JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
06
VOL.26

ISSN1006-8961
CN11-3758/TB

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第6期（总第302期）
2021年6月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年6月 第26卷 第6期（总第302期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

生物特征识别学科发展报告(第1254页)

图像处理与通信技术

大规模室外图像3维重建技术研究进展(第1429页)

三维视觉和图形技术

虚实融合场景中的深度感知研究综述(第1503页)

Volume 26, Number 6
Published June 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Overview of biometrics research (P1254)

Progress in the large-scale outdoor image 3D reconstruction (P1429)

Review of depth perception in virtual and real fusion environment (P1503)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2021)06-1389-40

论文引用格式: Long X X, Cheng X J, Zhu H, Zhang P J, Liu H M, Li J, Zheng L T, Hu Q Y, Liu H, Cao X, Yang R G, Wu Y H, Zhang G F, Liu Y B, Xu K, Guo Y L and Chen B Q. 2021. Recent progress in 3D vision. Journal of Image and Graphics, 26(06): 1389-1428(龙霄潇,程新景,朱昊,张朋举,刘浩敏,李俊,郑林涛,胡庆拥,刘浩,曹汛,杨睿刚,吴毅红,章国锋,刘烨斌,徐凯,郭裕兰,陈宝权. 2021. 三维视觉前沿进展. 中国图象图形学报,26(06):1389-1428)[DOI:10.11834/jig.210043]

三维视觉前沿进展

龙霄潇[1,2],程新景[2],朱昊[3,2],张朋举[4,5],刘浩敏[6],李俊[7],郑林涛[7],胡庆拥[8],刘浩[9],曹汛[3],杨睿刚[2],吴毅红[5,4],章国锋[10],刘烨斌[11],徐凯[7],郭裕兰[7],陈宝权[12*]

1. 香港大学, 香港　999077; 2. 际络科技(上海)有限公司, 上海　200000; 3. 南京大学, 南京　210023; 4. 中国科学院大学人工智能学院, 北京　100190; 5. 中国科学院自动化研究所, 北京　100190; 6. 商汤研究院, 杭州　311215; 7. 国防科技大学, 长沙　410073; 8. 牛津大学, 牛津　OX13QR; 9. 中山大学, 广州　510275; 10. 浙江大学, 杭州　310058; 11. 清华大学, 北京　100085; 12. 北京大学, 北京　100871

摘　要:　在自动驾驶、机器人、数字城市以及虚拟/混合现实等应用的驱动下,三维视觉得到了广泛的关注。三维视觉研究主要围绕深度图像获取、视觉定位与制图、三维建模及三维理解等任务而展开。本文围绕上述三维视觉任务,对国内外研究进展进行了综合评述和对比分析。首先,针对深度图像获取任务,从非端到端立体匹配、端到端立体匹配及无监督立体匹配3个方面对立体匹配研究进展进行了回顾,从深度回归网络和深度补全网络两个方面对单目深度估计研究进展进行了回顾。其次,针对视觉定位与制图任务,从端到端视觉定位和非端到端视觉定位两个方面对大场景下的视觉定位研究进展进行了回顾,并从视觉同步定位与地图构建和融合其他传感器的同步定位与地图构建两个方面对同步定位与地图构建的研究进展进行了回顾。再次,针对三维建模任务,从深度三维表征学习、深度三维生成模型、结构化表征学习与生成模型以及基于深度学习的三维重建等4个方面对三维几何建模研究进展进行了回顾,并从多视RGB重建、单深度相机和多深度相机方法以及单视图RGB方法等3个方面对人体动态建模研究进展进行了回顾。最后,针对三维理解任务,从点云语义分割和点云实例分割两个方面对点云语义理解研究进展进行了回顾。在此基础上,给出了三维视觉研究的未来发展趋势,旨在为相关研究者提供参考。

关键词:立体匹配;单目深度估计;视觉定位;同步定位与地图构建(SLAM);三维几何建模;人体动态重建;点云语义理解

Recent progress in 3D vision

Long Xiaoxiao[1,2], Cheng Xinjing[2], Zhu Hao[3,2], Zhang Pengju[4,5], Liu Haomin[6], Li Jun[7], Zheng Lintao[7], Hu Qingyong[8], Liu Hao[9], Cao Xun[3], Yang Ruigang[2], Wu Yihong[5,4], Zhang Guofeng[10], Liu Yebin[11], Xu Kai[7], Guo Yulan[7], Chen Baoquan[12*]

1. The University of Hong Kong, Hong Kong　999077, China; 2. Jiluo Technology (Shanghai) Co., Ltd., Shanghai　200000, China; 3. Nanjing University, Nanjing　210023, China; 4. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing　100190, China; 5. Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China; 6. SenseTime Research Institute, Hangzhou　311215, China; 7. National University of Defense Technology, Changsha　410073, China;

收稿日期:2021-01-21;**修回日期**:2021-02-02;**预印本日期**:2021-02-09

所有作者为共同一作,*通信作者:陈宝权　baoquan@pku.edu.cn

8. University of Oxford, Oxford OX13QR, United Kingdom; 9. Sun Yat-sen University, Guangzhou 510275, China;
10. Zhejiang University, Hangzhou 310058, China; 11. Tsinghua University, Beijing 100085, China;
12. Peking University, Beijing 100871, China

Abstract: 3D vision has numerous applications in various areas, such as autonomous vehicles, robotics, digital city, virtual/mixed reality, human-machine interaction, entertainment, and sports. It covers a broad variety of research topics, ranging from 3D data acquisition, 3D modeling, shape analysis, rendering, to interaction. With the rapid development of 3D acquisition sensors (such as low-cost LiDARs, depth cameras, and 3D scanners), 3D data become even more accessible and available. Moreover, the advances in deep learning techniques further boost the development of 3D vision, with a large number of algorithms being proposed recently. We provide a comprehensive review on progress of 3D vision algorithms in recent few years, mostly in the last year. This survey covers seven different topics, including stereo matching, monocular depth estimation, visual localization in large-scale scenes, simultaneous localization and mapping (SLAM), 3D geometric modeling, dynamic human modeling, and point cloud understanding. Although several surveys are now available in the area of 3D vision, this survey is different from few aspects. First, this study covers a wide range of topics in 3D vision and can therefore benefit a broad research community. On the contrary, most existing works mainly focus on a specific topic, such as depth estimation or point cloud learning. Second, this study mainly focuses on the progress in very recent years. Therefore, it can provide the readers with up-to-date information. Third, this paper presents a direct comparison between the progresses in China and abroad. The recent progress in depth image acquisition, including stereo matching and monocular depth estimation, is initially reviewed. The stereo matching algorithms are divided into non-end-to-end stereo matching, end-to-end stereo matching, and unsupervised stereo matching algorithms. The monocular depth estimation algorithms are categorized into depth regression networks and depth completion networks. The depth regression networks are further divided into encoder-decoder networks and composite networks. Then, the recent progress in visual localization, including visual localization in large-scale scenes and SLAM is reviewed. The visual localization algorithms for large-scale scenes are divided into end-to-end and non-end-to-end algorithms, and these non-end-to-end algorithms are further categorized into deep learning-based feature description algorithms, 2D image retrieval-based visual localization algorithms, 2D-3D matching-based visual localization algorithms, and visual localization algorithms based on the fusion of 2D image retrieval and 2D-3D matching. SLAM algorithms are divided into visual SLAM algorithms and multisensor fusion based SLAM algorithms. The recent progress in 3D modeling and understanding, including 3D geometric modeling, dynamic human modeling, and point cloud understanding is further reviewed. 3D geometric modeling algorithms consist of several components, including deep 3D representation learning, deep 3D generative models, structured representation learning and generative models, and deep learning-based 3D modeling. Dynamic human modeling algorithms are divided into multiview RGB modeling algorithms, single-depth camera-based and multiple-depth camera-based algorithms, and single-view RGB modeling methods. Point cloud understanding algorithms are further categorized into semantic segmentation methods and instance segmentation methods for point clouds. The paper is organized as follows. In Section 1, we present the progress in 3D vision outside China. In Section 2, we introduce the progress of 3D vision in China. In Section 3, the 3D vision techniques developed in China and abroad are compared and analyzed. In Section 4, we point out several future research directions in the area.

Key words: stereo matching; monocular depth estimation; visual localization; simultaneous localization and mapping (SLAM); 3D geometry modeling; dynamic human reconstruction; point cloud understanding

0 引 言

在自动驾驶、机器人、数字城市以及虚拟/混合现实等应用的驱动下,三维视觉在近年来得到了广泛的关注。三维视觉研究主要围绕深度图像获取、视觉定位与制图、三维建模及三维理解等问题而展开。本文针对上述问题,对立体匹配、单目深度估计、大场景下的视觉定位、同步定位与地图构建、三维几何建模、人体动态重建以及点云语义理解等研究方向的发展现状、前沿动态、热点问题和发展趋势进行系统综述。本文内容框架如图 1 所示。

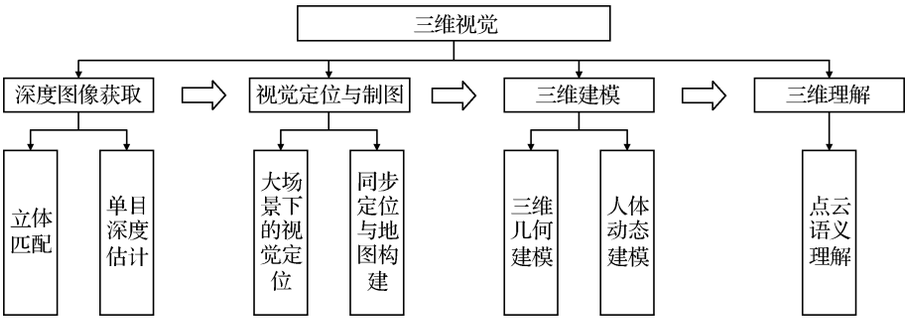


图 1　本文内容框架图

Fig. 1　The architecture of this work

单目深度估计是通过单幅亮度图像和相关先验恢复深度信息的过程，这种深度信息通常表示为深度图，即以二维数组表示每个图像像素所对应的深度值。深度图是三维模型的主要表征形式之一，能够应用于自动驾驶（Zhang 等，2019a；Liao 等，2020）、虚拟视频生成（Wang 等，2020h；Zhu 等，2018c）和增强现实等多个领域，具有重大的研究意义和应用价值。

双目立体视觉技术模仿人眼视觉系统对现实世界进行三维感知，通过两幅不同视角下的图像进行立体匹配获取视差/深度信息。相比于主动式感知技术（激光扫描、结构光扫描等），双目立体视觉技术具有设备简单、成本低和效率高的优势，因此双目立体匹配技术在数十年里一直是计算机视觉领域中的热点问题，并且获得了一系列的进展。目前双目立体视觉技术在实际中有着广泛的应用，包括智能机器人导航（Schmid 等，2013）、目标识别（Helmer 和 Lowe，2010）、遥感技术（Shean 等，2016）和自动驾驶（Menze 和 Geiger，2015）。

相机定位是三维计算机视觉的一个基本问题，它的任务是根据相机拍摄的图像估计相机在某个坐标系下的位置和朝向，即相机的姿态。相机定位在诸多领域，如：机器人、自动驾驶、增强现实和虚拟现实等，都有重要的应用价值，也一直是三维计算机视觉中的一个核心问题。视觉定位分为 3D 模型已知的定位和 3D 模型未知的定位（Wu 等，2018b）。本文大场景下的视觉定位部分阐述端到端的视觉定位和 3D 模型已知的视觉定位。同步定位与地图构建部分阐述模型未知的视觉定位。

同步定位与地图构建（simultaneous localization and mapping，SLAM）技术可以实时恢复移动设备的位姿，同时重建环境的三维信息，是增强现实、自动驾驶和移动机器人等应用领域中的关键技术（Cadena 等，2016）。早期的移动定位主要依靠专用硬件设备来实现，随着带有摄像头的移动设备的普及以及其计算能力的提高，基于低成本视觉传感器和 IMU（inertial measurement unit）的视觉惯性 SLAM 取得了重大突破，并已在一些产品上落地（Qin 等，2018；Campos 等，2020）。随着传感器、网络和云计算的迅速发展，SLAM 应用的场景规模也在不断扩大（Lynen 等，2015），通过 SLAM 技术实现基于低成本传感器的城市级甚至地球级的移动定位将有望成为现实。

三维数字内容是虚拟仿真、混合现实等的基本构成要素。创建三维内容的核心是三维几何建模，它是计算机图形学的重要基本问题。人工三维建模操作困难、繁琐，严重依赖于专业人员的技能和经验，未经训练的普通用户往往难以胜任。如何让普通大众方便快捷地创作和编辑三维内容，实现所谓“大众建模”，突破“三维内容生成瓶颈”，从而推动三维数据的规模化增长，一直是图形学领域的一个核心目标和关键挑战。

点云语义分割是指根据点云内部的空间几何结构和形状信息将点云分成具有不同语义标签的点集，而实例分割要求进一步对点云场景中的不同实例进行区分。相比而言，点云实例分割既要将具有不同语义标签的点进行区分，还要区分具有相同语义标签的不同实例，因而相比点云语义分割更具挑战性。点云的语义分割和实例分割是实现三维场景理解的重要基础，在视觉导航与定位、自动驾驶和增强现实（augmented reality，AR）/虚拟现实（virtual reality，VR）等许多领域有广泛应用前景。

1 国际研究现状

1.1 立体匹配

1.1.1 非端到端立体匹配

对于非端到端的立体匹配算法,卷积神经网络(convolutional neural network, CNN)通常用来代替传统匹配算法中的一部分或者多个部分。Žbontar和LeCun(2015)首先使用卷积神经网络MC-CNN来计算匹配代价。这一深度孪生网络由数个卷积层与全连接层组成,用来计算两个图像分块之间的相似度。这一方法在当年的KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)双目数据集上达到了最好的效果,证明了通过卷积神经网络提取到的图像特征比手工设计的特征算子更加准确。受此启发,大量的立体视觉工作(Gidaris和Komodakis, 2017; Barron和Poole, 2016;Brandao等,2019;Kim和Kim,2016;Taniai等,2018)利用卷积神经网络来计算匹配代价。Zagoruyko和Komodakis(2015)探索并提出了大量不同结构的卷积神经网络来表示两个图像分块之间的相似度测量函数并提升了最终效果。尽管这些方法(Žbontar和LeCun,2015;Zagoruyko和Komodakis, 2015)在一些双目数据集上,比如KITTI,取得了巨大的进步,但是这些方法往往需要耗费非常多的计算资源且十分耗时。Luo等人(2016)将立体匹配问题当做一个多类别的分类问题进行处理,使得模型可以通过学习在所有的候选视差值上的概率分布来隐性学习到不同视差下图像分块之间的差异。在这些方法中,通常会使用一些非学习的后处理技术来进一步优化通过神经网络得到的初始匹配代价,比如交叉特征聚合,半全局匹配和左右一致性检测、滤波等。

在传统立体视觉流程中除了匹配代价生成以外的部分也可以由神经网络取代。基于视差图局部平滑的假设,一些方法直接将平滑约束用在网络学习过程中。Seki和Pollefeys(2017)提出了SGM-Net(semi-global matching with neural network)框架来预测传统SGM算法中的惩罚代价。其提出了一种新型的代价函数,包括路径惩罚项和相邻惩罚项,该新型代价函数使得神经网络可以充分利用到实际中采集到的稀疏视差图标签,比如通过激光雷达采集到的稀疏视差图。然而,为了获取SGM惩罚代价的标签值需要复杂的处理过程,因此使得训练SGM-Net非常复杂并且耗时。

Knöbelreiter等人(2017)提出了一个神经网络和条件随机场结合的混合模型,并加入了平滑惩罚项。Gidaris和Komodakis(2017)使用了一个三阶段的网络来取代手工设计的视差优化函数。这一网络可以检测不正确的视差估计值,使用新的估计值替代错误的,然后再次优化新的视差值。然而这一检测、替代和优化的步骤需要额外的计算资源。因为大多数的立体匹配技术在反射和无纹理表面很难通过局部约束正确恢复出视差值。Güney和Geiger(2015)提出了Displets模型,利用物体的先验信息来解决反射和无纹理表面上视差估计的不确定性问题。该方法通过引入汽车的三维模型作为先验信息,其取得了当年KITTI双目视差估计数据集上的第1名。然而,引入三维模型极大增加了模型的计算负担。

这些非端到端的估计方法,相比传统方法有较大的效果提升,但其或多或少都依赖于手工设计的约束方程和后处理步骤来实现较好的结果。这些方法的效果受制于较高的计算负担、有限的感受野和缺乏图像全局信息等问题,逐渐被端到端的立体视觉方法所替代。

1.1.2 端到端立体匹配

端到端的视差估计网络可以无缝地结合立体视觉流程中的所有步骤,可以直接从双目图像中估计出完整且稠密的视差图。自从Mayer等人(2016)首先提出了端到端的视差估计网络,大量的后续工作都采用了这一结构并取得了优异的结果。一般而言,视差估计网络的结构可以分为两类:1)二维卷积层组成的编码器—解码器的层级优化结构;2)三维卷积层组成的正则化网络结构。一般而言,二维卷积网络的运行速度更快,而三维卷积网络的预测精度更高。

Mayer等人(2016)首先提出了一个光流和视差联合估计的网络。对于视差估计,该网络使用一维的相关性估计层沿着视差扫描线方向进行匹配代价计算,然后使用编码器—解码器的网络进行视差回归。这一端到端的网络结构使得完整视差图的估计过程变得简单,并且可以节省大量计算资源。

Dispnet与其拓展工作使用二维卷积网络对获

取的匹配代价进行正则化与优化，再对视差值进行回归。Kendall 等人（2017）提出了 GC-Net（global context network），首先使用三维卷积网络对匹配代价进行聚合，并取得了当时最先进的结果。GC-Net 的成功证明了三维卷积网络在匹配代价的聚合与正则化上，相对二维卷积网络有更好的泛化能力和更高的精度。由于额外的视差维度的引入，使用三维卷积网络的代价聚合过程更加规则化，有利于视差维度上不同体素之间的聚合。

然而，基于学习的视差估计网络往往在无纹理或弱纹理的区域表现较差。一些方法基于 Kusupati 等人（2020）提出的局部平面的先验假设，引入了额外的法线向量约束（normal constraint），使得视差/深度估计网络在无纹理区域也能有合理的估计值。Poggi 和 Mattoccia（2017）提出了一个置信度估计网络来衡量估计的视差值的可信度。Kim 等人（2018）也使用卷积神经网络来联合估计视差值以及其对应的置信度。Badki 等人（2020）将立体匹配问题重构为一个二值分类（binary classification）问题而不是直接回归视差值。

目前，端到端的立体匹配算法已经取代了非端到端的立体匹配算法成为主流。尽管端到端的立体匹配算法能够同时结合局部信息与图像的全局信息进行视差估计，但是这些算法依旧很难在无纹理区域、物体边缘和细小结构上取得令人满意的效果。另一方面，采用三维卷积网络的立体匹配算法需要占用巨量的 GPU（graphics processing unit）资源，限制了其在实际中的使用价值。近年来，一些算法试图改善这些情况，比如 Guo 等人（2019）使用群卷积来减少代价体的特征通道数，Lu 等人（2018）使用稀疏代价体替代稠密代价体，Gu 等人（2020）使用级联多代价体代替完整单一代价体。但是这些算法依旧依赖于三维卷积网络，因此高 GPU 显存消耗的问题并没有得到明显改善。同时，为了使得网络具有较好的性能与泛化性，这类端到端的立体匹配网络需要大量的视差/深度标签进行训练，但获得大量标签是非常昂贵的。

1.1.3 无监督立体匹配

训练立体视觉神经网络往往需要大量的数据标签，然而在实际中获取足够的视差或深度标签是非常昂贵的，尤其在室外场景下往往需要依赖高精度激光雷达对场景的三维结构进行稀疏采样。但是在实际中，大量没有数据标签的双目视频数据是非常廉价的，也能提供一定的信息。近年来，基于空间变换与视图生成技术，一些无监督立体视觉网络被提出，并取到了喜人的成果。

在没有数据标签的情况下，如何快速便捷地获取可靠的监督信号是无监督立体匹配算法最关键的问题。根据不同视角下的图片具有亮度一致性假设，一些工作通过生成新视图，使用图像重建损失函数来替代视差标签作为监督信号。该技术首先应用在图像生成领域，而后被用于无监督的视差深度估计问题。

Flynn 等人（2016）提出了一个新视图生成的网络 DeepStereo，这个网络通过选择相邻图片上的像素来生成新视角下的图片。Xie 等人（2016）对于新视图生成的问题提出了 Deep3D 网络，在他们的工作中，将一组双目图片中的左视图作为输入，生成右视图，并最小化生成的右视图与真实右图在像素级别上的图像重建损失函数。

基于图像损失函数的无监督深度估计技术首先是由 Garg 等人（2016）提出，其利用一对双目图像作为单目深度估计的监督信号。首先使用单目的深度估计网络预测一对双目图像中的左视图的视差图，再根据估计的视差图，将右视图利用反向图片扭曲技术生成左视图。这样即可使用生成的左视图与真实左视图做图像重建误差计算，作为网络训练的监督信号。进一步地，Godard 等人（2017）和 Luo 等人（2018a）使用左右一致性约束作为图像损失函数的额外补充。一致性约束的引入，极大提高了无监督深度估计技术的效果，Godard 等人（2017）的工作在 KITTI 测试集上甚至超过了一些有监督的深度估计网络，这也标志着基于图像重建损失函数的无监督深度估计技术的成熟。

尽管这些无监督的方法避免了昂贵的视差/深度标签的使用，但是这些方法本质上仍旧是使用神经网络直接对单目图像进行视差/深度回归，仅仅使用双目图像作为监督信号，并不像有监督的立体匹配技术利用极线约束去寻找双目图像的最佳匹配。Zhong 等人（2018）提出了一个循环神经网络用来估计连续的双目视频的视差值，利用双目视频中连续帧之间的时间相关性，来进一步提高模型的性能和鲁棒性。Aleotti 等人（2020）提出了一个新颖的无监督匹配网络，其首先利用传统的立体匹配技术对

一组双目图像进行视差估计,筛选出可靠的稀疏视差估计值,并使用一个单目视差补全网络来稠密化稀疏的视差估计值,使用补全的视差值作为后续匹配网络的训练数据标签。

一般而言,主流的无监督立体匹配技术的基本流程为:1)将一对双目图像作为输入,网络输出左图和右图的视差图;2)通过估计的视差图与原始的图像对,分别生成新的合成图像;3)使用图像重建损失函数作用于生成的图像与原始图像,作为补充,使用左右一致性损失函数(左、右之间的重建误差)作为网络的训练监督信号,使得在缺乏数据标签的情况下,立体匹配网络依旧可以获得不错的效果。但是,在无纹理或者弱纹理的区域,图像重建损失函数无法提供可靠的监督信号,同时无监督技术无法获取视差/深度图像的真实尺度,在一定程度上限制了无监督方法在实际中的使用。

1.2 单目深度估计

网络结构设计对深度图预测的精度起到至关重要的作用。早期的工作尝试使用有监督学习下的马尔可夫随机场做为预测模型(Saxena 等,2005,2009),该模型在精度和泛化性方面的性能有限。近年来,深度神经网络成为单目深度估计的主流框架,已有的工作可以根据网络结构的不同分为深度回归网络和深度补全网络。

1.2.1 深度回归网络

在深度回归网络中,单目深度估计问题被视做逐像素的深度值回归问题,其输入为单目彩色或灰度图像,输出为深度图。该方法的目的是通过神经网络训练等方式,获取逐像素深度预测器,即深度回归模型。根据网络结构的不同,深度回归网络方面的研究可分为编码解码网络和复合网络两大类。

1.2.1.1 编码解码网络

直接编码网络通常由负责特征提取的编码器和用于回归逐像素深度的解码器构成,其中编码器主要由多个卷积层以及池化层、激活函数等构成,解码器主要由逆卷积层或全连接层等构成。

Eigen 等人(2014)提出了编解码网络和优化网络级联的网络结构。在此工作的基础之上,Li 等人(2015)提出了通过不同尺度上的超像素回归深度。Liu 等人(2016)同样使用过分割方法将原图处理为多个超像素,而后将超像素周边的图像块作为输入,采用由卷积层和全连接层构成的神经网络实现逐个分块的深度预测,最终合成完整深度图。

在编码解码网络的整体框架确定后,后续工作尝试了不同的编码器和解码器结构,以提升单目深度预测的性能。在编码器方面,Garg 等人(2016)使用由卷积层和少量全连接层构成的深度卷神经网络实现单目深度预测。Eigen 和 Fergus(2015)发现在编码网络部分使用 VGG(Visugl Geometry Group)网络替换 AlexNet 可以有效提升单目深度预测网络的鉴别能力。此后,Laina 等人(2016)借鉴残差网络的思想,将 ResNet50 网络中的残差卷积模块应用于单目深度预测的特征提取网络。Fu 等人(2018)提出去除池化层,并将卷积层替换为空洞卷积(dilated convolutions),这一策略在 Liebel 和 Körner(2019)的工作中得到延用。空洞卷积的引入能够有效扩大卷积计算中的感受域,增加网络对全局特征的提取能力,同时避免了空间分辨率的下降以及模型参数尺寸的提升。

在解码器方面,早期的工作(Eigen 等,2014;Eigen 和 Fergus,2015;Liu 等,2016;Li 等,2015)直接使用多级全连接层作为解码器。全连接层的优点在于每个像素深度的预测可以利用来自图像全局的信息,但全连接层包含大量网络参数,同时降低了网络的训练/预测速度。Dosovitskiy 等人(2015)提出去除全连接层,使由卷积层构成的编码网络和由逆卷积层构成的解码网络直接相连。该结构中的编码网络和解码网络通常呈对称的“沙漏状”,并包含跳跃链接将编码网络中的隐含特征直接传递到解码网络中。这种结构被多个工作沿用(Tan 等,2020;Cheng 等,2018,2020a)。Liebel 和 Körner(2019)提出了任务指定型解码器网络,该网络使用了深度回归和多任务分类的金字塔池化层(Zhao 等,2017)。

1.2.1.2 复合网络

复合网络尝试组合、堆叠多个神经网络或算法模块,以提升单目深度预测的性能。

Roy 和 Todorovic(2016)将深度神经网络和随机森林结合,提出深度回归森林。与已有的编码解码网络结构相比,深度回归森林允许同时训练多个浅层卷积神经网络,而在某一个路径的处理过程等效于一个深层卷积神经网络。Chakrabarti 等人(2016)提出首先采用卷积神经网络提取图像中的局部特征,而后用全局优化算法协调局部特征,生成深度图。

Zhou 等人(2017b)借鉴“运动恢复结构”方法的思想,提出一种单视图深度估计和相机运动估计同时训练的网络框架。Zhou 等人(2020b)提出单目深度和光流联合训练的网络框架。Fu 等人(2018)提出了一种 3 层级联的网络结构,第 1 层网络是由卷积层构成的密集特征提取器,第 2 层由全图编码器、多尺度特征提取器和跨通道信息学习器并行组成,三者得到隐含表达张量在经过合并后进入第 3 层网络—场景理解模块,最终生成目标深度图。Patil 等人(2020)提出使用卷积长短期记忆网络实现对单目视频的深度预测,该方法将有监督的深度预测网络、基于视频的自监督深度预测网络和自监督深度补全网络融合在同一框架下,将上述主体网络结构和卷积长短期记忆网络结合,利用帧与帧之间的时空一致性提升深度预测精度。

1.2.2　深度补全网络

与深度回归网络不同,深度补全网络考虑在有稀疏的深度作为输入时,如何估计出密集的深度图。近年来,“彩色图像 + 激光雷达”的数据获取系统是自动驾驶等应用场景中的常见配置,其中激光雷达获取的深度图在深度维度精度较高,但在空间维度非常稀疏,因此这一研究具有较高实用价值。由于图像/深度填充(Doria 和 Radke, 2012; Ferstl 等, 2013)和超分辨率(Mac Aodha 等,2012;Kiechle 等, 2013;Matsuo 和 Aoki,2015)所采用的输入深度与实际应用情况相比更为密集,本文将不再回顾深度填充和超分辨率的工作,而是着重回顾以稀疏深度作为输入的深度补全网络。

传统的编码解码型卷积神经网络对稀疏深度的预测效果较差。为解决这一问题,Uhrig 等人(2017)提出了针对稀疏输入的稀疏不变卷积神经网络。该网络引入一种新型的稀疏卷积层,该层根据输入像素的有效性对卷积核的元素进行加权。此外,该像素的有效性将通过信息流传递到后续的卷积层中。Huang 等人(2019a)将系数不变卷积层与编解码网络相结合,并使用跳跃链接连通编解码网络,使得稀疏不变运算结果能够在不同尺度之间进行融合。Jaritz 等人(2018)同样采用带跳跃链接的编码解码网络来处理稀疏数据,不同的是在 NASNet(Zoph 等,2018)做为编码网络的基础之上去除了第 1 个卷积层后的批归一化层,这是由于稀疏数据用零表示不存在的像素值,这会使批处理层中的平均计算出错。Chen 等人(2018)在神经网络的训练中首先生成了二值掩膜用于表明真值深度图中的有效像素,而后生成了每个有效像素的最近邻填充以及二值掩膜的欧氏距离变换。这些信息与图像共同作为编码解码网络的输入,而神经网络的输出是真值密集深度图和输入稀疏深度图的残差部分。Chodosh 等人(2019)将从稀疏深度补全问题与压缩感知问题结合,提出使用一个端到端的多层字典学习算法作为网络模型,这种框架能够从内部实现优化,从输入中提取卷积后的稀疏特征,而后进行单目深度预测。

Ma 和 Karaman(2018)和 Liao 等人(2017)直接使用单个深度神经网络从 RGB 图像和稀疏的深度采样点中估计密集深度图。该网络具有与编码解码网络相同的结构,以深度残差网络作为编码网络,以逆卷积层构成解码网络,输入为稀疏深度和图像在通道维度上的叠加。在后续的工作中(Ma 等, 2019),该网络在第 1 层中转化为两个独立的卷积自网络,分别从稀疏深度和图像中提取特征,同时在编码网络和解码网络之间加入了跳跃链接。Shivakumar 等人(2019)在多分支输入编解码网络的基础之上进行了进一步改进,将图像、稀疏深度的分支扩展为两个完整的卷积神经网络,并在分支网络的末端增加空间金字塔池化层,而后合并输入解码网络。在解码网络部分,不同尺度的张量在经过逆卷积层后被分别提取、变形和堆叠,最终使用卷积层从堆叠的多尺度张量中回归达到深度图。Xu 等人(2019)提出的网络由预测网络、距离变换模块和优化网络三维部分构成。其中,预测网络采用多分支编码解码网络,以稀疏深度图和彩色图像作为输入,预测得到法向量、导向特征图、粗略深度图和置信度图。而后,输入的稀疏深度图和粗略深度图被变换到平面原点的距离空间。而后,该方法使用各向异性的扩散模块优化空间转化后的粗略深度图,并加强高置信位置上法向和深度之间的约束关系。最后,优化后的深度图通过逆变换过程从平面原点空间还原为初始空间。Hambarde 和 Murala(2020)提出使用一个端到端的神经网络从单目图像和稀疏深度图估计密集的深度图。

Lu 等人(2020)提出一种通过稀疏深度获取完整深度的方法,将与深度对应的彩色图像作为参考监督项。Park 等人(2020)针对深度补全问题提出

了一种非局部的空间传播网络。Cheng 等人(2020a)在卷积空间传播网络(Cheng 等,2018)的基础上提出一种以稀疏深度图和图像作为输入的深度预测网络。该网络能够根据内容调整每个像素对应的卷积核尺寸和迭代次数,衍生出更高性能的内容感知网络和更快速度的资源感知网络。

1.3 大场景下的视觉定位

大场景下的视觉定位大致可以分为端到端的视觉定位和非端到端的视觉定位。非端到端的视觉定位方法也就是传统的视觉定位方法,包括检测和描述关键点、建立 2D-3D 匹配和利用 RANSAC(random sample consensus) + PNP(perspective *N* points)估计位姿等步骤。端到端的视觉定位方法则利用神经网络来实现视觉定位框架中的所有模块。

1.3.1 端到端的视觉定位方法

端到端的视觉定位方法可以分为基于 3D 坐标回归的方法和基于位姿回归的定位方法。基于 GPU 强大的计算能力,端到端的方法十分高效。最具有代表性的基于 3D 坐标回归的方法包括 DSAC(differentiable RANSAC)(Brachmann 等,2017)、DSAC++(Brachmann 和 Rother,2018)等。Brachmann 等人(2017)首先利用神经网络预测图像中的 2D 点所对应的 3D 坐标,这样就得到了 2D-3D 匹配,并且,受强化学习理论的启发,他们提出了可导形式的 RANSAC,称之为 DSAC,这样,传统的视觉定位框架中所有的模块都可以用神经网络端到端地实现。在 DSAC++里,Brachmann 和 Rother(2018)提出了一种基于熵控制的内点计数方法来对假设模型进行打分,大大提高了 DSAC 的泛化性能。之后,受启发于集成学习的理论,Brachmann 和 Rother(2019)将 DSAC 集成到多专家模型中,在合成和真实的定位数据集都取得了不错的结果。基于位姿回归的方法则摒弃传统的视觉定位框架,在给定一组训练图像及其对应的位姿后,其通过训练卷积神经网络直接从输入图像回归相机的位姿。PoseNet(Kendall 等,2015)基于 GoogleNet 搭建了第 1 个端到端的 6DoF(six degrees of freedom)相机定位方法,之后有很多的其他方法是基于 PoseNet 的改进。另外比较有代表性的工作是 MapNet(Brahmbhatt 等,2018),其利用两幅图像间的相对位姿和每幅训练图像的绝对位姿来进行端到端的定位。Wang 等人(2020a)将注意力机制应用到全局位姿回归网络中,在室内外数据集上都取得了更加优越的性能。Sattler 等人(2019)深入研究了基于位姿回归的方法,发现基于位姿回归的图像定位更类似于图像检索,而不是基于 2D-3D 匹配的精准定位,这表明在大场景下,基于位姿回归的定位方法想要达到与基于 2D-3D 匹配的定位方法相同精度的话,还需要进一步的研究。

1.3.2 非端到端的视觉定位方法

基于深度学习的特征描述通常是用在非端到端的视觉定位中。因此本节第 1 部分描写基于深度学习的特征描述进展。然后对基于 2D 图像检索的方法、基于 2D-3D 匹配的方法以及二值融合的方法分别进行介绍。

1)基于深度学习的特征描述。近些年,基于深度学习端到端的描述子成为趋势。在视觉定位任务中,用深度学习描述子取代传统的描述子,往往能取得较好的鲁棒性。MatchNet 利用孪生网络提取特征和计算特征相似度(Han 等,2015)。DeepDesc 采用区分图像块的渐进采样策略来提升描述子性能(Simo-Serra 等,2015)。DeepCompare 研究多种网络结构来提高描述子性能(Zagoruyko 和 Komodakis,2015)。HardNet 的 loss 采用三元损失函数,并在 batch 内进行困难样本挖掘,最终取得了不错的效果(Mishchuk 等,2017)。DOAP(descriptors optimized for average precision)则是直接将描述子的性能度量标准 AP(average precision)作为损失函数(He 等,2018)。

值得注意的是,最近也出现了一些同时提取关键点和描述子的网络,如:SuperPoint, D2-Net, R2D2 等。为解决没有相关的数据集,Superpoint 首先在一些人造数据集上进行训练,然后利用 Homographic Adaptation 技术过渡到真实场景中训练网络(DeTone 等,2018)。D2-Net 则是模拟传统的提取关键点的方法,在损失函数中加入相关约束,在提取特征的同时筛选图像的关键点(Dusmanu 等,2019)。R2D2 网络同时输出 3 个 heads,一个表示关键点的重复性指标,一个表示关键点的可靠性指标,另外一个表示关键点的特征,然后用这 3 个 heads 完成关键点的检测和描述(Revaud 等,2019)。在训练描述子网络时,上述描述子大多优化的是描述子之间的匹配分数,而 Bhowmik 等人(2020)通过优化更高层任务——两幅图像间的相对位姿来训练描述子,获得了性能更好的描述子。在训练描述子

时，带有真实标签的数据集往往较难获得，为了解决这个问题，Wang 等人（2020g）仅仅使用两幅图像间的极线约束来训练描述子，也取得了不错的效果。Sarlin 等人（2020）利用图神经网络来学习特征匹配，与 Superpoint 相结合，在视觉定位、图像匹配等任务上取得了第 1 名的成绩。

2）基于 2D 图像检索的视觉定位方法。基于 2D 图像检索的方法首先在数据库图像里检索与查询图像最相似的图像，然后将检索到的数据库图像的位姿直接作为查询图像的位姿。传统的图像检索方法有 VLAD（vector of locally aggregated descriptors），Bag-of-words vector 和 FV（fisher vectors）。近年来，国外学者又提出了一些新的图像检索方法。DenseVLAD 利用深度图产生虚拟视角并提取稠密描述子，增加了数据的多样性，在查询的时候更容易找到查询图像的近似最近邻（Torii 等，2015）。NetVLAD 则使利用深度学习技术，提取局部特征，然后将深度学习学到的局部特征输入可学习的 VLAD 层来将局部特征聚合成向量，然后用合成向量来进行最近邻匹配（Arandjelovic 等，2016）。基于深度学习的图像检索方案可以提取图像的高层语义特征，所以在光照变化比较剧烈的情况下优势尤其明显。Liu 等人（2019a）在 NetVLAD 的基础上，通过优化 KL（Kullback Leibler）散度来学习判别性更强的图像特征。Radenović 等人（2019）提出可训练的广义平均池化层（generalized-mean pooling layer）来提高检索性能，这种广义平均池化层同时具有平均池化和最大池化层的优势。SOLAR（second-order loss and attention for image retrieval）（Ng 等，2020）将二阶注意力机制和二阶相似性用于图像检索任务中，并带来显著的检索性能改进。

3）基于 2D-3D 匹配的视觉定位方法。2D-3D 匹配的视觉定位方法需要借助场景的三维重建信息。首先建立查询图像中的二维点和数据库中的三维点之间的匹配关系，然后将这些 2D-3D 匹配输入到 PNP RANSAC 算法中来计算查询图像的 6DoF 位姿。早期的 2D-3D 匹配的视觉定位主要是研究如何有效地求解较少点个数下的非线性方程（Wu 和 Hu，2006）。2010 年之后，主要是研究大场景大数据量下的 2D-3D 的匹配。Wang 等人（2015）提出了基于场景平面图的室内大场景定位方法。Zeisl 等人（2015）提出了 O（n）复杂度的基于投票的匹配方法，可以得到更多的匹配。Lu 等人（2015）使用短视频构造了 3D 模型作为查询信息，并采用多任务点检索方式进行 3D 到 3D 查询定位。为了让 2D-3D 直接匹配的方法更加实用，Liu 等人（2017b）利用共视信息来确定哪一部分的 3D 点更有可能在查询图像上观察到，然后只在这些有较大概率能够在查询图像上看到的 3D 点中进行 2D-3D 匹配。Svärm 等人（2017）放松建立 2D-3D 匹配的标准，这样可以建立更多潜在的正确匹配，但同时也引入了大量的外点，然后利用具有更强判别性的滤波器处理大量的外点。这种方法在重力方向已知的情况下非常有效。Sattler 等人（2017）采用优先搜索的策略去查找一定数量的 2D-3D 匹配，然后就停止查找以此来提高定位效率。

4）基于 2D 图像检索和 2D-3D 匹配融合的视觉定位方法。不论是基于 2D 图像检索的方法还是 2D-3D 匹配的定位方法，都有各自的优点和不足。因此将二者融合，可以兼具二者的优势而克服各自的不足。HFNet 先用 NetVLAD 进行粗匹配，然后在图像检索的结果中用直接法进行更精细的 2D-3D 局部匹配，以此得到更准确的定位结果（Sarlin 等，2019）。Sarlin 等人（2018）还利用分层定位的方法来提高在机器人平台上定位的效率，同时还能保持较高的定位精度。

之后，很多研究学者基于 HFNet 改进其中的局部描述子来提高定位精度，如：将 HFNet 中使用的 SuperPoint（DeTone 等，2018）替换为 D2-Net（Dusmanu 等，2019），R2D2（Revaud 等，2019），ASLFeat（Luo 等，2020）等。

1.4 同步定位与地图构建

1.4.1 视觉 SLAM

MonoSLAM（Davison，2003）是最早提出的基于扩展卡尔曼滤波的视觉 SLAM 系统，通过跟踪稀疏特征估计位姿，但复杂度与三维点的数目成立方关系，只能局限在比较小的场景里运行。PTAM（parallel tracking and mapping）（Klein 和 Murray，2007）作为视觉 SLAM 的一个里程碑性的工作，首次提出了将跟踪和建图分为两个线程并行运行，保证了跟踪的实时性，并用集束调整优化替代了滤波框架，通过基于关键帧的地图管理方法提高了整体的精度。后来，Strasdat 等人（2010）提出使用基于滑动窗口集束调整来实现大规模场景的跟踪；Strasdat 等人

(2011)采用了双窗口优化以及可视图的设计框架将视觉跟踪的过程控制在常数时间。ORB-SLAM(oriented fast and rotated brief-simultaneous localization and mapping)(Mur-Artal 等, 2015)和 ORB-SLAM2(Mur-Artal 和 Tardós, 2017)进一步完善了 PTAM 的框架,采用前后端多线程以及基于关键帧的优化,并借助视觉词袋的方法(Gálvez-López 和 Tardos,2012)实现了高效的重定位以及回环检测。得益于本质图的设计,ORB-SLAM 能够高效地完成回环的优化。在最新的 ORB-SLAM3(Campos 等,2020)中,采用了自动生成新地图的方式来解决视觉丢失的情况,进一步提升了系统的鲁棒性。与上述基于特征点的方法不同,直接法直接使用图像的像素灰度,通过最小化光度误差来估计运动和结构,在弱纹理场景中通常更为鲁棒。代表性的工作有 LSD-SLAM(large-scale direct SLAM)(Engel 等,2014)、SVO(semi-direct risual odometry)(Forster 等,2014)、DSO(direct sparse odometry)(Engel 等,2017)。近几年有不少工作在 DSO 的基础上进行了改进。例如 Engel 等人(2015)实现了双目的 LSD-SLAM,Gao 等人(2018)、Lee 和 Civera(2019)为 DSO 系统添加了回环检测能力等;DSM(direct sparse mapping)(Zubizarreta 等,2020)将地图重用的思想引入到直接法中,提高了直接法的建图能力。

随着深度学习方法的迅速发展,有研究者将其与传统 SLAM 框架相结合,取得了较好的结果。DeepVO(Wang 等,2017b)采用 CNN 和 RNN(recurrent neural network)结合的方式实现了端到端的视觉里程计,通过 CNN 提取相邻帧间的特征,并结合 RNN 获取与之前状态的时空关系,从而估计出较为精确的位姿。在此基础上,Saputra 等人(2019)加入了规划学习的方式,提高了其泛化能力。SfmLearner(Zhou 等,2017a)利用连续帧间的一致性作为约束,提出了无监督的视觉里程计,大大减少了对于标注数据的依赖。基于这个框架,Li 等人(2018b)、Yin 和 Shi(2018)、Zhan 等人(2018)、Yang 等人(2018b)、Zhao 等人(2018)、Almalioglu 等人(2019)、Li 等人(2019d)、Li 等人(2019e)、Sheng 等人(2019)等后续工作针对绝对尺度问题和动态物体干扰等问题分别进行了改进。端到端的方式由于缺少传统的几何约束,以及存在过拟合等问题,通常无法达到传统方式的精度,而通过引入预训练的学习模块到传统 SLAM 框架中,往往可以达到更好的效果。Eigen 等人(2014)、Ummenhofer 等人(2017)、Garg 等人(2016)、Godard 等人(2017)通过引入深度估计模块到视觉里程计中,解决了绝对尺度的问题。CNN-SLAM(Tateno 等,2017)将基于深度学习的深度估计进一步融合到 LSD-SLAM 中,从而实现了一个具有绝对尺度信息的单目稠密 SLAM 系统。CodeSLAM(Bloesch 等,2018)将场景深度紧凑表达为定长的向量,并首次将 CNN 的反向传播机制与基于关键帧优化的传统 SLAM 框架结合,通过迭代对场景深度向量优化,解决了全局尺度一致性问题。其后续工作(Zhi 等,2019)在此基础上将语义信息也融入其中,提高了语义的一致性。除此之外,Zhan 等人(2020)引入了深度光流估计,提高了跟踪的鲁棒性。D3VO(Yang 等,2020)同时集成了深度、位姿估计和不确定性估计到直接法的视觉里程计中,显著提升了精度。

常见的 SLAM 系统只能提供一个局部坐标系,而对于大场景的定位导航应用来说,需要获得一个全局一致的位姿,通常采用基于高精度地图的定位技术来实现这一功能。Stewart 和 Newman(2012)、Wolcott 和 Eustice(2014)、Maddern 等人(2014)、Pascoe 等人(2015)、Wong 等人(2017)、Neubert 等人(2017)利用先验的高精度地图合成不同视角数据的方法进行定位。但由于从 3D 模型上合成数据需要较大的计算量,这些方法通常需要借助 GPU 来加速。还有一些工作则关注于如何将相机跟踪获得的点云与先验高精度地图匹配,从而获得相对位姿,如 Caselitz 等人(2016)、Gawel 等人(2016)、Kim 等人(2018)、Agamennoni 等人(2016)、Zuo 等人(2019)的工作。另外,在地图信息利用方面,Lu 等人(2018, 2020)、Park 等人(2019a)、Ye 等人(2020a)也从不同角度分别进行了尝试,从一开始加入手工的标记地面的约束,到加入其他平面信息的约束(例如 Surfel 面的约束)来改善融合效果。

1.4.2 融合其他传感器的 SLAM

视觉 SLAM 完全依赖视觉传感器,导致在视觉退化的场景,视觉 SLAM 精度和鲁棒性都会受到严重影响。在光照动态变化、弱纹理和快速运动等场景下,视觉 SLAM 容易失败。因此,探索视觉和多传感器信息融合对实现高精度和高鲁棒性 SLAM 具有重要意义和价值。其他常见的传感器包括 IMU(in-

ertial measurement unit)、TOF(time of flight)、红外、激光雷达、Wifi、蓝牙、地磁、超声波、里程计、GPS 和热成像相机等。

按照利用多传感器信息的方式,多传感器融合方法可以分为松耦合和紧耦合两类。无论是基于滤波还是优化的估计方法,松耦合分别处理视觉和其他传感器的运动约束,然后融合这些约束。该方法具有较高的计算效率,对于传感器失效更加容易处理,但视觉约束和其他传感器约束的解耦会导致一些信息丢失。紧耦合直接融合视觉和其他传感器测量信息,实现更高精度、更鲁棒的 SLAM。

松耦合的典型代表为 Weiss 等人(2012)、Lynen 等人(2013)提出的基于 EKF(extended kalman filter)的时延补偿的多传感器融合(视觉、IMU 和气压计等传感器)框架。紧耦合方面,Mourikis 和 Roumeliotis(2007)最早提出基于滤波的融合视觉和 IMU 的 MSCKF(multi-state constraint Ralman filter)紧耦合方法。MSCKF 联合估计视觉位姿和 IMU 状态,将视觉信息零空间化为视觉/IMU 位姿约束,不估计视觉特征点位置,从而降低优化复杂度。OKVIS(open keyframe-based visual-inertial SLAM)(Leutenegger 等,2015)采用非线性优化的紧耦合视觉和惯性观测,达到比基于滤波的方法更高的精度。Campos 等人(2020)对纯视觉的 ORB-SLAM2(Mur-Artal 和 Tardós,2017)加入 IMU 观测扩展为完整的 VISLAM(visual-inertial SLAM)系统 ORB-SLAM3,并新增多地图等功能,在公开数据集 EuRoC 上达到目前最优的精度和稳定性。目前,视觉和惯性紧耦合已成为大多商业化产品使用的方案,如谷歌 ARCore、Tango、苹果 ARKit 和微软 Hololens 等。

除了融合视觉和惯性观测以外,Zhang 和 Singh(2018)提出了一种序列数据处理流程来完成激光—视觉—惯性里程计和建图,能够应对环境退化和剧烈运动问题。Shao 等人(2019)提出融合视觉、惯导和激光雷达的 SLAM 系统,将双目视觉惯性里程计与激光雷达建图定位相结合,同时结合了基于视觉和激光雷达的回路闭合,实现了更高的准确性和鲁棒性。

Mascaro 等人(2018)提出融合视觉惯性里程计(visual-inertial odometry, VIO)和 3DoF 全局测量(GPS)方案,将融合问题转变为全局坐标系和局部坐标系(VIO 坐标系)对齐问题,通过包含最近多个状态的滑动窗口优化不断更新全局和局部坐标系变换关系。Lee 等人(2020)提出 GPS 和 VIO 紧耦合方案,解决 GPS 和 VIO 的初始化及在线标定问题,论证融合后全局 6DoF 的可观测性。

Kanhere 和 Rappaport(2019)和 Rappaport 等人(2019)借助无线信号引入更多有关定位技术。KO-Fusion(Houseago 等,2019)融合了视觉和轮式里程表,同年 Khattak 等人(2019)提出在视觉上退化的环境(如黑暗)中,使用带有 IMU 的热像仪。

在深度学习方面,Clark 等人(2017)首次提出使用神经网络实现端到端的视觉和 IMU 融合(VI-Net)。Shamwell 等人(2020)提出无监督的深度神经网络,融合 RGBD 图像与 IMU。它能学习在未知 IMU 内参或 IMU 和相机之间的外参下,执行视觉惯性里程计。Chen 等人(2019)提出策略性选择视觉和惯性有效信息融合的端到端网络,增加系统鲁棒性。

1.5 三维几何建模

深度学习为数据驱动三维建模带来了深刻变革和全新挑战。深度神经网络学习得到的三维几何和拓扑表征,可以用于从几何推断、结构推理到语义理解多个层次的数据驱动三维建模,实现更加智能、灵活和通用的三维内容生成。同时,端到端训练的深度网络可以将局部的几何合成和全局的结构—语义约束统一在一个生成模型中,实现从低层表征到高层推理的“全栈式”数据驱动建模,改变了以往方法对人为定义三维表示及相应约束规则的依赖。

1.5.1 深度三维表征学习

三维深度学习大约从 2015 年开始得到大量研究关注。为了将卷积神经网络应用于不同的三维表示,人们尝试在多种几何表示上设计卷积操作。最直接的做法是在三维体素表示上的三维卷积操作(Wu 等,2015)。但是体卷积的计算和存储开销十分大。因此,人们提出了基于八叉树空间划分的自适应体卷积技术(Wang 等,2017a)。另一种较直接的做法是将三维几何体表示为多视点投影的二维图像,这样就可以提取二维投影图像的卷积特征,并通过多视点融合来实现三维表征学习(Su 等,2015)。在几何处理领域中,曲面网格是最通用的三维几何表示。基于局部或全局网格参数化,可以定义三维曲面的卷积操作(Groueix 等,2018;Sinha 等,2016)。本质上来说,曲面参数化将三维曲面卷积

转化为二维参数域卷积。但曲面参数化本身是非常困难的,且难免引入参数化误差。三维数据获取的最原始形态是点云,直接对三维点云进行卷积操作,似乎是最通用灵活的三维表征学习方式。自从第1个点云深度网络PointNet(Qi,2017a)提出以来,点云深度学习已有大量研究工作。点云卷积的研究主要集中在两个方面:一是如何获取适合卷积计算的局部邻域;二是如何定义和学习有效的三维卷积核。

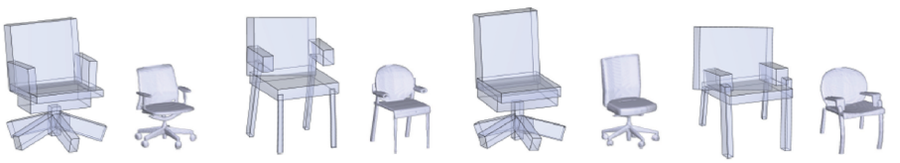
1.5.2 深度三维生成模型

深度表征学习为基于深度学习的三维几何生成奠定了基础。基于体卷积构建深度置信网络,美国普林斯顿大学的Wu等人(2015)提出了第1个深度三维生成模型。该网络可以用于基于单视角深度图像的三维物体补全。随后,美国麻省理工学院(Massachusetts Institute of Technology, MIT)的Wu等人提出基于体表示的三维对抗生成网络3D-GAN(3D generative adversarial network)(Wu等,2016)。该模型通过对抗训练的方式学习得到三维形状空间,实现了三维模型的随机生成。同样基于体表示,美国卡内基梅隆大学(Camegie Mellon University, CMU)的Girdhar等人(2016)研究了基于自编码器的体素生成模型,实现了基于单幅图像的三维物体重建。Kar等人(2017)提出了可微分的多视点立体视觉,用于学习从多视点图像生成三维几何。直接合成三维隐式场的深度网络模型得到了较多关注(Chen等,2019;Park等,2019b)。

1.5.3 结构化表征学习与生成模型

现有方法大多基于结构无关的几何表示,其只关注生成结果的几何外形,无法保证拓扑正确性和结构合理性。根据三维模型结构的定义(Mitra等,2014),结构相关的三维表示应是部件相关(part-aware)的,能够表达三维模型的部件构成及部件间的关系。结构相关的深度表征学习面向三维模型的结构特征,并实现结构推理,以保证合成三维模型的结构合理性。Li等人(2017)提出了第1个结构相关的三维生成模型。该模型采用递归神经网络来编码三维模型部件的层次结构,使网络学习到部件本身的几何和部件间的关系(包括连接和对称)特征。基于这种层次编解码网络设计的生成模型,可实现结构相关的模型自动生成。图2给出了基于结构无关表示(体素)和上述结构相关表示实现的三维模型生成结果对比。可以看出,后者通过直接表征三维模型部件结构,可以高质量地生成部件细节,并保证部件间关系的合理性,优于人为定义的启发式结构规则或约束(Xu等,2012;Averkiou等,2014;Jain等,2012)。但是,该方法只能生成包围盒表示的部件结构,而部件的几何细节需要训练额外的网络在每个包围盒内进行合成。随后,该模型还应用于基于单幅图像的结构合成(Han等,2017)、三维模型的保结构部件重组(Zhu等,2018b)、三维场景的结构化生成(Li等,2020c),以及基于多叉层次树的保结构生成(Mo等,2019b)。

(a) 基于体表示的三维模型生成(3D-GAN(Wu等人, 2016))



(b) 基于结构化表示的三维模型生成(GRASS (Li等人, 2016))

图2 不同方法三维模型生成效果对比

Fig. 2 Comparison of the effect of 3D model generation by different methods

((a) 3D-GAN(Wu et al., 2016); (b) GRASS(Li et al., 2016))

随后,人们研究了结合三维体素和部件表示的保结构生成模型(Wu等,2019b;Wang等,2020c),使网络在全局结构与局部几何之间进行迭代优化,实现由粗到精、由全局结构到局部细节的三维生成,

达到了高质量的合成结果。Genova 等人(2019)提出了结构化的隐式场表示,采用多个隐式函数的融合来表示三维几何,同时编码了隐式函数间的对称关系,因此该表示同时蕴含了几何和拓扑信息,可用于合成结构合理且细节丰富的三维模型。Mo 等人(2019a)采用层次化表示模型间的差异来实现几何和拓扑变化空间的学习。

1.5.4　基于深度学习的三维重建

三维深度学习强大的表征学习能力和几何推理能力,为基于单视点图像或不完整几何数据的三维重建或恢复带来实质性推动。目前主流方法大致有两种:一是基于几何推理的判别式模型;二是面向形状空间学习的生成模型。前者训练端到端神经网络,将输入图像或几何数据直接映射到目标三维模型。本质上,这种网络学习到了某类三维对象的形状先验。后者训练深度生成模型,学习三维对象的形状空间,然后基于度量学习将输入图像或几何数据嵌入到该形状空间中,最后从该嵌入向量解码出三维模型,从而实现对输入的三维重建。基于深度学习的二维草图三维重建得到了长足发展。最近,基于单幅图像的结构化三维重建也开始得到更多关注(Niu 等,2018;Wu 等,2019a)。

1.6　人体动态重建

人体静态三维重建方法中,静态刚体三维重建已有相对成熟的解决方案,但对于动态非刚体重建依旧面临重重挑战,实现人体动态三维重建依旧是极具挑战性的课题。

1.6.1　多视 RGB 重建方法

早期人体动态三维重建,常基于多视图 RGB 的方法。Matusik 等人(2000)使用多视图信息实时重建动态物体的凸包(visual hull);Vlasic 等人(2009)利用多视图信息重建人体骨骼运动,并且利用轮廓约束对人体模型表面非刚性变形处理,重建更为精细的几何表面结构,另外,他们还形变特定模板并拼接多视角模型片段,最终获得具有时间一致性拓扑结构重建。

1.6.2　单深度相机和多深度相机方法

单深度相机方面,基于 KinectFusion 提出的 DynamicFusion (Newcombe 等, 2015), 使用单个 Kinect 进行深度序列拍摄,创造性地将体融合三维重建技术和嵌入式变形图模型的表面非刚性跟踪技术糅合在一起,在 GPU 上演算,进而实现了实时单视角动态场景三维重建。VolumeDeform (Innmann 等, 2020) 算法引入 SIFT (scale-invariant feature transform) 特征点约束和新的体变形模型, 在 DynamicFusion 的框架下加入了彩色稀疏特征约束, 采用的由粗到精的分层优化策略提高了 Dynamic-Fusion 动态重建的精度。

单一视角下观测存在遮挡的严重缺陷,为人体模型动态重建带来巨大挑战,诸多研究者继而转向基于多深度相机方法。其中 Fusion4D (Dou 等, 2016)最为典范,Fusion4D 使用 8 个 Kinect,基于 ED (embedded deformation)变形模型提出了一个用来估计非刚性变形场的能量方程,并针对该方程能量优化的特点提出了基于 GPU 的实时演算方法。Fusion4D 通过复杂场景(快速动作、重建对象拓扑结构改变)的多个摄像机深度信息的融合(data fusion),在保留细节信息的同时获得良好的重建效果。

1.6.3　单视图 RGB 方法

单视图 RGB 方法(深度学习)通过数据驱动方式,捕捉到姿态和形态的先验信息,使人体的三维建模在可靠性和稳健性方面有显著进步。

Dynamic-HMR(human mesh recovery) (Kanazawa 等,2019)提出了一个端到端的 3D 人体动力学的理论模型,该模型由观察运动中人类的视觉序列所获得,通过简单但有效的图像特征时间编码,可以类似地从视频中学习人类 3D 动态的表示;PIFU (pixel-aligned implicit function) (Saito 等,2019) 基于一张图片,先预测一个连续的空间内/外概率场,通过提取 0.5 等值面得到人体的模型表面。然后纹理预测过程为人体几何表面的 3D 点预测一个 RGB 颜色值。在测试阶段网络支持单张或多张图片输入,最终得到一个带纹理的 3D 人体模型;PIFUHD (Saito 等,2020)提出了一个多层次的端到端网络来从分辨率为 1 000 像素的输入图像学习穿衣服的人体几何形状,不需要后处理就能保留原始输入图像中的细节信息。文中使用像素对齐的隐式函数表(pixel-aligned implicit function, PIFu),通过由粗糙到精细的方式逐层加入所缺失的图片细节信息,最后获得人体模型。

1.7　点云语义理解

1.7.1　点云语义分割

点云语义分割是指根据点云内部的空间几何结

构和形状信息将点云分成具有不同语义标签的点集。依据对输入点云数据的不同处理方式,可将现有的分割方法分为两大类,即:基于投影的方法和基于点的方法。

1)基于投影的方法。这类方法通常先将不规则点云投影到某种中间形式的规则表达(如投影图),而后借助成熟的二维或三维卷积神经网络实现点云分割。目前可用的中间表示可以分为以下6种:多视图表示、球面表示、体素表示、栅格表示、柱面表示和混合表示。

(1)多视图表示。Lawin 等人(2017)首先通过多个虚拟相机视角将3D点云投影到2D平面上,然后使用一个多分支全卷积网络以预测合成图上各个像素的类别分数,最后通过融合不同视角的预测分数以获得逐点的语义标签。Tatarchenko 等人(2018)引入切面卷积(tangent convolution)进行稠密的点云语义分割。该方法首先将各个点周围的局部曲面投影到虚拟的切平面上,然后使用切面卷积在虚拟切平面上预测点云的语义标签。总的来说,基于多视图表示方法的分割性能受视角选择和遮挡等因素的影响较大。

(2)球面表示。针对激光雷达获取的点云,Wu 等人(2018a)提出了一种基于 SqueezeNet 和条件随机场的端到端网络 SqueezeSeg,以实现快速精确的点云分割。随后,他们又提出了 SqueezeSegV2(Wu 等,2019a)以进一步提升语义分割的精确度。该方法利用一种基于无监督的域自适应以解决不同域差异大的问题。此后,他们发现 LiDAR 图像的特征分布在不同的图像位置会急剧变化,然而标准的卷积滤波器仅在图像中特定区域提取有效的局部特征。为此,他们进一步提出 SqueezeSegV3(Xu 等,2020)并提出空间自适应卷积(SAC),根据输入图像对不同位置采用不同的滤波器。与单视角投影相比,球面投影表示保持了更多的信息。同时,球面表示也不可避免地引入了离散化误差。

(3)体素表示。Tchapmi 等人(2017)提出了 SegCloud 来实现细粒度和全局一致的语义分割。该方法引入了三线性插值法,将3D-FCNN 预测的粗糙体素标签映射回点云,然后使用条件随机场来增强点云标签的空间一致性。然而,由于体素表示的稀疏性,直接使用卷积神经网络预测点云语义标签的方法是非常低效的。Graham 等人(2018)提出了一种子流形稀疏卷积网络,该网络通过限制卷积的输出只能与非零体素有关的方式,可以显著地减少基于体素分割方法的内存和计算资源的消耗。Choy 等人(2019)提出了一种 4D 时间—空间卷积网络 MinkowskiNet 以完成点云视频序列分割任务。

(4)栅格表示。Su 等人(2018)提出了一种基于双边卷积的稀疏栅格网络 SPLATNet。该方法首先将点云插值到栅格中,然后使用双边卷积对非零栅格进行卷积处理以得到逐点语义标签。Rosu 等人(2019)提出了 LatticeNet 以实现高效的大规模点云处理。该方法引入了一个名为 DeformsSlice 的插值模块以将栅格特征反投影到点云上。

(5)柱面表示。张艳国和李擎(2018)提出了一种针对激光点云雷达的神经网络 PolarNet。不同于球面投影及鸟瞰图(bird's-eye-view),该方法将点云投影到极坐标系均匀分布的网格中,在一定程度上缓解了激光雷达点云在径向维度上的长尾问题。此外,该方法不依赖于耗时的 K 最近邻搜索。Zhu 等人(2020)进一步提出对激光雷达点云采用圆柱划分,并引入非对称卷积网络以学习到有效的三维几何特征。相比于标准的体素划分,圆柱划分更好地遵循了激光雷达点云的特性,因此非空的区域更少,点在区块中的分布也更加均匀,尤其是对于距离传感器较远处。

(6)混合表示。为进一步利用所有的可用信息,许多方法试图利用多模态数据以预测精确的点云标签。Dai 和 Nießner(2018)提出了一种联合 RGB 图像和点云几何信息的端到端网络。Jaritz 等人(2019)提出了 MVPNet 来聚合 2D 多视图图像的外观特征和点云的空间几何信息。Liu 等人(2019d)提出了点—体素联合表征,将三维输入数据用点表示以减少内存消耗,同时在体素中进行卷积以减少不规则、稀疏数据访问并改善局部性(locality)。该方法有非常高的存储效率以及计算效率。此后,他们又进一步引入当前的 3D 神经架构搜索(3D network architecture, search, 3D-NAS),以实现在多样化的设计空间中高效地搜索最佳网络架构。此外,该模型还能够较好地分割三维空间中的小实例(例如行人等)。

2)基于点的方法。基于点的方法是直接在点云上进行操作的。然而,由于点云数据的无序性和不规则性,规则的卷积神经网络并不能直接应用到

非规则点云。PointNet(Qi 等,2017a)是该方向的先驱工作,它通过多层感知层学习逐点特征,并通过最大池化层聚合得到全局特征以预测点云语义标签。随后,为更好地建模每个点的局部上下文信息,近年来大量研究试图通过邻域特征池化,不规则点云卷积,循环神经网络建模或图网络等方法进一步提升分割性能。

(1)邻域特征池化。为了获取局部的几何表示,Qi 等人(2017b)提出了一种基于局部邻域点集合的方法 PointNet ++。该方法首先利用最远点采样和 K-最近邻聚类以获取各个点云的局部邻域,然后采用 PointNet 框架学习和聚合各点云邻域的特征。为进一步提升点云分割性能,注意力机制被引入点云分割以聚合逐点特征。Zhang 等人(2019b)提出使用来自同心球面壳的统计信息来定义代表性特征并解决点云顺序的模糊性,从而使传统的卷积可以在这些特征上执行。Zhao 等人(2019)提出了一种基于注意力机制的分数精化模块对点云分割结果进行后处理。Zhao 等人(2019)提出通过密集构建局部完全连接的网络来探索局部邻域中所有点对之间的关系。该方法提出了一种自适应特征调整(adaptive feature adjustment, AFA)模块来实现信息交换和特征细化,此聚合操作有助于网络学习判别式特征表示。为实现高效的大场景点云语义分割,Hu 等人(2020)提出了基于随机采样的轻量级点云语义分割框架 RandLA-Net,该方法不断地对点云进行随机采样以节约内存及计算消耗,并通过层次化局部特征聚合显著增加感受野,从而弥补随机采样带来的信息损失。RandLA-Net 在大规模点云语义分割任务中有显著优势。

(2)不规则点云卷积。这类方法旨在设计一个高效的卷积操作用于点云数据。PointCNN(Li 等,2018c)通过 X-conv 变换(通过 MLP(multilayer perceptron)实现)将输入的无序点转换为潜在有序的,然后在变换后的特征上采用典型的卷积运算以得到点分割结果。Wu 等人(2019c)将点云连续卷积定义为一个基于重要性采样的蒙特卡洛估计,其中卷积核是由加权函数(通过 MLP 层学习)和密度函数(通过核化密度估计和 MLP 层学习)组成的。Thomas 等人(2019)提出了一种基于 KPConv(Kernel point convolution)的核点全卷积网络(Kernel point fully convolutional nerual network, KP-FCNN)。KPConv 通过计算欧氏空间距离将逐点特征映射到预定义核点(kernel points),而后在规则核点上进行卷积操作。

(3)循环神经网络建模。Engelmann 等人(2017)首先将点云块转换为多尺度块和网格块以获得场景上下文,然后使用 PointNet 提取逐块特征并馈入循环单元以获得语义标签。Ye 等人(2018)提出了一种 pointwise pyramid pooling (P3P)模块以获取从粗到细的局部特征,并利用双向循环神经网络来实现端到端的学习。

(4)图网络。Landrieu 和 Simonovsky(2018)将点云表征成一组超点以及一系列内部相互连接的简单几何形状,并建立有向超点图来描述点云内部的几何结构及上下文信息。此后,Landrieu 和 Boussaha(2019)将点云过分割问题表征为邻接图的深度度量学习问题,并提出了有监督的过分割网络来提升几何划分性能。

此外,最近的一些工作开始尝试在弱监督或者自监督条件下实现点云语义分割。Wei 等人(2020)提出了基于点云场景标签或子点云标签的两阶段弱监督语义分割方法。Xu 等人(2020)研究了点云语义分割的几种弱监督方案。该方法在仅使用部分标记的点(例如 10%)进行训练的情况下,依然可以实现不错的效果。Xie 等人(2020)提出 PointContrast 框架,将点云配准作为辅助任务以实现点云网络的预训练。Wang 等 (2020b)提出 OcCo 框架,利用点云补全作为辅助任务以实现网络预训练。

1.7.2 点云实例分割

相比于语义分割,点云实例分割不仅需要区分具有不同语义标签的点,而且还需要分离具有相同语义信息的不同实例。总体而言,点云实例分割方法可以分为以下两类:基于候选目标框(proposal-based)的方法和无候选目标框(proposal-free)的方法。

1)基于候选目标框的方法。这类方法将实例分割问题转化为目标检测和实例掩码预测两个子任务。Hou 等人(2019)提出的 3D-SIS(3D semantic instance segmentation)方法首先使用三维候选目标区域建议网络及感兴趣区域池化层预测目标边界框,而后对每个目标框预测掩码以实现实例分割。不同于直接回归得到候选目标框,Yi 等人(2019)采用生成模型来描述目标在自然空间的分布,进而通过从

对象分布中采样来产生高质量且具有几何语义的目标候选区域。Yang 等人(2019)提出了一个单阶段、端到端且无锚点的实例分割算法框架 3D-BoNet。该方法根据点云全局特征显式地回归场景中目标的大致边界框,并将三维候选目标框生成问题建模为最优关联问题。得益于日趋成熟的三维目标检测框架,基于候选目标框的方法通常具有较好的实例分割效果。然而,此类方法通常需要两阶段训练及对冗余目标候选区域的精炼和剪枝,因而计算量较大。

2)无候选目标框的方法。这类方法通常假设属于同一实例的点在特征空间中分布更加紧密,进而将实例分割问题转化为特征空间中的聚类问题。Wang 等人(2018c)在 SGPN(similarity group proposal network)中首先引入相似性矩阵来度量点与点的特征相似性,并使用 double-hinge 损失函数来调整相似矩阵以获得更具区分度的点特征。考虑到点的语义和实例标签通常互相依赖,许多方法尝试将语义和实例分割相耦合以提升性能。Wang 等人(2019c)通过引入可端到端学习且相互关联的语义—实例分割模块,利用互补的语义和实例特征进一步提升联合分割的效果。Pham 等人(2019)首先引入基于点的多任务网络和判别损失来规范每个点在特征空间中的表达,而后采用条件随机场对点特征和预测得到的语义标签进行联合优化,最后使用均值场变分推理获得最终的语义及实例标签。Lahoud 等人(2019)则将多任务学习理念引入到点云实例分割中。Jiang 等人(2020)提出了名为 PointGroup 的实例分割网络,该网络由语义分割分支和偏移量预测分支组成。该方法还进一步采用了 dual set 聚类算法和 ScoreNet 以获取更好的分组结果。Engelmann 等人(2020)提出了一种以目标为中心的方法,每个点都为所属的实例中心投票,进而生成初步的候选区域。不同于以往的方法,该方法的最终对象实例是通过聚合多个候选区域而不是使用非极大值抑制来剪枝得到的。

Han 等人(2020)将“三维占用尺寸”定义为每个实例占用的体素数量,并提出一种名为 OccuSeg 的实例分割算法。该算法取得了在 ScanNet 数据集上的最佳性能。无候选目标框的方法虽然省去了计算代价高的目标候选框生成步骤,但通常依赖于启发式的聚类及非极大值抑制等后处理,且实例分割目标性(objectness)不好。

2 国内研究进展

2.1 立体匹配

2.1.1 非端到端立体匹配

Chen 等人(2015)提出了一个结合了多尺度特征的嵌入模型,直接对提取的特征向量进行内积操作来表示图像分块之间的相似度。内积的使用可以使网络不再需要多个全连接层,降低了网络的计算复杂度,其计算速度相比 MC-CNN(Žbontar 和 LeCun,2015)提高了 100 倍。

2.1.2 端到端立体匹配

Pang 等人(2017)提出了一个两阶段的层级残差学习网络(cascade residual learning, CRL),其第 1 阶段预测初始的视差估计值,然后在第 2 阶段的网络中从不同尺度上产生残差估计来对初始的视差值进行级联的优化与校正。Liang 等人(2018)进一步拓展了 Dispnet,提出了一个基于特征一致性假设的视差优化网络 iResNet。iResNet 利用特征相关性和重建误差来使得网络更加容易训练。尽管 CRL 与 iResNet 都采用了首先预测初始视差值,再进一步优化的策略,然而 iResNet 的层级残差优化网络利用了不同尺度的特征信息并且与初始估计网络有更多的信息交换,因此 iResNet 的效果相较于 CRL 更好。Song 等人(2018)提出了一个边界感知的视差估计网络, 充分利用物体边界的特征信息并使用了边界感知的平滑惩罚项作为额外监督信号,其在 KITTI 双目数据集上获得了优异结果。Yang 等人(2018a)提出了一个结合语义信息的视差估计网络,语义信息的引入提高了在已知物体类别上的视差估计效果。

在 GC-Net 的启发下,Chang 和 Chen(2018)提出了金字塔立体匹配网络(pyramid stereo matching network, PSMNet)。相较于 GC-Net,PSMNet 使用了一个空间金字塔特征编码器,将不同尺度下的特征进行充分融合,再构建四维的代价体,最后使用三维卷积网络对代价体进行聚合并估计视差值。不同尺度下特征的融合,使得匹配网络更加鲁棒,提高了其在弱纹理区域的效果。Cheng 等人(2020b)基于 PSMNet,引入了卷积空间传播网络(convolutional spatial propagation network, CSPN),利用像素间的关联性进一步提升了双目估计的精度,在当年 KITTI

数据集上排名第 1。Liang 等人(2018, 2021)充分利用多层次代价体及多尺度特征一致性,显著提升了双目立体匹配任务的精度,并获得了很好的泛化能力,在 CVPR 2018 Robust Vision Challenge 中获得了立体匹配赛道的冠军。

虽然三维卷积网络在代价聚合和视差估计上有着非常明显的优势,但是三维卷积的引入,导致其需要耗费大量的 GPU 内存资源,并且非常耗时。为了解决内存消耗过大的问题,Gu 等人(2020)提出了一个多层级由粗到细(from coarse to fine)的多分辨率立体匹配网络。不同于 PSMNet 仅仅对单一视差间距的代价体进行聚合优化,该网络使用多个视差间距,构建多个代价体。其可以根据上一级的视差估计值,使用更精细的视差间距进行代价体重构。由于避免使用 PSMNet 中完整的精细代价体,该网络显著减少了 GPU 内存的使用。

Jie 等人(2018)提出了一个左右对比的循环模型,在进行视差估计的同时执行左右一致性检测。虽然循环模块的使用提高了视差估计的精度,但是 LSTM(long short-term memory)结构相比普通卷积模块更加耗时。在实际应用中,人们不仅关注准确的视差估计值,也往往希望知道估计的视差值是否可靠。

2.1.3 无监督立体匹配

Zhou 等人(2017a)提出了一个迭代训练的神经网络,从随机初始化的网络参数开始,使用左右一致性检测来指导训练过程。在迭代训练的过程中,使用上一次更新的网络参数来生成估计的匹配值,再根据左右一致性检测挑选出可靠的匹配值作为下一次网络训练的数据标签。在迭代训练的过程中,随着网络被逐渐优化,生成的数据标签也逐步变得可靠,最终实现无监督的立体匹配网络的训练。虽然该迭代优化的网络结构取到了能够与有监督网络接近的效果,但是其迭代训练的过程需要消耗大量的计算资源与时间。

Luo 等人(2018a)提出了一个新颖的策略,首先使用一个图像生成网络,根据双目图像中的左视图来生成对应的右视图,再使用有监督立体匹配技术中使用的匹配网络对左视图与其对应的合成右视图进行匹配估计,进而估计其视差值。Wang 等人(2019c)提出了一个无监督光流与双目视差估计的神经网络,其考虑连续的双目图像对,对于每一对双目图像,其利用立体匹配网络估计对应的视差值,对于两张连续的左视图,则估计其相对的相机运动与对应的光流。光流与视差的联合估计,使其在 KITTI 双目视差估计数据集上表现出了优异的成绩。

Liang 等人(2019)提出了对于交叉谱(cross-spectral)图像对的无监督立体匹配技术,不同于对相同频谱下的一对双目图像进行视差估计,该工作聚焦于不同频谱下的双目图像。Wang 等人(2020d)为了估计有较大视差变化的双目图像,提出了视差注意力机制(parallax attention mechanism, PAM),直接学习双目匹配,避免了基于代价体方法的最大视差距离限制。Liu 等人(2020b)提出了一个无监督预测光流与视差的统一网络结构,将视差估计当做光流估计的特殊情况,利用立体视觉的三维几何信息指导同一网络来估计光流与视差。

在数据集方面,国内也有良好进展。Bao 等人(2020)开放了一个包含 2 050 对室内场景真实立体图像对的 InStereo2K 数据集。该数据集的真实视差图由一个结构光相机获得,是当前学术界较大的一个公共数据集。为促进无监督立体匹配的发展,Wang 等人(2020d)开放了一个包含 1 024 对立体图像对的 Flickr1024 数据集。该数据集覆盖了多样化的场景,包括城市、地表、人及人造物体等,可用于无监督立体匹配训练及双目超分辨等任务。

2.2 单目深度估计

在深度回归网络方面,国内开展了卓有成效的工作。Gan 等人(2018)认为现有的编码网络结构能够有效提取边缘、纹理等强特征,但忽视了相邻像素间的深度约束关系。为解决这一问题,他们在编码网络基础之上增加了上下文网络(context network),该网络包含了用于计算邻域像素间相关性的关联层(affinity layer),进而同时提取了局部和全局的特征信息。Ye 等人(2020b)改进了非局部空间关注模块,该模块能够显式地利用空间域的非局部相关性。同时,该工作使用了一种双分支深度估计网络结构,分别捕获低层和高层的特征表示。Cheng 等人(2018,2020a)同样使用包含跳跃链接的沙漏状编码解码网络,并引入卷积空间传播模块,该模块采用循环卷积的方式,使得卷积神经网络能够学习相邻像素之间的关联性。Liebel 和 Körner(2019)提出了任务指定型解码器网络,该网络使用了深度回归和多任务分类的金字塔池化层(Zhao 等,2017)。

当前大部分研究集中于透视图下的单目深度估计,而如何解决存在图像形变下的全景图单目深度估计是一个新的研究热点。Chen 等人(2021)将形变卷积和条带池化相结合,提出了一个具备形变感知能力的单目全景图深度估计网络 DAMO(distortion-aware monocular omnidirectional network),该网络在 360D 数据集上获得了优异的性能。

2.3 大场景下的视觉定位

小场景下 3D 结构信息已知进行视觉定位,早期国内也有较大影响力的工作(Wu 和 Hu,2006)。近几年来,在大场景视觉下重要的工作还较少,因此不再分为小类进行叙述,而是按照深度学习特征描述与视觉定位两大类来介绍。

2.3.1 基于深度学习的特征描述

在国内,中国科学院自动化研究所、香港科技大学和浙江大学等单位在基于深度学习的特征方面做出了一些较好工作。L2-Net(Tian 等,2017)将描述子的紧凑性约束加入到描述子的学习中,以此同时获取二进制和实值描述子。GeoDesc(Luo 等,2018b)采用了多视图重建的几何信息去约束描述子的学习,并提供了一个基于学习的描述子在 SfM(structure from motion)中应用的准则。SOSNet(second order similarity network)(Tian 等,2019)在二进制的学习中加入了二阶相似性的约束,并提出了基于 von Mises-Fischer distribution 的描述子评判方法。GIFT(group invariant feature transform)(Liu 等,2019c)从几何变换后的图像中提取出的特征可以看做是一个定义在变换组上的函数,并使用群卷积来提取图像特征的底层结构,以此增强特征的判别性。ASLfeat(Luo 等,2020)利用 DCN(deformable convolutional networks)和不同层次的特征能够更好地学习图像的形状信息,在一些有代表性的描述子数据集(Balntas 等,2017)和定位数据集(Sattler 等,2018)取得了不错的效果。Wu 和 Wu(2019)提出了哈希深度学习特征描述,对定位地点进行了识别,具有较高精度的同时在 CPU 端也可实时运行。Wang 等人(2020f)采用相机的位姿作为监督信息提取了图像局部描述子 CAPS(camera pose supervised)。

2.3.2 大场景视觉定位算法

Feng 等人(2016)利用数据库中 3D 点描述子的标记信息去训练随机树,然后这些随机树结构又用来索引 3D 点,加快了 2D-3D 匹配的速度。Zhang 等人(2020a)提出并行搜索的视觉定位方法,其同时利用局部描述子和全局描述子查找局部特征的近似最近邻,然后在查找到的近似最近邻里进行线性搜索,这样能找到更多正确的 2D-3D 匹配,进而提高定位精度。Shi 等人(2019)首先对查询图像和数据库图像进行语义标注,然后在 RANSAC 阶段,结合语义信息剔除外点。这种方法荣获 2019 年 CVPR Workshop "Long-Term Visual Localization Challenges" 冠军。进一步,Shi 等人(2020)使用混合的手工特征子以及学习的特征子,结合稠密的语义地图,实现了跨光照的长期视觉定位。香港科技大学的研究团队 Ye 等人(2020a)通过直接优化光度误差来实现持续的跟踪和定位。另外,百度、滴滴和商汤等公司也做了较多视觉定位相关的工作,百度在 2017 年提出了一个测评视觉定位算法的数据集(Sun 等,2017),并在 2020 年提出了一种基于注意力机制的端到端视觉定位方法(Zhou 等,2020b)。滴滴 MapVision 团队基于现有深度学习描述子,提出基于困难样本挖掘的二次合页损失函数改进方法,然后用改进的描述子进行相对位姿估计,并获得 CVPR 2020 Image Matching Challenge 挑战赛冠军。浙江大学和商汤科技在 ICCV2019 联合提出动态场景下端到端的视觉定位方法(Huang 等,2019b),之后他们还利用卷积神经网络学习二分图匹配,并将其应用在视觉定位任务中,取得了一定的性能提升(Yu 等,2020)。

2.4 同步定位与地图构建

2.4.1 视觉 SLAM

哈尔滨工业大学的研究人员提出了基于局部子图扩展卡尔曼滤波器的 SLAM 系统(Zheng 和 Zhang,2019),可以在保证传统 EKF-SLAM 的计算精度下,有效降低计算复杂度。香港科技大学研究团队提出了基于滑动窗口非线性优化的紧耦合单目视觉惯性里程计 VINS-Mono(Qin 等,2018),并在后续又推出了可以结合双目视觉乃至 GPS 的惯性里程计 VINS-Fusion(Qin 等,2019)。浙江大学研究团队提出了基于多平面先验的高效视觉惯导里程计 PVIO(plane-based visual-inertial odometry),有效提高了 VIO 在多平面场景下的鲁棒性(Li 等,2019b)。北京大学研究团队提出了基于线流表达的 SLAM,可以在人造的结构化场景中提升精度和鲁棒性(Wang 等,2020e)。中国科学院自动化研究所研究

团队提出了直接法与特征法融合的 FMD SLAM(fusing MVG and Direct SLAM)，兼具了速度与精度(Tang 等，2019)。上海交通大学研究团队提出了基于场景文字纹理的视觉 SLAM 系统 TextSLAM(Li 等,2020a),将场景中的文字信息转换为具有丰富纹理和语义信息的平面特征。

中国科学院大学研究团队提出将单级物体检测器加入到语义 SLAM 中,并且利用相邻帧速度预测来补偿物体漏检,有效提升了检测器的召回率,同时弥补了传统 SLAM 在动态场景中不能有效进行跟踪的缺陷(Xiao 等,2019)。北京大学研究团队在现有的深度学习视觉里程计中,嵌入了记忆和精化组件(Xue 等,2019),有效地保留了全局信息和上下文信息,提升了预测的轨迹精度。上海交通大学研究团队提出的 Attention-SLAM(Li 等,2020b),将视觉显著性检测模块应用于传统的 SLAM 中,加强了 SLAM 对于显著性特征的跟踪。

2.4.2　融合其他传感器的 SLAM

在松耦合方面,宁波大学的王泽华等人(2018)提出了磁力计视觉/IMU 松耦合 SLAM,通过 EKF 估计多源传感器位姿之间的尺度变换。北京理工大学的 Yan 等人(2017)提出了 VISO2 和 LOAM 的松耦合双目视觉激光雷达测距方法。

在紧耦合方面,东南大学 Zhu 等人(2018a)提出了一种紧耦合 SLAM 算法,通过计算流动站姿态估计器的 Cramér-Rao 下界来提高 SLAM 精度,实现估计车辆的轨迹及地图构建。Du 等人(2020)提出了一个基于 EKF 的针对惯导、雷达和相机等多种传感器的紧耦合 SLAM 框架。

针对视觉和惯导融合 SLAM,浙江大学的 Ye 等人(2019)提出了一种多相机和惯导紧耦合的 SLAM 系统,该系统利用惯导信息在初始化时估计尺度,同时在多个相机的帧之间进行闭环检测提高闭环检测成功率。上海交通大学的 Zuo 等人(2019)面向 Manhattan 场景引入线结构特征和惯导测量信息,适合于大规模室内场景定位。浙江大学的 Li 等人(2019a)通过挖掘人造环境中的竖直边缘特征,提出了视觉惯导系统的快速初始化方法。

针对视觉和激光雷达融合 SLAM,国防科技大学的 Chen 等人(2017)提出了激光视觉融合建图,将两个视觉关键帧之间的 2D 激光雷达点云数据变换到世界坐标系加入闭环修正,实现高精度建图。Chan 等人(2018)以轨迹匹配融合代替传统的特征匹配融合,适配了包括 2D 激光雷达、相机等在内的多种低精度传感器定位算法。国防科技大学的 Zhu 等人(2018d)在 3D 激光雷达的回环检测中引入了视觉词袋以避免计算点云相似性,达到了实时闭环检测效果。

此外,北京科技大学的 Zhang 和 Singh(2018)提出了一种基于 IMU 的激光雷点云融合 SLAM;南京航空航天大学的廖自威(2016)提出了基于激光雷达/微惯性融合 SLAM。

2.5　三维几何建模

深度表征学习为基于深度学习的三维几何生成奠定了基础。微软亚洲研究院(Microsoft Research Asia, MSRA)的 Wang 等人(2017a)研究了基于八叉树卷积神经网络的三维模型生成,以较低的时间和存储开销实现高质量三维表面模型的合成。基于 PointNet,Fan 等人(2017)提出了首个基于单幅图像的三维点云重建。国防科技大学徐凯教授团队(Li 等,2017)于 2017 年提出了第 1 个结构相关的三维生成模型。Li 等人(2020c)提出了一种同时学习部件生成和组装的结构化三维模型合成网络,该方法只能处理固定数量的部件。Wu 等人(2019b)提出了一种序列化生成部件及其组装变换的网络,可处理任意多个部件。

Li 等人(2018a)采用 CNN 推断二维草图的深度图和向量图以实现三维建模。Han 等人(2017)提出了从二维草图推断三维人脸模型的 CNN 网络。

2.6　人体动态重建

在人体动态三维重建方面,国内的研究单位较少,主要有清华大学,浙江大学,天津大学,东南大学和上海科技大学等。国内的工作大多基于 RGBD 相机进行动态三维重建,近些年均涌现出大量具有影响力的工作。

Liu 等人(2014)使用 3 个手持 Kinect 实现了单个或多个人体模型捕捉。Guo 等人(2015)利用基于 L0 的运动正则项,采用单个商用相机重建非刚性人体几何形状和运动。Guo 等人(2017)提出了一种利用单视点 RGB-D 输入,利用详细的几何模型、表面反照率、每帧非刚体运动和每帧低频光照重建人体场景的方法。Yu 等人(2017)提出了 BodyFusion,在人体表面模型中嵌入骨架,利用人体骨架对表面模型进行驱动,实现了更为鲁棒的人体动态重建。

Wang 等人(2018d)提出了一种使用手持摄像机进行户外无标记人体运动捕捉的方法。利用生成运动捕获方法,采用一种新的模型视图一致性方法,在跟踪阶段同时考虑了前景和背景。将背景建模为可变形的 2D 网格,通过全局局部优化来跟踪 3D 角色姿势。

Yu 等人(2018)提出了 DoubleFusion,该方法定义了一种双层人体三维信息表示方法:外层人体稠密形状以及内层的参数化人体体形(SMPL)。该方法通过合理地构造联合优化方程及参数化人体模型的变形形式,将双层人体三维信息进行了深度耦合,与深度观测对齐、融合,以实现单视点下实时人体动态重建。在此基础上,Zheng 等人(2018)将惯性测量单元与几何融合相结合,提出了 HybridFusion,实现了更为准确的动态人体重建,该方法在身体部位存在自遮挡的情况下可以产生更为精准的重建结果。Yu 等人(2019a)提出了一种人体双向运动捕捉方法,构建跟踪状态监测器与三维姿态检测器来校正错误跟踪。为了精细化重建人体表面衣物纹理,Yu 等人(2019b)提出了 SimulCap,采用弹簧质点模型,通过引入物理信息,实现人体及表面衣物的动态重建。

Su 等人(2020)提出了一种 MulayCap 方法,该方法使用多层表达实现几何重建和纹理渲染。在几何重建中,将穿衣服的人体重建分解为人体网格层和衣片层。对于纹理渲染部分,将每个输入图像帧分解为一个着色层和一个反照率层,并提出了一种融合固定反照率图和利用该阴影层求解服装几何细节的方法,实现了动态变化细节的逼真渲染。

Xu 等人(2018a)提出了 FlyCap,利用多架无人机实现了人体动态重建。Xu 等人(2020)提出了 UnstructuredFusion,使用 3 个无需预标定的商用 RGBD 相机,解决了多个异步视频的 4D 重建问题。Xu 等人(2019)提出了 FlyFusion,利用单个无人机搭载深度相机,实现了人体的动态三维重建,且对重建人体的拓扑变化有较强鲁棒性。之后,Xu 等人(2020)又提出了 RobustFusion,将数据驱动网络与传统人体运动跟踪相结合,实现了更为鲁棒的人体动态重建效果。

2.7 点云语义理解

2.7.1 点云语义分割

北京师范大学的 Liu 等人(2017a)受人类视觉认知模式的启发,提出了一种 3DCNN-DQN-RNN(三维卷积网络—深度 Q 网络—递归网络)的深度强化学习框架,克服了深层次网络难以充分训练的问题,实现了室内外场景点云的精确分割。中国科学院上海微系统与信息技术研究所的 Ye 等人(2018)提出了一种基于上下文和三维递归网络的点云分割方法。香港中文大学的 Jiang 等人(2019)利用每个点及其相邻点之间的语义关系来实现精确的点云场景分割。

三维卷积算子是点云语义分割的一个热门的研究方向,他致力于使卷积神经网络能够处理点云等非欧氏数据。为了捕获点云复杂的几何特征,清华大学的 Xu 等人(2018b)设计了一种以简单阶跃函数和泰勒多项式乘积为卷积核的网络框架 SpiderCNN。针对点云的无序、不规则分布等特点,山东大学的 Li 等人(2018c)提出了一种基于 X-变换的点云分割网络 PointCNN。上海交通大学的 Jiang 等人(2018)受 SIFT 算子的启发,设计了一种可编码点云方向信息和自适应物体尺度大小的网络模块 PointSIFT,该模块可以集成到基于 PointNet 的分割方法中以提升分割精度。中国科学院自动化研究所的 Liu 等人(2019b)提出了一种可学习点云中点与点之间的几何拓扑关系的分割网络 RSNet。受图卷积在非欧氏数据处理上取得成功的启发,武汉大学的 Wang 等人(2019a)提出了一种图注意力卷积,它可以自适应地适应目标的结构。香港中文大学的 Mao 等人(2019)设计了一种插值卷积神经网络,它是利用一组离散的核权值并通过插值函数将点特征插值到相邻的核权值坐标上进行卷积。

充分利用上下文信息是提升点云语义分割性能的有效途径。中山大学的 Liu 等人(2020a)提出了一个点云上下文编码模块 PointCE 来捕获点云的语义上下文信息,从而提升现有点云语义分割网络的性能。中山大学的 Ma 等人(2020)提出了一个点云上下文推理模块 PointGCR 来获得点云在通道维度的全局上下文信息。PointGCR 是一个即插即用且可端到端训练的模块,可比较好地整合到现有点云语义分割网络中以提升其性能。

2.7.2 点云实例分割

清华大学的 Han 等人(2020)提出了一种基于占用(Occupancy)的三维实例分割方法,该方法采用了一种自适应的聚类方案,同时考虑了占用信息和嵌入向量的距离。香港中文大学的 Jiang 等人

(2020)提出了一种自下向上(bottom-up)的点云实例分割框架,该框架是一个双臂网络以分别预测点的语义类别和偏移量。然后,聚类模块结合原始的点云坐标和偏移量以提升点的分组精度。

3 国内外研究进展比较

3.1 立体匹配

立体视觉技术一直是对空间进行三维感知的重要手段之一。相对于激光雷达等主动式感知技术,其设备简单、成本低及使用范围广泛,因此数十年来一直受到广大学者的关注。目前双目立体视觉技术在实际中有着广泛的应用,包括智能机器人导航、目标识别、遥感技术和自动驾驶。在传统的双目立体视觉技术中,关键在于寻找不同视角下图像的对应匹配点,这与光流估计问题高度相似。但是不同于光流估计,双目立体匹配需要满足极线约束。对于校正后的图像,基于极线约束假设,立体匹配的搜索空间可以从二维图像平面缩小到一维的扫描线上。进而视差/深度信息可以通过沿着扫描线方向对两张图片上的像素点进行匹配得到。近些年,得益于卷积神经网络技术的兴起,基于深度学习的立体视觉技术得到了蓬勃的发展。Mayer 等人(2016)首先提出了端到端的立体视觉神经网络,完全可导的网络框架使得视差匹配的过程可以充分结合局部信息与全局的图像信息,取得更好的效果。自此,端到端的神经网络成为基于学习的立体视觉技术的主流。尽管基于深度学习的立体视觉匹配技术对于遮挡、噪声和弱纹理等问题有着更优异的效果,但是数据驱动的方式往往需要大量的标签。在实际中,获取真实场景下的大量视差图或深度图是非常昂贵的,在室外场景中往往依赖于昂贵的高精度雷达进行采集稀疏数据标签的采集。为了解决标签缺乏的问题,近年来,一些无监督立体匹配算法利用空间变换以及视图生成技术来实现无标签下的网络学习,并取得了一定的进展。

3.2 单目深度估计

单目深度估计方法在算法实现上面临极大的挑战。从二维图像估计深度的问题是一个极为病态的问题,需要引入大量先验约束进行求解。早期的研究尝试利用阴影和光照的关系,以及引入外形模板实现从单幅图像中提取深度,但在精度和鲁棒性方面效果欠佳。人类的视觉理解系统为解决这一问题提供了很好的参考,通过获取图像,即三维世界在二维空间的投影,人类的视觉理解系统能够利用先验信息重构出丢失维度,即深度方面的信息。近年来的研究表明,这一过程同样可以通过神经网络实现,即以数据驱动的方式获得单目深度估计模型。近年来,随着深度学习的快速发展,该方法逐渐成为深度估计的主流方法,在网络模型构建、损失函数设计、数据提升和无监督训练等多个子研究课题上取得突破。从单目深度估计问题的研究内容出发,已有工作所提出的模型可分为深度回归模型和深度补全模型。前者着眼于仅从亮度图像预测深度图,而后者尝试结合亮度图像和稀疏的深度信息重构深度图。

3.3 大场景下的视觉定位

相比于 2D 视觉而言,视觉定位方向的研究人员较少。在深度学习快速发展和旺盛应用需求的驱动下,近几年越来越多的人加入了相关研究。当前,具有较大创新性和重要理论贡献的工作大多出自欧美国家,国内虽然在顶级会议或重要期刊上有系列工作或竞赛冠军,但其原始创新影响力还和欧美国家有明显差距。比如,大场景 AR 定位导航的研究思想于十多年前出自诺基亚芬兰研究中心。早在 2008 年,中国科学院自动化研究所就与诺基亚芬兰研究中心就 SLAM 及大场景 AR 定位导航进行研究攻关。目前,国内很多大公司在此方面有大量研究投入,在工程化和应用转化上发展强劲。比如,华为的河图系统在大场景下有较好的体验。

3.4 同步定位与地图构建

SLAM 算法方面而言,国内外的研究差距不大。德国慕尼黑工业大学提出的 DSO(Engel 等,2017),西班牙萨拉戈萨大学提出的 ORB-SLAM3(Campos 等,2020),香港科技大学提出的 VINS-MONO(monocular visual-inertial system)(Qin 等,2018),以及百度和浙江大学合作研发的 ICE-BA(incremental, consistent and efficient bundle adjustment)(Liu 等,2018),浙江大学提出的 PVIO(plane-based visual-inertial odometry)(Li 等,2019b),这些都是非常优秀的开源里程计或 SLAM 框架,其中以 ORB-SLAM3 和 VINS-MONO 影响力最大。

SLAM 最终的性能表现不仅取决于软件算法,还取决于移动终端的硬件适配。在软硬件一体化的发展道路上,国外企业进展显著,如苹果公司的

iPhone、微软的 HoloLens、Facebook 的 Oculus Quest 和高通的 XR 平台等,通过精心设计和标定的传感器配置提升 SLAM 表现,并将算法固化至芯片以提升计算效率、降低设备功耗,从而达到优异的性能表现。国内目前华为已采用了类似的方案,基于华为 P40 系列搭载的麒麟 990 5G SoC 芯片,华为终端云服务可以向用户提供华为 AR 地图服务。而国内其他 OEM(original equipment manufacturer)厂商,大多通过调用第三方的 SDK(software development kit)来实现 SLAM 算法在硬件上的部署和应用。

此外,在室内外大尺度的定位与导航应用中,需要场景高精地图的支持。而国外的谷歌、苹果、Facebook 等公司在高精地图的构建方面已经积累多年,已拥有高覆盖率的厘米级高精地图。相比之下,国内企业在高精地图方面的建设尚处于起步阶段,目前主要有华为河图(Cyberverse)、商汤科技 SenseMARS、百度的 Apollo/VPAS 等。总的来说,在高精地图方面,国内尚在起步发展阶段。

3.5 三维几何建模

参数方法建立三维模型几何或结构变化的概率模型,通过统计学习得到概率模型的参数,用于描述和约束三维模型的构建。非参数方法主要通过对三维模型集进行联合分析,构建模型之间的结构和语义关联,支持三维模型的几何形变和结构重组,实现智能化的模型构建与编辑。基于深度学习的三维几何生成方面在近年也取得了一系列进展。此外,目前国际上公开的三维数据集已有不少,但大多都是国外团队创建的。国内在数据集方面的贡献还有待加强。

3.6 人体动态重建

人体动态重建目前国内外差距较大,主要由于投入的科研人员的力量差异造成。在人体模型表征方面,德国马克斯—普朗克研究所提出的 SMPL 系列人体模板对计算机视觉的发展起到极大的推动作用。在基于深度表征的人体方面,德国马克斯—普朗克研究所、美国 USC、Facebook 等团队对基于深度学习的人体重建研究引领世界发展。在基于深度相机的人体重建方面,Google 团队的 Fusion4D 和国内清华大学的 DoubleFusion 处于领先地位。在单图像人体重建方面,目前已成为国内外的研究热点,各大高校和企业都有力量在这个方向上进行投入。总体来说,目前人体重建方面的基础和关键理论创新突破主要还是由国外科研机构主导。但长远来说,随着人体重建逐渐走向产业化应用,国内的资源投入必将加大,目前已初步看到奋起直追的态势。

3.7 点云语义理解

近年来,点云的语义分割及实例分割领域取得了明显的进展。Guo 等人(2020)对深度学习方法在点云特征学习、三维形状分类、三维目标检测与跟踪及点云分割等领域的工作做了系统综述和对比分析。国际上,美国斯坦福大学、麻省理工学院、加州伯克利大学、英国牛津大学、德国慕尼黑工业大学、波恩大学、亚琛工业大学、加拿大多伦多大学以及谷歌、Facebook、英伟达等团队在点云语义分割和实例分割等方面引领了该领域的研究。在国内,清华大学、香港中文大学、中国科学院自动化研究所、中山大学、国防科技大学和武汉大学等团队的研究工作紧随国际前沿。总体而言,国外在开创性方法(如 PointNet 等)、社区贡献(如 ScanNet 数据集等)以及重要理论创新等方面的领先优势依然明显。国内虽然在领域内的顶级会议或期刊有系列工作发表,但其原始创新和影响力与国外存在一定差距。

4 发展趋势与展望

4.1 立体匹配

1)高分辨率视差估计。目前大多数的立体视觉技术只能输出较低分辨率的视差图或深度图。如何设计有效的网络结构来估计高分辨的视差图,目前还是一个开放问题。同时,目前主流的双目数据集大多是低分辨率图像,高分辨率双目视觉数据集还十分匮乏。

2)弱光照下的视差估计。弱光照条件下拍摄的图像往往有较多的噪点,缺少足够的纹理特征,给立体匹配任务带来很大的困难。如何从硬件角度对夜间图片的成像过程进行改善,或使用后处理算法对夜间图像进行增强,进而实现弱光照下的视差估计还需要大量的探索。

3)模型泛化性能。深度卷积网络的性能受到训练数据的严重影响。受制于当前较少的立体视觉数据集,很多模型都存在着过拟合在特定数据集上和特定域上的问题。

4)物体边界深度的估计。在双目匹配中,一般物体的边界由于视差的原因会出现遮挡,即同一物

体的像素点不会同时出现在左右视角中，同时，由于卷积神经网络的过度平滑性，在物体的边界很难形成深度的跳变，就会带来在物体边界深度估计比较困难的情况。

5）实时视差估计。目前的基于学习的立体匹配网络通常利用三维或四维的代价体，然后使用二维卷积或者三维卷积对其进行聚合和正则化，计算量较大。

4.2 单目深度估计

1）单目深度估计模型在精度和泛化性方面仍有巨大的提升空间，网络结构优化、训练方法改进和数据集质量提升仍是主要的研究方向。

2）单目深度估计方法和相关视觉应用的结合也是未来的研究趋势。

4.3 大场景下的视觉定位

1）基于深度学习的视觉定位方法的泛化性。探索结合几何知识的深度学习视觉定位方法来提升泛化性能，是一个值得研究的方向。

2）大场景长时定位中变化较多情况下的可靠定位。长时定位会要求不同季节、天气、时间段和光照下都能够鲁棒和精准的定位，而目前还远远没有被解决。

3）精度与速度兼顾的轻量化大场景视觉定位。目前大场景下的视觉定位，在学术研究上都在追求较高的精度，采用深度学习在较高性能的 GPU 下进行实验，这样忽略了速度和功耗的问题。而视觉定位的终端目前对于多数的计算资源还是有限，不能和极高性能的 GPU 相适配。研究轻量化的神经网络以及大数据的快速索引方法不失为一种有效的解决方式。

4）室内弱纹理、相似纹理场景下的视觉定位。由于室内存在大量的弱纹理和相似纹理的区域，而室内又缺失卫星通讯，目前这样的场景下的定位精度还较低。结合多种传感信息，以及线、面的矢量信息，有望能够逐渐解决该问题。

4.4 同步定位与地图构建

单纯的 SLAM 技术仍面临应用场景规模偏小、在复杂环境下稳定性不足和功耗较大等问题，难以满足大尺度复杂多变场景下的长时间应用需求。而与高精度地图结合并充分利用云计算的优势，不但可以突破场景规模的限制，而且稳定性也会有提升。特别是随着终端传感器、计算芯片、5G 网络和云计算的快速发展，城市级甚至地球级现实世界的数字化和移动定位将有望实现。尤其我国的 5G 技术和云计算技术已走在世界前列，同时大力支持芯片技术的研发。如何抓住这一历史契机，进一步突破场景尺度、定位精度和设备功耗等瓶颈，是目前 SLAM 技术的重要发展趋势。

4.5 三维几何建模

1）短期内自动建模无法完全取代人工建模，为交互式建模提供智能化建模辅助应是当前数据驱动方法的主要努力方向。

2）发挥数据驱动方法的优势，研究智能化的三维获取与重建，需重点关注数据驱动的主动式三维获取，针对形状复杂、成像困难物体（如透明、反光物体）的三维重建，以及数据驱动的语义理解。

3）大规模三维数据集的构建是数据驱动三维建模发展的关键。目前国际上公开的三维数据集已有不少，单个物体和室内外场景都有覆盖，但大多都是国外团队创建的。国内在数据集方面的贡献还有待加强。

4）结构化三维表征学习是当前三维深度学习的热点。现有方法一般需要较强的监督信息，例如对训练数据进行的预分割和部件标注。如何设计无监督或自监督的深度网络，以无结构三维表示为输入，生成结构化的三维表示，是值得关注的研究课题。

5）人的交互式三维建模过程可以看成一个复杂的编辑操作序列。一个有趣的方向是采用强化学习训练一个三维建模智能体，使其具有编辑操作的序列决策能力，实现自动三维建模。另一种思路是基于模仿学习从美工人员的建模过程中学习编辑操作的序列决策。

6）应更多关注三维场景合成，特别是面向真实应用（如室内设计、虚拟现实等）的合成。本文总结的数据驱动分类方法同样适用于三维场景。但三维场景合成有其特有的难点，例如场景中物体的摆放既有关联性又有随意性，如何合成具有真实感的场景布局是值得研究的课题。

7）基于大规模三维数据离线训练的轻量级三维几何深度生成模型，可用于实时、在线的室内外场景导航、建图和语义理解、生成、预测及臆想。是目前三维视觉的研究热点，对于机器人自主三维建图定位和面向语义任务的导航规划具有重要意义。

4.6 人体动态重建

1)大趋势上,该领域逐渐向着商用化、实用化逐步迈进。对重建的实时性,重建质量,以及对运动和渲染的真实感的要求越来越高;同时逐步由室内简单环境下的人体三维重建,向着野外复杂环境下的人体的三维重建过渡;所用设备逐步简单化,从多台昂贵的摄像机向单目摄像机,继而向着消费者级别的单目摄像机,甚至是手机移动端相机发展;同时重建目标从单目标向着多目标的方向发展。

2)重建质量上,对于衣物的表面几何重建及表面运动捕捉任重道远。一个非常具有挑战性的问题是:如何重建衣物表面的剧烈形变和衣物的几何拓扑变化。由于衣物的运动已经很难满足人体运动的先验,因此使用人体语义信息进行约束反而容易造成重建失败,并且,现有动态场景三维重建方法也很难在单视角有限输入信息的情况下处理如此大幅度的运动。

3)基于单目(RGB 或者 RGBD)的重建,由于缺少足够的观测信息,挑战巨大。单视角采集的人体运动往往存在着严重的自遮挡现象,在这种情况下,如何推测或估计出合理的被遮挡运动仍是一个非常具有挑战性的问题。

4)对模型的压缩以及对存储数据的压缩也是未来一大发展趋势。目前对野外场景下的三维重建以及基于单目 RGB 的重建,大多采用基于深度学习的方法,但由于数据集难制作等限制,很难训练出鲁棒性和泛化性好的模型,同时由于深度学习方法的模型庞大,很难部署,也往往很难实时或者需要大量的计算资源才能支持实时。

5)对多目标人体的动态重建前景很大。现实生活中的视频数据,大多存在多人以及多人交互的场景。如何更简单有效地解决多人的动态重建问题,是未来值得关注和研究的一大发展趋势。

4.7 点云语义理解

1)大场景下的点云分割。现有语义分割及实例分割算法均在少量点的点云上进行,然而实际激光雷达扫描获得的点云规模在百万级左右。研究一个有效的大规模点云语义分割方法是十分具有实际意义的。

2)弱监督下的点云分割。现有语义分割及实例分割算法均需要大量密集的语义标注信息,且泛化性能不强,因此在一定程度上限制了其应用场景。如何设计弱监督及自监督学习框架以利用有限的标注数据并提升泛化性能,是未来的重要研究方向。

3)时序点云分割。现有语义分割及实例分割算法均关注单个时刻的单个场景,然而实际采集的点云是具有时序信息的。因此,如何设计一个能够处理时序点云的分割方法是一个重要的研究方向。

参考文献(References)

Agamennoni G, Fontana S, Siegwart R Y and Sorrenti D G. 2016. Point clouds registration with probabilistic data association//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South): IEEE: 4092-4098 [DOI: 10.1109/IROS.2016.7759602]

Aleotti F, Tosi F, Zhang L, Poggi M and Mattoccia S. 2020. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation [EB/OL]. [2020-10-10]. https://arxiv.org/pdf/2008.07130.pdf

Almalioglu Y, Saputra M R U, de Gusmão P P B, Markham A and Trigoni N. 2019. GANVO: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 5474-5480 [DOI: 10.1109/ICRA.2019.8793512]

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

Averkiou M, Kim V G, Zheng Y Y and Mitra N J. 2014. ShapeSynth: parameterizing model collections for coupled shape exploration and synthesis. Computer Graphics Forum, 33(2): 125-134 [DOI: 10.1111/cgf.12310]

Badki A, Troccoli A, Kim K, Kautz J, Sen P and Gallo O. 2020. Bi3D: stereo depth estimation via binary classifications//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1597-1605 [DOI: 10.1109/CVPR42600.2020.00167]

Balntas V, Lenc K, Vedaldi A and Mikolajczyk K. 2017. HPatches: a benchmark and evaluation of handcrafted and learned local descriptors//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3852-3861 [DOI: 10.1109/CVPR.2017.410]

Bao W, Wang W, Xu Y H, Guo Y L, Hong S Y and Zhang X H. 2020. InStereo2K: a large real dataset for stereo matching in indoor scenes. Science China Information Sciences, 63(11): #212101 [DOI: 10.1007/s11432-019-2803-x]

Barron J T and Poole B. 2016. The fast bilateral solver//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 617-632 [DOI: 10.1007/978-3-319-46487-9_38]

Bhowmik A, Gumhold S, Rother C and Brachmann E. 2020. Reinforced feature points: optimizing feature detection and description for a high-level task//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4947-4956 [DOI: 10.1109/CVPR42600.2020.00500]

Bloesch M, Czarnowski J, Clark R, Leutenegger S and Davison A J. 2018. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2560-2568 [DOI: 10.1109/CVPR.2018.00271]

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

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

Brachmann E and Rother C. 2019. Expert sample consensus applied to camera re-localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7524-7533 [DOI: 10.1109/ICCV.2019.00762]

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

Brandao P, Mazomenos E and Stoyanov D. 2019. Widening siamese architectures for stereo matching. Pattern Recognition Letters, 120: 75-81 [DOI: 10.1016/j.patrec.2018.12.002]

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I and Leonard J J. 2016. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Transactions on Robotics, 32(6): 1309-1332 [DOI: 10.1109/TRO.2016.2624754]

Campos C, Elvira R, Rodríguez J J G, Montiel J M M and Tardós J D. 2020. ORB-SLAM3: an accurate open-source library for visual, visual-inertial and multi-map SLAM [EB/OL]. [2020-07-23]. https://arxiv.org/pdf/2007.11898.pdf

Caselitz T, Steder B, Ruhnke M and Burgard W. 2016. Monocular camera localization in 3D LiDAR maps//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South): IEEE: 1926-1931 [DOI: 10.1109/IROS.2016.7759304]

Chakrabarti A, Shao J and Shakhnarovich G. 2016. Depth from a single image by harmonizing overcomplete local network predictions//Advances in Neural Information Processing Systems. Barcelona, Spain: Curran Associates, Inc.: 2658-2666

Chan S H, Wu P T and Fu L C. 2018. Robust 2D indoor localization through laser SLAM and visual SLAM fusion//Proceedings of 2018 IEEE International Conference on Systems, Man, and Cybernetics. Miyazaki, Japan: IEEE, 2018: 1263-1268 [DOI: 10.1109/SMC.2018.00221]

Chang J R and Chen Y S. 2018. Pyramid stereo matching network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5410-5418 [DOI: 10.1109/CVPR.2018.00567]

Chen C H, Rosa S, Miao Y S, Lu C X, Wu W, Markham A and Trigoni N. 2019. Selective sensor fusion for neural visual-inertial odometry//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10534-10543 [DOI: 10.1109/CVPR.2019.01079]

Chen H X, Li K H, Fu Z H, Liu M Y, Chen Z H and Guo Y L. 2021. Distortion-aware monocular depth estimation for omnidirectional images. IEEE Signal Processing Letters, 28: 334-338 [DOI: 10.1109/LSP.2021.3050712]

Chen M X, Yang S W, Yi X D and Wu D. 2017. Real-time 3D mapping using a 2D laser scanner and IMU-aided visual SLAM//Proceedings of 2017 IEEE International Conference on Real-time Computing and Robotics. Okinawa, Japan: IEEE: 297-302 [DOI: 10.1109/RCAR.2017.8311877]

Chen Z, Badrinarayanan V, Drozdov G and Rabinovich A. 2018. Estimating depth from rgb and sparse sensing//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 167-182 [DOI: 10.1007/978-3-030-01225-0_11]

Chen Z Y, Sun X, Wang L, Yu Y and Huang C. 2015. A deep visual correspondence embedding model for stereo matching costs//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 972-980 [DOI: 10.1109/ICCV.2015.117]

Cheng X J, Wang P, Guan C Y and Yang R G. 2020a. CSPN++: learning context and resource aware convolutional spatial propagation networks for depth completion. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 10615-10622 [DOI: 10.1609/aaai.v34i07.6635]

Cheng X J, Wang P and Yang R G. 2018. Depth estimation via affinity learned with convolutional spatial propagation network//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 108-125 [DOI: 10.1007/978-3-030-01270-0_7]

Cheng X J, Wang P and Yang R G. 2020b. Learning depth with convolutional spatial propagation network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2361-2379 [DOI:

10.1109/TPAMI.2019.2947374]

Chodosh N, Wang C Y and Lucey S. 2019. Deep convolutional compressed sensing for LiDAR depth completion//Proceedings of Asian Conference on Computer Vision. Cham: Springer: 499-513 [DOI: 10.1007/978-3-030-20887-5_31]

Choy C, Gwak J and Savarese S. 2019. 4D spatio-temporal convnets: minkowski convolutional neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3070-3079 [DOI: 10.1109/CVPR.2019.00319]

Clark R, Wang S, Wen H K, Markham A and Trigoni N. 2017. VINet: visual-inertial odometry as a sequence-to-sequence learning problem [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1701.08376.pdf

Dai A and Nießner M. 2018. 3DMV: joint 3D-multi-view prediction for 3D semantic scene segmentation//Proceedings of European Conference on Computer Vision. Munich, Germany: Springer: 458-474 [DOI: 10.1007/978-3-030-01249-6_28]

Davison A J. 2003. Real-time simultaneous localisation and mapping with a single camera//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 1403-1410 [DOI: 10.1109/ICCV.2003.1238654]

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

Doria D and Radke R J. 2012. Filling large holes in LiDAR data by inpainting depth gradients//Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, USA: IEEE: 65-72 [DOI: 10.1109/CVPRW.2012.6238916]

Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, van der Smagt P, Cremers D and Brox T. 2015. FlowNet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]

Dou M S, Khamis S, Degtyarev Y, Davidson P, Fanello S R, Kowdle A, Escolano S O, Rhemann C, Kim D, Taylor J, Kohli P, Tankovich V and Izadi S. 2016. Fusion4D: real-time performance capture of challenging scenes. ACM Transactions on Graphics, 35(4): #114 [DOI: 10.1145/2897824.2925969]

Du H, Wang W, Xu C W, Xiao R and Sun C Y. 2020. Real-time onboard 3D state estimation of an unmanned aerial vehicle in multi-environments using multi-sensor data fusion. Sensors, 20(3): #919 [DOI: 10.3390/s20030919]

Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A and Sattler T. 2019. D2-Net: a trainable CNN for joint description and detection of local features//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8084-8093 [DOI: 10.1109/CVPR.2019.00828]

Eigen D and Fergus R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2650-2658 [DOI: 10.1109/ICCV.2015.304]

Eigen D, Puhrsch C and Fergus R. 2014. Depth map prediction from a single image using a multi-scale deep network//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press: 2366-2374

Engel J, Schöps T and Cremers D. 2014. LSD-SLAM: large-scale direct monocular SLAM//Proceedings of European Conference on Computer Vision. Zurich, Switzerland: Springer: 834-849 [DOI: 10.1007/978-3-319-10605-2_54]

Engel J, Stückler J and Cremers D. 2015. Large-scale direct SLAM with stereo cameras//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany: IEEE: 1935-1942 [DOI: 10.1109/IROS.2015.7353631]

Engel J, Koltun V and Cremers D. 2017. Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(3): 611-625

Engelmann F, Bokeloh M, Fathi A, Leibe B and Nießner M. 2020. 3D-MPA: multi-proposal aggregation for 3D semantic instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9028-9037 [DOI: 10.1109/CVPR42600.2020.00905]

Engelmann F, Kontogianni T, Hermans A and Leibe B. 2017. Exploring spatial context for 3D semantic segmentation of point clouds//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 716-724 [DOI: 10.1109/ICCVW.2017.90]

Fan H Q, Su H and Guibas L. 2017. A point set generation network for 3D object reconstruction from a single image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2463-2471 [DOI: 10.1109/CVPR.2017.264]

Feng Y J, Fan L X and Wu Y H. 2016. Fast localization in large-scale environments using supervised indexing of binary features. IEEE Transactions on Image Processing, 25(1): 343-358 [DOI: 10.1109/TIP.2015.2500030]

Ferstl D, Reinbacher C, Ranftl R, Rather M and Bischof H. 2013. Image guided depth upsampling using anisotropic total generalized variation//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 993-1000 [DOI: 10.1109/ICCV.2013.127]

Flynn J, Neulander I, Philbin J and Snavely N. 2016. Deep stereo: learning to predict new views from the world's imagery//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recogni-

tion. Las Vegas, USA: IEEE: 5515-5524 [DOI: 10.1109/CVPR.2016.595]

Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry[C]//2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014: 15-22

Fu H, Gong M M, Wang C H, Batmanghelich K and Tao D C. 2018. Deep ordinal regression network for monocular depth estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2002-2011 [DOI: 10.1109/CVPR.2018.00214]

Gálvez-López D and Tardos J D. 2012. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics, 28(5): 1188-1197 [DOI: 10.1109/TRO.2012.2197158]

Gan Y K, Xu X Y, Sun W X and Lin L. 2018. Monocular depth estimation with affinity, vertical pooling, and label enhancement//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 232-247 [DOI: 10.1007/978-3-030-01219-9_14]

Gao X, Wang R, Demmel N and Cremers D. 2018. LDSO: direct sparse odometry with loop closure//Proceeding of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 2198-2204 [DOI: 10.1109/IROS.2018.8593376]

Garg R, Bg V K, Carneiro G and Reid I. 2016. Unsupervised cnn for single view depth estimation: geometry to the rescue//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 740-756 [DOI: 10.1007/978-3-319-46484-8_45]

Gawel A, Cieslewski T, Dubé R, Bosse M, Siegwart R and Nieto J. 2016. Structure-based vision-laser matching//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, Korea (South): IEEE: 182-188 [DOI: 10.1109/IROS.2016.7759053]

Ge Y X, Wang H B, Zhu F, Zhao R and Li H S. 2020. Self-supervising fine-grained region similarities for large-scale image localization [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2006.03926.pdf

Genova K, Cole F, Sud A, Sarna A and Funkhouser T. 2019. Deep structured implicit functions [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1912.06126.pdf

Gidaris S and Komodakis N. 2017. Detect, replace, refine: deep structured prediction for pixel wise labeling//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 7187-7196 [DOI: 10.1109/CVPR.2017.760]

Girdhar R, Fouhey D F, Rodriguez M and Gupta A. 2016. Learning a predictable and generative vector representation for objects//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 484-499 [DOI: 10.1007/978-3-319-46466-4_29]

Godard C, Mac Aodha O and Brostow G J. 2017. Unsupervised monocular depth estimation with left-right consistency//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6602-6611 [DOI: 10.1109/CVPR.2017.699]

Graham B, Engelcke M and van der Maaten L. 2018. 3D semantic segmentation with submanifold sparse convolutional networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9224-9232 [DOI: 10.1109/CVPR.2018.00961]

Groueix T, Fisher M, Kim V G, Russell B C and Aubry M. 2018. A papier-mache approach to learning 3D surface generation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 216-224 [DOI: 10.1109/CVPR.2018.00030]

Gu X D, Fan Z W, Zhu S Y, Dai Z Z, Tan F T and Tan P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2492-2501 [DOI: 10.1109/CVPR42600.2020.00257]

Güney F and Geiger A. 2015. Displets: resolving stereo ambiguities using object knowledge//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4165-4175 [DOI: 10.1109/CVPR.2015.7299044]

Guo K W, Xu F, Wang Y G, Liu Y B and Dai Q H. 2015. Robust non-rigid motion tracking and surface reconstruction using L0 regularization//Proceeding of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3083-3091 [DOI: 10.1109/ICCV.2015.353]

Guo K W, Xu F, Yu T, Liu X Y, Dai Q H and Liu Y B. 2017. Real-time geometry, albedo, and motion reconstruction using a single RGB-D camera. ACM Transactions on Graphics, 36(3): #32 [DOI: 10.1145/3083722]

Guo X Y, Yang K, Yang W K, Wang X G and Li H S. 2019. Group-wise correlation stereo network//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3268-3277 [DOI: 10.1109/CVPR.2019.00339]

Guo Y L, Wang H Y, Hu Q Y, Liu H, Liu L and Bennamoun M. 2020. Deep learning for 3D point clouds: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence: #3005434 [DOI: 10.1109/TPAMI.2020.3005434]

Hambarde P and Murala S. 2020. S2DNet: depth estimation from single image and sparse samples. IEEE Transactions on Computational Imaging, 6: 806-817 [DOI: 10.1109/TCI.2020.2981761]

Han L, Zheng T, Xu L and Fang L. 2020. OccuSeg: occupancy-aware 3D instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2937-2946 [DOI: 10.1109/CVPR42600.2020.00301]

Han X G, Gao C and Yu Y Z. 2017. DeepSketch2Face: a deep learning

based sketching system for 3D face and caricature modeling. ACM Transactions on Graphics, 36(4): #126 [DOI: 10.1145/3072959.3073629]

Han X F, Leung T, Jia Y Q, Sukthankar R and Berg A C. 2015. MatchNet: unifying feature and metric learning for patch-based matching//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3279-3286 [DOI: 10.1109/CVPR.2015.7298948]

He K, Lu Y and Sclaroff S. 2018. Local descriptors optimized for average precision//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 596-605 [DOI: 10.1109/CVPR.2018.00069]

Helmer S and Lowe D. 2010. Using stereo for object recognition//Proceedings of 2010 IEEE International Conference on Robotics and Automation. Anchorage, USA: IEEE: 3121-3127 [DOI: 10.1109/ROBOT.2010.5509826]

Hou J, Dai A and Nießner M. 2019. 3D-SIS: 3D semantic instance segmentation of RGB-D scans//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4416-4425 [DOI: 10.1109/CVPR.2019.00455]

Houseago C, Bloesch M and Leutenegger S. 2019. KO-fusion: dense visual SLAM with tightly-coupled kinematic and odometric tracking//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 4054-4060 [DOI: 10.1109/ICRA.2019.8793471]

Hu Q Y, Yang B, Xie L H, Rosa S, Guo Y L, Wang Z H, Trigoni N and Markham A. 2020. RandLA-Net: efficient semantic segmentation of large-scale point clouds//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Location: Seattle, USA: IEEE: 11105-11114 [DOI: 10.1109/CVPR42600.2020.01112]

Huang Z X, Fan J M, Cheng S G, Yi S, Wang X G and Li H S. 2019a. HMS-Net: hierarchical multi-scale sparsity-invariant network for sparse depth completion. IEEE Transactions on Image Processing, 29: 3429-3441 [DOI: 10.1109/TIP.2019.2960589]

Huang Z Y, Xu Y, Shi J P, Zhou X W, Bao H J and Zhang G F. 2019b. Prior guided dropout for robust visual localization in dynamic environments//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2791-2800 [DOI: 10.1109/ICCV.2019.00288]

Innmann M, Zollhöfer M, Nießner M, Theobalt C and Stamminger M. 2020. VolumeDeform: real-time volumetric non-rigid reconstruction//Proceedings of European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 362-379 [DOI: 10.1007/978-3-319-46484-8_22]

Jain A, Thormählen T, Ritschel T and Seidel H P. 2012. Exploring shape variations by 3D-model decomposition and part-based recombination. Computer Graphics Forum, 31: 631-640 [DOI: 10.1111/j.1467-8659.2012.03042.x]

Jaritz M, De Charette R, Wirbel E, Perrotton X and Nashashibi F. 2018. Sparse and dense data with CNNs: depth completion and semantic segmentation//Proceedings of 2018 International Conference on 3D Vision. Verona, Italy: IEEE: 52-60 [DOI: 10.1109/3DV.2018.00017]

Jaritz M, Gu J Y and Su H. 2019. Multi-view PointNet for 3D scene understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 3995-4003 [DOI: 10.1109/ICCVW.2019.00494]

Jiang L, Zhao H S, Liu S, Shen X Y, Fu C W and Jia J Y. 2019. Hierarchical point-edge interaction network for point cloud semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 10432-10440 [DOI: 10.1109/ICCV.2019.01053]

Jiang L, Zhao H S, Shi S S, Liu S, Fu C W and Jia J Y. 2020. PointGroup: dual-set point grouping for 3D instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE 4866-4875 [DOI: 10.1109/CVPR42600.2020.00492]

Jiang M Y, Wu Y R, Zhao T Q, Zhao Z L and Lu C W. 2018. PointSIFT: a SIFT-like network module for 3D point cloud semantic segmentation [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1807.00652.pdf

Jie Z Q, Wang P F, Ling Y G, Zhao B, Wei Y C, Feng J S and Liu W. 2018. Left-right comparative recurrent model for stereo matching//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3838-3846 [DOI: 10.1109/CVPR.2018.00404]

Kanazawa A, Zhang Y J, Felsen P and Malik J. 2019. Learning 3D human dynamics from video//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5607-5616 [DOI: 10.1109/CVPR.2019.00576]

Kanhere O and Rappaport T S. 2019. Position locationing for millimeter wave systems//Proceedings of 2018 IEEE Global Communications Conference. Abu Dhabi, United Arab Emirates: IEEE: 206-212 [DOI: 10.1109/GLOCOM.2018.8647983]

Kar A, Häne C and Malik J. 2017. Learning a multi-view stereo machine//Advances in Neural Information Processing Systems. Long Beach, USA: [s.n.]: 364-375

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

Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A and Bry A. 2017. End-to-end learning of geometry and context for deep stereo regression//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 66-75 [DOI: 10.1109/ICCV.2017.17]

Khattak S, Papachristos C and Alexis K. 2019. Keyframe-based direct

thermal-inertial odometry//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 3563-3569 [DOI: 10.1109/ICRA.2019.8793927]

Kiechle M, Hawe S and Kleinsteuber M. 2013. A joint intensity and depth co-sparse analysis model for depth map super-resolution//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 1545-1552 [DOI: 10.1109/ICCV.2013.195]

Kim K R and Kim C S. 2016. Adaptive smoothness constraints for efficient stereo matching using texture and edge information//Proceedings of 2016 IEEE International Conference on Image Processing. Phoenix, USA: IEEE: 3429-3433 [DOI: 10.1109/ICIP.2016.7532996]

Kim Y, Jeong J and Kim A. 2018. Stereo camera localization in 3D LiDAR maps//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 1-9 [DOI: 10.1109/IROS.2018.8594362]

Klein G and Murray D. 2007. Parallel tracking and mapping for small AR workspaces//Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. Nara, Japan: IEEE: 225-234 [DOI: 10.1109/ISMAR.2007.4538852]

Knöbelreiter P, Reinbacher C, Shekhovtsov A and Pock T. 2017. End-to-end training of hybrid CNN-CRF models for stereo//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1456-1465 [DOI: 10.1109/CVPR.2017.159]

Kusupati U, Cheng S, Chen R and Su H. 2020. Normal assisted stereo depth estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2186-2196 [DOI: 10.1109/cvpr42600.2020.00226]

Lahoud J, Ghanem B, Oswald M R and Pollefeys M. 2019. 3D instance segmentation via multi-task metric learning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9255-9265 [DOI: 10.1109/ICCV.2019.00935]

Laina I, Rupprecht C, Belagiannis V, Tombari F and Navab N. 2016. Deeper depth prediction with fully convolutional residual networks//Proceedings of the 4th International Conference on 3D Vision. Stanford, USA: IEEE: 239-248 [DOI: 10.1109/3DV.2016.32]

Landrieu L and Boussaha M. 2019. Point cloud oversegmentation with graph-structured deep metric learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7432-7441 [DOI: 10.1109/CVPR.2019.00762]

Landrieu L and Simonovsky M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4558-4567 [DOI: 10.1109/CVPR.2018.00479]

Lawin F J, Danelljan M, Tosteberg P, Bhat G, Khan F S and Felsberg M. 2017. Deep projective 3D semantic segmentation//Proceedings of International Conference on Computer Analysis of Images and Patterns. Ystad, Sweden: Springer: 95-107 [DOI: 10.1007/978-3-319-64689-3_8]

Lee S H and Civera J. 2019. Loosely-coupled semi-direct monocular SLAM. IEEE Robotics and Automation Letters, 4(2): 399-406 [DOI: 10.1109/LRA.2018.2889156]

Lee W, Eckenhoff K, Geneva P and Huang G Q. 2020. Intermittent GPS-aided VIO: online initialization and calibration//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 5724-5731 [DOI: 10.1109/ICRA40945.2020.9197029]

Leutenegger S, Lynen S, Bosse M. 2015. Keyframe-based visual-inertial odometry using nonlinear optimization. The International Journal of Robotics Research, 34(3): 314-334

Li B, Shen C H, Dai Y C, Van Den Hengel A and He M Y. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1119-1127 [DOI: 10.1109/CVPR.2015.7298715]

Li B Y, Zou D P, Sartori D, Pei L and Yu W X. 2020a. TextSLAM: visual SLAM with planar text features//Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 2102-2108 [DOI: 10.1109/ICRA40945.2020.9197233]

Li C J, Pan H, Liu Y, Tong X, Sheffer A and Wang W P. 2018a. Robust flow-guided neural prediction for sketch-based freeform surface modeling. ACM Transactions on Graphics, 37(6): #238 [DOI: 10.1145/3272127.3275051]

Li J Y, Bao H J and Zhang G. 2019a. Rapid and robust monocular visual-inertial initialization with gravity estimation via vertical edges//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China: IEEE: 6230-6236 [DOI: 10.1109/IROS40897.2019.8968456]

Li J, Niu C J and Xu K. 2020c. Learning part generation and assembly for structure-aware shape synthesis. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 11362-11369 [DOI: 10.1609/aaai.v34i07.6798]

Li J Q, Pei L, Zou D P, Xia S P C, Wu Q, Li T, Sun Z and Yu W X. 2020b. Attention-SLAM: a visual monocular SLAM learning from human gaze [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2009.06886.pdf

Li J, Xu K, Chaudhuri S, Yumer E, Zhang H and Guibas L. 2017. GRASS: generative recursive autoencoders for shape structures. ACM Transactions on Graphics, 36(4): #52 [DOI: 10.1145/3072959.3073637]

Li J Y, Yang B B, Huang K, Zhang G F and Bao H J. 2019b. Robust

and efficient visual-inertial odometry with multi-plane priors//Proceedings of Chinese Conference on Pattern Recognition and Computer Vision. Xi′an, China: Springer: 283-295 [DOI: 10.1007/978-3-030-31726-3_24]

Li M Y, Patil A G, Xu K, Chaudhuri S, Khan O, Shamir A, Tu C H, Chen B Q, Cohen-Or D and Zhang H. 2019c. GRAINS: generative recursive autoencoders for indoor scenes. ACM Transactions on Graphics, 38(2): #12 [DOI: 10.1145/3303766]

Li R H, Wang S, Long Z Q and Gu D B. 2018b. UnDeepVO: monocular visual odometry through unsupervised deep learning//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 7286-7291 [DOI: 10.1109/ICRA.2018.8461251]

Li S K, Xue F, Wang X, Yan Z K and Zha H B. 2019d. Sequential adversarial learning for self-supervised deep visual odometry//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2851-2860 [DOI: 10.1109/ICCV.2019.00294]

Li Y Y, Bu R, Sun M C, Wu W, Di X H and Chen B Q. 2018c. PointCNN: convolution on *X*-transformed points//Advances in Neural Information Processing Systems. Montréal, Canada: [s. n.]: 820-830

Li Y, Ushiku Y and Harada T. 2019e. Pose graph optimization for unsupervised monocular visual odometry//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 5439-5445 [DOI: 10.1109/ICRA.2019.8793706]

Liang M, Guo X, Li H, Wang X and Song Y. 2019. Unsupervised cross-spectral stereo matching by learning to synthesize//Proceedings of the AAAI Conference on Artificial Intelligence, 33: 8706-8713 [DOI: 10.1609/aaai.v33i01.33018706]

Liang Z F, Feng Y L, Guo Y L, Liu H Z, Chen W, Qiao L B, Zhou L and Zhang J F. 2018. Learning for disparity estimation through feature constancy//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2811-2820 [DOI: 10.1109/CVPR.2018.00297]

Liang Z F, Guo Y L, Feng Y L, Chen W, Qiao L B, Zhou L, Zhang J F and Liu H Z. 2021. Stereo matching using multi-level cost volume and multi-scale feature constancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 300-315 [DOI: 10.1109/TPAMI.2019.2928550]

Liao M, Lu F X, Zhou D F, Zhang S B, Li W and Yang R G. 2020. DVI: depth guided video inpainting for autonomous driving[EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2007.08854.pdf

Liao Y Y, Huang L C, Wang Y, Kodagoda S, Yu Y and Liu Y. 2017. Parse geometry from a line: monocular depth estimation with partial laser observation//Proceedings of 2017 IEEE International Conference on Robotics and Automation. Singapore, Singapore: IEEE: 5059-5066 [DOI: 10.1109/ICRA.2017.7989590]

Liao Z W. 2016. Research on Autonomous Mapping and Navigation Technology in Indoor Environment based on Lidar and MEMS Inertial Components. Nanjing: Nanjing University of Aeronautics and Astronautics (廖自威. 2016. 激光雷达/微惯性室内自主建图与导航技术研究. 南京: 南京航空航天大学)

Liebel L and Körner M. 2019. MultiDepth: single-image depth estimation via multi-task regression and classification//Proceedings of 2019 IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand: IEEE: 1440-1447 [DOI: 10.1109/ITSC.2019.8917177]

Liu F Y, Li S P, Zhang L Q, Zhou C H, Ye R T, Wang Y B and Lu J W. 2017a. 3DCNN-DQN-RNN: a deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5679-5688 [DOI: 10.1109/ICCV.2017.605]

Liu H M, Chen M Y, Zhang G F and Bao H J. 2018. Ice-ba: Incremental, consistent and efficient bundle adjustment for visual-inertial slam//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1974-1982

Liu F Y, Shen C H, Lin G S and Reid I. 2016. Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2024-2039 [DOI: 10.1109/TPAMI.2015.2505283]

Liu H, Guo Y L, Ma Y N, Lei Y J and Wen G J. 2020a. Semantic context encoding for accurate 3D point cloud segmentation. IEEE Transactions on Multimedia [DOI: 10.1109/TMM.2020.3007331]

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

Liu L, Li H D and Dai Y C. 2019a. Stochastic attraction-repulsion embedding for large scale image localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2570-2579 [DOI: 10.1109/ICCV.2019.00266]

Liu P P, King I, Lyu M R and Xu J. 2020b. Flow2Stereo: effective self-supervised learning of optical flow and stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6648-6657 [DOI: 10.1109/CVPR42600.2020.00668]

Liu Y C, Fan B, Xiang S M and Pan C H. 2019b. Relation-shape convolutional neural network for point cloud analysis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8887-8896 [DOI: 10.1109/CVPR.2019.00910]

Liu Y, Shen Z H, Lin Z X, Peng S D, Bao H J and Zhou X W. 2019c. GIFT: learning transformation-invariant dense visual descriptors via group CNNs//Advances in Neural Information Processing Systems. Vancouver, Canada: [s. n.]: 6992-7003

Liu Y B, Ye G Z, Wang Y G, Dai Q H and Theobalt C. 2014. Human performance capture using multiple handheld kinects//Computer Vision and Machine Learning with RGB-D Sensors. Switzerland: Springer: 91-108 [DOI: 10.1007/978-3-319-08651-4_5]

Liu Z J, Tang H T, Lin Y J and Han S. 2019d. Point-voxel CNN for efficient 3D deep learning//Advances in Neural Information Processing Systems. Vancouver, Canada: [s. n.]

Lu C H, Uchiyama H, Thomas D, Shimada A and Taniguchi R I. 2018. Sparse cost volume for efficient stereo matching. Remote Sensing, 10(11): #1844 [DOI: 10.3390/rs10111844]

Lu G Y, Yan Y, Ren L, Song J K, Sebe N and Kambhamettu C. 2015. Localize me anywhere, anytime: a multi-task point-retrieval approach//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2434-2442 [DOI: 10.1109/ICCV.2015.280]

Lu K Y, Barnes N, Anwar S and Zheng L. 2020. From depth what can you see? Depth completion via auxiliary image reconstruction//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11303-11312 [DOI: 10.1109/CVPR42600.2020.01132]

Luo W J, Schwing A G and Urtasun R. 2016. Efficient deep learning for stereo matching//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE: 5695-5703 [DOI: 10.1109/CVPR.2016.614]

Luo Y, Ren J, Lin M D, Pang J H, Sun W X, Li H S and Lin L. 2018a. Single view stereo matching//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 155-163 [DOI: 10.1109/CVPR.2018.00024]

Luo Z X, Shen T W, Zhou L, Zhu S Y, Zhang R, Yao Y, Fang T and Quan L. 2018b. GeoDesc: learning local descriptors by integrating geometry constraints//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 170-185 [DOI: 10.1007/978-3-030-01240-3_11]

Luo Z X, Zhou L, Bai X Y, Chen H K, Zhang J H, Yao Y, Li S W, Fang T and Quan L. 2020. Aslfeat: learning local features of accurate shape and localization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6588-6597 [DOI: 10.1109/CVPR42600.2020.00662]

Lynen S, Achtelik M W, Weiss S, Chli M and Siegwart R. 2013. A robust and modular multi-sensor fusion approach applied to MAV navigation//Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE: 3923-3929 [DOI: 10.1109/IROS.2013.6696917]

Lynen S, Sattler T, Bosse M, Hesch J, Pollefeys M and Siegwart R. 2015. Get out of my lab: large-scale, real-Time visual-inertial localization//Proceedings of Robotics: Science and Systems. Rome, Italy: [s. n.] [DOI: 10.15607/RSS.2015.XI.037]

Ma F C and Karaman S. 2018. Sparse-to-dense: depth prediction from sparse depth samples and a single image//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 4796-4803 [DOI: 10.1109/icra.2018.8460184]

Ma F C, Cavalheiro G V and Karaman S. 2019. Self-supervised sparse-to-dense: self-supervised depth completion from LiDAR and monocular camera//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 3288-3295 [DOI: 10.1109/ICRA.2019.8793637]

Ma Y N, Guo Y L, Liu H, Lei Y J and Wen G J. 2020. Global context reasoning for semantic segmentation of 3D point clouds//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 2920-2929 [DOI: 10.1109/WACV45572.2020.9093411]

Mac Aodha O, Campbell N D F, Nair A and Brostow G J. 2012. Patch based synthesis for single depth image super-resolution//Proceedings of European Conference on Computer Vision. Florence, Italy: Springer: 71-84 [DOI: 10.1007/978-3-642-33712-3_6]

Maddern W, Stewart A D and Newman P. 2014. LAPS-II: 6-DOF day and night visual localisation with prior 3D structure for autonomous road vehicles//Proceedings of 2014 IEEE Intelligent Vehicles Symposium Proceedings. Dearborn, USA: IEEE: 330-337 [DOI: 10.1109/IVS.2014.6856471]

Mao J G, Wang X G and Li H S. 2019. Interpolated convolutional networks for 3D point cloud understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1578-1587 [DOI: 10.1109/ICCV.2019.00166]

Mascaro R, Teixeira L, Hinzmann T, Siegwart R and Chli M. 2018. GOMSF: graph-optimization based multi-sensor fusion for robust UAV pose estimation//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 1421-1428 [DOI: 10.1109/ICRA.2018.8460193]

Matsuo K and Aoki Y. 2015. Depth image enhancement using local tangent plane approximations//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3574-3583 [DOI: 10.1109/CVPR.2015.7298980]

Matusik W, Buehler C, Raskar R, Gortler S J and McMillan L. 2000. Image-based visual hulls//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York, United States: ACM: 369-374 [DOI: 10.1145/344779.344951]

Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A and Brox T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4040-4048 [DOI: 10.1109/CVPR.2016.438]

Menze M and Geiger A. 2015. Object scene flow for autonomous vehicles//Proceedings of 2015 IEEE Conference on Computer Vision and

Pattern Recognition. Boston, USA: IEEE: 3061-3070 [DOI: 10.1109/CVPR.2015.7298925]

Mishchuk A, Mishkin D, Radenović F and Matas J. 2017. Working hard to know your neighbor's margins: local descriptor learning loss//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, United States: Curran Associates Inc.: 4829-4840

Mitra N J, Wand M, Zhang H, Cohen-Or D, Kim V and Huang Q X. 2014. Structure-aware shape processing//ACM SIGGRAPH 2014 Courses. Vancouver, Canada: ACM: 1-21 [DOI: 10.1145/2614028.2615401]

Mo K C, Guerrero P, Li Y, Su H, Wonka P, Mitra N and Guibas L J. 2019a. StructEdit: learning structural shape variations [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1911.11098.pdf

Mo K C, Guerrero P, Li Y, Su H, Wonka P, Mitra N and Guibas L J. 2019b. StructurEnet: hierarchical graph networks for 3D shape generation [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1908.00575.pdf

Mourikis A I and Roumeliotis S I. 2007. A multi-state constraint kalman filter for vision-aided inertial navigation//Proceedings of 2007 IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE: 3565-3572 [DOI: 10.1109/ROBOT.2007.364024]

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

Mur-Artal R and Tardós J D. 2017. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics, 33(5): 1255-1262 [DOI: 10.1109/TRO.2017.2705103]

Neubert P, Schubert S and Protzel P. 2017. Sampling-based methods for visual navigation in 3D maps by synthesizing depth images//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver BC, Canada: IEEE: 2492-2498 [DOI: 10.1109/IROS.2017.8206067]

Newcombe R A, Fox D and Seitz S M. 2015. DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 343-352 [DOI: 10.1109/CVPR.2015.7298631]

Ng T, Balntas V, Tian Y and Mikolajczyk K. 2020. SOLAR: second-order loss and attention for image retrieval [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2001.08972.pdf

Niu C J, Li J and Xu K. 2018. Im2Struct: recovering 3D shape structure from a single RGB image//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4521-4529 [DOI: 10.1109/CVPR.2018.00475]

Pang J H, Sun W X, Ren J S J, Yang C X and Yan Q. 2017. Cascade residual learning: a two-stage convolutional neural network for stereo matching//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 878-886 [DOI: 10.1109/ICCVW.2017.108]

Park C, Kim S, Moghadam P, Guo J D, Sridharan S and Fookes C. 2019a. Robust photogeometric localization over time for map-centric loop closure. IEEE Robotics and Automation Letters, 4(2): 1768-1775 [DOI: 10.1109/LRA.2019.2895262]

Park J J, Florence P, Straub J, Newcombe R and Lovegrove S. 2019b. DeepSDF: learning continuous signed distance functions for shape representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 165-174 [DOI: 10.1109/CVPR.2019.00025]

Park J, Joo K, Hu Z, Liu C K and Kweon I S. 2020. Non-local spatial propagation network for depth completion//Proceedings of the European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 120-136 [DOI: 10.1007/978-3-030-58601-0_8]

Pascoe G, Maddern W, Stewart A D and Newman P. 2015. FARLAP: fast robust localisation using appearance priors//Proceedings of 2015 IEEE International Conference on Robotics and Automation. Seattle, USA: IEEE, 2015: 6366-6373 [DOI: 10.1109/ICRA.2015.7140093]

Patil V, Van Gansbeke W, Dai D X and Van Gool L. 2020. Don't forget the past: recurrent depth estimation from monocular video. IEEE Robotics and Automation Letters, 5(4): 6813-6820 [DOI: 10.1109/LRA.2020.3017478]

Pham Q H, Nguyen T, Hua B S, Roig G and Yeung S K. 2019. JSIS3D: joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8819-8828 [DOI: 10.1109/CVPR.2019.00903]

Poggi M and Mattoccia S. 2017. Learning to predict stereo reliability enforcing local consistency of confidence maps//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4541-4550 [DOI: 10.1109/CVPR.2017.483]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Advances in Neural Information Processing Systems. Long Beach, USA: [s. n.]

Qin T, Li P L and Shen S J. 2018. VINS-mono: a robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 34(4): 1004-1020 [DOI: 10.1109/TRO.2018.2853729]

Qin T, Pan J, Cao S Z and Shen S J. 2019. A general optimization-based framework for local odometry estimation with multiple sensors

[EB/OL]. [2021-01-21]. https://arxiv. org/pdf/1901. 03638. pdf

Radenović F, Tolias G and Chum O. 2019. Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7): 1655-1668 [DOI: 10. 1109/TPAMI. 2018. 2846566]

Rappaport T S, Xing Y C, Kanhere O, Ju S H, Madanayake A, Mandal S, Alkhateeb A and Trichopoulos G C. 2019. Wireless communications and applications above 100 GHz: opportunities and challenges for 6G and beyond. IEEE Access, 7: 78729-78757 [DOI: 10. 1109/ACCESS. 2019. 2921522]

Revaud J, Weinzaepfel P, De Souza C, Pion N, Csurka G, Cabon Y and Humenberger M. 2019. R2D2: repeatable and reliable detector and descriptor [EB/OL]. [2021-01-21]. https://arxov. org/pdf/1906. 06195. pdf

Rosu R A, Schutt P, Quenzel J and Behnke S. 2019. LatticeNet: fast point cloud segmentation using permutohedral lattices [EB/OL]. [2021-01-21]. https://arxir. org/pdf/1912. 05905. pdf

Roy A and Todorovic S. 2016. Monocular depth estimation using neural regression forest//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5506-5514 [DOI: 10. 1109/CVPR. 2016. 594]

Saito S, Huang Z, Natsume R, Morishima S, Li H and Kanazawa A. 2019. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2304-2314 [DOI: 10. 1109/ICCV. 2019. 00239]

Saito S, Simon T, Saragih J and Joo H. 2020. PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 81-90 [DOI: 10. 1109/CVPR42600. 2020. 00016]

Saputra M R U, de Gusmao P P B, Wang S, Markham A and Trigoni N. 2019. Learning monocular visual odometry through geometry-aware curriculum learning//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 3549-3555 [DOI: 10. 1109/ICRA. 2019. 8793581]

Sarlin P E, Cadena C, Siegwart R and Dymczyk M. 2019. From coarse to fine: robust hierarchical localization at large scale//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12708-12717 [DOI: 10. 1109/CVPR. 2019. 01300]

Sarlin P E, Debraine F, Dymczyk M, Siegwart R and Cadena C. 2018. Leveraging deep visual descriptors for hierarchical efficient localization [EB/OL]. [2021-01-21]. https://arxiv. org/pdf/1809. 01019. pdf

Sarlin P E, DeTone D, Malisiewicz T and Rabinovich A. 2020. Superglue: learning feature matching with graph neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4937-4946 [DOI: 10. 1109/CVPR42600. 2020. 00499]

Sattler T, Leibe B and Kobbelt L. 2017. Efficient and effective prioritized matching for large-scale image-based localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9): 1744-1756 [DOI: 10. 1109/TPAMI. 2016. 2611662]

Sattler T, Maddern W, Toft C, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J and Kahl F. 2018. Benchmarking 6DOF outdoor visual localization in changing conditions//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8601-8610 [DOI: 10. 1109/CVPR. 2018. 00897]

Sattler T, Zhou Q J, Pollefeys M and Leal-Taixé L. 2019. Understanding the limitations of CNN-based absolute camera pose regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3297-3307 [DOI: 10. 1109/CVPR. 2019. 00342]

Saxena A, Sun M and Ng A Y. 2009. Make3D: learning 3D scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5): 824-840 [DOI: 10. 1109/TPAMI. 2008. 132]

Saxena A, Sung C H and Ng A Y. 2005. Learning depth from single monocular images//Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press: 1161-1168

Schmid K, Tomic T, Ruess F, Hirschmüller H and Suppa M. 2013. Stereo vision based indoor/outdoor navigation for flying robots//Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE: 3955-3962 [DOI: 10. 1109/IROS. 2013. 6696922]

Seki A and Pollefeys M. 2017. SGM-Nets: semi-global matching with neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6640-6649 [DOI: 10. 1109/CVPR. 2017. 703]

Shamwell E J, Lindgren K, Leung S and Nothwang W D. 2020. Unsupervised deep visual-inertial odometry with online error correction for RGB-D imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2478-2493 [DOI: 10. 1109/TPAMI. 2019. 2909895]

Shao W Z, Vijayarangan S, Li C and Kantor G. 2019. Stereo visual inertial LiDAR simultaneous localization and mapping//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China: IEEE: 370-377 [DOI: 10. 1109/IROS40897. 2019. 8968012]

Shean D E, Alexandrov O, Moratto Z M, Smith B E, Joughin I R, Porter C and Morin P. 2016. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 116: 101-117 [DOI: 10. 1016/j.

isprsjprs. 2016. 03. 012]

Sheng L, Xu D, Ouyang W L and Wang X G. 2019. Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep SLAM//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4301-4310 [DOI: 10.1109/ICCV.2019.00440]

Shi T X, Cui H N, Song Z and Shen S H. 2020. Dense semantic 3D map based long-term visual localization with hybrid features [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2005.10766.pdf

Shi T X, Shen S H, Gao X and Zhu L J. 2019. Visual localization using sparse semantic 3D map//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 315-319 [DOI: 10.1109/ICIP.2019.8802957]

Shivakumar S S, Nguyen T, Miller I D, Chen S W, Kumar V and Taylor C J. 2019. DFuseNet: deep fusion of RGB and sparse depth information for image guided dense depth completion//Proceedings of 2019 IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand: IEEE: 13-20 [DOI: 10.1109/ITSC.2019.8917294]

Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P and Moreno-Noguer F. 2015. Discriminative learning of deep convolutional feature point descriptors//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 118-126 [DOI: 10.1109/ICCV.2015.22]

Sinha A, Bai J and Ramani K. 2016. Deep learning 3D shape surfaces using geometry images//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 223-240

Song X, Zhao X, Hu H W and Fang L J. 2018. EdgeStereo: a context integrated residual pyramid network for stereo matching//Proceedings of 2018 Asian Conference on Computer Vision. Perth, Australia: Springer: 20-35 [DOI: 10.1007/978-3-030-20873-8_2]

Stewart A D and Newman P. 2012. LAPS-localisation using appearance of prior structure: 6-DoF monocular camera localisation using prior pointclouds//Proceedings of 2012 IEEE International Conference on Robotics and Automation. Saint Paul, USA: IEEE: 2625-2632 [DOI: 10.1109/ICRA.2012.6224750]

Strasdat H, Davison A J, Montiel J M M and Konolige K. 2011. Double window optimisation for constant time visual SLAM//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 2352-2359 [DOI: 10.1109/ICCV.2011.6126517]

Strasdat H, Montiel J M M and Davison A J. 2010. Scale drift-aware large scale monocular SLAM//Robotics: Science and Systems VI. Zaragoza, Spain: [s. n.]

Su H, Jampani V, Sun D Q, Maji S, Kalogerakis E, Yang M H and Kautz J. 2018. SPLATNet: sparse lattice networks for point cloud processing//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2530-2539 [DOI: 10.1109/CVPR.2018.00268]

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Su Z, Xu L, Zheng Z R, Yu T, Liu Y B and Fang L. 2020. RobustFusion: human volumetric capture with data-driven visual cues using a RGBD camera//Proceedings of European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 246-264 [DOI: 10.1007/978-3-030-58548-8_15]

Sun X, Xie Y F, Luo P and Wang L. 2017. A dataset for benchmarking image-based localization//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5641-5649 [DOI: 10.1109/CVPR.2017.598]

Svärm L, Enqvist O, Kahl F and Oskarsson M. 2017. City-scale localization for cameras with known vertical direction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(7): 1455-1461 [DOI: 10.1109/TPAMI.2016.2598331]

Tan F T, Zhu H, Cui Z P, Zhu S Y, Pollefeys M and Tan P. 2020. Self-supervised human depth estimation from monocular videos//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 647-656 [DOI: 10.1109/CVPR42600.2020.00073]

Tang F L, Li H P and Wu Y H. 2019. FMD stereo SLAM: fusing MVG and direct formulation towards accurate and fast stereo SLAM//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 133-139 [DOI: 10.1109/ICRA.2019.8793664]

Taniai T, Matsushita Y, Sato Y and Naemura T. 2018. Continuous 3D label stereo matching using local expansion moves. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(11): 2725-2739 [DOI: 10.1109/TPAMI.2017.2766072]

Tatarchenko M, Park J, Koltun V and Zhou Q Y. 2018. Tangent convolutions for dense prediction in 3D//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3887-389 [DOI: 10.1109/CVPR.2018.00409]

Tateno K, Tombari F, Laina I and Navab N. 2017. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6565-6574 [DOI: 10.1109/CVPR.2017.695]

Tchapmi L, Choy C, Armeni I, Gwak J and Savarese S. 2017. SEGCloud: semantic segmentation of 3D point clouds//Proceedings of 2017 International Conference on 3D Vision. Qingdao, China: IEEE: 537-547 [DOI: 10.1109/3DV.2017.00067]

Thomas H, Qi C R, Deschaud J E, Marcotegui B, Goulette F and Guibas L. 2019. KPConv: flexible and deformable convolution for point clouds//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6410-6419 [DOI: 10.1109/ICCV.2019.00651]

Tian Y R, Fan B and Wu F C. 2017. L2-net: deep learning of discriminative patch descriptor in euclidean space//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6128-6136 [DOI: 10.1109/CVPR.2017.649]

Tian Y R, Yu X, Fan B, Wu F C, Heijnen H and Balntas V. 2019. SOSNet: second order similarity regularization for local descriptor learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11008-11017 [DOI: 10.1109/CVPR.2019.01127]

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

Uhrig J, Schneider N, Schneider L, Franke U, Brox T and Geiger A. 2017. Sparsity invariant CNNs//Proceedings of 2017 International Conference on 3D Vision. Qingdao, China: IEEE: 11-20 [DOI: 10.1109/3DV.2017.00012]

Ummenhofer B, Zhou H Z, Uhrig J, Mayer N, Ilg E, Dosovitskiy A and Brox T. 2017. DeMoN: depth and motion network for learning monocular stereo//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5622-5631 [DOI: 10.1109/CVPR.2017.596]

Vlasic D, Peers R, Baran I, Debevec P, Popović J, Rusinkiewicz S and Matusik W. 2009. Dynamic shape capture using multi-view photometric stereo. ACM Transactions on Graphics, 28(5): #174 [DOI: 10.1145/1618452.1618520]

Wang B, Chen C H, Lu C X, Zhao P J, Trigoni N and Markham A. 2020a. AtLoc: attention guided camera localization. Proceedings of the AAAI Conference on Artificial Intelligence, 34(6): 10393-10401 [DOI: 10.1609/aaai.v34i06.6608]

Wang H C, Liu Q, Yue X Y, Lasenby J and Kusner M J. 2020b. Pre-training by completing point clouds [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/2010.01089.pdf

Wang H, Schor N, Hu R Z, Huang H B, Cohen-Or D and Huang H. 2020c. Global-to-local generative model for 3D shapes. ACM Transactions on Graphics, 37(6): #214 [DOI: 10.1145/3272127.3275025]

Wang L G, Guo Y L, Wang Y Q, Liang Z F, Lin Z P, Yang J G and An W. 2020d. Parallax attention for unsupervised stereo correspondence learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020: # 3026899 [DOI: 10.1109/TPAMI.2020.3026899]

Wang L, Huang Y C, Hou Y L, Zhang S M and Shan J. 2019a. Graph attention convolution for point cloud semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10288-10297 [DOI: 10.1109/CVPR.2019.01054]

Wang P S, Liu Y, Guo Y X, Sun C Y and Tong X. 2017a. O-CNN: octree-based convolutional neural networks for 3D shape analysis. ACM Transactions on Graphics, 36(4): #72 [DOI: 10.1145/3072959.3073608]

Wang P S, Sun C Y, Liu Y and Tong X. 2018a. Adaptive O-CNN: a patch-based deep representation of 3D shapes. ACM Transactions on Graphics, 37(6): #217 [DOI: 10.1145/3272127.3275050]

Wang Q Y, Yan Z K, Wang J Q, Xue F, Ma W and Zha H B. 2020e. Line flow based SLAM. [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/2009.09972.pdf

Wang Q, Zhou X, Hariharan B, Snavely N. 2020. Learning feature descriptors using camera pose supervision//Proceedings of 2020 European Conference on Computer Vision. [s. l.]: Springer, Cham: 757-774

Wang Q Q, Zhou X W, Hariharan B and Snavely N. 2020f. Learning feature descriptors using camera pose supervision [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/2004.13324.pdf

Wang S, Clark R, Wen H K and Trigoni N. 2017b. DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks//Proceedings of 2017 IEEE International Conference on Robotics and Automation. Singapore, Singapore: IEEE: 2043-2050 [DOI: 10.1109/ICRA.2017.7989236]

Wang S L, Fidler S and Urtasun R. 2015. Lost shopping! Monocular localization in large indoor spaces//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2695-2703 [DOI: 10.1109/ICCV.2015.309]

Wang S L, Suo S M, Ma W C, Pokrovsky A and Urtasun R. 2018b. Deep parametric continuous convolutional neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2589-2597 [DOI: 10.1109/CVPR.2018.00274]

Wang W Y, Yu R, Huang Q G and Neumann U. 2018c. SGPN: similarity group proposal network for 3D point cloud instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2569-2578 [DOI: 10.1109/CVPR.2018.00272]

Wang X L, Liu S, Shen X Y, Shen C H and Jia J Y. 2019b. Associatively segmenting instances and semantics in point clouds//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4091-4100 [DOI: 10.1109/CVPR.2019.00422]

Wang Y R, Huang Z H, Zhu H, Li W, Cao X and Yang R G. 2020g. Interactive free-viewpoint video generation. Virtual Reality and Intelligent Hardware, 2(3): 247-260 [DOI: 10.1016/j.vrih.2020.04.004]

Wang Y, Wang P, Yang Z H, Luo C X, Yang Y and Xu W. 2019c. UnOS: unified unsupervised optical-flow and stereo-depth estimation by watching videos//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8063-8073 [DOI: 10.1109/CVPR.2019.00826]

Wang Y G, Liu Y B, Tong X, Dai Q H and Tan P. 2018d. Outdoor markerless motion capture with sparse handheld video cameras. IEEE Transactions on Visualization and Computer Graphics, 24(5): 1856-1866 [DOI: 10.1109/TVCG.2017.2693151]

Wang Z H, Liang D T, Liang D, Zhang J C and Liu H J. 2018. A SLAM method based on inertial/magnetic sensors and monocular vision fusion. Robot, 40(6): 933-941 (王泽华, 梁冬泰, 梁丹, 章家成, 刘华杰. 2018. 基于惯性/磁力传感器与单目视觉融合的SLAM方法. 机器人, 40(6): 933-941) [DOI: 10.13973/j.cnki.robot.170683]

Wei J C, Lin G S, Yap K H, Hung T Y and Xie L H. 2020. Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4383-4392 [DOI: 10.1109/CVPR42600.2020.00444]

Weiss S, Achtelik M W, Lynen S, Chli M and Siegwart R. 2012. Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments//Proceedings of 2012 IEEE International Conference on Robotics and Automation. Saint Paul, USA: IEEE: 957-964 [DOI: 10.1109/ICRA.2012.6225147]

Wolcott R W and Eustice R M. 2014. Visual localization within lidar maps for automated urban drivingin///Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. Chicago, USA: IEEE: 176-183.

Wong D, Kawanishi Y, Deguchi D, Ide I and Murase H. 2017. Monocular localization within sparse voxel maps//Proceedings of 2017 IEEE Intelligent Vehicles Symposium (IV). Los Angeles, USA: IEEE: 499-504 [DOI: 10.1109/IVS.2017.7995767]

Wu B C, Wan A, Yue X Y and Keutzer K. 2018a. SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia: IEEE: 1887-1893 [DOI: 10.1109/ICRA.2018.8462926]

Wu B C, Zhou X Y, Zhao S C, Yue X Y and Keutzer K. 2019a. SqueezeSegV2: improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud//Proceedings of 2019 International Conference on Robotics and Automation. Montreal, Canada: IEEE: 4376-4382 [DOI: 10.1109/ICRA.2019.8793495]

Wu J J, Zhang C K, Xue T F, Freeman B T and Tenenbaum J B. 2016. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling//Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, United States: Curran Associates Inc.: 82-90

Wu L and Wu Y H. 2019. Similarity hierarchy based place recognition by deep supervised hashing for SLAM. IR-OS

Wu R D, Zhuang Y X, Xu K, Zhang H and Chen B Q. 2019b. PQ-NET: a generative part Seq2Seq network for 3D shapes [EB/OL]. [2021-01-21]. https://arxiv.org/pdf/1911.10949.pdf

Wu Y H and Hu Z Y. 2006. PnP problem revisited. Journal of Mathematical Imaging and Vision, 24(1) 131-141 [DOI: 10.1007/s10851-005-3617-z]

Wu Y H, Tang F L and Li H P. 2018b. Image-based camera localization: an overview. Visual Computing for Industry, Biomedicine, and Art, 1: #8 [DOI: 10.1186/s42492-018-0008-z]

Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O and Xiao J X. 2015. 3D ShapeNets: a deep representation for volumetric shapes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/cvpr.2015.7298801]

Wu Z J, Wang X, Lin D, Lischinski D, Cohen-Or D and Huang H. 2019c. SAGNet: structure-aware generative network for 3D-shape modeling. ACM Transactions on Graphics, 38(4): #91 [DOI: 10.1145/3306346.3322956]

Xiao L H, Wang J, Qiu X S, Rong Z and Zou X D. 2019. Dynamic-SLAM: semantic monocular visual localization and mapping based on deep learning in dynamic environment. Robotics and Autonomous Systems, 117: 1-16 [DOI: 10.1016/j.robot.2019.03.012]

Xie J Y, Girshick R and Farhadi A. 2016. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 842-857 [DOI: 10.1007/978-3-319-46493-0_51]

Xie S N, Gu J T, Guo D M, Qi C R, Guibas L and Litany O. 2020. PointContrast: unsupervised pre-training for 3D point cloud understanding//Proceedings of European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 574-591 [DOI: 10.1007/978-3-030-58580-8_34]

Xu C, Wu B, Wang Z, Tomizuka M. 2020. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation [C]// European Conference on Computer Vision. Springer, Cham: 1-19.

Xu K, Zhang H, Cohen-Or D and Chen B Q. 2012. Fit and diverse: set evolution for inspiring 3D shape galleries. ACM Transactions on Graphics, 31(4): #57 [DOI: 10.1145/2185520.2185553]

Xu L, Cheng W, Guo K W, Han L, Liu Y B and Fang L. 2021. FlyFusion: realtime dynamic scene reconstruction using a flying depth camera. IEEE Transactions on Visualization and Computer Graphics, 27(1): 68-82 [DOI: 10.1109/TVCG.2019.2930691]

Xu L, Liu Y B, Cheng W, Guo K W, Zhou G Y, Dai Q H and Fang L. 2018a. FlyCap: markerless motion capture using multiple autonomous flying cameras. IEEE Transactions on Visualization and Computer Graphics, 24(8): 2284-2297 [DOI: 10.1109/TVCG.2017.2728660]

Xu L, Su Z, Han L, Yu T, Liu Y B and Fang L. 2020. UnstructuredFusion: realtime 4D geometry and texture reconstruction using commercial RGBD cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2508-2522 [DOI: 10.1109/TPA-

MI. 2019. 2915229]

Xu Y F, Fan T Q, Xu M Y, Zeng L and Qiao Y. 2018b. SpiderCNN: deep learning on point sets with parameterized convolutional filters//Proceedings of European Conference on Computer Vision. Munich, Germany: Springer: 90-105 [DOI: 10.1007/978-3-030-01237-3_6]

Xu Y, Zhu X, Shi J P, Zhang G F, Bao H J and Li H S. 2019. Depth completion from sparse LiDAR data with depth-normal constraints//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2811-2820 [DOI: 10.1109/ICCV.2019.00290]

Xue F, Wang X, Li S K, Wang Q Y, Wang J Q and Zha H B. 2019. Beyond tracking: selecting memory and refining poses for deep visual odometry//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8567-8575 [DOI: 10.1109/CVPR.2019.00877]

Yan M, Wang J Z, Li J and Zhang C. 2017. Loose coupling visual-lidar odometry by combining VISO2 and LOAM//Proceedings of the 36th Chinese Control Conference. Dalian, China: IEEE: 6841-6846 [DOI: 10.23919/ChiCC.2017.8028435]

Yang B, Wang J, Clark R, Hu Q Y, Wang S, Markham A and Trigoni N. 2019. Learning object bounding boxes for 3D instance segmentation on point clouds//Advances in Neural Information Processing Systems. Vancouver, Canada: [s. n.]: 6737-6746

Yang G, Zhao H, Shi J, Deng Z and Jia J. 2018a. SegStereo: exploiting semantic information for disparity estimation//Proceedings of European Conference on Computer Vision. Munich, Germany: Springer: 660-676 [DOI: 10.1007/978-3-030-01234-2_39]

Yang N, von Stumberg L, Wang R and Cremers D. 2020. D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1278-1289 [DOI: 10.1109/CVPR42600.2020.00136]

Yang N, Wang R, Stückler J and Cremers D. 2018b. Deep virtual stereo odometry: leveraging deep depth prediction for monocular direct sparse odometry//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 835-852 [DOI: 10.1007/978-3-030-01237-3_50]

Ye H Y, Huang H Y and Liu M. 2020a. Monocular direct sparse localization in a prior 3D surfel map. [EB/DL]. [2021-02-03]. https://arxiv.org/pdf/2002.09923.pdf

Ye W L, Zheng R J, Zhang F Q, Ouyang Z Z and Liu Y. 2019. Robust and efficient vehicles motion estimation with low-cost multi-camera and odometer-gyroscope//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China: IEEE: 4490-4496 [DOI: 10.1109/IROS40897.2019.8968048]

Ye X C, Chen S D and Xu R. 2020b. DPNet: detail-preserving network for high quality monocular depth estimation. Pattern Recognition, 109: #107578 [DOI: 10.1016/j.patcog.2020.107578]

Ye X Q, Li J M, Huang H X, Du L and Zhang X L. 2018. 3D recurrent neural networks with context fusion for point cloud semantic segmentation//Proceedings of European Conference on Computer Vision. Munich, Germany: Springer: 415-430 [DOI: 10.1007/978-3-030-01234-2_25]

Yi L, Zhao W, Wang H, Sung M and Guibas L J. 2019. GSPN: generative shape proposal network for 3D instance segmentation in point cloud//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3942-3951 [DOI: 10.1109/CVPR.2019.00407]

Yin Z C and Shi J P. 2018. GeoNet: unsupervised learning of dense depth, optical flow and camera pose//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1983-1992 [DOI: 10.1109/CVPR.2018.00212]

Yu H L, Ye W C, Feng Y J, Bao H J and Zhang G F. 2020. Learning bipartite graph matching for robust visual localization//Proceedings of 2020 IEEE International Symposium on Mixed and Augmented Reality. Porto de Galinhas, Brazil: IEEE: 146-155 [DOI: 10.1109/ISMAR50242.2020.00036]

Yu T, Guo K W, Xu F, Dong Y, Su Z Q, Zhao J H, Li J G, Dai Q H and Liu Y B. 2017. BodyFusion: real-time capture of human motion and surface geometry using a single depth camera//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 910-919 [DOI: 10.1109/ICCV.2017.104]

Yu T, Zhao J H, Huang Y H, Li Y P and Liu Y B. 2019a. Towards robust and accurate single-view fast human motion capture. IEEE Access, 7: 85548-85559 [DOI: 10.1109/ACCESS.2019.2920633]

Yu T, Zheng Z R, Guo K W, Zhao J H, Dai Q H, Li H, Pons-Moll G and Liu Y B. 2018. DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7287-7296 [DOI: 10.1109/CVPR.2018.00761]

Yu T, Zheng Z R, Zhong Y, Zhao J H, Dai Q H, Pons-Moll G and Liu Y B. 2019b. SimulCap: single-view human performance capture with cloth simulation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5499-5509 [DOI: 10.1109/CVPR.2019.00565]

Zagoruyko S and Komodakis N. 2015. Learning to compare image patches via convolutional neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4353-4361 [DOI: 10.1109/CVPR.2015.7299064]

Žbontar J and LeCun Y. 2015. Computing the stereo matching cost with a convolutional neural network//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1592-1599 [DOI: 10.1109/CVPR.2015.7298767]

Zeisl B, Sattler T and Pollefeys M. 2015. Camera pose voting for large-

scale image-based localization//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2704-2712 [DOI: 10.1109/ICCV.2015.310]

Zhan H Y, Garg R, Weerasekera C S, Li K J, Agarwal H and Reid I M. 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 340-349 [DOI: 10.1109/CVPR.2018.00043]

Zhan H Y, Weerasekera C S, Bian J W and Reid I. 2020. Visual odometry revisited: what should be learnt? //Proceedings of 2020 IEEE International Conference on Robotics and Automation. Paris, France: IEEE: 4203-4210 [DOI: 10.1109/ICRA40945.2020.9197374]

Zhang J and Singh S. 2018. Laser-visual-inertial odometry and mapping with high robustness and low drift. Journal of Field Robotics, 35(8): 1242-1264 [DOI: 10.1002/rob.21809]

Zhang L, Chen W H, Hu C, Wu X M and Li Z G. 2019a. S&CNet: monocular depth completion for autonomous systems and 3D reconstruction [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/1907.06071.pdf

Zhang P J, Wu Y H and Liu B X. 2020a. Leveraging local and global descriptors in parallel to search correspondences for visual localization [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/2009.10891.pdf

Zhang Y G and Li Q. 2018. Multi-frame fusion method for point cloud of LiDAR based on IMU. Journal of System Simulation, 30(11): 4334-4339 (张艳国, 李擎. 2018. 基于惯性测量单元的激光雷达点云融合方法. 系统仿真学报, 30(11): 4334-4339)

Zhang Y, Zhou Z X, David P, Yue X Y, Xi Z R, Gong B Q and Foroosh H. 2020b. PolarNet: an improved grid representation for online LiDAR point clouds semantic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9598-9607 [DOI: 10.1109/CVPR42600.2020.00962]

Zhang Z Y, Hua B S and Yeung S K. 2019b. ShellNet: efficient point cloud convolutional neural networks using concentric shells statistics//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1607-1616 [DOI: 10.1109/ICCV.2019.00169]

Zhao C, Sun L, Purkait P, Duckett T and Stolkin R. 2018. Learning monocular visual odometry with dense 3D mapping from dense 3D flow//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 6864-6871 [DOI: 10.1109/IROS.2018.8594151]

Zhao H S, Jiang L, Fu C W and Jia J Y. 2019. PointWeb: enhancing local neighborhood features for point cloud processing//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5560-5568 [DOI: 10.1109/CVPR.2019.00571]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zheng B and Zhang Z X. 2019. An improved EKF-SLAM for Mars surface exploration. International Journal of Aerospace Engineering, 2019: #7637469 [DOI: 10.1155/2019/7637469]

Zheng Z R, Yu T, Li H, Guo K W, Dai Q H, Fang L and Liu Y B. 2018. HybridFusion: real-time performance capture using a single depth sensor and sparse IMUs//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 389-406 [DOI: 10.1007/978-3-030-01240-3_24]

Zhi S F, Bloesch M, Leutenegger S and Davison A J. 2019. SceneCode: monocular dense semantic reconstruction using learned encoded scene representations//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11768-11777 [DOI: 10.1109/CVPR.2019.01205]

Zhong Y R, Li H D and Dai Y C. 2018. Open-world stereo video matching with deep RNN//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 104-119 [DOI: 10.1007/978-3-030-01216-8_7]

Zhou C, Zhang H, Shen X Y and Jia J Y. 2017a. Unsupervised learning of stereo matching//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1576-1584 [DOI: 10.1109/ICCV.2017.174]

Zhou H, Zhu X, Song X, Ma Y C, Wang Z, Li H S and Lin D H. 2020a. Cylinder3D: an effective 3D framework for driving-scene LiDAR semantic segmentation [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/2008.01550.pdf

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

Zhou Y, Wan G W, Hou S H, Yu L, Wang G, Rui X F and Song S Y. 2020b. DA4AD: end-to-end deep attention-based visual localization for autonomous driving [EB/OL]. [2021-02-03]. https://arxiv.org/pdf/2003.03026.pdf

Zhu C, Giorgi G, Lee Y H and Günther C. 2018a. Enhancing accuracy in visual SLAM by tightly coupling sparse ranging measurements between two rovers//Proceedings of 2018 IEEE/ION Position, Location and Navigation Symposium. Monterey, USA: IEEE: 440-446 [DOI: 10.1109/PLANS.2018.8373412]

Zhu C Y, Xu K, Chaudhuri S, Yi R J and Zhang H. 2018b. SCORES: shape composition with recursive substructure priors. ACM Transactions on Graphics, 37(6): #211 [DOI: 10.1145/3272127.

3275008］

Zhu H, Su H, Wang P, Cao X and Yang R G. 2018c. View extrapolation of human body from a single image//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4450-4459［DOI: 10.1109/CVPR.2018.00468］

Zhu X, Zhou H, Wang T, Hong F Z. 2020. cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation［EB/OL］.［2021-01-21］. https://arxiv.org/pdf/2011.10033.pdf

Zhu Z L, Yang S W, Dai H D and Li F. 2018d. Loop detection and correction of 3D laser-based SLAM with visual information//Proceedings of the 31st International Conference on Computer Animation and Social Agents. Beijing, China: ACM: 53-58［DOI: 10.1145/3205326.3205357］

Zoph B, Vasudevan V, Shlens J and Le Q V. 2018. Learning transferable architectures for scalable image recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8697-8710［DOI: 10.1109/CVPR.2018.00907］

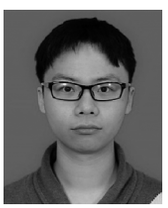
Zubizarreta J, Aguinaga I and Montiel J M M. 2020. Direct sparse mapping. IEEE Transactions on Robotics, 36(4): 1363-1370［DOI: 10.1109/TRO.2020.2991614］

Zuo X X, Geneva P, Yang Y L, Ye W L, Liu Y and Huang G Q. 2019. Visual-inertial localization with prior LiDAR map constraints. IEEE Robotics and Automation Letters, 4(4): 3394-3401［DOI: 10.1109/LRA.2019.2927123］

作者简介

龙霄潇，1996 年生，男，博士研究生，主要研究方向为深度估计，三维感知和三维重建。
E-mail：xxlong@hku.hk

程新景，1992 年生，男，博士研究生，主要研究方向为计算机视觉及三维感知。
E-mail：xinjing.cheng@inceptio.ai

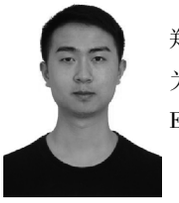


朱昊，1991 年生，男，特任副研究员，主要研究方向为计算机视觉、机器学习。
E-mail：zhuhaoese@nju.edu.cn

张朋举，1993 年生，男，博士研究生，主要研究方向是视觉定位，图像检索和特征描述。
E-mail：pengju.zhang@nlpr.ia.ac.cn

刘浩敏，1987 年生，男，商汤科技研究副总监，主要研究方向为运动恢复结构、同步定位与地图构建和增强现实。
E-mail：liuhaomin@sensetime.com

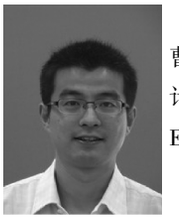
李俊，1986 年生，男，助理研究员，主要研究方向为计算机图形学，3D 视觉，机器学习。
E-mail：jun.johnson.li@gmail.com



郑林涛，1991 年生，男，讲师，主要研究方向为三维重建、3D 视觉、机器人智能感知。
E-mail：lintaozheng1991@gmail.com

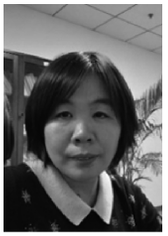
胡庆拥，1995 年生，男，博士研究生，主要研究方向是大规模点云语义理解与配准。
E-mail：qingyong.hu@cs.ox.ac.uk

刘浩，1994 年生，男，博士研究生，主要研究方向为三维场景语义理解、点云目标检测与跟踪。
E-mail：liuh327@mail2.sysu.edu.cn



曹汛，1983 年生，男，教授，主要研究方向为计算摄像学，图像和视频处理。
E-mail：caoxun@nju.edu.cn

杨睿刚,1973 年生,男,教授,主要研究方向为计算机视觉及三维感知。
E-mail: ryang@ inceptio. ai



吴毅红,1973 年生,女,研究员,主要研究方向为多视几何理论、相机标定与定位、SLAM、三维重建及在机器人定位与导航、AR、VR 中的应用。
E-mail: yhwu@ nlpr. ia. ac. cn

章国锋,1981 年生,男,教授,主要研究方向为同步定位与地图构建、计算机视觉和混合现实。
E-mail: zhangguofeng@ zju. edu. cn

刘烨斌,1981 年生,男,清华大学长聘副教授,主要研究方向为计算机视觉与计算机图形学。
E-mail: liuyebin@ tsinghua. edu. cn。

徐凯,1982 年生,男,教授,主要研究方向为计算机图形学、3D 视觉。
E-mail:kevin. kai. xu@ gmail. com

郭裕兰,1985 年生,男,副教授,主要研究方向为三维视觉与模式识别。
E-mail: yulan. guo@ nudt. edu. cn

陈宝权,1969 年生,通信作者,男,教授,主要研究方向为计算机视觉、计算机图形学与可视化。
E-mail: baoquan@ pku. edu. cn