1.10	–	0.74	**0.42**	1.02	1.04
C6	0.33	**0.26**	1.18	0.28	1.12	–	7.66	–	0.87	0.27	0.52
C7	1.53	3.22	4.09	1.52	2.90	–	–	**1.14**	1.46	–	3.76
平均值	0.60	1.266	2.68	0.59	2.01	1.27	8.23	0.69	**0.56**	0.93	1.39

注：加粗字体为每行最优值，“–”表示跟踪失败。

两段子序列轨迹间的一致性。在表 2 中观察到的源于手机摄像头的鲁棒性问题，在表 4 中有了量化的体现。由表 3 可以看出，ORB-SLAM3、Kimera 和 MSCKF 存在明显的鲁棒性问题。基于直接法的 DM-VIO 和 SVO-Pro 虽然在卷帘快门的数据上精度不高，但卷帘快门的问题并没有明显影响系统的鲁棒性。精度表现最好的 HybVIO，鲁棒性依然最好。而精度与 HybVIO 接近的 RNIN-VIO 和 RD-VIO，鲁棒性虽然也较为出色，但仍与 HybVIO 有明显差距，也不如精度相对较低的 OpenVINS。这一定程度上反映了这些方法在精度和鲁棒性之间存在着权衡。表 5 对比了在对视觉方法更具挑战的 ZJU-Sensetime-C 数据集下不同系统的鲁棒性。对比结果与表 3 基本一致。主要的区别在于，基于神经惯性网络的 RNIN-VIO 在视觉极端恶劣的条件下（尤其是开关灯、俯瞰地板情况下）具有明显的优势，其鲁棒性在所有系统中最高。图 3 采用 ZJU-

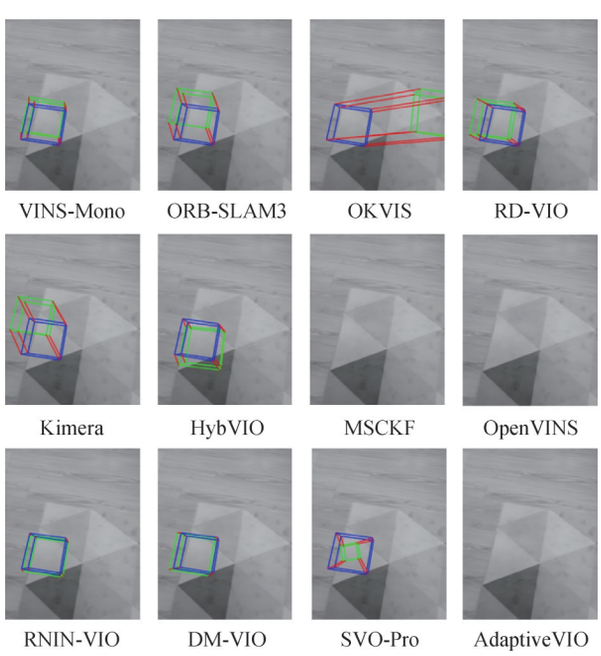

图 3　在 ZJU-Sensetime-C4 序列上不同方法的 AR 效果图

Fig. 3　AR effects of different methods on ZJU-Sensetime-C4 sequence

Sensetime-C4数据对比了不同方法下的AR效果。将一个虚拟立方体放置于地面，用不同方法求解的相机位姿绘制虚拟立方体(绿色)，以真值位姿绘制结果作为参照(蓝色)，对比两者偏差(红线)。其中没有任何框线的图像表示该方法在此时跟踪失败。对比结果与表5一致，表明该数据集能有效反映AR效果。

2.5.3 LSFB数据集结果分析

表6用大尺度场景数据集LSFB中的25个手机数据对比不同方法的精度表现，其中第1列表示该数据集各序列的名称，不同序列对应场景的环境类型描述如表7所示。对比结果与小场景数据集的结果基本一致。小场景中精度最高的RNIN-VIO、Hyb-VIO和RD-VIO，在大场景中精度依然较高。尤其是RNIN-VIO在大场景中的精度优势更为显著。大尺度场景中，很多特征点距离较远、缺少视差，过于依赖视觉约束的方法更容易累积误差。基于神经惯性网络的RNIN-VIO能更充分地利用IMU观测，从而缓解对视觉的依赖。此外，VINS-Mono在大场景中也有着明显的优势。基于滑动窗口优化的VINS-

表6 LSFB数据集上的跟踪精度对比

Table 6 Tracking accuracy on the LSFB dataset

	APE/m										
	VINS-Mono	ORB-SLAM3	OKVIS	RD-VIO	Kimera	HybVIO	MSCKF	OpenVINS	RNIN-VIO	DM-VIO	SVO-Pro
O0	4.687	4.6438 (96.64%)	9.473	3.567	–	–	–	6.484	**3.390**	10.439	4.2595
O1	3.409	1.6727 (93.03%)	3.174	**1.336**	–	2.262	–	6.298	–	–	4.406
O2	–	–	–	–	–	–	–	**5.288**	–	–	–
M0	0.667	**0.394** (74.02%)	0.784	1.112	2.009	0.757	–	1.026	0.835	3.880	0.984
M1	0.509	**0.474** (68.03%)	1.485	1.640	–	1.331	–	1.163	0.752	1.609	1.306
M2	0.794	**0.721** (60.23%)	1.513	1.336	1.683	2.575	–	1.102	0.785	4.532	1.535
L0	**1.160**	5.771 (97.49%)	9.554	4.078	9.167	4.105	–	2.224	2.421	5.482	2.311
L1	3.110	4.207 (48.21%)	8.700	–	–	8.098	–	9.860	8.194	9.560	**2.656**
L2	2.319	7.650 (92.90%)	11.565	4.841	–	3.002	–	3.707	5.095	8.387	**2.306**
L3	**1.581**	6.893 (61.11%)	4.376	3.956	7.781	3.821	–	3.267	1.920	–	3.990
E0	0.987	0.620 (76.38%)	1.713	1.081	–	0.800	–	0.556	**0.244**	5.024	0.962
E1	0.601	1.215 (88.52%)	2.992	1.506	1.919	0.918	–	7.523	**0.467**	3.296	1.296
E2	0.091	2.744 (93.33%)	0.618	0.159	0.253	0.081	–	–	**0.068**	0.883	0.427
A0	0.279	0.319 (55.91%)	2.486	0.247	–	0.570	–	0.393	**0.143**	3.771	0.479
A1	0.858	0.887 (77.38%)	4.291	5.242	2.272	**0.819**	–	1.758	1.308	11.087	3.457
A2	0.675	–	2.922	1.165	–	1.862	–	4.866	**0.598**	9.790	2.001
A3	0.510	**0.378** (62.17%)	1.461	0.442	0.939	1.123	–	0.469	0.453	6.992	0.817
A4	0.986	–	9.654	2.683	–	1.856	–	**0.764**	1.368	–	6.605
C0	–	**2.246** (70.32%)	3.397	–	–	3.014	–	5.819	–	–	4.005
C1	0.517	0.733 (83.40%)	1.549	1.302	2.138	0.740	–	**0.238**	0.337	4.884	1.416
C2	2.552	**0.469** (39.68%)	3.975	2.394	2.897	2.185	–	2.266	2.654	–	4.783
C3	7.513	2.531 (89.41%)	**2.077**	–	8.579	2.202	–	–	–	–	2.175
S0	0.236	0.361 (80.09%)	0.996	0.505	–	0.293	–	**0.130**	0.203	0.970	0.726
S1	**0.269**	0.387 (82.15%)	3.043	0.803	0.841	0.328	–	0.475	0.325	6.475	1.133
S2	0.385	**0.136** (80.39%)	0.635	0.563	0.666	0.422	–	0.227	0.447	5.030	0.404
平均值	**1.509**	2.066	3.851	1.903	3.165	1.877	–	2.865	1.524	5.672	2.268

注：加粗字体为每行最优值，“–”表示跟踪失败。

Mono能尽早地将小视差的特征点加入到滑动窗口进行优化，从而在大场景中更好地控制误差累积。相比之下，基于局部地图的ORB-SLAM3，需要确保特征点已经具有足够视差才加入局部地图，在远景环境中容易出现视觉约束不足的情况，造成误差累积甚至跟踪丢失。而基于滤波的MSCKF和Open-VINS，虽然也能尽早加入小视差特征点，但在出现足够视差前就被过早地线性化，产生误差累积。其他系统在大场景中的表现与在小场景中一致。值得注意的是，基于直接法且没有回路闭合机制的DM-VIO，卷帘快门造成的误差累积在大场景中表现地得为明显，误差明显高于其他系统。

表7　LSFB数据集中的各种环境类型

Table 7　Various types of environments on the LSFB dataset

环境	尺寸/(m × m)	序列	描述
Office park (outdoor)	100 × 100	O0	办公园区，包括大型办公楼和中型广场
Office park (indoor)	100 × 100	O1, O2	办公园区，包括带有小房间和过道的住宅以及宽敞的广场
Medium indoor office	15 × 35	M0 – M2	中型规模的室内办公场所
Large indoor office	135 × 30	L0 – L3	各种具有挑战性场景的大型室内办公场所
Exhibition hall	20 × 30	E0 – E2	具有动态变化内容屏幕的中等规模展厅
Atrium	15 × 60	A0 – A4	15 m高的宽敞艺术馆
Corridor (wide)	4 × 40	C0 – C2	具有良好纹理的宽敞走廊
Corridor (narrow)	2 × 30	C3	具有较弱纹理的狭窄走廊
Stairs	10 × 20	S0 – S2	具有丰富纹理的宽阔楼梯

3　研究热点与发展趋势

虽然VI-SLAM技术发展很快，但由于真实场景的复杂性，在实际应用中不可避免地会遇到一些问题。随着深度学习在诸多领域取得成功，近年来SLAM技术也开始与深度学习相结合，并朝着融合更多的传感器和端云协同的趋势发展。

3.1　深度学习与V-SLAM / VI-SLAM的结合

传统视觉SLAM框架已经比较成熟，也在实际应用中取得了巨大的成功。在传统SLAM的框架下，用基于深度学习的方法替换传统SLAM中各子模块是深度学习方法与SLAM系统结合的一种主要思路。在特征点上，DeTone等人(2018)提出了SuperPoint，即一个完全由网络学习得到的特征点，在各种挑战场景(弱纹理、光照、视角剧烈变化等)下都表现出比传统手工设计的特征点更稳定的效果。进一步地，SuperGlue(Sarlin等，2020)和LoFTR(Sun等，2021)直接实现了两幅图像间端到端特征点提取和匹配。除了特征点匹配之外，光流跟踪也是SLAM系统中常用的算法。传统的光流算法包含了一些人为设计的优化项和参数，难以适用到所有挑战场景下，Teed和Deng(2020)提出RAFT(recurrent all-pairs field transforms)，实现了端到端的稠密光流场估计，取得了比传统方法更好的精度，并表现出很好的泛化性。对于优化模块，BA-Net(Tang和Tan，2019)提出了BALayer，实现了基于深度基的端到端BA优化。DROID-SLAM(Teed和Deng，2021)使用RAFT-Flow作为视觉跟踪模块，并提出了DBA-Layer作为优化模块实现对地图点和相机位姿的端到端更新，实现了真正意义上的端到端SLAM系统。DPVO(Teed等，2023)则使用基于Patch的视觉匹配代替了稠密光流场的视觉匹配，实现了一个更轻量级的端到端VO系统。DROID-SLAM和DPVO生成的地图还只是稀疏地图，使用深度预测网络代替深度传感器来实现稠密SLAM建图则是SLAM与深度学习结合的另一个思路。Tandem(Koestler等，2022)将多帧深度预测网络MVSNet融入DSO。CodeVIO(Zuo等，2021)、CodeMapping(Matsuki等，2021)将深度补全网络预测的深度以编码的表达方式加入到传统SLAM的优化中，实现了对深度的持续优化。Xie等人(2023)则采用多深度基的深度表达方式将深度补全网络的输出融入到传统SLAM的优化器中。得益于深度基的表达形式，该方法能够在移动端实现在

线建图。

基于隐式表达的地图的出现，为深度学习与SLAM融合提供了新的思路。利用隐式表达地图的可微分特性，无需进行特征提取和匹配，可以通过直接渲染图像的像素误差最小化来实现端到端的训练，代表工作有iMap（Sucar等，2021）、NICE-SLAM（Zhu等，2022）、Vox-Fusion（Yang等，2022）、ESLAM（Johari等，2023）、Co-SLAM（Wang等，2023）、NICER-SLAM（Zhu等，2023）等。其中iMap、NICE-SLAM、Vox-Fusion、ESLAM、Co-SLAM都依赖于RGB-D输入。NICER-SLAM则使用其他网络来引入深度、法向以及光流预测信息，实现了基于单目图像的端到端SLAM系统。

3D高斯渲染（3D Gaussian splatting，3DGS）技术相比于隐式地图通常具备更快的渲染速度和更高的渲染质量，所以最近越来越多的工作将3DGS和SLAM融合。通过直接最小化渲染图像的光度误差，实现相机位姿估计和稀疏建图。代表工作有SplaTAM（splat，track and map）（Keetha等，2024）、GSSLAM（Gaussian splatting SLAM）（Matsuki等，2024）、Compact-GSSLAM（compact Gaussian splatting SLAM）（Deng等，2024）、HF-GS SLAM（high-fidelity SLAM using Gaussian splatting）（Sun等，2024a）、CG-SLAM（Hu等，2024）和MM3DGS-SLAM（multi-modal 3D Gaussian splatting for SLAM）（Sun等，2024b）等。其中，SplaTAM、Compact-GSSLAM、HF-GS SLAM和CG-SLAM的输入数据都是RGB-D序列，GSSLAM的输入是RGB序列，而MM3DGS-SLAM的输入数据是RGB-D序列和IMU数据。

基于学习的IMU神经网络对提升IMU惯性导航精度有明显优势。第一个基于学习的惯性导航系统是IoNet（Chen等，2018），用网络学习2D位移和方向，可以显著抑制IMU积分误差，取得比传统IMU导航更高的精度。RoNIN（Herath等，2020）用一个相似的网络预测2D速度。TLIO（Liu等，2020b）提出基于IMU神经网络和扩展卡尔曼滤波器（extended kalman filter，EKF）的相结合，实现纯IMU的3D导航。RNIN-VIO（Chen等，2021）提出结合IMU神经网络和VIO系统，提升原有VIO的鲁棒性和精度。Buchanan等人（2023）提出学习IMU的偏置，可以实现多运动模式泛化，并与VIO结合，提升VIO精度。Zhang等人（2021）提出学习IMU噪声，并将网络输出降噪的IMU数据提供给VIO，提升VIO精度。

总体而言，深度学习与视觉SLAM的融合还处于百花齐放的状态。很多现有端到端的SLAM系统还存在计算资源占用高、效率低的问题。如何将深度学习的方法高效地与SLAM系统结合是该方案能否落地的关键。

3.2 多传感器融合

基于传统相机的视觉方法，在弱纹理、重复纹理、光照变化和运动模糊等情况下，难以避免地存在鲁棒性挑战。通过融合IMU虽然可以在一定程度上提升鲁棒性，但如果视觉信息不可靠，只依靠IMU仍难以长时间跟踪，也无法建图。将SLAM方法向新型的视觉传感器延伸，是目前的一个研究热点与发展趋势。如事件相机就是一种异步、低延时的新型传感器，通过只记录亮度变化的像素，从而在高速运动和高动态范围的场景下提供稳定的视觉信息。基于事件相机的SLAM，代表性工作包括基于双目事件相机的VO（Zhou等，2021）、基于单目事件相机和光度三维地图的直接法SLAM（Bryner等，2019）、结合单目事件相机和IMU的VIO（Zhu等，2017）等。另外，偏振相机也是一种新型传感器，利用光的偏振特性，在不同偏振角度下捕捉物体反射光，从而获得物体的形状、纹理和材质等。代表性工作包括基于多视图几何的三维重建（Cui等，2017）和SLAM（Yang，2018），以及基于深度学习的偏振三维重建（Ba等，2020）等。

除视觉和惯性传感器外，如何通过融合更多类型的传感器，进一步突破视觉方法的局限，也是一个发展趋势。Qin等人（2019a，b）提出一种通用的融合定位框架，通过将相机、IMU、LiDAR等局部传感器与GPS（global positioning system）、磁力计、气压计等全局传感器融合，实现局部精准且全局无漂移的定位跟踪。GPS信号只在室外可用，室内则可以利用WiFi（Youssef和Agrawala，2005）、蓝牙（Faragher和Harle，2015）或UWB（ultra wide band）（Zhao等，2021）等无线信号提供全局定位信息。这些基于无线信号的定位方法都需要在环境中部署信号发射器，成本较高。磁力计通过利用建筑结构对地磁场的干扰也能在室内环境提供全局定位信息，且无需部署特殊设备，目前的技术水平已能实现AR导航等简单应用，但仍难以满足对定位精度要求更高的应用需求（Liu等，2023）。Liu等人（2023）首次提出

包含图像、IMU、WiFi、蓝牙和磁力计的大场景定位数据集，以推动多传感器融合的定位跟踪进一步发展。除此之外，从多传感器融合、多特征基元融合和多维度信息融合的角度出发，也为解决多源融合SLAM提供了新的思路和技术支持（王金科 等，2022）。这些融合策略不仅有助于提高SLAM系统的鲁棒性和精度，还可以增强其在复杂环境中的适应能力，从而更有效地处理不同传感器数据之间的异质性和不确定性。

3.3 端云协同

长时间运行的SLAM系统不可避免地会出现误差累积，一般需要通过回环检测来消除误差累积，而在实际应用场景中，不能保证轨迹上存在回环，即使可以通过回路闭合来消除误差，在检测到回环之前SLAM输出的实时位姿的误差已经存在，在回路闭合后必然会出现前端运动轨迹的跳变，从而造成不好的用户体验。移动终端上由于算力资源的限制，误差累积的问题会更加严重。对于一个特定场景，可以通过端云协同的方式来解决误差累积问题：预先对场景扫描拍摄构建三维地图，然后将预构建的三维地图部署在云端，利用预构建地图的信息来抑制移动端上SLAM的误差累积。端云结合通常包含3个模块：离线建图、云端定位以及云端定位结果与端上SLAM融合。离线建图可以采用SfM技术实现，例如开源软件COLMAP（Schonberger和Frahm，2016）、EC-SfM（Ye等，2024）等。近几年随着神经隐式表达的兴起，使用神经隐式表达来建图也逐渐成为一种重要的工作方向，如NeRF-SLAM（Rosinol等，2022）、Point-SLAM（Sandström等，2023）。基于高精度地图的定位需要解决大尺度场景、视角变化以及地图持久化等问题。提升特征匹配能力是一种方式，如使用SuperPoint这样的深度学习特征点。Yu等人（2020）提出了一种二分图匹配深度神经网络BGNet，通过K近邻搜索和网络预测的几何先验信息来消除在重复纹理和光照变化条件下导致局部特征区分度降低，从而难以匹配的问题，并通过匈牙利池化层来有效剔除错误匹配，可以显著提升场景和视角变化情况下的定位成功率和精度。Yan等人（2023）则使用了更多传感器信息来提升地图持久化的能力。在将预构建地图融入端上SLAM时，根据是否将地图结构信息或全局定位信息作为额外观测加入到SLAM中联合优化，可以将相关的工作分为松耦合（Platinsky等，2020；Qin等，2018；Yamaguchi等，2020）和紧耦合（Ye等，2020；Zuo等，2019；Bao等，2022；Katragadda等，2023）。松耦合这种直接基于定位位姿的融合通常很难排除定位误差的干扰，从而导致端上实时恢复的位姿出现频繁的跳动。Ye等人（2020）提出与预构建地图的Surfel结构进行耦合，改进实时位姿估计的精度，但这种方法依赖于高频的定位以及较高精度的地图结构。Bao等人（2022）使用Ray-Cast的方式从高精度地图提取结构点云与端上VIO融合，能够在较低频次和高时延的情况下有效抑制VIO的误差累积。NeRF-VINS（Katragadda等，2024）提出将VINS系统与预先构建的NeRF地图进行融合。

端云协同的SLAM目前还没有标准的范式。云端地图表达形式和定位方法以及与端上SLAM的融合策略是紧密相关的，而且在实际应用部署还需结合考虑运营成本、网络通信状况以及端侧运行速度等因素，因此还有较大的提升空间。

4 结 语

随着AR/VR和机器人等领域的蓬勃发展，作为其中关键技术的单目视觉惯性SLAM进展很快，涌现出一批出色的方法和系统。然而在实际复杂场景中，各类方法具有不同程度的局限性。本文的评测结果表明：基于优化的方法一般在跟踪精度和鲁棒性上比基于滤波的方法更有优势，但相对更耗时；直接法/半直接法在采用全局快门拍摄的情况下表现出了出色的性能，但容易受卷帘快门和光照变化的影响，尤其是大场景下误差累积较快；结合深度学习可以提高在复杂情况下的鲁棒性，如结合惯性神经网络能够显著提升在极端情况下的鲁棒性。

为了解决实际场景中遇到的挑战，近年来SLAM技术也朝着与深度学习结合、融合更多传感器以及端云协同的趋势发展。特别是基于深度学习的方法表现出很大的潜力，虽然在整个SLAM系统的层面尚未能完全取代传统方法，但通过与传统方法以合适的方式结合，可以突破传统方法的一些局限性，同时又能满足实际落地的轻量级计算要求。此外，基于视觉的SLAM难以克服对视觉信息的依赖，可以通过合适的方式（如通过深度学习和传统优化相结合）融合IMU、深度传感器、WiFi、蓝牙和地磁

等信号,能够显著增强系统的鲁棒性,从而能满足更复杂的实际落地应用要求。另外,通过预先构建环境地图以及端云协同的方式可以将移动终端和云端的各自优势充分结合,突破移动端算力瓶颈的限制,从而实现在大尺度场景下长时间的精准跟踪定位。

致谢:感谢陈信宇和占若豪同学协助测试多个SLAM系统在3个数据集上的性能;感谢刘心阳和沈奕辰同学协忙拍摄数据;感谢潘友琦同学在Adaptive VIO上完成的精度测试。

参考文献(References)

Almalioglu Y, Turan M, Saputra M R U, De Gusmão P P B, Markham A and Trigoni N. 2022. SelfVIO: self-supervised deep monocular visual-inertial odometry and depth estimation. Neural Networks, 150: 119-136 [DOI: 10.1016/j.neunet.2022.03.005]

Ba Y H, Gilbert A, Wang F, Yang J F, Chen R, Wang Y Q, Yan L, Shi B X and Kadambi A. 2020. Deep shape from polarization//Proceedings of 16th European Conference on Computer Vision-ECCV 2020. Glasgow, UK: Springer: 554-571 [DOI: 10.1007/978-3-030-58586-0_33]

Bao H J, Xie W J, Qian Q H, Chen D P, Zhai S J, Wang N and Zhang G F. 2022. Robust tightly-coupled visual-inertial odometry with pre-built maps in high latency situations. IEEE Transactions on Visualization and Computer Graphics, 28(5): 2212-2222 [DOI: 10.1109/TVCG.2022.3150495]

Bryner S, Gallego G, Rebecq H and Scaramuzza D. 2019. Event-based, direct camera tracking from a photometric 3D map using nonlinear optimization//Proceedings of 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE: 325-331 [DOI: 10.1109/ICRA.2019.8794255]

Buchanan R, Agrawal V, Camurri M, Dellaert F and Fallon M. 2023. Deep IMU bias inference for robust visual-inertial odometry with factor graphs. IEEE Robotics and Automation Letters, 8(1): 41-48 [DOI: 10.1109/LRA.2022.3222956]

Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, Achtelik M W and Siegwart R. 2016. The EuRoC micro aerial vehicle datasets. The International Journal of Robotics Research, 35(10): 1157-1163 [DOI: 10.1177/0278364915620033]

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I and Leonard J J. 2016. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Transactions on Robotics, 32(6): 1309-1332 [DOI: 10.1109/TRO.2016.2624754]

Campos C, Elvira R, Rodríguez J J G, Montiel J M M and Tardós J D. 2021. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Transactions on Robotics, 37(6): 1874-1890 [DOI: 10.1109/TRO.2021.3075644]

Chen C H, Lu X X, Markham A and Trigoni N. 2018. IONet: learning to cure the curse of drift in inertial odometry//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 6468-6476 [DOI: 10.1609/aaai.v32i1.12102]

Chen D P, Wang N, Xu R S, Xie W J, Bao H J and Zhang G F. 2021b. RNIN-VIO: robust neural inertial navigation aided visual-inertial odometry in challenging scenes//2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Bari, Italy: IEEE: 275-283 [DOI: 10.1109/ISMAR52148.2021.00043]

Clark R, Wang S, Wen H K, Markham A and Trigoni N. 2017. VINet: visual-inertial odometry as a sequence-to-sequence learning problem//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: 3995-4001 [DOI: 10.1609/aaai.v31i1.11215]

Cortés S, Solin A, Rahtu E and Kannala J. 2018. ADVIO: an authentic dataset for visual-inertial odometry//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 425-440 [DOI: 10.1007/978-3-030-01249-6_26]

Cui Z P, Gu J W, Shi B X, Tan P and Kautz J. 2017. Polarimetric multi-view stereo//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 369-378 [DOI: 10.1109/CVPR.2017.47]

Delmerico J and Scaramuzza D. 2018. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 2502-2509 [DOI: 10.1109/ICRA.2018.8460664]

Deng T C, Chen Y H, Zhang L Y, Yang J F, Yuan S H, Wang D W and Chen W D. 2024. Compact 3D Gaussian splatting for dense visual SLAM [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2403.11247.pdf

DeTone D, Malisiewicz T and Rabinovich A. 2018. SuperPoint: self-supervised interest point detection and description//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE: 337-349 [DOI: 10.1109/CVPRW.2018.00060]

Engel J, Koltun V and Cremers D. 2018. Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(3): 611-625 [DOI: 10.1109/TPAMI.2017.2658577]

Faragher R and Harle R. 2015. Location fingerprinting with bluetooth low energy beacons. IEEE Journal on Selected Areas in Communications, 33(11): 2418-2428 [DOI: 10.1109/JSAC.2015.2430281]

Forster C, Pizzoli M and Scaramuzza D. 2014. SVO: fast semi-direct monocular visual odometry//Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE: 15-22 [DOI: 10.1109/ICRA.2014.6906584]

Forster C, Zhang Z C, Gassner M, Werlberger M and Scaramuzza D. 2017a. SVO: semidirect visual odometry for monocular and multi-

camera systems. IEEE Transactions on Robotics, 33(2): 249-265 [DOI: 10.1109/TRO.2016.2623335]

Forster C, Carlone L, Dellaert F and Scaramuzza D. 2017b. On-manifold preintegration for real-time visual: inertial odometry. IEEE Transactions on Robotics, 33(1): 1-21 [DOI: 10.1109/TRO.2016.2597321]

Fu T M, Su S S, Lu Y R and Wang C. 2024. iSLAM: imperative SLAM [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2306.07894.pdf

Gálvez-López D and Tardos J D. 2012. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics, 28(5): 1188-1197 [DOI: 10.1109/TRO.2012.2197158]

Geneva P, Eckenhoff K, Lee W, Yang Y L and Huang G Q. 2020. OpenVINS: a research platform for visual-inertial estimation//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 4666-4672 [DOI: 10.1109/ICRA40945.2020.9196524]

Han L M, Lin Y M, Du G G and Lian S G. 2019. DeepVIO: self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China: IEEE: 6906-6913 [DOI: 10.1109/IROS40897.2019.8968467]

Harris C and Stephens M. 1988. A combined corner and edge detector//Proceedings of Alvey Vision Conference 1988. Manchester, UK: [s.n.]: 147-152

Herath S, Yan H and Furukawa Y. 2020. RoNIN: robust neural inertial navigation in the wild: benchmark, evaluations, and new methods//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE: 3146-3152 [DOI: 10.1109/ICRA40945.2020.9196860]

Hu J R, Chen X H, Feng B Y, Li G L, Yang L J, Bao H J, Zhang G F and Cui Z P. 2024. CG-SLAM: efficient dense RGB-D SLAM in a consistent uncertainty-aware 3D Gaussian field [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2403.16095.pdf

Johari M M, Carta C and Fleuret F. 2023. ESLAM: efficient dense SLAM system based on hybrid representation of signed distance fields//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 17408-17419 [DOI: 10.1109/CVPR52729.2023.01670]

Katragadda S, Lee W, Peng Y X, Geneva P, Chen C C, Guo C, Li M Y and Huang G Q. 2023. NeRF-VINS: a real-time neural radiance field map-based visual-inertial navigation system [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2309.09295.pdf

Keetha N, Karhade J, Jatavallabhula K M, Yang G S, Scherer S, Ramanan D and Luiten J. 2024. SplaTAM: splat, track and map 3D Gaussians for dense RGB-D SLAM [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2312.02126.pdf

Koestler L, Yang N, Zeller N and Cremers D. 2022. TANDEM: tracking and dense mapping in real-time using deep multi-view stereo//Proceedings of the 5th Conference on Robot Learning. London, UK: [s.n.]: 34-45

Leutenegger S, Chli M and Siegwart R Y. 2011. BRISK: binary robust invariant scalable keypoints//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 2548-2555 [DOI: 10.1109/ICCV.2011.6126542]

Leutenegger S, Furgale P, Rabaud V, Chli M, Konolige K and Siegwart R. 2013. Keyframe-based visual-inertial SLAM using nonlinear optimization//Proceedings of the International Conference on Robotics: Science and Systems IX (RSS 2013. Berlin, Germany: [s.n.])

Li J Y, Pan X K, Huang G, Zhang Z Y, Wang N, Bao H J and Zhang G F. 2024. RD-VIO: robust visual-inertial odometry for mobile augmented reality in dynamic environments. IEEE Transactions on Visualization and Computer Graphics: 1-14 [DOI: 10.1109/TVCG.2024.3353263]

Li J Y, Yang B B, Chen D P, Wamg N, Zhang G F and Bao H J. 2019. Survey and evaluation of monocular visual-inertial SLAM algorithms for augmented reality. Virtual Reality and Intelligent Hardware, 1(4): 386-410 [DOI: 10.1016/j.vrih.2019.07.002]

Liu H M, Jiang M X, Zhang Z, Huang X P, Zhao L S, Hang M, Feng Y J, Bao H J and Zhang G F. 2020a. LSFB: a low-cost and scalable framework for building large-scale localization benchmark//2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). Recife, Brazil: IEEE: 219-224 [DOI: 10.1109/ISMAR-Adjunct51615.2020.00065]

Liu H M, Xue H, Zhao L S, Chen D P, Peng Z and Zhang G F. 2023. MagLoc-AR: magnetic-based localization for visual-free augmented reality in large-scale indoor environments. IEEE Transactions on Visualization and Computer Graphics, 29(11): 4383-4393 [DOI: 10.1109/TVCG.2023.3321088]

Liu H M, Zhang G F and Bao H J. 2016. A survey of monocular simultaneous localization and mapping. Journal of Computer-Aided Design and Computer Graphics, 28(6): 855-868 (刘浩敏, 章国锋, 鲍虎军. 2016. 基于单目视觉的同时定位与地图构建方法综述. 计算机辅助设计与图形学学报, 28(6): 855-868) [DOI: 10.3969/j.issn.1003-9775.2016.06.001]

Liu H M, Zhao L S, Peng Z, Xie W J, Jiang M X, Zha H, Bao H J and Zhang G F. 2024. A low-cost and scalable framework to build large-scale localization benchmark for augmented reality. IEEE Transactions on Circuits and Systems for Video Technology, 34(4): 2274-2288 [DOI: 10.1109/TCSVT.2023.3306160]

Liu W X, Caruso D, Ilg E, Dong J, Mourikis A I, Daniilidis K, Kumar V and Engel J. 2020b. TLIO: tight learned inertial odometry. IEEE Robotics and Automation Letters, 5(4): 5653-5660 [DOI: 10.1109/LRA.2020.3007421]

Matsuki H, Murai R, Kelly P H J and Davison A J. 2024. Gaussian splatting SLAM [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2312.06741

Matsuki H, Scona R, Czarnowski J and Davison A J. 2021. CodeMap-

ping: real-time dense mapping for sparse SLAM using compact scene representations. IEEE Robotics and Automation Letters, 6(4): 7105-7112 [DOI: 10.1109/LRA.2021.3097258]

Mourikis A I and Roumeliotis S I. 2007. A multi-state constraint Kalman filter for vision-aided inertial navigation//Proceedings of 2007 IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE: 3565-3572 [DOI: 10.1109/ROBOT.2007.364024]

Pan X K, Liu H M, Fang M, Wang Z, Zhang Y and Zhang G F. 2023. Dynamic 3D scenario-oriented monocular SLAM based on semantic probability prediction. Journal of Image and Graphics, 28(7): 2151-2166 (潘小鹍, 刘浩敏, 方铭, 王政, 张涌, 章国锋. 2023. 基于语义概率预测的动态场景单目视觉SLAM. 中国图象图形学报, 28(7): 2151-2166) [DOI: 10.11834/jig.210632]

Pan Y Q, Zhou W G, Cao Y D and Zha H B. 2024. Adaptive VIO: deep visual-inertial odometry with online continual learning [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2405.16754.pdf

Pfrommer B, Sanket N, Daniilidis K and Cleveland J. 2017. PennCOSYVIO: a challenging visual inertial odometry benchmark//Proceedings of 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore, Singapore: IEEE: 3847-3854 [DOI: 10.1109/ICRA.2017.7989443]

Platinsky L, Szabados M, Hlasek F, Hemsley R, Del Pero L, Pancik A, Baum B, Grimmett H and Ondruska P. 2020. Collaborative augmented reality on smartphones via life-long city-scale maps//Proceedings of 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Porto de Galinhas, Brazil: IEEE: 533-541 [DOI: 10.1109/ISMAR50242.2020.00081]

Qin T, Cao S Z, Pan J and Shen S J. 2019a. A general optimization-based framework for global pose estimation with multiple sensors [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/1901.03642.pdf

Qin T, Li P L and Shen S J. 2018. VINS-Mono: a robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 34(4): 1004-1020 [DOI: 10.1109/TRO.2018.2853729]

Qin T, Pan J, Cao S Z and Shen S J. 2019b. A general optimization-based framework for local odometry estimation with multiple sensors [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/1901.03638.pdf

Rosinol A, Abate M, Chang Y and Carlone L. 2020. Kimera: an open-source library for real-time metric-semantic localization and mapping//Proceedings of 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE: 1689-1696 [DOI: 10.1109/ICRA40945.2020.9196885]

Rosinol A, Leonard J J and Carlone L. 2022. NeRF-SLAM: real-time dense monocular SLAM with neural radiance fields [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2210.13641.pdf

Sandström E, Li Y, Van Gool L and Oswald M R. 2023. Point-SLAM: dense neural point cloud-based SLAM//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 18387-18398 [DOI: 10.1109/ICCV51070.2023.01690]

Sarlin P E, DeTone D, Malisiewicz T and Rabinovich A. 2020. SuperGlue: learning feature matching with graph neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4937-4946 [DOI: 10.1109/CVPR42600.2020.00499]

Sarlin P E, Dusmanu M, Schönberger J L, Speciale P, Gruber L, Larsson V, Miksik O and Pollefeys M. 2022. LaMAR: benchmarking localization and mapping for augmented reality//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 686-704 [DOI: 10.1007/978-3-031-20071-7_40]

Schonberger J L and Frahm J M. 2016. Structure-from-motion revisited//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4104-4113 [DOI: 10.1109/CVPR.2016.445]

Seiskari O, Rantalankila P, Kannala J, Ylilammi J, Rahtu E and Solin A. 2022. HybVIO: pushing the limits of real-time visual-inertial odometry//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 287-296 [DOI: 10.1109/WACV51458.2022.00036]

Servières M, Renaudin V, Dupuis A and Antigny N. 2021. Visual and visual-inertial SLAM: state of the art, classification, and experimental benchmarking. Journal of Sensors, 2021: #2054828 [DOI: 10.1155/2021/2054828]

Sucar E, Liu S K, Ortiz J and Daviso A J. 2021. iMAP: implicit mapping and positioning in real-time//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6209-6218 [DOI: 10.1109/ICCV48922.2021.00617]

Sun J M, Shen Z H, Wang Y A, Bao H J and Zhou X W. 2021. LoFTR: detector-free local feature matching with Transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8922-8931 [DOI: 10.1109/CVPR46437.2021.00881]

Sun L C, Bhatt N P, Liu J C, Fan Z W, Wang Z Y, Humphreys T E and Topcu U. 2024b. MM3DGS SLAM: multi-modal 3D Gaussian splatting for SLAM using vision, depth, and inertial measurements [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2404.00923.pdf

Sun S, Mielle M, Liienthal A J and Magnusson M. 2024a. High-fidelity SLAM using Gaussian splatting with rendering-guided densification and regularized optimization [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/2403.12535.pdf

Tan W, Liu H M, Dong Z L, Zhang G F and Bao H J. 2013. Robust monocular SLAM in dynamic environments//2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Adelaide, Australia: IEEE: 209-218 [DOI: 10.1109/ISMAR.2013.6671781]

Tang C Z and Tan P. 2019. BA-Net: dense bundle adjustment network [EB/OL]. [2023-12-08]. https://arxiv.org/pdf/1806.04807.pdf

Teed Z and Deng J. 2020. RAFT: recurrent all-pairs field transforms for optical flow//Proceedings of the 16th European Conference on Computer Vision—ECCV 2020. Glasgow, UK: Springer: 402-419

［DOI：10.1007/978-3-030-58536-5_24］

Teed Z and Deng J. 2021. DROID-SLAM：deep visual SLAM for monocular，stereo，and RGB-D cameras//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual：Curran Associates Inc.：16558-16569

Teed Z，Lipson L and Deng J. 2023. Deep patch visual odometry［EB/OL］.［2023-12-08］. https://arxiv.org/pdf/2208.04726.pdf

Von Stumberg L and Cremers D. 2022. DM-VIO：delayed marginalization visual-inertial odometry. IEEE Robotics and Automation Letters，7(2)：1408-1415［DOI：10.1109/LRA.2021.3140129］

Von Stumberg L，Usenko V and Cremers D. 2018. Direct sparse visual-inertial odometry using dynamic marginalization//Proceedings of 2018 IEEE International Conference on Robotics and Automation（ICRA）. Brisbane，Australia：IEEE：2510-2517［DOI：10.1109/ICRA.2018.8462905］

Wang H Y，Wang J W and Agapito L. 2023. Co-SLAM：joint coordinate and sparse parametric encodings for neural real-time SLAM//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：13293-13302［DOI：10.1109/CVPR52729.2023.01277］

Wang J K，Zuo X X，Zhao X R，Lyu J J and Liu Y. 2022. Review of multi-source fusion SLAM：current status and challenges. Journal of Image and Graphics，27(2)：368-389（王金科，左星星，赵祥瑞，吕佳俊，刘勇. 2022. 多源融合SLAM的现状与挑战. 中国图象图形学报，27(2)：368-389）［DOI：10.11834/jig.210547］

Wenzel P，Yang N，Wang R，Zeller N and Cremers D. 2022. 4Seasons：benchmarking visual SLAM and long-term localization for autonomous driving in challenging conditions［EB/OL］.［2023-12-08］. https://arxiv.org/pdf/2301.01147.pdf

Wu K J，Ahmed A M，Georgiou G A and Roumeliotis S. 2015. A square root inverse filter for efficient vision-aided inertial navigation on mobile devices//Robotics：Science and Systems. Rome，Italy：MIT Press

Xie W J，Chu G Y，Qian Q H，Yu Y H，Li H，Chen D P，Zhai S J，Wang N，Bao H J and Zhang G F. 2023. Depth completion with multiple balanced bases and confidence for dense monocular SLAM［EB/OL］.［2023-12-08］. https://arxiv.org/pdf/2309.04145

Yamaguchi M，Mori S，Saito H，Yachida S and Shibata T. 2020. Global-map-registered local visual odometry using on-the-fly pose graph updates//Proceedings of the 7th International Conference on Augmented Reality，Virtual Reality，and Computer Graphics. Lecce，Italy：Springer：299-311［DOI：10.1007/978-3-030-58465-8_23］

Yan C，Qu D，Xu D，Zhao B，Wang Z，Wang D and Li X. 2024. GS-SLAM：dense visual SLAM with 3D Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：19595-19604［DOI：10.48550/arXiv.2311.11700］

Yan S，Liu Y，Wang L，Shen Z H，Peng Z，Liu H M，Zhang M J，Zhang G F and Zhou X W. 2023. Long-term visual localization with mobile sensors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：17245-17255［DOI：10.1109/CVPR52729.2023.01654］

Yang L W，Tan F T，Li A，Cui Z P，Furukawa Y and Tan P. 2018. Polarimetric dense monocular SLAM//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：3857-3866［DOI：10.1109/CVPR.2018.00406］

Yang X R，Li H，Zhai H J，Ming Y H，Liu Y Q and Zhang G F. 2022. Vox-Fusion：dense tracking and mapping with voxel-based neural implicit representation//Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality（ISMAR）. Singapore，Singapore：IEEE：499-507［DOI：10.1109/ISMAR55827.2022.00066］

Ye H Y，Huang H Y and Liu M. 2020. Monocular direct sparse localization in a prior 3D surfel map//Proceedings of 2020 IEEE International Conference on Robotics and Automation（ICRA）. Paris，France：IEEE：8892-8898［DOI：10.1109/ICRA40945.2020.9197022］

Ye Z C，Bao C，Zhou X，Liu H M，Bao H J and Zhang G F. 2024. EC-SfM：efficient covisibility-based structure-from-motion for both sequential and unordered images. IEEE Transactions on Circuits and Systems for Video Technology，34(1)：110-123［DOI：10.1109/TCSVT.2023.3285479］

Youssef M and Agrawala A. 2005. The Horus WLAN location determination system//Proceedings of the 3rd International Conference on Mobile Systems，Applications，and Services. Seattle，USA：ACM：205-218［DOI：10.1145/1067170.1067193］

Yu H L，Ye W C，Feng Y J，Bao H J and Zhang G F. 2020. Learning bipartite graph matching for robust visual localization//2020 IEEE International Symposium on Mixed and Augmented Reality（ISMAR）. Porto de Galinhas，Brazil：IEEE：146-155［DOI：10.1109/ISMAR50242.2020.00036］

Zuo X，Merrill N，Li W，Liu Y，Pollefeys M and Huang G. 2021. CodeVIO：visual-inertial odometry with learned optimizable dense depth//Proceedings of 2021 IEEE International Conference on Robotics and Automation（ICRA）. Xi'an，China：IEEE：14382-14388［DOI：10.1109/ICRA48506.2021.9560792］

Zeng Q H，Luo Y X，Sun K C，Li Y N and Liu J Y. 2022. Review on SLAM technology development for vision and its fusion of inertial information. Journal of Nanjing University of Aeronautics and Astronautics，54(6)：1007-1020（曾庆化，罗怡雪，孙克诚，李一能，刘建业. 2022. 视觉及其融合惯性的SLAM技术发展综述. 南京航空航天大学学报，54(6)：1007-1020）［DOI：10.16356/j.1005-2615.2022.06.002］

Zhang M，Zhang M M，Chen Y M and Li M Y. 2021. IMU data processing for inertial aided navigation：a recurrent neural network based approach//Proceedings of 2021 IEEE International Conference on Robotics and Automation（ICRA）. Xi'an，China：IEEE：3992-

3998［DOI：10.1109/ICRA48506.2021.9561172］

Zhao M H, Chang T, Arun A, Ayyalasomayajula R, Zhang C and Bharadia D. 2021. ULoc：low-power, scalable and cm-accurate UWB-tag localization and tracking for indoor applications. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(3)：#140［DOI：10.1145/3478124］

Zhou Y, Gallego G and Shen S J. 2021. Event-based stereo visual odometry. IEEE Transactions on Robotics, 37(5)：1433-1450［DOI：10.1109/TRO.2021.3062252］

Zhu A Z, Atanasov N and Daniilidis K. 2017. Event-based visual inertial odometry//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA：IEEE：5816-5824［DOI：10.1109/CVPR.2017.616］

Zhu Z, Peng S, Larsson V, Cui Z, Oswald M R, Geiger A and Pollefeys M. 2023. NICER-SLAM：neural implicit scene encoding for RGB SLAM［EB/OL］.［2023-12-08］. https://arxiv.org/pdf/2302.03594

Zhu Z H, Peng S Y, Larsson V, Xu W W, Bao H J, Cui Z P, Oswald M R and Pollefeys M. 2022. NICE-SLAM：neural implicit scalable encoding for SLAM//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA：IEEE：12776-12786［DOI：10.1109/CVPR52688.2022.01245］

作者简介

章国锋，男，教授，主要研究方向为SLAM、三维视觉和混合现实。E-mail：zhangguofeng@zju.edu.cn

鲍虎军，通信作者，男，教授，主要研究方向为计算机图形学、三维视觉和混合现实。E-mail：baohujun@zju.edu.cn

黄赣，男，博士研究生，主要研究方向为SLAM和混合现实。E-mail：huanggan@zju.edu.cn

谢卫健，男，博士研究生，主要研究方向为SLAM和混合现实。E-mail：11821126@zju.edu.cn

陈丹鹏，男，博士研究生，主要研究方向为SLAM和混合现实。E-mail：11921155@zju.edu.cn

王楠，男，硕士研究生，主要研究方向为SLAM和混合现实。E-mail：wangnan@sensetime.com

刘浩敏，男，博士研究生，主要研究方向为SLAM、运动恢复结构和混合现实。E-mail：liuhaomin@sensetime.com