首发于公众号：一点人工一点智能

两万字 | 视觉 SLAM 研究综述与未来趋势讨论

原文：Visual SLAM: What are the Current Trends and What to Expect?

地址：https://arxiv.org/abs/2210.10491

翻译：董亚微

摘要：近年来，基于视觉传感器在同时定位与地图构建（SLAM）系统中展示出了显著的性能、准确性和效率。在这里，视觉同时定位与地图构建（VSLAM）方法是指使用相机进行姿态估计和地图生成的 SLAM 方法。

我们可以看到许多研究表明，尽管视觉 SLAM 的成本较低，但是 VSLAM 是可以优于传统的仅仅依赖特定传感器方法的。VSLAM 方法利用不同的相机类型（例如，单目、立体和 RGB-D），在各种数据集（例如，KITTI、TUM RGB-D 和 EuRoC）和不同的环境（例如，室内和室外）中进行了测试，并采用多种算法和方法来更好地了解环境。

上述变化使这一研究主题受到科研人员的广泛关注，并产生了很多的 VSLAM 方法。在此基础上，本文的主要目的是介绍 VSLAM 系统的最新进展，并讨论现有的挑战和趋势。我们对在 VSLAM 领域发表的 45 篇有影响力的论文进行了深入的文献调研，根据不同的特点对这些论文进行了分类，包括方法创新性、领域应用新颖性、算法优化和语义层面，还讨论了目前的趋势和未来的方向，这可能有助于研究人员进行研究。

01 介绍

同时定位与地图构建（SLAM）是指在定位智能体位置的同时构建未知环境地图的过程[1]。在此处，智能体可以是家用机器人[2]、自动驾驶车辆[3]、行星漫步车[4]，甚至是无人机（UAV）[5]、[6]或无人车（UGV）[7]。在地图不可用或机器人位置未知的环境中，SLAM 有着非常广泛的应用。近些年，随着机器人技术应用的不断提升，SLAM 在产业圈和科研圈中获得了极大的关注[8]，[9]。

SLAM 系统可以使用各种传感器从环境中收集数据，这些传感器有基于激光的、声学的和视觉的[10]。基于视觉的传感器又有多种，包括单目（monocular）、立体（stereo）、基于事件（event-based）、广角（omnidirectional）和 RGB 深度（RGB-D）相机。带有视觉传感器的机器人，便是使用相机提供的视觉数据来估计机器人相对于其周围环境的位置和方向[11]。使用视觉传感器进行 SLAM 的过程即为视觉 SLAM（VSLAM）。

在 SLAM 中使用视觉数据具有以下优点：硬件更便宜，目标检测和跟踪更直观，并且能够提供丰富的视觉和语义信息[12]。其捕获的图像（或视频帧）还可以用于基于视觉的应用，包括语义分割和目标检测。上述特点使得 VSLAM 成为机器人学的热门方向，并促使机器人学和计算机视觉（CV）专家在过去几十年中进行了大量研究和调研。因此，VSLAM 已经存在于各种需要重建环境 3D 模型的应用中，例如：自动驾驶、增强现实（AR）和服务机器人[13]。

作为文献[14]引入的解决高计算成本的通用方法，SLAM 方法主要包含两个并行线程，即 tracking 和 mapping。因此，VSLAM 中使用的算法的分类是在表示研究人员如何在每个线程中使用不同的方法和策略。根据 SLAM 系统使用的数据类型，SLAM 方法可分为两类：直接法和间接法（基于特征的）[15]。

在使用场景中，间接方法从物体纹理中提取特征点（即关键点），并通过在连续帧中匹配描述子来跟踪它们。尽管特征提取和匹配阶段的计算成本很高，但这些方法对于每一帧中的光强度变化是精确和鲁棒的。另一方面，直接法直接依据像素级数据估计相机运动，并对光度误差进行最小化优化。依赖于摄影测量技术，这些方法利用所有相机输出像素，并根据其受约束的方面（如亮度和颜色）在连续帧中跟踪其替换的内容。这些特征使得直接法能够比间接法从图像中建模得到更多的信息，并且能够实现更高精度的 3D 重建。然而，尽管直接方法在纹理较少的环境中效果更好，并且不需要更多的计算来进行特征提取，但它们通常面临大规模优化问题[16]。每种方法的优缺点都鼓励研究人员考虑开发混合解决方案，同时考虑两种方法的组合。混合方法通常将间接和直接的检测阶段结合在一起，其中一个对另一个进行初始化和校正。

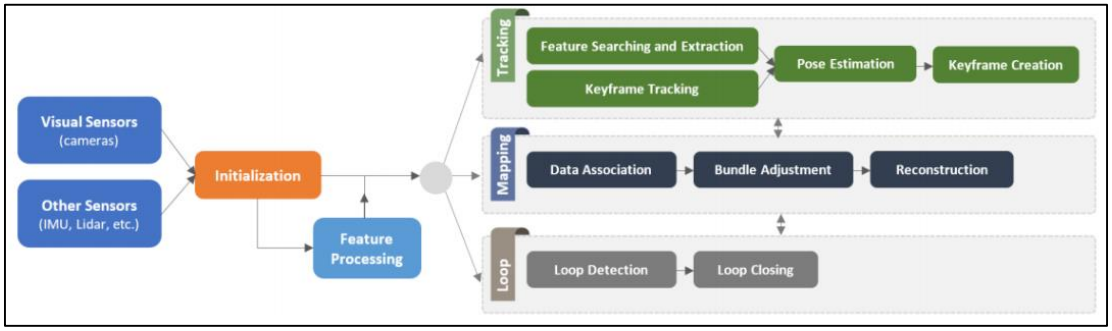

图 1 标准视觉 SLAM Pipeline。关于使用的直接/间接方法，其中一些模块的功能可能会改变或被忽略

此外，由于 VSLAM 主要包括一个视觉里程计（VO）前端（用于本地估计相机的轨迹），以及一个 SLAM 后端（用于优化创建的地图），因此每个部分使用的模块的多样性导致了实现的差异。VO 基于局部一致性提供机器人位姿的初步估计，并发送到后端进行优化。因此，VSLAM 和 VO 之间的主要区别是是否考虑地图和预测轨迹的全局一致性。一些最先进的 VSLAM 应用程序还包括两个附加模块：回环检测和建图[15]。他们负责检测之前访问过的位置，以便根据相机位姿进行更精确的 tracking 和 mapping。

图 1 展示了标准 VSLAM 方法的总体架构。因此，系统的输入也可以与其他传感器数据集成，例如惯性测量单元（IMU）和激光雷达，以提供更多信息，而不只是视觉数据。此外，关于 VSLAM Pipeline 中使用的直接法或间接法，视觉特征处理模块的功能可能会被更改或忽略。例如，“特征处理”阶段仅使用间接法。另一个因素是利用一些特定模块，例如环路闭合检测和束调整，以改进执行。

本文概括了 45 篇 VSLAM 论文，并根据不同方面将其分类为不同类别。我们希望我们的工作将为致力于优化 VSLAM 技术的机器人科研人员提供参考。

本文的其余部分结构如下：

第二节回顾了 VSLAM 算法的演化。

第三节介绍和讨论了 VSLAM 领域的其他综述。

第四节简要介绍了 VSLAM 各个模块。

第五节基于不同应用目标的 VSLAM 分类讨论。

第六节讨论该领域尚未解决的问题和潜在的研究趋势。

02 视觉 SLAM 的演化

VSLAM 系统在过去的几年中已经成熟，有几个框架在这个开发过程中发挥了重要作用。为了清晰表达总体情况，图 2 展示了使用广泛的 VSLAM 方法，这些方法影响了 SLAM 圈的发展，并被用作其他框架的标准参考。

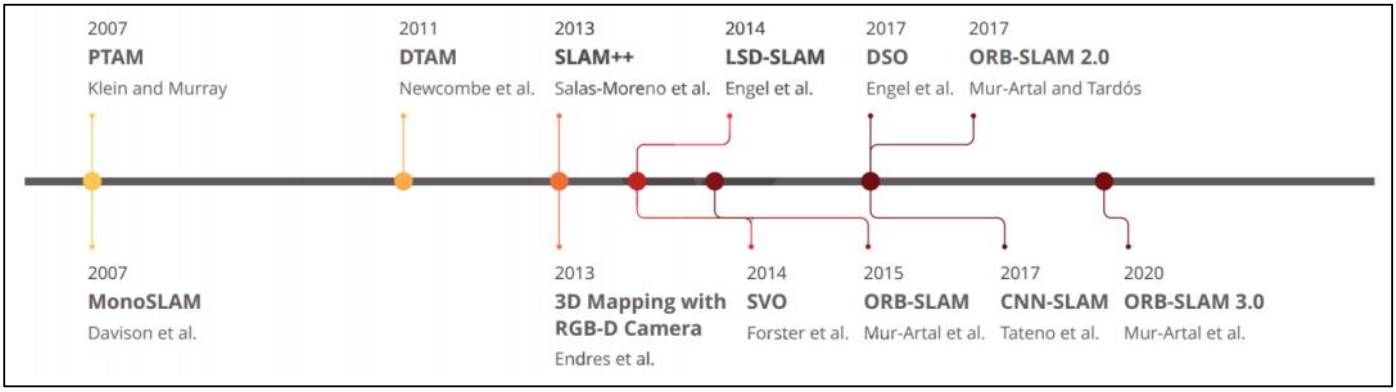

图 2 极具影响力的视觉 SLAM 方法

文献中首次尝试实现实时单目 VSLAM 系统是由 Davison 等人于 2007 年开发的，他们引入了一个名为 Mono-SLAM 的框架[17]。他们的间接法的框架可以使用扩展卡尔曼滤波(EKF)算法估计真实世界中的相机运动和 3D 物体[18]。尽管缺乏全局优化和回环检测检测模块，但 Mono-SLAM 开始在 VSLAM 域中发挥主要作用。不过，用这种方法重建的地图只包括地标，没有提供关于该地区的进一步详细信息。

Klein 等人[14]在同一年提出了并行 tracking 和 mapping（ PTAM ），他们将整个 VSLAM 系统分为两个主要线程：tracking 和 mapping。这一多线程标准在后续工作中得到了许多后续工作的认可，本文将对此进行讨论。他们的方法的主要思想是降低计算成本，并应用并行处理来实现实时性能。当 tracking 线程实时估计相机运动时，mapping 线

程预测特征点的 3D 位置。PTAM 也是第一个利用光束平差法（BA）联合优化相机姿态和创建 3D 地图的方法。它使用 FAST[19]角点检测器算法进行关键点匹配和跟踪。尽管该算法的性能优于 Mono-SLAM，但其设计复杂，在第一阶段需要用户手动设置。

Newcombe 等人于 2011 年推出了一种用于测量深度值和运动参数以构建地图的直接法，即 Dense Tracking and Mapping（DTAM）。DTAM 是一种配备稠密 mapping 和 tracking 模块的实时框架，可通过将整个帧与给定深度图对齐来确定相机姿态。为了构建环境地图，上述阶段分别估计场景的深度和运动参数。虽然 DTAM 可以提供地图的详细表示，但实时执行需要较高的计算成本。

作为 3D 建图领域和基于像素的优化的另一种间接方法，Endres 等人在 2013 年提出了一种基于 RGB-D 相机的方法。他们的方法是实时执行的，专注于低成本的嵌入式系统和小型机器人，但在无特征或具有挑战性的场景中无法产生准确的结果。同年，Salas Moreno 等人[22]提出了在实时 SLAM 框架中利用语义信息的第一个尝试，名为 SLAM++。他们的系统采用 RGB-D 传感器输出，并执行 3D 相机姿态估计和跟踪以形成位姿图（pose graph）。位姿图中的节点表示姿态估计，并由表示具有测量不确定性的节点之间的相对位姿的边进行连接[23]。然后，将通过合并从场景中的语义对象获得的相对 3D 姿态来优化预测位姿。

随着 VSLAM 基本框架的成熟，研究人员专注于提高这些系统的性能和精度。在这方面，Forster 等人在 2014 年提出了一种混合 VO 方法作为 VSLAM 架构的一部分，称为半直接视觉里程计（SVO）[24]。他们的方法可以结合基于特征的方法和直接法来进行传感器的运动估计和建图任务。SVO 可以与单目和立体相机一起工作，并配备了一个姿态细化模块，最小化重投影误差。然而，SVO 的主要缺点是采用短期数据关联，并且无法进行回环检测和全局优化。

LSD-SLAM[25]是 Engel 等人于 2014 年引入的另一种有影响力的 VSLAM 方法，包含跟踪、深度图估计和地图优化。该方法可以使用其姿态图估计模块重建大规模地图，并具有全局优化和回环检测功能。LSD-SLAM 的弱点在于其初始化阶段比较有挑战性，需要平面中的所有点，这使得它成为一种需要大量计算的方法。

Mur Artal 等人提出了两种精确的间接 VSLAM 方法 ,迄今为止吸引了许多研究人员的注意 :ORB-SLAM[26]和 ORB-SLAM 2.0[27]。这些方法可以在纹理良好的序列中完成定位和建图，并使用 Oriented FAST 和 Rotated BRIEF (ORB) 特征执行高性能的姿态检测。ORB-SLAM 的第一个版本能够使用从相机位置收集的关键帧来计算相机位置和环境结构。第二个版本是 ORB-SLAM 的扩展，具有三个并行线程，包括用于查找特征对应的 tracking、用于地图管理操作的本地 mapping 以及用于检测新循环和纠正漂移错误的回环检测。尽管 ORB-SLAM 2.0 可以与单目和立体相机设置一起使用，但由于重建地图数据比例未知，所以它不能直接用于自主导航。这种方法的另一个缺点是它无法在没有纹理的区域或具有重复图案的环境中工作。该框架的最新版本名为 ORB-SLAM 3.0，于 2021 提出[28]。它适用于各种相机类型，如单目、RGB-D 和立体视觉，并提供改进的姿态估计输出。

近年来，随着深度学习在各个领域的显著影响，基于深度神经网络的方法可以通过提供更高的识别和匹配率来解决许多问题。类似地，用 VSLAM 中的学习特征替换手工制作的特征是许多最近基于深度学习的方法提出的解决方案之一。

在这方面，Tateno 等人提出了一种基于卷积神经网络（CNN）的方法，该方法处理相机位姿估计的输入帧，并使用关键帧进行深度估计，命名为 CNN-SLAM[29]。将相机帧分割成较小的部分以更好地理解环境是 CNN-SLAM 中提供并行处理和实时性能的思想之一。

作为一种不同的方法，Engel 等人还引入了直接 VSLAM 算法中的一种新趋势，称为 Direct Sparse Odometry（DSO）[30]，它将直接法和稀疏重建相结合，以提取图像块中的最高强度点。通过跟踪稀疏像素集，它考虑了图像形成参数并使用间接跟踪方法。应注意的是，DSO 只能在光度标定相机时获取完美精度，无法使用常规相机获得高精度结果。

综上，在 VSLAM 系统演化的过程中，最近的方法侧重于多个专用模块的并行。这些模块形成了与多种传感器和环境兼容的通用技术和框架。上述特性使它们能够实时执行，并且在性能改进方面更加灵活。

03 相关综述

VSLAM 领域内有各种综述论文，对不同的现有方法进行了全面分析。每一篇论文都回顾了采用 VSLAM 方法的主要优点和缺点。

Macario Barros 等人[31]将视觉 SLAM 方案分为三个不同类别：纯视觉（单目）、视觉惯性（立体）和 RGB-D。他们还提出了简化分析 VSLAM 的各种标准。然而，它们并没有包括其他视觉传感器，比如我们稍后将在第四章第一节中讨论的基于事件的传感器。

Chen 等人[32]整理了大量的传统和语义 VSLAM 文献。他们将 SLAM 开发时代分为经典、算法分析和鲁棒感知阶段，并介绍了当时的热点问题。他们还总结了采用直接/间接方法的经典框架，并研究了深度学习算法在语义分割中的影响。尽管他们的工作提供了该领域高阶解决方案的全面论述，但方法的分类仅限于基于特征的 VSLAM 中使用的特征类型。

贾等人[33]调研了大量论文，并对基于图优化的方法和使用了深度学习的方法进行了简单对比。不过，尽管进行了适当的对比，但由于调研的论文数量有限，所以他们的结论无法合适概括。

在另一项工作中，Abaspur Kazerouni 等人[34]涵盖了各种 VSLAM 方法，利用了感官设备、数据集和模块，并模拟了几种间接方法进行比较和分析。不过，它们只对基于特征的算法做出了说明，例如 HOG、尺度不变特征变换（SIFT）、加速鲁棒特征（SURF）和基于深度学习的解决方案。Bavle 等人[35]分析了各种 SLAM 和 VSLAM 应用中的位姿感知方面，并讨论了它们的缺点。他们可以得出结论，操作缺乏语义场景的特征可以提高当前研究工作的结果。

其他综述研究了针对特定主题或趋势的最新 VSLAM 方法。例如，Duan 等人[15]研究了交通机器人视觉 SLAM 系统中的深度学习进展。作者在论文中总结了在 VO 和回环检测任务中使用各种基于深度学习的方法的优缺点。在 VSLAM 中使用深度学习方法的显著优点是在姿态估计和总体性能计算中准确提取特征。

在同一领域的另一项工作中，Arshad 和 Kim[36]重点研究了深度学习算法在使用视觉数据的回环检测中的影响。他们回顾了各种 VSLAM 论文，并分析了机器人在不同条件下的长期自主性。

Singandhupe 和 La[37]总结了 VO 和 VSLAM 对无人驾驶车辆的影响。他们整理了在 KITTI 数据集上评估的方法，使他们能够简要描述每个系统的优缺点。

Cheng 等人[32]在一份类似的文章中回顾了基于 VSLAM 的自动驾驶系统，并提出了此类系统的未来发展趋势。

其他一些研究人员调查了 VSLAM 在现实世界条件下的工作能力。例如，Saputra 等人[38]针对在动态和恶劣环境中运行的 VSLAM 技术的变化，讨论了线程的重建、分割、跟踪和并行执行问题。

这份综述与迄今为止的其他综述不同，对不同场地的 VSLAM 进行了全面分析。与其他 VSLAM 综述相比，本文的主要贡献是：

- **根据研究人员提出新解决方案的主要贡献、标准和目标，对 VSLAM 最近的各种出版物进行分类**
- **通过深入研究不同方面的不同方法，分析 VSLAM 的当前趋势**
- **介绍 VSLAM 潜在问题**

04 视觉 SLAM 的各个模块

综合各种视觉 SLAM 方法，我们将不同阶段的需求划分为以下几个模块：

4.1 传感器和数据采集

Davison 等人[17]引入的 VSLAM 算法的早期实施配备了用于轨迹恢复的单目摄像机。单目相机也是最常见的用于各种任务的视觉传感器，如物体检测和跟踪[39]。另一方面，立体相机包含两个或更多图像传感器，使其能够感知

捕获图像中的深度信息，从而在 VSLAM 应用中实现更优质的性能。这些相机配置能够为更高的精度要求提供信息感知，是值得的。RGB-D 相机是 VSLAM 中使用的视觉传感器的其他变体，可以提供场景中的深度信息和颜色信息。在拥有适当照明和适当运动速度的前提下，上述视觉传感器可以在直观的环境中提供关于环境的丰富信息，但它们通常难以应对光照条件差或场景动态范围大的情况。

近年来，事件相机也被用于各种 VSLAM 应用中。当检测到运动时，这些低延迟仿生视觉传感器（low latency bio-inspired vision sensors）可以产生像素级亮度变化，而不是标准强度帧，从而实现高动态范围输出，而不会产生运动模糊影响[40]。与标准相机相比，基于事件的传感器在高速运动和大范围动态场景中可以提供准确的视觉信息，但在运动速率较低时无法提供足够的信息。尽管事件摄像机在恶劣的照明和动态范围条件下可以优于标准视觉传感器，但它们主要提供关于环境的异步信息。这使得传统的视觉算法无法处理这些传感器的输出[41]。此外，使用事件的时空窗口以及从其他传感器获得的数据可以提供丰富的姿态估计和跟踪信息。

此外，一些方法使用多相机配置来解决在真实环境中工作的常见问题，来提高定位精度。利用多个视觉传感器有助于解决复杂问题，例如遮挡、伪装、传感器故障或可跟踪纹理稀疏等，为摄像机提供重叠视场。尽管多相机配置可以解决一些数据采集问题，但单纯相机的 VSLAM 可能会面临各种问题，例如遇到快速移动的对象时的运动模糊、低光照或高光照下的特征不匹配、高速变化场景下的动态对象遗漏等。因此，一些 VSLAM 应用方案可能会在相机旁边配备多种传感器。融合事件和标准帧（standard frames）[42]或将其他传感器（如激光雷达[43]和 IMU）集成到 VSLAM 是一些现有的解决方案。

4.2 应用场景

许多传统 VSLAM 实践中有一个有力假设：机器人在没有意料之外变化的相对静态的世界中工作。因此，尽管许多系统可以在特定环境中成功应用，但环境中的一些意外变化（例如，移动对象的存在）可能会导致系统复杂化，并在很大程度上降低状态估计质量。在动态环境中工作的系统通常使用诸如光流法或随机采样一致性（RANSAC）[44]

之类的算法来检测场景中的移动，将移动对象分类为异常值，并在重建地图时略过它们。这样的系统利用几何信息、语义信息或结合两种信息，来改进定位方案[45]。

此外，我们可以将环境分为室内和室外两类，作为一般分类。室外环境可以是具有结构地标和大规模运动变化（如建筑物和道路纹理）的城市区域，或具有弱运动状态（如移动的云和植被、沙子纹理等）的越野区域，这样的环境提升了定位和回环检测的风险。另一方面，室内环境包含具有完全不同的全局空间属性的场景，例如走廊、墙和房间。我们可以想到，虽然 VSLAM 系统可能在上述区域中的一个工作良好，但在其他环境中可能表现不出相同的性能。

4.3 视觉特征处理

如第一章所述，检测视觉特征并利用特征描述子信息进行姿态估计是间接 VSLAM 方法的一个不可避免的阶段。这些方法使用各种特征提取算法来更好地理解环境并跟踪连续帧中的特征点。特征提取阶段有很多的算法，包括 SIFT[46]、SURF[47]、FAST[19]、BRIEF[48]、ORB[49]等。其中，与 SIFT 和 SURF[50]相比，ORB 特征具有快速提取和匹配而不损失很大准确度的优点。

上述一些方法的问题是它们不能有效地适应各种复杂和不可预见的情况。因此，许多研究人员使用 CNN 来提取不同阶段图像的深层特征，包括 VO、姿态估计和回环检测。根据这些方法的设计功能，这些技术可以表示有监督或无监督的框架。

4.4 方案评估

虽然一些 VSLAM 方法，特别是那些能够在动态和挑战性环境中工作的方法，在真实世界条件下在机器人上进行了测试，但许多研究工作都使用了公开的数据集来证明其适用范围。

Bonarini 等人[51]的 RAWSEEDS 数据集是一个众所周知的多传感器标准测试工具，其包含室内、室外和混合机器人轨迹与地面实况数据。它是应用在机器人和 SLAM 目的的最早的公开标准测试工具之一。

McCormac 等人[52]的 Scenenet RGB-D 是场景理解问题的另一个受欢迎的数据集，例如语义分割和对象检测，其中包含 500 万个大规模渲染的 RGB-D 图像。该数据集还包含像素完整的地面真值标签和精确的相机姿态和深度数据，这些数据使其成为 VSLAM 应用的有力工具。

最近在 VSLAM 和 VO 领域的许多工作已经在 TUM RGB-D 数据集上测试了它们的方法[53]。上述数据集和标准测试工具包含由 Microsoft Kinect 传感器捕获的颜色和深度图像及其相应的地面真值传感器轨迹。

另外，Nguyen 等人[54]的 NTU VIRAL 是由配备 3D 激光雷达、相机、IMU 和多个超宽带（UWB）的无人机收集的数据集。该数据集包含室内和室外实例，旨在评估自动驾驶和空中操作性能。

此外，Burri 等人的 EuRoC MAV[55]是另一个受欢迎的数据集，包含由立体相机拍摄的图像以及同步的 IMU 测量和运动真值数据。根据环境条件，EuRoC MAV 中收集的数据分为易、中、难三类。

Shi 等人的 OpenLORIS Scene[56]是 VSLAM 工作的另一个公开数据集，包含由配备各种传感器的轮式机器人收集的大量数据。它为单目和 RGB-D 算法提供了适当的数据，以及来自车轮编码器的里程计数据。

作为 VSLAM 中使用的更通用的数据集，KITTI[57]是由移动车辆上的两个高分辨率 RGB 和灰度相机捕获的数据集。KITTI 使用 GPS 和激光传感器提供准确的地面信息，使其成为移动机器人和自动驾驶中非常受欢迎的数据集。

TartanAir[58]是另一个标准数据集，用于评估复杂场景下的 SLAM 算法。

此外，伦敦帝国理工学院（Imperial College London）和爱尔兰国立大学梅努斯分校（ICL-NUIM）[59]数据集是另一个包含手持 RGB-D 相机序列的 VO 数据集，该数据集已被用作许多 SLAM 的基准。

与之前的数据集不同，其他一些数据集包含使用特定相机而非常规相机获取的数据。例如，Mueggler 等人[60]引入的事件相机数据集是一个使用基于事件相机采集样本的数据集，用于高速机器人评估。数据集实例包含由运动捕捉系统捕捉的惯性测量和强度图像，使其成为配备事件相机的 VSLAM 的合适基准。

依据传感器设置、应用和目标环境，上述数据集用于多种 VSLAM 方法。这些数据集主要包含相机标定参数以及真值数据。表 1 和图 3 分别显示了数据集的总结特征和每个数据集的一些实例。

表 1 VSLAM 常用数据集；表中的 GT 是指真值的可用性

Dataset Name	Year	Environment		Utilized Sensors									GT
		indoor	outdoor	GPS	lidar	sonar	IMU	mono	stereo	RGB-D	event	omni	UWB	
RAWSEEDS [51]	2006	✓	✓	✓	✓	✓	✓		✓			✓		✓
KITTI [57]	2012		✓	✓	✓		✓	✓	✓					✓
ICL-NUIM [59]	2014	✓								✓				✓
TUM RGB-D [53]	2016	✓					✓			✓				✓
EuRoC MAV [55]	2016	✓					✓	✓	✓					✓
Event Camera Dataset [60]	2017		✓				✓				✓			✓
SceneNet RGB-D [52]	2017	✓								✓				✓
OpenLORIS-Scene [56]	2020	✓			✓		✓	✓	✓	✓				✓
TartanAir [58]	2020	✓	✓		✓			✓	✓	✓				✓
NTU VIRAL [54]	2021	✓	✓		✓		✓	✓					✓	✓

4.5 语义层

机器人需要语义信息才能理解周围的场景并做出更有利的决策。在许多最近的 VSLAM 工作中，将语义信息添加到基于几何的数据中比单纯基于几何的方法更好，使其能够提供周围环境的更多信息[61]。在这方面，预训练的物体识别模块可以将语义信息添加到 VSLAM 模型[62]。最新的方法之一是在 VSLAM 应用中使用 CNN。一般来说，语义 VSLAM 方法包含以下四个主要组成部分[43]：

Tracking：它使用从连续视频帧中提取的二维特征点来估计相机姿态并构建三维地图点云。相机姿态的计算和 3D 地图点云的构建分别建立了定位和建图过程的参考数据。

Local mapping ：通过处理两个连续视频帧，创建一个新的 3D 映射点，该点与 BA 模块一起用于优化相机姿态。

回环检测：通过将关键帧与提取的视觉特征进行比较并评估它们之间的相似性，它调整相机姿态并优化构建的地图。

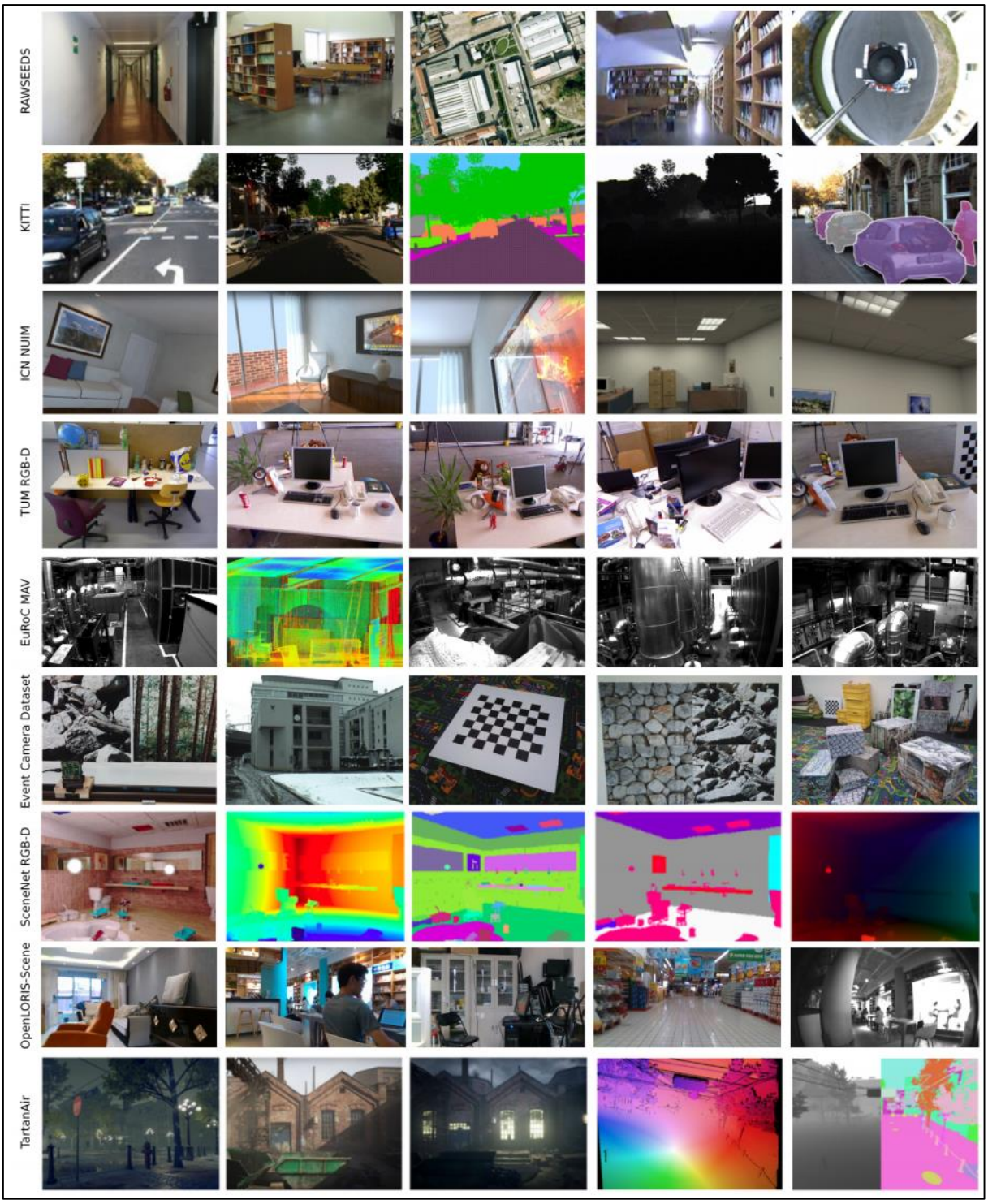

图 3 各种论文中用于评估的一些主流视觉 SLAM 数据集的实例。这些数据集的特征见表 1。

Non-Rigid Context Culling（NRCC）：使用 NRCC 的主要目的是从视频帧中过滤时态物体（temporal objects），以减少它们对定位和建图阶段的不利影响。它主要包含一个屏蔽/分割过程，用于分离帧中的各种不稳定实例，例如人。由于 NRCC 可以减少待处理的特征点的数量，因此简化了计算部分并获得了更鲁棒的性能。

因此，在 VSLAM 方法中利用语义层可以优化位姿估计和地图构建的不确定性。不过，在不会极大地影响计算成本的前提下正确使用提取的语义信息，是现在面临的一个挑战。

05 基于应用目标的 VSLAM 方法分类

为了精确查找能够实现优秀结果并具有稳定架构的 VSLAM 方法，我们从 Google Scholar 和著名的计算机科学书目数据库 Scopus 和 DBLP 中收集并筛选了近年来在顶级网站上发表的被高度引用的出版物。我们还研究了上述出版物中提到的论文，并选取了与 VSLAM 领域最相关的论文。在研究论文之后，我们可以根据主要解决的特定问题将收集的论文进行了分类，如下：

5.1 目标一：多传感器处理

这一类别涵盖了使用各种传感器更好地了解环境的 VSLAM 方法的内容。虽然一些技术只是使用相机作为传感器，但其他技术将各种传感器结合起来以提高算法的准确性。

1）使用多个相机：

由于用一台相机重建运动物体的 3D 轨迹比较困难，一些研究人员建议使用多台相机。例如，CoSLAM 4 是 Zou 和 Tan[63]推出的一个 VSLAM 系统，它使用部署在不同平台上的单独相机来重建鲁棒的地图。他们的系统整合了在动态环境中单独移动的多个相机，并根据它们重叠的视场重建地图。该过程通过整合相机内和相机间姿态估计和 mapping，使得在 3D 中重建动态点云更容易。CoSLAM 使用 KanadeLucas-Tomasi（KLT）算法跟踪视觉特征，并在室内/室外的静态和动态环境中运行，其中相对位置和方向可能会随时间变化。这种方法的主要缺点是需要复杂的硬件来解析大量相机输出的数据，并且由于增加了更多的相机使得计算成本也增加了。

对于具有挑战性的野外场景，Yang 等人[64]开发了一种多相机协同全景 VSLAM 方法。他们的方法需要每个相机都具有独立性，以提高 VSLAM 系统在具有困难条件下的性能，像是遮挡和纹理稀疏的环境。为了确定匹配范围，他

们从摄像机的重叠视场中提取 ORB 特征。此外，他们还采用了基于 CNN 的深度学习技术来识别回环检测的类似特征。在实验中，作者使用了由全景相机和集成导航系统生成的数据集。

MultiCol SLAM 是 Urban 和 Hinz 的另一个开源 VSLAM 框架，使用多相机配置[65]。他们使用之前创建的模型 MultiCol，利用支持多个鱼眼相机的基于关键帧的过程来增强 ORB-SLAM。他们在 ORB-SLAM 中添加了一个多关键帧（MKF）处理模块，该模块收集将图像转换为关键帧的图像。作者还提出了多相机回环的思想，其中回环是从 MKF 中检测出来的。尽管他们的方法是实时运行的，但由于几个线程必须同时运行，因此需要大量的计算能力。

2）Employing Multiple Sensors（使用多种传感器）

其他一些方法推荐融合多种传感器，并使用基于视觉、惯性的传感器输出以获得更好的性能。在这方面，Zhu 等人[66]提出了一种名为 CamVox 5 的低成本间接激光雷达辅助 VSLAM，并证明了其可靠的性能和准确性。他们的方法使用 ORB-SLAM 2.0，将 Livox 激光雷达作为高级深度传感器与 RGB-D 相机的输出相结合。作者使用 IMU 来同步和校正非重复扫描位置。他们的贡献是提出了一种在不受控制的环境中运行的自主激光雷达相机校准方法。在机器人平台上的真实测试表明，CamVox 在处理环境时能够实时运行。

文献[67]中的作者提出了一种名为 VIRAL（视觉惯性测距激光雷达）SLAM 的多模态系统，该系统将相机、激光雷达、IMU 和 UWB 耦合起来。他们还提出了一种基于激光雷达点云构建的局部地图的视觉特征地图匹配边缘化方案。使用 BRIEF 算法提取和跟踪视觉分量。该框架还包含用于所使用的传感器的同步方案和触发器。他们在模拟环境和生成的名为 NTU VIRAL[54]的数据集上测试了他们的方法，该数据集包含相机、激光雷达、IMU 和 UWB 传感器捕获的数据。然而，由于处理同步、多线程和传感器冲突，他们的方法计算量很大。

Vidal 等人[42]建议将事件相机、相机帧和 IMU 集成在并行配置中，以便在高速设置中进行可靠的姿态估计。他们的 Ultimate SLAM 6 系统是基于事件相机和文献[68]中引入的基于关键帧的非线性优化线程。他们分别使用 FAST 角点检测器和 Lucas Kanade 跟踪算法进行特征检测和跟踪。Ultimate SLAM 避免了高速活动带来的运动模糊问题，

并在具有不同照明条件的动态环境中运行。与其他纯事件相机和常规相机的配置相比，此技术在“事件相机数据集”上的效率是显而易见的。作者还在配备了事件相机的自主四旋翼无人机上测试了 Ultimate SLAM，以展示他们的系统如何处理常规 VO 平台无法处理的飞行条件。Ultimate SLAM 面临的主要问题是事件与标准帧输出的同步问题。

Nguyen 等人[69]为 VSLAM 提出了一种紧密耦合的单目相机和 UWB 距离传感器的方法。他们使用基于特征（可见）和无特征（UWB）地标的组合来创建地图。当 UWB 在拥挤环境中受到多径效应（multi-path effects）时，它能有效地工作。他们在 ORB-SLAM 基础上上构建了间接法，并使用 ORB 特征进行姿态估计。他们在一个数据集上测试了他们的系统，该数据集是使用手持的方式模拟空中机器人进行的数据采集。相机和 UWB 传感器的同步是这种情况下的一个大的困难，但通过为每个新图像使用具有相关时间戳的新相机姿态，已经克服了这一问题。

5.2 目标二：Pose Estimation（位姿估计）

这类方法的重点是如何使用各种算法优化 VSLAM 的位姿估计。

1）使用线/点数据：

在这方面，Zhou 等人[70]建议使用建筑结构线作为有用的特征来确定相机姿势。结构线与主导方向相关联，并编码全局方向信息，从而改善预测轨迹。上文说的 StructSLAM 是一种 6 自由度（DoF）VSLAM 技术，可在低特征和无特征条件下运行。它使用 EKF 根据场景中的当前方向估计变量。为了进行评估，使用了 RAWSEEDS 2009 的室内场景数据集和一组生成的序列图像数据集。

点和线 SLAM（PL-SLAM）是由 Pumarola 等人[71]提出的一种基于 ORB-SLAM 的 VSLAM 系统，针对非动态低纹理的场景进行了优化。该系统同时融合线和点特征以改进位姿估计，并帮助在特征点较少的情况下运行。作者在生成的数据集和 TUM RGB-D 上测试了 PL-SLAM。其方法的缺点是计算成本高，而使用其他几何元素（例如平面）则是为了获得更高的精度。

Gomez-Ojeda 等人[72]介绍了 PL-SLAM（不同于 Pumarola 等人[71]中同名的框架），这是一种间接 VSLAM 技术，使用立体视觉相机中的 point 和 line 来重建不可见的地图。他们将从所有 VSLAM 模块中的 point 和 line 获得的片段与从其方法中的连续帧获取的视觉信息合并。使用 ORB 和 Line 检测器（LSD）算法，在 PL-SLAM 中的后续立体帧中检索和跟踪 point 和 line。作者在 EuRoC 和 KITTI 数据集上测试了 PL-SLAM ,在性能方面可能优于 ORB-SLAM 2.0 的立体版本。PL-SLAM 的主要缺点之一是特征跟踪模块所需的计算时间，并且为了提取更多的环境信息，几乎要涵盖所有结构线。

Lim 等人[73]介绍了一种用于单目的基于点、线 VSLAM 的避免退化的技术。一个用于提取 line 特征的强大的基于光流的线路跟踪模块，过滤出每帧中的 short lines，并匹配先前识别的 line 特征，这是他们方法的另一个贡献。为了证明其技术的有效性，并证明其优于已建立的基于点的方法，他们在 EuRoC MAV 数据集上测试了他们的系统。尽管有很号的发现，但该系统缺乏识别正确优化参数的自适应方法。

2）使用其他特征：

文献[74]中提出了一种用于立体视觉相机的框架：双四元数视觉 SLAM（DQV-SLAM），该框架使用贝叶斯框架进行 6-DoF 姿态估计。为了防止非线性空间转换组的线性化，他们的方法使用渐进贝叶斯更新。对于地图和光流的点云，DQVSLAM 使用 ORB 功能在动态环境中实现可靠的数据关联。在 KITTI 和 EuRoC 数据集上，该方法可以可靠地估计实验结果。然而，它缺乏姿态随机建模的概率解释，并且对基于采样近似的滤波的计算要求很高。

Muñoz-Salinas 等人[75]开发了一种使用 artificial squared planar markers 来重建大规模室内环境地图的技术。如果每个视频帧中至少有两个标记是可以观察到的，他们的实时 SPM-SLAM 系统就可以使用标记解决姿态估计的歧义问题。他们创建了一个数据集，其中包含了放置在由一扇门链接的两个房间中的标记的视频序列。尽管 SPM-SLAM 具有很好的价值，但它仅在多个平面标记散布在该区域周围且至少有两个标记可用于标记连接识别时有效。此外，他们的框架处理场景中动态变化的能力并未进行判断。

3) 深度学习方法

Bruno 和 Colombini[76]提出了 LIFT-SLAM，它将基于深度学习的特征描述子与传统的基于几何的系统相结合。他们扩展了 ORB-SLAM 系统的 Pipeline，并使用 CNN 从图像中提取特征，使用学习到的特征提供更稠密和准确的匹配。为了检测、描述和方向估计，LIFT-SLAM 微调 LIFT 深度神经网络。使用 KITTI 和 EuRoC MAV 数据集的室内和室外实例进行的研究表明，LIFT-SLAM 在精度方面优于传统的基于特征和基于深度学习的 VSLAM 方案。但是，该方法的缺点是其计算密集的线程和未优化的 CNN 设计，当然，这也打造了其近乎实时的性能。

Naveed 等人[77]提出了一种基于深度学习的 VSLAM 方案，该方案具有可靠且一致的模块，即使在极其复杂的问题上也是如此。他们的方法优于几种 VSLAM，并使用了在真实模拟器上训练的深度强化学习网络。此外，它们还为主动 VSLAM 评估提供了基线，并可在实际的室内和室外环境中适当推广。网络路径规划器提供了理想的路径数据，由其基础系统 ORB-SLAM 接收。他们制作了一个数据集，其中包含了挑战性和无纹理环境中的实际导航问题，以供评估。

RWT-SLAM 是作者在文献[78]中针对弱纹理情况提出的基于深度特征匹配的 VSLAM 框架。他们的方法基于 ORB-SLAM，采用增强的 LoFTR[79]算法中的特征掩码进行局部图像特征匹配。使用 CNN 架构和 LoFTR 算法分别提取场景中的粗略级别和精细级别描述子。RWT-SLAM 在 TUM RGB-D 和 OpenLORIS 场景数据集以及作者收集的真实世界数据集上进行了测试。不过，尽管具有鲁棒的特征匹配结果和性能，他们的系统仍然需要大量的计算。

5.3 目标三：真实世界的可行性

这类方法的主要目标是在各种环境中使用，并在多种场景下工作。我们注意到，刚刚所提到的方法都对环境的语义信息进行了高度集成，并呈现了端到端的 VSLAM。

1) 动态环境

在这方面，Yu 等人[61]引入了一个名为 DS-SLAM 的 VSLAM 系统，该系统可用于动态环境，并为地图构建提供语义信息。该系统基于 ORB-SLAM 2.0，包含五个线程：跟踪（tracking）、语义分割（semantic segmentation）、局部建图（local mapping）、回环检测（loop closing）和稠密语义地图构建（dense semantic map construction）。为了在位姿估计过程之前排除动态项并提高定位精度，DS-SLAM 采用了具有实时语义分割网络 SegNet 的光流算法[80]。DS-SLAM 已经在真实环境中、RGB-D 相机以及 TUM RGB-D 数据集上进行了测试。不过，尽管它的定位精度很高，但它仍面临语义分割的限制和计算量大的特点。

语义光流 SLAM（SOF-SLAM）是基于 ORBSLAM 2.0 的 RGB-D 模式构建的间接 VSLAM 系统，是 Cui 和 Ma 提出的另一种高动态环境下的方法[45]。他们的方法使用了语义光流动态特征检测模块，该模块提取并跳过 ORB 特征提取提供的语义和几何信息中隐藏的动态特征。为了提供准确的相机姿态和环境信息，SOF-SLAM 使用了 SegNet 的像素级语义分割模块。在高度动态的情况下，在 TUM RGB-D 数据集和真实环境中的实验结果表明，SOF-SLAM 的性能优于 ORB-SLAM 2.0。然而，非静态特征识别的无效方法和仅依赖于两个连续帧的方法是 SOF-SLAM 的弱点。

Cheng 等人[81]使用光流方法分离和消除动态特征点，提出了一种用于动态环境的 VSLAM 系统。他们利用了 ORB-SLAM 的结构，并为其提供了由典型单目摄像机输出生成的固定特征点，用于精确的姿态估计。在无特征的情况下，该系统通过对光流值进行分类并将其用于特征识别。根据 TUM RGB-D 数据集的实验结果，该系统在动态室内环境下运行良好。

Yang 等人[82]发布了另一种 VSLAM 方案，该方案使用语义分割网络数据、运动一致性检测技术和几何约束重建环境地图。他们的方法基于 ORB-SLAM 2.0 的 RGB-D 变体，在动态和室内环境中表现良好。使用改进的 ORB 特征提取技术仅保留场景中的稳定特征，忽略动态特征。然后将特征和语义数据组合起来，以创建静态语义地图。牛津大学和 TUM RGB-D 数据集的评估结果证明了他们的方法在提高定位精度和使用大量数据创建语义地图方面的有效性。不过，他们的系统在走廊或信息较少的地方可能会遇到问题。

2）基于深度学习的解决方案

在 Li 等人[83]的另一个名为 DXSLAM 的工作中，深度学习用于找到与 SuperPoints 相似的关键点，并生成通用描述子和图像的关键点。他们训练了更强的 CNN HF-NET，从每一帧中提取局部和全局信息，生成基于帧和关键点的描述信息。他们还使用离线词袋模型（BoW）方法训练局部特征的视觉字典（Visual vocabulary），以实现精确的回环检测。DXSLAM 可以在不使用图形处理单元（GPU）的情况下实时运行，并且与 CPU 兼容。虽然没有特别强调，但它是有很强的抵抗动态环境中动态变化的能力。DXSLAM 已经在 TUM RGB-D 和 OpenLORIS 场景数据集以及室内和室外图像上进行了测试，可以获得比 ORBSLAM 2.0 和 DS-SLAM 更准确的结果。然而，这种方法的主要缺点是复杂的特征提取架构和深层特征与旧 SLAM 框架合并问题。

Li 等人[84]开发了一种实时 VSLAM 技术，用于在复杂情况下基于深度学习提取特征点。该方法是一个用于特征提取的具有自监督功能的多任务 CNN，可以在 GPU 上运行，支持创建 3D 稠密地图。CNN 的输出是固定长度为 256 的二进制代码串，这使得它可以被更传统的特征点检测器（如 ORB）所替代。它包括三个线程，用于在动态场景中实现准确和及时的性能：tracking、local mapping 和回环检测。这个方案支持使用单目和 RGB-D 相机的 ORB-SLAM 2.0 作为基线。作者分别在 TUM 数据集和自己采集的两个数据集（使用 Kinect 相机采集的走廊和办公室的数据集）上进行了测试。

Steenbeek 和 Nex 在文献[85]中介绍了一种实时 VSLAM 技术，该技术使用 CNN 进行精确的场景分析和地图重建。他们的解决方案利用无人机在飞行过程中的单目相机流，采用深度估计神经网络以获得稳定的性能。上述方法基于 ORB-SLAM 2.0，并利用从室内环境收集的视觉信息。此外，CNN 还接受了 48000 多个室内案例的训练，并操作姿态、空间深度和 RGB 输入来估计尺度和深度。使用 TUM RGB-D 数据集和使用无人机的真实世界测试来评估该系统，证明了姿态估计精度的提高。然而，系统在没有纹理的情况下会很困难，需要 CPU 和 GPU 资源来实现实时性能。

3）使用人工地标（Artificial Landmarks）

Muñoz-Salinas和 Medina Carnicer 开发的一种名为 UcoSLAM 11]的技术，通过结合自然和人造地标，并使用基准标记自动计算周围环境的比例 ,从而优于传统的 VSLAM 系统。UcoSLAM 的主要目的是解决自然地标的不稳定性、重复性和较差的跟踪质量。它可以在没有特征标记的环境中运行，因为它能在只有关键点、只有标记或者混合模式下运行。为了找出地图对应关系，优化重投影误差，并在跟踪失败时重新定位，UcoSLAM 设置了跟踪模式。此外，它有一个基于标记的回环检测系统，可以使用任何描述子描述特征，包括 ORB 和 FAST。尽管 UcoSLAM 有很多优点，但系统在执行了很多的线程，这使得它成为一种耗时的方法。

4）大范围的设置（Wide-range of Setups）

用于动态室内和室外环境的另一种 VSLAM 策略是 DMS-SLAM[87]，它支持单目、立体和 RGB-D 视觉传感器。该系统采用滑动窗口和基于网格的运动统计（GMS）[88]特征匹配方法来找到静态特征位置。DMS-SLAM 以 ORB-SLAM 2.0 系统为基础，跟踪 ORB 算法识别的静态特征。作者在 TUM RGB-D 和 KITTI 数据集上测试了他们提出的方法 ,其结果比一直效果很好的 VSLAM 算法更优。此外 ,由于在 tracking 步骤中删除了动态对象上的特征点 ,DMS-SLAM 比原始的 ORB-SLAM 2.0 执行得更快。尽管有上述优点，但这个方案在纹理少、运动快和高度动态环境的情况下会遇到困难。

5.4 目标四：资源限制（ Resource Constraint）

与具有理想条件的设备相比，一些 VSLAM 方法是为计算资源有限的设备构建的。例如，为移动设备和具有嵌入式系统的机器人设计的 VSLAM 就是这种情况。

1） 计算能力有限的设备：

EdgeSLAM 是 Xu 等人提出的用于移动和资源受限设备的实时、边缘辅助语义 VSLAM 系统[89]。它采用了一系列细粒度模块，由边缘服务器和相关移动设备使用，而不需要很复杂的线程。EdgeSLAM 中还包括基于掩码 RCNN 技术的语义分割模块，以优化目标分割和跟踪的效果。作者将他们的方法进行了实践，在一个边缘服务器上安装了一

些商用移动设备，如手机和开发板。通过重复使用目标分割的结果，他们使系统参数适应不同的网络带宽和延迟情况来避免重复处理。EdgeSLAM 已在 TUM RGB-D、KITTI 的单目视觉实例和为实验设置创建的数据集上进行了评估。

对于立体相机，Schlegel、Colosi 和 Grisetti[90]提出了一种轻量级的基于特征的 VSLAM 框架，名为 ProSLAM，其结果与效果很好的框架不相上下。他们这个方法包括了四个模块：三角测量模块，它创建 3D 点云和相关的特征描述子；增量运动估计模块，其处理两个帧以确定当前位置；地图管理模块，创建局部地图；重定位模块，基于局部地图的相似性更新全局地图。ProSLAM 使用单个线程检索点的 3D 位姿，并利用少量已知库来创建简单的系统。根据 KITTI 和 EuRoC 数据集的实验，他们的方法可以获得不错的结果。然而，它在旋转估计方面表现较弱，并且不包含任何 BA 模块。

Bavle 等人[91]提出了 VPS-SLAM，一种用于空中机器人的基于图的轻量级 VSLAM 框架。他们的实时系统集成了几何数据、多目标检测技术和 VO/VIO，以便于位姿估计和构建环境的语义地图。VPS-SLAM 使用低级特征、IMU 测量和高级平面信息来重建稀疏语义图和估计机器人状态。该系统利用基于 COCO 数据集[93]的轻量级版本 You Only Look Once v2.0（YOLO2）[92]进行目标检测，因为其实时性和计算效率高。他们使用了一个手持相机和一个装有 RGB-D 相机的空中机器人进行测试。TUM RGB-D 数据集的室内实例被用于测试其方法，它们能够提供与已知 VSLAM 方法相同的结果。但是，他们的 VSLAM 系统只能使用少量目标（例如椅子、书籍和笔记本电脑）来构建周围区域的语义地图。

Tseng 等人[94]提出了另一种满足低配条件的实时室内 VSLAM 方法。作者还提出了一种用于估计合理度的定位精度所需的帧数和视觉元素的技术。他们的方案是基于 OpenVSLAM[95]框架，并将其用于现实世界中出现的紧急情况，例如访问特定目标。该系统通过应用高效透视点（EPnP）和 RANSAC 算法获取场景的特征图以进行准确的姿态估计。根据室内测试结果，他们的设备可以在照明条件不佳时，获取准确的结果。

2）计算迁移（Computation Offloading）

Ben Ali 等人[96]提出使用边缘计算（edge computing）将资源密集操作迁移到云上，减少机器人的计算负担。他们在间接框架 Edge SLAM 14 中修改了 ORB-SLAM 2.0 的架构，在机器人上运行了 tracking 模块，并将其余部分迁移到边缘计算设备。通过在机器人和边缘设备之间划分 VSLAM Pipeline，系统可以同时维护局部地图和全局地图。在资源较少的情况下，它们仍然可以在不牺牲准确性的情况下正确运行。他们使用 TUM RGB-D 数据集和两个特定的室内环境数据集（使用不同的移动设备搭载 RGB-D 相机进行采集）进行了评估。然而，他们的方法的缺点之一是由于各种 SLAM 模块的解耦而导致的架构复杂性增大。另一个问题是，他们的系统只是短时间的工作中效果不错，而在长期场景（例如，多天）中使用 Edge SLAM 效果性能会下降。

5.5 目标五：多功能性（Versatility）

VSLAM 在这一类中的工作侧重于直接的开发、利用、适应和扩展。

Sumikura 等人[95]提出了 OpenVSLAM，这是一个适应能力强的开源 VSLAM 框架，主要用于快速开发，也可以被第三方程序调用。他们基于特征的方法与多种相机类型兼容，包括单目、立体和 RGB-D，并且可以存储或重用重建的地图以供以后使用。由于其强大的 ORB 特征提取器模块，OpenVSLAM 在 tracking 精度和效率方面优于 ORB-SLAM 和 ORB-SLAM2.0。然而，由于担心代码相似性侵权 ORB-SLAM 2.0，该系统的开源代码已经停止。

为了弥补实时性、准确性和弹性之间的差距，Ferrera 等人[97]开发了 *OV²SLAM*，可与单目相机和立体视觉相机配合使用。通过将特征提取限制在关键帧中，并通过消除 photo-metric 误差在后续帧中对其进行监控，这种方式减少了计算量。在这个意义上，*OV²SLAM* 是一种混合方案，它结合了 VSLAM 算法的直接法和间接法的优点。在室内和室外实验中使用包括 EuRoC、KITTI 和 TartanAir 在内的著名基准数据集，证明 *OV²SLAM* 在性能和准确性方面优于几种主流方案。

Teed 和 Deng 提出了另一种方法，名为 DROID-SLAM，这是一种基于深度学习的视觉 SLAM，适用于单目、立体和 RGB-D 相机[98]。它们可以获得比众所周知的单目和立体跟踪方法更高的精度和鲁棒性。他们的方案可以实时运

行，包括后端（用于 BA）和前端（用于关键帧收集和图优化）线程。DROID-SLAM 已经使用单目相机实例进行了训练，因此无需使用立体和 RGB-D 输入再次训练。与间接法一样，该方法将投影误差最小化，同时不需要对特征识别和匹配进行任何预处理。包括下采样层和残差块的特征提取网络处理每个输入图像以创建密集特征。DROID-SLAM 已在著名的数据集（包括 TartanAir、EuRoC 和 TUM RGB-D）上进行了测试，并可获得可接受的结果。

Bonetto 等人在文献[99]中提出了 iRotate，这是一种基于 RGB-D 相机的全方位机器人的主动技术。此外，在他们的方法中设置了一个模块，用于发现摄像机视野范围内的障碍物。iRotate 的主要目的是通过提供对未探索的地点和以前访问过的地点的勘察结果，缩短机器人绘制环境地图所需的距离。上述方法使用了一个 VSLAM 框架，以图特征作为其后端。通过在仿真和真实的三轮全向机器人上进行比较，作者可以获得与主流 VSLAM 方法相同的结果。这种方法的主要缺点是机器人可能会面临局部路径重新规划的启动-停止情况。

5.6 目标六：视觉里程计

此类方法旨在获取尽可能高的精度确定机器人的位姿。

1）深度神经网络

文献[100]中提出了动态 SLAM 框架，该框架利用深度学习进行精确的姿态估计和适当的环境理解。作为优化 VO 语义级模块的一部分，作者使用 CNN 来识别环境中的移动对象，这有助于他们降低由不正确的特征匹配带来的姿态估计误差。此外，动态 SLAM 使用选择性跟踪模块来忽略场景中的动态位置，并使用缺失特征校正算法来实现相邻帧中的速度不变性。尽管结果很好，但由于定义的语义类数量有限，该系统需要巨大的计算成本，并面临着对动态/静态对象进行错误分类的风险。

Bloesch 等人[101]提出了 Code-SLAM 直接技术，该技术提供了场景几何的浓缩和密集表示。他们的 VSLAM 系统是 PTAM 的增强版[14]，该系统仅依靠单目相机进行工作。他们将强度图像分成卷积特征，并使用基于 SceneNet

RGB-D 数据集的强度图像训练的 CNN 将其馈送到深度自动编码器。EuRoC 数据集的室内实例已被用于测试 Code-SLAM，结果在准确性和性能方面很有希望。

Wang 等人提出了 DeepVO，这是一种端到端 VO 框架，使用深度递归卷积神经网络（RCNN）架构进行单目设置。他们的方法使用深度学习来自动学习适当的特征，建模顺序动态和关系，并直接从颜色帧推断姿势。DeepVO 架构包括一个称为 FlowNet 的 CNN（可以通过连续帧计算光流），以及两个长期短期存储器（LSTM）层（用于基于 CNN 提供的馈送来估计时间变化）。该框架可以同时提取视觉特征并通过结合 CNN 和递归神经网络（RNN）进行顺序建模。DeepVO 可以将几何信息与为增强 VO 学习的知识模型结合起来。但是，它不能用来替代传统的基于几何的 VO 方法。

Parisotto 等人[103]提出了一种类似 DeepVO 的端到端系统，使用神经图优化（NGO）步骤代替 LSTM。他们的方法基于统一时间的不同姿态进行回环检测和校正。NGO 使用两种注意力优化方法来联合优化由局部姿态估计模块的卷积层做出的聚合估计，并提供全局姿态估计。他们在 2D 和 3D 迷宫上试验了他们的技术，并超过了 DeepVO 的性能和准确度水平。上述方法需要连接到 SLAM 框架以提供重新定位信号。

在另一项工作中，Czarnowski 等人[104]引入了最常见的名为 DeepFactors 的 VSLAM 框架，这个框架主要用于单目相机稠密重建环境地图。为了更稳定地重建地图，他们的实时方案会利用概率数据结合学习和基于模型的方法，进行姿态和深度的联合优化。作者修改了 CodeSLAM 框架，并添加了缺少的组件，如局部/全局回环检测。在对大约 140 万张 ScanNet[105]图像进行训练后，在 ICL-NUIM 和 TUM RGB-D 数据集上对系统进行了评估。DeepFactors 改进了 CodeSLAM 框架的思想，并专注于传统 SLAM Pipeline 中的代码优化。然而，由于模块的计算成本，这种方法需要使用 GPU 来保证实时性能。

2）深度帧间处理

在另一项工作中，通过减少用于相机运动检测的两张图片之间的光度和几何误差，文献[106]的作者为 RGB-D 相机开发了一种实时稠密 SLAM 方法，改进了他们的现有方法。他们基于关键帧的解决方案增强了 Pose SLAM（该方案仅保留非冗余姿态以生成稠密地图），增加了密集的视觉里程计特征，并有效地利用来自相机帧的信息进行稳定的相机运动估计。作者还采用了一种基于熵的技术来计算关键帧的相似性，用于回环检测和漂移避免。然而，他们的方法仍然需要在回环检测和关键帧选择质量方面进行工作。

在 Li 等人介绍的另一项工作中，使用基于特征的 VSLAM 方法（称为 DP-SLAM）实现实时动态目标移除。该方法使用贝叶斯传播模型，该模型依赖从运动目标导出的关键点的似然性。DP-SLAM 可以使用移动概率传播算法和迭代概率更新来克服几何约束和语义数据的变化。它与 ORB-SLAM 2.0 集成，并在 TUM RGB-D 数据集上进行了测试。尽管结果准确，但该系统仅在稀疏 VSLAM 中工作，并且由于迭代概率更新模块而面临较高的计算成本。

Dong 等人提出的室内导航系统 Pair Navi 重复使用智能体先前跟踪的路径，供其他智能体将来使用。因此，被称为 Leader 的上一个移动机器人会捕获跟踪信息，如转弯和特定的环境信息，并将其提供给需要前往同一目的地的后一个移动机器人（follower）。当 follower 使用重定位模块来确定其关于参考轨迹的位置时，Leader 结合了视觉里程计和轨迹创建模块。为了从视频特征集中识别和删除动态目标，系统采用了基于掩码区域的 CNN（Mask R-CNN）。他们在一组由几部智能手机收集的数据集上测试了 Pair-Navi 。

3）各种特征处理

此类别中的另一种方法是一种基于文本的 VSLAM 系统，称为 TextSLAM，由 Li 等人提出。它将使用 FAST 角点检测技术从场景中检索的文本项合并到 SLAM Pipeline 中。文本项包括各种纹理、模式和语义，使该方法更有效地使用它们来创建高质量的 3D 文本地图。TextSLAM 使用文本项作为稳定的视觉基准标记，在找到文本项的第一帧之后对其进行参数化，然后将 3D 文本对象投影到目标图像上以再次定位。他们还提出了一种新的三变量参数化技术，用于初始化瞬时文本项特征。使用单目相机和作者创建的数据集，在室内和室外环境中进行了实验，结果非常准确。在无文本环境中操作、解释短字母以及需要存储大量文本字典是 TextSLAM 的三大基本挑战。

Xu 等人[43]提出了一种基于改进 ORB-SLAM 的间接 VSLAM 系统，该系统使用占用网格映射（OGM）方法和新的 2D mapping 模块实现高精度定位和用户交互。他们的系统可以使用 OGM 重建环境地图，将障碍物的存在显示为等间距的变量字段，从而可以在规划路线时连续实时导航。对生成的数据集的实验检查显示了它们在 GPS-denied 下的接近函数。不过，他们的技术很难在动态和复杂的环境中很好地发挥作用，并且很难适当地匹配走廊和无特征条件下的特征。

Ma 等人提出了 CPA-SLAM 方法，一种用于 RGB-D 相机的直接 VSLAM 方法，其利用平面进行 tracking 和图优化。Frame-to-keyframe 和 frame-to-plane 对齐定期集成在其技术中。他们还引入了一种图像对齐算法，用于跟踪相机的参考关键帧和平面图像的对齐。关键帧数据由 CPA-SLAM 使用，用以查找要跟踪的最短时间和地理距离。在有平面设置和无平面设置的情况下，对系统的跟踪系统的实时性能进行了测试，并对 TUM RGB-D 和 ICL-NUIM 数据集以及室内和室外场景进行了分析。然而，它只支持少量的几何形状，即平面。

06 研究趋势

6.1 统计

关于上述各种综述论文的分类，我们在图 4 中可视化了处理后的数据，以发现 VSLAM 的当前趋势。在子图“a”中，我们可以看到，大多数提出的 VSLAM 系统都是独立的应用程序，它们从头开始使用视觉传感器实现整个定位和建图过程。虽然 ORB-SLAM 2.0 和 ORB-SLAM 是用于构建新框架的基础平台，但最小化方法是基于其它 VSLAM 系统的，如 PTAM 和 PoseSLAM。此外，就 VSLAM 的目标而言，子图“b”中最重要的是改进了视觉里程计模块。因此，最近的大多数 VSLAM 都试图解决当前算法在确定机器人位置和方向方面的问题。姿态估计和真实世界生存能力是提出新的 VSLAM 论文的进一步基本目标。关于被调查的论文中用于评估的数据集，子图“c”说明大多数工作都在 TUM RGB-D 数据集上进行了测试。该数据集已被用作已调研论文中评估的主要基线或多个基线之一。此外，许多研究人员倾向于对他们生成的数据集进行实验。我们可以假设生成数据集的主要动机是展示 VSLAM 方法在真实场景中的工作原理，以及它是否可以用作端到端应用程序。EuRoC MAV 和 KITTI 分别是 VSLAM 工作中下一个流

行的评估数据集。从子图“d”中提取的另一个有趣的信息是关于使用 VSLAM 系统时使用语义数据的影响。我们可以看到，大多数被调研的论文在处理环境时不包括语义数据。我们假设不使用语义数据的原因是：

- 在许多情况下，训练识别对象的模型并将其用于语义分割的计算成本相当大，这可能会增加处理时间。
- 大多数基于几何的 VSLAM 方案都被设计为即插即用设备，因此它们可以尽可能少地使用相机数据进行定位和建图。
- 从场景中提取的错误信息也会给过程中增加更多的噪声。

当考虑环境时，我们可以在子图“e”中看到，一半以上的方法也可以在具有挑战性条件的动态环境中工作，而其余的系统只关注没有动态变化的环境。此外，在子图“f”中，大多数方法都适用于“室内环境”或“室内和室外环境”，而其余论文仅在室外条件下进行了测试。应该提到的是，如果在其他场景中使用，只能在特定情况下工作的方法可能不会产生相同的准确性。这就是为什么有些方法只关注特定情况的主要原因之一。

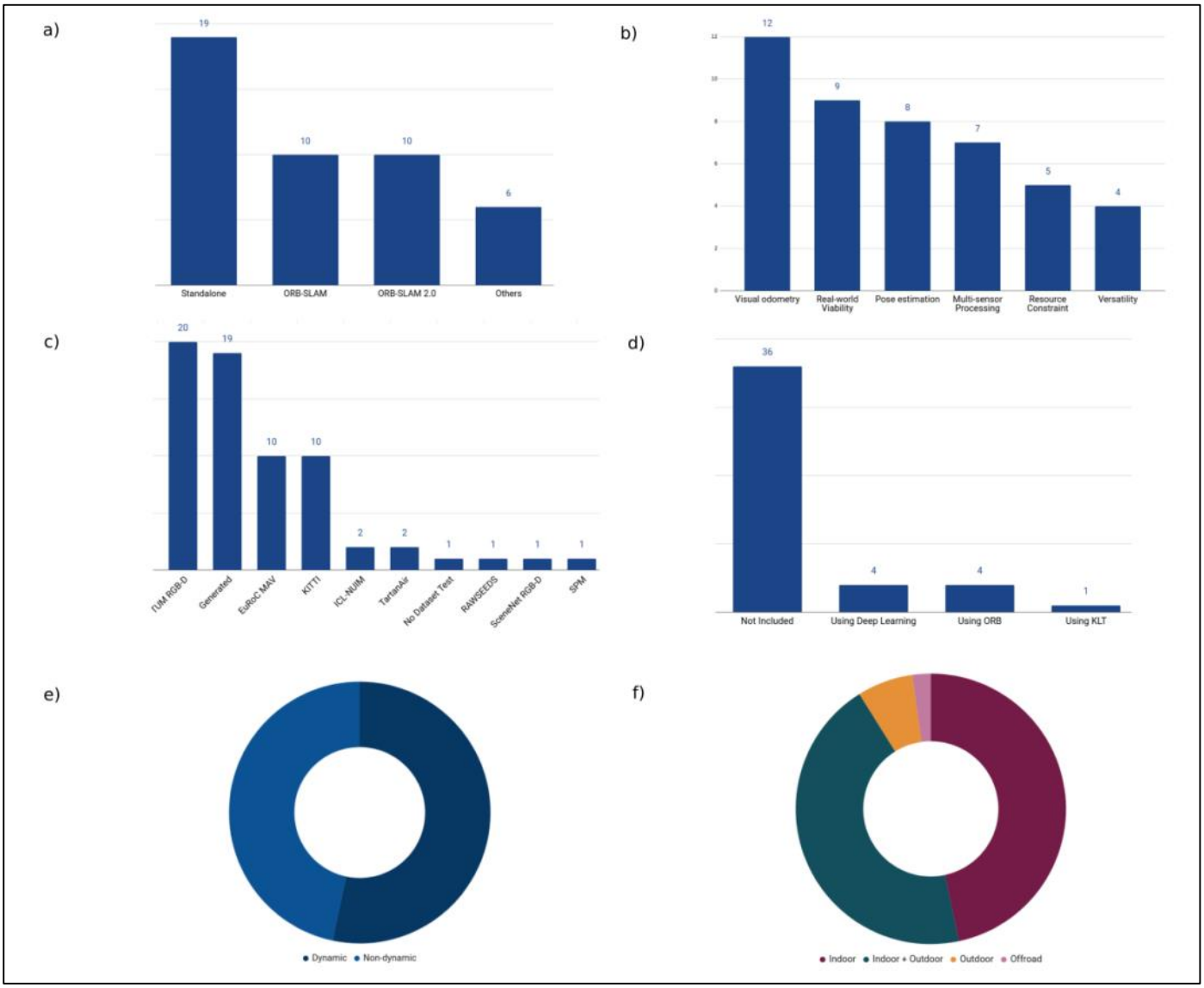

图 4 VSLAM 的当前研究趋势：a）用于实现新方法的基本 SLAM 系统；b）该方法的主要目的；c）提出的方法正在测试的各种数据集；d）在提出的方法中使用语义数据的影响；e）环境中存在的动态对象的数量；f）方案测试的各种环境。

6.2 趋势分析

当前的综述回顾了最新的广受关注的视觉 SLAM 方法，并说明了它们在该领域的主要贡献。尽管在过去的几年中，VSLAM 系统的各个模块有了广泛的稳定的解决方案和改进，但仍有许多高潜力领域和未解决的问题，这些领域的研究将为 SLAM 的未来发展带来更稳定的方法。鉴于视觉 SLAM 方法非常多，我们在此讨论当前趋势领域，并引入以下开放研究方向：

深度学习：深度神经网络在各种应用中显示出令人振奋的结果，包括 VSLAM[15]，使其成为多个研究领域的重要趋势。由于它们的学习能力，这些架构已经显示出相当大的潜力，可以用作不错的特征提取器，以解决 VO 和回环检测中的问题。CNN 可以帮助 VSLAM 进行精确的物体检测和语义分割，并且在正确识别 hand-crafted 特征方面可以优于传统的特征提取和匹配算法。不可不提的是，由于基于深度学习的方法是在具有大量多样化数据和有限对象类的数据集上训练的，因此总是存在对动态点进行错误分类并导致错误分割的风险。因此，它可能导致较低的分割精度和姿态估计误差。

信息检索和计算成本的平衡：通常，处理成本和场景中的信息量应该始终保持平衡。从这个角度来看，稠密地图允许 VSLAM 应用程序记录高维完整的场景信息，但实时执行将需要大量的计算。另一方面，尽管计算成本较低，但稀疏表示将无法捕获所有需要的信息。还应注意的是，实时性能与相机的帧速率直接相关，峰值处理时间的帧丢失会对 VSLAM 系统的性能产生负面影响，而与算法性能无关。此外，VSLAM 通常利用紧密耦合的模块，修改一个模块可能会对其他模块产生不利影响，这使得平衡任务更具挑战性。

语义分割：在创建环境地图的同时提供语义信息可以为机器人带来非常有用的信息。识别相机视场中的对象（例如门、窗、人等）是当前和未来 VSLAM 工作中的一个热门话题，因为语义信息可用于姿态估计、轨迹规划和回环检测模块。随着目标检测和跟踪算法的广泛使用，语义 VSLAM 无疑将是该领域的未来解决方案之一。

回环检测：任何 SLAM 系统都有一个关键问题：**漂移和**由于累积的定位误差而导致的**特征轨迹丢失**问题。漂移检测和回环检测，需要识别先前访问的位置信息，而这会导致 VSLAM 的高计算延迟和成本[89]。主要原因是回环检测的复杂度随着地图重建的大小而增加。此外，组合从不同位置收集的地图数据并细化估计位姿是非常复杂的任务。因此，回环检测模块的优化和平衡具有巨大的优化潜力。回环检测的常用方法之一是通过训练基于局部特征的视觉字典，然后将其聚合来优化图像检索。

特殊场景问题：在没有纹理的环境中工作，很少有明显的特征点，这通常会导致机器人的位置和方向出现漂移误差。作为 VSLAM 的主要挑战之一，此错误可能导致系统故障。因此，在基于特征的方法中考虑互补的场景理解方法，如对象检测或线特征，将是一个热门话题。

07 结论

本文介绍了一系列 SLAM 工作，其中从相机收集的视觉数据发挥了重要作用。我们根据 VSLAM 系统方法的各种特性对其最近的工作进行了分类，如实验环境、创新性领域、物体检测和跟踪算法、语义层、性能等。我们还根据作者观点、未来版本的优化以及其他相关方法中解决的问题，回顾了相关工作的关键贡献以及存在的缺陷和挑战。论文的另一个贡献是讨论了 VSLAM 系统的当前趋势以及研究人员将进一步研究的开放问题。