E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)10-2880-32

论文引用格式:Ye H Q, Liu Y D and Shen S H. 2024. Lightweight visual-based localization technology. Journal of Image and Graphics, 29(10): 2880-2911(叶翰樵, 刘养东, 申抒含. 2024. 轻量化视觉定位技术综述. 中国图象图形学报, 29(10):2880-2911)[DOI:10.11834/jig.230744]

轻量化视觉定位技术综述

叶翰樵[1,2],刘养东[2],申抒含[1,2*]

1. 中国科学院大学人工智能学院,北京　100049; 2. 中国科学院自动化研究所,北京　100190

摘　要:视觉定位旨在从已知的三维场景中恢复当前观测图像的相机位姿。视觉定位技术具备低成本、高精度和易于集成等优势,是实现计算设备与真实世界建立智能交互过程的关键技术之一,如今获得了混合现实、自动驾驶等应用领域的广泛关注。作为计算机视觉领域长期探索的基础任务之一,视觉定位方法至今已取得显著的研究进展,然而现有方法普遍存在计算开销和存储占用过大等不足,这些问题导致视觉定位在移动端的高效部署和场景模型的更新维护方面存在困难,并因此在很大程度上限制着视觉定位技术的实际应用。针对这一问题,部分研究工作开始聚焦于推动视觉定位技术的轻量化发展。轻量化视觉定位旨在研究更加高效的场景表达形式及其视觉定位方法,目前正逐渐成为视觉定位领域重要的研究方向。本文首先回顾早期视觉定位框架,随后从场景表达形式的角度对具备轻量化特性的现有视觉定位研究工作进行分类。在各个方法类别下,分析总结其特点优势、应用场景和技术难点,并同时介绍代表性成果。进一步地,本文对部分轻量化视觉定位的代表性方法在常用室内外数据集上的性能表现进行对比分析,评估指标主要包含离线建图的用时、场景地图的存储占用和定位精度3个维度。现有的轻量化视觉定位技术仍然面临着诸多的难题与挑战,场景模型的表达能力、定位方法的泛化性与鲁棒性尚存在较大的提升空间。最后,本文对轻量化视觉定位未来的发展趋势进行分析与展望。

关键词:视觉定位;相机位姿估计;三维场景表达;轻量化地图;特征匹配;场景坐标回归;位姿回归

Lightweight visual-based localization technology

Ye Hanqiao[1,2], Liu Yangdong[2], Shen Shuhan[1,2*]

1. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing　100049, China;

2. Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China

Abstract: Visual-based localization determines the camera translation and orientation of an image observation with respect to a prebuilt 3D-based representation of the environment. It is an essential technology that empowers the intelligent interactions between computing facilities and the real world. Compared with alternative positioning systems beyond, the capability to estimate the accurate 6DOF camera pose, along with the flexibility and frugality in deployment, positions visual-based localization technology as a cornerstone of many applications, ranging from autonomous vehicles to augmented and mixed reality. As a long-standing problem in computer vision, visual localization has made exceeding progress over the past decades. A primary branch of prior arts relies on a preconstructed 3D map obtained by structure-from-motion techniques. Such 3D maps, a. k. a. SfM point clouds, store 3D points and per-point visual features. To estimate the camera pose, these methods typically establish correspondences between 2D keypoints detected in the query image and 3D points of the

收稿日期:2023-10-23;**修回日期**:2023-11-24;**预印本日期**:2023-11-31

*通信作者:申抒含　shshen@nlpr.ia.ac.cn

基金项目:国家自然科学基金项目(U22B2055,62273345);北京市自然科学基金项目(L223003)

Supported by: National Natural Science Foundation of China (U22B2055, 62273345); Beijing Natural Science Foundation, China(L223003)

SfM point cloud through descriptor matching. The 6DOF camera pose of the query image is then recovered from these 2D-3D matches by leveraging geometric principles introduced by photogrammetry. Despite delivering fairly sound and reliable performance, such a scheme often has to consume several gigabytes of storage for just a single scene, which would result in computationally expensive overhead and prohibitive memory footprint for large-scale applications and resource-intensive platforms. Furthermore, it suffers from other drawbacks, such as costly map maintenance and privacy vulnerability. The aforementioned issues pose a major bottleneck in real-world applications and have thus prompted researchers to shift their focus toward leaner solutions. Lightweight visual-based localization seeks to introduce improvements in scene representations and the associated localization methods, making the resulting framework computationally tractable and memory-efficient without incurring a notable performance expense. For the background, this literature review first introduces several flagship frameworks of the visual-based localization task as preliminaries. These frameworks can be broadly classified into three categories, including image-retrieval-based methods, structure-based methods, and hierarchical methods. 3D scene representations adopted in these conventional frameworks, such as reference image databases and SfM point clouds, generally exhibit a high degree of redundancy, which causes excessive memory usage and inefficiency in distinguishing scene features for descriptor matching. Next, this review provides a guided tour of recent advances that promote the brevity of the 3D scene representations and the efficiency of corresponding visual localization methods. From the perspective of scene representations, existing research efforts in lightweight visual localization can be classified into six categories. Within each category, this literature review analyzes its characteristics, application scenarios, and technical limitations while also surveying some of the representative works. First, several methods have been proposed to enhance memory efficiency by compressing the SfM point clouds. These methods reduce the size of SfM point clouds through the combination of techniques including feature quantization, keypoint subset sampling, and feature-free matching. Extreme compression rates, such as 1% and below, can be achieved with barely noticeable accuracy degradation. Employing line maps as scene representations has become a focus of research in the field of lightweight visual localization. In human-made scenes characterized by salient structural features, the substitution of line maps for point clouds offers two major merits: 1) the abundance and rich geometric properties of line segments make line maps a concise option for depicting the environment; 2) line features exhibit better robustness in weak-textured areas or under temporally varying lighting conditions. However, the lack of a unified line descriptor and the difficulty of establishing 2D-3D correspondences between 3D line segments and image observations remain as main challenges. In the field of autonomous driving, high-definition maps constructed from vectorized semantic features have unlocked a new wave of cost-effective and lightweight solutions to visual localization for self-driving vehicle. Recent trends involve the utilization of data-driven techniques to learn to localize. This end-to-end philosophy has given rise to two regression-based methods. Scene coordinate regression (SCR) methods eschew the explicit processes of feature extraction and matching. Instead, they establish a direct mapping between observations and scene coordinates through regression. While a grounding in geometry remains essential for camera pose estimation in SCR methods, pose regression methods employ deep neural networks to establish the mapping from image observations to camera poses without any explicit geometric reasoning. Absolute pose regression techniques are akin to image retrieval approaches with limited accuracy and generalization capability, while relative pose regression techniques typically serve as a postprocessing step following the coarse localization stage. Neural radiance fields and related volumetric-based approaches have emerged as a novel way for the neural implicit scene representation. While visual localization based solely on a learned volumetric-based implicit map is still in an exploratory phase, the progress made over the past year or two has already yielded an impressive performance in terms of the scene representation capability and precision of localization. Furthermore, this study quantitatively evaluates the performance of several representative lightweight visual localization methods on well-known indoor and outdoor datasets. Evaluation metrics, including offline mapping time usage, storage demand, and localization accuracy, are considered for making comparisons. Results reveal that SCR methods generally stand out among the existing work, boasting remarkably compact scene maps and high success rates of localization. Existing lightweight visual localization methods have dramatically pushed the performance boundary. However, challenges still remain in terms of scalability and robustness when enlarging the scene scale and taking considerable visual disparity between query and mapping images into consideration. Therefore, extensive efforts are still required to promote the compactness of scene represen-

tations and improving the robustness of localization methods. Finally, this review provides an outlook on developing trends in the hope of facilitating future research.

Key words: visual localization; camera pose estimation; 3D scene representation; lightweight map; feature matching; scene coordinate regression; pose regression

0 引　言

相机位姿用于描述相机某一时刻在参考坐标系下的位置坐标和空间朝向，其在三维空间中具有6自由度，可由位移向量和旋转矩阵唯一表示。视觉定位任务则是在预先构建的三维场景模型中，根据当前观测图像估计相机位姿的过程。

视觉定位是一种基于场景信息的主动定位技术，其相比基于通信信号传输的被动定位技术，如全球导航卫星系统，具备低成本、抗干扰和易集成等优势，因此受到了广泛关注与研究。计算机视觉与移动机器人领域将视觉定位作为一项具有重要价值的基础任务，其研究至今已取得显著的发展成果。目前的视觉定位技术能够达到分米甚至厘米级定位精度，已经成为实现计算设备与真实世界之间智能化交互的重要纽带，在信息化生产、数字娱乐等各个领域发挥着重要作用（Sattler等，2018；Sarlin等，2022；Fang等，2022）。例如，在增强现实领域，为了以像素级精度将虚拟内容放置于物理三维世界中，并长期保留以实现多用户共享，需要依靠视觉定位在任意时刻或视角下准确确定成像设备的6自由度位姿；在移动机器人领域，当相对相机位姿跟踪发生较大偏差或失败时，视觉定位将用于恢复机器人系统在场景中的绝对位置及姿态，这对于保障机器人与复杂场景交互的鲁棒性、提升建图信息的可复用性有着重要意义（刘盛 等，2020）；在自动驾驶领域，视觉定位算法通过对周围环境进行感知识别并与高精地图进行信息匹配，为车辆计算出精确的位置与朝向信息，从而辅助自动驾驶系统进行后续的运动规划以及行为决策。

视觉定位技术的核心难点可概括为两部分，一是对三维场景进行理解与表达，二是建立查询图像与场景表达的信息匹配。在早期，传统的视觉定位技术基于多视图几何原理取得了巨大成功，其中一类典型的设计方案是：在离线构图阶段，使用从运动恢复结构（structure-from-motion，SfM）或即时定位与建图（simultaneous localization and mapping，SLAM）方法从预先采集的场景数据库图像中构建显式的三维点云地图；在定位阶段，首先利用关键点匹配算法建立查询图像与场景点云地图的2D-3D特征点匹配，随后使用成熟的几何求解器从特征点的匹配关系中进行对相机位姿的鲁棒估计。随着人工智能研究的不断发展，深度学习技术能够提取可辨性更强的视觉特征并建立更为可靠的特征匹配关系，进一步增强了该类基于结构的视觉定位方法的准确率与鲁棒性。然而，由于场景点云模型的尺寸将随场景规模而增大，对于大尺度场景定位任务而言，场景模型将造成庞大的存储压力以及高昂的计算开销，使得当前视觉定位技术难以部署至计算资源有限的移动端（Yang等，2022）。具体而言，传统技术所依赖的场景模型存在大量冗余，这些冗余不仅无法提供有助于视觉定位的关键信息，同时还将极大降低存储和计算效率，并为信息匹配过程引入干扰（Mera-Trujillo等，2020）。对此，研究者们开始关注视觉定位技术的轻量化探究工作。轻量化旨在保证定位精度的同时，降低视觉定位系统对存储及算力资源的要求，以此提升算法在场景规模上的可扩展性，并有助于移动端部署。轻量化视觉定位技术的研究工作可以概括为两大部分：1）场景模型的高效表达与构建；2）基于相应场景表达的视觉定位算法设计。近年来，国内外出现了诸多对视觉定位技术轻量化发展起到推动性作用的研究工作，这些工作相较于传统视觉定位方法，对场景的表达进行了改进与创新。目前已有轻量化场景表达包括点云压缩地图（Zhou等，2022）、结构化程度更高的线地图和高精地图（Hofer等，2017；Liu等，2023b；Zhang等，2021）以及各种隐式模型（Kendall等，2015；Liu等，2023a；Shotton等，2013）。

一方面，现有视觉定位技术的相关综述缺乏对现有方法轻量化特性的关注和分析，同时较少囊括近几年的新进展；另一方面，目前尚缺乏从轻量化角度出发对现有视觉定位技术所做出的综合性总结，研究者们难以全面了解当前视觉定位技术的轻量化

发展现状，也无法对不同轻量化视觉定位技术的优劣进行客观评估。本文将填补视觉定位领域综述的这一空缺，对当前视觉定位技术轻量化发展的研究现状进行综述。首先，本文将按照场景模型的表达方式，对轻量化视觉定位技术的现有研究工作进行划分，并进行逐类梳理与阐述。随后，本文将通过实验数据评估对比各类代表性算法的轻量特性以及定位性能，并于最后简要总结现有研究工作的成果与不足，探讨未来有价值的研究方向。

1 传统视觉定位技术

在具体介绍轻量化视觉定位研究现状前，本节首先对视觉定位技术的传统框架进行梳理。针对目前主流的视觉定位系统，通常可将其划分为3大环节，包括场景模型构建、环境信息匹配与相机位姿估计（陈宗海 等，2021），本节将采用该划分模式对现有视觉定位方法进行归纳介绍与横向对比。

场景模型包含满足视觉定位任务所需的场景信息，通常构建自预采集的场景数据库图像。在传统视觉定位系统中，场景模型包含显式的场景特征以及其在统一参考系下的坐标两部分，其中场景特征一般提取自数据库图像的外观细节、几何结构或高层语义。具体而言，传统视觉定位系统所采用的类型主要包括带位姿标签的图像数据库或三维点云模型两种。环境信息匹配过程则基于特征一致性，建立相机观测图像与场景模型之间的数据关联。关联形式主要为显式2D-2D匹配对或3D-2D匹配对，进一步还可从稠密或稀疏、局部或全局的角度对匹配方式进行划分。相机位姿估计环节由信息匹配提供的空间对应关系解算相机位姿，从是否依赖初始位姿的角度可将求解过程分为迭代法与非迭代法。

场景模型的表达形式决定了后续信息匹配与位姿估计的具体方式，并对整个视觉定位算法的准确率、鲁棒性以及效率起着关键性作用。按照场景表达方式的不同，可将传统视觉定位系统的实现框架划分为基于二维图像的视觉定位、基于点云地图的视觉定位以及二者混合的层次化视觉定位3类。

1.1 基于二维图像的视觉定位

基于二维图像的视觉定位用带有绝对位姿标签的数据库图像作为场景表达，其中位姿标签可通过SfM或SLAM等方法计算获得，也可借助外部传感器直接测量，如全球定位系统（global positioning system，GPS）的定位信息。该类方法首先从图像数据库中为查询图像搜索视觉相似性较高的近邻图像，然后根据近邻图像及其绝对位姿标签估计查询图像的位姿。

搜索视觉近似图像的过程也称为图像检索，该过程不依赖多视图几何原理，核心在于选取鲁棒的视觉相似性度量法则。具体做法通常先对整幅图像的视觉信息进行全局编码得到图像描述子，随后根据描述子在高维特征空间中的分布，确定图像间的视觉相似度。局部描述子聚合向量（vector of locally aggregated descriptors，VLAD）方法（Jégou 等，2010）通过聚合稀疏局部特征的方式来编码图像的全局特征。具体而言，聚合过程首先通过K均值聚类的方式构建局部描述子码本，然后对描述子到其对应码词的距离进行求和。过去一般采用手工设计的方法对局部特征进行提取与描述，代表工作包括尺度不变特征变换（scale-invariant feature transform，SIFT）（Lowe，2004）以及实时性更优的二值描述子ORB（oriented FAST and rotated BRIEF）方法（Rublee 等，2011）等。稠密VLAD方法（Torii 等，2015）在图像的不同尺度上均匀且稠密地采样SIFT描述子，以此可以构建对光照和视角变换具有更强鲁棒性的VLAD全局描述子。此外，该方法采用新视图合成方法对原始数据库图像进行数据增强，降低由于数据库图像采集视角稀疏而造成的检索失败。

随着深度学习的发展，图像检索算法的性能得到进一步提升。一方面，出现了以SuperPoint（DeTone 等，2018）和R2D2（Revaud 等，2019）等为代表的局部特征提取与描述网络，所提取的描述子在多项上层任务中普遍表现出比手工设计算法更为强大的鲁棒性和准确度；另一方面，基于弱监督训练的NetVLAD（Arandjelovic 等，2016，2018）网络设计了一种可微分的VLAD层，以端到端的方式编码图像全局特征。

对于原始数据库图像，一般采取离线方式预先为每幅图像进行全局特征编码，得到图像全局特征数据库。在线的视觉定位过程则首先将为查询图像生成全局特征描述子，随后通过最近邻算法，在高维特征空间中搜索与查询图像视觉最为相似的数据库图像。为提高邻域搜索效率，通常采用K维树、哈希表等高级数据结构组织图像特征数据库。

两幅图像的视觉相似性越高意味着它们在空间中处于相邻位置的可能性越大，鉴于这一假设，一部分基于二维图像的视觉定位方法在完成最近邻图像特征检索后，直接以最近邻图像的位姿标签近似查询图像的位姿。该方式通常应用于对定位精度要求不高的任务当中，如检测闭环（Mur-Artal等，2015）或场景识别（Lowry等，2016）。另外，也可进一步建立最近邻图像与查询图像的局部匹配关系，计算相对位姿变换，再结合最近邻图像的绝对位姿标签，更加精确地解算出查询图像在场景当中的全局位姿，这一类针对图像检索的后处理方法将在第2.6.2节进一步介绍。

基于二维图像的视觉定位方法直接以带位姿标签的数据库图像作为场景表达，并从中构建图像全局特征数据库。这一形式的不足之处在于缺乏直观的场景结构信息，这导致信息匹配过程只能依赖由全局描述子编码的底层视觉特征，因而该方法对相机视角、场景光照以及结构变化缺乏稳定性。另外，图像中包含的大量外观信息对于视觉定位任务而言存在较大冗余，大规模场景的数据库图像将为运行设备带来巨大的存储压力。

1.2 基于点云地图的视觉定位

利用SfM技术以及成熟的开源工具，如COLMAP（Schonberger和Frahm，2016），能够从多视角二维图像集合中重建稀疏的点云模型。SfM点云模型中的每一个稀疏关键点都包含其在场景参考坐标系下的坐标以及所对应的图像局部特征集合。相比于图像数据库，SfM点云不仅减少了场景模型的外观冗余，还包含了精确的场景结构信息，因此成为视觉定位任务中最为常用的场景表达形式之一。

SfM点云模型中的每个三维点均对应一个由图像局部描述子构成的特征集合，用于建立场景3D点与查询图像2D关键点的视觉特征匹配。相比于编码全局特征并基于整体视觉相似度构建全局匹配的图像检索方法，基于SfM点云模型的方法在本质上则是根据局部视觉相似度，建立图像间的特征点匹配。建立特征点匹配的过程同样采用最近邻搜索的方式，目前有一系列工作针对匹配搜索的效率和可靠性提出改进方案。Li等人（2010）将共视可见信息（co-visibility）纳入考虑，加快了2D-3D点的匹配效率。具体地，该方法以3D点在数据库图像中出现的频率作为匹配的优先级，从优先级高的3D点开始搜索查询图像中的匹配2D点，并在建立足够数量的匹配对后停止搜索。在大尺度场景模型下，从查询图像的2D点出发搜索与之相匹配3D点的方式效率较低，但一种基于视觉词汇库的搜索策略（Sattler等，2011）能够加快这一方式的匹配效率，并且对场景动态变化更为鲁棒。主动搜索（active search）机制（Sattler等，2017）将2D到3D、3D到2D双向的匹配策略相结合，首先从2D到3D搜索获得一组特征点匹配，然后从距离匹配点最近的3D点开始主动式地进行从3D到2D的反向匹配搜索。Liu等人（2017）提出可以在匹配过程中参考全局信息，即不再将2D-3D匹配对的搜索视为彼此孤立的过程，而是使用马尔科夫图，根据匹配对之间的视觉相似性以及全局兼容性进行推理决策。

除了上述手工设计的近邻匹配方法，基于注意力机制的图神经网络SuperGlue（Sarlin等，2020）将深度学习方法引入图像稀疏特征的匹配任务，其作为可端到端训练的中后端算法能够自由地与各种经典或深度学习特征相结合。相比传统匹配技术，SuperGlue在光照、视角剧烈变化的情况下所建立的特征点匹配具有更高的内点率，在各类上层任务中表现更优。然而，图神经网络中稀疏的注意力矩阵将造成大量冗余计算，ClusterGNN特征匹配网络（Shi等，2022）对此引入层级聚类模块，在前向过程中动态构建由粗到细的局部子图，使信息的聚合与传递更加高效。该方法在保持匹配性能表现的同时，将推理用时和内存占用缩小近60%。LightGlue方法（Lindenberger等，2023）根据匹配难易度，在推理过程中加入早停和外点过滤环节，使网络结构具备自适应能力，从而提高推理效率。

上述环节所建立的3D-2D特征点匹配本质上表示某个三维空间点与其投影点之间的对应关系。在已知相机内参矩阵的前提下，由n对空间点与投影点对应关系解算相机位姿的问题称为n点透视（perspective-n-point，PnP）问题。PnP问题的求解器大多通用且相对固定，常用的求解方法包括直接线性变换、P3P以及EPnP（Lepetit等，2009）等。在没有噪声的情况下，PnP几何求解器理论上能够准确解算相机位姿。然而，在实际场景中，由于描述子通常具有歧义性问题，输入求解器的匹配对中将不可避免地出现错误匹配，这些错误匹配通常也称为离群点。在对相机位姿进行解算前，有必要对错误的离

群点加以滤除，利用剩余的正确匹配对估计相机位姿。这种鲁棒估计思想能够从包含大量离群点的数据集中求解准确的相机位姿。

随机采样一致性（random sample consensus，RANSAC）是鲁棒估计环节的重要方法，其将不断从数据集中随机采样用于求解相机位姿的最小配置样本，计算对应的位姿实例，并统计在一定误差阈值内满足实例的有效点数目（Fischler 和 Bolles，1981）。随后，该算法以数目衡量每个位姿实例的优劣得分，并从中选取最优实例。抢占式 RANSAC 算法（Nistér，2003）在生成多个位姿实例后，随机选取数据集的子集来统计每个实例的得分，再按照得分对其进行排序，丢弃靠后的若干个，并优化剩余实例。该算法将不断重复上述排序与选取过程，直到仅剩唯一的最佳位姿实例，将其输出作为最终估计结果。NG-RANSAC 算法（Brachmann 和 Rother，2019b）用神经网络引导最小配置样本的采样过程，目标函数为实例集合的期望损失，通过自监督的方式学习各个数据点的采样概率。在更为复杂或更大尺度的场景中，错误匹配的占比甚至可高达 90%。对此，一些研究者从引入辅助信息或额外约束的角度出发进行尝试，提出了利用共视可见性（Li 等，2012）、考虑匹配点对的语义一致性（Shi 等，2019；潘小鹃 等，2023）或引入几何约束（Camposeco 等，2017）等方法强化离群点的滤除机制。

基于点云地图的视觉定位方法以存储视觉特征的稀疏 SfM 点云为场景模型，其离线构图过程主要利用三角化的方式来对特征点的空间坐标进行恢复。在线视觉定位过程则建立查询图像与 SfM 点云间的 2D-3D 匹配，将匹配对输入由 RANSAC 与 P*n*P 求解器组成的鲁棒估计环节，得到最终的相机位姿估计值。相比基于二维图像的视觉定位系统，该方法由于建立了 2D-3D 坐标匹配关系，可以借助几何求解器得到更加精确的定位结果。然而在另一方面，构建三维模型以及建立局部特征匹配的难度更高、耗时更长，并且 SfM 点云模型的存储效率较低，通常难以支撑大规模场景应用。

1.3　层次化视觉定位

尽管基于点云地图的诸多方法从多个方面对 2D-3D 匹配的效率及精度做出了提升，但由于匹配过程仍然是在整个点云模型中进行的，与查询图像实际位置相差甚远的 3D 点依然会参与匹配的判别。这种方式将始终限制匹配搜索的效率，而且难以进一步缓解由于重复结构或纹理而导致的特征歧义性。为了缩小 2D-3D 匹配的搜索空间，层次化视觉定位方法将基于二维图像的方法与基于点云地图的方法相结合，使用图像全局特征检索出候选近邻图像，然后在近邻图像所对应的局部点云模型中建立匹配关系（Middelberg 等，2014）。这种由粗粒度到细粒度的两阶段层次化匹配策略能够有效增强视觉定位在场景尺度方面的扩展能力。

经典的层次化定位策略主要在场景模型和环境信息匹配两个阶段对前述两种框架进行组合：同时使用图像全局特征数据库与 SfM 点云两种场景模型，在信息匹配环节按照图像检索、抽取局部场景模型以及局部特征点匹配 3 个步骤进行（Sarlin 等，2018）。如图 1 所示，第 1 步基于图像检索完成全局匹配，该步将在图像全局特征空间中搜索与查询图像整体视觉相似度最高的 k 幅图像；第 2 步根据该 k 幅图像之间共同可见的三维结构对图像进行聚类，并为每组图像抽取对应的局部场景点云模型，随后局部特征的匹配过程将分别在每一块局部点云模型中进行；第 3 步对 6 自由度相机位姿进行鲁棒估计，保留有效点数量最多的位姿实例作为最终估计结果。

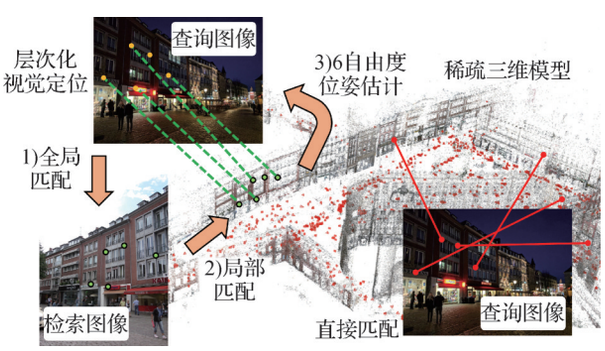


图 1　层次化视觉定位示意（Sarlin 等，2019）

Fig. 1　Illustration of hierarchical localization (Sarlin et al.，2019)

层次化定位策略同时要求提取图像的全局和局部特征，HLoc（Sarlin 等，2019）对此提出了一种基于 MobileNet（Sandler 等，2018）的卷积神经网络（convolutional neural network，CNN）结构，该网络结构集成了全局特征和局部稀疏特征的提取与编码过程，并联合知识蒸馏技术进行多任务学习。该方法不仅显著提升了大规模场景长时视觉定位的精度以及鲁棒性，还具备轻量的网络推理能力。

相比单阶段直接匹配的点云地图视觉定位方法，层次化策略的优势在于其首先进行图像检索并抽取局部点云模型，使后续2D-3D匹配的搜索效率以及匹配准确度得到大幅提升。此外，通过组合各个环节对应领域下的领先方法，层次化视觉定位框架的定位性能能够得到进一步提升。然而，层次化视觉定位过程由于同时依赖图像全局特征数据库以及SfM点云两种场景表达，因此场景地图的内存占用体积相比于前述两种方法更加庞大，难以满足大规模场景下的轻量化视觉定位需求。

2 轻量化视觉定位技术

轻量化对视觉定位方法提出了更高要求，所面临的挑战可概括为两大方面：如何构建轻量的场景表达以及如何基于相应场景表达进行高效、鲁棒的视觉定位。本文对现有视觉定位工作的划分模式如图2所示，其中左侧为上一节所介绍的传统视觉定位技术框架。本节则将按右侧所示的分类模式，从轻量化场景模型类别的角度，对现有轻量化视觉定位工作进行梳理。现有轻量化视觉定位技术所采用的场景模型可分为显式表达和隐式表达两种类型：显式场景表达由具备几何外观的结构特征及其在场景参考坐标系下的坐标构成，本节主要介绍压缩点云地图、线地图以及高精地图3类，其轻量化程度随场景特征的结构化程度而逐级提高。隐式场景表达则利用各类机器学习算法建立关于场景信息的不同映射关系，以数据库图像及其绝对位姿标签为训练数据，借助模型的拟合能力将场景特征及其坐标信息压缩编码至模型参数当中。

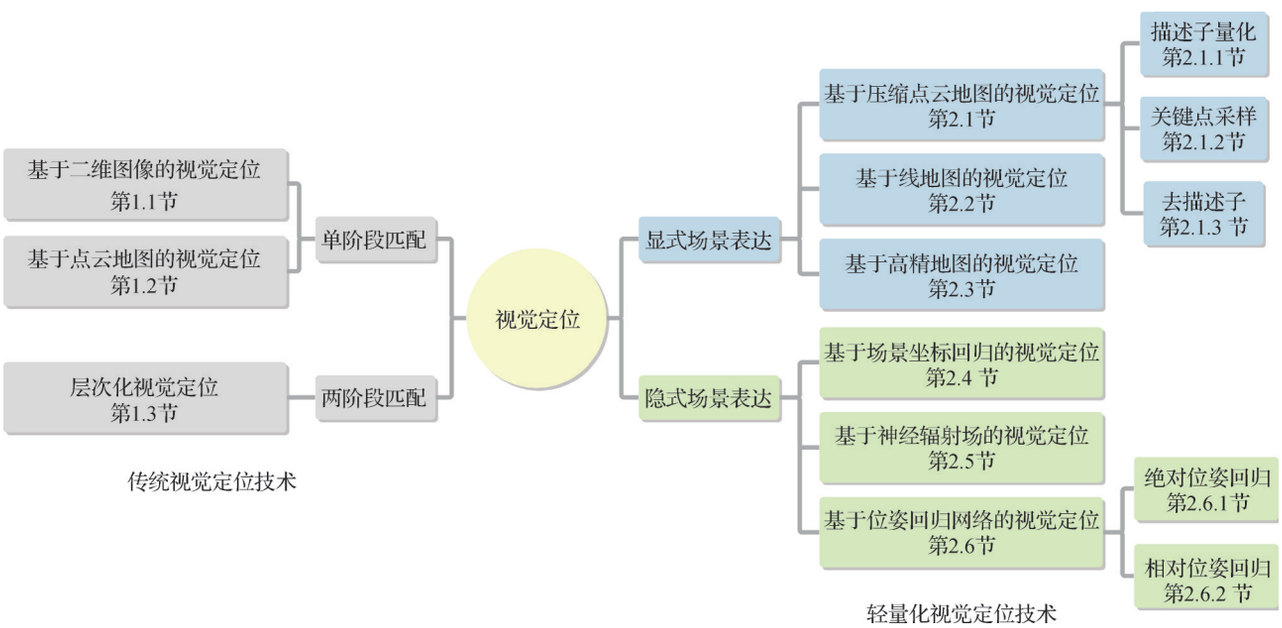


图2 本文对现有视觉定位工作的分类模式

Fig. 2 A taxonomy of existing works on visual-based localization

2.1 基于压缩点云地图的视觉定位

层次化视觉定位方法在一定程度上解决了由于搜索空间过大而导致的搜索效率低下和特征描述子的歧义性问题，但其仍然需要构建并存储庞大的点云模型。在实际应用中，大规模场景的SfM点云模型可能会包含上万亿个特征描述子，所消耗的存储空间甚至将超过原始图像数据库(Sattler等，2015)。对于视觉定位任务而言，SfM点云依然是一种存在冗余表达的场景模型。为了减少点云地图中的冗余信息，现有部分工作从3种角度提出了针对点云地图的压缩策略以及基于压缩模型的定位算法。

2.1.1 描述子量化

视觉词袋模型通过K-Means聚类算法对特征空间进行层级划分，所有特征描述子被组织为K维树的数据结构形式，叶节点所对应的特征描述子聚类被量化为一个视觉单词(word)。使用视觉词袋模型能够有效提升2D-3D匹配效率，但同时也引入了不利于匹配准确性的量化噪声。Sattler等人(2015)采

用大小为16 M的精细词袋模型，能够较好地取得训练和搜索效率的折中，而且量化噪声的占比相对合理，该工作还充分利用了共视关系图，在定位精度上超越当时同类方法领先水平的同时，场景模型的存储占用仅为原始模型的25%。Camposeco等人(2019)提出层次化的压缩策略，为判别性更强的关键点保留完整的特征描述子，而仅对剩余的普通点进行量化压缩。在相机位姿计算环节，具有高置信度的关键点匹配被RANSAC用于采样并生成位姿实例，而普通点匹配则用于对位姿实例进行评估。该方法的地图压缩率可达2%，同时定位精度相比同期地图压缩方法的领先水平提升15%以上。

为了降低量化噪声对匹配准确性的影响，Cheng等人(2019)从剔除错误匹配的角度考虑，提出了一种级联的离群点过滤流程，依次从特征、模型共视信息以及几何约束层面逐步排除错误匹配，并对匹配质量进行评估，在最后位姿解算过程中为质量更高的匹配赋予更高权重。该方法最终只需为每个视觉单词存储一个紧凑的二元特征向量，场景模型的存储占用不及Active Search方法(Sattler等，2017)的20%，但能够取得更高的定位精度。

随着深度学习技术的发展，过去基于视觉词袋模型的工作逐渐被网络结构所取代。Yang等人(2022)设计了一种轻型自编码器网络，采用可微分的软量化层结构(Gong等，2019)，使特征向量的降维与量化过程能以自监督形式实现端到端学习。

2.1.2 关键点采样

另一种点云地图压缩策略通过下采样方式选取点云地图中的关键点子集，该方式通过设计合理的下采样过程，还能够有效地剔除具有歧义性的特征点，提升匹配准确率。衡量下采样策略的合理性通常考虑采样覆盖率和关键点特殊性两个方面：覆盖率要求下采样后的点云尽可能均匀地覆盖整个场景，因为均匀分布的2D-3D匹配往往能够更为稳定地求解相机位姿(Irschara等，2009)；关键点的特殊性是指尽可能保留特征显著的3D点，如可以在多个视角下被观测到的点或者属于特殊场景结构的点，这些点往往为视觉定位任务提供着重要的场景信息(Donoser和Schmalstieg，2014)。

Li等人(2010)首先假设查询图像和数据库图像采样自同种概率分布，进而以点云地图中每个点在数据库图像中出现的次数作为其特殊性的评价指标，特殊性越高的点被观测到的可能性越高，因而与查询图像成功建立匹配的可能性也越高。另一方面，点云下采样的范围应当包含整个场景，而不能仅仅集中于场景的某个局部区间。基于以上两种假设，该方法将关键点采样的过程等效为K-覆盖问题，利用贪心算法增量式地选取出可以包含每幅数据库图像至少K次的最小3D点云子集。实验结果表明，该策略可获得下采样率约为8%的点云子集，并且其相比于未进行下采样压缩的完整点云地图，在视觉定位任务中具有更好表现。Cao和Snavely(2014)提出依概率的增量式采样方法，不再仅对每个三维点与数据库图像之间的关系做简单的二元划分，而是建立基于软划分的可视关系矩阵，取得了更高的定位准确率。Camposeco等人(2019)则在对关键点进行每一步采样决策时，比较该点与已经采样得到的点之间的特征差异，依次衡量点的特殊性，并在原有的增益函数基础上设计额外的特殊性加权策略。

后续有工作将点云的下采样过程建模为混合整数二次规划问题，在目标函数中额外考虑了3D点之间的二元关系，对空间间隔更近的3D点赋予更大的惩罚权重，以鼓励下采样保留的3D点之间尽可能分散(Park等，2013)。SceneSqueezer方法(Yang等，2022)通过引入注意力机制使网络模型学习点云地图中的一元特征显著性与二元分布权重系数，然后通过一种可微分的混合整数二次规划求解算法，将场景点云地图下采样变为一个完全端到端的过程。该方法在大规模场景公开数据集中压缩得到的点云体积仅不到Active Search方法的0.5%，但能够取得更加鲁棒和精确的定位结果。

2.1.3 去描述子

三维点的特征描述子在环境信息匹配环节发挥关键作用，然而也引入了两个关键性不足：1)逐点存储的高维特征描述子是点云地图占用庞大内存空间的主要原因，导致定位任务无法完全在端侧运行，而点云地图的特征描述子存储在服务器又将面临隐私泄露风险(Speciale等，2019)。2)对点云地图的描述子进行更新与维护的成本较高，如果需要更换鲁棒性更强的局部视觉特征，需对整个场景进行重新构图。为了避免上述问题，去描述子的点云地图压缩策略提出不依赖视觉特征，实现2D点与3D点之间的跨模态匹配。

去描述子的压缩点云地图视觉定位方法分为两类。一类方法遵循传统非迭代法的思路，首先进行跨模态的2D-3D点匹配，然后借助P*n*P几何求解器估计相机位姿。Donoser和Schmalstieg（2014）将跨模态2D-3D点匹配任务建模为分类问题，分类器采用随机厥模型（Ozuysal等，2010），在目标场景的点云地图中完成训练后，可建立2D关键点特征描述子与对应的空间3D点索引的映射关系。该模型仅需维护一定数量的稀疏概率查找表，存储占用较低。BP*n*PNet（Campbell等，2020a）在给定不带特征描述子和匹配关系2D/3D点集合的条件下，采用一种12层ResNet卷积神经网络提取2D、3D点特征，随后利用求解双边际最优传输问题的Sinkhorn算法，计算2D、3D点特征点两两之间的匹配权重，进而由可微分的加权Blind-P*n*P算法解算相机位姿。GoMatch（Zhou等，2022）在该项工作的基础上对网络的特征编码器做出改进。为了增强每个模态中上下文特征的融合，作者引入了基于自注意力机制的图神经网络，通过与空间中固定数量的最近邻特征进行交换信息来细化每个关键点的特征。作者接下来引入交叉注意力机制，对2D与3D跨模态特征进行融合，进而能够提升网络学习跨模态匹配的效果。另外，作者还采用分类器预测每对匹配关系的置信度，对置信度较低的匹配进行剔除。

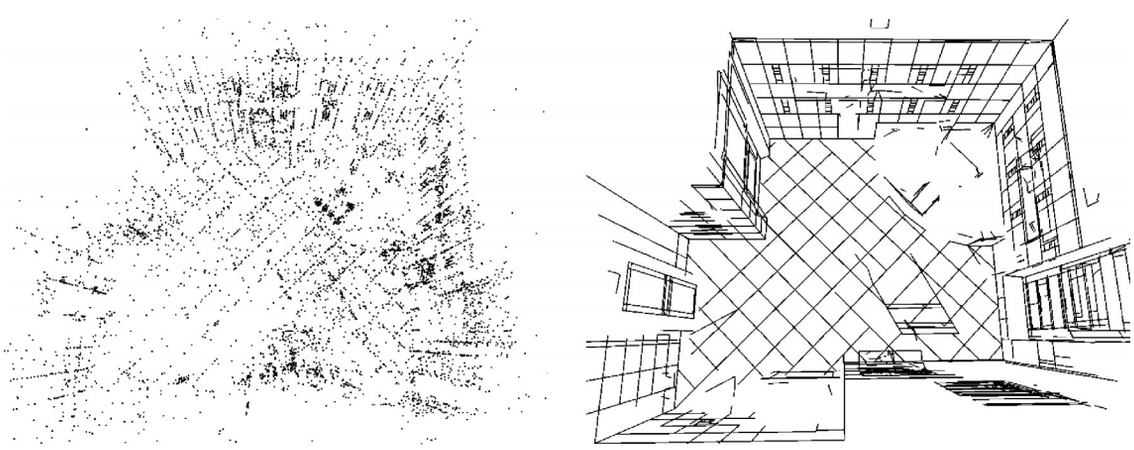
另一类方法采用迭代思路，对2D-3D特征点的匹配和相机位姿进行联合优化。SoftPosit（David等，2004）在给定相机位姿初始值的情况下，在相机位姿和特征匹配之间进行交替式的迭代优化。BlindP*n*P方法（Moreno-Noguer等，2008）将相机位姿先验用混合高斯分布进行建模，同时为每一个高斯成分初始化一个卡尔曼滤波器，随后建立2D-3D匹配并更新相机位姿估计值。这类方法由于需要提供初始位姿，应用场景有限。目前也存在方法基于全局优化算法，尽管无需提供初始位姿，但求解效率有限（Brown等，2015；Campbell等，2019，2020b）。

2.2 基于线地图的视觉定位

在直线特征显著的人造场景，如城市或室内场景中，用线地图替代点云地图作为场景模型主要存在两大优势。1）空间直线段包含比空间点更加丰富的几何信息。如图3所示，以直线段为几何基元构建的线地图能够更为简洁、直观地表征场景结构，因而在结构化程度较高的场景中，线地图相比点云地图具备更加突出的表达能力和轻量化优势。2）图像中的直线特征通常对应着场景中真实的物理结构（Lindeberg，1998），因此虽然人造场景中常常存在着大量弱纹理区域，但这些区域依然能够在图像上呈现丰富的直线段特征，并且直线特征在光照变化环境下通常比关键点特征更为鲁棒。因此，线地图的轻量化及鲁棒性优势使其成为结构化场景的视觉定位任务中一种具有发展潜力的场景表达形式。基于线地图视觉定位的具体方式与线地图的形式及构建过程相关，本节按照线地图构建方式的不同，对现有方法进行分类介绍。

2.2.1 SfM线地图

一类工作利用直线段SfM方法（Micusík和Wildenauer，2017；Wei等，2022；Liu等，2023b）构建



(a) 点云地图　　(b) 线地图

图3 由从运动恢复结构构建显式场景地图（Liu等，2023b）

Fig. 3 Explicit scene mapping via SfM (Liu et al., 2023b) ((a) point cloud map; (b) line map)

场景线地图。SfM线地图的每一条空间线段都对应一组来自多视角数据库图像的直线段特征，用于进行地图3D直线段与查询图像2D直线段的匹配。从3D和2D线的对应关系估计相机姿态可直接使用现有的n线透视（perspective-n-line，PnL）几何求解器（Xu等，2017）。然而，由于当前直线段特征提取与描述、直线段匹配方法研究的相对滞后，大部分工作与基于点云地图的视觉定位框架存在明显差异。Micusík和Wildenauer（2015）提出直接在相机位姿参数空间中进行采样，在每个采样位姿下为线地图生成对应的投影视图，逐一计算与查询图像直线段之间的倒角距离，最终以距离最小的采样位姿作为定位结果。该方法不借助图像特征描述子，也不建立直接的信息关联，但是采样过程消耗大量计算资源且定位精度有限。不同于仅考虑线匹配，一部分方法在环境信息关联环节同时建立点匹配，通过点线匹配相结合的方式求解相机位姿（Zhou等，2019a；Yoon和Kim，2021）。Gao等人（2022）在假设已知相机位姿先验的情况下，同时建立点匹配与线匹配，采用一种更加有效的点—线联合代价函数来优化求解最终的高精度位姿。具体而言，该方法提出了一种新的参数形式用于对直线段进行高效表示，并且设计了一种基于堆叠沙漏结构的卷积神经网络作为线提取器。该方法以一种由粗到细的搜索策略建立线匹配，通过引入对极几何约束并参考重投影误差的大小来选取高质量的线匹配对。

2.2.2 激光线地图

一部分工作采用从激光雷达点云地图中抽取线结构的方式构建场景线地图。与SfM线地图不同，该方式所获得的线地图不包含视觉特征，因此建立2D-3D信息关联的过程属于图像特征与激光点云之间的跨模态匹配。Yu等人（2020）提出先利用粗糙的位姿先验将3D直线段投影至图像平面，建立初步的2D-3D直线段匹配对，后续则通过迭代法最小化匹配对之间的投影误差并过滤外点，提高线匹配质量。LDL方法（Kim等，2023）采取了与Micusík和Wildenauer（2015）类似的相机位姿参数空间采样策略，但与后者提出的以倒角距离作为匹配衡量标准不同，作者设计了一种直线距离函数用于比较查询图像与投影视图之间的相似度。

2.2.3 CAD线地图

Goto等人（2018）直接从建筑图纸或CAD模型中抽取线地图，并以此提出一种基于室内线地图的球面相机定位方法。该方法不直接建立2D-3D匹配，而是设计了一种球面霍夫表示来对查询图像或线地图某一视角下的线段分布进行描述与匹配。具体而言，该过程分为两个阶段：1）从查询图像所提取的全部线段中选取3个主方向，并基于曼哈顿世界假设下将这3个主方向绑定为地图3个主轴方向，进而由此计算相机朝向；2）用查询图像的球面霍夫描述子与线地图进行鲁棒匹配，最终可估计相机的空间坐标。

线地图作为结构化程度更高的一种场景模型表达，适用于城市或室内等大规模人造场景的视觉定位任务当中。然而，由于直线段特征提取与匹配方法尚未成熟，当前大部分基于线地图的视觉定位方法仍然采用点线联合的方式，一部分工作还需借助初始位姿先验，通过迭代的方式逐步提高特征匹配的质量。因此，基于线地图的视觉定位方法在未来还存在较大的优化与发展空间。

2.3 基于高精地图的视觉定位

高精地图（high-definition map，HD map）是一种面向城市道路场景的专用电子地图，在自动驾驶、城市规划和安防等多个领域发挥着重要作用。基于高精地图的视觉定位方法是辅助车辆建立环境感知的重要环节：一旦确定无人系统在地图上所在位置，高精地图便可提供丰富的场景信息以进一步增强无人系统对环境的感知。自动驾驶领域所使用的高精地图一般包含3层结构，分别为定位层、车道层和路网层。定位层的地图特征为静态的矢量化语义地标，包括车道线、电线杆和交通标志等常见的道路元素，主要用于辅助车辆完成车道级的高精定位。相比于点或直线段等几何基元，高精地图定位层中的矢量化地标是一种更为结构化的地图特征，包含丰富的语义信息，因此相比点云地图或线地图，高精地图具有更加突出的鲁棒性与轻量化优势。目前有许多研究者致力于探究基于高精地图的视觉定位方法，提出了一系列自动驾驶车辆的轻量定位方案。

基于高精地图的视觉定位本质上需要解决的问题主要包含两部分：1）如何建立地图矢量特征与车载相机观测图像之间的跨模态匹配；2）如何利用匹配信息求解相机/车辆位姿。其中跨模态匹配过程通常需进行模态统一。根据模态统一方式的不同，可将现有方法划分为以下两类：

一类方法假设在相机俯仰角以及相对地面高度

已知的情况下，通过逆透视变换方法将相机观测图像变换至鸟瞰视图，以该视图作为中间模态完成图像与高精地图的特征关联。Ranganathan等人(2013)使用由特定路面标示组成的矢量高精地图，首先将前向单目视图逆透视变换至俯瞰视图，再通过模板匹配的算法检测该视图中的特殊路面标示，最后从检测结果中提取角点并与地图建立全局坐标匹配以直接求解相机绝对位姿。Wu和Ranganathan(2013)在上一项工作的基础上提出了一种基于立体视觉暗影增强的预处理方法，用于提升路面标示检测算法在光照变化情况下的鲁棒性。AVP-Loc视觉定位系统(Zhang等，2021)由4台环视的鱼眼相机组成，4幅观测图像在鸟瞰视图中完成融合后输入CNN网络进行语义分割。由于不同的语义类别具有不同的几何形状，作者分别针对各种类别设计了与地图矢量特征建立匹配的策略。该系统最终能够在多层地下停车场环境中实现厘米级定位精度，所使用的高精地图大小不到0.5 M。上述方法在进行逆透视变换时，需要假设道路相机对于地面的高度与俯仰角固定，否则逆透视变换的结果将存在较大畸变。为了摆脱这一限制，Poggenhans等人(2018)采用双目相机方案，利用视差生成稠密点云，将其投影至二维地面网格获得鸟瞰视图，接着使用基于ResNet的卷积神经网络检测鸟瞰图中的地图语义特征，最后采用霍夫投票的方式，从预先采样的一组位姿假说中确定与地图匹配程度最高的假说作为位姿的观测值。HDMI-Loc方法(Jeong等，2020)则首先用DeepLabv3+网络(Chen等，2018)对双目图像进行语义分割，随后生成带语义标签的稠密点云并投影得到一个8位鸟瞰特征图，每一位表示当前像素位置上各个语义特征是否存在；矢量地图同样经过一个粒子滤波器被转换为一个由8位特征图组成的数据库，最终通过搜索建立鸟瞰特征图与数据库之间的匹配来确定当前车辆位姿。

另一类方法则依靠相机位姿先验，将地图上的矢量特征投影至图像平面，随后在图像平面完成特征匹配。LaneLoc视觉定位系统(Schreiber等，2013)为搭载双目相机和惯导的车辆实现了车道级精度的实时定位。其手动构建了一个包含路面标示和路缘特征的高精语义地图，在视觉定位过程中，首先根据里程计输出的相机位姿估计值将地图特征投影至图像，投影结果在辅助图像特征提取的同时被用于与图像特征建立匹配，接着又将图像特征投影至地图坐标系，在地图上计算匹配特征之间的距离残差，最后使用卡尔曼滤波器对车辆当前位姿进行最优估计。Lu等人(2017)将车道线地图转换为一组离散且带有语义标签的3D点，同样利用里程计输出的相机位姿先验将这组点投影至相机图像。接着，通过最小化投影点集合与图像检测结果之间的倒角距离来获得更加准确的相机位姿，该方法在城市道路场景中定位的总体误差在1 m以内。廖文龙等人(2021)采用像平面距离及法向量夹角定义了一种新的重投影误差度量方式，该误差能够与相机位姿保持单调，从而避免优化过程陷入局部极小。TM3Loc方法(Wen等，2022)借鉴传统图像对齐算法，使用语义倒角匹配算法在图像平面建立与地图语义特征的匹配。其所提出的匹配损失对6DOF位姿的导数是可解析的，可较好地保证优化求解的效率。另外，为了解决矢量高精地图特征稀疏的问题，作者还设计了一种基于时间滑窗的优化策略来实现历史帧特征点匹配和地图语义匹配的紧耦合。

2.4 基于场景坐标回归的视觉定位

与基于点云地图的视觉定位方法一致，场景坐标回归方法通过建立2D-3D点坐标匹配的方式直接解算相机位姿。然而，场景坐标回归方法不再借助场景中的显式特征，而是以回归的方式端到端地建立数据库图像像素与场景坐标的映射关系，回归模型的训练过程可视为对场景信息的隐式编码。显式地图定位方法存在特征稀疏性问题，稀疏匹配的数量往往直接影响位姿估计的精度，而场景坐标回归方法则往往可根据需要生成稀疏或稠密的匹配。该方法相比点云地图存储空间占用量更小，并且与对应用场景有着较大限制的线地图或高精地图不同，场景坐标回归是一种更加通用的视觉定位方法。

早期的场景坐标回归方法主要使用彩色与深度(RGB-depth，RGB-D)图像，即数据库图像与观测图像均包含稠密的深度信息。此时的回归过程将对3D-3D坐标匹配关系进行预测，而获得3D坐标对应关系后，再利用Kabsch算法(Kabsch，1976)解算相机位姿。Shotton等人(2013)最早提出用随机回归森林模型预测RGB-D图像中各个点所对应的场景坐标，并使用鲁棒估计方法从稠密的3D-3D坐标匹配关系中解算相机位姿。具体而言，随机森林可视为一种聚类方法，其根据各个像素的深度及外观特

征，将对应空间中邻近区域的图像像素划分至同一叶节点下。为了进一步增强模型的判别能力，Guzman-Rivera 等人（2014）利用 Boosting 集成学习技术，训练一组坐标回归森林并分别估计位姿实例。针对每一个位姿实例，该方法将计算观测数据关于场景表面截断符号距离函数的误差值，随后从中选取最优结果。上述方法在训练回归树的过程中潜在地认为坐标匹配样本服从各向同性的单一分布，Valentin 等人（2015）指出该假设将限制模型回归能力，随后提出用异方差高斯混合模型对叶节点输出的不确定性进行建模，同时还引入一种连续的局部优化策略，使重定位精度相比过去方法提升约 40%。

后续出现了基于纯彩色图像的场景坐标回归工作。Brachmann 等人（2016）借鉴自动上下文（Auto-context）思想，使用级联的分类—回归森林模型，对场景中类别标签的离散分布以及空间坐标的条件概率分布进行联合估计，同时还根据物体空间坐标估计的不确定性，提出通过最大化有效点对数似然的方式，对鲁棒估计方法得到的位姿做进一步优化。该方法对于不含任何深度信息的纯图像输入，能够达到与同期领先工作相近的性能表现。在物体位姿估计任务中，Krull 等人（2015）提出用玻尔兹曼分布表示观测数据中物体位姿的后验概率，并用卷积神经网络拟合该概率模型中的能量函数。该网络基于合成并分析（analysis-by-synthesis）方法，对查询图像和基于位姿假设合成得到的图像进行对比，输出反映两者差异的能量值，以此计算该候选位姿后验概率的大小。该方法本质上以最大化样本数据集的概率似然为目标训练了一个评价网络，用于衡量候选位姿的优劣。DSAC 方法（Brachmann 等，2017）将这一工作引入视觉定位任务当中，如图 4 所示，该方法使用具有 13 层的 VGGNet（Visual Geometry Group network）卷积神经网络 $\boldsymbol{V}$ 对鲁棒估计环节中采样得到的相机位姿实例 $\boldsymbol{h}$ 进行评分，并另外提出用基于概率采样的方式代替 RANSAC 中不可微的假说选取环节。该方法以相机位姿估计误差的期望作为训练的损失函数，使场景坐标回归网络 $\boldsymbol{W}$ 和评分网络 $\boldsymbol{V}$ 能够通过端到端的方式进行联合训练。

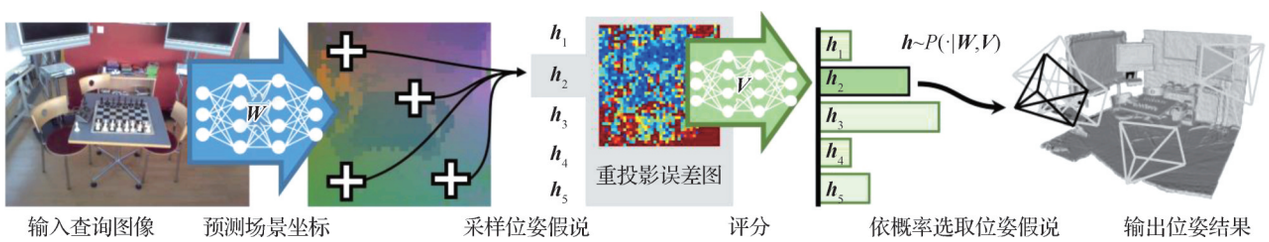


图 4　可微分视觉定位流程（Brachmann 等，2017）

Fig. 4　Differentiable camera localization pipeline (Brachmann et al., 2017)

由于评分卷积神经网络对于新视角图像的泛化能力较弱，并且存在容易使网络过拟合以及训练不稳定的问题，DSAC++ 方法（Brachmann 和 Rother，2018）将该部分替换为基于 Sigmoid 函数的位姿实例评分模块。该方法设计了一种三阶段的分步式训练方案，在不借助场景三维显式模型的条件下，仅用不带深度信息的色彩图像及其对应的位姿标签作为训练数据，在各个数据集上取得领先。为了简化回归网络的初始化过程，Li 等人（2019）提出一种基于角度的重投影误差作为网络训练的损失函数，提升网络训练的稳定性与收敛速度。DSAC*（Brachmann 和 Rother，2022）在 DSAC++ 工作的基础上对训练过程以及梯度计算等环节加以优化，同时将网络结构替换为参数量更少、学习能力更强的 ResNet，使内存占用下降 75%、推理速度提升 40%，并且在室内数据集上的定位精度提升近 30%。另外，相比 Active Search 方法，DSAC* 网络模型的体积比点云地图缩小近一个数量级，但可实现更高的定位精度。

上述工作未考虑目标场景的实际规模，而是选用结构、参数量固定的网络对场景进行隐式编码，因此对于场景规模的可扩展性有限。Brachmann 和 Rother（2019b）指出，仅通过增加网络参数量的方式无法有效提升方法在大尺度场景下的回归能力，于是采取训练多个坐标回归网络的方式，并引入混合专家模型（mixture of experts，MoE），将每个网络作为某一局部区域的专家模型，模型训练过程中将根据门网络（gating network）的输出值，在不同的专家模型下基于多项式分布采样得到不同数量的位姿假

说。该方法提升了场景坐标回归网络在大规模室外场景中的定位性能，并且取得了位姿估计精度与推理效率的折中。Li等人(2020)则设计了一种层级式的场景坐标回归网络HSCNet。如图5所示，该方法用分类网络为每个像素预测由粗粒度到细粒度的区域标签，回归网络则根据最后一个分类网络层的输出预测像素所对应的场景坐标。在该结构中，上一层网络的预测结果均经过控制层(conditioning layer)，由线性特征调制(feature-wise linear modulation，FiLM)方法(Perez等，2018)传递给下一层。该网络模型可以鲁棒地扩展到大规模场景当中。VS-Net方法(Huang等，2021)提出一种由两个网络分支所构成的解码器结构，分别用于划分场景区域以及建立区域地标点的2D-3D匹配。该方法采用一种基于原型的三元组损失(prototype-based triplet loss)，避免区域过多而导致计算区域分割损失的计算开销与内存占用过大的问题。

为了提高模型的跨场景迁移速度，部分方法提出将场景坐标回归过程解耦为两个环节：1)区域聚类环节与特定场景无关，模型根据外观特征将相近的像素划分至同一类别下，其被认为同属于空间中某处局部区域；2)坐标估计环节则根据特定的目标场景，为每一块局部区域估计其坐标分布。Cavallari等人(2017)提出用公开数据集对模型进行预训练，然后通过一种在线的训练方式将预训练模型迁移至目标场景。预训练的过程使回归森林能够根据像素外观及深度值将属于场景同一区域的像素划分至同一个样本集合当中，而迁移过程中则保持回归森林的结构与各节点的学习参数固定，只对叶节点下的样本集合进行动态更新。为了保证模型迁移的实时性，该方法使用水塘采样算法控制各个叶节点应保留样本的数量，并用轮询和并行的方式对叶节点下的样本分布进行更新。对于有效点占比较高的部分位姿实例，采用最近点迭代算法进行再次优化，并根据深度图合成误差对结果进行二次评估(Cavallari等，2020)。同一时期的ScoreNet方法(Cavallari等，2019)将类似策略应用到基于VGGNet网络结构的视觉定位任务当中。Dong等人(2022)则将场景坐标回归网络划分为与具体场景无关的图像特征提取骨干和与具体场景有关的坐标回归头两部分，其中骨干网络采用VGGNet结构，坐标回归头为层级分类网络，每一级分类子网络由两个卷积层与两个全连接层组成。另外，该方法针对数据库图像过于稀疏的情况，采取Reptile元学习方法(Nichol等，2018)进行少样本训练。该方法在训练数据仅为原始数据集0.5%～1.0%的情况下，可在几分钟内完成训练，定位效率约为HLoc方法(Sarlin等，2019)的两倍，同时达到较高的定位精度。

近年来最新的研究工作进一步推动了该类方法的轻量化发展。NeuMap方法(Tang等，2023)仅对由SuperPoint网络提取的关键点进行场景坐标回归，并且提出用一组可学习的低维嵌入向量对场景进行隐式编码，随后通过一组串联的Transformer解码器结构估计查询图像关键点的场景坐标。该方法的定位精度相较于同类工作有大幅提升。同时值得注意的是，该方法采用紧凑的隐式编码表达对场景坐标回归网络与场景模型进行了解耦，能够在编码体积不到Active Search点云地图25%的情况下，在长时定位数据集中取得更好的表现，而且模型在不同场景下具有较高的迁移效率。ACE(accelerated coordinate encoding)方法(Brachmann等，2023)将网络解耦为ResNet特征提取骨干与坐标回归头，在一轮训练过程中随机输入来自不同视角图像的特征块，降低梯度相关性，从而达到高学习率、快速收敛的目

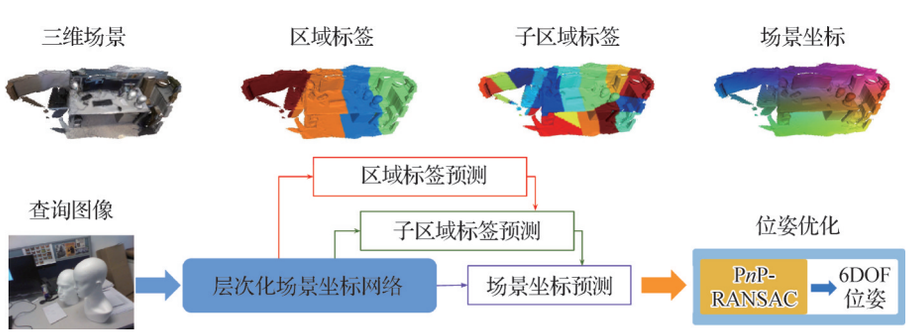


图5　层级场景坐标回归网络(Li等，2020)

Fig. 5　Hierarchical scene coordinate regression network (Li et al., 2020)

的。另外，该方法引入课程学习（curriculum training，CL）方法进一步提高了端到端学习效率，最终能够在 2 min 训练时间内达到与训练长达 15 h 的 DSAC*相当的精度，并且模型体积不足后者的 15%。

2.5　基于神经辐射场的视觉定位

基于神经辐射场（neural radiance fields，NeRF）的新视点图像合成工作（Mildenhall 等，2020）开启了计算机视觉与计算机图形学两大领域相结合的重要研究方向。该工作所采用的神经辐射场目前已成为一种备受关注的新型隐式场景表达形式。

辐射场用于描述三维场景中任意位置 $\boldsymbol{x}=(x,y,z)\in\mathbf{R}^3$ 处的体密度 σ 以及该位置上沿任意方向 $\boldsymbol{d}=(\theta,\varphi)$ 发散出的颜色向量 $\boldsymbol{c}=(r,g,b)$，可由一个五维的向量函数 $f:(\boldsymbol{x},\boldsymbol{d})\rightarrow(\sigma,\boldsymbol{c})$ 表示。辐射场本质上作为一类特殊的高维函数，可以使用诸如多层感知机等网络模型进行拟合。基于辐射场的可微分体渲染（volume rendering）技术允许仅用带位姿的场景图像作为辐射场拟合过程的监督信息。相比传统的显式三维场景模型如稀疏点云，使用神经辐射场作为场景表达的特点在于其能够渲染任意连续视点下的逼真场景图像。从输入与输出的形式上看，视觉定位是基于神经辐射场新视角渲染的逆过程：前者根据输入图像输出该图像相对于场景的 6 自由度相 机位姿；后者则根据输入的相机位姿，由体渲染合成该视点下的场景图像。

针对输入图像位姿带噪声的情况，有一系列方法利用神经辐射场的端到端特性，通过光度一致性损失（photometric loss）对神经辐射场和相机位姿进行联合优化（Chen 等，2023；Lin 等，2021；Meng 等，2021；Truong 等，2023；Wang 等，2021）。iNeRF 方法（Yen-Chen 等，2021）首次提出在给定的场景神经辐射场的条件下，固定辐射场模型参数，单独对查询图像的相机位姿进行迭代优化。如图 6 所示，由于在完整图像的光线集合 $\mathcal{R}$ 上进行采样将带来庞大的计算量，该工作提出一种仅在有效区域内进行的稀疏采样方法，选取部分关键光线进行渲染，以提高相机位姿的优化效率。其所选取的光线 $\boldsymbol{r}$ 能够为相机位姿参数的更新提供有效的梯度信息，在将前向采样次数降低两个数量级的同时，依然能够保证位姿估计结果的准确性。该方法在每一轮迭代过程中，首先通过体渲染技术合成当前相机位姿 $\boldsymbol{T}$ 下选定像素的彩色值 $\hat{C}(\boldsymbol{r})$，计算其与真实观测图像彩色值 $C(\boldsymbol{r})$ 之间的光度一致性损失，再将误差反向传播，以梯度下降的方式对相机位姿参数进行更新。在经典的图像对齐算法中，常采用模糊操作去除图像信号中的高频部分，以得到相对平滑的目标函数，避免优化过程陷入局部最优。这种由粗到细的配准方法同样适用于基于神经辐射场的视觉定位任务，Zhu 等人（2023）采用掩码作为低通滤波器，对神经辐射场的位置编码进行调节，随优化过程逐步激活位置编码的高频部分。

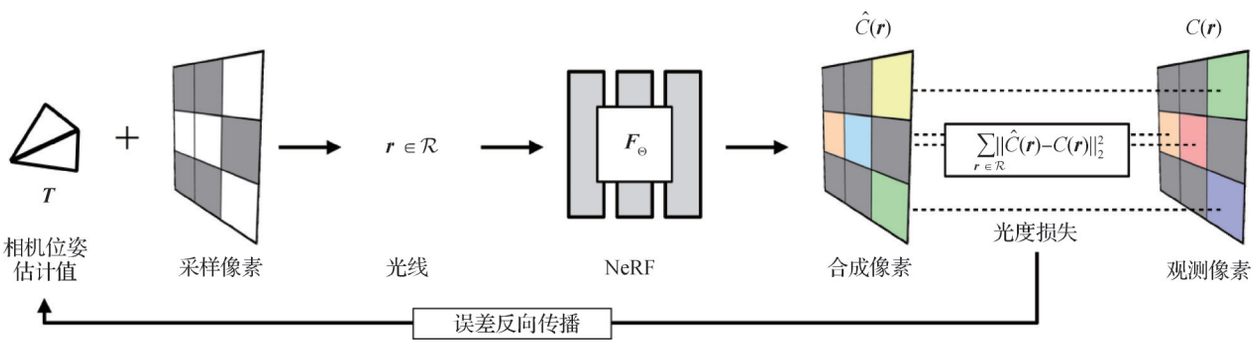


图 6　iNeRF 方法流程图：计算光度误差，随后进行反向传播优化相机位姿（Yen-Chen 等，2021）

Fig. 6　An overview of iNeRF pose estimation pipeline which backpropagates the photometric loss (Yen-Chen et al., 2021)

上述基于光度一致性损失对相机位姿进行迭代优化的方法主要存在两个缺点：1）以光度误差作为损失函数的优化模型高度非凸，这要求相机位姿初值不能与真值存在较大偏离；2）对位姿优化有效的光度一致性损失需依赖众多前提，例如场景表面应当满足漫反射模型，并且查询图像与构建神经辐射场所利用的训练图像应尽量保持一致的光照条件，因此仅仅依靠光度一致性约束优化相机位姿难以获得鲁棒且符合全局几何一致性的结果。针对前一项缺点，目前有方法利用蒙特卡洛采样策略以降低对

相机位姿初值的要求：首先在初值周围随机采样一组位姿实例，在每一轮迭代过程中，通过并行方式来为每一个位姿实例进行更新以及重采样（Lin等，2023；Maggio等，2023）。而对于后一项缺点，一类方法采用域迁移技术，学习查询图像与数据库图像之间的跨域统一特征（Chen等，2022；Liu等，2023a）。还有一类方法提出从神经辐射场中提取深度信息，在光度损失的基础上融合基于场景深度信息的几何损失（Sucar等，2021；Truong等，2023；Zhu等，2022）。

近年来的最新工作提出基于神经辐射场的非迭代方法，即不再依赖位姿初值，而是通过建立显式的3D-2D匹配，对相机位姿进行P*n*P-RANSAC鲁棒估计。NeRF-Loc方法（Liu等，2023a）采用可泛化神经辐射场能够获得空间任意位置的特征描述子，通过注意力机制对3D稀疏点特征与查询图像特征进行交叉融合，再计算相关性矩阵确立三维点云与查询图像之间的稀疏3D-2D匹配。Moreau等人（2023）利用神经辐射场设计了一种自监督的场景稠密特征表示。如图7所示，该方法利用由8层二维卷积神经网络组成的轻量特征提取器从查询图像中提取稠密的特征，另一方面基于位姿先验从神经辐射场中渲染视角对应的特征图，随后建立2D-3D特征点匹配并求解相机位姿。该方法的定位效率约为HLoc方法的两倍，存储占用则不到其2%。

借助神经网络强大的表达能力，用隐式模型代替经典的显式模型可实现连续、稠密场景信息的紧凑存储。目前有关神经辐射场的研究工作多集中于

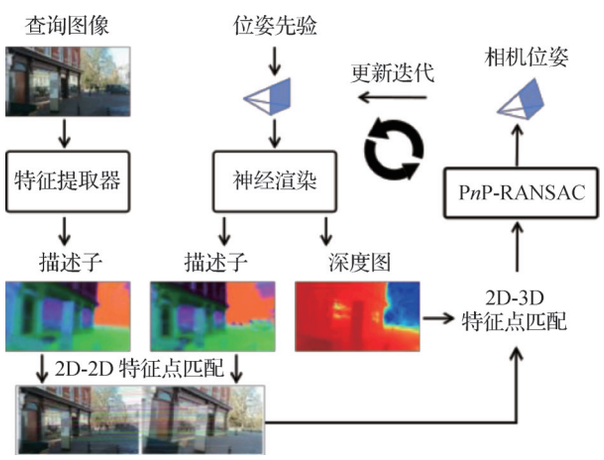


图7　基于隐式表达自监督合成稠密特征的视觉定位方法（Moreau等，2023）

Fig. 7　Camera relocalization on self-supervised features from an implicit representation (Moreau et al., 2023)

通用性问题，例如提高视觉合成质量和训练效率、动态场景的表示等，而神经辐射场的视觉定位作为面向实际应用场景的重要任务，仍有待深入研究。

2.6　基于位姿回归的视觉定位

不同于基于特征一致性建立环境信息匹配并对相机位姿进行解算或迭代优化的方式，相机位姿回归方法采用深度学习模型，将信息匹配与位姿解算环节隐式地包含在网络的推理过程中，以端到端的方式解决相机位姿估计任务。从回归过程与场景模型关系的角度，可将该类方法分为绝对相机位姿回归和相对相机位姿回归两类。

2.6.1　绝对位姿回归

绝对相机位姿回归（absolute pose regression，APR）方法将场景模型构建、场景信息匹配以及相机位姿解算3个环节融合为一个深度神经网络模型，该网络根据相机观测数据直接推断其在场景中的绝对位置与朝向。相机的绝对位姿取决于目标场景所定义的坐标系且网络包含着隐式场景模型，与相对位姿回归方法相比，绝对相机位姿回归网络理论上不具备跨场景泛化性。

PoseNet方法（Kendall等，2015）提供了从单目色彩图像回归相机绝对位姿的网络结构范式。如图8所示，位姿回归网络由3部分组成，包括编码器、全连接层以及线性层。具体而言，$H \times W$的彩色图像输入由卷积神经网络组成的编码器中得到$H' \times W' \times C$的特征图；全连接层随后将特征图映射为低维空间的特征向量；最后的线性层由特征向量回归相机旋转$\hat{\boldsymbol{q}}$与位移$\hat{\boldsymbol{t}}$。训练数据为带位姿标签的数据库图像，网络由包含位置误差与旋转误差两部分加权求和构成的损失函数对训练过程进行监督。其中编码器可以在大规模数据集中进行预训练，而全连接层与线性层则根据具体场景进行迁移学习，实现网络对目标场景信息的隐式编码。

与场景坐标回归方法一致，绝对位姿回归方法的存储占用不随目标场景的实际规模或数据库图像的体积动态变化，仅取决于网络参数量。PoseNet网络的体积不足50 M，单幅图像的定位耗时在100 ms以内。在定位精度方面，PoseNet相较于基于三维结构或层次化的定位方法相差了一个数量级，但是其在保持轻量化优势的同时能够处理一些富有挑战性的场景，例如光照变化剧烈的场景与弱纹理区域。

为了提高绝对相机位姿回归定位系统的可靠

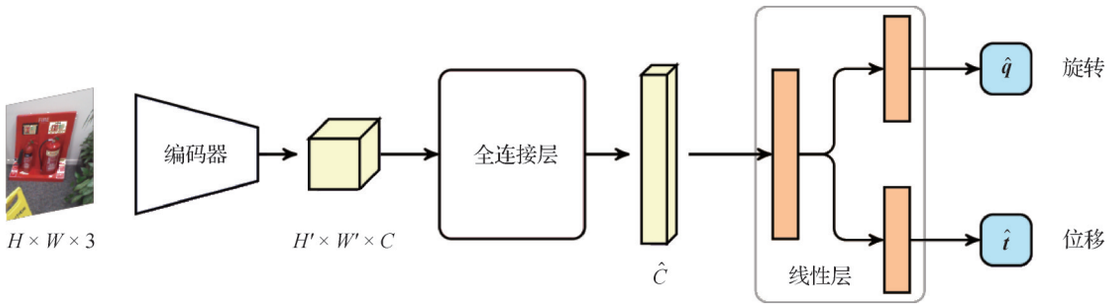


图 8　PoseNet 网络结构（Kendall 等，2015）

Fig. 8　The network structure of PoseNet (Kendall et al. , 2015)

性，Kendall 和 Cipolla（2016）进一步提出采用服从伯努利分布的贝叶斯卷积神经网络，为网络参数及预测位姿结果引入不确定性。在预测过程中根据蒙特卡洛丢弃（Monte Carlo dropout）法则进行多次网络前向采样过程，获得一组位姿预测值，再以它们的均值作为最终结果，并用方差衡量估计误差。

一些工作针对 PoseNet 在网络结构方面的不足做出了改进。Walch 等人（2017）发现特征嵌入向量的维数过高易导致线性层在回归相机位姿的过程中发生过拟合，于是提出在全连接层和线性层之间加入长短时记忆网络（long short-term memory，LSTM），作为特殊的图像特征关联环节，特征嵌入向量的维数由 2 024 进一步降至 128。实验结果表明，这种降低图像特征向量维数的策略能够大幅提高端到端位姿回归的性能。Melekhov 等人（2017）提出用一种基于对称编解码器结构（encoder-decoder）的沙漏神经网络（hourglass network）代替原来的编码器结构，引入解码器结构能够帮助网络更好地提取图像中的细粒度特征。Bui 等人（2019）将场景模型定义为 RGB 图像与相机位姿的联合分布，采用对抗网络的方式学习 RGB 图像与相机位姿之间的几何关联，并提出通过最小化判别器输出损失的方式对相机位姿估计进行优化。绝对位姿回归方法的泛化性仅限于训练数据所对应的目标场景，为了尽可能在不同场景下复用位姿估计模型参数，降低训练成本，APANet 方法（Chidlovskii 和 Sadek，2020）提出一种深度适应网络，用于学习与特定场景无关的图像特征，并基于域对抗网络（domain adversarial networks）结构进行跨场景的模型迁移。AtLoc 方法（Wang 等，2020）在编码器的全卷积骨干网络后加入基于自注意力机制的特征变换模块，使网络能够自适应地强化对环境中鲁棒结构的关注，在观测图像当中存在动态物体或光照剧烈变化的情况下，输出的特征编码依然保持稳定。

一部分工作为网络估计相机绝对位姿的过程引入额外的辅助信息。MapNet 方法（Brahmbhatt 等，2018）在训练数据中按一定间隔随机采样一组二元图像对，在原有损失函数的基础上引入相对位姿估计误差，从而引入更强的几何全局一致性以监督回归网络的训练过程。该方法还提出将图像序列作为输入，以辅助学习的方式来提升绝对相机位姿回归的表现。结合外部的辅助任务，如视觉里程计（visual odometry），以相对位姿作为辅助约束来学习绝对位姿，同时还可利用无位姿标签的视频序列对位姿回归网络进行半监督训练。VidLoc 方法（Clark 等，2017）以视频片段作为输入，提出用双向 LSTM 循环神经网络对时空信息进行建模，发现即使仅考虑较短的图片序列，网络也能输出平滑的位姿估计，而且可大大降低定位误差。VLocNet 方法（Valada 等，2018）同样引入相对位姿估计（视觉里程计）作为辅助任务，设计了参数共享的多任务学习模型同时对当前帧的绝对位姿和相对上一帧的相对位姿进行估计，采用几何一致性损失函数对网络进行监督训练。VLocNet++方法（Radwan 等，2018）采用多任务学习方法同时关联 3 项任务，包括语义分割、绝对和相对位姿估计。具体而言，其通过一种自适应加权层对相对位姿估计模块与语义分割模块输出的特征图进行聚合，聚合得到的特征图将辅助绝对位姿估计模块进行全局位姿回归。Xue 等人（2019）利用序列图像中的时空一致性，结合视觉里程计组件缓解视觉歧义所带来的不确定性，并设计了基于注意力机制的特征增强模块，其能较好地聚合局部地图中的共视特征。还可将视觉里程计估计的相对位姿作为运动约束，在位姿图中对绝对位姿进行细化。

网络回归相机旋转的能力与空间旋转的参数表达形式以及误差的度量方式有着密切关联。PoseNet系列方法对旋转回归网络分支输出的四维向量进行归一化后得到可用于描述相机朝向的单位四元数，通过计算单位四元数表示下估计值与真值之间的欧氏距离来衡量旋转误差。MapNet方法（Brahmbhatt等，2018）则采用单位四元数的对数形式，即旋转回归网络分支能够不加以归一化过程直接输出三维向量来表示相机朝向，同样可在该形式下计算估计值与真值之间的欧氏距离来测量旋转误差，该方法相比于过度参数化（over-parameterized）的单位四元数形式能够训练出回归旋转性能更好的网络。Zhou等人（2019b）在对旋转表示方法的连续性进行分析后指出，所有维数小于等于4的3D旋转表达，如欧拉角、旋转向量、单位四元数等，在实欧氏空间中均不连续，因而这些表达形式会降低网络的学习能力。一种连续表达方式是令网络回归6维向量，再经过格拉姆—施密特（Gram-Schmidt）正交化过程将其映射为一个3 × 3的旋转矩阵，实验验证表明，该方式可有效提升网络回归的精度。另一种方式是令网络回归9维向量，再经过正交普鲁克（orthogonal Procrustes）方法获得距离该9维向量最近的正交旋转矩阵（Chen等，2021）。Cai等人（2021）对欧拉角进行离散化，将角度回归问题化为分类问题。

对于不同尺度的场景，损失函数中的位置误差与旋转误差之间的量级大小将存在差异，用于平衡两者的权重超参数在很大程度上影响着网络的训练效率和最终的回归性能。为了减少超参数个数并提升模型在不同场景下的性能表现，可以引入基于同方差不确定性的（homoscedastic uncertainty）多任务学习方法，分别用可学习的回归不确定性变量对该两项误差的比重进行自适应动态调整（Cipolla等，2018）。另外一种思路是借助几何重投影过程，用图像平面上的像素距离作为代理损失（Kendall和Cipolla，2017）。Direct-PoseNet方法借鉴运动估计中的直接匹配策略，引入光度损失来监督位姿回归网络的训练过程。具体来讲，先在目标场景中分别对位姿回归网络和基于新视角合成的直接匹配模块进行单独训练，然后将其联合后进行网络微调。

为了提高网络在更大规模场景中的位姿回归能力，Blanton等人（2020）在PoseNet网络结构的基础上额外增加了场景区域预测分支，基于预测标签从网络权重数据库中检索对应的回归网络参数，该网络参数能够根据查询图像特征回归其在对应场景下的绝对相机位姿。MS-Transformer方法（Shavit等，2021）借鉴目标检测任务中的DeTR网络模型（Carion等，2020），将每个场景与一个查询向量（query）相关联。如图9所示，该模型首先采用CNN网络提取图像特征，随后用两个基于Transformer结构的分支获取各个场景编码与查询图像的特征关联，接着用全连接层作为分类器确定当前查询图像所在场景，两个分支最终分别通过一个多层感知机（multilayer perceptron，MLP）得到相机的位移$\boldsymbol{t}$与朝向$\boldsymbol{q}$。

Sattler等人（2019）经过理论分析与实验验证后指出，基于PoseNet系列或MapNet系列网络结构的绝对位姿回归方法无法保证在实际场景中应用的泛化性，无论是从本质原理还是实验验证结果来看，其都更接近基于二维图像检索的视觉定位方法。这些绝对位姿回归方法的定位精度在很大程度上取决于数据库图像的采集质量，无论网络结构如何改进，其定位精度都难以取得进一步突破。为了提高绝对位姿回归网络的精度与泛化性，Wu等人（2017）提出一种能够在一定程度的旋转扰动下为每幅训练图像合成新视角图像的方法，缓解训练数据位姿稀疏问题。Naseer和Burgard（2017）提出引入单目深度估计方法从训练图像中合成场景点云，通过对场景点云进行投影的方式合成新视角图像。基于神经辐射场的新视角合成方法为绝对位姿回归网络的训练提供了一种有效的数据增强的方式，进一步缓解了对数据库图像质量的依赖。Moreau等人（2022）首先从训练图像中重建目标场景的神经辐射场，然后通过对位姿进行稠密采样能够合成大量的新视角图像，实验表明用真实数据与合成数据混合训练获得的位姿回归网络具有更高的定位精度。光照变化以及动态物体均会导致在图像上出现较大的光度差别，为了降低光度不稳定性，DFNet方法（Chen等，2022）提出在更加鲁棒的特征图上逐像素计算特征误差作为位姿回归网络的训练损失，替代光度误差与几何误差。该方法还另外引入三元组损失项以缩小NeRF渲染图像与真实图像之间的域差距（domain gap）。

2.6.2 相对位姿回归

相对位姿表示两幅图像观测相机之间的刚体变换关系。相对位姿估计过程通常只需要建立图像间

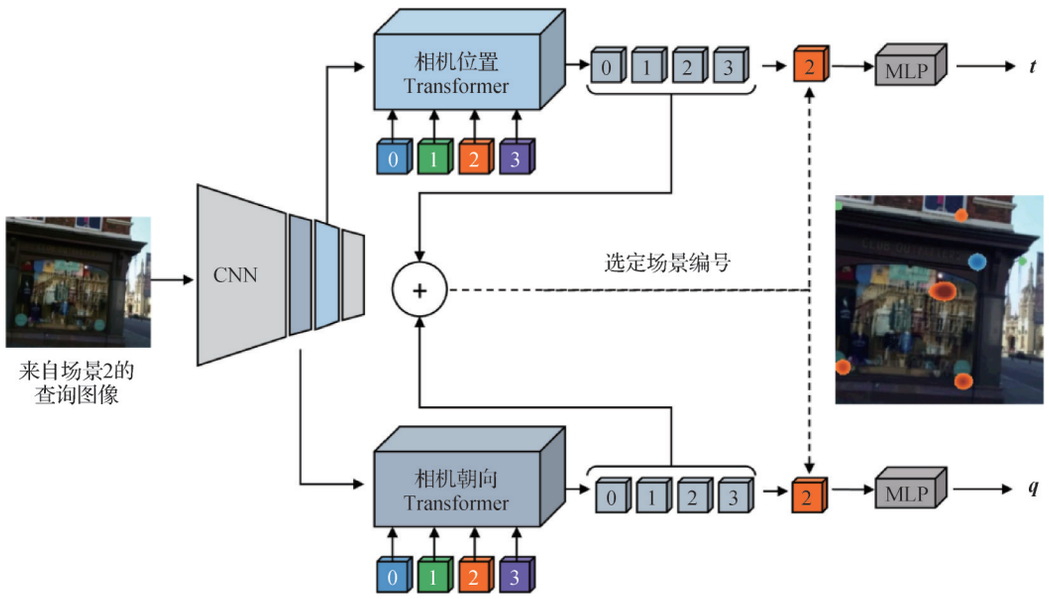


图 9　基于 Transformer 结构的多场景绝对位姿回归网络（Shavit 等，2021）

Fig. 9　Multi-scene absolute pose regression with Transformers（Shavit et al.，2021）

的信息关联，与特定场景、场景三维表达或场景坐标系无关，并且不具有绝对尺度信息。估计相机之间的相对位姿是计算机视觉中的一项基本任务，可以利用对极原理进行几何求解。具体而言，使用几何原理求解相对位姿时，通常首先建立图像间的 2D-2D 特征点匹配，随后可用五点法（Stewénius 等，2006）求解本质矩阵（essential matrix），最终通过对本质矩阵进行奇异值分解得到相机之间的刚体变换关系。SfMLearner 方法（Zhou 等，2017）将相机相对位姿估计与深度估计任务相结合，以自监督的方式对单目深度估计网络和相对位姿估计网络进行联合训练，使二者较好地发挥了协同作用。具体而言，深度估计网络可预测目标图像的深度图，而相对位姿估计网络则能够预测目标图像与前后帧图像之间的位姿变换关系。该方法随后根据深度图和相对位姿合成新视角图像，以合成图像与原图像间的光度误差作为损失函数对两个网络进行训练。

在视觉定位任务中，相对位姿估计通常作为图像检索方法的一种后处理过程，并有一系列方法提出采用网络回归的方式解决查询图像与最近邻数据库图像之间的信息关联以及相对位姿预测任务。Laskar 等人（2017）设计了一种基于孪生结构的 CNN 骨干网络，用于提取查询图像与近邻图像的特征向量，二者的特征向量在进行拼接后将输入回归网络以预测相对位姿。回归网络结构与 PoseNet 相似，由全连接层与双分支线性层组成。该方法将分别估计查询图像与检索得到的多幅近邻图像之间的相对位姿，随后经过旋转变换以及三角化后得到多个带尺度信息的绝对位姿估计值，并以一种最大化内点率的方式确定最终估计结果。RelocNet 方法（Balntas 等，2018）在此基础上引入连续的度量学习技术，并提出用视锥体重叠损失（frustum overlap loss）监督图像全局特征的学习过程，以提升图像检索效果。CamNet 方法（Ding 等，2019）在基于图像检索的环境信息匹配环节采用了由粗到细的层次化策略，编码器部分是由 3 个分支组成的孪生网络，其对应 3 个解码器分别对应基于图像的粗检索、基于位姿的细检索以及相对位姿回归 3 个步骤，该方法在很大程度上提升了检索质量与回归精度。不同于直接回归相对位姿，Zhou 等人（2020）提出用网络估计本质矩阵，随后对其进行分解获得相对位姿。该方法能够避免误差权重在不同场景下面临的不确定性，并且表现出更为精确的估计结果。

Sattler 等人（2019）指出相对位姿回归方法在本质上依然与图像检索方法存在内在关联，故同样面临精度不足和过度依赖数据库图像的问题。在相对位姿估计任务中，基于几何原理求解的方法仍然比深度网络直接回归的方法更为有效。Dong 等人（2023）指出在用五点法估计出相对相机位姿后，只需要使用一种经过特殊设计的运动平均（motion

averaging)算法,就能在定位精度上达到领先水平,且该方法无需预先构建真实尺度的3D场景模型,为视觉定位领域提供了一种仅依赖原始图像数据库的轻量化方案。

3 代表性方法性能对比与分析

为了客观、公平地评价视觉定位研究成果的有效性,目前领域内已出现一系列公开数据集和评价指标,以便研究者对方法进行训练、测试和综合比较。本节将对现有的一部分常用数据集以及评价指标进行介绍,并从定位精度以及方法轻量化角度,对该领域近年来的代表性方法进行对比分析。

3.1 常用公开数据集

目前针对视觉定位任务所提出的公开数据集已超过20个,其各自之间的差异主要体现在数据采集场景、场景规模、数据形式和观测条件等方面,覆盖了视觉定位任务在自动驾驶、增强现实等各个应用场景下的具体需求以及存在的困难与挑战。部分常见数据集的具体细节如表1所示,它们广泛应用于评估、验证视觉定位方法的过程之中。

在室内场景下,虽然相机观测条件受自然气候影响较小,但由于环境中存在重复性纹理、弱纹理区域以及对称结构等,并且需要考虑人或物体移动所导致的场景结构变化以及室内光照变化等因素,这些都是室内视觉定位任务所面临的挑战。7Scenes和12Scenes是面向室内场景的代表性公开数据集,其采集过程中相机的观测条件相对稳定,并且采集图像包含稠密的深度信息。两种数据集所提供的真值标签分别来自KinectFucion(Izadi等,2011;Newcombe等,2011)和BundleFusion(Dai等,2017)两种基于深度图的同步定位与建图算法。图10展示了数据集中包含的7个场景,从上至下依次为各场景示例图像、3D稠密重建模型以及采集轨迹真值。

表1 常见视觉定位公开数据集

Table 1 Overview of common datasets for visual localization

数据集	场景	图像样本量			参考点云点的数量/M	观测条件变化			真值来源
		总数	数据库	查询		天气	季节	昼夜	
7Scenes(Shotton等,2013)	室内	43 000	26 000	17 000	–	–	–	–	D-SLAM
Cambridge Landmarks(Kendall等,2015)	城市、街道	10 929	6 848	4 081	1.54	√	–	–	SfM
12Scenes(Valentin等,2016)	室内	246 673	240 002	6 671	–	–	–	–	D-SLAM
Aachen Day-Night(Sattler等,2018)	城市、街道	5 250	4 328	922	1.65	–	–	√	SfM+手动
RobotCar Seasons(Sattler等,2018)	城市、街道	38 055	26 121	11 934	6.77	√	√	√	ICP+手动
Extended CMU-Seasons(Toft等,2022)	城镇、郊区	117 550	60 937	56 613	3.37	√	√	–	SfM+手动

注:"–"表示未提供数据。

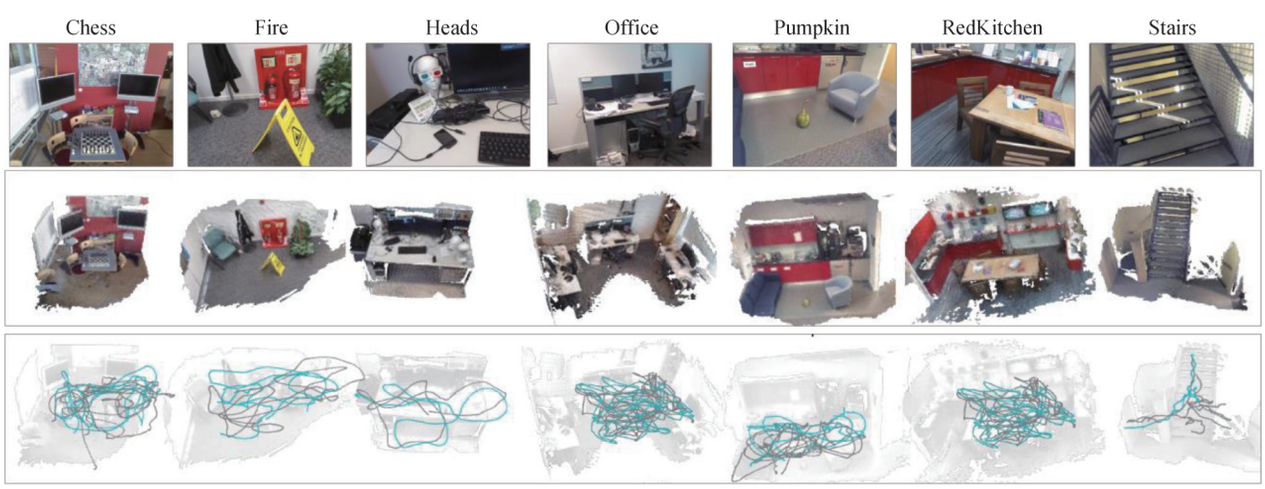


图10 7Scenes室内数据集

Fig. 10 7Scenes indoor dataset

室外场景的规模相较于室内场景更大，同样面临场景结构改变或地图歧义性等挑战，并且在长时视觉定位任务中，场景外观往往由于时间以及气候影响会发生变化。Cambridge Landmarks 数据集（Kendall 等，2015）针对室外中小型地标建筑，采用 SfM 算法获得图像的位姿真值，图 11 展示了该数据集 5 个场景所对应的示例图像以及采集轨迹真值。Aachen Day-Night 数据集（Sattler 等，2018）包含分别于白天和夜晚采集到的中等规模城市街景图像（如图 12（a）所示），能够考验算法根据白天图像定位夜间图像的泛化性能，该数据集为保证位姿真值尽可能准确，采取手动方式为夜间图像标注匹配点。RobotCar Seasons（Toft 等，2022）为跨越多个城市街区的大规模场景数据集，同时数据的采集时间跨度长达一年，对时段、气候和路况等场景观测变量进行了较为全面的覆盖，该数据利用同时采集到的激光点云间接计算图像的位姿标签。Extended CMU-Seasons（Toft 等，2022）是面向大规模郊外场景的数据集，相比于城市街道，郊外场景的植被覆盖率大，场景外观及结构将随气候更替发生更为剧烈的改变。

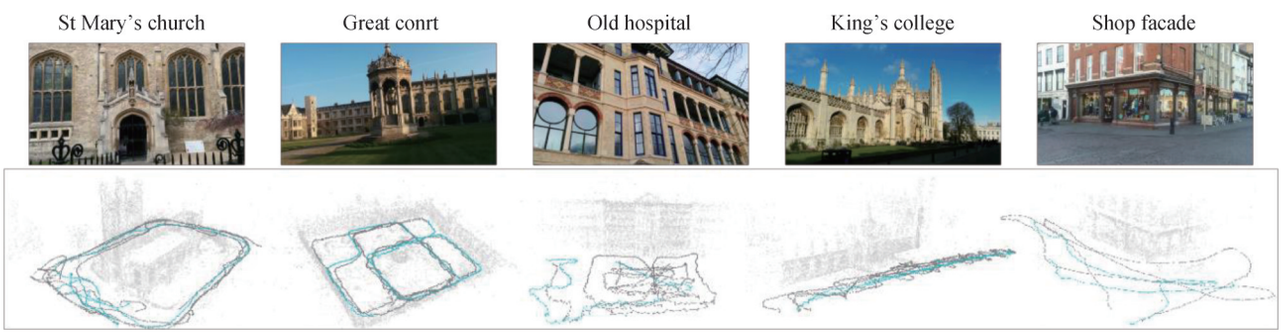


图 11　Cambridge Landmarks 室外数据集

Fig. 11　Cambridge Landmarks outdoor dataset

3.2　评价指标

在评估视觉定位方案的性能时，通常考虑的指标包括定位精度、定位效率、构建场景模型的效率以及场景模型的存储占用。理想的轻量化视觉定位方案在保持较高定位精度的同时，场景模型的构建以及存储占用更加高效。对于定位精度，通常由绝对定位误差统计量和召回率两种指标进行评估。

绝对定位误差包含位移误差 Δ_t 与旋转误差 Δ_q 两部分，计算方式为

$$\left(\Delta_t, \Delta_q\right) = \left(\left\| \boldsymbol{t} - \bar{\boldsymbol{t}} \right\|_2, 2\frac{180^\circ}{\pi}\arccos\left(\left\| \boldsymbol{q}\bar{\boldsymbol{q}} \right\| \right) \right) \quad (1)$$

式中，位移误差 Δ_t 为相机位移估计值 $\boldsymbol{t}$ 与位移真值 $\bar{\boldsymbol{t}}$ 之间的欧氏距离，常以 m 为单位；旋转误差 Δ_q 表示从相机朝向估计值 $\boldsymbol{q}$ 对齐真值 $\bar{\boldsymbol{q}}$ 所需的最小旋转角度，常以度为单位。式中的相机朝向均为四元数表示。

在计算某个场景下所有查询图像的绝对定位误差后，进一步进行统计学分析，通常报告这组误差的平均值、中位数或标准差。不同于直接比较绝对定位误差统计量，召回率是统计准确定位次数占查询总数百分比的间接评价指标，其数值越大代表算法的定位精度越高。目前研究者通常采用统一的误差阈值作为衡量查询图像是否被准确定位的依据：若查询图像的位移误差与旋转误差同时小于阈值，则认为该查询图像的定位结果准确。对于 7Scenes 等小规模室内数据集，通常选取（0.05 m，5°）作为误差阈值；对于 Aachen Day-Night 等中大型室外场景，Sattler 等人（2018）提出目前通常采用由粗至细的分层评价方式，确定了 3 组误差阈值：（5 m，10°）、（0.5 m，5°）、（0.25 m，2°）。

3.3　性能对比分析

7Scenes 室内数据集与 Cambridge Landmarks 室外数据集是目前最为常用的视觉定位任务公开数据集，Aachen Day-Night 是难度更大的长时定位数据集，本节将展示各类视觉定位方法在这 3 个数据集上的性能数据。由于线地图和高精地图视觉定位方法对应用场景具有特殊要求，并且目前尚缺乏针对这两类方法的统一数据集，因此该部分主要对其余的视觉定位方法进行对比分析。

表 2 列出了近年来视觉定位领域代表性方法在 7Scenes 小型室内数据集上的性能表现。不同于现有视觉定位的相关综述，本文提出将构建场景模型

的耗时以及场景模型的存储占用纳入评价指标以进行不同类方法间的纵向比较。其中，场景模型栏中的辅助信息表示对应方法在构建场景模型的过程中是否借助了外部传感器或算法提供的除单目RGB图像及位姿标签外的场景信息，例如，场景坐标回归类别下的一部分方法在训练回归模型的过程中需要借助深度图像，神经辐射场视觉定位类别下的两种方法目前均使用了额外深度估计算法获得关于场景深度的先验信息，而位姿回归下的MapNet + PGO（Brahmbhatt 等，2018）和 VlocNet++（Radwan 等，2018）方法则将序列图像作为训练输入，可利用视觉里程计输出的前后帧相对位姿进行自监督学习，并

表2　各方法在7Scenes室内定位数据集上的性能评估

Table 2　Holistic evaluation of well-exemplified methods on the 7Scenes indoor dataset

	典型方法	场景模型				中值误差	召回率/%
		辅助信息	耗时	存储空间占用	旋转/(°)	位移/m	(5°, 0.05 m)
基线方法	DenseVLAD（Torii等，2015）	/	～2.3 min	//	13.11	0.26	–
	Active Search (SIFT)（Sattler等，2017）	/	～3.3 h	～200 M	2.46	0.05	68.70
	HLoc (SP+SG)（Sarlin等，2019）	/	～3.3 h	～2 G	1.07	0.03	76.80
点云压缩	BP*n*PNet (SP)（Campbell等，2020a）	/	～3.3 h	～60 M	39.30	1.61	–
	GoMatch (SP)（Zhou等，2022）	/		～60 M	4.61	0.18	–
点线联合	PtLine（Gao等，2022）	/	–	–	1.09	0.03	72.70
		场景深度信息	–	–	0.93	0.03	80.60
	LIMAP（Liu等，2023b）	/	–	–	1	0.03	78.00
		场景深度信息	–	–	0.87	0.03	83.50
场景坐标回归	SANet（Yang等，2019）	场景深度信息	～3.7 min	～550 M	1.68	0.05	–
	HSCNet（Li等，2020）	场景深度信息	1.7 h	24 M	0.90	0.03	84.80
	DSAC*（Brachmann和Rother，2022）	/	15 h	28 M	1.41	0.03	81.10
		场景深度信息	15 h	28 M	1.25	0.02	85.20
	DSM（Tang等，2021）	场景深度信息	–	400 M	0.99	0.03	78.10
	SRC（Dong等，2022）	场景深度信息	2 min	40 M	0.79	0.03	–
	NeuMap（Tang等，2023）	/	–	～2 M	1.09	0.03	–
	ACE（Brachmann等，2023）	/	5 min	4 M	–	–	80.80
NeRF	NeRF-Loc（Liu等，2023a）	场景参考点云	–	–	1.33	0.02	89.50
	CROSSFIRE（Moreau等，2023）	/	～40 min	～50 M	1.39	0.04	–
绝对\相对位姿回归	PoseNet17（Kendall和Cipolla，2017）	/	4~24 h	～50 M	8.12	0.23	–
	MapNet+PGO（Brahmbhatt等，2018）	视觉里程计	–	–	6.55	0.18	–
	VlocNet++（Radwan等，2018）	视觉里程计+语义信息	–	–	1.39	0.02	96.40
	RelocNet（Balntas等，2018）	/	～3.7 min	//	6.73	0.21	–
	CamNet（Ding等，2019）	/	～3.7 min	//	1.69	0.04	–
	MS-Trans（Shavit等，2021）	/	～17 min	～10 M	7.28	0.18	–
	LENS（Moreau等，2022）	/	数日	50 M	2.96	0.08	–
	DFNet（Chen等，2022）	/	–	–	2.21	0.07	–

注："/"表示对应方法不依赖外部辅助信息；"//"表示对应方法不使用图像（特征）数据库以外的场景模型；"～"表示粗略测量值，"–"表示相关文献未报告该项指标。

且后者在损失函数中加入了语义损失项。

所有指标是对数据集中7个场景求取均值后的结果，其中场景模型的构建耗时与存储占用数据主要参考（Brachmann 等，2023）提供的统计信息。需要指出的是，由于构建耗时取决于实际硬件平台的运动性能，表中数据仅提供数量级上的参考。通过比较分析可知，在小型室内场景中基于场景坐标回归的方法整体表现更优。该类方法在保持较高定位精度的同时，所依赖的场景模型相比传统视觉定位方法更为轻量，并且回归模型的训练效率近两年得到了显著提升。其中，Brachmann 等人（2023）提出的ACE方法相比于层次化定位方法HLoc，在定位精度方面有所提高的同时，建图时间仅为后者的2.5%，模型体积仅为后者的0.2%。其余各类方法相比于基于点云地图的传统视觉定位方法尽管在定位精度方面不占优势，但场景模型的构建及存储成本方面普遍具有较大优势。

表3整理了各类方法在Cambridge Landmarks中

表3　典型方法在Cambridge室外定位数据集上的性能评估

Table 3　Holistic evaluation of well-exemplified methods on the Cambridge outdoor dataset

	典型方法	场景模型			中值误差	
		场景表达形式	耗时	存储空间占用/M	旋转/(°)	位移/m
基线方法	DenseVLAD（Torii 等，2015）	数据库图像	～1 min	/	7.12	2.56
	Active Search（Sattler 等，2017）+ SIFT（Lowe，2004）	SfM 点云	～35 min	～200	0.63	0.29
	HLoc（Sarlin 等，2019）+ SuperPoint（DeTone 等，2018）+ SuperGlue（Sarlin 等，2020）	数据库图像+SfM 点云	～35 min	～800	0.23	0.10
点云压缩	HybridSC（Camposeco 等，2019）	SfM 点云	～35 min	～1	1.73	0.56
	BP*n*PNet（Campbell 等，2020a）+ SuperPoint	点云	～35 min	～12	106.72	17.54
	QP+RootSIFT（Mera-Trujillo 等，2020）	SfM 点云	～35 min	～2	1.39	0.93
	GoMatch（Zhou 等，2022）+ SuperPoint	点云	～35 min	～12	5.85	1.73
	SceneSqueezer（Yang 等，2022）	SfM 点云	～35 min	～0.5	0.42	0.23
点线联合	PtLine（Gao 等，2022）	SfM 点云+线云	–	–	0.13	0.07
	LIMAP（Liu 等，2023b）	SfM 点云+线云	–	–	0.12	0.07
场景坐标回归	HSCNet（Li 等，2020）	网络模型	3 h	40	0.30	0.13
	DSAC*（Brachmann 和 Rother，2022）w/depth	网络模型	15 h	28	0.35	0.14
	DSAC*（Brachmann 和 Rother，2022）	网络模型	15 h	28	0.40	0.15
	DSM（Tang 等，2021）	网络模型	–	400	0.37	0.16
	SRC（Dong 等，2022）	网络模型	2 min	40	0.80	0.32
	NeuMap（Tang 等，2023）	隐式编码	–	～0.3	0.33	0.14
	ACE（Brachmann 等，2023）	网络模型	5 min	4	0.48	0.21
NeRF	NeRF-Loc（Liu 等，2023a）	神经辐射场	–	–	0.25	0.10
	CROSSFIRE（Moreau 等，2023）	神经辐射场	～4 h	～50	1	0.37
绝对位姿回归	PoseNet17（Kendall 和 Cipolla，2017）	网络模型	4～24 h	～50	2.85	1.63
	MS-Trans（Shavit 等，2021）	网络模型	～7 h	～18	2.75	1.28
	LENS（Moreau 等，2022）	网络模型	数日	50	1.15	0.39
	DFNet（Chen 等，2022）	网络模型	–	60	0.96	0.39

注：w/depth 表示不借助额外的场景深度信息；“/”表示对应方法不使用图像（特征）数据库以外的场景模型；“～”表示粗略测量值；“–”表示相关文献未报告该项指标。

小型室外数据集上的性能表现，并在场景模型栏中列出了各类方法所用场景模型的具体表达形式。在点云压缩方法类别中，BP*n*PNet（Campbell等，2020a）和GoMatch（Zhou等，2022）方法使用未存储视觉特征向量的3D点云地图，在定位精度上的表现相对劣势。而SceneSqueezer（Yang等，2022）方法通过端到端的方式学习点云的下采样策略，成为目前该类别下的领先工作。场景坐标回归方法在室外场景下依然保持着精度优势，NeuMap（Tang等，2023）方法对场景进行隐式编码，使回归模型与场景解耦，成为目前最高效的场景表达方法之一。

以神经辐射场作为场景模型完成视觉定位任务当前正属于较为前沿的研究方向，NeRF-Loc（Liu等，2023a）和CROSSFIRE（Moreau等，2023）两项工作在定位精度上已经达到了较高水平。早期绝对位姿回归方法，例如表中列出的PointNet17（Kendall和Cipolla，2017）与MS-Trans（Shavit等，2021），其在精度上相比于其他定位方法，有接近一个数量级的劣势。Sattler等人（2019）分析指出，对绝对位姿回归网络结构进行优化的工作已很难从泛化性和定位精度的角度作出明显进步。后来出现的LENS（Moreau等，2022）和DFNet方法（Chen等，2022）则借助神经辐射场提供的新视角图像合成方法，为网络训练进行数据增强，在一定程度上突破了该类方法在定位精度上的瓶颈。

表4比较了9种视觉定位方法在Aachen Day-Night室外场景数据集上的测试性能，从右至左依次对应由粗到细的3种误差阈值下的召回率统计结果。图12（b）所示的堆叠柱状图将数据结果可视化，侧面标注的百分比表示对应方法在最小的误差阈值（0.25 m，2°）下的召回率。对比白天（左）和夜间（右）的统计数据可知，由于该数据集仅提供白天图像作为构建场景的数据库，因此大部分方法在夜间的定位准确率相比于白天均发生大幅下降，但传统的层次化策略方法HLoc和基于相对位姿估计并采用运动平均的LazyLoc（Dong等，2023）方法凭借目前成熟的特征提取与匹配算法，取得了优异的鲁棒性。

表4　各方法在Aachen Day-Night室外定位数据集上的性能评估

Table 4　Evaluation of common methods on the Aachen Day-Night outdoor dataset

方法	白天定位召回率			夜间定位召回率		
	（2°，0.25 m）	（5°，0.5 m）	（10°，5.0 m）	（2°，0.25 m）	（5°，0.5 m）	（10°，5.0 m）
DenseVLAD（Torii等，2015）	0.0	0.1	22.8	0.0	1.0	19.4
Active Search (SIFT)（Sattler等，2012）	57.3	83.7	96.6	28.6	37.8	51.0
HLoc (SP+SG)（Sarlin等，2019）	**89.6**	**95.4**	**98.8**	**86.7**	**93.9**	**100.0**
Cascaded（Cheng等，2019）	76.7	88.6	95.8	33.7	48.0	62.2
QP+RootSIFT（Mera-Trujillo等，2020）	62.6	76.3	84.7	16.3	18.4	24.5
Squeezer（Yang等，2022）	75.5	89.7	96.2	50.0	67.3	78.6
ESAC（Brachmann和Rother，2019a）	42.6	59.6	75.5	6.1	10.2	18.4
HSCNet（Li等，2020）	71.1	81.9	91.7	40.8	56.1	76.5
NeuMap（Tang等，2023）	80.8	90.9	95.6	48.0	67.3	87.8
LazyLoc（Dong等，2023）	85.4	92.7	95.8	80.6	88.8	94.9

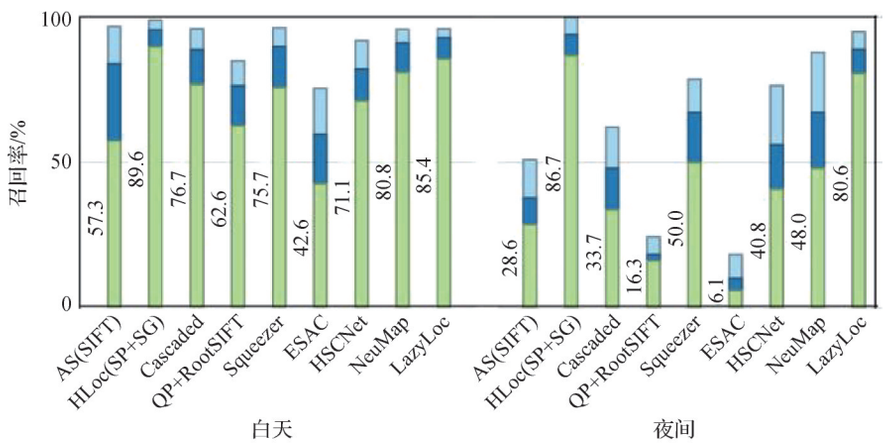
注：加粗字体表示各列最优结果。

图13以场景模型的体积为横轴、以在Aachen Day-Night定位数据集上的夜间定位精度为纵轴，比较了各种方法的场景表达效率。可见，尽管当前轻量化视觉定位相关工作在小型、外观恒定场景中取得了超越传统视觉定位框架的性能表现，但这些方法在大尺度、外观变化场景中的泛化性和鲁棒性仍有待提升。

4　结　语

轻量化对于视觉定位技术的实际应用具有重要价值。本文根据场景模型的表达形式对当前轻量化

(a) 数据集图像示例



(b) 典型方法的召回率

图 12　Aachen Day-Night 室外长时定位数据集

Fig. 12　Aachen Day-Night outdoor dataset for long-term localization

((a) example images in the dataset; (b) recall rates of common methods)

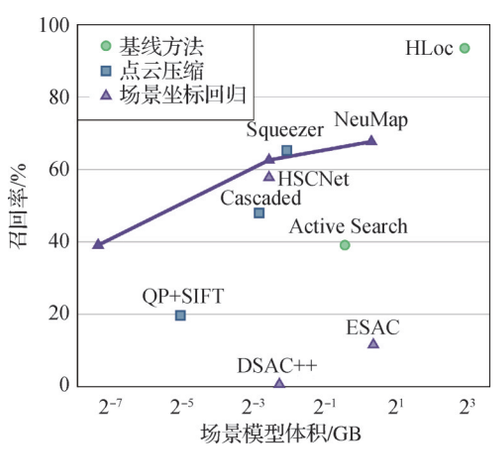


图 13　部分方法在 Aachen Day-Night 室外数据集上夜间定位性能对比

Fig. 13　Recall rate vs model size on the Aachen Day-Night outdoor dataset

视觉定位技术的相关研究进行了分类综述，分别介绍了基于显式模型和隐式模型的轻量化视觉定位方法。以压缩点云、线地图和高精地图为代表的显式模型通过降低传统点云地图信息冗余度或提升特征结构化程度等方式实现场景模型轻量化。而以场景坐标回归网络、神经辐射场、位姿回归网络为代表的隐式方法则通过梯度下降的方式将场景信息编码至网络参数当中。这些现有方法取得了显著的研究进展，基本能够满足轻量化定位需求。由目前的实验数据分析对比可以发现，在场景规模不大的情况下，场景回归方法整体上有着更优的性能表现，其场景模型存储占用普遍在几兆到几十兆不等，并且在 5°、2 cm 误差阈值范围内的定位成功率能够达到 80%。然而，如果进一步扩大场景规模，同时将场景外观及

结构的动态变化等实际情况纳入考量，现有方法在扩展性和鲁棒性方面仍然受到较大局限。

该领域在未来将继续探究如何基于一种更加高效的场景表达形式，实现兼具实时性、高精度、尺度泛化性以及鲁棒性的轻量化视觉定位系统。主要研究趋势包含如下几个方面：

1）探究高效的场景表达形式。如何设计场景表达形式以充分发挥场景信息在视觉定位任务中的作用将会是一个需要长期探索的话题。线地图、高精地图以及建筑白膜（Panek 等，2023）作为当前新兴的结构化显式模型，其相关的视觉定位研究工作目前尚未成熟，是未来具备潜力的研究方向。神经场（neural field）作为新兴的隐式场景表达形式，与其相关的通用性问题研究逐渐成为计算机视觉领域的热门探索方向，如何在视觉定位任务中发挥神经场的场景表达潜力将是较有价值的研究课题。另外还可探究各个场景模型的复合方式，设计多元化的场景表达形式。

2）提升泛化性与鲁棒性。当前基于神经网络进行场景坐标或绝对位姿回归的视觉定位方法由于网络模型的容量有限，难以将方法扩展至大规模场景，采用由粗到细的层级定位策略或集成学习技术是对现有方法作进一步扩展的有效途径。另一方面，由于视觉特征对光照、季节、动态物体等变化条件敏感，现有方法在长时定位任务下的鲁棒性较弱，可考虑将多元传感器信息以辅助学习的方式引入神经网络的训练过程，实现多传感器信息的互补融合。

3）探究无地图定位。现有方法普遍依赖预先构建的场景模型，而模型的构建及存储将会消耗大量资源。目前已经有工作指出，不依赖三维场景模型，仅利用两视图几何原理与运动平均算法，即可达到视觉定位数据集上的领先水平（Dong 等，2023）。Arnold 等人（2022）提出了一种基于重定位锚点的无地图定位方案，每个地点仅使用单幅图像作为参考，通过估计真实尺度的相对位姿实现查询图像定位。目前，无地图定位属于较新且有价值的研究方向，有待不同领域共同探讨。

参考文献（References）

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5297-5307 [DOI: 10.1109/CVPR.2016.572]

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J. 2018. NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1437-1451 [DOI: 10.1109/TPAMI.2017.2711011]

Arnold E, Wynn J, Vicente S, Garcia-Hernando G, Monszpart Á, Prisacariu V, Turmukhambetov D and Brachmann E. 2022. Map-free visual relocalization: metric pose relative to a single image//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 690-708 [DOI: 10.1007/978-3-031-19769-7_40]

Balntas V, Li S D and Prisacariu V. 2018. RelocNet: continuous metric learning relocalisation using neural nets//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 782-799 [DOI: 10.1007/978-3-030-01264-9_46]

Blanton H, Greenwell C, Workman S and Jacobs N. 2020. Extending absolute pose regression to multiple scenes//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 170-178 [DOI: 10.1109/CVPRW50498.2020.00027]

Brachmann E, Cavallari T and Prisacariu V A. 2023. Accelerated coordinate encoding: learning to relocalize in minutes using RGB and poses//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5044-5053 [DOI: 10.1109/CVPR52729.2023.00488]

Brachmann E, Krull A, Nowozin S, Shotton J, Michel F, Gumhold S and Rother C. 2017. DSAC — differentiable RANSAC for camera localization//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2492-2500 [DOI: 10.1109/CVPR.2017.267]

Brachmann E, Michel F, Krull A, Yang M Y, Gumhold S and Rother C. 2016. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3364-3372 [DOI: 10.1109/CVPR.2016.366]

Brachmann E and Rother C. 2018. Learning less is more - 6D camera localization via 3D surface regression//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4654-4662 [DOI: 10.1109/CVPR.2018.00489]

Brachmann E and Rother C. 2019a. Neural-guided RANSAC: learning where to sample model hypotheses//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4321-4330 [DOI: 10.1109/ICCV.2019.00442]

Brachmann E and Rother C. 2019b. Expert sample consensus applied to camera re-localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South):

IEEE：7524-7533［DOI：10.1109/ICCV.2019.00762］

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

Brahmbhatt S，Gu J W，Kim K，Hays J and Kautz J. 2018. Geometry-aware learning of maps for camera localization//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：2616-2625［DOI：10.1109/CVPR.2018.00277］

Brown M，Windridge D and Guillemaut J Y. 2015. Globally optimal 2D-3D registration from points or lines without correspondences//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago，Chile：IEEE：2111-2119［DOI：10.1109/ICCV.2015.244］

Bui M，Baur C，Navab N，Ilic S and Albarqouni S. 2019. Adversarial networks for camera pose regression and refinement//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul，Korea（South）：IEEE：3778-3787［DOI：10.1109/ICCVW.2019.00470］

Cai R J，Hariharan B，Snavely N and Averbuch-Elor H. 2021. Extreme rotation estimation using dense correlation volumes//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：14561-14570［DOI：10.1109/CVPR46437.2021.01433］

Campbell D，Liu L and Gould S. 2020a. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：244-261［DOI：10.1007/978-3-030-58536-5_15］

Campbell D，Petersson L，Kneip L and Li H D. 2020b. Globally-optimal inlier set maximisation for camera pose and correspondence estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence，42(2)：328-342［DOI：10.1109/TPAMI.2018.2848650］

Campbell D，Petersson L，Kneip L，Li H D and Gould S. 2019. The alignment of the spheres：globally-optimal spherical mixture alignment for camera pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：11788-11798［DOI：10.1109/CVPR.2019.01207］

Camposeco F，Cohen A，Pollefeys M and Sattler T. 2019. Hybrid scene compression for visual localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：7645-7654［DOI：10.1109/CVPR.2019.00784］

Camposeco F，Sattler T，Cohen A，Geiger A and Pollefeys M. 2017. Toroidal constraints for two-point localization under high outlier ratios//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：6700-6708［DOI：10.1109/CVPR.2017.709］

Cao S and Snavely N. 2014. Minimal scene descriptions from structure from motion models//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus，USA：IEEE：461-468［DOI：10.1109/CVPR.2014.66］

Carion N，Massa F，Synnaeve G，Usunier N，Kirillov A and Zagoruyko S. 2020. End-to-end object detection with Transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：213-229［DOI：10.1007/978-3-030-58452-8_13］

Cavallari T，Bertinetto L，Mukhoti J，Torr P and Golodetz S. 2019. Let's take this online：adapting scene coordinate regression network predictions for online RGB-D camera relocalisation//Proceedings of 2019 International Conference on 3D Vision. Québec City，Canada：IEEE：564-573［DOI：10.1109/3DV.2019.00068］

Cavallari T，Golodetz S，Lord N A，Valentin J，di Stefano L and Torr P H S. 2017. On-the-fly adaptation of regression forests for online camera relocalisation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：218-227［DOI：10.1109/CVPR.2017.31］

Cavallari T，Golodetz S，Lord N A，Valentin J，Prisacariu V A，Stefano L D and Torr P H S. 2020. Real-time RGB-D camera pose estimation in novel scenes using a relocalisation cascade. IEEE Transactions on Pattern Analysis and Machine Intelligence，42(10)：2465-2477［DOI：10.1109/TPAMI.2019.2915068］

Chen K F，Snavely N and Makadia A. 2021. Wide-baseline relative camera pose estimation with directional learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：3257-3267［DOI：10.1109/CVPR46437.2021.00327］

Chen L C，Zhu Y K，Papandreou G，Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：833-851［DOI：10.1007/978-3-030-01234-2_49］

Chen S，Li X H，Wang Z R and Prisacariu V A. 2022. DFNet：enhance absolute pose regression with direct feature matching//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv，Israel：Springer：1-17［DOI：10.1007/978-3-031-20080-9_1］

Chen Y，Chen X Y，Wang X，Zhang Q，Guo Y，Shan Y and Wang F. 2023. Local-to-global registration for bundle-adjusting neural radiance fields//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：8264-8273［DOI：10.1109/CVPR52729.2023.00799］

Chen Z H，Pei H Y，Wang J K and Dai D Y. 2021. Survey of monocular camera-based visual relocalization. Robot，43(3)：373-384（陈宗海，裴浩渊，王纪凯，戴德云. 2021. 基于单目相机的视觉重定位方法综述. 机器人，43(3)：373-384）［DOI：10.13973/j.cnki.robot.200350］

Cheng W T, Lin W S, Chen K and Zhang X F. 2019. Cascaded parallel filtering for memory-efficient image-based localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1032-1041 [DOI: 10.1109/ICCV.2019.00112]

Chidlovskii B and Sadek A. 2020. Adversarial transfer of pose estimation regression//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 646-661 [DOI: 10.1007/978-3-030-66415-2_43]

Cipolla R, Gal Y and Kendall A. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7482-7491 [DOI: 10.1109/CVPR.2018.00781]

Clark R, Wang S, Markham A, Trigoni N and Wen H K. 2017. VidLoc: a deep spatio-temporal model for 6-DoF video-clip relocalization//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2652-2660 [DOI: 10.1109/CVPR.2017.284]

Dai A, Nießner M, Zollhöfer M, Izadi S and Theobalt C. 2017. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM Transactions on Graphics, 36(3): #24 [DOI: 10.1145/3054739]

David P, DeMenthon D, Duraiswami R and Samet H. 2004. SoftPOSIT: simultaneous pose and correspondence determination. International Journal of Computer Vision, 59(3): 259-284 [DOI: 10.1023/B:VISI.0000025800.10423.1f]

DeTone D, Malisiewicz T and Rabinovich A. 2018. SuperPoint: self-supervised interest point detection and description//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE: 337-349 [DOI: 10.1109/CVPRW.2018.00060]

Ding M Y, Wang Z, Sun J K, Shi J P and Luo P. 2019. CamNet: coarse-to-fine retrieval for camera re-localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2871-2880 [DOI: 10.1109/ICCV.2019.00296]

Dong S Y, Wang S Z, Zhuang Y X, Kannala J, Pollefeys M and Chen B Q. 2022. Visual localization via few-shot scene region classification//Proceedings of 2022 International Conference on 3D Vision. Prague, Czech Republic: IEEE: 393-402 [DOI: 10.1109/3DV57658.2022.00051]

Dong S Y, Liu H, Guo H K, Chen B Q and Pollefeys M. 2023. Lazy visual localization via motion averaging [EB/OL]. [2023-08-19]. http://arxiv.org/pdf/2307.09981.pdf

Donoser M and Schmalstieg D. 2014. Discriminative feature-to-point matching in image-based localization//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 516-523 [DOI: 10.1109/CVPR.2014.73]

Fang Q H, Yin Y D, Fan Q N, Xia F, Dong S Y, Wang S, Wang J, Guibas L J and Chen B Q. 2022. Towards accurate active camera localization//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 122-139 [DOI: 10.1007/978-3-031-20080-9_8]

Fischler M A and Bolles R C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6): 381-395 [DOI: 10.1145/358669.358692]

Gao S, Wan J X, Ping Y S, Zhang X D, Dong S Z, Yang Y C, Ning H K, Li J J N and Guo Y D. 2022. Pose refinement with joint optimization of visual points and lines//Proceedings of 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan: IEEE: 2888-2894 [DOI: 10.1109/IROS47612.2022.9981420]

Gong R H, Liu X L, Jiang S H, Li T X, Hu P, Lin J Z, Yu F W and Yan J J. 2019. Differentiable soft quantization: bridging full-precision and low-bit neural networks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4851-4860 [DOI: 10.1109/ICCV.2019.00495]

Goto T, Pathak S, Ji Y, Fujii H, Yamashita A and Asama H. 2018. Line-based global localization of a spherical camera in manhattan worlds//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 2296-2303 [DOI: 10.1109/ICRA.2018.8460920]

Guzman-Rivera A, Kohli P, Glocker B, Shotton J, Sharp T, Fitzgibbon A and Izadi S. 2014. Multi-output learning for camera relocalization//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1114-1121 [DOI: 10.1109/CVPR.2014.146]

Hofer M, Maurer M and Bischof H. 2017. Efficient 3D scene abstraction using line segments. Computer Vision and Image Understanding, 157: 167-178 [DOI: 10.1016/j.cviu.2016.03.017]

Huang Z Y, Zhou H, Li Y J, Yang B B, Xu Y, Zhou X W, Bao H J, Zhang G F and Li H S. 2021. VS-net: voting with segmentation for visual localization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6097-6107 [DOI: 10.1109/CVPR46437.2021.00604]

Irschara A, Zach C, Frahm J M and Bischof H. 2009. From structure-from-motion point clouds to fast location recognition//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 2599-2606 [DOI: 10.1109/CVPR.2009.5206587]

Izadi S, Kim D, Hilliges O, Molyneaux D, Newcombe R, Kohli P, Shotton J, Hodges S, Freeman D, Davison A and Fitzgibbon A. 2011. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA: ACM: 559-568 [DOI: 10.1145/2047196.2047270]

Jégou H, Douze M, Schmid C and Pérez P. 2010. Aggregating local descriptors into a compact image representation//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 3304-3311 [DOI: 10.1109/CVPR.2010.5540039]

Jeong J, Cho Y and Kim A. 2020. HDMI-Loc: exploiting high definition map image for precise localization via bitwise particle filter. IEEE Robotics and Automation Letters, 5(4): 6310-6317 [DOI: 10.1109/LRA.2020.3013881]

Kabsch W. 1976. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A, 32(5): 922-923 [DOI: 10.1107/S0567739476001873]

Kendall A and Cipolla R. 2016. Modelling uncertainty in deep learning for camera relocalization//Proceedings of 2016 IEEE International Conference on Robotics and Automation. Stockholm, Sweden: IEEE: 4762-4769 [DOI: 10.1109/ICRA.2016.7487679]

Kendall A and Cipolla R. 2017. Geometric loss functions for camera pose regression with deep learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6555-6564 [DOI: 10.1109/CVPR.2017.694]

Kendall A, Grimes M and Cipolla R. 2015. PoseNet: a convolutional network for real-time 6-DOF camera relocalization//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2938-2946 [DOI: 10.1109/ICCV.2015.336]

Kim J, Choi C, Jang H and Kim Y. 2023. LDL: line distance functions for panoramic localization//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 17836-17846 [DOI: 10.1109/ICCV51070.2023.01639]

Krull A, Brachmann E, Michel F, Yang M Y, Gumhold S and Rother C. 2015. Learning analysis-by-synthesis for 6D pose estimation in RGB-D images//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 954-962 [DOI: 10.1109/ICCV.2015.115]

Laskar Z, Melekhov I, Kalia S and Kannala J. 2017. Camera relocalization by computing pairwise relative poses using convolutional neural network//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 920-929 [DOI: 10.1109/ICCVW.2017.113]

Lepetit V, Moreno-Noguer F and Fua P. 2009. EP*n*P: an accurate $O(n)$ solution to the P*n*P problem. International Journal of Computer Vision, 81(2): 155-166 [DOI: 10.1007/s11263-008-0152-6]

Li X T, Wang S Z, Zhao Y, Verbeek J and Kannala J. 2020. Hierarchical scene coordinate classification and regression for visual localization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11980-11989 [DOI: 10.1109/cvpr42600.2020.01200]

Li X T, Ylioinas J, Verbeek J and Kannala J. 2019. Scene coordinate regression with angle-based reprojection loss for camera relocalization//Proceedings of the 15th European Conference on Computer Vision Workshops. Munich, Germany: Springer: 229-245 [DOI: 10.1007/978-3-030-11015-4_19]

Li Y P, Snavely N, Huttenlocher D and Fua P. 2012. Worldwide pose estimation using 3D point clouds//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 15-29 [DOI: 10.1007/978-3-642-33718-5_2]

Li Y P, Snavely N and Huttenlocher D P. 2010. Location recognition using prioritized feature matching//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer: 791-804 [DOI: 10.1007/978-3-642-15552-9_57]

Liao W L, Zhao H Q and Yan J C. 2021. Online extrinsic camera calibration based on high-definition map matching on public roadway. Journal of Image and Graphics, 26(1): 208-217 (廖文龙, 赵华卿, 严骏驰. 2021. 开放道路中匹配高精度地图的在线相机外参标定. 中国图象图形学报, 26(1): 208-217) [DOI: 10.11834/jig.200432]

Lin C H, Ma W C, Torralba A and Lucey S. 2021. BARF: bundle-adjusting neural radiance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 5721-5731 [DOI: 10.1109/ICCV48922.2021.00569]

Lin Y Z, Müller T, Tremblay J, Wen B W, Tyree S, Evans A, Vela P A and Birchfield S. 2023. Parallel inversion of neural radiance fields for robust pose estimation//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, United Kingdom: IEEE: 9377-9384 [DOI: 10.1109/ICRA48891.2023.10161117]

Lindeberg T. 1998. Edge detection and ridge detection with automatic scale selection. International Journal of Computer Vision, 30(2): 117-156 [DOI: 10.1023/A:1008097225773]

Lindenberger P, Sarlin P-E and Pollefeys M. 2023. LightGlue: local feature matching at light speed//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 17581 - 17592 [DOI: 10.1109/ICCV51070.2023.01616]

Liu J L, Nie Q, Liu Y and Wang C J. 2023a. NeRF-loc: visual localization with conditional neural radiance field//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, United Kingdom: IEEE: 9385-9392 [DOI: 10.1109/ICRA48891.2023.10161420]

Liu L, Li H D and Dai Y C. 2017. Efficient global 2D-3D matching for camera localization in a large-scale 3D map//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2391-2400 [DOI: 10.1109/ICCV.2017.260]

Liu S, Zhang Y X, Xu J T, Zou D F, Chen S Y and Wang Z H. 2020. Visual prior-information-based map recovery slam in complex scenes. Journal of Image and Graphics, 25(1): 158-170 (刘盛, 张宇翔, 徐婧婷, 邹大方, 陈胜勇, 王振华. 2020. 复杂场景下视觉先验信息的地图恢复SLAM. 中国图象图形学报, 25(1): 158-170) [DOI: 10.11834/jig.190041]

Liu S H, Yu Y F, Pautrat R, Pollefeys M and Larsson V. 2023b. 3D

line mapping revisited//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21445-21455 [DOI: 10.1109/CVPR52729.2023.02054]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Lowry S, Sünderhauf N, Newman P, Leonard J J, Cox D, Corke P and Milford M J. 2016. Visual place recognition: a survey. IEEE Transactions on Robotics, 32(1): 1-19 [DOI: 10.1109/TRO.2015.2496823]

Lu Y, Huang J W, Chen Y T and Heisele B. 2017. Monocular localization in urban environments using road markings//Proceedings of 2017 IEEE Intelligent Vehicles Symposium. Los Angeles, USA: IEEE: 468-474 [DOI: 10.1109/IVS.2017.7995762]

Maggio D, Abate M, Shi J N, Mario C and Carlone L. 2023. Loc-NeRF: Monte Carlo localization using neural radiance fields//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, United Kingdom: IEEE: 4018-4025 [DOI: 10.1109/ICRA48891.2023.10160782]

Melekhov I, Ylioinas J, Kannala J and Rahtu E. 2017. Image-based localization using hourglass networks//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 870-877 [DOI: 10.1109/ICCVW.2017.107]

Meng Q, Chen A P, Luo H M, Wu M Y, Su H, Xu L, He X M and Yu J Y. 2021. GNeRF: GAN-based neural radiance field without posed camera//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6331-6341 [DOI: 10.1109/ICCV48922.2021.00629]

Mera-Trujillo M, Smith B and Fragoso V. 2020. Efficient scene compression for visual-based localization//Proceedings of 2020 International Conference on 3D Vision. Fukuoka, Japan: IEEE: 1-10 [DOI: 10.1109/3DV50981.2020.00111]

Micusík B and Wildenauer H. 2015. Descriptor free visual indoor localization with line segments//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3165-3173 [DOI: 10.1109/CVPR.2015.7298936]

Micusik B and Wildenauer H. 2017. Structure from motion with line segments under relaxed endpoint constraints. International Journal of Computer Vision, 124(1): 65-79 [DOI: 10.1007/s11263-016-0971-9]

Middelberg S, Sattler T, Untzelmann O and Kobbelt L. 2014. Scalable 6-DOF localization on mobile devices//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 268-283 [DOI: 10.1007/978-3-319-10605-2_18]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer International Publishing: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Moreau A, Piasco N, Bennehar M, Tsishkou D, Stanciulescu B and de La Fortelle A. 2023. CROSSFIRE: camera relocalization on self-supervised features from an implicit representation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 252-262 [DOI: 10.1109/ICCV51070.2023.00030]

Moreau A, Piasco N, Tsishkou D, Stanciulescu B and de La Fortelle A. 2022. LENS: localization enhanced by NeRF synthesis//Proceedings of the 5th Conference on Robot Learning. Auckland, New Zealand: PMLR: 1347-1356

Moreno-Noguer F, Lepetit V and Fua P. 2008. Pose priors for simultaneously solving alignment and correspondence//Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer: 405-418 [DOI: 10.1007/978-3-540-88688-4_30]

Mur-Artal R, Montiel J M M and Tardós J D. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 31(5): 1147-1163 [DOI: 10.1109/TRO.2015.2463671]

Naseer T and Burgard W. 2017. Deep regression for monocular camera-based 6-DoF global localization in outdoor environments//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada: IEEE: 1525-1530 [DOI: 10.1109/IROS.2017.8205957]

Newcombe R A, Fitzgibbon A, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A J, Kohi P, Shotton J and Hodges S. 2011. KinectFusion: real-time dense surface mapping and tracking//Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality. Basel, Switzerland: IEEE: 127-136 [DOI: 10.1109/ISMAR.2011.6092378]

Nichol A, Achiam J and Schulman J. 2018. On first-order meta-learning algorithms [EB/OL]. [2023-10-23]. https://arxiv.org/pdf/1803.02999v3.pdf

Nistér D. 2003. Preemptive RANSAC for live structure and motion estimation//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE: 199-206 [DOI: 10.1109/ICCV.2003.1238341]

Ozuysal M, Calonder M, Lepetit V and Fua P. 2010. Fast keypoint recognition using random ferns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3): 448-461 [DOI: 10.1109/TPAMI.2009.23]

Pan X K, Liu H M, Fang M, Wang Z, Zhang Y and Zhang G F. 2023. Dynamic 3D scenario-oriented monocular slam based on semantic probability prediction. Journal of Image and Graphics, 28(7): 2151-2166 (潘小鹍, 刘浩敏, 方铭, 王政, 张涌, 章国锋. 2023. 基于语义概率预测的动态场景单目视觉SLAM. 中国图象图形学报, 28(7): 2151-2166) [DOI: 10.11834/jig.210632]

Panek V, Kukelova Z and Sattler T. 2023. Visual localization using imperfect 3D models from the internet//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13175-13186 [DOI: 10.1109/CVPR52729.2023.01266]

Park H S, Wang Y, Nurvitadhi E, Hoe J C, Sheikh Y and Chen M. 2013. 3D point cloud reduction using mixed-integer quadratic programming//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA: IEEE: 229-236 [DOI: 10.1109/CVPRW.2013.41]

Perez E, Strub F, De Vries H, Dumoulin V and Courville A. 2018. FiLM: visual reasoning with a general conditioning layer//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press: 3942-3951 [DOI: 10.1609/aaai.v32i1.11671]

Poggenhans F, Salscheider N O and Stiller C. 2018. Precise localization in high-definition road maps for urban regions//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 2167-2174 [DOI: 10.1109/IROS.2018.8594414]

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

Ranganathan A, Ilstrup D and Wu T. 2013. Light-weight localization for vehicles using road markings//Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE: 921-927 [DOI: 10.1109/IROS.2013.6696460]

Revaud J, Weinzaepfel P, De Souza C and Humenberger M. 2019. R2D2: repeatable and reliable detector and descriptor//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 12414-12424

Rublee E, Rabaud V, Konolige K and Bradski G. 2011. ORB: an efficient alternative to SIFT or SURF//Proceedings of 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE: 2564-2571 [DOI: 10.1109/ICCV.2011.6126544]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Sarlin P E, Cadena C, Siegwart R and Dymczyk M. 2019. From coarse to fine: robust hierarchical localization at large scale//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12708-12717 [DOI: 10.1109/CVPR.2019.01300]

Sarlin P E, Debraine F, Dymczyk M, Siegwart R and Cadena C. 2018. Leveraging deep visual descriptors for hierarchical efficient localization//Proceedings of the 2nd Conference on Robot Learning. Zürich, Switzerland: PMLR: 456-465 [DOI: 10.3929/ETHZ-B-000318818]

Sarlin P E, DeTone D, Malisiewicz T and Rabinovich A. 2020. SuperGlue: learning feature matching with graph neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4937-4946 [DOI: 10.1109/CVPR42600.2020.00499]

Sarlin P E, Dusmanu M, Schönberger J L, Speciale P, Gruber L, Larsson V, Miksik O and Pollefeys M. 2022. LaMAR: benchmarking localization and mapping for augmented reality//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 686-704 [DOI: 10.1007/978-3-031-20071-7_40]

Sattler T, Havlena M, Radenovic F, Schindler K and Pollefeys M. 2015. Hyperpoints and fine vocabularies for large-scale location recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2102-2110 [DOI: 10.1109/ICCV.2015.243]

Sattler T, Leibe B and Kobbelt L. 2011. Fast image-based localization using direct 2D-to-3D matching//Proceedings of 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE: 667-674 [DOI: 10.1109/ICCV.2011.6126302]

Sattler T, Leibe B and Kobbelt L. 2012. Improving image-based localization by active correspondence search//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 752-765 [DOI: 10.1007/978-3-642-33718-5_54]

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

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

Sattler T, Zhou Q J, Pollefeys M and Leal-Taixé L. 2019. Understanding the limitations of CNN-based absolute camera pose regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3297-3307 [DOI: 10.1109/CVPR.2019.00342]

Schonberger J L and Frahm J M. 2016. Structure-from-motion revisited//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4104-4113 [DOI: 10.1109/CVPR.2016.445]

Schreiber M, Knöppel C and Franke U. 2013. LaneLoc: lane marking based localization using highly accurate maps//Proceedings of 2013 IEEE Intelligent Vehicles Symposium. Gold Coast, Australia: IEEE: 449-454 [DOI: 10.1109/IVS.2013.6629509]

Shavit Y, Ferens R and Keller Y. 2021. Learning multi-scene absolute pose regression with Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2713-2722 [DOI: 10.1109/ICCV48922.2021.00273]

Shi T X, Shen S H, Gao X and Zhu L J. 2019. Visual localization using

sparse semantic 3D map//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 315-319 [DOI: 10.1109/ICIP.2019.8802957]

Shi Y, Cai J X, Shavit Y, Mu T J, Feng W S and Zhang K. 2022. ClusterGNN: cluster-based coarse-to-fine graph neural network for efficient feature matching//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12507-12516 [DOI: 10.1109/CVPR52688.2022.01219]

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

Speciale P, Schonberger J L, Kang S B, Sinha S N and Pollefeys M. 2019. Privacy preserving image-based localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5488-5498 [DOI: 10.1109/CVPR.2019.00564]

Stewénius H, Engels C and Nistér D. 2006. Recent developments on direct relative orientation. ISPRS Journal of Photogrammetry and Remote Sensing, 60(4): 284-294 [DOI: 10.1016/j.isprsjprs.2006.03.005]

Sucar E, Liu S K, Ortiz J and Davison A J. 2021. IMAP: implicit mapping and positioning in real-time//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6209-6218 [DOI: 10.1109/ICCV48922.2021.00617]

Tang S T, Tang C Z, Huang R, Zhu S Y and Tan P. 2021. Learning camera localization via dense scene matching//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1831-1841 [DOI: 10.1109/CVPR46437.2021.00187]

Tang S T, Tang S C, Tagliasacchi A, Tan P and Furukawa Y. 2023. NeuMap: neural coordinate mapping by auto-transdecoder for camera localization//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 929-939 [DOI: 10.1109/CVPR52729.2023.00096]

Toft C, Maddern W, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J, Pajdla T, Kahl F and Sattler T. 2022. Long-term visual localization revisited. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4): 2074-2088 [DOI: 10.1109/TPAMI.2020.3032010]

Torii A, Arandjelović R, Sivic J, Okutomi M and Pajdla T. 2015. 24/7 place recognition by view synthesis//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1808-1817 [DOI: 10.1109/CVPR.2015.7298790]

Truong P, Rakotosaona M J, Manhardt F and Tombari F. 2023. SPARF: neural radiance fields from sparse and noisy poses//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: 4190-4200 [DOI: 10.1109/CVPR52729.2023.00408]

Valada A, Radwan N and Burgard W. 2018. Deep auxiliary learning for visual localization and odometry//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 6939-6946 [DOI: 10.1109/ICRA.2018.8462979]

Valentin J, Dai A, Niessner M, Kohli P, Torr P, Izadi S and Keskin C. 2016. Learning to navigate the energy landscape//Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford, USA: 323-332 [DOI: 10.1109/3DV.2016.41]

Valentin J, Niebner M, Shotton J, Fitzgibbon A, Izadi S and Torr P. 2015. Exploiting uncertainty in regression forests for accurate camera relocalization//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4400-4408 [DOI: 10.1109/CVPR.2015.7299069]

Walch F, Hazirbas C, Leal-Taixé L, Sattler T, Hilsenbeck S and Cremers D. 2017. Image-based localization using LSTMs for structured feature correlation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 627-637 [DOI: 10.1109/ICCV.2017.75]

Wang B, Chen C H, Lu C X, Zhao P J, Trigoni N and Markham A. 2020. AtLoc: attention guided camera localization//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 10393-10401 [DOI: 10.1609/aaai.v34i06.6608]

Wang Z R, Wu S Z, Xie W D, Chen M and Prisacariu V. 2021. NeRF--: neural radiance fields without known camera parameters [EB/OL]. [2023-09-09]. https://arxiv.org/pdf/2102.07064.pdf

Wei D, Wan Y, Zhang Y J, Liu X Y, Zhang B and Wang X Q. 2022. ELSR: efficient line segment reconstruction with planes and points guidance//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15786-15794 [DOI: 10.1109/CVPR52688.2022.01535]

Wen T P, Jiang K, Wijaya B, Li H Y, Yang M M and Yang D G. 2022. TM3Loc: tightly-coupled monocular map matching for high precision vehicle localization. IEEE Transactions on Intelligent Transportation Systems, 23(11): 20268-20281 [DOI: 10.1109/TITS.2022.3176914]

Wu J, Ma L W and Hu X L. 2017. Delving deeper into convolutional neural networks for camera relocalization//Proceedings of 2017 IEEE International Conference on Robotics and Automation. Singapore, Singapore: IEEE: 5644-5651 [DOI: 10.1109/ICRA.2017.7989663]

Wu T and Ranganathan A. 2013. Vehicle localization using road markings//Proceedings of 2013 IEEE Intelligent Vehicles Symposium. Gold Coast, Australia: IEEE: 1185-1190 [DOI: 10.1109/IVS.2013.6629627]

Xu C, Zhang L L, Cheng L and Koch R. 2017. Pose estimation from line correspondences: a complete analysis and a series of solutions. IEEE Transactions on Pattern Analysis and Machine Intelligence,

39(6)：1209-1222［DOI：10.1109/TPAMI.2016.2582162］

Xue F，Wang X，Yan Z K，Wang Q Y，Wang J Q and Zha H B. 2019. Local supports global：deep camera relocalization with sequence enhancement//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea（South）：IEEE：2841-2850［DOI：10.1109/ICCV.2019.00293］

Yang L W，Bai Z Q，Tang C Z，Li H H，Furukawa Y and Tan P. 2019. SANet：scene agnostic network for camera localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea（South）：IEEE：42-51［DOI：10.1109/ICCV.2019.00013］

Yang L W，Shrestha R，Li W B，Liu S C，Zhang G F，Cui Z P and Tan P. 2022. SceneSqueezer：learning to compress scene for camera relocalization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：8249-8258［DOI：10.1109/CVPR52688.2022.00808］

Yen-Chen L，Florence P，Barron J T，Rodriguez A，Isola P and Lin T Y. 2021. INeRF：inverting neural radiance fields for pose estimation//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague，Czech Republic：IEEE：1323-1330［DOI：10.1109/IROS51168.2021.9636708］

Yoon S and Kim A. 2021. Line as a visual sentence：context-aware line descriptor for visual localization. IEEE Robotics and Automation Letters，6(4)：8726-8733［DOI：10.1109/LRA.2021.3111760］

Yu H，Zhen W K，Yang W，Zhang J and Scherer S. 2020. Monocular camera localization in prior LiDAR maps with 2D-3D line correspondences//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas，USA：IEEE：4588-4594［DOI：10.1109/IROS45743.2020.9341690］

Zhang C，Liu H，Xie Z J，Yang K Y，Guo K，Cai R and Li Z W. 2021. AVP-Loc：surround view localization and relocalization based on HD vector map for automated valet parking//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague，Czech Republic：IEEE：5552-5559［DOI：10.1109/IROS51168.2021.9636746］

Zhou L P，Ye J M and Kaess M. 2019a. A stable algebraic camera pose estimation for minimal configurations of 2D/3D point and line correspondences//Proceedings of the 14th Asian Conference on Computer Vision. Perth，Australia：Springer：273-288［DOI：10.1007/978-3-030-20870-7_17］

Zhou Q J，Agostinho S，Ošep A and Leal-Taixé L. 2022. Is geometry enough for matching in visual localization?//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv，Israel：Springer：407-425［DOI：10.1007/978-3-031-20080-9_24］

Zhou Q J，Sattler T，Pollefeys M and Leal-Taixé L. 2020. To learn or not to learn：visual localization from essential matrices//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris，France：IEEE：3319-3326［DOI：10.1109/ICRA40945.2020.9196607］

Zhou T H，Brown M，Snavely N and Lowe D G. 2017. Unsupervised learning of depth and ego-motion from video//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：6612-6619［DOI：10.1109/CVPR.2017.700］

Zhou Y，Barnes C，Lu J W，Yang J M and Li H. 2019b. On the continuity of rotation representations in neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：5738-5746［DOI：10.1109/CVPR.2019.00589］

Zhu Z H，Peng S Y，Larsson V，Xu W W，Bao H J，Cui Z P，Oswald M R and Pollefeys M. 2022. NICE-SLAM：neural implicit scalable encoding for SLAM//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：12776-12786［DOI：10.1109/CVPR52688.2022.01245］

Zhu Z X，Chen Y T，Wu Z R，Hou C，Shi Y L，Li C X，Li P F，Zhao H and Zhou G Y. 2023. LATITUDE：robotic global localization with truncated dynamic low-pass filter in city-scale NeRF//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London，United Kingdom：IEEE：8326-8332［DOI：10.1109/ICRA48891.2023.10161570］

作者简介

叶翰樵，男，博士研究生，主要研究方向为三维计算机视觉、场景三维重建与感知、视觉定位。

E-mail：yehanqiao2022@ia.ac.cn

申抒含，通信作者，男，研究员，主要研究方向为三维计算机视觉理论与应用，包括大规模场景三维重建、智能机器人三维环境感知、场景三维语义理解。

E-mail：shshen@nlpr.ia.ac.cn

刘养东，男，助理研究员，主要研究方向为增强现实和三维视觉。E-mail：yangdong.liu@ia.ac.cn