液晶与显示
Chinese Journal of Liquid Crystals and Displays

文章编号:1007-2780(2021)XX-0001-18

视觉深度估计与点云建图研究进展

陈苑锋
(美的集团(上海)有限公司,上海 201799)

摘要:即时定位导航(SLAM)是无人驾驶和机器人实现自主移动的关键技术,而目前广泛应用于SLAM技术中的激光雷达传感器存在成本高昂、激光点云空间分辨率低及难以获得精确的语义信息等一系列问题。相比之下视觉传感器(摄像头等)可以有效避免以上问题,但是在深度预测和建图等方面需要更复杂的算法。近年来,随着处理器算力的提升、数据集的不断丰富,以及新的机器视觉算法的提出,视觉深度预测和建图算法的精度和效率都已经有了较大提升。本文对现有视觉深度预测与视觉建图方法进行了总结,从视觉数据的采集和算法设计等方面进行分类阐述,最后针对视觉深度估计和建图方案的应用场景和未来发展方向进行了分析。

关　键　词:单目深度;双目深度;深度学习;点云建图
中图分类号:TP394.;1;TH691.;9　　**doi**:10.37188/CJLCD.2021-0047

Progress of visual depth estimation and point cloud mapping

CHEN Yuan-feng

(*Midea Group* (*Shanghai*) *Co.*, *Ltd.*, *Shanghai* 201799, *China*)

Abstract: Simultaneous location and mapping (SLAM) is a key technology for autonomous driving vehicles and robots to realize autonomous movement. The lidar currently widely used in SLAM technology present a series of issues, including high cost, low spatial resolution of laser point clouds, and difficulty in obtaining accurate semantic information. In contrast, cameras can effectively avoid the above problems, but more complex algorithms are required in depth prediction and mapping. In recent years, with the increase in computing power, the continuous enrichment of data sets, as well as the introduction of new machine vision algorithms, the accuracy and efficiency of vision depth prediction and mapping algorithms have been greatly improved. This article summarizes the existing methods of visual depth prediction and point cloud mapping and classifies the methodology based on visual data collection approach and algorithm design, then analyzes the application scenarios and prospect of visual depth estimation and point cloud mapping.

Key words: monocular depth detection; binocular depth detection; deep learning; point cloud mapping

1　引言

是同步定位与地图构建(Simultaneous Localization And Mapping,SLAM)主要用于解决移动机器人在未知环境中运行时定位导航与地图构建的问题。SLAM通常包括如下几个部分,特

收稿日期:2021-02-24;**修订日期**:2021-03-20.
*通讯联系人,E-mail: yuanfeng_chen@hotmail.com

征提取,数据关联,状态估计,状态更新以及特征更新等。对于其中每个部分,均存在多种方法。SLAM方法大概分为3种形式:(1)在给定地图的情况下,估计机器人的位姿;(2)同时估计机器人的位姿和环境地图;(3)在给定机器人位姿的情况下,估计环境地图。

即时定位导航(SLAM)技术通常依赖激光雷达传感器,因其可提供高精度的3D点云信息。3D激光SLAM的帧间匹配方法通常包括以下3种:点云配准算法、Point-to-Plane ICP、Feature-based Method。而常用的3D激光SLAM的回环检测方法包括Scan-to-Scan、Scan-to-Map、Branch and Bound和Lazy Decision。目前主流激光SLAM算法框架包括:(1)LOAM-纯激光,匀速运动假设,无回环;(2)V-LOAM-视觉激光融合、漂移匀速假设,无回环;(3)VELO-视觉激光融合,无运动畸变假设,有回环。

但因激光雷达价格昂贵,某种程度上影响了其市场化的步伐。此外激光雷达由于受制于线数,在竖直方向的空间分辨率有限,难以精确反映目标物体轮廓形态,进而难以获得精确的语义信息。而近些年随着人工智能技术的快速发展,基于视觉的SLAM,即VSLAM逐渐成为研究热点[1-2]。VSLAM涉及两项核心技术,分别为视觉深度估计以及视觉建图技术。其中视觉建图以3D视觉点云图为输入,通过多视角特征匹配进行建图,其方法与逻辑与激光雷达点云建图非常类似,因此技术较为成熟,且难度可控[3-4]。而视觉深度估计相比较于激光雷达深度测量在测量精度方面面临着较大挑战。

2 视觉深度估计

视觉深度估计已逐渐成为当前计算机视觉领域的研究热点之一。无论是基于单目、双目还是多目的深度估计对于场景理解和实现自主导航定位均具有重要意义。以常用的集中视觉深度估计方法为例,基于双目视觉的深度估计受基线长度限制,导致设备体积与载具平台不能很好的匹配[5]。基于RGBD的深度估计量程较短、且价格也不菲,在实际应用中能力有限,同时在室外环境中的表现也不尽理想,受环境变化影响较大。而单目摄像头具有价格低廉,获取信息内容丰富,体积小等优点,可以有效克服上述传感器的诸多不足。然而现有的有监督和无监督方法均面临着巨大的挑战。有监督的方法需要大量的深度测量数据,这些数据通常很难获得,而无监督的方法通常在估计精度上受到限制。

表1对视觉深度估计方法进行了汇总,该表格从摄像头类型、计算模型(以深度学习模型为主)名称、所采用的数据集名称、数据量、深度学习模型监督类型和发表年份等方面进行了总结。从摄像头类型角度分析,近年来更多的研究集中于单目摄像头的深度估计,其主要原因一方面在于单目摄像头在硬件布置和成本上具有优势,另一方面在于神经网络加速芯片取得较大的进展,推动了单目深度神经网络算法的进展。本文将先从双目和多目深度估计入手进行总结,最后讨论单目深度估计。表1中神经网络的类型包括了有监督、半监督、自监督和无监督。表格中所列的文献主要发表于2017-2020年之间,是对近年来最新方法的总结,所涉及的具体方法将在2.1-2.3中展开介绍。

2.1 双目视觉深度预测

双目深度估计也称作视差估计(Disparity Estimation),或者立体匹配[31]。其输入是一对在同一时刻捕捉到的,经过极线校正的左右视图,而输出是依据摄像头焦距f、左右摄像头基线长度b、以及左右眼对于同一目标的视差,通过相似三角形计算出目标深度图d。视差是三维场景中某一点在左右图像中对应点位置的像素级差距。所以深度和视差是可以互相转换,相互等价的,如图1所示。

立体匹配作双目深度估计中的基本挑战,其任务是获得左右图片中像素的对应关系,从而计算出视差图。过去几十年,科研人员探索了多种双目立体视觉匹配算法,如SAD匹配算法、SURF算法、BM算法、SGBM算法、GC算法等[5,32]。代表性的工作包括Yao等人[33]提出的一个深度感知系统,该系统包含一个类似于Kinect的激光投影机,并在激光投影机的两侧安装两个红外摄像头,以获得更高的空间分辨率的深度信息。作者采用块匹配算法来估计视差。为了提高空间分辨率,减小了匹配块的大小,但是越小的匹配块产生的匹配精度越低。为了解决这一

表1 视觉深度预测方法汇总表

Tab. 1 Summary of visual depth prediction

摄像头类型	模型名称	数据集名称	数据量	监督类型	年份
双目	BDfF－Net[6]	FlyingThings3D	25 000	有监督	2020
	TFBD[7]	自采集	6 000	有监督	2018
	Stereo Cycle/ Stereo Half－Cycle[8]	KITTI＋CS＋ApolloScape	23 297＋24 498＋11 342	无监督	2019
	unsupervised binocular Resnet[9]	自采集	7 191(test)	无监督	2019
多目/多视角	DELTAS[10]	ScanNet	250万	有监督	2020
	MVSNet[11]	DTU	27 097	有监督	2018
	NNet[12]	SUN3D＋RGBD＋Scenes11	166 285	有监督	2020
	MVDepthNet[13]	SUN3D＋TUM RGB－D＋MVS＋SceneNN＋Scenes11	436 928(t)13 394(test)	有监督	2018
	MaskMVS[14]	SUN3D＋RGBD＋MVS＋Scenes11	92 558	有监督	2019
	无[15]	ScanNet	250万	有监督	2020
单目	Monodepth2[16]	KITTI	39 810(t)4 424(v)	自监督	2018
	MonoGAN[17]	CS＋KITTI	29 000＋23 000	无监督	2018
	无[18]	合成环境＋KITTI	70 000(t)10 000(v)	直接监督	2018
	UnDEMoN[19]	KITTI	27 116(t)2 979(v)	无监督	2018
	Struct2depth[20]	KITTI＋CS＋FIN	30 000＋4 500＋1 626	无监督	2019
	DORN[21]	KITTI＋Make3D＋NYU Depth v2	24 185＋534＋464	有监督	2018
	StereoNoFt→Mono StereoUnsupFt→Mono[22]	Scene Flow＋KITTI	43 000＋24 185	无监督	2018
	无[23]	KITTI＋CamVid＋CS	146＋701＋5 000	自监督	2018
	无[24]	KITTI	12 844	半监督	2017
	无[25]	KITTI＋Uncalibrated Bike Video	44 697＋91 866	无监督	2018
	DABC[26]	ScanNet＋KITTI	250万	有监督	2018
	Competitive Collaboration[27]	KITTI	无	联合无监督	2018
	PyD－Net[28]	KITTI	23 297	无监督	2018
	depth ConvNet[29]	KITTI	42 382	无监督	2018
	无[30]	KITTI	24 185	自监督	2018

注:* 表格中t代表training,v代表validation* CS表示cityscapes数据集,FIN表示获取室内导航数据集

问题,作者在视差估计过程中结合了两种匹配模式(双目模式和单眼模式)。实验结果表明,与Kinect相比,该方法可以在不影响距离图像质量的前提下获得更高的空间分辨率深度。然而传统的立体匹配算法在精度和速度仍需不断提升。

除了传统算法,深度学习算法在计算机视觉

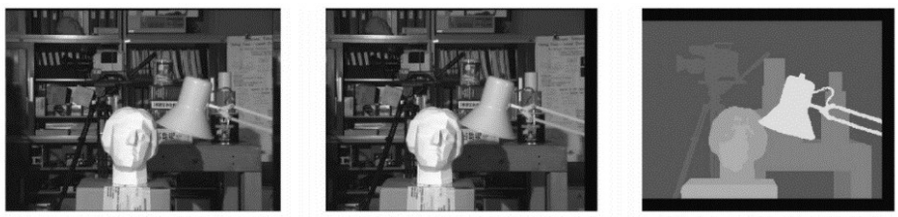


图1 双目摄像头数据。通过左右眼的差异,能够判断场景中物体离摄像头距离[5]。(a) 左眼图像 (b) 右眼图像 (c) 深度图

Fig. 1 Data of the binocular camera, Through the difference between the left and right images, the distance between the objects in the scene and the camera can be detected [5]. (a) Left image (b) Right image (c) Depth map

的诸多领域中获得了巨大成功,例如语义分割,物体检测,图像识别等。现如今深度学习在立体匹配上也取得了成功的应用。基于深度学习的立体匹配是将传统立体匹配方法的四个步骤,即代价计算、代价聚合、视差计算和视差细化,融入到卷积神经网络当中[34]。在KITTI数据集上,大多数排名靠前的方法都是基于深度学习的[35]。例如Song等人[7]在算法上对双目深度预测方法做了提升,提出了两种新的抗欺骗干扰的鲁棒性特征。第一种是基于双目摄像头的深度特征,称为模板人脸匹配双目深度特征;二是基于空间金字塔编码微纹理特征的高阶微纹理特征。在此基础上,提出了一种新的模板人脸配准算法和空间金字塔编码算法。基于这些特征实现了多模态人脸欺骗检测。

人类视觉系统既依赖双目立体信息,也依赖单眼聚焦线索来获得有效的三维感知。在计算机视觉中,这两个问题传统上是分开解决的。而Guo等人[6]同时使用这两种类型的线索进行深度推断。具体来说,作者使用一对焦距堆栈作为输入来模拟人类的感知。作者首先构建一个由深度引导光场渲染合成的综合焦堆栈训练数据集。然后构建了3个独立的网络:一个Focus-Net用于从单个焦堆栈中提取深度,一个EDoF-net用于从焦堆栈中获得扩展景深图像,一个stereo -net用于进行立体匹配。作者还展示了如何将它们集成到统一的BDfF-Net中,以获得高质量的深度图。

前述的深度学习算法里,基于监督回归的方法取得了显著的效果。然而,它们需要为算法训练进行昂贵的地面实况(Ground truth)注释。为了解决这一问题,Pilzer等人[8]提出了一种新的无监督深度学习方法来预测深度图。该方法采用一种新的网络结构,称为渐进融合网络(Progressive Fusion Network,PFN),它是专门为双目立体深度估计而设计的。该网络基于一种多尺度的细化策略,该策略结合了双目摄像头采集的两个立体视图。此外,文中还建议将这个网络堆叠两次,形成一个循环。这种循环方法可以被理解成数据增强的一种形式,因为在训练时,网络既可以从训练集图像(前半周)学习,也可以从合成图像(后半周)学习。该架构是与对抗型学习共同训练的。定性比较结果如图2所示。作为无监督学习的双目深度预测在医疗领域的重大应用,Xu等人[9]重建了双目立体腹腔镜的精确深度图。二维图像的三维腹腔镜成像让外科医生有深度感知,从而克服传统的二维腹腔镜成像缺乏深度感知、不能提供定量的深度信息、进而限制手术时的视野和范围等问题。

通过以上对于双目深度估计的代表性文献的总结,可以发现提高空间分辨率是提升深度信息质量的有效手段。就如同[33]中将两种匹配模式跟块匹配算法结合,更好地实现视差估计,从而得到深度信息。基于双目硬件分别实现单目算法和双目算法有可能在产品落地方面产生不错的应用效果。此外双目深度估计技术也在人脸欺骗检测和腹腔镜深度图预测上有了很好的效果与应用。学者们也正在探索人类视觉系统预测深度的原理,并且尝试通过双目立体线索和单眼聚焦线索获得有效的三维感知。此外在算法层面,渐进融合网络(PFN)与对抗性学习共同训练也是一个很好的研究方向。

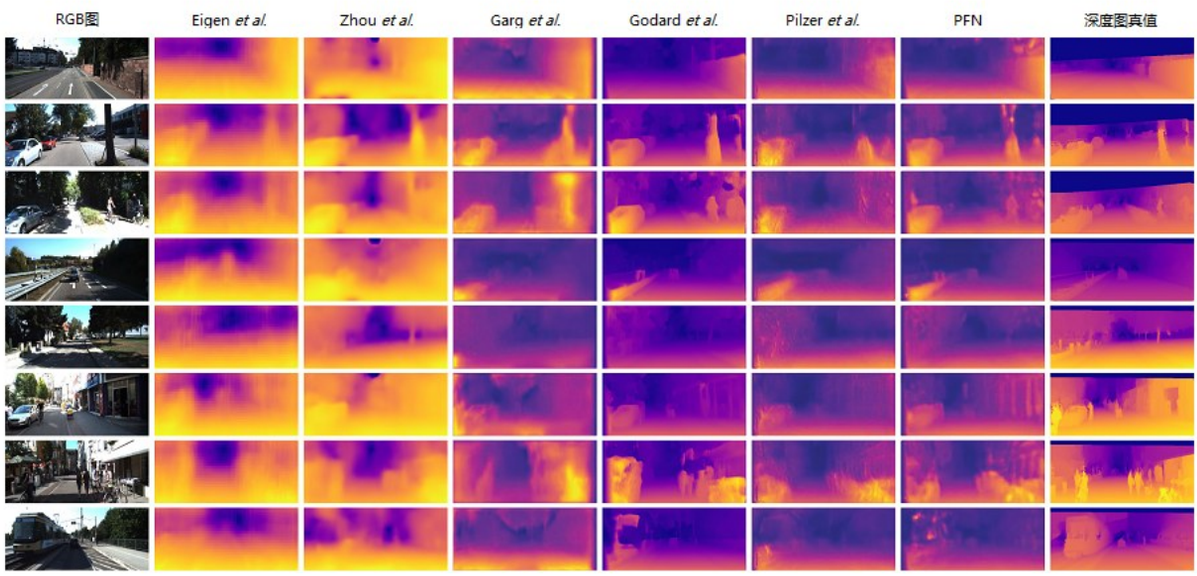


图2 各种方法定性比较[8]

Fig. 2 Qualitative comparison of various methods[8]

2.2 多目视觉深度预测

除双目视觉外，学界对多目视觉深度预测方法也开展了一系列研究，代表性的工作包括Anantrasirichai等人[36]提出的一种基于多视角图像的视差/深度估计算法。该算法基于窗口相关的动态规划方法和一种新的代价函数。视差/深度映射的平滑性被嵌入到动态规划方法中，而基于窗口的相关性增加了可靠性。增强方法包括自适应窗口大小和可移动窗口，以提高均匀区域的可靠性和提高目标边界的锐度。算法首先估计沿着单个摄像头轴的深度地图，然后将不同轴的深度估计值结合起来，得到适合多视图图像的深度图。该方案在并行和非并行的摄像头配置中都优于现有的方法。此外，Montserrat等人[37]提出了一种基于信念传播的多视图匹配与深度/颜色分割算法，并给出了一种信息传递压缩策略。在此基础上Liu等人[38]通过引入深度候选对象将多视图深度图合并生成3D模型。该模型首先将轮廓信息和外极约束集成到连续深度图的变分方法中，然后，基于多起始尺度(MSS)框架生成多个深度候选对象。从这些候选对象中，根据基于路径的归一化互相关度量合成每个视图的精细化深度图。

但上述方法并不适用于任意视角，Lee和Ho[39]提出了一种基于视点一致性的多视点深度估计算法。主要是在使用传统的深度估计方法获得左右视点的深度图后，将其投影到中心视点，并使用多视点图割算法进行误差最小化。此外，Mieloch等人[40]提出了一种适用于任意摄像头位置的多视点系统的深度估计方法。该方法利用了图割线法，图的顶点表示用于控制深度图质量和估计时间之间的权衡的段，同时保持深度图的原始分辨率。此外，通过在优化图中引入合适的连接，保证了对自由视点系统至关重要的深度图的视图间一致性。这使得该方法成为第一个允许使用基于分割的估计生成空间一致的多视图深度图的方法。如图3所示。

Facil等人[41]进一步利用了基于CNN的单视图深度估计的最新结果，并将其与多视图深度估计融合。这两种方法具有互补的优势。多视图深度是高度精确的，但仅在高纹理区域和高视差的情况下。单视图深度捕获了中层区域的局部结构，包括无纹理区域，但估计的深度缺乏全局一致性。该文献提出的单视点和多视点融合算法在几个方面具有挑战性。首先，这两个深度都与依赖于图像内容的变形有关。其次，对于低视差配置，高精度多视点的选择可能比较困难。

另一方面，Long等人[15]提出了一种基于单个视频的多视角深度估计方法。虽然以前的基于学习的方法已经证明了令人信服的结果，但大多数方法都是独立地估计单个视频帧的深度图，而没有考虑帧之间强烈的几何和时间一致性。此

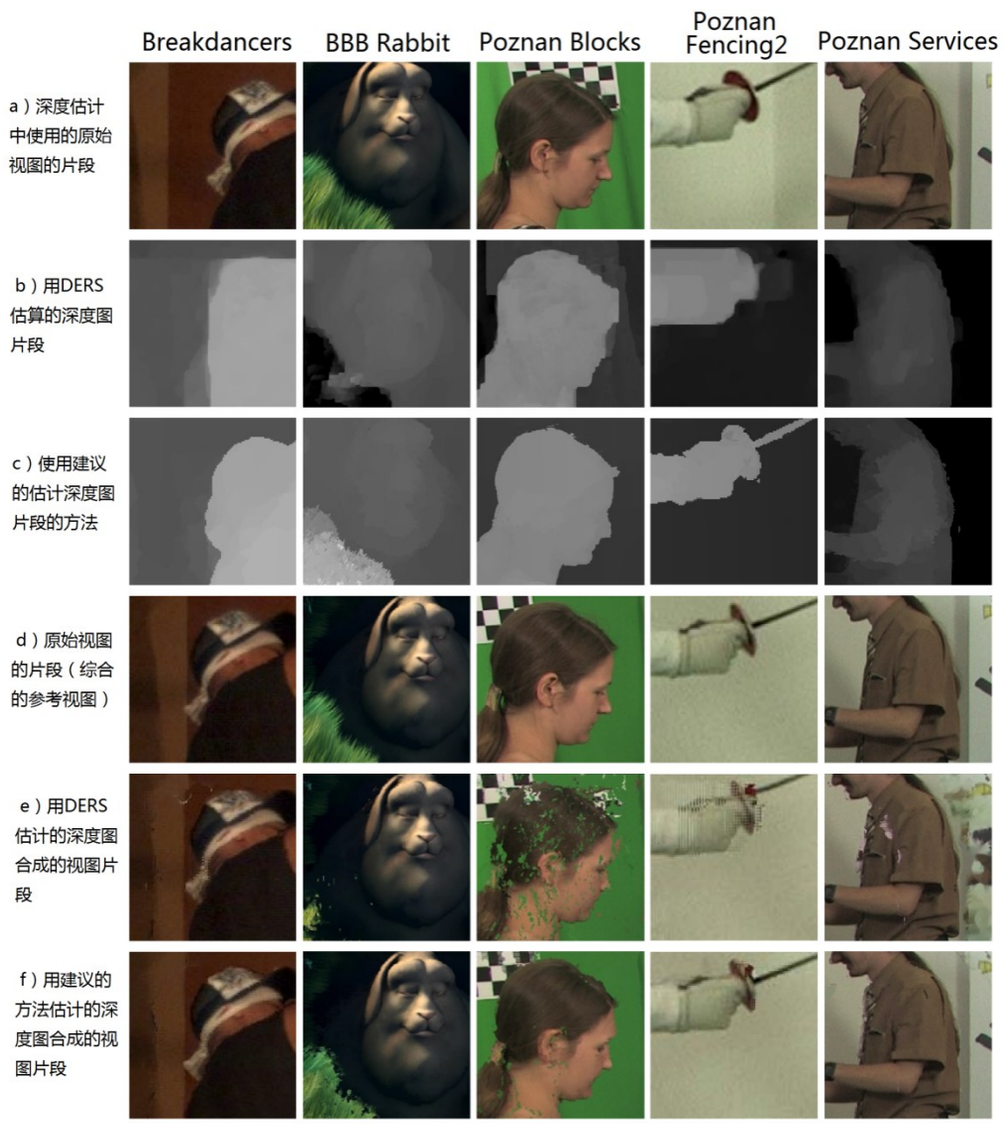


图3 深度图与虚拟视点合成的比较[40]

Fig. 3 Comparison of depth map and virtual viewpoint synthesis[40]

外，目前最先进的（SOTA）模型大多采用全3D卷积网络进行成本正则化，因此需要较高的计算成本，从而限制了其在现实应用中的部署。该文章的方法通过使用一个新的极时空（EST）变压器来实现时间相干深度估计结果，明确地关联几何和时间相关性与多个估计的深度地图。Yang等人[42]提出了一种从多视点同步和校准视频流中恢复空间和时间一致的深度图的方法。采用左右视图匹配和基于颜色的分割相结合的方法对深度图进行初始化。在此基础上，将色彩一致性和空间一致性引入优化框架，以保证单时刻的空间一致性。最后以时空一致性约束的形式加入深度和运动信息来细化和稳定深度视频，在每个瞬间的估计中不破坏原始的空间一致性。

为进一步提升深度估计得计算效率和计算精度，Ince等人[43]考虑了多视点视频编码中视点合成的深度估计，并证明了所提出的深度估计方法不仅可以有效地进行视图综合预测，而且可以生成编码比特数更少的深度图。文献[11]提出了一个端到端的深度学习架构MVSNet，用于从多视图图像进行深度映射推断。该网络首先提取深度视觉图像特征，然后通过可微分单应性翘曲在参考摄像头截锥上建立三维代价体。接下来，应用3D卷积对初始深度图进行正则化和回归，然后与参考图像进行细化，形成输出。Kusupati等人[12]研究了如何利用正态估计模型和预测的

法线图来提高深度质量。Hou等人[14]提出了一种求解非结构化多视角图像位姿对深度估计的新方法——MaskMVS。该方法在平面扫描过程中,通过直方图匹配对深度平面进行采样,确保覆盖感兴趣的深度范围。Sinha等人[10]研究了一种有效的深度估计方法,首先检测和评估兴趣点的描述符,然后学习匹配和三角化一小组兴趣点,最后使用CNN致密化这一稀疏的3D点集。端到端网络在深度学习框架内有效地执行所有这3个步骤,并通过中间2D图像和3D几何监督

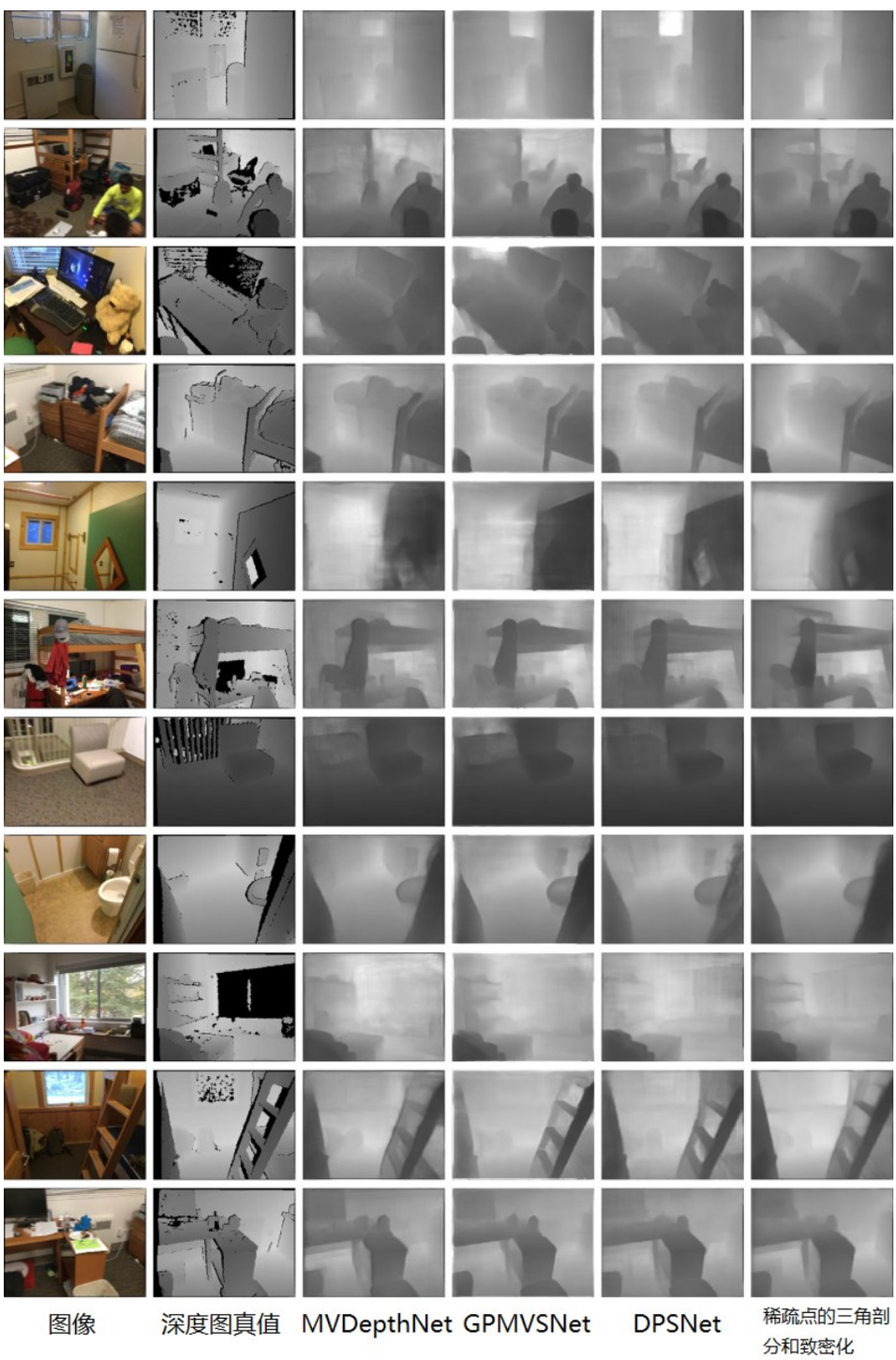


图4 深度预测定性表现[10]

Fig. 4 Qualitative performance of depth prediction[10]

以及深度监督进行训练。训练结果如图4所示。

Long等人的另一项工作[44]中引入了联合法向图(CNM)约束,以更好地保持高曲率特征和全局平面区域。为了进一步提高深度估计精度,作者引入了一种新的遮挡感知策略,将来自多个相邻视图的初始深度预测聚合到最终的深度图和当前参考视图的遮挡概率图中。Strecha和Gool[45]介绍了一种对多幅校正图像进行深度提取的方法,其重点在于匹配过程中多视图的融合。这个过程是由系统对来自不同视图的数据的相对置信度所引导的。这种权重是细粒度的,因为它是在每次迭代中为每个像素确定的。

通过以上对多目视觉深度预测的总结可以发现该方向的诸多新思路。例如,轮廓信息和外极约束集在连续深度图预测中的应用;适用于任意摄像头位置的多视点系统的深度估计方法;基于窗口相关的动态规划方法和一种新的代价函数的提出等。同时,多视点视频编码中视点合成的深度估计证明了该方法不仅可以有效地进行视图综合预测,而且可以生成编码比特数更少的深度图。另外,以下方法也为提升多目深度估计的准确率提供了很好的启发,包括端到端的深度学习架构MVSNet;利用正态估计模型和预测的法线图来提高深度质量;使用单个局部移动摄像头连续估计深度地图等。除此之外,采用单视图深度估计的最新结果,并将其与多视图深度估计融合,使二者优势互补,无疑会是未来一个很好的发展方向。在多视角图像融合方面,从多视角同步和校准视频流中恢复空间和时间一致的深度图已经取得不错的成果。在一致性问题上,Lee和Ho[39]考虑了视点一致性,而Liu等人[15]也考虑到了帧间强烈的几何和时间一致性。为了更好地保持高曲率特征和全局平面区域,Liu等人[44]还引入了联合法向图(CNM)约束。这些方法为后续的视觉深度估计方案创新提供了启发。

2.3 单目视觉深度预测

从彩色图像生成高质量的深度图,这项研究是十分具有吸引力的,因为它以低廉的价格成本实现深度建图。通过使用大量未标注数据集求解深度,也可以达到为下游识别任务的深度神经网络进行预训练的目的。然而,为监督学习收集具有精确标签的训练数据集本身就是一个巨大的挑战。本节仅针对自监督和无监督方案进行分析探讨。

近期有几个自监督的方法被提出,并且已经证明可以只使用双目摄像头的左右视图[46-47]或单目视频[48]来训练单目深度估计模型。在这两种自我监督的方法中,基于单目视频训练是一种有吸引力的替代立体图像监督的方法,但它也带来了一系列挑战。除了估计深度外,模型还需要估计训练过程中时序图像对之间的帧间运动。这通常涉及到训练一个以有限帧序列作为输入,并输出相应的摄像头变换的位姿估计网络。相反,使用立体图像对数据进行训练,使得摄像头姿态估计成为一次性离线校准,但可能会导致与遮挡和纹理复制等相关的问题。

基于以上自监督方法以及出现的问题,在文章[16]中,作者采用单目自监督的monodepth2模型对每个像素的深度进行学习。该论文采用最小的重投影损失设计来稳健地处理遮挡,并采用全分辨率多尺度减少视觉伪影的采样方法以及自动掩膜以忽略违反摄像头运动假设的训练像素。该模型在KITTI数据集中体现了高精度的深度估计结果,如图5所示。

除了最经典的单目深度无监督网络monodepth2之外,以下文献也提出了其他不同形式的无监督算法。例如Aleotti等人[17]提出在GAN范式下进行无监督单目深度估计,生成器

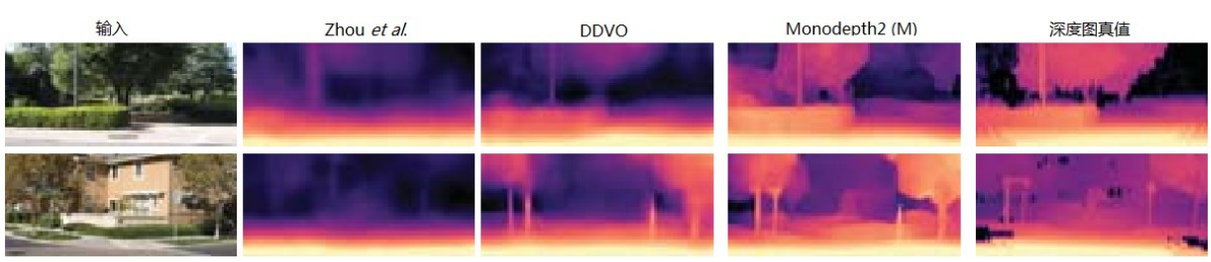


图5 Make3D定性结果(基于KITTI)[16]

Fig. 5 Qualitative results of Make3D (based on KITTI)[16]

网络从参考图像推断深度以生成扭曲的目标图像。在训练时，鉴别器网络学习如何区分由生成器生成的假图像和通过立体装备获取的目标帧，所使用的数据集也是KITTI。预测效果如图6所示。Amir等人[18]利用风格转换和对抗性训练，在对大量合成环境数据进行训练的基础上，从单个真实彩色图像中预测像素深度。实验结果表明，该方法相比当前最先进的技术具有有效性。文献Casser等人[20]在学习过程中引入几何结构，通过对场景和单个物体的建模，以单目视频为输入学习摄像头的自我运动和物体运动。此外，还引入了一种在线求精方法来适应未知领域的动态学习。所提出的方法优于所有最新的方法，包括处理运动（例如通过学习流）的方法，结果如下图7所示。Mahjourian等人[25]提出了一种新的单目视频深度和自我运动的无监督学习方法。作者的主要贡献是明确地考虑整个场景的推断3D几何，并加强估计的3D点云和连续帧的自我运动的一致性。作者还采用了有效掩蔽，以避免无有效信息存在的区域惩罚。论文在KITTI数据集和在未校准的手机摄像头捕获的视频上测试了算法。提出的方法持续改进了这两个数据集的深度估计，在深度和自我运动方面的性能都超过了最先进的水平。在[26]中，作者提出了基于深度注意的分类（DABC）网络，用于鲁棒单一图像深度预测。首先将深度预测作为一个多类分类任务，并应用softmax分类器对每个像素的深度标签进行分类。然后引入全局池化层和通道关注机制，自适应地选择特征的区分通道，并通过赋予重要通道更高的权重来更新原始特征。此外，为了减少量化误差的影响，作者采用了软加权求和推理策略来进行最终预测。Ranjan等人[27]解决了低层次视觉中几个相互关联的问题的无监督学习：单视图深度预测、摄像头运动估计、光流以及将视频分割到静态场景和移动区域。文章的主要见解是，这4个基本的视觉问题是通过几何约束耦合的。因此，学习一起解决它们可以简化问题，因为解决方案可以互相加强。为此，作者引入了竞争协作，这是一个促进多个专门神经网络协调训练以解决复杂问题的框架。竞争协作的工作原理很像期望最大化，但神经网络既扮演着解释与静态或移动区域对应的像素的竞争对手的角色，也扮演着通过调节者将像素分配为静态或独立移动的协作者的角色。Wang等人[13]提出了MVDepthNet，一种卷积网络来解决局部单目摄像头在相邻视点的图像对下的深

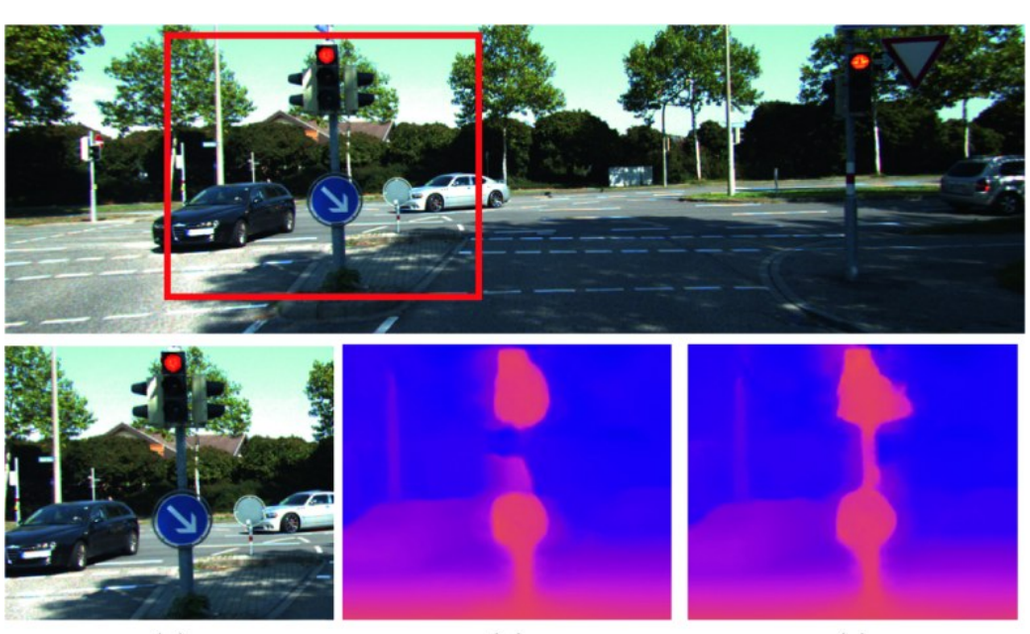


图6 GAN架构与他人论文预测结果对比；(b)由Godard et al预测；(c)由作者的GAN架构预测[17]。

Fig. 6 Comparison of GAN architecture and prediction results of other papers; (b) predicted by Godard *et al.*; (c) predicted by the author's GAN architecture[17].

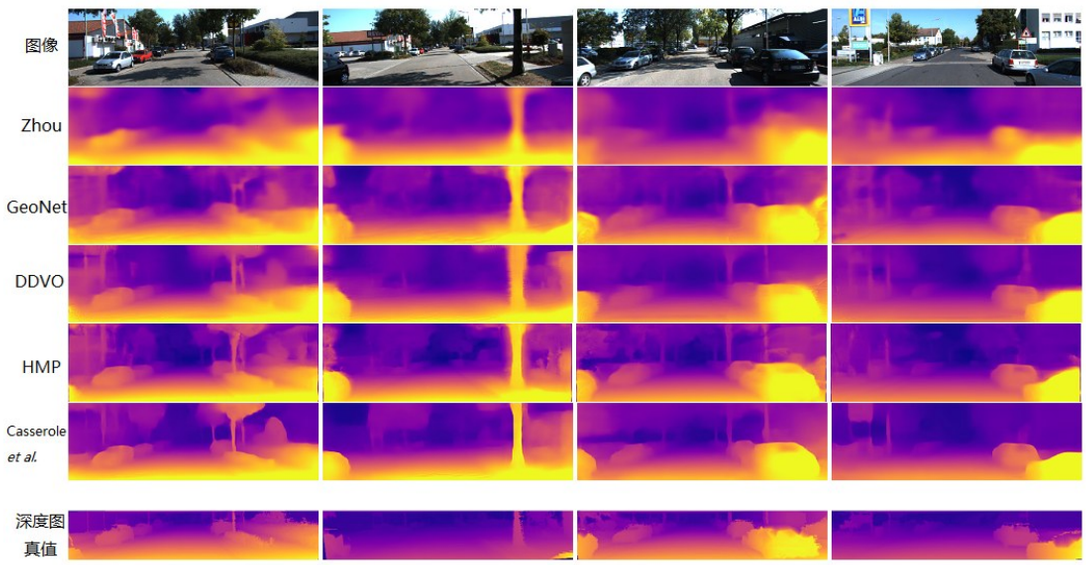


图7 各种方法预测结果比较[20]

Fig. 7 Comparison of prediction results of various methods[20]

度估计问题。多视图观测被编码在一个成本体积中,然后与参考图像结合,使用编码器-解码器网络来估计深度映射。通过将多视角观测信息编码到成本量中,该方法实现了传统方法的实时性和灵活性。采用几何数据增强技术对MVDepthNet进行训练。然后进一步将MVDepthNet应用于单目密集测绘系统,实现通过单个局部移动摄像头连续估计深度地图。

为了增强模型迁移学习的能力,在Eigen和Fergus[49]中,作者使用一个单一的多尺度卷积网络架构来处理三种不同的计算机视觉任务:深度预测、表面法线估计和语义标记,如图8。开发的网络只需稍加修改,就能自然地适应每个任务,直接从输入图像回归到输出地图。该方法使用一系列的尺度逐步细化预测,并捕获许多图像细节而不需要任何超像素或低水平分割。这种能够在多任务中迁移运用并且还能取得不错成绩的深度网络在未来会有很好的前景。Chen等人[50]研究了野外单幅图像的深度感知,即从无约束环境下拍摄的单幅图像中恢复深度。作者引入了一个新的数据集"野外深度",由野外图像组成,如图9,并在随机点对之间标注相对深度,还提出了一种新的算法,通过使用注释的相对深度学习估计度量深度。

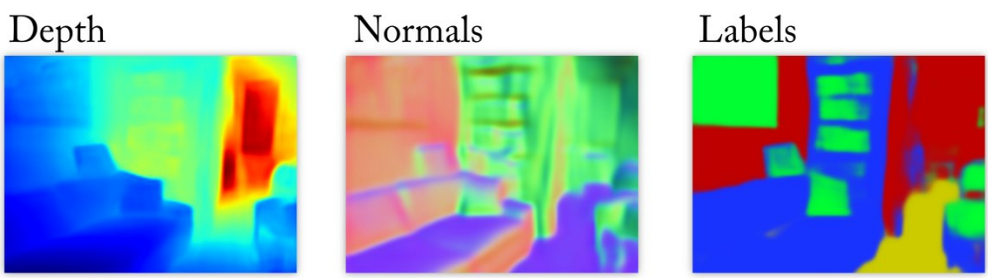


图8 3种任务的预测[49]

Fig. 8 Prediction of three tasks[49]

单目深度估计是一个不适定问题,对理解三维场景几何结构起着至关重要的作用。最近的方法通过从深度卷积神经网络(Deep Convolutional Neural Network, DCNN)中挖掘图像级信息和层次特征,取得了显著的改进,这些方法将深度估计建模为一个回归问题,并通过最小化均

图9 各种数据集(采用的为最右边的数据集)[50]

Fig. 9 Various data sets (the one used is the rightmost data set)[50]

方误差来训练回归网络,收敛速度慢,局部解不理想。此外,现有的深度估计网络采用重复的空间池操作,导致不需要的低分辨率特征图。为了获得高分辨率的深度图,需要跳转连接或多层反褶积网络,这使得网络训练复杂,计算量大。为了消除或至少在很大程度上减少这些问题,Fu等人[21]引入了一种间距递增离散化(Spacing-Increasing Discretization, SID)策略,将深度离散化,并将深度网络学习作为一个有序回归问题进行重构。通过使用普通的回归损失训练网络,Fu等人[21]的方法获得了更高的精度和更快的同步收敛速度。此外,作者采用多尺度网络结构,避免了不必要的空间池,并行地获取多尺度信息。Guo等人[22]采用图形引擎生成的合成数据收集大量深度数据,使用立体匹配网络作为代理从合成数据中学习深度,并使用预测立体视差图来监控单目深度估计网络。Jiang等人[23]的工作中,从训练一个深度网络开始,使用全自动监控,从单个图像中预测相对场景深度。相对深度训练图像是从简单的视频中自动获得的,视频中的汽车通过一个场景,使用最新的运动分割技术,没有人为提供的标签。从一幅图像中预测相对深度的代理任务在网络中引入了一些特征,这些特征使得一组下游任务(包括语义分割、联合道路分割和车辆检测以及单目(绝对)深度估计)在从头开始训练的网络上有了很大的改进。在语义切分任务上的改进比其他任何自动监督方法都要大。此外,对于单目深度估计,该文章实现的无监督预训练方法甚至优于ImageNet的有监督预训练。Kendall等人[51]提出了一种新的深度学习架构,用于从校正后的立体图像对中回归视差。

有监督的深度学习往往缺乏足够的训练数据。特别是在单目深度图预测的背景下,在真实的动态室外环境中,几乎不可能确定密集的地面真实深度图像。例如,当使用激光雷达传感器时,距离测量中存在噪声,传感器之间的校准不可能完美,并且测量通常比摄像头图像稀疏得多。Kuznietsov等人[24]提出了一种基于半监督学习的单目图像深度图预测方法。当我们使用稀疏的地面真深度进行监督学习时,我们也使用直接的图像对齐损失在立体设置中执行我们的深度网络来产生光一致的密集深度图。Li和Snavely[52]建议使用多视角互联网照片集,这是一个几乎无限的数据源,通过现代结构-从运动和多视角立体(Multi-View Stereo, MVS)方法生成训练数据,并基于此想法提出一个名为MegaDepth的大型深度数据集。来自MVS的数据有它自己的挑战,包括噪音和不可重构的对象。作者通过新的数据清洗方法来解决这些挑战,以及使用语义分割生成的顺序深度关系来自动增强数据。作者验证了大量互联网数据的使用,通过显示在MegaDepth上训练的模型具有很强的泛化能力——不仅适用于新场景,而且适用于其他不同的数据集,包括Make3D、KITTI和DIW,即使在训练期间没有看到这些数据集的图像。在文献[53]中,作者关注于解决从单一图像中估计场景深度的问题。这是一个具有挑战性的任务,因为单个图像本身不提供任何深度提示。为了解决这个问题,作者利用已知深度的图像池。更具体地说,作者将单目深度估计表述为一个离散-连

续优化问题，其中连续变量编码输入图像中超像素的深度，离散变量表示相邻超像素之间的关系。然后利用粒子信念传播在图形模型中进行推理，得到这个离散-连续优化问题的解。

为了解决功耗高的GPU不允许在功耗有限的应用领域中推断深度映射的问题，在文献[28]中，作者提出了一种新的架构，利用从单个输入图像中提取的特征金字塔，能够在CPU上，甚至在嵌入式系统上，快速推断出准确的深度地图。类似于最先进的状态，作者以无监督的方式训练网络，将深度估计作为一个图像重建问题。Montiel等人[54]介绍了一种基于特征的单目SLAM系统，该系统可在大、小、室内和室外环境中实时运行。该系统对严重的运动杂波具有鲁棒性，允许较宽的基线环路闭合和重新定位，并包括完整的自动初始化。Zhan等人[29]探讨了使用立体序列学习深度和视觉里程测量。立体序列的使用使空间（左右对之间）和时间（前向后）光度偏差的使用成为可能，并限制场景深度和摄像头运动在一个共同的，现实世界的尺度。在测试时，该框架能够从单目序列估计单视点深度和双视点里程。Wu等人[55]提出了一种新的视角来解决单目图像的深度估计问题：尺寸到深度。该方法需要稀疏的关于真实世界尺寸物体的标签，而不是原始的深度。然后根据尺寸标签的几何关系推断出一个粗糙的深度映射。然后对条件随机场（CRF）进行能量函数优化，对深度图进行细化。

近年来，自监督单目深度估计技术的性能已接近监督方法，但仅适用于低分辨率。研究表明，高分辨率是实现高保真自监督单目深度预测的关键。受近期用于单幅图像超分辨率的深度学习方法的启发，Pillai等人[30]提出了一种用于深度超分辨率的亚像素卷积层扩展，该方法可以从相应的低分辨率卷积特征中精确地合成高分辨率差异。此外，作者引入了一个可微的翻转增强层，可以准确地融合来自图像及其水平翻转版本的预测，减少由于遮挡而产生的左右阴影区域的影响。在公共KITTI基准的自我监督深度和姿态估计方面，这两项贡献都提供了显著的性能提高。该方法预测效果如图10所示。Yang等人[56]介绍了一种用于无监督深度估计框架的表面法线表示方法。估算深度被限制为与预测法线兼容，从而产生更稳健的几何结果。具体来说，作者制定了一个边缘感知的深度-法线一致性项，并通过在深度卷积网络DCN内部构造一个深度-法线层和一个法线-深度层来解决此问题。

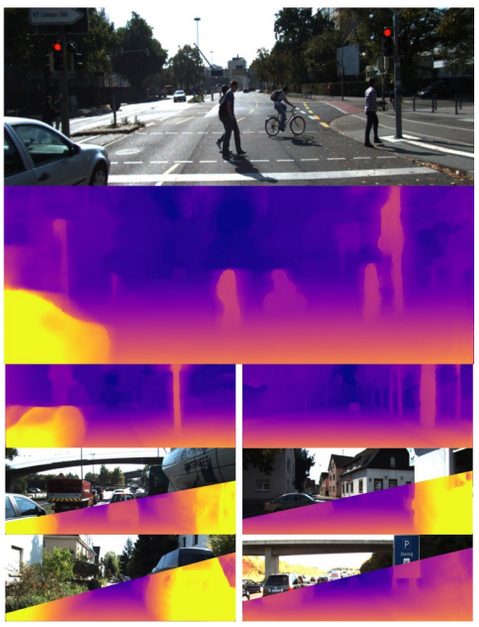


图10 深度效果图[30]

Fig. 10 Depth maps[30]

正如上文中提到的，单目视觉深度估计的难度在这3类中是最大的，但由于存在成本优势而被广泛研究。近年来，随着深度神经网络的迅速发展，基于深度学习的单目深度估计得到了广泛的研究，并取得了良好的精度。比如利用深度神经网络对单个图像进行端到端的稠密深度图估计。为了提高深度估计的精度，研究人员们提出了不同的网络结构、损失函数和训练策略，如单目自监督的monodepth2，在GAN范式下进行无监督深度估计的MonoGAN，基于深度注意的DABC网络等，这些工作推动了单目深度估计的快速发展。此外，有的作者还利用风格转换和对抗性训练，在对大量合成环境数据进行训练的基础上，从单个真实彩色图像中预测像素完美深度。Madhu等人[19]使用未标记的双目立体图像对训练，提出了一种基于深度网络的无监督视觉里程计系统，用于六自由度摄像头姿态估计和单目密集深度图的获取。此外，可以在单目图像的深度估计策略中增加真实世界物体尺寸的标签，建立尺寸标签与深度的映射关系，降低粗糙深度

预测的计算量。另外，也可以通过限制无监督深度估计与预测法线兼容来增加模型的鲁棒性。

3 视觉建图

SLAM框架类算法中，定位是主体，通常需要定位能实时响应，而稠密地图的构建通常规模和计算量都较大，因此地图的构建却不一定需要实时。当然，地图比较稀疏时，也能实时建图，但并不满足实际需求。另一种思路是以建图为主体，定位次之，旨在构建稠密准确的高质量地图，这种高质量地图可以提供给SLAM算法定位使用，因此基于深度视觉的建图便是满足上述需求的一种方法。

视觉建图需要与视觉里程计(VO)，回环检测，后端非线性优化配合以形成精确的建图。Schneider等人[1]提出了maplab，一个开放的、面向研究的视觉惯性映射框架，用于处理和操作多会话映射。一方面，maplab可以看作是一种随时可用的视觉惯性测绘定位系统。另一方面，maplab为研究社区提供了一组多会话映射工具，包括映射合并、视觉惯性批优化和环路闭合。此外，它还包括一个在线前端，可以创建视觉-惯性地图，并在定位地图中跟踪一个全局无漂移姿态。下面将按照图11的模块进行展开。

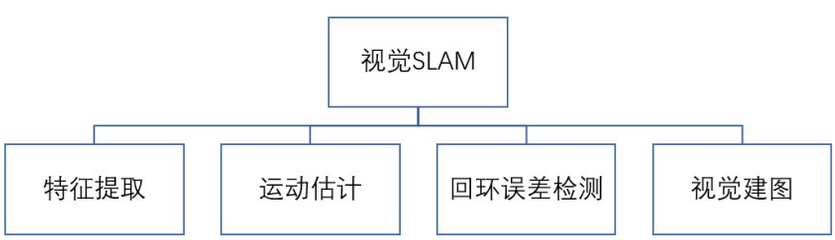


图11 SLAM模块

Fig. 11 SLAM architecture

在特征提取方面，Konolige等人[2]用大量的点特征匹配视觉帧，使用来自计算视觉的经典束调整技术，但只保留相对的帧姿态信息(骨架)，如下图12所示。Blake等人[57]文章中的S3GP混合使用了不同的图像特征用以提高映射的准确性和一致性。Qin等人[58]利用鲁棒语义特征来构建停车场的地图和车辆定位，如图13所示，右下角的图是地下停车场常见的场景，环境狭窄，光线不足，没有GPS定位。在这样的环境下自动驾驶具有挑战性。较大的图形是该停车场的语义视觉地图，由语义特征(引导标志、停车线、减速带)组成。这张地图可用于以厘米级精度对车辆进行定位。语义特征包括路标、停车线、减速带等，这些特征通常出现在停车场。与传统特征相比，这些语义特征对透视和光照变化具有长期的稳定性和鲁棒性。Xavier等人[3-4]提出了用人工标记特征实现SLAM的方法。

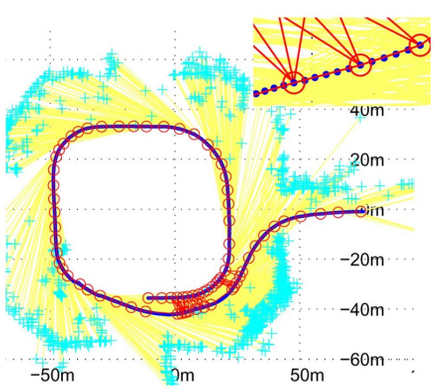


图12 一个100 m城市场景的骨架缩小。完整贝叶斯图是700帧和约100 K的特征[2]

Fig. 12 Scaled-down map of a 100 m city. The complete Bayesian diagram is 700 frames and about 100 K[2]

在运动估计上，Fernandez等人[4]通过一组分布在整个环境中的智能标记，如图14所示，所提出的映射方法根据一组校准图像和PMS单元收集的方向/距离测量数据来估计标记的姿态。在此基础上，本文所提出的定位方法可以对具有正确比例尺的单目摄像头进行定位，直接得益于该方法提高的定位精度。Schneider等人[1]中的maplab包含一个在线前端，可以创建视觉-惯性地

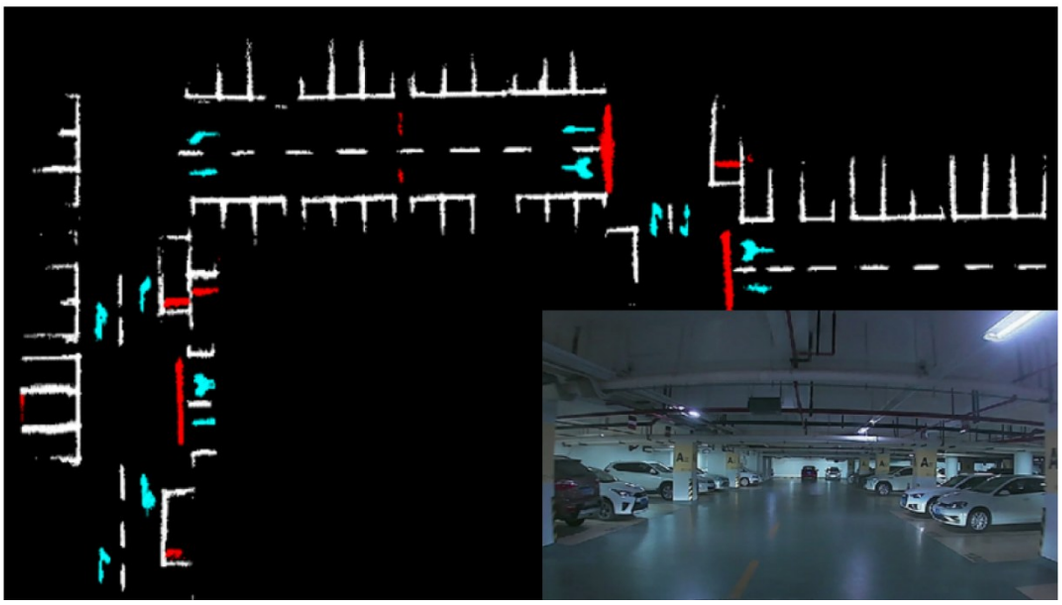
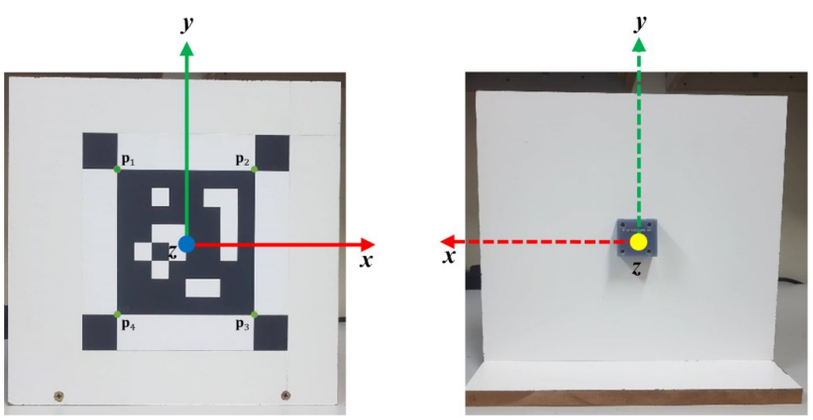


图13 语义视觉地图[58]

Fig. 13 Semantic visual map[58]



(a) (b)

图14 智能标记:前面一个正方形平面基准标记(a)以及一个嵌入式姿态测量系统(PMS)单元,其坐标系统在后面(b)[4]。

Fig. 14 Smart marker: a square plane fiducial mark in the front (a) and an embedded attitude measurement system (PMS) unit with its coordinate system in the back (b)[4].

图,并在定位地图中跟踪一个全局无漂移姿态。而Saeedi等人[59]通过开发新的度量不依赖任何SLAM或运动估计算法的情况下评估轨迹和环境。

回环误差检测方面,Usenko等人[60]提出了利用非线性因子恢复从视觉惯性里程测量中提取相关信息来进行视觉惯性映射。该方法重建一组非线性因素,使VIO积累的轨迹上的信息的最佳近似。为了获得全局一致的映射,我们使用bundle调整将这些因素与循环闭合约束结合起来。VIO因子使全局映射的横倾角和俯仰角可以观测到,提高了映射的鲁棒性和精度。Xiao等人[61]在跟踪线程中通过选择性跟踪算法对动态目标的特征点进行处理,显著减少了由于不正确匹配而导致的姿态估计误差。

在建图方面,正如激光建图技术是利用激光扫描的帧对帧匹配来生成详细的局部映射以及大回路的闭合,视觉建图也可以采用同样的策略。Qin等人[58]采用4个全景摄像头来增加感知范围。该系统在惯性测量单元和车轮编码器的

辅助下，生成全局视觉语义图。Hong和Kim[62]的主要思想是模型适度弯曲船体表面的结合分段平面面板，并通过调整在一个二维坐标系的局部图像生成一个全局地图，并作出适当的纠正以反映角度的信息预测的3D板，如图15所示。

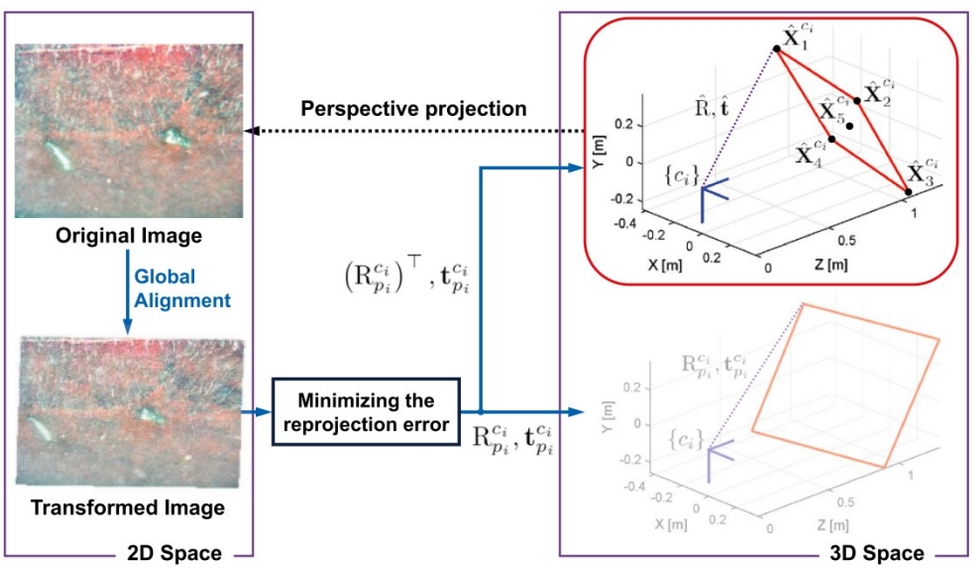


图15 3D面板的姿态估计示例。实际面板的方向可以通过对旋转矩阵Rci进行逆(转置)来估计[62]

Fig. 15 An example of pose estimation for a 3D panel. The direction of the actual panel can be estimated by inverting (transposing) the rotation matrix Rci[62]

4 结束语

本文从视觉SLAM的两项核心技术——视觉深度的预测以及视觉建图技术入手进行研究分析。视觉深度预测部分包含了视觉数据的采集方式、算法的监督设计展开，按照视觉数据的采集视觉技术分为单目、双目以及多目，按照算法的设计分为，全监督、半监督和无(自)监督方法。视觉建图部分包含了特征提取，运动估计，回环检测和建图等方面的最新方法综述。通过对以上方法的总结分析可以发现：

在视觉深度感知方面，双目和多目深度估计可以实现较单目更高的精度，然而出于对降低硬件配置的成本及复杂性的需求，学界对于单目深度估计方法也开展了大量研究，而以monodepth2为代表的最新单目深度估计算法也在数据集上体现了较好地预测精度，并且通过时序的前后帧实现训练过程的自监督。

未来的视觉深度感知策略仍然需要在硬件配置、算力需求和预测精度间寻求最优，而单目双目融合估计可能是发展方向之一，因为该方法在成本和算力方面均有潜在优势，并且可以同时实现对静态和动态目标的三维重建(因为单目对于动态目标的深度估计具有局限性)。此外，视觉语义建图由于可以提供更高层的语义特征，因而可以提供更鲁棒的特征用来解决SLAM建图中的匹配以及闭环问题，也可以用来改善定位精度，因此是视觉建图的发展方向。但由于语义分割算法本身对算力提出了较高的要求，该方法需要与性能优越的处理终端配合使用。

在视觉建图方面，特征提取是核心环节之一，按照计算量有小到大、精度由低到高可以分为点特征、图像特征和语义特征匹配，因此需要根据对计算量和精度的要求选择合适的特征匹配策略。而视觉建图的另一项核心技术是运动估计，该环节既可以通过视觉帧匹配本身来完成，也可以通过视觉融合惯性测量单元和车轮编码器共同完成，而后者的由于提高了定位精度，可以生成更精准的三维地图。

参考文献：

[1] Schneider T. Maplab：An Open Framework for Research in Visual-inertial Mapping and Localization [J]. *IEEE Robotics & Automation Letters*, 2018. 3(3):1418-1425.

[3] Konolige K, Agrawal M. FrameSLAM：From Bundle Adjustment to Real-Time Visual Mapping [J]. *IEEE Transactions on Robotics*, 2008. 24(5):1066-1077.

[4] Xavier R S, da SilvaB M, GoncalvesL M. Accuracy analysis of augmented reality markers for visual mapping and localization [C]// 2017 Workshop of Computer Vision (WVC). 2017, IEEE.

[5] Ortiz-Fernandez L E. Smart Artificial Markers for Accurate Visual Mapping and Localization [J]. *Sensors*, 2021. 21(2)：625.

[6] Chai Y, Cao X. Stereo Matching Algorithm Based on Joint Matching Cost and Adaptive Window [C]// 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). 2018.

[7] Guo X. Deep Eyes：Binocular Depth-from-Focus on Focal Stack Pairs. 2019.

[8] Song X, Zhao X, Lin T. Face Spoofing Detection by Fusing Binocular Depth and Spatial Pyramid Coding Micro-Texture Features [C]// 2017 IEEE International Conference on Image Processing (ICIP). 2017.

[9] Pilzer A. Progressive fusion for unsupervised binocular depth estimation using cycled networks [J]. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 2019. 42(10)：2380-2395.

[J] Xu K, Chen Z, Jia F, Unsupervised binocular depth prediction network for laparoscopic surgery [J]. *Computer Assisted Surgery*, 2019:1-6.

[10] Sinha A. DELTAS：Depth Estimation by Learning Triangulation and densification of Sparse points. 2020.

[11] Yao Y. Mvsnet：Depth inference for unstructured multi-view stereo [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[12] Kusupati U. Normal Assisted Stereo Depth Estimation. 2019.

[13] Wang K, Shen S. Mvdepthnet：Real-time multiview depth estimation neural network [C]// 2018 International conference on 3d vision (3DV). 2018. IEEE.

[14] Hou Y, Solin A, Kannala J. Unstructured Multi-View Depth Estimation Using Mask-Based Multiplane Representation [C]// 2019：Springer, Cham.

[15] Long X. Multi-view Depth Estimation using Epipolar Spatio-Temporal Networks. 2020.

[16] Godard C. Digging Into Self-Supervised Monocular Depth Estimation [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2020.

[17] Aleotti F. Generative Adversarial Networks for Unsupervised Monocular Depth Prediction [C]// Proceedings, Part I. Munich, Germany, 2019.

[18] Atapour-Abarghouei A. Real-time monocular depth estimation using synthetic data with domain adaptation [C]// IEEE/CVF Conference on Computer Vision & Pattern Recognition. 2018.

[19] Madhu B V. A Deeper Insight into the UnDEMoN：Unsupervised Deep Network for Depth and Ego-Motion Estimation [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018.

[20] Casser V. Depth Prediction without the Sensors：Leveraging Structure for Unsupervised Learning from Monocular Videos [C]// Thirty-Third AAAI Conference on Artificial Intelligence (AAAI′

[20] 2019.

[21] Fu H. Deep ordinal regression network for monocular depth estimation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[22] Guo X. Learning monocular depth by distilling cross-domain stereo networks [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[23] Jiang H. Self-supervised relative depth learning for urban scene understanding [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[24] Kuznietsov Y, Stückler J, Leibe B. Semi-Supervised Deep Learning for Monocular Depth Map Prediction [C]// IEEE Conference on Computer Vision & Pattern Recognition. 2017.

[25] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[26] Li R. Deep attention-based classification network for robust depth prediction [C]// Asian Conference on Computer Vision. 2018. Springer.

[27] Ranjan A. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[28] Poggi M. Towards Real-Time Unsupervised Monocular Depth Estimation on CPU [C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018.

[29] Zhan H. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[30] Pillai S, Ambrus R, Gaidon A. SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation [C]// 2019 International Conference on Robotics and Automation (ICRA). 2019.

[31] 程明明, 王. 贺, 安. 平, 基于特征点匹配的多视图像校正[J]. 液晶与显示, 2010. 25(4):593-597.
Cheng Mingming, Wang. He, An. Ping, Multi-view image correction based on feature point matching [J]. *Chin*, J. Liq. *Cryst. Displays*, 2010. 25(4):593-597. (in Chinese)

[32] 张建业,朴燕, 基于改进稳态匹配概率的立体匹配算法研究[J]. 液晶与显示, 2018. 33(4):357-364.
Zhang Jianye, Park Yan, Research on stereo matching algorithm based on improved steady-state matching probability [J]. *Chin*, *J. Liq. Cryst. Displays*, 2018. 33(4):357-364. (in Chinese)

[33] Huimin Yin. A High Spatial Resolution Depth Sensing Method Based on Binocular Structured Light [J]. *Sensors*, 2017. 17(4): 805-811.

[34] Ende W. Stereo matching algorithm based on the combination of matching costs [C]// 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). 2017.

[35] 汪伟, 基于深度学习的双目立体视觉研究[D]. 合肥:合肥工业大学,2020,
Wang Wei, Research on Binocular Stereo Vision Based on Deep Learning[D] Hefei:Hefei University of Technology, 2020,(in Chinese)

[36] Anantrasirichai N. Dynamic Programming for Multi-View Disparity/Depth Estimation [C]// 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. 2006.

[37] Montserrat T. Depth estimation based on multiview matching with depth/color segmentation and memory efficient belief propagation [C]// Image Processing (ICIP), 2009 16th IEEE International Conference on. 2009.

[38] Liu Y. Continuous depth estimation for multi-view stereo [C]// IEEE Conference on Computer Vision & Pattern Recognition. 2009.

[39] Lee S B, Ho Y S. View-consistent multi-view depth estimation for three-dimensional video generation [C]// 3dtv-conference: the True Vision-capture. 2010.

[40] Mieloch D. Graph-based multiview depth estimation using segmentation [C]// IEEE International Conference on Multimedia & Expo. 2017.

[41] Facil J M. Single-View and Multi-View Depth Fusion. *IEEE Robotics & Automation Letters*, 2017. 2(4):1994-2001.

[42] Yang M, Cao, X, Dai Q. Multiview video depth estimation with spatial-temporal consistency [C]// British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, 2010.

[43] Ince S. Depth Estimation for View Synthesis in Multiview Video Coding [C]// 3dtv Conference. 2007.

[44] Long X. Occlusion-Aware Depth Estimation with Adaptive Normal Constraints. 2020.

[45] Strecha C. Gool L V. PDE-based Multi-view Depth Estimation [C]// International Symposium on 3d Data Processing Visualization & Transmission. 2002.

[46] Garg R. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue [C]// European Conference on Computer Vision. 2016.

[47] Godard C, Mac Aodha O, Brostow G J. Unsupervised Monocular Depth Estimation with Left-Right Consistency

[C]// Computer Vision & Pattern Recognition. 2017.
[48] Zhou T. Unsupervised Learning of Depth and Ego-Motion from Video. *in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
[49] Eigen D, Fergus R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture [C]// IEEE International Conference on Computer Vision. 2015.
[50] Chen W. Single-image depth perception in the wild [J]. *arXiv preprint arXiv*:*1604*.03901, 2016.
[51] Kendall A. End-to-end learning of geometry and context for deep stereo regression [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017.
[52] Li Z, Snavely N. Megadepth: Learning single-view depth prediction from internet photos [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[53] Liu M, Salzmann M, He X. Discrete-Continuous Depth Estimation from a Single Image [C]// IEEE Conference on Computer Vision & Pattern Recognition. 2014.
[54] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. *IEEE Transactions on Robotics*, 2015. 31(5):1147-1163.
[55] Wu Y, Ying S, Zheng L, Size-to-depth: a new perspective for single image depth estimation [J]. *arXiv preprint arXiv*:1801.04461, 2018.
[56] Yang Z. Unsupervised learning of geometry with edge-aware depth-normal consistency [J]. *arXiv preprint arXiv*: *1711*.03665, 2017.
[57] Williams O, Blake A, Cipolla R. Sparse and Semi-supervised Visual Mapping with the S^3GP [C]// 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2006.
[58] Qin T. AVP-SLAM: Semantic Visual Mapping and Localization for Autonomous Vehicles in the Parking Lot [J]. *arXiv preprint arXiv*:2007.01813, 2020.
[59] Saeedi S. Characterizing visual localization and mapping datasets [C]// 2019 International Conference on Robotics and Automation (ICRA). 2019. IEEE.
[60] Usenko V. Visual-Inertial Mapping With Non-Linear Factor Recovery [J]. *IEEE Robotics and Automation Letters*, 2020. 5(2):422-429.
[61] Xiao L. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment [J]. *Robotics & Autonomous Systems*, 2019.
[62] Hong S. Kim J, Three-Dimensional Visual Mapping of Underwater Ship Hull Surface using View-based Piecewise-Planar Measurements [J]. *IFAC—PapersOnLine*, 2019. 52(21): 384-389.

作者简介：

陈苑锋(1979—),男,福建安溪人,博士,工程师,2006年于复旦大学获得博士学位,主要从事机器人方向的研究。
E—mail: yuanfeng_chen@hotmail.com