计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009710] [CSTR: 32024.14.csa.009710]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多视图立体深度学习的堆叠工件三维重建①

姬田杰[1,3], 郑飂默[2], 曹克让[1,3], 王诗宇[2], 周淞杰[2]

[1](沈阳化工大学 计算机科学与技术学院, 沈阳 110142)
[2](中国科学院 沈阳计算机技术研究所, 沈阳 110168)
[3](辽宁省化工过程工业智能化技术重点实验室, 沈阳 110142)
通信作者: 郑飂默, E-mail: zhengliaomo@sict.ac.cn

摘　要: 随着工业自动化的不断发展, 工件的三维重建技术在制造业中扮演着越来越重要的角色. 在实际的工作环境下, 工件普遍存在堆叠问题, 对后续的机器人识别抓取等工作存在较大影响. 目前三维重建技术对于一些具有弱纹理区域的工件重建, 仍存在图像特征点提取难度大、特征配准精度低的问题. 针对以上问题, 本文提出了一种基于多视图立体匹配深度学习的堆叠工件三维重建方法. 首先, 输入多张不同视角的图像经过融合 DCNv2 的特征金字塔网络, 进行特征提取; 然后, 进行单应性变换构建代价体, 再使用方差聚合为一个统一的代价体; 接着在代价体正则化部分, 引入 SE 通道注意力机制模块来提高网络的特征表达能力, 增强模型的性能和泛化能力; 此方法在 DTU (Danish Technical University) 数据集上具有较好的表现, 并且运用该方法生成的堆叠工件点云模型对以后的工业自动化开展具有重要意义.

关键词: 多视图三维重建; DCNv2; 级联架构; 通道注意力机制

引用格式: 姬田杰,郑飂默,曹克让,王诗宇,周淞杰.基于多视图立体深度学习的堆叠工件三维重建.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9710.html

Stacked Workpiece 3D Reconstruction Based on Multi-view Stereo Deep Learning

JI Tian-Jie[1,3], ZHENG Liao-Mo[2], CAO Ke-Rang[1,3], WANG Shi-Yu[2], ZHOU Song-Jie[2]

[1](College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Liaoning Provincial Key Laboratory of Intelligent Technology for Chemical Process Industry, Shenyang 110142, China)

Abstract: With the continuous development of industrial automation, the three-dimensional reconstruction technology of workpieces is playing an increasingly important role in the manufacturing industry. In actual working environments, there is a common problem of stacking workpieces, which significantly impacts subsequent work including robot recognition and grasping. Currently, it is hard for 3D reconstruction to extract image feature points and achieve accurate feature registration in workpieces with weak textures. To address the above issues, this study proposes a 3D reconstruction method for stacked workpieces based on deep learning with multi-view stereo matching. Firstly, multiple images from different perspectives are input through a DCNv2-based feature pyramid network for feature extraction. Then, homography transformation is performed to construct cost volumes, and a unified cost volume is obtained through variance aggregation. In the regularization section of the cost volume, an SE channel attention module is introduced to improve the feature expression ability of the network and enhance the performance and generalization ability of the model. This method exhibits good performance on the Danish Technical University (DTU) dataset. The point cloud model of stacked workpieces generated by this method is of great significance for future applications of industrial automation.

① 基金项目: “兴辽英才计划”产业高端人才项目 (XLYC2207077); “兴沈英才计划”高水平技术创新人才项目; 沈阳市中青年科技创新人才支持计划 (RC210475)
收稿时间: 2024-04-18; 修改时间: 2024-05-20, 2024-07-03; 采用时间: 2024-07-11; csa 在线出版时间: 2024-10-25

Key words: multi-view 3D reconstruction; DCNv2; cascade architecture; channel attention mechanism

1 引言

近年来, 随着工业生产自动化的不断发展, 螺钉、齿轮、轴承等常见工件是工业生产中的重要组成部分, 为整个工业加工流程有着重要作用. 在实际工业生产过程中经常面临工件摆放堆叠的情况, 与理想情况下单个工件出现在工作场景存在显著差异. 工件堆叠引发了繁杂的视角和相互遮挡关系, 给生产制造过程带来了挑战, 常见的堆叠问题对工件识别、定位和跟踪等工作造成极大影响, 使工作难以顺利进行. 与传统的单一工件重建不同, 在现实场景中需要处理的是多个工件之间的相互遮挡、视角变化等复杂情况. 因此需要通过利用多视角图像和相应的相机位姿信息, 以解决工件堆叠场景下的重建难题. 多视图立体 (multi-view stereo, MVS) 重建[1]旨在以立体匹配为主要线索, 从一系列图像、相应的相机位姿和已知参数中重建出场景三维模型. 本研究旨在实现对工件堆叠场景的高质量三维重建, 为实际生产制造提供准确可靠的场景还原和模型重建. 这将有助于优化生产过程中的工件识别、定位和跟踪等关键环节, 提高生产效率和质量水平, 为工业制造领域的发展带来新的技术突破和应用前景.

在传统 MVS 方法中, Snavaly 等人[2]设计了首个增量式的运动恢复结构 (structure from motion, SFM) 系统, 估计相机位姿并获得目标的三维稀疏点云, 然后利用 MVS 生成目标的三维稠密点云模型, 后续的论文研究都遵循了 Snavaly 等人[2]设计的流程. 2016 年, Schonberger 等人[3]提出了 COLMAP, 在原有增量式 SFM 的基础上引入了一种几何验证策略, 用信息增强场景图, 从而提高初始化和三角测量组件的鲁棒性, 并进行一系列优化, 使系统在鲁棒性和完整性方面明显优于当时现有技术的同时保持其效率. 随着深度学习的出色表现, 许多学者开始将深度学习运用到 MVS 算法中, 2018 年, Yao 等人[4]提出了一种端到端的深度学习架构 MVSNet, 用于从多视图图像中推断深度图. 该网络首先提取深度视觉图像特征, 然后通过可微单应性变换构建 3D 成本体积. 然后应用 3D 卷积来进行代价体正则化和回归初始深度图, 最后用参考图像对其进行细化以生成最终输出. 2019 年, Yao 等人[5]又在原来的基础上对正则化、深度推断、后处理 3 个部分进行优化, 提出了 R-MVSNet, 在提高性能的同时减少了内存消耗, 但相应的训练时间大大增加. 2020 年, Gu 等人[6]首次提出了一种级联网络结构 CasMVSNet, 通过一个较小的代价体估计低分辨率的深度图, 然后根据上一级输出的深度图, 缩减当前尺度的深度假设范围. 2020 年, Yang 等人[7]提出的 CVP-MVSNet, 以粗略到精细的方式构建成本体积金字塔, 而不是以固定分辨率构建成本体积, 从而实现紧凑、轻量级的网络, 并推断高分辨率深度图, 获得了更好的重建结果. 同年, Yu 等人[8]提出了 Fast-MVSNet, 利用数学高斯牛顿迭代法来优化, 其中的模块都是轻量级的, 保证了工作效率, 但是重建数据一般. 2021 年, Yi 等人[9]提出 PVA-MVSNet, 通过引入两种新的自适应视图聚合: 逐像素视图聚合和逐体素视图聚合, 以较小的额外内存消耗合并了不同视图中的成本方差, 在重建方面有了显著的改进. 同年, Wang 等人[10]首次在端到端可训练架构中引入迭代多尺度 Patchmatch, 并在每次迭代中使用一种新的、经过学习的自适应传播和评估方案来改进 Patchmatch 核心算法, 在内存和速度上有很大的提升, 但重建效率有所下降. 2022 年, Ding 等人[11]提出了 TransMVSNet 网络结构, 是首次将 Transformer 运用到 MVS 任务中, 在特征匹配部分取得了不错的效果, 但整体性能一般, 受设备性能影响较大. 2023 年, Chang 等人[12]提出 RC-MVSNet, 通过施加深度渲染一致性损失, 以约束靠近对象表面的几何特征, 从而减轻遮挡影响, 比许多有监督的方法具有竞争力, 在训练周期上有着较好的性能.

目前多视图立体三维重建技术对于具有弱纹理区域[13–15]的工件重建工作中, 多个工件堆叠带来了复杂的视角和遮挡关系, 存在图像特征点提取难度大、特征配准精度低的诸多情况. 本文为实现对特定区域中重建稀疏的表面保留更多的结构信息和纹理细节, 同时丰富堆叠工件细节以及提高重建精度, 提出一种多视图立体重建算法, 在级联架构的基础上进行优化并完成 MVS 任务, 使用 DTU 数据集进行训练和测试, 并自己采集堆叠工件图像做出数据集用来实现工件的三维重建. 本文的主要工作如下: (1) 在特征提取部分, 在金字塔网络 FPN[16]后添加可变形卷积 DCNv2 模块[17]来增大卷积的范围提升模型的表现力, 控制好偏移的范围, 避免提取

不相关的特征信息; (2) 在代价体正则化部分, 引入 SE (squeeze-and-excitation) 通道注意力机制模块[18]来提高网络的特征表达能力, 增强模型的性能和泛化能力; (3) 最后根据实际项目工程场景, 采集并制作了一个堆叠工件的数据集, 包括各角度的图像以及每张图片对应的相机位姿信息, 并用该数据集较好地实现了堆叠工件场景的三维点云恢复, 验证了本文算法的实用性.

2 方法

本文设计的多视图立体匹配深度学习网络架构依照 CasMVSNet 采用级联结构, 利用多张堆叠工件图像和对应的相机位姿信息对实际场景进行三维模型重建. 将融合可变卷积 DCNv2 的金字塔特征提取模块和融合 SE 通道注意力机制模块整合到基准级联结构网络中, 以增强特征表达和注意力集中能力. 并按照标准的特征提取、可微单应变换、代价体正则化、深度图回归 4 大步骤完成场景的三维重建工作, 以确保重建结果的准确性和稳定性. 整体而言, 本文所提出的网络架构在图 1 中得以清晰展示, 为实现高效三维重建工作奠定了坚实基础.

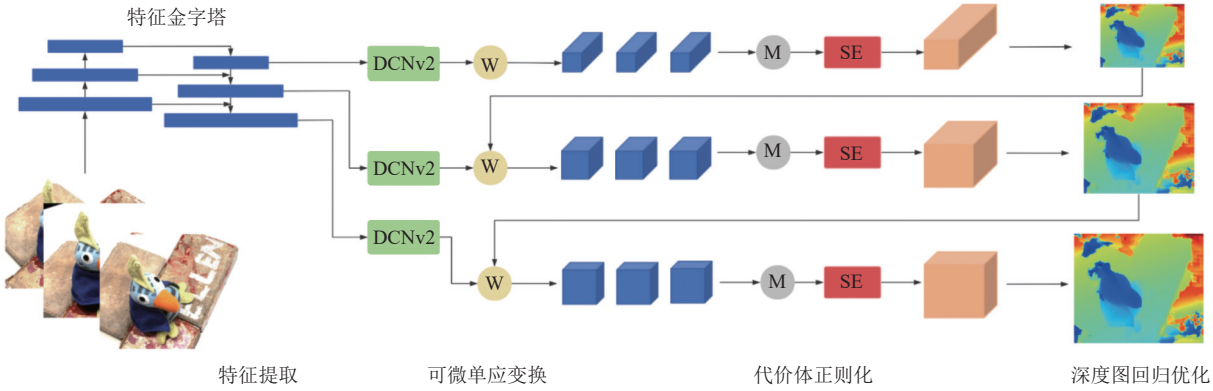

图 1 本文网络架构

2.1 基于可变卷积 DCNv2 的金字塔特征提取网络

在多视角图像匹配的复杂过程中, 为了全面把握目标信息, 通过构建多尺度特征金字塔, 以涵盖不同尺度下的细节和特征. 然而, 在这一过程中, 传统的卷积操作却显得力不从心, 因为传统卷积核所采用的固定大小卷积核和感受野在捕捉多样的尺度和形态特征上表现欠佳. 这种局限性会引起网络对特征处理的不足, 因为卷积核和感受野的设定可能无法灵活适配不同场景的大小和形状变化, 从而导致部分关键特征信息的遗漏.

针对以上问题, 本文引入了可变卷积 DCNv2, 可变卷积允许网络在学习过程中动态地调整卷积核的形状和位置, 从而扩大了卷积操作的感受野, 集中于感兴趣的区域或目标, 以更有效地捕捉不同纹理区域的特征. 传统卷积操作和可变卷积操作采样的方式大不相同, 如图 2 所示.

图 2 中, (a) 代表了标准卷积的采样方式, 用圆点表示. (b) 则展示了可变形卷积中, 通过增强偏移的变形采样位置, 用圆点和箭头来呈现. (c)和 (d) 是对 (b) 的特殊情况的描述, 说明可变形卷积技术能够适应各种尺度、纵横比和旋转的变换. 对于输出特征图 y 上的每个位置 p_0 计算由式 (1) 所示:

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n) \tag{1}$$

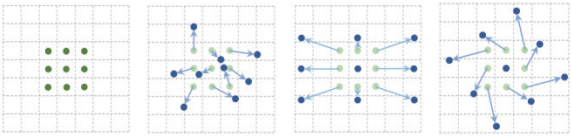

图 2 标准和可变卷积中 3×3 采样位置

在可变卷积中, 规则网格 R 通过偏移 $\{\Delta p_n|n=1,\cdots,N\}$ 进行扩充, 其中 $N=|R|$. 式 (1) 变为:

$$y(p_0)=\sum_{p_n\in R} w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{2}$$

采样是在不规则和偏移位置 $p_n+\Delta p_n$ 上进行的. 由于偏移量 Δp_n 通常是分数的, 因此通过双线性插值 Δp_n 将式 (2) 实现为:

$$x(p)=\sum_{q} G(q,p)\cdot x(q) \tag{3}$$

可变卷积 DCNv2 在此基础上为了进一步增强可变形卷积神经网络对空间支持区域的控制能力, 引入了一种调制机制. 调制机制为网络模块提供了另一个自由维度来调整其空间支持区域. $x(p)$ 和 $y(p)$ 分别表示来自输入特征图 x 和输出特征图 y 的位置处 p 的特征. 调制的可变卷积可以表示为:

$$y(p)=\sum_{k=1}^{K} w_k \cdot x(p+p_k+\Delta p_k)\cdot \Delta m_k \tag{4}$$

其中, Δp_k 和 Δm_k 分别是第 k 个位置的可学习偏移量和调制标量, 调制标量位于范围 [0, 1]内.

2.2 基于 SE 通道注意力机制的代价体正则化网络

在构建三维代价体及后续重建部分中, 代价体积是指从多个视角获取的图像中重建的三维场景中的每个点的深度候选集合. 由于非朗伯面、弱纹理及遮挡等因素会包含噪声信息, 需要对其进行正则化. 而代价体积正则化的目标是优化代价体积, 以使其对应于真实深度值. 然而普遍网络无法自适应地调整通道间的权重, 且每个分支产生的特征图重要性存在差异, 从而出现特征选择不当的问题. 在此过程中, 本文引入 SE (squeeze-and-excitation) 通道注意力机制模块, SE 模块是一种轻量级的门控机制, 通过学习特征通道之间的关系, 动态地调整每个通道的权重, 提取和强调最重要的特征信息, 并抑制不太有用的特征信息. 从而使得网络能够更好地捕获关键信息, 提高网络对输入数据的表达能力. 其主要流程由图 3 所示.

首先是压缩操作, 对输入的图像特征 u 进行全局平均池化, 得到当前特征的全局压缩特征量. 通过缩小空间维度 $H\times W$ 来生成数据, 其计算公式如式 (5) 所示:

$$z_c=F_{\mathrm{sq}}(u_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{5}$$

然后进行激励操作, 用于完全捕获通道依赖关系, 使其能够学习通道之间的非线性相互作用, 并确保多个通道被强调. 将经过上一步压缩得到的 z 进行处理, 获得最终的激励权重, 其计算公式如式 (6) 所示. 其中 δ 表示 ReLU 函数.

$$s=F_{\mathrm{ex}}(z,W)=\sigma(g(z,W))=\sigma(W_2\delta(W_1 z)) \tag{6}$$

接着, 将上一步生成的激励权重 s 对特征图 u 进行权重赋值, 获得最终的特征图, 使其尺寸大小与输入的原特征图完全一样, SE 通道注意力机制模块不改变特征图的大小. 其最终输出使通过式 (7) 激活重新缩放变换输出获得的.

$$\tilde{X}_c=F_{\mathrm{scale}}(u_c,s_c)=s_c\cdot u_c \tag{7}$$

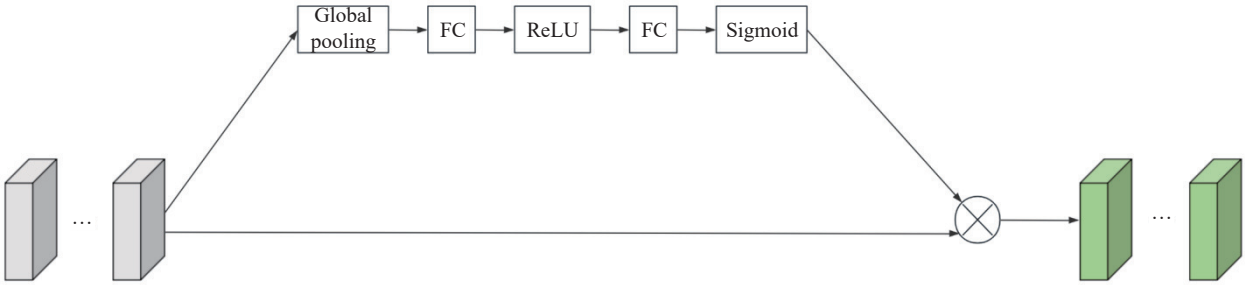

图 3 SE 模块

3 实验

3.1 实验环境

本文实验所使用计算平台的主要参数: 32 GB 内存、NVIDIA GeForce RTX 3080 显卡. 运行系统环境: Windows 10, 编程语言: Python, 深度学习框架: PyTorch.

3.2 DTU 数据集实验分析

3.2.1 DTU 数据集

DTU 数据集[19]是一个涵盖多种对象的室内公开数据集, 具有广泛的应用价值, 尤其在计算机视觉和三维重建领域. 它涵盖了多种室内对象, 为研究者们提供了丰富的实验数据. 这个数据集的采集过程十分精细, 采用了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. 这种拍摄方式能够捕捉到物体的各个细节, 从而生成高质量的三维模型.

数据集中一共 124 个场景, 每个场景都取 49 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, 可以更好地应对

光照变化对三维重建的影响, 提高算法的鲁棒性和准确性. 数据集中包括每张 RGB 图像的摄像头参数、拍摄图像、真实深度图和掩膜.

3.2.2 实验细节

本文将所提出的方法与 CasMVSNet 及其他方法在 DTU 数据集上进行实验对比, 在训练过程中, 将输入图片数量设为 N=3, 其中包括 1 张主要参考图像和 2 张其邻近图像, 将图像分辨率设置为 640×512. 采用三级级联成本体积, 从第 1 阶段到第 3 阶段, 每个阶段的平面扫描深度假设的数量分别为 48、32、8, 对应的深度间隔被设置为 4、2、1. 使用 Adam 优化器训练 16 个 epoch. 初始学习率为 0.001, β_1=0.9, β_2=0.999. 在第 10、12、14 个 epoch 后将其缩小 2 倍, batchsize 大小设为 1. 在测试过程中, 输入图片数量设为 N=5, 其中包括 1 张主要参考图像和 4 张其邻近图像, 将图像分辨率设置为 1600×1200, 深度采样次数 D=192.

3.2.3 实验结果

实验使用 DTU 数据集提供的官方 Matlab 脚本评估点云的准确性 *Acc*、完整性 *Comp* 以及整体性能 *Overall* 这 3 个指标[20,21]. *Acc* 衡量的是重建点云到真实值的距离, *Comp* 衡量的是真实值到重建点云的距离, 需要计算 precision 值和 recall 值, 其中的指标越低越好, 表示重建值 R 与真值 G 之间的距离越小, 恢复的点云越准确. 首先定义 R 中的一个点 r 到 G 的距离, r 到 G 中所有点的距离中, 最小的距离就是 r 到整个 G 的距离, 计算公式如式 (8) 所示:

$$e_r \to G = \min_{g \in G} \|r - g\| \tag{8}$$

并在式 (8) 基础上定义一个阈值 d, 统计所有距离小于 d 的点的个数, 然后通过式 (9) 得到最终百分比结果.

$$P(d) = \frac{100}{|R|} \sum_{r \in R} [e_r \to G < d] \tag{9}$$

Recall 的计算与前者相似, 用于衡量 G 到 R 的距离. 计算公式如式 (10)、式 (11) 所示:

$$e_g \to R = \min_{r \in R} \|g - r\| \tag{10}$$

$$R(d) = \frac{100}{|G|} \sum_{g \in G} [e_g \to R < d] \tag{11}$$

在计算出以上的指标后, 就可以计算出衡量 precision 和 recall 的整体指标 F-score, 具体计算公式如式 (12) 所示:

$$F(d) = \frac{2P(d)R(d)}{P(d) + R(d)} \tag{12}$$

其中, *Overall* 是 3 个指标中最重要的, 是准确性 *Acc* 和完整性 *Comp* 的总和平均值, 代表着重建整体性误差. 本文评价指标 *Acc*、*Comp* 以及整体性能 *Overall* 的关系如式 (13) 所示:

$$Overall = \frac{Acc + Comp}{2} \tag{13}$$

由表 1 中数据可得, 本文所提出的方法完整性 *Comp* 和整体性能 *Overall* 指数都是最低的, 虽然准确性 *Acc* 没有达到最低, 但是整体上来说本方法在与其他方法对比三维重建中性能最好, 证明了所提方法的有效性. 完整性 *Comp* 和整体性 *Overall* 的提高主要得益于可变卷积 DCNv2 模块和 SE 通道注意力机制模块, 其中可变卷积 DCNv2 模块通过灵活调整卷积核的形状和大小, 有效地捕获了不同尺度下的特征, 从而提升了模型对于图像中物体边缘和细节的感知能力. 与此同时, SE 通道注意力机制模块能够动态地调整不同通道的重要性权重, 有针对性地增强对于关键特征的抽取, 从而使模型更加聚焦于关键信息, 进一步提高了整体的性能表现.

表 1 DTU 数据集重建指标结果 (mm)

方法	准确性*Acc*	完整性*Comp*	整体性*Overall*
MVSNet	0.6542	0.4831	0.5686
Fast-MVSNet	0.4165	0.4706	0.4435
RC-MVSNet	0.4675	0.3629	0.4152
PVA-MVSNet	**0.3721**	0.4182	0.3952
CasMVSNet	0.3856	0.3611	0.3733
TransMVSNet	0.3797	0.3599	0.3698
Ours	0.3747	**0.3575**	**0.3661**

具体在点云数据生成方面, 该方法在细节处理过程中对比其他方法也有较好的表现, 选用 scan34 和 scan62 两个场景进行对比, 点云效果对比如图 4 所示. 从方框标注的部分, 可清晰地看出, 本文算法点云信息更加丰富, 尤其是在 scan34 对比中可以看到, CasMVSNet 砖缝中间还存在着大面积的空白, 而本文算法在物体边缘和细节的感知能力较强, 已经基本实现了填补空洞的目标.

3.2.4 消融实验

本节进行了一系列的消融实验, 以验证本文提出

算法中关键模块的有效性, 该部分采用与第 3.2 节相同的参数在 DTU 测试集上进行以下 4 组消融研究: (a) Baseline, 即采用基准网络, 不添加任何额外的模块; (b) Baseline+DCNv2 模块; (c) Baseline+SE 模块; (d) Baseline+DCNv2 模块+SE 模块; 文中算法通过比较展现出了优秀的性能水平. 各个模块的性能结果如表 2 所示.

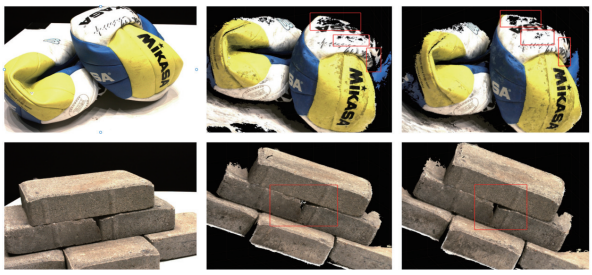

图 4 实验的对比

表 2 定量消融实验分析对比 (mm)

组别	方法	准确性*Acc*	完整性*Comp*	整体性*Overall*
(a)	Baseline	0.3856	0.3611	0.3733
(b)	Baseline+DCNv2	0.3445	0.3955	0.3700
(c)	Baseline+SE	0.3693	0.3650	0.3671
(d)	Baseline+DCNv2+SE	0.3747	0.3575	**0.3661**

通过深入对比表 2 中的定量结果, 观察了它们对模型性能的影响. 首先, 单独引入 DCNv2 后, 准确性指标显著下降, 表明 DCNv2 模块提升了网络对复杂几何形状和细节的捕捉能力, 从而提高了模型的准确性. 然而, 完整性指标上升, 这是由于 DCNv2 的卷积核可以动态变形, 在某些区域的特征提取过程中产生了更多的噪声或不连续性, 导致完整性下降. 其次, 单独引入 SE 模块后, 准确性指标同样下降, 这表明 SE 模块通过自适应调整通道间的权重, 增强了重要特征的表示能力, 从而提高了模型的准确性. 然而, 完整性指标上升, 这是因为 SE 模块在增强重要特征时忽略了一些边缘或细节特征, 导致完整性下降.

当同时引入 DCNv2 和 SE 模块时, 准确性和完整性指标均下降, 这表明两者的结合能够在捕捉复杂几何形状和增强特征表示上互补, 进一步降低了准确性误差. SE 模块在全局上自适应调整通道间的权重, 能够平衡 DCNv2 引入的噪声和不连续性, 从而改善了完整性. 这使得整体性指标显著下降, 说明这种结合方式在处理复杂几何形状和增强特征表示方面具有显著优势, 显著提升了模型的整体性能.

综上所述, 单独引入 DCNv2 或 SE 模块分别提升了模型的某些方面, 但同时也引入了一些新的问题. 而同时引入 DCNv2 和 SE 模块能够互相补充, 平衡特征提取过程中出现的问题, 从而显著提升模型的整体性能. 我们可以清晰地看到每个模块在提升点云重建整体性能方面所起到的积极作用. 这些模块不仅各自独具特色, 而且相互协作, 共同构成了本文所提算法的核心架构.

3.3 自建数据集实验分析

3.3.1 堆叠工件图像数据集

为了验证本文算法在重建堆叠工件的实用性、对待纹理稀疏、纹理重复和细节处理区域的性能, 自行采集并制作了堆叠工件图像数据集用于测试. 该数据集是从多角度环绕拍摄一组实际工厂项目中出现的实际堆叠工件场景, 仿照 DTU 数据集的部分场景需求, 在堆叠工件这一场景共拍摄了 49 张场景图片. 为了确保目标场景立体匹配信息的完整性和目标模型重建的质量, 原始图片的分辨率为 3024×4032, 以尽可能捕获场景的细节和特征. 在数据采集后, 利用 COLMAP 对每个图片进行标定, 得到了对应的一套相机参数以及视觉选定. 具体的场景图片如图 5 所示.

图 5 COLMAP 中每个图片的机位

这一步骤是为了准确地还原出每张图片所对应的拍摄视角和相机参数, 为后续的立体匹配和三维重建提供了基础. 随后, 使用了 MVSNet 架构中提供的 colmap2mvsnet 脚本工具, 对 COLMAP 的输出文件进行处理, 将其转化为本文架构可输入的数据集格式. 这

个过程包括了将视觉几何信息转换为深度信息, 并进行了适当的格式转换和数据处理, 以满足本文算法的输入要求. 通过这一步骤, 成功将原始的图片数据转换为了适合于本文算法的输入数据集, 最终数据集包括, 49 张去畸变后的图片、49 个对应的相机参数文件以及一个用于定义视觉图像对及其相对关系的配置文件. 该数据集提供了丰富的场景信息和几何结构, 为本文算法提供了充足的数据基础, 从而确保了重建模型的质量和准确性.

3.3.2 实验结果

本文的训练参数设置与上文相同, 针对自建的堆叠工件图像数据集进行三维重建. 该实验比较了不同算法在堆叠工件三维重建场景中的表现, 如图 6 所示.

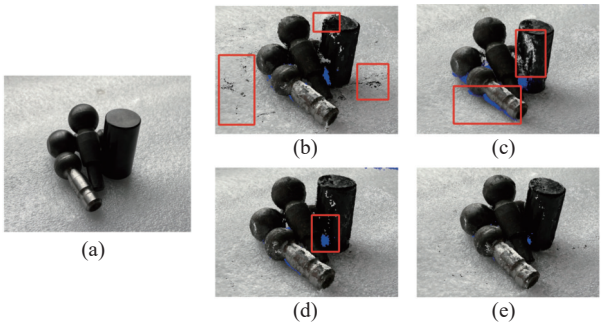

图 6 不同方法得到的堆叠工件场景的三维重建结果

图 6 中, 标准场景 (图 6(a)): 标准堆叠工件场景图像. RC-MVSNET (图 6(b)): 在重建过程中产生了较多噪点和错误的细节区域. CasMVSNet (图 6(c)): 重建质量有所提升, 但仍有一些缺失的重建点云细节. Trans-MVSNet (图 6(d)): 表现良好, 但仍存在少量重建点云细节错误.

实验结果表明, 本文的算法 (图 6(e)) 能够在保持重建细节的同时, 较为精确地重建出工件的三维模型, 重建的细节更精确, 噪点显著减少, 通过扩大卷积范围和增强特征表达能力, 整体重建效果更好. 值得注意的是, 在对堆叠工件进行重建时, 由于场景本身存在大量的重复纹理, 工件堆叠后, 颜色、形状等局部特征非常相似. 然而, 本文算法在重建过程中能够准确地区分不同工件之间的差异, 实现重建目标.

总体而言, 我们的改进算法通过更有效地捕捉和表达特征信息, 在完整性和细节上明显优于其他算法, 重建出来的效果与真实场景高度相似, 展示了更高的重建质量和模型性能, 为项目的后续进行, 奠定了基础.

4 总结

本文提出了一种基于多视图立体深度学习的堆叠工件三维重建技术, 旨在解决工业自动化中常见的堆叠问题对工件识别、定位和跟踪等工作的影响. 针对当前现有方法在处理具有弱纹理区域的工件重建时存在的问题, 本文提出了一种优化的多视图立体重建算法. 该算法通过优化级联结构、引入可变形卷积 DCNv2 和 SE 通道注意力机制等关键技术, 成功地完成了对堆叠工件的三维模型重建. 在实验方面, 首先在 DTU 数据集上进行了训练和测试, 并验证了算法的有效性. 结果显示, 所提出的方法在完整性以及整体性能方面均表现出色, 优于基准方法. 进一步地, 在自建堆叠工件图像数据集上也取得了良好的重建效果. 这些结果验证了所提出方法的可行性和有效性, 为工业自动化领域提供了一种更加高效、精确的数字化解决方案. 未来, 我们将继续改进和优化算法, 进一步推动堆叠工件三维重建技术与自动化技术的融合, 以满足智能自动化中对于精准、高效生产的需求.

参考文献

1 鄢化彪, 徐方奇, 黄绿娥, 等. 基于深度学习的多视图立体重建方法综述. 光学 精密工程, 2023, 31(16): 2444–2464.

2 Snavely N, Seitz SM, Szeliski R. Photo tourism: Exploring photo collections in 3D. ACM SIGGRAPH 2006 Papers. Boston: ACM, 2006. 835–846. [doi: 10.1145/1179352.1141964]

3 Schönberger JL, Frahm JM. Structure-from-motion revisited. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4104–4113. [doi: 10.1109/CVPR.2016.445]

4 Yao Y, Luo ZX, Li SW, *et al*. MVSNet: Depth inference for unstructured multi-view stereo. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 785–801.

5 Yao Y, Luo ZX, Li SW, *et al*. Recurrent MVSNet for high-resolution multi-view stereo depth inference. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5520–5529. [doi: 10.1109/CVPR.2019.00567]

6 Gu XD, Fan ZW, Dai ZY, *et al*. Cascade cost volume for high-resolution multi-view stereo and stereo matching.

Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2019. 2492–2501. [doi: 10.1109/CVPR42600.2020.00257]

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

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

9 Yi HW, Wei ZZ, Ding MY, *et al*. Pyramid multi-view stereo net with self-adaptive view aggregation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 766–782.

10 Wang FJH, Galliani S, Vogel C, *et al*. PatchmatchNet: Learned multi-view patchmatch stereo. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14189–14198. [doi: 10.1109/CVPR46437.2021.01397]

11 Ding YK, Yuan WT, Zhu QT, *et al*. TransMVSNet: Global context-aware multi-view stereo network with Transformers. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8575–8584. [doi: 10.1109/CVPR52688.2022.00839]

12 Chang D, Božič A, Zhang T, *et al*. RC-MVSNet: Unsupervised multi-view stereo with neural rendering. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 665–680. [doi: 10.1007/978-3-031-19821-2_38]

13 栗博, 何红艳, 王钰, 等. 面向弱纹理空间目标的特征点匹配方法. 航天返回与遥感, 2024, 45(1): 99–110. [doi: 10.3969/j.issn.1009-8518.2024.00.009]

14 徐一成, 里鹏, 李帅, 等. 基于区域的弱纹理零件三维跟踪方法. 计算机集成制造系统, 1–21. https://doi.org/10.13196/j.cims.2023.0222. [2024-04-17].

15 陈蔓菲. 基于多视图的工件三维重建技术研究 [硕士学位论文]. 长春: 长春理工大学, 2023. [doi: 10.26977/d.cnki.gccgc.2023.000265]

16 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

17 Zhu XZ, Hu H, Lin S, *et al*. Deformable ConvNets V2: More deformable, better results. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 9300–9308. [doi: 10.1109/CVPR.2019.00953]

18 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

19 Aanaes H, Jensen RR, Vogiatzis G, *et al*. Large-scale data for multiple-view stereopsis. International Journal of Computer Vision, 2016, 120(2): 153–168.

20 Seitz SM, Curless B, Diebel J, *et al*. A comparison and evaluation of multi-view stereo reconstruction algorithms. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2006. 519–528. [doi: 10.1109/CVPR.2006.19]

21 Knapitsch A, Park J, Zhou QY, *et al*. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (TOG), 2017, 36(4): 78. [doi: 10.1145/3072959.3073599]

(校对责编: 孙君艳)