视觉 SLAM 的研究现状与展望 *

吴 凡，宗艳桃，汤霞清

(陆军装甲兵学院 兵器与控制系，北京 100072)

摘 要：针对自主定位与环境构建问题，基于视觉传感器的同时定位与地图构建(SLAM)成为现阶段研究的热点。为深入分析视觉 SLAM 的现状，综述其相关算法与成果。首先简要概述了视觉 SLAM 的概念、特点与研究意义；然后深入分析帧间估计算法，详细描述经典的帧间估计方法，其中包含基于特征点的方法、基于光流的方法、与直接法，并介绍经典视觉 SLAM 算法的标志性成果；之后按照有监督的学习与无监督的学习两种方式介绍深度学习在视觉 SLAM 中的研究进展，并对算法进行归纳总结；此外分析了视觉 SLAM 和惯性导航的融合；最后展望了视觉 SLAM 的未来发展趋势。

关键词：SLAM；帧间估计；深度学习；神经网络

中图分类号：TP242.6 doi: 10.19734/j.issn.1001-3695.2019.02.0035

Research status and prospect of vision SLAM

Wu Fan, Zong Yantao, Tang Xiaqing

(*School of Weapons & Control, Army Academy of Armored Forces, Beijing* 100072, *China*)

Abstract: Aiming at the problem of autonomous localization and environmental reconstruction, simultaneous localization and mapping (SLAM) based on visual sensor has become a hot research topic at this stage. In order to deeply analyze the current situation of visual SLAM, the related algorithms and results are summarized. Firstly, the concept, characteristics and research significance of visual SLAM are briefly summarized; then, the visual odometry algorithm is analyzed in depth, and the classical visual odometry methods are described in detail, which include feature-based method, optical flow-based method and direct method, and the landmark achievements of classical visual SLAM algorithm are introduced; secondly, Introduce the research progress visual SLAM based on deep learning according to supervised learning and unsupervised learning ; thirdly, summarizes the algorithms; in addition, analyses the integration of visual SLAM and inertial navigation; finally, the future development trend of visual SLAM is proposed.

Key words: simultaneous localization and mapping; frame to frame estimation; deep learning ; neural network

0 引言

视觉 SLAM (simultaneous localization and mapping) [1]是指通过视觉传感器的方式实现同时定位与地图构建。如图 1 所示，视觉 SLAM 首先从视觉传感器得到信息；视觉里程计[3]，即帧间估计估算相邻图片之间的相机运动；通过回环检测[5]判断相机是否到过之前的位置；将视觉里程计与回环检测的内容送入后端进行优化；根据估计的相机轨迹与姿态实现场景重建。

视觉传感器具有体积小、重量轻、小功耗特点，不仅可以用于自定位，还可以用于目标检测、跟踪、障碍物识别等任务。通过视觉传感器实现相机姿态回归、地图构建成为目前研究的热点内容。

视觉传感器按照工作的方式主要分为单目相机、双目相机[8]和 RGB-D[9]相机。双目相机由两个单目相机组成。深度相机需要的红外结构光在外界环境中易失效或受到影响。在一些应用场景下，视觉 SLAM 的研究聚焦在单目视觉[10,11]上，通过单目相机实现相机姿态回归与地图构建。本文主要介绍单目相机的帧间估计方法。

视觉 SLAM 主要应用于如无人机[12]、无人车等无人作战平台上，以及增强现实(AR)与虚拟现实(VR)中。经典的视觉 SLAM 方法包括基于特征点的方法、光流法、直接法。随着信息时代的发展，基于深度学习的视觉 SLAM 得到了进一步的重视与研究。为了提高视觉 SLAM 的导航精度，将视觉 SLAM 与 INS 进行组合[13]也成为视觉 SLAM 的重要研究内容。

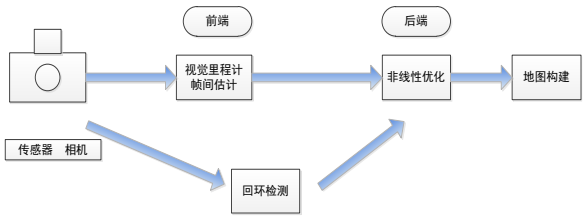

图 1 视觉 SLAM 工作流程

Fig. 1 Visual SLAM workflow

1 几何法

帧间估计依据的图像特征不同，回归相机位姿的方法也不同。经典的帧间估计方法主要包括基于特征点匹配的方法、光流法和直接法。特征点法在相邻的帧图像间采用多视几何的方法提取特征点，由特征点间几何关系回归相机的位姿。

收稿日期：2019-02-10；**修回日期：**2019-03-23 **基金项目：**武器装备军内重点科研资助项目

作者简介：吴凡(1995-)，女，河北保定人，硕士研究生，主要研究方向为深度学习、机器视觉；宗艳桃(1983-)，男，河北石家庄人，讲师，硕导，博士，主要研究方向为图像处理、目标跟踪与识别、深度学习(240089545@qq.com)；汤霞清(1965-)，男，湖南长沙人，教授，博导，博士，主要研究方向为导航定位技术.

光流法以追踪光流的方法辅助特征点的提取。直接法根据图片中像素的亮度信息直接回归相机的位姿。

1.1　特征点法

基于特征点匹配的帧间估计首先在相邻的图像中选取角点、边缘点、区块等比较有代表性的点，在这些显著可重复的特征基础上估计相机的运动。研究人员设计了很多具有稳定性的特征提取与匹配的算法。

1)SIFT (scale-invariant feature transform) [14]算法

在最早设计并大范围应用的 SIFT 特征中，虽充分的考虑了图像的光照、尺度等变化，却有很大的计算量。它的特征提取步骤如下：

a)对图像的位置与尺度进行探测，计算两个相邻的由乘法因子 k 分离的尺度空间的差值，构造高斯差分尺度空间。

$$\begin{aligned} &D(\mathrm{x},\ \mathrm{y},\ \sigma)\\ &=(G(\mathrm{x},\ \mathrm{y},\ \mathrm{k}\sigma)\text{-}G(\mathrm{x},\ \mathrm{y},\ \sigma))*I(\mathrm{x},\ \mathrm{y})\\ &=\mathrm{L}(\mathrm{x},\ \mathrm{y},\ \mathrm{k}\sigma)\text{-}\mathrm{L}(\mathrm{x},\ \mathrm{y},\ \sigma) \end{aligned} \tag{1}$$

其中：$G(\mathrm{x},\ \mathrm{y},\ \sigma)$ 为高斯函数，$G(\mathrm{x},\ \mathrm{y},\ \sigma)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{-(\mathrm{x}^2+\mathrm{y}^2)/2\sigma^2}$；$I(\mathrm{x},\ \mathrm{y})$ 为输入图像；$L(\mathrm{x},\ \mathrm{y},\ \sigma)$ 为图像的尺度空间函数。

b)检测尺度空间的极值点，将每个采样点与相邻点比较，选取极值点作为特征点。

c)对关键点实现精确的定位，基于局部图像的梯度分配方向信息。

d)将关键点附近区域上测量的梯度转换为一个允许局部形状和光照变化的关键点描述。

2)SURF (speeded up robust features) [15]算法

SURF 算法是一种新型的具有尺度和旋转不变性的关键点检测和描述方法。该方法与之前的特征提取与匹配的方法如 SIFT 相接近甚至更好，计算和匹配的速度更快。该方法首先构建 Hessian 矩阵：

$$H(\mathrm{f}(\mathrm{x},\ \mathrm{y}))=\begin{bmatrix} \frac{\partial^2 \mathrm{f}}{\partial \mathrm{x}^2} & \frac{\partial^2 \mathrm{f}}{\partial \mathrm{x}\partial \mathrm{y}} \\ \frac{\partial^2 \mathrm{f}}{\partial \mathrm{x}\partial \mathrm{y}} & \frac{\partial^2 \mathrm{f}}{\partial \mathrm{y}^2} \end{bmatrix} \tag{2}$$

计算原图像的 H 矩阵的行列式的近似值构成变换图，在变换图上利用非极大值抑制初步选择关键点，由三维线性插值法精确定位关键点，选取特征点的主方向，在关键点周围取方框统计小波特征建立关键点的描述。

3)ORB (oriented FAST and rotated BRIEF) [16]算法

特征点的选取与匹配一直依赖于功耗巨大的特征向量。ORB 算法采用 FAST 关键点和 BRIEF 描述子，具有旋转不变性和抗噪声的特性。ORB 在标准的 CPU 计算上，计算效率高于 SIFT 与 ORB。其选定一个合适的阈值，在图像中选择一个像素，如果该像素周围 16 个像素与其相比，连续 12 个都大于阈值，则选取为一个关键点。BRIEF 描述子采用快速二值特征向量，在关键点周围取 n 个点对，将 n 个点对的比较结果组合起来作为描述子。例如以关键点 p 为圆心，以 d 为半径做圆 O，在圆 O 内以某一模式选取 4 个点对标记为 $P_1(A,B)$，$P_2(A,B)$，$P_3(A,B)$，$P_4(A,B)$，定义操作 T：

$$T(P(A,B))=\begin{cases} 1 & I_A \rangle I_B \\ 0 & I_A \le I_B \end{cases} \tag{3}$$

其中：I_A 表示点 A 的灰度，分别对已选取的点进行 T 操作，将得到的结果进行组合得到最终的描述子。将两个二进制串组成的描述子进行异或操作即可得到描述子之间的相似度。

特征匹配后进行误匹配、运动目标等外点排除[17]，与运动估计。单目相机基于特征点方法估计相机运动最早是 Davison 等人[20~23]的 MonoSLAM 在扩展卡尔曼滤波的方法上，追踪非常稀疏的特征点。随后，Klein [24]提出了 PTAM 方法将跟踪与建图并行化，使用非线性优化，达到增强现实的效果。目前，处于顶峰优势的是 ORB-SLAM[25]，系统包含了 SLAM 的共有模块跟踪、建图、重定位、闭环检测，对剧烈运动鲁棒性好。针对单目相机基于特征点进行帧间估计的思路为：将获得的 2D 图片流据对极几何约束[26]进行初始化，由针孔相机模型得到两个像素点 p_1, p_2 的位置坐标，转换得出对极约束关系，求出基础矩阵 E 及旋转 R 和平移 T；然后通过三角测量在两处观察同一个点的夹角得到特征点的相对深度；初始化完成后，利用得到的三维图像与下一帧二维图像采用 P3P[27]的方法进行相机位姿的估计，据三角形相似的特点用余弦定理组成二元二次方程组，求解投影点的 3D 坐标。近年来 Camposeco 等人[28]还提出了 2D-3D 匹配和 2D-2D 匹配的混合位姿估计方法。Mur-Artal 等人[29]提出 ORB-SLAM2 在 ORB-SLAM 基础上支持标定后的双目及 RGB-D 相机，选取不用 GPU 加速就可以实时且具有视点和光照不变性的 ORB 特征描述符，通过每一帧图像定位相机选择是否加入关键帧，使用局部调整处理新的关键帧，使用 Covisibility Graph 进行跟踪建图， Essential Graph 优化位姿并回环检测。陆建伟等人[30]基于 ORB-SLAM2 实现了实时网格地图构建。

但是基于特征点进行位姿估计在给定的配对点多的情况下无法利用更多的信息，造成信息的浪费，在有噪声影响的情况下会造成匹配错误导致算法失效。设计算法提取合适的特征点并实时准确的匹配是基于特征点的位姿估计方法的基石。

1.2　光流法

基于特征点的方法中关键点的提取和描述子的计算非常耗时；将提取出的特征点作为图像的代表也忽略了很多信息；有些图片中纹理单一，没有明显的特征点可以提取。面对这些问题时，使用光流法追踪并描述像素在光流之间的运动，其中稀疏光流法选取图像中的部分像素点进行运动跟踪；稠密光流法对图像中的所有像素点进行跟踪。其中，以 LK 稀疏光流法[31]为代表，示意如图 2 所示。光流法基于灰度不变假设：

$$I(x+dx,y+dy,t+dt)=I(x,y,t) \tag{4}$$

对上述进行泰勒展开：

$$\frac{\partial I}{\partial x}\frac{dx}{dt}+\frac{\partial I}{\partial y}\frac{dy}{dt}=-\frac{\partial I}{\partial t} \tag{5}$$

其中：$\frac{dx}{dt}$ 是 x 方向的速度，记为 u；$\frac{dy}{dt}$ 是 y 方向上的速度，记为 v；$\frac{\partial I}{\partial x}$ 为图像在该点处 x 方向的梯度，$\frac{\partial I}{\partial y}$ 为图像在该点处 y 方向的梯度，分别记为 I_x, I_y；$\frac{\partial I}{\partial t}$ 为图像对时间的变化量 I_t，则得到

$$\begin{bmatrix} I_x & I_y \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix}=-I_t \tag{6}$$

假设某一框内的 w^2 个像素具有相同的运动，得到像素点的运动的线性方程：

$$\begin{bmatrix} u \\ v \end{bmatrix}=-(A^TA)^{-1}A^Tb \tag{7}$$

其中：A 为所有像素的 I_x, I_y 组成的集合，利用最小二乘法求解像素在图像间运动的速度。在图像平面单一或者特征点跟踪失败，光流可以追踪角点等特征点。通过光流追踪算法过

程可以得到特征点的对应关系，减少描述子的计算与匹配过程。

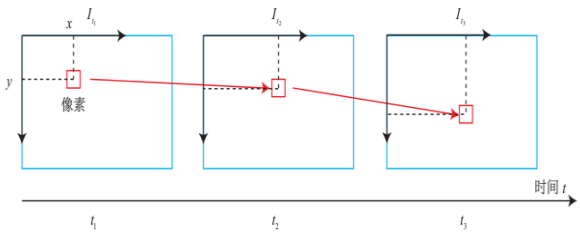

图 2　光流示意图

Fig. 2　Optical flow diagram

Maity 检测图像中的边缘点使用光流法跟踪并利用三视点几何关系来细化、优化对应匹配关系[32]。边缘点具有很强的鲁棒性，在无纹理与纹理较少的环境下可以可靠地工作。王泽民等人[33]也提出了一种基于 LK 光流的 SLAM 新方法，据光流大小与阈值的比较减少移动物体的影响。

1.3　直接法

直接法不提取关键点、不计算描述子，可以在图像有光度的变化的无纹理情况直接回归相机的姿态。空间中的某点 P 在两帧中的投影 P_1,P_2，依据于 P_1 的亮度 $I_1(P_1)$ 与 P_2 的亮度 $I_2(P_2)$，基于灰度不变的假设，求得与相机位姿相关的光度误差二范数作为优化目标，将其转换为误差最小化下的相机位姿估计问题。其中，P 来源于稀疏的关键点进行稀疏的重构为稀疏直接法；P 来源于有梯度的像素点，舍弃图像中像素梯度不明显的像素点为半稠密直接法(semi-dense)；P 来源于图像中的所有像素点为稠密直接法。

LSD-SLAM (large scale direct monocular SLAM)[34]是直接法在单目 SLAM 中成功的典例，其算法流程如图 3 所示。LSD-SLAM 分为跟踪(tracking)、深度图预测(depth map estimation)和地图优化(map optimization)三部分。在跟踪部分，使用方差归一化的光度误差，求解参考关键帧和新图像帧之间的相机运动李代数；而后，通过与被跟踪的图像帧的距离判断是否构建新的关键帧，若构建关键帧则构建新的深度图；若不构建关键帧则更新当前的深度图，根据标准选出一些像素进行小基线立体匹配，对深度地图正则化。地图优化，也称为建图一致性约束，使用关键点提供初始值，通过帧与帧之间双向跟踪实现闭环检测，解决尺度漂移的问题并进行全局优化。

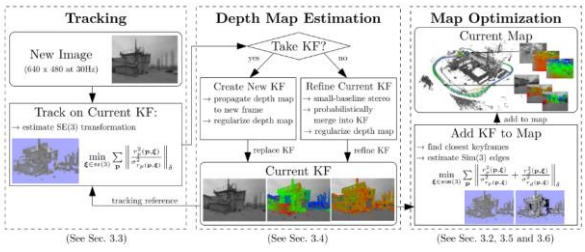

图 3　LSD-SLAM 算法流程[21]

Fig. 3　LSD-SLAM algorithm

2　深度学习法

深度学习采用端到端(end-to-end)的方式实现并进行大量的自动完成。基于深度学习的帧间估计实用性更强，不依赖于关键点的选取与特征的描述，可以规避图像中方向、形状、光照等变化带来的影响。基于深度学习的视觉 SLAM[35~38]引起了研究人员极大的重视。

从学习方式分为有监督学习(supervised learning)与无监督学习(unsupervised learning)。在有监督学习中，每一个样本都有着明确的标签，训练总结出样本与标签之间的映射关系；而无监督学习，在训练数据没有标签的情况下，采用聚类的思想找出其内在蕴涵的关系，把数据分组为多个类的过程。

2.1　有监督的深度学习法

最早，Roberts 等人[39]提出基于记忆的学习来估计移动机器人自我运动的技术，学习稀疏光流到速度和旋转的关系。将图片分为几个单元，计算每个单元的光流作为输入。虽然此方法还不能像几何法那样精确地估计相机的自我运动，但是在场景结构不明显或没有相机标定的情况下，依然可以学习相机和环境之间的关系。随后，Roberts 等人[40]还提出了用 EM 算法优化之前的结果，减少相机位姿的误差，并研究广义成像系统的概率线性子空间约束估计稠密光流和相机自运动。VGuizilini 等人[41,42]还提出了输入光流图像，采用高斯耦合的方法对相机的姿态进行回归。

PoseNet[43]第一次利用卷积神经网络(CNN)实现相机姿态回归，训练卷积神经网络从一个单一的彩色图像，用端到端的方式回归六自由度相机的姿态，并且不需要额外的工程或图形优化。其采用从运动估计结构创建大的回归数据集 Cambridge Landmarks，并通过迁移学习首先预训练为一个分类器来训练姿态回归器。该算法可以实时地在室内和室外进行操作，以每帧 5 ms 的速度达到在室外大型场景约 2 m 和 3° 的精度，在室内 7Scences 数据集[44]上 0.5 m 和 5° 的精度。网络从高层特征进行定位，对于光照变化、运动模糊及特征点匹配失败等情况具有鲁棒性。算法同时学习位置和旋转，采用随机梯度下降来训练欧几里德损失，得到网络参数。目标损失函数如下：

$$Loss(I)=\left\|\hat{x}-x\right\|_2+\beta\left\|\hat{q}-\frac{q}{\|q\|}\right\|_2 \tag{8}$$

PoseNet 对 GoogLeNet[45]网络结构改变如图 4 所示。用回归替换分类，对每个最终全连接层进行改进，以输出一个 7 维的姿态向量，其中 3 维表示位置 4 维表示旋转；在特征尺寸为 2 048 的最终回归层之前插入另一个完全连接的层，形成定位的特征向量；在测试时将四元数的方向向量归一化为单位长度。

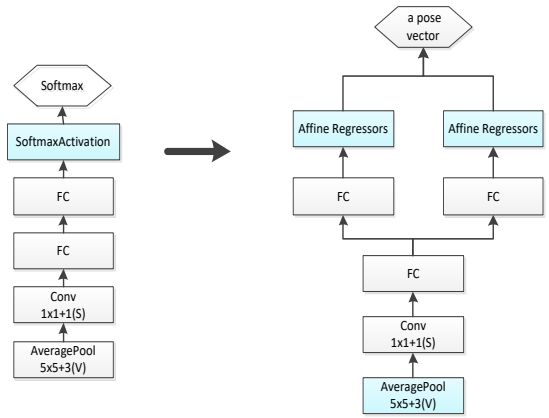

图 4　更改部分

Fig. 4　Change section

Walch 等人[46]提出了一个新的循环卷积神经网络架构实现相机姿态回归，算法结构如图 5 所示。其对运动模糊和光照变化具有鲁棒性，在大部分无纹理区域可以得到不错的相机姿态回归效果。卷积部分基于改进的 GoogLeNet 网络结构获得一个 2 048 维的输出向量，实现对特征向量的捕捉。卷积神经网络并不能得到特征向量之间的空间依赖关系，在 Posenet 基础上，使用向上、向下、向左、向右方向上 4 个

LSTM 扩大每个像素的感受野，使它有足够的上下文信息来准确定位图像。将 4 个 LSTM 输出到两个全连接层分别得出相机的位置和旋转。网络损失函数为欧氏距离与 PoseNet 相同，在超参数和数据集相同的情况下比 PoseNet 和 Bayesian Posenet[48]大大提高了定位性能。

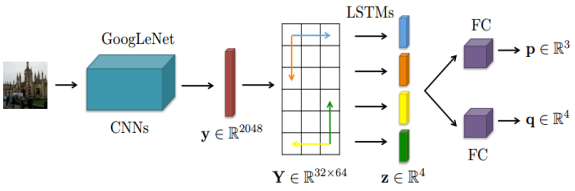

图 5　CNN+LSTM 算法流程

Fig. 5　CNN+LSTM algorithm

DeepVO 由两个并行 AlexNet[49]的级联卷积层组成，提取图片的低级特征到高级特征，在末端串联全连接，如图 6 所示。与 Posenet 不同，DeepVO 将两张彩色(RGB)图片与其间姿态标签同时输入网络，与传统的帧间估计方法更为相似。其在 KITTI 数据集中进行训练与测试，损失函数为欧氏距离。训练环境与测试环境相同，网络对环境了解越多，视觉里程计预测的效果就越好，在不相同环境测试性能还有待提高。笔者将 FAST 特征作为网络的先验附加到 RGB 数据中，在相同的网络结构进行训练与测试，体现了有效性。在 CNN 产生更高层次的特征之前，进一步探讨传统可追踪特征的融合也是发展的一个方向。

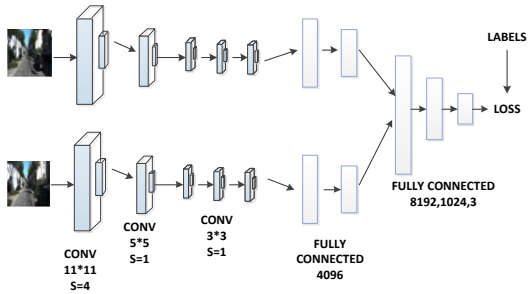

图 6　DeepVO 网络结构

Fig. 6　Deepvo structure

Costante 等人[50]提出了一种组合的深度学习神经网络，首先对图片采用 Brox 算法[51]提取密集的光流信号，输入光流图片。与前面基于深度学习的算法输入不同，采用光流图片代替 RGB-D 图片，淡化了外观对神经网络的影响，并且包含了图片之间的信息。测试三种不同的体系结构，如图 7 所示，并比较了它们的性能。

a)CNN-1b VO：对图像进行 8 次下采样后进行平均池化。

b)CNN-4b VO：将图像分成四个子图像，对每个子图像下采样 4 次，然后通过一系列类似于 CNN-1b 的 CNN 滤波器，在最后一层使用四个 CNN 网络的输出给出全局的帧到帧估计。

c)P-CNN VO：使用 CNN -1b 和 CNN-4b 的输出馈送到完全连接的网络，综合全局信息与局部信息，得到回归的位姿。

许多传统的基于深度学习的姿态估计是针对每一帧独立产生的。如果只提供单个图像输入网络，没有利用时间约束来平滑估计的相机运动，会导致误差较大。与 PoseNet 等纯卷积神经网络相比，采用循环卷积神经网络替代卷积神经网络，循环神经网络架构融入了时间依赖性，考虑为帧间估计应用更为适宜。VidLoc[52]是使用双向 LSTM 的循环卷积神经网络(RCNN),从单目图像序列进行有效的 6 自由度全局定位，将姿态估计瞬时协方差的计算方法集成到网络中。VidLoc 输入有时间连续性的图像流，通过时间规律性获得大量的姿态信息，使网络能够克服训练过程中的梯度消失问题。VidLoc 的 CNN 部分采用了 GoogLeNet 中的 inception 结构，使用其中的卷积层和池化层，删除完全连接的层，其 RNN 部分使用双向 LSTM 分别沿正向和反向建模捕获更丰富的数据，算法结构如图 8 所示。其在室内室外数据集上进行评估，获得了更高的位姿定位准确性。

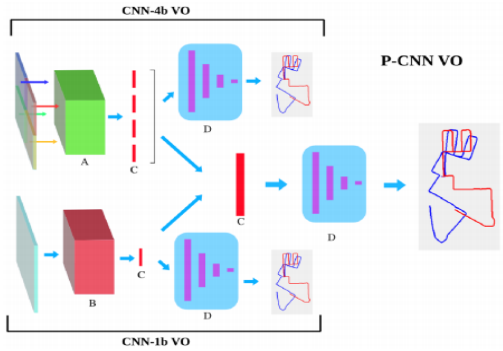

图 7　P-CNN 网络结构[36]

Fig. 7　P-CNN structure

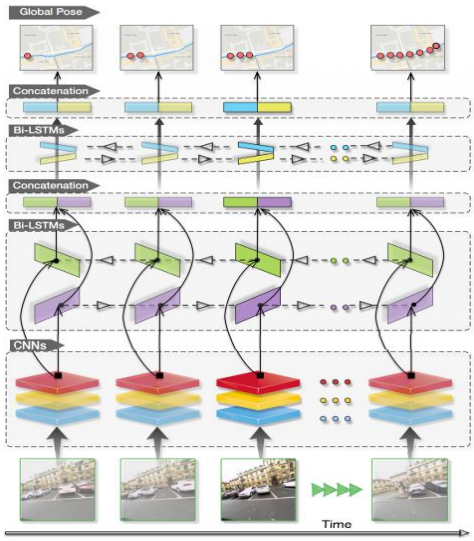

图 8　VidLoc 算法流程

Fig. 8　Vidloc algorithm

2.2　无监督的深度学习法

为避免使用大量人工标记的数据集进行训练，无监督框架在深度学习领域中兴起。在 SLAM 中无监督的深度学习方法最初用于深度估计，后来开始尝试在六自由度姿态估计方面及两者相结合方面的应用。

其中 UnDeepVO[53]是估计单目相机的 6 自由度姿态和深度的深度卷积神经网络的典例。UnDeepVO 有两个显著特点：a)是通过利用空间和时间的几何约束以无监督的方式实现；b)绝对规模的恢复。网络输入立体图像对进行训练，通过连续的单目图像进行测试。网络的损失函数是基于空间和时间的光度的一致性、视差的一致性、姿态的一致性、几何配准一致性来定义的，由于旋转具有很高的非线性，所以与平移相比通常更难以训练，损失函数中给旋转给予更大的权重。网络基于 VGG 卷积神经网络结构，为了更好地训练，在最后一个卷积层之后，将平移和旋转解耦为两组完全连接的层。

位姿估计网络使用某种形式的图像特征，而深度估计网络可能识别场景和对象的共同结构特征，更详细地研究系统学习的特征，对网络执行新任务如对象检测和语义分割很有效果。将无监督深度学习系统扩展到一个视觉 SLAM 系统里，以减少漂移也是发展的一大方向。

3　算法对比归纳

经典的帧间估计方法中应用最多的为特征点法，在很长一段时间，研究者们致力于研究大量算法如 SIFT、ORB 等提取关键点，计算描述子并匹配，提取特征点并寻找快速匹配的方法是帧间估计经典方法提升的关键点。在发生光照变化、存在雨雾等天气状况、场景中存在动态物体等情况下，提取具有鲁棒性的关键点和描述子是视觉 SLAM 的先决条件，对后续的运动估计、闭环检测、地图优化都有直接的影响。

光流法可以看做基于特征点方法的一种补充，通过光流跟踪特征点，减少了计算描述子的时间，从而也减少了计算量。通过跟踪取代匹配的过程，误匹配的情况会迅速减少，在大多情况下发生的应为丢失的情况，一定程度上增加了算法的鲁棒性。

对整张图像来说，特征点提取的过程实际上损失了很多图像信息。针对图像信息利用不充分的问题，直接法的计算方式可以有效地避免此问题；但是稠密的计算所有像素点，计算量过大。直接法在半稠密及稀疏方向也有很好的发展，在计算资源有限或要求高速的计算相机位姿时，可以采用稀疏直接法或者半稠密直接法，在建立完整地图等应用场合，可以采用稠密直接法。

基于深度学习的帧间估计将图像的特征蕴涵在深度神经网络的神经元中，是由低层次特征逐渐到高层次特征的一种特征学习的过程。与经典的帧间估计方法相比，基于深度学习的帧间估计需要大量的图像数据库进行训练，训练时间较长。但是在训练结束后得到具有权重的神经网络结构，其以端到端的架构实现，测试时间短可以快速地得到相机位姿信息。其因网络结构参数众多容易发生过拟合，但引入随机失活、提前停止等方法减少结构的复杂性，稀疏化参数，模型的泛化能力强。基于深度学习的帧间估计可以通过迁移学习有力的进行成果共享，但是这样的相机位姿估计过程少了直观性。表 1 中列举了经典几何方法与深度学习实现帧间估计的不同项目的对比。

基于深度学习的帧间估计方法，在最开始将相机位姿问题归纳为分类问题，无法在未知环境中应用，并且无法得到相机的旋转信息。后将帧间估计问题归纳为回归问题得以重视，最开始通过单纯的卷积神经网络得出相机的位置与姿态，在发展过程中还加入了光流、特征点等特征向量提高回归的精度。后来，将卷积神经网络结构转换为循环卷积神经网络结构，加入了时间依赖性，对帧间估计问题适用性非常强，大大地提升了帧间估计的位置和旋转精度，甚至超越了一些经典的单目帧间估计方法。表 2 中列举了基于深度学习实现视觉里程计的经典的网络结构及其定位结果与性能的比较。

表 1　优缺点对比

Table 1　Comparison of advantages and disadvantages

项目	经典的几何方法	基于深度学习的 SLAM
需要的图像数据	小规模数据	大规模数据
模型物理意义	明确的特征提取的过程	端到端的实现过程
计算量	很大	较小
模型迁移能力	较弱	较强

表 2　经典的网络结构

Table 2　Classical network architecture

结构名称	数据集	数据集简介	定位结果
PoseNet	Cambridge Landmarks	Alex 等人在伦敦使用手机拍摄，包含了雨雾等天气变化，大型室外城市定位数据集	精度 2m，3°
	7Scences	包含无纹理等室内场景数据集，包含 7 个场景	精度 0.5m，5 读
CNN+LSTM (WALCH)	Cambridge Landmarks	见上	PoseNet 2.08m，6.83°；CNN+LSTM 1.30m，5.52°
	7Scences	见上	PoseNet 0.44m，10.4°；CNN+LSTM 1.30m，9.85°
	LSI	大规模室内定位数据集，数据集分横跨相连的大学校园三个序列	PoseNet 1.72m，5.68°；CNN+LSTM 1.07m，3.59°
P-CNN	KITTI	由立体相机行驶在街道上拍摄，包含农村及郊区的场景，其中前 11 个序列提供了地面真实值	VISO2-M 18.55%,0.0376°/m SVR VO 13.81%,0.0302°/m P-CNN 8.96%,0.0235°/m
Vidloc	7 Scences	见上	PoseNet 0.44m；Vidloc 0.25m
	RobotCar	包含充满道路和树木的室外数据集	Vidloc 定位均方误差有效减小，验证了输入连续图像帧的必要性

4　视觉与惯导的融合

为提高视觉 SLAM 精度，研究视觉 SLAM 与惯性导航融合的算法成为视觉 SLAM 研究的重要内容。视觉 SLAM 通过惯导辅助提高精度，惯导通过 SLAM 修正积累的误差，两者相互补偿。

4.1　卡尔曼滤波法

松耦合将视觉里程计解算出的位置、旋转信息直接与 SINS 解算出的位置、旋转作差作为观测量，将惯导得到的信息作为估计量，通过卡尔曼系数，对估计量和观测量做融合，得到最后的结果[54~55]。组合的结构如图 9 所示。

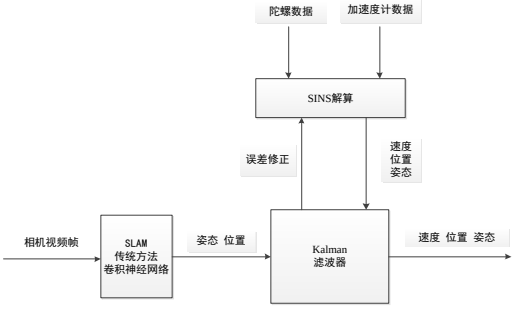

图 9　SINS/SLAM 组合结构

Fig. 9　Sins/slam composite structure diagram

其中：姿态量测为

$$Z_\varphi(t)=\begin{bmatrix}\varphi_{IE}-\varphi_{BE}\\ \varphi_{IN}-\varphi_{BN}\\ \varphi_{IU}-\varphi_{BU}\end{bmatrix}=\begin{bmatrix}\delta\varphi_E+M_E\\ \delta\varphi_N+M_N\\ \delta\varphi_U+M_U\end{bmatrix}+H_\varphi(t)+V_\varphi(t) \tag{9}$$

其中：I 表示 SINS；B 表示视觉里程计；$H_\varphi(t)=[0_{3\times3}\quad diag[1\quad 1\quad 1]\quad 0_{3\times9}]_{3\times15}$；$V_\varphi(t)$ 为姿态量测噪声。

位置量测为

$$\begin{aligned}Z_P&=\begin{bmatrix}(L_I-L_B)(R_M+h)\\ (\lambda_I-\lambda_B)(R_N+h)\cos L\\ h_I-h_B\end{bmatrix}\\&=\begin{bmatrix}(R_M+h)\delta L+N_N\\ (R_N+h)\cos L\delta\lambda+N_E\\ \delta h+N_U\end{bmatrix}=H_PX+V_P\end{aligned} \tag{10}$$

其中：V_P 为位置量测噪声；$H_P=[0_{3\times6}\quad diag[(R_M+h)\quad (R_N+h)\cos L\quad 1]\quad 0_{3\times6}]_{3\times15}$。

4.2　深度学习法

深度学习方法可以消除相机和 IMU 手动同步、手动校准的需要，进一步结合了特定信息，显著地减少漂移。

VINet[56]是一种以帧到帧学习的方法融合视觉和惯性传感器实现运动估计的深度神经网络，传统方法与 VINet 对比如图 10 所示。由 CNN-RNN 神经网络组成，视觉与惯导信息在中间特征级别上执行数据融合。网络的输入是单目 RGB 图像和 IMU 数据，IMU 数据是加速度 x、y、z 分量和陀螺仪测量的角速度的 6 维矢量，输出是一个 3 维平移和 4 维旋转的 7 维向量，将图像和 IMU 数据的输入序列转换为姿态向量。整个网络使用反向传播时间 (BPTT) 进行训练，使用具有 RMSProp 自适应学习速率的随机梯度下降法更新网络的权重。实验得到 VINet 在位置和旋转上都超越了仅使用图像数据的神经网络结果。

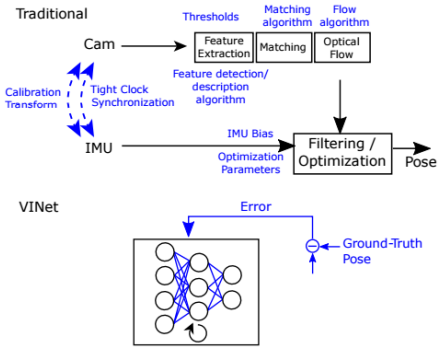

图 10　传统方法与 VINet 思想对比

Fig. 10　Comparison between traditional methods and vinet ideas

5　发展趋势

5.1　传感器类型增进与融合

视觉 SLAM 传感器主要有单目相机、双目相机、深度相机，不同的应用环境及不同的实验要求采用的传感器类型不同。视觉 SLAM 的发展有一个新的趋势，也就是采用新的传感器，如最近新出现的 Event Camera 视觉动态传感器，小巧迅速，与传统相机记录一个场景不同，Event Camera 记录一个场景的变化，仅检测像素的变化，并在像素基础上高频呈现，对低光也很敏感。使用 Event Camera 进行位姿估计与三维重建已有成果[57]，探索新的传感器类型取代经典相机是发展的趋势。还有 Polarimetric 相机，进行稠密的单目 SLAM 重建[58]。

前面的章节已经介绍了视觉 SLAM 与惯性传感器相结合，修正惯性传感器的累计误差以及提高视觉 SLAM 精度，已经取得了很好的成果。视觉 SLAM 多传感器融合方案多样，如与激光探测器的融合[59]，探索新的传感器融合方案[61]，是提高视觉 SLAM 鲁棒性与精度的很好的研究方向。

5.2　语义 SLAM

将物体识别与视觉 SLAM 结合起来，构建带有物体种类标签的地图是视觉 SLAM 发展的大趋势，即语义 SLAM[62~67]。物体识别需要对要识别的物体在各个角度拍摄并进行人工标定，生成大量的训练数据，而视觉 SLAM 可以自动计算物体在图像中的位置，生成带有标注的样本数据；将物体位姿及其所处区域类别标签融入到优化的目标函数，作为双重的约束来提高视觉 SLAM 的准确度。

Li 等人[68]提出了构建半密集三维语义地图，利用大规模直接单目 SLAM（LSD-SLAM）提供室内和室外场景中的 3 维空间信息，并结合卷积神经网络建立一个 3 维场景理解系统。使用深层卷积神经网络预测语义信息，将语义信息从单目 SLAM 系统投射到全局一致的 3 维地图中，由计算出的深度信息作为跟踪参考以递增的方式构造 3 维地图。

对于语义视觉 SLAM 的研究还处于起步阶段，还需进行深一步的研究使其更具有实用性。

5.3　基于深度学习的算法发展

深度神经网络的计算是特征提取的过程，将光流场分布作为辅助特征加入可以提高位置与姿态估计的精度，融合辅助特征丰富神经网络输入是基于深度学习进行帧间估计的有效途径。

前文中主要介绍的是实现帧间估计的深度学习算法，其算法得到的精度和经典的闭环视觉 SLAM 算法还有距离，这使得基于深度学习的 SLAM 算法还难以进行工程应用。

将 SLAM 中的一或几个模块选择用深度学习的方法代替，减少计算量，如将后端优化，闭环检测[69]通过深度神经网络进行计算，或者将整体过程通过深度学习进行表述，都可以提高基于深度学习估计相机的位置与姿态的精度和鲁棒性，基于这方面的研究还相对较少。

5.4　动态环境下的 SLAM

传统的 SLAM 在动态环境下定位与建图的能力有限，需要手动更新，没有办法实时的分析环境的变化，也使得其对动态障碍物识别不够准确。针对动态场景一直在探索如何检测与处理[70]。

高仙提出动态环境下位姿估计与地图构建的 SLAM2.0 方案。其将语义信息与 SLAM 相融合，增强对环境的理解力。具有广度建图的功能，在大范围全场景环境中成功应用，高精度的全局定位达到 2 cm。采用激光、单双目视觉、超声等多种传感器动态的提取特征回传数据，与原地图对比分析，完成对地图动态实时更新，显著地提高了 SLAM 的多项关键性能指标，目前成功应用于清洁、警用、配送等领域，并将进一步扩展。

其采用多种传感器融合，实现了对环境的感知，但大的传感器数量限制了其在小型设备环境下的应用。大量的不同数据类型，需要高性能的计算设备进行计算与融合，成本较高，对实时性也有所影响。在低成本和便携的应用需求与技术开发环境需求下，仍然需要进一步讨论研究。目前基于深度学习的方法对动态物体有规避的效果，其也可以为动态环境下的 SLAM 提供另一种思路。

参考文献：

[1]　Fuentes-Pacheco J, Ruiz-Ascencio J, Rendon-Mancha J M. Visual

simultaneous localization and mapping: a survey [J]. Artificial Intelligence Review, 2015, 43 (1): 55-81.

[2] Birk A, Pfingsthorn M. Simultaneous localization and mapping (SLAM) [M]// John Wiley & Sons, Wiley Encyclopedia of Electrical and Electronics Engineering. 2016.

[3] Nister D, Naroditsky O, Bergen J R. Visual odometry [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2004: 652-659.

[4] Scaramuzza D, Fraundorfer F. Visual odometry: part I: the first 30 years and fundamentals [J]. IEEE Robotics Automation Magazine, 2011, 18 (4): 80-92.

[5] Paul R, Newman P. FAB-MAP 3D: Topological mapping with spatial and visual appearance [C]// Proc of IEEE International Conference on Robotics and Automation. Piscataway. Piscataway, NJ: IEEE Press, 2010: 2649-2656.

[6] Pinies P, Paz L M, Galvez-Lopez D, *et al.* CI-graph simultaneous localization and mapping for three-dimensional reconstruction of large and complex environments using a multicamera system [J]. Journal of Field Robotics, 2010, 27 (5): 561-586.

[7] Galvez-Lopez D, Tardos J D. Real-time loop detection with ´ bags of binary words [C]// Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway. Piscataway, NJ: IEEE Press, 2011: 51-58.

[8] 张国良，姚二亮，林志林，*et al.* 融合直接法与特征法的快速双目 SLAM 算法 [J]. 机器人, 2017, 39 (6): 879-888.) . (Zhang Guoliang, Yao Erliang, Lin Zhilin, *et al.* Fast binocular SLAM algorithm combining the direct method and the feature-based method [J]. Robot, 2017, 39 (6): 879-888.)

[9] 姜山. 基于 RGB-D SLAM 的视觉定位与路径规划方法研究 [D]. 哈尔滨: 哈尔滨工业大学, 2017. (Jiang Shan. Visual location and path planning based on RGB-D SLAM [D]. Harbin: Harbin Institute of Technology, 2017.)

[10] Younes G, Asmar D, Shammas E, *et al.* Keyframe-based monocular SLAM: Design, survey, and future directions [J]. Robotics and Autonomous Systems, 2017, 98 (4): 67-88.

[11] 曹恒. 基于单目视觉的 SLAM 算法研究 [D]. 武汉: 华中科技大学, 2016. (Cao Heng. Research of SLAM algorithm based on monocular vision [D]. Wuhan: Huazhong University of Science and Technology, 2016.)

[12] 郑伟. 基于视觉的微小型四旋翼飞行机器人位姿估计与导航研究 [D]. 合肥: 中国科学技术大学, 2014. (Zheng Wei. Vision based pose estimation and navigation for a quadrotor [D]. Hefei: University of Science and Technology of China, 2014.)

[13] 陈丁，马跃龙，曹雪峰, *et al.* 融合 IMU 与单目视觉的无人机自主定位方法 [J]. 系统仿真学报, 2017 (S1): 9-14. (Chen Ding, Ma Yuelong, Cao Xuefeng, *et al.* A method of autonomous localization for UAV with fusion of IMU and Mono SLAM [J]. Journal of System Simulation, 2017 (S1): 9-14.)

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

[15] Bay H, Ess A, Tuytelaars T, *et al.* Speeded-up robust features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110 (3): 346-359.

[16] Rublee E, Rabaud V, Konolige K, *et al.* ORB: An efficient alternative to SIFT or SURF [C]// Proc of International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2011: 2564-2571.

[17] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24 (6): 381-395.

[18] Maimone M, Cheng Y, Matthies L. Two years of visual odometry on the mars exploration rovers [J]. Journal of Field Robotics, 2007, 24 (3): 169-186.

[19] Deigmoeller J, Eggert J. Stereo visual odometry without temporal filtering [C]// Proc of the 38th German Conference on Pattern Recognition. Berlin, German: Springer Press, 2016: 166-175.

[20] Davison A J. SLAM with a single camera [C]// Proc of Workshops on Concurrent Mapping an Localization for Autonomous Mobile Robots in Conjunction with ICRA. Piscataway, NJ: IEEE Press, 2002: 18-27.

[21] Davison A J. Real-time simultaneous localisation and mapping with a single camera [C]// Proc of IEEE Computer Society. Piscataway, NJ: IEEE Press 2003: 1403-1410.

[22] Davison A J, Reid I D, Molton N D, *et al.* MonoSLAM: real-time single camera SLAM [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29 (6): 1052-1067.

[23] Civera J, Davison A J, Montiel J. M M. Inverse depth parametrization for monocular SLAM [J]. IEEE Trans on Robotics, 2008, 24 (5): 932-945.

[24] Klein G. Parallel tracking and mapping for small AR workspaces [C]// Proc of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality . New York: ACM Press, 2008: 250-259.

[25] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. IEEE Trans on Robotics, 2015, 31 (5): 1147-1163.

[26] Hartley R, Zisserman A. Multiple view geometry in computer vision [M].Cambridge: Cambridge University Press, 2003.

[27] Lepetit V, Moreno-Noguer F, FUA P. EPnP: an accurateO (n) solution to the PnP problem [J]. International Journal of Computer Vision, 2009, 81 (2): 155-166.

[28] Camposeco F, Cohen A, Pollefeys M, *et al.* Hybrid camera pose estimation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 136-144.

[29] Mur-Artal R, Tardos J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras [J]. IEEE Trans on Robotics, 2017: 1-8.

[30] 陆建伟，王耀力. 基于 ORB-SLAM2 的实时网格地图构建 [J/OL]. 计算机应用研究, 2019, 36 (11) . (2018-06-19) [2018-08-10]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180811. 1328. 054. html. (Lu Jianwei, Wang Yaoli. Real-time grid map construction based on ORB-SLAM2 [J/OL]. Application Research of Computers, 2019, 36 (11) . (2018-06-19) [2018-08-10]. http://kns. cnki. net/kcms/detail/51. 1196. TP. 20180811. 1328. 054. html.)

[31] Horn B K P, Schunck B G. Determining optical flow [C]// Proc of SPIE. Bellingham, USA: SPIE Press, 1981: 319-331.

[32] Maity S, Saha A, Bhowmick B. Edge SLAM: edge points based monocular visual SLAM [C]// Proc of ICCV Workshops. Piscataway, NJ: IEEE Press, 2017: 2408-2417.

[33] 王泽民，李建胜，王安成，等. 一种基于 LK 光流的动态场景 SLAM 新方法 [J]. 测绘科学技术学报, 2018, 35 (2): 80-83. (Wang Zeping, Li Jiansheng, Wang Ancheng, et al. A new SLAM method for dynamic scene based on LK optical flow [J]. Journal of Surveying and Mapping Science and Technology, 2018, 35 (2): 80-83.)

[34] Engel J, Thomas S, Cremers D. LSD-SLAM: large-scale direct monocular SLAM [C]// Proc of European Conference on Computer

Vision. Cham: Springer Press, 2014: 834-849.

[35] Henriques J F, Vedaldi A. Mapnet: an allocentric spatial memory for mapping environments [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 8476-8484.

[36] 逄淑超. 深度学习在计算机视觉领域的若干关键技术研究 [D]. 长春：吉林大学，2017. (Pang Shuchao. Research on some key technologies of deep learning in the field of computer vision [D]. Changchun: Jilin University, 2017.)

[37] 李策，魏豪左，卢冰，等. 基于深度视觉的 SLAM 算法研究与实现 [J]. 计算机工程与设计, 2017, 38 (4): 1062-1066. (Li Ce, Wei Haozuo, Lu Bing, *et al*. Research and implementation of SLAM algorithm based on depth vision [J]. Computer Endineering and Design, 2017, 38 (4): 1062-1066.)

[38] 徐利义. 基于深度学习和视觉图像的机器人定位与地图构建 [D]. 北京：北京理工大学, 2016 (Xu Liyi. Research on localization and mapping for mobile robot based on deep learning and visual image [D]. Beijing: Beijing Institute of technology, 2016.)

[39] Roberts R, Nguyen H, Krishnamurthi N, *et al.* Memory-based learning for visual odometry [C]// Proc of Robotics and Automation. Piscataway, NJ: IEEE Press, 2008: 47-52.

[40] Roberts R, Potthast C, Dellaert F. Learning general optical flow subspaces for egomotion estimation and detection of motion anomalies [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 57-64.

[41] Guizilini V, Ramos F. Semi-parametric models for visual odometry [C]// Proc of IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2012: 3482-3489.

[42] Guizilini V, Ramos F. Semi-parametric learning for visual odometry [J]. The International Journal of Robotics Research, 2013, 32 (5): 526-546.

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

[44] Shotton J, Glocker B, Zach C, *et al*. Scene coordinate regression forests for camera relocalization in RGB-D images [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society. Piscataway, NJ: IEEE Press, 2013: 2930-2937.

[45] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going Deeper with convolutions [C]// Proc of ImageNet Large-Scale Visual Recognition Challenge . Piscataway,NJ: IEEE Press, 2014: 1-9.

[46] Walch F, Hazirbas C, Leal-Taixé L, *et al.* Image-based localization with spatial LSTMs [C]// Proc of International Conference on Computer Vision. 2016.

[47] Kendall A, Cipolla R. Modelling uncertainty in deep learning for camera relocalization [C]// Proc of IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2016: 4762-4769.

[48] Mohanty V, Agrawal S, Datta S, *et al.* DeepVO: a deep learning approach for monocular visual odometry [C]// Proc of Computer Vision and Pattern Recognition. 2016.

[49] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25 (2) .

[50] Costante G, Mancini M, Valigi P, *et al.* Exploring representation learning with CNNs for frame-to-frame ego-motion estimation [J]. IEEE Robotics & Automation Letters, 2015, 1 (1): 18-25.

[51] Brox T, Andrés B, Papenberg N, *et al.* High accuracy optical flow estimation based on a theory for warping [J]. Eccv, 2004, 2004 (pp): 25-36.

[52] Clark R, Wang S, Markham A, *et al.* VidLoc: 6-DoF video-clip relocalization [C]// Proc of Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2-6.

[53] Li Ruihao, Wang Sen, Long Zhiqiang, *et al.* UnDeepVO: monocular visual odometry through unsupervised deep learning [J]. arXiv. arXiv preprint: 1709. 06841, 2017.

[54] 周绍磊，吴修振，刘刚，等. 一种单目视觉 ORB-SLAM/INS 组合导航方法 [J]. 中国惯性技术学报，2016 (5): 633-637. (Zhou Shaolei, Wu Xiuzhen, Liu Gang, *et al.* Integrated navigation method of monocular ORB-SLAM/INS [J]. Journal of Chinese Inertial Technology, 2016 (5): 633-637.)

[55] 熊敏君，卢惠民，熊丹，等. 基于单目视觉与惯导融合的无人机位姿估计 [J]. 计算机应用，2017 (S2): 127-133. (Xiong Minjun, Lu Huimin, Xiong Dan, *et al.* Pose estimation of UAV based on monocular vision and inertial navigation. [J]. Journal of Computer Applications, 2017 (S2): 127-133.)

[56] Clark R, Wang Sen, Wen Hongkai, *et al.* VINet: visual-inertial odometry as a sequence-to-sequence learning problem [J]. arXiv. arXiv preprint: 1701. 08376, 2017.

[57] Kim H, Leutenegger S, Davison A J. Real-time 3D reconstruction and 6-DoF tracking with an event camera [C]// Proc of European Conference on Computer Vision. Cham: Springer Press, 2016: 349-364.

[58] Yang Luwei, Tan Feitong, Li Ao, *et al.* Polarimetric dense monocular SLAM [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 3857-3866.

[59] Nicolai A, Skeele R, Eriksen C, *et al*. Deep learning for laser based odometry estimation [EB/OL]. [2016-6-17]. http://research. e ngr. oregonstate. edu. pdf.

[60] 张毅，杜凡宇，罗元，等. 一种融合激光和深度视觉传感器的 SLAM 地图创建方法 [J]. 计算机应用研究, 2016, 33 (10): 2970-2972,3006. (Zhang Yi, Du Fanyu, Luo Yuan, *et al*. A SLAM map creation method combining laser and depth vision sensors [J]. Computer Applied Research, 2016, 33 (10): 2970-2972,3006.)

[61] Wang Peng, Yang Ruigang, Cao Binin, *et al.* Dels-3d: deep localization and segmentation with a 3d semantic map [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 5860-5869.

[62] Salas-MORENO, Renato F. Dense semantic SLAM [D]. London, UK: Imperial College, 2014.

[63] 于金山，吴皓，田国会，等. 基于云的语义库设计及机器人语义地图构建 [J]. 机器人, 2016, 38 (4): 410-419. (Yu Jinshan, Wu Hao, Tian Guohui, *et al.* Semantic database design and semantic map construction of robots based on the cloud [J]. Robot, 2016, 38 (4): 410-419.)

[64] Tenorth M, Kunze L, Jain D, *et al.* KNOWROB-MAP-knowledge-linked semantic object maps [C]// Proc of IEEE-RAS International Conference on Humanoid Robots. Piscataway, NJ: IEEE Press, 2010: 430-435.

[65] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2014: 1449-1456.

[66] McCormac J, Handa A, Davison A, *et al.* SemanticFusion: dense 3D semantic mapping with convolutional neural networks [C]// Proc of

IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2016: 4628-4635.

[67] Saarinen J P, Andreasson H, Stoyanov T, *et al.* 3D normal distributions transform occupancy maps: an efficient representation for mapping in dynamic environments [J]. The International Journal of Robotics Research, 2013, 32 (14): 1627-1644.

[68] Li Xuanpeng, Belaroussi R. Semi-dense 3D semantic mapping from monocular SLAM [J]. arXiv. arXiv preprint: 1611. 04144. 2016.

[69] Gao Xiang, Zhang Tao. Unsupervised learning to detect loops using deep neural networks for visual SLAM system [J]. Autonomous Robots, 2017, 41 (1): 1-18.

[70] Saarinen J P, Andreasson H, Stoyanov T, *et al.* 3D normal distributions transform occupancy maps: an efficient representation for mapping in dynamic environments [J]. International Journal of Robotics Research, 2013, 32 (14): 1627-1644.

[71] Einhorn E, Gross H M. Generic NDT mapping in dynamic environments and its application for lifelong SLAM [M]. [S. l.] : North-Holland Publishing, 2015.