液晶与显示
Chinese Journal of Liquid Crystals and Displays
ESCI/Scopus收录
中文核心期刊
主管：中国科学院
主办：中国科学院长春光学精密机械与物理研究所
京东方科技集团股份有限公司
中国物理学会液晶分会
中国光学光电子行业协会液晶分会
主编：郭海成

视觉深度估计与点云建图研究进展

陈苑锋

引用本文:
陈苑锋. 视觉深度估计与点云建图研究进展[J]. 液晶与显示, 2021, 36(6): 896-911.
CHEN Yuan-feng. Progress of visual depth estimation and point cloud mapping[J]. *Chinese Journal of Liquid Crystals and Displays*, 2021, 36(6): 896-911.

在线阅读 View online: https://doi.org/10.37188/CJLCD.2020-0047

您可能感兴趣的其他文章

Articles you may be interested in

基于笔划三维深度特征的签名识别

Signatures recognition based on strokes 3D depth feature

液晶与显示. 2019, 34(10): 1013-1020 https://doi.org/10.3788/YJYXS20193410.1013

基于电磁追踪系统的点云配准方法研究

Point cloud registration method based on electro-magnetic tracking system

液晶与显示. 2020, 35(12): 1309-1314 https://doi.org/10.37188/YJYXS20203512.1309

基于深度学习的木材表面缺陷图像检测

Image detection of wood surface defects based on deep learning

液晶与显示. 2019, 34(9): 879-887 https://doi.org/10.3788/YJYXS20193409.0879

纯相位全息图优化算法

Review of computer-generated phase-only hologram optimization algorithm

液晶与显示. 2021, 36(6): 810-826 https://doi.org/10.37188/CJLCD.2021-0035

三维枪弹痕点云数据处理及特征提取研究

Processing and feature extraction for three-dimensional bullet point cloud data

液晶与显示. 2016, 31(9): 889-896 https://doi.org/10.3788/YJYXS20163109.0889

文章编号:1007-2780(2021)06-0896-16

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。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。视觉传感器可以有效避免上述问题,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、数据集的丰富和新机器视觉算法的出现,视觉深度预测和建图算法的精度和效率都有了明显提升。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 **文献标识码**:A **doi**:10.37188/CJLCD.2020-0047

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 used in SLAM technology presents 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 progress in computing hardware, the continuous expansion of data sets, as well as the introduction of new computer vision algorithms, the accuracy and efficiency of visual depth prediction and mapping algorithms have been significantly improved. This article summarizes the existing methods of visual depth prediction and point cloud mapping and classifies the methodology based on 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

收稿日期:2020-02-27;**修订日期**:2021-03-29.

* 通信联系人,E-mail:yuanfeng_chen@hotmail.com

1 引　　言

同步定位与地图构建(Simultaneous Localization and Mapping,SLAM)主要用于实现移动机器人在未知环境中运行时定位导航与地图构建功能,通常包括特征提取、数据关联、状态估计、状态更新以及特征更新等。一般分为 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。SLAM 技术依赖于激光雷达传感器,因其可提供高精度的 3D 点云信息。3D 激光 SLAM 的帧间匹配方法包括以下 3 种:点云配准算法、Point-to-Plane ICP、Feature-based Method。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。

因激光雷达价格昂贵,影响了其市场化的进展。加上受制于线数,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,无法获得精确的语义信息。相较之下,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 涉及视觉深度估计和视觉建图两项核心技术。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,其方法逻辑与激光雷达点云建图类似,技术较为成熟,且难度可控[3-4]。视觉深度估计则比激光雷达深度测量在测量精度方面面临着较大挑战,成为研究界的热门方向。

2 视觉深度估计

基于单目、双目和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。以常用的几种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难以良好匹配[5]。基于 RGBD 的深度估计量程较短、价格不菲,在实际应用中能力有限,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。当然,现有的单目摄像头里,有监督方案和无监督方案均面临着巨大的挑战。有监督方案需要大量的深度测量数据,这些数据通常很难获得,而无监督方案在估计精度上受到限制。

表 1 对业界视觉深度估计方法进行了汇总,从摄像头类型、计算模型(以深度学习模型为主)、所采用的数据集名称、数据量、深度学习模型监督类型和发布时间等方面进行了对比。从摄像头类型角度看,近年来更多的研究集中于单目摄像头的深度估计,主要原因是一方面单目摄像头在硬件布置和成本上具有优势,另一方面神经网络加速芯片的性能提升进一步推动了单目算法的进展。本文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含了有监督、半监督、自监督和无监督,所列的文献主要发表于 2017～2020 年间,是对近年来最新方法的总结。

表 1　视觉深度估计方案表

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

续 表

摄像头类型	模型名称	数据集名称	数据量	监督类型	年份
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	27116(t)2979(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:t 代表 training,v 代表 validation;CS 表示 cityscapes 数据集,FIN 表示室内导航数据集。

2.1 双目视觉深度估计

双目深度估计,又称视差估计(Disparity Estimation)[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,输出是通过摄像头焦距 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等参数计算出的目标深度图 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,通过深度和视差的相互转换关系来得到结果。双目摄像头的左右眼原始图像以及双目匹配获得的深度图如图 1 所示。

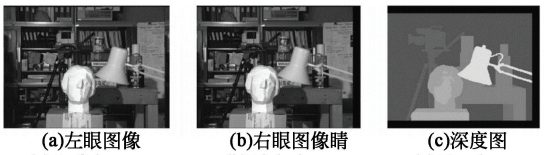


图 1 双目摄像头的原始数据,以及通过左右眼的视差计算出的深度图[5]。

Fig.1 Raw images of the binocular camera, as well as the depth map calculated through the difference between the left and right images[5].

立体匹配是深度估计中的基本模块,通过获得左右图片中像素的对应关系计算出视差图。过去几十年间,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，在一个类似于 Kinect 的激光投影机的两侧安装两个红外摄像头，采用双目模式和单眼模式两种匹配模式，可在不影响深度图像质量的前提下获得更高精度的视差图。

深度学习算法在立体匹配领域也有进展。立体匹配的深度学习方案将传统立体匹配方法的 4 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]。大多数在 KITTI 数据集上排名靠前的方法均基于深度学习[35]，例如 Song 等人[7]在算法上对双目深度预测方法做了提升，提出了两种新抗欺骗干扰的鲁棒性特征：基于双目深度的模板人脸匹配特征和基于空间金字塔编码的高阶微纹理特征，配以新的模板人脸配准算法和空间金字塔编码算法，实现了多模态人脸欺骗检测。

在计算机视觉中，双目立体信息和单眼聚焦线索通常是分开解决的。但 Guo 等人[6]同时使用这两种类型的线索进行深度推断，构建了 3 个独立的网络：Focus-Net 用于单个焦堆栈中提取深度，EDoF-net 用于焦堆栈中获得扩展景深图像，stereo-net 用于立体匹配，最后集成到统一 BDfF-Net 中以获得高质量的深度图。

使用监督回归的深度学习算法在视觉处理中能取得显著的效果，但监督学习需要为算法训练进行昂贵的真值(Ground truth)注释。为解决这一问题，Pilzer 等人[8]专门为双目立体深度估计设计了一个新的计算框架——渐进融合网络(Progressive Fusion Network，PFN)，该网络结合了双目摄像头采集的两个立体视图，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。该架构定性比较结果如图 2 所示。作为无监督学习方案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，让外科医生获得了深度感知，从而克服了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。

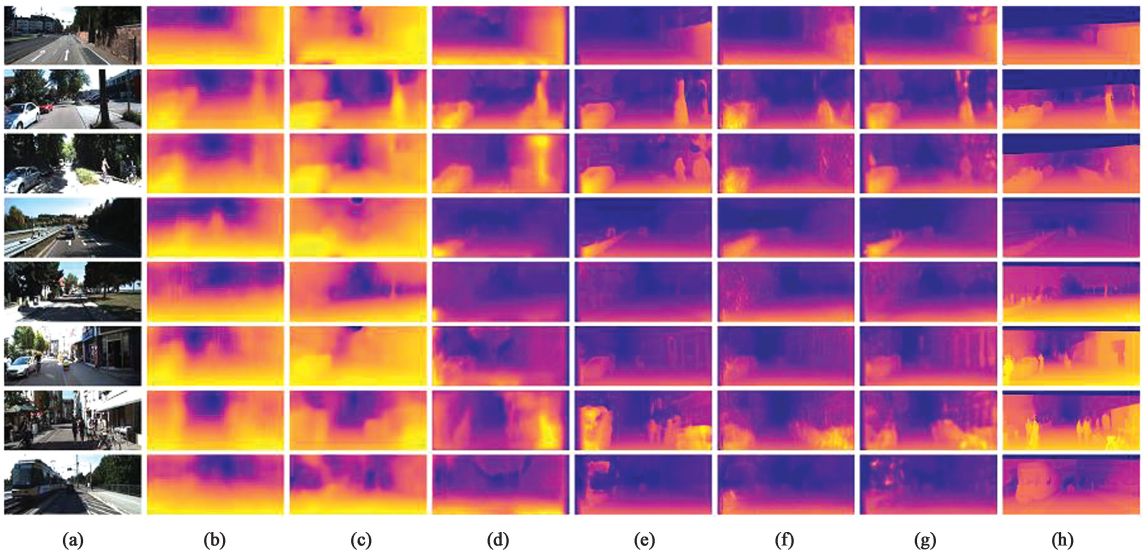


图 2　各种方法定性比较[8]。(a) RGB 图；(b) Eigen 等人的方法；(c)Zhou 等人的方法；(d)Garg 等人的方法；(e)Godard 等人的方法；(f)Pilzer 等人的方法；(g)PFN；(h)深度图真值。

Fig.2　Qualitative comparison of various methods[8]. (a) RGB images; (b) Eigen *et al.*; (c) Zhou *et al.*; (d) Garg *et al.*; (e) Godard *et al.*; (f) Pilzer *et al.*; (g) PFN; (h) Ground truth.

通过上述代表性文献可以发现，利用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，通过双目立体线索和单眼聚焦线索获得有效的三维感知、采用渐进融合网络(PFN)与对抗性学习共同训练也是值得探索的研究方向。

2.2　多目视觉深度预测

除双目视觉外，学术界对多目视觉深度预测方法也开展了一系列研究。基于多视角图像的视差/深度估计算法，Anantrasirichai 等人[36]提出使用窗口相关的动态规划方法和新的代价函数，以视差/深度映射的平滑性和窗口的相关性为约

束，得到适合多视图图像的深度图。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，将轮廓信息和外极约束集成到连续深度图的变分方法中，通过多起始尺度框架生成多个深度候选对象，实现了路径归一化互相关度量合成到每个视图的精细化深度图。

因上述方法并不适用于任意视角，Lee 和 Ho 等[39]提出了一种基于视点一致性的多视点深度估计算法，使用传统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，采用多视点图割算法使误差最小化。Mieloch 等人[40]提出了一种适用于任意摄像头位置多视点系统的深度估计方法，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使用基于分割估计和与生成空间一致的多视图深度图的方法，如图 3 所示。

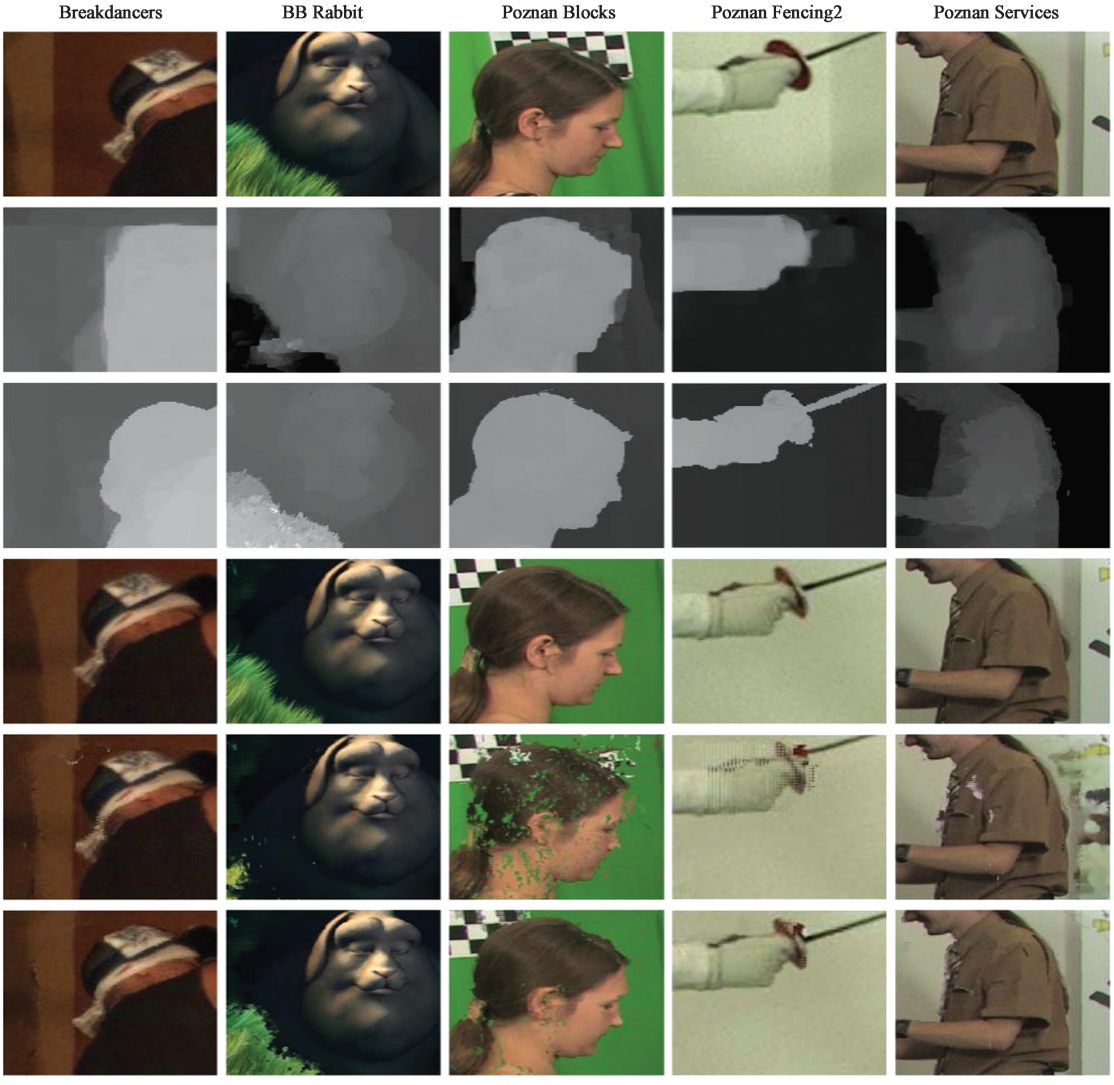


图 3 深度图与虚拟视点合成的比较[40]。(a)深度估计中使用的原始视图的片段；(b)用 DERS 估算的深度图片段；(c)使用建议的估计深度图片段的方法；(d)原始视图的片段(综合的参考视图)；(e)用 DERS 估计的深度图合在的视图片段；(f)用建议的方法估计的深度图合成的图片段。

Fig.3 Comparison of depth map and virtual viewpoint synthesis[40].(a) Fragment of the original view used in the depth estimation; (b) Fragment of the depth map estimated with DERS; (c) Fragment of the depth map estimated using the proposed method; (d) Fragment of the original view (the reference view for the synthesis); (e) Fragment of the view synthesized with depth maps estimated with DERS; (f) Fragment of the view synthesized with depth maps estimated using the proposed method.

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。Facil 等人[41]进一步利用了基于 CNN 的单视图深度估计与多视图深度估计进行了融合。

另一方面,虽然之前的基于学习的方法已经有了令人信服的结果,但大多数方法都是独立地估计单个视频帧的深度图,而没有考虑帧间强烈的几何和时间一致性。而且,目前最先进的(SOTA)模型大多采用全 3D 卷积网络,需要较高的计算成本,从而限制了其在现实应用中的部署。Long 等人[15]通过使用一个新的极时空变压器来实现时间相干深度估计,明确地关联几何和时间相关性,取得不错的深度估计结果。Yang 等人[42]则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,将左右视图匹配和基于颜色的分割相结合对深度图进行初始化,并将色彩一致性和空间一致性引入优化框架,以保证单一时刻的空间一致性。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的效率和精度,Ince 等人[43]考虑了多视点视频编码中视点合成的深度估计,可以有效地进行视图综合预测和生成编码比特数更少的深度图。Kusupati 等人[12]利用正态估计模型和预测的法线图提高了深度质量。Hou 等人[14]提出了一种求解非结构化多视角图像位姿以实现深度估计的新方法—MaskMVS,在平面扫描过程中,通过直方图匹配对深度平面进行采样,确保覆盖感的深度范围。Sinha 等人[10]提出了一种有效的步骤用于深度估计方法:(a)检测和评估兴趣点的描述符;(b)学习匹配和三角化一小组兴趣点;(c)使用 CNN 致密化这一稀疏的 3D 点集。他们采用端到端网络在深度学习框架执行所有这 3 个步骤,并通过中间 2D 图像和 3D 几何监督以及深度监督进行训练,训练结果如图 4 所示。

Long 等人[44]通过引入了联合法向图(CNM)约束来保持高曲率特征和全局平面区域,将多个相邻视图初始深度预测聚合到最终深度图和当前参考视图的遮挡概率图中,提高了深度估计的精度。Strecha 和 Gool[45]研究了一种对多幅校正图像进行深度提取的方法,通过系统对来自不同视图数据的相对置信度分配不同权重,在匹配过程中对不同权重的视图数据进行融合,从而取得了较高的深度信息精度。

综上所述,多目视觉深度预测主要有以下几个方向:(1)轮廓信息和外极约束集在连续深度预测;(2)任意摄像头位置的多视点系统的深度估计;(3)多视点视频编码中视点合成的深度估计。另外,为提升深度预测准确率,可以参考以下几种方法:(1)端到端的深度学习架构 MVSNet;(2)利用正态估计模型和预测的法线图来提高深度质量;(3)使用单个局部移动摄像头连续估计深度地图;(4)将单视图深度估计与多视图深度估计融合。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进展;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。这些方法为后续视觉深度估计方案创新提供了建设性的思路。

2.3　单目视觉深度预测

通过彩色图像生成高质量深度图的研究有望以较低的成本实现深度建图,通过使用大量未标注数据集求解深度,可为下游具有识别任务的深度神经网络实现预训练的目的,但具有精确标签的训练数据集本身就是一个巨大的挑战,故本节仅针对自监督和无监督方案进行分析探讨。

目前已有自监督的方法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。这就需要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。Godard[16]采用单目自监督的 Monodepth2 模型对每个像素的深度进行学习,使用损失小的重投影设计来处理遮挡,采用多尺度的采样方法以及忽略明显异常的训练像素,在 KITTI 数据集中实现了高精度的深度估计,如图 5 所示。

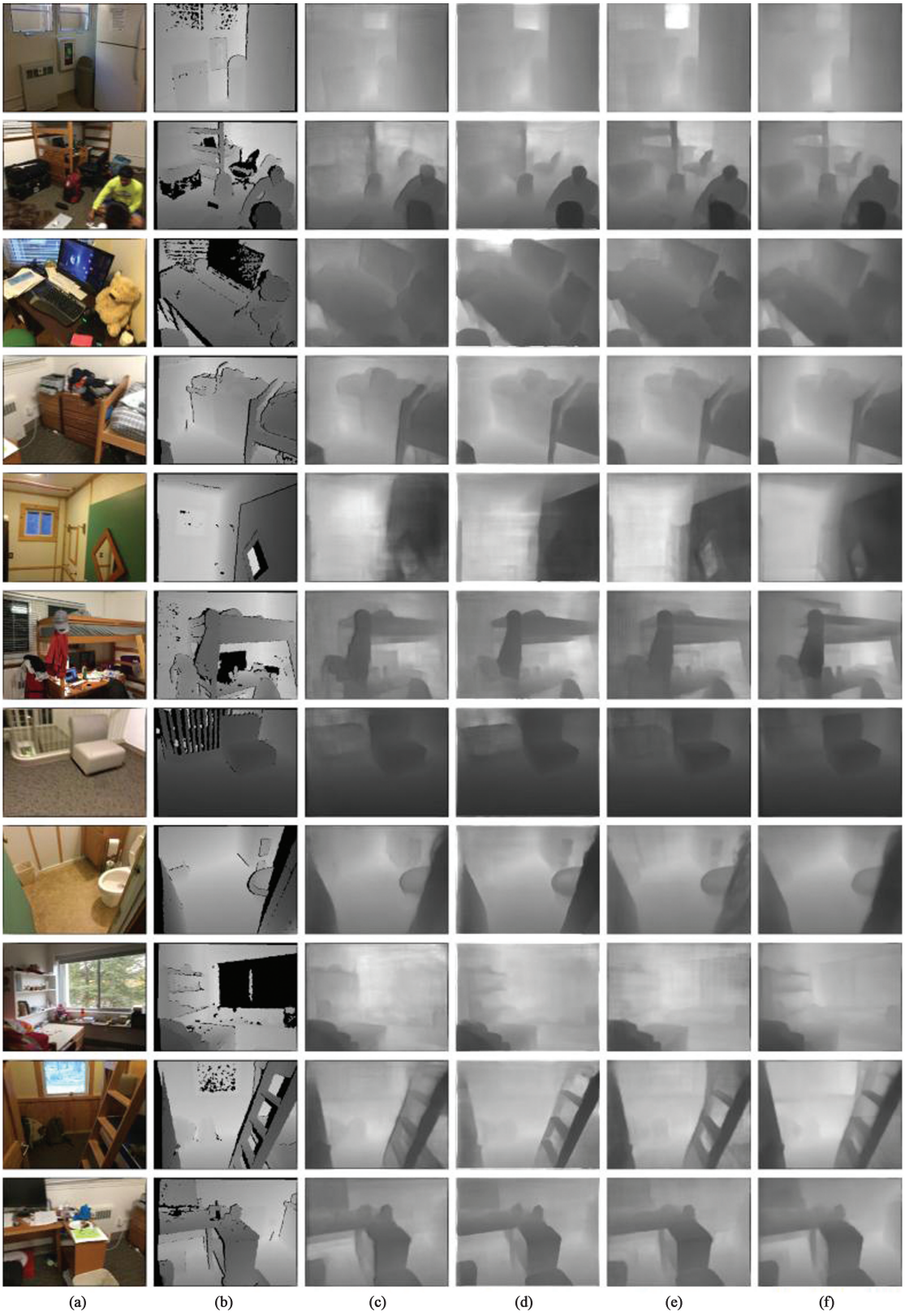


图 4 深度预测定性表现[10]。(a)图像;(b)深度图真值;(c)MVDepthNet;(d)GPMVSNet;(e)DPSNet;(f)DELAS。

Fig.4 Qualitative performance of depth prediction[10]. (a)Image; (b) Ground truth; (c) MVDepthNet; (d) GPMVSNet; (e)DPSNet; (f) DELAS.

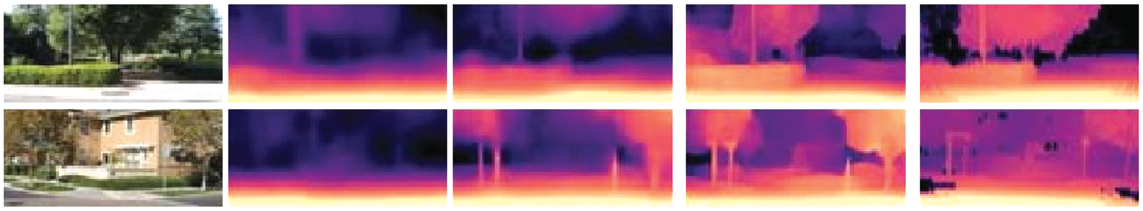


图 5　Make3D 定性结果(基于 KITTI)[16]。(a)输入;(b)Zhou 等;(c)DDVO;(d)Monodepth2(M);(e)深度图真值。

Fig.5　Qualitative results of Make3D (based on KITTI)[16].(a) Input; (b) Zhou *et al.*; (c) DDVO; (d) Monodepth2(M); (e) Ground truth.

Aleotti 等人[17]提出在 GAN 范式下进行无监督单目深度估计,通过生成器网络从参考图像推断深度以生成目标图像,训练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,预测效果如图 6 所示。

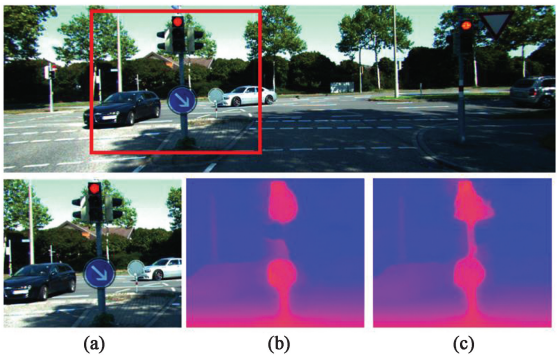


图 6　GAN 架构与他人论文预测结果对比。(a)原始图片;(b)由 Godard 等预测的深度图;(c)由作者的 GAN 架构预测的深度图[17]。

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

Amir 等人[18]利用风格转换和对抗性训练,实现了技术的优化。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,引入了一种在线求精方法来适应未知领域的动态学习,结果如图 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,不仅强化 3D 点云和连续帧自我运动的一致性,而且采用了有效掩蔽,在 KITTI 数据集和在未校准的手机摄像头上验证了深度和自我运动。Li 等人[26]提出了分类(DABC)网络,将深度预测作为一个多类分类任务,应用 Softmax 分类器对每个像素的深度标签进行分类,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,并通过赋予权重进行预测,可显著提高单一图像深度预测的鲁棒性。低层次视觉中有几个相互关联问题的无监督学习: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。Ranjan 等人[27]引入了竞争协作框架,即采用多个专门神经网络协调训练以解决复杂问题,其中神经网络既扮演着静态或移动区域对应像素的竞争对手角色,也扮演着将像素分配为静态或独立移动的协作者的角色。Wang 等人[13]通过引入 MVDepthNet 卷积网络,采用几何数据增强技术,多视图被编码后与参考图像结合,解决了局部单目摄像头在相邻视点的图像对下的深度估计问题,提高了实时性和灵活性。

Eigen 和 Fergus[49]使用单一的多尺度卷积网络架构来处理 3 种不同的视觉任务:深度预测、表面法线估计和语义标记,如图 8 所示。深度网络不仅可适应不同任务,且使用一系列尺度逐步细化预测,能捕获许多图像细节信息而不需要任何超像素或低水平分割,在未来有很好的应用前景。

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,可通过使用注释的相对深度估计度量深度的方法来得到预测结果,见图 9。

单目深度估计对理解三维场景几何结构起着至关重要的作用。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,获得了更高的精度和更

快的同步收敛速度。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，使用立体匹配网络从合成数据中学习深度，并预测立体视差图来监控单目深度估计网络。Jiang 等人[23]

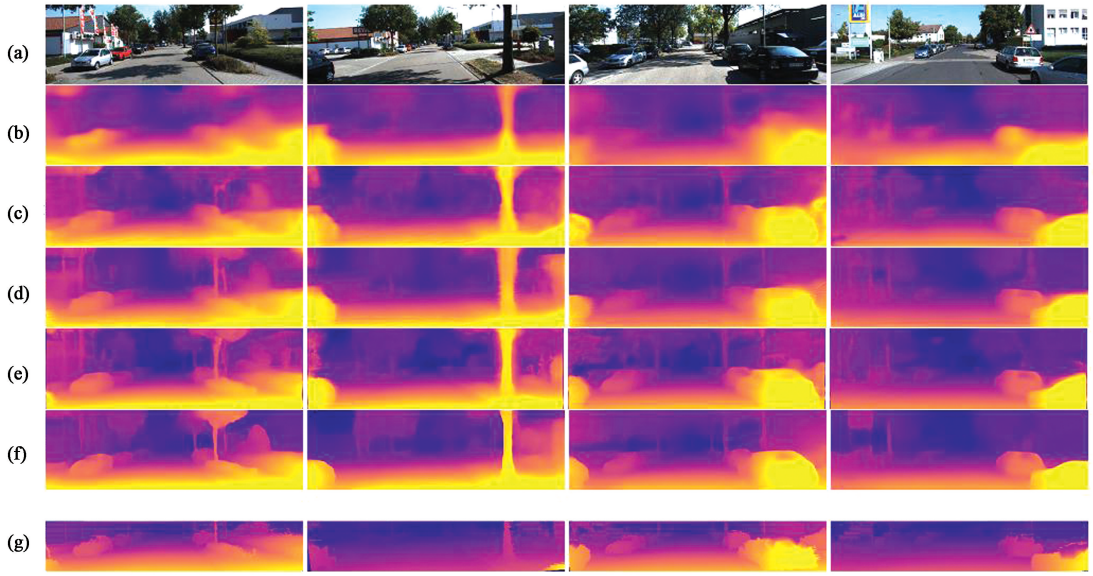


图 7 各种方法预测结果比较[20]。(a)图像；(b)Zhou 等；(c)GeoNet；(d)DDVO；(e)HMP；(f)Casserole 等；(g)深度图真值。

Fig.7 Comparison of prediction results of various methods[20]. (a) Images; (b) Zhou *et al.*; (c) GeoNet; (d) DDVO; (e) HMP; (f) Casserole *et al.*; (g) Ground truth.

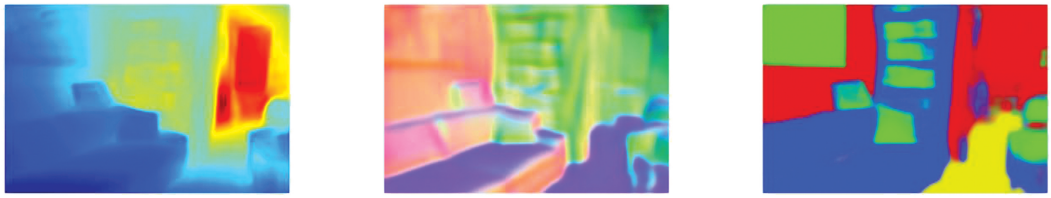


图 8 3 种任务的预测[49]。(a)深度图；(b)正常图；(c)标签。

Fig.8 Prediction of three tasks[49]. (a) Depth; (b) Normals; (c) Labels.

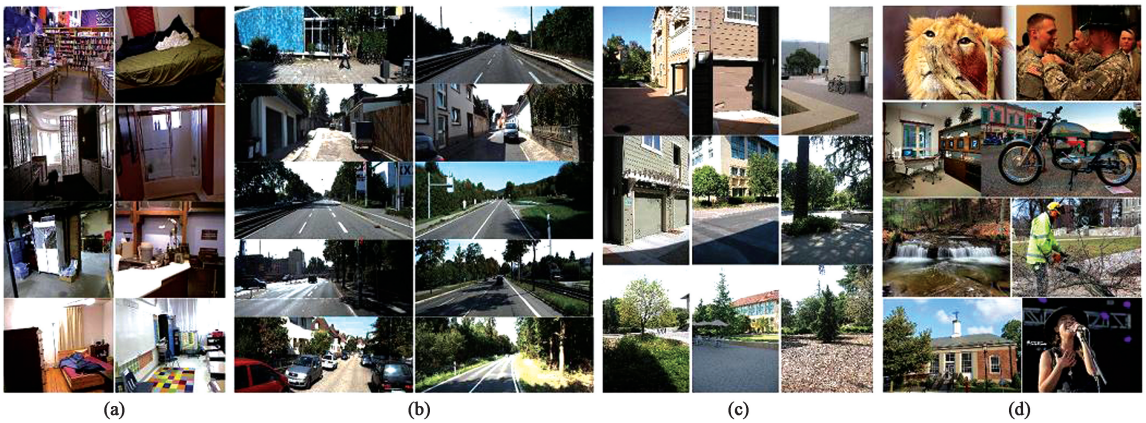


图 9 各种数据集(采用的为最右边的数据集)[50]。(a)NYU V2 数据集；(b)KITTI 数据集；(c)Make3D 数据集；(d)野外深度数据集。

Fig.9 Various data sets (the one used is the rightmost data set)[50]. (a) NYU V2 data set; (b) KITTI adta set; (c) Make 3D data set; (d) Field depth data set.

为了从单个图像中预测相对场景深度,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;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]也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情况下,在真实的室外环境中,几乎不可能确定地面深度信息。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,使用稀疏的真值进行监督学习,同时通过图像对齐损失函数来产生密集深度图。Li 和 Snavely[52] 提出了 MegaDepth 的大型深度数据集,从运动和多视角立体(Multi-View Stereo, MVS)方法生成训练数据,建议使用多视角互联网照片集;通过验证了大量互联网数据,验证了模型具有很强的泛化能力:不仅可用于新场景,而且可用于其他不同的数据集,包括 Make3D、KITTI 和 DIW。Liu[53]将单目深度估计表述为一个离散-连续优化问题(其中连续变量编码表示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),利用粒子信念传播在图形模型中进行推理,获得离散-连续优化问题的解,解决了单一图像中估计场景深度的问题。

为了解决 GPU 功耗高的问题,利用从单个输入图像中提取的特征金字塔,可在 CPU 或嵌入式系统上,实现快速推断出准确深度图的功能[28]。Montie 等人[54]介绍了一种基于特征的单目 SLAM 系统,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,可在各种操作系统环境中实时运行。Zhan 等人[29]使用立体序列学习深度和视觉里程测量,使空间(左右对之间)和时间(前向后)光度偏差的应用成为可能,并限制场景深度和摄像头运动在一个共同的尺度,可使单目序列获取不错的单视点深度和双视点里程。Wu 等人[55]从提取真实世界物体尺寸的标签并根据尺寸标签的几何关系推断出一个粗糙的深度映射,同时对条件随机场(CRF)进行能量函数优化并对深度图进行细化,对单目图像深度估计问题进行了新探索。

高分辨率是实现高保真自监督单目深度预测的关键。Pillai 等人[30]提出了一种采用深度超分辨率的亚像素卷积层扩展的方案,从相应的低分辨率卷积特征中精确地合成出高分辨率特征,同时引入了一个翻转增强层,可以准确地融合来自图像及其水平翻转版本的预测结果,减少由于遮挡而产生的左右阴影区域的影响,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标识方法,估算深度被限制为与预测法线兼容,从而产生更稳健的深度估计结果。

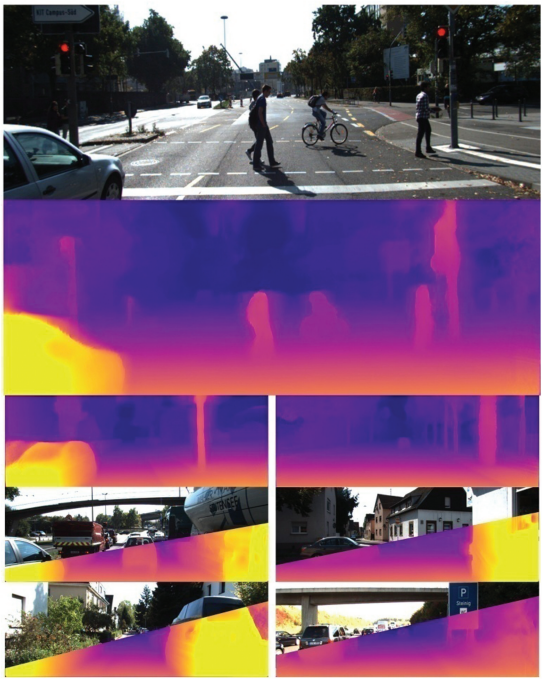


图 10　深度效果图[30]

Fig.10　Depth maps[30]

单目视觉深度估计的难度很大,但由于存在成本优势而被广泛研究。近年来,基于深度学习的单目深度估计得到了有效提升,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,学术界提出了多种网络结构、损失函数和训练策略,如单目自监督的 Monodepth2、GAN 范式下无监督深度估计的 MonoGAN 和基于深度注意的 DABC 网络等,这些工作推动了单目深度估计的快速发展。Madhu 等人[19]使用未标记的双目立体图像对训练,提出了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。

3 视觉建图

通常，构建 SLAM 稠密地图的规模和计算量都较大，导致地图的构建很难满足实时性要求；而基于深度视觉的建图方案旨在快速构建稠密准确的高质量地图，可以实时提供给 SLAM 算法用于定位。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。下文将按照图 11 所列举的视觉 SLAM 模块进行展开。

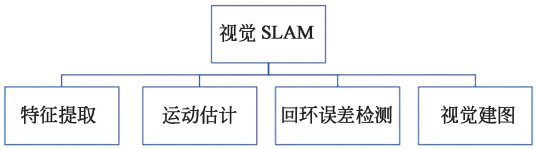


图 11 SLAM 模块

Fig.11 SLAM architecture

Schneider 等人[1]提出 Maplab 视觉惯性测绘定位系统，包括映射合并、视觉惯性批优化和环路闭合，通过可创建视觉-惯性地图的在线前端，实现在定位地图中跟踪一个全局无漂移姿态的处理和操作多会话映射。Konolige 等人[2]用大量的点特征匹配视觉帧以实现特征提取，通过经典束调整技术，简化保留相对的帧姿态信息(骨架)，获得了较好的建图效果，如图 12 所示。

Blake 等人[57]采用混合不同图像特征的方法，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，构建了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，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。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]则提出了用人工标记特征实现 SLAM 的方法。

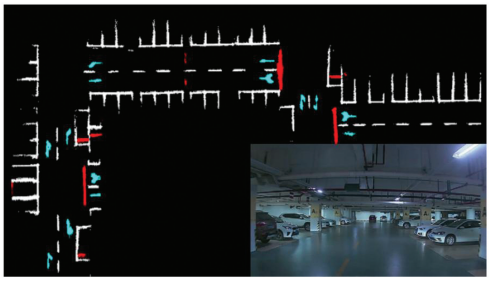


图 13 语义视觉地图[58]

Fig.13 Semantic visual map[58]

如图 14 所示，Fernandez 等人[4]通过智能标记系统实现运动估计，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，可以对具有正确比例尺的单目摄像头进行高精度的定位。Saeedi 等人[59]通过开发新的度量，在不依赖任何 SLAM 或运动估计算法的情况下正确地评估轨迹和环境。

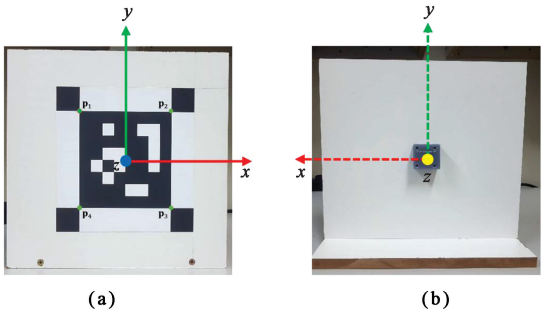


图 14 智能标记：前面一个正方形平面基准标记(a)，以及一个嵌入式姿态测量系统(PMS)单元(b)[4]。

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

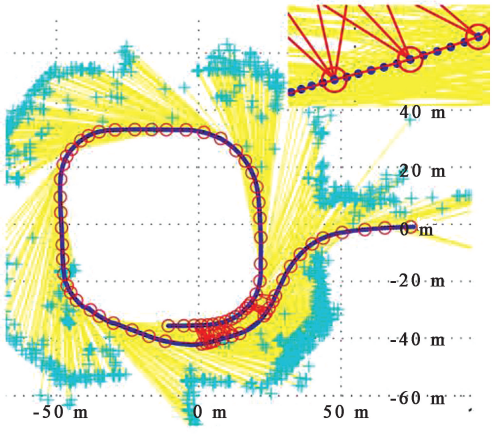


图 12 一个 100 m 城市场景的骨架缩小。完整贝叶斯图是 700 帧和约 100 K 的特征[2]。

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

回环误差检测方面，Usenko 等人[60]提出了利用非线性因子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，通过重建一组非线性

因子，使 VIO 积累的轨迹上的信息成为最佳近似，可使用 Bundle 调整这些因素以获得全局一致的映射。VIO 因子使全局映射的横倾角和俯仰角变得可见，从而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，视觉建图也采用帧对帧匹配来生成详细的局部映射以及大回路的闭合。Qin 等人[58]采用 4 个全景摄像头来增加感知范围，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]主要通过调整一个二维坐标系的局部图像以生成一个全局地图，并做出适当的纠正以生成 3D 面板，如图 15 所示。

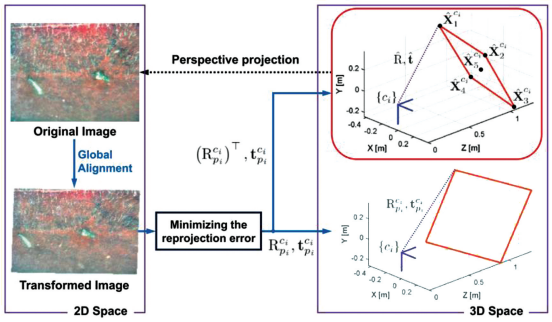


图 15　3D 面板的姿态估计示例。实际面板的方向可以通过对旋转矩阵 Rci 进行逆（转置）来估计[62]。

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　　论

本文从视觉 SLAM 视觉深度的预测及视觉建图两项核心技术入手，进行研究分析。视觉深度预测部分的分析包含了视觉数据的采集方式和算法的监督设计，按照视觉数据的视觉采集技术从单目、双目以及多目的不同角度进行探讨，根据算法的设计方案分别从全监督、半监督和无（自）监督等角度进行梳理；视觉建图部分则包含了特征提取、运动估计、回环检测和建图等方面的最新方法综述。研究表明：在视觉深度感知方面，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，单目双目融合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，是未来的发展方向之一。在视觉语义建图方面，由于可以提供更高层的语义特征，因而在算法上更具备鲁棒性，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，算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，所以需要根据计算量和精度的要求选择合适的特征匹配策略；视觉建图的另一项核心技术是运动估计，其既可以通过视觉帧匹配来完成，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；后者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, DYMCZYK M, FEHR M, *et al*. Maplab: an open framework for research in visual-inertial mapping and localization [J]. *IEEE Robotics and Automation Letters*, 2018, 3(3): 1418-1425.

[2] KONOLIGE K, AGRAWAL M. FrameSLAM: from bundle adjustment to real-time visual mapping [J]. *IEEE Transactions on Robotics*, 2008, 24(5): 1066-1077.

[3] XAVIER R S, DA SILVA B M F, GONCALVES L M G. Accuracy analysis of augmented reality markers for visual mapping and localization [C]//2017 *Workshop of Computer Vision* (*WVC*). Natal, Brazil: IEEE, 2017: 73-77.

[4] ORTIZ-FERNANDEZ L E, CABRERA-AVILA E V, DA SILVA B M F, *et al*. Smart artificial markers for accurate visual mapping and localization [J]. *Sensors*, 2021, 21(2): 625.

[5] CHAI Y, CAO X J. Stereo matching algorithm based on joint matching cost and adaptive window [C]//2018 *IEEE 3rd Advanced Information Technology*, *Electronic and Automation Control Conference* (*IAEAC*). Chongqing, China: IEEE, 2018: 442-446.

[6] GUO X Q, CHEN Z, LI S Y, *et al*. Deep eyes: binocular depth-from-focus on focal stack pairs [J]. *arXiv*, 1711.10729, 2019.

[7] SONG X, ZHAO X, LIN T W. Face spoofing detection by fusing binocular depth and spatial pyramid coding micro-texture features [C]//2017 *IEEE International Conference on Image Processing* (*ICIP*). Beijing, China: IEEE, 2017: 96-100.

[8] PILZER A, LATHUILIÈRE S, XU D, *et al*. Progressive fusion for unsupervised binocular depth estimation using cycled networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(10): 2380-2395.

[9] XU K, CHEN Z Y, JIA F C. Unsupervised binocular depth prediction network for laparoscopic surgery [J]. *Computer Assisted Surgery*, 2019, 24(S1): 30-35.

[10] SINHA A, MUREZ Z, BARTOLOZZI J, *et al*. DELTAS: depth estimation by learning triangulation and densification of sparse points [C]//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer, 2020: 104-121.

[11] YAO Y, LUO Z X, LI S W, *et al*. MVSNet: depth inference for unstructured multi-view stereo [C]//*Proceedings of the 15th European Conference on Computer Vision* (*ECCV*). Munich, Germany: Springer, 2018: 785-801.

[12] KUSUPATI U, CHENG S, CHEN R, *et al*. Normal assisted stereo depth estimation [C]//2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, WA, USA: IEEE, 2020: 2186-2196.

[13] WANG K X, SHEN S J. MVDepthNet: real-time multiview depth estimation neural network [C]//2018 *International Conference on 3D Vision* (*3DV*). Verona, Italy: IEEE, 2018: 248-257.

[14] HOU Y X, SOLIN A, KANNALA J. Unstructured multi-view depth estimation using mask-based multiplane representation [C]//*Proceedings of the 21st Scandinavian Conference on Image Analysis*. Norrköping, Sweden: Springer, 2019: 54-66.

[15] LONG X X, LIU L J, LI W, *et al*. Multi-view depth estimation using epipolar spatio-temporal networks[J]. *arXiv*, 2011.13118, 2020.

[16] GODARD C, AODHA O M, FIRMAN M, *et al*. Digging into self-supervised monocular depth estimation [C]//2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Seoul, Korea (South): IEEE, 2019: 3827-3837.

[17] ALEOTTI F, TOSI F, POGGI M, *et al*. Generative adversarial networks for unsupervised monocular depth prediction [C]//*Computer Vision-ECCV* 2018 *Workshops*. Munich, Germany: Springer, 2018: 337-354.

[18] ATAPOUR-ABARGHOUEI A, BRECKON T P. Real-time monocular depth estimation using synthetic data with domain adaptation *via* image style transfer [C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 2800-2810.

[19] BABU V M, DAS K, MAJUMDAR A, *et al*. UnDEMoN: unsupervised deep network for depth and ego-motion estimation [C]//2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Madrid, Spain: IEEE, 2018: 1082-1088.

[20] CASSER V, PIRK S, MAHJOURIAN R, *et al*. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos [C]//*Thirty-Third AAAI Conference on Artificial Intelligence* (*AAAI*'19). Hawaii, USA: AAAI, 2019.

[21] FU H, GONG M M, WANG C H, *et al*. Deep ordinal regression network for monocular depth estimation [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 2002-2011.

[22] GUO X Y, LI H S, YI S, *et al*. Learning monocular depth by distilling cross-domain stereo networks [C]//*Proceedings of the 15th European Conference on Computer Vision* (*ECCV*). Munich, Germany: Springer, 2018: 506-523.

[23] JIANG H Z, LARSSON G, MAIRE M, *et al*. Self-supervised relative depth learning for urban scene understanding [C]//*Proceedings of the 15th European Conference on Computer Vision* (*ECCV*). Munich, Germany: Springer, 2018: 20-37.

[24] KUZNIETSOV Y, STÜCKLER J LEIBE B. Semi-supervised deep learning for monocular depth map prediction [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017:

2215-2223.

[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/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 5667-5675.

[26] LI R B, XIAN K, SHEN C H, *et al*. Deep attention-based classification network for robust depth prediction [C]//*Proceedings of the* 14*th Asian Conference on Computer Vision*. Perth, Australia: Springer, 2019: 663-678.

[27] RANJAN A, JAMPANI V, BALLES L, *et al*. 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*. Long Beach, CA, USA: IEEE, 2019: 12232-12241.

[28] POGGI M, ALEOTTI F, TOSI F, *et al*. Towards real-time unsupervised monocular depth estimation on CPU [C]//2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Madrid, Spain: IEEE, 2018: 5848-5854.

[29] ZHAN H Y, GARG R, WEERASEKERA C S, *et al*. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 340-349.

[30] PILLAI S, AMBRUŞ R, GAIDON A. SuperDepth: self-supervised, super-resolved monocular depth estimation [C]//2019 *International Conference on Robotics and Automation* (*ICRA*). Montreal, QC, Canada: IEEE, 2019: 9250-9256.

[31] 程明明,王贺,安平,等.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 M M, WANG H, AN P, *et al*. Multi-view image correction based on feature point matching [J]. *Chinese Journal of Liquid Crystals and 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 J Y, PIAO Y. Research on stereo matching algorithm based on improved steady-state matching probability [J]. *Chinese Journal of Liquid Crystals and Displays*, 2018, 33(4): 357-364. (in Chinese)

[33] YAO H M, GE C Y, XUE J R, *et al*. A high spatial resolution depth sensing method based on binocular structured light [J]. *Sensors*, 2017, 17(4): 805.

[34] WANG E D, ZHU Y L, PENG L Y, *et al*. 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*). Honolulu, HI, USA: IEEE, 2017: 1001-1004.

[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 W. Research on binocular stereo vision based on deep learning [D]. Hefei: Hefei University of Technology, 2020. (in Chinese)

[36] ANANTRASIRICHAI N, CANAGARAJAH C N, REDMILL D W, *et al*. Dynamic programming for multi-view disparity/depth estimation [C]//2006 *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Toulouse, France: IEEE, 2006.

[37] MONTSERRAT T, CIVIT J, ESCODA O D, *et al*. Depth estimation based on multiview matching with depth/color segmentation and memory efficient belief propagation [C]//2009 16*th IEEE International Conference on Image Processing* (*ICIP*). Cairo, Egypt: IEEE, 2009: 2353-2356.

[38] LIU Y B, CAO X, DAI Q H, *et al*. Continuous depth estimation for multi-view stereo [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE, 2009: 2121-2128.

[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, Transmission and Display of 3D Video*. Tampere, Finland: IEEE, 2010: 1-4.

[40] MIELOCH D, DZIEMBOWSKI A, GRZELKA A, *et al*. Graph-based multiview depth estimation using segmentation [C]//*IEEE International Conference on Multimedia and Expo*. Hong Kong, China: IEEE, 2017: 217-222.

[41] FÁCIL J M, CONCHA A, MONTESANO L, *et al*. Single-view and multi-view depth fusion [J]. *IEEE Robotics*

and 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*. Aberystwyth, UK: BMVA Press, 2010: 1-11.

[43] INCE S, MARTINIAN E, YEA S, *et al*. Depth estimation for view synthesis in multiview video coding [C]//*3DTV Conference*. Kos, Greece: IEEE, 2007: 1-4.

[44] LONG X X, LIU L J, THEOBALT C, *et al*. Occlusion-aware depth estimation with adaptive normal constraints [C]//*16th European Conference on Computer Vision*. Glasgow, UK: Springer, 2020: 640-657.

[45] STRECHA C, VAN GOOL L. PDE-based multi-view depth estimation [C]//*First International Symposium on 3D Data Processing Visualization and Transmission*. Padua, Italy: IEEE, 2002: 416-425.

[46] GARG R, VIJAY KUMAR B G, CARNEIRO G, *et al*. Unsupervised CNN for single view depth estimation: geometry to the rescue [C]//*14th European Conference on Computer Vision*. Amsterdam, Netherlands: Springer, 2016: 740-756.

[47] GODARD C, AODHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 6602-6611.

[48] ZHOU T H, BROWN M, SNAVELY N, *et al*. Unsupervised learning of depth and ego-motion from video [C]//2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu, HI, USA: IEEE, 2017: 6612-6619.

[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*. Santiago, Chile: IEEE, 2015: 2650-2658.

[50] CHEN W F, FU Z, YANG D W, *et al*. Single-image depth perception in the wild [C]//*Proceedings of the 30th International Conference on Neural Information Processing Systems*. Barcelona: Curran Associates Inc., 2016: 730-738.

[51] KENDALL A, MARTIROSYAN H, DASGUPTA S, *et al*. End-to-end learning of geometry and context for deep stereo regression [C]//*Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017: 66-75.

[52] LI Z Q, SNAVELY N. MegaDepth: learning single-view depth prediction from internet photos [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018: 2041-2050.

[53] LIU M M, SALZMANN M, HE X M. Discrete-continuous depth estimation from a single image [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014: 716-723.

[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 R, YING S H, ZHENG L M. Size-to-depth: a new perspective for single image depth estimation[J]. *arXiv*, 1801.04461, 2018.

[56] YANG Z H, WANG P, XU W, *et al*. Unsupervised learning of geometry with edge-aware depth-normal consistency[J]. *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*). New York, NY, USA: IEEE, 2006: 230-237.

[58] QIN T, CHEN T Q, CHEN Y L, *et al*. AVP-SLAM: semantic visual mapping and localization for autonomous vehicles in the parking lot [C]//2020 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Las Vegas, NV, USA: IEEE, 2020: 5939-5945.

[59] SAEEDI S, CARVALHO E D C, LI W B, *et al*. Characterizing visual localization and mapping datasets [C]//2019 *International Conference on Robotics and Automation* (*ICRA*). Montreal, QC, Canada: IEEE, 2019: 6699-6705.

[60] USENKO V, DEMMEL N, SCHUBERT D, *et al*. Visual-inertial mapping with non-linear factor recovery [J]. *IEEE Robotics and Automation Letters*, 2020, 5(2): 422-429.

[61] XIAO L H, WANG J G, QIU X S, *et al*. Dynamic-SLAM: semantic monocular visual localization and mapping based on deep learning in dynamic environment [J]. *Robotics and Autonomous Systems*, 2019, 117: 1-16.

[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，主要从事机器人方面的研究。E-mail：yuanfeng_chen@hotmail.com