

CMCL-DFR: 复杂背景下工业小目标的高精度 6D 位姿估计

刘星宇^{1,2}, 夏仁波^{1†}, 陈月玲¹, 赵昊^{1,2}, 孟祥宇^{1,2}, 赵明宇¹

(1. 中国科学院沈阳自动化研究所机器人学国家重点实验室, 沈阳 110016; 2. 中国科学院大学, 北京 100049)

摘要: 精确的 6D 位姿估计对于柔性制造、机器人抓取与智能装配至关重要, 但仍面临三个主要局限: 复杂背景下小目标检测困难; 传统配准方法对初始值敏感、收敛域小, 而深度学习对工业小目标泛化不足; 现有神经渲染方法主要围绕 RGB-D 图像设计, 旨在优化视图合成质量, 难以直接满足工业场景中对高精度点云几何信息进行无损、精确位姿解算的需求。为此, 本文提出一种“跨模态粗定位与差异化几何精配准”(CMCL-DFR) 的协同估计框架。第一阶段提出基于虚拟渲染的神经位姿估计方法 (VR-NPE), 通过可微渲染将点云桥接至图像域, 并设计几何感知多尺度网络 (GMS-Net) 融合多模态特征, 提升小目标检测与粗定位鲁棒性。第二阶段提出位姿引导的多尺度几何感知配准方法 (PG-MSGAR), 通过曲率分析实现点云自适应区域分割, 为不同几何显著性区域赋予差异化约束权重, 并利用 TEASER++ 抑制离群点, 实现高精度位姿精化。在自建工业零件数据集 (IPD) 上的实验表明, 本文方法平均距离 (ADD) 误差为 0.95mm, 成功率为 91.8%, 与 FoundationPose 相比 ADD 误差降低 48.6%。

关键词: 柔性制造; 小目标检测; 6D 位姿估计; 几何感知神经网络; 点云配准; 多尺度特征融合

中图分类号: TP391 文献标志码: A

DOI: 10.13195/j.kzyjc.2026.0102

引用格式: 刘星宇, 夏仁波, 陈月玲, 等. CMCL-DFR: 复杂背景下工业小目标的高精度 6D 位姿估计 [J]. 控制与决策.

CMCL-DFR: High-accuracy 6D pose estimation for industrial small objects in cluttered scenes

LIU Xing-yu^{1,2}, XIA Ren-bo^{1†}, CHEN Yue-ling¹, ZHAO Hao^{1,2}, MENG Xiang-yu^{1,2}, ZHAO Ming-yu¹

(1. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Accurate 6D pose estimation is crucial for flexible manufacturing, robotic grasping, and intelligent assembly. However, it still faces three major limitations: First, it is difficult to detect small targets against complex backgrounds; Second, traditional registration methods are sensitive to initial estimates, and have narrow convergence basins, while deep learning has poor generalization for industrial small objects; Third, primarily designed based on RGB-D images for view synthesis, existing neural rendering approaches struggle to meet the industrial demand for lossless, precise pose estimation from geometry of high-accuracy point cloud. To address these challenges, a collaborative estimation framework termed "Cross-Modal Coarse Localization and Differentiated Fine Registration" (CMCL-DFR) is proposed. In the first stage, a Virtual Rendering-based Neural Pose Estimation (VR-NPE) method is introduced. Differentiable rendering is used to bridge the point cloud to the image domain. A designed Geometry-aware Multi-Scale Network (GMS-Net) fuses multimodal features to enhance the robustness of small-target detection and coarse localization. In the second stage, a Pose-Guided Multi-Scale Geometric-Aware Registration (PG-MSGAR) method is proposed. In this method, adaptive region segmentation of the point cloud is achieved through curvature analysis. Differential constraint weights are assigned to regions with varying geometric saliency, and TEASER++ is utilized to suppress outliers, thereby enabling high-precision pose refinement. Experimental results on a self-built Industrial Parts Dataset (IPD) demonstrate that the proposed method achieves an Average Distance (ADD) error of 0.95 mm with a success rate of 91.8%, reducing the ADD error by 48.6% in comparison with FoundationPose.

Keywords: flexible manufacturing; small object detection; 6D pose estimation; geometry-aware neural network; point cloud registration; multi-scale feature fusion

收稿日期: 2026-01-29; 录用日期: 2026-04-21.

基金项目: 国家自然科学基金项目 (52505582); 辽宁省自然科学基金项目 (2024-BSBA-55).

责任编辑: 王琦.

†通信作者. E-mail: xiab@sia.cn.

0 引言

精确的 6D 位姿估计是视觉引导机器人操作的核心技术,在柔性制造、智能装配和自动化分拣等领域具有重要应用价值.与传统固定夹具方案相比,基于视觉的位姿估计使机器人能够适应多品种、小批量的生产模式,显著提升产线柔性.然而,工业场景对位姿精度和鲁棒性提出了双重挑战:一方面需要在复杂背景下可靠地检测和定位目标,另一方面需要尽可能提升位姿估计精度以满足装配需求.

针对上述需求,现有方法虽从不同技术路线进行了探索,但在面向工业小目标的高精度估计时仍存在根本性局限.点云配准方法虽直接利用三维几何信息、理论上精度上限更高,但其收敛域有限且对初始化敏感.同时,现有配准方法普遍对所有观测点赋予相同权重,忽视了角点、边缘与平面区域在几何约束能力上的本质差异.另一方面,基于神经渲染的方法虽在复杂背景下鲁棒性较强,但其主要围绕 RGB-D 图像设计,位姿精度受限于深度传感器分辨率,难以直接利用高精度点云的几何信息.对于工业场景中常见的小尺寸目标,其有限的几何结构进一步加剧了判别性特征提取的难度,导致检测鲁棒性与定位精度难以兼顾.

为解决上述挑战,本文提出了一种两阶段的位姿估计方法.该方法通过虚拟渲染实现点云域到图像域的跨模态表征转换,并利用几何差异化约束提升配准精度.第一阶段提出基于虚拟渲染的神经位姿估计方法 (VR-NPE),通过将高精度点云渲染为多视角图像,使神经渲染框架能够处理工业级点云数据;设计几何感知多尺度神经网络 (GMS-Net),通过多模态几何特征和多尺度特征金字塔融合,增强对小目标几何结构的感知能力.第二阶段提出位姿引导的多尺度几何感知配准方法 (PG-MSGAR),通过曲率驱动的自适应区域分割显式建模不同几何区域的差异化约束强度,并基于 TEASER++^[1] 截断最小二乘框架实现对离群对应的鲁棒抑制.

本文的主要贡献如下:

(1) 提出点云-图像域桥接的两阶段位姿估计框架.通过虚拟渲染神经位姿估计方法 (VR-NPE),首次将高精度点云适配至神经渲染框架,解决了其无法直接利用无损几何信息的根本瓶颈;所设计的几何感知多尺度网络 (GMS-Net) 显著提升了复杂背景下小目标的检测与粗定位鲁棒性.

(2) 提出一种曲率驱动的自适应多尺度区域分割方法 (AMRS).该方法在点云配准中首次显式建模

了角点、边缘与平面区域的差异化几何约束强度,突破了传统方法对所有点赋予等权重的局限,为高精度配准提供了结构化的几何先验.

(3) 提出基于 TEASER++ 的鲁棒配准与递进式精化架构 (PG-MSGAR).针对传统 ICP 方法收敛域小且易陷入局部最优的问题,引入截断最小二乘框架实现可证明的全局最优解,结合三层递进式精化架构,降低了对初始位姿质量的依赖.

本文结构安排如下:第 2 节从 6D 位姿估计、点云配准两个维度综述相关工作;第 3 节详细描述所提方法,包括基于虚拟渲染的神经位姿估计 (VR-NPE) 和位姿引导的多尺度几何感知配准 (PG-MSGAR);第 4 节在自建工业零件数据集上进行实验验证,包括与现有方法的对比实验和消融实验;第 5 节总结全文并展望未来工作.

1 相关工作

6D 位姿估计的核心是从观测数据中恢复物体的三维平移与旋转.为满足工业场景对高精度的要求,现有研究普遍采用“直接估计-迭代精化”的两阶段范式:第一阶段从数据中预测初始位姿,第二阶段则通过点云配准进行几何优化以提升精度.本章将依次综述这两个阶段的关键方法,并在此框架下阐述本文工作的贡献.

1.1 6D 位姿估计方法

根据技术路线,6D 位姿估计方法可分为基于模板匹配、基于特征回归和基于渲染比对三类.

基于模板匹配的方法通过构建多视角模板库实现位姿检索.Hinterstoisser 等^[2]提出的 LineMOD 采用梯度方向和表面法向量构建多模态模板,对纹理缺失物体具有较好的鲁棒性;Hodan 等^[3]提出的 EPOS 通过表面片段编码来处理对称物体的位姿歧义问题.此类方法的主要局限在于模板库的存储开销随视角采样密度呈指数增长,且对遮挡场景的鲁棒性不足.

基于特征回归的方法通过深度网络建立图像到位姿的端到端映射.Xiang 等^[4]提出的 PoseCNN 通过语义分割与坐标回归的联合学习实现实例级位姿估计;Wang 等^[5]提出的 DenseFusion 引入 RGB-D 异构特征的逐像素融合机制;He 等^[6]提出的 FFB6D 通过全流双向融合网络在编码-解码各阶段深度融合 RGB 外观与深度几何特征,增强了特征表示能力;Tang 等^[7]提出 DG-6D,通过双流几何特征融合模块提取旋转等变特征与局部几何特征,并引入全局增强机制优化关键点对应关系.孙先涛等^[8]

针对姿态任意、尺寸不一的物体抓取场景, 提出基于语义分割与旋转目标检测的单目位姿估计方法, 采用语义分割与旋转检测双阶段策略实现高效抓取位姿估计. 此类方法虽具有较高的推理效率, 但在特征稀疏或几何复杂的工业场景中, 其精度仍面临固有上限.

基于渲染比对的方法通过可微渲染建立模型与观测的显式对应. Yen-Chen 等^[9]提出的 iNeRF 首次将神经辐射场^[10]应用于位姿优化; Kerbl 等^[11]提出的 3D Gaussian Splatting 因其实时渲染能力受到关注, 但主要面向场景重建. Wen 等^[12]提出的 FoundationPose 结合渲染-比对策略与对比学习实现零样本位姿估计; Labbé 等^[13]提出的 MegaPose 进一步扩展了类别级泛化能力. 近期的 SAM6D^[14]和 Gen6D^[15]将视觉基础模型引入位姿估计, 在开放词汇场景下取得进展. 此类方法在复杂背景下表现出较强的鲁棒性, 但输入模态局限于 RGB-D 图像, 位姿精度受制于深度传感器分辨率. 此外, 张文安等^[16]针对传统目标位姿跟踪在处理旋转运动时精度低、稳定性差的问题, 提出分布式目标位姿跟踪的序列无关融合估计方法, 通过多源信息融合策略提升位姿估计的鲁棒性.

上述方法侧重于从数据中直接预测 6D 位姿, 为进一步充分挖掘三维几何信息、实现更精确的对齐, 点云配准作为关键的迭代精化步骤被广泛应用于位姿估计流程中.

1.2 点云配准方法

前文所述方法构成了 6D 位姿估计的通用框架. 然而, 要充分利用高精度点云数据实现最优的几何对齐, 常需依赖点云配准这一专门的优化过程进行最终精化. 点云配准旨在求解两组点云之间的最优刚体变换, 可进一步提升位姿估计精度. 与基于图像的方法相比, 点云配准直接利用三维几何信息, 其精度理论上仅受限于点云采集精度.

传统方法以迭代最近点 (ICP) 算法^[17]为基础. 经典 ICP 通过交替执行最近点搜索和变换估计实现点云对齐, 但其收敛性强依赖于初始化质量. Chen 等^[18]提出 Point-to-Plane ICP, 利用局部切平面约束加速收敛; Segal 等^[19]提出 G-ICP, 通过概率模型统一建模点到点和点到面距离. 然而, 这些方法的收敛域仍然有限, 在初始误差较大时易陷入局部最优. 针对离群点问题, Yang 等提出 TEASER++, 采用截断最小二乘框架结合 GNC 优化策略, 在理论上保证了全局最优收敛性.

深度学习为点云配准带来了新的可能. Wang 等^[20]提出 DCP, 通过注意力机制学习软对应关系, 规避了硬匹配的组合爆炸问题; Aoki 等^[21]提出 PointNetLK, 将经典 Lucas-Kanade 框架与深度特征结合; Yew 等^[22]提出 RPM-Net, 通过可微 Sinkhorn 层实现端到端的对应学习; Qin 等^[23]提出 GeoTransformer, 引入几何结构编码的 Transformer 架构, 在低重叠和大旋转场景下取得显著突破. Chen 等^[24]提出 LDGR, 采用自适应点卷积与局部图形特征感知机制, 在低重叠场景下实现鲁棒配准. Pu 等^[25]针对点云配准提出层级化策略, 通过多几何特征统计直方图匹配与马氏距离约束优化配准精度. 然而, 这些方法的训练数据主要来自 ModelNet40 等合成数据集, 在工业小目标上面临域迁移挑战, 且精度受限于网络回归能力.

2 CMCL-DFR 协同估计框架

本章详细阐述所提出的两阶段位姿估计方法. 首先给出问题形式化定义和方法总体框架, 然后分别介绍基于虚拟渲染的神经位姿估计方法 (VR-NPE) 和位姿引导的多尺度几何感知配准方法 (PG-MSGAR).

2.1 问题描述

给定目标工件的 CAD 模型 M 和包含该工件的场景点云 P_{scene} , 本文的目标是求解最优刚体变换 $T^* = [R^* | t^*] \in SE(3)$, 使得 CAD 模型经该变换后与场景中的目标工件达到最佳对齐:

$$T^* = \arg \min_{T \in SE(3)} \mathcal{L}(T \cdot P_{\text{model}}, P_{\text{target}}). \quad (1)$$

其中 P_{model} 为从 CAD 模型采样的点云, P_{target} 为从场景中提取的目标点云, $\mathcal{L}(\cdot, \cdot)$ 为配准损失函数. 这一问题的核心挑战在于: 场景点云包含目标工件和背景干扰, 需要首先分割出目标点云; 由于缺乏初始位姿, 直接求解上述优化问题容易陷入局部最优; 小目标点云特征稀疏, 配准精度难以保证.

所提方法采用两阶段架构解决上述挑战, 整体框架如图 1 所示. 第一阶段提出基于虚拟渲染的神经位姿估计方法 (VR-NPE) 实现目标检测、分割和粗定位, 输出目标点云 P_{tar} 和粗定位位姿 T_{init} ; 第二阶段提出位姿引导的多尺度几何感知配准方法 (PG-MSGAR) 基于 T_{init} 进行初始化, 通过多尺度几何感知配准实现位姿精化, 输出最终位姿 T^* .

两阶段策略的设计基于最优化理论的考虑. 位姿估计本质上是求解非凸优化问题, 目标函数 $\mathcal{L}(T)$ 在 $SE(3)$ 流形上存在大量局部极小值. ICP 类方法的收敛性强依赖于初始化质量: 当初始位姿接近真实

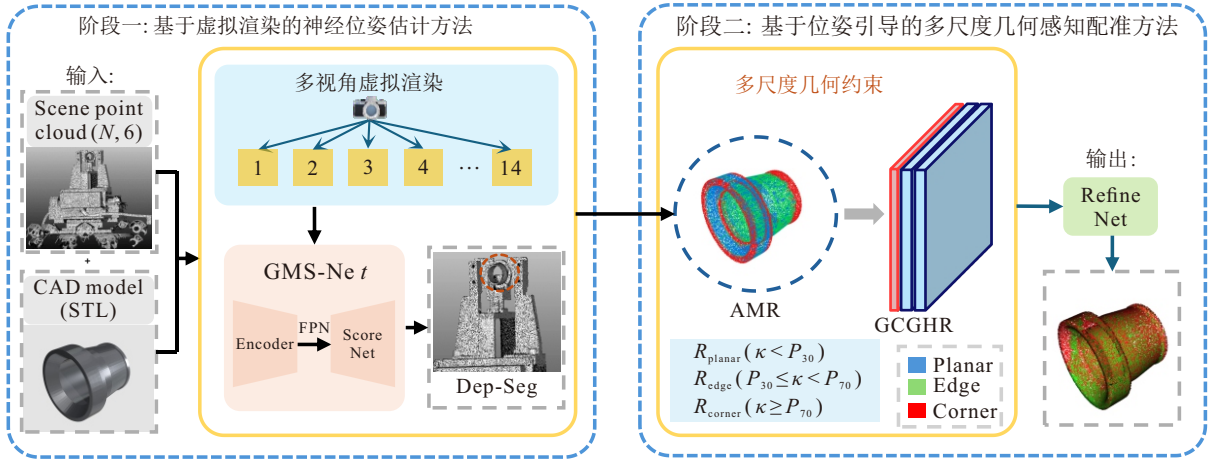


图1 两阶段方法总体流程图

解时,算法能够快速收敛到全局最优;而当初始误差超出收敛域时,则易陷入局部最优.本文的两阶段策略通过 VR-NPE 提供误差在 3mm 以内的粗定位,为后续精配准提供良好的初始化.从互补性的角度,神经渲染方法利用多模态特征在复杂背景下表现鲁棒,但精度受限于传感器分辨率;几何配准方法直接利用点云结构,能够进一步提升位姿精度.两阶段策略融合了两类方法的优势.

2.2 基于虚拟渲染的神经位姿估计 (VR-NPE)

VR-NPE 模块的核心思想是通过虚拟渲染技术将点云配准问题转化为图像域位姿估计问题.神经渲染方法在处理复杂背景时表现出较强的鲁棒性,但主要针对 RGB-D 图像输入设计;高精度点云数据包含丰富的几何信息,但缺乏有效的初始化手段.通过虚拟渲染技术,可以将点云数据转换为 RGB-D 图像,从而利用成熟的神经渲染位姿估计框架.

2.2.1 自适应多视角虚拟渲染策略

在复杂工业场景中,目标工件可能被其他物体部分遮挡,单一视角难以获得目标的完整信息.传统方法采用固定视角配置(如正交 6 视角),但这种配置对非凸目标的覆盖度不足.为确保目标在复杂场景中的完整可见性,本文提出基于信息熵最大化的自适应多视角渲染策略.

给定场景点云 $\mathbf{P}_{\text{scene}}$, 首先计算其几何中心 $\mathbf{c}_{\text{scene}}$ 和最大尺寸 d_{max} , 其中 $\mathbf{c}_{\text{scene}}$ 定义为场景点集 $\{\mathbf{p}_i\}_{i=1}^N$ 的均值. 然后在以 $\mathbf{c}_{\text{scene}}$ 为中心的球面上采样 K 个视角, 相机位置计算为 $\mathbf{e}_k = \mathbf{c}_{\text{scene}} + r \cdot \mathbf{v}_k$, 其中 \mathbf{v}_k 为第 k 个视角方向, $r = 2.0 \cdot d_{\text{max}}$ 为相机到场景中心的距离. 视角采样策略基于覆盖率最大化原则. 定义点 \mathbf{p}_i 在视角 \mathbf{v}_k 下的可见性指示函数:

$$V_i(\mathbf{v}_k) = \mathbb{1}[(\mathbf{p}_i - \mathbf{c}_{\text{scene}}) \cdot \mathbf{v}_k > 0] \mathbb{1}[\neg \text{Occluded}(\mathbf{p}_i, \mathbf{v}_k)], \quad (2)$$

其中第一项判断点是否位于相机前方(朝向相机半球), 第二项通过 Z-buffer 深度测试检测是否被其他点遮挡. 视角 \mathbf{v}_k 的覆盖率定义为可见点占总点数的比例:

$$C(\mathbf{v}_k) = \frac{1}{N} \sum_{i=1}^N V_i(\mathbf{v}_k). \quad (3)$$

最优视角集合通过最大化累积覆盖率并惩罚视角冗余获得:

$$\{\mathbf{v}_k^*\}_{k=1}^K = \arg \max_{\{\mathbf{v}_k\} \subset \mathbb{S}^2} \left[\sum_{k=1}^K C(\mathbf{v}_k) - \lambda \sum_{k=1}^{K-1} \sum_{j=k+1}^K |\mathbf{v}_k \cdot \mathbf{v}_j| \right]. \quad (4)$$

其中第一项最大化总覆盖度, 第二项惩罚视角间的冗余 ($|\mathbf{v}_k \cdot \mathbf{v}_j|$ 越大表示两视角越接近), λ 为平衡系数. 该优化问题中的覆盖函数 $f(V) = |\bigcup_{\mathbf{v}_k \in V} \{\mathbf{p}_i : V_i(\mathbf{v}_k) = 1\}|$ 具有子模性质(边际收益递减), 根据 Nemhauser 等^[26] 的经典结论, 贪心算法可获得不低于 $(1 - 1/e) \approx 63.2\%$ 最优解的近似保证. 具体地, 每次迭代选择边际覆盖增益最大的视角加入集合: $\mathbf{v}_{k+1}^* = \arg \max_{\mathbf{v} \in \mathbb{S}^2 \setminus V_k} [f(V_k \cup \{\mathbf{v}\}) - f(V_k)]$, 直至达到目标视角数 K . 视角数量 K 的选取基于覆盖度与计算效率的权衡. 根据子模优化理论, 覆盖度函数具有边际收益递减性质. 消融实验(表 5)表明, $K = 14$ 时成功率达 91.8%, $K = 18$ 仅提升 0.7% 而渲染耗时增加 28.6%, 因此 $K = 14$ 为最优平衡点. 平衡系数 λ 通过网格搜索确定为 0.5, 此时覆盖率 91.2%, 平均视角间夹角 25.7° , 其几何意义在于确保视角间距不低于 60° , 与球面均匀采样的理论最优间距(约 63.4°) 相近.

渲染过程采用法向量着色方法, 将点云表面法向量 $\mathbf{n} = (n_x, n_y, n_z)$ 映射为 RGB 颜色值: $r = (n_x +$

1)/2, $g = (n_y + 1)/2$, $b = (n_z + 1)/2$. 这种着色方式使得不同朝向的表面呈现不同颜色, 有效增强了几何特征的区分度, 且对光照变化不敏感.

2.2.2 几何感知多尺度神经位姿估计网络 (GMS-Net)

现有神经渲染位姿估计方法主要使用 RGB-XYZ 6 维特征, 对曲率、法向量等高阶几何特征利用不足. 工业零件通常缺乏纹理特征, RGB 信息的区分度有限. 此外, 现有方法采用单一尺度的特征提取架构, 难以同时捕获小目标的局部细节和全局结构.

针对上述问题, 本文提出几何感知多尺度神经位姿估计网络 (GMS-Net), 核心设计思想是“几何特征增强+多尺度融合”. 如图 2 所示, GMS-Net 采用渲染-比对范式, 整体架构包含几何特征编码器、特征金字塔融合模块和位姿评分头三个部分. GMS-Net 的多尺度特征金字塔旨在解决工业小目标检测中局部细节与全局语义难以兼顾的固有矛盾. 工业小目标 ($< 100 \text{ mm}$) 在图像中占据像素有限, 单一尺度的特征存在根本局限: 高分辨率特征图感受野不足, 低分辨率特征图则易稀释细节. 几何特征编码器基于 ResNet-34 骨干网络构建, 为适应 10 通道多模态几何特征输入, 将第一层卷积核从 $7 \times 7 \times 3$ 扩展为 $7 \times 7 \times 10$. 特征金字塔融合模块采用 FPN 架构, FPN 通过自顶向下路径将高层语义信息传递至低层, 同时以横向连接保留空间细节, 使网络在同一前向传播中同时获取多尺度特征表达: Level-1(1/4 分辨率) 捕获微观几何细节, Level-2(1/8 分辨率) 强化边缘轮廓, Level-3(1/16 分辨率) 提供全局位置约束, 实现了局部细节与全局结构的互补融合, 各层特征通过双线性插值上采样后进行通道拼接, 再经 1×1 卷积融合为 256 维特征向量. 在此基础上, 法向量特征编码表面朝向变化, 对边缘与角点敏感; 曲率特征量化局部几何复杂度, 可显式区分平面、边缘与角点三类区域. 几何特征与多尺度金字塔的结合, 使网络能够从局部到全局逐层感知目标的几何结构, 在纹理

贫乏的工业场景中建立起纯 RGB 特征难以获得的判别性表示. 位姿评分头基于 Transformer 注意力机制, 将融合特征与候选位姿的 6 维表示作为输入, 输出标量置信度得分. 为增强对几何结构的感知能力, GMS-Net 构建了 10 维多模态几何特征向量:

$$\mathbf{f}_i = [\tilde{r}_i, \tilde{g}_i, \tilde{b}_i, \tilde{x}_i, \tilde{y}_i, \tilde{z}_i, n_x^i, n_y^i, n_z^i, \kappa_i]^T. \quad (5)$$

其中各分量经过归一化处理: $(\tilde{r}, \tilde{g}, \tilde{b}) \in [0, 1]$ 为法向量着色的 RGB 值, $(\tilde{x}, \tilde{y}, \tilde{z}) = (x, y, z)/d_{\max}$ 为归一化三维坐标, (n_x, n_y, n_z) 为单位表面法向量 ($\|\mathbf{n}\| = 1$), $\kappa \in [0, 1/3]$ 为归一化局部曲率. 曲率特征的引入使网络能够感知局部几何复杂度, 消融实验 (表 4) 表明 10 维特征结合多尺度 FPN 相比 6 维基线, 粗定位精度提升 24.4%.

位姿评分函数综合考虑 RGB 外观相似度、深度一致性和法向量一致性:

$$s_i = w_{\text{rgb}} \cdot s_i^{\text{rgb}} + w_{\text{depth}} \cdot s_i^{\text{depth}} + w_{\text{normal}} \cdot s_i^{\text{normal}}. \quad (6)$$

其中各子评分定义如下: RGB 相似度 $s_i^{\text{rgb}} = \exp(-\|\mathbf{I}_{\text{rgb}}^{\text{render}} - \mathbf{I}_{\text{rgb}}^{\text{obs}}\|_2^2 / \sigma_{\text{rgb}}^2)$ 衡量渲染图像与观测图像的外观一致性; 深度一致性 $s_i^{\text{depth}} = \exp(-|d^{\text{render}} - d^{\text{obs}}|^2 / \sigma_d^2)$ 衡量深度图的匹配程度; 法向量一致性 $s_i^{\text{normal}} = (\mathbf{n}^{\text{render}} \cdot \mathbf{n}^{\text{obs}} + 1) / 2$ 衡量表面朝向的一致性. 公式 (6) 中 RGB、深度、法向量三项的权重依据各自模态的域差异程度进行分配. 工业零件普遍纹理贫乏, 虚拟渲染的 RGB 图像与真实场景外观存在显著差异, 故赋予较低权重. 深度与法向量是几何属性的直接度量, 在虚拟渲染数据与真实传感器获取的深度数据之间具有更好的统计一致性, 因此赋予较高权重. 该设计使网络在位姿评分时优先依赖几何特征, 弱化外观差异的影响, 从而提升跨域泛化能力.

图 3 展示了 GMS-Net 在不同输入配置下的网络激活热力图对比. 其中, 图 3(a) 为 6 维 RGB-XYZ 基线特征的激活响应, 图 3(b) 为增加法向量与曲率后的 10 维几何增强特征响应, 图 3(c) 为在此基础上进一步引入多尺度特征金字塔的响应分布. 对比可

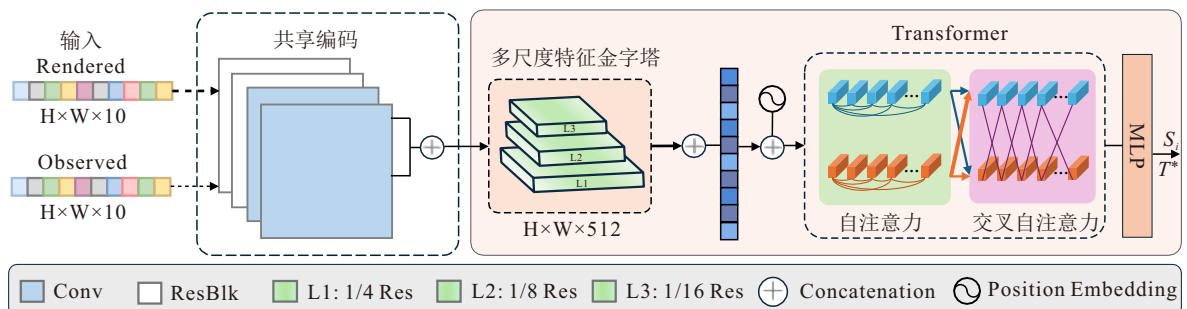


图2 GMS-Net 网络架构图

知, 基线特征的激活响应较为分散, 易受背景杂波干扰; 引入法向量与曲率特征后, 网络的关注点显著集中于目标边缘与角点等几何显著区域; 结合多尺度融合后, 模型对目标整体结构的感知更加完整, 在纹理缺失的工业场景中建立了更强的判别性表示。

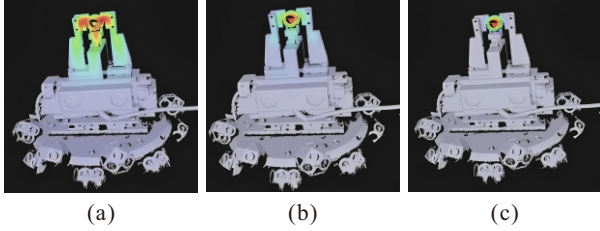


图3 几何特征增强的响应热力图对比

2.2.3 深度一致性目标分割网络

基于粗定位位姿, 需要从场景点云中分割出目标点云。传统的聚类方法在杂乱堆叠场景下容易将相邻物体误分割为同一目标。本文采用深度一致性原理实现目标分割: 若候选位姿正确, 则 CAD 模型渲染的深度图与观测深度图在目标区域应当一致。定义像素 (x, y) 的目标掩模为:

$$M_{\text{target}}(x, y) = \mathbb{1}(|I_{\text{depth}}^{\text{obs}}(x, y) - I_{\text{depth}}^{\text{render}}(x, y)| < \tau_{\text{depth}}). \quad (7)$$

其中 $I_{\text{depth}}^{\text{obs}}$ 为观测深度图, $I_{\text{depth}}^{\text{render}}$ 为基于候选位姿渲染的深度图, $\tau_{\text{depth}} = 3 \text{ mm}$ 基于粗定位误差统计确定。将分割掩模应用于深度图反投影得到的 3D 点云, 即可提取目标区域点云。VR-NPE 模块的输出包括目标点云和粗定位位姿 (平移误差约 $\pm 1-3 \text{ mm}$, 旋转误差约 $\pm 1-3^\circ$), 为后续配准提供良好的初始化。

2.3 位姿引导的多尺度几何感知配准

位姿引导的多尺度几何感知配准 PG-MSGAR 模块以 VR-NPE 提供的粗定位位姿为初始化, 通过多尺度几何分析和层次化配准策略进一步提升位姿精度。粗定位位姿提供了良好的初始化, 使得配准问题从全局搜索转化为局部优化; 同时, 点云中不同几何区域具有不同的配准约束强度, 高曲率区域的几何特征更加显著, 应当优先利用。

2.3.1 自适应多尺度区域分割 (AMRS)

传统配准方法采用统一的点权重策略, 未能充分利用不同几何区域的差异化约束作用。角点和边缘等高曲率区域的法向量变化剧烈, 对位姿变化更加敏感, 能够提供更精确的约束; 平面区域虽然几何约束较弱, 但覆盖面积大, 能够提供全局的配准约束。AMRS 模块基于曲率分析将点云分层, 系统化地捕获不同几何层次的特征。图 4 为点云自适应区域分割可视化图。

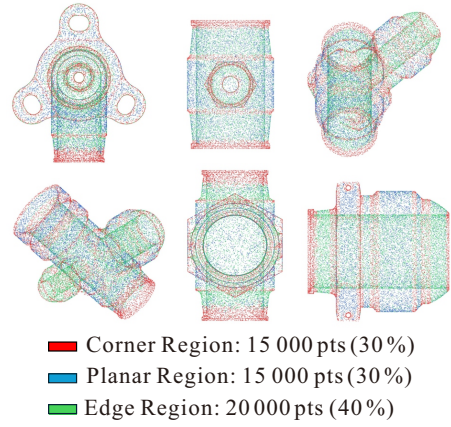


图4 点云自适应区域分割可视化图

对于点云中的每个点 \mathbf{p}_i , 基于其 k 近邻 $\mathcal{N}_i = \{\mathbf{p}_j : \|\mathbf{p}_j - \mathbf{p}_i\| < r_k\}$ 计算局部曲率。首先计算近邻点的质心 $\bar{\mathbf{p}}_i = \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{p}_j \in \mathcal{N}_i} \mathbf{p}_j$, 然后构建协方差矩阵:

$$\mathbf{C}_i = \frac{1}{|\mathcal{N}_i|} \sum_{\mathbf{p}_j \in \mathcal{N}_i} (\mathbf{p}_j - \bar{\mathbf{p}}_i)(\mathbf{p}_j - \bar{\mathbf{p}}_i)^\top. \quad (8)$$

对协方差矩阵进行特征值分解 $\mathbf{C}_i = \mathbf{V}_i \mathbf{\Lambda}_i \mathbf{V}_i^\top$, 得到特征值 $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ 及对应的特征向量 $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ 。特征值的几何意义如下: λ_1 和 λ_2 表征局部表面的主方向延展程度, λ_3 表征沿法向量方向的离散程度; 对应最小特征值的特征向量 \mathbf{v}_3 即为局部表面的法向量估计。基于特征值定义曲率指标:

$$\kappa_i = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \epsilon}. \quad (9)$$

其中 $\epsilon = 10^{-8}$ 为数值稳定性正则化项。该定义与微分几何中的经典曲率概念存在对应关系: 对于光滑曲面, 高斯曲率 $K = \kappa_1 \kappa_2$ 和平均曲率 $H = (\kappa_1 + \kappa_2)/2$ 需要二阶导数信息, 而本文采用的 PCA 曲率仅需一阶邻域统计, 对噪声更加鲁棒。当点位于理想平面上时, $\lambda_3 \rightarrow 0, \kappa_i \rightarrow 0$; 当点位于边缘时, $\lambda_2 \gg \lambda_3, \kappa_i$ 取中等值; 当点位于角点时, $\lambda_1 \approx \lambda_2 \approx \lambda_3, \kappa_i \rightarrow 1/3$ 。

基于曲率分布的百分位数自适应确定分割阈值, 如图 4 所示将点云分为三类区域: 平面区域 $\mathcal{R}_{\text{planar}} (\kappa < P_{30})$, 边缘区域 $\mathcal{R}_{\text{edge}} (P_{30} \leq \kappa < P_{70})$, 角点区域 $\mathcal{R}_{\text{corner}} (\kappa \geq P_{70})$ 。选择 30% 和 70% 分位数作为分割阈值是基于以下考虑: 工业工件通常由平面、边缘和角点三类几何元素组成, 三分法能够较好地捕获这三类特征。消融实验 (表 4) 验证了该分割策略相比其他阈值配置优越性。对于角点处法向量变化剧烈, 对位姿变化最敏感, 应给予最高权重; 边缘处法向量单向变化, 提供一维约束, 给予中等权重; 平面处法向量基本不变, 约束较弱, 但覆盖面积大,

给予低权重.

为保证区域质量, 本文提出综合稳定性评估指标, 结合法向量一致性 $s_{\mathcal{R}}^{\text{norm}}$ 和空间紧凑度 $s_{\mathcal{R}}^{\text{compact}}$:

$$s_{\mathcal{R}} = w_n \cdot s_{\mathcal{R}}^{\text{norm}} + w_c \cdot s_{\mathcal{R}}^{\text{compact}}. \quad (10)$$

仅保留满足稳定性 $s_{\mathcal{R}} > 0.5$ 且点数 $|\mathcal{R}| > \max(50, 0.01N)$ 的候选区域, 以确保后续配准所采用的点集兼具较强的几何特征显著性和较高的结构稳定性. 其中, 法向量一致性 $s_{\mathcal{R}}^{\text{norm}} = 1 - \frac{1}{|\mathcal{R}|} \sum_{\mathbf{p}_i \in \mathcal{R}} \frac{\arccos(|\mathbf{n}_i \cdot \bar{\mathbf{n}}_{\mathcal{R}}|)}{\pi/2}$ 用于表征区域内法向量的一致程度, 其中 $\bar{\mathbf{n}}_{\mathcal{R}}$ 为区域平均法向量. 空间紧凑度 $s_{\mathcal{R}}^{\text{compact}} = 1 - \sigma_{\mathcal{R}}/d_{\max}$ 用于表征区域点云的空间聚集程度, 其中 $\sigma_{\mathcal{R}}$ 为区域点云的标准差.

2.3.2 几何一致性引导的分层配准 (GCGHR)

传统 ICP 方法采用单层配准策略, 对初始值敏感且容易陷入局部最优. 本文设计三层递进式配准架构, 从高特征区域到全局点云逐步精化, 遵循"先鲁棒、后精确"的原则.

L1 层为基于 TEASER++ 的鲁棒初始配准. 在高特征点集 $\mathbf{P}_{\text{high}} = \mathcal{R}_{\text{corner}} \cup \mathcal{R}_{\text{edge}}$ 上, 首先计算 FPFH 特征描述子并建立初始对应关系. 传统 RANSAC 方法对离群对应敏感, 当错误匹配比例较高时易产生错误估计. 本文采用 TEASER++ 的截断最小二乘框架实现鲁棒估计. TEASER++ 将配准问题解耦为旋转和平移两个子问题: 首先通过尺度不变约束构建平移不变量 (TIMs), 对于对应点对 $(\mathbf{p}_i^{\text{src}}, \mathbf{q}_i^{\text{tgt}})$ 和 $(\mathbf{p}_j^{\text{src}}, \mathbf{q}_j^{\text{tgt}})$, 定义 $\bar{\mathbf{p}}_{ij} = \mathbf{p}_i^{\text{src}} - \mathbf{p}_j^{\text{src}}$ 和 $\bar{\mathbf{q}}_{ij} = \mathbf{q}_i^{\text{tgt}} - \mathbf{q}_j^{\text{tgt}}$; 然后通过求解截断最小二乘问题估计旋转矩阵:

$$\mathbf{R}^* = \arg \min_{\mathbf{R} \in \text{SO}(3)} \sum_{(i,j)} \min(\|\bar{\mathbf{q}}_{ij} - \mathbf{R}\bar{\mathbf{p}}_{ij}\|^2, \bar{c}^2). \quad (11)$$

其中 \bar{c} 为截断阈值, 通过 GNC (Graduated Non-Convexity) 算法从凸松弛逐步收紧求解. TEASER++ 在一定噪声与内点假设下具有鲁棒的初始化能力, 能够在存在较多错误对应时输出稳定的初始位姿, 从而降低匹配误差对估计结果的影响, 为后续精配准提供可靠的初始变换 \mathbf{T}_{L1} .

L2 层以 \mathbf{T}_{L1} 为初值, 在扩展点集 $\mathbf{P}_{\text{high}} \cup \mathbf{P}_{\text{planar}}$ 上执行 Point-to-Plane ICP 优化:

$$\mathbf{T}_{L2} = \arg \min_{\mathbf{R} \in \text{SO}(3), \mathbf{t} \in \mathbb{R}^3} \sum_i ((\mathbf{R}\mathbf{p}_i^{\text{src}} + \mathbf{t} - \mathbf{q}_i^{\text{tgt}}) \cdot \mathbf{n}_{q_i})^2. \quad (12)$$

其中 \mathbf{n}_{q_i} 为目标点 $\mathbf{q}_i^{\text{tgt}}$ 处的单位法向量. 切平面约束使算法在平滑表面上具有超线性收敛速度, 平面区域的引入提供全局几何约束以抑制配准漂移.

L3 层在完整点云上执行 Generalized ICP, 利用

点和协方差信息实现最终精化:

$$\mathbf{T}_{L3} = \arg \min_{\mathbf{R} \in \text{SO}(3), \mathbf{t} \in \mathbb{R}^3} \sum_i \mathbf{d}_i^T (\mathbf{C}_i^{\text{src}} + \mathbf{R}\mathbf{C}_i^{\text{tgt}}\mathbf{R}^T)^{-1} \mathbf{d}_i. \quad (13)$$

其中 $\mathbf{d}_i = \mathbf{R}\mathbf{p}_i^{\text{src}} + \mathbf{t} - \mathbf{q}_i^{\text{tgt}}$ 为配准残差, $\mathbf{C}_i^{\text{src}}, \mathbf{C}_i^{\text{tgt}} \in \mathbb{R}^{3 \times 3}$ 分别为源点和目标点的局部协方差矩阵. 三层递进架构通过逐步引入更多几何约束, 有效提升配准的鲁棒性.

该架构通过三层递进设计, 在数据范围、初值依赖与优化目标三个维度上逐层深化: 首先, L1 层仅利用角点与边缘点集进行鲁棒初始配准, 最大限度容忍离群对应; 其次, L2 层以 L1 层结果为初值, 扩展至包含平面区域的点集, 通过切平面约束实现快速几何精化; 最后, L3 层在完整目标点云上进行全局微调, 实现最终精化. 三层之间依次衔接, 前一层的输出作为后一层的输入初值, 共同构成从鲁棒初始化到几何精化再到全局微调的递进优化流程.

3 实验结果与分析

本章通过系统实验验证所提方法的有效性. 首先介绍实验设置 (3.1 节), 然后进行主实验对比 (3.2 节), 接着进行消融实验 (3.3 节).

3.1 实验设置

实验在配备双路 Intel Xeon Gold 6330 CPU (2.00 GHz) 和 NVIDIA RTX A6000 GPU (48 GB 显存) 的服务器上进行. GMS-Net 使用 PyTorch 1.12 实现, 采用 Adam 优化器, 初始学习率为 10^{-4} , 权重衰减为 10^{-5} , 批大小为 32, 训练 50 个 epoch. PG-MSGAR 基于 Open3D 0.16 实现, 关键超参数: 曲率近邻数 $k = 20$, 体素大小 $v = 0.5$ mm. 本文方法的计算开销主要来自 VR-NPE 和 PG-MSGAR 两个阶段. VR-NPE 模块的计算负担源于多视角渲染 ($K = 14$) 与 GMS-Net 网络推理; PG-MSGAR 模块的计算开销集中于曲率计算与三层递进配准. 在 NVIDIA RTX A6000 GPU 上, 单样本处理时间约为 8–10 秒, 满足离线高精度估计的需求. 主要计算瓶颈在于多视角渲染的串行执行, 后续可通过并行化与网络轻量化进一步优化.

现有公开位姿估计数据集主要面向 RGB-D 图像场景设计, 在点云精度和目标特性方面与本文问题设定存在显著差异. YCB-Video 和 LineMOD 等主流数据集的深度精度约为 1–3 mm, 目标物体以日常用品为主, 多数在 100–300 mm 范围, 难以满足本文针对高精度点云和工业小目标的研究需求. 此外, ModelNet40 等合成数据集虽然提供了标准化的点云数据, 但缺乏工业场景下常见的观测不完整性 (如遮

挡/自遮挡导致的点缺失)、点密度变化以及复杂背景干扰等因素,与实际应用仍存在较大域差异.鉴于上述原因,本文构建了专门面向高精度点云位姿估计的工业零件数据集 (Industrial Parts Dataset, IPD).

IPD 数据集包含 8 类共 72 个工业零件:管道接头 (16 个)、紧固件 (18 个)、夹具 (10 个)、密封件 (8 个)、弹簧 (7 个)、轴承 (6 个)、齿轮 (4 个) 和连接件 (3 个).零件 CAD 模型来源于实际工业生产线,涵盖平面、圆柱面、螺纹、法兰、齿形等典型几何特征.我们使用 BlenderProc 构建包含复杂背景的合成场景,并通过随机视角与干扰物布置控制遮挡程度在 0%–50% 范围内.针对每个零件生成 25 个合成场景样本,共计 1800 个合成样本.图 5 为仿真 IPD 数据集示例.

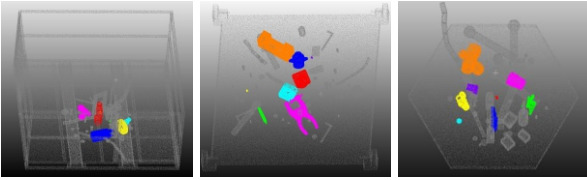


图5 仿真 IPD 数据集示例

点云由虚拟扫描流程生成,以近似实际结构光采集中“可见性约束、遮挡缺失与点密度变化”等主要退化因素.具体地,我们仅保留从观测视角可见的表面点,并通过深度一致性判定实现遮挡与自遮挡下的点缺失建模;同时采用随机下采样控制点密度,以模拟不同距离与反射条件下有效回波数量的变化.本文不额外叠加参数化坐标扰动(如高斯噪声),以避免在缺乏目标传感器误差标定统计的情况下引入不可靠的噪声假设.

评价指标采用 ADD(Average Distance of Model Points) 与 ADD-S(ADD-Symmetric) 以及成功率 (Success Rate, SR). 本文所有距离量均以毫米 (mm) 计. ADD 定义为模型点在预测位姿和真实位姿下的平均对应点距离:

$$ADD = \frac{1}{|M|} \sum_{x \in M} \|(\mathbf{R}x + \mathbf{t}) - (\hat{\mathbf{R}}x + \hat{\mathbf{t}})\|. \quad (14)$$

其中 M 为模型点集, (\mathbf{R}, \mathbf{t}) 和 $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ 分别为真实和预测位姿. ADD-S 针对对称物体设计,计算预测点云到真实点云的最近邻距离:

$$ADD-S = \sum_{x_1 \in M} \min_{x_2 \in M} \|(\mathbf{R}x_1 + \mathbf{t}) - (\hat{\mathbf{R}}x_2 + \hat{\mathbf{t}})\| \quad (15)$$

成功率定义为 $ADD-S < 2 \text{ mm}$ 的样本比例,用于评估完整框架在复杂背景下的整体性能:

$$SR_{\text{pose}} = \frac{N_{ADD-S < 2\text{mm}}}{N_{\text{total}}} \times 100\%. \quad (16)$$

配准实验中收敛率定义为 $RMSE < 2\text{mm}$ 的样本比例,用于评估配准方法的鲁棒性,并仅在收敛样本上统计 RMSE:

$$CR = \frac{N_{RMSE < 2\text{mm}}}{N_{\text{total}}} \times 100\%. \quad (17)$$

3.2 对比实验

本实验评估各方法在复杂背景完整场景中的端到端位姿估计能力.输入为包含目标和背景的完整场景,输出为目标 6D 位姿.工业零件中存在大量回转体等对称物体,对于这类物体,ADD 指标会因对称等效位姿而产生误判(如圆柱体绕轴旋转 180° 实际位姿等效),而 ADD-S 通过最近邻匹配规避了这一问题.本文同时报告 ADD 和 ADD-S 两种指标:ADD 反映非对称物体的精确配准能力,ADD-S 反映对称物体的有效配准能力.

实验设计说明:为在统一基准下评估各方法从输入到最终位姿输出的端到端流程性能,表 1 在统一相机参数与可见性约束下,将端到端对比方法(如 FoundationPose、DenseFusion 等)的输入统一为由本文 IPD 数据集高精度点云渲染生成的模拟 RGB-D 观测,以实现不同端到端方法在同一输入接口上的可比评估.需指出的是,该模拟观测的 RGB 分量由几何信息构造,与真实成像的纹理与光照统计存在域差异,因此相关结果仅用于统一协议下的参考性对照,不作为对 RGB-D 方法原生输入域性能的最终结论.本文方法以高精度点云为起点,VR-NPE 旨在学习几何一致的视觉表征,从而增强遮挡与复杂背景条件下的位姿估计稳定性.

表1 端到端位姿估计方法在完整场景下的性能对比

方法	ADD/mm	ADD-S/mm	成功率/%
FoundationPose	1.85	0.92	86.5
MegaPose	3.52	1.45	32.2
DenseFusion	3.21	1.54	72.6
FFB6D	2.46	1.13	80.1
Ours	0.95	0.76	91.8

如表 1 所示,在本文统一渲染 RGB-D 协议与遮挡设置下,现有 RGB-D 方法(如 FoundationPose)的精度与成功率仍存在提升空间.尤其在亚毫米级精度需求下,基于 RGB-D 的端到端回归范式可能对输入域差异、深度量化与遮挡缺失更为敏感,从而使最终误差呈现一定下限.相比之下,本文方法以高精度点云为起点,VR-NPE 阶段显式利用几何信息生成几何一致的视觉表征,并在 PG-MSGAR 阶段结合鲁

棒初始化与几何一致性细化, 实现了更稳定的收敛率与更高的精度. 综上, 在柔性制造的高精度场景中, 充分利用原始几何并结合神经表征与几何优化的混合范式更具潜力.

此外, 点云配准方法 (包括传统与深度学习范式) 本质上被设计为精化模块, 其运行以已知的目标分割和初始位姿为前提, 因此不适用于本表的端到端场景评估. 图 6 给出了真实工业场景下两阶段小目标分割配准的可视化结果, 每行展示一个测试案例, 从左到右依次为输入场景点云、VR-NPE 分割结果、PG-MSGAR 配准结果, 配准后模型点云与目标子点云实现高一一致性对齐.

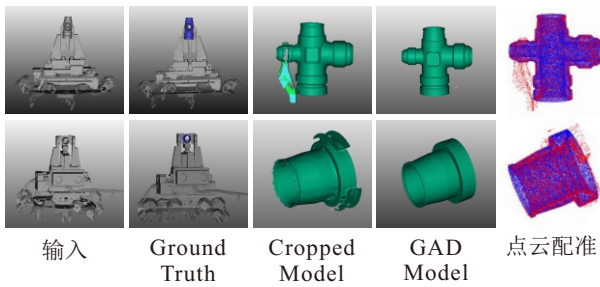


图6 真实工业场景测试可视化结果

为公平评估各配准方法在理想条件下的优化潜力与精度上限, 我们在 Oracle 分割与统一初始位姿 (来自 VR-NPE, 误差 1–3mm) 下进行对比. 本实验中, RMSE 仅对收敛样本 ($RMSE < 2$ mm) 进行计算, 旨在剥离方法在失败样本上的差异性, 聚焦评估其在成功收敛时所能达到的极限精度; 收敛率则独立衡量其鲁棒性. 结果表明, 传统方法 (如 Point-to-Plane ICP) 在收敛时可达到较高的精度上限, 但其收敛率受限于局部最优; 而部分深度方法 (如 RPM-Net) 则受训练数据域差异及端到端回归的毫米级精度瓶颈制约. 本文方法通过融合鲁棒对应估计与几何迭代优化, 在此基准上同时实现了最高的精度上限与收敛率, 验证了其设计在潜力与稳定性上的双重优势.

从表 2 可以看出, PG-MSGAR 在各项指标上均达到最优. 与次优方法 Geo-Transformer 相比, 其 RRE 由 1.12° 降至 0.72° , 降幅为 35.7%; RTE 降至 0.18 mm, 降幅为 35.7%; 收敛样本 RMSE 降至 0.168 mm, 降幅为 31.4%. 传统优化方法如 Point-to-Plane ICP 的收敛样本精度尚可, 但收敛率不超过 85.2%, 制约了其整体表现. 本文方法在无需训练数据的前提下, 实现了 94.2% 的最高收敛率与最优精度的统一. 这一优势源于三层递进架构的功能分工: L1 层提供鲁棒初始化, L2 层实现快速收敛, L3 层完

成最终微调. 三者逐层递进, 互为前提, 使方法同时具备了单阶段方法难以兼顾的鲁棒性与高精度.

表2 配准方法性能对比 (Oracle 分割条件)

方法	RRE($^\circ$)	RTE(mm)	RMSE(mm)	收敛率/%
ICP	1.85	0.42	0.352	81.8
G-ICP	1.52	0.38	0.318	82.5
Point-to-Plane	1.28	0.32	0.285	85.2
RPM-Net	2.40	0.65	0.580	75.0
DCP	2.15	0.58	0.520	70.5
GeoTransformer	1.12	0.28	0.245	88.5
PG-MSGAR	0.72	0.18	0.168	94.2

3.3 消融实验

为验证各组件的有效性, 通过移除关键模块进行消融研究.

表 3 的对比验证了差异化权重设计的有效性. 仅用角点区域时, RRE 为 1.35° , RMSE 为 0.268 mm, 相对较低, 收敛率达 82.5%, 说明角点区域具有最强的几何约束能力, 但因覆盖不均匀导致收敛率未能达到最优; 仅用平面区域时, RTE 为 0.38 mm, 尚可, 但 RRE 为 1.52° , 较差, 表明平面区域能提供全局位置约束但对旋转不敏感; 仅用边缘区域时各项指标均不理想, 说明边缘区域单独使用时约束质量有限. 三类联合 (AMRS) 在所有指标上均达到最优, 相比统一权重配准, RMSE 降低 31.4%, 收敛率提升 6.0%. 这一结果证明: 角点区域提供强旋转约束, 平面区域提供全局位置约束, 边缘区域起过渡连接作用, 三类区域的协同作用产生了显著的改进效果, 单一区域均无法达到同等性能.

表3 区域配准对比实验

配置	RRE($^\circ$)	RTE(mm)	RMSE(mm)	收敛率/%
仅用角点区域	1.35	0.35	0.268	82.5
仅用平面区域	1.52	0.38	0.318	75.2
仅用边缘区域	1.95	0.45	0.385	68.5
三类联合(AMRS)	0.85	0.22	0.168	94.2
统一权重配准	1.18	0.32	0.245	88.2

表 4 系统级消融实验结果表明: (1) 多视角渲染模块对系统鲁棒性影响最为显著, 其缺失导致成功率下降 13.3%, 证明了多视图信息在复杂遮挡背景下稳定检测目标的关键作用; (2) 精配准模块 (PG-MSGAR) 是达成高精度的核心, 移除后, 成功样本的 ADD-S 误差从 0.76 mm 升至 1.90 mm (增幅 150%), 同时成功率下降 15.8%, 验证了渐进式迭代优化对提升位姿精度的决定性贡献; (3) 几何特征增强 (GMS-Net) 与分层配准策略 (GCGHR) 分别贡献约

表4 系统级消融实验

配置	ADD(mm)	ADD-S(mm)	成功率/%	Δ 成功率
完整方法	0.95	0.76	91.8	-
w/o GMS-Net (用FoundationPose替代)	1.52	1.12	85.8	-6.0%
w/o多视角渲染(单视角)	1.70	1.30	78.5	-13.3%
w/o PG-MSGAR(仅粗定位)	2.35	1.95	75.0	-16.8%
w/o AMRS(统一权重配准)	1.18	0.85	88.2	-3.6%
w/o GCGHR(单层ICP)	1.35	1.01	85.5	-6.3%

6.0%与6.3%的成功率提升及相应的精度改善,体现了针对性设计在提升整体性能中的有效性。

表5关键模块的消融实验结果进一步阐释了各设计细节的贡献:(1)几何特征通道对粗定位精度贡

献显著,法向量通道提升11.2%,曲率通道进一步提升7.8%,多尺度FPN额外提升5.4%;(2)视角数 $K=14$ 在覆盖度和计算效率之间取得最优平衡,当 $K=14$ 时成功率达91.8%, $K=18$ 仅再提升0.7%而渲染耗时增加28.6%;(3)曲率分割阈值30%–70%分位数配置的RMSE为0.168 mm,显著优于20%–60%配置(0.224 mm)和25%–75%配置(0.195 mm),验证了该阈值设置的合理性;(4)三层递进配准架构使RMSE从L1层(0.52 mm)到L1+L2层(0.32 mm)再到完整三层(0.168 mm)逐层降低67.7%,验证了三层递进架构的有效性。

表5 关键模块消融实验

模块	消融变量	变体	性能	最优配置
GMS-Net	输入特征	RGB+XYZ (6D)	Top-1: 68.5%	
		+Normal (9D)	Top-1: 76.2% (+11.2%)	
		+Normal+Curv (10D)	Top-1: 81.5% (+19.0%)	
		+多尺度FPN	Top-1: 85.2% (+24.4%)	✓
VR-NPE	视角数 K	6	成功率: 78.5%	
		10	成功率: 85.2%	
		14	成功率: 91.8%	✓
		18	成功率: 92.5%	
PG-MSGAR	分割阈值	20%–60%	RMSE: 0.225 mm	
		25%–75%	RMSE: 0.195 mm	
		30%–70%	RMSE: 0.168 mm	✓
		L1 only	RMSE: 0.52 mm	
配准层数	L1+L2	RMSE: 0.32 mm		
	L1+L2+L3	RMSE: 0.168 mm	✓	

4 结论

本文针对柔性制造场景下复杂背景小目标位姿估计的难题,提出了一种“跨模态粗定位与差异化几何精配准”(CMCL-DFR)的协同估计框架,第一阶段VR-NPE通过虚拟渲染策略桥接点云域与图像域,结合几何感知多尺度神经网络实现复杂背景下的目标检测与粗定位;第二阶段PG-MSGAR通过曲率驱动的差异化约束建模和三层递进配准架构,在粗定位基础上进一步提升位姿精度.两阶段策略融合了神经渲染方法的鲁棒性与几何配准方法的高精度优势。

在自建工业零件数据集IPD上的实验表明,在模拟RGB-D输入的统一条件下,所提方法达到0.95 mm平均ADD误差和91.8%成功率,相比FoundationPose的ADD误差降低48.6%,相比GeoTransformer的RMSE降低63.3%.消融实验验证了各关键模块的有效性:10维几何特征使粗定位精度提升24.4%,AMRS模块使ADD降低24.2%,

三层递进配准使RMSE从0.52 mm降至0.168 mm.

本研究提出的方法在处理已知CAD模型的工业零件时展现了高精度优势,但其当前流程仍依赖于CAD模型以及质量较高的几何观测(点云或深度),并包含候选生成与几何优化等步骤,因此在部署便捷性与实时性方面仍存在一定权衡.未来工作将集中于:第一,开发轻量化与加速版本,通过减少渲染与候选规模、引入并行化与早停策略,并探索与消费级深度相机或稀疏点云输入的结合,以兼顾精度与工程可用性;第二,在保持几何一致性优势的基础上,引入更具泛化能力的特征学习与鲁棒建模机制,以提升对复杂遮挡、形变以及更广泛零件类型的适应性,并进一步拓展至弱先验条件下的位姿估计任务。

参考文献 (References)

- [1] Yang H, Shi J N, Carlone L. TEASER: Fast and certifiable point cloud registration[J]. *IEEE Transactions on Robotics*, 2021, 37(2): 314-333.

- [2] Hinterstoisser S, Lepetit V, Ilic S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C]. *Computer Vision – ACCV 2012*. Berlin, Heidelberg: Springer, 2013: 548-562.
- [3] Hodaň T, Baráth D, Matas J. EPOS: Estimating 6D pose of objects with symmetries[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2020: 11700-11709.
- [4] Xiang Y, Schmidt T, Narayanan V, et al. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[J/OL]. 2017, arXiv: 1711.00199.
- [5] Wang C, Xu D F, Zhu Y K, et al. DenseFusion: 6D object pose estimation by iterative dense fusion[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2020: 3338-3347.
- [6] He Y S, Huang H B, Fan H Q, et al. FFB6D: A full flow bidirectional fusion network for 6D pose estimation[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, 2021: 3002-3012.
- [7] Tang Y H, Shi J L, Liu C S, et al. DG-6D: Dual-stream geometric feature fusion and global enhanced key-point feature extraction for category-level 6D pose estimation[J]. *Measurement*, 2026, 257: 118921.
- [8] 孙先涛, 闻勇, 陈文杰, 等. 基于语义分割与旋转目标检测的机器人抓取位姿估计[J]. *控制与决策*, 2024, 39(9): 2913-2922.
(Sun X T, Wen Y, Chen W J, et al. Robot grasping pose estimation based on semantic segmentation and rotating target detection[J]. *Control and Decision*, 2024, 39(9): 2913-2922.)
- [9] Yen-Chen L, Florence P, Barron J T, et al. iNeRF: Inverting neural radiance fields for pose estimation[C]. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Prague, 2021: 1323-1330.
- [10] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[J/OL]. 2020, arXiv: 2003.08934.
- [11] Kerbl B, Kopanas G, Leimkuehler T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. *ACM Transactions on Graphics*, 2023, 42(4): 1-14.
- [12] Wen B W, Yang W, Kautz J, et al. FoundationPose: Unified 6D pose estimation and tracking of novel objects[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2024: 17868-17879.
- [13] Labbé Y, Manuelli L, Mousavian A, et al. MegaPose: 6D pose estimation of novel objects via render & compare[J/OL]. 2022, arXiv: 2212.06870.
- [14] Lin J H, Liu L H, Lu D K, et al. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2024: 27906-27916.
- [15] Liu Y, Wen Y L, Peng S D, et al. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images[C]. *Computer Vision – ECCV 2022*. Cham: Springer, 2022: 298-315.
- [16] 张文安, 张怀政, 孙虎, 等. 分布式目标位姿跟踪的序列无关融合估计方法[J]. *控制与决策*, 2025, 40(7): 2125-2134.
(Zhang W A, Zhang H Z, Sun H, et al. Sequence-independent fusion estimation method for distributed target pose tracking[J]. *Control and Decision*, 2025, 40(7): 2125-2134.)
- [17] Besl P J, McKay N D. A method for registration of 3-D shapes[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992, 14(2): 239-256.
- [18] Chen Y, Medioni G. Object modelling by registration of multiple range images[J]. *Image and Vision Computing*, 1992, 10(3): 145-155.
- [19] Segal A, Haehnel D, Thrun S. Generalized-ICP[C]. *Proceedings of Robotics: Science and Systems*. Seattle: RSS, 2009: 21.
- [20] Wang Y, Solomon J. Deep closest point: Learning representations for point cloud registration[C]. *IEEE/CVF International Conference on Computer Vision*. Seoul, 2020: 3522-3531.
- [21] Aoki Y, Goforth H, Srivatsan R A, et al. PointNetLK: Robust & efficient point cloud registration using PointNet[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2020: 7156-7165.
- [22] Yew Z J, Lee G H. RPM-net: Robust point matching using learned features[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2020: 11821-11830.
- [23] Qin Z, Yu H, Wang C J, et al. GeoTransformer: Fast and robust point cloud registration with geometric transformer[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(8): 9806-9821.
- [24] Chen Y L, Mei Y, Lu T, et al. Adaptive spatial feature extraction and graphical feature awareness for robust point cloud registration[J]. *Neural Networks*, 2026, 193: 107966.
- [25] Pu D D, Chai H Z, Dong C, et al. Hierarchical strategy and correspondence optimization for ALB point cloud registration[J]. *Measurement Science and Technology*, 2026, 37(1): 015201.
- [26] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of approximations for maximizing submodular set functions — I[J]. *Mathematical Programming*, 1978, 14(1): 265-294.

作者简介

刘星宇 (2000–), 男, 硕士生, 从事点云分割点云配准的研究, E-mail: liuxingyu@sia.cn;

夏仁波 (1977–), 男, 研究员, 博士, 从事计算机视觉、工业光学测量与检测的研究, E-mail: xiarb@sia.cn;

陈月玲 (1987–), 女, 副研究员, 博士, 从事计算机视觉、模式识别方向的研究, E-mail: chenyueling@sia.cn;

赵昊 (1995–), 男, 博士研究生, 从事多目视觉测量、图像处理的研究, E-mail: zhaohao@sia.cn;

孟祥宇 (1999–), 男, 硕士生, 从事光学测量、三维重建的研究, E-mail: mengxiangyu@sia.cn;

赵明宇 (2001–), 男, 硕士生, 从事深度学习、目标检测的研究, E-mail: zhao mingyu@sia.cn.