# 基于深度学习的图像匹配:方法、应用与挑战

孔庆群[1),2)] 吴福朝[1)] 樊 彬[3)]

[1)](中国科学院自动化研究所 北京 100190)

[2)](中国科学院大学 北京 100049)

[3)](北京科技大学 智能科学与技术学院 北京 100083)

**摘 要** 图像匹配旨在建立图像之间的点对应关系,是许多计算机视觉任务的关键环节.近年来,随着深度学习技术的发展,图像匹配方法已从以手工设计特征为主转变为基于深度网络的方法,基于深度学习的图像匹配方法在多个标准数据集上展现出卓越的性能,推动着多个相关应用的发展.围绕图像匹配涉及的若干关键问题,如:特征点检测、特征点描述、稠密点匹配、误匹配去除,本文对深度学习图像匹配方法进行了系统性总结.首先分析了领域内基于深度学习的典型方法和关键技术,随后介绍了与图像匹配密切相关的几个典型应用并给出其现状分析,最后,根据对图像匹配领域技术发展的分析总结,结合作者在该领域的长期研究积累,本文给出了目前图像匹配所面临的主要挑战以及未来发展趋势.

**关键词** 图像匹配;特征点匹配;稠密匹配;三维重建;视觉定位;同时定位与建图;深度学习

**中图法分类号** TP18 **DOI号** 10.11897/SP.J.1016.2024.01485

# Image Matching in Deep Learning Era: Methods, Applications and Challenges

KONG Qing-Qun[1),2)] WU Fu-Chao[1)] FAN Bin[3)]

[1)](*Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

[2)](*University of Chinese Academy of Sciences, Beijing* 100049)

[3)](*School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing* 100083)

**Abstract** Image matching is a crucial technique within the field of computer vision, primarily focused on identifying and establishing point correspondences between two different images depicting the same scene. It seeks to find points in one image that correspond to points in another, thus enabling a wide range of computer vision tasks that rely on the analysis of multiple images of the same object or scene from different viewpoints or at different times, including but not limited to 3D reconstruction, motion tracking, image stitching for panoramic views, and visual localization. Traditionally, this process has leaned heavily on the use of hand-crafted keypoint detectors and local descriptors, i.e., algorithms and methodologies designed to pinpoint and describe discriminative features within a local image region, aiming to achieve invariance to scale, rotation, and changes in lighting and perspective. In recent years, with the revolutionary development of deep learning in many areas of computer vision, image matching methods have switched from handcrafted design style to relying on deep learning. The advent of deep learning technologies has catalyzed significant advancements in the area of image matching, and numerous deep learning based image matching techniques have emerged, showcasing promising results

收稿日期:2023-05-26;在线发布日期:2024-04-25. 本课题得到北京市自然科学基金(4202073)、国家自然科学基金(62222302,61876180)资助. **孔庆群**,博士,副研究员,主要研究领域为人工智能、类脑视觉理解. E-mail: qingqun.kong@ia.ac.cn. **吴福朝**,研究员,主要研究领域为计算机视觉、模式识别. **樊 彬**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为计算机视觉、模式识别、人工智能. E-mail: bin.fan@ieee.org.

across a wide range of benchmarks. This has also significantly accelerated the development of many downstream applications of image matching, notably including structure from motion, visual localization, and simultaneous localization and mapping (SLAM), among others.

This paper aims to provide a comprehensive overview of deep learning-based image matching methods that have emerged in recent years. By delving into the core challenges of image matching, including keypoint detection, local feature description, dense matching, and mismatch removal, it offers a detailed summary of the innovative deep learning approaches devised to tackle these issues. This systematic review not only highlights the advancements in the field but also sheds light on how these cutting-edge methods have redefined the landscape of image matching, setting new benchmarks for accuracy, efficiency, and reliability. Specifically, it first delineates the problem definition of image matching and describes the main challenges. Then, it proceeds to dissect each problem associated with image matching, offering a thorough analysis of typical and emblematic methods. Additionally, it delves into the critical techniques employed by deep learning to address these issues, providing an in-depth exploration of how these innovative approaches can effectively solve the challenges inherent in image matching. Moreover, some highly related downstream tasks of image matching are described along with a detailed introduction of their state of the art. These downstream tasks include 3D reconstruction/structure from motion, image based localization, and simultaneous localization and mapping. Besides exploring these downstream applications, this paper provides a comprehensive description of popular benchmarks for image matching and its downstream tasks. Finally, the paper discusses the remaining challenges and future research directions. In conclusion, this paper presents itself as an invaluable resource for researchers and engineering technicians within related fields, enabling the swift assimilation of knowledge concerning the fundamentals, challenges, key technological advancements, and the current state of the art in image matching. As such, it can be served as a comprehensive resource for researchers venturing into this field, providing references in terms of research directions and dataset resources. Through its detailed exposition, the paper aims to catalyze further exploration and innovation, thereby could contributing significantly to the advancement of image matching and its application in advancing the frontiers of computer vision.

**Keywords** image match; feature point match; dense match; 3D reconstruction; visual localization; simultaneous localization and mapping; deep learning

## 1 引 言

图像匹配旨在建立不同图像之间相同物理点[1,2]或者相同语义点之间的对应关系[3,4],其中后者亦称为语义匹配,本文主要讨论面向前者的图像匹配方法,两者的具体定义和区别详见第 2 节. 建立同一实际物体在不同图像之间的点对应关系,是三维计算机视觉的基本出发点[5,6],许多三维计算机视觉的理论都建立在已知图像点对应关系基础上,三维重建[7,8]、相机姿态计算[9]、视觉定位[10,11]、图像拼接[12]、增强现实[13]、同步定位与地图绘制(Simultaneous Localization and Mapping, SLAM)[14-16]等三维计算机视觉应用都离不开高质量的图像匹配算法. 此外,高质量的图像匹配算法还可直接应用于物体识别[1,17]、目标跟踪[18,19]等经典计算机视觉问题,而且遥感图像和医学影像处理中的图像融合与变化检测等应用方向[20-23]均离不开图像匹配. 可以说,图像匹配是计算机视觉和图像处理领域极具应用价值的一个研究方向,得到了研究人员的广泛关注.

早期的图像匹配方法以手工设计的特征为主,其中最具代表性的工作是 SIFT (Scale Invariant

Feature Transform)[1]和 SURF (Speeded Up Robust Features)[24],尤其是 SIFT,不仅推动了图像匹配领域的技术进步,还影响了图像识别、目标检测等众多计算机视觉技术的发展.例如:在 SIFT 基础上提出的 HoG (Histogram of Gradients)[25,26]特征在深度学习出现之前一直都是行人检测领域的主流方法,并在一般性的目标检测领域也得到了广泛应用[27,28],而基于 SIFT 这种局部图像特征发展起来的视觉词袋(Bag of Visual Words)[29,30]方法则在很长一段时间都主导着图像分类技术的发展.比 SIFT 计算更加高效的 SURF 方法则推动了许多对实时性图像特征点匹配有要求的应用技术发展,如:目前广泛使用的视觉 SLAM 方法,即 ORB-SLAM 系列[14,31,32],依然是基于手工特征 ORB (Oriented FAST and Rotated BRIEF)[33]的方法.

图 1 概括了前深度学习时代图像匹配领域典型方法的发展历程,包括最早期的基于灰度统计量的方法[34-36]、后续出现的基于梯度统计量[1,2,24,37]、基于灰度大小关系[38-40]和基于二进制特征表示的方法[33,41,42],以及在深度学习出现之前使用传统机器学习方法进行数据驱动的图像匹配方法的一些尝试[43-45],更多关于手工设计的图像特征匹配方法的介绍可参考综述论文[7,46-48],本文聚焦于深度学习时代的图像匹配.

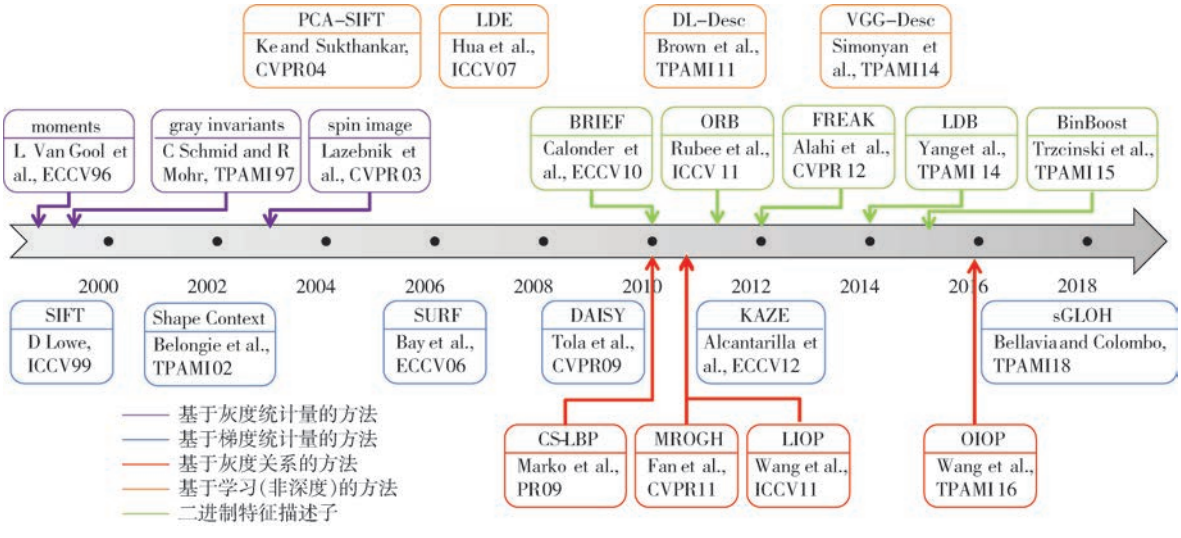



图 1　前深度学习时代的图像匹配领域的典型方法发展历程

随着深度学习技术的不断发展[49-52],图像匹配领域也取得了长足的进步,出现了许多优秀的基于深度学习的方法,在图像匹配涉及的多个方面均取得了显著效果,如:图像特征点检测[53]、图像特征点描述[54]、稠密图像点匹配[55]、错误匹配点滤除[56]等.相比传统的手工设计方法以及非深度学习的机器学习方法,基于深度学习的图像匹配方法不仅在图像匹配相关的数据集上取得了卓越的性能提升[47,57],而且在多个以特征匹配为基础的下游任务上展现出强大的应用潜力,包括基于图像的大场景三维重建[7,58]、基于图像的定位[11,59]、视觉 SLAM[60]、多模态融合[22,61,62]等.尽管文献中提出的许多方法在不同方面促进了图像匹配技术的进步,已有的综述主要聚焦于总结稀疏特征点匹配中不同的特征点检测与特征点描述方法[47,48,63]、或者基于某个特定任务/数据集对不同的特征点检测与描述方法的组合进行性能对比分析[7,61,64,65],本文从稀疏特征点匹配与稠密点匹配两个角度,聚焦于近年来深度学习在这两个领域相关技术问题上的突破、对已有方法进行了深入总结与分析,并且对误匹配去除、典型的下游应用技术进展进行了详细介绍,给出了相关研究所涉及的数据集,力求给读者展现出图像匹配这一领域的全貌和最新的技术突破点,帮助进入该领域的研究人员快速掌握图像匹配的内涵、难点、关键技术与数据集资源等.

本文首先在第 2 节给出图像匹配问题的正式定义和研究难点;之后,将在第 3 节至第 5 节对近年来该领域的代表性方法进行综述,总结分析现有方法的特点、内在联系、发展历程,以及关键技术等,涵盖稀疏特征点匹配、稠密像素点匹配和错误点滤除三个方向,试图给读者一个关于该领域的发展、现状和关键技术的全面了解.另一方面,深度学习的出现也革新了许多计算机视觉问题的技术路线,如:基于局部图像特征的视觉词袋模型[29]在深度学习出现之

前占据了图像分类方法绝对的主导地位,而深度神经网络端到端学习能力使得图像分类这一典型的计算机视觉问题不再依赖于局部图像特征的设计[49,52],目标检测也不再依赖繁琐的特征工程[66,67].换句话说,随着深度学习技术的进步,图像特征匹配以及相关的局部图像特征提取的应用范畴也发生了较大变化,本文将在系统总结分析基于深度学习的图像匹配方法基础上,在第 6 节给出一些目前仍然极度依赖局部图像点对应关系的典型应用,并介绍其现状,第 7 节对图像匹配及其下游应用任务的研究中经常使用的数据集进行介绍.最后,值得指出的是,尽管深度学习技术在图像匹配问题的成功应用,使得图像匹配技术的发展取得了可喜的进步,实际应用需求的不断延申也对其提出了新的要求,本文第 8 节将对该领域的现有挑战与未来发展方向进行展望.

## 2 图像匹配问题定义与研究难点

给定两幅图像 $I_A$ 和 $I_B$,图像匹配任务的目的在于通过计算机算法,自动获得两幅图像之间的对应点集合$\{(P_A^i, P_B^i), i=1,2,\cdots,N\}$,其中 $P_A/P_B$ 分别表示图像 $I_A/I_B$ 中点的坐标值,$N$ 表示获取的匹配点个数.狭义地讲,一对图像匹配点指同一物理空间点在两幅不同图像中的位置,因此匹配图像是同一物理场景/物体在不同拍照条件(如:拍照的时间不同、拍照的相机姿态不同、拍照的相机传感器不同、拍照的相机参数不同等)下获取的图像,如图 2(b)(c)所示;更广义的图像匹配点还包含两幅图像中具有相同语义信息的对应点,这些点通常来自于不同的物理空间点,例如:不同汽车轮子上的点、不同鸭子嘴巴上的点,如图 2(a)所示.后者又称为语义匹配,主流解决方案为将图像特征点表示成图上的节点、点与点之间的关系表示成图上的边,使用图匹配技术得到图上节点的对应关系[68,69],近年来基于深度学习的图匹配技术也得到了显著进步[70,71];前者则是三维计算机视觉领域的核心,在摄像机标定、三维重建、相机定位、视觉 SLAM 等应用中均具有不可或缺的作用,本文后续提到的图像匹配均指前者较为狭义的定义,且相关应用也主要指显著依赖于这类同名物理点匹配的应用,为了与语义匹配相区分,也有文献将这种同名物理点的匹配称为几何匹配[72,73].

根据匹配对象是具有某种特性的点还是一般性的图像像素点,图像匹配可分为稀疏特征点匹配与稠密点匹配.稀疏特征点匹配包含特征点检测、特征点描述、匹配三个过程:特征点检测旨在从图像中检测得到若干点用于后续匹配,这些点可以在不同图像中被重复检测出,是特征点匹配的前提;特征点描述根据特征点周围的图像信息使用向量作为特征描述子对其进行表述,基本准则是期望不同图像中同一物理空间点对应的特征点具有距离相近的特征描述子,而不同物理空间点的对应特征点具有距离较远的特征描述子,是特征点匹配的核心;匹配则在特征描述子空间,以最近邻为基本原则建立两幅图像中特征点之间的对应关系,是特征点匹配的最终目的.与稀疏特征点匹配相比,稠密点匹配以图像中所有像素点为匹配对象,通过对所有像素点建特征描述子,在特征描述子空间基于最近邻基本原则进行匹配或者设计匹配层[2,74].这类稠密点匹配方法通常针对极几何关系已知或者经过粗匹配初步对齐的图像对展开,以将匹配的搜索空间限制在极线附近或者是范围较小的局部区域[75,76],搜索空间的减少不仅节省了匹配时间同时降低了对特征描述子区分力的要求.近年来研究较多的宽基线稠密匹配方法将稠密特征点学习与匹配统一至一个端到端深度网络中,在网络中使用 sinkhorn 算法构建匹配层[59]或者对输出的相似度矩阵采用松弛的互最近邻策略(dual-softmax)[55,77]计算相对匹配得分实现稠密点匹配,这种基于全局特征相似性的稠密匹配方法也常作为初始粗匹配用在由粗到细的稠密坐标变换场预测网络中[72,73,78-80],以从回归的角度实现宽基线图像稠密对应关系求解.

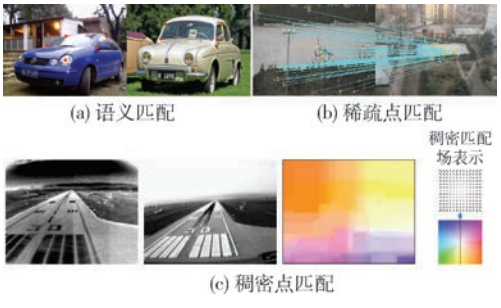



图 2 图像匹配问题分类

图像匹配的难点主要来自于两幅待匹配图像拍照条件不同引起的图像差异,主要可以分为以下三大类:

(1)图像光谱差异,由于图像拍照时的自然光照条件的不同以及人工灯光的影响(如图 3(a)所示)、

以及不同成像传感器特性的影响(如图 3(b)所示)、或者气候/季节不同导致物体景物外观的改变(如图 3(c))等,同一景物/目标会呈现出剧烈的光谱差异,使得同一物理点对应的局部区域图像呈现出巨大的色彩差异,如何提取出具有光谱变化鲁棒性的图像特征是图像匹配的主要挑战之一.

(2)图像内容几何变化,该类图像差异仍然主要来自于拍照条件的不同,如拍照时的焦距、距离物体的远近、拍照的相机姿态,这些拍照相机的内外参数差异使得拍照的物体在图像中会呈现出很大的几何差异(如图 3(d)、图 3(f)),即:相同内容在图像中呈现出涵盖区域的几何形状差异,如何在匹配中克服这种局部几何形状差异是图像匹配的另一个挑战.

(3)图像内容变化,在图像匹配相关的实际应用中,需要对间隔时间较长的图像进行匹配,长时间隔除了会因为季节的不同引起图像光谱变化,人工活动的影响也会改变图像中部分景物/目标,如桌椅的移动、建筑物装饰的改变(如图 3(f)所示),由于三维计算机视觉中的多视角几何学理论[5,6]是建立在静态目标的图像点对应关系基础上,这类动态移动物体上的匹配点会影响后续以多视角几何学为基础理论的下游应用任务、甚至导致错误结果,如:三维重建、视觉定位、图像拼接等,因此,图像匹配方法需要在正确匹配无变化物体的基础上去除变化物体上的匹配点.

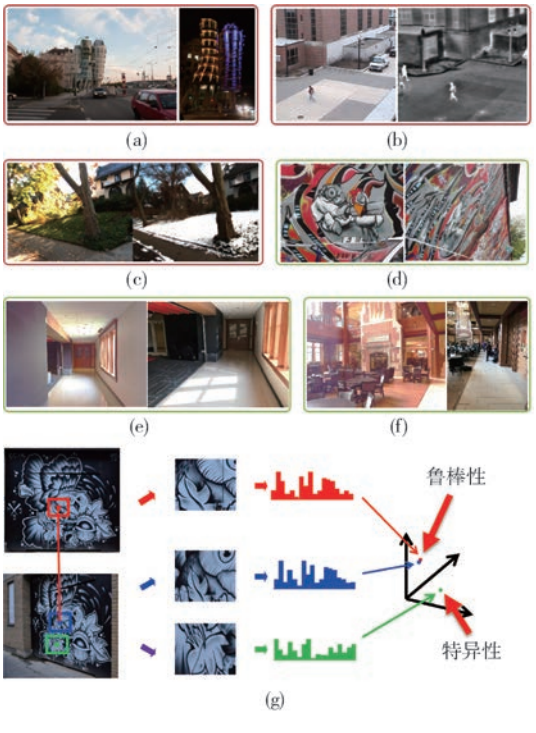



图 3　图像匹配难点

为了解决上述难点,图像匹配的根本挑战在于设计一个从图像到向量空间的映射,建立一个特征描述子空间使得匹配点对应的特征描述子距离近,同时不匹配点特征描述子之间的距离尽可能远,如图 3(g)所示.使得匹配点特征描述子距离近对应了图像匹配的鲁棒性需求,即:不同成像条件下的同一物理点应该具备相同或相近的特征描述子,这是匹配的基本前提.与此同时,使得不匹配点特征描述子之间的距离尽可能远对应了特征描述子的鉴别能力,即:不同物理点对应的图像特征描述子应该是各异的,这是消除混淆/错误匹配的保障.图像匹配归根结底要同时提升特征描述子的鲁棒性和特异性,从模式识别的角度,图像匹配是一个具有无限类别,类内差异小于类间差异的易区分情形、以及类内差异大于类间差异的难区分情形两者共存,并兼具部分长尾分布与零样本特性的复杂模式分类问题.

## 3　稀疏特征点匹配

### 3.1　基于深度学习的特征点检测

稀疏匹配的前提是进行图像特征点检测,通过计算图像的某种局部属性极值等操作得到可以在不同视角、光照条件下可重复检测得到的图像点,以获得稀疏点匹配的基本处理对象.然而,目前依然没有一种普遍适合的图像属性和相关理论可用于高重复性、强鲁棒特征点检测[63,81,82],研究人员转而开始研究基于学习的特征点检测方法,尤其是基于深度学习的方法.

Yarnick 等人[83]构造了一个涵盖白天-夜晚、不同气候、不同季节图像匹配的数据集,并在此基础上提出使用分段线性回归以及浅层 CNN(Convolutional Neural Networks)学习对不同拍照时间具有很强稳定性的特征点检测算法,可在白天和夜晚、不同季节、不同气候等不同时相图像匹配问题取得了较好的性能,但在匹配图像中同时存在视角等变化情况下的匹配性能依然欠佳.随后,Lenc 和 Vedaldi[84]对图像特征点的协变性进行分析、建模,将特征点检测问题转化为一个像素点至最近特征点偏移量的回归问题,并在"局部特征点应该在不同的仿射变换下都应该被检测到"(即:特征点在不同仿射图像变换下的高重复性)这一基本思想指导下设计了一种特征点协变误差,在此基础上,实现了一个类似 LeNet[85]的 7 层 CNN,利用神经网络强大的函数拟合能力学习

得到一个特征点位置回归器,最后通过投票的形式确定最终的特征点热度图进行特征点检测. Zhang 等人[86]在 Lenc 和 Vedaldi 的特征协变性分析的基础上,增加了关于标准特征点的正则项约束,解决了 Lenc 和 Vedaldi 的方法存在二义性解的问题,此外,他们采用的 5 层 CNN 进行特征点检测也更加高效. Savinov 等人[87]提出特征点的高重复性意味着不同点的相对排序应该保持不变,基于此提出了基于四元组排序的无监督特征点学习目标,但他们的方法仅在简单的三层神经网络以及单层线性神经网络上进行学习,未充分利用深度神经网络强大的表示能力.

受手工设计的仿射不变特征区域检测算法启发,Mishkin 等人[88]提出利用神经网络强大的区分能力学习具有判别性的局部仿射区域,在此区域上提取特征描述子进行匹配以实现大视角变化图像的鲁棒匹配. Barroso-Laguna 等人[53]将手工设计的特征点检测方法中常用的滤波器组与 CNN 通过数据驱动学习得到的卷积特征图进行组合,提出在规整化网格中通过特征点响应分数加权平均进行特征点检测,他们提出的 Key. Net 仅利用较少的参数获得了领先的特征点检测性能.

在特征点检测器的研究过程中,研究人员发现通过深度网络学习得到的特征描述子实际上已经具备了较好的鲁棒性和区分性,可在其基础上进行特征点检测. 基于这个思路,Tian 等人[89]仿照手工设计的特征点检测算法,提出 D2D 方法对预训练的 CNN 网络提取的特征响应图计算其信息熵与自区分性,检测信息熵高并且自区分能力强的点作为特征点. Brauch 和 Keller[90]直接使用特征描述子最大响应进行特征点检测,他们对最后一层特征响应图进行反向传播,利用多个通道特征中最大值进行传播直至输入图像,获取对应的特征点位置和响应分数,该方法检测得到的特征点在跨模态图像匹配中展现出了较高的重复性.

除了上述使用手工设计的方法对深度学习特征图进行处理,在特征描述子网络基础上进一步训练特征点检测器也展示出良好的性能,Barroso-Laguna 等人[91]提出的 HDD-Net 使用了 Key. Net[53]的特征点检测网络结构,通过在给定的稠密特征点描述子基础上最大化特征点匹配性能进行特征点检测器训练. Li 等人[92]提出的 PosFeat 在特征描述子基础上训练一个轻量化特征点检测器,通过基于匹配奖励的强化学习对特征点检测器进行训练.

总体而言,利用深度学习单独进行特征点检测网络训练的工作较少,而且这类方法的网络结构设计比较简单以保证计算的高效性以及对任意尺寸图像的适应性,现有工作中以浅层的全卷积 CNN 为主并融合了部分手工设计的卷积滤波器. 在学习方法方面,以有监督/自监督学习方法为主,因此需要依赖标注的特征点信息或者从图像中生成的高重复性点信息进行学习,这使得目前学习得到特征点检测器存在一定的偏好性和局限性,如:SuperPoint[93]利用合成角点图像学习得到的特征点检测器在仿射变换图像中对重复度高的点进行标注用于训练,其输出的检测结果倾向于图像中的角点,对于其他类型的特征点检测效果欠佳. 现有的无监督特征点学习方法[87,89,94]在性能上距离有监督方法仍有一定差距.

值得一提的是,近年来该领域逐渐发展为结合特征点描述子同时学习特征点检测器,通过将图像匹配的这两个核心模块同时学习,相比单独的特征点检测器学习获得了重复率更高、更易匹配的特征点,这部分工作的研究进展与关键技术将在 3.3 节中详细介绍.

### 3.2 基于深度学习的特征点描述

2015 年 CVPR 会议上最早出现了利用深度学习进行特征点描述和匹配的工作,Han 等人提出的 MatchNet[95]以及 Zagoruyko 和 Komodakis 提出的 DeepCompare[96]针对局部匹配图像验证任务,利用孪生网络(Siamese Network)结构,通过参数共享的两个深度网络学习得到一对局部图像块对应的特征描述子,并将其作为后续度量网络的输入,在标准数据集(UBC Patches)[44]上取得了优越的性能. 然而,研究人员很快便发现这类方法提取的特征描述子如果在欧式空间使用最近邻匹配策略并不能取得令人满意的效果,需要搭配对应的度量网络将特征匹配问题转化为二分类才取得到较好的性能,而在许多依赖图像匹配的实际应用中,基于特征欧式空间的最近邻匹配具有不可替代的地位,如在大规模特征匹配问题中经常使用的快速最近邻匹配方法(KDTree[97]、FLANN[98]),这一缺点极大地限制了这类方法在图像匹配相关应用中的使用.

注意到此缺陷之后,研究人员开始关注使用深度学习在欧式空间直接学习特征描述子,早期的方法有 Simo-Serra 等人提出的 DeepDesc[99]和 Balntas 等人提出的 TFeat[100]. DeepDesc 提出使用 5 层卷积的 CNN 提取 128 维特征描述子,通过同时最小化

匹配局部图像对的描述子距离和最大化非匹配局部图像对的描述子距离进行 CNN 网络参数的学习. 此外,DeepDesc 提出了一种困难负样本挖掘策略来应对局部图像特征匹配训练数据集中匹配/非匹配样本极度不平衡问题,在每个数据批(batch)训练的时候,仅使用误差最大的一部分非匹配样本对应的误差梯度进行网络的参数更新. TFeat 提出仅包含 3 个卷积层的轻量化网络用于特征点描述,使用三元组合页损失同时优化匹配样本和不匹配样本之间的距离差来进行 CNN 网络的训练,同时将困难样本挖掘的思路应用于三元组合页损失:计算该损失时使用三元组样本中较难区分的一对非匹配样本. 尽管这些方法可直接替代传统的特征描述子在三维重建等应用中直接使用,也在 UBC Patches 数据集中的局部匹配图像验证任务上取得了不错的精度,但在该任务上与之前的结合度量学习网络的深度学习方法仍有一定差距. 另一方面,这两个直接在欧式空间学习特征描述子的早期工作均采用了困难样本挖掘技术,该思想对深度学习特征点描述领域的发展具有深远影响,目前该领域的主流方法大多采用了后续提出的困难三元组挖掘技术[101]或者其变种[58,102,103].

Tian 等人[54]提出的 L2Net 首次在局部匹配图像验证任务上(UBC Patches 数据集[44])展现了深度学习特征描述子在欧式空间可具备超越度量空间的匹配性能,并在 2016 年 ECCV 会议上举办的特征描述子竞赛中取了图像匹配、检索、验证三个任务的第一①(使用 2016 年 ECCV 会议新发布的 HPatches 数据集和评测标准[47]). 如图 4(a)所示,L2Net 采用一个 7 层全卷积的网络结构,输入为 32×32 的局部图像,前 6 个卷积层的卷积核大小为 3×3,其中第 2、4 层卷积步长设置为 2 以减少特征图的大小,使得提取的特征图减小为 8×8,最后通过一个 8×8 的卷积输出得到 128 维向量作为输入局部图像的特征描述子. 与之前方法不同的是,L2Net 使用 Local Response Normalization 对输出层进行模长归一化,使得输出的 128 维特征描述子规范化为单位模长,其实验结果表明该方法能有效减少特征描述子的学习难度. 在训练误差方面,L2Net 通过优化结构化损失函数使得一对匹配样本的特征描述子距离比同一个数据批内所有的不匹配样本的描述子距离均要小,同时将该结构化损失推广至网络的中间层特征以此提升最终网络提取的特征描述子的匹配性能.

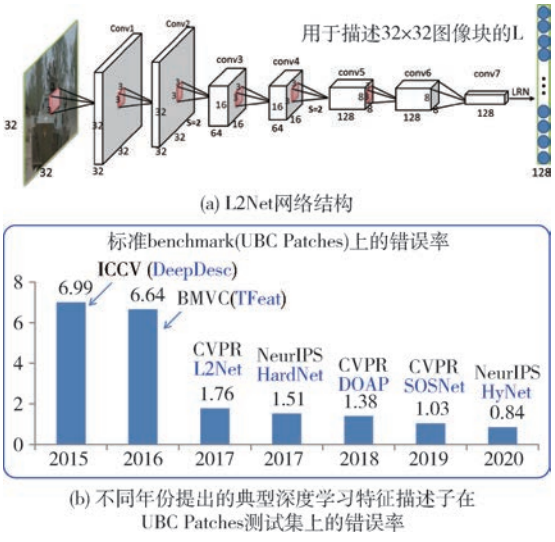



图 4　基于 CNN 的典型特征描述子网络及性能

L2Net 的优异性能和相对轻量化的网络结构,很快便引起了领域内研究人员的广泛关注,并在 L2Net 网络结构基础上提出了许多优秀的特征描述子. Mishchuk 等人[101]提出的 HardNet 采用了 L2Net 的网络结构,并提出基于最难负样本的三元组合页损失进行网络参数的学习,即:HardNet 在 L2Net 网络训练的时候只考虑优化一批训练数据里正样本以及最难区分的负样本之间的相对距离,其原理如图 5(a)所示. Ebel 等人[104]提出使用极坐标采样生成待描述局部区域,并将其输入 HardNet 得到了更好的匹配性能. Wang 等人[103]提出对一个训练数据批内所有样本同时挖掘困难的匹配样本和不匹配样本进行深度网络学习,同时对三元组损失与平方和三元组损失的数学特性进行了分析,认为平方和三元组损失更适合特征描述子学习. 以上方法均在 L2Net 网络结构上,基于机器学习中最基本的三元组合页损失,通过改善学习目标中的困难样本挖掘方法以训练得到更好的 L2Net 网络用于特征点描述,图 5(a)概括对比了特征描述子发展历程中出现的困难样本挖掘技术的原理.

在特征描述子学习目标方面,He 等人[105]认为三元组合页损失仅直接考虑了三个样本相互之间的描述子距离特征,而在基于特征描述子的匹配和检索任务中,经常会遇见多个匹配样本的情况,因此他们提出最大化平均准确率的特征描述子学习目标,对所有样本相互之间的距离特征都进行了建模,学习得到的特征描述子称为 DOAP(Descriptor Optimized

① http://icvl.ee.ic.ac.uk/DescrWorkshop/#Challenge

by Average Precision). Zhang 和 Rusinkiewicz[102] 认为不同的三元组应该使用不同的正负样本间隔以解决三元组合页损失中固定间隔带来的后期学习乏力问题,提出的 DSM(Dynamic Soft Margin)方法可构建动态边界三元组距离损失进行 L2Net 网络参数的学习.

除了直接优化匹配/不匹配样本之间的描述子距离特性,有研究人员提出对描述子特征空间的分布特性进行优化,以获得性能更好的特征描述子. Zhang 等人[106] 提出 Spread out 正则项,期望通过深度网络学习得到的 128 维特征描述子均匀分布在一个 128 维的单位球面上. Tian 等人[107] 提出特征描述子学习的二阶距离约束,其思想为任意两个样本之间的距离应该与它们对应匹配样本之间的距离保持相同,他们使用该二阶距离约束结合平方和三元组合页损失,在 UBC Patches 数据集上取得了非常不错的性能. Tian 等人[58] 比较系统地分析了特征描述子欧式距离和余弦相似性在 HardNet 的三元组合页损失中对网络梯度计算的影响,提出 HyNet 对两者进行了有机结合,取得了 UBC Patches 数据集上目前公开报道出来最好的局部匹配图像验证精度. 图 5(b)展示了一些文献中常用的特征描述子学习目标.

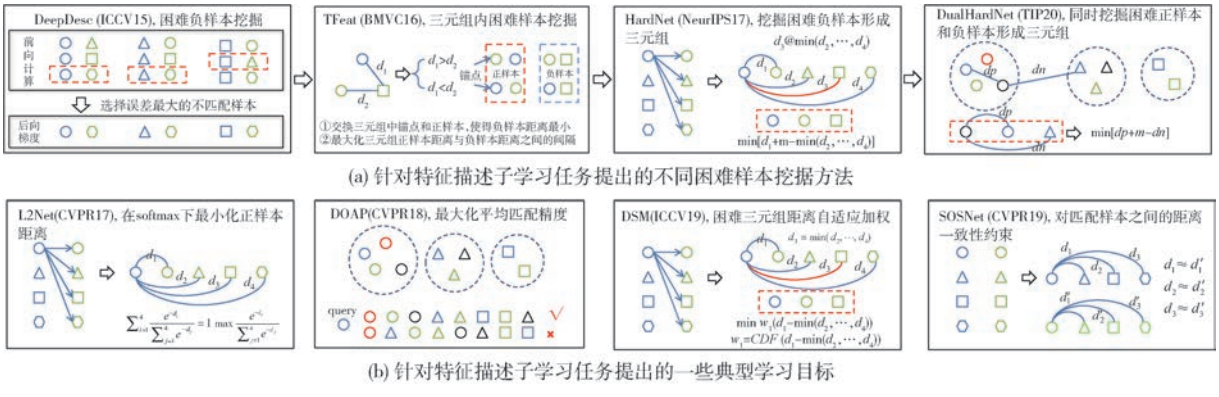



图 5 基于深度学习的特征点描述子方法概述

上述方法均采用了 L2Net 网络结构进行特征描述子提取,图 4 下部展示了历年来在 UBC Patches 数据集上错误率最低的深度学习特征描述子,可以看出,L2Net 在该领域占据了主导地位,自它在 2017 年提出以后历年在该数据集上的最好性能方法均采用了 L2Net 作为网络结构. 此外,从 2020 年和 2021 年报道的最好结果来看,该数据集基本进入了饱和状态,因此在近两年使用该数据集进行深度学习特征描述子方法的性能评测较少,目前该数据集的使用主要集中在对无监督特征描述子学习方法的性能评测,如 GraphBit[108]、UDBD[109]、BLCD[110]、D-GraphBit[111] 等.

在上述基于 L2Net 的方法中,研究人员发现对得到的浮点特征点描述子直接根据符号量化获得的二值描述子依然可以取得非常好的匹配性能. 最早进行这方面尝试的工作是 L2Net,而之后提出的 DOAP 和 DSM 都有对应的二进制版本. Fan 等人[110] 从二进制描述子学习所期望的一般特性出发,针对 L2Net 这一具体的网络结构,分析了已有方法通过学习浮点特征描述子再直接进行二值化的合理性.

为了获得更有难度的局部图像匹配对进行特征描述子训练,Luo 等人[112] 提出 GeoDesc 摒弃了 UBC Patches 数据集提供的用于训练的局部图像匹配对,而是利用 SfM(Structure from Motion)算法重建得到的相机姿态、匹配图像之间的极几何关系,以及 SfM 重建留下的 SIFT 对应点生成了更难区分的匹配局部图像块用于 L2Net 网络的训练,此外还在 HardNet 的网络学习目标中增加了一项根据匹配图像块射影畸变程度自适应调整的正样本距离损失. 除了 L2Net,深度学习领域广泛采用的 ResNet[52] 和 VGG[49] 也被用于特征描述子提取网络. Wang 等人[113] 使用 ResNet-50 进行图像特征提取,并在其基础上根据 SfM 重建得到的结果提出了基于极几何约束的弱监督特征点匹配学习目标,在图像匹配以及与之相关的多个下游任务中都取得了较好的性能. Germain 等人[114] 提出一种“稀疏-稠密”的非对称稀疏特征点匹配方式,仅需要对一幅图像提取特征点计算它与另一幅图像所有像素位置的特征描述子进行匹配,他们提出的 S2DNet 使用 VGG 作为骨

干网络进行特征描述子计算,通过“稀疏-稠密”匹配方式将特征点匹配转化为分类问题、使用交叉熵损失对特征描述子网络 VGG 进行训练,得益于这种非对称匹配方式,S2DNet 得到的匹配点具有很高的位置精度.

在一些挑战场景如重复纹理、弱纹理等,由于局部图像包含的信息量有限并且具有二义性,利用更大范围的信息增强特征描述子已成为一种共识. Luo 等人[115]利用 ResNet-50 以及特征点的位置分布将半全局图像信息和全局几何信息融入 L2Net 所刻画的局部信息,得到的 ContextDesc 综合了单幅图像里更大范围的图像信息以及特征点的分布信息,取得了很好的图像匹配性能. Sarlin 等人[59]提出 SuperGlue 在 SuperPoint 特征点基础上构建关系图,利用图注意力机制(多头注意力[50])对图像内部特征点的自注意力和图像间的特征点交叉注意力进行建模,学习得到具有全局感受野和跨图像上下文敏感的特征描述子,既有效建模了同一图像中不同点之间的关系,又建模了两幅图像中不同点之间的关联性,通过对这种增强的特征描述子进行二分图匹配可以得到高质量特征点匹配,在许多挑战场景下的图像匹配中均展现出优异的性能. 基于 SuperGlue 框架,研究人员进一步通过改进特征之间的注意交互方式和注意力网络,进一步提升了 SuperGlue 的特征匹配性能和效率[116-118]. 相比领域内一些基于局部图像块的特征点描述网络,基于大范围上下文信息增强的特征描述子需要在整幅图像上进行操作,这既给特征点描述子学习相关的研究提供了新思路,也是 UBC Patches、HPatches 这类仅基于局部图像信息的数据集在该领域逐渐被淘汰的原因之一,即:研究人员越来越关注从原始的图像中获取特征点的描述子信息,这也符合特征点描述子的实际应用场景.

### 3.3　兼具特征点检测与描述的深度学习方法

尽管基于深度学习的特征描述子极大提升了特征描述子的区分能力,并在图像匹配相关的标准数据集上(如:Oxford VGG[46]、UBC Patches[44]、HPatches[47])展示了对于各种复杂图像变化的稳定性,然而,在特征点匹配相关的下游视觉三维感知任务中(如:三维重建、视觉定位),多个独立的评测性工作均表明,通过组合不同的特征点检测和描述方法依然存在很大的局限性[7,47,65].

由于特征点匹配任务需要配合特征点检测和特征描述共同完成,为了更好地适应特征点匹配的下游应用任务,研究人员逐渐开始关注如何利用深度学习同时进行特征点检测和描述子提取. 最早提出的同时具有特征点检测和描述功能的 CNN 网络是 LIFT(Learned Invariant Feature Transform)[119],其流程模仿经典的 SIFT 方法,在整个网络结构中设计了基于 CNN 的特征点检测模块、主方向估计模块、特征区域提取与规范化模块、以及规范化区域的特征描述模块,但是该方法训练需要分步骤进行,并不能联合优化特征点和描述子的匹配性能,并且训练难度大、计算复杂度高,类似的方法还有 LF-Net[120]、RF-Net[121]. 这些早期探索主要思路仍然局限于传统特征点检测和描述子构建的框架,未能完全发挥深度神经网络端到端学习特性,而且在特征点描述子的匹配性能及其可被检测性两方面的联合考虑不充分,导致它们在图像匹配的下游任务中表现欠佳.

近年来,研究人员越来越关注使用同一个神经网络骨干结构进行特征提取,进而在提取的特征图上分别进行特征点检测和特征点描述,并且在网络训练时联合优化关于特征点检测的学习目标以及特征描述子的学习目标,其基本流程如图 6 所示.

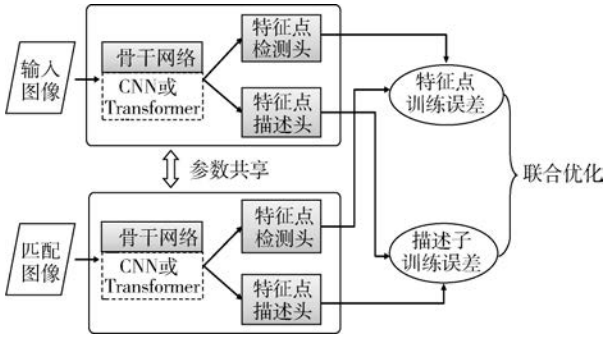



图 6　兼具特征点检测与描述的深度学习方法的基本结构

在这类方法中,特征描述子几乎都是通过对骨干网络提取得特征图进行插值或者多尺度融合获得,主要区别点在于特征点检测方式以及学习优化目标. 在特征点检测方式方面,目前常用的范式有分类和回归两大类.

由于骨干网络提取的特征图空间尺寸通常小于原始图像的分辨率,其上每个点对应了原始图像中一个局部区域,分类的特征点检测范式利用特征图上的特征进行分类,预测原始图像局部区域中具体哪个位置为特征点或者该区域无特征点,SuperPoint[93]是这类方法中的典型代表. 由于将特征点检测转化为

分类问题,因此这类方法需要给定训练图像中的特征点标注信息,目前采用较多的方法是 SuperPoint 提出的单应变换适应技术,其基于一个给定的特征点检测器(如:SuperPoint 使用的角点检测器),对自然图像在多次变换中均能稳定检测的图像点自动标注为特征点,用于特征点检测头的训练.在这种特征点检测范式下,采用的特征点训练误差通常为交叉熵损失,考虑到特征点检测是一个类别极度不平衡的问题,即:特征点样本(正样本)个数显著少于非特征点样本(负样本),因此会利用一个较大的权值加大正样本交叉熵项对总体损失的影响.

基于回归的特征点检测范式则保留了传统特征点检测的思路,其对输入图像利用深度网络计算得到一个特征点置信度图,通过在置信度图上进行求局部极大值得到特征点,该技术路线的主要挑战在于解决求取局部极大值的不可导问题.针对该问题,Dusmanu 等人[122]提出了 D2-Net 方法,这也是该领域第一个可端到端学习的联合特征点检测与描述网络,D2-Net 使用 soft-nms 在空间维度计算不同通道特征图的极值点,并针对每个通道特征计算其与通道最大值之比,这两者相乘作为每个网格点的特征点似然分数.具体地,D2-Net 采用 VGG 网络的第 conv4_3 层输出特征图进行特征点置信度计算,特征图上的每个网格点通过公式(1)分别计算它在局部空间区域的"峰值性"$\alpha$ 和在特征通道维度上的"峰值性"$\beta$:

$$\alpha_{ij}^{k}=\frac{\exp(D_{ij}^{k})}{\sum_{(i',j')\in N(i,j)}\exp(D_{i'j'}^{k})}$$
$$\beta_{ij}^{k}=D_{ij}^{k}/\max_{t}(D_{ij}^{t}) \tag{1}$$

其中,$(i,j)$是特征图上的空间位置索引,$k$ 表示特征通道索引,$N(i,j)$表示空间位置$(i,j)$的局部邻域(通常取 3×3 区域).如公式(2),每个网格点在空间区域的"峰值性"与特征通道上的"峰值性"相乘即得到了该点作为特征点的似然:

$$\gamma_{ij}=\max_{k}(\alpha_{ij}^{k}\beta_{ij}^{k}) \tag{2}$$

由于 VGG 网络的第 conv4_3 层输出特征图空间分辨率只有原始输入图像的 1/8,即使 D2-Net 在网络应用阶段采用修改最后一层卷积步长的方式将特征图空间分辨率提升到了输入图像的 1/4,其得到的特征点空间位置精度仍然较低,导致它在一些对匹配位置精度要求高的应用中表现欠佳.

Luo 等人[57]在提出的 ASLFeat 方法中,使用 L2Net 不同层次的特征图进行特征点似然计算,并将它们映射至原始图像分辨率以提高获得的特征点空间位置精度,相比公式(1)中定义的空间位置和特征通道"峰值性",ASLFeat 提出了另一种同样可导的计算方法,如公式(3)所示:

$$\alpha_{ij}^{k}=softplus\left(D_{ij}^{k}-\frac{1}{|N(i,j)|}\sum_{(i',j')\in N(i,j)}D_{ij}^{k}\right)$$
$$\beta_{ij}^{k}=softplus\left(D_{ij}^{k}-\frac{1}{C}\sum_{t=1}^{C}D_{ij}^{t}\right) \tag{3}$$

基于上述峰值性定义代入公式(2)即可计算得到对应特征图上的特征点似然.ASLFeat 通过似然图上采样和加权融合的方式,将原始分辨率、1/2 分辨率、1/4 分辨率三个不同分辨率特征图上计算得到的特征点似然进行融合,得到最终特征点检测置信度.上述方法均需要对提取的特征图计算峰值性以获得对应位置处特征点的置信度,文献中也出现了直接通过神经网络预测特征点似然的方法,典型的方法有 DISK[123]、R2D2[124]和 PosFeat[92],它们直接使用卷积层对骨干网络提取的特征响应图进行处理以输出对应位置处特征点似然.

除了上述基于局部极值和直接回归的特征点似然计算方式,还有直接回归特征点坐标的方法.一类主流方法是利用 softmax 操作对局部窗口内的检测分数进行归一化操作,之后使用归一化分数对局部窗口内位置进行加权求和得到该窗口内特征点位置,该加权求和的特征点位置可通过公式(4)计算获得:

$$(\Delta x,\Delta y)=\sum_{(i,j)\in N(x,y)}s_{ij}[i,j]^{\mathrm{T}} \tag{4}$$

其中,$(x,y)$为局部窗口左上角图像坐标,$N(x,y)$表示对应的局部窗口区域,$(i,j)$表示局部窗口内的偏移量,$s_{ij}$表示经过归一化后的检测分数,$(\Delta x,\Delta y)$是特征点相对于窗口左上角的偏移量,即特征点坐标为$((\Delta x+x,\Delta y+y))$.这种基于回归的特征点检测方法最早在 Key.Net[53]中提出,其优点在于可微性,因此可以直接嵌入至整个网络学习过程直接最大化特征点的重复率,ALIKE[125]使用该特征点检测范式,结合神经投影误差对深度网络的特征点检测和描述性能进行联合优化.使用卷积层直接输出位置偏移量的方式在 UnsuperPoint[94]中使用,为了避免网络收敛至奇异解并且得到具有均匀分布的特征点,需要结合一种基于数值排序的平均位置分布

约束进行学习.

在特征描述子优化目标方面,这类方法基本沿用特征描述子学习领域的损失函数,常用的有三元组合页损失[122]和三元组对比损失[57],如公式(5)和公式(6)所示:

$$L_h = \sum_{i=1}^{N} \max\{p_i + m - n_i, 0\} \tag{5}$$

$$L_c = \sum_{i=1}^{N} (\max\{p_i - m_p, 0\} + \max\{m_n - n_i, 0\}) \tag{6}$$

在网络学习策略方面,如公式(7)所示的特征点检测与特征描述子联合优化方法将特征点检测与描述子性能进行耦合,

$$L_{\text{联合}} = \sum_{i \in C} \frac{s_i^{(1)} s_i^{(2)}}{\sum_{j \in C} s_j^{(1)} s_j^{(2)}} L_{\text{描述子}} \tag{7}$$

其中 $s_i^{(1)}, s_i^{(2)}$ 表示第 $i$ 对真实匹配点在两幅图像上的特征点检测似然值,$L_{\text{描述子}}$ 是特征描述子损失.

该联合误差的物理意义在于它为每一对标注的图像对应点计算了特征描述子匹配损失,同时根据特征点检测的似然对描述子损失进行加权,由于网络的学习目标是最小化联合损失,因此会驱使网络在特征描述子匹配误差较小的对应点位置处输出较大的特征点检测似然,类似地,也会在特征点检测似然值较大的位置驱使相应的特征描述子具有更小的匹配误差,因此该学习误差具有联合优化特征点和描述子的功能. D2-Net[122]采用了公式(7)进行联合优化,在最大化匹配特征点置信度的同时最小化三元组合页损失,而 ASLFeat[57]则在相同的联合优化训练框架下最小化三元组对比损失. Fan 等人[126,127]在 ASLFeat 联合训练的基础上引入辅助损失项以提升特征图的鲁棒性,例如,ASLFeat-GAN[126]将对抗学习引入 ASLFeat 的训练过程,采用额外的白天/黑夜图像判别器以促使原有的 ASLFeat 网络学习得到具有白天/黑夜光谱不变性的局部特征用于图像匹配,有效提升了 ASLFeat 特征的鲁棒性、尤其是在夜晚图像的视觉定位任务上取得了更好的精度. SeLF[127]在学习局部图像特征时融入语义特征信息,通过在 ASLFeat 网络学习过程中增加一个语义蒸馏分支,使得 ASLFeat 学习得到的特征图具备一定的语义推断能力,有效提高了所学特征对于复杂光谱变化的鲁棒性.

文献中另一种较多采用的联合优化误差为如公式(8)所示的多个误差项直接相加方式,代表性的方法有 SuperPoint[93]、MTLDesc[128]、ALIKE[125]等,

$$L_{\text{联合}} = L_{\text{特征点}} + \lambda L_{\text{描述子}} \tag{8}$$

公式(7)和公式(8)以显式的方式对特征点检测及描述子性能优化,强化学习则通过设计基于匹配或者下游任务相关的激励隐式地对特征点与描述子的联合性能进行优化. Tyszkiewicz 等人[123]提出使用强化学习策略通过最大化正确匹配奖励对基于 U-Net[129]的特征点检测器和描述子进行学习. Bhowmik 等人[130]使用强化学习策略将特征点匹配任务与具体的下游任务(如:相对位姿估计)相结合,得到与任务适配的 CNN 特征点检测器和描述子.

表 1 展示了典型的同时进行特征点检测与描述的深度网络及训练方法,并在图像匹配和视觉定位两个任务上对现有方法的性能进行了对比,具体的评价指标(即:图像匹配使用的 MMA、视觉定位使用的定位正确率)定义详见第 7 节.

**表 1　同时进行特征点检测与描述的深度学习采用的网络结构、优化学习目标及性能对比**

| 方法出处 | 基础网络 | 学习目标 | 监督信息 | 室内定位 | | 视外定位 | 图像匹配 |
|---|---|---|---|---|---|---|---|
| | | | | DUC1 | DUC2 | | |
| LF-Net[120] NeurIPS'18 | ResNet style | 特征点:最大化特征点分数一致性+匹配特征点描述子相似性<br>描述子:三元组合页损失<br>联合训练:否 | 图像匹配点 | 23.7<br>36.9<br>52.0 | 14.5<br>22.9<br>33.6 | 58.3<br>73.2<br>84.7 | 55.6 |
| SuperPoint[93] CVPRW'18 | VGG style | 特征点:交叉熵损失<br>描述子:加权三元组对比损失<br>联合训练:是 | 图像匹配点+特征点位置 | 39.9<br>55.6<br>67.2 | 37.4<br>57.3<br>70.2 | 68.0<br>85.1<br>95.4 | 65.37 |
| D2-Net[122] CVPR'19 | VGG | 特征点:最大化特征描述子鉴别力强位置处的点检测概率<br>描述子:最小化高置信度对应点的三元组合页损失<br>联合训练:是 | 图像匹配点 | 39.9<br>57.6<br>67.2 | 36.6<br>53.4<br>61.8 | 68.8<br>85.4<br>96.4 | 37.29 |

（续表）

| 方法出处 | 基础网络 | 学习目标 | 监督信息 | 室内定位 | | 视外定位 | 图像匹配 |
|---|---|---|---|---|---|---|---|
| | | | | DUC1 | DUC2 | | |
| R2D2[124] NeurIPS'19 | L2Net | 特征点：最大化检测分数一致性＋局部峰值响应性<br>描述子：最大化高置信度对应点的平均匹配精度<br>联合训练：是 | 图像匹配点 | 38.9<br>67.2<br>75.8 | 43.5<br>61.8<br>71.0 | 65.6<br>81.7<br>95.1 | 68.6 |
| ASLFeat[57] CVPR'20 | L2Net＋DCN | 特征点：最大化鉴别力强的特征描述子被检测概率<br>描述子：最小化高置信度对应点的三元组对比损失<br>联合训练：是 | 图像匹配点 | 37.4<br>56.1<br>69.7 | 37.4<br>58.8<br>64.9 | 70.4<br>85.2<br>96.3 | 72.26 |
| ASLFeat-GAN[126] IEEE TMM'22 | L2Net＋DCN | 特征点：同 ASLFeat＋特征图对抗学习<br>描述子：同 ASLFeat＋对抗学习<br>联合训练：是 | 图像匹配点＋白天/黑夜图像 | | | 71.3<br>86.5<br>97.3 | 68.6 |
| SeLF[127] IEEE TIP'22 | L2Net＋DCN | 特征点：同 ASLFeat＋特征图蒸馏学习<br>描述子：同 ASLFeat＋蒸馏学习<br>联合训练：是 | 图像匹配点＋图像语义分割网络 | 41.4<br>61.6<br>73.2 | 44.3<br>61.1<br>68.7 | 75.4<br>86.8<br>97.6 | — |
| DISK[123] NeurIPS'20 | UNet style | 特征点：最大特征匹配奖励<br>描述子：最大特征匹配奖励<br>联合训练：是 | 图像匹配点 | 34.8<br>57.1<br>72.7 | 29.8<br>46.6<br>60.3 | 73.8<br>86.2<br>97.4 | 77.59 |
| ALIKE[125] IEEE TMM'22 | Transformer | 特征点：特征点重投影误差＋特征点峰值性<br>描述子：神经重投影误差<br>联合训练：是 | 图像匹配点 | | | | 70.5 |
| MTLDesc[128] AAAI'22 | ResNet | 特征点：交叉熵损失<br>描述子：注意力加权三元组合页损失<br>联合训练：是 | 图像匹配点＋特征点位置 | 41.9<br>61.6<br>72.2 | 45.0<br>61.1<br>70.2 | 74.3<br>86.9<br>96.9 | 78.66 |
| PosFeat[92] CVPR'22 | ResNet style | 特征点：最大特征匹配奖励<br>描述子：最小匹配特征的极线误差<br>联合训练：否 | 图像极几何关系 | | | 73.8<br>87.4<br>98.4 | 76.43 |

注：室内定位指在 InLoc[131] 数据集上（DUC1、DUC2 两个子集）使用 HLoc 定位框架[132]、在 3 个不同定位误差（(0.25m，2°)、(0.5m，5°)、(1m，10°)）下的定位正确率（正确定位图像/所有定位图像，单位%），室外定位指在 Aachen Day-Night v1.1[10] 数据集上使用官方为评测不同图像特征性能而提供的定位框架[133]、在 3 个不同定位误差（(0.25m，2°)、(0.5m，5°)、(5m，10°)）下的定位正确率（单位%），图像匹配指在 HPatches[47] 数据集上匹配误差为 3 像素时的平均匹配正确率（MMA，Mean Matching Accuracy，单位%）. LF-Net、DISK、R2D2、SuperPoint、D2-Net、ASLFeat 定位结果来自论文 SeLF[127]，其余方法的定位结果来自于原始论文；LF-Net、D2-Net、SuperPoint、DISK 匹配结果来自论文 ALIKE[125]，其余方法的匹配结果来自于原始论文，空白表示文献中未报道相应结果.

## 4 稠密点匹配

稀疏特征点匹配的质量很大程度依赖于特征点检测的结果，一方面，如果在同一场景不同图像中检测得到的特征点重复性较低，那么即使特征点描述子具有很好的区分能力，也很难得到足够多的匹配特征点；另一方面，在待匹配图像特征点重复性较低的情况下，其对特征描述子区分能力的要求也更高，因为存在很多干扰点. 在图像匹配领域，与稀疏特征点匹配对应的是稠密像素点匹配. 稠密特征点匹配跳过特征点检测阶段，将所有像素点都认为是可匹配的点，通过比较像素点特征描述子的欧式距离建立匹配对应关系. 由于所有像素都需要进行匹配，如果不对待匹配区域进行约束，那么匹配的搜索空间非常巨大，不仅增加计算量，而且对特征描述子的区分能力也是一个大的挑战. 因此，早期稠密像素点匹配通常用在立体匹配[2,134,135] 以及具有窄基线特性的光流估计[72,136] 中，在立体匹配中，待匹配图像的极几何关系已知或者初步对齐，因此匹配主要限制在对应的极线附近或者较小的局部窗口范围，在光流估计中则采用由粗到细的匹配策略对匹配范围进行限制.

Yao 等人[137] 结合传统多视角立体匹配的流程，将其使用端到端可学习的深度网络进行实现以此提出了 MVSNet 方法，该方法利用相机的内外参数将

多视角图像像素进行对应,计算不同深度平面上的像素 CNN 特征匹配代价并对不同视角的映射结果进行聚合.由于 MVSNet 采用 3D 卷积进行匹配代价聚合,复杂度高,难以处理高分辨率图像,为此,MVSNet 的作者进一步提出 R-MVSNet[138],采用循环神经网络渐进式的方法在不同深度平面的代价体上进行 2D 卷积,极大地减少了 MVSNet 的内存需求,实现了高分辨率多视角图像的深度计算.MVSNet 开启了利用深度学习进行端到端稠密像素三维重建的先河,引起了三维视觉领域内研究人员的广泛重视,产生了许多跟进研究,例如:Point-MVSNet[139]、Cascade-MVSNet[140]、EPP-MVSNet[141].这些工作在特征匹配时几乎都是采用了一般性的 CNN 特征,主要创新工作集中在如何对同一深度平面上的多视角匹配代价进行聚合、以及网络的结构设计和学习训练等方面.

上述立体匹配问题需要已知待匹配图像之间的极几何关系,深度学习在立体匹配问题上的成功及其强大的特征表达能力和端到端学习方式,进一步激发了研究人员利用端到端 CNN 网络解决一般性的稠密像素点匹配(宽基线立体匹配)问题的研究热情.根据是否显式计算最相似特征描述子进行匹配,已有的稠密匹配方法可以分为两类:基于特征描述子匹配的方法与基于稠密坐标变换回归的方法.基于稠密坐标变换回归的方法将稠密匹配问题建模成已知查询图像 $I$ 和匹配图像 $J$、求解 $I$ 相对 $J$ 的坐标变换场问题,即:

$$F_{I\to J}(x)\in R^{W\times H\times 2}:I(x)=J(x+F_{I\to J}(x))$$

其中 $x$ 表示图像像素坐标.

这类方法的特点在于根据最底层特征进行全局匹配实现查询图像与匹配图像的粗对齐,再构建由粗到细的网络结构不断精细化粗匹配得到的坐标变换场.Melekhov 等人[78]提出 DGC-Net 使用 CNN 特征对降采样后图像进行全局粗匹配,之后逐层使用上一层的坐标变换场与两幅图像的粗对齐特征进行叠加、构建卷积神经网络学习局部特征关系以输出更精细的稠密坐标变换场,在高层精细化稠密匹配过程中不显式计算图像特征之间的相似度.受 DGC-Net 结构启发,Truong 等人[72]提出一种全局粗匹配与局部精细化相结合的稠密坐标场计算网络,实现对高分辨率图像宽基线光流估计、宽基线几何/语义稠密特征匹配的统一处理;他们提出的 GLU-Net 将输入图像降采样至预定的低分辨率进行全局稠密匹配,之后结合低分辨率全局匹配结果在特征金字塔更精细尺度上计算局部特征相关性,通过逐层解码得到精细化的稠密坐标对应点,在稠密坐标变换场计算中融入局部特征相关性计算有效提升了 GLU-Net 对多种稠密特征匹配任务的统一处理性能.上述稠密坐标变换场估计网络由于自身复杂度限制只能在粗匹配结果上基于局部相关性进行精细对应计算,依赖局部相关性使得这类方法在解决重复性纹理图像等全局二义性匹配问题中性能欠佳,Truong 等人[79]进一步提出融入全局优化的全局相关稠密匹配层,在网络训练优化过程中可有效考虑多种不同的匹配先验约束(如:唯一性、平滑性).此外,Truong 等人[80]还在基于特征相关性的稠密匹配框架中增加匹配不确定性解码器,以实现稠密匹配对应和置信度估计,置信度的引入使得匹配下游任务中的模型拟合算法可以克服低置信度匹配对拟合模型的负面影响.Truong 等人[73]将一对图像中的任意一幅根据随机选择的变换生成新图像组成图像三元组,在三元组中构建变换一致性误差函数实现稠密对应估计网络的无监督学习,是上述稠密变换场估计网络结构的重要补充.近期,Edstedt 等人[142]提出的 DKM 在图像匹配以及姿态估计、视觉定位等多个下游任务上均取得了领先的性能,DKM 使用核函数与高斯过程回归提高粗匹配精度,在坐标变换场逐层精细化过程中使用深度可分离卷积实现以粗对齐特征、对应坐标嵌入向量、局部特征相关性的组合为输入的稠密对应场估计,同时提出基于深度一致性预测的匹配不确定性估计与稠密匹配点均匀采样方法,在稠密匹配及其下游应用的全流程均进行了有效改善.

另一方面,由于直接对高分辨率图像特征进行全局匹配需要巨大的存储空间和计算量,基于特征描述子匹配的方法同样采用了由粗到细的总体框架,将全局匹配限制在低分辨率特征图,之后基于低分辨率粗匹配结果在局部区域进行高分辨率特征匹配实现微调,这类方法与前述基于回归的方法区别在于这类方法的细匹配阶段仍然基于特征描述子空间的最近邻匹配完成(有的方法使用了最近邻匹配的松弛版本 dual softmax 或者 sinkhorn 匹配层).具体地,Rocco 等人[77]提出 NC-Net,该方法首先穷举式计算两幅图像中 1/8 分辨率网格点两两之间的匹配分数 $C\in R^{w_1\times h_1\times w_2\times h_2}$(即:ResNet-101 第 conv4_23 层输出特征之间的相似性,其中 $w_1,h_1$ 为第一幅图像经过 conv4_23 输出特征图的宽和高,$w_2,h_2$ 为第二幅图像特征图的宽和高),并对其进行软最近邻

互匹配规范化，如公式(9)所示：

$$\hat{c}_{ijkl}=\frac{c_{ijkl}}{\max_{ab}c_{abkl}}\times\frac{c_{ijkl}}{\max_{ab}c_{ijab}}\times c_{ijkl} \tag{9}$$

之后使用4维卷积对互匹配规范化之后的相似性矩阵 $\hat{C}$ 进行操作，以学习相邻候选匹配点之间的一致性关系并用其调整网格点的相似性矩阵，最后再通过软最近邻互匹配规范化得到最终的相似性匹配矩阵用以计算误差. 此外，由于对图像对进行稠密对应点标注非常困难，NC-Net 仅使用匹配图像中的稀疏匹配点标注进行误差计算，从而保证了该方法在实际应用中的可行性.

NC-Net 中所有网格点的相似性矩阵以及所采用的4维卷积非常占用显存，而网格点对应原始匹配图像经过 CNN 得到的特征图上的空间点，因而限制了网格点的个数，所以得到的稠密匹配结果在图像像素上的精度非常差(可达16像素). 针对 NC-Net 空间复杂度和计算复杂度高的问题，Rocco 等人[143]进一步提出了 Sparse NC-Net，利用相似性矩阵内在的稀疏性减少4维卷积的空间复杂度，加快计算速度，并提出两阶段匹配点重定位方法以提高最终输出匹配点的精度，相比 NC-Net 可以处理4倍以上大小的特征图. Li 等人[144]提出 DRC-Net，基于双分辨率网格点、由粗到细的匹配思想，有效解决了 NC-Net 匹配精度与图像分辨率难以兼顾的问题.

类似于 DRC-Net 由粗到细稠密匹配的策略，并且注意到 Transformer 对像素点全局相互关系的强大建模能力，Sun 等人[55]提出 LoFTR 方法，在 ResNet 提取的特征图基础上，融入位置编码并使用 Transformer 对输入的两幅图像交替进行自注意力和跨图注意力学习，提升匹配图像的特征表示能力. 为了提高匹配特征点的位置精度，同时考虑到 Transformer 注意力学习的高复杂度，LoFTR 采用了由粗到细的匹配策略，在原始图像 1/8 大小的特征图上使用上述 Transformer 注意力学习获取匹配性能强的特征，并在 1/8 特征图上的匹配基础上，使用局部区域内的 1/2 大小的特征图特征进行精细位置的特征提取和匹配，在细匹配阶段，同样采用了 Transformer 注意力学习提升特征的匹配性能. Bokman 和 Kahl[145]受 GIFT[146]启发，提出使用具有旋转不变性的组滤波器替换 LoFTR 中的卷积层，改善了 LoFTR 针对大旋转图像匹配性能退化的问题，该技术同样可应用于其他基于 CNN 的特征点提取和描述子学习网络. Chen 等人[147]提出 ASpanFormer 对 LoFTR 中的注意力特征提取方法进行改善，提出全局-局部注意力模块(Global Local Attention)自适应确定交叉注意力的作用范围. Tang 等人[148]利用二次树数据结构构建视觉令牌(token)金子塔实现由粗到细的高效注意力计算方法，将传统注意力计算的复杂度由二次降为线性，应用于 LoFTR 提升了计算效率和匹配精度. Wang 等人[149]提出 MatchFormer，从特征提取阶段使用基于 Transformer 的自注意力和交叉注意力，通过匹配图像相互之间的上下文关系提升提取特征的匹配能力，结合 LoFTR 提出的由粗到细匹配策略和高效的 Transformer 操作得到了相比 LoFTR 更加轻量的图像匹配网络. Zhou 等人[150]提出 Patch2Pix 方法，对 NC-Net 获得的粗匹配使用级联的两级精修网络进行像素级位置的精调整，同时对匹配结果给出置信度，不仅可以获得像素级位置精度的点匹配结果，而且匹配正确率相比 NC-Net 也有明显提升. 由于上述方法在精细匹配阶段通常也只能使用 1/2 原图大小的特征图进行匹配，最终得到的匹配点并非是逐像素的，因此有些文献中也称此类方法为半稠密匹配.

由于稠密点匹配方法无须进行特征点检测，在弱纹理、复杂几何形变等特征点检测重复性不好的情况下，依然可以取得较好的点匹配效果，但这类方法计算复杂度较高、处理高分辨率图像的能力有限. 此外，稠密点匹配方法应用至多幅图像匹配时会面临匹配点不一致的问题，即：同一幅图像与其他不同视角图像匹配时会出现输出的匹配点不一致情况，这个缺点使得在 SfM 等依赖多幅图像点对应关系的应用中，需要特征点近似等后处理，也会影响实际的应用效果[55]，在三维重建、视觉定位等图像匹配下游任务中的应用受到一定限制.

## 5 误匹配去除

在图像特征点匹配方面，近几年出现了一些基于深度学习的误匹配去除方法，在具有较为复杂的几何变化图像上取得了较好的特征点匹配效果，这类方法的基本思路在于利用深度网络拟合正确匹配点的内在几何流形，或者利用正确匹配点相互之间的一致性关系去除错误匹配点.

文献中利用深度网络处理特征点匹配集合滤除错误匹配点的最早尝试为 Yi 等人[151]提出 LFGC-Net，该方法将两幅图像之间的初始匹配点串接组成一个4维向量，构建多层感知器网络对4维坐标空

间的内在几何一致关系进行学习,并提出 Context Normalization 技术对每组匹配点提取的特征嵌入全局的上下文信息,该方法需要已知相机内参数进行网络训练.由于 LFGC-Net 没有考虑局部点与点之间的关系,提取的匹配点特征表示能力有限,从而限制了其对正确/错误匹配点的分类能力,Zhao 等人[152]进一步提出通过构建具有局部一致性的图结构进行匹配点潜在特征表示的学习,获得了相对 LFGC-Net 更好的错误匹配点滤除性能,他们提出的 NM-Net 基于"局部区域内的正确匹配点对应的仿射关系具有一致性"的假设,获得一致性高的匹配点构成 4 维向量空间的局部近邻图,在图上进行学习获得了良好的匹配点特征表示,NM-Net 采用极几何关系确定的点对应关系作为交叉熵损失的真实标签信息进行网络训练.Liu 等人[153]利用图拉普拉斯方法对嵌入运动一致性约束的匹配关系图进行解析求解,将其融入神经网络提出可差分的一致残差层实现稀疏特征点匹配中的运动一致性学习.Zhang 等人[154]引入可差分池化层(Differentiable Pooling Layer)以自适应地确定对应匹配点的邻域点挖掘其局部上下文关系,并在此基础上提出顺序敏感的滤波模块对不同局部区域的特征点对应关系进行融合,提取对应匹配点蕴含的潜在全局上下文关系,最终通过对应的顺序敏感可差分上采样层获得与输入一样大小的样本个数特征以预测输入对应点的置信度,所提出的 OA-Net 同 LFGC-Net 一样,使用极几何关系引导的匹配关系以及基本矩阵估计的误差作为监督信号进行整个网络的训练.Sun 等人[155]从最小二乘的角度对 LFGC-Net 中的 Context Normalization 特性进行分析,并在此基础上受加权最小二乘法对外点稳定性高的特性,提出具有注意力机制的 Context Normalization,提升了 LFGC-Net 对错误匹配点占比较大情况下的性能.针对初始特征点匹配集合中存在极大比例的错误匹配点问题,Zhao 等人[156]通过学习正确匹配点内在关系一致性,提出渐进剪枝网络 CLNet 对点匹配集合逐步去除错误匹配点.最近,Ma 等人[157]使用多头注意力机制对 LFGC-Net 进行了改良,通过在原始 LFGC-Net 中的特征提取层之后串接 Transformer 注意力层实现上下文敏感的特征提取,以使得网络的学习更加聚焦于正确匹配点相互之间的几何关系.

上述方法依赖初始特征点匹配作为输入,为了最终获取较多的正确匹配点,通常采用一对多的匹配方式获得初始特征点匹配集合,这无疑增加了初始匹配点集合的外点率,过大的外点率又会给误匹配去除带来挑战.另一方面,误匹配去除的重要应用场合在于匹配点正确率较低时,通过误匹配去除提升内点率,以满足后续模型估计的要求,在这些场景下,使用一对多的匹配方法生成的初始特征点匹配集合会面临更大的外点率,甚至超出这类方法的处理能力.因此,上述误匹配去除方法对于解决实际应用中的挑战性图像匹配问题依然很难取得令人满意的结果,目前在各类下游任务中并未得到广泛应用.

## 6　典型应用

纵观计算机视觉的发展历程,图像匹配技术发挥了重要作用,很多计算机视觉任务都可以通过建立图像之间的点对应关系解决,例如:稀疏特征点匹配的经典方法 SIFT 提出时就是为了解决在图像中识别并定位出给定模板中物体的问题[1],以 SIFT 为代表的图像特征点匹配技术推动了图像拼接技术的应用和成熟[12],实例级目标检索也是基于图像特征匹配发展起来的应用[158],除此之外,图像匹配相关技术在目标跟踪[159]、相机姿态估计[160]、SLAM[14,161,162]、三维重建[11,65,163]、视觉定位[57,122,164,165]等任务中均有重要应用.在这诸多应用中,图像拼接已经出现了很多商用产品,随智能手机进入了千家万户,而目前图像匹配技术难以克服的困难挑战在图像拼接相关应用中也较少出现,因此图像拼接领域目前主要关注在拼接方法优化和新的应用场景方面[166,167],而物体识别/检测、目标跟踪、图像检索作为计算机视觉的主要研究内容之一,目前已经呈现出深度学习端到端方案大一统的局面[168-171],基于图像特征点匹配的方法在这些应用中逐渐淡出历史舞台.现阶段,三维重建、视觉定位、SLAM 作为对多视角几何学具有强依赖性的应用,在各类实际应用条件(如室外剧烈光照变化、室内弱纹理场景)下获得高质量的图像点对应关系依然是其从实验室走向实用的核心问题,图 7 展示了这三个典型应用示例,下面对其逐一展开介绍.

### 6.1　图像三维重建

一个完整的图像三维重建通常包括稀疏点云重建、稠密点云重建和三维模型生成三个部分.

稀疏点云重建的核心技术是从运动恢复结构(Structure from Motion,SfM),它基于具有重叠视场图像之间的点对应关系,根据多视角几何学理论,通过大规模非线性优化,计算出所有图像匹配点所

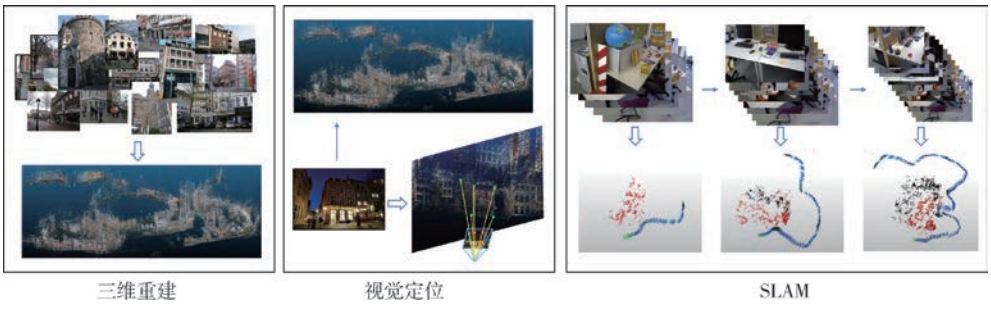



图 7 图像匹配典型应用

三维重建:利用同一场景的多幅图像,计算场景对应的三维点;视觉定位:基于给定的场景三维点云,计算查询图像相对三维场景的位姿;SLAM:针对视频,实时进行场景三维重建、计算当前帧相对重建场景的位姿.

对应的空间三维点以及图像对应的相机矩阵,由于获得的空间三维点对应于图像中的稀疏特征匹配点,该步获得的场景三维重建是以稀疏的三维点云进行表示的.

常用的 SfM 程序有 Blundler[172]、Visual SFM[173]、COLMAP[163]、OpenMVG[174]. Blundler 可以说是最早出现的得到广泛应用的 SfM 开源程序,COLMAP 在多方面对 SfM 算法流程进行了优化,是目前综合性能最好的开源 SfM 程序,得到了广泛应用,但运行速度较慢,而 Visual SFM 采用线性增量式 SfM 算法可以快速完成大规模三维场景重建,并且对 SfM 重建的过程进行实时的展示,便于应用者理解 SfM 算法原理和运行过程,缺点是对于比较困难的大规模场景,重建的完整度和精度相对 COLMAP 较低,而且整个系统没有开源,仅有部分核心算法提供了源代码. 稠密点云重建通常在稀疏点云重建基础上进行,它依赖 SfM 获得的图像相机内外参数等信息,基于多视角图像之间的极几何关系,在较小的范围内根据多视角图像的局部一致性,建立图像像素点之间的密集对应关系,以此获得拍摄场景所对应的稠密空间三维点. 常用的稠密点云重建方法有结合 Blundler 一起使用的基于局部图像的 PMVS 算法[175]以及近年提出的基于深度学习的多视图立体重建算法[141,176],此外,COLMAP 也集成了稠密点云重建方法,可以便捷地结合 COLMAP 中的 SfM 方法直接使用.

图像匹配在三维重建中的应用主要体现在稀疏三维重建环节,图像特征点匹配是 SfM 算法的输入,也是其核心环节,因此,图像特征点匹配的相关算法也经常通过 SfM 来验证其性能的优劣[11,57,64,107],此外,由于 SfM 的输出质量直接影响到后续的稠密点云重建质量,所以图像特征点匹配一直被认为是整个图像三维重建系统中的关键问题. 近年来提出的许多深度学习特征点检测、描述和匹配方法都在 COLMAP 或 Visual SFM 等框架下通过代替其中的 SIFT 模块进行了验证,在重建出的稀疏点云规模、重建的图像数量、重建的重投影误差等指标上展现出超越传统手动设计特征描述子的性能[57,58,177,178],但优势并非十分显著,而且验证的数据集相对比较有限(主要集中在 ETH 3D 数据集[59]和 IMC 比赛数据集[179]),因而,在实际应用中全面替代传统的 SIFT 方法仍然需要更多的验证. 究其原因,主要是因为 SfM 方法在其自身的发展中,已经建立了一套比较完善的鲁棒优化方法,对于特征点匹配质量具有较强的容错性,而近年来提出的深度学习图像匹配方法虽然相比传统方法在图像匹配性能上取得了明显的改善,但是在极端困难条件(如:大视角变化、尺度变化、复杂光照条件变化、人为活动等因素,尤其是综合了多种因素的复杂图像变化)下的图像匹配性能依然较低,因此在下游的 SfM 任务中性能提升不够显著. 这也反映出现有的基于深度学习的图像匹配方法依然存在较大的局限性,如要在对应的下游任务中获得全面应用,仍然有待更具突破性的方法创新和更多的应用验证.

### 6.2 基于图像的定位

近年来,随着手机定位导航、自动驾驶、增强现实等应用需求的不断扩大,基于图像的定位技术得到了研究人员的广泛关注. 基于图像的定位亦称为视觉定位,它通过手机拍摄的照片计算得到拍照时手机相对于已知三维环境(三维地图)的位置和姿态,是基于卫星[180]、WLAN(Wireless Local Area Networks)[181]等定位手段的有效补充,在地下停车场、商场等室内环境定位中起着至关重要的作用,而且相对 WLAN 无线定位等技术手段具有更高的稳定性. 此外,由于视觉定位可以获取照片拍摄时的相机姿态信息,可以根据相机成像原理方便地在照片

中绘制虚拟三维物体信息,在增强现实应用中具有不可替代的地位[182].

基于图像的定位大致可以分为基于图像检索(全局特征)的方法[183-185]、基于学习的方法[186]、基于三维结构(局部特征)的方法[132,187].基于图像检索的方法构建全局化的图像特征描述向量,通过比较待定位图像全局描述符与数据库图像全局描述符的相似性,找到数据库中相似性最大的图像,使用其位置信息作为定位结果,该方法对数据库图像依赖较大,而且只能进行粗略定位,因此在实际应用很少单独使用,而是结合局部特征的方法使用,用于筛选出较少的可能具有视场重叠的图像用于后续图像特征点匹配,这类方法中的代表性工作有 NetVLAD[183]、DenseVLAD[188]及其改进的方法[184].基于学习的方法旨在利用 CNN 网络直接学习图像对应的 6D 姿态进行定位,如 PoseNet[186],但是这类方法在对场景的适应性方面存在较大局限性[189].目前最常用的视觉定位是基于局部特征的方法,它通过局部图像特征匹配建立待定位图像与已知的场景三维模型之间的 2D-3D 匹配关系,通过 PnP 算法[5,190]计算得到定位图像对应的相机位姿.为了提高定位速度和定位的鲁棒性,基于局部特征的方法通常会与基于全局特征的方法相结合[132,187],利用全局图像特征通过图像检索或者位置识别技术,获取与待定位图像具有潜在匹配关系的少量图像,之后利用图像特征点匹配方法建立图像之间的点对应关系,目前取得领先性能的 HLoc[132]和 PixLoc[191]都属于这类,这类方法也称为层次化定位方法.

值得一提的是,基于三维结构的方法具有很大的灵活性,其中的基于全局特征的图像检索模块可以测试不同的全局特征用于视觉定位的性能,而替换 HLoc 中的局部图像特征和匹配可以测试不同的图像特征点匹配方法在视觉定位中的性能,通过在 HLoc 中集成 SuperGlue[59],Paul-Edouard Sarlin 等人获得了 CVPR 2019、CVPR 2020、ECCV 2020 举办的连续三届长时视觉定位竞赛①的冠军,以及 CVPR 2020 举办的图像匹配竞赛②冠军.此外,KAPTURE③对视觉定位中的不同数据以及常用的性能测试基准协议进行了统一表示[192],并给出了一个简单易用的视觉定位流程,可以方便集成不同的图像匹配方法进行定位效果的测试.

得益于深度学习图像匹配技术的发展,目前针对相同气候、光照条件下的室外场景定位可以取得较好的结果,当下研究领域更关注具有挑战性的长时视觉定位与室内视觉定位问题.

### 6.3　视觉 SLAM

SLAM 的全称为同步地图构建与定位,同时包含了三维环境地图重建与定位两部分,顾名思义,视觉 SLAM 指以视觉传感器(如:摄像头)输入为主的 SLAM 技术,它是机器人在未知环境中行走的核心技术[193-195].然而,虽然视觉 SLAM 的理论基础同图像三维重建技术和定位技术一样,都是多视角几何学理论,但是它并非前述基于图像的三维重建和定位的简单组合.

一般来说,视觉 SLAM 通常包含地图构建和定位两个并行的模块.建图模块通过三角化相邻帧之间的对应特征点获得场景三维点,并在必要的时候或者系统空闲时进行捆绑调整以不断优化所构建地图的精度;定位模块需要根据当前帧图像特征点与已有三维场景点之间的对应关系计算当前帧的相机姿态与位置.此外,还需要进行闭环检测以消除长期建图过程中的累积误差.图像匹配方法在视觉 SLAM 系统中的上述三个环节中起着至关重要的作用.

由于 SLAM 需要对视频输入进行实时处理,它对所采用算法的计算速度要求很高,而且一旦定位误差超出一定范围就会导致整个 SLAM 系统失败,因此它对特征点匹配算法的精度要求也很高.另一方面,视频的连续性可以有效降低图像特征点匹配的难度,所以,尽管视觉 SLAM 对图像特征点匹配方法的性能要求高,它所面临的图像匹配问题相比图像三维重建或者定位中的匹配问题更简单一点.

尽管如此,视觉 SLAM 的图像特征点匹配依然是一个开放性问题,现阶段的视觉 SLAM 更像是一个整体的复杂系统,在图像特征点匹配问题未完全解决的情况下,通过实时跟踪、回环检测、深度估计以及一些复杂的工程化实现技巧,有效弥补了图像特征点匹配能力不足对整个 SLAM 系统的影响.值得一提的是,尽管快速、准确的图像特征点匹配方法在 SLAM 系统中起着非常关键的作用,然而已有的轻量化图像特征点匹配方法在许多 SLAM 应用场景中都不能获得高质量的特征点匹配结果,这方面能力的不足促使 SLAM 领域的研究人员在整体的系统鲁棒性方面做了很多创新[14,31,32,196],有效推动

① https://www.visuallocalization.net/
② https://image-matching-workshop.github.io/
③ https://github.com/naver/kapture

SLAM 技术的发展与应用，也提出了一些不依赖图像特征点的 SLAM 方法[197-200] 以及基于语义的 SLAM 方法[201]. 可以预见，随着轻量化特征点匹配方面的技术突破，视觉 SLAM 系统会更加稳定、简洁，推动更多的相关应用发展. 目前使用较多的视觉 SLAM 系统当属基于轻量化局部特征 ORB 的 ORB-SLAM 系列[14,31,32]，从 2014 年第一版 ORB-SLAM[14] 发布，目前已经发展到了 ORB-SLAMv3[32]，可实现稳定的基于单目相机、双目相机、RGB-D 相机的纯视觉 SLAM、视觉-惯导结合的 SLAM、以及多地图 SLAM 系统，并且适用于小孔和鱼眼相机两种常用相机成像模型，此外，融合多传感器的 SLAM 系统也是该领域目前的一大发展趋势，对于提高应用场景的系统稳定性具有重要作用[202-204].

## 7 数据集

UBC Patches[44]：该数据集由 Matthew Brown 等人提出，广泛应用于数据驱动的特征描述子学习方法的训练与测试. 该数据集以局部图像块的形式展示，包含了三个子集，每个子集的局部图像块来自于三个不同场景(Liberty、NotreDame、Yosemite)的图像，这些局部图像块由 SfM 算法重建得到的三维点云根据估计得到相机矩阵在对应图像上重投影得到，它们之间的对应关系也由于之对应的三维点确定，每个三维点对应了 2 至 15 个局部图像块. 为了方便测试，Matthew Brown 等人在构建数据集时为每个子集提供了包含 200,000 对和 100,000 对局部图像块的训练子集与测试子集，其中匹配的局部图像块和不匹配的局部图像块各占一半. 该数据集仅包含了局部图像块以及它们之间的对应关系，所以它只能用于特征描述子深度学习方法的训练与测试，不能用于图像特征点检测算法的评测以及包含特征点和特征描述子的图像匹配方法评测，此外，由于局部图像块由重建三维点重投影得到，而该数据集采用的 SfM 重建方法使用了 SIFT 特征进行图像特征点匹配，因此该数据集提供的局部图像块为图像中 SIFT 特征点周围的局部图像信息，不利于测试更一般性的特征点描述方法. 现阶段，该数据集在有监督的深度学习方法中，已经表现出性能饱和，因此近几年主要用于无监督深度学习特征描述子的训练和评测.

HPatches[47]：该数据集在 ECCV 2016 首次提出，随 ECCV 2016 举办的特征描述子竞赛一同发布，是竞赛使用的数据集. 该数据集基于图像匹配领域之前提出的若干小规模数据集并进行了扩展，主要关注平面场景的视角变化和一般场景的光照变化，包含了 2 个子集，分别对应不同光照变化的图像集合与不同视角变化的图像集合，其中光照变化子集包含了 57 个场景，每个场景包含 1 张参考图像和 5 张光照变化图像，而视角变化子集包含了 59 个平面场景，每个场景包含 1 张参考图像和 5 张不同视角下拍摄的图像，在进行图像匹配测试时，标准的做法是使用参考图像和另外 5 张图像分别组成 5 对图像对. 由于不同视角的平面场景图像之间的点对应关系有单应变换确定，而光照变化图像不存在像素点位置的变化，图像之间的变换关系由单位阵确定，因此，该数据集内同一场景图像之间的对应均由单应变换(Homograph)确定，该数据集也因此命名为 HSequences.

针对应用特征描述子的三个常见任务：局部图像块的验证、匹配、检索，HSquences 的提出者在上述 116 个场景的参考图像中使用 Hessian、Harris 和 DoG 特征点检测算法进行特征点提取并进行聚合、去除重复检测点，在此基础上，模拟不同程度的变换扰动将参考图像中的特征点映射至对应的 5 幅匹配图像中，在这些图像中，根据特征点位置和尺度得到 $65 \times 65$ 大小的标准化局部图像块，得到 HPatches 数据集. 由于对应关系已知，根据扰动剧烈程度，这些对应的局部图像块划分为容易、难、困难三个档次，并根据验证、匹配、检索三个任务的特性，划分出不同任务对应的测试图像块. 由于任务本身的难易程度不同，相对来说，目前该数据集上验证任务的性能较高，大部分主流方法的平均正确率均在 90%以上，而匹配和检索任务的平均正确率分别只有 60%和 70%左右的水平[58,107]. 由于 HPatches 仅包含局部图像块，因此它只能用于测试特征描述子的性能，而 HSequences 提供了原始的匹配图像用于测试，它在基于深度学习的特征点检测、端到端特征点检测和描述方法的测试评估中使用较多，并且可以使用单应变换参数估计精度评估特征点匹配方法在下游位姿估计任务中的性能，目前该数据集的使用主要集中在 HSequences.

IMC2020(Image Matching Challenge)[64]：该数据集包含了 25 个场景的约 3 万张图像，其中 2 个场景用做验证集、9 个用作测试集，其余 14 个场景用作训练集，每个测试场景仅公开了 100 幅图像. 数据集发布者使用 COLMAP 对每个场景图像进行重

建获得相机的姿态、图像的深度图等信息,这些信息用于确定图像之间的点对应关系,以此评估不同的特征点匹配算法的性能.该数据集给出了一个完整的评测流程,包括两个基于图像匹配的下游任务(立体匹配和多视角三维重建)以及对应的测试图像,并使用下游任务的姿态估计精度用于评估图像匹配过程所涉及到的算法性能.该数据集的发布者近几年均在计算机视觉的顶级会议上举办竞赛以评估该领域的最新进展.

MegaDepth[205]:该数据集包含了来自 196 个场景的超过 100 万张互联网图像,每个场景通过 COLMAP[163]程序运行 SfM 算法和 MVS 算法重建得到对应的稠密三维点云以及图像对应的相机内外参数.在所有的重建图像中,有 102,681 张图像重建成功提供了对应的深度图、相机内参数、姿态,在使用 MegaDepth 进行算法训练和测试时,具体使用的是这些重建成功的图像.由相机的内外参数可以计算得到两幅视场重叠图像之间的极几何关系,基于图像之间的极几何关系以及对应的深度信息,可以确定图像之间的点对应关系,因此该数据集提供了大规模的图像点对应关系,随着 HPatches 数据集上性能的饱和,近年来越多越多的方法使用 MegaDepth 数据集进行训练,如 D2-Net、CAPS、LOFTR 等,通常使用其中的 118 个场景图像用于训练,剩下的 78 个场景图像用于验证.由于该数据集提供真实的位姿参数,因此也用于测试图像匹配的下游任务:位姿变换参数估计.

GL3D[206]:该数据集是一个类似 MegaDepth 的数据集,共包含了来自 378 个场景的 90,590 张高分辨率图像,每个场景包含 50 至 1000 张不等的具有大视野重叠的无人机航拍图像,378 个场景涵盖了城市、乡村、景点等多样化场景,在场景类型、相机视角等方面与 MegaDepth 具有较大差异,是 MegaDepth 图像类型的重要补充.该数据集随机选择了 338 个场景共 81,222 张图像进行三维重建[207,208],与 MegaDepth 一样,通过三维重建获得了这些场景对应的三维点云模型以及重建成功图像对应的深度图、内外参数,由这些信息可以计算得到具有视野重叠图像之间的极几何关系以及图像点对应关系.目前该数据集在文献中主要用于模型的训练,基于该数据集训练获得的代表性方法有 ASLFeat.

YFCC100M[209]:该数据集包含了 Yahoo 公司研究人员发布的在 Flickr 网站上获取的约 1 亿张全球地标公开图像,Heinly 等人[210]提出大规模 SfM 算法对其进行稀疏三维重建,最终重建出包含约 150 万张图像的 72 个地标场景所对应的三维模型,类似于 MegaDepth 和 GL3D,重建成功的 72 个地标场景对应了 72 个子集,每个子集内的图像点对应关系可以通过三维重建结果获得.该数据集在基于深度学习的误匹配滤除问题上应用较多,如 LFGC-Net、OA-Net、CLNet.

ScanNet[211]:该数据集是一个室内三维扫描数据集,包含了 1613 个单目视频序列,并且提供了对应相机姿态和深度图.该数据集根据场景不同划分了训练集、验证集和测试集,一般也用于测试在其他数据集训练得到的特征点匹配方法应用于相机姿态估计任务的性能.

InLoc[131]:ScanNet 之外较为常用的用于评估室内定位和位姿估计算法性能的测试数据集,包含了华盛顿大学圣路易斯分校内的 2 栋大楼内的 5 个楼层图像,共有 277 张彩色深度(RGB-D)全景图像,这些图像通过与给定的建筑物平面图进行对齐切分,并根据所属物理区域划分为 5 个子集,共包含 9972 张 1600×1200 分辨率图像用做数据库图像,另外还使用 iphone7 在不同时间采集了 2 个楼层的 356 张高分辨率图像用做查询图像.

Aachen Day-Night[212,213]:该数据集包含德国亚琛老城区的图像,涵盖了 4328 幅白天自然光场景下的图像用于三维重建,以及 824 幅其他时间白天光照和 98 幅夜晚灯光条件下的查询图像[212],查询图像的姿态未公开,通过将定位结果上传至测试服务器① 的方式计算三个不同误差阈值下(即:(0.25m, 2°)、(0.5m, 5°)、(5m, 10°),其中前者为相机姿态的位置误差,后者为旋转误差)的定位精度.该数据集 V1.1 版本[213]通过额外增加 2369 幅亚琛城区白天图像,对定位场景进一步进行了扩充,同时增加了 93 张使用不同手机拍照获得的夜间查询图像.

RobotCar Seasons[10]:该数据集通过车载摄像头收集了英国牛津的 26580 张图像用于三维场景重建,同时在相同道路行驶轨迹、不同时间、不同季节采集了 5616 张图像用于视觉定位性能测试.同 Aachen Day-Night 一样,该数据集用于对视觉定位算法在不同季节、不同气候条件下的性能进行评估,但是数据集侧重城市道路驾驶场景,而且采集的时

① https://www.visuallocalization.net/

间跨越长达一年,因此覆盖了不同季节和更多的气候条件.同 Aachen Day-Night 一样,数据集发布者没有公开查询图像的真实位姿,通过上传至测试服务器的方式进行定位精度计算.

Extended-CMU Seasons[10]:包含了城区与市郊场景的图像,利用汽车上安装的 2 个摄像头采集了不同季节、不同光照的图像,植被随季节变化引起的外观差异在图像中非常普遍.该数据集共包含 60937 张数据库图像,查询图像涵盖了晴天、阴天、多云三种不同的气候条件,城区、郊区、公园三种不同的场景,以及三种不同的植被条件,共计 56613 张查询图像,同样地,查询图像的真实位姿没有公开,需要通过上传至测试服务器的方式进行定位精度计算.

ETH 3D[59]:该数据集通常指 Schonberger 等人[59]的特征点和描述子评测工作所使用的数据集,由于 Schonberger 等人开源了一个标准易用的评测框架,该数据集得到了广泛的应用,用于比较不同的特征点匹配算法对于三维重建任务的优劣.该数据集包含了来源于 Strecha 等人[214]发布的多视角立体重建数据集和康奈尔大学研究人员的发布的大规模 SfM 数据集[215,216].Strecha 等人[214]发布的多视角立体重建数据集包含了“Fountain”和“Herzjesu”两个场景,分别包含 11 幅和 8 幅针对拍摄景物的环视高分辨率图像,并且提供了激光扫描数据和相机的真实姿态用于评估三维重建精度.大规模 SfM 数据集包含了在互联网搜集的 14 个地标场景图像,每个地标场景包含从 1000 至 7000 不等的图像,Schonberger 等人[59]在评测不同特征描述子时选择了该数据集其中 5 个地标场景.此外,此数据集还包含了文献[217]提供的康奈尔大学数据集,含有 6514 张图像以及通过高质量差分 GPS 获得的相机姿态作为真实值.

TUM RGB-D[218]:由德国慕尼黑技术大学的研究人员发布的一个室内环境 SLAM 测试数据集,共包含了不同纹理、光照、结构等多样性室内环境下的 RGB-D 视频序列(通过微软 Kinect 获取),视频分辨率为 640×480(30Hz),同时提供了通过外部运动捕捉系统获得的相机位姿作为真实值.

KITTI[219]:包含了 11 段在城区和高速路上拍摄获得的双目视频序列,双目相机架设在车辆顶端,视频分辨率为 1240×376(10Hz),其中有 6 个视频序列包含了 SLAM 系统中经常遇见的回环问题,此外,该数据集包含了 LiDAR 获取的周围环境三维点云数据以及通过 GPS 获得的精确位置信息,广泛用于验证视觉 SLAM 算法在自动驾驶场景的性能.

EuRoC[220]:包含了 11 段 20Hz 双目视频序列,视频序列由微飞行器机载的双目传感器获得,视频内容涵盖了两个不同的室内场景和一个大规模的工业场景,该数据集同时提供了惯性数据,因此可用于视觉-惯性融合的 SLAM 算法测试,真实的位姿通过运动捕捉系统获得.

图 8 展示了这些典型数据集中的部分图像示例,表 2 概括了上述图像匹配相关任务常用的数据集.总体而言,通过人工收集的图像匹配数据集体量小,不适合进行机器学习方法尤其是基于深度学习的图像匹配方法训练,而现有大规模图像匹配数据集都依赖 SfM 算法重建获得场景的三维模型以及图像对应的相机参数和姿态数据,在此基础上建立图像之间的点对应关系,在一定程度上,这些大规模图像匹配数据集有力推动了深度学习图像匹配方法的进展.然而,现有的基于 SfM 算法得到的数据集均以 SIFT 特征点及其匹配结果为基础,因此仍然存在一定的偏向性,即:在这些数据集上很难学到非常具有挑战场景的图像匹配特征.此外,通过随机生成单应变换和改变图像对比度、图像风格迁移等方式生成匹配图像对用于网络训练也被一些研究人员作为自监督学习策略采用[72,78,94,124,221],但是单应变换仅能模拟平面场景在不同视角下的成像,对于一般性的三维场景建模能力弱,合成的图像不能很好的反应三维场景中物体的遮挡情况,学习得到的特征在解决遮挡等引起的错误匹配方面也存在一定的局限性,另一方面,修改对比度以及图像风格迁移仍然不能很好的模拟不同光照等成像条件的变化,且图像风格迁移技术仍处理发展中,因此这类方法对于复杂光谱和视角变化下的图像匹配问题很难提供高质量训练数据以学习得到适用的图像特征.

对于图像匹配任务,常用的评价指标有特征点重复率[47]、特征点匹配精度[47]、平均匹配精度(Mean Matching Accuracy,MMA)[122]、正确关键点比率(Percentage of Correct Keypoints,PCK)[78]、平均端点误差(Average Endpoint Error,AEPE)[78].特征点重复率指同一个点在两幅匹配图像中被同时检测出来的概率,定义为两幅图像中同时检测的特征点个数与共视区域中检测出的所有特征点个数之和,如果图像 A 中的点经过真值(groundtruth)变换映射至图像 B 之后,在某个阈值(如:2 像素)范围内存在从 B 中检测的特征点,则认为该点为可重复检


图 8　图像匹配常用数据集中的图像示例

**表 2　图像匹配及下游任务相关的常用数据集**

| 数据集名称 | 数据集规模用途 | 特性 | 使用该数据集的文献 |
|---|---|---|---|
| | | 图像匹配 | |
| UBC Patches[44] | 场景:3<br>图像:未提供<br>图像块:约 15 万 | ·利用 SfM 对图像进行场景三维重建,根据重建结果确定图像中 DoG 特征点对应关系,取出对应的局部图像块;<br>·包含三个不同的三维场景,每个场景包含约 150k 三维点,每个三维点对应 2～15 个局部图像块. | [39,42,44,45,54,58,86,88,95,96,99-103,105-111] |
| HPatches[47] | 场景:116<br>图像:696<br>图像块:约 90 万 | ·仅包含平面场景图像或远景固定机位图像,给定了图像之间的单应关系;<br>·提供了 patch-level 的数据,分为 easy、hard、tough 三个难度;<br>·基于 patch-level 数据,定义了验证、匹配、检索三个标准任务;<br>·提供了所有 116 个场景的匹配图像,分为光照变化子集和视角变化子集. | [11,47,53-55,57,58,72,88,89,91-93,101-105,107-115,121-126,128,130,132,143,144,149,150,222-226] |
| IMC2020[64] | 场景:25<br>图像:约 3 万 | ·每个场景仅公开 100 幅图像用作测试,使用所有图像利用 SfM 重建获得场景三维点云和图像位姿真值;<br>·两个下游任务(立体匹配和多视角三维重建)用作性能评. | [53,64,123,125,147,178,222,226] |
| MegaDepth[205] | 场景:196<br>图像:约 100 万 | ·使用 SfM 算法和 MVS 算法计算得到场景图像的真实位姿和深度图;<br>·通常划分 118 个场景用于训练,78 个用于测试和验证;<br>·196 个场景中,共有 102,681 张图像重建成功提供了对应的深度图和相机内外参. | [11,55,59,92,113,118,122-125,128,142,144,147,149,150,178,210,223,224,226] |
| GL3D[206] | 场景:338<br>图像:81222 | 与 MegaDepth 类似,含 378 个场景,场景图像通过无人机航拍获取. | [55,115-117,126,185] |
| YFCC100M[209] | 场景:72<br>图像:约 150 万 | 与 MegaDepth 类似,原始图像数据来源于 Flickr 网站的约 1 亿张全球地标图像,重建出的 72 个地标的 150 万张图像用于评估图像匹配算法. | [56,59,115-117,120,130,147,151,153-157,227] |

（续表）

| 数据集名称 | 数据集规模用途 | 特性 | 使用该数据集的文献 |
|---|---|---|---|
| | | 下游任务 | |
| ScanNet[211] | 1613 段 RGB/深度视频<br>约 250 万帧图像<br>任务：姿态估计 | ·室内办公场景，涵盖 707 个不同区域；<br>·RGB 视频：1296×968(30Hz)，深度视频：640×480(30Hz)，两者经过标定进行了对齐；<br>·提供划分好的训练、测试、验证集，包含相机姿态真值和图像深度信息. | [55，59，113，116，117，120，142，147-149，212，226] |
| Aachen Day-Night (v1.1)[212,213] | 数据库图像：4328 (6697)<br>查询图像：922 (1015)<br>任务：视觉定位 | ·室外步行街道场景；<br>·数据库图像为白天自然光场景下的图像，用于场景三维重建；<br>·查询图像包含白天光照和夜晚人造灯光图像. | [10,11,53,55,57,59,89,91,92,113,116,117,122-128，132，143，144，147,149,150,210-213,226] |
| RobotCar Seasons[10] | 数据库图像：26580<br>查询图像：5516<br>任务：视觉定位 | ·城市道路场景；<br>·数据库图像为白天自然光场景下的图像，用于场景三维重建；<br>·查询图像包含跨越一年不同季节、光照、气候图像；<br>·图像通过汽车上安装的 3 个摄像头采集，具有运动模糊. | [10，126，127，132，164，184，187，189，191，192，201] |
| Extended CMU Seasons[10] | 数据库图像：60937<br>查询图像：56613<br>任务：视觉定位 | ·城区与市郊场景，图像利用汽车上安装的 2 个摄像头采集；<br>·数据库图像为白天自然光场景下的图像，用于场景三维重建；<br>·查询图像包含不同季节、不同光照图像，植被随季节变化引起的外观差异在图像中常见. | [10，132，164，165，184，187,191,192,201] |
| InLoc[131] | 数据库图像：9972<br>查询图像：356<br>任务：视觉定位 | ·室内场景，弱纹理区域多，动态物体、重复/相似结构干扰大；<br>·数据库图像包含深度，无需场景三维重建；<br>·高分辨率查询图像，查询图像与数据库图像分辨率差异大. | [55，120，122，127，128，77，143，144，147，149，150,192] |
| ETH 3D[59] | 小场景：8/11 张图像<br>大场景：1000～7000 图像<br>任务：三维重建 | ·2 个小场景包含激光扫描器获得的场景三维真值；<br>·大场景为 14 个地标场景，图像通过互联网搜索获得，无真值；<br>·额外提供一个康奈尔大学的图像集，包含通过高精度差分 GPS 获取的真值. | [7,11,53,54,57,58,72,89，91，92，107，113-115，122，123，127，130，206，224] |
| TUM RGB-D[218] | 近 100 段 RGB-D 视频<br>640×480(30Hz)<br>任务：SLAM | ·包含不同纹理、光照、结构等多样性室内环境；<br>·通过微软 Kinect 采集 RGB-D 视频；<br>·通过高精度动补系统提供真实位姿. | [14,31,60,161,162,193-197,199] |
| KITTI[219] | 11 段视频双目<br>1240×376(10Hz)<br>任务：SLAM | ·汽车顶端双目相机采集视频，涵盖城区和高速公路场景；<br>·包含周围环境的 LiDAR 三维点云数据；<br>·视频路径包含回环；<br>·通过高精度 GPS 提供真实位姿. | [14,15,31,162,194,195,200,201,203,204] |
| EuRoC[220] | 11 段双目视频<br>752×480(20Hz)<br>任务：SLAM | ·机载双目相机采集视频，涵盖 2 个室内场景和 1 个工业场景；<br>·通过高精度动补系统提供真实位姿；<br>·包含惯性传感器数据. | [31,32,60,162,198,200,204] |

测的特征点. 特征点匹配精度指一对图像的匹配结果中正确匹配点与所有匹配点个数之比，如果一对匹配点经过真值变换映射后的像素距离小于给定阈值，则认为其为正确匹配点，匹配精度在所有测试图像对上的均值称为平均匹配精度(MMA)，该指标在有些论文中也称为 PCK. 平均端点误差(AEPE)通常用来评估稠密点匹配方法，由估计的坐标变换场与真实值之间的均方误差进行定义.

对于图像匹配的下游任务，评价指标主要是衡量两幅图像之间的变换关系估计是否准确，根据具体的应用分为单应变换估计的精度和相对位姿估计的精度. 单应变换由一个 3×3 矩阵表示，代表了两幅平面图像之间的像素映射关系，无法通过矩阵元素的比较来衡量估计精度，因此，分别利用估计得到

的单应矩阵和真值单应矩阵将图像 $A$ 的 4 个角映射至图像 $B$ 中,计算映射之后 4 个点的平均欧式距离作为单应矩阵估计的误差,通过与给定阈值相比判断单应矩阵估计是否正确,整个测试数据集上单应矩阵估计正确的图像对所占比率即为正确率(Homography Accuracy,HA)[128],不同精度下的平均召回率为相应的 AUC 指标[55].对于位姿估计,相机位姿可以分解为 $x$、$y$、$z$ 三个方向上的旋转和平移量,这些量可以直接与真实值进行对比得到位姿估计的精度,在相对位姿估计问题中,使用向量夹角定义旋转和平移量的误差[59],在绝对位姿估计(如:视觉定位)问题中,使用向量夹角定义旋转误差、向量距离定义平移误差,通过给定阈值判定位姿估计是否正确,同样可以计算不同误差下的平均召回率为相应的 AUC 指标[59].视觉定位和 SLAM 本质是相机位姿估计,因此通过位姿估计的正确率和精度进行衡量,特别地,视觉定位使用给定旋转和平移阈值下位姿计算正确的图像所占比例定义为定位正确率[10],SLAM 通常报道逐帧的旋转误差和平移误差[190].

## 8　现有挑战与展望

图像匹配方法从上世纪 90 年代基于局部区域内像素灰度矩等统计量的方法[34,35,228]开始发展,经过 20 多年的历程,其间出现了 SIFT、SURF、ORB 等耳熟能详的方法有力推动了相关应用技术的发展,例如目前 SLAM 中广泛使用的 ORB-SLAM 系列依然采用了 ORB 特征进行图像匹配、三维重建领域著名的开源软件 COLMAP 采用 SIFT 进行图像匹配.2015 年之后,基于深度学习提出的诸多图像匹配方法极大地促进了图像匹配技术的发展,总体而言,条件可控、特定场景下的图像匹配问题目前已基本得到解决,现有的挑战在于解决挑战环境下的图像匹配问题,包括弱纹理和重复纹理图像之间的匹配、极端视角变化下的图像匹配、以及不同模态图像之间的匹配.此外,由于特征匹配在诸多任务中的适用性,笔者来看,面向特定下游任务优化的图像特征点匹配方法也将是未来的一个重要发展方向.

### 8.1　弱纹理、重复纹理图像匹配

弱纹理、重复纹理图像匹配问题早在 2010 年左右就引起了研究人员的关注[229],迄今为止一直是图像匹配领域的难点问题.弱纹理指白墙等缺少局部细节信息的区域,而重复纹理指局部细节高度相似的区域(如建筑物的窗户),这两类区域中点与点之间的对应关系均不能依据点周围的局部图像之间的相似性进行确定,基于人类认知的启发,需要结合更大范围内其他容易匹配的点,根据与这些点之间的几何关系确定弱纹理或重复纹理区域中的点对应关系,以此克服这类区域内的点局部图像信息不足于支撑匹配关系建立的问题.在该思路下,早期方法通过图匹配技术利用点与点之间的几何关系一致性获得特征点对应关系[230],基于该思路也出现了一些利用深度图卷积网络学习图像对应点关系的方法以进行图像匹配[116,117]或者误匹配去除[227],而近几年出现的 transformer[50]在关系建模方面表现出更好的能力,也在图像匹配任务中得到了验证[222,223],初步展示了具有处理弱纹理区域图像点匹配的能力[55,147-149],但这些工作仍属于浅尝辄止,总体而言,现有方法对弱纹理和重复纹理图像的匹配在鲁棒性和精确性方面仍然有待提高,如何更好地利用 transformer 结构,设计兼具计算高效性与匹配鲁棒性的特征点匹配方法是未来值得探索的一个方向.

值得一提的是,弱纹理在室内场景中比较常见,而这类场景具有较好的结构化信息,一种替代策略是检测直线并对直线进行匹配以完成后续的相关任务[231,232],近期也出现了一些基于深度学习的直接检测与匹配方法[233,234],然而直线相比特征点具有多一维的自由度,利用其对应关系进行图像配准、三维重建、视觉定位等相比利用点对应关系的方法难度更大,因此,直接建立弱纹理、重复纹理图像之间的点对应关系的方法在未来应该具有更大的应用潜力.

### 8.2　极端视角变化下的图像匹配

极端视角差异(视角差异定义为两幅图像成像的光心到图像中心连线的夹角,一般来说,大于 60 度可认为是极端视角差异)引起图像局部内容呈现出很大的差异,一直以来都是图像特征点匹配的难点之一,在深度学习技术有力推动图像匹配发展的今天也不例外.从现有方法报道的结果来看,利用深度学习联合进行特征点检测和描述可以较好地解决一定程度上的图像光谱差异引起的变化(不同模态图像引起的光谱差异依然是一个挑战),但是对于尺度、大视角等变化处理能力依然比较有限.一方面,现有深度学习方法在处理不同尺度的图像匹配问题时,仍然以采用对图像进行多尺度缩放再提取特征的方式为主[57,122,124],这种方式计算量大,而且通常会提取得到大量的特征点,对于后续的匹配等操作也会增加计算负担,如何设计类似于特征金字塔网

络结构[235]以在一次前向传递计算中提取尺度自适应的特征点和描述子,值得进一步研究,近几年也出现了一些利用深度网络学习特征尺度和图像共视区域的尝试[224-226,236,237].另一方面,现有方法处理大视角变化的策略主要还是依赖匹配特征点之间的几何一致性在特征匹配策略和误匹配去除方面进行解决,提取的特征描述子区分能力自身还不足于满足这类图像的匹配需求,由于特征匹配策略以及误匹配会进一步增加计算量,如何利用深度网络提取特征描述子以具备对这种大视角图像变化的稳定性值得进一步研究.

值得说明的是,在现有的一些应用(如:三维重建、SLAM)中,通过结合应用特点在一定程度上避开了极端视角差异下的图像匹配,从而缓解了现有方法在这方面能力不足对实际应用的影响,例如,多视角三维重建通过对拍照视角进行设计以避开大视角图像匹配问题、或者通过采集冗余图像减少大视角图像匹配缺少正确匹配点对该部分场景重建完整度的影响;SLAM 则将图像匹配限制在相邻帧,通常不涉及大视角图像匹配.然而,在一些特定的应用中,由于应用条件限制必须面临大视角差异图像的匹配问题,例如利用无人机对一些不可达区域进行探测时,需要将斜视的无人机图像与正射视角下的卫星图像进行匹配.因此,针对大视角差异的图像,提取具有对视角差异引起的局部射影变化和遮挡等鲁棒的特征描述子有望从本质上解决这一问题,以此满足一些特定应用的需求,扩大图像匹配的应用领域.

### 8.3 跨模态图像匹配

单一图像传感器受物理特性限制对环境感知能力存在一定的局限性,融合多传感器信息提升机器对客观世界的感知能力是一种有效的手段.由于不同传感器具有不同的物理特性,其成像反映了观测场景的不同特性,通过跨模态图像匹配对这些不同特性融合无疑可以获得更强的感知能力,然而,不同的成像特性也给这些不同模态图像之间的匹配带来了新挑战:如何在具有显著视觉差异的图像中提取出具有模态不变性的内在鉴别性特征,建立不同模态之间的图像点对应关系? 广义上讲,特征的模态不变性可以认为是特征点描述鲁棒性的延申,但具有更大的挑战,因为不同模态图像光谱特性差异更大,有的甚至展现出完全不同的特性(如红外图像中目标的亮度关系可与可见光图像中的亮度关系完全相反),因此跨模态图像匹配既是一个非常有应用潜力的研究方向,又极具挑战性.

针对这个问题,现有研究主要集中在遥感领域的全色与多光谱图像配准[238]、光学图像与 SAR 图像配准[239,240],医学图像处理领域的各种模态医学影像之间的配准[22,241,242],以及针对自然场景感知的可见光与红外图像之间的匹配[90,243-245].总体而言,基于深度学习的技术在该方面也展示出了卓越的性能,如:SDNet[239]、CoMIR[242]等方法,然而,现有的技术进步与实际应用需求仍有不小的差距,例如:针对光学遥感图像和 SAR 遥感图像的配准方法在处理较大几何形变时能力不足,现有的方法大多在光学与 SAR 图像成像视角差异不大的情况下能取得较好的配准结果.另一方面,跨模态图像匹配的公开数据集较少,尤其缺少具有大规模标注的真实多模态图像匹配数据集,这在一定程度上限制了该领域的发展,受限于医学影像和遥感等具体的应用领域,如何在有限标注数据下学习得到高性能的跨模态图像特征点和描述子以及高精匹配方法值得更多的研究人员关注.

### 8.4 结合具体任务的图像特征点匹配

如第 3 节所述,尽管深度学习的出现改变了一些计算机视觉任务对特征点匹配的依赖,目前依然有多种不同的任务与图像匹配性能密切相关,但是这些任务之间又存在一定的差异,正如机器学习中的“无免费午餐”(No Free Lunch)理论,笔者认为,我们也不能获得对任何问题都具有最优匹配能力的图像匹配方法,举个简单的例子,当我们面对的任务都是正常拍摄角度的图像,那么我们采用具有旋转不变性的特征描述子进行特征点匹配就不如不考虑旋转不变性的特征描述子,因为旋转不变性特征描述子在设计时为了提升其对于旋转的不变性会牺牲其对于具有旋转差异的局部图像内容的区分能力.因此,结合图像匹配具有的终端任务进行特征点和描述子学习具有重要的意义,在笔者看来,从“通用化”的特征匹配到“定制化”转变也将是该领域未来的重要发展方向之一.实际上,已有的多篇关于特征描述子的综述论文已经发现了这一问题[7,65,64],并且已经有研究人员开始提出面向下游任务的局部图像特征学习[130]与在线选择方法[246],将来有望出现更多这一方向的相关工作,针对不同的下游任务学习得到特定的图像匹配方法以在不同的应用中获得更好的性能.

### 8.5 自监督图像匹配方法学习

为了提高图像特征表达能力,不断发展的深度

网络结构需要越来越多的训练数据,显然,仅靠人工标注的方式很难满足深度学习日益增长的对高质量训练数据的要求.为此,越来越多的研究人员开始关注如何从无标签数据中学习深度网络的参数[73,93,94],其中,构建辅助任务从海量无标签图像进行网络预训练的自监督学习方法已经在图像识别等任务中展现出较好的性能[247,248],在有些情况甚至展现出超越全监督学习方法的性能.不可否认,依赖辅助任务构建的自监督学习方法得到的预训练深度网络可以直接作为图像匹配方法中的骨干特征提取网络,如现有方法中广泛采用的 VGG、ResNet 预训练网络,如何针对图像匹配及其下游任务构建更有针对性的代理辅助任务进行网络预训练值得期待.

另一方面,由于图像对应点级别的标注难以获得,图像匹配领域已有研究人员提出基于图像几何变换和内容风格迁移的自监督学习方法[72,110,124,221],这些方法展现出的图像匹配性能证明了这一技术路线的可行性.现阶段,这类方法的缺点主要在于仿射/单应等几何变换会丢失图像中的三维信息导致部分监督信息无效或者难以学习处理遮挡问题,而且现有的风格迁移技术仍不够成熟难以建模实际场景因不同气候、时间、人类活动等引起的复杂图像光度学变化,此外,如何结合几何变换和内容风格迁移生成符合真实场景成像效果的图像的技术仍不成熟.幸运的是,近年来以神经辐射场(Neural Radiance Fields, NeRF)为代表的三维场景建模[249]、人工智能内容生成(AIGC)[250,251]得到了广泛研究,将来有望利用这些技术生成高质量匹配图像以更好地进行自监督图像匹配方法学习.

## 9 总　结

本文围绕图像匹配方法的三个方面,即:稀疏特征点检测与描述、稠密像素点匹配、误匹配点滤除,系统总结了近年来该领域的研究进展,对比分析了典型代表性方法的特点、关键技术及性能,介绍了现阶段基于图像匹配的典型应用并给出了现状分析,总结了图像匹配与下游任务相关的数据集,展望了图像匹配领域现存的难点及其未来发展趋势.本文可帮助诸多相关领域的研究人员和工程技术人员快速了解和掌握图像匹配的内涵、难点、关键技术、研究现状等,并为进入该领域的研究人员提供研究方向与数据集资源等方面的参考.

## 参　考　文　献


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

[2] Tola Engin, Lepetit Vincent, Fua Pascal. DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intellignece, 2010, 32(5): 815-830

[3] Jeon Sangryul, Kim Seungryong, Min Dongbo, Sohn Kwanghoon. Pyramidal semantic correspondence networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9102-9118

[4] He Liu, Tao Wang, Yidong Li, Congyan Lang, Yi Jin, Haibin Ling. Joint graph learning and matching for semantic feature correspondence. Pattern Recognition, 2023, 134: 109059

[5] Hartley Richard, Zisserman Andrew. Multiple view geometry in computer vision (2nd Edition). Cambridge, USA: Cambridge University Press, 2004

[6] Wu Fuchao. Mathematical methods in computer vision. Beijing: Science Press, 2008(in Chinese)
(吴福朝. 计算机视觉中的数学方法. 北京: 科学出版社, 2008)

[7] Fan Bin, Kong Qingqun, Wang Xinchao, Wang Zhiheng, Xiang Shiming, Pan Chunhong, Fua Pascal. A performance evaluation of local features for image based 3D reconstruction. IEEE Transactions on Image Processing, 2019, 28(10): 4774-4789

[8] Li Zhao-Xin, Jiang Hao, Liu Yan-Qing, Wang Zhao-Qi. Research on multi-view stereo 3D reconstruction in virtual reality system of silk road cultural inheritance. Chinese Journal of Computers, 2022, 45(3): 500-512(in Chinese)
(李兆歆, 蒋浩, 刘衍青, 王兆其. 丝路文化虚拟体验中的多视角立体重建技术研究. 计算机学报, 2022, 45(3): 500-512)

[9] Liu Jinbo, Guo Pengyu, Li Xin, Zhang Xiaohu. Evaluation strategy for camera pose estimation algorithm based on point correspondences. Acta Optica Sinica, 2016, 36(5): 129-138(in Chinese)
(刘进博, 郭鹏宇, 李鑫, 张小虎. 基于点对应的相机姿态估计算法性能评价. 光学学报, 2016, 36(5): 129-138)

[10] Tofl Carl, Maddern Will, Torii Akihiko, Hammarstrand Lars, Stenborg Erik, Safari Daniel, Okutomi Masatoshi, Pollefeys Marc, Sivic Josef, Pajdla Tomas, Kahl Fredrik, Sattler Torsten. Long-term visual localization revisited. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(4): 2074-2088

[11] Liu Chuanjin, Liu Hongmin, Zhang Lixin, Zeng Hui, Luo Lufeng, Fan Bin. Learning task-aligned local features for visual localization. IEEE Robotics and Automation Letters, 2023, 8(6): 3366-3373

[12] Brown Matthew, Lowe David. Automatic panoramic image

stitching using invariant features. International Journal of Computer Vision, 2007, 74: 59-73

[13] Sarlin Paul-Edouard, Dusmanu Mihai, Schönberger Johannes L, Speciale Pablo, Gruber Lukas, Larsson Viktor, Miksik Ondrej, Pollefeys Marc. LaMAR: Benchmarking localization and mapping for augmented reality // Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 686-704

[14] Mur-Artal Raul, Montiel J. M. M., Tardos Juan D. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163

[15] Xiu Haixing, Liang Yiyou, Zeng Hui, Li Qing, Liu Hongmin, Fan Bin, Li Chen. Robust self-supervised monocular visual odometry based on prediction-update pose estimation network. Engineering Applications of Artificial Intelligence, 2022, 116(105481):1-11

[16] Zhao Hong-Ru, Qiao Xiu-Quan, Tan Zhi-Jie, Li Yan, Sun Heng. Loosely coupled visual-inertial odometry based on spatial-temporal two-stream convolution and long short-term memory networks. Chinese Journal of Computers, 2022, 45(8): 1674-1686(in Chinese)
(赵鸿儒, 乔秀全, 谭志杰, 李研, 孙恒. 基于时空双流卷积和长短期记忆网络的松耦合视觉惯性里程计. 计算机学报, 2022, 45(8): 1674-1686)

[17] Zhang Tian-Xu, Liu Jin. Moment invariant stability and its features for object recognition. Chinese Journal of Computers, 2004, 27(10): 1335-1340(in Chinese)
(张天序, 刘进. 不变矩稳定性及基于多级特征模型的目标识别研究. 计算机学报, 2004, 27(10): 1335-1340)

[18] Jiang Wen-Tao, Liu Wan-Jun, Yuan Heng. Research of object tracking based on soft feature theory. Chinese Journal of Computers, 2016, 39(7): 1334-1355(in Chinese)
(姜文涛, 刘万军, 袁姮. 基于软特征理论的目标跟踪研究. 计算机学报, 2016, 39(7): 1334-1355)

[19] Tran Son, Davis Larry. Robust object tracking with regional affine invariant features // Proceedings of the IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007:1-8

[20] Li Hong, Liu Fang, Yang Shu-Yuan, Zhang Kai. Remote sensing image fusion based on deep support value learning networks. Chinese Journal of Computers, 2016, 39(8): 1583-1596(in Chinese)
(李红, 刘芳, 杨淑媛, 张凯. 基于深度支撑值学习网络的遥感图像融合. 计算机学报, 2016, 39(8): 1583-1596)

[21] Fan Bin, Huo Chunlei, Pan Chunhong, Kong Qingqun. Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT. IEEE Geoscience and Remote Sensing Letters, 2013, 10(4): 657-661

[22] Goubran Maged, Leuze Christoph, Hsueh Brian, Aswendt Markus, Ye Li, Tian Qiyuan, Cheng Michelle Y, Crow Ailey, Steinberg Gary K, McNab Jennifer A, Deisseroth Karl, Zeineh Michael. Multimodal image registration and connectivity analysis for integration of connectomic data from microscopy to MRI. Nature Communication, 2019, 10: 5504

[23] Zhou Yuan, Rangarajan Anand, Gader Paul D. An integrated approach to registration and fusion of hyperspectral and multispectral Images. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(5): 3020-3033

[24] Bay Herbert, Ess Andreas, Tuytelaars Tinne, Gool Luc Van. SURF: Speeded up robust features. Computer Vision and Image Understanding, 2008, 110(3): 346-359

[25] Dalal Navneet, Triggs Bill. Histograms of oriented gradients for human detection // Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

[26] Pang Yanwei, Yuan Yuan, Li Xuelong, Pan Jing. Efficient HOG human detection. Signal Processing, 2011, 91(4): 773-781

[27] Felzenszwalb Pedro F., Girshick Ross B., McAllester David, Ramanan Deva. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645

[28] Zhang Junge, Huang Kaiqi, Yu Yinan, Tan Tieniu. Boosted local structured HOG-LBP for object localization // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1393-1400

[29] Sivic Josef, Zisserman Andrew. Video google: A text retrieval approach to object matching in videos // Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 1470-1477

[30] Grauman K, Darrell T. The pyramid match kernel: discriminative classification with sets of image features // Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005: 1458-1465

[31] Mur-Artal Raul, Tardos Juan D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Transactions on Robotics, 2017, 31(5): 1147-1163

[32] Campos Carlos, Elvira Richard, Rodríguez Juan J. Gómez, Montiel José M. M., Tardos Juan D. ORB-SLAM3: An accurate open-source library for visual, visual-inertial. and multimap SLAM. IEEE Transactions on Robotics, 2021, 37(6): 1874-1890

[33] Rublee Ethan, Rabaud Vincent, Konolige Kurt, Bradski Gary. ORB: An efficient alternative to SIFT or SURF // Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2564-2571

[34] Gool Luc Van, Moons T, Ungureanu D. Affine/photometric invariants for planar intensity patterns // Proceedings of the European Conference on Computer Vision. Berlin, Germany, 1996: 642-651

[35] Schmid C, Mohr R. Local gray value invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(5): 530-535

[36] Lazebnik S, Schmid C, Ponce J. A sparse texture representation using affine-invariant regions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003: 319-324

[37] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509-522

[38] Heikkilä Marko, Pietikäinen Matti, Schmid Cordelia. Description of interest regions with local binary patterns. Pattern Recognition, 2009, 42(3): 425-436

[39] Fan Bin, Wu Fuchao, Hu Zhanyi. Rotationally invariant descriptors using intensity order pooling. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2012, 34(10): 2031-2045

[40] Wang Zhenhua, Fan Bin, Wang Gang, Wu Fuchao. Exploring local and overall ordinal information for robust feature description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(11): 2198-2211

[41] Alahi Alexandre, Ortiz Raphael, Vandergheynst Pierre. FREAK: Fast retina keypoint // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 510-517

[42] Trzcinski Tomasz, Christoudias Mario, Lepetit Vincent. Learning image descriptors with boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 597-610

[43] Ke Y., Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004:1-8

[44] Brown Matthew, Hua Gang, Winder Simon. Discriminative learning of local image descriptors. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2011, 33(1): 43-57

[45] Simonyan Karen, Vedaldi Andrea, Zisserman Andrew. Learning local feature descriptors using convex optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1573-1585

[46] Mikolajczyk K., Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615-1630

[47] Balntas Vassileios, Lenc Karel, Vedaldi Andrea, Mikolajczyk Krystian. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3852-3861

[48] Ma Jiayi, Jiang Xingyu, Fan Aoxiang, Jiang Junjun, Yan Junchi. Image matching from handcrafted to deep features: a survey. International Journal of Computer Vision, 2021, 129: 23-79

[49] Simonyan Karen and Zisserman Andrew. Very deep convolutional networks for large-scale image recognition // Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-14

[50] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz, Polosukhin Illia. Attention is all you need // Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 1-11

[51] Zhang Zhao, Wei Yanyan, Zhang Haijun, Yang Yi, Yan Shuicheng, Wang Meng. Data-Driven Single Image Deraining: A Comprehensive Review and New Perspectives. Pattern Recognition, 2023, 143: 109740

[52] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[53] Barroso-Laguna Axel and Mikolajczyk Krystian. Key.Net: Keypoint detection by handcrafted and learned CNN filters revisited. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 698-711

[54] Tian Yurun, Fan Bin, Wu Fuchao. L2-Net: Deep learning of discriminative patch descriptor in Euclidean space // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6128-6136

[55] Sun Jiaming, Shen Zehong, Wang Yuang, Bao Hujun, Zhou Xiaowei. LoFTR: Detector-free local feature matching with transformers // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 8918-8927

[56] Liu Xin, Xiao Guobao, Chen Riqing, Ma Jiayi. PGFNet: Preference-guided filtering network for two-view correspondence learning. IEEE Transactions on Image Processing, 2023, 32: 1367-1378

[57] Luo Zixin, Zhou Lei, Bai Xuyang, Chen Hongkai, Zhang Jiahui, Yao Yao, Li Shiwei, Fang Tian, Quan Long. ASLFeat: Learning local features of accurate shape and localization // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6588-6597

[58] Tian Yurun, Barroso-Laguna Axel, Ng Tony, Balntas Vassileios, Mikolajczyk Krystian. HyNet: Local descriptor with hybrid similarity measure and triplet loss // Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2020: 7401-7412

[59] Sarlin Paul-Edouard, DeTone Daniel, Malisiewicz Tomasz, Rabinovich Andrew. SuperGlue: Learning feature matching with graph neural networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4938-4947

[60] Teed Zachary and Deng Jia. DROID-SLAM: Deep visual SLAM for monocular, stereo, RGB-D cameras // Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 16558-16569

[61] Jiang Xingyu, Ma Jiayi, Xiao Guobao, Shao Zhenfeng, Guo

Xiaojie. A review of multimodal image matching: Methods and applications. Information Fusion, 2021, 73: 22-71

[62] Lin Su-Zhen, Han Ze. Image fusion based on deep stack convolutional neural network. Chinese Journal of Computers, 2017, 40(11): 2506-2518(in Chinese)
(蔺素珍, 韩泽. 基于深度堆叠卷积神经网络的图像融合. 计算机学报, 2017, 40(11): 2506-2518)

[63] Jing Junfeng, Gao Tian, Zhang Weichuan, Gao Yongsheng, Sun Changming. Image feature information extraction for interest point detection: A comprehensive review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4694-4712

[64] Jin Yuhe, Mishkin Dmytro, Mishchuk Anastasiia, Matas Jiri, Fua Pascal, Yi Kwang Moo, Trulls Eduard. Image matching across wide baselines: From paper to practice. International Journal of Computer Vision, 2021, 129: 517-547

[65] Schönberger Johannes L, Hardmeier Hans, Sattler Torsten, Pollefeys Marc. Comparative evaluation of hand-crafted and learned local features // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6959-6968

[66] Redmon Joseph, Divvala Santosh, Girshick Ross, Farhadi Ali. You only look once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[67] Zou Zhengxia, Chen Keyan, Shi Zhenwei, Guo Yuhong, Ye Jieping. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023, 111(3): 257-276

[68] Zhang Zhen and Lee Wee Sun. Deep graphical feature learning for the feature matching problem // Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5087-5096

[69] Rolinek Michal, Swoboda Paul, Zietlow Dominik, Paulus Anselm, Musil Vit, Martius Georg. Deep graph matching via blackbox differentiation of combinatorial solvers // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020:407-424

[70] Gao Quankai, Wang Fudong, Xue Nan, Yu Jin-Gang, Xia Gui-Song. Deep graph matching under quadratic constraint // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 5069-5078

[71] Liu He, Wang Tao, Li Yidong, Lang Congyan, Jin Yi, Ling Haibin. Joint graph learning and matching for semantic feature correspondence. Pattern Recognition, 2023, 134: 109059

[72] Truong Prune, Danelljan Martin, Timofte Radu. GLU-Net: Global-local universal network for dense flow and correspondences // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6258-6268

[73] Truong Prune, Danelljan Martin, Yu Fisher, Gool Luc Van. Warp consistency for unsupervised learning of dense correspondences // Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 10346-10356

[74] Zbontar Jure and LeCun Yann. Computing the stereo matching cost with a convolutional neural network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1592-1599

[75] Liang Zhengfa, Feng Yiliu, Guo Yulan, Liu Hengzhu, Chen Wei, Qiao Linbo, Zhou Li, Zhang Jianfeng. Learning for disparity estimation through feature constancy // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2811-2820

[76] Chen Zhuoyuan, Sun Xun, Wang Liang, Yu Yinan, Huang Chang. A deep visual correspondence embedding model for stereo matching costs // Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 972-980

[77] Rocco Ignacio, Cimpoi Mircea, Arandjelovic Relja, Torii Akihiko, Pajdla Tomas, Sivic Josef. Neighbourhood consensus networks // Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 1658-1669

[78] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, Juho Kannala. DGC-Net: Dense geometric correspondence network // Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2019: 1034-1042

[79] Truong Prune, Danelljan Martin, Gool Luc Van, Timofte Radu. GOCor: Bringing globally optimized correspondence volumes into your neural network // Proceedings of the Advances in Neural Information Processing Systems, Virtual, 2020: 14278-14290

[80] Truong Prune, Danelljan Martin, Gool Luc Van, Timofte Radu. Learning accurate dense correspondences and when to trust them // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2021: 5714-5724

[81] Zhang Weichuan, Sun Changming, Gao Yongsheng. Image intensity variation information for interest point detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9883-9894

[82] Zhang Weichuan, Sun Changming. Corner detection using multi-directional structure tensor with multiple scales. International Journal of Computer Vision, 2019, 128: 438-459

[83] Verdie Yannick, Yi Kwang Moo, Fua Pascal, Lepetit Vincent. TILDE: A temporally invariant learned detector // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5279-5288

[84] Lenc Karel and Vedaldi Andrea. Learning covariant feature detectors // Proceedings of the European Conference on Computer Vision Workshops. Amsterdam, The Netherlands, 2016: 100-117

[85] LeCun Y, Boser B, Denker J S, Henderson D, Howard R E,

Hubbard W, Jackel L D. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4):541-551

[86] Zhang Xu, Yu Felix X, Karaman Svebor, Chang Shi-Fu. Learning discriminative and transformation covariant local feature detectors// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6818-6826

[87] Savinov Nikolay, Seki Akihito, Ladicky Lubor, Sattler Torsten, Pollefeys Marc. Quad-networks: Unsupervised learning to rank for interest point detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1822-1830

[88] Mishkin Dmytro, Radenovic Filip, Matas Jiri. Repeatability is not enough: Learning affine regions via discriminability// Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 287-304

[89] Tian Yurun, Balntas Vassileios, Ng Tony, Barroso-Laguna Axel, Demiris Yiannis, Mikolajczyk Krystian. D2D: Keypoint extraction with describe to detect approach// Proceedings of the Asian Conference on Computer Vision. Kyoto, Japan, 2020: 223-240

[90] Elad Baruch Ben and Yosi Keller. Joint detection and matching of feature points in multimodal images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6585-6593

[91] Barroso-Laguna Axel, Verdie Yannick, Busam Benjamin, Krystian Mikolajczyk. HDD-Net: Hybrid detector descriptor with mutual interactive learning// Proceedings of the Asian Conference on Computer Vision. Kyoto, Japan, 2020: 500-516

[92] Li Kunhong, Wang Longguang, Liu Li, Ran Qing, Xu Kai, Guo Yulan. Decoupling makes weakly supervised local feature better// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15838-15848

[93] DeTone Daniel, Malisiewicz Tomasz, Rabinovich Andrew. Superpoint: Self-supervised interest point detection and description// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 224-236

[94] Peter Hviid Christiansen, Mikkel Fly Kragh, Yury Brodskiy, Henrik Karstoft. Unsuperpoint: End-to-end unsupervised interest point detector and descriptor. arXiv preprint arXiv: 1907.04011, 2019

[95] Han Xufeng, Leung Thomas, Jia Yangqing, Sukthankar Rahul, Berg Alexander C. MatchNet: Unifying feature and metric learning for patch-based matching// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3279-3286

[96] Zagoruyko Sergey and Komodakis Nikos. Learning to compare image patches via convolutional neural networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4353-4361

[97] Silpa-Anan C. and Hartley R. Optimised kd-trees for fast image descriptor matching// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008:1-8

[98] Muja M. and Lowe D. Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2227-2240

[99] Simo-Serra Edgar, Trulls Eduard, Ferraz Luis, Kokkinos Iasonas, Fua Pascal, Moreno-Noguer Francesc. Discriminative learning of deep convolutional feature point descriptors// Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 118-126

[100] Balntas Vassileios, Riba Edgar, Ponsa Daniel, Mikolajczyk Krystian. Learning local feature descriptors with triplets and shallow convolutional neural networks// Proceedings of the British Machine Vision Conference. New York, USA, 2016:1-11

[101] Mishchuk Anastasiya, Mishkin Dmytro, Radenovic Filip, Matas Jiri. Working hard to know your neighbor's margins: local descriptor learning loss// Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 4829-4840

[102] Zhang Linguang and Rusinkiewicz Szymon. Learning local descriptors with a CDF-based dynamic soft margin// Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 2969-2978

[103] Wang Song, Guo Xin, Tie Yun, Qi Lin, Guan Ling. Deep local feature descriptor learning with dual hard batch construction. IEEE Transactions on Image Processing, 2020, 29: 9572-9583

[104] Ebel Patrick, Mishchuk Anastasiia, Yi Kwang Moo, Fua Pascal, Trulls Eduard. Beyond cartesian representations for local descriptors// Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 253-262

[105] He Kun, Lu Yan, Sclaroff Stan. Local descriptors optimized for average precision// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 596-605

[106] Zhang Xu, Yu Felix X., Kumar Sanjiv, Chang Shih-Fu. Learning spread-out local feature descriptors// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4605-4613

[107] Tian Yurun, Yu Xin, Fan Bin, Wu Fuchao, Heijnen Huub, Balntas Vassileios. SOSNet: Second order similarity regularization for local descriptor learning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11008-11017

[108] Lin Kevin, Lu Jiwen, Chen Chu-Song, Zhou Jie, Sun Ming-

Ting. Unsupervised deep learning of compact binary descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1501-1514

[109] Wu Gengshen, Lin Zijia, Ding Guiguang, Ni Qiang, Han Jungong. On aggregation of unsupervised deep binary descriptor with weak bits. IEEE Transactions on Image Processing, 2020, 29: 9266-9278

[110] Fan Bin, Liu Hongmin, Zeng Hui, Zhang Jiyong, Liu Xin, Han Junwei. Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness. IEEE Transactions on Multimedia, 2021, 23: 2770-2781

[111] Wang Ziwei, Xiao Han, Duan Yueqi, Zhou Jie, Lu Jiwen. Learning deep binary descriptors via bitwise interaction mining. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022

[112] Luo Zixin, Shen Tianwei, Zhou Lei, Zhu Siyu, Zhang Runze, Yao Yao, Fang Tian, Quan Long. GeoDesc: Learning local descriptors by integrating geometry constraints // Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 170-185

[113] Wang Qianqian, Zhou Xiaowei, Hariharan Bharath, Snavely Noah. Learning feature descriptors using camera pose supervision // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 757-774

[114] Germain Hugo, Bourmaud Guillaume, Lepetit Vincent. S2DNet: Learning accurate correspondences for sparse-to-dense feature matching // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 626-643

[115] Luo Zixin, Shen Tianwei, Zhou Lei, Zhang Jiahui, Yao Yao, Li Shiwei, Fang Tian, Quan Long. ContextDesc: Local descriptor augmentation with cross-modality context // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2522-2531

[116] Cai Youcheng, Li Lin, Wang Dong, Li Xinjie, Liu Xiaoping. HTMatch: An efficient hybrid transformer based graph neural network for local feature matching. Signal Processing, 2023, 204: 108859

[117] Chen Hongkai, Luo Zixin, Zhang Jiahui, Zhou Lei, Bai Xuyang, Hu Zeyu, Tai Chiew-Lan, Quan Long. Learning to match features with seeded graph matching network // Proceedings of the International Conference on Computer Vision. Montreal, Canada, 2021: 6301-6310

[118] Shi Yan, Cai Jun-Xiong, Shavit Yoli, Mu Tai-Jiang, Feng Wensen, Zhang Kai. ClusterGNN: Cluster-based coarse-to-fine graph neural network for efficient feature matching // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 12517-12526

[119] Yi Kwang Moo, Trulls Eduard, Lepetit Vincent, Fua Pascal. LIFT: Learned invariant feature transform // Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 467-483

[120] Ono Yuki, Trulls Eduard, Fua Pascal, Yi Kwang Moo. LF-Net: Learning local features from images // Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 6237-6247

[121] Shen Xuelun, Wang Cheng, Li Xin, Yu Zenglei, Li Jonathan, Wen Chenglu, Cheng Ming, He Zijian. RF-Net: An end-to-end image matching network based on receptive field // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8124-8132

[122] Dusmanu M., Rocco I., Pajdla T., Pollefeys M., Sivic J., Torii A., Sattler T. D2-Net: A trainable CNN for joint detection and description of local features // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8084-8093

[123] Tyszkiewicz Michał J., Fua Pascal, Trulls Eduard. DISK: Learning local features with policy gradient // Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 14254-14265

[124] Revaud Jerome, Weinzaepfel Philippe, de Souza Csar Roberto, Humenberger Martin. R2D2: Repeatable and reliable detector and descriptor // Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 12414-12424

[125] Zhao Xiaoming, Wu Xingming, Miao Jinyu, Chen Weihai, Chen Peter C. Y., Li Zhengguo. ALIKE: Accurate and lightweight keypoint detection and descriptor extraction. IEEE Transactions on Multimedia, 2023,25: 3101-3112

[126] Fan Bin, Yang Yuzhu, Feng Wensen, Wu Fuchao, Lu Jiwen, Liu Hongmin. Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features. IEEE Transactions on Multimedia, 2023, 25: 1713-1726

[127] Fan Bin, Zhou Junjie, Feng Wensen, Pu Huayan, Yang Yuzhu, Kong Qingqun, Wu Fuchao, Liu Hongmin. Learning semantic-aware local features for long term visual localization. IEEE Transactions on Image Processing, 2022, 31: 4842-4855

[128] Wang Changwei, Xu Rongtao, Zhang Yuyang, Xu Shibiao, Meng Weiliang, Fan Bin, Zhang Xiaopeng. MTLDesc: Looking wider to describe better // Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 2388-2396

[129] Ronneberger, O., Fischer, P., Brox, T. U-net: Convolutional networks for biomedical image segmentation // Proceedings of the Medical Image Computing and Computer Assisted Intervention. 2015: 234-241

[130] Bhowmik Aritra, Gumhold Stefan, Rother Carsten, Brachmann Eric. Reinforced feature points: Optimizing feature detection and description for a high-level task // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4947-4956

[131] Taira H., Okutomi M., Sattler T., Cimpoi M., Pollefeys

M., Sivic J., Pajdla T., Torii A. InLoc: Indoor visual localization with dense matching and view synthesis// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1293-1307

[132] Sarlin Paul-Edouard, Cadena Cesar, Siegwart Roland, Dymczyk Marcin. From coarse to fine: Robust hierarchical localization at large scale // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12716-12725

[133] Visual localization benchmark-local features, https: // github. com/tsattler/visuallocalizationbenchmark/tree/master/local_feature_evaluation

[134] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, y Ding, Francis X Creighton, Russell H Taylor, Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers // Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 6197-6206

[135] Luo Wenjie, Schwing Alexander G., Urtasun Raquel. Efficient deep learning for stereo matching // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5695-5703

[136] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, cost volume// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8934-8943

[137] Yao Yao, Luo Zixin, Li Shiwei, Fang Tian, Quan Long. MVSNet: Depth inference for unstructured multi-view stereo// Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 785-801

[138] Yao Yao, Luo Zixin, Li Shiwei, Shen Tianwei, Fang Tian, Quan Long. Recurrent MVSNet for high-resolution multi-view stereo depth inference// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5520-5529

[139] Chen Rui, Han Songfang, Xu Jing, Su Hao. Point-based multi-view stereo network// Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 1538-1547

[140] Gu Xiaodong, Fan Zhiwen, Zhu Siyu, Dai Zuozhuo, Tan Feitong, Tan Ping. Cascade cost volume for high-resolution multi-view stereo and stereo matching // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 2492-2501

[141] Ma Xinjun, Gong Yue, Wang Qirui, Huang Jingwei, Chen Lei, Yu Fan. EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo// Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021:5732-5740

[142] Edstedt Johan, Athanasiadis Ioannis, Wadenback Marten, Felsberg Michael. DKM: Dense kernelized feature matching for geometry estimation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 17765-17775

[143] Rocco Ignacio, Arandjelovic Relja, Sivic Josef. Efficient neighbourhood consensus networks via submanifold sparse convolutions // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 605-621

[144] Li Xinghui, Han Kai, Li Shuda, Prisacariu Victor. Dual-resolution correspondence networks // Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 17346-17357

[145] Bokman Georg and Kahl Fredrik. A case for using rotation invariant features in state of the art feature matchers// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, USA, 2022: 5110-5119

[146] Liu Yuan, Shen Zehong, Lin Zhixuan, Peng Sida, Bao Hujun, Zhou Xiaowei. GIFT: Learning transformation-invariant dense visual descriptors via group cnns// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 6992-7003

[147] Chen Hongkai, Luo Zhixin, Zhou Lei, Tian Yurun, Zhen Mingmin, Fang Tian, McKinnon David, Tsin Yanghai, Quan Long. ASpanFormer: Detector-free image matching with adaptive span transformer// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 20-36

[148] Tang Shitao, Zhang Jiahui, Zhu Siyu, Tan Ping. Quadtree attention for vision transformers// Proceedings of the International Conference on Learning Representations. Virtual, 2022:1-16

[149] Wang Qing, Zhang Jiaming, Yang Kailun, Peng Kunyu, Stiefelhagen Rainer. Matchformer: Interleaving attention in transformers for feature matching// Proceedings of the Asian Conference on Computer Vision, 2022: 2746-2762

[150] Zhou Qunjie, Sattler Torsten, Leal-Taixe Laura. Patch2Pix: Epipolar-guided pixel-level correspondences // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 4667-4676

[151] Yi Kwang Moo, Trulls Eduard, Ono Yuki, Lepetit Vincent, Salzmann Mathieu, Fua Pascal. Learning to find good correspondences // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2666-2674

[152] Zhao Chen, Cao Zhiguo, Li Chi, Li Xin, Yang Jiaqi. NM-Net: Mining reliable neighbors for robust feature correspondences// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 215-224

[153] Liu Yuan, Liu Lingjie, Lin Cheng, Dong Zhen, Wang Wenping. Learnable motion coherence for correspondence pruning// Proceedings of the IEEE Conference on Computer Vi-

sion and Pattern Recognition. Nashville, USA, 2021: 3237-3246

[154] Zhang Jiahui, Sun Dawei, Luo Zixin, Yao Anbang, Zhou Lei, Shen Tianwei, Chen Yurong, Quan Long, Liao Hongen. Learning two-view correspondences and geometry using order-aware network// Proceedings of the IEEE International Conference on Computer Vision. Long Beach, USA, 2019: 5845-5854

[155] Sun Weiwei, Jiang Wei, Trulls Eduard, Tagliasacchi Andrea, Yi Kwang Moo. ACNe: Attentive context normalization for robust permutation-equivariant learning // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11283-11292

[156] Zhao Chen, Ge Yixiao, Zhu Feng, Zhao Rui, Li Hongsheng, Salzmann Mathieu // Progressive correspondence pruning by consensus learning. Proceedings of the IEEE International Conference on Computer Vision. Nashville, USA, 2021: 6444-6453

[157] Ma Jiayi, Wang Yang, Fan Aoxiang, Xiao Guobao, Chen Riqing. Correspondence attention transformer: A context-sensitive network for two-view correspondence learning. IEEE Transactions on Multimedia, 2023, 25: 3509-3524

[158] Philbin James, Chum Ondrej, Sivic Josef, Isard Michael, Coto Ernesto, Zisserman Andrew. Object retrieval with large vocabularies and fast spatial matching// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[159] Kalal Z., Matas J., Mikolajczyk K. Tracking-Learning-Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422

[160] Lepetit Vincent and Fua Pascal. Keypoint recognition using randomized trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(9): 1465-1479

[161] Li Dongjiang, Shi Xuesong, Long Qiwei, Liu Shenghui, Yang Wei, Wang Fangshi, Wei Qi, Qiao Fei. DXSLAM: A robust and efficient visual SLAM system with deep features // Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, USA, 2020: 4958-4965

[162] Munoz-Salinas Rafael and Rafael Medina-Carnicer. UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers. Pattern Recognition, 2020, 101: 107193

[163] Schonberger Johannes L. and Frahm Jan-Michael. Structure-from-Motion revisited // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4104-4113

[164] Doan Anh-Dzung, Turmukhambetov Daniyar, Latif Yasir, Chin Tat-Jun, Bae Soohyun. Learning to predict repeatability of interest points// Proceedings of the IEEE International Conference on Robotics and Automation. Xi'an, China, 2021: 10294-10301

[165] Chang Ming-Fang, Zhao Yipu, Shah Rajvi, Engel Jakob J., Kaess Michael, Lucey Simon. Long-term visual map sparsification with heterogeneous GNN// Proceedings of the International Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 2406-2415

[166] Lee Kyu-Yul and Sim Jae-Young. Warping residual based image stitching for large parallax// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8195-8203

[167] Wan Qi, Chen Jun, Luo Linbo, Gong Wenping, Wei Longsheng. Drone image stitching using local mesh-based bundle adjustment and shape-preserving transform. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(8): 7027-7037

[168] Liu Li, Ouyang Wanli, Wang Xiaogang, Fieguth Paul, Chen Jie, Liu Xinwang, Pietikainen Matti. Deep learning for generic object detection: A survey. International Journal of Computer Vision, 2020, 128: 261-318

[169] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 2687-2704

[170] Li Bo, Yan Junjie, Wu Wei, Zhu Zheng, Hu Xiaolin. High performance visual tracking with siamese region proposal network// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8971-8980

[171] Fu Zhihong, Liu Qingjie, Fu Zehua, Wang Yunhong. STMTrack: Template-free visual tracking with space-time memory networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 13769-13778

[172] Snavely Noah, Seitz Steven M., Szeliski Richard. Photo tourism: Exploring image collections in 3D// Proceedings of the International Conference on Computer Graphics and Interactive Techniques. New York, USA, 2006: 835-846

[173] Wu Changchang. Towards linear-time incremental structure from motion // Proceedings of the International Conference on 3D Vision. Seattle, USA, 2013: 127-134

[174] Moulon Pierre, Monasse Pascal, Perrot Romuald, Marlet Renaud. OpenMVG: Open multiple view geometry // Proceedings of the International Workshop on Reproducible Research in Pattern Recognition. Cancún, Mexico, 2016: 60-74

[175] Furukawa Yasutaka and Ponce Jean. Accurate, dense, robust multi-view stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376

[176] Zhang Zhe, Peng Rui, Hu Yuxi, Wang Ronggang. GeoMVSNet: Learning multi-view stereo with geometry perception// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 21508-21518

[177] Wang Changwei, Xu Rongtao, Lu Ke, Xu Shibiao, Meng Weiliang, Zhang Yuyang, Fan Bin, Zhang Xiaopeng. At-

tention weighted local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10632-10649

[178] Zhao Xiaoming, Wu Xingming, Chen Weihai, Chen Peter C Y, Xu Qingsong, Li Zhengguo. ALIKED: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 5014016

[179] https://image-matching-workshop.github.io

[180] Wang Jian-Shu, Yin Kui-Ying, Yin Jin-Rong, Huang Zhao-You, Wang Zhong-Bao. A navigation method based on the inertial navigation information, satellite positioning and scene matching points. Command Control and Simulation, 2023, online(in Chinese)
(王剑书, 尹奎英, 尹锦荣, 黄照悠, 王中宝. 基于惯导信息卫星定位与景象匹配的导航方法. 指挥控制与仿真, 2023)

[181] Li Meng-Yuan, Zhou Feng-Yu, Tian Tian, Yin Lei, Shen Dong-Dong, Wang Shu-Qian. Design of the indoor WiFi cloud positioning system based on GAN and its application to service robots. Robot, 2018, 40(5): 693-703(in Chinese)
(栗梦媛, 周风余, 田天, 尹磊, 沈冬冬, 王淑倩. 基于 GAN 的服务机器人室内 WiFi 云定位系统设计与实现. 机器人, 2018, 40(5): 693-703)

[182] Sarlin Paul-Edouard, Dusmanu Mihai, Schönberger Johannes L, Speciale Pablo, Gruber Lukas, Larsson Viktor, Miksik Ondrej, Pollefeys Marc. Lamar: Benchmarking localization and mapping for augmented reality // Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 686-704

[183] Arandjelovic Relja, Gronat Petr, Torii Akihiko, Pajdla Tomas, Sivic Josef. NetVLAD: CNN architecture for weakly supervised place recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5297-5307

[184] Hausler Stephen, Garg Sourav, Xu Ming, Milford Michael, Fischer Tobias. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition // Proceedings of the International Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 14141-14152

[185] Hou Qianbao, Xia Rui, Zhang Jiahuan, Feng Yu, Zhan Zongqian, Wang Xin. Learning visual overlapping image pairs for SfM via CNN fine-tuning with photogrammetric geometry information. International Journal of Applied Earth Observation and Geoinformation, 2023, 116: 103162

[186] Kendall Alex, Grimes Matthew, Cipolla Roberto. PoseNet: a convolutional network for real-time 6-DOF camera relocalization // Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2938-2946

[187] Yin Jiahao, Zhou Xinyu, Xiao Huahui, Liu Zhili, Li Wei, Li Xue, Fan Shengyin. HAPOR: Hierarchical-features aligned projection optimization for relocalization. IEEE Robotics and Automation Letters, 2023, 8(3): 1447-1454

[188] Torii Akihiko, Arandjelovic Relja, Sivic Josef, Okutomi Masatoshi, Pajdla Tomas. 24/7 place recognition by view synthesis // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1808-1817

[189] Sattler Torsten, Zhou Qunjie, Pollefeys Marc, Leal-Taixe Laura. Understanding the limitations of CNN-based absolute camera pose regression // Proceedings of the International Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3302-3312

[190] Lepetit Vincent, Moreno-Noguer Francesc, Fua Pascal. EPnP: Efficient perspective-n-point camera pose estimation. International Journal of Computer Vision, 2009, 81: 155-166

[191] Sarlin Paul-Edouard, Unagar Ajaykumar, Larsson Mans, Germain Hugo, Toft Carl, Larsson Viktor, Pollefeys Marc, Lepetit Vincent, Hammarstrand Lars, Kahl Fredrik, Sattler Torsten. Back to the feature: Learning robust camera localization from pixels to pose // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021:3247-3257

[192] Humenberger Martin, Cabon Yohann, Guerin Nicolas, Morat Julien, Leroy Vincent, Revaud Jerome, Rerole Philippe, Pion Noe, Souza Cesar de, Csurka Gabriela. Robust image retrieval-based visual localization using kapture. arXiv: 2007.13867v3

[193] Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M. R. and Pollefeys, M. NICE-SLAM: Neural implicit scalable encoding for slam // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 12786-12796

[194] Zou Yuliang, Ji Pan, Tran Quoc-Huy, Huang Jia-Bin, Chandraker Manmohan. Learning monocular visual odometry via self-supervised long-term modeling // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 710-727

[195] Tiwari Lokender, Ji Pan, Tran Quoc-Huy, Zhuang Bingbing, Anand Saket, Chandraker Manmohan. Pseudo RGB-D for self-improving monocular slam and depth prediction // Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 437-455

[196] Sucar, E., Liu, S., Ortiz, J. and Davison, A. J. iMAP: Implicit mapping and positioning in real-time // Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 6229-6238

[197] Engel J., Schops T., Cremers D. LSD-SLAM: Large-scale direct monocular SLAM. Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 834-849

[198] Zubizarreta J., Aguinaga I., Montiel J. M. M. Direct sparse mapping. IEEE Transactions on Robotics, 2020, 36(4): 1363-1370

[199] Di Giammarino, Luca, Leonardo Brizi, Tiziano Guadagni-

no, Cyrill Stachniss, Giorgio Grisetti. MD-SLAM: Multi-cue direct SLAM// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, Japan, 2022: 11047-11054

[200] Yang Nan, Stumberg Lukas von, Wang Rui, Cremers Daniel. D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry// Proceedings of the International Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1281-1292

[201] Wu Junjun, Shi Qingwu, Lu Qinghua, Liu Xilin, Zhu Xiaoman, Lin Zeqin. Learning invariant semantic representation for long-term robust visual localization. Engineering Applications of Artificial Intelligence, 2022, 111: 104793

[202] Yin Jie, Li Ang, Li Tao, Yu Wenxian, Zou Danping. M2DGR: A multi-sensor and multi-scenario SLAM dataset for ground robots. IEEE Robotics and Automation Letters, 2022, 7(2): 2266-2273

[203] Jia Yupeng, Luo Haiyong, Zhao Fang, Jiang Guanlin, Li Yuhang, Yan Jiaquan, Jiang Zhuqing, Wang Zitian. Lvio-Fusion: A self-adaptive multi-sensor fusion SLAM framework using actor-critic method// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic, 2021: 286-293

[204] Chen Changhao, Rosa Stefano, Miao Yishu, Lu Chris Xiaoxuan, Wu Wei, Markham Andrew, Trigoni Niki. Selective sensor fusion for neural visual-inertial odometry// Proceedings of the International Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 10542-10551

[205] Li Zhengqi and Snavely Noah. Mega depth: Learning single-view depth prediction from internet photos// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2041-2050

[206] Shen Tianwei, Luo Zixin, Zhou Lei, Zhang Runze, Zhu Siyu, Fang Tian, Quan Long. Matchable image retrieval by learning from surface reconstruction// Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 415-431

[207] Zhu Siyu, Zhang Runze, Zhou Lei, Shen Tianwei, Fang Tian, Tan Ping, Quan Long. Very large-scale global SfM by distributed motion averaging// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4568-4577

[208] Li Shiwei, Siu Sing-Yu, Fang Tian, Quan Long. Efficient multi-view surface refinement with adaptive resolution control// Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 349-364

[209] Thomee B., Shamma D. A., Friedland G., Elizalde B., Ni K., Poland D., Borth D., Li L.-J. YFCC100M: The new data in multimedia research. Communications of the ACM, 2016, 59(2): 64-73

[210] Heinly Jared, Schonberger Johannes, Dunn Enrique, Frahm Jan-Michael. Reconstructing the world * in six days * (as captured by the yahoo 100 million image dataset)// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015:3287-3295

[211] Dai Angela, Chang Angel X., Savva Manolis, Halber Maciej, Funkhouser Thomas, Nießner Matthias. ScanNet: Richly-annotated 3D reconstructions of indoor scenes// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2432-2443

[212] 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 6DOF outdoor visual localization in changing conditions// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8601-8610

[213] Zhang Z., Sattler T., Scaramuzza D. Reference pose generation for visual localization via learned features and view synthesis. International Journal of Computer Vision, 2021, 129(4): 821-844

[214] Strecha Christoph, Hansen Wolfgang Von, Gool Luc Van, Fua Pascal, Thoennessen Ulrich. On benchmarking camera calibration and multi-view stereo for high resolution imagery// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008:1-8

[215] Hane C., Zach C., Cohen A., Angst R., Pollefeys M. Joint 3D scene reconstruction and class segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 97-104

[216] Wilson K. and Snavely N. Robust global translations with 1dsfm// Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 61-75

[217] Crandall D., Owens A., Snavely N., Huttenlocher D. P. Discrete-continuous optimization for large-scale structure from motion// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011:3001-3008

[218] Sturm J., Engelhard N., Endres F., Burgard W., Cremers D. A benchmark for the evaluation of RGB-D SLAM systems// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve, Portugal, 2012: 573-580

[219] Geiger A., Lenz P., Stiller C., Urtasun R. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 2013, 32(11): 1231-1237

[220] Burri M., Nikolic J., Gohl P., Schneider T., Rehder J., Omari S., Achtelik M. W., Siegwart R. The EuRoC micro aerial vehicle datasets. International Journal of Robotics Research, 2016, 35(10): 1157-1163

[221] Jerome Revaud, Vincent Leroy, Philippe Weinzaepfel, Boris Chidlovskii. PUMP: Pyramidal and uniqueness matching priors for unsupervised learning of local descriptors// Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3926-3936

[222] Jiang Wei, Trulls Eduard, Hosang Jan, Tagliasacchi Andrea, Yi Kwang Moo. COTR: Correspondence transformer for matching across images // Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 6207-6217

[223] Tan Dongli, Liu Jiang-Jiang, Chen Xingyu, Chen Chao, Zhang Ruixin, Shen Yunhang, Ding Shouhong, Ji Rongrong. ECO-TR: Efficient correspondences finding via coarse-to-fine refinement // Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 317-334

[224] Yan Pei, Tan Yihua, Xiong Shengzhou, Tai Yuan, Li Yansheng. Learning soft estimator of keypoint scale and orientation with probabilistic covariant loss // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19406-19415

[225] Quan Dou, Wang Shuang, Huyan Ning, Li Yi, Lei Ruiqi, Chanussot Jocelyn, Hou Biao, Jiao Licheng. A concurrent multiscale detector for end-to-end image matching. IEEE Transactions on Neural Networks and Learning Systems, 2024,35(3): 3560-3574

[226] Barroso-Laguna Axel, Tian Yurun, Mikolajczyk Krystian. ScaleNet: A shallow architecture for scale estimation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 12808-12818

[227] Guo Junwen, Xiao Guobao, Tang Zhimin, Chen Shunxing, Wang Shiping, Jiayi Ma. Learning for feature matching via graph context attention. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61:5102714

[228] Mohr Roger, Gros Patrick, Schmid Cordelia. Efficient matching with invariant local descriptors. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, 1998

[229] Fan Bin, Wu Fuchao, Hu Zhanyi. Towards Reliable matching of images containing repetitive patterns. Pattern Recognition Letters, 2011, 32(14): 1851-1859

[230] Zhou Feng and Torre Fernando De la. Factorized graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1774-1789

[231] Baillard Caroline, Schmid Cordelia, Zisserman Andrew, Fitzgibbon Andrew. Automatic line matching and 3D reconstruction of buildings from multiple views. ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery. Munich, Germany, 1999: 69-80

[232] Fan Bin, Wu Fuchao, Hu Zhanyi. Robust line matching through line-point invariants. Pattern Recognition, 2012, 45(2): 794-805

[233] Pautrat Remi, Lin Juan-Ting, Larsson Viktor, Oswald Martin R., Pollefeys Marc. SOLD2: Self-supervised occlusion-aware line description and detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 11368-11378

[234] Lange Manuel, Schweinfurth Fabian, Schilling Andreas. DLD: A deep learning based line descriptor for line feature matching // Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Macao, China, 2019: 5910-5915

[235] Lin Tsung-Yi, Dollar Piotr, Girshick Ross, He Kaiming, Hariharan Bharath, Belongie Serge. Feature pyramid networks for object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 936-944

[236] Fu Yujie, Zhang Pengju, Liu Bingxi, Rong Zheng, Wu Yihong. Learning to reduce scale differences for large-scale invariant image matching. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1335-1348

[237] Chen Ying, Huang Dihe, Xu Shang, Liu Jianlin, Liu Yong. Guide local feature matching by overlap estimation // Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 365-373

[238] Cao Si-Yuan, Shen Hui-Liang, Chen Shu-Jie, Li Chunguang. Boosting structure consistency for multispectral and multimodal image registration. IEEE Transactions on Image Processing, 2020, 29: 5147-5162

[239] Quan Dou, Wei Huiyuan, Wang Shuang, Lei Ruiqi, Duan Baorui, Li Yi, Hou Biao, Jiao Licheng. Self-distillation feature learning network for optical and SAR image registration. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18

[240] Xiong Xin, Xu Qing, Jin Guowang, Zhang Hongmin, Gao Xin. Rank-based local self-similarity descriptor for optical-to-SAR image matching. IEEE Geoscience and Remote Sensing Letters, 2020, 17(10): 1742-1746

[241] Hu Jing, Luo Ziwei, Wang Xin, Sun Shanhui, Yin Youbing, Cao Kunlin, Song Qi, Lyu Siwei, Wu Xi. End-to-end multimodal image registration via reinforcement learning. Medical Image Analysis, 2021, 68: 101878

[242] Pielawski Nicolas, Wetzer Elisabeth, Öfverstedt Johan, Lu Jiahao, Wählby Carolina, Lindblad Joakim, Sladoje Natasa. CoMIR: contrastive multimodal image representation for registration // Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 18433-18444

[243] Quan Dou, Fang Shuai, Liang Xuefeng, Wang Shuang, Jiao Licheng. Cross-spectral image patch matching by learning features of the spatially connected patches in a shared space // Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 115-130

[244] Liu Xiangzeng, Ai Yunfeng, Tian Bin, Cao Dongpu. Robust and fast registration of infrared and visible images for electro-optical pod. IEEE Transactions on Industrial Electronics, 2019, 66(2): 1335-1344

[245] Deng Yuxin and Ma Jiayi. ReDFeat: Recoupling detection and description for multimodal feature learning. IEEE Transactions on Image Processing, 2023, 32: 591-602

[246] Pautrat Remi, Larsson Viktor, Oswald Martin R., Pollefeys Marc. Online invariance selection for local feature descriptors// Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 707-724

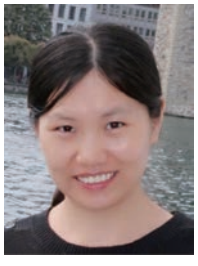
[247] He Kaiming, Chen Xinlei, Xie Saining, Li Yanghao, Dollar Piotr, Girshick Ross. Masked autoencoders are scalable vision learners// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16000-16009

[248] Chen Ting, Kornblith Simon, Norouzi Mohammad, Hinton Geoffrey. A simple framework for contrastive learning of visual representations // Proceedings of the International Conference on Machine Learning. Virtual, 2020: 1597-1607

[249] Mildenhall Ben, Srinivasan Pratul P., Tancik Matthew, Barron Jonathan T., Ramamoorthi Ravi, Ng Ren. NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2021, 65(1): 99-106

[250] Wei Yanyan, Zhang Zhao, Wang Yang, Xu Mingliang, Yang Yi, Yan Shuicheng, Wang Meng. DerainCycleGAN: Rain attentive cycleGAN for single image deraining and rainmaking. IEEE Transactions on Image Processing, 2021, 30: 4788-4801

[251] Gafni Oran, Polyak Adam, Ashual Oron, Sheynin Shelly, Parikh Devi, Taigman Yaniv. Make-a-scene: Scene-based text-to-image generation with human priors// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 89-106



**KONG Qing-Qun**, Ph. D., associate professor. Her research interest covers artificial intelligence and brain-inspired visual understanding.

**WU Fu-Chao**, professor. His research interests are in computer vision and pattern recognition.

**FAN Bin**, Ph. D., professor. His research interests are in computer vision, pattern recognition, and artificial intelligence.

**Background**

Image matching is a fundamental problem in computer vision. It aims to establish point correspondences across images of a same scene captured at different imaging conditions. Traditional methods for image matching rely on designing keypoint detectors and descriptors based on expert knowledge. Recently, many deep learning based methods have been proposed with promising results, covering various aspects of image matching. These include learning keypoint detectors and local patch descriptors with CNNs, as well as joint learning of detectors and descriptors based on deep learning. In addition, the excellent context modeling ability of graph neural networks and transformers has advanced using deep learning techniques in developing detector-free or semi-dense image matching methods as well as mismatch removal methods. Although there are survey papers in this area, few are focused on recent techniques, especially on how the deep learning advances this area. In addition, it lacks a systematically review of downstream tasks relying on image matching quality and those datasets/benchmarks that are widely used for evaluation of image matching methods under different prospects. Viewing of these drawbacks in existing surveys, this paper aims to provide an up-to-date summarization and analysis of key techniques used in recent methods for image matching. It also provides a thorough review of existing applications driven by image matching methods. Meanwhile, this paper summarizes the main datasets and benchmarks for evaluating image matching related methods and their downstream tasks, with detailed description about the properties of these datasets, as well as indexes to literature for reader easily finding state of the art performance on these benchmarks.