DOI: 10.19650/j.cnki.cjsi.J2311081

基于深度学习的视觉同时定位与建图研究进展*

张 耀,吴一全,陈慧娴
(南京航空航天大学电子信息工程学院 南京 211106)

摘 要:随着机器视觉的不断发展,视觉传感器其小巧轻便、价格低廉等优势,使得视觉同时定位与建图(VSLAM)越来越受人们关注,深度学习为处理 VSLAM 问题提供了新的方法与思路。本文综述了近年来基于深度学习的 VSLAM 方法。首先回顾了 VSLAM 的发展历程,系统阐释了 VSLAM 的基本原理与组成结构。然后从视觉里程计(VO)、回环检测与建图 3 个方面分析各类基于深度学习的方法,从特征提取与特征匹配、深度估计与位姿估计及关键帧选择等 3 个部分阐述了深度学习在 VO 中的应用;基于场景表达方式的不同,总结了几何建图、语义建图及广义建图中的深度学习方法。接着介绍了目前 VSLAM 常用的各种数据集以及性能评估指标。最后指出了目前 VSLAM 面临的难题与挑战,展望未来深度学习与 VSLAM 结合的研究趋势与发展方向。

关键词: 同时定位与建图;机器视觉;深度学习;视觉里程计;回环检测;数据集;评估指标
中图分类号: TP242.6 TH89 **文献标识码**: A **国家标准学科分类代码**: 510.8050

Research progress of visual simultaneous localization and mapping based on deep learning

Zhang Yao, Wu Yiquan, Chen Huixian
(*College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*)

Abstract: With the continuous development of machine vision, visual sensors has advantages of lightweight and low cost. Thus, visual simultaneous localization and mapping (VSLAM) is attracting more and more attention and becoming a research hotspot. Deep learning has provided new methods and ideas to deal with VSLAM problems. This article reviews the deep learning-based VSLAM methods in recent years. Firstly, the development history of VSLAM is reviewed, and the basic principle and composition structure of VSLAM are systematically explained. Then, various methods based on deep learning are summarized and analyzed from three aspects, including visual odometry (VO), loop closure detection and mapping. The application of deep learning in visual odometry is described in three parts, which are feature extraction and feature matching, depth estimation and pose estimation and keyframes selection. Based on the different manner of scene representation, deep learning-based methods in geometric mapping, semantic mapping and general mapping are summarized. Thirdly, it introduces various datasets and performance evaluation metrics commonly used in VSLAM at present. Finally, the challenges of VSLAM are pointed out, and the future research trends and development directions of combining deep learning with VSLAM are forecasted.

Keywords: simultaneous localization and mapping; machine vision; deep learning; visual odometry; loop closure detection; dataset; evaluation metrics

0 引 言

同时定位与建图(simultaneous localization and mapping, SLAM)是为了在未知环境中,机器人能够在移动过程中实现自我定位及构造地图而提出的技术。SLAM 早期大多是采用激光雷达、声纳等,虽然具备较高的精度,然而其造价高昂,体积庞大,且容易损坏。而以

收稿日期:2023-02-22 Received Date: 2023-02-22
*基金项目:国家自然科学基金(61573183)项目资助

摄像头为主的视觉传感器不仅轻巧便携,成本低廉、安装方便,而且拍摄的图像包含了丰富的各种环境信息,适用范围广,图像的表示也类似于人眼感知环境的形式。

通过使用视觉传感器的视觉同时定位与建图技术(visual simultaneous localization and mapping, VSLAM)感知周围环境,在复杂的三维空间中进行定位以及导航。这种 VSLAM 技术在诸如智能机器人、自动驾驶汽车、无人机、无人航行器、军用探测车、增强现实(augmented reality, AR)[1]、虚拟现实(virtual reality, VR)等领域发挥着重要作用。智能工厂的机器人对货物的自动分拣与搬运,实现生产线的全自动化生产。对于全球定位系统(global positioning system, GPS)失效的特殊场景(如山洞、隧道、深水、电磁干扰等),自动驾驶车、救援机器人、无人水下航行器等在复杂环境下完成救援、长距离巡航、探测等任务。在军事危险场所中,军用探测车进行侦察、排雷、危化品处理等。此外,近年兴起的 AR、VR 技术将虚拟网络空间与现实世界交织一起,VSLAM 重建的三维地图可以提供场景几何信息,对叠加的虚拟物体做相应渲染保持其与现实场景的几何一致性,从而使得虚拟空间更加真实。随着这些领域对 VSLAM 技术需求的日益增大,各种新颖的方法和技术不断涌现,VSLAM 的研究也逐渐成为热点。

目前 VSLAM 的方法大致可以分为基于滤波器的方法、基于优化的方法、基于深度学习的方法等。基于滤波器的方法包括了卡尔曼滤波器、扩展卡尔曼滤波器、无迹卡尔曼滤波器、粒子滤波器等。但是这种方法无法胜任大尺度的室外场景,随着人们对 VSLAM 技术要求的不断提高,出现了基于优化的方法,以图优化为主。而随着深度学习的迅速发展,机器视觉与深度学习的结合在精度、鲁棒性及运算效率等方面都获得较大提升,人们开始考虑数据驱动学习的方法来解决 VSLAM 的问题。基于深度学习的 VSLAM 方法有着如下优势:

1) 较强的泛化能力。通过神经网络自动提取特征,相比于手工设计,能更充分地利用丰富的图像信息,更好地适应各种复杂的场景,如运动模糊、动态环境、大尺度场景等。

2) 提取高级语义信息。可以提取到更高级的语义信息,构建语义 SLAM 以及地图理解。此外,易将抽象元素(如语义标签)与人为理解的术语建立联系,而这种联系用数学公式却难以描述。

3) 深度学习这种数据驱动的方式更符合人类与环境交互的形式,有着更大的发展空间与研究前景。

4) 能充分利用海量的数据与日益提高的硬件算力。随着 VSLAM 系统的运行,传感器产生的大量数据与神经网络的大量优化参数都会导致巨大的运算量,硬件算力的提高为系统实时性提供了保证。

5) 深度学习可以从过去的经验学习。通过构建通用的网络模型,在面对新的场景自动训练发现新的解决方案,从而进一步改进自身模型。

然而,这方面的综述还不够系统、全面和深入。文献[2]仅总结了早期的深度学习 VSLAM 方法,至于深度学习与建图的结合没有涉及。文献[3]从宏观的角度,阐述了 SLAM 的发展与传统算法。文献[4]则详细介绍了传统基于滤波和基于图优化的 VSLAM 算法,并没有对各类算法进行深入对比分析,也不涉及深度学习的方法。文献[5-6]主要是针对动态场景 VSLAM 的介绍。文献[7-9]按各自的分类方法,梳理了部分基于深度学习的 VSLAM 算法,但是不够全面且没有包含最新进展。文献[10]主要包括视觉与激光雷达融合的解决方案。文献[11]从 6G 无线网络角度,总结了激光雷达和 VSLAM 技术,对基于深度学习的 VSLAM 只是简单罗列。文献[12]详细介绍了 VSLAM 的经典组成结构以及最新算法,但是并不是以深度学习为主要部分。

综上,对于基于深度学习的 VSLAM 算法,还没有较为系统、全面而深入的综述。本文总结了近年来最新的基于深度学习的 VSLAM 方法,旨在帮助相关领域的研究人员更快速地了解基于深度学习的 VSLAM 的研究现状与发展前景。

1　VSLAM 的应用领域

VSLAM 技术在全球范围已经广泛应用于智能物流、智能家居、自动驾驶、AR/VR 等领域,产生了巨大的经济效益并深刻地影响着社会的发展。

随着网购的普及和快速增长,传统的物流模式依靠人工物流作业不仅效率低下,而且成本高昂,已经难以满足市场需求。VSLAM 技术通过让机器人自主识别物流场地、货架和物品的特征,以及辅助设备科学高效地展开分拣、搬运、装箱等物流作业,从而使物流作业更加高效、快捷,减少物品丢失等意外问题。直接促进网购和物流行业蓬勃发展的同时,也间接地推动了互联网和经济的繁荣。

在智能家居领域,VSLAM 技术通过房间的布局和特征,构建出室内地图,并根据地图自主地规划路径,从而为用户提供更好的服务。最为典型的应用就是智能扫地机器人,VSLAM 技术可以帮助扫地机器人实时感知环境的变化,并在移动的过程中动态更新地图,自主定位房间内的位置和识别障碍物与垃圾,从而能够更准确地规划清扫路线。

相比于居家和工厂的室内环境,户外的世界存在着更多的不确定性因素且环境地图规模更大。在自动驾驶领域,VSLAM 技术可以实现高清晰度地图的构建和车辆

位置的准确定位,帮助无人驾驶汽车实现更高效的行驶路线规划和避让障碍物的能力,带来更智能、高效、安全的驾驶体验,也能帮助智能交通系统更加准确地规划交通,优化路径规划,提高交通效率和安全性,对于国家未来智能交通的建设有着重大的意义。

在 AR/VR 领域中,都需要场景重建和位置跟踪来实现沉浸式的效果。VSLAM 技术可以快速地捕捉现实世界中的环境特征,实时更新 VR 场景的位置和姿态,从而保持虚拟场景和现实环境的同步,增强了虚拟场景的真实感以及用户的沉浸感。在 AR 应用中,则可以帮助用户更好地实现与虚拟场景的交互。例如在现实环境中合成虚拟物体,实现在现实场景中与远在千里的朋友互动。VSLAM 技术与 AR/VR 的结合,提高了交互性和真实性,也推动着相关产业的创新和发展。

除此以外,VSLAM 技术已经应用到各个行业领域中,极大地促进相关领域繁荣的同时,也深刻地影响着社会乃至世界的发展方向。上至航空航天,军事行动,自然灾害搜救等国家国防之基础,下至物流工厂,家居智能,医疗健康,交通出行等民主民生之根本。VSLAM 的一些典型应用如图 1 所示。可以预见,随着技术的不断创新攀升,VSLAM 技术将在更多领域得到应用,为未来社会带来更多新的机遇和可能性。

(a) 自动驾驶
(a) Autonomous driving

(b) AR/VR 技术
(b) AR/VR technology

(c) 分拣机器人
(c) Sorting robots

(d) 扫地机器人
(d) Sweeping robots

图 1 VSLAM 的典型应用

Fig. 1 Typical applications of VSLAM

2 VSLAM 发展历程与基本结构

2.1 VSLAM 发展历程

SLAM 的概念最早出现在 1986 年的 IEEE 机器人与自动化会议(IEEE Robotics and Automation Conference)中,Smith 等[13]在会议上提出将基于估计理论的方法应用到机器人的定位与建图中,是机器人领域与人工智能领域结合的开端。迄今为止已经走过了 30 多年,Cadena 等[14]将其发展历程大致分为 3 个时代,如表 1 所示。

早期的 SLAM 研究大多是围绕激光雷达展开,激光雷达传感器测量距离较为准确,误差模型简单,基于多线束激光雷达构造出的三维点云地图,不仅在帧间匹配上具备更多的匹配手段,而且容易融合物理模型、图像等信息,鲁棒性更强,定位精度较高。Ayache 等[15]最早进

表 1 SLAM 发展的 3 个时代

Table 1 Three eras of development of SLAM

年代	时代	主要研究工作
1986~2004 年	传统时代	主要是提出问题和理清框架。引入了概率形式推导框架,包括基于扩展卡尔曼滤波器、Rao-Blackwellised 粒子滤波器和极大似然估计的方法,提出了效率问题和数据关联的鲁棒性问题。
2004~2015 年	算法分析时代	主要进行算法的分析以及深入研究 SLAM 的可观测性、收敛性和一致性。同时人们理解了增量方程中矩阵结构的稀疏性对于效率提升有着至关重要的作用,开发出一些开源的 SLAM 框架。
2015 年至今	预测性-鲁棒性时代	未来的 SLAM 系统拥有: 1)非常强的鲁棒性能:具备失效保护机制,能自动调整系统的参数以适应各种的场景。 2)更高级别的理解能力:实现基本的地图重建,理解环境信息以完成相应的任务。 3)优化资源与任务驱动:根据实际机器人任务所需,生成自适应复杂程度的地图。

行视觉导航的研究。文献[16]提及 Chatila 和 Laumond 早期基于卡尔曼滤波器研究移动机器人视觉导航。直到 2000 年左右,大部分的 VSLAM 还是采用滤波器的方法。虽然也出现使用光束平差法(bundle adjustment, BA)优化来实现 VSLAM,如 PTAM,但非线性优化的误差与巨大的计算量使其并不受学者们青睐。随着 2009 年 BA 的稀疏性的发现[17],基于图优化的方法才开始逐渐增多,涌现了 DTAM[18]、LSD-SLAM[19]、SVO[20]、ORB-SLAM[21]等各种优秀算法。随着计算机性能的大幅提升,视觉传感器结构小巧,成本低廉,能获得丰富的形状、颜色、语义等辅助信息。以各种摄像头等传感器的 SLAM 逐渐成为近年来的研究热点,传统滤波的方法存在不可避免的环境适应性问题,且随着人工智能技术和深度学习的快速发展,人们在研究基于优化方法的同时,也在思考着如何把深度学习应用到传统滤波器方法与图优化方法中,利用深度学习来解决 SLAM 的相关问题已成为主流趋势。随着硬件算力的不断提升,人们对便携性与实时性要求越来越高。视觉传感器结构小巧,成本低廉,能获得丰富的形状、颜色、语义等辅助信息。使用各种视觉传感器的 VSLAM 逐渐成为近年来的研究热点。

2.2　VSLAM 基本结构

和经典 SLAM 系统一样,VSLAM 系统同样包含前端、后端、回环检测以及建图 4 个主要组成部分[22]。前端主要通过图像帧间估计得到相机的运动关系,完成对相机位姿的估计与局部地图的三维重建。后端则是对前端得到的结果进行优化,通过滤波器或者非线性优化求解得到最优的位姿估计与更精确的地图。回环检测用于判断是否回到曾经来过的位置。后端优化能根据回环检测的结果,更新和调整轨迹和地图,消除累计误差,得到全局一致性轨迹和地图。建图是根据自身定位与环境信息感知的结果,构建出与任务需求对应的地图。

1)前端

前端又称为视觉里程计(visual odometry, VO),是利用运动中的相机在不同时刻获取到的图像帧,对相邻图像帧进行帧间估计得到相机位姿变化,并融合图像帧重建出局部的三维地图。目前主要的方法包括特征点法与直接法。

(1)特征点法

目前最常用的特征点有尺度不变特征变换(scale-invariant feature transform, SIFT)[23]、加速鲁棒特征(speeded up robust features, SURF)[24]、快速特征检测与旋转不变性描述子(oriented FAST and rotated BRIEF, ORB)[25]等。SIFT 对旋转、尺度缩放与亮度变化具有不变性且信息量大,然而其大量的浮点运算以及较高的数据存储复杂度,使其无法实现实时处理。SURF 是对 SIFT 的进一步优化,基于 Hessian 矩阵进行特征点检测,构造尺度空间时利用箱式滤波器简化 2 维高斯滤波,大大加快了检测的速度且稳定性更好。ORB[25]结合了加速分段测试特征(features from accelerated segment test, FAST)[26]与二进制鲁棒独立基本特征(binary robust independent elementary features, BRIEF)[27],具备尺度不变性与旋转不变性。然而它对特征的缺失较为敏感,在纹理较弱甚至无纹理的环境中,鲁棒性较差。

接着需要进行特征匹配。在相邻帧之间的运动与外观变化较大时,需要计算描述子之间的距离度量(如汉明距离、欧氏距离)进行特征点匹配。匹配方法有 K 最近邻(K nearest neighbors, KNN)匹配、暴力匹配、交叉匹配等。而在运动与外观变化较小的时,可以采用光流跟踪实现关键点的匹配。通过跟踪关键点使图像两个对应位置的光度误差最小从而得到相机的运动关系,无需计算描述子,因此运算速度较快。但是光流跟踪遵循灰度不变假设,对光度变化极其敏感,在光度变化较大时容易出现跟踪丢失。

(2)直接法

直接法无需提取特征点与计算描述子,直接根据两帧图像中的像素灰度信息,基于梯度搜索计算求解相机位姿,估计相机的运动关系,构建光度误差优化函数并使其最小化进行位姿优化,因此在纹理特征较弱或者无纹理环境中,具有较强的鲁棒性。同样,直接法也是基于灰度不变假设,因此受光照与快速模糊影响很大,要求相机运动较慢或者采集频率较高。完成位姿估计后,将新图像帧中的点云映射到全局坐标系中,与重建的地图融合,逐帧更新最终得到局部的三维地图。

由于特征点法与光流法都提取了特征点,仅利用到了极少部分的像素点,丢失大量可能有用的图像信息,因此只能构建出稀疏的地图。直接法直接对图像像素点进行处理,不需要提取特征点,可以构建出半稠密与稠密的地图。但是直接法基于光度不变性假设,非常容易受到光照与模糊的影响,单个像素点没有区分度而且图像的非凸性容易使优化陷入到局部最优中。

2)后端

由于帧间估计仅考虑了相邻的两帧图像帧,每两张图像间的运动都必定存在一定误差,相邻帧间估计的误差经过多次的传递导致误差累积,从而出现轨迹漂移。因此需要后端优化与回环检测来消除累计误差。目前主要的优化方法可分为基于滤波理论的优化与基于非线性优化。

(1)基于滤波器优化

早期的后端优化均是以滤波器为基本框架。卡尔曼滤波器适用于运动方程与观测方程均为线性方程,且状态与噪声均服从高斯分布。但是很多实际系统均存在不同程度的非线性,因此出现了扩展卡尔曼滤波器

(extended Kalman filter, EKF),其性能很大程度上取决于局部非线性化程度与原始不确定度,且随着系统的运行,需要更新和维护的均值以及协方差矩阵的规模呈现平方增长,占用大量的存储空间,导致 EKF 不适用于大规模环境。为此出现了无迹卡尔曼滤波器(unscented Kalman filter, UKF)。UKF 避免了求解复杂的 Jacobian 矩阵,不会随着系统模型的复杂而增加算法的实现难度。但其参数选择问题没有得到完全解决,且与 EKF 算法一样,滤波初值会极大影响滤波效果。

粒子滤波(particle filter, PF)采用蒙特卡罗模拟方法实现贝叶斯最优估计,通过重要性采样在状态空间得到一组不断更新的粒子,适用于任何状态空间模型。使用重采样策略来舍弃权值较小的粒子,复制权值较大的粒子,从而使粒子更逼近真实状态的概率分布,通过粒子群加权平均计算得到系统状态估计值。然而,重采样导致权值越大的粒子子代越多,权值小的粒子子代越少甚至无子代。而且容易导致样本有效性和多样性的缺失,从而出现样本贫化现象,因此同样不适合在大规模环境中应用。

SLAM 的研究逐渐从室内的小场景转变到室外的大场景,基于滤波优化方法的局限性使其无法在大场景的 SLAM 问题中胜任。另外,基于滤波器优化假设状态间具备马尔科夫性,这也导致后续的回环问题非常难以处理。基于滤波器优化仅使用了部分历史数据,而基于非线性优化方法使用了所有的历史数据,理论上能得到更好的性能。

(2)基于非线性优化

BA 的总体思想就是根据视觉图像估计位姿与空间点的位置,通过构造相应的最小二乘问题进行优化求解。不同的求解方法有高斯牛顿下降法,梯度下降法以及列文伯格-马夸尔特法。无论哪种方法,最终的增量方程求解都得到了 $\boldsymbol{H}\Delta x = g$ 或者 $(\boldsymbol{H}\lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})\Delta x = g$ 的形式。因此在问题规模较大的时候,求解 $\boldsymbol{H}$ 矩阵非常困难。BA 的优化问题计算量大,无法应用于实时计算。随着 SLAM 问题的发展,人们逐渐认识到 $\boldsymbol{H}$ 矩阵的稀疏结构,并且该结构可以自然、显式地用图优化表示,得以将 BA 应用于实时场景。不同于滤波理论优化的马尔可夫假设,仅考虑了相邻两个状态,非线性优化对之前的状态全部进行优化,使用到全部的历史信息,而且状态矩阵的稀疏性也让求解速度更快,因此对于回环检测问题更容易求解,也能在实时场景中实现,使其成为目前的主流方法。

3)回环检测

现阶段应用最广的回环检测方法是词袋模型(bag-of-words, BoW)。通过对图像计算生成描述子,用 K 均值聚类算法对描述子聚类得到词典,词典中词的频率即可描述该帧图像。通过对不同关键帧的词袋计算相似度进行回环检测。然而,使用 SIFT、SURF 等特征描述子进行特征描述,存在计算量大、相似度计算能力低的问题。词袋模型本质上也是一种无监督的机器学习过程,回环检测计算两帧图像相似性也可视为一个分类问题。人们开始将深度学习与回环检测结合起来,通过 Superpoint、D2、R2D2 等深度学习特征替换传统的 ORB、SIFT 特征来实现闭环检测。Li 等[28]将各种深度学习特征点提取与匹配网络(Superpoint、D2NET、HF-NET 等)替换 ORB-SLAM 中的相关结构,用于重定位与回环检测。Gao 等[29]提出了一种基于堆叠去噪自动编码器(stack denoising autoencoder, SDA)的多层神经网络,通过无监督方式从原始图像数据学习图像的特征。随着 SLAM 技术的不断发展,使用深度学习结合回环检测目前已经成为一种趋势。

4)建图

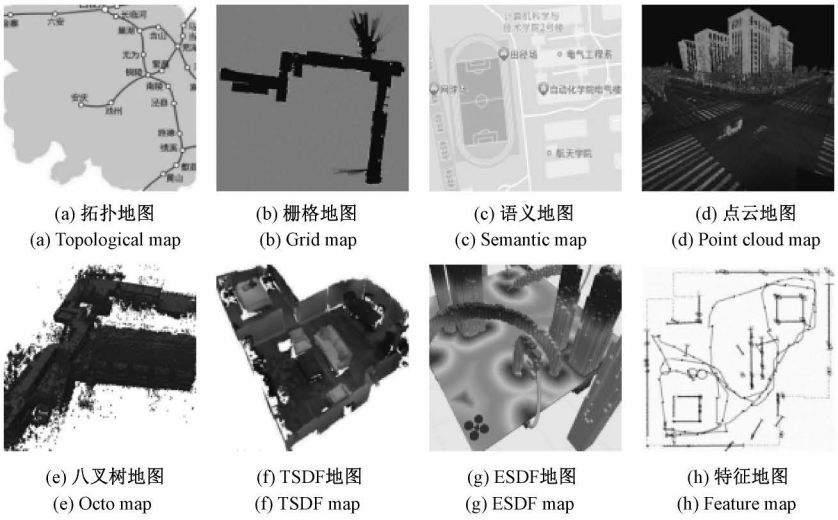
根据前端不同的处理方法以及不同的任务需求,需要构建出与相对应形式与复杂程度的地图,不仅能准确描述环境特征,还要在保证精度的同时,减少地图的复杂度。根据不同维度,地图表示形式可以分为二维和三维。

二维地图分为几何地图、栅格地图与拓扑地图[30]。几何地图使用稀疏的点、线段、曲线等路标来描述场景环境[31]。栅格地图将环境均等地划分为若干个栅格,通过概率值表示每个栅格是否被物体占据,每个栅格单元具有占据、空闲和未知 3 种状态,从而区分可通过区域与障碍物区域。拓扑地图使用节点与节点之间的连接线组成拓扑结构图来表示场景,其中节点为实际环境中的地点,节点之间的连接线表示不同地点间的关系。

三维地图中,点云地图是使用范围最广的地图。虽然点云地图保留了原始环境的详细信息,然而点云地图规模一般较大,很多任务无需的细节占用了大量空间。基于八叉树结构可构建出三维栅格地图,又称为八叉树地图(Octomap)[32]。相比于二维栅格地图,八叉树地图对环境更有效描述,歧义性更少,相比于点云地图大大节省空间,但是相应的计算复杂度要大,因此对于实时路径搜索与规划存在较大难度。此外,根据具体任务需求以及前端处理方式不同,不同类型的地图还有特征地图、欧氏符号距离场(Euclidean signed distance functions, ESDF)地图、截断符号距离场(truncated signed distance field, TSDF)地图、语义地图[33]等。各种不同类型的地图如图 2 所示。

3 结合深度学习的 VSLAM 方法

深度学习与 VSLAM 结合已是大势所趋,通常的做法是利用深度学习代替 VSLAM 系统的某一个或者某一些模块和步骤。现根据 VSLAM 的各个基本结构与深度学



(a) 拓扑地图 (a) Topological map　(b) 栅格地图 (b) Grid map　(c) 语义地图 (c) Semantic map　(d) 点云地图 (d) Point cloud map

(e) 八叉树地图 (e) Octo map　(f) TSDF地图 (f) TSDF map　(g) ESDF地图 (g) ESDF map　(h) 特征地图 (h) Feature map

图 2　各种地图类型

Fig. 2　Various map types

习结合的情况,对 VO、回环检测、建图 3 个部分中的深度学习方法进行总结。

3.1　基于深度学习的 VO

VO 是通过跟踪不同时刻相邻两帧图像进行帧间估计得到相机位姿与深度图等,特征检测与提取过程是其中重要步骤之一。传统手工特征非常依赖于设计者的先验知识,人为调整参数耗时耗力,难以利用大量数据。深度学习具有强大的特征提取能力,通过具有高度表达能力的深度神经网络作为通用拟合器,自动提取到任务所需特征。这一特性使得深度学习模型可以很好地适应各种环境,特别是对于动态环境、复杂场景、运动模糊等手动建模困难的情况。

1)特征提取与特征匹配

(1)SuperPoint

Detone 等[34]提出一种适用于多视角几何的自监督框架,引入单应自适应方法提取特征点并生成特征描述子。用预训练好的 MagicPoint 网络提取角点,生成输入图像的伪标签。但是当视角发生改变,其表现就容易变差。单应自适应对图像进行多次不同尺度、角度变换,使得在不同视角与尺度都能观测到特征点。传统的特征提取方法通常是先检测出特征点,再计算描述符,缺乏在两个任务之间共享计算与表征的能力。SuperPoint 使用一个类似视觉几何组(visual geometry group,VGG)的共享编码器降低图像维度,而后分成特征点解码与描述子解码学习各自权值。特征点解码输出每个像素的概率图。描述子解码学习半稠密描述子并进行双三次插值与 L2 标准化得到单位长度的描述子。该方法能在原始图像中提取到像素级的特征点位置与描述子,且两者同时进行训练与输出。虽然 SuperPoint 有着不错的精度,但存在耗时太长的问题。MobileSP[35]在原本框架的基础上进行了改进,提出了局部共享检测、描述编码架构,基于预排序的非极大值抑制(non-maximum suppression, NMS)引擎以及软硬件混合的计算技术,通过算法与硬件协同设计,保证原有精度的同时,大大提高了运行速度。为减少 SuperPoint 模型的参数量,文献[36]采用 MobileNetV2 与 GhostNet 轻量化网络替换原始类似 VGG 的架构,而且在描述子计算中没有采用三次插值,真实环境下该系统也能实时运行。

(2)Guided Feature Selecion

Xue 等[37]提出了一种基于深度卷积双分支递归神经网络的端到端 VO 体系结构来引导特征选择。双分支递归网络分别学习旋转与平移,通过卷积神经网络进行特征表示,递归神经网络进行图像序列推理。为增强特征选择能力,引入一种有效的上下文感知引导机制,允许网络自适应地抑制无用信息,强制每个分支提取特定运动模式的相关信息。

(3)GCNv2

GCNv2[38]是一种针对 3D 投影几何的特征提取算法。GCNv2 建立在几何对应网络(geometric correspondence network, GCN)[39]的基础上,设计了与 ORB 相同格式的二进制描述子,使其能够在 ORB-SLAM2 等系统轻松替换 ORB 的特征提取部分。原始的 GCN 网络结构由两个部分组成,ResNet-50 为主干的全卷积网络(fully convolutional networks, FCN)与一个双向卷积网络组成。FCN 进行稠密特征提取,双向卷积网络确定关键点位置。存在的问题是网络架构较大,对于资源有限的硬件平台存在局限性,而且双向网络结构要求两帧或多帧之间同时完成匹配,极大增加了系统的复杂度。借鉴

SuperPoint 只对单帧图像进行检测的思想，GCNv2 对单幅低分辨率的原始图像进行独立预测，参数更少，尺度更小，大大简化了网络结构。而且特征向量二进制化与二进制描述子结合使用，使得 GCNv2 在保证与 GCN 相当精度的同时，大大提高了计算效率，实时性较好。同样基于 GCN 进行改进，王启来等[40]提出了一种更为轻量级的 GCN-L 网络，能提取到有着更均匀的空间分布的关键点，具备较强的鲁棒性与实时性。

(4) DF-SLAM

大部分深度学习的方法严重依赖训练所用的数据集，无法很好地适应位置环境，有时牺牲效率以获得更高的精度，降低了系统的实用性。DF-SLAM[41]提出采用具有 3 重网络结构的 TFeat，经过两个卷积层，第 1 个卷积层后进行最大池化，减少参数进一步加快网络传递，第 2 个卷积层后为全连接层，输出一个 128 维的描述符，最后 L2 标准化。借鉴 HardNet 的负样本采样策略，通过 L2 成对距离矩阵选择同一批次中最接近的非匹配图像块，最大化最近正负样本之间的距离。通过浅层神经网络获得的局部特征描述子替代传统手工特征描述子，TFeat 与 HardNet 的融合不仅提高了效率和稳定性，而且在光照强烈变化的场景鲁棒性较强，可移植性与实时性具有更大优势。

(5) LIFT-SLAM

LIFT-SLAM[42]是一种基于特征的深度学习单目 VSLAM 系统，使用深度神经网络提取 ORB-SLAM 中的特征。学习不变特征变换(learned invariant feature transform, LIFT)适用于图像块，网络主要包括 3 个基于卷积神经网络(convolutional neural network, CNN)的模块：检测模块、方位估计模块与描述模块，以端到端的监督式方法分别实现局部特征检测、方位估计与描述。某一图像块作为输入，检测网络将给出该图像块的分数图，在此分数图上执行 softargmax 操作找到潜藏特征点位置，通过方位估计器以该位置为中心进行裁剪并预测方位。根据所估计的方位旋转图像块，最终描述网络从旋转的图像块输出特征向量。原始 LIFT 是在不同于 VO 数据集的摄影旅游图像集上训练。因此，LIFT-SLAM 使用迁移学习对 VO 数据集的网络进行微调，提高了跨数据集的性能。

(6) YOLO

只需看一次(you only look once，YOLO)系列网络是神经网络中最经典的目标检测网络，能在保证较高精度的同时，对视频或图像进行实时检测。文献[43]用 YOLOv1 检测目标，Mask-RCNN 进行实例分割。当场景中检测到有人存在，才会切换到实例分割，删除属于人的关键点。但是仅考虑了人这一移动物体。文献[44]在 LeGO-LOAM 基础上，采用 YOLOv3 进行多目标识别，并在重建出的室内 3D 地图上描述多目标的位置。相比于 YOLOv3，YOLOv4 的主干网络架构从 Darknet53 更换为 CSP Darknet53，引入了空间金字塔池化(spatial pyramid pooling, SPP)，更好地提取到不同尺度的特征，如图 3 的 YOLO 网络结构图所示。文献[45-47]均采用 YOLOv4 作为目标检测网络，文献[45]应用对极几何约束与光流约束过滤动态特征点以更好地完成特征匹配，大幅度提高在动态环境下的定位精度。文献[46]将动态对象检测与动态对象概率(dynamic object probability, DOP)模型结合。文献[47]利用从物体中提取的 ORB 特征来更新输入图像特征。文献[48-49]对 YOLO 系列的网络进行了不同的改进，对动态场景 VO 的精度性能都有不错提升。

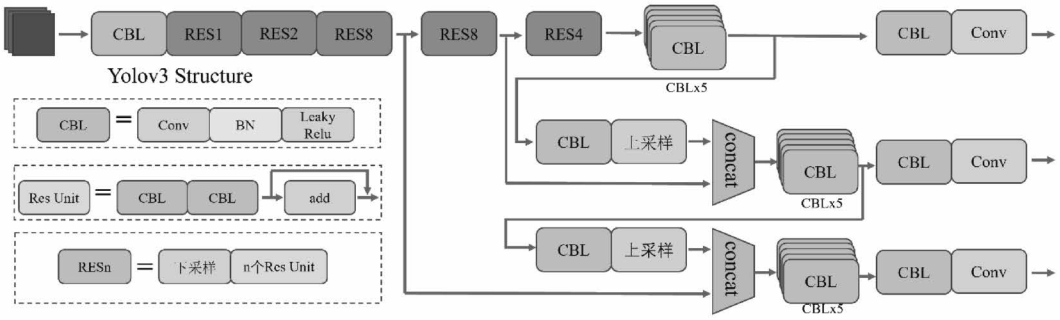
(7) Attention mechanism

VO 问题更注重的是两幅图像之间的几何特征信息对比，通过在深度神经网络中引入有效的注意力机制，增强网络架构的表征能力，提高对特征区域的权重，从而更易于学习到任务所需的重要特征信息。Attention-SLAM[50]模拟人类导航模式，提出了 SalNavNet 来预测图像中的显著区域，并以预测区域的重要性大小作为权重。还引入了一个自适应的指数移动平均(exponential moving average, EMA)模块，减轻模型的中心偏置问题。采用了一种加权的 BA 方法，更加关注显著区域中提取到的特征，有效减小轨迹误差。张再腾等[51]提出了一种类似于 VGGNet 结构的 CNN 网络，仅包含 9 层卷积层。受到 CBAM[52]的思想启发，提出了一种基于卷积的注意力模块，该方法可以实现实时 VO 的估计，具备较高精度的同时，拥有更低的网络复杂度。俎晨洋等[53]提出了一种基于注意力机制的特征点匹配网络，将图神经网络与注意力机制结合以得到每个特征点的匹配描述子，相比于传统的算法不仅能处理视角不稳定的情况，精度也显著提升。存在的问题就是运行耗时较大，难以实时应用。特征提取与特征检测方法总结如表 2 所示。

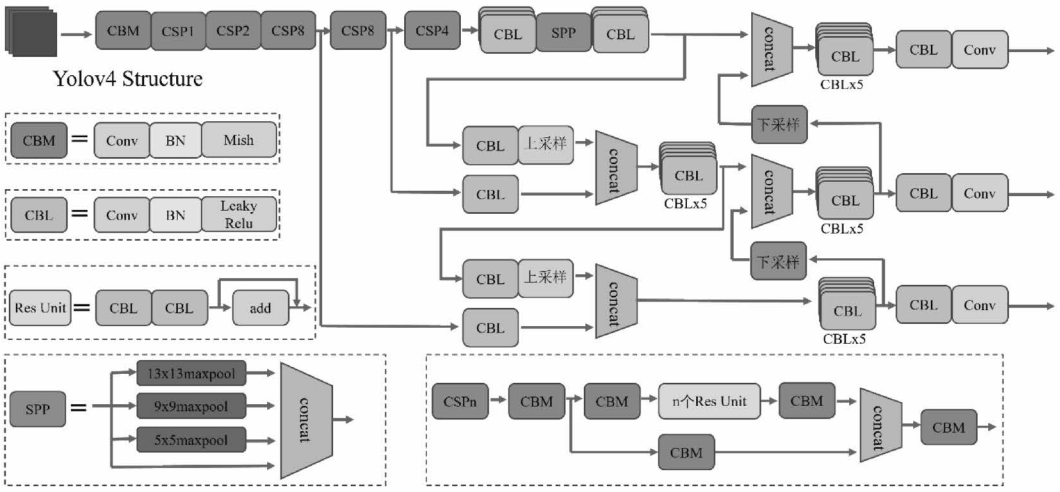
2) 深度估计与位姿估计

(1) SFMLearner

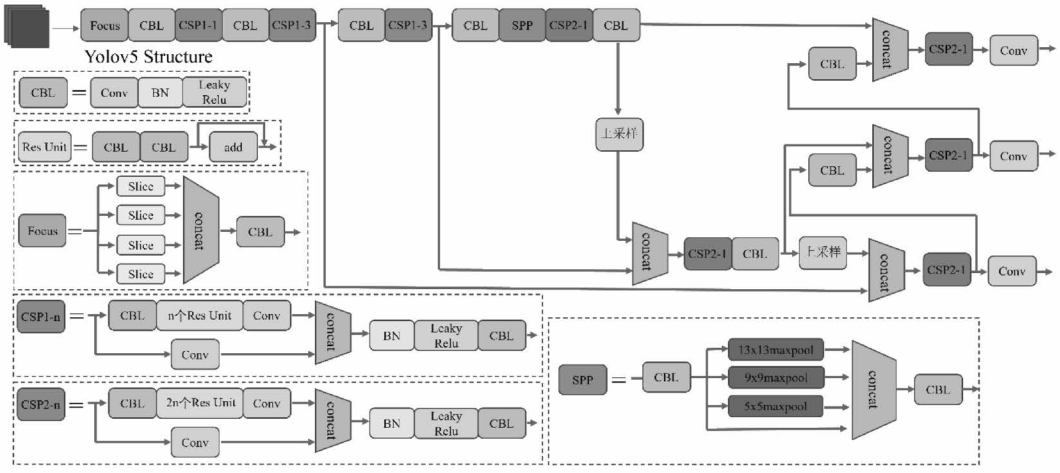
Zhou 等[54]提出了一种完全无监督的端到端网络框架 SFMLearner，包括深度、位姿与可解释性 3 个部分。将不同相机位姿的视图合成新的图像作为监督。合成过程用 CNN 以完全可微的双线性插值方式实现。单视图深度预测采用 DispNet 架构，输出目标视图的深度图。位姿网络与可解释性预测网络均以目标视图以及附近的源视图为输入，位姿网络输出目标视图与每个源视图之间的 6 个自由度的相对位姿。对于动态场景、有遮挡以及非朗伯表面的情况，容易出现梯度破坏问题而导致网络训练失败。可解释性预测网络与位姿网络共用前 5 个卷积层，输出多尺度目标图像与源图像对中每个像素的可解释性预掩码，表明每个目标像素建模成功的概率。



(a) YOLOv3网络结构图
(a) The network structure of YOLOv3



(b) YOLOv4网络结构图
(b) The network structure of YOLOv4



(c) YOLOv5网络结构图
(c) The network structure of YOLOv5

图 3　YOLO 网络结构图

Fig. 3　The network structure of YOLO

与 SFMLeaner 直接生成深度图不同，Godard 等[55]认为仅用图像重建的结果会产生较差的深度图，提出了一个无监督单目视差估计网络，将深度估计视为图像重建问题，受 DispNet 的启发，最小化图像重建损失训练网络生成视差图，再通过视差重建像素深度。同时提出一种新的损失函数，加入左右一致性约束，提高了左右视差图一致性与视差估计精度。

表 2 特征提取与特征检测方法总结

Table 2 A summary of the feature extraction and feature detection methods

年份	文献	贡献	局限
2018 年	[34]	单一网络完成特征点提取与描述子生成 自动标注图像伪真实标签	没有实现从实际数据中学习 特征点重复率较高
2018 年	[37]	引入上下文感知引导特征选择机制 双分支递归网络分别学习运动的旋转与平移	没有考虑实时性问题
2019 年	[38]	设计与 ORB 相同格式的二进制描述子 训练过程加入特征向量二值化加速	预测投影几何而不是特征匹配 没有对室内和动态环境进行测试
2019 年	[43]	YOLOv1 进行检测对象,Mask-RCNN 进行实例分割	仅考虑了人这一移动物体
2019 年	[41]	学习的局部特征描述子代替传统手工描述子 光照变化下鲁棒性较强,且能保证系统实时性	对比算法较少
2021 年	[36]	改进 SuperPoint 的类 VGG 架构,能做到实时运行	描述符匹配精度在特定情况下较低
2021 年	[40]	轻量级的网络架构	输入图像尺寸固定
2021 年	[42]	深度学习特征描述子结合传统几何 VSLAM 评估了迁移学习与微调对系统性能的影响	无法检测到闭环时性能将变得很差 运行速度慢,无法做到实时
2021 年	[44]	YOLOv3 进行目标识别	仅实现单目标识别
2021 年	[45]	3 种约束共同辅助剔除动态特征点,且能兼顾精度与实时性	对比的算法较少
2021 年	[48]	在常见的路面情况能较好地提高精度	对于更复杂的路面情况有较大挑战
2021 年	[50]	SalNavNet 预测图像显著区域	对运动模糊和快速旋转表现不佳
2021 年	[51]	具备较高精度的同时,拥有更低的网络复杂度。	对于动态场景会有较大误差
2022 年	[53]	不仅能处理视角不稳定的情况,精度也显著提升	运行耗时较大,难以实时应用
2022 年	[49]	在特定条件下准确率较高	局限室内,且可识别物体较少时容易误匹配

(2) GeoNet

GeoNet[56] 包含 DepthNet、PoseNet 与 ResFlowNet 共 3 个子网络。DepthNet 与 PoseNet 组成刚性重建器,用于静态场景,根据预测得到的深度与位姿融合推导出全局刚性流。ResFlowNet 组成非刚性运动定位器作为动态对象的补偿,不仅可以纠正动态对象的错误预测,还能改善前一阶段的预测结果。将学习到的残差非刚性流与刚性流结合得到最终的流预测。GeoNet 是一种联合学习单目深度、光流与相机位姿的无监督学习框架,采用分而治之的策略,分别学习刚性流与物体运动,对于光度一致性假设无法完美处理遮挡与非朗伯表面的问题,引入了自适应几何一致性损失,应用前后一致性检查的同时,也考虑了左右一致性损失。

(3) GANVO

GANVO[57] 网络首次在单目 VO 中使用生成对抗网络(generative adversarial network, GAN)与循环无监督学习联合估计姿态与深度图,由目标图像与源图像组成原始的 RGB 图像序列堆叠作为深度网络输入。深度网络包含生成网络、鉴别网络与编码网络。编码网络将输入目标图像映射成一个特征向量,生成网络再将其映射成深度图。位姿估计网络对原始图像经过 CNN 与两个长短期记忆(long short term memory, LSTM)模块输出 6 个自由度的位姿参数,包括平移与旋转。视图重建模块通过深度图、位姿参数以及源图像的颜色值合成目标图像。鉴别网络对合成的目标图像与原始目标图像博弈对抗,而不是直接比较生成网络输出的深度图。

SGANVO[58] 则是首次使用堆叠生成对抗性网络进行自我运动与深度估计。每层 GANs 大体组成也和 GANVO 相同,包括生成网络与鉴别器,生成网络由深度估计网络,自我运动估计网络以及视图重建网络组成,同时加入了 卷积长短时记忆网络(convolutional long short term memory network, ConvLSTM)在各层之间连接。

(4) D3VO

D3VO[59] 提出一种自监督单目深度估计网络,使用视频序列作为输入,通过 DepthNet 预测深度,PoseNet 学习相邻两帧的位姿信息,通过最小化时间立体图像与静态立体图像之间的光度重投影误差来桥接两个网络,从而在训练过程中融入时间信息。针对训练图像对之间光照不一致的问题,网络在训练过程中预测光度变换参数,利用该参数将源图像与目标图像对齐到

相似光照条件。然而单纯建模仿射光度变换参数并不足以囊括所有光度恒定假设失败的情况,因此需要对源图像像素建模,根据每个像素的光度值概率分布来预测得到源图像的光度不确定性图,这不仅提高了深度估计的准确性,也为后续提供了一个光度残差的可学习加权函数。

(5)DDL-SLAM

DDL-SLAM[60]是一种结合深度学习的新型 RGB-D SLAM 框架,以减少移动物体对相机姿态估计的影响。采用 DUNET 进行语义分割,并结合多视图几何作为预处理,以过滤掉与动态目标相关的特征点,再得到不含动态物体的合成 RGB 帧及相对应的深度图。在高动态场景中的准确度较高,但是实时性难以保证,且构造出的八叉树地图未利用到语义信息。

(6)VIOLearner

Shamwell 等[61]提出了一种无监督的在线纠错模块以及深度神经网络(visual-inertial-odometry learner, VIOLearner),学习惯性测量单元(inertial measurement unit, IMU)的测量结果并生成多步轨迹估计,通过对多个空间尺度上的雅可比矩阵进行卷积处理,根据像素坐标空间网格的缩放图像投影误差的 Jacobian 进行在线校正。VIOLearner 可以使用视图合成方法在相机帧中生成相机姿态变化,其训练的基础是目标图像和重建目标图像之间的欧几里得损失,重建目标图像根据学习到的 3D 仿射变换确定的位置,对源图像中的像素进行采样生成。

(7)WF-SLAM

WF-SLAM[62]采用 Mask-RCNN[63]网络对原始输入的 RGB-D 图像进行语义分割生成语义掩码,对动态物体进行分类。但是对于语义上静态而实际上动态的物体难以检测到,因此采用语义与对极几何约束紧密耦合使用,能更好地剔除动态特征点。同时还提出一种对动静态特征点权重的加权优化算法,联合优化特征点与位姿的权重,大大提高了在动态场景下的定位精度与准确性。但是该系统无法实时运行。

(8)DynaSLAM II

和 DynaSLAM[64]一样,DynaSLAM II[65]也是采用 Mask-RCNN[64]进行语义分割作为实例语义先验。对于单目和立体图像,DynaSLAM 用 Mask-RCNN 进行语义分割,划分出先验动态像素。在 RGB-D 图像中,则结合了多视图几何进行动态物体检测。在完成相机定位和全动态物体检测后,利用先前视图的静态信息重建出当前帧的遮挡背景。与 DynaSLAM 忽视动态对象信息不同,DynaSLAM II 则是利用语义结果对场景中的不同动态对象进行多目标跟踪,最终实验也证明了跟踪动态对象不仅为场景理解提供了丰富的信息,完善所构建的地图细节,也有助于提高定位、轨迹精度。

与大多数采用 Mask-RCNN 或者 YOLO 进行分割与检测的工作不同,文献[66]采用了较为新颖的 YOLACT++[67]网络进行图像的实例分割,结合 MaskFlownet 光流预测网络的光流场检测动态区域掩码,并过滤掉区域内的动态特征点以此提高定位精度。文献[68]采用 YOLACT 网络分割先验动态对象,并采用 MobileNetV3 分割车道区域,结合多视图几何、区域特征与相对距离来剔除动态对象的动态特征点。

(9)CubeSLAM

不同于其他需要先验信息的方法,CubeSLAM[69]无需先验的对象模型,将二维与三维的对象检测与位姿估计相结合。CubeSLAM 使用 YOLOv2 用于室内,室外则采用 MS-CNN 来进行二维对象检测,并生成二维边界框。根据基于视点(view point, VP)的二维边界框生成高质量的立方体框,并用不同的代价函数对其进行评分,VP 是投影透视图后平行线的交点。为了共同优化相机、物体和点的位姿,提出了带有新物体检测的多视图 BA。深度估计与位姿估计方法总结如表 3 所示。

3)关键帧选择

视频由一系列以一定时间间隔出现的图像帧构成,大部分相邻帧的重复率通常较高。当运动很慢的时候,序列中的图像帧存在太多冗余信息,给 VSLAM 任务增加了很多不必要的计算量。根据相对位移与追踪特征点数量,从中选择一帧关键帧代表局部帧以减少待优化的帧数。但是对每帧都执行特征点提取会大大增加计算量,且容易出现误选。

文献[70]提出一种专门为关键帧选择设计的深度网络,以端到端的方式同时学习关键帧选择和视觉里程任务。这也是第一次在同一个深度框架内联合优化这两个互补任务。关键帧选择网络学习观测图像和关键帧之间的联合视觉和几何相似性。如果二者相似性低于某一个阈值,该观测图像将被视为新的关键帧,添加到关键帧集中。关键帧选择网络包含视觉流与几何流,两个流的网络结构相同但不共享网络参数。基本结构来源于 ResNet18,提取到的视觉特征与几何特征通过跨模态注意力模块进行融合,自适应地将视觉和几何相似性组合生成最终的相似性分数。文献[71]则是受到 LSTM 成功应用于结构化预测问题的启发,考虑了视频图像序列的前向与后向信息,提出了一种 vsLSTM 网络。LSTM 能对长距离依赖关系进行建模。整体网络由两个 LSTM 与一个多层感知机(multilayer perceptron, MLP)网络构成,此外,还加入了行列式点过程(determinantal point processes, DPP)对网络进行增强,确保关键帧集的差异性大。vsLSTM 用于预测远近不同帧的重要程度,DPP 输出不同的候选关键帧子集为关键帧集的概率,能在有充足标注样本的数据集中获得很好的结果。

表 3 深度估计与位姿估计方法总结

Table 3 A summary of the depth estimation and pose estimation methods

年份	文献	贡献	局限
2017 年	[54]	完全无监督、端到端方式 3 个独立网络分别估计深度、位姿、解释性	对于动态场景、有遮挡、非朗伯表面情况容易失败
2017 年	[55]	端到端无监督 引入左右视差一致性约束	遮挡区域边界有伪影 单视图数据集无法用于训练
2018 年	[56]	引入前后一致性检查与左右一致性约束 分别学习物体运动与刚性流	在一些数据集中表现不是很好
2019 年	[57]	首次使用对抗网络进行深度、位姿估计 无监督学习,且不需要严格的参数调整	在使用多数据集时的表现不如其他方法
2019 年	[58]	首次使用堆叠对抗网络联合自我运动预测与深度估计 学习到一个可捕捉时间动态特征的递归表示	没有回环检测,有一定的漂移
2019 年	[69]	二维三维对象检测生成高质量长方体 多视图 BA 调整	对于一些场景效果较差
2020 年	[59]	时间信息融入到训练过程中 综合深度、位姿、光度不确定性	泛化能力不是很好,在不同数据集训练和测试表现差距较大
2020 年	[60]	语义分割与多视图几何组合过滤动态特征点	实时性有待提高
2020 年	[61]	无监督学习 能在线校正位姿变化,提高 VO 精度	仅接受单个源图像与目标图像,无法执行任何类型的 BA 调整
2021 年	[65]	利用动态对象先验语义信息进行多目标跟踪	对比结果较少
2021 年	[68]	YOLACT 分割先验动态对象 MobileNetV3 分割车道区域	提高了系统复杂度和计算度
2022 年	[66]	YOLACT++实例分割,MaskFlownet 光流预测	网络推理时间较长
2022 年	[62]	语义信息与对极几何约束结合 动静态特征点加权优化	运行时间较长

在考虑图像质量与语义信息的基础上,文献[72]提出了一种关键帧选择算法。组合基于拉普拉斯能量[73]计算的模糊度分数与基于图像像素亮度的亮度分数作为图像质量标准。语义分数基于语义分割,通过 MiniNet 完成其计算。受到如 ERFNet[74]、Deeplab-v3[75]等轻量型 CNN 架构启发,MiniNet 考虑到小型的硬件平台,其中的下采样块与上采样块之间串联两个卷积分支,且具有类似 Unet 的跳层连接。在此图像上计算目标类的像素比率,即为语义分数。综合图像质量标准与语义分数确定最佳的图像作为关键帧。对于一些耗时较长的语义分割网络,假如按照顺序分割每个关键帧,在跟踪过程中当前帧可能无法获得新的语义信息导致跟踪失败。为此 Liu 等[76]提出了基于 ORB-SLAM3 构建的实时视觉动态 SLAM 算法 RDS-SAM,采用一种基于语义分割的关键帧选择策略,使用双向模型而不是顺序模型,缩短了语义延迟,尽可能获取最新的语义信息,适配不同处理速度的分割方法。关键帧选择方法总结如表 4 所示。关键帧选择的大体流程与回环检测的较为相似,区别仅在于其更关注于两帧之间的差异程度,而回环检测更关注相似程度,可以参考图 4 回环检测的流程图。

3.2 基于深度学习的回环检测

回环检测主要判断是否曾经到达过某个位置,对于 SLAM 的重定位与消除累计误差有着重要的作用。大体的思路流程为:提取图像帧的特征信息,并衡量当前帧与以前关键帧的相似程度,当相似度达到一定程度时,就认为检测到闭环。

Gao 等[29,77]提出使用 SDA 来解决 VSLAM 中的回环检测问题。SDA 是一种无监督深度神经网络,由多个去噪自动编码器(denoise autoencoder, DA)组成,每个 DA 层输出作为下一层的输入。SDA 最终输出的特征响应与中间层的深度特征通过相似性矩阵计算相似性分数,在当前帧与以前关键帧两者之间的相似性分数超过给定阈值时,则检测到回环。张云洲等[78]则是采用了无监督栈式卷积自编码器(convolutional autoencoder, CAEs))提取

表 4　关键帧选择方法总结

Table 4　A summary of the selection of keyframe methods

年份	文献	贡献	局限
2016 年	[71]	利用了前向后向信息 展示了如何应对标注不足的情况	在场景变化较快时表现较差
2019 年	[70]	首次在一个框架联合优化关键帧选择与 VO 能对关键帧检测、更新、管理、定位新帧	对于动态物体的预测存在偏差
2019 年	[72]	结合语义信息进行关键帧选择,运行速度快,能在 CPU 运行	分割的每个个体之间没有建立联系
2021 年	[76]	不需要等待语义信息,可以适配不同处理速度的分割算法	未部署到实际应用系统中测试

原始图像特征用于回环检测,通过夹角余弦的方法计算两帧图像的相似程度。

Chen 等[79]提出一种基于多尺度深度特征融合的回环检测方法,分为特征提取、特征融合与决策 3 个部分。当前帧与以前的关键帧对应两个分支,在特征提取与特征融合的结构上是相同的。特征提取使用预训练好的 AlexNet[80]提取深度特征,对于图像尺寸不符合 AlexNet 固定输入图像尺寸时,需要进行切割裁剪或者压缩,导致图像信息缺失。因此特征融合使用 SPP[81]按不同比例将特征图划分成不同大小的块,将提取到的不同特征组合一起,从而实现多尺度深度特征融合。决策层根据两个分支的输出特征向量进行相似度计算,检测是否达到闭环。

文献[82]将超级字典与深度学习结合进行闭环检测,主要包含深度 CNN 分类器、自动编码器、超级字典与相似性函数 4 个部分。深度 CNN 分类器基于 VGG16[83]架构,将图像对象分类成静态和动态对象,并提取动态对象的深度特征。使用超级字典与 BoW 字典协同,加快特征匹配过程,减小丢失循环的概率。超级字典仅保存总帧数的 10% 的关键帧信息,BoW 字典保存所有帧的信息。CNN 分类器最终输出的特征向量输入到自动编码器,这是一种无监督网络,基于均方误差计算重构误差,检测当前场景是否已经被访问过。由于许多场景有着很多相似之处,已经访问过该场景并不意味着检测到闭环。相似性函数结合超级字典与 BoW 字典计算两帧特征的相似性以检测闭环。

与大多数直接使用 CNN 输出提取到的特征不同,Zhang 等[84]对 CNN 特征进行预处理,使用 overfeat[85]网络生成整个图像描述符,进行主成分分析(principal component analysis, PCA)、白化等预处理步骤提高表征图像的能力,再进行后续的操作。应用对特征后处理的想法,除了白化处理,Wang 等[86]还引入可选择的压缩比对特征信息进行压缩,消除动态环境冗余信息,同时加入时间相似性约束降低非闭环相邻图像间的相似程度。

杨馨竹等[87]借鉴 ResNet,基于 CSP-DarkNet 网络进行改进,实现快速回环检测,在网络最后一层引入 NetVLAD 池化层,从而更好地利用图像的局部空间信息。同样基于残差网络,占浩等[88]通过预训练好的 ResNet18 网络对图像序列进行全局特征提取,将相应的特征组合作为当前帧的特征,同时设计了一种双层查询方法以获得相似度最高的图像帧,并对其进行时间与空间一致性检验,以保证得到准确的回环。

针对回环检测算法效率较低而无法应用于轻量级设备的问题,郭烈等[89]采用 SqueezeNet 网络提取原始图像的特征向量,对特征向量计算余弦相似度,判断是否出现回环。SqueezeNet 网络包含多个 Fire 模块,结构简单轻量,在简化网络复杂度的前提下,同时达到一般网络的精度,大大减少了计算时间,十分适合应用在小型移动机器人等轻量级设备上。考虑到系统实时性的要求,赵浩苏等[90]使用 EdgeBoxes 方法提取图像的标志区域,并对 ResNet18 改进,使用 CNN 提取出标志区域的二进制特征完成快速检索,通过浮点特征完成选择最优匹配。通过增量式字典进一步计算不同图像的相似性。相比于浮点运算,二进制计算速度非常快,能很好地兼顾了实时性。

深度学习与回环检测的结合大多集中在特征提取,根据学习到的不同尺度特征信息、语义信息、动态信息等,结合相似度函数,计算两帧之间的相似程度,通过是否达到一个设置的阈值来确定是否检测到回环。同时光照敏感性与动态场景也都会影响到系统的鲁棒性,在设计网络结构或者相似度函数时也需要加以考虑,图 4 展示了回环检测的大体流程。结合深度学习的回环检测的方法总结如表 5 所示。表中列举了近年来回环检测中深度学习方法的主要贡献及其存在的局限性。

3.3　基于深度学习的地图构建

建图是为了根据任务所需,构建出一个能够描述周遭环境的地图,具有以下作用:1) 可以从复杂的周围环境中,为使用者提供能够理解的地图参考;2) 可以根据构建的地图信息完成目标任务,如路径规划、定位、导航等;3) 作为先验模型为全局定位提供参考。

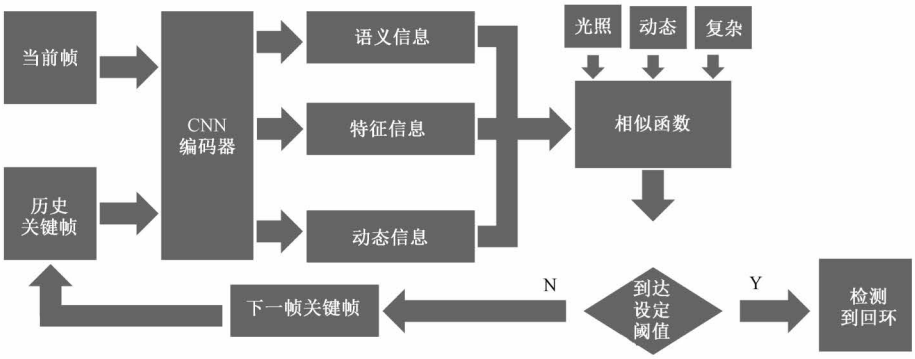


图 4 回环检测流程图

Fig. 4 The flow chart of loop closure detection

表 5 回环检测方法总结

Table 5 A summary of the loop closure detection method

年份	文献	贡献	局限
2017 年	[29]	应用堆叠自动编码器 探讨了学习速率、稀疏约束、网络维度等超参数影响	仅在高召回率时精度较高 训练网络时间长,仅在同一数据集进行训练
2017 年	[77]	对 CNN 提取到的特征进行 PCA、白化预处理	没有额外约束,仅在高召回率时精度较高
2019 年	[79]	空间金字塔处理不同尺寸图像时不会丢失信息 光照鲁棒性强	结果对比的算法太少
2020 年	[90]	很好兼顾了实时性与准确性	没有集成到完整系统中,具体效果有待测试
2020 年	[86]	压缩特征冗余信息 引入时间相似性约束	不好集成到其他系统 没有说明如何选择确切的压缩比
2020 年	[82]	神经网络与超级字典结合 动态对象、不同角度与遮挡的情况鲁棒性强	仅与传统字典的方法做对比
2021 年	[89]	网络轻量,保证精度的同时运行时间较短 适合集成到小型轻量级设备	对系统鲁棒性没有做过多改进
2021 年	[88]	设计双层查询方法,时间与空间一致性检验 比较了不同序列长度与采样数量的影响	仅与传统词袋方法对比
2022 年	[87]	充分利用图像中的局部空间信息 能保证实时性	性能提升幅度不大

在 VSLAM 中,构建完整地图通常不是一个独立的线程,需要前端 VO、回环检测等方法,生成地图所需的环境信息。根据获取到的位姿信息、像素深度、语义信息等,将其映射到二维图像或三维点云中,构建出与之对应的地图。根据场景表达方式的不同,将地图表示分为几何建图、语义建图以及广义建图。

1)几何建图

几何地图主要检测场景的形状和结构的描述,场景表示通常包括深度、体素以及网格等。

(1)深度表示

深度是理解场景几何和结构信息的最常用的表达方式。准确的深度估计有助于 VSLAM 的绝对尺度恢复。通常的做法是将深度估计作为输入图像的映射函数,利用带有深度标签的大量数据,训练深度神经网络预测源图像的每个像素深度。虽然比传统基于结构的方法准确度更高,但是非常依赖模型训练好坏,且在缺乏标签数据的情况下,鲁棒性和泛化能力都会严重下降。在无监督的研究中,深度预测往往重新定义为视图合成的问题。

Godard 等[55]以光度一致性损失作为自监督信号。通过左右视差图,最小化合成图像与真实图像的损失,引入了空间一致性,利用视差重建像素深度。Zhou 等[54]还将时间一致性作为监督信号,从源时间帧合成了目标时间帧中的图像作为监督信号。在完成深

度估计同时,还实现了自运动的恢复。在深度与自运动的结构下,文献[56, 59, 70, 91-96]在不同程度上进行扩展,在深度预测与自运动估计中获得了不错的效果。文献[92, 94-95, 97]针对损失函数,在网络框架中添加了各种不同的约束,以提高网络性能。

(2)体素表示

对于三维几何体,基于体素的表达方式最为常见,类似于二维图像中的像素。SurfaceNet[98]是一种用于多视图立体视觉的端到端学习框架。该框架可以自动学习表面结构的照片一致性(photo-consistency)和几何关系,它采用一组图像及其相应的相机参数作为输入,并直接推断 3D 模型。将相机参数与图像信息一起编码,提出一种彩色体素立方体(colored voxel cube, CVC)的表示方式,通过投影每个立方体的体素到图像中且每个体素存储 RGB 的值,从而将图像信息转换到 CVC 中。SurfaceNet 是一个完全 3D 的卷积网络,以一对不同视点的 CVC 作为输入,预测每个体素的表面置信度从而重建出 2D 图像的表面。还将该网络推至多视图,获得不错的结果。RayNet[99]将 CNN 与马尔科夫随机场(Markov random fields, MRF)结合,CNN 用于学习视图不变特征表示,MRF 能模拟透视几何、遮挡等物理过程。提取视图不变特征的同时添加几何约束,从而重建场景几何。

基于体素的表达存在的最大局限性就是计算量巨大,在高分辨率的地图中通常难以保证实时性。Tatarchenko 等[100]提出了一种深度卷积解码器架构,该架构可以通过使用八叉树表示,以计算和内存高效的方式生成 3D 模型。与作用于常规体素网格的标准解码器相比,该架构没有以前立方体的复杂性。因此在有限的内存下有着更高分辨率的输出。

(3)网格表示

基于网格的表示对三维模型的底层结构进行编码,如边、顶点、面等。Pixel2Mesh[101]是一种端到端的深度学习架构,在基于级联图的卷积神经网络中表示 3D 网格,通过原始图像中提取到的特征,将初始的椭球体逐渐变形成正确的几何形状。Scan2Mesh[102]以三维对象扫描点云数据为输入,通过卷积神经网络和基于图神经网络架构的组合,实现预测和实际数据点之间一对一的离散映射,生成紧凑的网格表示。但是这两种方法都只能重构单个对象,且结构简单的模型。文献[103]以 2.5D 的三角网格作为场景几何的紧凑表示,通过神经网络直接从源图像中预测平面的顶点坐标,并将顶点深度作为自由变量进行优化。

2)语义建图

语义建图将语义概念与环境几何形状结构联系一起,构造出的地图不仅包含场景几何信息,同时还融入了更高层次的语义信息。根据场景分割类型分为语义分割、实例分割与全景分割 3 个部分。图 5 中展示了不同分割方式的结果对比。

(a) 原始图像
(a) Original image

(b) 语义分割
(b)Semantic segmentation

(c) 实例分割
(c) Instance segmentation

(d) 全景分割
(d) Panoptic segmentation

图 5　三种分割方式结果对比

Fig. 5　The result comparison of three segment methods

(1)语义分割

语义分割通过对图像中的每个像素点关联对应的语义类别标签进行图像像素点分类。SemanticFusion[104]是将深度神经网络得到的语义分割标签与 SLAM 中密集场景几何结合的早期作品之一。从反卷积语义分割网络中获得像素级语义信息,以 RGB-D 为输入,基于带有语义标签的 ElasticFusion SLAM 重建密集 3D 地图。但是该系统要求 RGB-D 输入具有可靠的深度测量。在 SemanticFusion 的基础上,文献[105]提出了一种针对道路场景的 3D 语义地图重建的方法,使用高效的 CNN 网络预测语义信息,并在实时单目 LSD-SLAM 系统中将这些信息映射到全局一致的 3D 地图中。受 MobileNet[106]编码器的启发,二维语义分割过程中采用金字塔网络对关键帧进行语义分割,使用深度可分离卷积简化了模型,减少了计算量,金字塔池化结合了多尺度的特征图作为全局上下文先验,实现对关键帧中的像素分类。根据关键帧序列的位姿图在三维点云中建立标记像素和体素之间的对应关系,重建出 3D 地图,从而将关键帧的语义标签信息融合到全局一致的三维地图中。而齐少华等[107]采用的是 mobilenet-v2-ssdlite 目标检测网络对二维关键帧图像进行目标检测,在 ORB-SLAM2 的基础上,提出了基于光流的动态点检测方法与基于点云分割的三维目标信息获取方法,根据三维点云分割的目标信息,动态更新八叉树地图,但是其动态点检测精度较低,而且对于密集物体堆放的情况,无法准确分割出不同物体。

针对三维点云地图会占用大量存储资源的问题,张荣芬等[108]提出了一种可去除动态物体,表征室内静态环境的语义八叉树地图。通过 Fast-SCNN 网络提取语义信息,使用多尺度随机采样一致性(random sample consensus, RSC)进行特征点采样,并根据对极几何约束剔除动态特征点,最后通过体素滤波降低点云冗余。相比于原始的点云地图,最终构建的地图仅占 4% 的存储空间,大大减少了内存资源。

Ma 等[109]提出了一种自监督的深度神经网络预测 RGB-D 图像的多视图一致性语义分割。通过使用 SLAM 轨迹将多视图的 CNN 特征图转化为公共参考视图,在多个视图中对语义预测的一致性施加约束,最终预测地图的一致性语义标签。对于立体图像,STDyn-SLAM[110]使用立体相机捕获立体图像对以及深度图像,采用了一种类似于 VGG16 的编码-解码的 SegNet[111]网络架构,对左侧的图像进行动态对象的语义分割。结合分割结果、光流与几何约束检测出动态特征点,并根据当前位姿的深度图、左侧图像的语义结果以及 VO 信息构建局部点云,并更新到完整点云的局部点云中,最终构建出八叉树地图,完成三维重建。该方法运行速度较快,能在室内室外环境中实时运行,但是对于大规模场景面临较大挑战。

胡美玉等[112]对以 ResNet101 为基础的 DeepLab 算法进行改进,包含了 ResNet101 网络、SPP、深度值门控模块以及上采样网络,并通过稠密条件随机场进行了后处理。根据语义分割结果、深度图以及帧间相对位姿,将图像从二维空间反投影到三维空间中,从而完成三维语义地图的构建。通过上采样卷积层替换原本双线性插值的上采样方式,缓解了其缺失细节、太过粗糙的问题。利用深度图控制金字塔的空洞卷积核选择,对不同远近的物体能更好地保留细节。最终结果在 CityScapes、NYUv2、Pascal VOC2012 共 3 个数据集上进行了训练与测试,还与 PSPNet、Mask R-CNN 的分割效果进行了对比,所提出的方法在细节还原与精度提升上都有不错的效果。

(2)实例分割

语义分割虽然能标记每个像素的类别,但是却无法区分相同标签的像素所属个体,实例分割则可以做到。相比于语义分割,实例分割不需要对每个像素进行标记,只需要找到物体的边缘或者轮廓即可分割成不同的个体。

Fusion++[113]提出了一种在线且近乎实时的面向对象的 SLAM 系统,可生成一个长期地图,专注于场景中最重要的物体,而且具有可变的、与对象大小相关的分辨率。即使在真实杂乱的室内场景中,依然有着不错的性能表现。MaskFusion[114]是一个实时语义动态的 RGB-D SLAM 系统,该系统可以对运动对象进行精确分割并分配语义标签,而且能实现 Fusion++所没有考虑的非刚性与动态场景的重建。

文献[115]提出了一种 SLAM、对象检测、实例级分割、数据关联和模型更新的新组合,以获得语义建图系统。通过基于边界框的对象检测模块和无监督三维几何分割模块的结合识别单个对象,从而实现实例级语义建图。此外,还研究将检测到的对象如何作为语义地标,实现闭环检测以创建完整的语义 SLAM 系统,从而提高准确性。文献[116]提出了一种提出了一种组合的几何语义分割方案,可检测到以前从未见过的对象类别。通过地图整合策略,三维的形状、位置以及语义信息等以增量方式融入到最终的全局地图中。文献[117]在对关键帧图像提取 ORB 特征点的同时,采用改进的 Mask R-CNN 网络进行实例分割,通过提取到的语义信息辅助剔除误匹配的特征点,同时在三维地图中融入了语义信息。文献[118]则是使用 YOLOv4 目标检测网络与融合了条件随机场(conditional random field, CRF)的 Mask R-CNN 对关键帧图像进行实例分割与边缘优化,将处理后的图像融合到最终的三维点云地图。

(3)全景分割

实例分割仅对图像中的对象类别进行检测,并对同一类别对象进行分割。与实例分割不同,全景分割

则是将整张图像的全部物体以及背景检测与分割出来。

PanopticFusion[119]提出了一种在线的全景分割三维重建系统,能够实现全景重建和密集语义标记,并具有区分单个对象的能力。将无定形区域定义为 stuff 类,如地板、墙壁、天空和道路等。将可解释性对象定义为 things 类,如椅子、桌子、人和车辆。以 RGB 帧图像输入 2D 语义和实例分割网络,融合两个网络的输出获得像素级的全景标签。此外,使用一个关于全景标签的全连通 CRF 模型对地图进行正则化。还提出了一种地图分割策略,可以在不降低精度的情况下显著减少计算时间。AVP-SLAM[120]系统利用强大的语义特征来构建三维地图并定位停车场中的车辆。采用 4 个环视摄像头来增加感知范围,4 个摄像头的图像通过逆透视映射(inverse perspective mapping, IPM)投影到鸟瞰图并合成一个全向图像,使用改进的 Unet 提取语义特征,包括引导标志、停车线、减速带等。与传统特征相比,这些语义特征对透视和照明变化具有长期稳定性和鲁棒性。IMU 与车轮编码器组成 VO,利用其提供的相对位姿参数,并将语义特征投影到全局坐标中构建全局语义地图。此地图还用于在厘米级别对车辆进行定位,在地下昏暗的停车场也有着非常不错的鲁棒性与精确度。

3)广义建图

几何建图与语义建图是显性建图,利用深度学习模型将整个场景编码为隐性场景表达,称之为广义建图。广义的场景表达以一种人为无法直接理解的方式,却包含了场景的各种重要信息以解决实际的任务需求。通过深度自动编码器、神经渲染模型、神经辐射场等方法,可以将地图表示为隐性表达。

深度自动编码器可以压缩数据,将高维数据以一个高级紧凑的方式表示。CodeSLAM[121]通过在强度图像上训练一个变分自动编码器来生成密集场景几何的更通用且可优化的紧凑表示。这种广义的表示进一步地被用于基于关键帧的 SLAM 系统,以推断位姿估计和关键帧深度图。由于这种表示的大小有限,CodeSLAM 能高效联合优化跟踪相机运动和场景几何,获得全局一致性。在 CodeSLAM 的基础上,CodeMapping[122]对变分自动编码器进行了拓展,引入了 DeepFactors[123]中的基于因子图的优化,进一步提高了全局一致性。这种变分自动编码器很容易被集成到基于关键帧的 SLAM 系统中,而不会耽误系统的主线程。文献[124]引入了一种连续符号距离函数表示,(signed distance function, SDF),称之为 DeepSDF。用连续隐式表面的 SDF 表示三维物体,通过概率自动编码器学习三维形状,这也是首次利用 SDF 进行三维场景表示的文章。

神经渲染模型是通过视图合成作为自监督信号,进行隐式三维场景重建。GQN[125]将从不同视角获得的场景图像编码为一个场景表征,并利用这个表征来预测当前场景在新视角下的外观。同时 GQN 引入几何感知注意机制对系统扩展,可以生成更复杂的环境建模,而且多模态数据也适用于场景推断。文献[126]提出了一种连续的、三维结构感知的场景表示,即场景表示网络(scene representation network, SRN),它隐式地将场景表示为一个连续可微函数,将全局坐标映射到局部场景属性的特征表示。在联合形状和外观插值、视图合成等方面具有更大的潜力。文献[127]利用编码-解码的网络架构学习动态场景的潜在表示,将输入图像转换为 3D 体积表示,并将基于表面的隐式表示合并到自身的框架中。

最近兴起的神经辐射场(neural radiance fields, NeRF)也是一种不错的渲染器,学习 3D 空间信息与 2D 视角,并投影到 RGB 颜色值上。NeRF[128]通过单个 MLP 对整个场景进行连续场景建模,多层感知器与位置编码相结合,生成连续的 5D 神经辐射场将场景表示为空间中点的体素密度与视角相关的 RGB 颜色值。位置编码可以将每个 5D 坐标映射到更高维的空间中,可更好地表达场景细节内容。这种连续紧凑的表示虽然相对比于网格、体素等离散的表示表现出较大优势,但是通常需要针对每个新场景进行较长时间的优化,且随着场景复杂程度的增大,所占用的内存空间也就越大。

神经辐射场的基础上,生成辐射场(generative radiance fields, GRAF)[129]引入了生成对抗网络,对未设定姿势的 2D 图像训练生成一种神经辐射场的变种场景表示,具有以前 3D 生成模型所不具备的高度视角一致性。但是这种方法仅适用于只有单个对象的简单场景,在复杂的实际世界中表现较差。为解决 GRAF 的局限性,推广至多目标的复杂场景,GIRAFFE[130]不同于 NeRF 与 GRAF 的 MLP 输出空间中点的颜色,GIRAFFE 输出的是抽象的特征,根据这个特征经过后续处理还原到真实的像素中,因此其场景表示方式称之为生成神经特征场(generative neural feature fields)。并对场景中的每个对象与背景都引入一个仿射变换,以控制物体的位置,能增减多个物体。这种方法的局限性是有时会出现解耦失败的情况,如前景中附着了背景的元素,或者背景中含有前景的元素,且难以解决数据偏差的问题。后续也出现很多对神经辐射场不同方面的改进与优化[131-137],如 pi-GAN[131]提出了正弦表示网络(sinusoidal representation network, SIREN)周期性激活函数的 NeRF 变体,ShapeGAN[132]不仅考虑不同视角,还兼顾了不同光照条件。

根据几何建图、语义建图及广义建图 3 种建图方法中不同类型的场景表达方式进行分类,如表 6 所示。

表 6 建图方法分类

Table 6 A classification of mapping methods

建图方式	场景表达方式	文献
几何建图	深度表示	[54-56, 59, 70] [91-97]
	体素表示	[98-100]
	网格表示	[101-103]
语义建图	语义表示	[104-105, 107-110, 112, 138]
	实例表示	[113-118]
	全景表示	[119-120]
广义建图	自动编码器	[121-122, 124]
	神经渲染模型	[125-127]
	神经辐射场	[128-137]

4 常用的 VSLAM 数据集与评估指标

4.1 VSLAM 常用的数据集

KITTI 数据集由德国卡尔斯鲁厄理工学院(Karlsruher Institut fürTechnologie, KIT)和丰田工业大学芝加哥分校(Toyota Technological Institute at Chicago, TTIC)联合创立,包含了市区、高速公路等场景采集到的真实图像数据。

EuRoC 数据集是通过微型飞行器(micro air vehicle, MAV)采集的,于欧洲机器人挑战赛时公布,主要包含了苏黎世联邦理工学院的一个机房与一个普通房间两个场景采集到的双目数据和 IMU 数据。

TUM RGB-D 数据集由慕尼黑工业大学发布,通过 RGB-D 传感器捕获数据,包含两个不同的室内场景,总共 39 个序列,支持 SLAM 验证与闭环检测,提供高速准确的 6D 位姿真实地面标签。同时该数据集提出了一套评价标准与指标。

OXFORD ROBOTCAR DATASET 数据集包含超过 2 000 万张由 6 台车载相机拍摄的图片,以及激光测距、GPS 和惯性导航收集的地貌资料,常用于室外动态环境中自动驾驶车辆的定位和地图映射的研究。由于采集时间较长,涵盖了各种时间段、天气与场景变化,包括行人、车辆、施工等各种场景,大雨、小雨、下雪等不同天气以及阳光直射、夜晚、黄昏不同时间段。

RGB-D 对象数据集是一个包含 300 个常见家居用品,还提供了所有 300 个对象的地面真实姿势信息,除了 300 个对象的独立视图外,还包括了 22 个包含数据集中对象的自然场景的注释视频序列,涉及办公空间、会议室和厨房区域等常见的室内场景。

更多的数据集整理如表 7 所示。文献[139]对激光雷达、VSLAM 等常用的数据集做了更为详细的整理。

表 7 VSLAM 常用数据集

Table 7 Datasets commonly used in VSLAM

数据集	年份	类别	链接
PanoraMIS[140]	2020 年	单目	https://home.mis.u-picardie.fr/~panor
KITTI[141]	2012 年	双目、立体	http://www.cvlibs.net/datasets/kitti/index.php
Oxford RobotCar[142]	2016 年	单目、立体	https://robotcar-dataset.robots.ox.ac.uk/
EuRoC[143]	2016 年	单目、立体	https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets
TartanAir[144]	2020 年	单目、立体	http://theairlab.org/tartanair-dataset/
MVSEC[145]	2018 年	双目、立体	https://daniilidis-group.github.io/mvsec/
Complex Urban[146]	2019 年	双目、立体	https://www.complexurban.com/
Málaga Urban[147]	2014 年	立体	https://www.mrpt.org/MalagaUrbanDataset
Cityscapes[148]	2016 年	立体	https://www.cityscapes-dataset.com/
Apollo	2019 年	立体	https://www.cifasis-conicet.gov.ar/robot/doku.php
Rosario[149]	2020 年	立体	https://etsin.fairdata.fi/dataset/06926f4b-b36a-4d6e-873c-aa3e7d84ab49
FinnForest[150]	2021 年	立体	https://dsec.ifi.uzh.ch/
DSEC[151]	2018 年	立体、RGB-D	https://interiornet.org/
InteriorNet[152]	2011 年	RGB-D	http://rgbd-dataset.cs.washington.edu/

4.2 VSLAM 系统的性能评估指标

回环检测中通常使用精确率 *Precision* 以及召回率 *Recall* 来评估结果的正确性。其中真阳性(true positive, TP)、假阴性(false negative, FN)、假阳性(false positive, FP)、真阴性(true negative, TN)之间的关系如表 8 混淆矩阵所示。通常来说,*Precision* 与 *Recall* 是矛盾的,一个指标高,则另一个降低。召回率越高则系统对正样本识别能力更强,精确率越高则系统对负样本区分能力更强。在 VSLAM 系统中,更强调精确率,因此可以适度牺牲一下召回率。

表 8　混淆矩阵

Table 8　Confusion matrix

预测	实际	
	是回环	非回环
是回环	*TP*	*FN*
非回环	*FP*	*TN*

精确率又称为查准率,描述的是预测的所有回环是真实回环的概率,公式表示如下:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

召回率描述的是预测的所有真实回环是真实回环的概率,公式表示如下:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

绝对轨迹误差(absolute trajectory error, ATE),直接计算预测位姿与实际位姿之间的差值,能直观地反映出算法精度与轨迹全局一致性,非常适用于 VSLAM 系统的性能评估。*ATE* 公式表示如下:

$$ATE_{all} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| \log_e(T_{g,i}^{-1} \times T_{esti,i})^{V} \|_2^2} \tag{3}$$

其中,

$$\log_e(T_{g,i}^{-1} \times T_{esti,i})^{V} \tag{4}$$

式中:表示第 i 个点误差李群对应的李代数,$T_{esti,i}$ 表示估计轨迹第 i 个点的欧氏变换, $T_{g,i}$ 表示真实轨迹第 i 个点的欧氏变换,N 表示所取点的个数。*ATE* 公式相当于求两个点误差李群对应的李代数的均方根误差。一般的 *ATE* 求解中只需要考虑平移部分,不需要求出对应李代数,在欧氏变换中取平移量求解均方根误差即可。对应公式则变为:

$$ATE_{trans} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| trans(T_{g,i}^{-1} \times T_{esti,i}) \|_2^2} \tag{5}$$

其中,*trans* 表示欧氏变换的平移部分。

相对位姿误差(relative pose error, RPE),主要描述在相隔固定时间差下,两帧位姿变化的差,相当于直接计算里程计的误差,适用于估计系统的漂移。*RPE* 包含了旋转误差与平移误差。*RPE* 公式表示如下:

$$RPE = \sqrt{\frac{1}{N-\Delta}\sum_{i=1}^{N-\Delta} \| \log_e((T_{g,i}^{-1} \times T_{g,i+\Delta})^{-1}(T_{esti,i}^{-1} \times T_{esti,i+\Delta}))^{V} \|_2^2} \tag{6}$$

式中:Δ 表示所取点之间的时间差。

计算出相对误差和绝对误差后,通常需要对其进行统计,使用均方根误差衡量系统整体的精度,也可以用中位数或者平均值表示。*RPE* 的均方根误差如下所示,也可以取平移或者旋转部分单独计算。

$$RPE\ RMSE = \left(\frac{1}{m}\sum_{i=1}^{m} \| E_i \|^2\right)^{\frac{1}{2}} \tag{7}$$

$$RPE_{trans}\ RMSE = \left(\frac{1}{m}\sum_{i=1}^{m} \| trans(E_i) \|^2\right)^{\frac{1}{2}} \tag{8}$$

式中:E_i 表示计算得到的相对位姿误差,m 表示 *RPE* 的个数。而绝对轨迹误差通常只评估平移部分,其均方根误差公式如下所示:

$$ATE_{trans}\ RMSE = \left(\frac{1}{n}\sum_{i=1}^{n} \| F_i \|^2\right)^{\frac{1}{2}} \tag{9}$$

式中:F_i 表示计算得到的绝对位姿误差,n 表示 ATE 的个数。评估指标总结如表 9 所示。

表 9　评估指标

Table 9　Evaluation criterion

指标类别	名称	公式
VSLAM 算法精度评估指标	绝对轨迹误差	$ATE_{all} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| \log_e(T_{g,i}^{-1} \times T_{esti,i})^{V} \|_2^2}$
	绝对轨迹误差(平移部分)	$ATE_{trans} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| trans(T_{g,i}^{-1} \times T_{esti,i}) \|_2^2}$
	绝对轨迹误差(均方根误差)	$ATE_{trans}\ RMSE = \left(\frac{1}{n}\sum_{i=1}^{n} \| F_i \|^2\right)^{\frac{1}{2}}$
	相对轨迹误差	$RPE = \sqrt{\frac{1}{N-\Delta}\sum_{i=1}^{N-\Delta} \| \log_e((T_{g,i}^{-1} \times T_{g,i+\Delta})^{-1}(T_{esti,i}^{-1} \times T_{esti,i+\Delta}))^{V} \|_2^2}$
	相对轨迹误差(平移部分)	$RPE_{trans}\ RMSE = \left(\frac{1}{m}\sum_{i=1}^{m} \| trans(E_i) \|^2\right)^{\frac{1}{2}}$
回环检测评估指标	精确率	$Precision = \frac{TP}{TP + FP}$
	召回率	$Recall = \frac{TP}{TP + FN}$

5 面临的问题与挑战

虽然基于深度学习的 VSLAM 方法在近些年的发展中获得了较大的成果,但是目前 VSLAM 与深度学习结合的工作,还存在很多方面的问题没有解决,未来研究人员仍然面临着种种困难挑战。

1) 不同传感器数据处理

目前基本上所有的 VSLAM 系统都会使用多个传感器联合优化,不同传感器的共同使用,能让 VSLAM 系统更好地应对复杂的场景与环境。但是不同传感器的数据类型、数值范围大小、坐标系表示方式、时间戳表达形式等不尽相同,在使用这些数据前需要对其进行统一处理。否则容易增大系统运算量,影响运行效率与实时性。传感器的数量越多,则需要处理的数据也就越多,对不同传感器数据进行处理是一个值得探讨的问题。

2) 真实世界的突发情况

虽然训练网络所用的数据集大都是人为标定好的,但是在真实世界的环境中会出现各种突发情况与不确定性因素。例如,路面不平整导致车身剧烈晃动使摄像头拍摄的图像出现严重的运动模糊;环境明暗交替很大而影响相机曝光成像。这些不确定因素都可能使得整个 VSLAM 系统无法自定位或者回环检测,从而导致整个系统失败,影响系统鲁棒性。

3) 跨场景下的应用

目前的 VSLAM 系统的评估大多是在一些特定的场景,例如地下停车场、城市道路等。这也导致针对一个场景就需要训练一个全新的网络,这在实际应用中将会非常耗费计算量与内存。如何利用不同类型的数据集,训练网络使得能在任意不同场景下都能完成 VSLAM 任务,提高网络的泛化能力,也值得令人深思。此外,现有的场景重建工作仅限于单个对象、合成数据或者房间的级别,如何将其扩展至更复杂和大规模的重建仍然是一个需要解决的问题。

4) 深度学习的可解释性

虽然深度学习的火热促进了各个领域的发展,但是长期以来深度学习一直被调侃为“黑盒”,最大的局限性就是深度学习的可解释性差,这也使得一些学者不愿使用深度学习的方法,而更青睐于有着严谨数学公式推理以及明确对应关系的传统方法。我们只知道将一组数据送入神经网络后得到一个预测结果,但是不知道其原因。我们无法解释整个模型的运作机制,也不能理解网络训练和动态的行为,不知道网络学习到了什么,从而难以做出针对性的优化。近年来也出现了很多对可解释性研究的总结,文献[153]集中于视觉可解释性,例如特征可视化、热图等。文献[154-155]不仅介绍神经网络,还包括其他人工智能(artificial intelligence, AI)模型,而文献[156]则做了更为全面细致的总结。深度学习的可解释性对于深度学习的发展至关重要,一旦取得突破必然对 AI 领域产生巨大影响。

6 未来的发展趋势

1) 多传感器融合

多传感器融合是 VSLAM 未来发展的必然趋势,现在常用的传感器包括摄像头、IMU、激光雷达、GPS 等。不同传感器的信息结合,可实现各自优势互补。新型的传感器采集到更精确、鲁棒性更强的感知数据,能促进系统更好地完成任务,提高性能精度。新型的传感器有事件相机[157]、毫米波雷达[158]、热敏相机[159]、磁传感器[154]、偏振传感器[160]等。目前对这些不太常用的传感器的研究工作还处于初级阶段,相信未来会有非常出色的表现,同时也会有更多的新型传感器出现,帮助更好地完成 VSLAM 任务。

2) 更高级的地图表达方式

虽然目前的地图表达方式多种多样,复杂程度也不尽相同。但是现有的 VSLAM 系统建图大多是建立在提前设计好的建图方式,对于不同任务和不同场景采用统一的建图方式可能会造成资源浪费和计算量增大的问题。如何根据实际任务需要和场景复杂程度,让系统自主选择最适合当前场景与需求的地图复杂程度与地图类型,构建出能满足当前任务所需的地图,这也是未来的研究趋势之一。这一问题的解决将会大大减小计算量,减少资源内存浪费的同时,也能更好地保持实时性。

3) 终身学习

大部分的深度学习工作都是基于大量简单的封闭性数据集得到预训练的模型,根据实际应用的数据集再进行微调或者重新训练,最终得到的模型通常会忘记以前已经学习到的东西(比如物体的类别、动态对象等),这种现象称为”灾难性遗忘”。对于实际开放的世界,不断变化的外界环境与各种动态对象因素,这也要求深度学习模型需要不断地连续学习且适应世界的变化,实现终身学习以得到持续移动与观测的 VSLAM 系统。

4) 多机器人协同控制。

随着 VSLAM 系统对大规模环境与大尺度场景的需求,多机器人共同协作能同时在大规模环境中独立行动与感知,有效提高效率与系统稳定性。在搜索、救援、行星探测、军事行动等各个方面都能极大加快任务速度。但是多机器人协作旋即带来的最大问题就是如何将各个机器人的感知数据,集合到统一的框架中。例如,多个机器人构建的局部地图如何将其拼凑成一个整体的地图,并对各个机器人进行精准定位。虽然在多机协作上已经

有大量的研究尝试[161-163],但是目前还尚不成熟,是未来发展一大趋势。

5) 语义与 VSLAM 结合

单纯的 VSLAM 系统缺乏场景理解的能力,语义信息的引入使得以往像素级的数据关联上升到物体级别,不仅提高系统的建图与定位精度,还可以让机器人像人类一样感知真实的世界。最直观的结合为语义地图,相比于构建出的点云地图中一个个无法人为理解的点,语义地图能更直观地展示地图上的各种对象与物体,实现更高级别的场景理解,对于 VSLAM 系统的智能化有着重要的意义。语义与 VSLAM 优势互补,相辅相成,在物体识别、目标检测、语义分割、语义地图等方面获得不错的进展,在未来的研究中,二者的结合也必然发挥巨大作用。

7　结　　论

本文总结了近些年来在基于深度学习的 VSLAM 方面所取得的一些研究成果,简要介绍了 VSLAM 的发展历程与基本结构,从深度学习分别与 VO、回环检测以及建图 3 个方面的结合现状详细展开叙述。然后搜集了目前常用的各种 VSLAM 数据集,给出了评估 VSLAM 系统的性能指标,最后指出了 VSLAM 的现存问题,并对未来研究方向进行了展望。

虽然 VSLAM 与深度学习的结合起步较晚,但是其惊人的发展速度足见其巨大的潜力与研究空间。目前大多数的工作都是将深度学习应用到 VSLAM 架构的某一个模块或步骤,如特征提取、深度估计、回环检测、三维重建等。语义信息的引入不仅能让系统获得更丰富的特征信息,也更容易实现高层次的人机交互与理解。基于深度学习的 VSLAM 不仅提供了一种数据驱动的可选方法,也为下一代 AI 空间感知拓展了新的思路,未来整个 VSLAM 系统也或将由深度学习直接完成。随着深度学习的发展,相信不久的将来就会有相应的方法解释深度学习其原理,而 VSLAM 的发展也必然更上一个台阶。期望本文有助于促进 VSLAM 技术的发展与应用。

参考文献

[1] 刘佳, 徐闯, 陈大鹏, 等. 基于 Marker-SLAM 的视触觉增强现实交互算法[J]. 仪器仪表学报, 2022, 43(8): 26-38.
LIU J, XU CH, CHEN D P, et al. Visuo-haptic augmented reality interaction algorithm based on Marker-SLAM[J]. Chinese Journal of Scientific Instrument, 2022, 43(8):26-38.

[2] 赵洋, 刘国良, 田国会, 等. 基于深度学习的视觉 SLAM 综述[J]. 机器人, 2017, 39(6): 889-896.
ZHAO Y, LIU G L, TIAN G H, et al. A survey of visual SLAM based on deep learning[J]. Robot, 2017, 39(6): 889-896.

[3] 吴建清, 宋修广. 同步定位与建图技术发展综述[J]. 山东大学学报(工学版), 2021, 51(5): 16-31.
WU J Q, SONG X G. Review on development of simultaneous localization and mapping technology[J]. Journal of Shandong University (Engineering Science), 2021, 51(5): 16-31.

[4] 周彦, 李雅芳, 王冬丽, 等. 视觉同时定位与地图创建综述[J]. 智能系统学报, 2018, 13(1): 97-106.
ZHOU Y, LI Y F, WANG D L, et al. A survey of VSLAM[J]. CAAI Transactions on Intelligent Systems, 2018, 13(1): 97-106.

[5] 高兴波, 史旭华, 葛群峰, 等. 面向动态物体场景的视觉 SLAM 综述[J]. 机器人, 2021, 43(6): 733-750.
GAO X B, SHI X H, GE Q F, et al. A survey of visual SLAM for scenes with dynamic objects[J]. Robot, 2021, 43(6): 733-750.

[6] 王柯赛, 姚锡凡, 黄宇, 等. 动态环境下的视觉 SLAM 研究评述[J]. 机器人, 2021, 43(6): 715-732.
WANG K S, YAO X F, HUANG Y, et al. Review of visual SLAM in dynamic environment[J]. Robot, 2021, 43(6): 715-732.

[7] CHEN C, WANG B, LU C, et al. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence[J]. ArXiv Preprint, 2020, ArXiv:2006.12567.

[8] 李少朋, 张涛. 深度学习在视觉 SLAM 中应用综述[J]. 空间控制技术与应用, 2019, 45(2): 1-10.
LI SH P, ZHANG T. A survey of deep learning application in visual SLAM[J]. Aerospace Control and Application, 2019, 45(2): 1-10.

[9] 刘瑞军, 王向上, 张晨, 等. 基于深度学习的视觉 SLAM 综述[J]. 系统仿真学报, 2020, 32(7): 1244-1256.
LIU R J, WANG X SH, ZHANG CH, et al. A survey on visual SLAM based on deep learning[J]. Journal of System Simulation, 2020, 32(7): 1244-1256.

[10] DEBEUNNE C, VIVET D. A review of visual-lidar fusion based simultaneous localization and mapping[J]. Sensors, 2020, 20(7): 2068.

[11] HUANG B, ZHAO J, LIU J. A survey of simultaneous localization and mapping with an envision in 6G wireless

networks[J]. Journal of Global Positioning Systems, 2021, 17(2): 208-238.

[12] JIA G, LI X, ZHANG D, et al. Visual-SLAM classical framework and key techniques: A review[J]. Sensors, 2022, 22(12): 4582.

[13] SMITH R, CHEESEMAN P. On the representation and estimation of spatial uncertainty[J]. The International Journal of Robotics Research, 1987, 5(4): 56-68.

[14] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332.

[15] AYACHE N, FAUGERAS O D. Building, registrating, and fusing noisy visual maps[J]. The International Journal of Robotics Research, 1988, 7(6): 45-65.

[16] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]. 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Piscataway, 2007: 225-234.

[17] LOURAKIS M, ARGYROS A. SBA: A software package for generic sparse bundle adjustment[J]. ACM Trans. Math. Softw., 2009, 36(1): 2, DOI: 10.1145/1486525.1486527.

[18] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]. 2011 International Conference on Computer Vision, 2011: 2320-2327.

[19] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C]. Computer Vision-ECCV 2014: 834-849.

[20] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: Fast semi-direct monocular visual odometry[C]. 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014: 15-22.

[21] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[22] 郭金辉, 陈秀万, 王媛. 视觉惯性 SLAM 研究进展[J]. 火力与指挥控制, 2021, 46(1): 1-8.
GUO J H, CHEN X W, WANG Y. A review of visual inertial SLAM research development[J]. Fire Control & Command Control, 2021, 46(1): 1-8.

[23] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[24] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[25] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF[C]. 2011 International Conference on Computer Vision, Barcelona, Spain, 2011: 2564-2571.

[26] ROSTEN E, DRUMMOND T. Machine learning for high-speed corner detection[C]. Computer Vision-ECCV 2006, 2006: 430-443.

[27] CALONDER M, LEPETIT V, OZUYSAL M, et al. BRIEF: Computing a local binary descriptor very fast[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1281-1298.

[28] LI D, SHI X, LONG Q, et al. DXSLAM: A robust and efficient visual SLAM system with deep features[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: 4958-4965.

[29] GAO X, ZHANG T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. Autonomous Robots, 2017, 41(1): 1-18.

[30] BEESON P, MODAYIL J, KUIPERS B. Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy[J]. I. J. Robotic Res., 2010, 29(4): 428-459.

[31] ARSHAD S, KIM G W. Role of deep learning in loop closure detection for visual and lidar SLAM: A survey[J]. Sensors, 2021, 21(4): 1243.

[32] HORNUNG A, WURM K M, BENNEWITZ M, et al. OctoMap: An efficient probabilistic 3D mapping framework based on octrees[J]. Autonomous Robots, 2013, 34(3): 189-206.

[33] QIN T, ZHENG Y, CHEN T, et al. A light-weight semantic map for visual localization towards autonomous driving[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 11248-11254.

[34] DETONE D, MALISIEWICZ T, RABINOVICH A. Superpoint: Self-supervised interest point detection and description[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018: 337-33712.

[35] LIU Y, LI J, HUANG K, et al. MobileSP: An FPGA-based real-time keypoint extraction hardware accelerator for mobile VSLAM[J]. IEEE Transactions on Circuits

and Systems I, 2022, 69(12):4919-4929.

[36] 余洪山, 郭丰, 郭林峰, 等. 融合改进 SuperPoint 网络的鲁棒单目视觉惯性 SLAM[J]. 仪器仪表学报, 2021, 42(1): 116-126.
YU H SH, GUO F, GUO L F, et al. Robust monocular visual-inertial SLAM based on the improved superpoint network[J]. Chinese Journal of Scientific Instrument, 2021, 42(1): 116-126.

[37] XUE F, WANG Q, XIN W, et al. Guided feature selection for deep visual odometry[C]. Asian Conference on Computer Vision, 2018: 293-308.

[38] TANG J, ERICSON L, FOLKESSON J, et al. GCNv2: Efficient correspondence prediction for real-time SLAM[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 3505-3512.

[39] TANG J, FOLKESSON J, JENSFELT P. Geometric correspondence network for camera motion estimation[J]. IEEE Robotics and Automation Letters, 2018, 3(2): 1010-1017.

[40] 王启来, 董朝轶, 刘晓阳, 等. 一种改进 GCN 深度学习算法 AGV 视觉 SALM 的研究[J]. 小型微型计算机系统, 2021, 42(10): 2116-2120.
WANG Q L, DONG CH Y, LIU X Y, et al. Research on improved GCN deep learning algorithm for AGV visual SALM[J]. Journal of Chinese Computer Systems, 2021, 42(10): 2116-2120.

[41] KANG R, SHI J, LI X, et al. DF-SLAM: A deep-learning enhanced visual SLAM system based on deep local features[J]. ArXiv Preprint, 2019, ArXiv: 1901.07223.

[42] BRUNO H M S, COLOMBINI E L. LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method[J]. Neurocomputing, 2021, 455:97-110.

[43] SOARES J C V, GATTASS M, MEGGIOLARO M A. Visual SLAM in human populated environments: Exploring the trade-off between accuracy and speed of YOLO and Mask R-CNN[C]. 2019 19th International Conference on Advanced Robotics (ICAR), Belo Horizonte, Brazil, 2019: 135-140.

[44] KIM J, NAM S, OH G, et al. Implementation of a mobile multi-target search system with 3D SLAM and object localization in indoor environments[C]. 2021 21st International Conference on Control, Automation and Systems (ICCAS), 2021: 2083-2085.

[45] 李博, 段中兴. 室内动态环境下基于深度学习的视觉里程计[J]. 小型微型计算机系统, 2023, 44(1): 49-55.
LI B, DUAN ZH X. Visual odometer based on deep learning in dynamic indoor environment[J]. Journal of Chinese Computer Systems, 2023, 44(1):49-55.

[46] AI Y B, RUI T, YANG X Q, et al. Visual SLAM in dynamic environments based on object detection[J]. Defence Technology, 2021, 17(5): 1712-1721.

[47] BALA J A, ADESHINA S, AIBINU A M. A modified visual simultaneous localisation and mapping (V-SLAM) technique for road scene modelling[C]. 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 2022:1-5.

[48] 吴丽凡, 魏东岩, 袁洪. 基于 YOLO 的复杂环境视觉 SLAM 优化方法[J]. 计算机应用, 2021, 41(S2): 208-213.
WU L F, WEI D Y, YUAN H. YOLO-based SLAM optimization method for complex environment vision[J]. Journal of Computer Applications, 2021, 41(S2): 208-213.

[49] 方娟, 方振虎. 基于目标检测网络的动态场景下视觉 SLAM 优化[J]. 北京工业大学学报, 2022, 48(5): 466-475.
FANG J, FANG ZH H. Vision SLAM optimization in dynamic scene based on object detection network[J]. Journal of Beijing University of Technology, 2022, 48(5): 466-475.

[50] LI J, PEI L, ZOU D, et al. Attention-SLAM: A visual monocular SLAM learning from human gaze[J]. IEEE Sensors Journal, 2021, 21(5): 6408-6420.

[51] 张再腾, 张荣芬, 刘宇红. 一种基于深度学习的视觉里程计算法[J]. 激光与光电子学进展, 2021, 58(4): 324-331.
ZHANG Z T, ZHANG R F, LIU Y H. Visual odometry algorithm based on deep learning[J]. Laser & Optoelectronics Progress, 2021, 58(4): 324-331.

[52] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018:3-19.

[53] 俎晨洋, 刘凤连, 汪日伟. 基于注意力机制的特征点匹配网络的 SLAM 方法[J]. 光电子·激光, 2022, 33(1): 14-22.
ZU CH Y, LIU F L, WANG R W. A SLAM method for

feature point matching network based on attention mechanism[J]. Journal of Optoelectronics·Laser, 2022, 33(1): 14-22.

[54] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu Hawaii, USA, 2017: 6612-6619.

[55] GODARD C, AODHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 2017: 6602-6611.

[56] YIN Z, SHI J. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, 2018: 1983-1992.

[57] ALMALIOGLU Y, SAPUTRA M R U, GUSMÃO P P B D, et al. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 5474-5480.

[58] FENG T, GU D. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4431-4437.

[59] YANG N, STUMBERG L V, WANG R, et al. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, United States, 2020: 1278-1289.

[60] AI Y, RUI T, LU M, et al. DDL-SLAM: A robust RGB-D SLAM in dynamic environments combined with deep learning[J]. IEEE Access, 2020, 8:162335-162342.

[61] SHAMWELL E J, LINDGREN K, LEUNG S, et al. Unsupervised deep visual-inertial odometry with online error correction for RGB-D imagery [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2478-2493.

[62] ZHONG Y, HU S, HUANG G, et al. WF-SLAM: A robust VSLAM for dynamic scenarios via weighted features[J]. IEEE Sensors Journal, 2022, 22(11): 10818-10827.

[63] HE K, GKIOXARI G, DOLLáR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[64] BESCOS B, FÁCIL J M, CIVERA J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.

[65] BESCOS B, CAMPOS C, TARDÓS J D, et al. DynaSLAM II: Tightly-coupled multi-object tracking and SLAM [J]. IEEE Robotics and Automation Letters, 2021, 6(3): 5191-5198.

[66] 徐陈, 周怡君, 罗晨. 动态场景下基于光流和实例分割的视觉 SLAM 方法[J]. 光学学报, 2022, 42(14): 147-159.
XU CH, ZHOU Y J, LUO CH. Visual SLAM method based on optical flow and instance segmentation for dynamic scenes [J]. Acta Optica Sinica, 2022, 42(14): 147-159.

[67] BOLYA D, ZHOU C, XIAO F, et al. YOLACT++ better real-time instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 1108-1121.

[68] 冯明驰, 刘景林, 李成南, 等. 一种多焦距动态立体视觉 SLAM[J]. 仪器仪表学报, 2021, 42(11): 200-209.
FENG M CH, LIU J L, LI CH N, et al. A multi-focal length dynamic stereo vision SLAM[J]. Chinese Journal of Scientific Instrument, 2021, 42(11): 200-209.

[69] YANG S, SCHERER S. CubeSLAM: Monocular 3D object SLAM[J]. IEEE Transactions on Robotics, 2019, 35(4): 925-938.

[70] SHENG L, XU D, OUYANG W, et al. Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep SLAM [C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), South Korea, 2019: 4301-4310.

[71] ZHANG K, CHAO W L, SHA F, et al. Video summarization with long short-term memory [C]. Computer Vision-ECCV 2016: 766-782.

[72] ALONSO I, RIAZUELO L, MURILLO A C. Enhancing V-SLAM keyframe selection with an efficient convnet for semantic analysis[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 4717-4723.

[73] PERTUZ S, PUIG D, GARCIA M A. Analysis of focus measure operators for shape-from-focus [J]. Pattern Recognition, 2013, 46(5): 1415-1432.

[74] ROMERA E, ÁLVAREZ J M, BERGASA L M, et al.

ERFNet: Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272.

[75] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. ArXiv Preprint, 2017, ArXiv: 1706.05587.

[76] LIU Y, MIURA J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods[J]. IEEE Access, 2021, 9:23772-23785.

[77] GAO X, ZHANG T. Loop closure detection for visual SLAM systems using deep neural networks[C]. 2015 34th Chinese Control Conference (CCC), 2015: 5851-5856.

[78] 张云洲, 胡航, 秦操, 等. 基于栈式卷积自编码的视觉 SLAM 闭环检测[J]. 控制与决策, 2019, 34(5): 981-988.
ZHANG Y ZH, HU H, QIN C, et al. Loop closure detection for visual SLAM based on stacked convolutional autoencoder[J]. Control and Decision, 2019, 34(5): 981-988.

[79] CHEN B F, YUAN D, LIU C, et al. Loop closure detection based on multi-scale deep feature fusion[J]. Applied Sciences, 2019, 9(6):1120.

[80] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Commun ACM, 2017, 60(6): 84-90.

[81] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[82] MEMON A R, WANG H, HUSSAIN A. Loop closure detection using supervised and unsupervised deep neural networks for monocular SLAM systems[J]. Robotics and Autonomous Systems, 2020, 126:103470.

[83] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv Preprint, 2014, ArXiv:1409.1556.

[84] ZHANG X, SU Y, ZHU X. Loop closure detection for visual SLAM systems using convolutional neural network[C]. 2017 23rd International Conference on Automation and Computing (ICAC), 2017: 1-6.

[85] SERMANET P, EIGEN D, ZHANG X, et al. OverFeat: Integrated recognition, localization and detection using convolutional networks[J]. International Conference on Learning Representations (ICLR), 2014. DOI: 10.48550/arXiv.1312.6229.

[86] WANG S, LYU X, LIU X, et al. Compressed holistic convnet representations for detecting loop closures in dynamic environments[J]. IEEE Access, 2020, 8: 60552-60574.

[87] 杨馨竹, 张建勋, 郭纪志. Darknet-NVPP 视觉 SLAM 快速闭环检测方法[J]. 小型微型计算机系统, 2022, 44(4): 832-837.
YANG X ZH, ZHANG J X, GUO J ZH. Darknet-NVPP rapid closed-loop detection method for visual SLAM[J]. Journal of Chinese Computer Systems, 2022, 44(4): 832-837.

[88] 占浩, 朱振才, 张永合, 等. 基于残差网络的图像序列闭环检测[J]. 激光与光电子学进展, 2021, 58(4): 315-323.
ZHAN H, ZHU ZH C, ZHANG Y H, et al. Loop-closure detection using image sequencing based on resNet[J]. Laser & Optoelectronics Progress, 2021, 58(4): 315-323.

[89] 郭烈, 葛平淑, 王肖, 等. 基于卷积神经网络优化回环检测的视觉 SLAM 算法[J]. 西南交通大学学报, 2021, 56(4): 706-712,768.
GUO L, GE P SH, WANG X, et al. Visual simultaneous localization and mapping algorithm based on convolutional neural network to optimize loop detection[J]. Journal of Southwest Jiaotong University, 2021, 56(4): 706-712,768.

[90] 赵浩苏, 邢凯, 宋力. 基于 CNN 特征提取和增量式字典的 VSLAM 回环检测[J]. 计算机应用与软件, 2020, 37(1): 157-164.
ZHAO H S, XING K, SONG L. VSLAM loop closure detction based on CNN feature extraction and incremental dictionary[J]. Computer Applications and Software, 2020, 37(1): 157-164.

[91] ZOU Y, LUO Z, HUANG J B. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 36-53.

[92] ALMALIOGLU Y, TURAN M, SAPUTRA M R U, et al. SelfVIO: Self-supervised deep monocular visual-inertial odometry and depth estimation[J]. Neural Networks, 2022, 150:119-136.

[93] LI Y, USHIKU Y, HARADA T. Pose graph optimization

for unsupervised monocular visual odometry[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 5439-5445.

[94] WANG R, PIZER S M, FRAHM J. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth [C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 5550-5559.

[95] ZOU Y, JI P, TRAN Q H, et al. Learning monocular visual odometry via self-supervised long-term modeling[C]. European Conference on Computer Vision (ECCV), 2020: 710-727.

[96] ZHAO C, SUN L, PURKAIT P, et al. Learning monocular visual odometry with dense 3D mapping from dense 3D flow [C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 6864-6871.

[97] SHEN T, LUO Z, ZHOU L, et al. Beyond photometric loss for self-supervised ego-motion estimation[C]. 2019 International Conference on Robotics and Automation (ICRA), 2019: 6359-6365.

[98] JI M, GALL J, ZHENG H, et al. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis[C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 2326-2334.

[99] PASCHALIDOU D, ULUSOY A O, SCHMITT C, et al. RayNet: Learning volumetric 3D reconstruction with ray potentials[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 3897-3906.

[100] TATARCHENKO M, DOSOVITSKIY A, BROX T. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs[C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 2107-2115.

[101] WANG N, ZHANG Y, LI Z, et al. Pixel2Mesh: Generating 3D mesh models from single RGB images[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 55-71.

[102] DAI A, NIEßNER M. Scan2Mesh: From unstructured range scans to 3D meshes [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 5574-5583.

[103] BLOESCH M, LAIDLOW T, CLARK R, et al. Learning meshes for dense visual SLAM [C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 5854-5863.

[104] MCCORMAC J, HANDA A, DAVISON A, et al. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks [C]. 2017 IEEE International Conference on Robotics and Automation (ICRA), Marina Bay Sands, Singapore, 2017: 4628-4635.

[105] LI X, AO H, BELAROUSSI R, et al. Fast semi-dense 3D semantic mapping with monocular visual SLAM[C]. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017: 385-390.

[106] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications [J]. ArXiv Preprint, 2017, ArXiv: 1704.04861.

[107] 齐少华, 徐和根, 万友文, 等. 动态环境下的语义地图构建[J]. 计算机科学, 2020, 47(9): 198-203.
QI SH H, XU H G, WANG Y W, et al. Construction of semantic mapping in dynamic environments [J]. Computer Science, 2020, 47(9): 198-203.

[108] 张荣芬, 袁文昊, 卢金, 等. 面向室内动态场景的视觉同时定位与地图构建语义八叉树地图构建方法[J]. 激光与光电子学进展, 2022, 59(18): 190-204.
ZHANG R F, YUAN W H, LU J, et al. Visual simultaneous localization and mapping method of semantic octree map toward indoor dynamic scenes[J]. Laser & Optoelectronics Progress, 2022, 59(18): 190-204.

[109] MA L, STÜCKLER J, KERL C, et al. Multi-view deep learning for consistent semantic mapping with RGB-D cameras[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017: 598-605.

[110] ESPARZA D, FLORES G. The STDyn-SLAM: A stereo vision and semantic segmentation approach for VSLAM in dynamic outdoor environments[J]. IEEE Access, 2022, 10:18201-18209.

[111] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[112] 胡美玉, 张云洲, 秦操, 等. 基于深度卷积神经网络的语义地图构建[J]. 机器人, 2019, 41(4):

452-463.
HU M Y, ZHANG Y ZH, QIN C, et al. Semantic map construction based on deep convolutional neural network[J]. Robot, 2019, 41(4): 452-463.

[113] MCCORMAC J, CLARK R, BLOESCH M, et al. Fusion++: Volumetric object-Level SLAM[C]. 2018 International Conference on 3D Vision (3DV), 2018: 32-41.

[114] RUNZ M, BUFFIER M, AGAPITO L. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects[C]. 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2018: 10-20.

[115] SÜNDERHAUF N, PHAM T T, LATIF Y, et al. Meaningful maps with object-oriented semantic mapping[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017: 5079-5085.

[116] GRINVALD M, FURRER F, NOVKOVIC T, et al. Volumetric instance-aware semantic mapping and 3D object discovery[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 3037-3044.

[117] 何召兰, 何乃超, 张庆洋, 等. 基于实例分割的视觉 SLAM 算法[J]. 计算机工程与设计, 2020, 41(10): 2791-2796.
HE ZH L, HE N CH, ZHANG Q Y, et al. Visual SLAM algorithm based on instance segmentation[J]. Computer Engineering and Design, 2020, 41(10): 2791-2796.

[118] 邓晨, 李宏伟, 张斌, 等. 基于深度学习的语义 SLAM 关键帧图像处理[J]. 测绘学报, 2021, 50(11): 1605-1616.
DENG CH, LI H W, ZHANG B, et al. Research on key frame image processing of semantic SLAM based on deep learning[J]. Acta Geodaetica et Cartographica Sinica, 2021, 50(11): 1605-1616.

[119] NARITA G, SENO T, ISHIKAWA T, et al. PanopticFusion: Online volumetric semantic mapping at the level of stuff and things[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 4205-4212.

[120] QIN T, CHEN T, CHEN Y, et al. AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: 5939-5945.

[121] BLOESCH M, CZARNOWSKI J, CLARK R, et al. CodeSLAM-Learning a compact, optimisable representation for dense visual SLAM[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 2560-2568.

[122] MATSUKI H, SCONA R, CZARNOWSKI J, et al. CodeMapping: Real-time dense mapping for sparse SLAM using compact scene representations[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 7105-7112.

[123] CZARNOWSKI J, LAIDLOW T, CLARK R, et al. DeepFactors: Real-time probabilistic dense monocular SLAM[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 721-728.

[124] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: Learning continuous signed distance functions for shape representation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 165-174.

[125] ESLAMI S, JIMENEZ REZENDE D, BESSE F, et al. Neural scene representation and rendering[J]. Science, 2018, 360:1204-1210.

[126] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G. Scene representation networks: Continuous 3D-structure-aware neural scene representations[C]. Advances in Neural Information Processing Systems, 2019, 32: 1119-1130.

[127] LOMBARDI S, SIMON T, SARAGIH J, et al. Neural volumes: Learning dynamic renderable volumes from images[J]. ACM Trans. Graph., 2019, 38(4): 65. 1-65. 14.

[128] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[J]. Commun ACM, 2021, 65(1): 99-106.

[129] SCHWARZ K, LIAO Y, NIEMEYER M, et al. GRAF: Generative radiance fields for 3D-aware image synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 20154-20166.

[130] NIEMEYER M, GEIGER A. GIRAFFE: Representing scenes as compositional generative neural feature fields[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 11448-11459.

[131] CHAN E R, MONTEIRO M, KELLNHOFER P, et al. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 5795-5805.

[132] PAN X, XU X, LOY C C, et al. A shading-guided generative implicit model for shape-accurate 3D-aware image synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34: 20002-20013.

[133] PENG S, ZHANG Y, XU Y, et al. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 9050-9059.

[134] SRINIVASAN P P, DENG B, ZHANG X, et al. NeRV: Neural reflectance and visibility fields for relighting and view synthesis[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 7491-7500.

[135] LI Z, NIKLAUS S, SNAVELY N, et al. Neural scene flow fields for space-time view synthesis of dynamic scenes[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 6494-6504.

[136] MARTIN-BRUALLA R, RADWAN N, SAJJADI M S M, et al. NeRF in the wild: Neural radiance fields for unconstrained photo collections[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 7206-7215.

[137] PARK K, SINHA U, BARRON J T, et al. Nerfies: Deformable neural radiance fields[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 5845-5854.

[138] LIU Y, FU Y, CHEN F, et al. Simultaneous localization and mapping related datasets: A comprehensive survey[J]. ArXiv Preprint, 2021, ArXiv:2102.04036.

[139] BENSEDDIK H, MORBIDI F, CARON G. PanoraMIS: An ultra-wide field of view image dataset for vision-based robot-motion estimation[J]. The International Journal of Robotics Research, 2020, 39(9): 1037-1051.

[140] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32:1231-1237.

[141] MADDERN W, PASCOE G, LINEGAR C, et al. 1 year, 1000 km: The oxford robotcar dataset[J]. The International Journal of Robotics Research, 2016, 36(1): 3-15.

[142] BURRI M, NIKOLIC J, GOHL P, et al. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[143] WANG W, ZHU D, WANG X, et al. TartanAir: A dataset to push the limits of visual SLAM[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: 4909-4916.

[144] ZHU A Z, THAKUR D, ÖZASLAN T, et al. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception[J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2032-2039.

[145] JEONG J, CHO Y, SHIN Y S, et al. Complex urban dataset with multi-level sensors from highly diverse urban environments[J]. The International Journal of Robotics Research, 2019, 38(6): 642-657.

[146] BLANCOCLARACO J L, MORENODUEÑAS F Á, GONZÁLEZJIMéNEZ J. The málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario[J]. The International Journal of Robotics Research, 2014, 33(2): 207-214.

[147] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 3213-3223.

[148] PIRE T, MUJICA M, CIVERA J, et al. The rosario dataset: multisensor data for localization and mapping in agricultural environments[J]. The International Journal of Robotics Research, 2019, 38(6): 633-641.

[149] ALI I, DURMUSH A, SUOMINEN O, et al. FinnForest dataset: A forest landscape for visual SLAM[J]. Robotics and Autonomous Systems, 2020, 132:103610.

[150] GEHRIG M, AARENTS W, GEHRIG D, et al. DSEC: A stereo event camera dataset for driving scenarios[J]. IEEE Robotics and Automation Letters, 2021, 6(3): 4947-4954.

[151] LI W, SAEEDI S, MCCORMAC J, et al. InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset[C]. In British Machine Vision Conference

(BMVC), Northumbria University, 2018.

[152] ZHANG Q S, ZHU S C. Visual interpretability for deep learning: A survey [J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(1): 27-39.

[153] ADADI A, BERRADA M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)[J]. IEEE Access, 2018, 6:52138-52160.

[154] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2018, 51(5):1-42.

[155] FAN F L, XIONG J, LI M, et al. On interpretability of artificial neural networks: A survey [J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, 5(6): 741-760.

[156] REBECQ H, HORSTSCHAEFER T, GALLEGO G, et al. EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time[J]. IEEE Robotics and Automation Letters, 2017, 2(2): 593-600.

[157] XIAOXUAN LU C, ROSA S, ZHAO P, et al. See through smoke: Robust indoor mapping with low-cost mmwave radar [C]. Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, Toronto Ontario, Canada, 2020: 14-27.

[158] SAPUTRA M R U, GUSMAO P P B D, LU C X, et al. DeepTIO: A deep thermal-inertial odometry with visual hallucination [J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1672-1679.

[159] 夏琳琳, 张晶晶, 初妍, 等. 融合天空偏振光的视觉 SLAM 研究进展与展望[J]. 兵工学报, 2022: 1-15.
XIA L L, ZHANG J J, CHU Y, et al. Progresses and prospects of polarized skylight fused visual SLAM[J]. Acta Armamentarii, 2022, 1-15.

[160] LAJOIE P, RAMTOULA B, CHANG Y, et al. DOOR-SLAM: Distributed, online, and outlier resilient SLAM for robotic teams[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1656-1663.

[161] TCHUIEV V, INDELMAN V. Distributed consistent multi-robot semantic localization and mapping [J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4649-4656.

[162] TIAN Y, CHANG Y, ARIAS F H, et al. Kimera-Multi: Robust, distributed, dense metric-semantic SLAM for multi-robot systems[J]. IEEE Transactions on Robotics, 2022, 38(4):2022-2038.

[163] 阴贺生, 裴硕, 徐磊, 等. 多机器人视觉同时定位与建图技术研究综述[J]. 机械工程学报, 2022, 58(11): 11-36.
YIN H SH, PEI SH, XU L, et al. Review of research on multi-robot visual simultaneous localization and mapping [J]. Journal of Mechanical Engineering, 2022, 58(11): 11-36.

作者简介

张耀,2021 年于广西师范大学获得学士学位,现为南京航空航天大学电子信息工程学院信号与信息处理专业学术型硕士研究生,主要研究方向为图像处理与机器视觉。

E-mail: yaozhang_yuan@163.com

Zhang Yao received his B. Sc. degree from Guangxi Normal University in 2021. He is currently pursuing his academic postgraduate degree in Department of Signal and Information Processing, College of Electronic and Information Engineering at Nanjing University of Aeronautics and Astronautics. His main research interests include image processing and machine vision.

吴一全(通信作者),1998 年于南京航空航天大学获得博士学位,现为南京航空航天大学教授、博士生导师,主要研究方向为视觉检测与图像测量、遥感图像处理与理解、红外目标检测与识别、视频处理与智能分析等。

E-mail: nuaaimage@163.com

Wu Yiquan (Corresponding author) received his Ph. D. degree from Nanjing University of Aeronautics and Astronautics in 1998. He is currently a professor and a Ph. D. advisor at Nanjing University of Aeronautics and Astronautics. His main research interests include visual detection and image measurement, remote sensing image processing and understanding, infrared target detection and recognition, video processing and intelligent analysis, etc.