文章编号　1004-924X(2023)16-2444-21

基于深度学习的多视图立体重建方法综述

鄢化彪[1]，徐方奇[1]，黄绿娥[2*]，刘词波[1]，林初欣[1]
(1. 江西理工大学 理学院，江西 赣州 341000；
2. 江西理工大学 电气工程与自动化学院，江西 赣州 341000)

摘要：多视图立体重建(Multi-view stereo Reconstruction，MVS Reconstruction)的目标是根据一组已知摄像机参数的多视角图像来重建场景的三维模型，是近年来三维重建的一类主流方法。本文针对最新的近百个基于深度学习的MVS方法做了较为系统的算法评估对比。首先，对现有的基于监督学习的MVS方法，按照特征提取、代价体构建、代价体正则化和深度回归的重建流程对各算法进行梳理，重点对代价体构建和正则化这两阶段的改进策略进行归纳总结，对于无监督的MVS方法，主要分析各算法损失项的设计，并按照其训练方式进行分类；其次，总结了MVS方法常用的实验数据集及其对应的性能评价指标，进一步研究特征金字塔结构、注意力机制、由粗到精等策略的引入对MVS网络性能的影响；此外，介绍了MVS方法的具体应用场景，包括数字孪生、自动驾驶、机器人技术、遗产保护、生物科学等领域；最后，提出关于MVS改进方向的建议，并对多视图三维重建未来的技术难点与研究方向进行探讨。

关　键　词：多视图立体；三维重建；深度学习；深度估计；单应性变换
中图分类号：TP394.1　　**文献标识码**：A　　**doi**：10.37188/OPE.20233116.2444

Review of multi-view stereo reconstruction methods based on deep learning

YAN Huabiao[1], XU Fangqi[1], HUANG Lü'er[2*], LIU Cibo[1], LIN Chuxin[1]

(1. *School of Science, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*;
2. *School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*)
* *Corresponding author, E-mail*: 9320080310@*jxust.edu.cn*

Abstract: The goal of Multi-view stereo (MVS) Reconstruction is to reconstruct a 3D model of a scene based on a set of multi-view images with known camera parameters, which is a mainstream method of 3D reconstruction in recent years. This paper provides a algorithm evaluation comparison for the latest hundreds of MVS methods based on deep learning. First, we sorted out the existing supervised learning-based MVS methods according to the reconstruction process of feature extraction, cost volume construction, cost volume regularization and depth regression, focusing on the summary of improvement strategies in the two stages of cost volume construction and cost volume regularization. For the unsupervised MVS

收稿日期：2022-11-14；**修订日期**：2022-12-26.
基金项目：国家自然科学基金资助项目(No. 11765008)；江西省自然科学基金资助项目(No. 20224BAB202036)；江西省教育厅科学技术重点研究项目资助(No. GJJ23005)；江西理工大学研究生创新计划资助项目(No. XY2021-S153)

methods, we mainly analyzed the design of the loss terms of each algorithm. It is classified according to its training mode. Secondly, we summarized the common datasets of MVS methods and their corresponding performance evaluation indexes, and further studied the introduction of strategies such as feature pyramid network, attention mechanism, coarse-to-fine strategy on the performance of MVS networks. In addition, it introduced the specific application scenarios of MVS methods, including digital twin, autonomous driving, robotics, heritage conservation, bioscience and other fields. Finally, we made some suggestions for the improvement direction of MVS methods, and also discussed the future technical difficulties and the research directions of MVS 3D reconstruction.

Key words: multi-view stereo; 3D reconstruction; deep learning; depth estimation; homography transformation

1　引　言

多视图立体重建(Multi-view stereo Reconstruc-tion,MVS Reconstruction)旨在根据从多视角拍摄的一系列图像中重建出场景三维模型,是三维重建的一类主流方法[1-2],被广泛应用于自动驾驶、增强现实、文物保护、智慧城市等领域。与使用激光雷达、深度相机[3]等设备的主动式三维重建方法相比,MVS这种基于图像的被动式三维重建方法具有重建精度高、视野大、成本低、易于推广应用等优点。

传统的MVS方法[2,4-6]通过使用多个相机视图之间的投影关系来优化深度值。例如,Schonberger等人提出了COLMAP[2,4],该方法在特征匹配阶段采用手工制作的特征,COLMAP会利用光度一致性同时估计视角的深度值和法向量值,并利用几何一致性进行深度图优化。Xu等人[5]提出具有多尺度几何一致性、自适应棋盘采样和多假设联合视图选择的ACMM。传统的MVS方法在理想的Lambertian场景下取得了一定成功,但在处理场景的弱纹理区域和反射表面的密集匹配时,重建完整度有待提升,且重建效果受光照强度和采样角度等外部因素影响严重。因此,基于深度学习的MVS算法应运而生。

基于深度学习的MVS方法可分为两种类型:基于体素的MVS和基于深度图的MVS。基于体素的方法[7-8]使用训练的网络回归每个体素的占用率,但体积表示方法存在巨大的内存消耗。另一种重建方法是首先估计每个视图的深度,然后回归并融合深度图以形成最终的3D点云模型。使用深度图作为中间层可以得到比基于体素的方法更精确的3D模型[9-12]。基于深度学习的方法对场景的全局和局部信息进行编码并提取特征,大大提高了对多视图立体特征匹配的鲁棒性,能够考虑镜面性、反射和环境光照变化等影响因素,有利于低纹理区域和非朗伯表面区域的重建,极大地提高了重建的完整度和整体质量。在之前的综述文章中,Zhu等人[13]介绍了MVS算法的代价体构建原理、深度图后处理方法和相关数据集等,重点梳理了MVS方法相关背景和原理,并按照特征提取、代价体构建和代价体正则化三个步骤进行方法的概述。Wang等人[14]对MVS方法的进展做了综述,根据3D表示形式将MVS方法分为基于深度图的方法和基于体素的方法。本文将重点比较分析基于深度图的MVS方法的最新进展,通过对近百篇基于深度学习的MVS算法文章的搜集整理,将这些方法做了更深入的归类分析,主要贡献可概括如下:

(1)对最新的基于深度学习的MVS方法进行了系统性归纳总结和比较分析;

(2)总结了MVS方法常用的公开数据集和性能评价指标;

(3)分析了基于学习的MVS方法中不同改进方式对模型性能的影响,按照重建流程对其进行归类,并对比分析典型MVS方法的综合重建性能;

(4)探讨了该方向当前研究所面临的挑战与核心技术难点,指出未来可考虑的研究方向。

2　基于深度学习的MVS方法

基于深度学习的MVSNet是一种端到端的根据多视角图像进行逐视图深度估计并融合生成点云模型的方法[11],该方法主要包括四个步骤:特征提取、代价体构建、代价体正则化和深度回归。图1为MVSNet的基本网络结构。其中,特征提取模块从一张参考图像和几张源图像中提取深层特征;代价体构建是将源图像的特征图单应性变换到参考图像的平行平面上,并采用基于方差的代价度量构建匹配代价体;代价体正则化将经3D CNN得到的代价体沿深度方向进行Softmax操作,得到像素级深度分布的概率体;深度回归则是根据不同平面的匹配结果判断参考视图像素所在深度,通过求深度的加权平均产生初始深度图。由于正则化时感受野较大,初始深度图的边界可能过度平滑,因此,通过2D CNN得到深度残差,加至初始深度图上得到细化深度图,分别对初始深度图和细化深度图进行L1损失计算并以权重系数λ相加,完成深度图的优化。最后,将不同视图的深度图进行过滤,再经深度融合产生最终三维点云模型。

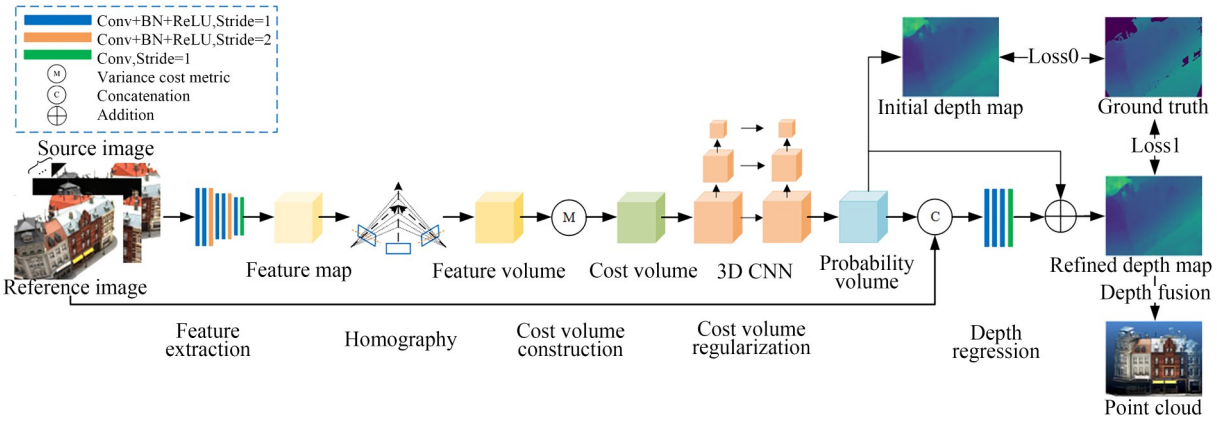


图1　MVSNet网络结构

Fig. 1　Overall structure of MVSNet

基于学习的MVS方法将基于单应性变换的平面扫描算法[15]引入代价体构建阶段。单应性变换可以隐式编码相机与物体间的几何关系,利用二维图像特征构建三维代价体。首先将源特征图经单应性变换映射到参考图像所在的相机坐标中,得到一个包含物体多角度信息的特征体,再输入到后续网络模块进行深度图的生成与细化。假设在世界坐标系下,参考图像的内参、旋转矩阵和位移矩阵分别为$\boldsymbol{K}_1$,$\boldsymbol{R}_1$和$\boldsymbol{t}_1$,源图像的内参、旋转矩阵和位移矩阵分别为$\boldsymbol{K}_i$,$\boldsymbol{R}_i$和$\boldsymbol{t}_i$,$\boldsymbol{n}^{\mathrm{T}}$为目标平面法向量且指向光源,则第$i$个源特征图$F_i$和深度为$d$的参考特征图$F_0$之间的单应性矩阵为:

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

基于深度学习的MVS方法有着重建精度高、方便高效、成本低廉和易于推广应用等优点[16-20]。许多学者以MVSNet为基准进行网络的改进,在重建效率、准确性和完整性等方面都获得了极大提升。本文主要将它们分为基于监督学习的方法和基于无监督学习的方法,并将监督学习方法按照其重建流程对改进方案做进一步细分,分析影响重建性能的主要因素,对于无监督学习方法,主要根据训练方式和损失函数进行了归纳总结,概括了提升算法性能的几种改进策略。

2.1　基于监督学习的MVS方法

针对基于监督学习的MVS方法,根据重建流程对最新改进方法进行分类,如图2所示。

2.1.1　特征提取模块的改进策略

前期大多数算法在特征提取模块中使用通用的CNN作为骨干网络,如U-Net[21]。一些方法[18,22-23]使用U-Net提取融合全局信息和局部信息的深度特征。通过多次下采样成倍增大感受野,使特征包含更多的全局信息。同时,从浅层到深层的跳跃连接有助于保留丰富的局部信息。以往基于深度学习的MVS方法[11,16,24]在特征提取模块通常利用下采样扩大感受野,同时降低分辨率以满足内存限制,并将经下采样次数最多的最后一层特征图输入到后续网络。这些方法可能会造成纹理信息丢失[25-26],影响重建结果的准确性。

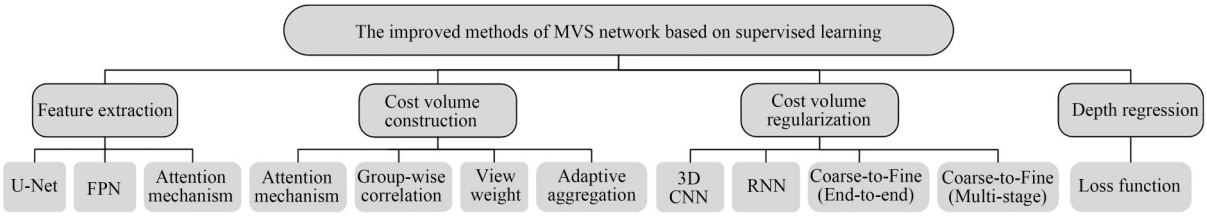


图2　基于监督学习的MVS网络改进方法分类

Fig. 2　Classification of MVS Network Improvement Methods Based on Supervised Learning

为了提取到更好、更丰富的特征,一些网络[19-20,27]采用特征金字塔网络(Feature Pyramid Net-work, FPN)[26]来进行特征提取。CVP-MVSNet[20],PVA-MVSNet[24]和DRI-MVSNet[28]采用图像金字塔,能够对多尺度的图像进行特征提取,并且所有尺度的特征图都具有较强的语义信息,但训练时内存消耗大、耗时长。而其他大部分方法采用FPN融合多个不同尺度的特征,只增加了较少的计算量,却能够融合低分辨率语义信息较丰富的特征图和高分辨率空间信息较丰富的特征图,并对多个尺度的特征图都进行后续的单应性变换,以促进下一步代价体的构建。

引入注意力机制可以使提取到的特征具有更强的表达能力。金字塔注意力网络(Pyramid Attention Network, PAN)[29]是注意力机制在FPN上的应用。PA-MVSNet[30]引入了PAN,利用多尺度特征金字塔注意力机制,引入尺度不可知注意力模块来捕获自上而下路径中的长距离特征对应,提取更丰富的特征信息。一些方法[31-32]在特征提取阶段引入独立自注意力机制[33]使网络更聚焦于重要信息,捕获像素之间的相互依赖关系。对于纹理信息丰富的区域,我们期望使用局部感受野,而弱纹理区域应该在更大的范围内匹配。不同于以往通过逐像素匹配的工作,LANet[34]引入一个远程注意网络,捕捉像素之间的远程相关性以增强图像特征,聚集更多的信息来度量图像之间的相似性。由于U-Net顶层特征图经多次下采样会存在更多细节的损失,HSF-MVSNet[35]引入一个特征增强的卷积块注意力模块(Convolutional Block Attention Module,CBAM)[36],其关注重要特征,同时抑制不重要特征,与在每一层中执行注意力机制过程相比,采用CBAM能够减少训练时间,更加灵活和高效。对于MVS中的反射和弱纹理区域的问题,MVSFormer[37]使用预训练的Vision Transformer(ViT)[38]来增强FPN,可以提供对MVS模型的全局理解。

ASPPMVSNet[39]将空洞空间卷积池化金字塔(Atrous Spatial Pyramid Pooling, ASPP)[40]引入MVS方法中,利用空洞卷积在不丢失信息的同时进行多尺度特征提取。AA-RMVSNet[41]、D-CasMVS Net[42]、ADIM-MVSNet[43]在特征提取阶段采用可变形卷积(Deformable Convolutional Networks, DCN)[44],能够根据局部上下文自适应地扩大感受野,使网络更好地学习边界和无纹理区域。TransMVSNet [45]在FPN后插入一个由DCN实现的自适应感受野模块,以自适应调整提取特征的范围。当一个物体的尺度在图像中变化很大时,传统方法提取的特征会导致

匹配代价的质量较低，CDSFNet[46]提出一种曲率引导的动态尺度特征提取网络，该网络可以适应各种对象尺度和图像分辨率。D2HC-RM-VSNet[47]提出一个密集接收扩展模块，融合不同扩展卷积层生成的多尺度特征信息，在不损失分辨率的同时增大感受野，实现稠密深度估计。

2.1.2　代价体构建阶段的改进策略

将源视图特征图单应性变换到参考视图的不同深度平面中，得到各视图的特征体。代价体构建是把特征体转化为代价体的过程，为适应任意数目的图像输入，许多方法[16,17,19-20,22,24,27,34-35,41,47,48-51]采用MVSNet[11]中基于方差的代价度量衡量视图间的相似性。所得代价体上一点是所有图像在该点深度值上特征的方差，方差越小，说明在该深度上置信度越高。然而，基于方差的代价体构建包含冗余信息，内存消耗大，许多学者针对这一问题提出了改进。

2.1.2.1　注意力机制的引入

Transformer[52]模型利用自注意力机制来捕获特征的内部相关性，将其引入MVS方法中有助于感知全局上下文信息[23,37,45,53-55]。Trans-MVSNet[45]引入了特征匹配Transformer（Feature Matching Transformer，FMT），利用内部（自身）和外部（交叉）注意力来加强图像内和图像间的远程全局上下文信息聚合。MVSTR[54]设计了全局上下文Transfor- mer和三维几何Transformer模块，以便提取具有全局上下文的密集特征，实现特征的三维一致性，促进视图间信息交互。考虑到效率问题，一些方法[23,53,55]采用极线Transformer（Epipolar Transformer，ET）[56]的交叉注意力，并利用几何知识沿极线建立多视图三维相关性，避免关注不必要的特征相关性。Liao等人[55]在特征匹配阶段引入了一种基于窗口的ET来减少匹配冗余信息。LANet[34]引入了一个远程注意力网络，选择性地聚合每个位置的参考特征，以捕捉整个空间的长期相互依赖。MVS-Net++[57]提出基于深度的注意力机制，并引入课程学习（Curriculum Learning，CL）[58]训练策略进行代价体构建，将深度掩膜作为先验知识并逐渐减少掩膜信息的提供，通过增强学习特征的强度，更好地关注前景对象的深度，得到准确的深度值。通过性能对比发现，引入注意力机制能够提升算法的准确性。

2.1.2.2　组相关相似性的引入

Guo等人[59]提出分组相关立体网络（Gwc-Net），在将源视图的深层特征变换到参考图像的坐标中后，引入分组相关相似性度量，减少代价体的通道数，从而减少内存使用量。首先将参考特征图F_0和转换特征图$F_i(d_j)$的特征通道均匀地划分为G组，然后如下计算第G组相似性$S_i^g(d_j)$：

$$S_i^g(d_j)=\frac{1}{Ch/G}\left\langle F_0^g, W_i^g(d_j)\right\rangle. \tag{2}$$

2.1.2.3　可见性的引入

尽管采用组相关可以在正则化中衰减不可见像素，但是其对场景内容的变化较为敏感，因此也限制了重建性能。许多方法通过学习视图权重[22,41,49,61-62,67]进行代价体的构建和聚合，进一步考虑了像素可见性。可见性是指一个三维点在给定的图像中是否可见。被遮挡的像素容易在代价聚合时出错，影响重建准确度。一些方法[22,24,49,61,68]采用自适应加权的代价聚合方法，利用可见性信息抑制不匹配代价的影响，能够提高重建点云的准确性和完整性。EPP-MVSNet[67]在代价体聚合上参考Vis-MVSNet[22]采用了加权聚合的方式，同时为了节省计算量，只在粗阶段生成权重可视图，后续阶段通过上采样的方式复用权重。ACINR-MVSNet[63]设计了一种用于自适应聚合的体素视图权重计算网络，利用平均组相关相似性度量来衡量匹配代价，以自适应的方式衡量总代价，有效地抑制了无效信息的负面影响。CDS-MVSNet[46]利用曲率信息估计像素可见性，法向曲率可以隐式地提供曲面的层次细节信息，通过去除错误匹配的像素，提高匹配代价的质量。这些迭代方法反映了引入可见性遮挡推理能够提高MVS算法的准确性，但由于代价体是一个4维张量，用传统方法进行代价体聚合时，会引入大量参数限制了效率。针对该问题，中国科学技术大学团队提出了一种高效立体匹配网络——自适应聚合网络（Adaptive Aggregation Network，AA-Net）[69]。AAModules包括同尺度聚合（Adaptive Intra-Scale Aggregation，ISA）模块和跨尺度聚合（Adaptive Cross-Scale Aggregation，CSA）模块。对于弱纹理甚至无纹

理区域,利用下采样更能提取高级语义信息,而对于纹理丰富的区域,又需要较高分辨率的深度估计来获得纹理信息。AA-RMVSNet[41]首先引入视图内聚合模块,利用上下文感知卷积和多尺度自适应聚合提取图像特征,并提出了一种像素级视图间ISA模块。这两种自适应聚合模块提高了低纹理区域重建性能,缓解了复杂场景中的遮挡问题。

2.1.3　代价体正则化网络结构

代价正则化是利用空间上下文信息将匹配代价体转化为深度假设的概率分布,输出为概率体,其每一点的取值为该像素点处在对应深度的概率,代价体正则化是实现精确深度预测的关键。如表1所示,本文根据代价体正则化策略以及网络结构的不同将所有MVS算法分为以下4类。

表1　采用不同代价体正则化策略的MVS方法的主要特点及存在问题

Tab. 1　Main characteristics and problems of MVS methods with different cost volume regularization strategies

Type		Characteristic	Problem	Method
End-to-end	CNN	Cost volume regularization using 3D CNN	Usually slow in training and reasoning, with large memory consumption	[18],[70],[57],[55],[71]
	RNN	Combining the accuracy of 3D CNN and the efficiency of RNN, greatly reducing memory consumption	Reduced memory consumption but increased runtime	[16],[47],[72],[73],[41],[74]
	Coarse-to-fine	Refine by other methods after obtaining the initial depth map	Need some prior knowledge	[11],[17],[66],[49],[22],[48],[34],[50],[30],[75],[45],[53],[64],[63],[43],[76]
Multi-stage	Coarse-to-fine	Build the cost volume over the entire depth range with coarse resolution, and calculate the reduced sample range based on the coarse depth map	Prediction accuracy is highly dependent on the initial depth map. The cost volume characteristics of different stages are not fully considered	[60],[61],[27],[20],[62],[19],[24],[54],[51],[31],[67],[42],[68],[77],[78],[23],[79],[39],[28],[32],[65]

2.1.3.1　基于3D CNN的代价体正则化

传统的MVS方法采用3D U-Net进行代价体正则化生成概率体。为了提高精度,MVSCRF[18]在深度图估计阶段第一次引入了条件随机场优化(Conditional Random Field, CRF)[80]正则化,它将深度估计看作一个多标签分类问题,每个深度假设对应一个标签,对象内部区域的邻近像素往往具有相似的标签,而边界附近的像素可能具有显著不同的标签。已有工作证明[81],RNN形式的CRF通过显式约束逐像素预测的输出,滤除潜在概率体中的噪声,可以极大地增强性能。并且CRF可以集成到模型中实现具有反向传播的端到端训练。在后续工作中,BP-MVSNet[70]对MVS中的CRF做了改进,采用了基于信念传播(Belief Propagation, BP)[82]的可微分CRF正则化层,也取得了不错的性能。3D DCN可以根据输入特征自适应改变感受野以适应局部几何形状,SPGNet[71],PatchMatchNet[62]通过利用DCN以自适应方式执行假设传播。LANet[34]使用3D ASPP代替3D U-Net,使网络有效扩大感受野,以纳入远程上下文,并缓解对象边界丢失问题。Att-MVSNet[83]引入了一个注意力引导的正则化模块,以自适应地聚合代价体。该模块由多层射线融合模块组成,可以分层聚合和正则化代价体。代价体本质上在深度和空间方向应该都是各向异性的,P-MVSNet[66]在提出的混合3D U-Net中利用了两种各向异性卷积在空间和深度方向上进行代价体聚合,充分

利用代价体的上下文信息，根据概率体推断深度概率分布。

2.1.3.2　基于RNN的代价体正则化

在代价体正则化阶段，利用3D CNN会消耗大量内存，特别是对于高分辨率图像。一些工作[16,41,47]用2D CNN和RNN代替3D CNN，以减少内存负担。Yao等人[16]提出用2D门控递归单元（GRU）递归网络沿深度方向顺序正则化代价体。空间上使用2D CNN，深度方向使用GRU聚合代价，效率显著提高，但缺少多尺度上下文信息的聚合。D2HC-RMVSNet[47]和AA-RMVSNet[41]结合3D CNN和RNN的优点，提出一种混合递归正则化网络U-LSTM，可以聚合多尺度上下文信息，同时能够高效处理原始大小的代价体。RED-Net[72]引入循环编码器-解码器（Recurrent Encoder-Decoder，RED）架构来顺序正则化代价体，实现了更高的效率和准确性，同时保持分辨率，有利于大规模的重建。现有的递归方法仅关注深度域中的局部依赖关系，大大限制了沿深度维度获取全局上下文的能力。为解决该问题，Xu等人[73]提出了一种非局部递归正则化网络NR2-Net，设计一个深度注意模块来捕捉非局部深度交互，以封闭的循环方式更新，对不同块之间的全局场景上下文进行建模，捕获沿深度维度的长期依赖关系以促进代价体正则化。BH-RMVSNet[84]采用基于双向混合长期记忆的结构来进行代价体正则化，在性能与3D CNN相当的同时节省运行内存。

2.1.3.3　采用多阶段由粗到精策略的MVS算法

利用RNN来调整代价体可在一定程度上减少内存消耗，但运行时间较长。为了使存储效率和重建精度之间达到良好的平衡，提出使用由粗到精的结构范式进行重建，在低分辨率特征上进行粗略全局深度范围下的预测，使用粗略深度图自适应地调整深度假设的采样范围，构建高分辨率代价体，逐步回归高质量的深度图。该类方法在内存和运行时间上都很高效[19-20,27]。Gu等人[19]提出在由粗到精的深度推断过程中构建金字塔结构并缩小深度搜索范围进行细粒度预测，可以估计高分辨率深度图，提高重建精度。类似的，Yang等人[20]提出用由粗到精的策略推断深度图，并提出一种自适应深度范围确定方法，在像素深度残差上迭代构建新的代价体来进行深度图细化。由粗到精的级联网络分散了网络的复杂性，能够在增加较少计算量的情况下提升深度预测精度，但其重建质量仍受到分辨率和深度假设范围的限制。在粗略深度预测较差的情况下，用事先确定的固定因子缩小深度假设范围可能导致错误预测或者引入冗余。Ma等人[67]提出一种合理设置深度假设的由粗到精算法，分别针对粗阶段和精阶段提出了极线聚集模块（Epipolar Assembling Module，EAM）和熵细化（Entropy Refining，ER）模块。EAM模块首先根据原始采样点的分布间隔自适应地插入新采样点，再采用卷积提取插值后代价体的信息，通过最大池化使代价体变回插值前的尺寸。ER模块通过计算深度图上每个点对应的熵来自适应地确定下一阶段合适的深度假设范围，进一步细化深度预测。

2.1.3.4　采用端到端由粗到精策略的MVS算法

一些端到端的MVS算法也采用了由粗到细的策略，通过在预测出初始深度图之后添加一个细化模块来得到精细深度图。Chen等人[17]提出的Point-MVSNet首先生成粗深度图，将其转换为点云，迭代预测深度残差，在预定义的局部空间范围内对粗略点云进行迭代细化。类似的，VA-Point MVSNet[49]根据从预测的点云推断出的3D几何先验信息和从多视图输入图像动态获取的2D图像信息来估计3D点云流。LA-Net[34]引入了一个新的损失来监督概率体，约束它的分布合理集中在真实深度处。Fast-MVSNet[48]对估计出的高分辨率稀疏深度图进行卷积，对局部区域内像素的深度依赖进行编码以加密该深度图，在得到深度图后添加一个高斯-牛顿层作为深度图细化模块，使重建效率大幅度提高。在此基础上，ACINR-MVSNet[63]对特征提取网络进行改进，设计了一个增强型高斯-牛顿层，明显提高了重建精度。为了降低内存消耗，GBi-MVS Net[75]将MVS定义为一个二值搜索问题，每一步通过执行分类来确定真实深度，大大降低深度假设数量，在加速模型训练的同时性能也得到提升。

2.1.4　深度回归和后处理策略

深度回归的目的是从概率体中获取深度图。选择合适的损失函数能够提高重建的准确度，DDR-Net[51]提出一种新的损失策略，利用学习到

的动态深度范围生成细化深度图,以保持下一阶段范围假设中覆盖的每个像素的真值。MVS-Net++[57]设计了三个损失函数,提出绝对相对损失以使模型专注于估计前景的深度,进一步设计了几何相似性损失和结构相似性损失来正则化图像和特征空间中多视图之间的相似性。DRI-MVSNet[28]提出多阶段深度残差预测模块,使用非均匀深度采样策略来构造假设的深度平面,生成高精度深度图。为了减少高分辨率场景重建的内存消耗并保持重建准确度,ADR-MVSNet[79]提出自适应深度减小模块,使用置信区间来逐渐减小最后两个阶段的深度范围。焦点损失[85]是目标检测领域中提出的常见解决方案,它是针对传统的离散标签定制的,Trans-MVSNet[45]采用焦点损失来加强监督,可以更好地处理模糊预测的问题。类似地,UniMVS-Net[68]采用统一焦点损失能够捕获更多细粒度指标,以重新平衡样本,并合理地处理连续标签。大多数由粗到精的方法通过计算当前预测深度与真实深度之间的残差,迭代细化点云。UCS-Net[27]在三种分辨率下应用了L1损失。

一些方法[16,34,41,47,72,86]将深度回归任务视为多分类任务,并在概率体和深度图中使用交叉熵损失函数。Ding等人[87]提出了基于特征相似性的对比度匹配损失和加权焦点损失,减小不重要区域中低置信度像素的权重。Point-MVSNet[17]和VA-PointMVS Net[49]提出PointFlow模块将输入深度图细化到更高的分辨率,并提高精度。对于每个点,PointFlow模块通过在所有视图中观察其相邻点来估计其沿参考相机方向到真实曲面的位移,推动这些点流向目标曲面,迭代细化预测深度图,从而提高时间和内存效率。DDL-MVS[88]联合估计深度图和边界图,提出边缘深度损失项来定义估计的边缘与真实深度变化之间的均方误差,利用边界图进一步细化深度图。

2.2　基于无监督学习的MVS方法

基于监督学习的MVS方法容易在进行合适的改进后达到较好的重建结果,预测结果可控,优化目标明确,损失函数设计较为简单。目前基于学习的MVS方法在一定程度上依赖于训练数据的丰富程度,大多数MVS算法依赖于用大规模真实三维数据作为监督以达到更好的重建效果,但是用于训练的真实数据标签(如点云、深度图等)的获取成本较高,并且基于监督学习的模型泛化能力较弱,在其他场景数据集上难以取得较好的重建效果。因此,对基于无监督学习的MVS方法的研究具有重要价值,在没有真实标签的情况下,研究如何利用数据本身先验信息自监督是方法改进的关键。Knot等人[89]提出第一个基于无监督学习的MVS框架,利用参考图像与单应性变换后的源图像之间的光度一致性进行监督。光度一致性损失如式(3)所示:

$$L_{PC}=\sum_{i=2}^{N}\frac{\left\|(I'_i-I_0)\odot M_i\right\|_2+\left\|(\nabla I'_i-\nabla I_0)\odot M_i\right\|_2}{\left\|M_i\right\|_1},\tag{3}$$

其中:∇表示梯度算子,$\odot$为点积。考虑到光度损失对照明条件和拍摄角度敏感以及多视图之间存在遮挡和光照信息不同的问题,在计算光度损失时融合多个图像对之间的匹配误差图,再将光度一致性损失结合深度平滑损失和结构相似性损失一起作为网络训练的监督信号。光滑度损失表示为:

$$L_{\text{smooth}}=\frac{1}{N}\sum_{i=1}^{N}\left(e^{-\alpha_1\left|\nabla I^i_{\text{ref}}\right|}\left|\nabla D_i\right|+e^{-\alpha_2\left|\nabla^2 I^i_{\text{ref}}\right|}\left|\nabla^2 D_i\right|\right),\tag{4}$$

其中:N为像素数量,∇^2表示二阶导数,D为深度。深度平滑损失可促进预测深度图中的平滑度。结构相似性损失表示为:

$$L_{SSIM}=\frac{1}{N}\sum_{i=1}^{N}\frac{1-SSIM(I^i_{\text{ref}},I'^i_{\text{src}})}{2}M_{\text{ref}}.\tag{5}$$

结构相似性损失通过亮度、对比度、结构来测量两个图像之间的相似性。当处理亮度剧烈变化的区域时,结构损失约束可以提高鲁棒性。Dai等人[90]提出了一种同时预测所有视图深度的对称网络,进一步丰富了损失函数,在实现自监督的同时通过学习视图遮挡掩膜避免遮挡区域的点参加损失计算,提高算法性能。Mallick等人[91]利用模型不可知元学习[92]框架来学习自适应特征表示,用于基于视图合成的自监督MVS重建。

采用视图合成损失进行自监督学习的前提是一个点在不同视图中具有相同颜色,但这在环境光照条件时刻变化的真实世界中无法实现,因此仅使用光度一致性约束不够准确,需要引入更多的约束来解决纹理模糊问题。Huang等人[93]

在特征提取阶段采用FPN结构，并结合基于像素和基于特征的损失，像素级考虑光度一致性、结构一致性和深度平滑约束，特征级采用预训练的VGG16网络对中间层提取的特征进行一致性约束。此外，在三维点云中引入新的法向深度一致性来细化初始深度图，以提高深度图的准确性和连续性。Xu等人[94]在损失函数中引入语义一致性和数据增强一致性，将预训练的VGG网络提取的特征经非负矩阵分解进行多视图间的无监督协同分割。语义一致性损失表示为：

$$L_{SC}=-\sum_{i=2}^{N}\left[\frac{1}{\left\|M_i\right\|_1}\sum_{j=1}^{HW}f(S_{1,j})\log(S'_{i,j})M_{i,j}\right],\quad(6)$$

其中：S'_i为变换的分割图，S_1为参考分割图转换的真实标签，$f(S_{1,j})=\mathrm{onehot}(\mathrm{argmax}(S_{1,j}))$，计算它们之间每像素交叉熵损失作为语义一致性损失$L_{SC}$。$M_i$是从第i个视图到参考视图有效像素的二进制掩码。数据增强一致性损失表示为：

$$L_{DA}=\frac{1}{\left\|M'\right\|_1}\sum\left\|(D-D')\odot M'\right\|_2,\quad(7)$$

其中：M'表示变换下的掩膜，原始图像I的预测表示为D，增强图像I'的预测表示为D'。通过最小化D和D'之间的差异来确保数据增强一致性。由此，从多视图中挖掘相互语义以指导语义一致性。对原始多视图进行随机数据增强并输入到网络中，以深度估计分支预测的深度图作为伪标签来监督数据增强分支的预测结果，构建数据增强一致性损失，同时加入结构相似性和深度平滑度约束，提高自监督信号对图像中光线颜色干扰的鲁棒性。

由于基于无监督的MVS方法在复杂区域的监督信号较弱，因此提出了基于伪标签的方法来增强约束。以CVP-MVSNet[20]为骨干网络，Yang等人[95]提出了一个用于自监督网络的伪标签合成和优化策略，首先用视图合成损失初始化网络，将得到的深度图作为初始伪标签监督网络的训练，利用多视图几何一致性对初始伪标签进行过滤以确保伪标签的可靠性，将过滤后的深度图进行点云融合，再通过泊松表面重建得到重建点云的表面网格模型，渲染出每个视角下的深度图作为网络模型自监督训练的伪标签。现有方法缺乏对自监督MVS任务有效性的全面解释。为此，Xu等人[96]提出自监督MVS方法中的认知不确定性。通过计算光流深度一致性损失减轻前景中的模糊监督，采用蒙特卡洛方法获取不确定性图，并进一步过滤背景的无效监督。通过对多个采样深度图求平均值来计算每个视图的伪标签，以再次训练网络。

Qi等人[97]采用渐进式多级架构，并引入多视图相关先验，不仅提高了深度图的分辨率，还减少了内存消耗。Dong等人[98]用新的损失函数对弱纹理表面进行重建，其中，逐片光度一致性扩展了特征的感受域，视图几何一致性有效提高弱纹理重建的完整性。几何一致性损失表示为：

$$L_{\mathrm{geometric}}=\frac{1}{N}\sum_{i=1}^{N}\left|I_{\mathrm{ref}}-I^{i}_{\mathrm{ref}\rightleftarrows\mathrm{src}}\right|\odot M_{\mathrm{ref}},\quad(8)$$

其中，$I^{i}_{\mathrm{ref}\rightleftarrows\mathrm{src}}$为交叉渲染的参考视图。因此，几何一致性可以利用冗余的交叉视图信息来评估深度图的质量。

真实场景中的多视图非朗伯表面并受遮挡。Chang等人[99]结合视图渲染和结构化深度表示的优点，提出了一种新的神经呈现方法(RC-MVS-Net)来解决视图间对应关系模糊的问题。提出一种基于神经体渲染的参考视图合成损失以监督视图的光度效果，并引入高斯均匀混合采样来学习相似物体表面的几何特征，以克服遮挡问题。引入深度渲染一致性损失优化初始深度图，确保预测的稳健性。NeRF[100]将神经场与体渲染有效结合，首次利用隐式表示实现了图像级的视图合成。不同于以往MVS方法，MVS-NeRF[101]不用代价体进行深度推断，而是利用代价体进行几何感知场景推理，并将其与基于物理的体渲染相结合进行神经辐射场重建，实现视图合成。用真实像素颜色来作为唯一的监督信息，既解决了NeRF推理时间慢的问题，又充分利用了NeRF在无监督情况下的高质量视图合成能力以及MVS可进行跨视图相关性推断的优点，具有很强的泛化能力。然而，该网络需要输入固定数量的图像。Point-NeRF[102]可以融合任意数量视图的神经点，并实现360°完整辐射场的快速重建。该方法利用点云信息避免在空域上做无效操作，提高了效率，但其在训练时需要真实深度图作为监督。为了实现无监督，Zhang等人[103]在连续二维可变形表面上定义辐射亮度场，局部区域中的显式表面可以通过可微渲染沿视相关的相机射线逐渐变形，不需要像NeRF那样拟合密集空间中的几何形状和纹理信息，对于大规模

场景可以保持形状完整和纹理逼真。知识蒸馏(Knowledge Distillation,KD)[104]是一种将知识从预训练的教师模型转移到学生模型的方法。最近,Ding等人[105]将KD引入MVS方法中,首先以自监督的方式用光度一致性和特征一致性训练教师模型,通过交叉视图一致性和概率编码生成伪概率分布,再将教师模型的知识提取到学生模型中,获得与监督训练方法相当的重建质量。对基于无监督学习的MVS方法损失函数的总结如表2所示。

表2　基于无监督学习的MVS方法的损失函数汇总

Tab. 2　Summary of loss functions of MVS methods based on unsupervised learning

Method			Loss function								
Type	Network	Training methods	L_{SSIM}	L_{smooth}	L_{PC}	$L1$	$L_{feature}$	View	Image gradient	L_{DA}	Other
Unsupervised	UnsupMVS[89]	End-to-End	√	√	√						
	MVS2[90]	End-to-End	√	√	√	√		√	√		
	M3VSNet[93]	End-to-End	√	√	√		√				Normal depth
	PatchMVS[98]	End-to-End	√	√	√		√				$L_{geometric}$
	RC-MVS[99]	End-to-End	√	√	√					√	L_{render}
	MS-MVS[97]	End-to-End	√	√	√				√		
Self-supervised	JDACS[94]	End-to-End	√	√	√					√	L_{SC}
	Meta MVS[91]	Multi-stage	√	√		√		√			
	SelfsupCVP[95]	Multi-stage	√	√					√		$L_{perception}$
	U-MVSNet[96]	Multi-stage			√						Optical flow
	KD-MVS[105]	Multi-stage			√		√	√			

与端到端训练的方法不同,多阶段方法包括多个训练过程,并引入了许多附加的监督,例如重建网格模型的光流或渲染深度图。通过引入额外的监督可以提高生成的深度图质量,但会增加模型复杂性,训练过程耗时长。因此,如何利用先验信息得到合适的监督信号是无监督学习方法改进的关键。

3　MVS常用数据集和评价指标

3.1　数据集

对MVS方法常用数据集的总结如表3所示。

DTU数据集[9]是一个涵盖多种对象的室内数据集,利用一个安装有结构光扫描仪的工业机器臂对物体进行多视角的拍摄,可获取每个视角的相机内、外参数。数据集中每个场景都取49或64个相机位置,对应于每个场景RGB图像和结构光标签的数量,每个视角有7种不同亮度的图像。数据集中包括摄像头参数、拍摄图像、真实深度图和掩膜。

由于DTU数据集的场景相对简单,为了提升模型的泛化能力,Tanks and Temples数据集[106]采集了较为复杂的室外场景,包含雕塑、车辆、建筑物以及具有复杂几何布局大型室内室外场景。该数据集由美国英特尔公司提出,场景的真实点云通过工业激光扫描仪获取,使用高清数码单反相机(Digital Single Lens Reflex,DSLR)对真实场景进行视频拍摄,提供视频序列作为输入。

ETH3D数据集[107]由瑞士苏黎世理工大学提出,包括使用DSLR和多相机装置拍摄的图像和使用高精度激光扫描仪获取的真实点云,包括复杂自然场景和具有较大视点变化的人造环境,两种模态的数据通过一种基于光度一致性的优化算法对齐。其图像具有强烈视点变化,用于高分辨率MVS重建问题。训练数据集包括庭院、办公室、操场等中型建筑,测试数据集包括桥、门、客厅等场景。

由于大规模真实模型的扫描成本很高,BlendedMVS[108]通过将纹理化的三维网格模型渲染到不同视点得到训练图像和深度图,是一个

表3 MVS方法常用数据集

Tab. 3 Dataset of MVS methods

Dataset		Scene	Resolution	Scale	Number of scenes	Online benchmark	Website
DTU[9]		indoor	1 600×1 200 pixels	27 097	124		http://roboimagedata.compute.dtu.dk/
Tanks and Temples[106]	Intermediate	outdoor	8-megapixel	≈5600	8	√	https://www.tanksandtemples.org/
	Advanced				6		
ETH3D[107]	high resolution	indoor and outdoor	24-megapixel	454	13(train)	√	https://www.eth3d.net/
				443	12(test)		
	low resolution		0.4-megapixel	4 796	5(train)		
				5 212	5(test)		
BlendedMVS[108]		outdoor	1 536×2 048 pixels	17 818	106(train) 7(test)		https://github.com/YoYo000/BlendedMVS
GigaMVS[109]		outdoor	gigapixel	3 599	13		http://www.gigamvs.com/

大规模MVS合成数据集。为了在模型训练中引入环境的光照信息,将渲染得到的彩色图像和输入图像进行混合,生成训练真实值,包括城市、建筑、雕塑和小物体等各种场景。

清华大学成像与智能技术实验室构建了国际首个十亿像素级室外超大规模场景三维重建数据集GigaMVS[109],既包含大规模场景的3D几何又包含局部高分辨率纹理细节,其图像分辨率比ETH3D数据集[107]大10倍左右,可以清楚地观察到场景结构和局部细节。该数据集的真实点云由多达15亿个融合的激光点组成。所使用的扫描仪可覆盖最大面积为32007 m^2的场景,平均比Tanks and Temples数据集[106]大20倍。如今,MVS算法相关数据集规模越来越大,在新的数据集上进行评估可以验证网络模型的泛化能力。

3.2 评价指标

MVS算法的评价指标可分为距离度量和百分比度量两种形式。距离度量包括准确性(Accuracy)和完整性(Completeness)。其中,重建结果的精确度是指重建结果与真实点云的接近程度,即每个重建点到真实点云中最近点绝对距离的均值或中值。重建结果的完整性是看真实模型中有多少点被重建结果覆盖,即从真实点云到重建点云的绝对距离的均值或中值。通常将准确性和完整性的平均值作为总分(Overall Score)来表示重建结果的总体质量。距离度量的总分值越低表示重建效果越好。评价指标的另一种形式是百分比度量。利用精确度(Precision)和召回率(Recall)来衡量重建点云与真实点云的一致性,与距离度量不同的是,百分比度量需要给定一个距离阈值d,重建点云的精确度为重建点云中所有点到真实点云的绝对距离小于d的点的个数除以重建点云中点的总个数,因此,其结果为百分比形式,召回率则是真实点云到重建点云的绝对距离小于d的点的个数除以真实点云中点的总个数。为了采用唯一衡量标准对整体重建性能进行评价,取Precision和Recall的调和平均值作为综合指标,用F-score表示。因此,做好准确性和完整性的权衡可以提高F-score,百分比度量的值越高则表示重建效果越好。

其中,DTU数据集[9]与Blended MVS数据集[108]采用距离度量,Tanks and Temples数据集[106]、ETH3D数据集[107]与GigaMVS[109]数据集采用百分比度量。特别地,Blended MVS数据集[108]在定量评估中还使用了2D深度图评估指标,包括预测深度和真实深度之间的L1距离、深度误差大于给定1像素阈值和3像素阈值的像素比值。由于单一的几何量化指标不能很好地反映3D模型的视觉质量,GigaMVS数据集增加了峰值信噪比、归一化互相关和Completeness作为定量纹理评估。准确性评估使用PSNR和NCC来衡量图像之间的相似性。联合考虑PSNR,NCC和Completeness是一种鲁棒的纹理评价度量。

4　基于深度学习的MVS算法性能分析

本节将展示大部分MVS算法在两个主流数据集DTU[9]和Tanks and Temples[106]的性能。每种方法在DTU上的性能总结如表4所示,在Tanks and Temples的Intermediate数据集上的性能总结如图3所示。经过对性能的对比分析,总结出MVS算法性能的主要影响因素,有利于寻找更优的算法改进策略。

表4　不同MVS算法在DTU数据集[9]上的定量结果(Overall越低越好)

Tab. 4　Quantitative results of different MVS methods on DTU datasets[9] (lower is better)

Type	Method	Acc. /mm	Comp. /mm	Overall /mm	Method	Acc. /mm	Comp. /mm	Overall /mm
Supervised	MVSNet[11]	0.396	0.527	0.462	AACVP-MVSNet[31]	0.357	0.326	0.341
	R-MVSNet[16]	0.383	0.452	0.417	NR2-Net[73]	0.370	0.332	0.351
	P-MVSNet[66]	0.406	0.434	0.420	AA-RMVSNet[41]	0.376	0.339	0.357
	Point-MVSNet[17]	0.342	0.411	0.376	EPP-MVSNet[67]	0.413	0.296	0.355
	MVSCRF[18]	0.371	0.426	0.398	PatchmatchNet[62]	0.427	0.277	0.352
	CVP-MVSNet[20]	0.296	0.406	0.351	DRI-MVSNet[28]	0.432	0.327	0.379
	CasMVSNet[19]	0.325	0.385	0.355	IterMVS[74]	0.373	0.354	0.363
	UCSNet[27]	0.338	0.349	0.344	SPGNet[71]	0.320	0.382	0.351
	BP-MVSNet[70]	0.333	0.320	0.327	D-CasMVSNet[42]	0.348	0.350	0.349
	Fast-MVSNet[48]	0.336	0.403	0.370	ASPPMVSNet[39]	0.334	0.360	0.347
	PVSNet[61]	0.337	0.315	0.326	Effi-MVS[78]	0.314	0.334	0.324
	Vis-MVSNet[22]	0.369	0.361	0.365	GBi-Net[75]	0.315	0.262	0.289
	VA-Point-MVSNet[49]	0.359	0.358	0.359	TransMVSNet[45]	0.321	0.289	0.305
	PVA-MVSNet[24]	0.379	0.336	0.357	RayMVSNet[23]	0.341	0.319	0.330
	Att-MVS[83]	0.383	0.329	0.356	MVSTER[53]	0.350	0.276	0.313
	D2HC-RMVSNet[47]	0.395	0.378	0.386	UniMVSNet[68]	0.352	0.278	0.315
	MVSNet++[57]	0.407	0.345	0.376	NP-CVP-MVSNet[77]	0.356	0.275	0.315
	REDNet[72]	0.456	0.326	0.391	ACINR-MVSNet[63]	0.306	0.364	0.335
	CIDER[60]	0.417	0.437	0.427	ADR-MVSNet[79]	0.354	0.317	0.335
	LANet[34]	0.32	0.349	0.335	MSCVP-MVSNet[65]	0.379	0.278	0.328
	DDR-Net[51]	0.339	0.320	0.329	FPSA-MVSNet[32]	0.363	0.283	0.323
	PA-MVSNet[30]	0.313	0.437	0.375	ADIM-MVSNet[43]	0.344	0.298	0.321
	DDL-MVS[88]	0.405	0.267	0.336	HSF-MVSNet[35]	0.378	0.353	0.365
	BH-RMVSNet[84]	0.368	0.303	0.335	MVSFormer[37]	0.327	0.251	0.289
	CER-MVS[110]	0.359	0.305	0.332	CDSFNet[46]	0.352	0.280	0.316
	HighRes-MVSNet[50]	0.354	0.393	0.373	MFNet[64]	0.339	0.304	0.321
	MVSTR[54]	0.356	0.295	0.326	WT-MVSNet[55]	0.309	0.281	0.295
Unsupervised	UNSUP-MVS[89]	0.881	1.073	0.977	SurRF[103]	0.388	0.390	0.389
	Meta MVS[91]	0.594	0.779	0.687	CLD-MVS[111]	0.335	0.430	0.383
	MVS2[90]	0.760	0.515	0.637	Self-sup-CVP-MVSNet[95]	0.308	0.418	0.363
	M3VSNet[93]	0.636	0.531	0.583	JDACS-MS[94]	0.398	0.318	0.358
	JDACS[94]	0.571	0.515	0.543	U-MVSNet[96]	0.354	0.354	0.354
	PatchMVSNet[98]	0.538	0.365	0.451	RC-MVSNet[99]	0.396	0.295	0.345
	MS-MVS[97]	0.383	0.415	0.399	KD-MVS[105]	0.359	0.295	0.327

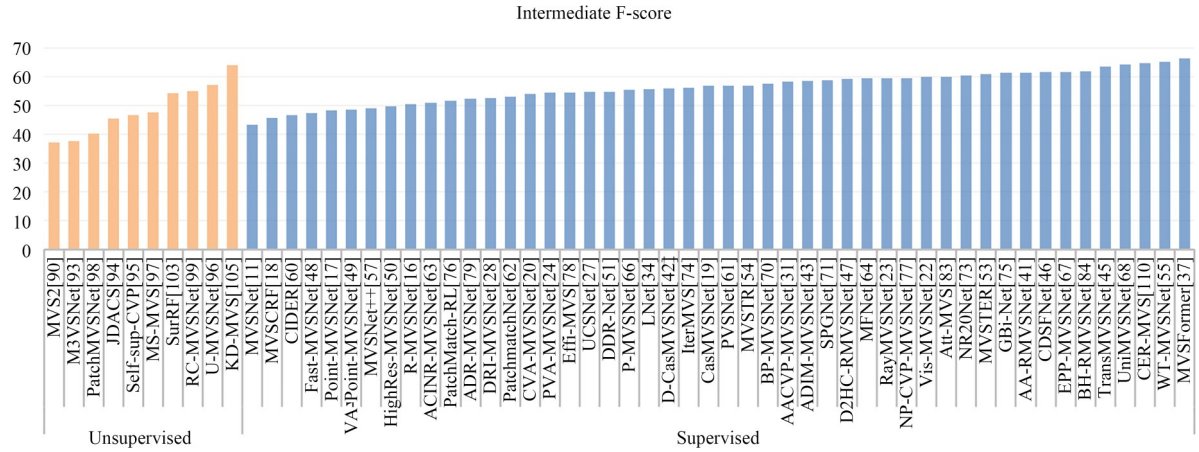


图 3　Tanks and Temples[106]数据集上 MVS 方法的可视化性能比较(F-score 越高越好)

Fig. 3　Visualized performance comparisons of MVS methods on Tanks and Temples benchmark[106] (higher is better)

在 DTU 数据集上,基于监督学习的总分最高的几个方法 MFNet[64],NP-CVP-MVSNet[77],Uni-MVSNet[68],MVSTER[53],TransMVS-Net[45],WT-MVSNet[55],GBi-Net[75]中,全部都在特征提取模块采用了 FPN 结构。其中,Trans-MVSNet[45],MVSTER[53]和 WT-MVSNet[55]都引入 Transformer 进行代价体的构建,充分利用了 Transformer 的优点,能够在多个视图中捕获和聚合全局上下文,极大地改善了重建效果。组相关相似性度量[32,43,64]也能明显提升模型性能。准确率最高的几个方法 CVP-MVSNet[20],Effi-MVS[78],PA-MVSNet[30],WT-MVSNet[55],ACINR-MVSNet[63]和 GBi-Net[75]都采用了由粗到精的结构,可以看出,采用由粗到精的结构可以显著提升方法的准确性和整体性能,但无法保证重建结果的完整性。而 GBi-Net[75]和 WT-MVSNet[55]则既保证了重建的准确性又保证了完整性。IterMVSNet[74]和 Fast-MVSNet[48]等方法为了节省内存消耗和运行时间,牺牲了准确度。对于无监督学习方法,多阶段方法都取得了不错的结果,该策略利用额外的后处理阶段来训练模型,利用来自第三方输入的辅助,例如真实数据标签、渲染网格或光流,能够有效提高重建质量,但会增加模型训练成本。在 Tanks and Temples 数据集上,良好的重建性能依然离不开由粗到精架构和特征金字塔的作用,可见性推理[41,67,22]也有效提升了重建精度,适合的损失函数[45,68,55]也是提升重建质量的重要因素之一。最近,一些新方法的引入大大提升了基于自监督学习 MVS 方法的重建性能,其中,多阶段的 KD-MVSNet[105]性能最好,端到端的 RC-MVSNet[99]性能其次。在监督学习方法中,GBi-Net[75],TransMVSNet[45],MVSTER[53],UniMVSNet[68]和 WT-MVSNet[55]在 DTU 数据集上表现较好的同时,在 Tanks and Tem- ples 数据集上也取得了不错的 F1-score,而[43,78,64]仅在 DTU 数据集上具有良好的性能,可能只适用于小范围的室内对象重建,模型泛化能力较差。无监督学习方法中,KD-MVS[105],RC-MVSNet[99],U-MVSNet[96]在两个数据集上的表现都处在前三。但无论是在 DTU 数据集上还是 Tanks and Temples 数据集上,基于无监督学习的重建方法的性能与基于监督学习的重建性能之间都有着较大的差距,该类方法仍需要进一步的深入研究。

5　多视图三维重建的应用

近年来,各领域的三维模型需求越来越大,多视图三维重建被广泛应用于不同领域:(1)数字孪生:为实现数字孪生建立物理实体的数字化三维模型,创建面向数字孪生的场景智能理解、动态数据感知及可视化等[112]。通过数字孪生技术可以把物理世界映射到三维虚拟空间来进行交互和感知。(2)自动驾驶:多视图三维重建可利

用视图间的几何约束准确预测深度值。车辆通过摄像头捕获道路场景图像，可提取出实时交通相关特征信息，并在车端完成导航定位、设备控制等操作。(3)机器人技术：基于摄像头重建出所处场景的三维结构信息，实现机器人准确定位、抓取目标物体或规划自身运动轨迹[113]。在自动驾驶与机器人技术领域中，重建出目标的三维模型后还需对目标物体进行检测和识别[114]，例如自动驾驶中对道路环境中的行人进行三维重建和目标检测。因此，重建出精确的三维模型是促进目标识别的关键。(4)虚拟现实与增强现实：目的是提供逼真的数字化三维模型，导入虚拟现实场景实现交互沉浸式体验，其三维模型的精确构建和几何交互是提供真实沉浸感的重要基础[115-116]。(5)数字城市与数字地图：通过无人机航拍地面数据，重建出地表建筑、山川等几何模型，可大量减少三维地图构建工作量，以极少的人力成本快速、准确地实现大场景地图数字化[72,117-119]。(6)遗产保护：大量文物、遗迹可用照片拍摄的方式快速存储并数字化为三维模型数据，即便日后文物因为老化等原因损坏，也可以从数据库中快速重构出文物的原貌。(7)生物科学：可通过多视角扫描内脏器官重建出三维模型，为医疗人员的分析诊断提供帮助，甚至在未来5G技术更成熟之后，给远程手术提供数据基础[120]，对植物进行三维重建也能通过提取叶片长度、株高等参数建立相关植物生长预测模型[121]。(8)航空航天：对空间目标进行三维重建能为卫星在轨服务提供服务对象的结构信息，是提高系统自主性的关键技术[122-124]。因此，来自众多领域的需求推动着三维重建技术的高质量发展。

6　总结与展望

本文对最新的基于深度学习的MVS方法做了综述，根据它们的网络结构和改进方案，按照MVS重建流程进行了分析和归类，根据实验结果比较了大多数方法的重建性能，总结出能够提升网络性能的一些改进策略，例如可见性分析、FPN、注意力机制和粗到精策略等。目前，最新的MVS算法能基本解决MVS综述文献[13-14]中提到的特征提取器的设计、数据集的多样性和模型复杂度等问题，无监督的MVS方法性能也在逐步接近监督学习方法。尽管MVS方法已经取得了巨大进步，但仍存在一些问题限制其在各领域中的应用。

首先，如何在控制内存成本和时间成本的同时提高端到端三维重建方法在大规模场景中的高分辨率重建精度是一个待解决的问题。其次，弱纹理会导致匹配模糊，感受野较小，难以确保像素匹配。将基于几何信息约束的三维重建和场景语义信息识别相结合，融合高级语义信息和低级语义信息，实现大规模场景的重建和感知是未来研究方向之一。另外，现有的MVS方法主要针对静态物体进行重建，对动态物体或视频对象的重建关注较少，对于非刚性、透明、镜面等物体的重建问题也有待进一步研究。最后，基于学习的MVS方法在一定程度上依赖于训练数据的丰富程度，模型的泛化能力仍存在较大提升空间，在没有真实标签的情况下，研究如何利用数据本身先验信息自监督是方法改进的关键。同时，深度学习网络的一些最新进展也有可能提高MVS算法性能，如Transformer[52]、NeRF[100]、KD[104]等。因此，如何合理地引入一些新思想实现无监督的MVS重建任务以提升重建效果也是一个值得研究的问题。

参考文献：

[1] FURUKAWA Y, HERNÁNDEZ C. Multi-view stereo: a tutorial[J]. *Foundations and Trends® in Computer Graphics and Vision*, 2015, 9(1/2): 1-148.

[2] SMITH M W, CARRIVICK J L, QUINCEY D J. Structure from motion photogrammetry in physical geography[J]. *Progress in Physical Geography: Earth and Environment*, 2016, 40(2): 247-275.

[3] 刘东生，陈建林，费点，等. 基于深度相机的大场景三维重建[J]. 光学 精密工程，2020，28(1)：234-243.
LIU D S, CHEN J L, FEI D, *et al*. Three-dimensional reconstruction of large-scale scene based on depth camera[J]. *Opt. Precision Eng.*, 2020, 28

(1): 234-243. (in Chinese)

[4] SCHÖNBERGER J L, ZHENG E L, FRAHM J M, *et al*. *Pixelwise View Selection for Unstructured Multi-View Stereo*[M]. Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 501-518.

[5] XU Q S, TAO W B. Multi-scale geometric consistency guided multi-view stereo[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 5478-5487.

[6] 张宝祥, 王振明, 杨秋慧. 基于Harris-SIFT算法和全卷积深度预测的显微镜成像的三维重建研究[J]. 光学 精密工程, 2022, 30(14): 1669-1681.
ZHANG B X, YU Z M, YANG Q H. Research on 3D reconstruction of microscope imaging based on Harris-SIFT algorithm and full convolution depth prediction[J]. *Opt. Precision Eng.*, 2022, 30(14): 1669-1681. (in Chinese)

[7] JI M Q, GALL J, ZHENG H T, *et al*. SurfaceNet: an End-to-End 3D neural network for multiview stereopsis[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 2326-2334.

[8] KAR A, HÄNE C, MALIK J. Learning a Multi-View Stereo Machine [EB/OL]. 2017: *arXiv*: 1708.05375. https://arxiv.org/abs/1708.05375

[9] AANÆS H, JENSEN R R, VOGIATZIS G, *et al*. Large-scale data for multiple-view stereopsis [J]. *International Journal of Computer Vision*, 2016, 120(2): 153-168.

[10] KNAPITSCH A, PARK J, ZHOU Q Y, *et al*. Tanks and temples: benchmarking large-scale scene reconstruction[J]. *ACM Transactions on Graphics*, 36(4): 1-13.

[11] YAO Y, LUO Z X, LI S W, *et al*. *MVSNet: Depth Inference for Unstructured Multi-View Stereo*[M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 785-801.

[12] LI L Y, LI X Y, JIANG L Y, *et al*. A review on deep learning techniques for cloud detection methodologies and challenges[J]. *Signal, Image and Video Processing*, 2021, 15(7): 1527-1535.

[13] ZHU Q, MIN C, WEI Z, *et al*. Deep Learning for Multi-View Stereo via Plane Sweep: a Survey [EB/OL]. 2021: *arXiv*: 2106.15328. https://arxiv.org/abs/2106.15328

[14] WANG X, WANG C, LIU B, *et al*. Multi-view stereo in the deep learning era: a comprehensive review[J]. *Displays*, 2021, 70: 102102.

[15] GALLUP D, FRAHM J M, MORDOHAI P, *et al*. Real-time plane-sweeping stereo with multiple sweeping directions[C]. 2007 *IEEE Conference on Computer Vision and Pattern Recognition*. 17-22, 2007, *Minneapolis*, *MN*, *USA*. IEEE, 2007: 1-8.

[16] YAO Y, LUO Z X, LI S W, *et al*. Recurrent MVSNet for high-resolution multi-view stereo depth inference[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 5520-5529.

[17] CHEN R, HAN S F, XU J, *et al*. Point-based multi-view stereo network[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 1538-1547.

[18] XUE Y Z, CHEN J S, WAN W T, *et al*. MVSCRF: learning multi-view stereo with conditional random fields[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 4311-4320.

[19] GU X D, FAN Z W, ZHU S Y, *et al*. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 2492-2501.

[20] YANG J Y, MAO W, ALVAREZ J M, *et al*. Cost volume pyramid based depth inference for multi-view stereo[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 4876-4885.

[21] RONNEBERGER O, FISCHER P, BROX T. *U-net: Convolutional Networks for Biomedical Image Segmentation*[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[22] ZHANG J, YAO Y, LI S, *et al*. Visibility-Aware Multi-View Stereo Network[EB/OL]. 2020: *arXiv*: 2008.07928. https://arxiv.org/abs/2008.07928

[23] SHI Y F, XI J H, HU D W, *et al*. RayMVSNet: learning ray-based 1D implicit fields for accurate multi-view stereo[C]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 17, 2023, IEEE, 2023: 1-17.

[24] YI H W, WEI Z Z, DING M Y, *et al*. *Pyramid Multi-View Stereo Net with Self-Adaptive View Aggregation* [M]. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 766-782.

[25] YU F, KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions[EB/OL]. 2015: *arXiv*: 1511.07122. https://arxiv.org/abs/1511.07122

[26] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 936-944.

[27] CHENG S, XU Z X, ZHU S L, *et al*. Deep stereo using adaptive thin volume representation with uncertainty awareness[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 2521-2531.

[28] LI Y, LI W Y, ZHAO Z J, *et al*. DRI-MVSNet: a depth residual inference network for multi-view stereo images [J]. *PLoS One*, 2022, 17(3): e0264721.

[29] LI H, XIONG P, AN J, *et al*. Pyramid Attention Network for Semantic Segmentation [EB/OL]. 2018: *arXiv*: 1805.10180. https://arxiv.org/abs/1805.10180

[30] ZHANG K, LIU M Y, ZHANG J L, *et al*. PA-MVSNet: sparse-to-dense multi-view stereo with pyramid attention [J]. *IEEE Access*, 2021, 9: 27908-27915.

[31] ANZHU, YU, *et al*. Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, 175: 448-460.

[32] LI J J, BAI Z Y, CHENG W, *et al*. Feature pyramid multi-view stereo network based on self-attention mechanism[C]. *Proceedings of the* 2022 5*th International Conference on Image and Graphics Processing*. *January* 7 - 9, 2022, *Beijing*, *China*. New York: ACM, 2022: 226-233.

[33] PARMAR N, RAMACHANDRAN P, VASWANI A, *et al*. Stand-alone self-attention in vision models[C]. 2019 *Neural Information Processing Systems* (*NeurIPS*). *Vancouver*, *Canada*. ACM, 2019: 13.

[34] ZHANG X D, HU Y T, WANG H C, *et al*. Long-range attention network for multi-view stereo [C]. 2021 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*). *January* 3-8, 2021, *Waikoloa*, *HI*, *USA*. IEEE, 2021: 3781-3790.

[35] LIU W J, WANG J K, QU H C, *et al*. Hierarchical MVSNet with cost volume separation and fusion based on U-shape feature extraction[J]. *Multimedia Systems*, 2023, 29(1): 377-387.

[36] WOO S, PARK J, LEE J Y, *et al*. *CBAM: Convolutional Block Attention Module*[M]. Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[37] CAO C, REN X, FU Y. MVSFormer: multi-view stereo with pre-trained vision transformers and temperature-based depth [J]. *arXiv preprint arXiv*:2208.02541, 2022.

[38] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. 2020: *arXiv*: 2010.11929. https://arxiv.org/abs/2010.11929

[39] SAEED S, LEE S, CHO Y, *et al*. ASPPMVS-Net: a high-receptive-field multiview stereo network for dense three-dimensional reconstruction [J]. *ETRI Journal*, 2022, 44(6): 1034-1046.

[40] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834-848.

[41] WEI Z Z, ZHU Q T, MIN C, *et al*. AA-RMVS-Net: adaptive aggregation recurrent multi-view stereo network[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 6167-6176.

[42] MASSON J E N, PETRY M R, COUTINHO D F, *et al*. Deformable convolutions in multi-view stereo[J]. *Image and Vision Computing*, 2022,

118: 104369.

[43] CHENG W, BAI Z Y, LI J J, *et al*. ADIM-MVSNet: adaptive depth interval multi-view stereo network for 3d reconstruction[C]. *Proceedings of the* 2022 5*th International Conference on Image and Graphics Processing. January* 7-9, 2022, *Beijing*, *China*. New York: ACM, 2022: 281-287.

[44] DAI J F, QI H Z, XIONG Y W, *et al*. Deformable convolutional networks[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 764-773.

[45] DING Y K, YUAN W T, ZHU Q T, *et al*. TransMVSNet: global context-aware multi-view stereo network with transformers[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8575-8584.

[46] GIANG KT, SONG S, JO S. Curvature-guided dynamic scale networks for multi-view stereo[J]. *arXiv preprint arXiv*:2112.05999, 2022.

[47] YAN J F, WEI Z Z, YI H W, *et al*. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking[C]. *Computer Vision - ECCV* 2020: 16*th European Conference*, *Glasgow*, *UK*, 23-28, 2020, *Proceedings*, *Part IV*. New York: ACM, 2020: 674-689.

[48] YU Z H, GAO S H. Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 1946-1955.

[49] CHEN R, HAN S F, XU J, *et al*. Visibility-aware point-based multi-view stereo network[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 43(10): 3695-3708.

[50] WEILHARTER R, FRAUNDORFER F. HighRes-MVSNet: a fast multi-view stereo network for dense 3D reconstruction from high-resolution images[J]. *IEEE Access*, 2021, 9: 11306-11315.

[51] YI P, TANG S, YAO J. DDR-Net: Learning Multi-Stage Multi-View Stereo with Dynamic Depth Range [EB/OL]. 2021: *arXiv*: 2103.14275. https://arxiv.org/abs/2103.14275

[52] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems. December* 4 - 9, 2017, *Long Beach*, *California*, *USA*. New York: ACM, 2017: 6000-6010.

[53] WANG X F, ZHU Z, HUANG G, *et al*. *MVSTER*: *Epipolar Transformer for Efficient Multi-View Stereo*[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 573-591.

[54] ZHU J, PENG B, LI W, *et al*. Multi-view Stereo with Transformer [EB/OL]. 2021: *arXiv*: 2112.00336. https://arxiv.org/abs/2112.00336

[55] LIAO J, DING Y, SHAVIT Y, *et al*. WT-MVSNet: Window-Based Transformers for Multi-View Stereo [EB/OL]. 2022: *arXiv*: 2205.14319. https://arxiv.org/abs/2205.14319

[56] HE Y H, YAN R, FRAGKIADAKI K, *et al*. Epipolar transformers[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 7776-7785.

[57] CHEN P H, YANG H C, CHEN K W, *et al*. MVSNet: learning depth-based attention pyramid features for multi-view stereo[J]. *IEEE Transactions on Image Processing*, 2020, 29: 7261-7273.

[58] BENGIO Y, LOURADOUR J, COLLOBERT R, *et al*. Curriculum learning[C]. *Proceedings of the* 26*th Annual International Conference on Machine Learning. June* 14 - 18, 2009, *Montreal*, *Quebec*, *Canada*. New York: ACM, 2009: 41-48.

[59] GUO X Y, YANG K, YANG W K, *et al*. Group-wise correlation stereo network[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 3268-3277.

[60] XU Q S, TAO W B. Learning inverse depth regression for multi-view stereo with correlation cost volume[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 12508-12515.

[61] XU Q, TAO W. PVSNet: Pixelwise Visibility-Aware Multi-View Stereo Network [EB/OL]. 2020: *arXiv*: 2007.07714. https://arxiv.org/abs/2007.07714

[62] WANG F, GALLIANI S, VOGEL C, *et al*.

PatchmatchNet: learned multi-view patchmatch stereo[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 14189-14198.

[63] SONG B, HU X, XIAO J, *et al*. Implicit neural refinement based multi-view stereo network with adaptive correlation[J]. *Image and Vision Computing*, 2022, 124: 104511.

[64] CAI Y C, LI L, WANG D, *et al*. MFNet: Multi-level fusion aware feature pyramid based multi-view stereo network for 3D reconstruction[J]. *Applied Intelligence*, 2023, 53(4): 4289-4301.

[65] GAO S Y, LI Z X, WANG Z Q. *Cost Volume Pyramid Network with Multi-Strategies Range Searching for Multi-View Stereo*[M]. Advances in Computer Graphics. Cham: Springer Nature Switzerland, 2022: 157-169.

[66] LUO K Y, GUAN T, JU L L, *et al*. P-MVSNet: learning patch-wise matching confidence aggregation for multi-view stereo[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27 - *November* 2, 2019, *Seoul*, *Korea* (*South*). IEEE, 2020: 10451-10460.

[67] MA X J, GONG Y, WANG Q R, *et al*. EPP-MVSNet: epipolar-assembling based depth prediction for multi-view stereo[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 5712-5720.

[68] PENG R, WANG R J, WANG Z Y, *et al*. Rethinking depth estimation for multi-view stereo: a unified representation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8635-8644.

[69] XU H F, ZHANG J Y. AANet: adaptive aggregation network for efficient stereo matching [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 1956-1965.

[70] SORMANN C, KNÖBELREITER P, KUHN A, *et al*. BP-MVSNet: belief-propagation-layers for multi-view-stereo[C]. 2020 *International Conference on 3D Vision* (*3DV*). *November* 25-28, 2020, *Fukuoka*, *Japan*. IEEE, 2021: 394-403.

[71] QI Y, SU W, XU Q, *et al*. Sparse prior guided deep multi-view stereo[J]. *Computers & Graphics*, 2022, 107: 1-9.

[72] LIU J, JI S P. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 6049-6058.

[73] XU Q, OSWALD M R, TAO W, *et al*. Non-local recurrent regularization networks for multi-view stereo [J]. *arXiv preprint arXiv*: 2110.06436, 2021.

[74] WANG F, GALLIANI S, VOGEL C, *et al*. IterMVS: iterative probability estimation for efficient multi-view stereo[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8596-8605.

[75] MI Z X, DI C, XU D. Generalized binary search network for highly-efficient multi-view stereo[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 12981-12990.

[76] LEE J Y, DEGOL J, ZOU C H, *et al*. PatchMatch-RL: deep mvs with pixelwise depth, normal, and visibility[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 6138-6147.

[77] YANG J Y, ALVAREZ J M, LIU M M. Non-parametric depth distribution modelling based depth inference for multi-view stereo[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8616-8624.

[78] WANG S Q, LI B, DAI Y C. Efficient multi-view stereo by iterative dynamic cost volume[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8645-8654.

[79] LI Y, ZHAO Z, FAN J, *et al*. ADR-MVSNet: a cascade network for 3D point cloud reconstruction with pixel occlusion [J]. *Pattern Recognition*, 2022, 125: 108516.

[80] LAFFERTY J, MCCALLUM A, PEREIRA

FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. *Proc. 18th International Conf. on Machine Learning*. 2001.

[81] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, *et al*. Conditional random fields as recurrent neural networks[C]. 2015 *IEEE International Conference on Computer Vision* (*ICCV*). 7-13, 2015, *Santiago*, *Chile*. IEEE, 2016: 1529-1537.

[82] KNÖBELREITER P, SORMANN C, SHEKHOVTSOV A, *et al*. Belief propagation reloaded: learning BP-Layers for labeling problems[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 7897-7906.

[83] LUO K Y, GUAN T, JU L L, *et al*. Attention-aware multi-view stereo [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 1587-1596.

[84] WEI Z Z, ZHU Q T, MIN C, *et al*. Bidirectional hybrid LSTM based recurrent neural network for multi-view stereo[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2022, PP (99): 1.

[85] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318-327.

[86] ZHOU H Z, ZHAO H L, WANG Q, *et al*. Miper-MVS: multi-scale iterative probability estimation with refinement for efficient multi-view stereo [J]. *Neural Networks*, 2023, 162: 502-515.

[87] DING Y K, LI Z Y, HUANG D H, *et al*. Enhancing Multi-View stereo with contrastive matching and weighted focal loss[C]. 2022 *IEEE International Conference on Image Processing* (*ICIP*). *October* 16-19, 2022, *Bordeaux*, *France*. IEEE, 2022: 821-825.

[88] IBRAHIMLI N, LEDOUX H, KOOIJ J, *et al*. DDL-MVS: Depth Discontinuity Learning for MVS Networks [EB/OL]. 2022: *arXiv*: 2203.01391. https://arxiv.org/abs/2203.01391

[89] KHOT T, AGRAWAL S, TULSIANI S, *et al*. Learning unsupervised multi-view stereopsis via robust photometric consistency [J]. *arXiv preprint arXiv*: *1905.02706v2*, 2019.

[90] DAI Y C, ZHU Z D, RAO Z B, *et al*. MVS2: Deep unsupervised multi-view stereo with multi-view symmetry[C]. 2019 *International Conference on 3D Vision* (*3DV*). *September* 16-19, 2019, *Quebec City*, *QC*, *Canada*. IEEE, 2019: 1-8.

[91] MALLICK A, STÜCKLER J, LENSCH H. Learning to adapt multi-view stereo by self-supervision[J]. *arXiv preprint arXiv*:2009.13278, 2020.

[92] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]. *Proceedings of the 34th International Conference on Machine Learning - Volume 70. August 6 - 11, 2017, Sydney, NSW, Australia*. New York: ACM, 2017: 1126-1135.

[93] HUANG B C, YI H W, HUANG C, *et al*. M3VSNET: unsupervised multi-metric multi-view stereo network[C]. 2021 *IEEE International Conference on Image Processing* (*ICIP*). 19-22, 2021, *Anchorage*, *AK*, *USA*. IEEE, 2021: 3163-3167.

[94] XU H B, ZHOU Z P, QIAO Y, *et al*. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(4): 3030-3038.

[95] YANG J Y, ALVAREZ J M, LIU M M. Self-supervised learning of depth inference for multi-view stereo[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 7522-7530.

[96] XU H B, ZHOU Z P, WANG Y L, *et al*. Digging into uncertainty in self-supervised multi-view stereo[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 6058-6067.

[97] QI S, SANG X, YAN B, *et al*. Unsupervised multi-view stereo network based on multi-stage depth estimation [J]. *Image and Vision Computing*, 2022, 122: 104449.

[98] DONG H, YAO J. PatchMVSNet: patch-wise unsuper-vised multi-view stereo for weakly-textured surface reconstruction [J]. *arXiv preprint arXiv*:2203.02156, 2022.

[99] CHANG D, BOŽIČ A, ZHANG T, *et al*. *RC-MVSNet: Unsupervised Multi-View Stereo with*

Neural Rendering[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 665-680.

[100] MILDENHALL B, SRINIVASAN P P, TANCIK M, *et al*. NeRF[J]. *Communications of the ACM*, 2022, 65(1): 99-106.

[101] CHEN A P, XU Z X, ZHAO F Q, *et al*. MVSNeRF: fast generalizable radiance field reconstruction from multi-view stereo[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2022: 14104-14113.

[102] XU Q G, XU Z X, PHILIP J, *et al*. Point-nerf: point-based neural radiance fields[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 5428-5438.

[103] ZHANG J Z, JI M Q, WANG G Y, *et al*. SurRF: unsupervised multi-view stereopsis by learning surface radiance field[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7912-7927.

[104] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[EB/OL]. 2015: *arXiv*: 1503.02531. https://arxiv.org/abs/1503.02531

[105] DING Y K, ZHU Q T, LIU X Y, *et al*. *KD-MVS: Knowledge Distillation Based Self-Supervised Learning for Multi-View Stereo*[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 630-646.

[106] UMMENHOFER, BENJAMI, VLADLEN KOLTUN, *et al*. Adaptive surface reconstruction with multiscale convolutional kernels[J]. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*: 5631-5640.

[107] SCHÖPS T, SATTLER T, POLLEFEYS M. BAD SLAM: bundle adjusted direct RGB-D SLAM[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 134-144.

[108] YAO Y, LUO Z X, LI S W, *et al*. BlendedMVS: a large-scale dataset for generalized multi-view stereo networks[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 1787-1796.

[109] ZHANG J N, ZHANG J Z, MAO S, *et al*. GigaMVS: a benchmark for ultra-large-scale gigapixel-level 3D reconstruction[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(11): 7534-7550.

[110] MA Z Y, TEED Z, DENG J. *Multiview Stereo with Cascaded Epipolar RAFT*[M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 734-750.

[111] LI Z X, ZUO W M, WANG Z Q, *et al*. Confidence-based large-scale dense multi-view stereo[J]. *IEEE Transactions on Image Processing*, 2020, 29: 7176-7191.

[112] LI W, ZHU D, WANG Q. A single view leaf reconstruction method based on the fusion of ResNet and differentiable render in plant growth digital twin system[J]. *Computers and Electronics in Agriculture*, 2022, 193: 106712.

[113] DENG X P, QIU S, JIN W Q, *et al*. Three-dimensional reconstruction method for bionic compound-eye system based on MVSNet network[J]. *Electronics*, 2022, 11(11): 1790.

[114] 郝雯, 张雯静, 梁玮, 等. 面向三维点云的场景识别方法综述[J]. 光学 精密工程, 2022, 30(16): 1988-2005.
HAO W, ZHANG W J, LIANG W, *et al*. Scene recognition for 3D point clouds: a review[J]. *Opt. Precision Eng.*, 2022, 30(16): 1988-2005. (in Chinese)

[115] EBNER T, FELDMANN I, RENAULT S, *et al*. Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications[J]. *Journal of the Society for Information Display*, 2017, 25(3): 151-157.

[116] 李兆歆, 蒋浩, 刘衍青, 等. 丝路文化虚拟体验中的多视角立体重建技术研究[J]. 计算机学报, 2022, 45(3): 500-512.
LI Z X, JIANG H, LIU Y Q, *et al*. Research on multi-view stereo 3D reconstruction in virtual reality system of silk road cultural inheritance[J]. *Chinese Journal of Computers*, 2022, 45(3): 500-512. (in Chinese)

[117] 余加勇, 薛现凯, 陈昌富, 等. 基于无人机倾斜

摄影的公路边坡三维重建与灾害识别方法[J]. 中国公路学报, 2022, 35(4): 77-86.
YU J Y, XUE X K, CHEN C F, *et al*. Three-dimensional reconstruction and disaster identification of highway slope using unmanned aerial vehicle-based oblique photography technique[J]. *China Journal of Highway and Transport*, 2022, 35(4): 77-86. (in Chinese)

[118] HU Z, HOU Y, TAO P, *et al*. IMGTR: Image-triangle based multi-view 3D reconstruction for urban scenes[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, 181: 191-204.

[119] ORSINGHER M, ZANI P, MEDICI P, *et al*. Revisiting patchmatch multi-view stereo for urban 3D reconstruction[C]. 2022 *IEEE Intelligent Vehicles Symposium* (*IV*). 4-9, 2022, *Aachen*, *Germany*. IEEE, 2022: 190-196.

[120] ZHOU Y X, EIMEN R L, SEIBEL E J, *et al*. Cost-efficient video synthesis and evaluation for development of virtual 3D endoscopy[J]. *IEEE Journal of Translational Engineering in Health and Medicine*, 2021, 9: 1-11.

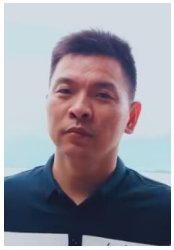
[121] 何东健, 熊虹婷, 芦忠忠, 等. 基于多视角立体视觉的拔节期玉米水分胁迫预测模型[J]. 农业机械学报, 2020, 51(6): 248-257.
HE D J, XIONG H T, LU Z Z, *et al*. Predictive model of maize moisture stress during jointing stage based on multi-view stereo vision [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2020, 51(6): 248-257. (in Chinese)

[122] 王思启, 张家强, 李丽圆, 等. MVSNet在空间目标三维重建中的应用[J]. 中国激光, 2022, 49(23): 2310003.
WANG S Q, ZHANG J Q, LI L Y, *et al*. Application of MVSNet in 3D reconstruction of space objects [J]. *Chinese Journal of Lasers*, 2022, 49(23): 2310003. (in Chinese)

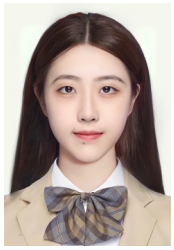
[123] GÓMEZ A, RANDALL G, FACCIOLO G, *et al*. An Experimental comparison of multi-view stereo approaches on satellite images[C]. 2022 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). 3-8, 2022, *Waikoloa*, *HI*, *USA*. IEEE, 2022: 707-716.

[124] LU J C, LI Y X, ZUO Z C. *SatMVS: A Novel 3D Reconstruction Pipeline for Remote Sensing Satellite Imagery*[M]. Lecture Notes in Electrical Engineering. Singapore: Springer Nature Singapore, 2022: 521-538.

作者简介：



鄢化彪(1978—),男,江西丰城人,副教授,硕士生导师,分别于2002年、2008年在江西理工大学获得学士、硕士学位,主要从事复杂系统建模及深度学习方面的研究。E-mail: yanhuabiao@jxust.edu.cn



徐方奇(2000—),女,河南长垣人,硕士研究生,2017年于上海海洋大学获得学士学位,现就读于江西理工大学,主要从事深度学习、三维重建方面的研究。E-mail:xufangqi777@163.com

通讯作者：

黄绿娥(1981—),女,江西井冈山人,博士,副教授,2008年于北京交通大学获得硕士学位,2019年于南昌大学获得机械工程博士学位,主要从事图像处理及深度学习方面的研究。E-mail: 9320080310@jxust.edu.cn