DOI：10.13973/j.cnki.robot.220426

深度学习下的视觉 SLAM 综述

黄泽霞，邵春莉
（安徽大学，安徽 合肥 230601）

摘 要：本综述涵盖了深度学习技术应用到 SLAM（同步定位与地图创建）领域的最新研究成果，重点介绍和总结了深度学习在前端跟踪、后端优化、语义建图和不确定性估计中的研究成果，展望了深度学习下视觉 SLAM 的发展趋势，为后继者了解与应用深度学习技术、研究移动机器人自主定位和建图问题的可行性方案提供助力。

关键词：深度学习；同步定位与建图；前端跟踪；后端优化；语义建图；不确定性估计
中图分类号：TP24 **文献标识码**：A **文章编号**：1002-0446(2023)-06-0756-13

Survey of Visual SLAM Based on Deep Learning

HUANG Zexia，SHAO Chunli
(*Anhui University*, *Hefei* 230601, *China*)

Abstract: The review covers the latest research results of deep learning techniques applied to the field of SLAM (simultaneous localization and mapping), focusing on and summarizing the research results of deep learning in front-end tracking, back-end optimization, semantic mapping and uncertainty estimation, and looking forward to the development trends of visual SLAM under deep learning. This work can help the successors to understand and apply the deep learning techniques to studying the feasible solutions to the problem of autonomous localization and mapping for mobile robots.

Keywords: deep learning; simultaneous localization and mapping; front-end tracking; back-end optimization; semantic mapping; uncertainty estimation

1 引言（Introduction）

随着机器人技术的发展，越来越多的机器人被用来代替人类完成简单重复或危险的工作。移动机器人由于具有较强的灵活性和可靠性，已逐渐成为机器人领域的研究焦点。在没有人干预的情况下，通过自身所带的传感器感知环境，获取未知环境的信息，并对环境进行建模，实现自主导航和定位是移动机器人的核心任务。目前，同步定位与地图创建（SLAM）技术是实现移动机器人这一任务的主流技术方案。

SLAM 技术最早被应用在机器人领域，是希望在没有任何先验知识的情况下，机器人能依据传感器的信息实时构建周围环境地图，同时，根据这个地图推测自身的位置[1]。根据所使用的传感器类型的不同，可以把 SLAM 分为基于雷达的 SLAM 和基于视觉的 SLAM。如图 1 所示，一个完整的视觉 SLAM 系统主要由传感器数据流、前端跟踪模块（视觉里程计）、后端优化模块、回环检测模块和地图构建模块组成[2]。

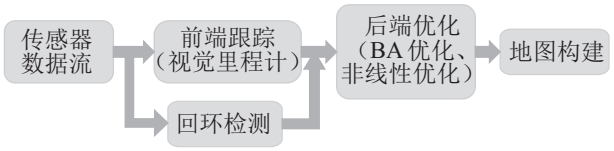

图 1 视觉 SLAM 系统框架
Fig.1 Visual SLAM framework

随着深度学习技术的兴起，计算机视觉的许多传统领域都取得了突破性进展，例如目标的检测、识别和分类等领域。近年来，研究人员开始在视觉 SLAM 算法中引入深度学习技术，使得深度学习 SLAM 系统获得了迅速发展，并且比传统算法展现出更高的精度和更强的环境适应性。

从 2015 年 Kendall 等[3] 提出在视觉里程计中引入深度学习方法开始，经过近十年的发展，基于深度学习的视觉 SLAM 系统框架已日趋成熟。同时，深度学习与视觉 SLAM 结合发展方面也取得了很多

基金项目：国家自然科学基金（61903001）；安徽省自然科学基金（1908085QF269）.
通信作者：黄泽霞，huangzexia1977@163.com 收稿/录用/修回：2022-11-18/2023-01-11/2023-01-17

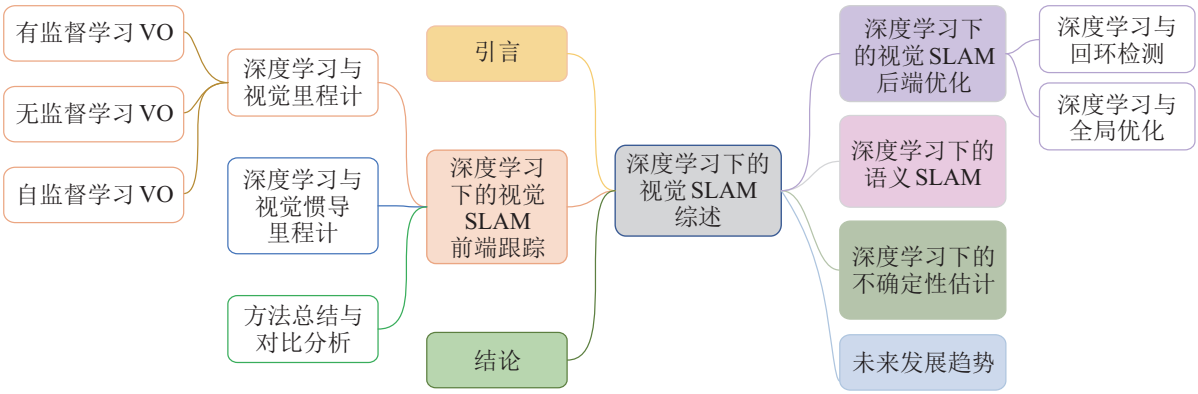

图 2　基于深度学习的视觉 SLAM 现有方法的分类

Fig.2　Classification of the existing deep-learning based visual SLAM

进展[4-8]。其中，文 [4] 较早地对深度学习与 SLAM 融合方法进行了深入细致的调研，并展望了几个未来的方向。但由于当时对语义 SLAM 领域的研究刚刚起步，文中只进行了简要讨论，没有办法进行全面总结。此外，多数综述都只对 SLAM 系统的某几个方面进行归纳与总结，如，对视觉里程计和回环检测研究成果的总结[5]，对视觉里程计、回环检测和地图重建的调研[6-7] 等。值得注意的是，虽然也有专门讨论不确定性估计算法的综述[9-10]，然而，它们大部分的关注点主要集中在基于神经网络方法对不确定性的建模、深度模型下不确定性方法之间的对比等。

基于上述分析及广泛调研，本文对深度学习下视觉 SLAM 方法涵盖的几大模块（视觉里程计、回环检测、全局优化、语义 SLAM 以及不确定性估计）当前采用的算法的性能特点、应用环境等方面进行分类讨论，如图 2 所示。同时，论述了现有模型的局限性，并指出该领域未来可能的发展方向。

2　深度学习下的视觉 SLAM 前端跟踪（Visual SLAM front-end tracking based on deep learning）

SLAM 前端跟踪也称作视觉里程计（VO），可以通过传感器获得的不同帧之间的感知信息估计出移动机器人的运动变化[11]。VO 估计最核心的任务是利用传感器的测量数据准确地预测移动机器人的运动并输出相对位姿。对 SLAM 系统而言，在初始状态已知的情况下，可通过这些相对位姿重构全局轨迹。因此，保证输出位姿估计精度是移动机器人实现高精度定位的关键因素[8]。

2.1　深度学习与视觉里程计

传统的 VO 估计通常包括相机标定、特征提取、特征匹配/跟踪、异常值剔除、运动估计、尺度估计和局部优化几部分，系统架构如图 3 所示[12]。

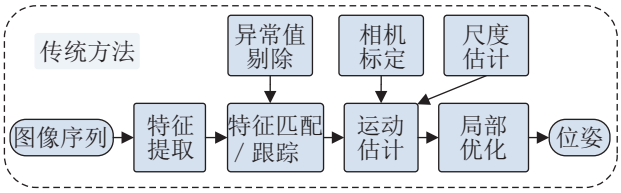

图 3　传统单目 VO 的框架图

Fig.3　Architecture of the conventional monocular VO

卷积神经网络（CNN 或 ConvNet）在图像识别任务中获得的巨大成功，使得利用 CNN 来处理 VO 问题成为了可能。和传统的 VO 估计方法相比，深度学习方法可以自动对图像特征进行提取，而不需要繁重的人工特征标注过程，使得整个估计过程更加直观简洁。根据网络的训练方式和数据集是否使用标签，本节主要集中对有监督学习（supervised learning）VO、无监督学习（unsupervised learning）VO 和自监督学习（self-supervised learning）VO 三种情况进行讨论和总结。

2.1.1　有监督学习 VO

有监督学习 VO 的目的是通过在标记数据集上训练一个深度神经网络模型，直接构造出从连续图像到运动变换的映射函数。模型的输入是一对连续的图像，输出是包含了平移信息和旋转信息的矩阵。

2015 年，Konda 等[13] 提出了基于端到端的卷积神经网络架构来预测相机速度和输入图像的方向变化的方法，整个预测过程主要包括图像序列深度和运动信息的提取、图像序列速度和方向变化估计 2 个步骤，是将深度学习融入到 VO 研究领域中最早的研究成果之一。

Costante 等[14] 通过学习图像数据的最优特征表示，对视觉里程计进行了估计。该方案将稠密光流

特征作为 CNN 网络的输入，探索和设计了 3 种不同的 CNN 深度网络架构，基于全局特征的 CNN-1b、基于局部特征的 CNN-4b，以及结合前 2 种架构的 P-CNN。所提方案虽然在应对图像运动模糊、光照变化方面具有较强的鲁棒性，但当图像序列帧间速度过快时，算法误差会较大，准确性会下降。

在有监督学习 VO 的模型中，DeepVO[12] 是目前效果最好且应用较为广泛的。该算法采用将 ConvNet 和递归神经网络（RNN）相结合的方法来实现视觉里程计的端到端学习。该网络的框架如图 4 所示，它不采用传统 VO 中的任何模块，而是直接从一系列原始 RGB 图像或视频中推断出姿态。DeepVO 框架不仅能通过 CNN 自动学习 VO 问题的有效特征表示，而且能够利用 RNN 隐式地学习图像间的内在联系及动力学关系。

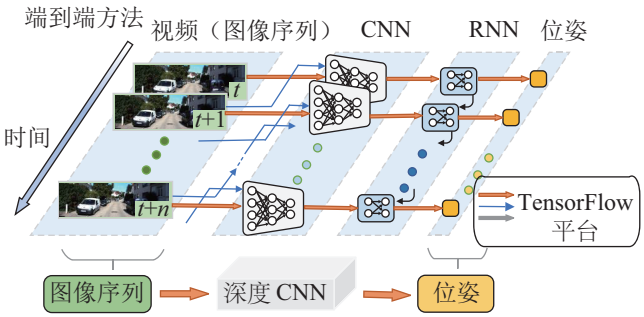

图 4 DeepVO 网络的框架结构图

Fig.4 Architectures of the DeepVO network

与传统方法相比，DeepVO 模型在精度上没有绝对的优势，但是，因其学习的是各帧之间的位姿关系，具有较好的泛化能力，因而得到了广泛的关注。例如，VINet 算法[15]和 Deep EndoVO 算法[16]等都是在此基础上进行的改进，并获得了较好的效果。

随着研究人员对高效的小规模网络的深入研究，知识蒸馏作为一种新兴的轻量化小模型，已成为深度学习领域又一个被关注的重点。2019 年，Saputra 等[17] 首次利用知识蒸馏来预测位姿回归，提出了一种基于对教师模型结果的“信任”程度来附加蒸馏损失的方案。该方法有效地减少了网络的参数量，增强了移动机器人的实时操作性。其他相关方法还有很多，比如，Saputra 等[18] 在 ICRA 会议上探讨了将课程学习（curriculum learning，CL）应用到复杂几何任务上的问题，设计了 CL-VO 网络。该网络利用新的课程学习策略来学习单目视觉里程计中的几何信息，通过几何感知目标函数，在训练的过程中逐步提升训练数据的复杂度。

总之，由于机器学习技术、数据存储量和计算速度等方面的飞速发展，这些有监督学习方法可以从输入图像中自动获取相机的位姿变换，从而解决实际场景中视觉里程计估计难的问题。

2.1.2 无监督学习 VO

无监督学习所学习的数据不需要标注，学习的目标通常是找出数据与数据之间的关系。随着深度学习技术在计算机视觉领域中的优势凸显，人们对探索无监督学习在视觉里程计中的应用越来越感兴趣，研究者也逐步把侧重点放在了该领域上。

2017 年，Godard 等[19] 在 CVPR 会议上提出了采用无监督学习的方法来进行单一图像的深度估计。该方法的基本思路是利用图像的多重目标损失来训练神经网络，使得光度误差最小化，从而得到很好的视差图。特别值得注意的是，文 [19] 是在已知相机参数的情况下进行训练的。为了解决相机参数未知且左右相机不在同一个平面的问题，Zhou 等[20] 提出了一种既不需要双目相机，也不用知道相机参数的改进算法。其核心思想是通过深度 CNN 和位姿 CNN 两个网络分别生成深度图和图像间的位姿，根据深度图与位姿将原图像投射到目标图像上，最后通过比较真实目标图像与投射产生的目标图像的重建误差来训练网络。该学习方式在网络结构设计、初始值设定和训练方法上都采用了较为合适的策略，是目前效果最好的无监督学习方法之一。然而在文献中，作者提到还存在几个尚待解决的问题：1) 该方法存在绝对尺度问题。由于文中的深度预测不够完整，因而无法重建环境的全局轨迹，降低了其在全局范围内定位的精度。2) 文中的光度一致性计算没有考虑实际场景中可能出现的物体移动和遮挡。

对于上述尺度一致性问题，学者们进行了讨论和研究，并提出了许多不同的改进方案[21-23]。例如，Li 等[21] 在文 [20] 的基础上作了相应的改进，提出一种基于无监督学习方法来得到相机位姿绝对尺度的单目视觉里程计估计网络 UnDeepVO。该方法通过左右图像分别估计出相机左右序列的位姿值和深度值，然后再利用输入的立体图像对得到真实尺度的深度图，与大多数单目无监督的学习方案相比，该方法能够真实地恢复相机位姿的尺度。文 [22] 提出利用几何一致性损失函数来满足深度估计和位姿估计之间的尺度一致性约束。该方法将预测的图像深度图转换到 3D 空间，然后将局部深度重投影作为损失函数，以此来保持深度预测的尺度一致性，从而保持位姿估计的尺度一致性。

在改善位姿估计精度方面，Yin 等[24] 提出了一种可以联合学习单目深度、光流和相机姿态的 GeoNet 无监督网络学习框架。该学习过程通过刚性结构重建器和非刚性运动定位器 2 个子任务，分别学习刚性流和目标物体的运动。除此之外，GeoNet 还引入了自适应几何一致性损失，增强了对相机遮挡和非朗伯区域的异常值的鲁棒性，提升了相机位姿估计的精度。此外，Zhao 等[25] 同样也在改善位姿估计精度方面进行了改进和扩展。

自 2014 年 Goodfellow 等[26] 提出生成式对抗网络（GAN）以来，由于其强大的生成能力，该方法在计算机视觉、自然语言处理等领域越来越受到学术界和工业界的重视。GANVO 算法[27] 正是在 GAN 基础上提出的一种生成式无监督学习框架，该算法通过在单目 VO 中使用生成式对抗神经网络和循环无监督学习方法来预测相机运动姿态和单目深度图。SGANVO （叠加生成式对抗网络）[28] 是继 GANVO 之后出现的一种改进算法，其整体是由一堆 GAN 层堆叠组成。系统在对抗性学习过程中进行深度估计和自我运动预测，并对算法的前、后层网络进行递归表示，从而有效地捕捉各层的时间动态特征。SGANVO 通过增加网络层数的方式，使得深度估计效果得到了很大的改善。

传统的无监督深度估计需要利用双目图片进行自监督，而文 [29] 提出的 SfM-Net 网络却只需要单目的视频流就能恢复深度图和相机位姿的估计。首先，通过输入的单个图像生成对应深度图像；然后，融合生成深度点云；最后，通过输入连续两帧的图像计算输出图像间的位姿关系，识别并分割出（以掩模的形式）场景中的运动物体。

相比于有监督学习 VO，无监督学习 VO 学习到的特征更加具有适应性和丰富性，因此，在性能上虽然与前者还有一定差距，但其在提供未知场景位姿信息方面具有更佳的可拓展性和可解释性。

2.1.3　自监督学习 VO

在传统的 VO 中，想要获得场景像素点的深度真值比较困难，而自监督学习方法集成了深度学习框架和经典的几何模型，给这一难题指明了方向。

第一种自监督学习法是以立体相机拍摄的图像对作为训练样本，根据视差与场景深度的关系，预测出目标图像的视差图，并转换为深度图[30-32]。如，文 [30] 提出用立体图像作为训练网络的输入，以自监督的方式在图像对上进行模型训练。文中以左右视差之间的双循环一致性作为目标函数，同时引入自适应正则化损失函数，以此排除立体图像中的遮挡区域。Godard 等[19] 使用单个图像作为卷积神经网络的输入，在全局范围内预测得到每个像素的场景深度；然后利用左右图像一致性损失，增强左右视差图的一致性，可以使结果更准确。此外，Chen 等[31] 和 Choi 等[32] 从训练策略着手，基于双目深度估计的结果来估计单目图像的深度。通过这种方式获得的网络模型可以获得最佳性能。

另一种基于自监督估计深度的思路是将视频序列中的连续帧作为训练样本[33-37]。由于连续帧之间的相机运动是未知的，因此，该方法既要估计目标图像的深度，还需要预测相机位姿。伦敦大学 Godard 等[33] 利用深度估计和姿态估计网络得到图像的逆深度估计和相机位姿估计，然后把相机位姿与视差计算的光度投影误差作为损失函数，利用梯度下降这种优化方法对损失函数中的每个误差进行优化或更新，以此来提升算法处理遮挡场景的鲁棒性。Li 等[34] 利用连续帧之间的时序约束进行自监督学习，该算法将自监督学习 VO 表示为一个序列学习问题，将帧间相关性表示为一个压缩码，并通过长短期记忆（LSTM）网络来集成序列信息。通过对抗学习这种方法，很好地解决了位姿估计过程中造成的误差积累，给系统后端提供了更精确的深度和更准确的位姿估计。Zhan 等[36] 将学习到的深度和光流预测整合到传统的 VO 测量模型中，获得了比其他算法更具竞争力的性能表现。此外，Li 等[37] 提出了基于元学习的在线自监督学习方法。

研究表明，与传统的单目 VO 或视觉惯导里程计相比，将深度学习与传统方法相结合的自监督方法在性能上更加优越[38]。这一结论从侧面说明了自监督领域发展的巨大潜力和无限可能。

2.2　深度学习与视觉惯导里程计

高精度的导航和定位是自动驾驶汽车的核心技术之一。传统的视觉里程计方法由于遮挡、尺度不确定性、相对位置偏移和低帧率等一系列问题，很难达到实际场景的应用需求；相比而言，惯性测量单元（IMU）定位设备价格低廉，可以直接获得运动主体的角速度和加速度的测量数据，达到理想的定位效果。因此，为了提升导航定位系统的精度和稳定性，在传统的 VO 中融入惯性信息是行之有效的方案，并已取得了一定成果[39-41]。

深度学习是一种端到端的学习方式，在模型训练时直接学习从输入的原始数据到期望输出的映射。与传统方法相比，基于深度学习的视觉惯性里程计（VIO）方法最大的优点是无需手动提取特征，完全依靠数据驱动，能利用数据本身蕴含的信息实

现深度预测。近年来，对该领域的探索与研究开始引起许多研究者的关注。

VINet 网络[15] 首次提出结合 IMU 的信息，通过深度神经网络的框架来解决 VIO 的问题。整个 VINet 网络利用 CNN 网络从 2 个相邻帧图像中提取视觉运动特征，同时使用 LSTM 网络来建模 IMU 的惯导特征。然后利用特殊欧氏群 $SE(3)$ 把视觉运动特征和惯导特征进行结合，以此实现对相机位姿的预测。通过 VINet 方法，既减少了对手动同步和校准的依赖，同时在同步误差方面也表现出了更强的鲁棒性。

文 [42] 利用在线纠错 OEC 模块进行了 VIO 无监督网络学习方法的设计。该方法在没有惯性测量单元内在参数或缺失 IMU 和相机之间的外部校准的情况下，将 RGB-D 图像与惯性测量直接相结合，根据像素的缩放图像投影误差的雅可比行列式生成相机运动的估计轨迹。DeepVIO 是 Han 等[43] 提出的一种端到端自监督深度学习网络框架，该框架主要使用双目序列来估计每个场景的深度和密集的 3D 几何约束并作为监督信号，结合 IMU 数据来获取绝对轨迹估计值。与传统方法相比，DeepVIO 减少了相机与 IMU 之间校准不正确、数据不同步和丢失的影响，与其他基于 VO 和 VIO 系统的最新学习方法相比，该算法在准确性和数据适应性方面的表现也更为突出。

基于深度学习的视觉惯性里程计方法已经被证明是成功的，然而，这些方法在设计过程中并没有完全解决多传感数据的鲁棒融合策略问题。针对这一问题，Chen 等[44] 提出一种新的单目端到端 VIO 多传感器选择融合策略。该策略融合了单目图像和惯性测量单元，根据外部环境和内部传感器的动态数据来估计运动轨迹，提高了对应用场景的鲁棒性。此外，还提出了不同掩码策略下的融合网络模式，在数据损坏的情况下，该融合策略表现出更优的性能。

在很多室内和室外场景中，面对不同的场景尺度因子，单目的 SLAM 系统需要对相机和 IMU 之间的空间变换和时间偏移进行标定。对这一限制问题，Lee 等[45] 利用光流神经网络的思想，以连续的 2 个相邻帧作为网络的输入，提出了一种不需要标定的 VIO 学习框架，该方法适用于计算能力不高且需要实时处理信息的 VIO 系统。为了解决单目视觉 SLAM 系统实时重构真实尺度场景困难的问题，浙江大学左星星博士提出了一种实时的 CodeVIO 方法[46]，采用一种新的、实时的单目相机惯导定位与稠密深度图重建的策略。该策略结合了深度神经网络与传统的状态估计器，利用轻量级的条件变分自动编码器（conditional variational autoencoder，CVAE），把高维度的稠密深度图在神经网络中编码为低维度的深度码，以增加稠密深度估计的准确性。CodeVIO 方法一方面利用 VIO 稀疏深度图的信息，以稀疏视觉特征点的深度作为神经网络的输入；另一方面使用了一种高效的网络雅可比矩阵计算方法，使网络在实时单线程运行的同时，具有了很强的泛化能力和高了一个数量级的计算效率。

此外，Liu 等[47] 提出 InertialNet 网络，训练端到端模型来推导图像序列和 IMU 信息之间的联系，预测相机旋转角度。Kim 等[48] 将不确定性建模引入无监督的损失函数中，在不需要用真值协方差作为标签的情况下学习多传感器间深度与位姿的不确定性。通过这种方法，克服了学习单个传感器时的不确定性和局限性。文 [49] 提出了一种新的基于深度学习模型的相机和 IMU 传感器融合的算法，以预测无人机系统的 3D 运动。

2.3 方法总结与对比分析

近年来，结合深度学习的视觉 SLAM 方法越来越受到研究者的高度关注。现有基于深度学习的 VO 估计方法的性能对比如表 1 所示。由于各算法的测试数据集和评估性能各有差异，难以对算法性能进行精确对比，因此表中仅列出了各算法在特定测试条件下的定位误差作为参考指标。特别强调，表中所列误差的性能指标值越小，说明算法的尺度一致性越佳，定位越准确。

结合表 1 性能，从现有的成果来看，深度学习在 SLAM 领域取得了一定的成果。与无监督学习方法相比，有监督学习方法表现出的尺度漂移误差更小、跟踪鲁棒性更佳。从算法的深度估计结果来说，目前提出的基于无监督 / 自监督的 VO 算法都能达到较好的预测效果。

值得一提的是，无监督学习是通过学习数据之间的规律来提取输入图像的特征，因此，能学习到更加丰富多样的图像特征表征，在未知的场景下具有更佳的适应性和泛化能力。

自监督学习方法既保留了传统算法的特点，又融合了深度学习的优势，能够较好地恢复场景的尺度，与无监督学习相比，具有更大的优势。如，自监督模型 D3VO 方法[38] 的跟踪精度甚至超过了现有的单目深度视觉里程计或视觉惯导里程计系统。当然，在特定限制的任务环境中，具体可以采用哪种学习方式还需要根据具体情况来决定。

表 1　现有基于深度学习的 VO 估计方法的性能对比

Tab.1　Performance comparison of existing VO estimation methods based on deep learning

模型 / 作者	传感器	监督方式	数据集	平均平移误差百分比 /%	旋转误差 /(°)	相对误差绝对值	相对误差平方	均方根误差
DeepVO [12]	单目	有监督	KITTI 10	9.04	0.0391	–	–	–
Costante 等 [14]	单目	有监督	KITTI 10	17.57	0.0319	–	–	–
Saputra 等 [17]	单目	有监督	KITTI 10	–		–	–	–
CL-VO [18]	单目	有监督	KITTI 10	8.29	0.0294	–	–	–
Godard 等 [19]	立体	无监督	KITTI 2015	–	–	–	–	–
SfMLearner [20]	单目	无监督	KITTI 07	17.52	5.38	0.208	1.768	6.856
UnDeepVO [21]	立体	无监督	KITTI 07	3.15	2.48	0.183	0.173	6.570
Bian 等 [22]	单目	无监督	KITTI 10	10.1	4.96	0.128	1.047	5.234
Zhan 等 [23]	单目	无监督	KITTI 10	12.45	3.46	0.135	0.905	4.366
GeoNet [24]	单目	无监督	KITTI 2015	35.6	–	0.147	0.936	4.348
GANVO [27]	单目	无监督	KITTI 2015	–	–	0.150	1.141	5.448
SGANVO [28]	单目	无监督	KITTI 10	5.89	3.56	0.065	0.673	4.003
Wong 等 [30]	立体	自监督	KITTI 2015	–	–	0.128	0.856	4.201
Chen 等 [31]	立体	自监督	KITTI 2015	–	–	0.052	0.558	3.733
Godard 等 [33]	单目	自监督	KITTI 2015	–	–	0.115	0.903	4.863
Li 等 [34]	单目	自监督	KITTI	5.564	–	0.146	0.927	4.107
D3VO [38]	单目	自监督	KITTI	0.62	–	–	–	4.485
Zhang 等 [36]	立体	自监督	KITTI 10	2.29	–	–	1.74	0.030

表 2　现有视觉惯导里程计融合算法的简要比较

Tab.2　Brief comparison of existing methods on visual-inertial odometry fusion algorithms

模型 / 作者	提出时间	性能指标	主要贡献
ORB-SLAM [2]	2015	KITTI 数据集测试，09 序列：t_{rel} 为 7.62；10 序列：t_{rel} 为 8.68	基于几何的 SLAM 方法，在大规模、小规模、室内室外的环境都可以运行
VINET [15]	2017	–	首次提出 VO 与 IMU 融合方案
VIOLearner [42]	2020	KITTI 数据集测试，t_{rel} 为 1.74；ATE 为 0.012	在没有 IMU 内在参数或外部校准的情况下生成运动轨迹
DeepVIO [43]	2019	KITTI 数据集测试，t_{rel} 为 0.85	在准确性和数据适应性方面表现突出
Chen 等 [44]	2019	KITTI 数据集测试，直接融合 RE 为 0.163；确定性软融合 RE 为 0.129；随机性硬融合 RE 为 0.140	提高了对实际应用场景的鲁棒性
Lee 等 [45]	2019	–	不需标定相机参数，适用于计算能力不高且需要实时处理信息的 VIO 系统
CodeVIO [46]	2021	EuRoC 数据集测试，RMSE 为 0.24	可提高稠密深度估计的准确性
Kim 等 [48]	2021	KITTI 数据集测试，未知的环境中，t_{rel} 为 6.41	在不需要真值协方差作为标签的情况下学习多传感器间深度和位姿的不确定性
HVIONet [49]	2022	EuRoC 数据集测试，RMSE 为 0.167	适用于未知室内环境，无需对输入图像进行校准调整

近年来，将惯性单元数据与相机的地标信息进行融合已成为构建高精度、高鲁棒 SLAM 系统的重要途径。部分现有基于视觉 / 惯性融合的视觉 SLAM 算法的总结如表 2 所示。表 2 中，t_{rel} 表示平均平移误差百分比，ATE 表示绝对轨迹误差，RE 表示旋转误差，RMSE 表示均方根误差。不难看出，基于学习的 VIO 的研究虽然才起步，但与传统的 SLAM 系统相比，其在定位精度、尺度一致性以及生成运动轨迹等方面的能力很突出。另外，IMU 和相机之间具有较强的互补性，将两者进行融合是提

升 SLAM 系统精度和鲁棒性的重要途径。

综上所述，深度学习在 SLAM 领域中的实际应用效果虽然还不是很理想，但是随着深度学习研究的深入，该领域已成为近年来的研究热门。

3 深度学习下的视觉 SLAM 后端优化（Visual SLAM backend optimization based on deep learning）

SLAM 的后端优化主要是对不同时刻视觉里程计预测得到的相机位姿信息以及局部地图进行优化调整。在 VO 中，不管是位姿估计还是建图，都是利用相邻帧之间的运动来完成的，这容易导致误差逐帧累积，最终产生较大的累积漂移[11]。在对这些区域进行地图重构时，将导致与同一区域已建图不重合，出现重影现象；同时，也有必要把所有地图数据放到一起再做一次全局的优化，以降低系统各部分的误差，提高系统的准确性。因此，为了降低误差漂移对 SLAM 系统性能带来的影响，后端优化就显得至关重要。

3.1 深度学习与回环检测

在视觉 SLAM 领域中，回环检测（loop closure detection）是又一个值得关注和研究的热点问题。其主要解决机器人位姿估计的累积漂移问题，以实现在大规模复杂环境下的精确导航。准确的回环检测可以进一步优化移动机器人的运动估计，建立全局一致的地图，反之则可能导致地图重建失败。因此，回环检测算法的好坏对整个视觉 SLAM 系统精度与鲁棒性的提升至关重要[11]。

早期的回环检测方法是手工标注特征点，应用词袋（BoW）模型来达到图像匹配的目的。随着深度学习、目标识别、语义分割等领域的迅速发展，研究者更倾向于使用先进技术来更好地实现回环检测。2015 年，国防科技大学张宏等[50]较早地将深度学习应用在回环检测中，利用 Caffe 深度学习框架下已经提前训练好的 AlexNet 模型产生一种适合回环检测的描述符。该方法先将图像输入到 CNN 中，以每个中间层的输出作为一个特征值，用来描述整幅图像，然后利用二范数进行特征匹配来确定是否存在回环。仿真结果表明在光照变化明显的环境下这种深度学习的特征描述符比传统的 BoW 和随机蕨法等方法更稳定、鲁棒性更强，并且产生描述符的用时更短。

自动编码器是一种无监督学习模型，能够自动提取数据中的有效特征，具有较强的泛化性。近些年，该方法受到了广泛的关注，且已成功应用于诸多领域。清华大学高翔等[51]提出采用堆叠去噪自动编码器（stacked denoising auto-encoder，SDA）的无监督学习方式描述整幅图像来进行图像的匹配，最终得到了较好的回环检测效果。此外，如文 [52] 也是在自动编码器结构的基础上，以无监督学习的方式压缩场景数据来提取紧凑的特征表示向量。

随着 CNN 训练的飞速发展，针对光照变化、天气变化和物体快速移动等复杂场景，有不少研究者开始考虑采用 CNN 网络学习特征与人工设计特征相结合的方式进行场景识别。文 [53] 在局部特征聚合描述子（VLAD）的基础上进行了扩展，提出了一种端对端的场景识别 NetVLAD 算法。此算法将传统的 VLAD 结构与 CNN 网络结构相结合，利用卷积网络的反向传播对网络进行算法优化，提高了对同类别图像的表达能力，同时大大地提高了图像的匹配精度。Bampis 等[54]提出了新的回环检测方法，主要通过旋转不变和尺度不变的局部特征描述向量以及动态序列识别技术来提高系统的性能。除此之外，文中还引入时间一致性过滤器来进一步提升所产生序列的相似性度量结果。参照文 [54] 的思路与方法，Memon 等[55]提出了有监督学习与无监督学习相结合的回环检测方法。文中利用深度学习在特征提取方面的优势，引入超级字典的概念，加快了场景比较的速度。同时，结合自动编码器对新场景进行回环检测，提高了回环检测的效率。

虽然，基于深度学习的回环检测方法可以从原始数据中自动地学习特征，能更充分地表达图像信息，对复杂的环境变化有更好的适应性和更强的鲁棒性，但是，如何针对不同场景自动选择不同隐含层的结果、如何找到更好的用于场景识别的特征、如何寻找合适的回环检测的性能评估基准等诸多问题依然是未来研究的重点。

3.2 深度学习与全局优化

SLAM 全局优化需要考虑的问题是如何利用不准确的关键帧建立起全局约束，以优化各帧的相机位姿。为了实现全局优化，可以通过建立和优化位姿图来求解各帧的相机位姿。位姿图是以关键帧的全局位姿作为图的节点，以关键帧之间的相对位姿误差作为图的边的权重，通过令整个图的所有边的权重值总和最小，来优化得到每个图节点的值。也可通过另一种目前比较主流的图优化方法来获得全局最优解。不论是何种优化方法，一般采用的求解器都是高斯－牛顿法或 LM 算法[11]。

深度学习的实质是利用观察到的相机位姿和场景表征来提取图像特征并构建映射函数。近年来，

研究者们针对如何将深度学习融入到全局优化问题中进行了探索与尝试，获得了比较好的性能优化结果。文 [56] 提出的 CNN-SLAM 法将 CNN 预测的稠密深度地图引入到直接单目 SLAM 法获得的深度测量值中，该方法使得 SLAM 系统在回环检测和图形优化方面具有更强的鲁棒性和更高的准确性。Zhou 等 [57] 提出了 DeepTAM 学习方法，其核心在于将来自 CNN 的相机位姿和深度估计引入到经典 DTAM 系统 [58] 中，然后通过后端全局优化，来实现更精确的相机位姿估计和场景重构。

基于无监督学习的单目视觉里程计，由于缺少累积误差的校正技术，在大规模里程计估计方面的精确度达不到预期目标。针对这一局限性，Li 等 [59] 将无监督学习的单目 VO 与图优化后端集成在一起，提出了一种混合的视觉里程计系统。以时间和空间光度损失作为主要监督信号，在系统后端，根据估计得到的局部闭环 6 自由度约束构建全局位姿图并进行优化，从而改善系统的定位精度和鲁棒性。除了文 [59] 的方法之外，DeepFactors 算法 [60] 也值得一提。文 [60] 中提出的深度 SLAM 系统是将学习到的稠密地图与 3 种不同类型的后端概率因子图相结合来实现的。该系统在概率框架中整合了一致性度量、先验学习等算法，在对位姿和深度变量进行联合优化的同时还能保持系统的实时性能。

目前，深度学习方法在全局优化中的应用处于初步探索阶段，随着各种深入研究的解决方案的提出与实现，深度学习在该领域的应用将会引来更多的关注。基于深度学习的全局优化方案也会得到进一步的提升和改进。

4　深度学习下的语义 SLAM（Semantic SLAM based on deep learning）

语义 SLAM 是语义信息和视觉 SLAM 的相互融合，其研究的核心就是对目标物体进行检测与识别。而深度学习算法是当前主流的物体识别算法。因此，在语义 SLAM 系统中引入深度学习成为 SLAM 系统发展的必然趋势。

而真正意义上的语义 SLAM（即语义建图和 SLAM 定位相互促进）发展相对较晚。2017 年，Bowman 等 [61] 引入了期望最大值方法来动态估计物体与观测的匹配关系。作者把语义 SLAM 转换成概率问题，利用概率模型计算出来的物体中心在图像上重投影时应该接近检测框的中心这一思想来优化重投影误差。虽然文 [61] 解决了语义特征的数据关联问题和如何用语义信息获取路标和摄像头位姿的问题，但是没有考虑语义元素之间的互斥关系，以及连续多帧的时序一致性。Lianos 等 [62] 提出的视觉语义里程计（VSO）方法是在文 [61] 的基础上，使用距离变换将分割结果的边缘作为约束，同时利用投影误差构造约束条件，从而实现中期连续点跟踪。

为了提高语义 SLAM 系统识别动态物体的准确性，清华大学的 Yu 等 [63] 在 2018 年 IROS 会议上提出了一种动态环境下鲁棒的语义视觉 SLAM 系统（DS-SLAM）。在 DS-SLAM 中，将语义分割网络放在一个单独运行的线程之中，结合语义信息和运动特征点检测，来剔除每一帧中的动态物体，从而提高位姿估计的准确性和系统运行的效率。动态环境下，此系统降低了对动态目标的影响，极大地提高了定位精度。同时，生成的密集语义八叉树地图可用于执行高级任务。但此方法要求所使用的语义网络运行速度足够快。

Kaneko 等 [64] 借用语义分割能将图像中每一类物体进行分类和标注这一特点，利用语义分割产生的掩模来排除不可能找到正确对应的区域。在检测特征点阶段，添加了“不检测掩蔽区域中的特征点”的操作，可以排除大部分获得的不准确的对应关系，减小了随机一致性采样误差。该方法引入了语义分割的全局信息，可以弥补视觉 SLAM 局部信息的不足，故具有较高的精度。

为了解决实际应用中的动态遮挡问题，文 [65] 提出了一种新颖的动态分割方法，从而实现对相机自我运动的准确跟踪。该方法首先将语义信息与对象级的几何约束相结合，快速提取出场景中的静态部分，再对静态部分从粗到细分两步实现精确跟踪。另外，对动态部分，提出了利用分层次掩码的动态物体掩码策略。相比于其他动态视觉 SLAM 方法，文 [65] 的方法在效率和动态跟踪精度等方面都有了明显的提升。

随着语义分割技术的发展，借助语义信息，将数据关联升级到物体级别，使得提升复杂场景下的识别精度成为了可能。目前，有许多研究（如文 [66-69]）都是基于物体级别关联的语义 SLAM 算法。2019 年，Yang 等 [66] 提出用于联合估计相机位姿和动态物体轨迹的 CubeSLAM 方法。该算法针对静态物体和动态物体分别采用不同的关联方法：对于静态物体，将 SLAM 提取到的特征点和 2D 检测框检测的对象关联起来；而对于动态物体，直接用稀疏光流算法来跟踪像素，动态特征的 3D 位置通过三角化测量来得到。数据关联过程

中，采用立方体在地图中表示物体。除了上述描述，还有学者提出用椭圆体（特殊双曲面）来表示物体[67-68]。但是椭圆体的物体表示只是一种近似，它的检测框和实际测量的检测框不可能完全重合，因此 QuadricSLAM 算法[67] 对精度提升并没有帮助，但采用 CubeSLAM 方法对其精度提升很大。DSP-SLAM 算法[69]的基础框架也是把一个物体级的 3 维重建算法加到一个传统 SLAM 算法中，其数据关联还是要用到特征点，也是在地图优化中加入物体与相机以及物体与地图点的约束。

在复杂多变的环境下，基于深度学习的语义信息具有光线不变性，因此语义分割下的定位比较稳定[70-71]。如，Stenborg 等[70] 通过结合深度学习去解决 SLAM 中的位置识别问题。其核心思想是在已有 3D 地图的基础上利用图像语义分割后得到的描述子代替传统描述子，然后再去建模，同时考虑 2D 点到 3D 点的映射关系。

虽然对语义 SLAM 已有不少初步探索，但由于其发展较晚，因此许多工作还仅处于起步阶段，很多问题还没有考虑，但可以预见未来几年这方面的研究会越来越多。

5 深度学习下的不确定性估计（Uncertainty estimation based on deep learning）

尽管深度神经网络在无人驾驶车辆控制或医学图像分析等高风险领域非常有吸引力，但它们在重视安全的现实生活中的应用仍然有限。而造成这种限制的主要原因是模型给出的预测结果并不总是可靠的。例如，在无人驾驶等对安全性要求较高的领域中，完全依赖深度模型进行决策有可能导致灾难性的后果。为此，有必要对基于深度学习的移动机器人的不确定性进行预测，以确保安全性。

一般地，模型中预测的不确定性大致可分为由模型引起的认知不确定性（模型不确定性）和由数据引起的任意不确定性（数据不确定性）[10]。近年来，很多研究者对捕捉深度神经网络（DNN）中的不确定性表现出越来越大的兴趣。贝叶斯模型就是预测认知不确定性的重要方法之一[72]。该方法使用随机失活方法（dropout 方法）来训练 DNN，训练得到的均值是预测值，而方差就是不确定度。本节重点讨论定位与建图过程中的不确定性估计和运动跟踪过程中的不确定性估计，以及这些不确定性估计的用途，表 3 对现有的深度学习下不确定性估计算法进行了总结。

在视觉 SLAM 系统中，定位或场景识别的不确定性是影响系统可信度的重要因素。语义分割是进行长期视觉定位或者场景理解的重要工具，有意义的不确定性度量对于决策至关重要。随着技术的发展，越来越多的工作对上述问题进行了探讨（如文 [73-76]），并获得了较理想的性能。

文 [73] 提出了一种基于信息理论的视觉 SLAM 特征选择方法 SIVO（semantically informed visual odometry and mapping），该方法将语义分割和神经网络不确定性引入到特征选择过程中，利用贝叶斯神经网络把特征的分类熵加到新的特征中，每一个被选择的特征都显著降低了车辆状态的不确定性，并多次被检测为静态对象（建筑物、交通标志等），且具有较高的置信度。根据这种选择策略生成稀疏地图，可以促进长期定位。

贝叶斯 SegNet 网络[74] 能够通过对场景模型不确定性的度量来预测场景像素级的不确定性，其核心思想是在 SegNet 网络结构的基础上增加随机失活层与贝叶斯决策。算法通过多次的前向运算得到多个输出结果，对这些结果求均值得到最终预测的分割结果；求对应位置像素的方差，得到模型的不确定性图。此外，该算法还可以使用蒙特卡洛算法来生成像素类标签的后验分布，并在多个预测的结果中找到最优的结果。

在实际应用中除了需要进行模型预测之外，也需要预测结果的置信度。利用神经网络学习给定输

表 3 现有的深度学习下不确定性估计算法

Tab.3 Existing methods of uncertainty estimation based on deep learning

模型 / 作者	提出时间	不确定性估计方法	预测对象	作用
SIVO[73]	2019	贝叶斯神经网络	模型不确定性预测	定位与建图
SegNet[74]	2015	贝叶斯 SegNet 网络	模型不确定性预测	定位与建图
Mcallister 等[75]	2017	贝叶斯深度学习	模型不确定性预测	定位与建图
ESP-VO[77]	2018	深度递归卷积神经网络	预测结果的不确定性	运动跟踪
Poggi 等[78]	2020	自学范式建模方法	预测结果的不确定性	运动跟踪
Eldesokey 等[79]	2020	自监督的概率归一化的卷积网络	预测结果的不确定性	运动跟踪

入的不确定性估计已受到越来越多研究者的重视。2018 年，Wang 等[77] 从深度学习的角度出发探讨了视觉里程计估计的不确定性，针对基于深度递归卷积神经网络的单目 VO，提出了一种端到端的序列间概率视觉里程计（ESP-VO）框架。通过这种方法，在不引入太多额外计算的情况下，可以有效地预测运动变换的不确定性。为了验证算法的有效性，文 [77] 在代表驾驶、飞行和步行情景的几个数据集上进行了广泛的验证实验。结果表明，基于这些最小化误差函数进行全局优化能减少系统的累积漂移，与其他先进的方法相比，所提出的 ESP-VO 具有竞争优势。

鉴于单目自监督网络在深度估计时不需要深度标注，因而越来越多的研究者开始致力于理解和量化自监督网络预测中深度不确定性的估计。2020 年，Poggi 等[78] 提出了一种新颖的不确定性估计方法，该方法用到 2 个网络：一个网络用于重建，主要利用翻转图像输入的方法和多个不同的模型对同一张图片的深度不确定性进行预测；另一个网络用来模拟重建网络生成的分布，通过自监督的方式学习一个可以预测不确定度的模型，其输出为不确定度。在位姿未知的情况下，该方法可以始终提高深度的估计准确度。另外，文 [79] 提出一种自监督的概率归一化的卷积网络，该方法可同时对深度与不确定度进行预测。一方面，对输入数据的不确定度进行估计，使得该网络可以基于数据可靠性进行针对性的学习；另一方面，提出概率归一化的卷积神经网络（NCNN），将训练过程转变为最大化似然估计问题，实现对输出不确定度的估计。

综上所述，在视觉 SLAM 中引入不确定估计后，可知模型对于预测结果的置信程度，有助于提高模型在实际场景环境中的应用性能。但目前关于该理论的研究才刚刚起步，其学习的方法较少，在实际场景下的适应性还有待进一步验证。

6　未来发展趋势（Future trends）

尽管基于深度学习的 SLAM 技术在精度和鲁棒性上已经表现出比传统 SLAM 方法更优的性能，解决方案也变得更有吸引力。但目前的研究仍处于初级阶段，所设计的模型还存在不足，故无法完全解决当前的问题。为了提高实际应用中的适用性和安全性，研究人员还将面临许多挑战。为此，文中讨论了几点可能助力该领域进一步发展的思路。

1) 适应性更强的数据集标注

深度学习严重依赖于海量的数据，如果想用这些数据来训练深度学习的模型，首先需要对它们进行处理与标注。从理想的角度看，标注的数据数量越多，训练得到的模型效果也会越好。但是，在实际标注过程中，不但需要结合实际的硬件资源与时间，还需要注意数据量的增大给模型效果提升带来的负面影响。数据标注的质量将直接影响训练得到的深度学习模型的可靠性。

综上所述，提高数据标注的质量也成为了该领域的研究重点。数据标注是一个耗费成本与时间的过程，经济、高效地完成数据标注，这是研究人员必须面对和解决的难题。如何在成本与质量这两者之间找到一个平衡就显得尤为重要。同时，期望未来能够利用 SLAM 方法来构建图像之间存在对应关系的大规模的数据集，这可能有助于解决数据标注问题。

2) 深度学习模型的拓展

目前，许多基于深度学习的模型，如卷积神经网络、长短期记忆网络和自动编码器等都是端到端的学习方式。尽管这些模型的快速发展提升了系统的鲁棒性和准确性，但在实际应用场景中，许多数据是从非欧氏空间生成的，而传统的端到端的深度学习方法对此类数据的处理能力却难以使人满意。

近几年，越来越多的学者对深度学习方法在图数据上的扩展产生了浓厚的兴趣。用于处理图数据的图神经网络（graph neural network）[80] 由此应运而生。从本质上讲，图神经网络是几何深度学习的一部分，主要是将端到端学习与归纳推理相结合，研究具有结构属性、拓扑性质的数据的学习和预测任务。因此，对于图神经网络结构的深入研究有助于解决深度学习无法处理的关系推理和组合泛化的问题，是未来一个新的研究热点。

3) 多传感器融合算法的研究

在现实生活中，移动机器人或硬件设备往往不仅仅只携带一种传感器，而是多种传感器相互配合使用。不同传感器的最远探测距离、精度、功能等各不相同，因此在使用多种传感器的情况下，要想保证系统决策的可靠性和快速性，就必须对传感器进行信息融合。例如，手机 VIO 系统就是通过融合 IMU 数据和相机信息，弥补了单一传感器的不足，为实现 SLAM 的小型化和低成本提供了行之有效的研究方向。DeLS-3D 设计[81] 融合了相机视频、运动传感器（GPS/IMU）等数据和 3 维语义地图，可以提升 SLAM 系统的鲁棒性和效率。上述例子表明，将多种具有互补性的传感器进行融合是提升 SLAM 系统精度和鲁棒性的重要途径。

多传感器融合的软硬件难以分离。当前，在硬件层面实现多传感器融合并不难，重点和难点在于如何实现算法和传感器之间的融合。另外，动态与未知环境下的融合问题也将是多传感器融合面临的另一个难题。相信随着技术的不断发展，算法融合问题将会得到很好的解决，多传感器融合技术也许很快就会在实际生活中得到广泛应用。

7 结论（Conclusion）

从已有的大量研究可以看出，基于深度学习的 SLAM 方法虽然是一个刚起步且在不断发展的研究领域，但是已逐渐引起了研究者的广泛关注。到目前为止，深度学习与 SLAM 的结合已经在视觉里程计、场景识别与全局优化等各种任务中取得了显著的成果。同时，由于深度神经网络具有强大的非线性拟合能力，可以任意逼近人工建模难以模拟的非线性函数，因此在实际应用中鲁棒性更佳。此外，语义信息与传统视觉 SLAM 算法的集成有助于提高对图像特征的理解，对构建高精度的语义图也产生了重要影响。基于深度学习的 SLAM 技术的快速发展为移动机器人向实用化、系列化、智能化发展提供了助力。

参考文献（References）

[1] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part I[J]. IEEE Robotics & Automation Magazine, 2006, 13(2): 99-110.

[2] 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.

[3] Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2015: 2938-2946.

[4] 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.

[5] 敬学良，王晨升，杨光，等. 深度学习在视觉 SLAM 研究中的应用综述[J]. 中国科技论文在线精品论文，2019，12(6)：872-878.
Jing X L, Wang C S, Yang G, et al. Application review of deep learning in visual SLAM research[J]. Highlights of Sciencepaper Online, 2019, 12(6): 872-878.

[6] 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 S P, Zhang T. A survey of deep learning application in visual SLAM[J]. Aerospace Control and Application, 2019, 45(2): 1-10.

[7] 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 S, Zhang C, et al. A survey on visual SLAM based on deep learning[J]. Journal of System Simulation, 2020, 32(7): 1244-1256.

[8] Chen C H, Wang B, Lu C X, et al. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence[DB/OL]. (2020-02-09) [2022-10-11]. https://doi.org/10.48550/arXiv.2006.12567.

[9] Gawlikowski J, Tassi C R N, Ali M, et al. A survey of uncertainty in deep neural networks[DB/OL]. (2022-01-18) [2022-10-11]. https://doi.org/10.48550/arXiv.2107.03342.

[10] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision?[C]//31st International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017: 5580-5590.

[11] 高翔，张涛，刘毅，等. 视觉 SLAM 十四讲：从理论到实践[M]. 北京：电子工业出版社，2017.
Gao X, Zhang T, Liu Y, et al. 14 lectures on visual SLAM: From theory to practice[M]. Beijing: Publishing House of Electronics Industry, 2017.

[12] Wang S, Clark R, Wen H K, et al. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 2043-2050.

[13] Konda K, Memisevic R. Learning visual odometry with a convolutional network[C]//10th International Conference on Computer Vision Theory and Applications. Setubal, Portugal: SciTePress, 2015: 486-490.

[14] Costante G, Mancini M, Valigi P, et al. Exploring representation learning with CNNs for frame-to-frame ego-motion estimation [J]. IEEE Robotics and Automation Letters, 2016, 1(1): 18-25.

[15] Clark R, Wang S, Wen H K, et al. VINET: visual-inertial odometry as a sequence-to-sequence learning problem[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 3995-4001.

[16] Turan M, Almalioglu Y, Araujo H, et al. Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots[J]. Neurocomputing, 2018, 275: 1861-1870.

[17] Saputra M R U, Gusmao P, Almalioglu Y, et al. Distilling knowledge from a deep pose regressor network[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 263-272.

[18] Saputra M R U, de Gusmao P P B, Wang S, et al. Learning monocular visual odometry through geometry-aware curriculum learning[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 3549-3555.

[19] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 6602-6611.

[20] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 6612-6619.

[21] Li R H, Wang S, Long Z Q, et al. UnDeepVO: Monocular visual odometry through unsupervised deep learning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 7286-7291.

[22] Bian J W, Li Z C, Wang N Y, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video [C]//33rd Annual Conference on Neural Information Processing Systems. Vancouver，USA: Curran Associates, 2019: 35-45.

[23] Zhan H Y, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 340-349.

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

[25] Zhao C, Sun L, Purkait P, et al. Learning monocular visual odometry with dense 3D mapping from dense 3D flow[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 6864-6871.

[26] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//27th Conference on Neural Information Processing Systems. New York, USA: ACM, 2014: 2672-2680.

[27] Almalioglu Y, Saputra M R U, de Gusmão P P B, et al. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 5474-5480.

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

[29] Vijayanarasimhan S, Ricco S, Schmid C, et al. SfM-Net: Learning of structure and motion from video[DB/OL]. (2017-04-25) [2022-11-22]. https://doi.org/10.48550/arXiv.1704.07804.

[30] Wong A, Soatto S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 5637-5646.

[31] Chen Z, Ye X Q, Yang W, et al. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 15509-15518.

[32] Choi H, Lee H, Kim S. Adaptive confidence thresholding for monocular depth estimation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 12788-12798.

[33] Godard C, Aodha O M, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 3827-3837.

[34] Li S K, Xue F, Wang X, et al. Sequential adversarial learning for self-supervised deep visual odometry[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 2851-2860.

[35] Zou Y L, Ji P, Tran Q H, et al. Learning monocular visual odometry via self-supervised long-term modeling[M]//Lecture Notes in Computer Science, Vol.12359. Berlin, Germany: Springer, 2020: 710-727.

[36] Zhan H Y, Weerasekera C S, Bian J W, et al. Visual odometry revisited: What should be learnt?[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 4203-4210.

[37] Li S K, Wang X, Cao Y D, et al. Self-supervised deep visual odometry with online adaptation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 6338-6347.

[38] Yang N, von Stumberg L, Wang R, et al. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 1278-1289.

[39] Bloesch M, Omari S, Hutter M, et al. Robust visual inertial odometry using a direct EKF-based approach[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2015: 298-304.

[40] Qin T, Li P L, Shen S J, et al. VINS-Mono: A robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[41] Mur-Artal R, Tardós J D. Visual-inertial monocular SLAM with map reuse[J]. IEEE Robotics and Automation Letters, 2017, 2(2): 796-803.

[42] 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.

[43] Han L M, Lin Y M, Du G G, et al. DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2019: 6906-6913.

[44] Chen C H, Rosa S, Miao Y S, et al. Selective sensor fusion for neural visual-inertial odometry[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 10534-10543.

[45] Lee H Y, McCrink M, Gregory J W. Visual-inertial odometry for unmanned aerial vehicle using deep learning[C]//AIAA Scitech 2019 Forum: Intelligent/Autonomous Guidance and Navigation. Reston, USA: AIAA, 2019: 1410-1431.

[46] Zuo X X, Merrill N, Li W, et al. CodeVIO: Visual-inertial odometry with learned optimizable dense depth[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2021: 14382-14388.

[47] Liu T A, Lin H Y, Lin W Y. InertialNet: Toward robust SLAM via visual inertial measurement[C]//IEEE Intelligent Transportation Systems Conference. Piscataway, USA: IEEE, 2019: 1311-1316.

[48] Kim Y, Yoon S, Kim S, et al. Unsupervised balanced covariance learning for visual-inertial sensor fusion[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 819-826.

[49] Aslan M F, Durdu A, Yusefi A, et al. HVIONet: A deep learning based hybrid visual-inertial odometry approach for unmanned aerial system position estimation[J]. Neural Networks, 2022, 155: 461-474.

[50] Hou Y, Zhang H, Zhou S L. Convolutional neural network-based image representation for visual loop closure detection[C] //IEEE International Conference on Information and Automation. Piscataway, USA: IEEE, 2015: 2238-2245.

[51] Gao X, Zhang T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. Autonomous Robots, 2017, 41: 1-18.

[52] Merrill N, Huang G Q. Lightweight unsupervised deep loop closure[C]//Robotics: Science and Systems. Cambridge, USA: MIT, 2018: 26-30.
[53] Arandjelović R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1437-1451.
[54] Bampis L, Amanatiadis A, Gasteratos A. Fast loop-closure detection using visual-word-vectors from image sequences[J]. International Journal of Robotics Research, 2018, 37(1): 62-82.
[55] Memon A R, Wang H S, Hussain A. Loop closure detection using supervised and unsupervised deep neural networks for monocular SLAM systems[J]. Robotics and Autonomous Systems, 2020, 126. DOI: 10.1016/j.robot.2020.103470.
[56] Tateno K, Tombari F, Laina I, et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 6565-6574.
[57] Zhou H Z, Ummenhofer B, Brox T. DeepTAM: Deep tracking and mapping with convolutional neural networks[J]. International Journal of Computer Vision, 2020, 128: 756-769.
[58] Newcombe R A, Lovegrove S J, Davison A J. DTAM: Dense tracking and mapping in real-time[C]//International Conference on Computer Vision. Piscataway, USA: IEEE, 2011: 2320-2327.
[59] Li Y, Ushiku Y, Harada T. Pose graph optimization for unsupervised monocular visual odometry[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 5439-5445.
[60] 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.
[61] Bowman S L, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic SLAM[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 1722-1729.
[62] Lianos K N, Schönberger J L, Pollefeys M, et al. VSO: Visual semantic odometry[C]//Lecture Notes in Computer Science, Vol.11208. Berlin, Germany: Springer, 2018: 246-263.
[63] Yu C, Liu Z X, Liu X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 1168-1174.
[64] Kaneko M, Iwami K, Ogawa T, et al. Mask-SLAM: Robust feature-based monocular SLAM by masking using semantic segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, USA: IEEE, 2018: 371-379.
[65] Bao R Q, Komatsu R, Miyagusuku R, et al. Stereo camera visual SLAM with hierarchical masking and motion-state classification at outdoor construction sites containing large dynamic objects[J]. Advanced Robotics, 2021, 35(3/4): 228-241.
[66] Yang S C, Scherer S. CubeSLAM: Monocular 3-D object SLAM[J]. IEEE Transactions on Robotics, 2019, 35(4): 925-938.
[67] Nicholson L, Miford M, Sünderhauf N. QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM[J]. IEEE Robotics and Automation Letters, 2019, 4(1): 1-8.
[68] Hosseinzadeh M, Latif Y, Pham T. Structure aware SLAM using quadrics and planes[M]//Lecture Notes in Computer Science, Vol.11363. Berlin, Germany: Springer, 2018: 410-426.
[69] Wang J W, Rünz M, Agapito L. DSP-SLAM: Object oriented SLAM with deep shape priors[C]//International Conference on 3D Vision. Piscataway, USA: IEEE, 2021: 1362-1371.
[70] Stenborg E, Toft C, Hammarstrand L. Long-term visual localization using semantically segmented images[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 6484-6490.
[71] Liu Y, Petillot Y, Lane D, et al. Global localization with object-level semantics and topology[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 4909-4915.
[72] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning[C]//33rd International Conference on Machine Learning. Cambridge, USA: MIT, 2016: 1050-1059.
[73] Ganti P, Waslander S L. Network uncertainty informed semantic feature selection for visual slam[C]//16th Conference on Computer and Robot Vision. Piscataway, USA: IEEE, 2019: 121-128.
[74] Kendall A, Badrinarayanan V, Cipolla R. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding[DB/OL]. (2016-10-10) [2022-11-11]. https://doi.org/10.48550/arXiv.1511.02680.
[75] McAllister R, Gal Y, Kendall A, et al. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning[C]//26th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann, 2017: 4745-4753.
[76] Klodt M, Vedaldi A. Supervising the new with the old: Learning SFM from SFM[M]//Lecture Notes in Computer Science, Vol.11214. Berlin, Germany: Springer, 2018: 713-728.
[77] Wang S, Clark R, Wen H K, et al. End-to-end, sequence-to- sequence probabilistic visual odometry through deep neural networks[J]. International Journal of Robotics Research, 2018, 37(4/5): 513-542.
[78] Poggi M, Aleotti F, Tosi F, et al. On the uncertainty of self-supervised monocular depth estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 3224-3234.
[79] Eldesokey A, Felsberg M, Holmquist K, et al. Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 12011-12020.
[80] Battaglia P W, Hamrick J B, Bapst V, et al. Relational inductive biases, deep learning, and graph networks[DB/OL]. (2018-10-17) [2022-11-11]. https://doi.org/10.48550/arXiv.1806.01261.
[81] Wang P, Yang R G, Cao B B, et al. DeLS-3D: Deep localization and segmentation with a 3D semantic map[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 5860-5869.

作者简介：

黄泽霞（1977 –），女，博士，讲师。研究领域：人工智能，图像识别等。
邵春莉（1989 –），女，博士，讲师。研究领域：检测技术，信号处理等。