软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(2):975−1009 [doi: 10.13328/j.cnki.jos.006946]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于先验地图的视觉重定位方法综述*

蔡旭东, 王永才, 白雪薇, 李德英

(中国人民大学 信息学院, 北京 100872)

通信作者: 王永才, E-mail: ycw@ruc.edu.cn

摘　要: 在自动驾驶、增强现实和智能移动机器人领域, 视觉重定位是非常重要的基础问题. 视觉重定位是指根据视觉传感器实时拍摄的数据, 在已有先验地图中确定位置和姿态的问题. 过去数十年间, 该问题受到广泛关注, 涌现出种类繁多的先验地图构建方法和视觉重定位方法. 这些工作差异大, 涉及范围广, 技术概括和总结尚缺乏. 因此, 对视觉重定位领域进行综述具有重要的理论和应用价值. 尝试为视觉重定位相关方法建立一个统一的蓝图, 从图像数据在大规模地图数据库中查询的角度对相关工作进行分析和总结. 综述不同类型地图数据库构建方法、不同特征匹配、重定位和位姿计算方法, 总结目前视觉重定位的主流数据集, 最后分析视觉重定位存在的挑战和潜在发展方向.

关键词: 先验地图; 地图构建; 位姿估计; 视觉重定位

中图法分类号: TP391

中文引用格式: 蔡旭东, 王永才, 白雪薇, 李德英. 基于先验地图的视觉重定位方法综述. 软件学报, 2024, 35(2): 975–1009. http://www.jos.org.cn/1000-9825/6946.htm

英文引用格式: Cai XD, Wang YC, Bai XW, Li DY. Survey on Visual Relocalization in Prior Map. Ruan Jian Xue Bao/Journal of Software, 2024, 35(2): 975–1009 (in Chinese). http://www.jos.org.cn/1000-9825/6946.htm

Survey on Visual Relocalization in Prior Map

CAI Xu-Dong, WANG Yong-Cai, BAI Xue-Wei, LI De-Ying

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: In the fields of autonomous driving, augmented reality, and intelligent mobile robots, visual relocalization is a crucial fundamental issue. It refers to the issue of determining the position and attitude in an existing prior map according to the data captured in real time by visual sensors. In the last decades, visual relocalization has received extensive attention, and numerous kinds of prior map construction methods and visual relocalization methods have come to the fore. These efforts vary considerably and cover a wide scope, but technical overviews and summaries are still unavailable. Therefore, a survey of the field of visual relocalization is valuable both theoretically and practically. This study tries to construct a unified blueprint for visual relocalization methods and summarize related studies from the perspective of image data querying from large-scale map databases. This study surveys various types of construction methods for map databases and different feature matching, relocalization, and pose calculation approaches. It then summarizes the current mainstream datasets for visual relocalization and finally analyzes the challenges ahead and the potential development directions of visual relocalization.

Key words: prior map; mapping; pose estimation; visual relocalization

1 引　言

近年来, 随着新一代信息技术和智能传感器的不断发展, 智能移动机器人[1], 自动驾驶[2]和增强现实

* 基金项目: 国家自然科学基金 (61972404, 12071478)
收稿时间: 2022-08-16; 修改时间: 2023-02-28; 采用时间: 2023-04-06; jos 在线出版时间: 2023-09-13
CNKI 网络首发时间: 2023-09-14

(augmented reality)[3]等技术成为学术界和工业界的热点. 在上述应用中, 智能设备在复杂环境中感知自身位置, 即自定位能力是众多应用的基础. 视觉传感器由于其结构简单, 成本低, 往往是各类智能设备自定位的首选传感器[4], 但是视觉传感器无法直接获取深度信息[5], 易受光照等环境因素影响[6], 视觉里程计存在累计误差[7]等问题. 而在预先制备的高精度先验地图中基于视觉进行重定位可以很好地克服视觉里程计的缺点, 提升视觉定位的可靠性和准确性[8]. 视觉重定位是指根据实时采集的视觉观测数据, 在已有的先验地图中确定自身在世界坐标系下位置和姿态的问题[9], 其关键包括: (1) 如何建立先验地图数据库[10,11]; (2) 如何基于图像查询先验地图建立准确的数据关联[12−14]; (3) 如何基于关联的数据计算当前准确位姿[15].

过去的 20 年间, 视觉重定位问题受到学术界和工业界广泛关注, 不同领域的研究者从多个角度对这个问题进行了探索, 其本质上都是根据视觉传感器采集的图像在大规模地图数据库中进行查询的问题[16]. 但由于不同的应用场景对视觉重定位的精度, 运行效率, 地图体积和查询数据有不同的要求, 这些差异用到不同类型的视觉传感器和不同模态的先验地图, 从而衍生出一系列基于不同视觉传感器和先验地图的视觉重定位方法[17−21]. 现有一些前期工作也对视觉重定位领域的工作进行了一些总结. Piasco 等人[22]将视觉重定位方法分为了直接方法和间接方法两大类, Chen 等人[23]介绍了基于深度学习的建图和定位方法, 但这些工作对各种视觉重定位方法的总结仍不够全面. Chen 等人[24]对基于单目相机的重定位算法进行了总结, 但是忽略了其他类型的视觉传感器. 本文将从视觉图像在地图数据库中查询的角度展开综述. 同传统的数据库查询不同, 视觉重定位中查询信息为图像, 而地图数据库的数据类型却多种多样, 存储方式也各不相同. 为了更好地总结现有方法和发展趋势, 本文尝试为视觉重定位领域归纳一个统一的蓝图.

首先, 从查询图像上, 图像可能来自单目相机[25]、双目相机[26]或 RGB-D 相机[27]. 单目相机使用最为广泛, 具有成本低、体积小、结构简单、部署方便等优点[28]. 但单目相机采集的图像没有深度信息[29], 基于单目相机的视觉里程计等方法在光照变化大、纹理信息少、运动较快等情况下会产生较大的偏差[30]. 双目相机一般是由左右水平放置的两个单目相机组成, 和人眼的工作原理类似, 双目相机可以通过计算左右相机图像之间的视差来估计图像中每个像素的深度信息, 但视差的计算量比较大, 对硬件条件有一定要求[31]. 和单目相机类似, 双目相机在光照变化大、纹理信息少时也效果不佳, 得到的深度信息精度较差. RGB-D 相机则是利用红外结构光 (structured light) 或飞行时间 (time of flight) 的原理主动测量图像中每个像素点的深度信息. RGB-D 相机使用物理方式进行测距, 其深度信息不需要复杂的计算且比较准确, 在光照变化大, 快速运动等场景下都可以进行测距. 但是其测距范围较短, 结构光易受日光干扰, 对于一些特殊表面材料效果不佳[32]. 本文将上述差异视为查询数据的差异.

从先验地图数据库的角度, 地图精度, 体积大小和检索效率是地图数据库的 3 个重要指标, 且它们之间相互制约, 需要根据应用场景有所取舍, 先验地图的设计对重定位算法有较大影响[33,34]. 目前, 按照地图数据库的形态不同, 主流的地图数据库大致上可以划分为 5 种: 图像数据库地图[10]、点云表示地图[35]、稠密边界表示地图[36]、语义地图[37]和高分辨率地图[11]. (1) 图像数据库地图由一系列带有全局位姿信息的图像组成, 采集较为方便但是容易受到光照变化, 季节更替等环境因素影响, 其地图精度受图像采样间隔限制[38]; (2) 点云表示地图由空间中一系列带有三维坐标的无序点组成, 其对纹理, 环境外观变化等不敏感, 但点云是无序且缺乏特征信息的[39]; (3) 稠密边界表示地图尝试准确的表示场景表面和结构, 主要借助计算机图像学的方法来构建场景的几何结构并渲染其表面纹理, 构建时往往比较占用空间和消耗计算资源[40]; (4) 语义地图是在一些现有地图表示形式的基础之上, 提取出更高级的语义信息并保存在地图中的一种地图形式, 语义地图更加鲁棒, 但构建时需要消耗额外的算力和存储资源[41]; (5) 高分辨率地图是专为无人驾驶相关应用服务的地图, 其主要以矢量形式存储道路、建筑、路标等交通元素, 具有轻量化的特点[42].

不同查询数据和不同地图数据库对应的查询方法各不相同, 本文首次将多种基于不同类型视觉传感器和先验地图数据库的视觉重定位方法放在一个框架下进行回顾和比较, 使得视觉重定位的相关方法介绍更为全面和立体. 本文的主要贡献可以归纳为如下几点.

(1) 对视觉重定位方法中常见的先验地图数据库进行了总结, 详细介绍了每种地图形式的构建方法和特点, 对

几种地图进行了对比.

(2) 通过将视觉重定位看作图像在大规模地图数据库中的查询问题, 对不同先验地图数据库中基于图像的查询、匹配和位姿计算方法进行详细介绍, 对各类方法进行总结归纳.

(3) 重点介绍基于单目相机的视觉重定位的同时, 涵盖了双目相机、RGB-D 相机的相关算法; 总结并概括了常用的数据集, 最后提出了视觉重定位目前亟待解决的问题和未来重点发展方向.

从地图数据库类型、查询方法角度, 本文对视觉重定位方法进行分类介绍, 整体结构如图 1 所示.

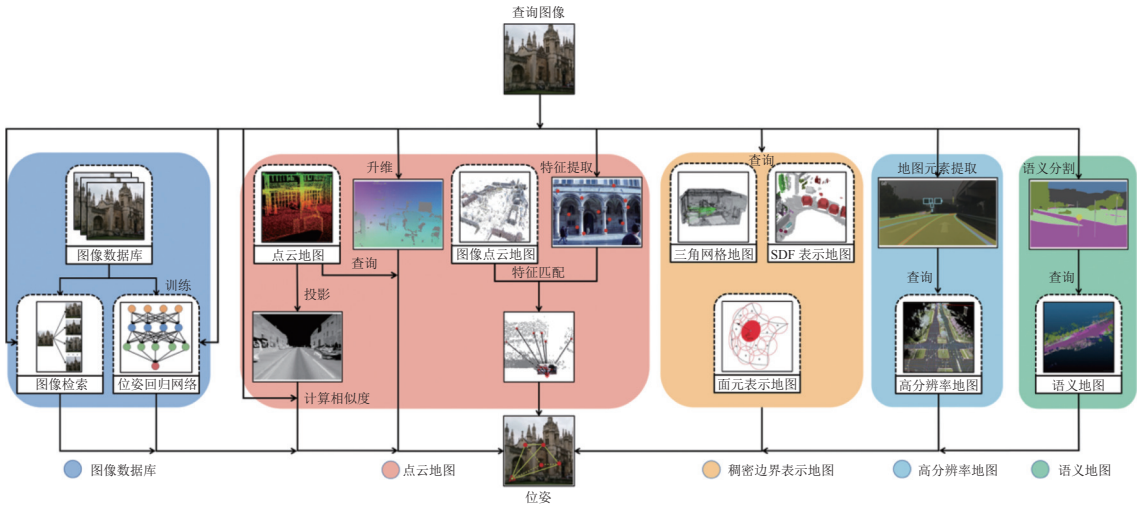

图 1 基于先验地图的视觉重定位方法整体架构图

(1) 基于图像数据库的视觉重定位: 由于地图数据与查询图像属于相同模态, 主要重定位方法分为: 1) 使用图像检索方法直接在图像数据库中查询[43]; 2) 使用深度学习方法在位姿回归网络中直接进行位姿回归计算[44].

(2) 基于点云数据库的视觉重定位: 点云数据库与查询图像处于不同模态, 无法直接比较, 目前主流方法分为: 1) 基于点云投影的方法先将点云投影为 2D 图像, 再借助图像相似度进行查询[45]; 2) 基于场景升维方法先将查询图像升到三维, 再在点云地图数据库中进行查询[46]; 3) 基于特征点匹配的方法利用 SfM (structure from motion)[47−49] 构建带有图像特征的点云地图数据库, 再基于图像特征匹配关系解算位姿[50].

(3) 稠密边界表示地图的视觉重定位: 稠密边界表示地图主要包括三角网格地图、SDF 表示地图、面元表示地图. 稠密边界地图中进行查询的方法主要利用场景的几何结构和丰富的纹理信息与查询图像建立数据关联[51].

(4) 高分辨率地图数据库的视觉重定位: 高分辨率地图构建时主要进行地图元素提取, 以矢量形式保存地图元素. 查询时先从查询图像中提取出同种地图元素之后, 再进行匹配与查询[52].

(5) 语义地图数据库的视觉重定位: 语义数据库中包含了地图元素的语义信息, 可以利用查询图像的语义信息作为全局特征进行查询[53], 也可以利用语义标签之间的对应关系进行查询[54], 语义信息还可以用于指导算法选择更鲁棒的特征进行查询[55].

本文第 2 节介绍目前视觉重定位中主流先验地图数据库表示形式. 第 3–7 节详细介绍上述视觉重定位的方法. 第 8 节总结视觉重定位领域常用数据集. 第 9 节探讨目前视觉重定位领域面临的挑战与机遇. 第 10 节对全文进行总结.

2 不同形式的先验地图

先验地图数据库的类型决定查询图像的定位方式. 地图数据库应尽可能多地包含真实世界中的显著信息, 并能够被高效查询和存储. 在本节中, 我们分析和总结视觉重定位中目前主流地图表示形式, 分为 5 类介绍: (1) 点云

表示地图; (2) 图像数据库地图; (3) 稠密边界表示地图; (4) 语义地图; (5) 高分辨率地图.

2.1 点云表示地图

点云表示地图由 3D 空间中一系列带有三维坐标的无序点组成. 点云地图可以进一步划分为: (1) 雷达点云地图; (2) 图像点云地图; (3) 体素化点云地图.

(1) 雷达点云地图, 由激光雷达 (LiDAR) 扫描后使用 3D 激光 SLAM 算法[39,56,57]得到, 现有 3D 激光 SLAM 算法已可以构建起高精度 3D 雷达点云地图. 雷达点云地图通常只包含每个点的 3 维坐标系下的位置信息以及脉冲光反射强度信息.

(2) 图像点云地图, 使用相机采集的环境图片或视频数据做为输入, 使用 SfM 算法[47−49]从相机拍摄的场景图片或视频中重建得到. 现有 SfM 算法如 VisualSFM[47,48]和 COLMAP[49]可以构建出带有图像特征信息的图像点云地图. 相比于雷达点云地图, 使用 SfM 算法重建的点云中的 3D 点包含该点对应的 2D 图像中特征点的描述子 (如 SIFT[58]). 由于可能存在多个 2D 图像特征点对应一个 3D 点的情况, 通常的做法是使用这些特征点描述子的 Mean shift[59]聚类作为该 3D 点的描述子[60].

(3) 体素化点云地图, 原始的点云地图丢失了环境中的三维结构信息, 可以通过将其体素化 (Voxel) 的方式进行改进[61]. 体素化是用一定大小的三维立方体 (体素网格) 对原始点云进行分割, 最终使用分割得到的各个体素网格来表示整个环境. 对点云进行体素化的方法有很多, 最常见的方法是使用哈希表的结构来存储体素化地图[62].

上述 3 类点云地图各有优缺点. (1) 雷达点云地图不容易受到环境光照条件变化的影响, 在纹理较少的环境中也能较好的对环境进行重建. 然而, 由于点云地图是由数量众多的无序点构成, 点云地图缺失了很多环境的纹理信息并且比较难以处理. (2) 图像点云地图保留了一些图像的特征, 但是对于 SIFT[58]等特征点的计算却非常耗时, 且这种点云地图的构建方法非常容易受到环境光照条件以及季节变化的影响. (3) 相比于原始点云, 体素化的地图可以保存一些环境三维结构信息, 同时减少原始点云的计算量, 但由于体素化本质上是对点云进行下采样处理, 因此会导致地图中一些信息的丢失. 体素化的粒度 (体素网格的大小) 对定位精度也有着较大的影响, 需要在精度和计算量之间寻找平衡.

2.2 图像数据库地图

图像数据库地图由一系列带有 6-DoF 位姿信息 (旋转和平移) 的图像组成. 图像数据库中的图像可以由不同内参的相机拍摄得到. 由于图像数量众多, 人工标定每张图像的 6-DoF 位姿信息是不现实的. 与图像点云地图的构建方法类似, 图像数据库中图像的位姿通常使用 SfM 算法[47−49]进行标定或者使用 GPS 等外部测量设备获得全局定位信息. 不同的是, 图像数据库不保存整个场景点云, 只需保存图片和其对应的位姿即可.

受相机拍摄视野的限制, 基于图像数据库的定位方法的精度都非常依赖于图像的采样间隔. 图像的采样频率越高, 图像采集地点间距越短, 相机的定位便能更精确. 但图像数据库也比较容易受到环境光照条件, 季节更替的影响. 对于表示同一场景的图片, 不同拍摄时间, 季节和天气都可能导致图像差距较大, 引起错误的定位结果[63,64]. 另一方面, 图像数据库可以包含比较丰富的环境纹理和颜色信息, 并应用较为广泛.

2.3 稠密边界表示地图

稠密边界表示地图尝试准确的表示场景表面和结构, 其主要是通过计算机图形学的方法来对环境表面进行表示并保持其原有的几何结构, 常见的表现形式有三角网格 (Mesh) 地图[19]、有向距离场 (signed distance function, SDF)[65]和面元 (Surfel) 表示地图[66].

(1) 三角网格 (Mesh) 地图: 三角网格结构简单, 可以方便且快速的生成, 是目前使用最为广泛的网格表现形式, 也被称为三角剖分. 三角网格地图在空间中进行表示时, 每个三角网格的 3 条边都和其他三角形相连接, 在存储时每个三角网格都需要存储 3 个顶点, 3 条边和 1 个面的信息. 主要是使用基于 Delaunay 三角剖分的方法, 通过处理点云构建, 该算法在 PCL 库中有对应的实现[67].

(2) 有向距离场 (SDF) 表示地图: 有向距离场最早由 Osher 等人[65]于 2003 年提出, 主要用于场景稠密重建方法中[68,69]. 基本的有向距离场把一个场景划分为 $W\times H\times L$ 个体素进行表示. 每个体素除了拥有全局的三维坐标,

还保存有该体素到最近表面的距离 SDF 值和 RGB 信息. 如果体素相对于表面在靠近相机的一侧 (表面之外) 则为正值, 反之则为负值. 截断有向距离场 (truncated signed distance function) 则是给有向距离场中距离表面太近或者太远的体素赋值为固定值. 其余体素的 SDF 值只需要归一化到 $[-1, 1]$ 区间内即可. 有向距离场地图在构建时需要同时利用 RGB 图像和深度图像信息, 一般使用 RGB-D 相机采集数据进行构建. 这种方式重建的场景比较精密, 但代价是需要使用 GPU 进行加速并对显存要求较高, 因此一般情况下用于小规模场景重建. 另外, 整个场景的大小需要提前固定, 超出场景的部分将无法进行重建, 场景的精细程度也和每个体素的大小相关.

(3) 面元 (Surfel) 表示地图: 面元表示地图最早由 Pfister 等人[66]于 2000 年左右提出, Whelan 等人[70]最先将其应用到了场景稠密重建的相关工作中. 面元是三维空间中的一个圆形平面, 由中心点三维坐标, 面元的半径, 面元的法向量, 面元的 RGB 颜色以及时间戳组成. 建图时, 对于每一帧深度图, 使用中心差分 (central difference) 的方式计算出法向量图, 然后在图像中采样得到一部分点作为面元的中心点. 对于部分重叠的面元, 可以使用高斯核对其进行卷积, 然后取加权平均即可. 面元表示地图中的面元之间没有任何连接关系, 无需维护面元之间的拓扑关系, 使得地图表示形式更加简单. 但一个面元元素中保存的信息较多, 面元表示地图的体积随着场景的增大将会快速增长.

2.4 语义地图

语义地图并不能完全算作一种全新的数据形式, 而是在一些现有的数据形式 (如 3D 点云) 上提取出更高级的语义信息并保存的一种地图形式, 可以让机器人更深层次的理解周围的环境. 语义地图的思想在很早之前就已经被提出过[71], 但受限于当时图像和点云的检测和分割技术的困难, 无法在大规模地图中加入准确而有效的语义信息. 近年来基于深度学习的图像和点云的语义分割技术发展迅速, 语义地图也随之发展起来.

目前主流的语义地图主要是基于点云的 3D 语义地图, 其构建方法大致可以分为离线处理和在线建图两种. (1) 离线处理: 对于已有的点云地图, 可以采用语义分割算法[72]对其进行语义分割, 将每个点都标记上对应的语义标签, 结合人工调整提高地图精度. 在此基础上还可以将地图处理为包含语义信息的体素形式表示[21]. (2) 在线建图: 另一种是基于语义 SLAM 方法[41]进行实时建图, 通常是在现有 SLAM 系统中集成语义分割模块, 然后增量式地融到 3D 地图中去.

由于图像在相同场景下因光照条件, 视角不同, 季节变化等外部环境的不同会产生较大的变化, 目前大部分相机重定位算法都会受到影响. 而语义信息不会受到这些因素干扰, 因此它能够提供更加稳定和丰富的信息用于相机重定位, 是跨模态定位的理想载体, 但是语义地图的制备也更加复杂.

2.5 高分辨率地图

对于一个可靠的自动驾驶系统, 车牌、车道线、路标等信息必不可少, 单纯的基于图像或者点云的先验地图无法满足自动驾驶的要求, 所以精度和体积更具优势的高分辨率地图 (high definition map, HD map) 成为自动驾驶地图的主流方案. 高分辨率地图专为无人驾驶相关应用服务, 主要通过多传感器融合方法构建. 目前无人驾驶技术正在高速发展, 高分辨率地图的定义和技术规范随之不断变化, 学界并没有一个统一的定义. 概括的来说, 高分辨率地图主要存储以下信息[73]: 1) 以坐标形式存储的点、线、面等几何元素所描述的道路, 建筑以及边界的抽象表示; 2) 以矢量的形式存储的车道的行车路径和方向; 3) 包含语义信息的车道标记, 交通标志和交通信号灯等地图元素; 4) 部分高分辨率地图有用于近距离精确感知和定位的 3D 点云图, 一般使用时动态加载. 由于高分辨率地图对地图精度、地图覆盖范围以及地图中包含的信息种类要求较高, 现阶段主流地图构建方案是通过安装有集成了差分 GNSS、惯性导航 IMU、相机和三维激光雷达等高精度测量设备的移动测量系统的专用地图采集车进行数据采集工作[42]. 主流高精度地图的构建流程大致分为: 1) 根据车载 GNSS/INS 采集的数据获取采集车的高精度 6-DoF 位姿信息和轨迹; 2) 从车载相机和雷达采集的数据中提取出道路、车道线、路标和交通信号等特定地图元素; 3) 将多种模态的数据进行融合, 最终建立起完整的高精度地图.

相比于其他类型的地图, 高分辨率地图具有如下特点: 1) 高分辨率地图高度结构化, 内存占用低, 易于部署和更新; 2) 高分辨率地图的绝对位置精度高; 3) 地图中包含的信息更全面. 近年来, 随着 OpenDRIVE[74]和 NDS 等各种高精度地图标准的发布, 主流地图厂商 (如 Tomtom 和 HERE) 的高分辨率地图正在蓬勃发展.

2.6 先验地图小结

表 1 对现有先验地图进行了总结. 点云地图数据库和图像数据库由于其构建过程相对简单, 技术比较成熟而成为基于手机重定位等应用中最广泛使用的地图. 但是点云地图包含的信息较少, 缺失场景纹理信息并且体积较大, 图像数据库则不够鲁棒, 其地图数据对光照变化, 季节更替等场景外观变化比较敏感. 稠密边界表示地图对场景重建的精度更高, 包含了更多的纹理信息, 但也需要占据更多存储空间, 因此往往用于室内场景的表示. 语义地图由于保存了地图元素的语义信息, 对环境变化等因素不敏感, 同时也是跨模态匹配的理想载体. 高分辨率地图则主要是专为无人驾驶应用而设计, 具有体积小, 精度高等特点, 但是其制备较为复杂. 总之, 不同类型的先验地图具有不同的特点, 为提高查询速度, 各类地图主要采用 Octree, KD-Tree, Hash 表等方式进行存储. 视觉重定位算法在设计时需要根据实际应用场景进行合理的先验地图选择. 本文在第 3 节中首先介绍基于图像数据库地图的重定位方法.

表 1 常见先验地图数据库总结表

先验地图		传感器	基本元素	存储方式	查询速度	优势	劣势	场景
点云地图	雷达点云地图[39,56,57]	激光雷达	3D点	Octree/KD-Tree	O(log$N$)	精度相对高 不易受外部环境变化的影响 纹理较少的环境也可以工作	对不平滑的运动, 极端天气敏感 缺失纹理信息 体积较大	室内/室外
	图像点云地图[47–49]	相机	带图像特征的3D点	KD-Tree	O(log$N$)	数据采集方便 保留了一定的图像特征	建图精度易受光照等环境变化干扰 体积较大	
	体素点云地图[61,62]	—	体素	哈希表	O(1)	包含三维结构信息 计算量较小	存在精度损失 地图精度受体素大小影响	
图像数据库地图[43]		相机	带有位姿标签的图像	—	—	数据采集方便 纹理信息丰富	地图精度依赖数据采集频率 易受光照变化, 季节更替的影响	室内/室外
稠密边界表示地图	三角网格地图[19]	相机	三角网格	Octree/KD-Tree	O(log$N$)	地图精细度高, 纹理信息丰富	计算复杂, 计算资源消耗大 地图体积较大 需要维护顶点间的拓扑关系	室内
	TSDF地图[68,69]	深度相机	体素	—	—	直接记录距离信息	计算复杂, 计算资源消耗大 地图体积较大 建图精度取决于体素大小 场景大小需要提前固定	
	面元表示地图[66,70]	深度相机	面元	LDC (layered depth cube) Tree	O(log$N$)	无需维护地图元素间拓扑关系 包含较多纹理信息	保存信息较多, 地图体积大	
语义地图[21,41,71]		多传感器	带有语义标签的地图元素	—	—	鲁棒性较高	制备较为复杂	室内/室外
高分辨率地图[42,73,74]		多传感器	矢量化地图元素	—	—	高度结构化, 地图体积较小 地图精度较高 地图中包含的信息较多	制备较为复杂 主要用于自动驾驶	室外

3 图像数据库地图中的视觉重定位

得益于图像数据库采集的便利性以及场景信息丰富的特点, 图像数据库在视觉重定位中得到了广泛的应用. 早期的使用图像数据库的定位方法主要是基于图像检索的工作. 随着深度学习技术在图像处理领域的不断发展, 使用位姿回归网络直接预测相机位姿的方法也逐渐涌现出来. 本节将从这两个方面分别进行介绍.

3.1 基于图像检索的定位方法

基于图像检索的定位方法输入信息为单张 RGB 图像, 其核心思想是首先从图像数据库中搜索到与查询图像

最相似的参考图像, 并根据查询图像与参考图像的相对位姿关系以及参考图像的位姿标签计算查询图像的位姿信息[43]. 本节将从图像检索方法、特征点的匹配与位姿解算以及图像外观归一化 3 个方面来进行介绍.

(1) 图像检索方法

目前主流视觉重定位算法使用的图像检索方法主要有 3 种: 基于视觉词袋模型 (bag of visual words) 的方法[75,76]、基于局部特征聚合向量 (vector of locally aggregated descriptors, VLAD)[77,78]的方法和基于神经网络的特征提取方法[79].

1) 视觉词袋模型借用文本搜索技术的思想, 通过对大量视觉特征进行聚类得到一个视觉词典, 并将图像根据视觉词典进行量化. 具体地, 对于一个图像数据库, 首先对每张图像都进行局部特征提取 (如 SIFT[58]、ORB[80]), 将得到的所有特征构建一个有 M 个叶子节点的 KD-Tree, 每个叶子节点则称为一个视觉词, 每张图像便可以使用一个 M 维的向量进行描述. 在使用时, 只需比较查询图像的词向量和数据库图像的词向量之间的相似度即可[81]. 为了加速搜索速度, 每个视觉词还维护一个逆向索引表, 记录包含该单词的图像, 在进行词向量比较时只需要比较共享相同视觉词的图像即可[75].

2) VLAD[77]方法受 Fisher Vector[82]启发, 将局部图像特征聚合为一个向量来表示整张图像. 具体的, 首先对每张图像都提取 N 个局部 D 维特征点 (如 SIFT[58]、ORB[80]等). 然后对全部的视觉特征进行 K-means 聚类, 得到 K 个聚类中心, 记为 C_k. 然后根据如下公式累加每个维度上特征与聚类中心的差, 抹去图像本身的特征分布差异, 将每个 $N\times D$ 的特征图转化为一个全局特征 $V\in R^{K\times D}$.

$$V(k,j)=\sum_{i=1}^{N}a_k(x_i)(x_i(j)-c_k(j)),k\in K,j\in D \tag{1}$$

其中, x_i 表示第 i 个局部特征, c_k 表示第 k 个聚类中心. $a_k(x_i)$ 函数用来判断 x 是否属于聚类 k, 如果属于, $a_k(x_i)$=1, 如果不属于则 $a_k(x_i)$=0. VLAD 特征提取方法有多种后续变形. DenseVLAD[83]是通过聚合在每个图像的规则网格上密集采样的 RootSIFT[84]描述符来构建 VLAD 特征. NetVLAD[78]将 VLAD 特征的提取流程转化为了可以使用神经网络进行端到端训练的网络. 其主要改动是将经典 VLAD 算法[77]中不可导的 $a_k(x_i)$ 函数转化为了可导的权重函数, 即 x_i 与 c_k 越相近, 则函数值越接近 1, 反之则接近 0. NeXtVLAD[85]对 NetVLAD[78]进行了进一步的改进, 增加了 VLAD 网络结构中的非线性参数, 在保证性能的同时降低了 VLAD 的输出的特征维度. 除了基于 VLAD 的全局特征之外, 还有一些算法使用 GIST 描述子[86]来描述图像全局特征并基于此进行图像检索和定位[87−89].

3) 深度神经网络能很好地提取图像高维特征, 很多研究者将其用于图像检索. Laskar 等人[79]基于 ResNet-34[90]设计了一个孪生网络来同时处理一对图像, 使用图像之间的位姿差作为损失函数来进行迁移训练. 然后使用训练好的网络的一个分支对所有数据库图像提取高维特征用于图像检索. RelocNet[91]对损失函数进行了改进, 提出了使用相机截头体的重合度作为损失来进行训练. Radenović等人[92]则提出了一种无监督的对比学习方法对现有的图像分类网络进行微调并用于图像检索. 算法首先对一个大规模的无标注图像数据集进行聚类, 然后对每个聚类的图像使用 SfM 方法进行三维重建. 在训练时, 将同一类中距离较近的图片作为正样例, 将不同类中的图片作为负样例进行训练. TransGeo[93]将图像分割成小块, 并使用一种注意力引导的非均匀裁剪策略去除无信息的图像块, 重点关注信息量大的图像块进行计算, 在降低计算成本的同时提高了性能. Berton 等人[94]实现了一个完整基于图像数据库的定位框架, 将定位流程中的各个部分 (如特征提取, 特征聚合等) 进行了模块化, 可以对其中任意的模块进行更换和测试. 为了在不断变化的环境中进行稳定的视觉重定位, DrosoNet[95]提出了一种投票机制, 利用多个小而高效的分类器进行定位并对结果进行投票, 实现了比单一分类器更稳健和一致的定位效果, 提高了查询图像在视角变化和外观变化等情况下的定位准确性.

(2) 特征点的匹配与位姿解算

根据图像检索方法得到的图像之间的匹配关系, 便可以计算得到查询图像和参考图像之间的相对位姿. 传统做法是对查询图像和参考图像提取局部特征点, 寻找它们之间的匹配关系. 然后再通过对极几何得到相对位姿, 最后根据参考图像的位姿标签便可以推断出查询图像的位姿[38]. 使用单个图像对计算得到的相对位姿可能偏差较

大, Sattler 等人[16]提出使用由图像检索方法得到的 Top-K 张相似图片进行 SfM 重建, 再将这些图像的全局位姿标签和局部位姿对齐, 最终推断出查询图像的位姿, 解决了单个图像对相对位姿误差较大的问题. Zamir 等人[43]则使用了一种基于 GMCP (generalized minimum clique graphs) 的多重最近邻特征匹配算法来进行特征点的匹配.

很多研究者尝试使用深度学习的方法进行特征匹配和位姿求解. SuperGlue[96]使用图神经网络来解决图像之间特征点匹配的问题. 部分研究者[79,91]则使用深度神经网络对图像进行特征提取并直接预测图像之间的相对位姿变换. SuperGlue 虽然取得了很好的效果, 但是计算成本较高. 为了解决这个问题, ClusterGNN[97]对特征点进行聚类将其划分成不同的子图, 极大地降低了计算量. Zhou 等人[98]则使用基于 ResNet-34 网络[90]的孪生网络结构直接预测出图像对之间的基础矩阵 (essential matrix), 再使用基于 RANSAC 的流程即可求解出查询图像的精确位姿. 大部分图像之间的特征匹配方法都是从局部特征点到局部特征点, S2DNet[99]则从查询图像中提取出局部特征点, 然后使用网络直接预测出参考图像中对应的特征点. LoFTR[100]通过建立像素级别的密集匹配来实现无检测器方式的图像匹配. LoFTR 采用分层匹配的思想, 先通过 CNN 网络从查询图像和参考图像中提取出低分辨率特征图并预测得到像素级的匹配结果, 然后将低分辨率特征图中匹配的特征根据位置关系映射到重新提取的高分辨率特征图中, 再利用高分辨率特征图计算所有匹配特征的相似度, 得到最终匹配结果. 3DG-STFM[101]在 LoFTR 的基础上加入了深度匹配关系约束, 进一步提高了特征匹配的效果.

(3) 图像外观归一化

即使图像的拍摄位姿相同, 环境光照变化, 季节更替等因素也会使得图像之间的差异较大, 这会严重影响重定位算法的准确性. 为了解决图像外观变化过大的情况, 一些研究者尝试将不同拍摄条件的图像转化为正常拍摄条件下的图像. ToDayGAN[6]使用深度学习的方法, 根据夜间拍摄的图像生成其对应的白天的图像, 再使用 VLAD[77]提取图像特征用于图像定位. LE-net[102]同样基于 CNN 的网络架构来对弱光条件下的图像进行增强以用作视觉定位. Tang 等人[103]则尝试将图像中的位置信息和外观信息进行解耦合, 实现在图像外观变化较大情况下的定位.

3.2 基于位姿回归网络的重定位方法

基于位姿回归网络的重定位方法通过一个端到端的深度神经网络提取出查询图像的高维特征信息, 然后使用高维特征进行回归直接得到相机位姿预测结果[104]. 单张图像中包含的信息有限, 很多研究者探索了使用序列图像和 RGB-D 图像作为输入来预测相机位姿. 本节将从单张图像、序列图像和 RGB-D 图像做输入这 3 个角度进行介绍.

(1) 基于单张图像的位姿回归

PoseNet 是这种方法的开山之作[17]. PoseNet 是一个鲁棒并且实时的 6-DoF 位姿回归网络, 它使用 GoogLeNet[105]作为骨干网络提取图像的高维特征, 在基于 ImageNet 预训练的模型上进行迁移学习. 为了实现端到端的相机位姿预测, PoseNet 使用仿射回归器替换了 Softmax 分类器. PoseNet 的输出为一个向量, 包含相机 3D 位置 x 和旋转四元数 q, 训练时使用这两者和真实值的误差的加权和作为位姿损失函数, 后续很多工作都沿用了这种损失函数的设计.

$$loss(I)=\|\hat{x}-x\|_2+\beta\left\|\hat{q}-\frac{q}{q}\right\|_2 \tag{2}$$

其中, $\hat{x}$ 和 x 表示 3D 位置的预测值和真值, $\hat{q}$ 和 q 表示旋转的预测值和真值. 与其他方法相比较, PoseNet 的推理时间短且不受场景大小影响, 不需要进行特征提取, 对于运动模糊和环境变化具有更好的鲁棒性. 但是它也存在着模型泛化性差, 定位精度不如传统方法等问题.

基于 PoseNet, 很多研究者提出了改进方法, 并取得了不错的效果. Bayesian PoseNet[106]使用贝叶斯卷积神经网络作为其主干网络, 其使用 dropout 的方式来对模型权重进行采样, 最终在预测查询图像的位姿的同时给出了结果的不确定性, 并且在小样本数据集上取得不错的效果. Geometric PoseNet[25]提出可自适应的损失函数, 使用方差不确定性来自适应相机 3D 位置误差和旋转误差之间的权重, 相比于 PoseNet[17]中人工设定的权重, 能够对不同尺度的场景有更好的适应性. Geometric PoseNet 还提出使用重投影误差作为损失函数, 但是这种损失函数难以收敛, 只适合用来对其他已有损失函数的训练结果做进一步提升. Hourglass PoseNet[107]则对 PoseNet 的主干网络进行较

大的改进, 它使用编码器-解码器的网络结构来对图像特征进行提取. 编码器部分使用 ResNet-34[90]作为主干网络, 解码器部分使用上采样层和卷积层组成. LSTM PoseNet[44]把 PoseNet 最后的全连接层输出的高维向量重塑为二维矩阵的形式, 并在上、下、左、右 4 个方向上应用 LSTM[108]来提取特征. Naseer 等人[109]使用 VGG16 结构[110]替换了 PoseNet 中使用的 GoogLeNet[105]结构作为主干网络.

BranchNet[111]基于 GoogLeNet[105]网络设计了一个多任务 CNN 结构来分别对相机的旋转和平移进行估计. 同时, BranchNet 提出一种名为 Euler6 的全新的旋转表达形式, 使用 sin 和 cos 同时约束一个坐标轴上的旋转来解决欧拉角存在的周期性的问题. AtLoc[112]将注意力机制引入特征的提取过程中, 使得网络在提取特征时能够忽略场景中的动态部分, 更关注图像中比较稳定的部分, 从而提高定位的效果. AD-PoseNet[113]使用先验知识引导的 dropout 模块与自注意力模块相结合, 使得网络在提取图像高维特征时自动忽略对相机重定位没有指导作用的前景特征. 为了减弱光照变化对定位效果的影响, DFNet[114]使用 NeRF[115]模型生成正常曝光的图像来辅助网络训练, 提高了对光线变化的鲁棒性, 但这种方式额外需要每个场景对应的 NeRF 模型. SC-wLS[116]通过利用场景坐标回归[9]对最小二乘姿态回归进行加权, 以提高精度和效率.

上述方法虽然已经取得不错的定位效果, 但仍存在对新场景泛化性差, 需要重新训练模型的缺点. 为了解决这个问题, MSPN[117]提出一种多场景的位姿估计网络. MSPN 使用基于 ResNet-34[90]作为骨干网络, 首先对查询图像提取高维特征, 然后对特征按照场景进行分类, 对于特定场景的特征使用对应的全连接层来对位姿进行回归. MS-Transformer[118]与其思路相似, 使用两个 Transformer 网络[119]来对图像进行特征提取, 分别处理位置和旋转的信息特征. DiffPoseNet[120]提出了用于估计图像光流的 NFlowNet, 再通过可微的 cheirality 约束层来学习相机位姿估计, 避免了传统方法的场景依赖性和光流估计误差的问题.

(2) 基于序列图像的位姿回归

单张 RGB 图像进行位姿回归无法利用时序信息对相机的位姿做进一步约束. 因此很多学者研究如何利用序列图像作为输入来估计相机位姿. Melekhov 等人[121]使用两张连续有重叠的图像作为输入, 直接预测两张图像之间的相对位姿. 作者使用两个基于 AlexNet 结构[122]的网络组成一个孪生网络来分别对两张图片进行特征提取, 然后使用这两张图片的相对位姿的真实值和预测值的误差作为损失函数实现端到端的训练流程. VidLoc[123]则直接使用较长的图像序列作为输入, 然后对图像序列中图像的相机位姿进行估计. 为了能够更好地利用序列图像中的时序信息, VidLoc 将每一张图像使用 GoogLeNet[105]网络提取出来的高维特征按照时间顺序输入一个双向的 LSTM[108]网络结构中, 对每一张图像进行位姿估计.

VLocNet[124]借助辅助学习的方法, 使用两张连续图像作为输入, 以相机相对位姿估计任务来辅助相机全局位姿估计任务, 将相对位姿估计中的时序信息融合到了全局位姿估计的过程中, 同时完成里程计和相机重定位的任务. VLocNet++[125]延续 VLocNet[124]的思路, 在网络中加入语义分割任务作为辅助, 进一步提高相机重定位的精度, 在 7Scene 数据集[9]上取得和传统方法相媲美的定位效果.

Map-Net[126]则将深度学习的方法与传统的视觉里程计相结合来获得更加精准的绝对相机位姿. 其首先使用由两个 ResNet-34[90]网络组成的孪生网络来对两帧相邻图像的相对位姿进行预测, 然后将网络的预测结果输入到位姿图中进一步优化[127]. Xue 等人[128]沿用这种设计思想, 将网络的输入从图像对扩展为图像序列. GTCaR[129]则提出一种 Graph Transformer 网络对多视角/序列图像定位问题进行建模. 输入序列查询图像, GTCaR 将相应的图像姿势、图像特征和图像间相对运动建模为图并输入 Graph Transformer 求解相应的位姿.

(3) 基于 RGB-D 图像的位姿回归

除了利用序列图像作为输入, 部分工作还使用 RGB-D 图像作为输入. Li 等人[130]使用两个 GoogLeNet[105]网络组成一个孪生网络结构, 分别对 RGB 图像和深度图像进行处理, 最终联合两者的特征对相机的绝对位姿进行预测. 由于将深度图像直接输入到网络中得到的定位效果并不理想, 作者提出一种称为 MND (minimized normal + depth) 的新的深度图像编码方式, 同时能够保留深度图像中的相对结构信息和绝对深度信息.

(4) 基于位姿回归网络的重定位方法总结

表 2 对现有基于位姿回归网络的重定位方法进行总结, 报告了相关算法在 7Scenes 数据集[9]和 Cambridge 数据集[66]上的中位数定位精度. 表 2 及后续表中精度比较包括平移误差 (m) 及旋转误差 (°). 基于位姿回归网络的重定位方法大多是从主干网络、损失函数和查询信息的角度对 PoseNet[17]的改进, 近期的工作主要是融入了注意力机制, 进一步提升算法的精度和泛化性.

表 2 基于位姿回归网络的重定位方法总结表

查询信息	方法	主干网络	泛化性	精度 (平移误差/旋转误差)		特点
				7Scenes	Cambridge	
单张图像	PoseNet[17]	GoogLeNet	差	0.44 m/10.44°	2.09 m/6.84°	开创性工作
	Bayesian PoseNet[106]	GoogLeNet	差	0.47 m/9.81°	1.92 m/6.28°	引入了贝叶斯不确定性
	Geometric PoseNet[25]	GoogLeNet	差	0.23 m/8.12°	1.63 m/2.86°	提出了自适应的损失
	Hourglass PoseNet[107]	ResNet-34	差	0.23 m/9.53°	—	使用了编码器-解码器的网络架构
	LSTM PoseNet[44]	GoogLeNet	差	0.31 m/9.85°	1.30 m/5.52°	使用LSTM网络结构
	Naseer等人[109]	VGG16	差	—	1.33 m/5.17°	将主干网络替换为VGG16
	BranchNet[111]	GoogLeNet	差	0.29 m/8.30°	—	将选择和平移分开学习
	AtLoc[112]	ResNet-34	差	0.20 m/7.56°	—	引入注意力机制
	AD-PoseNet[113]	ResNet-34	差	—	1.60 m/4.21°	利用先验知识指导特征提取
	MSPN[117]	ResNet-34	好	0.20 m/8.41°	2.47 m/5.34°	通过多场景训练提升泛化性
	MS-Transformer[118]	Transformer	好	0.18 m/7.28°	1.28 m/2.73°	引入Transformer网络, 增加模型泛化性
	DFNet[114]	VGG16	好	0.12 m/3.71°	0.39 m/0.96°	使用NeRF[116]模型辅助训练
	SC-wLS[116]	—	好	0.03 m/0.99°	0.12 m/0.28°	融合场景坐标回归方法
	DiffPoseNet[120]	VGG16	好	—	—	结合光流信息约束相机位姿
序列图像	Melekhov等人[121]	ResNet-34	差	—	—	预测相对位姿变化
	VidLoc[123]	LSTM	差	0.25 m/—	—	对图像序列进行定位
	VLocNet[124]	ResNet-50	差	0.05 m/3.80°	0.78 m/2.82°	同时进行定位和里程计
	VLocNet++[125]	ResNet-50	差	0.02 m/1.39°	—	在VLocNet基础上加入语义分割
	Map-Net[126]	ResNet-34	差	0.21 m/7.77°	1.63 m/3.64°	引入位姿图进行优化
	Xue等人[128]	ResNet-34	差	0.19 m/7.47°	—	将Map-Net扩展至图像序列
	GTCaR[129]	Graph Transformer	差	0.18 m/5.13°	0.98 m/1.65°	引入Graph Transformer网络
RGB-D图像	Li等人[130]	GoogLeNet	差	0.35 m/10.22°	—	提出深度图像编码方式MND

3.3 图像数据库地图视觉重定位小结

对比基于图像检索和直接位姿回归的重定位方法, 图像检索类方法对新场景有一定扩展性, 只需要对新的场景图像提取特征并加入特征数据库即可. 但图像检索类方法的定位精度受场景图像采集间隔影响较大, 并且在图像外观变化较大的情况下, 定位效果会受到很大影响. 基于位姿回归网络的方法具有推理速度快, 对外观变化有一定的鲁棒性等特点. 但该类方法泛化性较差, 对新场景通常需要进行重新训练.

4 点云地图中的视觉重定位

点云地图包含雷达点云地图、图像点云地图和体素化点云地图, 查询图像和点云地图的模态不同, 无法进行直接比较, 需要首先将两者转换到同一模态下再进行查询和比较. 图像点云地图中包含图像特征信息, 可借助图像特征建立查询图像和图像点云地图之间的数据关联; 雷达点云地图中只有坐标和颜色信息, 可通过两种方式进行转换: (1) 3D→2D, 即将 3D 点云投影为图像再进行定位[35]; (2) 2D→3D, 即将 2D 图像转换为三维场景数据再进行定位[18,28]. 本节将首先介绍在图像点云地图中基于特征点匹配的重定位方法, 之后介绍在雷达点云地图中不基于特征匹配的 3D→2D 以及 2D→3D 的两类重定位方法. 本节整体结构如图 2 所示.

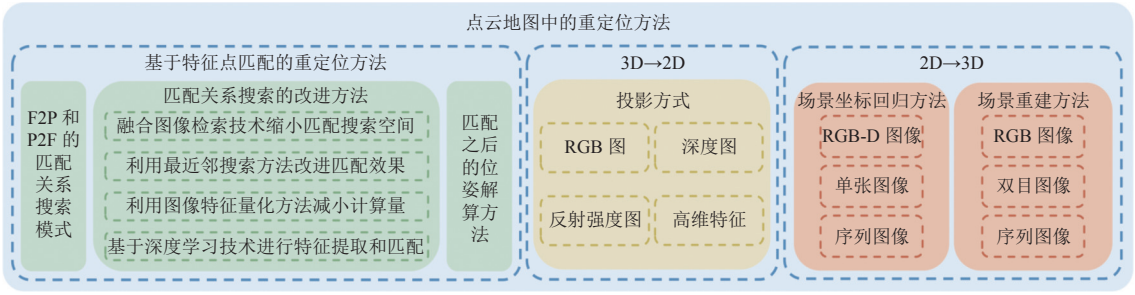

图 2　点云地图中的重定位方法

4.1　基于特征点匹配的重定位方法

图像点云地图带有图像特征, 可以利用这些特征信息进行重定位, 其重定位流程主要包括两个关键步骤: 1) 匹配关系搜索; 2) 位姿关系解算. 具体的, 给定需要查询的图像和图像点云地图, 定位时首先对查询图像提取出一定数量的特征点, 然后将二维图像特征点和图像点云根据描述子之间的相似性建立起对应关系, 再使用 PnP+RANSAC 进行位姿解算[1]. 其中位姿关系解算方式较为一致, 匹配关系搜索方法是各类方法研究的关键, 所以本节我们首先介绍 F2P 和 P2F 的匹配关系搜索模式; 然后介绍匹配关系搜索的各类改进方法; 最后介绍基于匹配结果的位姿解算方法.

(1) F2P (feature to point) 和 P2F (point to feature) 的匹配关系搜索模式

早期工作从查询图像中提取出特征点后, 直接通过暴力搜索的方式将查询图像的特征在图像点云中进行比较来确定匹配关系[1], 这种搜索模式也被称为 F2P (feature to point) 模式. Irschara 等人[60]使用所有 3D 点的图像特征构建词汇树进行图像特征的匹配. 当场景较大时, 使用查询图像的特征在整个点云中进行搜索效率很低, 研究者尝试利用 3D 点之间的共视关系提高搜索效率. Li 等人[13]提出使用视觉点云在图像特征中进行搜索的方法, 即 P2F (point to feature) 模式. 算法将 3D 点按照场景重建时出现的频率从高到低进行排序, 然后按顺序与查询图像中的特征进行匹配. 一旦某个 3D 点匹配成功, 则所有和该 3D 点存在共视关系的点的优先级都会上调. 得到一定数量匹配关系后, 通过 DLT 算法[131]结合光束平差法即可求解相机位姿.

Sattler 等人[50]借鉴了上述基于优先级的搜索思路, 将 3D 点按照其图像特征存储在一个视觉词汇表中, 并将所有视觉词按照该视觉词下包含的 3D 点数量从低到高进行排序. 对每个查询图像的特征点按顺序在视觉词汇表中进行搜索. Li 等人[15]先通过最近邻搜索查找图像 2D 特征对应的 3D 点, 再用与该 3D 点具有共视关系的 3D 点在 2D 图像特征中查找匹配. 在使用 PnP+RANSAC 求解相机位姿时, 共视关系还用于指导 RANSAC 算法的采样. Liu 等人[12]用马尔可夫图编码所有 3D 点间的共视关系用于匹配图像特征点和 3D 点.

Active Search[14]对前人的方法进行了综合, 提出一种高效且有效的基于大规模图像点云的定位方法, 其算法流程如图 3 所示. 算法核心是一种基于优先级的 2D-3D 匹配查找策略. 算法首先通过最近邻搜索为每个查询图像中的特征查找可能匹配的 3D 点, 根据每个特征需要匹配的 3D 点的数量作为搜索代价, 对所有特征按照搜索代价从低到高的顺序进行排序并依次进行搜索. 为了恢复在 2D-3D 搜索期间由于量化失真而丢失的匹配, 算法在 F2P 匹配之后又加入 P2F 搜索机制. 最后, 算法利用点之间的共视关系排除不可能存在匹配的点, 提高了检索效率. 查询图像的精确位姿使用 PnP+RANSAC 流程进行求解.

(2) 匹配关系搜索的改进方法

在上述匹配关系搜索方法基础上, 研究人员采用不同的技术提升匹配的效果和效率.

● 融合图像检索技术缩小匹配搜索空间: 除了利用共视关系为特征搜索提供优先级, 部分研究者借鉴图像检索的思路来缩小 2D-3D 匹配的搜索空间. HFNet[132]通过单个编码器和 3 个预测头的网络同时输出图像的全局描述符、密集的局部描述符以及关键点的检测分数. 定位时, HFNet 采用分层定位策略, 先使用全局描述符缩小搜索范围, 再使用局部描述符进行 2D-3D 匹配. Rubio 等人[133]使用神经网络提取图像特征, 并使用该特征检索查询图

像的 TOP-K 个最相似的参考图像, 使用这些图像对应的 3D 点与查询图像的特征点进行匹配. Azzi 等人[88]对其进行改进, 使用 GIST 全局描述符[86]来表示图像, 并引入一种新的相似性度量的方法来进行图像检索, 提高了算法求解速度和定位精度.

● 利用最近邻搜索方法改进匹配效果: 由于现有搜索方法 (层次聚类树和局部敏感哈希) 对二进制特征效果不佳, 研究者对其进行了改进. Donoser 等人[134]将 2D 特征与 3D 点之间的匹配问题看作一个分类问题, 使用随机蕨分类器替换传统算法中使用的最近邻搜索方法, 使得匹配查找时间与点云模型的大小没有关系. 同时由于不需要同时加载所有 3D 点进行查找, 减少了算法的内存占用. Feng 等人[135]对其进行改进, 使用二进制特征点来构建图像点云, 并提出一种基于随机树[136]的更有效的图像特征最近邻搜索方法.

● 利用图像特征量化方法减小计算量: 部分研究者通过对图像点云本身进行优化来减少计算量, 提高匹配效率. Irschara 等人[60]对图像点云中每个 3D 点的所有图像特征应用 Mean shift 聚类[59]进行了量化, 减少了每个 3D 点所存储的图像特征的数量. Li 等人[15]则使用 3D 点对应的所有特征的均值作为该 3D 点的特征, 极大地减少了特征数量, 但是损失了较多的信息. Sattler 等人[137]提出新的 3D 点描述符量化方式来对模型进行压缩, 相比于对所有描述符取平均值或者删除部分特征, 该方法能够保留更多场景信息, 大大降低了内存占用. 这些方式在降低了点云特征数量的同时, 难免会增加图像特征的歧义性, 影响特征匹配的准确性. Sattler 等人[138]则提出了一种特征加权方案, 为存在歧义的图像特征分配较低的权重来提高整体的匹配效果.

● 基于深度学习技术进行特征提取和匹配: PixLoc[139]是一种场景无关的端到端相机位姿预测网络, 利用 CNN 网络对图像提取多尺度的特征图, 通过将特征图和点云地图进行对齐预测相机位姿. DA4AD[140]使用图像和其对应的局部点云图作为输入, 利用深度注意力机制从中提取出显著, 独特和稳定的关键点, 并使用 L3-Net 算法[141]进行位姿解算. Yu 等人[142]使用深度学习的方法分别从点云地图和查询图像中提取出 3D 和 2D 的线特征, 然后使用这些线特征来对齐查询图像和点云地图. 2D3D-MatchNet[143]则利用对比学习的方法训练神经网络, 将从图像中提取的 2D 特征 SIFT[58]特征点和从点云中提取的 ISS[144]特征点转化到相同维度从而直接进行直接比较和匹配. LCD[145]则对其进行了改进, 不需要先提取 SIFT[58]或 ISS[144]特征, 而是使用双流神经网络结构分别处理查询图像和点云, 直接提取跨模态的特征描述符用于匹配.

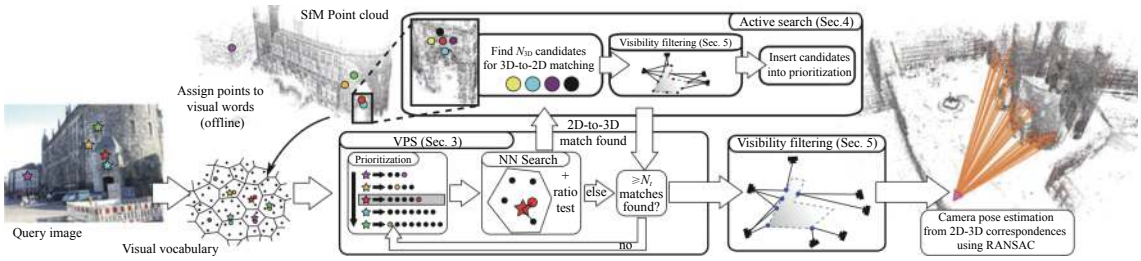

图 3 Active Search 算法流程图[14]

(3) 匹配之后的位姿解算方法

在获得查询图像的特征点与存储在地图中的图像的特征点的 2D-3D 匹配关系之后, 通过 PnP (perspective-n-point) 系列算法即可解算出相机最终位姿. P3P 算法[146]使用最为广泛, 只需 3 个点对和一对验证点即可求解出精确位姿, 与之类似的还有 P4P 算法[147]. EPnP 算法[148]则将世界坐标系下的点通过 4 个控制点进行表示, 通过求解控制点在相机坐标系下的坐标推断出位姿. DLT (direct linear transformation) 算法[131]则是构建一个增广矩阵, 通过投影矩阵构建一个 12 维线性方程组, 使用 6 个点对来进行求解. 此外, 还可以将相机位姿看作优化变量, 使用重投影误差构建一个 BA (bundle adjustment) 问题[149]进行求解.

在现实场景中得到的 2D-3D 匹配关系可能存在很多错误匹配的情况, 此时使用 PnP 方法求解的位姿会存在很大误差, 通常情况下会结合 RANSAC (random sample consensus) 算法[150]排除外点, 得到最终的相机位姿. RANSAC 算法通过迭代的方式, 不断的采集 2D-3D 点对子集并计算位姿矩阵, 使得内点数量尽量多, 重投影误差

尽量小, 以此得到最优结果. PROSAC[151]、MLESAC[152]、USAC[153]等则对经典 RANSAC 方法进行了改进, 节省了计算量, 提高了运行速度.

表 3 从特征点, 搜索机制, 检索方法, 位姿解算方法等角度对基于特征点匹配的重定位方法进行了详细的总结和对比. 从表中可以看出, 现有方法大多基于 SIFT 特征点构建图像点云地图, 使用 F2P 的方式建立数据关联, 使用 KD-Tree 加速特征检索速度, 并使用 PnP+RANSAC 模式解算相机位姿.

表 3 基于特征点匹配的重定位方法总结表

方法	特征点	搜索机制	检索方法	位姿解算	特点
Royer等人[1]	Harris角点	F2P	暴力匹配	P3P+RANSAC	早期工作
Irschara等人[60]	SIFT	F2P	词汇树	P3P+RANSAC	引入词汇树的检索方法
Li等人[13]	SIFT	P2F	词汇树	DLT+BA	提出基于优先级的P2F搜索机制
Sattler等人[50]	SIFT	F2P	KD-Tree	DLT+RANSAC	提出基于优先级的F2P搜索机制
Li等人[15]	SIFT	F2P	KD-Tree	P3P+RANSAC	利用了共视关系
Liu等人[12]	SIFT	F2P	KD-Tree	P4P+RANSAC	使用马尔可夫图编码共视关系
Active Search[14]	SIFT	F2P+P2F	KD-Tree	DLT+RANSAC	综合使用F2P和P2F的搜索机制 同时利用了共视关系
HFNet[132]	HFNet特征	F2P	—	P3P+RANSAC	使用神经网络同时提取全局和局部特征
Rubio等人[133]	SIFT	F2P	图像检索	REPPnP	使用图像检索的方法缩小特征搜索空间
Azzi等人[88]	SIFT	F2P	图像检索	PnP+RANSAC	使用GIST特征进行图像检索
Donoser等人[134]	SIFT	F2P	随机蕨分类器	PnP+PROSAC	使用随机蕨分类器搜索2D-3D匹配
Feng等人[135]	BRISK	F2P	随机树	DLT+RANSAC	使用二进制特征点和基于随机树的方法搜索2D-3D匹配
Sattler等人[138]	RootSIFT	F2P	KD-Tree	PnP+RANSAC	提出新的描述符量化方式
PixLoc[139]	—	—	—	—	场景无关端到端相机位姿预测
DA4AD[140]	DA4AD特征	—	—	L3-Net	利用深度注意力机制提取显著, 独特和稳定的关键点用于定位

4.2 3D→2D: 将待选点云投影为图像的重定位方法

这类方法输入信息为查询图像, 粗略的相机位姿以及点云地图, 其基本思想是首先根据某种方法 (GPS/IMU) 得到的先验相机位姿确定大致的搜索范围, 然后在该范围内产生一系列的位姿假设, 根据相机针孔模型将点云投影为一系列不同相机姿态下的 2D 图像. 得到周围 3D 点云地图的 2D 投影之后, 用查询图像再和这些投影进行比较, 根据相似度最高的投影图像对应的位姿即可得到查询图像的位姿.

Mastin 等人[35]使用点云地图对鸟瞰图进行定位. 定位时, 算法首先使用 GPS/INS 获得粗略的初始位姿, 在该位姿附近选取 100 个不同的位姿假设, 按照这些位姿将点云地图按照高程着色和按照反射强度着色两种方式投影为 2D 图像. 然后使用查询图像分别同上述两种投影图像计算互信息 (mutual information), 将两者相加作为相似度分数, 相似度最大的投影对应的位姿即为最终的相机位姿预测. 一些研究者[154,155]提出使用归一化信息距离 (normalised information distance) 来度量查询图像和投影图像之间的相似度. 在 3D 空间中进行 6-DoF 的位姿假设并生成投影图像需要大量的计算, Wolcott 等人[4]从两个方面减少了计算量: 首先是根据车辆定位的应用场景将问题简化为 3-DoF 的定位问题, 减少位姿假设的数量; 其次是在进行点云投影时, 算法在每个位置只对一个角度使用针孔模型生成投影, 并对该投影图像应用仿射变换来模拟生成多角度的图像, 其投影图像如图 4 所示. Neubert 等人[45]认为导致深度变化的特征也可能会产生视觉梯度, 因此在进行相似度比较之前, 算法先对查询图像和投影图像提取基于图像梯度的特征, 在归一化之后再计算两者的相似度.

CMRNet[156]则使用深度学习的方法直接预测生成的投影图像与查询图像之间的相对位姿, 并且使用了层次定位的策略逐步缩小定位误差. Sun 等人[157]使用深度神经网络生成出查询图像的深度图, 再将该深度图和从不同相

机位姿假设得到的深度图输入到相同的特征提取网络提取特征, 使用提取得到的特征之间的欧氏距离进行相似度计算. 在动态环境下的图像和静态的点云地图可能存在较大差异, 为了使定位结果更稳定, CPO[158]将点云投影为 RGB 图像, 并提出了一种 Score Map 机制用于衡量图像和场景中的动态程度, 重点关注稳定的位置用于定位. Cattaneo 等人[159]则利用 PointNet 网络[160]直接提取点云特征, 使用 VGG-16 网络[110]来对查询图像进行特征提取, 使用对比学习的方法使同一场景下的图像和点云所提取的高维特征尽可能相似, 而不同场景下的图像和点云的高维特征之间的距离要超过一个阈值. 查询时使用训练好的网络对查询图像和点云进行特征提取再比较特征之间的相似度即可.

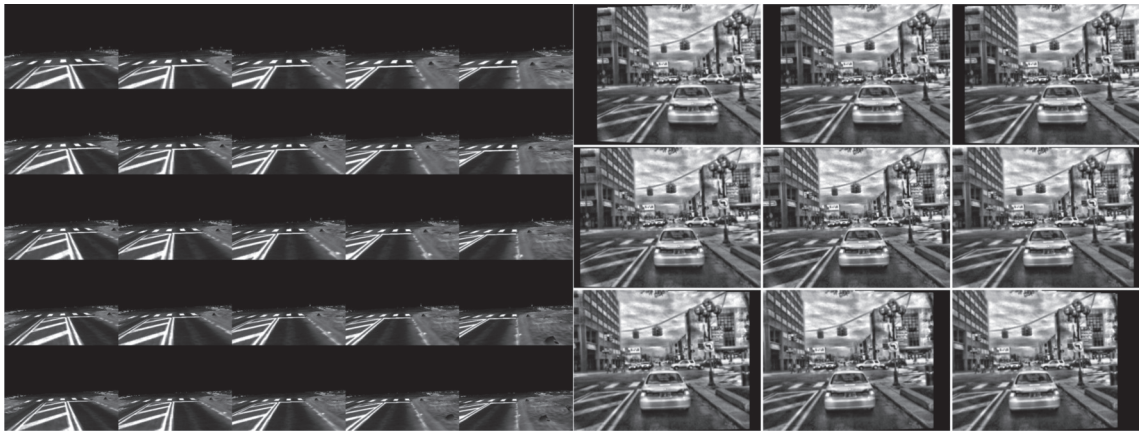

图 4 模拟生成多角度投影图像[4]

表 4 对现有 3D→2D, 基于候选点云投影到图像的重定位方法进行了详细的总结. 现有方法的区别主要在于投影方式, 相似度评价指标等.

表 4 现有基于点云投影的定位方法总结表

方法	投影方式	相似度评价	定位粒度	特点
Mastin等人[35]	RGB图	互信息	6-DoF	对鸟瞰图进行定位
Stewart等人[154]	RGB图	归一化信息距离	6-DoF	引入归一化信息距离
CPO[158]	RGB图	—	6-DoF	适用于动态场景
Pascoe等人[155]	反射强度图	归一化信息距离	6-DoF	使用GPU加速渲染
Wolcott等人[4]	反射强度图	归一化互信息	3-DoF	大幅降低了计算量
Neubert等人[45]	深度图	梯度互投影	6-DoF	使用图像的视觉梯度计算相似度
CMRNet[156]	深度图	—	6-DoF	利用神经网络计算查询和投影图像的相对位姿
Sun等人[157]	深度图	欧氏距离	3-DoF	通过神经网络提取图像高维特征来计算相似度
Cattaneo等人[159]	高维特征	欧氏距离	3-DoF	利用神经网络直接提取点云和图像高维特征计算相似度

4.3 2D→3D: 将 2D 查询图像升维到点云的重定位方法

将 2D 查询图像升维的重定位方法先将二维查询图像进行升维操作, 再和 3D 点云地图建立数据关联从而完成重定位. 现有研究工作主要有两种思路: (1) 使用随机森林或者神经网络模型直接预测出查询图像像素级的 3D 世界坐标, 即场景坐标回归方法; (2) 通过视觉里程计, 神经网络等方式恢复出查询图像对应的局部点云, 即查询图像场景重建的方法. 本节将从这两个方面对现有方法进行分析和总结.

(1) 场景坐标回归方法

场景坐标回归 (scene coordinate regression) 方法的输入信息包括查询图像和场景点云地图, 其核心思路是对 2D 查询图像进行升维操作, 使用随机森林或者深度神经网络的方法预测出查询图像中各像素对应的世界坐标, 建

立起图像与场景之间的 2D-3D 匹配关系, 再采用 PnP+RANSAC 的方法求解相机位姿.

场景坐标回归方法最初用于 RGB-D 图像在室内环境中的位姿估计[9]. 算法先使用 RGB-D 相机的真实位姿计算出所有深度图像中点的世界坐标, 然后根据 RGB 图像和深度图像之间像素级的对应关系, 训练一个随机森林来预测 RGB 图像中每个像素的世界坐标. 查询时, 使用 PnP+RANSAC 算法求解位姿.

Guzman 等人[27]对上述工作进行了改进, 提出一种混合多个预测器的架构, 使用多个随机森林对查询图像的场景坐标进行预测, 并设计一个选择器选择其中一个作为最终结果. Valentin 等人[32]将不确定性引入随机森林的预测中, 并将预测的不确定性考虑在内来进行连续姿态优化, 提高了模型的预测精度. Massiceti 等人[161]将已经训练好的随机森林转化为两层的神经网络结构[162], 再用一部分训练数据进行微调, 取得了更好的定位效果, 并且推理速度更快, 内存占用更少. 上述方法虽然取得了不错的定位精确度, 但是针对每个特定的场景都需要单独的离线训练. Cavallari 等人[163]将一个在通用场景中训练好的随机森林的叶子节点全部移除, 然后在新环境中定位时, 使用一个跟踪器提供相机的初始位姿, 然后使用该位姿结合在新场景中采集的数据来填充该随机森林的叶子结点, 从而达到适应新场景的目的. Meng 等人[164]则提出在场景坐标回归的流程中加入点特征和线特征的约束.

Brachmann 等人[18]提出可微的 RANSAC 算法—DSAC, 将传统的基于随机森林的场景坐标回归方法转换为基于 CNN 的可以进行端到端训练的方法, 其算法流程如图 5 所示. 输入一张 RGB 图像, DSAC 使用基于 VGG-16 的网络来预测图像中像素的 3D 场景坐标. 它将查询图像裁切为多个 40×40 的图像块并输入到场景坐标预测网络中, 便可以得到每个图像块中心点对应的 3D 场景坐标预测值. 和传统方法[9]类似, DSAC 也需要通过采样的方法得到一个相机位姿假设集合并且通过评价函数来对位姿假设进行评价. DSAC 使用一个基于 VGG-16 的网络结构来对位姿假设进行评价. 在选择位姿假设时, DSAC 使用一个可微的概率选择方式代替使用最高评分选择模型的确定性选择操作 argmax, 使得整个定位流程可以进行端到端的训练.

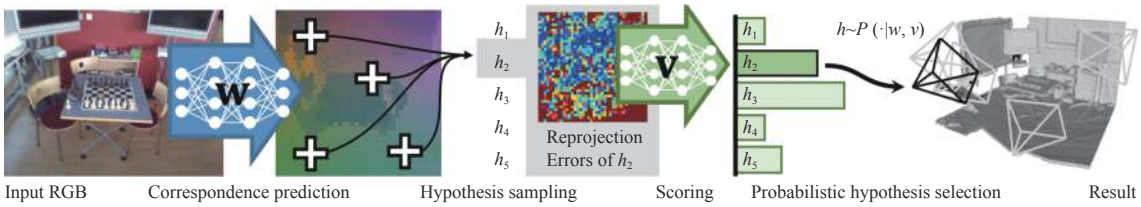

图 5 DSAC 算法流程图[18]

DSAC 虽然将场景坐标回归方法的整个流程转化为端到端的方式并取得不错的定位效果, 但仍存在着容易过拟合、必须使用深度信息进行训练以及位姿评价不够稳定的问题. DSAC++[28]对上述问题进行改进, 证明了将定位流程中所有的部分都设为学习目标并不一定是一种高效的方式. DSAC++以整张图像作为场景坐标预测网络的输入, 直接生成 60×80 的场景坐标预测图, 减少了 DSAC 中的重复性计算. DSAC++使用 Soft Inlier Counting 的方式对所有位姿假设进行评分, 替换了 DSAC 中基于 VGG-16[110]的评分网络, 减轻了模型过拟合的问题. 为了降低模型的评分数值不稳定性, DSAC++使用香农熵对概率分布进行处理, 使模型参数泛化性更好. Li 等人[46]对 DSAC++进一步进行改进, 提出基于射线角度的重投影误差作为损失函数, 使 DSAC++能在不仔细初始化的情况下对网络进行端到端的训练, 并可以提高定位精度. Cai 等人[165]将多视角几何约束加入 DSAC++的定位流程中.

ESAC[166]将 DSAC++算法与多专家模型 (MoE)[167]相结合, 训练了一组 DSAC++专家网络, 每个 DSAC++网络专门负责对某一特定场景下的相机位姿进行估计, 再使用一个门网络为查询图像分配专家网络来进行场景坐标预测, 一定程度上解决了 DSAC++在面对大规模或存在较多歧义的场景时效果不好的问题. HSC-Net[168]使用 K-means 算法将 3D 点云地图划分为不同层级的多个区域, 从粗到细的预测查询图像的场景坐标, 而不是直接得到预测结果, 提升了 DSAC++的泛化性以及在大场景下的表现. DSAC*[29]对 DSAC++进行进一步的提升, 把对数据的要求

降低到训练时最少只需要场景图像数据库以及对应的相机位姿, 查询时只需要单张 RGB 图像. 同时, DSAC*简化了训练流程, 将 DSAC++中的分步训练流程进行了统一. DSAC*还基于 ResNet-34[90]结构对场景坐标回归网络进行改进, 提升了网络的性能. KFNet[169]则将场景坐标回归方法的输入从单张图像扩展到序列图像. SANet[170]构建了一个全新的网络结构来预测查询图像的场景坐标, 尝试让网络能够与场景无关, 具有更好的泛化性, 它将输入的查询图像通过 NetVLAD[78]方法从数据库中检索出参考图像和其对应的场景坐标, 然后从粗到细的提取特征并进行融合来生成查询图像的场景坐标. Tang 等人[171]对 SANet 进行了改进, 使用深度学习的方法[172]进行图像检索, 并且网络并不直接输出最终的场景坐标预测, 而是使用稠密场景匹配模块 (DSM) 从粗到细的预测各级场景坐标图, 在最后一级得到精确的场景坐标图. SLD[173]不直接回归稠密的场景坐标图, 而是通过神经网络得到场景中稀疏的 Landmark 及其对应的方向向量进行定位. 在重复或稀疏纹理的场景, 直接使用高纬特征回归场景坐标容易导致性能下降. Xie 等人[174]尝试利用低维特征图中丰富的细节信息来解决这一问题.为了消除不同级别特征表示之间的差异, Xie 等人提出一种带有通道注意力模块 (channel attention module) 的深度特征聚合模块, 来学习鲁棒并独特的特征表示用于场景坐标预测, 提高了在重复或稀疏纹理区域中的特征判别能力.

表 5 对现有基于场景坐标回归的定位方法进行了总结, 报告了各方法在 7Scenes[9]和 Cambridge[17]数据集上的定位精度, 其中百分比数据表示定位误差小于 0.05 m/5°的结果的占比, 其余结果表示算法在相应场景下的中位数定位误差.

表 5　现有场景坐标回归方法总结表

查询信息	方法	主干网络	泛化性	性能		特点
				7Scenes	Cambridge	
RGB-D图像	Shotton等人[9]	随机森林	差	67.60%	—	开创性工作
	Guzman等人[27]	随机森林	差	79.30%	—	混合多个随机森林进行预测
	Valentin等人[32]	随机森林	差	89.50%	—	引入不确定性
	Massiceti等人[161]	多层感知机	差	0.04 m/1.9°	—	将随机森林转化为神经网络
	Cavallari等人[163]	随机森林	好	84.90%	—	对新场景具有适应能力
	Meng等人[164]	随机森林	差	92.70%	0.27 m/0.39°	使用点特征和线特征约束
	DSAC[18]	VGG	差	0.20 m/6.3°	0.32 m/0.78°	提出了可微的RANSAC算法
单张图像	DSAC++[28]	VGG	差	0.08 m/2.40°	0.28 m/0.50°	将DSAC改为仅使用RGB图像训练
	Li等人[46]	VGG	差	0.06 m/1.47°	0.24 m/0.40°	提出了基于射线角度的损失函数
	Cai等人[165]	VGG	差	0.05 m/1.63°	0.26 m/0.4°	加入了多视角几何约束
	ESAC[166]	多专家模型	差	0.034 m/1.5°	—	使用多专家模型预测场景坐标
	HSC-Net[168]	HSC-Net	差	0.03 m/0.90°	0.16 m/0.3°	使用了分层定位机制
	DSAC*[29]	ResNet	差	0.03 m/1.41°	0.18 m/0.4°	简化了DSAC++的训练流程
	SANet[170]	SANet	好	0.05 m/1.68°	0.84 m/0.8°	网络能够与场景无关
	Tang等人[171]	DSM	好	0.03 m/0.92°	0.20 m/0.3°	从粗到细预测场景坐标图
	SLD[173]	ResNet	好	0.03 m/1.07°	—	只预测稀疏Landmark及其方向向量
	Xie等人[174]	ResNet18	差	0.026 m/0.89°	—	利用低维特征图
序列图像	KFNet[169]	KFNet	差	0.03 m/0.88°	0.13 m/0.3°	将输入扩展为序列图像

(2) 基于查询图像场景重建的重定位方法

另一些研究者通过视觉里程计等方式来重建查询图像对应的局部点云并获得粗略的相机位姿, 再使用 ICP 方法和全局点云进行对齐来估计位姿. Li 等人[5]使用深度神经网络来预测单张 RGB 图像的深度图, 并使用 ICP 算法与全局点云地图进行匹配来估计相机位姿. Akai 等人[175]使用神经网络从 RGB 图像预测包含不确定性的深度图, 并提出一种新的概率模型来处理深度回归结果的不确定性, 提高了定位的鲁棒性. Sun 等人[7]使用 DSO[176]从单目相机图像中重建半稠密的 3D 局部点云, 使用改进的基于特征的点云配准方法和全局点云地图进行匹配, 并通过

采用可更新的尺度估计来解决尺度漂移的问题. Caselitz 等人[8]使用基于 ORB-SLAM[177]的视觉里程计从视频流重建出当前场景的局部点云并提供相机的粗略位姿, 然后使用 ICP 的方法在相机粗略位姿附近和全局点云地图进行对齐. 除了使用单目视觉里程计的方式重建局部点云, 部分研究者使用双目相机重建局部点云, 并使用正态分布变换 (normal distributions transform) 的方式和全局点云地图进行匹配来估计相机位姿[26,31]. DeepI2P[178]将图像在点云地图中的定位看作一个分类问题. DeepI2P 设计了一个深度神经网络, 根据查询图像对点云数据进行分类: 将属于查询图像视野范围内的点分为一类, 不在查询图像视野范围内的点分为另一类. 得到在视野范围内的点云之后, 相机的精确位姿便可以使用标准的高斯牛顿求解器对此无约束的连续优化问题进行求解. 表 6 对现有基于查询图像场景重建的方法进行总结, 报告了基于 KITTI[179]数据集的平均定位误差.

表 6　现有基于查询图像场景重建的方法总结表

查询信息	方法	重建方法	泛化性	KITTI定位精度 (平移误差/旋转误差)	特点
RGB图像	Li等人[5]	神经网络	差	—	使用神经网络预测深度图进行定位
	DeepI2P[178]	—	差	3.28 m/7.56°	将定位问题建模为分类问题
	Akai等人[175]	神经网络	差	—	预测包含不确定性的深度图
双目图像	Kim等人[26]	视差计算	好	0.18 m/0.34°	使用视差计算深度图进行定位
	Lin等人[31]	视差计算	好	1.94 m/1.06°	对点云地图下采样加速匹配
序列图像	Sun等人[7]	DSO	好	0.11 m/1.42°	改进点云配准方法减小尺度漂移
	Caselitz等人[8]	ORB-SLAM	好	0.3 m/1.65°	使用里程计重建局部点云用于定位

4.4 点云地图重定位小结

点云地图鲁棒性较好, 不容易受到环境光照, 季节更替等环境因素影响, 非常适合用于重定位任务. 但由于查询信息为图像, 和点云无法直接比较, 现有的研究主要关注如何将两者进行统一并进行比较. 基于点云投影的定位方法将点云投影为图像再和查询图像进行比较, 这类方法比较直观且易于操作, 但是由于生成多个位置, 多个视角的投影图像计算量较大, 往往需要结合其他方法获得一个粗略的位置来缩小采样空间. 基于查询图像升维的方法通过场景坐标回归或查询图像局部场景重建的方式对查询图像进行升维操作, 从而和全局点云地图建立数据关联. 场景坐标回归的方法通常用于室内等小规模场景, 虽然精度较高但模型泛化性较差; 查询图像局部场景重建的方式往往需要粗略的相机位姿以减少位姿搜索空间. 另一方面, 基于特征点匹配的方法则利用图像点云中带有的图像特征和查询图像建立起 2D-3D 匹配关系, 从而计算精确的相机位姿. 这类方法需要在大量特征中进行匹配搜索, 因此由于特征量化而引起的歧义性和搜索算法的效率会对定位效果产生较大的影响, 很多工作通过缩小特征搜索空间来提高算法的效果. 部分研究者尝试提取出图像和点云通用的特征点用于特征匹配, 但是实验效果仍有待提升.

5 稠密边界表示地图中的视觉重定位

稠密边界表示地图在表示场景的几何结构的同时保持了环境的表面信息, 使场景更加稠密和精细. 但由于稠密边界表示地图用于较大规模场景重建和定位的研究还处于发展阶段, 目前用这类地图进行视觉重定位研究的工作较少. 本节将从三角网格地图, SDF 表示地图和面元表示地图 3 个方面介绍几个代表性前沿工作.

5.1 三角网格地图

由于三角网格 (mesh) 地图可以连续的表示场景表面, Pascoe 等人[19]使用基于点云投影的重定位方法的思路, 根据一个粗略的相机位姿采样相机投影的位置, 然后将三角网格地图投影为 2D 图像, 使用归一化信息距离来和查询图像比较相似度从而对相机进行定位. 作者使用视角较广而畸变较大的广角相机采集图像, 为了减少图像去畸变时的信息损失, 算法使用具有相同畸变效果的相机参数来产生投影图像再和查询图像计算相似度. Mock 等人[36]构建了一个带有图像特征点的三角网格地图用于视觉重定位. 具体的, 作者先使用激光雷达和一个绑定的相

机采集环境的点云和图像信息, 然后使用 LVR2[180]根据雷达数据构建出场景的三角网格地图. 对于采集的图像, 作者根据相机和激光雷达的相对位姿计算出图像的位姿, 然后提取图像特征点并通过逆相机模型通过射线将特征点投射到三角网格地图上, 从而构建出带有图像特征点三角网格地图.

5.2 面元表示地图

面元 (Surfel) 表示地图在提供了场景的几何信息的同时, 还带有一定的图像信息和方向信息, 可以为现有的重定位算法提供更多约束. Ye 等人[51]利用 Surfel 地图渲染顶点和法线图来获得图像帧中特征点的全局平面信息, 然后将几何平面约束引入直接光度残差中, 使用紧耦合的光束平差法对相机位姿进行优化. Ye 等人[181]将图像数据和 Surfel 地图结合起来, 构建了带有几何信息的视觉数据库. 具体的, 作者先收集了场景的图像数据库和对应 Surfel 地图, 先对图像数据库中的图像提取图像特征点, 然后将特征点和 Surfel 地图中的面元建立关联, 利用面元重投影误差便可以优化视觉数据库中的图像姿势. 定位时, 首先在图像数据库中检索得到查询图像的参考图像, 然后在两者之间建立起特征点之间的匹配关系, 即可求解查询图像位姿.

5.3 SDF 表示地图

SDF 地图可以为场景中每个 3D 点提供其到场景表面的距离信息, 目前主流的研究都是基于场景对齐的思路展开研究. Huang 等人[30]使用 ORB-SLAM2[76]对相机进行追踪并构建局部 3D 特征图. 由于视觉里程计构建的 3D 点一定是场景表面上的点 (SDF 值为 0), 因此可以用该特征点的 SDF 值作为约束使用光束平差法对关键帧的相机位姿进行优化, 可以很好地解决视觉里程计的漂移问题. Millane 等人[182]则提出了一种基于 SDF 地图中的特征点用于定位. 具体的, 作者选择 SDF 地图中 SDF 值变化曲率较大的点作为关键点, 然后通过将特征点周围的平均 SDF 值添加到 SIFT 描述符的方向直方图中来构建其描述符. 定位时, 先使用 Voxgraph[183]构建出局部 SDF 地图, 然后提取局部 SDF 地图中的特征点, 并和全局 SDF 地图进行对齐从而计算相机位姿.

5.4 稠密边界表示地图重定位小结

表 7 对稠密边界表示地图中的重定位方法进行了详细的总结和对比, 并报告了算法的测试场景和对应的定位精度. 稠密边界表示地图对于场景的几何结构表示有着很大的优势, 逐渐引起越来越多的研究者的关注, 在路径规划、视觉导航等领域得到广泛的应用, 但在视觉重定位方面的研究还较少. 目前, 稠密边界表示地图在视觉重定位中的应用主要还是借鉴其他类型地图中的定位方法, 还需要进行更加深入的研究.

表 7 稠密边界表示地图中的定位方法总结表

地图	方法	查询信息	地图使用方式	实验数据集	实验效果	特点
三角网格地图	Pascoe等人[19]	单张图像	投影为2D图像	自制数据集	均方根误差: 0.57 m/1.03°	通过对齐查询图像和投影图像进行定位
	Mock等人[36]	序列图像	加上图像特征	EuRoC MAV[184]	均方误差: 0.0007 m	提取查询图像的特征并与地图构建数据关联
面元表示地图	Ye等人[51]	序列图像	渲染为法线图	EuRoC MAV[184]	平均定位误差: 0.12 m/0.29°	在光束平差法中加入平面约束
	Ye等人[181]	序列图像	构建有几何信息的视觉数据库	EuRoC MAV[184]	平均绝对轨迹误差: 4.55 m	提取查询图像特征点并与视觉数据库构建数据关联
SDF表示地图	Huang等人[30]	序列图像	用特征点SDF值作为约束	EuRoC MAV[184]	均方根误差为: 0.072 m	将特征点SDF值作为约束加入光束平差法优化位姿
	Millane等人[182]	RGB-D图像	提出SDF专用的特征点和描述子	3DMatch[185]	准确率为0.8时召回率: 35%	基于特征匹配的方法对齐子图和全局SDF地图

6 高分辨率地图中的视觉重定位

在高分辨率地图中的定位方法的核心思路是将图像中的道路元素和高分辨率地图中的道路元素进行对齐. 大部分主要遵循如下流程. 在初始化阶段, 系统通过车载 GPS 或里程计获得粗略位姿来缩小定位范围, 裁剪全局高

分辨率地图得到局部地图. 对于一张或者多张的车载相机拍摄的图像, 先通过图像处理手段得到图像中的道路元素集合, 和高分辨率地图中保存的地图元素集合建立对应关系, 通过最小化道路元素之间的重投影误差来对位姿进行优化.

早期由于目标检测和语义分割技术发展不够成熟, 高精度地图的制备以及在高精度地图中定位只能使用传统方法获得易于检测的道路标记等元素[11]. 地图制备时, 使用 active snake algorithm[186]在图像上检测道路标记的轮廓, 并记录轮廓内这些角点的相对像素位置, 再结合车载 GPS 信号确定这些标记点的全局位置, 从而构建包含大量地图标记的导航地图. 定位时同样使用车载相机拍摄图片, 然后将检测出的地图标记与地图进行对齐来实现定位. 这种方法仅适用于带有清晰彩绘标记的道路上的定位. 类似的, Schreiber 等人[187]通过检测图像中的道路标记和道路边缘并结合 GPS 信号来构建高精度地图, 在查询图像中检测相同类型的元素作为观测数据, 通过卡尔曼滤波的方式进行定位. Yu 等人[188]使用线分割算法[189]提取图像中的线特征, 并将线特征按方向分成了纵向, 横向和垂直方向 3 类, 然后分别和地图中的元素进行对齐.

近年来, 由于深度学习技术在图像处理领域的高速发展, 从 RGB 图像中进行像素级语义分割成为可能. 精确的语义分割能够提供更可靠和精确的信息用于相机位姿估计. 部分研究者利用双目相机采集图像并从中提取柱状路标用于和高分辨率地图进行匹配[190,191]. Xiao 等人[2]用 OpenDRIVE[74]标准来构建高分辨率地图. 算法通过语义分割提取出图像中的道路元素, 并将其转化为点特征和线特征: 对于交通信号牌等块形元素, 提取其中心点作为点特征; 对于车道线, 路灯等线型特征使用最小二乘法拟合为线特征. 在线定位时, 算法使用视觉里程计提供粗略位姿, 使用 RANSAC 算法的思想, 从高分辨率地图和图像特征集合中随机采样 3 个相同语义的特征作为对应关系, 生成位姿假设. 然后根据该位姿假设将高分辨率地图中的地图元素投影到相机平面, 计算和图像特征之间的距离, 如果距离小于一定阈值则认为是内点, 选择产生内点数量最多的相机位姿作为最终的相机位姿. Guo 等人[20]使用车道线, 交通信号板和柱状路标作为地图元素来构建高分辨率地图. 定位时, 算法先用轮式里程计获得车辆的粗糙位置, 然后在此位置将高分辨率地图中的地图元素投影到图像平面, 再使用基于深度学习的图像分割方法从查询图像中提取出相同类型的地图元素, 使用距离变换操作来计算投影图像和查询图像的语义标签之间的损失并进行优化, 最终得到相机的精确位姿. 相似的, Pauls 等人[33]使用一个多头神经网络分割出图像中的物体和车道, 然后和高精度地图中的地图元素的投影图像使用距离变换操作计算出损失图用于位姿优化. Li 等人[192]同样使用基于投影的方法. 算法首先使用视觉里程计给出粗略的相机位姿, 然后在此位姿附近对相机位姿进行采样, 然后将对高分辨率地图根据采样的位姿假设进行投影, 并和查询图像计算地图元素之间的距离, 从而推断相机位姿. Lu 等人[52]使用基于随机森林的边缘检测器[193]来提取图像中的道路标记的边缘特征, 并将边缘线特征进行采样, 使用一系列稀疏点来进行表示并以此构建高精度地图. 基于里程计提供的粗略位姿, 算法使用 Chamfer matching 方法[194]对齐图像中提取的线特征和地图中使用稀疏点表示的道路元素, 并基于此使用 Levenberg-Marquardt 算法进行位姿优化. Choi 等人[34]利用地图和图像的车道虚线的端点将车辆定位到车道的正确位置, 但只使用车道虚线限制了其应用场景和效果. 为了处理数据关联高度模糊的情况, LTSR[195]提出了一种基于新的本地特征描述符的鲁棒语义特征匹配方法, 以及一种准确、高效且简单的异常值剔除方法, 提高了算法的准确性和鲁棒性.

由于高分辨率地图的定义尚未统一, 因此目前的研究工作对于高分辨率地图的定义和其所用的定位算法的耦合关系较强, 表 8 对现有工作进行了分析和总结, 并给出了算法的平均定位误差: 对于定位粒度为 3-DoF 的算法, 我们给出了平移误差 (m); 对于定位粒度为 6-DoF 的算法, 我们给出了平移误差 (m) 和旋转误差 (°). 虽然地图元素之间存在差别, 总的来说都是矢量化形式存储的交通相关元素, 现有定位方法的核心思想也基本是将从图像中提取的道路元素和地图进行对齐从而进行进一步的优化. 随着 NDS (navigation data standard)、OpenDRIVE[74]等高分辨率地图数据标准的不断发展和推广, 高分辨率地图的定义将会逐渐趋于统一, 基于高分辨率地图的定位方法也会在此基础上取得更好的定位效果.

表 8 高分辨率地图中的定位方法总结表

查询	方法	地图元素	位姿估计方法	定位粒度	实验数据集	实验效果
单张图像	Ranganathan等人[11]	道路标记	P3P+RANSAC	3-DoF	自制数据集	—
	Schreiber等人[187]	道路标记, 道路边缘	卡尔曼滤波	3-DoF	自制数据集	0.11 m
	LTSR[195]	车道线, 交通信号牌	GNC[196]	3-DoF	自制数据集	—
双目图像	Sefati等人[190]	柱状路标	蒙特卡罗定位	3-DoF	KITTI[179]	—
	Spangenberg等人[191]	柱状路标	扩展卡尔曼滤波	3-DoF	自制数据集	0.167 m
序列图像	Xiao等人[2]	OpenDRIVE标准	RANSAC-based	6-DoF	自制数据集	—
	Guo等人[20]	车道线, 交通信号牌, 柱状路标	位姿图优化	6-DoF	自制数据集	0.29 m/0.48°
	Pauls等人[33]	车道线, 交通信号牌, 道路边缘	位姿图优化	6-DoF	自制数据集	0.30 m/0.54°
	Lu等人[52]	道路标记	LM算法	3-DoF	自制数据集	—
	Choi等人[34]	车道线	粒子滤波	3-DoF	自制数据集	0.248 m

7 语义地图中的视觉重定位

语义地图的地图元素中带有对应的语义标签, 可以为重定位算法提供更高级的信息. 一般情况下, 场景语义信息不会受到传感器的不同、环境光照变化、季节更替等因素的影响, 因此具有很强的鲁棒性, 可以根据语义标签之间的对应关系进行定位. 同时, 根据场景的语义信息可以指导重定位算法重点关注对定位更有利的元素. 场景所包含的语义信息本身也能看作是场景的一种全局特征, 可以对场景进行快速的检索. 本章将根据语义信息的使用方式对现有方法分类进行介绍.

7.1 全局语义特征

相比于传统的图像特征, 不同场景中所包含的语义信息之间的差异更加显著和鲁棒, 因此可以用场景的语义信息构建全局描述符. Schonberger 等人[21]使用 RGB-D 图像重建出包含语义信息的体素地图, 然后根据体素的语义和位置信息聚类得到多个 subvolume. 再使用神经网络对 subvolume 提取特征并使用基于 BoW 的方法构建特征词典, 从而构建描述整个场景的全局特征. 查询图像通过类似的流程得到全局特征便可以定位. Cinaroglu 等人[37]对语义分割图像提取 NetVLAD 全局特征, 将全局特征输入 CNN 网络进一步提取特征并使用对比学习的方法进行训练, 使得相同地点的特征更相似, 使用基于图像检索的方法进行定位. Orhan 等人[53]则使用对比学习的方法判断两张语义分割图像之间的相似性, 使用相似性分数对已有图像检索方法的定位结果进行验证和更新. Wang 等人[197]使用基于投影的方法在语义点云地图中进行定位. 算法根据相机的粗略位置将语义点云地图根据其语义标签投影为平面语义图, 然后和查询图像一起输入卷积神经网络中预测查询图像和平面语义图之间的相对位姿, 最终通过优化推断出查询图像的全局位姿. Wang 等人[198]使用文字信息作为地图中的语义信息, 基于室内的平面结构图以及图像信息构建了一个大型商场的语义地图用于视觉定位. 定位时同样从查询图像中提取出文本, 然后基于马尔可夫随机场的方法推断查询图像的位置. Radwan 等人[199]同样利用文字信息作为场景的全局特征. 作者从城市街道场景图像中提取出文字信息并加上位置标签, 查询时使用查询图像中的文字信息在地图中进行检索来完成定位.

7.2 语义对应关系

一般情况下, 查询图像中某个特征点所属的语义标签应当与其在先验地图中的对应点的语义标签相同, 利用这种对应关系可以缩小特征匹配的搜索范围或排除不稳定的匹配关系, 从而提高重定位算法的效率和精度. Stenborg 等人[54]在带有语义标签的 3D 点云地图中进行定位, 以解决长时间跨季节情况下的定位问题. 定位时, 先对查询图像进行语义分割, 通过语义标签建立起图像和点云之间的对应关系, 然后使用粒子滤波的方式来估计查询图像的位姿. Larsson 等人[200]使用 SfM 方法构建场景的图像点云地图, 同时对场景图像进行语义分割, 为图像点云加上语义特征. 定位时也使用相同的网络对查询图像进行语义分割, 定位时将 2D-3D 匹配中语义标签不相同的匹配删除以提高定位算法的准确性. Toft 等人[201]同样通过检查一对 2D-3D 匹配点的语义标签是否相同来为匹配关系评分, 根据此评分对 RANSAC 流程进行指导, 从而提升定位算法的表现. SemLoc[202]提出了一种混合约束,

使用 Dirichlet 分布来紧耦合地图中的语义和几何结构信息. 在前端跟踪局部地标及其语义状态的情况下, 通过期望最大化算法共同优化相机姿态和数据关联.

7.3 鲁棒特征选择

在现实场景中, 场景中往往包含很多动态的事物 (如行人, 车辆等), 从这些物体上提取出的特征很难找到正确的匹配关系. 而从一些静态事物 (如建筑) 上提取的特征往往更加鲁棒, 更有利于重定位算法建立正确的匹配. Yu 等人[55]对场景图像进行语义分割, 从语义分割图像中提取出场景的边缘信息, 将这些边缘转化为矢量表示并且用于定位. Seymour 等人[203]提出一种基于注意力机制的 CNN 网络来提取图像的全局特征, 网络将 RGB 图像和其语义分割图像进行融合, 使用语义信息指导模型关注更鲁棒的特征. Mousavian 等人[204]和 Naseer 等人[63]使用语义信息指导算法更关注人造建筑上的特征, 不属于人造建筑的图像特征会被降低权重或不用于定位. PixSelect[205]则使用神经网络直接检测出图像中稳定的特征点和其对应的世界坐标, 再使用 PnP+RANSAC 进行位姿计算.

7.4 语义地图重定位小结

表 9 对语义地图中的定位方法进行了详细的总结和对比, 并给出算法的测试场景和报告的定位精度. 对于相同的场景, 不同类型的传感器所采集到的数据的语义信息是相同的, 因此语义地图是跨模态定位方法的理想载体. 现阶段, 在语义地图上的定位方法的核心思想还是使用语义信息对传统方法进行增强, 如利用语义信息获取更好的全局信息或者寻找更稳定的数据关联.

表 9 语义地图中的定位方法总结表

语义信息利用方式	方法	语义地图形式	测试场景	性能	特点
全局特征	Schonberger等人[21]	体素地图	KITTI[179]	—	使用BoW方法描述场景全局特征
	Cinaroglu等人[37]	图像数据库	RobotCar Seasons[206]	Recall@1>70%	用NetVLAD特征描述场景语义信息
	Orhan等人[53]	图像数据库	Google Street View[10]	Recall@1=90%	使用对比学习的方法判断语义图像之间的相似性
	Wang等人[197]	点云地图	自制数据集	平均定位误差: 0.95 m/0.64°	使用神经网络预测投影语义图像和查询图像的相对位姿
	Wang等人[198]	图像数据库	自制数据集	Recall@1=46.43%	使用场景中文字信息进行定位
	Radwan等人[199]	图像数据库	Google Street View[10]	平均定位误差: 10.9 m	使用带有位置标签的文字进行定位
语义对应关系	Stenborg等人[54]	点云地图	CMU Seasons[206]	—	在语义标签相同的点之间搜索匹配
	Larsson等人[200]	点云地图	CMU Seasons[206]	93.1%的定位误差小于0.5 m/5°	删除语义标签不同的2D-3D匹配
	Toft等人[201]	点云地图	CMU Seasons[206]	57.9%的定位误差小于0.5 m/5°	根据语义标签是否相同为2D-3D匹配进行打分
	SemLoc[202]	点云地图	KITTI[179]	平均定位误差: 1.96 m	将语义和几何结构信息相结合
鲁棒特征选择	Yu等人[55]	图像数据库	KAIST Day/Night[207]	Recall@1=73%	从语义分割图中提取边缘信息
	Seymour等人[203]	图像数据库	RobotCar Seasons[206]	PR曲线的AUC: 77.9	用注意力机制融合语义和原始图像
	Mousavian等人[204]	图像数据库	自制数据集	Recall@1=70.6%	降低不属于人造建筑上特征的权重
	Naseer等人[63]	图像数据库	自制数据集	平均$F1$分数: 0.61	不属于人造建筑的特征不用于定位
	PixSelect[205]	图像数据库	Cambridge[17]	平均定位误差: 0.11 m/0.45°	用神经网络直接检测出稳定特征点

8 视觉重定位的常用数据集

为了能对重定位算法的效果进行比较全面的评估, 往往需要在较好的数据集上进行测试, 这样也方便和其他

相关工作进行对比. 另外, 对于基于深度学习的方法来说, 数据集的好坏和规模会对最终效果产生很大的影响. 因此, 我们整理了目前在视觉重定位领域被广泛使用的数据集, 总结了这些数据集的发表情况, 使用的传感器类型, 数据集大小, 采集场景, 采集地点以及采集的数据是否有较大的环境外观变化, 如表 10 所示.

表 10 视觉重定位常用数据集

数据集	传感器	大小	场景	环境变化	地点	语义标签
Google Street View[10]	相机	约10万张图像	室外	—	—	无
San Francisco[208]	IMU, GPS, 相机	约170万张图像	室外	—	San Francisco	无
Aachen Day-Night[206]	相机	3 047张图像	室外	日夜变化	Aachen	无
TUM-RGBD[209]	IMU, RGB-D相机	39条记录	室内	—	—	无
Dubrovnik 6K [15]	相机	约6 800张图像	室外	—	Dubrovnik	无
7Scenes[9]	RGB-D相机	7个场景	室内	—	—	无
KITTI[179]	IMU, GPS, LiDAR, 相机, 双目相机	39.2 km道路数据	室外	—	Karlsruhe	无
CMU Seasons[206]	GPS, 相机	约8万张图像	室外	季节变化	Pittsburgh	无
Cambridge[17]	相机	5个场景	室外	—	Cambridge	无
Oxford RobotCar[210]	IMU, GPS, LiDAR, 相机	约2 000万张图像	室外	多天气, 季节变化	Oxford	无
EuRoC MAV [184]	IMU, 双目相机	两个场景	室内	—	—	无
NCLT[211]	IMU, GPS, LiDAR, 相机	共147.4 km轨迹数据	室外	多天气, 季节变化	Michigan	无
Cityscape[212]	GPS, 相机	共25 000张图像	室外	—	多个城市	有
SceneNN[213]	RGB-D相机	上百个场景	室内	—	—	有
3DMatch[185]	RGB-D相机	62个场景	室内	—	—	无
SceneNet RGB-D[214]	RGB-D相机	约530万张RGB-D图像	合成场景	—	—	有
KAIST Day/Night[207]	IMU, GPS, LiDAR, 相机	约20万张图像	室外	日夜变化	Daejeon	无
Apollo Scape[215]	IMU, GNSS, LiDAR, 相机	约14万张图像	室外	天气变化	多个城市	有
ADVIO[216]	IMU, GNSS, 相机	4.5 km室内数据	室内	—	—	无
SemanticKITTI[217]	IMU, GPS, LiDAR, 相机, 双目相机	39.2 km道路数据	室外	—	Karlsruhe	有
nuScenes[218]	IMU, GPS, LiDAR, 相机	约140万张图像	室外	天气变化	多个城市	有
LaMAR[219]	IMU, LiDAR, 相机, WiFi, 蓝牙	45 000 m^2场景	室内/室外	天气, 日夜变化	—	无
SF-XL[220]	GPS, 相机	4 000万张图像	室外	天气, 日夜变化	San Francisco	无

9 机遇与挑战

在过去的几十年里, 大量研究者们对基于先验地图的视觉重定位方法进行了广泛而深入的研究, 取得了很大的成功. 表 11 对现有主流视觉重定位工作按照其使用的查询信息类型和地图类型进行了归纳总结. 表中第 1 行是基于重定位方法所使用的地图类型进行分类; 第 2 行基于对应的定位方法分类; 第 1 列是重定位方法所使用的查询信息类型. 从查询类型方面看, 基于单张图像的定位方案由于查询信息方便获取并且信息丰富, 在大部分主流地图中都是研究重点. 序列图像作查询的方法情况类似, 同样受到了研究者的重视. RGB-D 图像由于包含深度信息, 因此主要作为在点云地图中进行定位时的查询信息. 而双目图像作查询的研究工作目前较少. 从地图类型的角度来看, 图像数据库地图和点云地图作为目前最常见的两种地图形式, 可以满足大部分应用场景要求而得到了广泛的研究. 而稠密边界表示地图和高分辨率地图由于其自身特点而应用场景受限, 分别在室内和室外场景下得到了研究者的重视. 语义地图是一种新兴的地图形式, 其包含了更高级的地图信息, 相关方面的工作正在不断涌现, 但目前还主要集中在以单张图像作查询的定位方法. 总的来看, 目前视觉重定位领域的研究重点主要集中在以单张

图像作为查询的方法, 尤其是在图像数据库地图和点云地图中的定位方法. 随着 AR, 自动驾驶等技术对地图的要求不断提高, 基于稠密边界表示地图和高分辨率地图的定位技术正在快速发展中. 而语义地图中的语义信息使地图信息更加丰富和鲁棒, 具有很大的发展潜力.

表 11　视觉重定位方法分类

查询类型		单张图像	序列图像	RGB-D图像	双目图像
图像数据库	图像检索	文献[6,16,38,43,75–79, 81–103]	—	—	—
	位姿回归网络	文献[17,25,44,104–119]	文献[121–129]	文献[130]	—
点云地图	特征点匹配	文献[1,12–15,50,60,132–142, 139–145]	—	—	—
	3D→2D	文献[4,35,45,154–160]	—	—	—
	2D→3D	文献[5,28,29,46,166,168, 170,171,173,175,178]	文献[7,8,169]	文献[9,18,27,32, 161,163,164]	文献[26,31]
稠密边界表示地图	三角网格	文献[19]	文献[36]	—	—
	面元表示	—	文献[51,181]	—	—
	SDF表示	—	文献[30]	文献[182]	—
高分辨率地图	—	文献[11,187,195]	文献[2,20,33,34,52]	—	文献[190,191]
语义地图	全局语义特征	文献[21,37,53,197–199]	—	—	—
	语义对应关系	文献[54,200–202]	—	—	—
	鲁棒特征选择	文献[55,203–205]	—	—	—

但近年来自动驾驶, 增强现实等新兴技术的飞速发展对定位算法的精度, 效率和鲁棒性提出了更高的要求, 现有视觉重定位算法仍然面临巨大挑战, 也迎来新的机遇. 通过对现有方法的分析和总结, 我们认为以下问题仍然亟待解决.

(1) 语义信息的利用

目前, 大多数视觉重定位算法对语义信息的利用还停留在比较浅的层次, 相关的研究工作相对来说还较少. 语义信息可以从更高维度对场景进行表示和描述, 对语义信息的利用是未来视觉重定位算法设计的必然发展趋势. 如何以更好的方式利用语义信息, 还需要更多研究者继续探索.

(2) 大规模场景

目前, 很多工作都建立在提供相机粗略位姿这一假设基础上开展研究, 而这一假设在现实的大规模复杂场景中往往很难保证, 这会导致相关算法定位误差很大甚至失效. 基于图像检索的定位方法虽然可以在大规模环境下进行定位, 但是其定位精度却受图像数据库的采样间隔, 环境变化等因素的限制. 以 DSAC 等为代表的方法虽然在小规模室内数据集上取得了不错的成绩, 但在大规模室外场景下却很难收敛. 为了解决这个问题, 部分研究者提出由粗到细的层级定位机制, 逐步确定相机位姿, 在很多场景都取得了不错的成绩, 具有很大的研究潜力.

(3) 模型泛化性

越来越多研究者使用基于深度学习的方法来解决视觉重定位的问题, 并取得了很好的成绩. 但是大部分基于深度学习的方法都需要在场景相对应的数据上进行训练, 对数据集中未包含的新场景往往效果很差. 同时, 当场景范围过大时, 数据的收集和模型本身的容量都会成为问题. 因此, 如何提升基于深度学习方法的泛化性将会是一个非常值得研究的问题.

(4) 复杂环境条件

由于视觉传感器本身的特性, 即使是在相同场景采集的图像数据, 也会由于场景外观变化, 运动模糊等因素导致图像之间存在较大差异. 而在现实场景中, 复杂多变的天气状况, 早晚不同的光照条件以及季节更替等因素都会造成场景的外观发生变化, 这给基于视觉的重定位算法带来了很大的问题. 在图像外观发生较大变化的情况下进行精确的定位的问题, 将会成为一个值得关注的热点.

(5) 实时性

虽然目前很多研究工作已经取得了不错的精度, 但是其往往需要较为复杂的计算, 或者需要使用比较复杂的神经网络模型进行推理, 对部署平台的硬件条件具有一定要求, 因此无法很好地在性能较差的平台 (如移动设备等) 上实现实时性定位. 如何减少现有算法的计算成本, 使得其在保持定位精度的情况下具有较好的实时性也是一个需要继续探索的问题.

(6) 与 XR 应用结合

过去数十年间, VR, AR 和 MR 技术发展非常迅速, 这类融合现实技术统称为 XR (extended reality), 即扩展现实. 无论是基于哪种方法的 XR 应用, 都离不开 6-DoF 视觉重定位技术的支持. 然而, 受限于 XR 应用特殊的使用场景和硬件设备, 大部分现有视觉重定位算法无法直接应用到这些场景中. 如何更好地在 XR 应用场景下实现高精度的视觉重定位算法, 将会是未来的一个重点研究方向.

10 总 结

基于先验地图的视觉重定位研究受到了学术界和工业界广泛的关注, 并取得了很大的发展. 本文首先介绍了几种目前常见的先验地图形式, 分析了它们各自的特点. 然后将现有视觉重定位算法按照其先验地图的形态不同进行重新划分, 并根据各类方法自身的特点进行了全面的分析和总结, 并且整理了常用的数据集. 最后, 我们讨论了目前视觉重定位领域面临的新的挑战以及新的机遇. 随着新一代信息技术的不断进步, 我们相信视觉重定位领域具有广阔的发展前景.

致谢 感谢中国人民大学校级计算平台为本文提供技术支持.

References:

[1] Royer E, Lhuillier M, Dhome M, Lavest JM. Monocular vision for mobile robot localization and autonomous navigation. Int'l Journal of Computer Vision, 2007, 74(3): 237–260. [doi: 10.1007/s11263-006-0023-y]

[2] Xiao ZY, Yang DG, Wen TP, Jiang K, Yan RD. Monocular localization with vector HD map (MLVHM): A low-cost method for commercial IVs. Sensors, 2020, 20(7): 1870. [doi: 10.3390/s20071870]

[3] Li J, Wang CY, Kang XJ, Zhao Q. Camera localization for augmented reality and indoor positioning: A vision-based 3D feature database approach. Int'l Journal of Digital Earth, 2020, 13(6): 727–741. [doi: 10.1080/17538947.2018.1564379]

[4] Wolcott RW, Eustice RM. Visual localization within LiDAR maps for automated urban driving. In: Proc. of the 2014 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Chicago: IEEE, 2014. 176–183. [doi: 10.1109/IROS.2014.6942558]

[5] Li Q, Zhu JS, Liu J, Cao R, Fu H, Garibaldi JM, Li QQ, Liu BZ, Qiu GP. 3D map-guided single indoor image localization refinement. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 161: 13–26. [doi: 10.1016/j.isprsjprs.2020.01.008]

[6] Anoosheh A, Sattler T, Timofte R, Pollefeys M, Van Gool L. Night-to-day image translation for retrieval-based localization. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 5958–5964. [doi: 10.1109/ICRA.2019.8794387]

[7] Sun MH, Yang SW, Liu HZ. Scale-aware camera localization in 3D LiDAR maps with a monocular visual odometry. Computer Animation and Virtual Worlds, 2019, 30(3–4): e1879. [doi: 10.1002/cav.1879]

[8] Caselitz T, Steder B, Ruhnke M, Burgard W. Monocular camera localization in 3D LiDAR maps. In: Proc. of the 2016 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Daejeon: IEEE, 2016. 1926–1931. [doi: 10.1109/IROS.2016.7759304]

[9] Shotton J, Glocker B, Zach C, Izadi S, Criminisi A, Fitzgibbon A. Scene coordinate regression forests for camera relocalization in RGB-D images. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2930–2937. [doi: 10.1109/CVPR.2013.377]

[10] Zamir AR, Shah M. Accurate image localization based on Google maps street view. In: Proc. of the 11th European Conf. on Computer Vision. Heraklion: Springer, 2010. 255–268. [doi: 10.1007/978-3-642-15561-1_19]

[11] Ranganathan A, Ilstrup D, Wu T. Light-weight localization for vehicles using road markings. In: Proc. of the 2013 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Tokyo: IEEE, 2013. 921–927. [doi: 10.1109/IROS.2013.6696460]

[12] Liu L, Li HD, Dai YC. Efficient global 2D-3D matching for camera localization in a large-scale 3D map. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 2391–2400. [doi: 10.1109/ICCV.2017.260]

[13] Li YP, Snavely N, Huttenlocher DP. Location recognition using prioritized feature matching. In: Proc. of the 11th European Conf. on Computer Vision. Heraklion: Springer, 2010. 791–804. [doi: 10.1007/978-3-642-15552-9_57]

[14] Sattler T, Leibe B, Kobbelt L. Efficient & effective prioritized matching for large-scale image-based localization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1744–1756. [doi: 10.1109/TPAMI.2016.2611662]

[15] Li YP, Snavely N, Huttenlocher D, Fua P. Worldwide pose estimation using 3D point clouds. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 15–29. [doi: 10.1007/978-3-642-33718-5_2]

[16] Sattler T, Torii A, Sivic J, Pollefeys M, Taira H, Okutomi M, Pajdla T. Are large-scale 3D models really necessary for accurate visual localization? In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6175–6184. [doi: 10.1109/CVPR.2017.654]

[17] Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DoF camera relocalization. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 2938–2946. [doi: 10.1109/ICCV.2015.336]

[18] Brachmann E, Krull A, Nowozin S, Shotton J, Michel F, Gumhold S, Rother C. DSAC—Differentiable RANSAC for camera localization. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2492–2500. [doi: 10.1109/CVPR.2017.267]

[19] Pascoe G, Maddern W, Stewart AD, Newman P. FARLAP: Fast robust localisation using appearance priors. In: Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA). Seattle: IEEE, 2015. 6366–6373. [doi: 10.1109/ICRA.2015.7140093]

[20] Guo CC, Lin MJ, Guo HY, Liang PP, Cheng EK. Coarse-to-fine semantic localization with HD map for autonomous driving in structural scenes. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021. 1146–1153. [doi: 10.1109/IROS51168.2021.9635923]

[21] Schönberger JL, Pollefeys M, Geiger A, Sattler T. Semantic visual localization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6896–6906. [doi: 10.1109/CVPR.2018.00721]

[22] Piasco N, Sidibé D, Demonceaux C, Gouet-Brunet V. A survey on visual-based localization: On the benefit of heterogeneous data. Pattern Recognition, 2018, 74: 90–109. [doi: 10.1016/j.patcog.2017.09.013]

[23] Chen CH, Wang B, Lu CX, Trigoni A, Markham A. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. arXiv:2006.12567, 2020.

[24] Chen ZH, Pei HY, Wang JK, Dai DY. Survey of monocular camera-based visual relocalization. Robot, 2021, 43(3): 373–384 (in Chinese with English abstract). [doi: 10.13973/j.cnki.robot.200350]

[25] Kendall A, Cipolla R. Geometric loss functions for camera pose regression with deep learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6555–6564. [doi: 10.1109/CVPR.2017.694]

[26] Kim Y, Jeong J, Kim A. Stereo camera localization in 3D LiDAR maps. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 1–9. [doi: 10.1109/IROS.2018.8594362]

[27] Guzman-Rivera A, Kohli P, Glocker B, Shotton J, Sharp T, Fitzgibbon A, Izadi S. Multi-output learning for camera relocalization. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1114–1121. [doi: 10.1109/CVPR.2014.146]

[28] Brachmann E, Rother C. Learning less is more—6D camera localization via 3D surface regression. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4654–4662. [doi: 10.1109/CVPR.2018.00489]

[29] Brachmann E, Rother C. Visual camera re-localization from RGB and RGB-D images using DSAC. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5847–5865. [doi: 10.1109/TPAMI.2021.3070754]

[30] Huang HY, Sun YX, Ye HY, Liu M. Metric monocular localization using signed distance fields. In: Proc. of the 2019 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Macao: IEEE, 2019. 1195–1201. [doi: 10.1109/IROS40897.2019.8968033]

[31] Lin XH, Wang FH, Yang BS, Zhang WW. Autonomous vehicle localization with prior visual point cloud map constraints in GNSS-challenged environments. Remote Sensing, 2021, 13(3): 506. [doi: 10.3390/rs13030506]

[32] Valentin J, Nießner M, Shotton J, Fitzgibbon A, Izadi S, Torr P. Exploiting uncertainty in regression forests for accurate camera relocalization. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 4400–4408. [doi: 10.1109/CVPR.2015.7299069]

[33] Pauls JH, Petek K, Poggenhans F, Stiller C. Monocular localization in HD maps by combining semantic segmentation and distance transform. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 4595–4601. [doi: 10.1109/IROS45743.2020.9341003]

[34] Choi K, Suhr JK, Jung HG. In-lane localization and ego-lane identification method based on highway lane endpoints. Journal of Advanced Transportation, 2020, 2020: 8684912. [doi: 10.1155/2020/8684912]

[35] Mastin A, Kepner J, Fisher J. Automatic registration of LiDAR and optical images of urban scenes. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 2639–2646. [doi: 10.1109/CVPR.2009.5206539]

[36] Mock A, Wiemann T, Hertzberg J. Monocular localization in feature-annotated 3D polygon maps. In: Proc. of the 2021 European Conf. on Mobile Robots (ECMR). Bonn: IEEE, 2021. 1–7. [doi: 10.1109/ECMR50962.2021.9568810]

[37] Cinaroglu I, Bastanlar Y. Long-term image-based vehicle localization improved with learnt semantic descriptors. Engineering Science and Technology, an Int'l Journal, 2022, 35: 101098. [doi: 10.1016/j.jestch.2022.101098]

[38] Zhang W, Kosecka J. Image based localization in urban environments. In: Proc. of the 3rd Int'l Symp. on 3D Data Processing, Visualization, and Transmission (3DPVT 2006). Chapel Hill: IEEE, 2006. 33–40. [doi: 10.1109/3DPVT.2006.80]

[39] Zhang J, Singh S. LOAM: LiDAR odometry and mapping in real-time. In: Proc. of the Robotics: Science and Systems X. Berkeley, 2014. 1–9. [doi: 10.15607/RSS.2014.X.007]

[40] Mukasa T, Xu J, Bjorn S. 3D scene mesh from CNN depth predictions and sparse monocular SLAM. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 912–919. [doi: 10.1109/ICCVW.2017.112]

[41] McCormac J, Handa A, Davison A, Leutenegger S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 4628–4635. [doi: 10.1109/ICRA.2017.7989538]

[42] Editorial Department of China Journal of Highway and Transport. Review on China's automotive engineering research progress: 2017. China Journal of Highway and Transport, 2017, 30(6): 1–197 (in Chinese with English abstract). [doi: 10.3969/j.issn.1006-3897.2017.06.001]

[43] Zamir AR, Shah M. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1546–1558. [doi: 10.1109/TPAMI.2014.2299799]

[44] Walch F, Hazirbas C, Leal-Taixé L, Sattler T, Hilsenbeck S, Cremers D. Image-based localization using LSTMs for structured feature correlation. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 627–637. [doi: 10.1109/ICCV.2017.75]

[45] Neubert P, Schubert S, Protzel P. Sampling-based methods for visual navigation in 3D maps by synthesizing depth images. In: Proc. of the 2017 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 2492–2498. [doi: 10.1109/IROS.2017.8206067]

[46] Li XT, Ylioinas J, Verbeek J, Kannala J. Scene coordinate regression with angle-based reprojection loss for camera relocalization. In: Proc. of the 2019 European Conf. on Computer Vision. Munich: Springer, 2019. 229–245. [doi: 10.1007/978-3-030-11015-4_19]

[47] Wu CC. Towards linear-time incremental structure from motion. In: Proc. of the 2013 Int'l Conf. on 3D Vision. Seattle: IEEE, 2013. 127–134. [doi: 10.1109/3DV.2013.25]

[48] Wu CC, Agarwal S, Curless B, Seitz SM. Multicore bundle adjustment. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Colorado Springs: IEEE, 2011. 3057–3064. [doi: 10.1109/CVPR.2011.5995552]

[49] Schönberger JL, Frahm JM. Structure-from-motion revisited. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 4104–4113. [doi: 10.1109/CVPR.2016.445]

[50] Sattler T, Leibe B, Kobbelt L. Fast image-based localization using direct 2D-to-3D matching. In: Proc. of the 2011 Int'l Conf. on Computer Vision. Barcelona: IEEE, 2011. 667–674. [doi: 10.1109/ICCV.2011.6126302]

[51] Ye HY, Huang HY, Liu M. Monocular direct sparse localization in a prior 3D Surfel map. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 8892–8898. [doi: 10.1109/ICRA40945.2020.9197022]

[52] Lu Y, Huang JW, Chen YT, Heisele B. Monocular localization in urban environments using road markings. In: Proc. of the 2017 IEEE Intelligent Vehicles Symp. (IV). Los Angeles: IEEE, 2017. 468–474. [doi: 10.1109/IVS.2017.7995762]

[53] Orhan S, Guerrero JJ, Baştanlar Y. Semantic pose verification for outdoor visual localization with self-supervised contrastive learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans: IEEE, 2022. 3988–3997. [doi: 10.1109/CVPRW56347.2022.00444]

[54] Stenborg E, Toft C, Hammarstrand L. Long-term visual localization using semantically segmented images. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 6484–6490. [doi: 10.1109/ICRA.2018.8463150]

[55] Yu X, Chaturvedi S, Feng C, Taguchi Y, Lee TY, Fernandes C, Ramalingam S. VLASE: Vehicle localization by aggregating semantic edges. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 3196–3203. [doi: 10.1109/IROS.2018.8594358]

[56] Shan TX, Englot B. LeGO-LOAM: Lightweight and ground-optimized LiDAR odometry and mapping on variable terrain. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 4758–4765. [doi: 10.1109/IROS.2018.

8594299]

[57] Shan TX, Englot B, Meyers D, Wang W, Ratti C, Rus D. LIO-SAM: Tightly-coupled LiDAR inertial odometry via smoothing and mapping. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 5135–5142. [doi: 10.1109/IROS45743.2020.9341176]

[58] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

[59] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603–619. [doi: 10.1109/34.1000236]

[60] Irschara A, Zach C, Frahm JM, Bischof H. From structure-from-motion point clouds to fast location recognition. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 2599–2606. [doi: 10.1109/CVPR.2009.5206587]

[61] Muglikar M, Zhang ZC, Scaramuzza D. Voxel map for visual SLAM. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 4181–4187. [doi: 10.1109/ICRA40945.2020.9197357]

[62] Nießner M, Zollhöfer M, Izadi S, Stamminger M. Real-time 3D reconstruction at scale using voxel hashing. ACM Trans. on Graphics, 2013, 32(6): 169. [doi: 10.1145/2508363.2508374]

[63] Naseer T, Oliveira GL, Brox T, Burgard W. Semantics-aware visual localization under challenging perceptual conditions. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 2614–2620. [doi: 10.1109/ICRA.2017.7989305]

[64] Germain H, Bourmaud G, Lepetit V. Sparse-to-dense hypercolumn matching for long-term visual localization. In: Proc. of the 2019 Int'l Conf. on 3D Vision (3DV). Québec City: IEEE, 2019. 513–523. [doi: 10.1109/3DV.2019.00063]

[65] Osher S, Fedkiw R. Signed distance functions. In: Osher S, Fedkiw R, eds. Level Set Methods and Dynamic Implicit Surfaces. New York: Springer, 2003. 17–22. [doi: 10.1007/0-387-22746-6_2]

[66] Pfister H, Zwicker M, van Baar J, Gross M. Surfels: Surface elements as rendering primitives. In: Proc. of the 27th Annual Conf. on Computer Graphics and Interactive Techniques. ACM Press, 2000. 335–342. [doi: 10.1145/344779.344936]

[67] Rusu RB, Cousins S. 3D is here: Point cloud library (PCL). In: Proc. of the 2011 IEEE Int'l Conf. on Robotics and Automation. Shanghai: IEEE, 2011. 1–4. [doi: 10.1109/ICRA.2011.5980567]

[68] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras. In: Proc. of the 2013 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Tokyo: IEEE, 2013. 2100–2106. [doi: 10.1109/IROS.2013.6696650]

[69] Oleynikova H, Taylor Z, Fehr M, Siegwart R, Nieto J. Voxblox: Incremental 3D euclidean signed distance fields for on-board MAV planning. In: Proc. of the 2017 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 1366–1373. [doi: 10.1109/IROS.2017.8202315]

[70] Whelan T, Leutenegger S, Salas-Moreno RF, Glocker B, Davison AJ. ElasticFusion: Dense SLAM without a pose graph. In: Proc. of the Robotics: Science and Systems XI. Rome: MIT Press, 2015. [doi: 10.15607/RSS.2015.XI.001]

[71] Vasudevan S, Gächter S, Nguyen V, Siegwart R. Cognitive maps for mobile robots—An object based approach. Robotics and Autonomous Systems, 2007, 55(5): 359–371. [doi: 10.1016/j.robot.2006.12.008]

[72] Wu ZF, Shen CH, van den Hengel A. Wider or deeper: Revisiting the ResNet model for visual recognition. Pattern Recognition, 2019, 90: 119–133. [doi: 10.1016/j.patcog.2019.01.006]

[73] Li L, Li DG, Xing XY, Yang F, Rong W, Zhu HH. Extraction of road intersections from GPS traces based on the dominant orientations of roads. ISPRS Int'l Journal of Geo-information, 2017, 6(12): 403. [doi: 10.3390/ijgi6120403]

[74] Dupuis M, Strobl M, Grezlikowski H. OpenDRIVE 2010 and beyond-status and future of the de facto standard for the description of road networks. In: Proc. of the 2010 Driving Simulation Conf. Europe. Paris: INRETS Arcueil, 2010. 231–242.

[75] Gálvez-López D, Tardós JD. Real-time loop detection with bags of binary words. In: Proc. of the 2011 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. San Francisco: IEEE, 2011. 51–58. [doi: 10.1109/IROS.2011.6094885]

[76] Mur-Artal R, Tardós JD. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Trans. on Robotics, 2017, 33(5): 1255–1262. [doi: 10.1109/TRO.2017.2705103]

[77] Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3304–3311. [doi: 10.1109/CVPR.2010.5540039]

[78] Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 5297–5307. [doi: 10.1109/CVPR.2016.572]

[79] Laskar Z, Melekhov I, Kalia S, Kannala J. Camera relocalization by computing pairwise relative poses using convolutional neural network. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 920–929. [doi: 10.1109/ICCVW.2017.113]

[80] Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. In: Proc. of the 2011 Int'l Conf. on Computer Vision. Barcelona: IEEE, 2011. 2564–2571. [doi: 10.1109/ICCV.2011.6126544]

[81] Mur-Artal R, Tardós JD. Fast relocalisation and loop closing in keyframe-based SLAM. In: Proc. of the 2014 IEEE Int'l Conf. on Robotics and Automation (ICRA). Hong Kong: IEEE, 2014. 846–853. [doi: 10.1109/ICRA.2014.6906953]

[82] Perronnin F, Liu Y, Sánchez J, Poirier H. Large-scale image retrieval with compressed fisher vectors. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3384–3391. [doi: 10.1109/CVPR.2010.5540009]

[83] Torii A, Arandjelović R, Sivic J, Okutomi M, Pajdla T. 24/7 place recognition by view synthesis. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1808–1817. [doi: 10.1109/CVPR.2015.7298790]

[84] Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 2911–2918. [doi: 10.1109/CVPR.2012.6248018]

[85] Lin RC, Xiao J, Fan JP. NeXtVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification. In: Proc. of the 2019 European Conf. on Computer Vision (ECCV) Workshops. Munich: Springer, 2019. 206–218. [doi: 10.1007/978-3-030-11018-5_19]

[86] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int'l Journal of Computer Vision, 2001, 42(3): 145–175. [doi: 10.1023/A:1011139631724]

[87] Russell BC, Sivic J, Ponce J, Dessales H. Automatic alignment of paintings and photographs depicting a 3D scene. In: Proc. of the 2011 IEEE Int'l Conf. on Computer Vision Workshops (ICCV Workshops). Barcelona: IEEE, 2011. 545–552. [doi: 10.1109/ICCVW.2011.6130291]

[88] Azzi C, Asmar DC, Fakih AH, Zelek JS. Filtering 3D keypoints using GIST for accurate image-based localization. In: Proc. of the 2016 British Machine Vision Conf. York: British Machine Vision Association, 2016. 127.1–127.12.

[89] Hays J, Efros AA. IM2GPS: Estimating geographic information from a single image. In: Proc. of the 2008 IEEE Conf. on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008. 1–8. [doi: 10.1109/CVPR.2008.4587784]

[90] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[91] Balntas V, Li SD, Prisacariu V. RelocNet: Continuous metric learning relocalisation using neural nets. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 782–799. [doi: 10.1007/978-3-030-01264-9_46]

[92] Radenović F, Tolias G, Chum O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 3–20. [doi: 10.1007/978-3-319-46448-0_1]

[93] Zhu SJ, Shah M, Chen C. TransGeo: Transformer is all you need for cross-view image geo-localization. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 1152–1161. [doi: 10.1109/CVPR52688.2022.00123]

[94] Berton G, Mereu R, Trivigno G, Masone C, Csurka G, Sattler T, Caputo B. Deep visual geo-localization benchmark. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 5386–5397. [doi: 10.1109/CVPR52688.2022.00532]

[95] Arcanjo B, Ferrarini B, Milford M, McDonald-Maier KD, Ehsan S. An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments. IEEE Robotics and Automation Letters, 2022, 7(2): 2527–2534. [doi: 10.1109/LRA.2022.3140827]

[96] Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A. SuperGlue: Learning feature matching with graph neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 4937–4946. [doi: 10.1109/CVPR42600.2020.00499]

[97] Shi Y, Cai JX, Shavit Y, Mu TJ, Feng WS, Zhang K. ClusterGNN: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 12507–12516. [doi: 10.1109/CVPR52688.2022.01219]

[98] Zhou QJ, Sattler T, Pollefeys M, Leal-Taixé L. To learn or not to learn: Visual localization from essential matrices. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 3319–3326. [doi: 10.1109/ICRA40945.2020.9196607]

[99] Germain H, Bourmaud G, Lepetit V. S2DNet: Learning accurate correspondences for sparse-to-dense feature matching. arXiv:

2004.01673, 2020.

[100] Sun JM, Shen ZH, Wang Y, Bao HJ, Zhou XW. LoFTR: Detector-free local feature matching with transformers. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 8918–8927. [doi: 10.1109/CVPR46437.2021.00881]

[101] Mao RY, Bai C, An YT, Zhu FQ, Lu C. 3DG-STFM: 3D geometric guided student-teacher feature matching. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 125–142. [doi: 10.1007/978-3-031-19815-1_8]

[102] Li GF, Yang YF, Qu XD, Cao DP, Li KQ. A deep learning based image enhancement approach for autonomous driving at night. Knowledge-based Systems, 2021, 213: 106617. [doi: 10.1016/j.knosys.2020.106617]

[103] Tang L, Wang Y, Luo QH, Ding XQ, Xiong R. Adversarial feature disentanglement for place recognition across changing appearance. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 1301–1307. [doi: 10.1109/ICRA40945.2020.9196518]

[104] Sattler T, Zhou QJ, Pollefeys M, Leal-Taixé L. Understanding the limitations of CNN-based absolute camera pose regression. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 3297–3307. [doi: 10.1109/CVPR.2019.00342]

[105] Szegedy C, Liu W, Jia YQ, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1–9. [doi: 10.1109/CVPR.2015.7298594]

[106] Kendall A, Cipolla R. Modelling uncertainty in deep learning for camera relocalization. In: Proc. of the 2016 IEEE Int'l Conf. on Robotics and Automation (ICRA). Stockholm: IEEE, 2016. 4762–4769. [doi: 10.1109/ICRA.2016.7487679]

[107] Melekhov I, Ylioinas J, Kannala J, Rahtu E. Image-based localization using hourglass networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 870–877. [doi: 10.1109/ICCVW.2017.107]

[108] Shi XJ, Chen ZR, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015, 802–810.

[109] Naseer T, Burgard W. Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In: Proc. of the 2017 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 1525–1530. [doi: 10.1109/IROS.2017.8205957]

[110] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[111] Wu J, Ma LW, Hu XL. Delving deeper into convolutional neural networks for camera relocalization. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 5644–5651. [doi: 10.1109/ICRA.2017.7989663]

[112] Wang B, Chen CH, Lu CX, Zhao PJ, Trigoni N, Markham A. AtLoc: Attention guided camera localization. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 10393–10401. [doi: 10.1609/aaai.v34i06.6608]

[113] Huang ZY, Xu Y, Shi JP, Zhou XW, Bao HJ, Zhang GF. Prior guided dropout for robust visual localization in dynamic environments. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 2791–2800. [doi: 10.1109/ICCV.2019.00288]

[114] Chen S, Li XH, Wang ZR, Prisacariu VA. DFNet: Enhance absolute pose regression with direct feature matching. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 1–17. [doi: 10.1007/978-3-031-20080-9_1]

[115] Martin-Brualla R, Radwan N, Sajjadi MSM, Barron JT, Dosovitskiy A, Duckworth D. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 7206–7215. [doi: 10.1109/CVPR46437.2021.00713]

[116] Wu X, Zhao H, Li SK, Cao YD, Zha HB. SC-wLS: Towards interpretable feed-forward camera re-localization. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 585–601. [doi: 10.1007/978-3-031-19769-7_34]

[117] Blanton H, Greenwell C, Workman S, Jacobs N. Extending absolute pose regression to multiple scenes. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2020. 170–178. [doi: 10.1109/CVPRW50498.2020.00027]

[118] Shavit Y, Ferens R, Keller Y. Learning multi-scene absolute pose regression with transformers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 2713–2722. [doi: 10.1109/ICCV48922.2021.00273]

[119] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[120] Parameshwara CM, Hari G, Fermüller C, Sanket NJ, Aloimonos Y. DiffPoseNet: Direct differentiable camera pose estimation. In: Proc.

of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 6835–6844. [doi: 10.1109/CVPR52688.2022.00672]

[121] Melekhov I, Ylioinas J, Kannala J, Rahtu E. Relative camera pose estimation using convolutional neural networks. In: Proc. of the 18th Int'l Conf. on Advanced Concepts for Intelligent Vision Systems. Antwerp: Springer, 2017. 675–687. [doi: 10.1007/978-3-319-70353-4_57]

[122] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[123] Clark R, Wang S, Markham A, Trigoni N, Wen HK. VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2652–2660. [doi: 10.1109/CVPR.2017.284]

[124] Valada A, Radwan N, Burgard W. Deep auxiliary learning for visual localization and odometry. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 6939–6946. [doi: 10.1109/ICRA.2018.8462979]

[125] Radwan N, Valada A, Burgard W. VLocNet++: Deep multitask learning for semantic visual localization and odometry. IEEE Robotics and Automation Letters, 2018, 3(4): 4407–4414. [doi: 10.1109/LRA.2018.2869640]

[126] Brahmbhatt S, Gu JW, Kim K, Hays J, Kautz J. Geometry-aware learning of maps for camera localization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2616–2625. [doi: 10.1109/CVPR.2018.00277]

[127] Grisetti G, Kummerle R, Stachniss C, Burgard W. A tutorial on graph-based SLAM. IEEE Intelligent Transportation Systems Magazine, 2010, 2(4): 31–43. [doi: 10.1109/MITS.2010.939925]

[128] Xue F, Wang X, Yan ZK, Wang QY, Wang JQ, Zha HB. Local supports global: Deep camera relocalization with sequence enhancement. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 2841–2850. [doi: 10.1109/ICCV.2019.00293]

[129] Li XY, Ling HB. GTCaR: Graph transformer for camera re-localization. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 229–246. [doi: 10.1007/978-3-031-20080-9_14]

[130] Li RH, Liu Q, Gui JJ, Gu DB, Hu HS. Indoor relocalization in challenging environments with dual-stream convolutional neural networks. IEEE Trans. on Automation Science and Engineering, 2018, 15(2): 651–662. [doi: 10.1109/TASE.2017.2664920]

[131] Abdel-Aziz YI, Karara HM. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Photogrammetric Engineering & Remote Sensing, 2015, 81(2): 103–107. [doi: 10.14358/PERS.81.2.103]

[132] Sarlin PE, Cadena C, Siegwart R, Dymczyk M. From coarse to fine: Robust hierarchical localization at large scale. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 12708–12717. [doi: 10.1109/CVPR.2019.01300]

[133] Rubio A, Villamizar M, Ferraz L, Penate-Sanchez A, Ramisa A, Simo-Serra E, Sanfeliu A, Moreno-Noguer F. Efficient monocular pose estimation for complex 3D models. In: Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA). Seattle: IEEE, 2015. 1397–1402. [doi: 10.1109/ICRA.2015.7139372]

[134] Donoser M, Schmalstieg D. Discriminative feature-to-point matching in image-based localization. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 516–523. [doi: 10.1109/CVPR.2014.73]

[135] Feng YJ, Fan LX, Wu YH. Fast localization in large-scale environments using supervised indexing of binary features. IEEE Trans. on Image Processing, 2016, 25(1): 343–358. [doi: 10.1109/TIP.2015.2500030]

[136] Lepetit V, Fua P. Keypoint recognition using randomized trees. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1465–1479. [doi: 10.1109/TPAMI.2006.188]

[137] Sattler T, Havlena M, Radenovic F, Schindler K, Pollefeys M. Hyperpoints and fine vocabularies for large-scale location recognition. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 2102–2110. [doi: 10.1109/ICCV.2015.243]

[138] Sattler T, Havlena M, Schindler K, Pollefeys M. Large-scale location recognition and the geometric burstiness problem. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 1582–1590. [doi: 10.1109/CVPR.2016.175]

[139] Sarlin PE, Unagar A, Larsson M, Germain H, Toft C, Larsson V, Pollefeys M, Lepetit V, Hammarstrand L, Kahl F, Sattler T. Back to the feature: Learning robust camera localization from pixels to pose. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 3246–3256. [doi: 10.1109/CVPR46437.2021.00326]

[140] Zhou Y, Wan GW, Hou SH, Yu L, Wang G, Rui XF, Song SY. DA4AD: End-to-end deep attention-based visual localization for autonomous driving. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 271–289. [doi: 10.1007/978-3-

030-58604-1_17]

[141] Lu WX, Zhou Y, Wan GW, Hou SH, Song SY. L3-Net: Towards learning based LiDAR localization for autonomous driving. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 6382–6391. [doi: 10.1109/CVPR.2019.00655]

[142] Yu H, Zhen WK, Yang W, Zhang J, Scherer S. Monocular camera localization in prior LiDAR maps with 2D-3D line correspondences. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 4588–4594. [doi: 10.1109/IROS45743.2020.9341690]

[143] Feng MD, Hu SX, Ang MH, Lee GH. 2D3D-MatchNet: Learning to match keypoints across 2D image and 3D point cloud. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 4790–4796. [doi: 10.1109/ICRA.2019.8794415]

[144] Zhong Y. Intrinsic shape signatures: A shape descriptor for 3D object recognition. In: Proc. of the 12th IEEE Int'l Conf. on Computer Vision Workshops. Kyoto: IEEE, 2009. 689–696. [doi: 10.1109/ICCVW.2009.5457637]

[145] Pham QH, Uy MA, Hua BS, Nguyen DT, Roig G, Yeung SK. LCD: Learned cross-domain descriptors for 2D-3D matching. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11856–11864. [doi: 10.1609/aaai.v34i07.6859]

[146] Gao XS, Hou XR, Tang JL, Cheng HF. Complete solution classification for the perspective-three-point problem. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003, 25(8): 930–943. [doi: 10.1109/TPAMI.2003.1217599]

[147] Bujnak M, Kukelova Z, Pajdla T. A general solution to the P4P problem for camera with unknown focal length. In: Proc. of the 2008 IEEE Conf. on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008. 1–8. [doi: 10.1109/CVPR.2008.4587793]

[148] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem. Int'l Journal of Computer Vision, 2009, 81(2): 155–166. [doi: 10.1007/s11263-008-0152-6]

[149] Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW. Bundle adjustment—A modern synthesis. In: Proc. of the 2000 Int'l Workshop on Vision Algorithms: Theory and Practice. Corfu: Springer, 2000. 298–372. [doi: 10.1007/3-540-44480-7_21]

[150] Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981, 24(6): 381–395. [doi: 10.1145/358669.358692]

[151] Chum O, Matas J. Matching with PROSAC—Progressive sample consensus. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005). San Diego: IEEE, 2005, 1: 220–226. [doi: 10.1109/CVPR.2005.221]

[152] Torr PHS, Zisserman A. MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding, 2000, 78(1): 138–156. [doi: 10.1006/cviu.1999.0832]

[153] Raguram R, Chum O, Pollefeys M, Matas J, Frahm JM. USAC: A universal framework for random sample consensus. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(8): 2022–2038. [doi: 10.1109/TPAMI.2012.257]

[154] Stewart AD, Newman P. LAPS—Localisation using appearance of prior structure: 6-DoF monocular camera localisation using prior pointclouds. In: Proc. of the 2012 IEEE Int'l Conf. on Robotics and Automation. Saint Paul: IEEE, 2012. 2625–2632. [doi: 10.1109/ICRA.2012.6224750]

[155] Pascoe G, Maddern W, Newman P. Robust direct visual localisation using normalised information distance. In: Proc. of the 2015 British Machine Vision Conf. Swansea: British Machine Vision Association, 2015. 70.1–70.13. [doi: 10.5244/C.29.70]

[156] Cattaneo D, Vaghi M, Ballardini AL, Fontana S, Sorrenti DG, Burgard W. CMRNet: Camera to LiDAR-map registration. In: Proc. of the 2019 IEEE Intelligent Transportation Systems Conf. (ITSC). Auckland: IEEE, 2019. 1283–1289. [doi: 10.1109/ITSC.2019.8917470]

[157] Sun MH, Yang SW, Liu HZ. Convolutional neural network-based coarse initial position estimation of a monocular camera in large-scale 3D light detection and ranging maps. Int'l Journal of Advanced Robotic Systems, 2019, 16(6): 1729881419893518. [doi: 10.1177/1729881419893518]

[158] Kim J, Jang H, Choi C, Kim YM. CPO: Change robust panorama to point cloud localization. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 176–192. [doi: 10.1007/978-3-031-20077-9_11]

[159] Cattaneo D, Vaghi M, Fontana S, Ballardini AL, Sorrenti DG. Global visual localization in LiDAR-maps through shared 2D-3D embedding space. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 4365–4371. [doi: 10.1109/ICRA40945.2020.9196859]

[160] Charles RQ, Su H, Kaichun M, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 77–85. [doi: 10.1109/CVPR.2017.16]

[161] Massiceti D, Krull A, Brachmann E, Rother C, Torr PHS. Random forests versus neural networks—What's best for camera localization? In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 5118–5125. [doi: 10.1109/ICRA.2017.7989598]

[162] Sethi IK. Entropy nets: From decision trees to neural networks. Proc. of the IEEE, 1990, 78(10): 1605–1613. [doi: 10.1109/5.58346]

[163] Cavallari T, Golodetz S, Lord NA, Valentin J, Di Stefano L, Torr PHS. On-the-fly adaptation of regression forests for online camera relocalisation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 218–227. [doi: 10.1109/CVPR.2017.31]

[164] Meng LL, Tung F, Little JJ, Valentin J, de Silva CW. Exploiting points and lines in regression forests for RGB-D camera relocalization. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 6827–6834. [doi: 10.1109/IROS.2018.8593505]

[165] Cai M, Zhan HY, Weerasekera CS, Li KJ, Reid I. Camera relocalization by exploiting multi-view constraints for scene coordinates regression. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision Workshop (ICCVW). Seoul: IEEE, 2019. 3769–3777. [doi: 10.1109/ICCVW.2019.00469]

[166] Brachmann E, Rother C. Expert sample consensus applied to camera re-localization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 7524–7533. [doi: 10.1109/ICCV.2019.00762]

[167] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation, 1991, 3(1): 79–87. [doi: 10.1162/neco.1991.3.1.79]

[168] Li XT, Wang SZ, Zhao Y, Verbeek J, Kannala J. Hierarchical scene coordinate classification and regression for visual localization. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11980–11989. [doi: 10.1109/CVPR42600.2020.01200]

[169] Zhou L, Luo ZX, Shen TW, Zhang JH, Zhen MM, Yao Y, Fang T, Quan L. KFNet: Learning temporal camera relocalization using Kalman filtering. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 4918–4927. [doi: 10.1109/CVPR42600.2020.00497]

[170] Yang LW, Bai ZQ, Tang CZ, Li HH, Furukawa Y, Tan P. SANet: Scene agnostic network for camera localization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 42–51. [doi: 10.1109/ICCV.2019.00013]

[171] Tang ST, Tang CZ, Huang R, Zhu SY, Tan P. Learning camera localization via dense scene matching. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 1831–1841. [doi: 10.1109/CVPR46437.2021.00187]

[172] Gordo A, Almazán J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 241–257. [doi: 10.1007/978-3-319-46466-4_15]

[173] Do T, Miksik O, DeGol J, Park HS, Sinha SN. Learning to detect scene landmarks for camera localization. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 11122–11132. [doi: 10.1109/CVPR52688.2022.01085]

[174] Xie T, Dai K, Wang K, Li RF, Wang JH, Tang XY, Zhao LJ. A deep feature aggregation network for accurate indoor camera localization. IEEE Robotics and Automation Letters, 2022, 7(2): 3687–3694. [doi: 10.1109/LRA.2022.3146946]

[175] Akai N. Mobile robot localization considering uncertainty of depth regression from camera images. IEEE Robotics and Automation Letters, 2022, 7(2): 1431–1438. [doi: 10.1109/LRA.2021.3140062]

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

[177] Mur-Artal R, Montiel JMM, Tardós JD. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans. on Robotics, 2015, 31(5): 1147–1163. [doi: 10.1109/TRO.2015.2463671]

[178] Li JX, Lee GH. DeepI2P: Image-to-point cloud registration via deep classification. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 15955–15964. [doi: 10.1109/CVPR46437.2021.01570]

[179] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. The Int'l Journal of Robotics Research, 2013, 32(11): 1231–1237. [doi: 10.1177/0278364913491297]

[180] Wiemann T, Mitschke I, Mock A, Hertzberg J. Surface reconstruction from arbitrarily large point clouds. In: Proc. of the 2nd IEEE Int'l Conf. on Robotic Computing (IRC). Laguna Hills: IEEE, 2018. 278–281. [doi: 10.1109/IRC.2018.00059]

[181] Ye HY, Huang HY, Hutter M, Sandy T, Liu M. 3D Surfel map-aided visual relocalization with learned descriptors. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 5574–5581. [doi: 10.1109/ICRA48506.2021.9561005]

[182] Millane AJ, Oleynikova H, Lanegger C, Delmerico J, Nieto J, Siegwart R, Pollefeys M, Cadena Lerma C. Freetures: Localization in signed distance function maps. IEEE Robotics and Automation Letters, 2021. [doi: 10.1109/LRA.2021.3052388]

[183] Reijgwart V, Millane A, Oleynikova H, Siegwart R, Cadena C, Nieto J. Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps. IEEE Robotics and Automation Letters, 2020, 5(1): 227–234. [doi: 10.1109/LRA.2019.2953859]

[184] Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, Achtelik MW, Siegwart R. The EuRoC micro aerial vehicle datasets. The

Int'l Journal of Robotics Research, 2016, 35(10): 1157–1163. [doi: 10.1177/0278364915620033]

[185] Zeng A, Song SR, Nießner M, Fisher M, Xiao JX, Funkhouser T. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 199–208. [doi: 10.1109/CVPR.2017.29]

[186] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. Int'l Journal of Computer Vision, 1988, 1(4): 321–331. [doi: 10.1007/BF00133570]

[187] Schreiber M, Knöppel C, Franke U. LaneLoc: Lane marking based localization using highly accurate maps. In: Proc. of the 2013 IEEE Intelligent Vehicles Symp. (IV). Gold Coast City: IEEE, 2013. 449–454. [doi: 10.1109/IVS.2013.6629509]

[188] Yu YF, Zhao HJ, Davoine F, Cui JS, Zha HB. Monocular visual localization using road structural features. In: Proc. of the 2014 IEEE Intelligent Vehicles Symp. Dearborn: IEEE, 2014. 693–699. [doi: 10.1109/IVS.2014.6856539]

[189] von Gioi RG, Jakubowicz J, Morel JM, Randall G. LSD: A fast line segment detector with a false detection control. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(4): 722–732. [doi: 10.1109/TPAMI.2008.300]

[190] Sefati M, Daum M, Sondermann B, Kreisköther KD, Kampker A. Improving vehicle localization using semantic and pole-like landmarks. In: Proc. of the 2017 IEEE Intelligent Vehicles Symp. (IV). Los Angeles: IEEE, 2017. 13–19. [doi: 10.1109/IVS.2017.7995692]

[191] Spangenberg R, Goehring D, Rojas R. Pole-based localization for autonomous vehicles in urban scenarios. In: Proc. of the 2016 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Daejeon: IEEE, 2016. 2161–2166. [doi: 10.1109/IROS.2016.7759339]

[192] Li HP, Xue CL, Wen F, Zhang HB, Gao W. BSP-MonoLoc: Basic semantic primitives based monocular localization on roads. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021. 5470–5475. [doi: 10.1109/IROS51168.2021.9636321]

[193] Dollár P, Zitnick CL. Fast edge detection using structured forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(8): 1558–1570. [doi: 10.1109/TPAMI.2014.2377715]

[194] Barrow HG, Tenenbaum JM, Bolles RC, Wolf HC. Parametric correspondence and chamfer matching: Two new techniques for image matching. In: Proc. of the 5th Int'l Joint Conf. on Artificial Intelligence. Cambridge: Morgan Kaufmann Publishers Inc., 1977. 659–663.

[195] Wang HY, Xue CL, Tang Y, Li WL, Wen F, Zhang HB. LTSR: Long-term semantic relocalization based on HD map for autonomous vehicles. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation (ICRA). Philadelphia: IEEE, 2022. 2171–2178. [doi: 10.1109/ICRA46639.2022.9811855]

[196] Yang H, Antonante P, Tzoumas V, Carlone L. Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection. IEEE Robotics and Automation Letters, 2020, 5(2): 1127–1134. [doi: 10.1109/LRA.2020.2965893]

[197] Wang P, Yang RG, Cao BB, Xu W, Lin YQ. DeLS-3D: Deep localization and segmentation with a 3D semantic map. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5860–5869. [doi: 10.1109/CVPR.2018.00614]

[198] Wang SL, Fidler S, Urtasun R. Lost shopping! Monocular localization in large indoor spaces. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 2695–2703. [doi: 10.1109/ICCV.2015.309]

[199] Radwan N, Tipaldi GD, Spinello L, Burgard W. Do you see the bakery? Leveraging geo-referenced texts for global localization in public maps. In: Proc. of the 2016 IEEE Int'l Conf. on Robotics and Automation (ICRA). Stockholm: IEEE, 2016. 4837–4842. [doi: 10.1109/ICRA.2016.7487688]

[200] Larsson M, Stenborg E, Toft C, Hammarstrand L, Sattler T, Kahl F. Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 31–41. [doi: 10.1109/ICCV.2019.00012]

[201] Toft C, Stenborg E, Hammarstrand L, Brynte L, Pollefeys M, Sattler T, Kahl F. Semantic match consistency for long-term visual localization. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 391–408. [doi: 10.1007/978-3-030-01216-8_24]

[202] Liang SW, Zhang YZ, Tian R, Zhu DL, Yang LH, Cao ZZ. SemLoc: Accurate and robust visual localization with semantic and structural constraints from prior maps. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation (ICRA). Philadelphia: IEEE, 2022. 4135–4141. [doi: 10.1109/ICRA46639.2022.9811925]

[203] Seymour Z, Sikka K, Chiu HP, Samarasekera S, Kumar R. Semantically-aware attentive neural embeddings for 2D long-term visual localization. In: Proc. of the 30th British Machine Vision Conf. Cardiff: BMVA Press, 2019.

[204] Mousavian A, Košecká J, Lien JM. Semantically guided location recognition for outdoors scenes. In: Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA). Seattle: IEEE, 2015. 4882–4889. [doi: 10.1109/ICRA.2015.7139877]

[205] Altillawi M. PixSelect: Less but reliable pixels for accurate and efficient localization. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation (ICRA). Philadelphia: IEEE, 2022. 4156–4162. [doi: 10.1109/ICRA46639.2022.9812345]

[206] Sattler T, Maddern W, Toft C, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J, Kahl F, Pajdla T. Benchmarking 6-DoF outdoor visual localization in changing conditions. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8601–8610. [doi: 10.1109/CVPR.2018.00897]

[207] Choi Y, Kim N, Hwang S, Park K, Yoon JS, An K, Kweon IS. KAIST multi-spectral day/night data set for autonomous and assisted driving. IEEE Trans. on Intelligent Transportation Systems, 2018, 19(3): 934–948. [doi: 10.1109/TITS.2018.2791533]

[208] Chen DM, Baatz G, Köser K, Tsai SS, Vedantham R, Pylvänäinen T, Roimela K, Chen X, Bach J, Pollefeys M, Girod B, Grzeszczuk R. City-scale landmark identification on mobile devices. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 737–744. [doi: 10.1109/CVPR.2011.5995610]

[209] Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems. In: Proc. of the 2012 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Vilamoura-Algarve: IEEE, 2012. 573–580. [doi: 10.1109/IROS.2012.6385773]

[210] Maddern W, Pascoe G, Linegar C, Newman P. 1 year, 1000 km: The Oxford RobotCar dataset. The Int'l Journal of Robotics Research, 2017, 36(1): 3–15. [doi: 10.1177/0278364916679498]

[211] Carlevaris-Bianco N, Ushani AK, Eustice RM. University of Michigan north campus long-term vision and LiDAR dataset. The Int'l Journal of Robotics Research, 2016, 35(9): 1023–1035. [doi: 10.1177/0278364915614638]

[212] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 3213–3223. [doi: 10.1109/CVPR.2016.350]

[213] Hua BS, Pham QH, Nguyen DT, Tran MK, Yu LF, Yeung SK. SceneNN: A scene meshes dataset with annotations. In: Proc. of the 4th Int'l Conf. on 3D Vision (3DV). Stanford: IEEE, 2016. 92–101. [doi: 10.1109/3DV.2016.18]

[214] McCormac J, Handa A, Leutenegger S, Davison AJ. SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 2697–2706. [doi: 10.1109/ICCV.2017.292]

[215] Huang XY, Cheng XJ, Geng QC, Cao BB, Zhou DF, Wang P, Lin YQ, Yang RG. The apolloscape dataset for autonomous driving. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City: IEEE, 2018. 954–960. [doi: 10.1109/CVPRW.2018.00141]

[216] Cortés S, Solin A, Rahtu E, Kannala J. ADVIO: An authentic dataset for visual-inertial odometry. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 425–440. [doi: 10.1007/978-3-030-01249-6_26]

[217] Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, Gall J. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 9296–9306. [doi: 10.1109/ICCV.2019.00939]

[218] Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O. nuScenes: A multimodal dataset for autonomous driving. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11618–11628. [doi: 10.1109/CVPR42600.2020.01164]

[219] Sarlin PE, Dusmanu M, Schönberger JL, Speciale P, Gruber L, Larsson V, Miksik O, Pollefeys M. LaMAR: Benchmarking localization and mapping for augmented reality. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 686–704. [doi: 10.1007/978-3-031-20071-7_40]

[220] Berton G, Masone C, Caputo B. Rethinking visual geo-localization for large-scale applications. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 4868–4878. [doi: 10.1109/CVPR52688.2022.00483]

附中文参考文献:

[24] 陈宗海, 裴浩渊, 王纪凯, 戴德云. 基于单目相机的视觉重定位方法综述. 机器人, 2021, 43(3): 373–384. [doi: 10.13973/j.cnki.robot.200350]

[42] 《中国公路学报》编辑部. 中国汽车工程学术研究综述·2017. 中国公路学报, 2017, 30(6): 1–197. [doi: 10.3969/j.issn.1006-3897.2017.06.001]

蔡旭东(1999－), 男, 博士生, 主要研究领域为视觉重定位.

白雪薇(1997－), 女, 博士生, 主要研究领域为网络定位算法, 图优化与应用.

王永才(1978－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为智能感知, 网络定位, 定位建图算法与应用.

李德英(1965－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为物联网, 智能网络算法设计与分析.