软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(4):1692−1714 [doi: 10.13328/j.cnki.jos.007248] [CSTR: 32375.14.jos.007248]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的多视图立体视觉综述*

樊铭瑞[1,2], 申冰可[1,2], 牛文龙[1,2], 彭晓东[1,2,3], 谢文明[1,2], 杨 震[1]

[1](中国科学院 国家空间科学中心, 北京 100190)

[2](中国科学院大学, 北京 100049)

[3](国科大杭州高等研究院, 浙江 杭州 310024)

通信作者: 牛文龙, E-mail: niuwenlong@nssc.ac.cn

摘 要: 多视图立体视觉在自动驾驶、增强现实、遗产保护和生物医学等领域得到广泛应用. 为了弥补传统多视图立体视觉方法对低纹理区域不敏感、重建完整度差等不足, 基于深度学习的多视图立体视觉方法应运而生. 对基于深度学习的多视图立体视觉方法的开创性工作和发展现状进行综述, 重点关注基于深度学习的多视图立体视觉局部功能改进和整体架构改进方法, 深入分析代表性模型. 同时, 阐述目前广泛使用的数据集及评价指标, 并对比现有方法在数据集上的测试性能. 最后对多视图立体视觉未来有前景的研究发展方向进行展望.

关键词: 深度学习; 计算机视觉; 三维重建; 多视图立体视觉

中图法分类号: TP18

中文引用格式: 樊铭瑞, 申冰可, 牛文龙, 彭晓东, 谢文明, 杨震. 基于深度学习的多视图立体视觉综述. 软件学报, 2025, 36(4): 1692–1714. http://www.jos.org.cn/1000-9825/7248.htm

英文引用格式: Fan MR, Shen BK, Niu WL, Peng XD, Xie WM, Yang Z. Survey on Multi-view Stereo Based on Deep Learning. Ruan Jian Xue Bao/Journal of Software, 2025, 36(4): 1692–1714 (in Chinese). http://www.jos.org.cn/1000-9825/7248.htm

Survey on Multi-view Stereo Based on Deep Learning

FAN Ming-Rui[1,2], SHEN Bing-Ke[1,2], NIU Wen-Long[1,2], PENG Xiao-Dong[1,2,3], XIE Wen-Ming[1,2], YANG Zhen[1]

[1](National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China)

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

[3](Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China)

Abstract: Multi-view stereo (MVS) is widely used in fields such as autonomous driving, augmented reality, heritage conservation, and biomedicine. To address the limitations of traditional MVS methods, such as insensitivity to low-texture regions and poor reconstruction integrity, deep learning-based MVS methods have been proposed. This study reviews the pioneering work and current development of deep learning-based MVS methods. In particular, it focuses on methods for local functional improvement and overall architectural improvement and analyzes representative models. Meanwhile, the study describes widely used datasets and evaluation metrics and compares the test performance of existing methods on the datasets. Finally, promising research directions for MVS are presented.

Key words: deep learning; computer vision; 3D reconstruction; multi-view stereo (MVS)

三维重建能够通过传感器获取目标场景信息并恢复目标的三维几何结构, 是计算机视觉领域中一个非常重要的研究方向. 近年来, 三维重建因其可应用场景的凸显和采样技术的发展受到了越来越多的关注. 它所涉及的场景主要包括: 生物医学[1]、增强/虚拟现实[2]、机器人视觉导航[3]、遗产考古[4]等. 同时, 采样传感器的便捷性有助于人们更好地捕获场景信息, 如深度传感器、激光扫描仪和摄像机等. 然而深度传感器捕获的深度图无法满足三维重建的细节需求, 激光扫描仪的价格也较为昂贵. 相较于深度和激光传感器, 图像数据的获取更为简单方便, 同时

* 基金项目: 中国科学院基础前沿科学研究计划 (22E0223301); 中国科学院青年创新促进会项目 (E1213A02)
收稿时间: 2023-06-28; 修改时间: 2024-02-08; 采用时间: 2024-07-10; jos 在线出版时间: 2024-12-31
CNKI 网络首发时间: 2025-01-02

它也包含除深度之外的诸多环境信息, 适合更多样的场景[5,6]. 因此, 基于图像的三维重建是当前计算机视觉领域的研究热点.

多视图立体视觉 (multi-view stereo, MVS) 是基于图像三维重建的基础, 能够实现根据摄像机参数从多视图视角中恢复场景的三维几何形状, 从而完成二维平面到三维立体的转化. 摄像机参数包括内参和外参, 主要由摄像机本身及拍摄时的位姿决定. 摄像机从多个角度对目标场景拍摄获取多视角图像. 利用多视角图像通过 MVS 方法实现三维场景的重建, 流程如图 1 所示[7].

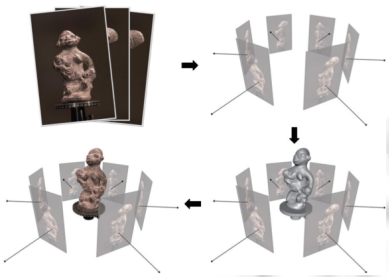


图 1 MVS 流程图[7]

传统的 MVS 方法通常可以分为 4 种类型: 基于面片的方法、多边形网格法、体素法和基于深度图的方法, 如表 1 所示. 基于面片的方法采用面片的方式表示场景, 并通过将重建纹理区域的补丁传播到低纹理区域, 实现场景重建. 基于多边形网格的方法通过三角剖分构建初始稠密点云, 然后根据点云生成多边形网格, 并对网格优化, 最后将纹理映射到网格上, 实现重建. 该方法为轻量级方法, 其优势在于能够更清晰地表示目标物体的形状细节信息, 但是只能完成表面重建, 无法对物体的内部进行处理. 基于体素的方法是将场景划分为由块状体素组成的结构, 并通过二值的方式对每块体素标记为空或满, 但是该方法消耗内存大, 且无法满足高分辨率和高精度要求. 而基于深度图的方法根据稀疏重建获得的三维点位置和相机轨迹, 对每个视角的图像逐个计算深度图, 然后通过对深度信息融合获得点云. 因此与其他 MVS 方法相比, 基于深度图的方法更为灵活, 可以保持场景的平滑性.

表 1 传统多视图立体视觉方法对比

方法	特点	优点	缺点
基于面片的方法	以面片的方式表示场景	能够恢复复杂结构	存在不可靠问题
多边形网格法	利用多边形对曲面分割	轻量级方法	需要良好的初始点
体素法	将模型划分为块表示	易于提取网格	占用内存大, 精度受限于块大小
基于深度图的方法	融合为点云表示	适合大规模场景	需要对深度图融合

尽管传统 MVS 方法在重建结果方面精度高, 但仍然存在一些不足.

(1) 消耗时间长, 效率低

大量特征点匹配和复杂的几何计算导致计算过程的复杂度增加.

(2) 完整度不高, 难以处理弱纹理区域以及反射和投射的表面

传统 MVS 依赖于几何和光度一致性, 在朗伯场景下能够生成精度较高的模型, 但是对于有弱纹理和反射区域的场景容易出现匹配错误的情况, 因此给重建带来困难.

(3) 需要手工设定相似性度量

传统方法中, 为了实现对特征的密集匹配, 需要手工设定度量的标准来测量图像中两点的相似性, 这种方式具有一定的局限性.

(4) 无法实现大规模场景的重建

重建过程对内存资源要求高且建模时间长. 在处理大规模场景时, 由于图像数量多, 需要处理大量的信息, 易导致内存不足和耗时过长的问题.

因此, 在非理想的环境下, 传统方法难以取得较好的重建效果. 随着计算机视觉技术的发展和算力的提升, 深度学习方法开始着手解决以上问题. 与传统方法相比, 基于深度学习的方法具有以下优势.

(1) 重建耗时短. 基于深度学习的方法利用强大的特征提取能力避免了繁琐的特征匹配和视差计算过程, 能够提高重建的效率.

(2) 对场景的鲁棒性强. 基于深度学习的方法通过学习更丰富的特征表示[6]和利用上下文信息能够增强对非朗伯场景的鲁棒性.

(3) 泛化性强. 基于深度学习的方法能够提取更深层次的特征, 学习利用不同场景的特征对度量优化调整, 减少了手工设计度量方法的局限性.

(4) 能够实现大规模场景的重建. 基于深度学习的方法减少对特征点和视差信息的存储需求降低了内存消耗, 能够更好地应对大规模场景的重建需求.

最初, 有研究者[8−13]使用深度学习模型结合传统 MVS 方法中的部分模块, 实现部分集成以提高单个模块的性能. 在这个阶段, 仍然存在需要手工设计的处理步骤. 目前基于深度学习的多视图立体视觉 (MVS) 网络主要分为 3 类: 基于深度图、基于体素和基于辐射场. 这些方法避免了繁琐的手工设计, 实现了端到端的学习. 这 3 种方法都是根据一组图像及其相机参数, 通过端到端网络结构获得三维模型. 其中, 基于深度图的方法需要先获得深度图, 然后通过融合深度图得到三维点云模型, 基于体素的方法可以直接获得由体素表示的三维模型, 而基于辐射场的方法则通过距离函数或高斯函数的表示三维表面重建结果.

近几年来, 深度学习方法取得了许多突破性进展. 端到端的多视图立体视觉网络经历了多个里程碑. 因此本文梳理归纳了近年来该领域的发展历程和代表方法. 图 2 列举了 3 类方法的发展历程, 上方区域是基于体素的方法, 中间区域是基于深度图的方法并分为有监督的方法和自监督/无监督的方法, 下方区域是基于辐射场的方法. 2017 年, 出现了第 1 个基于深度学习的端到端 MVS 重建系统 SurfaceNet[14]. 随后, 为了解决大规模重建问题, Yao 等人开创性地提出了基于深度图的 MVSNet[15]. 后续的多个网络在 MVSNet 的基础上相继被提出. 鉴于三维模型真实数据难以获取, 无监督的方法逐渐引起关注, Khot 等人[16]提出了第 1 个无监督 MVS 框架. 基于辐射场的方法发展较晚, 2021 年, MVSNeRF[17]将 MVS 和 NeRF 相结合, 实现三维场景的隐式表达. 2024 年, GaussianPro[18]的提出, 开创了辐射场新的研究方向.

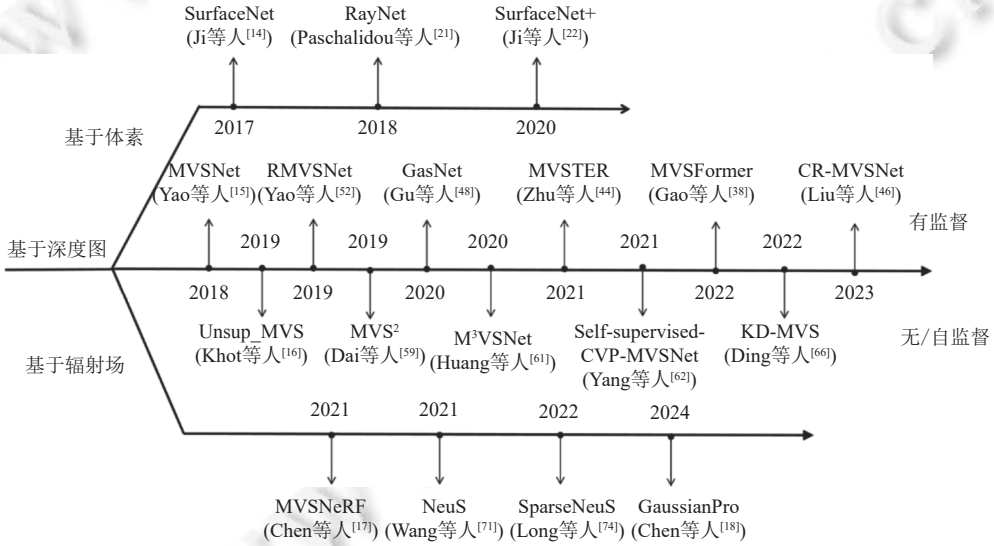


图 2　2018–2024 年基于深度学习的 MVS 整体框架改进算法的发展历程

本文系统性地回顾了截至目前基于深度学习的多视图立体视觉发展现状, 并对其相关的概念、方法及效果进行了总结和讨论. 本文第 1 节介绍针对传统多视图立体视觉方法的局部模块改进. 第 2 节总结基于深度学习的多

视图立体视觉整体架构改进, 包括基于体素的深度学习 MVS 方法、基于深度图的深度学习 MVS 方法和基于辐射场的深度学习 MVS 方法. 第 3 节介绍各类常用数据集和重建结果的评价指标. 第 4 节对各种方法在数据集上进行分析对比. 第 5 节讨论该领域现有的难点和未来的研究方向.

1 基于深度学习的 MVS 局部模块改进

多视图立体视觉通用流程是提取特征点, 进行特征匹配, 并计算深度图, 然后通过深度图融合得到三维模型. 卷积神经网络 (CNN) 在图像识别、目标检测和语义分割等领域应用广泛. 在多视图立体视觉任务中, 前人也提出了多种基于 CNN 改进流程中单个步骤的方法. 本文重点关注影响最终的准确率和消耗内存资源大小的匹配代价计算部分.

匹配代价是指特征像素点或像素窗口之间的差值, 常用于视图之间的两两匹配. 为了解决手工设计的描述符无法最佳地表示特征的问题, 有研究者受深度学习进展的启发进行了一系列的工作. Žbontar 等人[8]摒弃了使用手工设计的特征来计算立体匹配代价, 提出 MC-CNN 来预测两个图像块之间的匹配程度. 与文献 [8] 类似, 文献 [9] 提出了多种基于 CNN 的架构, 能够直接从图像的像素中学习到相似函数, 从而实现对图像补丁之间相似度的比较. 此外, 针对基于面片的匹配方式无法充分匹配图像块的问题, 受新一代学习描述符的启发, Han 等人[10]利用度量学习定义相似性, 从而计算出小图像块之间的匹配代价, 但是无法寻找到图像之间的对应部分. 文献 [11] 构建了一个包含相似和不相似块对的二元分类数据集, 并根据速度和准确率的不同设计了两种 CNN 结构, 以学习图像块上的相似性度量并将其应用于立体匹配问题. 针对立体匹配过程中传统方法在有反射性和无纹理的平面容易产生歧义的问题, Güney 等人[12]引入基于稀疏视差估计和图像语义分割的逆向图形技术, 通过分类确定目标视差位置, 在更大距离范围上进行正则化, 有效地解决了此问题. 实际上, 上述工作将问题视作二分类问题来学习匹配网络的参数, 产生了昂贵的计算代价. 为了实现快速的 GPU 计算, 文献 [13] 通过内积层计算连接层中两个特征的内积, 并将问题视作多分类问题来训练网络.

尽管此阶段的模型已经取得了一定的成效, 但是也存在一些局限性. 采用多阶段方法的模型复杂度高, 可能会导致数据传输过程中的不一致问题, 影响模型的精度和鲁棒性. 因此, 后续工作逐渐尝试采用端到端的结构解决问题. 根据场景表示的不同, 将其分为基于体素和基于深度图两类框架体系结构.

2 基于深度学习的 MVS 整体架构改进

过去的工作集中于对 MVS 流程中部分步骤进行改进后集成, 但由于缺乏上下文几何知识, 它们的性能在具有挑战性的场景中受到限制. 只有将整个流程设计为端到端的学习框架, 才能激发出多视图立体视觉更大的潜力. 基于深度学习的端到端 MVS 架构分为 3 种方法: 基于体素、基于深度图和基于辐射场. 下面详细介绍这 3 种方法的研究现状.

2.1 基于体素的方法

为了应对传统方法在缺乏纹理或宽基线情况下导致的重建失败问题, 受长短期记忆网络 (long short-term memory, LSTM) 的启发, 3D-R2N2[19]将单视图或多视角图像作为输入, 建立 2D 图像和 3D 体素之间的映射关系, 以三维占用网格的形式输出重建模型. 但是, 重建模型的精度低, 无法反映出目标物体的细节信息. Kar 等人[20]提出一种名为立体学习机 (learnt stereo machine, LSM) 的方法. LSM 直接利用相机参数投影形成代价体, 将像素特征向上投影到 3D 体素, 并根据体素是否被曲面占据进行分类. 虽然上述方法都已经开始将卷积神经网络应用到多视图立体视觉的研究中, 但是生成的三维模型都比较粗糙, 缺乏细节信息. 第 1 个基于深度学习的端到端 MVS 重建系统是 SurfaceNet 网络[14]. 它以一组图像和对应的摄像机参数作为输入, 预测体素表面概率, 再转换为曲面, 直接获得三维模型.

值得注意的是, SurfaceNet 是第 1 个端到端的, 也是第 1 个基于体素的多视图立体视觉深度学习框架. 在 SurfaceNet 之前, 所有结合深度学习的多视图立体视觉方法虽然已经取得了很大的进步, 但是缺点也很明显: 分块

运行, 缺乏上下文几何知识. 为此, SurfaceNet 建立了彩色体素立方体 (colored voxel cube, CVC), 通过编码表示相机参数与体素. SurfaceNet 的思想主要是通过学习光度一致性和表面结构的几何关系, 从三维体素空间推断重构出二维表面. 用三维体素表示隐式编码摄像机参数. 将体素投影到每个视图的图像上, 转换为 RGB 值来表示彩色值的三维彩色体素, 将两个不同视角的彩色体素立方体 $I^C_{V_i}$ 和 $I^C_{V_j}$ 经过多组卷积层 l 聚合到输出层 y, 从而推断出体素表面的概率. 最后, 通过二值化和细化的后处理操作, 将概率转换成表面体素, 以完成完整的重建. SurfaceNet 网络结构如图 3 所示[14].

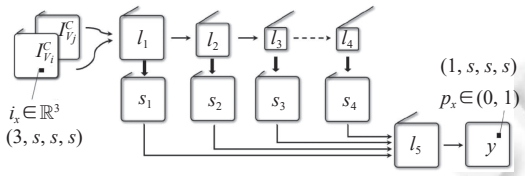


图 3 SurfaceNet 网络结构[14]

此后, 考虑到卷积神经网络和带有射线势的马尔可夫随机场 (Markov random field, MRF) 的优势, Paschalidou 等人提出了 RayNet[21]. 该网络利用 CNN 学习视图的不变特征, 通过 MRF 对透视投影编码并处理遮挡的问题. 但是, 仍然存在无法在大型场景中捕捉细节问题. 在 SurfaceNet 的基础上, Ji 等人提出了 SurfaceNet+[22], 解决了 SurfaceNet 中视图选择复杂和模型存在大孔洞等问题. 使用可训练的视图选择方式对预测的粗糙曲面进行几何验证, 逐步迭代细化曲面. SurfaceNet+最终预测出不同视图下的子体素在曲面上的概率, 其网络结构表示如图 4 所示[22].

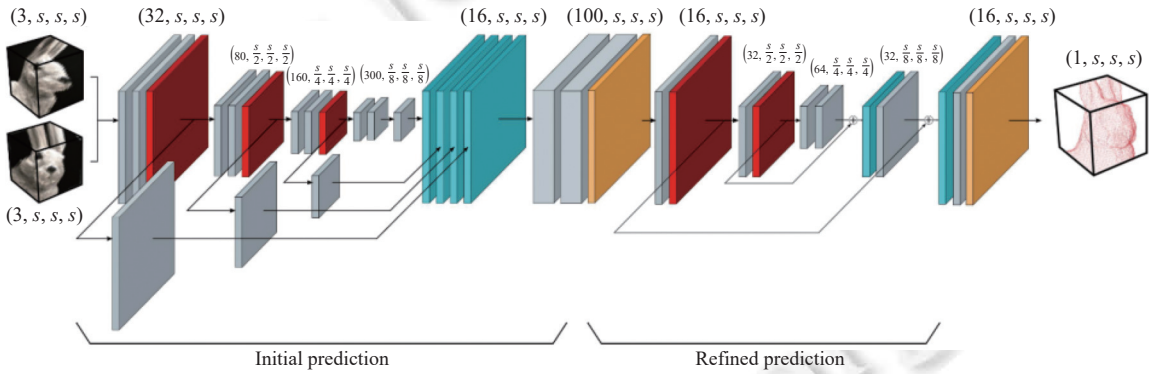


图 4 SurfaceNet+网络结构[22]

2.2 基于深度图的方法

尽管基于体素的算法可以实现重建三维模型, 但使用体素表示三维模型的方式会占有大量内存. 一旦需要提高精度, 会增加大量的运算时间. 因此, 无法扩展到更多样化和更大规模的场景中. 在应用过程中, 高精度模型和更高的分辨率意味着需要消耗大量的计算时间. 此外, 根据最近的 MVS 数据集评价[23], 基于深度图的算法重建精度优于基于体素的算法. 因此, 研究基于深度图的多视图立体视觉方法具有重要意义. 基于深度图的多视图立体视觉架构一般包括特征提取、构造代价体、深度图计算和优化模块等, 并在后处理步骤中通过深度图融合完成稠密重建. 根据不同的学习策略, 本文将基于深度图的算法分为监督学习和无监督学习.

2.2.1 监督学习

虽然 SurfaceNet 开创了端到端训练, 但是受到内存消耗的限制, 无法适应大规模重建. 因此没有被广泛应用. MVSNet[15]是由 Yao 等人在 2018 年提出的. 它解决了基于体素的 SurfaceNet 和 LSM 只能用于小规模重建的问题. 将直接计算整个场景转化为先根据每个视图生成深度图再进行深度图融合的两阶段方法. MVSNet 开创了基于深

度图多视图立体视觉的先河, 后续方法大多是以 MVSNet 为基础的变体.

MVSNet 由 5 个模块组成, 包括特征提取、代价体构建、深度图计算、深度图优化和后处理模块. 虽然后续的网络在精度以及效率上超过了 MVSNet, 但由于深受其网络结构的影响, 大都遵循此范例进行设计. 因此, MVSNet 被认为是多视图立体视觉的里程碑之一. MVSNet 的结构表示如图 5 所示[15]. 其基本流程如下.

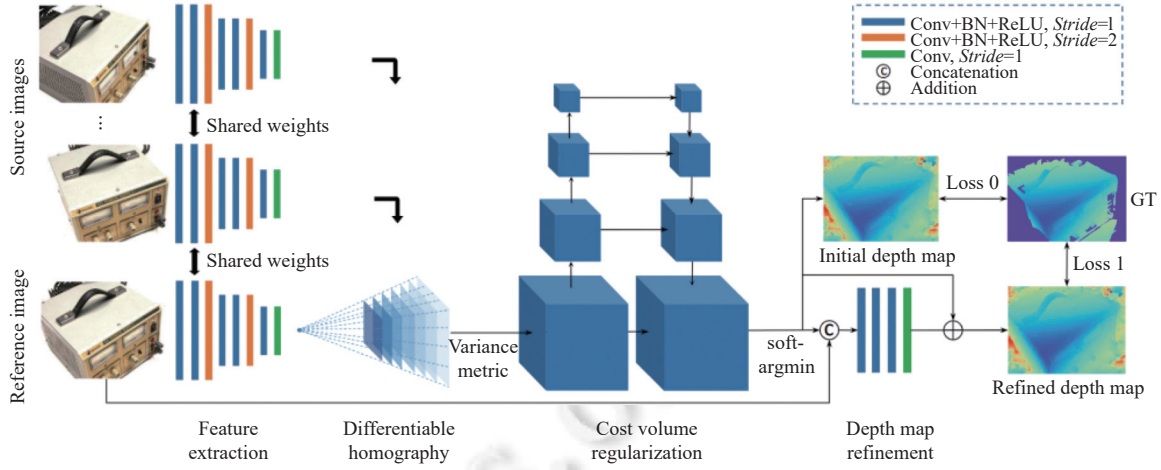


图 5 MVSNet 网络结构[15]

(1) 特征提取. 选取数据集中图像依次作为参考图, 其他图像作为邻域帧. 每个图像通过 8 层 2D 卷积神经网络得到 32 通道特征图, 长宽均为原图像的 1/4. 为了提高学习效率, 特征提取的操作过程中共享参数. 其中第 3 层和第 6 层的卷积步长设置为 2, 用于提取更高维的图像特征. 虽然特征图尺寸减小, 但由于最后输出的是 32 通道的特征图, 因此缩小后的特征图中的像素仍包含其周围像素的信息.

(2) 代价体构建. 构建 3D 代价体分为构建单应性矩阵、代价累积和代价体正则化这 3 个过程. 首先, 结合相机的几何关系, 通过可微单应性矩阵实现 2D 特征图到 3D 特征体的转换. 基于参考相机视锥体中若干不同深度的平行平面, 将特征图投影到参考图像下:

$$\boldsymbol{H}_i(d)=\boldsymbol{K}_i\boldsymbol{R}_i\left(I-\frac{(\boldsymbol{t}_1-\boldsymbol{t}_i)n_1^{\mathrm{T}}}{d}\right)\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{R}_1^{-1} \tag{1}$$

其中, $\boldsymbol{K}$ 表示相机内参, $\boldsymbol{R}$ 和 $\boldsymbol{t}$ 分别表示相机外参的旋转矩阵和平移矩阵. n 为参考相机的主轴, d 为深度. $\boldsymbol{H}(d)$ 为获得的单应性矩阵, 表示源视图和参考视图之间的坐标映射关系.

代价累积是将所有特征体合并成一个代价体的过程. 代价体是由代价图在深度方向上构建而成的. 采用基于方差的方法, 逐个像素进行计算:

$$C=M(V_1,\ldots,V_2)=\frac{\sum_{i=1}^{N}(V_i-\overline{V}_i)^2}{N} \tag{2}$$

其中, V 表示特征体, $\overline{V}$ 表示所有特征体的均值, N 表示特征体的数量. 计算获得方差后, 得到锥形体, 并对锥形体进行可微双线性插值, 以得到长宽一致的方形代价体. 代价体衡量了视图之间的相似性. 由于代价体中可能存在噪声, 因此对代价体采用 3D 四级 UNet 结构进行正则化, 最终出路形成通道数为 1 的概率体. 概率体表示每个像素深度值的可能性. 为了减少计算成本, 在正则化过程中, 将 32 通道减少到 8 通道, 并将卷积层改为两层.

(3) 深度图计算. 在传统方法中, 选择最大概率深度图 (即赢者通吃原则) 作为结果, 但是在非理想情况下效果一般, 而且无法产生亚像素级别的结果. 因此, MVSNet 网络通过计算数学期望生成初始深度图:

$$D=\sum_{d=d_{\min}}^{d_{\max}}d\times P(d) \tag{3}$$

其中, $P(d)$ 表示每个像素在深度值 d 处的概率估计值, $[d_{\min}, d_{\max}]$ 是深度值的范围.

(4) 深度图优化. 初始深度图的边缘信息模糊, 因此需要借助参考图像优化. 将参考图变为原始尺寸的 1/4 后与初始深度图拼接为 4 通道图, 之后输入到 4 层深度残差结构中. 其中, 残差结构包括 3 层 32 通道的 2D 卷积层和 1 层 1 通道的卷积层. 为了学习残差, 最后一层去除了 BN 层和 ReLU 层. 同时, 将初始深度图的深度范围设定为 [0, 1], 在完成优化后再转换回来. 由参考网络输出的深度残差结果和初始深度图相加得到最终深度图.

(5) 后处理. 由于拍摄图像容易被遮挡或受到噪声影响, 预测的深度图可能会存在误差. 因此在生成三维模型之前, 需要对深度图进行过滤. 根据光度和几何一致性过滤深度图, 确定每个像素的可见性, 并生成可见视图. 然后, 对可见视图重投影估计出像素的深度值. 最后, 基于融合算法将深度图重投影到三维空间, 生成稠密点云.

MVSNet 网络为基于深度图的深度学习 MVS 算法构建了一个完整的模块流程, 后续的方法针对网络中模块的问题进行了改进.

(1) 特征提取

针对 MVSNet 中直接使用卷积层序列提取特征时存在特征不能有效提取和利用的问题, 一些研究者提出修改特征提取模块的结构.

多尺度特征提取能够获得场景下的宏观和细节信息, 对不同视角和不同尺度的图像具有更好的鲁棒性. Xue 等人[24]使用 6 种尺寸的 U 型结构和跳跃连接结构提取图像中的深度特征, 这样能够包含更多的全局信息和丰富的局部信息. 而金字塔结构常被应用在图像的特征提取模块中以实现提取不同尺度图像的特征. CVP-MVSNet[25]分别对图像进行采样形成图像金字塔, 然后对所有层级图像分别提取特征图. 而许多网络[26–28]提出的另一个策略是实现一个特征金字塔. 金字塔结构能够从粗到细地提取高层和低层的语义特征以构建级联代价体, 融合多层次的信息. 然而, 由于不同尺度特征图之间存在语义差异, 直接对特征金字塔提取的特征图进行融合会忽略全局上下文信息和图像之间的特征关系, 导致多尺度特征的表达能力降低. 为了恢复场景更多细节, Yan 等人[29]提出了 D2HC-RMVSNet. 它采用轻量级的 DRENet 提取密集特征, 连接不同的扩张卷积层, 在不丢失分辨率的基础上聚合多尺度的上下文信息. ADIM-MVSNet[30]也采用了多尺度特征聚合模块 (MFA), 如图 6 所示[30], 通过局部感知域感知纹理丰富的区域. 类似地, MG-MVSNet[31]使用多粒度特征融合 (multiple granularity feature fusion, MGFF) 模块, 在提取不同尺度特征图后, 通过密集特征自适应连接模块实现对细粒度特征的融合.

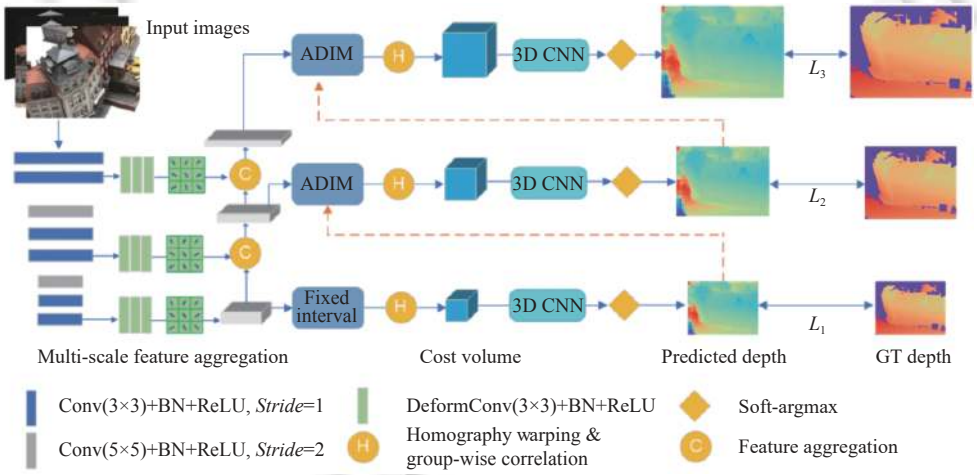


图 6 ADIM-MVSNet 网络结构[30]

注意力机制在一些视觉任务中取得了更好的表现. 文献 [32] 构建了图像金字塔分层提取特征, 并引入自注意力层用于学习聚焦更重要的特征信息. DRI-MVSNet[33]使用结合通道注意力机制和基于空间池化网络的 CSCP 模块, 获得通道和空间信息. ATLAS-MVSNet[34]基于 UNet 结构, 在第 2 阶段引入混合注意力块 (HAB), 通过卷积和局部注意力层的组合, 实现对密集特征和重要特征的提取. 而 Transformer 以自注意力机制作为编码器和解码器中

重要的组成部分, 利用注意力和位置机制感知全局上下文信息, 能够表示每个像素位置的重要性, 并可以学习到不同位置之间的依赖关系. TransMVSNet[35]通过特征匹配 Transformer 来聚合图像内部和图像之间的背景信息, 从而降低特征匹配的不确定性. 使用 FMT (feature matching Transformer) 对特征图进行位置编码, 并将其在空间维度上压平, 最后通过注意力块处理特征, 如图 7 所示[36]. MVSTR[36]在全局上下文和三维几何的共同约束下提取特征. 文献 [37] 基于特征金字塔构造了远程注意网络 (long-range attention network, LANet), 有选择地聚合每个位置的特征, 利用像素之间的远程依赖关系, 实现无纹理和遮挡区域的匹配. 鉴于特征提取器 Vision Transformer (ViT) 在二维视觉任务中的表现, 为了更好地学习表示特征, Cao 等人[38]提出了预训练的 ViT 增强 MVS 网络 MVSFormer-H 和 MVSFormer-P, 实现与 FPN 的特征互补. 其中, MVSFormer-P 采用 DINO 作为主干, 通过自监督的方式训练, 能够获得良好的泛化能力. MVSFormer-H 采用 Tiwns 作为主干, 设计多层的 ViT 模型.

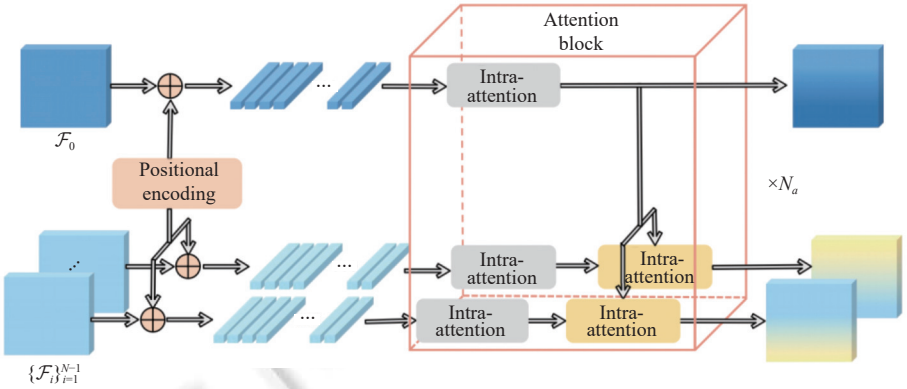


图 7 特征匹配 Transformer 的体系结构[36]

(2) 代价体构建

MVSNet 通过像素方差构建代价体, 构建过程中对每个像素对给予相同的关注, 可能会产生不匹配像素对, 因此无法有效地克服复杂场景下存在遮挡物和无纹理区域的困难. PVSNet[39]估计相邻图像像素水平上的可见性信息, 利用像素的可见性构建加权代价体. Wang 等人在 PatchmatchNet[40]中提出在多尺度特征提取的基础上引入自适应代价聚合模块, 将像素空间窗口组织为网格. 这种方法能够在较大空间中聚合代价体, 进而减少模糊. AA-RMVSNet[26]提出了视图间自适应聚合模块, 为了抑制不可靠匹配, 对易产生混淆的像素赋予小的权重, 并对包含关键上下文信息的像素赋予大的权重, 如图 8 所示[41]. ADR-MVSNet[41]的主要贡献是引入多代价体聚合模块, 具体包含 3 个步骤: 首先构造视图间的代价体, 根据基尼系数反映像素的遮挡程度, 并通过 CNN 获得遮挡图, 最后聚合多个可视代价体为聚合代价体. PVA-MVSNet[42]提出了自适应元素聚合模块, 包括像素聚合和体素聚合两部分. 像素聚合和体素聚合分别在高度宽度和深度方向上通过加权注意力进行平均. UniMVSNet[43]类似于 PVA-MVSNet, 采用自适应聚合扭曲后的特征体方法来处理非朗伯面的不可靠匹配. DRI-MVSNet[33]提出用于特征映射融合的 CVSF 模块, 获得二维相似加权图, 直观地反映了视图之间的相似度. 但逐个寻找每个像素之间的关系容易造成效率低下的问题.

上述方法仅关注二维的局部相似性, 忽略了空间信息. MVSTER[44]利用 3D 的深度信息, 通过 epipolar Transformer 聚合特征体以生成代价体. 通过可微单应性恢复特征图的深度信息, 在 epipolar 对极约束下计算三维关联关系, 并聚合不同视图的特征体. 但没有极线约束容易产生匹配冗余的问题. WT-MVSNet[45]提出基于窗口的 epipolar Transformer (WET), 包括注意力内模块和注意力间模块. 通过不同窗口的交互实现全局特征聚合, 并根据极线约束减少了匹配冗余. 此外, CR-MVSNet[46]提出了共可见性代价聚合模块 (CRCA), 利用多视图之间的可见性关系和单视图的空间上下文信息, 实现更可靠的匹配.

不同于以往基于方差的方法, 有一些研究通过相似度度量生成代价体. AACVP-MVSNet[32]使用相似性度量聚

合两两图像之间的代价. 受群体关联的启发, 文献 [47] 通过群体关联相似度构造轻量级代价体. 对提取的特征微分扭曲, 使用组相关计算相似度得分, 对得分取平均值后计算得到最终代价体. MFNet[27]也采用平均组相关相似性的方式生成代价体, 分组计算特征图的相似性, 减少了代价体的通道数量, 实现代价体的轻量化.

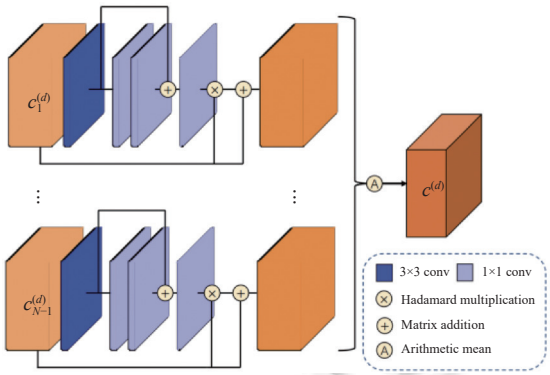


图 8 视图间 AA 模块[41]

MVSNet 的正则化工作是从 3D UNet 结构正则化 3D 代价体开始的. 然而直接使用 3D CNN 结构构建代价体会产生大量的内存消耗和计算需求, 尤其代价体的体积会随着分辨率的增长而增长. 后续研究主要使用两类不同的网络结构缓解该问题, 包括由粗到细的多阶段方法和基于 RNN 的递归方法.

由粗到细的多阶段方法顾名思义, 首先预测较粗的深度分辨率, 在粗分辨率的基础上迭代细化出更精细分辨率的深度图, 减少了搜索范围. CasMVSNet[48]提出了代价-体积公式, 首先生成代价分辨率, 然后使用预测结果自适应的调整深度间隔, 构建出精细的代价体. 将单个代价体使用多阶段级联的方式分解, 通过每个阶段的深度映射影响下一阶段的深度范围. 根据更高空间分辨率的代价体生成精细的输出, 如图 9 所示[49]. UCSNet[49]也提出一种类似思想的方法, 通过 3 个级联阶段, 预测不同尺寸大小的深度图, 并构建 ATV 模块根据上个阶段的输出细化深度. CVP-MVSNet[25]基于特征图像金字塔, 选取最粗分辨率的图像构建代价体积, 然后迭代构建新的代价体, 实现深度预测. 针对均匀采样和假设平面的局限性, SuperMVS[28]在深度范围内非均匀采样建立非均匀代价体, 在采样更精细的同时也降低了平面数量, 以降低计算成本. EPP-MVSNet[50]提出一种集成 Pseudo-3D CNN 的轻量级网络, 分为粗阶段和精细阶段. 采用由粗到细 (coarse-to-fine) 的思想, 在 coarse 阶段采用对极组装模块 (epipolar assembling module, EAM) 以获取高分辨率特征, 从而提高了对高分辨率图像信息的利用. 它还在 fine 阶段引入了一个基于熵的精炼策略 (entropy refining strategy, ER), 以帮助减少构建细代价体的信息冗余. 同时, 在空间和深度维度上, 采用伪 3D 卷积对代价体进行卷积. 同样地, 在 MG-MVSNet[31]中, 也使用分布式 3D 卷积 (D3D) 代替传统卷积, 降低了计算代价. 为了应对多阶段方法在粗阶段可能出现错误预测粗糙深度范围, 从而导致后续细化深度的阶段无法修正的问题, NP-CVP-MVSNet[51]采用无参数概率分布模型来描述深度假设值的概率分布, 并使用稀疏代价体细化深度图.

基于 CNN 的多阶段方法能够有效利用局部信息和多尺度上下文信息, 但是对于处理高分辨率图像会受到内存限制的影响. 卷积门控递归单元 (gate recurrent unit, GRU) 由重置门和更新门组成, 网络的参数量小. 鉴于 GRU 能够实现与 3D CNN 同样的作用, 因此可以采用 GRU 进行递归正则化, 提高重建的效率. R-MVSNet[52]提出通过 GRU 对 2D 代价图进行正则化, 2D 代价图相对于 3D 代价体能够更好地用于高分辨率场景. 图 10 展示了 RMVSNet 的结构[53]. 该网络应用 GRU 的卷积聚合代价图在深度方向上的时间和上下文信息. 为了提高正则化能力, 堆叠了 3 层 GRU 单元. 利用 CNN 处理的结果精度高, 但计算成本高. 利用 GRU 处理, 虽然降低了对内存限制的要求, 但也损失了重建的完整性和精度. 因此, D2HC-RMVSNet[29]结合 3D CNN 和 GRU 的优点, 基于 LSTM 和 UNet 进行改进, 提出了 HU-LSTM 模块, 采用二维 UNet 结构对每层构建更为强大的 LSTMConvCell, 以聚合多尺度上下文

信息. AA-RMVSNet[26]采用 RNN-CNN 的混合网络, 在深度方向上切片. 对每个切片通过编码器-解码器结构的 CNN 正则化, 并采用 RNN 传递 ConvLSTMCell.

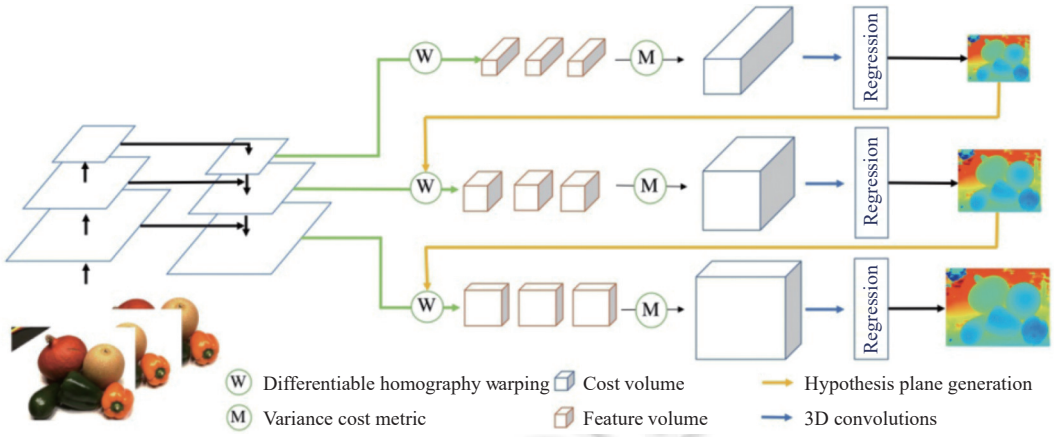


图 9 CasMVSNet 网络结构[49]

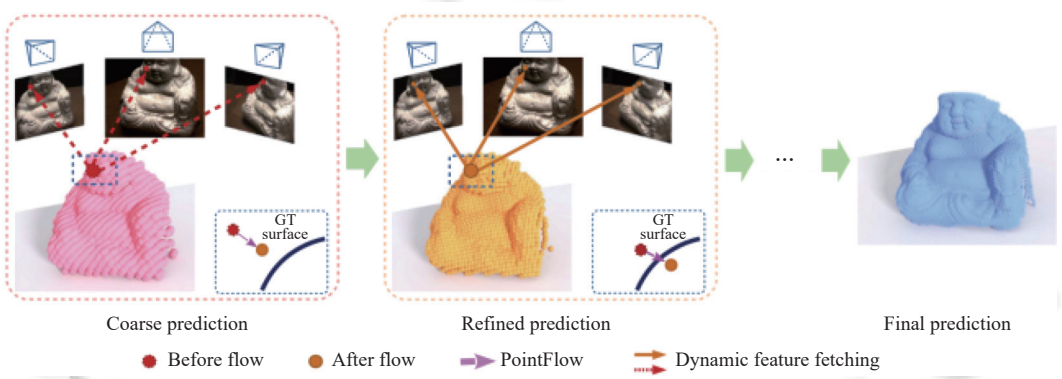


图 10 VA-Point-MVSNet 概述[53]

此外, 有研究尝试了其他改进方法以实现降低内存需求的目标. HighRes-MVSNet[54]使用解码器结构对 4 个阶段进行输出, 每个阶段都与下一阶段的初始代价体融合, 并通过构建级联代价体的方式控制代价体的大小. 此外, 该网络还在深度预测阶段限制每一层次的搜索范围.

(3) 深度估计

模型中常用的深度估计方法包括回归、分类和二者相结合的方法. 深度回归是对概率体沿深度方向执行 soft-argmin 操作, 实现亚像素水平上的深度预测, 即沿深度方向上的期望值. MVSNet 使用 L1 损失函数, 考虑初始深度图和优化后深度图两个方面的损失. 后续大多数方法[25,40,48,53,55,56]也采取回归的方式进行预测. 分类方法将问题看作交叉熵损失的多分类任务, 能够直接约束代价体. 文献 [26,29,35,52] 将问题转化为像素级别的概率分布, 通过 one-hot 编码真值. 但是它们忽略了对深度距离的考虑, 对不同深度平面给予同样的关注. MVSTER[44]将问题定义为深度感知分类问题, 预测深度分布和真值之间的距离, 并据此计算损失.

回归方法容易产生过拟合问题, 而分类方法具有离散性无法实现准确的预测. 因此, 基于传统分类方法和回归方法的优点, 提出了二者相结合的方法. 文献 [57] 中的方法不仅对多模态分布具有鲁棒性, 而且实现了亚像素水平的估计. Peng 等人[43]将深度估计定义为多标签分类任务, 首先分类出最佳的深度范围, 然后回归预测最终的深度值.

不同于上述深度推断方法, 为改进回归任务鲁棒性不强的弱点, 文献 [47] 将问题表述为深度反回归任务. 在深度空间中采样并记录序数, 获取代价映射切片. 通过回归获得亚像素序数, 最后将其转化为深度值. 在回归、分类和二者结合方法的基础上, MVSFormer[38]采用基于温度的深度预测方法, 对不同方法进行选择, 解决了 argmax 不能提供精准的深度值问题.

(4) 其他方法

为了突破 MVS 目前的瓶颈, 一些网络设计不同于典型 MVSNet 的模型被提出来. 有研究使用将深度图转化为点云的方式表示场景. 其中, 3DVNet[58]结合基于深度图和体素的方法, 首先提取图像特征并预测初始的粗深度图, 投影形成特征点云, 通过提出的 3D 场景建模网络和改进的 PointFlow 不断细化深度图. VA-Point-MVSNet[53]将目标场景处理为点云. 如图 10 所示, 它通过 3D 代价体生成粗糙深度图并转化为初始点云, 通过 PointFlow 模块将初始点云迭代回归以得到精细和稠密的点云.

综上, 在开创性工作 MVSNet 的基础上, 监督学习方法主要针对改进网络中的各个阶段模块, 缓解了 MVSNet 网络生成模型过程中存在的问题. 同时, 也出现一些与 MVSNet 结构不同的其他网络, 旨在提高重建 MVS 模型的精度和完整度.

2.2.2 无监督学习

基于监督学习的方法, 在训练阶段依赖于真实数据. 而真值往往难以获取. 目前这类的公开数据集也较为有限, 可能出现模型泛化能力差的问题. 因此, 出现了一系列以无监督的方式学习的方法.

Khot 等人[16]提出一种仅依赖可用的多视角图像作为监督信号的无监督学习方法. 该方法使用光度一致性损失来训练深度预测 CNN, 将原始视图和可用新视图中像素强度之间的差异作为惩罚项. 由于遮挡以及视图间的光照信息不同等问题的存在, 只使用光照一致性不足以约束模型. 为此, 在损失函数中增加了梯度衡量项增强模型的鲁棒性. 文献 [59] 提出第 1 个具有对称网络结构的无监督 MVS 网络, 称为 MVS^2. 该网络遵循 MVSNet 模型, 以一种对称的方式同时预测所有视图的深度图, 包括多尺度特征提取、代价体构建、代价体正则化、优化深度图以及无监督损失评估几个阶段. 为了解决物体的重建边界可能会由于上采样操作而导致的过度平滑的问题, 使用了空间传播网络 (spatial propagation network, SPN) 优化深度图, 并通过特征提取模块引导细化深度图. 针对多视图的遮挡问题, 该网络在深度图中引入交叉视图一致性约束, 并提出损失函数度量一致性. 但是这种方法的缺陷是会消耗大量的 GPU 显存.

为了减少环境的变化对重建结果的影响并且适应于对新目标域的重建, 文献 [60] 提出首先使用元学习在 BlendedMVS 数据集上训练, 之后对获得的模型进行微调. 使用目标域 DTU 数据集的训练数据进行自监督学习, 提高了自监督网络的性能.

M3VSNet[61]扩展了文献 [16] 的工作, 构建一个多度量无监督网络. 其中, 提出一种多度量的损失函数, 结合像素损失函数和特征损失函数, 同时考虑了从纹理和语义信息多个层次进行匹配.

然而, 无监督方法在训练中依赖于视图之间的颜色一致性, 对光照变化高度敏感, 难以应用于光照变化明显的场景. Yang 等人[62]应用 CVP-MVSNet 作为骨干网络, 根据输入图像数据生成伪深度标签, 以实现自监督的深度估计. 该算法由用于初始伪标签估计的无监督学习和用于自训练的迭代伪标签改进两个阶段组成.

通过比较监督和自监督方法, U-MVS[63]被提出. 它采用代表性的 MVSNet 作为主干网络, 主要包括两个阶段: 自监督训练前阶段和伪标签训练后阶段. 针对前景监督模糊的问题, 该方法提出了流深度一致性损失. 针对后台监督无效的问题, U-MVS 使用 MonteCarlo-Dropout 来估计不确定性映射, 并从监督中过滤掉不确定性部分.

针对无监督重建算法容易无法寻找到不同视图之间的精确对应关系的问题, PatchMVSNet[64]引入基于块的光度一致性损失函数, 以帮助减少模糊匹配.

无监督方法大都依赖于不同视图之间的对应点共享相同颜色的假设, 但这并不适用于所有现实情况. 因此, Xu 等人[65]提出一种联合数据增强和协同分割的自监督 MVS 框架 JDACS. 该网络在自监督损失中加入语义一致性和数据增强一致性约束, 以解决颜色恒常不稳定的问题.

知识蒸馏将复杂教师模型的输出作为简单学生模型的训练目标, 能够有效地提升模型性能. KD-MVS[66]提出

了基于知识蒸馏的自监督网络, 包括自监督教师训练和基于蒸馏的学生训练. 其中, 教师模型根据光度一致性和特征一致性进行自监督训练, 并将教师模型中的知识转移到学生模型中.

表 2 中总结了无监督多视图立体视觉的方法. 无监督方法的关键是准确对应视图之间的关系. 目前, 无监督学习的方法已经能够有效缓解监督模糊和光照变化敏感等问题.

表 2　无监督和自监督多视图立体视觉方法对比

方法	特点	优点	缺点
Unsup_MVS[16]	利用多视图间光度一致性	不再依赖有标签数据	对光照变化敏感
MVS^2[59]	对所有视图对称	使深度图有相同的3D几何结构	消耗大量的GPU内存
M^{3}VSNet[61]	加入语义信息的特征损失	能够通过特征感知场景	对光照变化敏感
U-MVSNet[63]	引入流深度一致性损失和不确定性感知的自训练损失	处理前景监督模糊和后景监督无效问题	需要预训练语义特征
KD-MVS [66]	采用知识蒸馏的思想	结果超越教师模型	不适用于小规模数据集

2.3　基于辐射场的方法

随着隐式表示在三维重建领域的发展, 一些研究者尝试将隐式表示引入 MVS 中. 神经辐射场 (neural radiance fields, NeRF) 通过全连接多层感知机隐式的学习三维场景, 能够实现对三维场景的连续表达. 与 MVS 相比, NeRF 利用光线直接建模的方式, 能够更准确地捕捉场景细节. 2020 年, Mildenhall 等人[67]提出了 NeRF, 根据场景中 3D 点的位置和视角方向预测点的密度和颜色, 然后通过体渲染的方法沿相机光线积分得到像素颜色, 从而合成新视角下的图像. 其网络结构如图 11 所示[67]. NeRF 的出现使得图像的生成和渲染更加真实. 由于 NeRF 能够直接对场景进行推理, 有研究将其与 MVS 结合以解决 MVS 无法有效处理无纹理区域和场景细节的问题. Chen 等人[17]提出一种神经渲染方法 MVSNeRF, 利用神经辐射场, 通过构建的代价体对几何感知场景进行推理, 并结合基于物理的体渲染方法实现神经辐射场重建. 通过 3D 卷积构建神经编码体, 同时采用多层感知器将编码体的插值特征回归, 以获得体密度和 RGB 亮度, 进而用于最后的渲染. 但 MVSNeRF 的局限性在于代价体的大小固定, 不能对不可视区域进行渲染. 因此, NeuralMVS[68]提出将附近像素特征聚合到射线上, 可以实现场景建模不受输入视图的影响. 此外, NeuralMVS 还提出一种从粗到细恢复深度的方法. 该方法基于球体跟踪, 能够显著提高恢复几何形状的速度. 为了减少代价计算的成本耗费, RayMVSNet[69]通过基于光线的深度优化表示 MVS, 在相机光线上学习一维隐式场, 并估计出采样点的有向距离场 (signed distance field, SDF) 和过零点的位置. 为了获得高质量的深度估计, Change 等人[70]提出一种神经渲染方法 RC-MVSNet, 由基于 CasMVSNet 的主干和基于神经辐射场的辅助分支组成. 通过引入高斯均匀混合采样学习靠近物体曲面的几何特征以减轻遮挡. 同时, 为了减少光度监督模糊程度, 通过参考视图合成损失直接学习场景的几何特征.

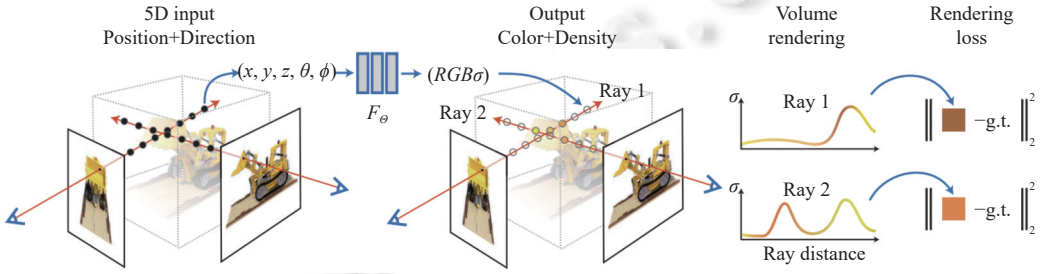


图 11　NeRF 网络结构图[67]

虽然 NeRF 能够连续表示体场景, 但是通过隐式表示结合 MVS 构建三维场景的方法无法精细地表达表面细节, 因此之后又有一系列研究对基于神经体渲染方法进行改进. Wang 等人[71]提出一种神经表面重建方法 NeuS, 利用有符号距离函数的零水平集表示曲面, 通过隐式 SDF 渲染图像, 提高渲染图像和输入图像的一致性, 从而实

现表面的重建. 同样, Yariv 等人[72]将体积密度建模为到场景表面的有符号距离函数, 利用累积分布函数 cumulative distribution function, CDF) 作为 SDF 和体密度之间的变换函数, 不仅能够提供更精确的几何近似, 还能约束不透明估计误差, 避免表面无采样点的问题. Michael 等人[73]提出了统一神经隐式表面和辐射场 UNISURF (unified neural implicit surface and radiance fields), 构建了一种近似表面渲染的体渲染方法, 利用占用网格对体积和表面进行渲染, 从而更有效地细化物体表面. UNISURF 的流程图如图 12 所示[73]. 虽然上述方法解决了重建表面粗糙的问题, 但无法有效地应对稀疏视角下的表面重建问题. SparseNeuS[74]构建了几何编码器, 通过学习图像特征中的可泛化先验预测重建曲面的网络编码符号距离函数. 虽然 SparseNeuS 能够实现跨场景的泛化, 但会受到特征体分辨率的影响. S-VolSDF[75]利用全局一致性约束将 VolSDF 整合到 MVS 中, 通过 MVS 的预测值对神经体积表面进行正则化优化, 提高场景重建质量. VolRecon[76]通过有符号射线距离函数 (SRDF) 实现可泛化的隐式重建, 构建全局特征体, 引入插值特征和沿射线采样点的投影特征计算 SRDF, 进而对密度函数建模.

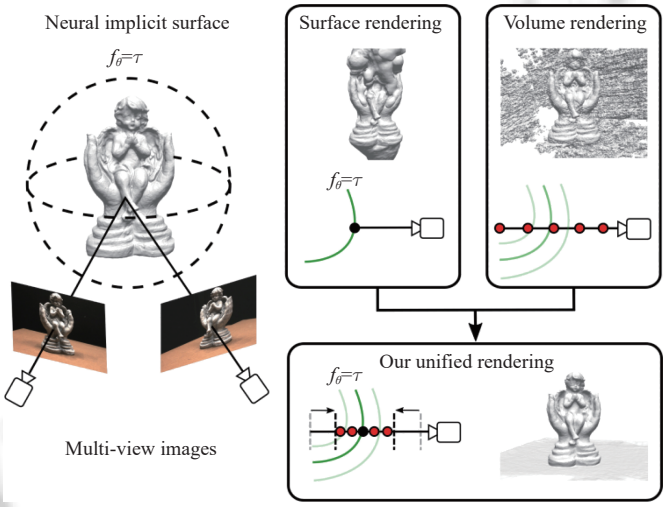


图 12 UNISURF 流程图[73]

NeRF 利用隐式辐射场表示场景, 虽然能够更加逼真地实现对场景的呈现, 但其高昂的计算成本也不可忽视. 3D GS (Gaussian splatting) 利用 3D 高斯结合隐式辐射场和显式辐射场的优势对场景建模, 在实现辐射场优化的同时也减少了计算开销. 2023 年, Kerbl 等人[77]提出了 3D Gaussian Splatting, 引入 3D 高斯函数表示场景. 该方法将高斯函数参数优化和高斯分布密度控制交替进行, 并在优化过程中实现快速的渲染, 从而实现新视图的合成. 因其高效的渲染效率, 3D GS 被应用到 MVS 领域. Cheng 等人[18]提出了 GaussianPro, 通过渐进传播的方式构建三维高斯分布, 从而能够实现将建模精细区域的几何信息传递到建模稀疏区域. 网络结构如图 13 所示[18].

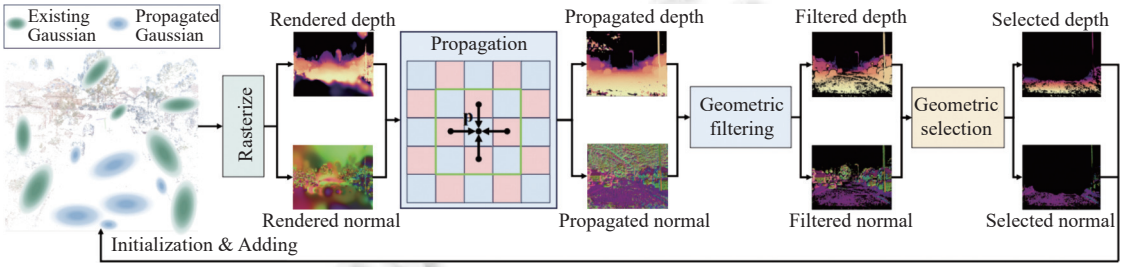


图 13 高斯函数渐进传播网络结构图[18]

3 常用数据集及评价指标

为了促进多视图立体视觉领域的快速发展, 并建立性能基准, 领域内已经制作生成了几个广泛使用的数据集.

表 3 总结了常用数据集, 包括 Middlebury MVS[78]、EPFL benchmark[79]、DTU[23]、Tanks and Tampels[80]、ETH3D[81]、BlendedMVS[82]等. 这些数据集涵盖了室内和室外场景以及合成数据. 本节分别从数据集的场景、数据规模、分辨率等方面进行介绍. 与此同时, 本文提供了每个数据集用于评估模型性能的基本指标.

表 3 多视图立体视觉常用数据集

数据集名称	场景	评估指标	年份
Middlebury MVS[78]	寺庙、恐龙	精度、完整度	2006
EPFL benchmark[79]	建筑物	精度	2008
DTU[23]	室内小目标场景	精度、完整度、*F-score*	2016
Tanks and Tampels[80]	室内场景和室外环境	精度、完整度、*F-score*	2017
ETH3D[81]	自然和人造的室内外环境	精度、完整度、*F-score*	2017
BlendedMVS[82]	室内和室外的合成场景	精度、完整度、*F-score*	2020

3.1 数据集

3.1.1 Middlebury MVS

Middlebury 是最早用于多视图立体视觉评估的数据集, 如图 14 所示[78]. 数据集在室内收集了 6 种视角的 790 张分辨率为 640×480 像素的图像以及对应的 3D 网格模型真值. 图像由定位在机械臂上相机拍摄获得. 参考 3D 模型则采用激光扫描仪获取.

图 14 Middlebury 数据集示例[78]

3.1.2 EPFL benchmark

EPFL benchmark 是建筑外侧的场景集合, 如图 15 所示[79], 包括对重建结果的评估. 数据集提供分辨率为 3072×2028 像素的高分辨率图像, 同时使用激光扫描 (LiDAR) 获取户外场景的真值. 使用激光雷达数据, 通过摄像机校准的平均值和方差生成图像的真值.

图 15 EPFL benchmark 数据集示例[79]

3.1.3 DTU

DTU 是在实验室环境中获得的小目标场景数据集. 如图 16 所示[23], 它具有 128 个不同反射率、纹理和几何属性的室内场景. 每个场景在 7 种不同的光照条件下, 由工业机械臂设定的 49 或 64 个镜头位置 (即固定摄像机轨迹) 进行多视角拍摄, 并由结构光扫描获得表面点云. 最终, 生成分辨率为 1200×1600 像素的 RGB 彩色图像, 并提供结构光标注. 数据集通过泊松重建来重建曲面以获取网格模型, 并由网络模型渲染生成的深度图真值.

图 16 DTU 数据集示例[23]

3.1.4 Tanks and Temples

Tanks and Temples 数据集的提出是为了推动大规模场景重建的研究. 数据集包括室内场景和室外环境, 并使用工业激光扫描仪获得真值数据. 同时, 该数据集提供了高分辨率的视频, 分辨率为 800 万像素. 相较于图像, 视频的数据冗余能够覆盖场景中更广的范围. 制作者根据场景规模, 从多个视角位置对场景进行多次扫描, 注册获得真值模型. 根据场景的规模、复杂性和其他复杂因素将数据集分为中级组 (8 个场景) 和高级组 (6 个场景). 其中, 中级组包括雕塑、车辆和房屋建筑, 高级组包括大型室内室外场景, 如图 17 所示[80].

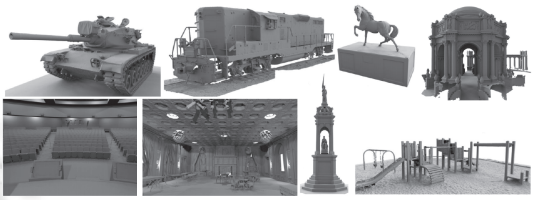


图 17 Tanks and Temples 数据集示例[80]

3.1.5 ETH3D

ETH3D 基准涵盖了包括自然、人造、室内外的多种场景类型, 如图 18 所示[81]. 数据集包括单反相机记录的高分辨率多视角立体视觉图像、多摄像机设备录制的低分辨率的多视图立体视觉视频和多摄像机视频帧上的低分辨率的双视图立体视觉图像. 利用单反相机获得 25 个场景的 2 400 万像素的高分辨率图像. 利用摄像机设备获得 10 个场景的 40 万像素的低分辨率图像. 高精度激光扫描仪记录了室内和室外场景的真值点云. 该数据集提供在线评估网站.

图 18 ETH3D 数据集示例[81]

3.1.6 BlendedMVS

BlendedMVS 的建立旨在推进大规模数据集任务, 如图 19 所示[82]. 该数据集属于合成数据集, 不包含真值点

云. 具体来说, 数据集包含 17818 张的高分辨率图像, 由 113 个包括城市、建筑、雕塑和小物体等室外场景的模型组成. 每个模型包含 20–1 000 张不等的图像, 由沿非结构化轨迹的摄像机捕获. 为了实现视景照明, 将原始图像混合真实的环境光照信息. 制作者利用场景图像恢复纹理网络模型, 并据此渲染生成分辨率为 1536×2048 的彩色图像和深度图.

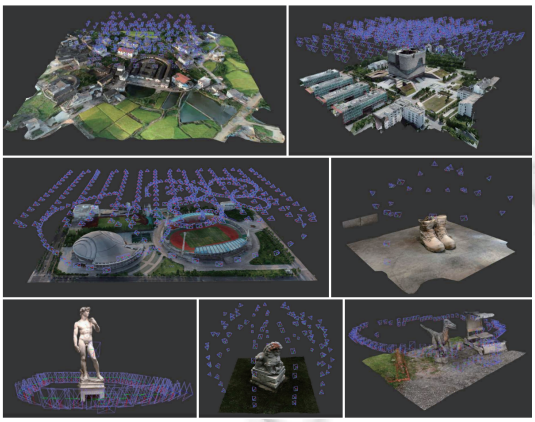


图 19 BlendedMVS 数据集示例[82]

3.2 MVS 重建效果的评价指标

常见的数据集评价指标包括精度、完整度和 *F-score* 综合评价. 针对不同的数据集, 目前评价指标还没有统一的标准. 本节将介绍上述数据集中所使用的评价标准.

按照距离阈值区间范围评估重建的精度和完整度. 其中, 精度代表待评估模型与真值模型之间的接近程度, 定义为真值点与重建点距离小于给定阈值的分数, 表示如下:

$$P(d)=\frac{100}{|R|}\sum_{r\in R}[e_{r\to R}<d] \tag{4}$$

其中, R 为重建点集合, $e_{r\to R}$ 为从重建点到真值的距离, d 为距离阈值, $[\cdot]$ 为艾弗森括号.

完整度表示待评估模型与真值模型中重合度, 定义为真值点云中的点到距离其最近重建点的距离低于阈值的数量, 表示如下:

$$R(d)=\frac{100}{|G|}\sum_{g\in G}[e_{g\to G}<d] \tag{5}$$

其中, G 为真值, $e_{g\to G}$ 表示从重建点到真值的距离, d 为距离阈值.

同时, 引入 *F-score* 综合评价重建质量, 定义为精度和召回率的调和平均值, 表示如下:

$$F(d)=\frac{2P(d)R(d)}{P(d)+R(d)} \tag{6}$$

其中, $P(d)$ 表示对任意距离阈值 d 的重建精度, $R(d)$ 表示为重建的完整度. 它结合了两者的特性, 只有准确又完整的重建才能获得较高的 *F-score*. 调和平均值是算术平均值的一个变形, 能够有效地避免极端值的影响, 从而更好地反映重建质量.

4 算法性能对比

本节列出了前面描述的典型模型的评价结果, 并进行了总结和讨论. 本文使用该领域内目前最为广泛使用的数据集: Tanks and Temples 和 DTU 数据集对算法性能进行对比, 并在两个数据集上总结出表现最好的模型. 表 4 和表 5 分别列出上述典型的基于监督学习和无监督学习方法的基线模型在 Tanks and Temples 和 DTU 数据集上的测试结果.

表 4 有监督多视图立体视觉方法在数据集上的性能

方法	Tanks and Temples	DTU		
	Mean (%)	*Accuracy* (mm)	*Completeness* (mm)	*Overall score* (mm)
MVSNet[15]	43.48	0.396	0.527	0.462
MVSCRF[24]	45.73	0.371	0.426	0.398
P-MVSNet[55]	55.62	0.406	0.434	0.420
RMVSNet[52]	50.55	0.385	0.459	0.422
CasMVSNet[48]	56.42	0.325	0.385	0.355
CVP-MVSNet[25]	54.03	0.296	0.406	0.351
Fast-MVSNet[56]	47.39	0.336	0.403	0.370
MVSNet++[83]	49.12	0.376	0.345	0.407
PatchmatchNet[40]	53.15	0.427	0.277	0.352
D^2HC-RMVSNet[29]	59.20	0.395	0.378	0.386
UCSNet[49]	54.83	0.338	0.349	0.344
SurfaceNet+[22]	49.38	0.385	0.448	0.416
HighRes-MVSNet[54]	49.81	0.354	0.393	0.373
AA-RMVSNet[26]	61.51	0.376	0.339	0.357
MVSTR[36]	56.93	0.356	0.295	0.326
LANet[37]	55.70	0.320	0.349	0.335
TransMVSNet[35]	63.52	0.321	0.289	0.305
MVSTER[44]	60.92	0.350	0.276	0.313
IterMVS[57]	56.22	0.373	0.354	0.363
UniMVSNet[43]	64.36	0.352	0.278	0.315
WT-MVSNet[45]	65.34	**0.309**	0.281	0.295
MVSFormer-H[38]	**66.37**	0.327	**0.251**	**0.289**

表 5 无监督和自监督多视图立体视觉方法在数据集上的性能

方法	Tanks and Temples	DTU		
	Mean (%)	*Accuracy* (mm)	*Completeness* (mm)	*Overall score* (mm)
Unsup_MVS[16]	—	0.881	1.073	0.997
MVS^2[59]	37.21	0.760	0.515	0.637
M^3VSNet[61]	37.67	0.636	0.531	0.583
Self-supervised-CVP-MVSNet[62]	46.71	**0.308**	0.418	0.363
SurRF[84]	54.36	0.388	0.390	0.389
JDACS[65]	45.48	0.571	0.515	0.543
U-MVS[63]	57.15	0.354	0.353 5	0.353 7
PatchMVSNet[64]	40.26	0.538	0.365	0.451
RC-MVSNet[70]	55.04	0.396	**0.295**	0.345
KD-MVS (CasMVSNet)[66]	**64.14**	0.359	**0.295**	**0.327**

表 4 以时间顺序列出有监督的模型, 可以看出以下几点.

(1) 近几年提出的大多数模型的精度和完整度较高, *F-score* 与最初的模型相比也有显著的进步, 这表明该领域的发展十分迅速.

(2) 多视图立体视觉的一个难点之一是如何有效地重建弱纹理或无纹理区域. 这部分区域的特征较为不明显, 通过多尺度上下文信息和图像的特征关系, 并结合空间维度上的信息, 能够增强对特征的提取.

(3) 复杂场景下存在大量遮挡和无纹理区域, 这些区域难以被处理. 在构建代价体阶段, 能够通过学习可见性信息调整来实现可靠的匹配.

(4) 通过 Transformer 方法, 不仅能够更好地表示特征, 而且可以有效地缓解对无纹理和遮挡区域的特征匹配问题. 从表 4 可以看出, 结合 Transformer 的方法普遍优于其他方法, 尤其是 MVSFormer-H 模型取得了最优的 *F-score*, 比最初的 MVSNet 模型有 34.4% 的提升.

(5) 复杂的大规模场景信息庞大, 导致构建 3D 代价体造成大量的内存和时间消耗. 提出和引入的多阶段方法能够从粗到细地预测深度, 或利用 GRU 结构实现递归正则化, 这使得模型的精度和完整度有了很大的提升.

在数据集匮乏的情况下, 无监督方法更为适用. 表 5 中列出了一系列无监督方法, 可以看出以下几点.

(1) 无监督方法尽管没有达到有监督方法的性能, 但是在训练数据缺失的情况下, 重建结果能够达到一定的精度要求.

(2) 早期提出的网络使用光照一致性作为损失, 导致对光照敏感度较高, 无法泛化到真实的环境中. 后续网络提出其他一致性损失方式, 帮助实现有效的监督.

(3) 无监督方法精度和完整度无法达到有监督方法的主要原因是监督模糊和监督无效, 对此进行改进能使模型的性能进一步提升.

(4) KD-MVS 采用知识蒸馏的思想, 利用教师模型的输出引导学生模型, 在 DTU 和 Tanks and Temples 数据集中都取得了目前最好的效果.

(5) RC-MVSNet 的表现仅次于 KD-MVS, 引入神经渲染的, 有效避免了监督模糊, 并为多视图立体视觉带来新的活力.

(6) 目前, 相对于有监督学习的方法, 基于无监督学习的相关方法较少, 还有进一步提升的空间.

5 总结与展望

在过去几十年里, 基于深度学习的多视图立体视觉得到了迅速发展, 取得了显著成就. 为了取得更广的应用范围和更高质量的应用效果, 要求算法具有更高精度的点云模型和更小的内存消耗. 为了在该领域实现更高的精度和效率, 研究学者们提出了一系列方法. 随着在医疗领域、建筑领域、自动驾驶领域等的应用逐渐广泛, 多视图立体视觉也有了更大的发展空间. 本文回顾了近年来基于深度学习多视图立体视觉具有代表性的方法、数据集、评价指标等的研究进展. 在文中首先介绍了基于传统方法的多视图立体视觉的发展现状, 提出结合深度学习的必要性. 描述了针对传统方法的局部功能改进方法, 以及基于深度图、基于体素和基于辐射场深度学习方法的整体架构改进方法. 为了更好地组织文献, 将现有的模型分类. 对每个类别, 介绍了优化网络的不同方法, 并总结了部分贡献的主要思想. 随后描述了广泛使用的数据集, 并梳理了评估这些数据集所用的指标. 最后, 在两个数据集上对多视图立体视觉领域的多个方法进行了对比分析.

基于深度学习的方法对比传统重建方法取得了巨大的突破, 但是仍然存在一些困难有待进一步深入研究. 本文也总结了多视图立体视觉技术目前所面临的挑战, 同时展望未来可能的研究方向.

(1) 现有方法在个别区域和情况下重建效果不佳, 如低特征、反射、遮挡、弱光照或透明表面等造成的重建模型表面空洞; 同时, 相邻视角之间变化剧烈、可见性差的情况也会影响重建效果. 因此, 有必要对这些情况下的重建问题进行进一步研究.

(2) 现有重建方法未能充分利用全局信息. 由于全局上下文信息 (视图内部和视图之间) 利用的不足和 3D 表示的不一致, 导致仍然存在特征匹配歧义和不匹配的问题, 而不同视图中不匹配像素则会造成重建的不完全.

(3) 目前基于高分辨率图像重建精细化场景的效率不高. 虽然相关研究工作在减少内存消耗方面已有所进展, 但当使用高分辨率图像重建时, 仍会产生较大的内存和计算需求. 然而当使用效率较高的模型时则会损失高分辨率图像中的细节信息, 影响模型精度. 在应用阶段想要实现方法的真正落地, 需要平衡效率和精度, 在保证空间分辨率同时使用网络复杂度较低、内存效率较高的模型.

(4) MVS 公开数据集的数量有限且评价指标不一致. 现有的实拍数据集由于受到设备限制, 可用于训练测试的数量有限同时场景也有局限性, 无法支撑具有更强泛化性的通用模型训练, 也无法获得如天空等难以重建区域的标签. 而合成数据集虽然降低了采样的成本, 但无法真实地反映自然图像的光照效果和噪声. 同时, 多种类型的数据集也造成了不同数据集下的评估指标不一致的问题. 如果能够创建更具有全面性、真实性和多样性的大规模数据集供模型训练和测试, 预期三维重建的性能指标会进一步提升.

随着计算机技术的快速发展和模型的不断完善, 这些年也涌现出如 Transformer 等功能更强大的神经网络结构. NeRF 和 GS 的出现也使场景表征技术得到了跨越式发展, 与之结合的多视角三维重建具有巨大的发展潜力. 此外, 如何创建更为广泛、通用的标准数据集并设立一致的评估指标也是之后发展的方向. 可以预见, 未来多视图立体视觉发展会更加成熟, 也必将应用于更多领域.

References:

[1] Peng Y, Wang AD, Wang TT, Li JL, Wang ZQ, Zhao Y, Wang ZL, Zhao Z. Three-dimensional reconstruction of carp brain tissue and brain electrodes for biological control. Journal of Biomedical Engineering, 2020, 37(5): 885–891 (in Chinese with English abstract). [doi: 10.7507/1001-5515.201911011]

[2] Nicholson DT, Chalk C, Funnell WRJ, Daniel SJ. Can virtual reality improve anatomy education? A randomised controlled study of a computer-generated three-dimensional anatomical ear model. Medical Education, 2006, 40(11): 1081–1087. [doi: 10.1111/j.1365-2929.2006.02611.x]

[3] Qu YF, Huang JY, Zhang X. Rapid 3D reconstruction for image sequence acquired from UAV camera. Sensors, 2018, 18(1): 225. [doi: 10.3390/s18010225]

[4] Carvajal-Ramírez F, Navarro-Ortega AD, Agüera-Vega F, Martínez-Carricondo P, Mancini F. Virtual reconstruction of damaged archaeological sites based on unmanned aerial vehicle photogrammetry and 3D modelling. Study case of a southeastern Iberia production area in the Bronze Age. Measurement, 2019, 136: 225–236. [doi: 10.1016/j.measurement.2018.12.092]

[5] Gao ZP, Zhai GT, Deng HW, Yang XK. Extended geometric models for stereoscopic 3D with vertical screen disparity. Displays, 2020, 65: 101972. [doi: 10.1016/j.displa.2020.101972]

[6] Wang X, Wang C, Liu B, Zhou XQ, Zhang L, Zheng J, Bai X. Multi-view stereo in the deep learning era: A comprehensive review. Displays, 2021, 70: 102102. [doi: 10.1016/j.displa.2021.102102]

[7] Furukawa Y, Hernández C. Multi-view stereo: A tutorial. Foundations and Trends® in Computer Graphics and Vision, 2015, 9(1–2): 1–148. [doi: 10.1561/0600000052]

[8] Žbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1592–1599. [doi: 10.1109/CVPR.2015.7298767]

[9] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4353–4361. [doi: 10.1109/CVPR.2015.7299064]

[10] Han XF, Leung T, Jia YQ, Sukthankar R, Berg AC. MatchNet: Unifying feature and metric learning for patch-based matching. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3279–3286. [doi: 10.1109/CVPR.2015.7298948]

[11] Murphy K, Schölkopf B, Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. The Journal of Machine Learning Research, 2016, 17(1): 2287–2318.

[12] Güney F, Geiger A. Displets: Resolving stereo ambiguities using object knowledge. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4165–4175. [doi: 10.1109/CVPR.2015.7299044]

[13] Luo WJ, Schwing AG, Urtasun R. Efficient deep learning for stereo matching. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5695–5703. [doi: 10.1109/CVPR.2016.614]

[14] Ji MQ, Gall J, Zheng HT, Liu YB, Fang L. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2326–2334. [doi: 10.1109/ICCV.2017.253]

[15] Yao Y, Luo ZX, Li SW, Fang T, Quan L. MVSNet: Depth inference for unstructured multi-view stereo. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 785–801. [doi: 10.1007/978-3-030-01237-3_47]

[16] Khot T, Agrawal S, Tulsiani S, Mertz C, Lucey S, Hebert M. Learning unsupervised multi-view stereopsis via robust photometric consistency. arXiv:1905.02706, 2019.

[17] Chen AP, Xu ZX, Zhao FQ, Zhang XS, Xiang FB, Yu JY, Su H. MVSNeRF: Fast generalizable radiance field reconstruction from multi-

view stereo. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 14104–14113. [doi: 10.1109/ICCV48922.2021.01386]

[18] Cheng K, Long XX, Yang KZ, Yao Y, Yin W, Ma YX, Wang WP, Chen XJ. GaussianPro: 3D Gaussian splatting with progressive propagation. arXiv:2402.14650, 2024.

[19] Choy CB, Xu DF, Gwak J, Chen K, Savarese S. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 628–644. [doi: 10.1007/978-3-319-46484-8_38]

[20] Kar A, Häne C, Malik J. Learning a multi-view stereo machine. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 364–375.

[21] Paschalidou D, Ulusoy AO, Schmitt C, van Gool L, Geiger A. RayNet: Learning volumetric 3D reconstruction with ray potentials. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3897–3906. [doi: 10.1109/CVPR.2018.00410]

[22] Ji MQ, Zhang JZ, Dai QH, Fang L. SurfaceNet+: An end-to-end 3D neural network for very sparse multi-view stereopsis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(11): 4078–4093. [doi: 10.1109/TPAMI.2020.2996798]

[23] Jensen R, Dahl A, Vogiatzis G, Tola E, Aanæs H. Large scale multi-view stereopsis evaluation. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 406–413. [doi: 10.1109/CVPR.2014.59]

[24] Xue YZ, Chen JS, Wan WT, Huang YQ, Yu C, Li TP, Bao JY. MVSCRF: Learning multi-view stereo with conditional random fields. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4311–4320. [doi: 10.1109/ICCV.2019.00441]

[25] Yang JY, Mao W, Alvarez JM, Liu MM. Cost volume pyramid based depth inference for multi-view stereo. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4748–4760. [doi: 10.1109/TPAMI.2021.3082562]

[26] Wei ZH, Zhu QT, Min C, Chen YS, Wang GP. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 6167–6176. [doi: 10.1109/ICCV48922.2021.00613]

[27] Cai YC, Li L, Wang D, Liu XP. MFNet: Multi-level fusion aware feature pyramid based multi-view stereo network for 3D reconstruction. Applied Intelligence, 2023, 53(4): 4289–4301. [doi: 10.1007/s10489-022-03754-3]

[28] Zhang T. SuperMVS: Non-uniform cost volume for high-resolution multi-view stereo. arXiv:2203.14331, 2022.

[29] Yan JF, Wei ZZ, Yi HW, Ding MY, Zhang RZ, Chen YS, Wang GP, Tai YW. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 674–689. [doi: 10.1007/978-3-030-58548-8_39]

[30] Cheng W, Bai ZY, Li JJ, Liu HJ, Yang LF. ADIM-MVSNet: Adaptive depth interval multi-view stereo network for 3D reconstruction. In: Proc. of the 5th Int'l Conf. on Image and Graphics Processing. Beijing: ACM, 2022. 281–287. [doi: 10.1145/3512388.3512429]

[31] Zhang XD, Yang FZ, Chang M, Qin XF. MG-MVSNet: Multiple granularities feature fusion network for multi-view stereo. Neurocomputing, 2023, 528: 35–47. [doi: 10.1016/j.neucom.2023.01.062]

[32] Yu AZ, Guo WY, Liu B, Chen X, Wang X, Cao XF, Jiang BC. Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 175: 448–460. [doi: 10.1016/j.isprsjprs.2021.03.010]

[33] Li Y, Li WY, Zhao ZJ, Fan JH. DRI-MVSNet: A depth residual inference network for multi-view stereo images. PLoS One, 2022, 17(3): e0264721. [doi: 10.1371/journal.pone.0264721]

[34] Weilharter R, Fraundorfer F. ATLAS-MVSNet: Attention layers for feature extraction and cost volume regularization in multi-view stereo. In: Proc. of the 26th Int'l Conf. on Pattern Recognition. Montreal: IEEE, 2022. 3557–3563. [doi: 10.1109/ICPR56361.2022.9956633]

[35] Ding YK, Yuan WT, Zhu QT, Zhang HT, Liu XY, Wang YJ, Liu X. TransMVSNet: Global context-aware multi-view stereo network with transformers. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8575–8584. [doi: 10.1109/CVPR52688.2022.00839]

[36] Zhu J, Peng B, Li WQ, Shen HF, Zhang Z, Lei JJ. Multi-view stereo with transformer. arXiv:2112.00336, 2021.

[37] Zhang XD, Hu YT, Wang HC, Cao XB, Zhang BC. Long-range attention network for multi-view stereo. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3781–3790. [doi: 10.1109/WACV48630.2021.00383]

[38] Cao CJ, Ren XL, Fu YW. MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth. arXiv:2208.02541, 2022.

[39] Xu QS, Tao WB. PVSNet: Pixelwise visibility-aware multi-view stereo network. arXiv:2007.07714, 2020.

[40] Wang FJH, Galliani S, Vogel C, Speciale P, Pollefeys M. PatchmatchNet: Learned multi-view patchmatch stereo. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14189–14198. [doi: 10.1109/CVPR46437.2021.01397]

[41] Li Y, Zhao ZJ, Fan JH, Li WY. ADR-MVSNet: A cascade network for 3D point cloud reconstruction with pixel occlusion. Pattern Recognition, 2022, 125: 108516. [doi: 10.1016/j.patcog.2021.108516]

[42] Yi HW, Wei ZZ, Ding MY, Zhang RZ, Chen YS, Wang GP, Tai YW. Pyramid multi-view stereo net with self-adaptive view aggregation. In: Proc. of the 2020 European Conf. on Computer Vision. Glasgow: Springer, 2020. 766–782. [doi: 10.1007/978-3-030-58545-7_44]

[43] Peng R, Wang RJ, Wang ZY, Lai YW, Wang RG. Rethinking depth estimation for multi-view stereo: A unified representation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8635–8644. [doi: 10.1109/CVPR52688.2022.00845]

[44] Wang XF, Zhu Z, Huang G, Qin FB, Ye Y, He YJ, Chi X, Wang XG. MVSTER: Epipolar transformer for efficient multi-view stereo. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 573–591. [doi: 10.1007/978-3-031-19821-2_33]

[45] Liao JL, Ding YK, Shavit Y, Huang DH, Ren SH, Guo J, Feng WS, Zhang K. WT-MVSNet: Window-based Transformers for multi-view stereo. Advances in Neural Information Processing Systems, 2022, 35: 8564–8576.

[46] Liu YM, Rao Y, Rigall E, Fan H, Dong JY. Incorporating co-visibility reasoning into surface depth measurement. IEEE Trans. on Instrumentation and Measurement, 2023, 72: 5009912. [doi: 10.1109/TIM.2023.3250231]

[47] Xu QS, Tao WB. Learning inverse depth regression for multi-view stereo with correlation cost volume. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 12508–12515. [doi: 10.1609/aaai.v34i07.6939]

[48] Gu XD, Fan ZW, Zhu SY, Dai ZZ, Tan FT, Tan P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2492–2501. [doi: 10.1109/CVPR42600.2020.00257]

[49] Cheng S, Xu ZX, Zhu SL, Li ZW, Li LE, Ramamoorthi R, Su H. Deep stereo using adaptive thin volume representation with uncertainty awareness. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2521–2531. [doi: 10.1109/CVPR42600.2020.00260]

[50] Ma XJ, Gong Y, Wang QR, Huang JW, Chen L, Yu F. EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5712–5720. [doi: 10.1109/ICCV48922.2021.00568]

[51] Yang JY, Alvarez JM, Liu MM. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8616–8624. [doi: 10.1109/CVPR52688.2022.00843]

[52] Yao Y, Luo ZX, Li SW, Shen TW, Fang T, Quan L. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5520–5529. [doi: 10.1109/CVPR.2019.00567]

[53] Chen R, Han SF, Xu J, Su H. Visibility-aware point-based multi-view stereo network. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3695–3708. [doi: 10.1109/TPAMI.2020.2988729]

[54] Weilharter R, Fraundorfer F. HighRes-MVSNet: A fast multi-view stereo network for dense 3D reconstruction from high-resolution images. IEEE Access, 2021, 9: 11306–11315. [doi: 10.1109/ACCESS.2021.3050556]

[55] Luo KY, Guan T, Ju LL, Huang HP, Luo YW. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 10451–10460. [doi: 10.1109/ICCV.2019.01055]

[56] Yu ZH, Gao SH. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1946–1955. [doi: 10.1109/CVPR42600.2020.00202]

[57] Wang FJH, Galliani S, Vogel C, Pollefeys M. IterMVS: Iterative probability estimation for efficient multi-view stereo. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8596–8605. [doi: 10.1109/CVPR52688.2022.00841]

[58] Rich A, Stier N, Sen P, Höllerer T. 3DVNet: Multi-view depth prediction and volumetric refinement. In: Proc. of the 2021 Int'l Conf. on 3D Vision. London: IEEE, 2021. 700–709. [doi: 10.1109/3DV53792.2021.00079]

[59] Dai YC, Zhu ZD, Rao ZB, Li B. MVS^2: Deep unsupervised multi-view stereo with multi-view symmetry. In: Proc. of the 2019 Int'l Conf. on 3D Vision. Quebec City: IEEE, 2019. 1–8. [doi: 10.1109/3DV.2019.00010]

[60] Mallick A, Stückler J, Lensch HPA. Learning to adapt multi-view stereo by self-supervision. In: Proc. of the 31st British Machine Vision Conf. BMVA Press, 2020. [doi: 10.5555/3600270.3600893]

[61] Huang BC, Yi HW, Huang C, He YJ, Liu JB, Liu X. M^{3}VSNET: Unsupervised multi-metric multi-view stereo network. In: Proc. of the 2021 IEEE Int'l Conf. on Image Processing. Anchorage: IEEE, 2021. 3163–3167. [doi: 10.1109/ICIP42928.2021.9506469]

[62] Yang JY, Alvarez JM, Liu MM. Self-supervised learning of depth inference for multi-view stereo. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7522–7530. [doi: 10.1109/CVPR46437.2021.00744]

[63] Xu HB, Zhou ZP, Wang YL, Kang WX, Sun BG, Li H, Qiao Y. Digging into uncertainty in self-supervised multi-view stereo. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 6058–6067. [doi: 10.1109/ICCV48922.2021.00602]

[64] Dong HN, Yao J. PatchMVSNet: Patch-wise unsupervised multi-view stereo for weakly-textured surface reconstruction. arXiv:2203.02156, 2022.

[65] Xu HB, Zhou ZP, Qiao Y, Kang WX, Wu QX. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 3030–3038. [doi: 10.1609/aaai.v35i4.16411]

[66] Ding YK, Zhu QT, Liu XY, Yuan WT, Zhang HT, Zhang C. KD-MVS: Knowledge distillation based self-supervised learning for multi-view stereo. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 630–646. [doi: 10.1007/978-3-031-19821-2_36]

[67] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: Representing scenes as neural radiance fields for view synthesis. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 405–421. [doi: 10.1007/978-3-030-58452-8_24]

[68] Rosu RA, Behnke S. NeuralMVS: Bridging multi-view stereo and novel view synthesis. In: Proc. of the 2022 Int'l Joint Conf. on Neural Networks. Padua: IEEE, 2022. 1–7. [doi: 10.1109/IJCNN55064.2022.9892024]

[69] Xi JH, Shi YF, Wang YJ, Guo YL, Xu K. RayMVSNet: Learning ray-based 1D implicit fields for accurate multi-view stereo. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8585–8595. [doi: 10.1109/CVPR52688.2022.00840]

[70] Chang D, Božič A, Zhang T, Yan QS, Chen YC, Süsstrunk S, Nießner M. RC-MVSNet: Unsupervised multi-view stereo with neural rendering. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 665–680. [doi: 10.1007/978-3-031-19821-2_38]

[71] Wang P, Liu LJ, Liu Y, Theobalt C, Komura T, Wang WP. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 2081.

[72] Yariv L, Gu JT, Kasten Y, Lipman Y. Volume rendering of neural implicit surfaces. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 367.

[73] Oechsle M, Peng SY, Geiger A. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5569–5579. [doi: 10.1109/ICCV48922.2021.00554]

[74] Long XX, Lin C, Wang P, Komura T, Wang WP. SparseNeuS: Fast generalizable neural surface reconstruction from sparse views. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 210–227. [doi: 10.1007/978-3-031-19824-3_13]

[75] Wu HY, Graikos A, Samaras D. S-VolSDF: Sparse multi-view stereo regularization of neural implicit surfaces. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 3533–3545. [doi: 10.1109/ICCV51070.2023.00329]

[76] Ren YF, Wang FJH, Zhang T, Pollefeys M, Süsstrunk S. VolRecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 16685–16695. [doi: 10.1109/CVPR52729.2023.01601]

[77] Kerbl B, Kopanas G, Leimkuehler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. on Graphics (TOG), 2023, 42(4): 139. [doi: 10.1145/3592433]

[78] Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. New York: IEEE, 2006. 519–528. [doi: 10.1109/CVPR.2006.19]

[79] Strecha C, von Hansen W, Van Gool L, Fua P, Thoennessen U. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Proc. of the 2008 IEEE Conf. on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008. 1–8. [doi: 10.1109/CVPR.2008.4587706]

[80] Knapitsch A, Park J, Zhou QY, Koltun V. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. on Graphics (TOG), 2017, 36(4): 78. [doi: 10.1145/3072959.3073599]

[81] Schöps T, Schönberger JL, Galliani S, Sattler T, Schindler K, Pollefeys M, Geiger A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2538–2547. [doi: 10.1109/CVPR.2017.272]

[82] Yao Y, Luo ZX, Li SW, Zhang JY, Ren YF, Zhou L, Fang T, Quan L. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1787–1796.

[doi: 10.1109/CVPR42600.2020.00186]

[83] Chen PH, Yang HC, Chen KW, Chen YS. MVSNet++: Learning depth-based attention pyramid features for multi-view stereo. IEEE Trans. on Image Processing, 2020, 29: 7261–7273. [doi: 10.1109/TIP.2020.3000611]

[84] Zhang JZ, Ji MQ, Wang GY, Xue ZW, Wang SJ, Fang L. SurRF: Unsupervised multi-view stereopsis by learning surface radiance field. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7912–7927. [doi: 10.1109/TPAMI.2021.3116695]

附中文参考文献:

[1] 彭勇, 王爱迪, 王婷婷, 李京龙, 王占秋, 赵洋, 王子霖, 赵政. 面向生物控制的鲤鱼脑组织及脑电极三维重建. 生物医学工程学杂志, 2020, 37(5): 885–891. [doi: 10.7507/1001-5515.201911011]

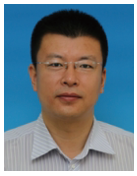
樊铭瑞(1995－), 女, 博士, 主要研究领域为小行星自主导航, 三维重建.

彭晓东(1981－), 男, 博士, 研究员, 博士生导师, 主要研究领域为航天任务演示与仿真, 卫星态势分析与演示, 海量数据管理与可视化, 场景感知与重建.

申冰可(2000－), 女, 硕士生, 主要研究领域为态势感知, 场景重构.

谢文明(1984－), 男, 研究员, 主要研究领域为复杂航天系统仿真.

牛文龙(1988－), 男, 博士, 副研究员, 主要研究领域为运动目标检测, 智能目标识别, 图像处理, 信号处理, 航天系统仿真与评估.



杨震(1972－), 男, 博士, 研究员, 博士生导师, 主要研究领域为复杂系统仿真, 空间任务协同与演示, 空间信息服务, 分布式空间系统.