第 49 卷 第 23 期/2022 年 12 月/中国激光 研究论文

中国激光

MVSNet 在空间目标三维重建中的应用

王思启[1,3]，张家强[1,3]，李丽圆[1,3*]，李潇雁[1,2]，陈凡胜[1,2]

[1]中国科学院上海技术物理研究所中国科学院智能红外感知重点实验室，上海 200083；

[2]国科大杭州高等研究院，浙江 杭州 310024；

[3]中国科学院大学，北京 100049

摘要 对空间目标进行三维重建能够为在轨服务卫星提供服务对象的结构信息，是提高系统自主性的关键技术。受空间目标的结构对称性以及成像非朗伯特性的影响，传统的重建方法存在特征点匹配错误或特征点匹配不足的问题，重建精度低。针对该问题，提出了一种基于 MVSNet 深度学习网络实现空间目标三维重建的方法，利用深度学习提取图像高层语义，提高了立体匹配的鲁棒性。首先，基于空间目标的成像特点，分析了模型的几何结构和材质对重建结果的影响，设计了搭建在 Blender 平台上的空间目标多视图采集系统。然后，基于 MVSNet 深度学习网络，采用多尺度卷积充分提取了图像的深度特征，并通过编码解码结构融合和规整上下文信息进行了立体匹配，有效解决了传统方法重建卫星的弱纹理、反射、重复纹理等区域时对特征点的高度依赖问题。最后通过残差网络解决了多次卷积造成的边界过平滑问题，进一步提升了重建效果。实验结果表明，所设计的重建模型的平均准确度误差为 0.449 mm，平均完整度误差为 0.379 mm，误差综合评价为 0.414 mm，精度较经典开源软件 COLMAP 提升了 20%。该方法为空间操作自动化提供了技术参考，进一步推动了三维重建在相关领域中的应用。

关键词 遥感；深度学习；多视图；空间目标三维重建；卷积神经网络；编码解码结构

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/CJL202249.2310003

1 引 言

随着科技的发展，人类对太空的探索活动日益频繁。为了满足空间在轨服务的更高要求，需要提升卫星的探测感知能力，在执行任务之前获取目标的光学特性、材料和三维结构等信息[1]。三维重建算法可以帮助导航卫星获取目标本体、帆板以及感兴趣部位的结构尺寸等先验信息，辅助完成目标运动状态的估计，是实现自动化空间操作的关键技术[2]。

三维重建的方法按照重建目标深度信息获取的方式可以分为主动式和被动式。主动式方法通过解析经激光、超声波或者红外线等能量源照射后物体返回的信号来计算物体表面的深度信息。常见的方法有结构光法[3-5]、三角测距法[6]和飞行时间法[7-9]等。然而，空间目标处在复杂的太空环境中，难以通过主动测量的方式获取目标的深度信息。被动式方法则使用图像集作为数据源，基于几何原理计算得到物体的深度信息。近年来，基于该方法的重建软件层出不穷。Moulon 等[10]提出了基于自适应参数的运动恢复结构算法，虽然重建结果的精度较高，但是时间复杂度较高。Bao 等[11]提出了语义式运动恢复结构算法，通过对三维场景中区域、物体等高层语义信息的识别与估计，提高了算法的鲁棒性。Wu 等[12]提出了 VisualSFM 算法，通过预处理共轭梯度法进行改进，在保证精度的前提下降低了时间复杂度。Schönberger 等[13]提出了 COLMAP 算法，该算法主要对几何校正、视角选择、三角化等多个关键步骤进行改进，在重建精度与完整度方面取得了较大的进展，是目前传统被动式方法中性能最佳的算法之一。

虽然这些传统的被动式方法在理想朗伯反射体模型中显示出了很好的效果，但它们依赖人工设计的相似性指标和正则化恢复三维点，在度量弱纹理、重复纹理、非朗伯特性区域的相似性时准确度较低。针对空间目标的重建任务，存在的问题主要包括两个方面：1）常见的空间目标如卫星会在光照条件下呈现非朗伯特性，导致不同角度视图下物体表面同一点的辐射能量不同，无法成功匹配；2）由于人造卫星的对称几何特性，多组像素块高度相似，无法区分，立体匹配发生错误。

随着学者们对深度学习方法的深入研究，人们发现通过卷积操作可以引入全局语义信息，在有监督的条件下针对性地提取特征，解决弱纹理、非朗伯特性等

收稿日期：2022-03-14；**修回日期**：2022-03-26；**录用日期**：2022-04-07

基金项目：国家自然科学基金（61975222）、中国科学院地球微卫星热红外光谱仪项目（XDA19010102）

通信作者：*liliyuan@mail.sitp.ac.cn

不利条件下的立体匹配难题[14-16]。Eigen 等[17]基于两尺度深度网络进行单视图的深度估计，利用粗尺度网络预测图片的全局深度，利用细尺度网络优化局部细节。基于上述工作，该团队随后采用更深的基础网络 VGGNet 设计了三尺度网络框架，利用第三个细尺度网络进一步增添了细节信息，提高了重建分辨率[18]。Choy 等[19]采用卷积神经网络(CNN)与 长短期记忆(LSTM)网络结合的方式，提出了一种同时适用于单视角和多视角重建的网络，但该算法高度依赖三维计算机辅助设计模型和训练集质量。Ji 等[20]提出的 SurfaceNet 算法和 Kar 等[21]提出的 LSM 算法，将图像信息与相机参数结合并融入网络中，基于 3D CNN 算法逐体素重建物体，提高了重建精度。但以上算法生成的都是体素网格，在网格分辨率比较大的情况下显存开销较大，学习效率低。香港科技大学的权龙团队提出了 MVSNet 网络模型，实现了多视图的深度估计[22]。该方法基于相机视锥结构建立体匹配代价体，并且每次只处理一张视图，极大减小了显存开销，使得大规模重建成为可能。但当前实验研究多集中在日用品和大规模建筑的重建方面，并未对目标特性分析、目标多角度视图采集、目标建模方法和目标建模效果的评价方法等展开系统研究，且缺少公开的大规模数据集[23]。

针对以上问题，本文提出了一种基于 MVSNet 深度学习网络实现空间目标三维重建的方案。利用多尺度不同结构的卷积网络，充分提取图像的高层语义信息，提高了空间目标在纹理稀疏、纹理重复等不利环境下的重建精度。本文仿真制作了空间目标模型的多角度视图数据集，利用主观目视评价和客观量化评价两种方式，通过对比基于传统方法的经典重建软件 VisualSFM 和 COLMAP 及 SurfaceNet 的重建效果，证明了该方案的有效性。此外，本文还对立体匹配中所使用的视图数量对建模效果的影响和不同重建算法的系统运行速度进行了实验分析。

2 空间目标的成像特性

在本文研究的在轨服务领域中，空间目标是指在轨工作的航天器。相较于自然图像，空间目标的成像图在重建时往往会产生较多错误，其主要原因有以下三点：

1)纹理稀疏。多数空间目标的外观颜色单一，常为白色和灰色，难以提取有效特征点进行匹配。

2)纹理重复。空间目标中往往存在帆板等多个重复图像的部件，这部分图像的特征点差异较小，在匹配时容易发生混淆。

3)镜面反射。部分金属材质的空间目标舱体在光照条件下呈现非朗伯特性，不同视角下的高光点位置不同，造成邻接视图相差过大，匹配错误。

3 基于 MVSNet 模型的三维重建网络流程

本文基于 MVSNet 模型实现多视图的深度估计，网络流程如图 1 所示，其中 X_1、X_2、X_3 为物体空间中的三个点，X'_1、X'_2、X'_3 分别为 X_1、X_2、X_3 在参考图像相机平面上的三个投影点。首先，利用三种不同尺度的 8 层二维卷积网络，提取重建目标的不同层次的图像特征。随后，基于立体匹配的视差理论构造代价体。接着，使用三维卷积网络对代价体进行正则化处理，并加强上下文语义，解决了某些视角下遮挡造成的信息缺失问题。最后，通过 SoftMax 函数将正则化后的代价体转换为概率体，遍历像素点，确定最优深度估计结果。为了解决多次卷积造成的图像边界深度特征损失问题，通过设计残差网络结构，对估计结果进行了优化，得到了最终的深度图。

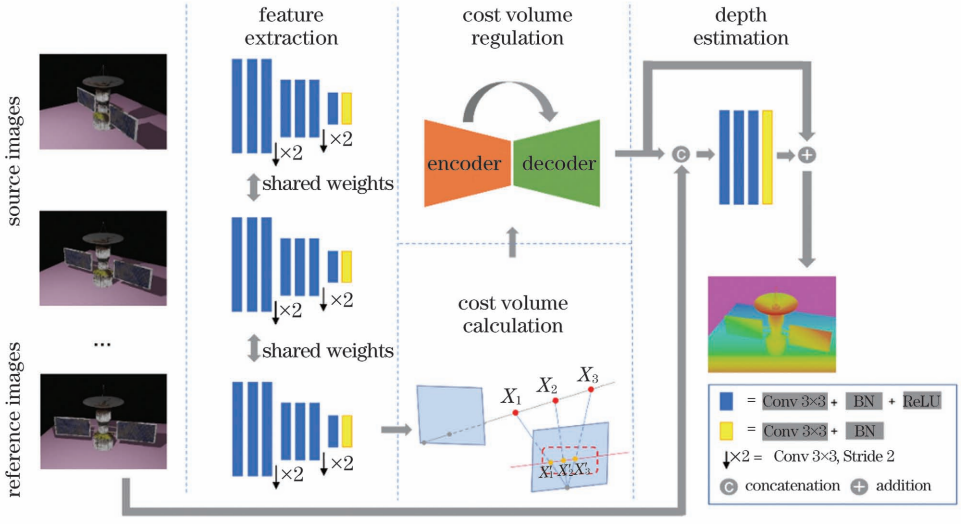

图 1 MVSNet 网络结构示意图

Fig. 1 Structural diagram of MVSNet network

3.1 代价体的构造方法

传统的立体匹配方法是先对图像进行极线纠正，再沿着水平极线扫描匹配像素。处理纹理重复的场景时，由于一个视图里的像素与另一视图的多个像素相似，这种方法会发生匹配错误。

MVSNet 网络是一种基于平面扫描算法[24]的描述不同视图像素间映射关系的方法。平面扫描算法的基本思想是用一组划分得足够密集的平行平面对物体进行等距切割，那么物体表面上的任意一点一定位于某平面上。对于物体表面上的点，不同相机显示的颜色必定相同或相近；对于不在物体表面上的点，不同相机显示的颜色差异较大。因此，对于每一个像素点，可以先假设它的深度。如果深度假设正确，通过单应变换公式求出的对应位置处的像素点与原像素点有很小的差异或没有差异，采用方差对该差异值进行表征，从而使得输入图像数量不受限制。首先，通过 2D CNN 网络提取图像在不同空间尺度上的深度特征信息，如图 2 所示。网络共有 8 层，通过设置网络的第三层和第六层的步长为 2，构建三个不同尺度的特征。对于每一个尺度，用两个卷积层提取图像的高层语义信息；在每个卷积层后，使用一个批归一化（BN）层和一个线性修正单元（ReLU）层提高模型的拟合能力。

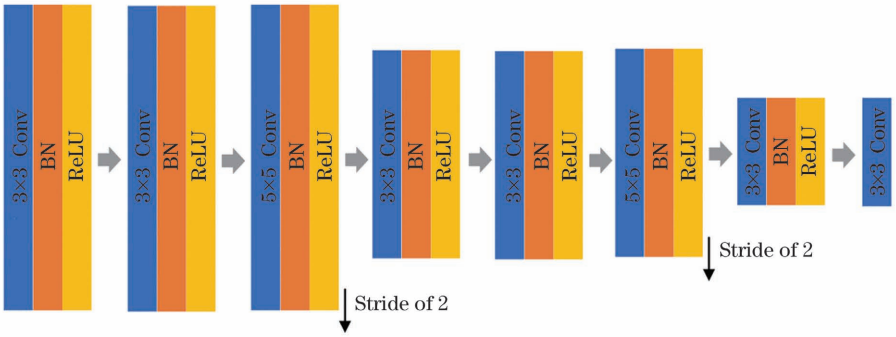

图 2 特征提取网络的架构图

Fig. 2 Architecture diagram of feature extraction network

然后，基于单应变换公式，将其他图像特征映射到参考图上，形成匹配代价体。

$$\boldsymbol{H}_i(d)=\boldsymbol{K}_i\cdot\boldsymbol{R}_i\cdot\left[\boldsymbol{F}-\frac{(\boldsymbol{t}_1-\boldsymbol{t}_i)\cdot\boldsymbol{n}_1^{\mathrm{T}}}{d}\right]\cdot\boldsymbol{R}_1^{\mathrm{T}}\cdot\boldsymbol{K}_1^{-1},\tag{1}$$

$$V_i=\frac{W}{4}\cdot\frac{H}{4}\cdot M\cdot D,\tag{2}$$

式中：$\boldsymbol{F}$ 为图像的特征图集合；$\boldsymbol{H}_i(d)$ 为第 i 个特征图变换到参考图像相机视锥深度 d 处平面时的单应矩阵；$\boldsymbol{n}_1$ 为参考图像相机的主光轴方向向量；$\boldsymbol{K}_i$，$\boldsymbol{R}_i$，$\boldsymbol{t}_i(i=1,2,\cdots,N)$ 为各个图像对应相机的内参数、旋转矩阵和位移向量，N 为特征体个数；V_i 为以第 i 个特征图作为参考图时形成的特征体；W、H、M、D 分别为特征图分辨率宽度、特征图分辨率长度、深度采样数和特征图通道数。

最后，使用方差计算代价体 C：

$$C=\frac{\sum_{i=1}^{N}(V_i-\overline{V}_i)^2}{N},\tag{3}$$

式中：$\overline{V}_i$ 为以第 i 个特征图作为参考图时形成的所有特征体的均值。

3.2 代价体正则化

对于不同视角下的图像，会出现视觉遮挡和非朗伯反射体表面高光点变化等问题，不能实现逐个像素点的一一映射，所以还需要编码译码结构对上面的代价体进行正则化处理，以加强上下文语义信息。MVSNet 网络模仿语义分割问题中经典的网络 UNet[25]，具有基于 3D CNN 的“编码-解码”架构。首先，通过三次下采样操作，编码抽取出四个尺度的抽象特征。随后，通过逐步解码，恢复位置信息。每个对应尺度的编码与解码间有直接的跳跃连接，可实现多尺度特征信息的聚合。最终得到通道数为 1 的正则化代价体，如图 3 所示。

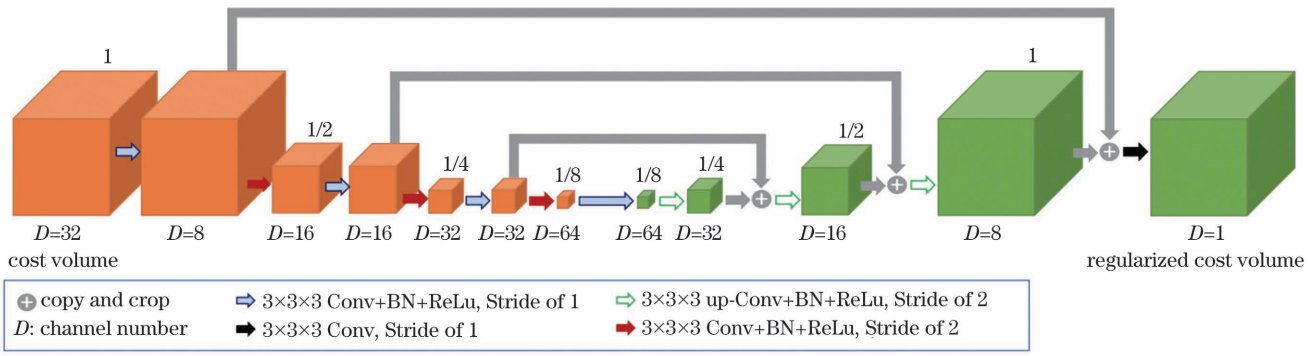

图 3 代价体正则化网络架构图

Fig. 3 Architecture diagram of cost volume regularization network

3.3 深度估计

为了更好地进行深度值估计，使用 SoftMax 分类操作将正则化的代价体 C 转换为概率体 P，即沿深度采样维度，将每个像素点的匹配代价的特征离散程度值转化为不同深度采样值下的概率分布值。

接下来是遍历像素点，根据其对应的深度采样值的概率分布确定最优的深度估计结果 d_{E}。网络采用深度值回归时的 Soft Argmin 操作[26]来估计像素点的深度值：

$$d_{\mathrm{E}}=\sum_{d=d_{\min}}^{d_{\max}} d\times P(d), \tag{4}$$

式中：$P(d)$为像素在深度 d 处的概率估计值；$d_{\max}$ 和 $d_{\min}$ 分别为深度采样的最大值和最小值，一般是采用 SFM 算法对图片进行预处理后得到的。本文通过 COLMAP 软件对图片进行预处理，确定深度采样范围。

由于进行了多次卷积操作，图像边界的深度特征会因为较大的感受野而损失，出现边界过平滑的现象。针对这一问题，基于图像抠图算法[27]中残差网络的设计，利用原始参考图的边界信息优化深度估计结果。首先，将三通道的初始参考图等比例缩放四分之一，然后与单通道的深度图拼接，形成四通道特征图，并将其输入到残差网络中。最后，为了保证负残差值也能进行学习，将残差值与初始估计的深度图逐像素进行相加后直接输出，得到优化后的深度图。

3.4 损失函数设计

为了能够同时评估初始深度图和优化后深度图的损失，本文使用两者与地面真值的绝对差值之和作为损失函数。由于测量方法的限制，图像的地面真值深度图并不是完整的，因此本文设计的损失函数只计算包含有效地面真值标签的像素点深度估计误差：

$$L=\sum_{p\in P_{\mathrm{valid}}}\left\|\boldsymbol{d}(p)-\hat{\boldsymbol{d}}_{\mathrm{i}}(p)\right\|_1+\lambda\cdot\left\|\boldsymbol{d}(p)-\hat{\boldsymbol{d}}_{\mathrm{r}}(p)\right\|_1, \tag{5}$$

式中：L 表示损失函数；P_{valid} 表示有效的地面真值像；$\boldsymbol{d}(p)$表示像素 p 的地面真值深度值；$\hat{\boldsymbol{d}}_{\mathrm{i}}(p)$表示初始深度图估计值；$\hat{\boldsymbol{d}}_{\mathrm{r}}(p)$表示优化深度图估计值；$\lambda$ 为权重系数，可以通过调整其数值来控制初始估计的深度图和微调后的深度图对网络训练的影响程度。通过将 λ 分别设置为 0、0.01、0.05 和 1.00 来开展实验研究，可知 λ 设置为 1.00 时，神经网络的效果最好。因此，本文参数 λ 设置为 1.00。

4 实　　验

4.1 实验数据集

本实验使用的主要数据集来自 DTU 数据集[28]和自建的空间目标数据集。其中，DTU 是被广泛应用在多视图三维重建中的经典大规模数据集。它一共包含 124 个场景，删除部分质量较低的场景后，共计 119 个有效场景。每个场景含有 7 种不同光照条件下从 49 或 64 个不同视角采集到的二维图像，每张图片的分辨率为 1600 pixel×1200 pixel。

由于相机拍摄的真实空间目标图片十分有限，为了验证本文方法在空间目标重建方面的性能，本文自建了多角度空间目标视图集。采用了 62 种来自多个开源网站的不同卫星模型，并基于 Blender 软件，通过严格模拟真实场景进行采集。数据集的具体采集方法如下：

1)在模型的前后左右 4 个方向上布置相同高度的光源，确保获得一致均匀的自然环境光。

2)设定球形环绕的相机机位，每一层以 6°为间隔在不同方位采集目标缩比模型的图像序列。探测相机焦距为 35 mm，相机光轴始终指向目标中心，满足重建照片的多角度需求，相机机位示意图如图 4 所示。

3)对每一个模型分别采集上百张照片，图像分辨率为 1600 pixel×1200 pixel，确保立体匹配信息的完整性，提高目标模型重建的质量。

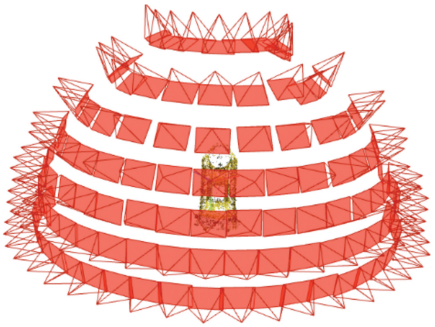

图 4 相机机位示意图

Fig. 4 Camera location map

为了公平地与其他深度学习方法进行对比，数据集的划分与 SurfaceNet 算法一样，按照 0.70∶0.15∶0.15 的数据量大小比例划分训练集、验证集和测试集。为了验证我们提出的重建方案的可靠性，采用主观目视评价和客观量化评价两种方法，对比了本文所提重建方案、基于传统方法的经典软件 VisualSFM 和 COLMAP 及深度学习网络 SurfaceNet 的重建效果。

4.2 网络模型训练

训练数据集的各项参数设置如下：匹配视图数量设为 3，图片的分辨率为 1600 pixel×1200 pixel，深度采样数设为 192，深度间隔大小设为 1.06。单应变换所需的每张图片对应的相机参数和深度估计需要的最近距离 $d_{\min}$ 是通过 COLMAP 软件对图片进行预处理而得到的。深度采样范围为 425 mm$\leqslant d\leqslant$935 mm。批大小(Batch size)设置为 1，使用 Adam 策略优化网络。当学习率过大时，则可能模型不收敛，损失值不断上下振荡；而当学习率过小时，模型收敛速度偏慢，需要更长的时间训练。因此，本实验在训练时动态调整学习率。模型共有 16 轮训练，学习率初始值设为 0.001，在第 10、12 和 14 轮后，将学习率降低为原来的 1/2。测试数据集的各项参数设置与训练数据集的各

项参数设置相同。模型的搭建依托于 Pytroch 框架，训练过程约持续 3 d。

5 实验结果分析

5.1 主观目视评估

DTU 数据集是在严格控制的实验室环境中拍摄的，缺乏与航天相关的模型，并不能很好地验证本文算法在空间目标重建方面的能力，因此本文采集 62 种空间目标模型的多角度视图进行三维重建实验，部分模型如图 5 所示。

下面以最具代表性的 Clementine 卫星模型为例，对比分析传统软件 VisualSFM 和 COLMAP、深度学习网络 SurfaceNet 及本文方法的重建结果。

5.1.1 非朗伯反射体区域的重建

如图 6 所示，卫星舱体由金属材质包裹，在光照下存在严重的非漫反射现象。传统方法立体匹配的基本

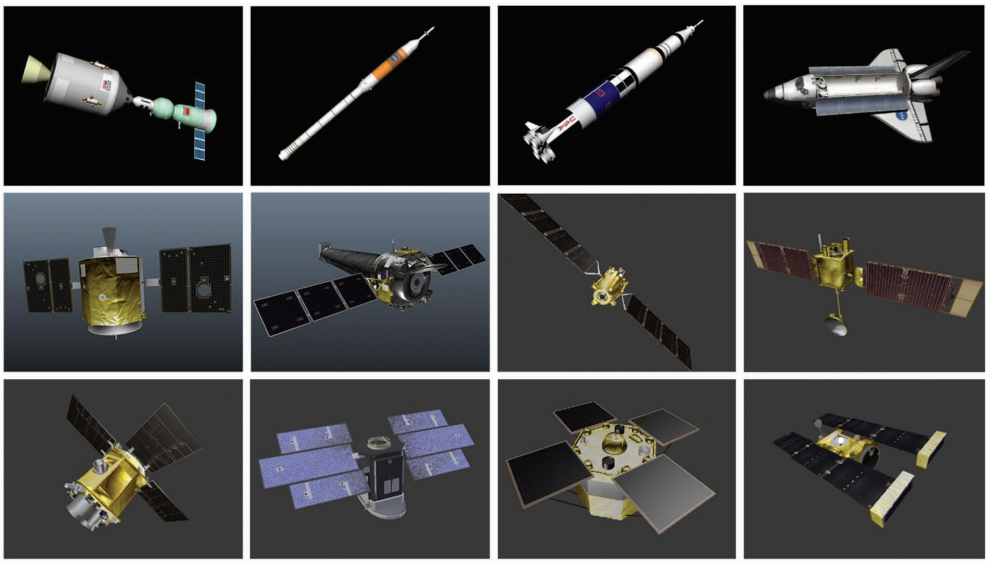

图 5 空间目标数据集

Fig. 5 Space target dataset

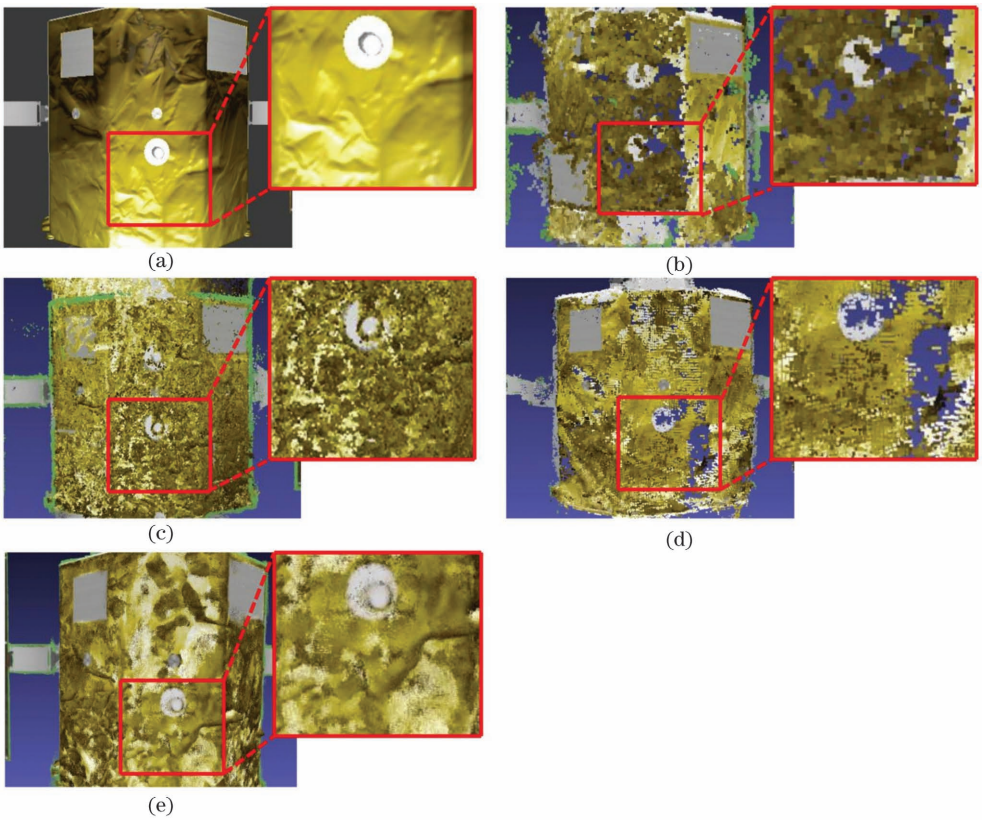

图 6 不同方法得到的 Clementine 卫星非朗伯反射体区域的重建结果。(a)原始图像；(b)VisualSFM；(c)COLMAP；(d)SurfaceNet；(e)本文方法

Fig. 6 Reconstruction results for non-Lambertian reflector region of Clementine satellite obtained by different methods. (a)Ground truth; (b)VisualSFM; (c)COLMAP; (d)SurfaceNet; (e) our method

假设是观测物体为理想的朗伯反射体，即物体为完全漫反射。观察该区域的重建效果可知，传统方法的性能退化明显。在 VisualSFM 的重建结果中，出现大量空洞，几乎不能提供表面细节信息。COLMAP 虽然恢复出了一部分高光点信息，但是杂点较多。基于体素重建的 SurfaceNet 虽然恢复出了较多的高光点信息，但是重建模型的表面仍存在一些空洞。本文算法使用了多尺度二维卷积结构，具有强大的特征表征能力，特征的强语义性有助于镜面反射区域的稳定立体匹配，所以本文方法重建的物体表面完整，杂点少，表面的细节也得到恢复。

5.1.2 弱纹理区域的重建

如图 7 所示，卫星的顶部呈银白色，纹理稀少。在卫星顶部的弱纹理区域，由于传统算法所依赖的图像特征没有足够的鲁棒性与表征能力，因此 VisualSFM 和 COLMAP 的重建结果中都出现了空洞（如图 7 虚线框所示）。而本文方法使用全局 3D CNN 进行回归，尽可能搜索周围的信息以找到更多正确配对的同名点，其推理的深度图在弱纹理区域的连续性更强。

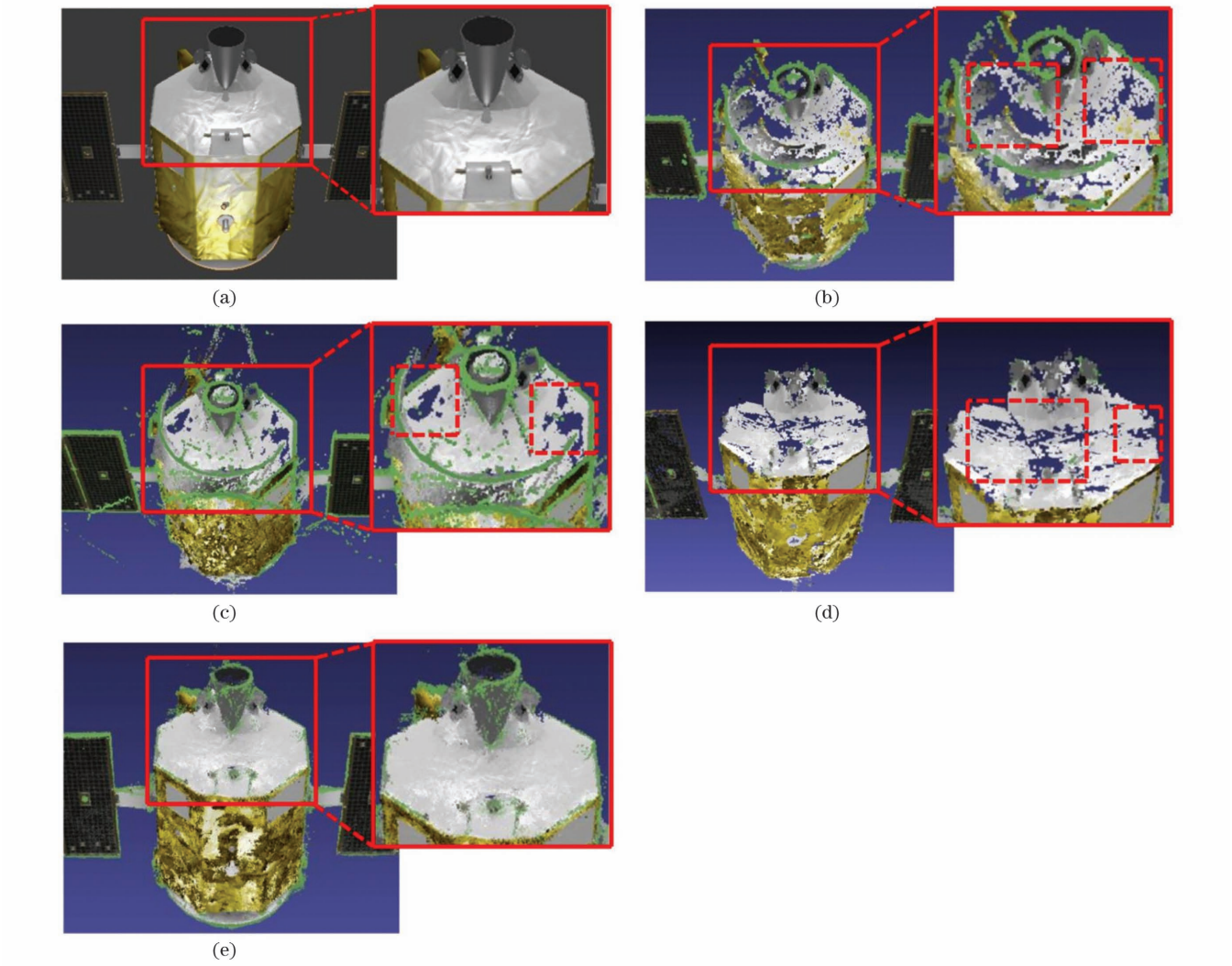

图 7 不同方法得到的 Clementine 卫星弱纹理区域的重建结果。(a)原始图像；(b)VisualSFM；(c)COLMAP；(d)SurfaceNet；(e)本文方法

Fig. 7 Reconstruction results for low-textured region of Clementine satellite obtained by different methods. (a)Ground truth; (b)VisualSFM; (c)COLMAP; (d)SurfaceNet; (e) our method

5.1.3 重复纹理区域的重建

如图 8 所示，卫星上下左右结构高度对称。观察该区域的重建效果可知，由于存在大量重复纹理，局部特征高度相似，VisualSFM 和 COLMAP 在立体匹配时均产生了错误。卫星的底端结构被重建在卫星的顶端上。深度学习方法 SurfaceNet 可以提取高层、全局的语义信息，从而有效地规避了局部特征相似带来的误匹配问题。本文方法借助于强推理的“编码-解码”结构，相较于 SurfaceNet 更加完整地恢复出了卫星的顶部和底部结构。

此外，由于传统方法局限于像素级的匹配，而本文方法通过可微的 Soft Argmin 操作实现了亚像素级别的深度估计，在重建结果的精度方面有更好的表现，具有更细的重建粒度。综合以上定性分析结果可知，相

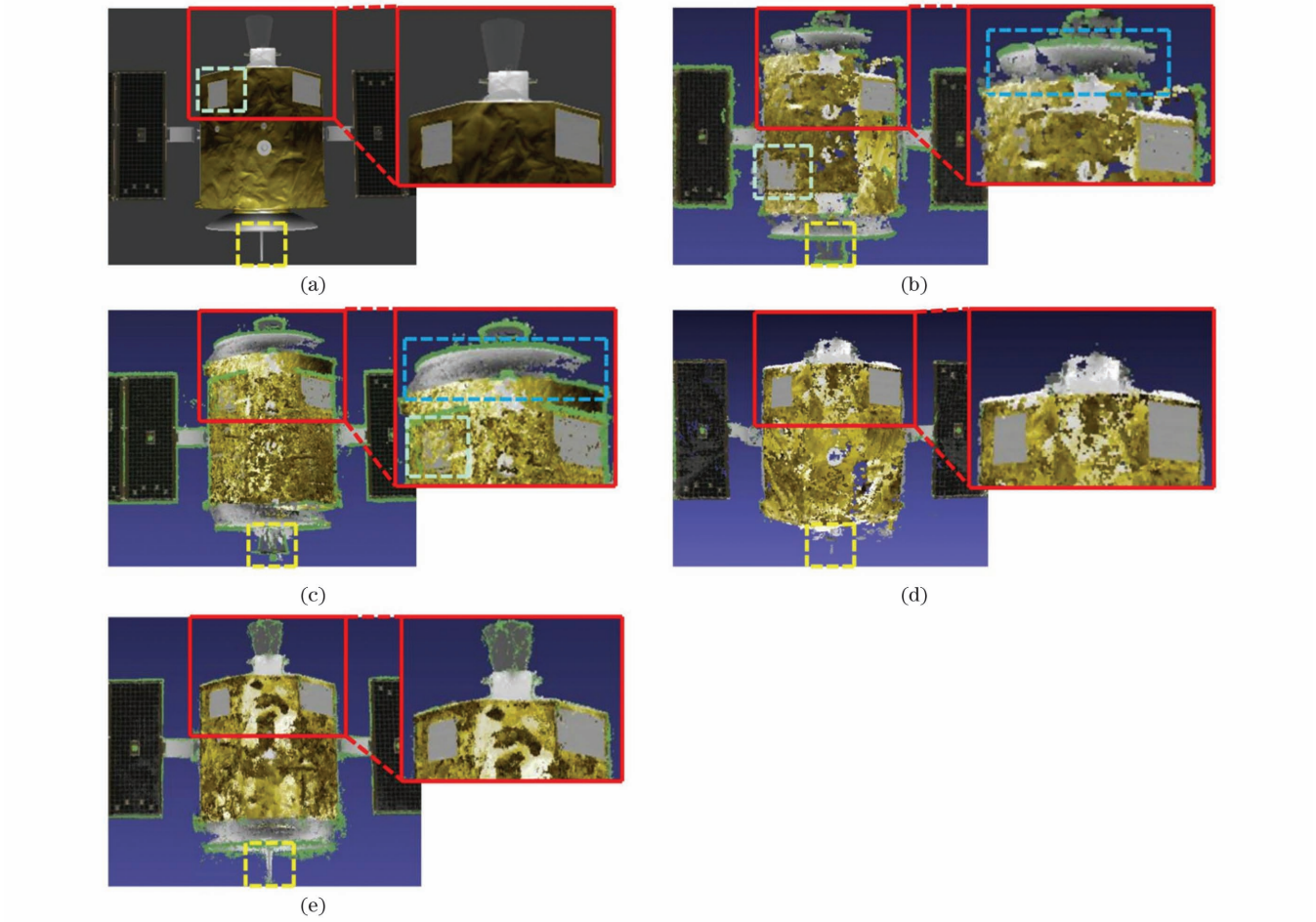

图 8 不同方法得到的 Clementine 卫星重复纹理区域的重建结果。(a)原始图像;(b)VisualSFM;(c)COLMAP;(d)SurfaceNet;(e)本文方法

Fig. 8 Reconstruction results for repetitive texture region of Clementine satellite obtained by different methods. (a)Ground truth;(b)VisualSFM;(c)COLMAP;(d)SurfaceNet;(e) our method

比传统算法,本文算法与真实模型更加接近,尤其是在纹理稀疏、非漫反射和纹理重复等具有挑战的空间目标场景的重建中。

5.2 客观量化评估

本实验使用 DTU 数据集提供的 Matlab 代码,分别计算 SurfaceNet、COLMAP、VisualSFM 和本文方法的重建模型的平均准确度误差、平均完整度误差和误差综合评价。在这三个评价指标里,平均准确度误差指标是指实验重建点云到地面真值点云的距离,体现了算法重建出的三维点云的质量。平均完整度误差指标是指地面真值点云到实验重建点云的距离,体现了算法恢复三维点的能力。理想情况下这两个指标应该同时具有高准确性和高完备性,但它们存在相互制约的关系,因此,为了进行更公平的比较和分析,本文同时考虑这两个评价指标,并进一步计算两个指标的平均值,将平均值作为衡量算法误差综合评价的指标。这三个指标的取值越低,重建质量越高。具体来说,计算一个点云到目标点云的距离时,先遍历计算该点云每一个点到目标点云最近点的距离,并将其作为该点到目标点云的距离。随后,考虑到外点的干扰,剔除距离大于 20 mm 的点。最后,计算所有剩余点与目标点云距离的平均值,将其作为该点云到目标点云的距离。

如表 1 所示,数值越低,代表误差越小,算法性能越好。可知,本文算法虽然在重建的平均准确度误差方面略低于 COLMAP,但在重建的平均完整度误差和综合性能方面都具有很明显的优势。

表 1 DTU 数据集上不同算法性能的对比

Table 1 Performance comparison of different algorithms on DTU dataset

Method	Mean accuracy error /mm	Mean completeness error /mm	Overall error /mm
Our method	0.449	0.379	0.414
SurfaceNet	0.510	0.748	0.629
COLMAP	0.400	0.644	0.522
VisualSFM	0.613	0.835	0.724

5.3 系统运行时间

模型重建的速度是三维重建系统的重要评估指标之一。在 4.2 节所示的实验环境下,测试本文算法、VisualSFM、COLMAP、SurfaceNet 在 DTU 验证集单组数据(每组数据包含 49 张图片,图片分辨率为

1600 pixel×1200 pixel)上的三维重建时间。其中,为了保证重建精度,SurfaceNet 重建时的体素立方体大小设定为 32^3。由表 2 可知,在这些算法中,本文算法和 VisualSFM 的运行速度最快。但由 5.2 节可知,VisualSFM 重建出的模型空洞较多,而本文算法能获得更为稠密精细的重建效果。在传统方法中,COLMAP 的重建效果最好,但其运行速度最慢,重建所需的运行时间是本文算法的 100 倍以上。深度学习方法 SurfaceNet 重建所需的运行时间是本文算法的 160 倍以上。

表 2 运行耗时

Table 2 Running time

Method	Our method	SurfaceNet	COLMAP	VisualSFM
Running time /s	230	36802	2700	223

5.4 匹配视图数量

为了研究匹配视图数量对重建模型精度的影响,本实验分别选用 2、3、4、5 张匹配视图进行实验,并计算所得重建模型的平均准确度误差、平均完整度误差和误差综合评价指标。实验结果如表 3 所示,可知随着匹配视图数量的增多,重建模型的准确度、完整度和综合性能都有所上升。值得注意的是,多视图虽然能提供冗余信息,提升重建精度,但也容易出现遮挡问题,这会严重影响立体匹配效果。因此,匹配的视图并不是越多越好,一般选取最接近参考视图的几幅图进行匹配即可。由表 3 可知,使用三张视图进行重建时,精度提升最为明显,故本实验设定匹配视图数量为 3 进行模型训练。

表 3 不同匹配视图数量下的算法性能对比

Table 3 Performance comparison of algorithms under different numbers of matching views

Number of matching views	Mean accuracy error /mm	Mean completeness error /mm	Overall error /mm
2	0.477	0.431	0.454
3	0.449	0.379	0.414
4	0.432	0.354	0.393
5	0.429	0.345	0.387

6 结 论

提出了一种基于 MVSNet 深度学习网络获知目标三维立体信息的方案。首先提出了通过严格模拟真实场景制作多视图数据集的方法。接着介绍了基于立体匹配的视差理论构建相机视锥视角下代价体的方法。然后研究了 SoftMax 代价体正则化方法以及利用残差网络恢复边缘特征的算法。最后,给出了评估重建模型质量的主客观两种方法,并对比了所提方法、VisualSFM、COLMAP 和 SurfaceNet 的重建效果。实验结果证明:所提方法的平均准确度误差为 0.449 mm,平均完整度误差为 0.379 mm,误差综合评价为 0.414 mm。基于深度学习的方法能更好地恢复空间目标的尺寸大小和材质特征,在今后仍有很好的研究和应用价值。

参 考 文 献

[1] 牟金震, 郝晓龙, 朱文山, 等. 非合作目标智能感知技术研究进展与展望[J]. 中国空间科学技术, 2021, 41(6):1-16.
Mu Y Z, Hao X L, Wei C L, et al. Review and prospect of intelligent perception for non-cooperative targets[J]. Chinese Space Science and Technology, 2021, 41(6): 1-16.

[2] 杨洪飞. 空间目标多源数据三维重建技术研究[D]. 上海: 中国科学院上海技术物理研究所, 2018: 1-2.
Yang H F. Research on multi-sources data 3D reconstruction technology for space targets [D]. Shanghai: Institute of Technical Physics, Chinese Academy of Sciences, 2018: 1-2.

[3] 杨帆, 刘斌, 初录, 等. 基于网格结构光的双目测量方法[J]. 中国激光, 2021, 48(23): 2304004.
Yang F, Liu B, Chu L, et al. Binocular measurement method using grid structured light[J]. Chinese Journal of Lasers, 2021, 48(23): 2304004.

[4] Wu Z C, Wei X X, Song L M, et al. Solution for vision occlusion based on binocular line-structured light [J]. Optoelectronics Letters, 2021, 17(7): 432-437.

[5] 杨帆, 丁晓剑, 曹杰. 基于彩色结构光的自由曲面三维重建方法[J]. 光学学报, 2021, 41(2): 0212001.
Yang F, Ding X J, Cao J. 3D reconstruction of free-form surface based on color structured light[J]. Acta Optica Sinica, 2021, 41(2): 0212001.

[6] 井李强, 郑刚, 孙彬, 等. 基于调频连续波干涉技术的运动目标距离测量[J]. 中国激光, 2019, 46(12): 1204001.
Jing L Q, Zheng G, Sun B, et al. Measurement of distance to moving target using frequency-modulated continuous-wave interference technique[J]. Chinese Journal of Lasers, 2019, 46(12): 1204001.

[7] 吴冠豪, 周思宇, 杨越棠, 等. 双光梳测距及其应用[J]. 中国激光, 2021, 48(15): 1504002.
Wu G H, Zhou S Y, Yang Y T, et al. Dual-comb ranging and its applications[J]. Chinese Journal of Lasers, 2021, 48(15): 1504002.

[8] Vázquez-Arellano M, Reiser D, Paraforos D S, et al. 3-D reconstruction of maize plants using a time-of-flight camera[J]. Computers and Electronics in Agriculture, 2018, 145: 235-247.

[9] Micheletto M, Zubiaga L, Santos R, et al. Development and validation of a LiDAR scanner for 3D evaluation of soil vegetal coverage[J]. Electronics, 2020, 9(1): 109.

[10] Moulon P, Monasse P, Marlet R. Adaptive structure from motion with a Contrario model estimation[M]//Fitzgibbon A, Lazebnik S, Perona P, et al. Computer vision-ACCV 2012. Lecture notes in computer science. Heidelberg: Springer, 2013, 7727: 257-270.

[11] Bao S Y, Bagra M, Chao Y W, et al. Semantic structure from motion with points, regions, and objects [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 2703-2710.

[12] Wu C C. Towards linear-time incremental structure from motion [C]//2013 International Conference on 3D Vision-3DV 2013, June 29-July 1, 2013, Seattle, WA, USA. New York: IEEE Press, 2013: 127-134.

[13] Schönberger J L, Frahm J M. Structure-from-motion revisited [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4104-4113.

[14] 叶春凯, 万旺根. 基于特征金字塔网络的多视图深度估计[J].

电子测量技术, 2020, 43(11): 91-95.
Ye C K, Wan W G. Feature pyramid network for multi-view depth estimation [J]. Electronic Measurement Technology, 2020, 43(11): 91-95.
[15] Zhu Q T. Deep learning for multi-view stereo via plane sweep: a survey[EB/OL]. (2021-06-18)[2021-04-06]. https://arxiv.org/abs/2106.15328v2.
[16] Wang X, Wang C, Liu B, et al. Multi-view stereo in the deep learning era: a comprehensive review[J]. Displays, 2021, 70: 102102.
[17] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[EB/OL]. (2014-06-09)[2021-05-04]. https://arxiv.org/abs/1406.2283.
[18] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C] // 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2650-2658.
[19] Choy C B, Xu D F, Gwak J, et al. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 628-644.
[20] 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, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2326-2334.
[21] Kar A, Häne C, Malik J. Learning a multi-view stereo machine [EB/OL]. (2017-08-17)[2021-04-05]. https://arxiv.org/abs/1708.05375.
[22] Yao Y, Luo Z X, Li S W, et al. MVSNet: depth inference for unstructured multi-view stereo [M] // Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11212: 785-801.
[23] 李睿. 空间目标图像三维重建关键技术研究[D]. 西安: 西北工业大学, 2019.
Li R. Research on the key technologies of space debris 3D reconstruction [D]. Xi'an: Northwestern Polytechnical University, 2019.
[24] Collins R T. A space-sweep approach to true multi-image matching [C] // Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 18-20, 1996, San Francisco, CA, USA. New York: IEEE Press, 1996: 358-363.
[25] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[EB/OL]. (2015-05-18)[2021-04-05]. https://arxiv.org/abs/1505.04597.
[26] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression[C]// 2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 66-75.
[27] Xu N, Price B, Cohen S, et al. Deep image matting[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 311-320.
[28] Jensen R, Dahl A, Vogiatzis G, et al. Large scale multi-view stereopsis evaluation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 406-413.

Application of MVSNet in 3D Reconstruction of Space Objects

Wang Siqi[1,3], Zhang Jiaqiang[1,3], Li Liyuan[1,3*], Li Xiaoyan[1,2], Chen Fansheng[1,2]

[1] *Key Laboratory of Intelligent Infrared Perception, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai* 200083, *China*;

[2] *Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou* 310024, *Zhejiang, China*;

[3] *University of Chinese Academy of Sciences, Beijing* 100049, *China*

Abstract

Objective 3D reconstruction of space targets can provide prior structural information for space services, which is a key technology for improving system autonomy. Conventional 3D reconstruction methods rely on handcrafted features to recover the 3D structure of objects by dense matching. Therefore, affected by the symmetrical structure and non-Lambert imaging of spatial targets, conventional 3D reconstruction methods often suffer from mismatching and insufficient matches of feature points, resulting in a low reconstruction accuracy. In recent years, with continuous developments in deep learning technology, convolution neural networks (CNNs) have been widely used in computer vision. Compared with the handcrafted features used by conventional 3D reconstruction methods, the deep features extracted by CNNs can introduce high-level semantics of images for more robust matching. Inspired by this, a 3D reconstruction method based on MVSNet for space targets is proposed. This algorithm organically applies CNNs with different structures to improve the accuracy and completeness of 3D reconstruction. We hope that our basic strategy and findings will be beneficial to the 3D reconstruction of space targets.

Methods The space-target 3D-reconstruction algorithm model (Fig. 1) is described as follows. First, in view of the imaging characteristics of a space target, the influence of the geometric structure and material of the model on reconstruction results is analyzed, and a multi-view acquisition system for space targets based on the Blender platform is designed. Subsequently, deep visual image features are fully extracted via multi-scale convolution based on MVSNet. The coder and decoder are then used to gather and regularize the spatial context information for stereo matching, which effectively avoids the heavy dependence of conventional methods on the feature points in the reconstructions of low-

textured, reflective, and repetitive texture regions. Finally, the residual network is used to solve the boundary smoothing problem caused by the multiple convolutions to further improve the reconstruction results. The model is tested on both the DTU dataset and our self-collected space-target dataset. Its performance is compared with those of VisualSFM, COLMAP, and SurfaceNet through both qualitative and quantitative evaluations. The running time of the proposed algorithm and other methods is also measured to verify the efficiency improvement. In addition, the influence of different numbers of matching views on the accuracy of the reconstructed model is studied to discuss the most appropriate settings.

Results and Discussions The proposed method outperforms conventional 3D reconstruction methods in handling low-textured, specular, and reflective regions, which can completely restore typical structures, such as satellite cabins and roofs (Figs. 6 and 7). The mean accuracy, mean completeness, and overall errors of the proposed algorithm are 0.449, 0.379, and 0.414 mm, respectively. The proposed algorithm has the best accuracies among all four compared algorithms (Table 1). In particular, the accuracy of the proposed method is 20% higher than that of the advanced open source software COLMAP. The running time study shows that our method is faster with an average time cost of approximately 230 s for reconstructing one scan (Table 2). The running speed of the proposed method is 100 times faster than that of COLMAP and 160 times faster than that of SurfaceNet. In addition, the performance study on different numbers of matching views shows that more views result in a better performance (Table 3). However, the accuracy improvement is the greatest when three matching views are used for 3D reconstruction. Thus, the matching view number is set as 3 for model training. In general, the proposed model is optimal in terms of reconstruction accuracy and speed.

Conclusions This study proposes a method for the 3D reconstruction of spatial targets based on the MVSNet deep learning network. First, deep visual image features are fully extracted by the multi-scale 2D convolution, and then spatial context information for stereo matching is fully gathered and regularized through the skip connection between the coding and decoding paths. Subsequently, the matching cost is converted into the depth value probability using the SoftMax function, and the expectation is calculated as the initial estimation value. Finally, the final depth estimation map is obtained by strengthening the edge semantic information through the residual network. Experimental results show that the accuracy of the proposed method is 20% higher than that of the advanced open source software COLMAP. Moreover, the running speed is 100 times faster than that of COLMAP and 160 times faster than that of SurfaceNet. In general, this model can effectively introduce the high-level semantics of images for more robust matching and has the low system running time, which can provide a technical reference for space operation automations and further promote the application of 3D reconstruction in this field.

Key words remote sensing; deep learning; multiple views; 3D reconstruction of space targets; convolutional neural networks; encoder-decoder architecture