工程科学学报
Chinese Journal of Engineering

面向抓取检测的位姿估计数据集自动采集标注系统

陈鹏 白勇 孙翰翔

Automatic data collection and annotation system for a pose estimation dataset designed for grasping detection

CHEN Peng, BAI Yong, SUN Hanxiang

引用本文:
陈鹏, 白勇, 孙翰翔. 面向抓取检测的位姿估计数据集自动采集标注系统[J]. 北科大：工程科学学报, 2024, 46(8): 1458–1468. doi: 10.13374/j.issn2095–9389.2023.09.28.001
CHEN Peng, BAI Yong, SUN Hanxiang. Automatic data collection and annotation system for a pose estimation dataset designed for grasping detection[J]. *Chinese Journal of Engineering*, 2024, 46(8): 1458–1468. doi: 10.13374/j.issn2095–9389.2023.09.28.001

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.09.28.001

工程科学学报, 第 46 卷, 第 8 期: 1458−1468, 2024 年 8 月
Chinese Journal of Engineering, Vol. 46, No. 8: 1458−1468, August 2024
https://doi.org/10.13374/j.issn2095-9389.2023.09.28.001; http://cje.ustb.edu.cn

面向抓取检测的位姿估计数据集自动采集标注系统

陈　鹏，白　勇✉，孙翰翔

河北工业大学人工智能与数据科学学院, 天津 300401
✉通信作者，E-mail:b18234912627@163.com

摘　要　机器人抓取在物流分拣、自动装配和医疗手术等领域中具有广泛的应用. 抓取检测是机器人抓取中的重要步骤之一, 随着三维传感器的成本逐渐降低, 抓取检测任务中越来越多地使用深度相机采集彩色图像和深度图像对 (RGB-D), 并采用基于位姿估计的方法实现机器人抓取. 然而, 目前已经公开的基于 RGB-D 图像的位姿估计数据集, 大多需要借助价格昂贵的三维激光扫描仪获得目标物体的三维模型, 而且标注过程依赖人工操作, 费时费力, 不利于大规模数据集的制作. 为此, 本文设计并实现了一个面向位姿估计的数据集自动采集标注系统. 该系统无需使用三维激光扫描仪, 只通过采集、分析由深度相机获得的 RGB-D 图像序列即可重建出目标物体的三维模型, 并自动标注目标物体的位姿信息, 生成二维图像中的分割掩码. 实验中, 使用该系统制作了包含 84 个物体、8400 张 RGB-D 图像的位姿估计数据集, 并将自动标注数据与手动标注数据进行了对比, 发现两者分割掩码重合率可以达到 98%, 并且自动标注的位姿信息能够使模型点云与场景点云的对齐率达到 100%, 充分说明了所提系统自动标注结果的准确性与可靠性.

关键词　抓取检测; 自动标注; 三维重建; 位姿估计; 分割掩码
分类号　TP391.4

Automatic data collection and annotation system for a pose estimation dataset designed for grasping detection

CHEN Peng, *BAI Yong*✉, *SUN Hanxiang*

School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
✉Corresponding author, E-mail: b18234912627@163.com

ABSTRACT　Robotic grasping has extensive applications in fields such as logistics sorting, automated assembly, and medical surgery. Grasping detection is an important step in robotic grasping. Recently, with the decrease in their costs, depth cameras have been gradually applied for grasping detection, which has promoted the application of pose estimation-based methods for robotic grasping. However, most publicly available RGB-D image-based pose estimation datasets rely on equipment such as expensive 3D laser scanners to obtain 3D models of target objects. Meanwhile, the annotation process relies heavily on manual operation, which is time-consuming, labor-intensive, and unfavorable for the creation of large-scale datasets. To address these issues, this study implements a dataset automatic acquisition and annotation system aimed at developing RGB-D image-based pose estimation methods for robotic grasping. The proposed system deploys easily and does not require an expensive 3D laser scanner. RGB-D image sequences are obtained only by an off-the-shelf depth camera, and the system can automatically acquire the reconstructed 3D model of the target object, annotated pose information, and 2D image segmentation masks. During the process of developing the automatic annotation algorithm for the proposed system, a novel minimum spanning tree-based normal propagation method is proposed to guarantee that consistent normal directions can be acquired so that deformations or tearing on the reconstructed 3D surface caused by inconsistent normal directions can be avoided. During the

收稿日期: 2023−09−28
基金项目: 国家自然科学基金资助项目(U20A20201); 河北省高等学校科学技术研究项目(QN2022048)

experiments, the proposed system created a pose estimation dataset containing 84 objects with 8400 RGB-D images. 3D models, image segmentation masks, and 6D poses were annotated by the system in every RGB-D image for each object. To evaluate the accuracy of the annotated segmentation masks, the annotated segmentation masks and the corresponding manually labeled results were compared. Furthermore, the accuracy of the annotation results was also assessed from the performance of an instance segmentation network trained by the annotated image masks. To evaluate the accuracy of the annotated poses, a point cloud registration mission was launched to align the model point cloud and the scene point cloud using the annotated pose parameters. Furthermore, a category-level pose estimation network was trained using the annotated pose parameters, and its performance can directly reflect the accuracy of the annotation results. The experimental results show that the overlapped area between the annotated mask and the manually labeled mask is greater than 98%. Additionally, a 100% alignment rate can be achieved, meaning that the model point cloud can be aligned to any scene point cloud through the corresponding annotated pose parameters. These results demonstrate that the designed and implemented system in this paper can be used to sufficiently create a high-quality dataset for developing real pose estimation-related solutions. A solid data foundation can be provided on the basis of the proposed system for future research and application of deep learning models aimed at robotic grasping detection.

KEY WORDS　grasp detection; automatic labeling; 3D reconstruction; pose estimation; segmentation mask

机器人抓取在物流分拣、自动装配和医疗手术等领域具有广泛的应用，是实现智能化、自动化的关键技术. 随着人工智能和机器学习技术的应用和推广，机器人抓取越来越受到人们的广泛关注. 通常机器人抓取流程可以细分为三个子任务，即抓取检测、抓取规划和机器人控制，其中抓取检测是完成后面两个子任务的基础. 随着视觉传感器技术的发展，特别是三维传感装置的成本逐渐降低，在抓取检测任务中越来越多地使用深度相机，通过视觉计算的方式估计场景中抓取目标的位姿信息，并由位姿信息引导机器人完成抓取. 这种基于位姿估计的抓取方法也被认为是目前比较稳定可靠的抓取策略. 例如，徐进等[1]利用固定于环境中的深度相机搭建了一个无序抓取系统，可以完成零件抓取的部署和作业. 夏浩宇等[2]将深度相机应用于纺纱企业络筒工序中管纱的自动抓取和上料过程中，提出了基于关键点区域卷积神经网络改进模型的物体抓取检测算法. 茅凌波等[3]仅利用单张 RGB 彩色图像，设计了一种单视图两阶段物体位姿估计算法，在自制抓取数据集上，通过投票机制筛选目标表面关键点，再通过求解 P*n*P(Perspective-*n*-point) 问题，得到目标位姿信息. Tremblay 和 To [4]针对家居用品图像，使用深度神经网络估计图像二维关键点的置信图，然后采用标准 P*n*P 算法估计物体实例的位姿.

不难发现，在物体六自由度位姿估计过程中基于深度学习的方法得到了越来越多的应用[5−6]. 然而，基于深度学习的方法往往需要大量数据来训练神经网络，数据集的丰富程度和标注质量会在很大程度上影响神经网络模型的性能. 为此，专门用于物体六自由度位姿估计的数据集应运而生. 可以将这些数据集分为两类，即合成数据集与真实数据集.

合成数据集利用图像渲染引擎对物体的三维模型进行渲染，模拟深度相机得到目标物体的彩色图与深度图[7−8]. 合成数据集由于不涉及真实场景，可以容易地获得大量数据用于模型训练与测试，但是所获得的数据往往与真实场景中采集的数据存在一定差异. 例如，在真实场景中，深度相机发射的光线照射在目标物体表面常会发生多重反射，在深度图中形成错误的深度数据；深颜色的表面会吸收光线，导致深度相机无法接收到这些区域的反射光线，从而在深度图中形成无效的深度数据. 显然，这些现象无法利用图像渲染引擎进行模拟，图像渲染引擎同样无法完全模拟真实深度图中存在的噪声. 因此，合成数据集无法替代真实数据集来训练深度神经网络模型以取得最佳的效果.

目前，基于真实 RGB-D 图像的位姿估计数据集主要包括 LineMOD (Multimodal-LINE) 数据集[9]、YCB-Video (Yale-CMU-Berkeley-video) 数据集[10]、HomebrewedDB 数据集[11]、HOPE (Household objects for pose estimation) 数据集[12]和 NOCS (Nomalized object coordinate space) 数据集[13]等. LineMOD 数据集是应用范围比较广的位姿估计数据集之一. 其中的 RGB-D 数据通过 PrimeSense Carmine RGB-D 传感器获得，共包括 15 个物体. 数据集为每个物体手动标注了大约 1000 个位姿和分割标签. 由于 LineMOD 中，物体姿态和分割真值都是人工标注的，因此数据量较小，不适用于大规模深度神经网络的训练.

YCB-Video 数据集是一个包含 92 段 RGB-D 视频序列的数据集, 每个序列从 21 个物体中随机选取、任意摆放, 然后移动相机完成拍摄. 整个数据集共包含 133827 帧图像, 每帧图像均包含 RGB 信息和深度信息. 与 LineMOD 数据集不同, YCB-Video 数据集使用了半自动化的位姿标注方式, 即首先在每个视频序列的第一帧中对目标物体进行人工标注; 然后, 利用深度视频跟踪物体的姿态, 并根据先前的人工标注信息和姿态跟踪结果初始化相机的运动轨迹和物体的位姿; 最后, 采用全局优化方法对相机轨迹和位姿信息进行调整, 以获得更加精确的标注结果. 然而, YCB-Video 数据集在估计目标位姿时需要利用由 3D 扫描设备获得的目标物体三维模型, 这导致了数据集制作成本的增加.

不难发现, 目前公开的基于真实 RGB-D 图像的位姿估计数据集具有一定的局限性. 首先, 这些数据集大都需要使用三维扫描仪获得目标物体的三维模型, 这必然会导致数据集制作成本的增加; 其次, 这些数据集都需要或多或少进行人工标注, 显然不利于大量标注数据的制作.

为此, 本文设计并实现了一种面向抓取检测的基于 RGB-D 图像的位姿估计数据集自动采集标注系统. 该系统由 RGB-D 图像数据自动采集平台与数据集自动标注算法两部分构成. 其中, RGB-D 图像数据自动采集平台使用深度相机作为 RGB-D 图像采集设备. 在自动标注算法方面, 通过检测场景中的 ArUco(Augmented reality university of cordoba) 标记获得 RGB-D 图像间的姿态变换关系, 在此基础上分割出目标点云并标注其 6D 位姿信息, 再将目标物体的三维点云进行表面重建, 并投影到图像平面, 从而获得目标物体的准确的二维分割掩码. 在点云的三角化过程中, 为了避免由于法线方向不一致而导致的三维模型变形或撕裂, 提出了一种基于最小生成树的法向传播算法. 为了验证所设计系统的性能, 在实验中使用日常生活物品, 制作了一个基于 RGB-D 图像的 6D 位姿估计数据集, 并将自动标注数据与人工标注数据进行了对比, 发现两者分割掩码重合率可以达到 98%, 并且自动标注的位姿信息能够使模型点云与场景点云的对齐率达到 100%, 充分说明了自动标注结果的准确性与可靠性.

1　RGB-D 图像数据自动采集平台

RGB-D 图像数据自动采集平台主要由机器人、工控机、深度相机和电动转台组成. 深度相机固定到机器人末端, 电动转台放置在深度相机的视野范围内, 深度相机、机器人和电动转台连接到工控机的通信接口. 工控机中安装数据集自动标注软件, 软件可以控制深度相机触发、机械臂运动和电动转台的旋转, 整个数据自动采集平台如图 1 所示.

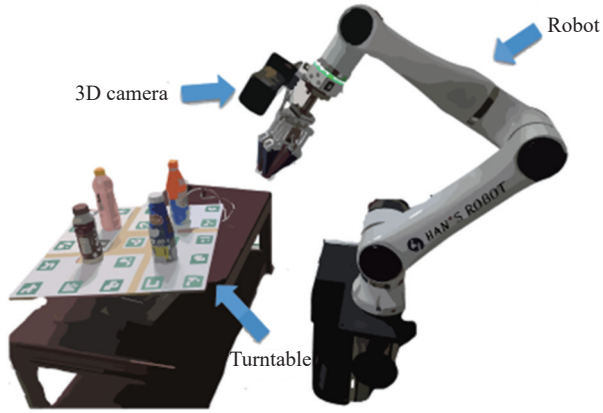

图 1　RGB-D 图像数据自动采集平台

Fig.1　RGB-D image data automatic acquisition platform

在图 1 中, 粘贴到转台上的 ArUco 标记是一种二进制方形基准标记, 可用于视觉定位. 它最早由 Garrido-Jurado 等[14] 提出, 其外观如图 2 所示.

图 2　ArUco 标记示例

Fig.2　ArUco markup example

RGB-D 图像数据的采集过程为: 电动转台每转过一定角度, 则触发深度相机拍摄一张 RGB-D 图像. 当转台转动一周后, 就获得了围绕目标 360° 视角拍摄的 RGB-D 图像序列. 随后, 控制机器人末端运动, 改变深度相机与转台参考面的夹角, 再次转动转台并进行拍摄, 得到第二组 RGB-D 图像序列. 如此获取多组图像序列, 即可用于后续数据集的自动标注过程, 如图 3 所示.

2　数据集自动标注算法

2.1　算法总体流程

采集到 RGB-D 图像序列后, 对 RGB-D 图像数据进行自动标注的算法流程如图 4 所示. 假设共获得了 n 帧 RGB-D 图像, 每帧图像中包含 b 个 ArUco 标记, q 个目标物体. 由 RGB-D 图像中的深度信息得到对应 n 帧图像的点云集合, 进而在每一帧点

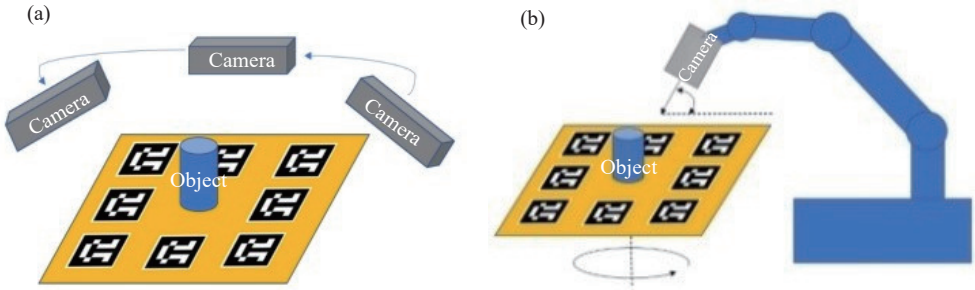

图 3　数据集采集示意图. (a) 多视角拍摄物体; (b) 旋转相机和地面的夹角

Fig.3　Schematic of data collection: (a) shooting objects from multiple angles; (b) rotating the angle between the camera and the ground

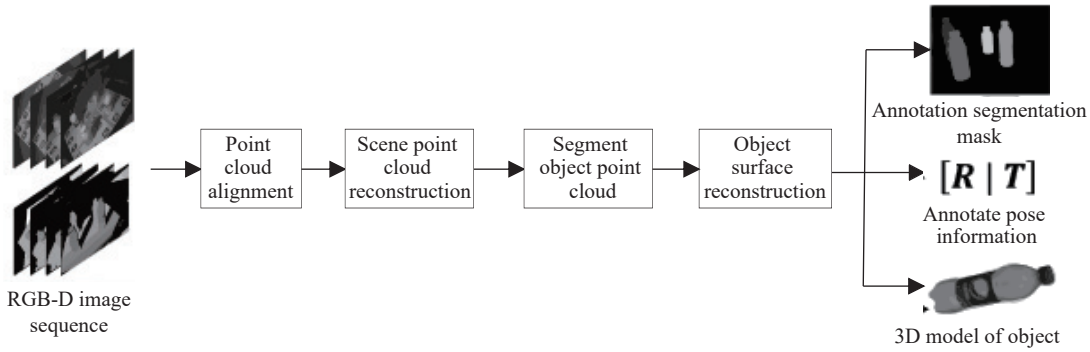

图 4　数据集自动标注流程图

Fig.4　Dataset automatic labeling flowchart

云中找出 ArUco 标记的角点的三维坐标. 以第 k 帧 RGB-D 图像对应的点云为例, 其中的 b 个 ArUco 标记的角点集合表示为 $\boldsymbol{A}_k = \{a_{ki} \mid i = 1, 2, \cdots, 4b\}$. 利用不同帧点云中 ArUco 标记的对应关系, 通过点云配准能够计算出任意两帧点云间的位姿变换关系. 以第一帧点云作为全局坐标系, 那么对于由 n 帧点云构成的点云集合来说, 就能够形成一个位姿变换集合 $\boldsymbol{T}_1 = \{T_{1k} \mid k = 1, 2, \cdots, n\}$, 其中, 第 k 帧点云经过 T_{1k} 变换后就与第 1 帧点云对齐.

采用随机采样一致性 (RANSAC) 算法对 $\boldsymbol{A}_k$ 进行平面拟合, 去除转台参考面上的点. RANSAC 算法是一种鲁棒估计模型参数的迭代方法, 特别适用于存在大量异常值的数据集, 在计算机视觉、计算机图形学、机器学习等领域得到了广泛应用. 以平面拟合为例, 其核心步骤是: 首先从数据点中选择最少数量的观测点以估计平面参数, 然后计算所有数据点到估计平面的距离, 将距离小于阈值的点标记为内点, 然后重复上述步骤, 经过多次随机采样和迭代后, 选择具有最多内点的模型作为最佳拟合平面, 在此基础上, 使用全部数据点对最佳平面的参数进行优化, 从而获得最终的平面拟合结果.

使用 T_{1k} 变换第 k 帧点云, 利用投票机制将变换结果与第一帧点云进行融合并采用点云平滑方法去除噪声, 从而得到场景的完整重建点云 $\boldsymbol{P}_{\mathrm{r}}$. 进一步使用欧氏聚类分割算法对重建点云 $\boldsymbol{P}_{\mathrm{r}}$ 进行分割, 获得场景中 q 个物体的点云, 记为 $\boldsymbol{P}_{\mathrm{O}} = \{P_{\mathrm{O}i} \mid i = 1, 2, \cdots, q\}$.

对 $\boldsymbol{P}_{\mathrm{O}}$ 中的每个物体点云 $P_{\mathrm{O}i}$ 计算法向, 并利用基于最小生成树的法向传播方法对各个物体点云的法向量进行一致化处理. 使用泊松表面重建算法对 $P_{\mathrm{O}i}$ 进行曲面重建, 从而得到每个物体的三维曲面模型 M_i, i=1,2, ⋯, q.

将 M_i 中的三角面逐个投影到图像平面, 获得目标物体的分割掩码. 由 M_i 计算出相对于全局坐标系的方向包围盒及其位姿变换关系 T_{M_i}. 由前面已经获得了第一帧点云与第 k 帧点云的位姿变换关系 T_{1k}, 计算出第 k 帧点云中物体模型 M_i 的 6D 位姿信息 T_k^{po}. 接下来, 将对算法流程中的重要步骤进行更加详细的阐述.

2.2　点云配准与对齐

在数据集自动标注算法中, 通过利用不同帧点云数据中 ArUco 标记的角点对应关系得到了不同帧点云中所有物体间的粗配准结果, 即如果将第 i、j 帧点云表示为 $\boldsymbol{P}_i$ 和 $\boldsymbol{P}_j$, $\boldsymbol{P}_i$ 与 $\boldsymbol{P}_j$ 间的粗略位姿变换关系可以由 ArUco 标记的三维角点集合 $\boldsymbol{A}_i$ 与 $\boldsymbol{A}_j$ 的对应关系得到. 在此基础上, 使用 ICP (Iterative Closest Point) 算法[15] 来获得 $\boldsymbol{P}_i$ 与 $\boldsymbol{P}_j$ 的准

确位姿关系 $\boldsymbol{T}_{ij}$, 即可使 $\boldsymbol{T}_{ij}\boldsymbol{P}_j$ 与 $\boldsymbol{P}_i$ 对齐. ICP 算法是一种经典的点云配准算法, 用于将两个或多个点云数据进行对齐, 以便它们在同一坐标系中能够精确对应或重叠. ICP 算法需要计算初始点云中所有点与目标点云的距离, 保证这些点和目标点云的最近点相互对应, 基于最小二乘法对残差平方构成的目标函数进行最小化处理, 反复迭代, 直到均方误差小于设定的阈值. 由于在数据集自动标注算法中, 将第一帧点云所处坐标系作为全局坐标系, 因此需要将后续帧点云与第一帧点云对齐, 这里采用了 Choi 等[16] 提出的位姿图优化方法实现点云的多向配准, 从而在全局坐标系中实现不同帧点云的对齐.

在 Choi 的方法中, 构建位姿图时, 将第 i 帧点云 $\boldsymbol{P}_i$ 作为位姿图的第 i 个节点, 将相邻帧点云的变换关系 $\boldsymbol{T}_{ij}$(i=1, 2, ⋯, n−1, j=i+1)作为位姿图的里程计边, 将不相邻点云的变换矩阵 $\boldsymbol{T}_{ij}$(i=1, 2, ⋯, n−1, j≠i+1)作为位姿图的回环边. 采用 Choi 提出的鲁棒性优化算法对位姿图进行优化. 在优化后的位姿图中, 将每一帧点云相对于第一帧点云的变换矩阵取出, 构成位姿变换集合 $\boldsymbol{T}_1 = \{T_{1k} \mid k = 1, 2, \cdots, n\}$.

2.3　获取完整场景点云

采用 RANSAC 算法对第 k 帧点云 $\boldsymbol{P}_k$ 中的 $\boldsymbol{A}_k$ 进行平面拟合, 将 $\boldsymbol{P}_k$ 中参考平面上的点去除, 得到点云 P_{sk}. 在此基础上, 通过位姿变换集合 $\boldsymbol{T}_1$ 对点云 P_{sk} 进行位姿变换使其与第 1 帧点云对齐, 通过投票机制来融合经过位姿变换的各帧点云, 得到完整的重建点云 $\boldsymbol{P}_r$, 如图 5 所示.

图 5　场景点云 $\boldsymbol{P}_r$

Fig.5　Scene point cloud $\boldsymbol{P}_r$

不难发现, $\boldsymbol{P}_r$ 中不可避免地存在着空洞等点云局部不完整情况. 故采用滑动最小二乘算法[17] 对场景点云 $\boldsymbol{P}_r$ 进行重采样, 通过对局部点云数据进行高阶多项式插值来重建 $\boldsymbol{P}_r$ 中的表面缺失部分. 这样还可以解决由于同一区域存在多个扫描点而出现的“双墙”现象, 从而使 $\boldsymbol{P}_r$ 更加光滑.

上述场景点云构建过程总结为如下伪代码算法.

Input: Point cloud set $\boldsymbol{P}_s = \{P_{si}, i = 1, 2, \cdots, n\}$, pose transformation set $\boldsymbol{T}_1 = \{T_{1i} \mid i = 1, 2, \cdots, n\}$, distance threshold D_{new}, voting distance threshold D_{vote}

Output: Reconstructed point cloud $\boldsymbol{P}_r$

$\boldsymbol{V}$←Zero vector with the same length as P_{s1}

For i←2 To n Do

　P_{1i}←$T_{1i}\cdot P_{si}$

　　For j←1 To length of P_{1i} Do

　　　k←Index of the nearest point from P_{1i} to P_{s1}

　　　d←Distance of the nearest point from P_{1i} to P_{s1}

　　　If d<D_{vote} Then

　　　　$\boldsymbol{V}[k]$←$\boldsymbol{V}[k]$+1

　　　Else If d>D_{new} Then

　　　　Add $P_{1i}[j]$ to the end of P_{s1}

　　　　Add 0 to the end of $\boldsymbol{V}$

　　　End If

　　End For

　End For

$\boldsymbol{k}_{in}$←Index set of elements in $\boldsymbol{V}$ greater than 4

$\boldsymbol{P}_r$←$P_{s1}[\boldsymbol{k}_{in}]$

$\boldsymbol{P}_r$←Sliding least squares processing for $\boldsymbol{P}_r$

Return $\boldsymbol{P}_r$

2.4　目标物体表面重建

在本文提出的数据集自动采集标注系统中, 拍摄场景中可以放置多个目标物体, 因此在进行物体表面重建之前, 需要对场景点云 $\boldsymbol{P}_r$ 进行分割来得到单个物体点云 P_{Oi}(i=1, 2, ⋯, q). 本文采用了欧氏聚类点云分割算法[18], 其步骤简单描述为:

Step1: 初始化两个空点集, 一个作为备选点的点集, 一个作为聚类的结果点集;

Step2: 在点云中随机选择一个点作为种子,按照给定的半径对最临近的点进行搜索;

Step3: 如果最近点小于给定欧式距离的阈值, 将该点纳入备选点和相应的聚类结果点集. 如果最近点大于给定半径, 则对当前点的搜索结束;

Step4: 重复 Step3, 直至对当前点搜索结束;

Step5: 从原点云中删除聚类结果点集, 重复步骤 Step2 至 Step4, 直至原点云中所有点都被剔除.

获得了单个物体点云 P_{Oi} 后, 采用泊松表面重建算法[19] 来进行物体表面重建. 泊松表面重建算法能够将带有法向量属性的三维点数据转换为三

角网格模型, 从而给出一个平滑的物体表面估计. 在泊松表面重建算法中, 需要准确且具有一致性的法向作为输入数据[20], 否则表面重建算法在处理法向不一致的点云时会出现表面撕裂、变形等情况. 通常曲面上任意一点的法向可通过主成分分析法 (Principal component analysis, PCA) 进行估计, 根据该点及其邻近点的位置信息反映的曲面局部形状的逼近进行法向估计. 然而, 在使用 PCA 对点云进行法向估计时, 较小的邻域半径虽然会使估计出的法向更加精确, 但也会使相邻点的法向变化过大, 容易导致法向传播失败; 而较大的半径尽管会提高法向传播的成功概率, 但不精确的法向会使表面重建的结果也不精确. 而且, 经过 PCA 方法估计所得到的法向不保证具备一致性, 即任意一点的法向与其邻近点的法向可能相反, 如图 6 所示.

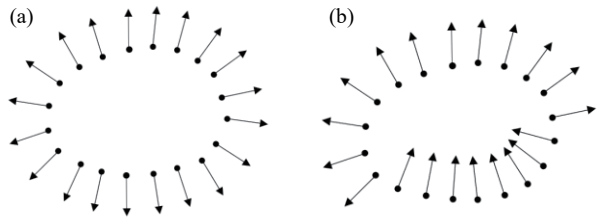

图 6　点云法向一致性示意图. (a) 具有法向一致性的点云; (b) 不具有法向一致性的点云

Fig.6　Schematic of point cloud normal direction consistency: (a) point clouds with normal consistency; (b) point clouds without normal consistency

为此, 本文提出了一种基于最小生成树的法向传播方法来对点云法向估计进行一致化. 该方法首先用大半径计算出的法向进行法向传播, 然后用小半径计算出的法向进行精细化, 从而在保证法向传播一致性的同时提高表面重建的精度. 采用本文方法在物体点云上构建的最小生成树如图 7 所示, 基于最小生成树的法向传播方法具体步骤是:

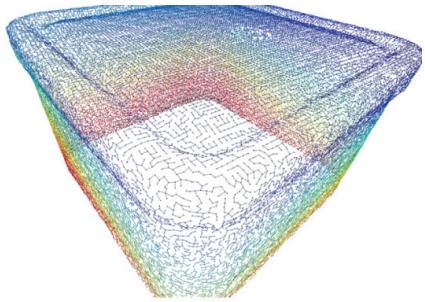

图 7　最小生成树结果图

Fig.7　Minimum spanning tree result graph

Step1: 在主成分分析方法中, 使用较大的邻域半径对点云进行法向估计, 得到初始法向量.

Step2: 选择曲率最小的点作为最小生成树的根节点, 并将其标记.

Step3: 计算所有未标记点到树的距离, 找到距离树最近的点 $p_{\min}$, 并且找到距离 $p_{\min}$ 最近的树节点.

Step4: 将 $p_{\min}$ 作为新的节点加入树中, 并连接到距离它最近的树节点上, 从而生成一个新的边, 并标记 $p_{\min}$.

Step5: 重复 Step3 和 Step4, 直到点云中所有的点都被加入到最小生成树中.

Step6: 遍历最小生成树的所有边, 计算其两端连接节点的法向量间的夹角. 如果夹角大于 90°则反转后加入节点的法向量, 从而获得方向一致的粗略法向量.

Step7: 再次使用主成分分析方法并采用较小的邻域半径对点云进行法向估计, 得到精确法向量.

Step8: 计算每个点的精确法向量与粗略法向量间的夹角, 如果夹角大于 90°则将精确法向量反向, 最终得到方向一致的精确法向量.

图 7 是以一个包装盒的点云数据为基础, 应用本文提出的最小生成树法向传播算法, 得到的最小生成树示意图.

得到方向一致的法向量估计结果后, 采用泊松表面重建算法对 P_{Oi} 进行表面重建, 可以得到三角化后的三维曲面模型 $M_i(i=1,2,\cdots,q)$.

2.5　获取分割掩码

利用改进的泊松算法得到三维曲面模型 M_i, 可以将 M_i 中的三角面投影到图像平面上, 从而获得目标物体的分割掩码. 如果将相机的内参数表示为 f_x、f_y、c_x、c_y, 三角面的顶点三维坐标表示为 (X, Y, Z), 那么其所对应的二维图像坐标 (u, v), 就可以由式 (1) 得到:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{Z}\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{1}$$

将 M_i 中各个三角面的顶点分别投影到图像平面后, 对投影点组成的三角形进行填充, 即可得到目标物体在图像平面中的分割掩码. 图 8 展示了深度相机拍摄到的图像以及由算法自动获得的目标物体分割掩码.

2.6　标注目标物体 6D 位姿

给定 M_i, 通过主成分分析法确定其方向包围

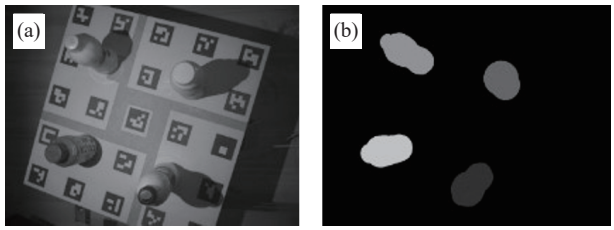

图 8　目标物体分割效果. (a) 原图像; (b) 物体对应的分割掩码

Fig.8　Target object segmentation effect: (a) original image; (b) segmentation mask corresponding to the objects

盒以及方向包围盒相对于全局坐标系的位姿变换关系 T_{M_i}. 因此, 通过 $T_{M_i}^{-1}$ 对模型 M_i 进行姿态变换就可以将其与全局坐标系对齐, 从而得到最终输出的目标物体三维模型.

根据位姿变换集合 $\boldsymbol{T}_1=\{T_{1k}\mid k=1, 2, \cdots, n\}$ 以及模型 M_i 与全局坐标系的变换关系 T_{M_i}, 通过式 (2) 计算得到第 k 帧点云中目标物体的位姿变换关系, 即需要标注的目标物体 6D 位姿信息 $T_k^{M_i}$:

$$T_k^{M_i}=T_{1k}^{-1}T_{M_i} \tag{2}$$

3　实验结果与分析

实验过程中, 搭建了文中提出的基于 RGB-D 图像的 6D 位姿数据集自动采集标注系统, 并利用日常生活物品制作了一个 6D 位姿估计数据集, 标注了这些物品的三维模型、图像分割掩码以及 6D 位姿信息. 为了验证标注的分割掩码数据的准确性, 将分割掩码标注结果与其手工标注结果进行了对比, 并使用实例分割模型 Mask RCNN(Mask regional-based convolutional neural network) 进行进一步验证. 为了验证位姿标注的准确性, 使用由标注得到的位姿参数对目标物体点云进行位姿变换, 将变换后的点云与拍摄的场景点云进行对齐, 通过点云对齐效果反映位姿参数标注的准确性, 进而使用类别级位姿估计网络对标注位姿的准确性进行了进一步验证.

3.1　自制数据集

数据集中共涉及 84 个物体, 其中易拉罐 28 个、饮料瓶 28 个、包装盒 28 个. 这些物体被放置于 21 个场景中, 每个场景包含 4 个物体. 在每个场景中拍摄 400 张 RGB-D 图像, 一共得到 8400 张 RGB-D 图像, 如图 9 所示. 对这 8400 张图像进行划分, 6000 张图像作为训练集, 涉及 15 个场景, 60 个物体, 另外 2400 张图像作为测试集, 涉及 6 个场景, 24 个物体. 由数据集自动采集标注系统获得的目标物体三维模型如图 10 所示.

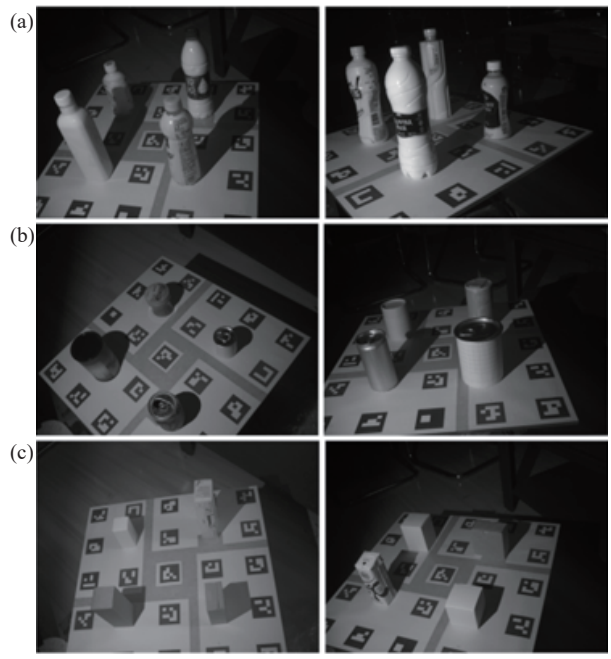

图 9　自制数据集中不同物体的 RGB-D 图像. (a) 饮料瓶; (b) 易拉罐; (c) 包装盒

Fig.9　RGB-D images of different objects in a homemade dataset: (a) bottle; (b) can; (c) box

图 10　由数据集自动标注系统获得的三维模型

Fig.10　3D model obtained from the automatic labeling system of the dataset

3.2　分割掩码标注准确度实验

本节对由数据集自动采集标注系统获得的图像分割掩码进行测试. 以真实分割掩码与自动标注掩码的重合率, 即交并比 (IoU) 作为主要评价指标, 来评价掩码标注的准确性[21−23]. 其中, 真实分割掩码来自于对自制数据集中随机选出的 100 张图像的人工标注结果.

表 1 展示了 IoU 指标的计算结果, 可见自动标注的图像分割掩码与人工标注的图像分割掩码的重合率达到 98% 以上, 说明自动标注结果与人工标注的结果是非常吻合的.

在图 11 中, 将自动标注的分割掩码和人工标注的分割掩码分别用红色和黄色的轮廓线绘制在

表 1　IoU 计算结果

Table 1　IoU calculation results　%

Bottle	Can	Box	Average
98.2	97.3	99.1	98.3

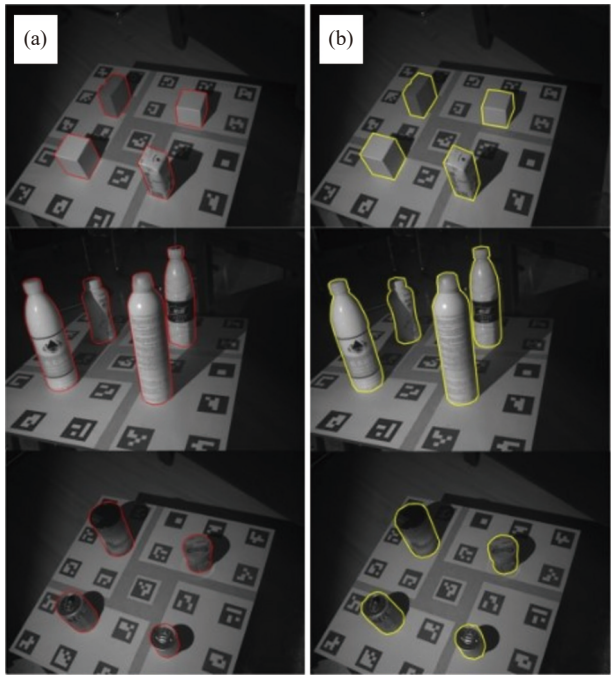

图 11　分割标注可视化. (a) 手动标注; (b) 自动标注

Fig.11　Segmentation annotation visualization: (a) manual annotation; (b) automatic annotation

了左右两侧的图像当中. 可见, 无论自动标注的分割掩码还是人工标注的分割掩码都与图像中目标物体的真实位置吻合, 不存在分割掩码偏离目标物体的情况. 而且, 自动标注的分割掩码与人工标注的分割掩码非常接近.

3.3　位姿标注准确度实验

在验证位姿标注的准确性时, 借助标注的位姿信息, 进行场景点云与模型点云的对齐, 通过点云对齐的效果来反映标注位姿的准确性, 即使用由标注得到的位姿参数对目标物体点云进行位姿变换, 将变换后的点云与拍摄的场景点云进行对比, 分别计算两片点云的均方误差 (MSE) 和内点率 (Inlier rate), 通过这两个指标来判断所标注的位姿参数是否准确[24−26]. 实施时, 首先使用自动标注得出的变换矩阵 $T_k^{M_i}$ 将模型点云 $P_{\mathrm{O}i}$ 对齐到第 k 个场景中, 得到

$$\boldsymbol{P}'_k = T_k^{M_i} P_{\mathrm{O}i} \tag{3}$$

然后, 分别对点云 $\boldsymbol{P}'_k$ 和第 k 帧场景点云 $\boldsymbol{P}_k$ 进行体素下采样, 以使点云能够尽可能保留足够的信息同时又可以降低在对应点云中搜索最近点对的计算量. 在 $\boldsymbol{P}'_k$ 的所有点中搜索 $\boldsymbol{P}_k$ 的最近点, 组成点对. 由于受到拍摄视角的限制, 场景点云通常只包含模型点云的一部分, 即模型点云中会存在无法与场景点云匹配的三维点. 为此, 筛选出点对中两点距离小于 5 mm 的点对记入集合 $\boldsymbol{Q}=\{(h'_j, h_j) \mid h'_j \in \boldsymbol{P}'_k, h_j \in \boldsymbol{P}_k, j = 1, 2, \cdots, w\}$, w 表示 $\boldsymbol{P}'_k$ 中距离 $\boldsymbol{P}_k$ 小于 5 mm 的点数, 参与后续对齐指标的计算. 于是, 可以按照式 (4) 计算集合 $\boldsymbol{Q}$ 中点对的均方误差指标:

$$\mathrm{MSE} = \frac{1}{w}\sum_{j=1}^{w} \mathrm{dist}\left(h'_j, h_j\right)^2 \tag{4}$$

进一步, 内点率指标 inlier_rate 的计算公式如下:

$$\mathrm{inlier\_rate} = \frac{w}{\#\boldsymbol{P}'_k} \times 100\% \tag{5}$$

其中, $\#\boldsymbol{P}'_k$ 表示点云 $\boldsymbol{P}'_k$ 中点的数量.

考虑到第 k 帧点云 $\boldsymbol{P}_k$ 由于受到拍摄视角的限制, 往往只包含一部分模型点云, 因此认为如果 MSE 指标小于 5 倍体素大小的平方, 并且内点率大于 10%, 则 $\boldsymbol{P}'_k$ 与 $\boldsymbol{P}_k$ 是对齐的. 实验中, 对自制数据集中所有模型点云按照自动标注出的位姿参数进行位姿变换, 并记录经过位姿变换的点云与场景点云的 MSE 指标、内点率指标和对齐率(表 2). 可以看出, 自动标注出的位姿参数非常准确. 这也可以从图 12 所示的定性实验中看出. 图中红色的为经过位姿参数变换的模型点云, 黑白色的为场景点云. 同时, 也能够看出由自动标注算法得出的目标物体的三维模型具有较高的精度.

表 2　自制数据集的 MSE、内点率和对齐率

Table 2　MSE, inlier_rate, and alignment rate of the self-made datasets

Object	MSE/mm^2	Inlier_rate/%	Alignment/%
Bottle	5.25	39.5	100
Can	7.48	32.6	100
Box	4.12	56.4	100

3.4　基于 GPV 模型的测试

GPV (Geometry-guided point-wise voting) 模型[27]是一个类别级位姿估计网络, 它利用 RGB-D 图像数据来估计物体的位姿信息和尺寸信息. 采用 GPV 模型对自制数据集进行测试时, 将自制数据集中的图像分割掩码重投影到由 RGB-D 图像获得的点云数据当中, 分割出目标物体点云作为 GPV 模型的输入. GPV 模型的输出为目标物体的尺寸信息和位姿信息, 其中位姿信息用来与自制数据集

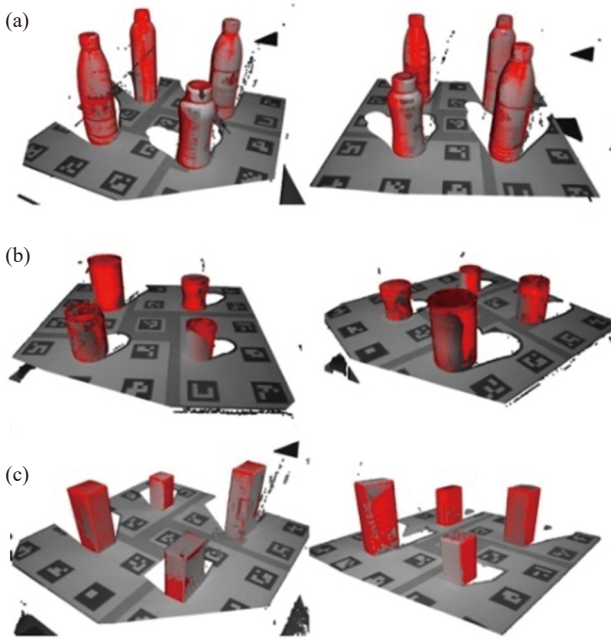

图 12　位姿标注的可视化实验结果. (a) 饮料瓶; (b) 易拉罐; (c) 包装盒

Fig.12　Visualization of pose labeling experimental results: (a) bottle; (b) can; (c) box

中标注的位姿信息进行比较, 从而验证标注数据的有效性.

实验中采用 3DIoU 和 *d*°*e* cm 来评价 GPV 模型输出位姿与数据集标注位姿的差异. 其中, 3DIoU 指标主要考察 GPV 模型输出的目标物体包围框与数据集标注的目标物体包围框的重合程度, 即如果两个目标物体包围框交集的体积在两个包围框并集体积中的占比超过了给定阈值, 就说明数据集标注结果与 GPV 模型的输出结果相一致. *d*°*e* cm 指标则是通过判断 GPV 模型输出的 6D 位姿估计结果与数据集标注的位姿信息间的旋转误差是否小于 *d*°, 平移误差是否小于 *e* cm 来判断两个位姿数据是否相吻合.

使用自制数据集训练 GPV 模型, 得到的 6D 位姿估计结果如表 3 所示. 可以看到, 如果将 3DIoU 指标的阈值设定为 75%, 则 GPV 模型输出的目标物体包围框与数据集标注的物体包围框的一致率可以达到 95% 左右, 如果将 3DIoU 指标阈值降低为 50%, 则包围框的一致率可以达到 98% 左右, 这说明绝大多数情况下, 数据集的标注结果与 GPV 模型的输出结果都是相一致的. 从旋转误差与平移误差的角度看, GPV 模型的估计位姿与数据集标注位姿的差异在 10°5 cm 的范围内, 再次验证了在自制数据中标注的位姿参数精度较高, 可以进一步应用于机器人抓取等实际任务当中.

3.5　基于 Mask RCNN 模型的测试

Mask RCNN 模型[28] 是实例分割模型, 它通过在 Faster R-CNN (Faster region-based convolutional neural network) 模型中添加一个与目标检测框回归并行的, 用于预测目标掩码的分支来扩展 Faster R-CNN 模型, 实现目标检测与目标分割同时进行. 实验使用不同置信度阈值下的平均精度 (AP) 指标作为目标检测和实例分割的评价指标, 其计算过程是: 首先计算神经网络预测目标与真实目标的 IoU, 当 IoU 指标大于给定置信度阈值时认为预测正确; 然后, 统计正确预测数量 TP, 错误预测数量 FP, 漏检测数量 FN, 由此可以计算出准确率 Pr 和召回率 Re, 计算公式如下:

$$\begin{cases} \mathrm{Pr} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ \mathrm{Re} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \end{cases} \tag{6}$$

记录不同置信度阈值下的 Pr 值和 Re 值, 并绘制为 PR 曲线图, 则 PR 曲线下的面积就是 IoU 大于 *t* 时的 AP 值, 记作 AP_t. 表 4 和表 5 分别给出了以 AP_{50}、AP_{75} 和 AP_{95} 作为评价指标的 Mask RCNN 网络在自制数据集上的实例分割结果和目标检测结果.

从表 4 和表 5 中不难看出, 经过自制数据集训练、测试的 Mask RCNN 网络可以具有高于 90% 的目标检测与实例分割精度. 说明由数据集自动采集标注系统获得的数据集具有非常高的数据标注质量, 这也反映在了图 13 所示的 Mask RCNN 网络的预测结果当中.

表 3　基于 GPV 模型的测试结果

Table 3　Test results based on the GPV model

Object	3DIoU75/%	3DIoU50/%	Pose matching accuracy/%			
			5°2 cm	5°5 cm	10°2 cm	10°5 cm
Bottle	95.5	98.6	78.7	82.4	93.6	99.6
Can	94.0	97.9	84.7	89.9	100	100
Box	95.2	98.1	85.2	90.6	100	100

表 4　实例分割实验结果

Table 4　Instance segmentation experimental results

Object	AP_{50}/%	AP_{75}/%	AP_{95}/%
Bottle	97.5	96.5	91.8
Can	98.2	96.8	92.5
Box	98.6	96.5	92.9

表 5　目标检测实验结果

Table 5　Target detection experimental results

Object	AP_{50}/%	AP_{75}/%	AP_{95}/%
Bottle	97.5	96.5	89.2
Can	98.2	96.8	90.9
Box	98.6	95.2	90.4

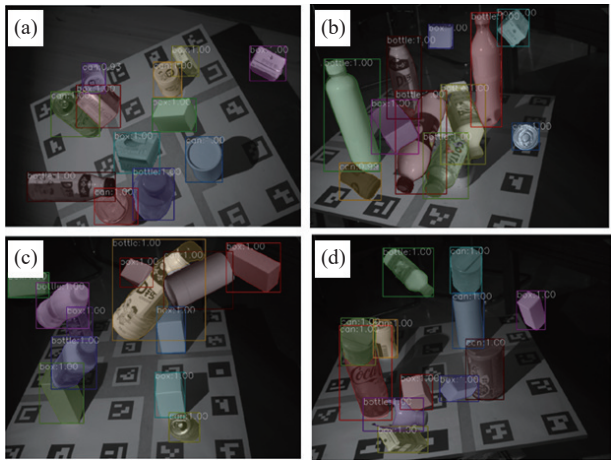

图 13　Mask RCNN 预测结果可视化. (a) 预测结果 1; (b) 预测结果 2; (c) 预测结果 3; (d) 预测结果 4

Fig.13　Visualization of Mask RCNN prediction results: (a) prediction result 1; (b) prediction result 2; (c) prediction result 3; (d) prediction result 4

4　结论

本文设计并实现了一种面向抓取检测的基于 RGB-D 图像的位姿估计数据集自动采集标注系统. 该系统由 RGB-D 图像数据自动采集平台与数据集自动标注算法两部分构成. 其不需要使用价格昂贵的三维激光扫描仪来获取目标物体的三维模型, 仅使用深度相机采集 RGB-D 图像序列, 并通过对该 RGB-D 图像序列的分析, 即可自动得出目标物体的三维模型、位姿参数, 以及图像分割掩码. 此外, 在设计自动标注算法时, 为了避免由于法线方向不一致而导致的三维表面重建结果出现变形或撕裂的情况, 本文还提出了一种基于最小生成树的法向传播算法. 实验中使用本文设计的数据集自动采集标注系统, 以日常生活中常见的易拉罐、饮料瓶、包装盒作为目标对象, 制作了一个具有 8400 张 RGB-D 图像的数据集, 分别对该数据集中分割掩码、位姿参数的标注质量进行了评估, 发现自动标注结果能够媲美人工标注结果, 实验还将文中自制数据集应用于 GPV 模型和 Mask RCNN 模型的训练和测试过程中, 发现由此获得的网络模型具有非常高的预测精度, 说明由本文设计的系统能够实现数据集的高质量采集与标注, 能够为面向抓取检测的深度学习模型的研究、应用提供良好的数据基础.

参　考　文　献

[1] Xu J, Liu N, Li D P, et al. A grasping pose detection algorithm for industrial parts based on grasping clusters and collision voxels. *Robotics*, 2022, 44(2): 153
(徐进, 柳宁, 李德平, 等. 一种基于抓取簇和碰撞体素的工业零件抓取姿态检测算法. 机器人, 2022, 44(2): 153)

[2] Xia H Y, Suo S F, Wang Y, et al. Object grasping detection algorithm based on keypoint RCNN improved model. *Chin J Sci Instrum*, 2021, 42(4): 236
(夏浩宇, 索双富, 王洋, 等. 基于 Keypoint RCNN 改进模型的物体抓取检测算法. 仪器仪表学报, 2021, 42(4): 236)

[3] Mao L B, Shi J L, Zhou Z Q, et al. Robot grabbing method based on single-view key point voting. *Comput Integr Manuf Syst*, 2023, 29(11): 3572
(茅凌波, 史金龙, 周志强, 等. 基于单视图关键点投票的机器人抓取方法. 计算机集成制造系统, 2023, 29(11): 3572)

[4] Tremblay J, To T, Sundaralingam B, et al. Deep object pose estimation for semantic robotic grasping of household objects //*2nd Conference on Robot Learning* (*CoRL*). Zurich, 2018: 306

[5] Wan Y D. *Research on Stable Grasping Pose Estimation Based on Deep Learning* [Dissertation]. Harbin: Harbin Institute of Technology, 2022
(万延多. 基于深度学习的稳定抓取位姿估计研究[学位论文]. 哈尔滨: 哈尔滨工业大学, 2022)

[6] Li S F, Shi Z L, Zhuang C G. Deep learning-based 6D object pose estimation method from point clouds. *Comput Eng*, 2021, 47(8): 216
(李少飞, 史泽林, 庄春刚. 基于深度学习的物体点云六维位姿估计方法. 计算机工程, 2021, 47(8): 216)

[7] He X, Li J C, Jin L, et al. A synthetic dataset and performance evaluation for 3D template tracking. *Chin J Comput*, 2022, 45(3): 585
(何弦, 李佳宸, 金立, 等. 三维模板跟踪的基准合成数据集构建及算法评估. 计算机学报, 2022, 45(3): 585)

[8] Wu F D. *Research on Image Enhancement Algorithms for Uneven Illumination and Low Illumination at Night Based on Deep Learning* [Dissertation]. Hangzhou: Zhejiang Gongshang University, 2022

（吴凡丁. 基于深度学习的非均匀光照和夜间低照度图像增强算法研究[学位论文]. 杭州：浙江工商大学, 2022）

[9] Hinterstoisser S, Lepetit V, Ilic S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes // *Computer Vision–ACCV* 2012: 11*th Asian Conference on Computer Vision*. Daejeon, 2012: 548

[10] Li H Y, Lin J H, Jia K. DCL-net: Deep correspondence learning network for 6D pose estimation // *European Conference on Computer Vision*. Tel Aviv, 2022: 369

[11] Kaskman R, Zakharov S, Shugurov I, et al. HomebrewedDB: RGB-D dataset for 6D pose estimation of 3D objects // *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. Seoul, 2019: 15

[12] Tyree S, Tremblay J, To T, et al. 6-DoF pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark // *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Kyoto, 2022: 13081

[13] Wang H, Sridhar S, Huang J W, et al. Normalized object coordinate space for category-level 6D object pose and size estimation // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 2642

[14] Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas F J, et al. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit*, 2014, 47(6): 2280

[15] Li W T, Wang Y C, Shao Y, et al. TrackPuzzle: Efficient registration of unlabeled PDR trajectories for learning indoor route graph. *Future Gener Comput Syst*, 2023, 149: 171

[16] Choi S, Zhou Q Y, Koltun V. Robust reconstruction of indoor scenes // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, 2015: 5556

[17] Niedzwiedzki J, Lipinski P, Podsedkowski L. IDTMM: Incremental direct triangle mesh mapping. *IEEE Robot Autom Lett*, 2023, 8(9): 5416

[18] Li J L, Saydam S, Xu Y Y, et al. Class-aware tiny object recognition over large-scale 3D point clouds. *Neurocomputing*, 2023, 529: 166

[19] Han J L, Liu Y Z, Rong M, et al. FloorUSG: Indoor floorplan reconstruction by unifying 2D semantics and 3D geometry. *ISPRS J Photogramm Remote Sens*, 2023, 196: 490

[20] Kazhdan M, Hoppe H. Screened poisson surface reconstruction. *ACM Trans Graph*, 2013, 32(3): 1

[21] Ma X L, Xue H R. Point cloud normal vector estimation method based on iterative least squares. *Comput Simul*, 2023, 40(7): 363
（马学磊, 薛河儒. 基于迭代最小二乘的点云法向量估计方法. 计算机仿真, 2023, 40(7)：363）

[22] Cheng J K. *Remote Sensing Image Feature Classification Based on Deep Learning* [Dissertation]. Xi'an: Xidian University, 2022
（成金凯. 基于深度学习的遥感图像地物分类[学位论文]. 西安：西安电子科技大学, 2022）

[23] Wang Q M. *Research on Improved Methods of Intersection and Union Prediction in Target Detection* [Dissertation]. Wuhan: Huazhong University of Science and Technology, 2022
（王启萌. 目标检测中交并比预测的改进方法研究[学位论文]. 武汉：华中科技大学, 2022）

[24] Duan D Y, Qiu W G, Cheng Y J, et al. Reconstruction of shield tunnel lining using point cloud. *Autom Constr*, 2021: 103860

[25] Lyu Z, Kong Z, Xu X, et al. A conditional point diffusion-refinement paradigm for 3D point cloud completion // *International Conference on Learning Representations*. Vienna, 2021

[26] Li W, Wang C, Lin C, et al. Inlier extraction for point cloud registration via supervoxel guidance and game theory optimization. *ISPRS J Photogramm Remote Sens*, 2020, 163: 284

[27] Di Y, Zhang R D, Lou Z Q, et al. GPV-pose: Category-level object pose estimation *via* geometry-guided point-wise voting // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, 2022: 6781

[28] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN // *Proceedings of the IEEE International Conference on Computer Vision*. Venice, 2017: 2961