Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1319-1330
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116135

Hans汉斯

基于深度学习的6D位姿估计方法最新研究

杨　涵，陶佳林，怀珍豪

上海理工大学，上海

收稿日期：2022年11月7日；录用日期：2022年12月7日；发布日期：2022年12月28日

摘　要

6D位姿估计技术在工业机器人、虚拟现实和餐饮服务等领域已经成为关键性的技术。它的发展逐渐由离线方式发展到端到端的方式。这主要因为随着深度学习技术的发展，促进了6D位姿估计理论不断地完善，进一步导致了对该方面的技术总结和时效性不强。因此，本文基于最新的6D位姿估计技术的方法进行调研。在本文中，我们首先介绍了6D位姿估计技术的评估指标、数据集和6D位姿估计方法。其中6D位姿估计方法以数据的输入方式进行划分，我们将方法划分为基于D、RGB和RGBD的方法。最后，我们就这些各类方法给出了我们的启发和未来可能的研究方向，希望给予相关人员一定的帮助。

关键词

深度学习，位姿估计，RGB，RGBD

Recent Research on 6D Pose Estimation Method Based on Deep Learning

Han Yang, Jialin Tao, Zhenhao Huai

University of Shanghai for Science and Technology, Shanghai

Received: Nov. 7th, 2022; accepted: Dec. 7th, 2022; published: Dec. 28th, 2022

Abstract

6D pose estimation technology has become critical in industrial robotics, virtual reality, and food service areas. It has gradually evolved from an offline approach to an end-to-end approach. This is mainly due to the fact that with the development of deep learning techniques, it has facilitated the continuous improvement of 6D pose estimation theory, which has further led to a poor summary and timeliness of the techniques in this area. Therefore, this paper investigates the approach based on the latest 6D pose estimation techniques. In this paper, we first introduce the evaluation metrics, datasets, and 6D pose estimation methods for 6D pose estimation techniques.

文章引用: 杨涵, 陶佳林, 怀珍豪. 基于深度学习的 6D 位姿估计方法最新研究[J]. 软件工程与应用, 2022, 11(6): 1319-1330. DOI: 10.12677/sea.2022.116135

Where the 6D pose estimation methods are classified by the input method of the data, we classify the methods into D-based, RGB, and RGBD-based methods. Finally, we give our inspiration and possible future research directions on these various methods, which we hope can help related people.

Keywords

Deep Learning, Pose Estimation, RGB, RGBD

1. 引言

随着计算机视觉的发展，6D 位姿估计理论逐渐被完善。它在智能驾驶[1] [2]、工业机器人[3] [4]、虚拟现实和餐饮服务等领域[5] [6]逐渐成为关键技术。在智能驾驶中可根据物体的 6D 位姿信息测算车距，从而避免车辆直接的碰撞。在工业机器人领域，获取物体的 6D 位姿，从而使其机器臂准确地抓取物体。在餐饮服务中主要获取物体的位姿状态。

传统的 6D 位姿估计方法主要有模板匹配、特征点匹配和相似度匹配等方式[7] [8]。对于模板匹配任务，一般需要手工制造模板特征，将制造完的特征与待检测的特征匹配计算物体的 6D 位姿。因此，该种方式的弊端是随着模板库中手工模板的增加，模板匹配的时间复杂度成倍数的增加，并且手工制造模板需要大量的人力。特征点匹配的方法在待匹配的物体中寻找 SIFT 或 ORB 等特征点与库中的模板进行匹配。但是，待匹配物体的点云质量较差，一般会出现空洞和空缺等点云质量的问题。对于相似度匹配方式与 RGB 图像方式类似，不过计算的是 3D 点云的相似度。

随着深度学习的发展，将计算机视觉领域的技术[9] [10] [11] [12]不断地被结合起来，进一步推动了位姿估计理论的完善。我们根据网络输入信息的不同，将网络进行划分。它主要被分为基于深度点云(D)、基于 RGB 和基于 RGBD 的方法。一般基于 D 的方式采集到的点云受物体性质影响，导致采集到的点云弥散和空缺。基于 RGB 方式采集到图像信息受光照和遮挡等因素的影响。相比较基于 RGBD 的方式结合两种数据信息一定程度上增加了网络的精度，但是数据的预处理过大也增加了网络的时间复杂度。因此，针对上述的不同任务中，我们要对各种方法进行选择。

位姿估计技术已经被调研研究很多年[13] [14] [15] [16]。而本文与之前的研究不同的是：1) 近几年的研究是着重调研传统方式，而最近两年位姿估计技术的发展迅猛，因此，本文着重调研基于深度学习的位姿估计方法。2) 在之前的调研研究中，主要以传统和非传统的方式进行划分，无法具体区分各个方法从属的细类。本文以 D、RGB 和 RGBD 输入信息进行划分方便读者理解各类方法的研究进展。3) 在本文的调研中，增加了近几年最新推出新的具体挑战的数据集和更为合理的评估指标。4) 在以前的工作中很少分析当前的挑战和未来的发展方向。因此，在我们的工作中会指出当前存在的问题及其未来发展的方向。

2. 评估指标

在 6D 位姿估计任务中，期望给出的真值与网络预测物体的 6D 位姿值保持一致。一般的评估方法有 ADD、ADD-S、n◦n cm、ACPD 和 MCPD 等方法。

ADD 度量方法[17]主要用于估计非对称对象。一般计算真实点集和预测点集之间的平均距离小于模型直径的 10%时候[18]认为网络预测的正确。公式(1)即为度量公式，其中 M 来自于三维模型的点集，R 和 T 分别代表地面真实的三维旋转和三维平移，R_p 和 T_p 分别代表预测的三维旋转和三维平移。

$$E_{ADDS} = \underset{x \in M}{avg} \left\| \left(Rx_1 + T\right) - \left(R_p x_2 + T_p\right) \right\|_2 \tag{1}$$

ADD-S 度量方法主要应用于对称对象。在对称的对象中预测出的标签值和真实值可能存在二义性[19]，但是由于物体是对称，并不一定代表网络预测是有所误差的。

$$E_{ADDS} = \underset{x_1 \in M}{avg} \min_{x_2 \in M} \left\| \left(Rx_1 + T\right) - \left(R_p x_2 + T_p\right) \right\|_2 \tag{2}$$

n°n cm 的方法[20]。该方法的主要意思是如果旋转误差和位移误差都小于 n，则认为该位姿是正确的。对于对称对象，n°n cm 为所有可能的地面真实姿态的最小误差。最常用的阈值设置包括 5°5 cm，5°10 cm 和 10°10 cm。

一般物体的姿态是模糊的[21]，如杯体，它可能是完全对称的，也有可能是非严格意义上单面对称。因此，可以通过可见表面差异进行计算。公式(3)是公式(1)的扩充，它可以用来评估具有或不具有不可区分视图的对象的结果，从而允许它们的公正的比较。公式(4)与机器人抓取方面的研究进行评估。因为，它反映最大表面的偏差，这表明机器人是否成功的抓住物体。

$$E_{ACPD} = \min \underset{x_1 \in M}{avg} \left\| \left(Rx_1 + T\right) - \left(R_p x_2 + T_p\right) \right\|_2 \tag{3}$$

$$E_{MCPD} = \min \max_{x_1 \in M} \left\| \left(Rx_1 + T\right) - \left(R_p x_2 + T_p\right) \right\|_2 \tag{4}$$

3. 用于位姿估计的数据集

对于 6D 位姿估计的数据集有很多应用场景如生活、工作和工业制造等场景。针对场景的不同，数据集的注释方式有所不同，又可以分为单个物体的注释和多个物体的注释。对于多个物体注释而言，往往需要较大的网络读入多个物体的信息进行 6D 回归。而单物体的输入则需要较大标注数据提供网络的学习。如表 1 所示，总结了一般场景的数据集，图 1 展示了数据集的个别图像案例。

3.1. 生活工作场景数据集

在生活工作场景中，一般被注释的对象包含各种生活用具或者玩具。如杯子、纸盒、小黄鸭等。LINEMOD [17]采集的场景是生活办公场景，被采集的对象是玩具。该数据集采集了 15 个玩具目标，在办公场景中注释了 18,000 张真实场景的图像，20,000 张利用虚拟合成技术合成的图像。每张图像具有遮掩等挑战。每一帧图像分辨率为 640 × 480。后来，研究者为了研究如何使 6D 位姿估计网络克服遮挡场景的挑战，将 LINEMOD 数据集进行整理，开源了 Occlusion LINEMOD [22]。数据集增加了遮掩、截断和光照变化等挑战，使其具有挑战的数据集之一。该数据集整理了 8 个目标对象，真实注释了 1214 张图像数据，并且每帧分辨率与 LINEMOD 类似。BigBIRD [23]采集的场景是单一的背景，被采集的对象是生活用品。该数据集采集了 100 个生活用品目标，在单一背景中注释了 6000 张真实场景的图片。但是数据集采集的场景缺少遮挡因素，使位姿估计方法有着较好的表现。后来，Calli 等[24]中采用 BigBIRD 的数据采集平台策略，采用单一背景的采集场景。数据集包含食品、厨房、工具、具体形状等总共 88 个目标对象，每个目标对象收集 600 帧 RGB-D 图像和物体的位姿信息。YCB 数据集[25]采集的场景是办公场景。该数据集采集了 21 个深度可见性好的生活目标对象。在该场景采集了 92 个视频共记 133,827 帧，并且每帧平均有 2 或 3 个目标对象同时存在，每个对象都具有完整的 6D 位姿注释。每帧 RGB-D 图像的

分辨率为 640 × 480。而该数据集以多个目标、遮挡、拥挤和光照变换成为具有挑战的数据集之一。在NOCS [26]中提出了基于上下文感知混合现实技术。可以从真实的图像中合成对象，从而生成一个大的注释数据集。数据集包含 6 个对象类别的 18 个不同场景和地面真实 6D 姿势和大小注释，以及总共 42 个独特的实例。

3.2. 工业应用场景数据集

在工业制造场景中，一般被注释的对象包含各种工业零件或者无纹理物体。如齿轮、轴承和箱体等。T-LESS [27]采集的场景是多个单一背景的场景，被采集的对象是工业无纹理零件。这些对象具有对称和相似等特性。该数据集在真实场景中共采集 38,000 幅图像，外加 20 个场景 10,000 张测试图像。数据集的分辨率主要是由 400 × 400，1900 × 1900 和 720 × 540 组成。而该数据集以多目标、遮挡、拥挤、无纹理、对称和相似等特性成为具有挑战的数据集之一。还有用于工业箱体拾取场景的数据集。如Doumanoglou 等人[28]提供了工业装箱和家庭环境两个场景的数据集，总共 8 个目标对象，在一幅图像中可以出现的同一对象类型范围内多个实例。Siléane 数据集[29]选取 8 个目标对象，但是每个对象有 300 帧左右的图像，该数据集的数量对于使用先进的深度学习方法是远远不够的，为此在 Bin-Picking [30]的工作中通过增加数据集的数据量和引入了两个新的工业对象(齿轮轴和环形螺杆)扩展了 Siléane 等人的工作，使其能够适用于深度学习的方法。

Table 1. Datasets for pose estimation
表 1. 用于位姿估计的数据集

数据集	类别	数量	真实/合成	遮挡	拥挤	年份
LINEMOD	单	15	18,000/20,000	Yes	No	2012
Occlusion LINEMOD	单	8	1214/-	Yes	Yes	2014
BigBIRD	单	100	6000/-	No	No	2014
YCB Model Set	单	88	52,800/-	No	No	2015
YCB	多	21	133,827/-	Yes	Yes	2018
T-LESS	多	30	48,000/77,000	Yes	Yes	2018
Doumanoglou	多	8	536/-	Yes	Yes	2016
Siléane	多	8	678/1922	Yes	Yes	2017
Bin-Picking	多	10	8000/198,000	Yes	Yes	2019
NOCS	多	6	8000/300,000	Yes	Yes	2019

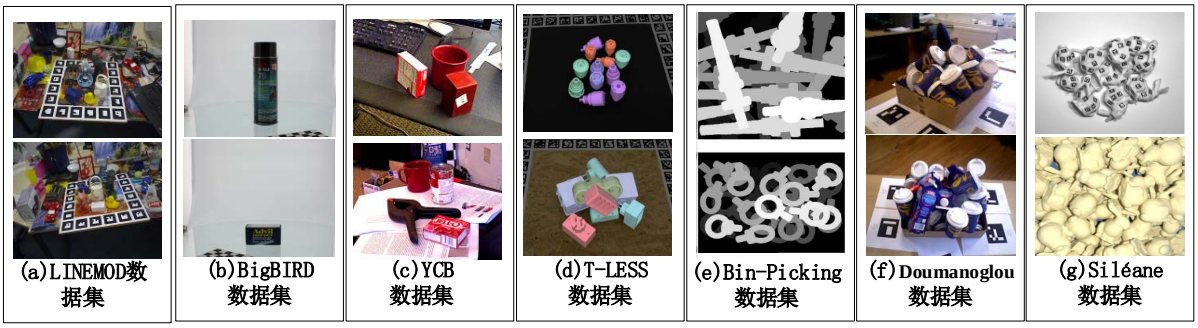

Figure 1. Image examples of commonly used datasets
图 1. 常用数据集的图像示例

4. 位姿估计

在本文中为了分类介绍各类位姿估计方法，根据输入模型数据的不同，将网络划分为基于 D、基于 RGB 和基于 RGBD 的方法。

4.1. 基于 D 的方法

基于点云的位姿估计方法传统地使用点云配准技术还原物体的 6D 位姿。一般的点云配准流程有粗配准和精配准。粗配准提供物体初始的 6D 位姿，经过精配准进行微调。该种方式需要手工制造待匹配的模板。当新的物体需要估计 6D 位姿的时候，与模板库的模板进行匹配[7] [8]。该种方式对模板和被采集的点云要求严格。并且后期随着点云模板库的增加，模板的匹配时间和估计速度成几何倍数的增加。因此，该方式适用于固定工步且物体的深度可见性较好的场景。

随着深度学习的发展，基于点云的位姿估计方式逐渐成熟。文献[31]使用自动编码器提取点云的深度信息，将其进行编码通过网络来学习这些信息的高维语义表达。文献[32]提出点云分割模型，将需要的点云从背景中分离出来并进行位姿的估计。但是，对于遮掩、拥挤场景的点云，该种方式是失效的。

4.2. 基于 RGB 的方法

基于 RGB 的方法以图像信息为输入。一般输入的图像受光照变化、遮挡和拥挤等因素挑战。导致后续网络主要基克服这些困难展开研究。一般将这些方法分为关键点、坐标对应和其他方法。关于 RGB 方法的位姿估计流程图，具体如图 2 所示。如表 2 所示，呈现了各个方法在不同数据集的表现，以供根据不同场景任务选择合适的网络。

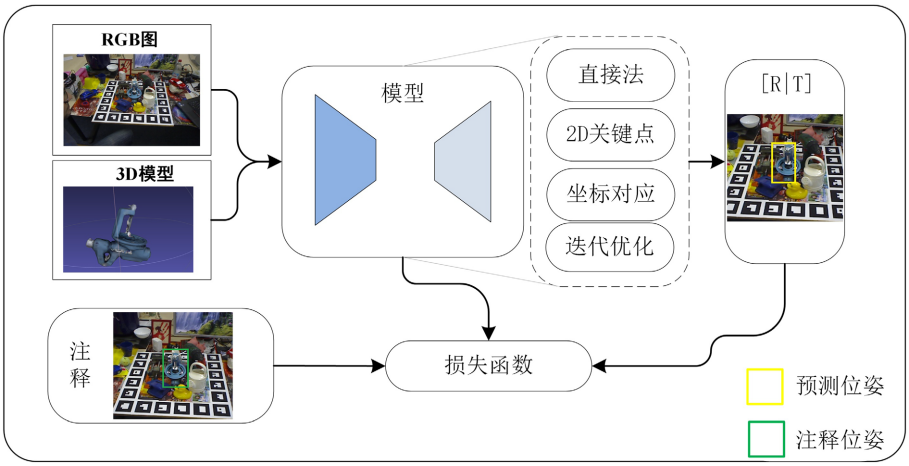

Figure 2. Image examples of commonly used datasets
图 2. 常用数据集的图像示例

基于关键点的方法。该种方法会提前在 3D 模型中设置物体的关键点，然后通过 RGB 图像去预测这些关键点，将 3D 关键点与预测的关键点通过 PnP 算法恢复物体的 6D 位姿。一般该种方式研究的是如何鲁棒地获取图像中物体特征的关键点。文献[33]通过分割网络提取物体的最大和最小边框，将边框作为关键角点进行 2D 投影的方式预测物体的 6D 位姿。但是该种方式在大场景工作中需要消耗较大的时间复杂度。文献[34]通过 Yolo 检测物体边框，边框的角点作为物体的关键点配合 3D 模型角点恢复物体的 6D 位姿。文献[35]通过热力图的方式获取物体关键点，同样通过 Perspective-n-Point (PnP)法恢复物体的 6D 位

姿。文献[36]过霍夫投票的方式获取图像中物体的关键点，将得分最高的关键点通过 PnP 算法恢复位姿。进一步，文献[37]同样通过霍夫投票还原物体的6D位姿，不过论文中它提取的特征不仅仅局限于关键点，还包括边向量和对称关系。基于关键点的方式虽然一定程度上可以提高模型恢复 6D 位姿的速度，但是透视投影导致的几何信息的丢失是无可避免的问题。

基于坐标对应的方法。该种方法主要通过密集的坐标关系获取 2D-3D 的对应关系，再次通过 PnP 算法恢复物体的 6D 位姿。文献[38]为了解决预测物体坐标系的不确定性，通过随机森林的方法来提高模型的鲁棒性。但是，该种方式对于处理截断和遮掩场景能力不强。文献[39]通过密集坐标系的方式构建 2D 与 3D 之间的对应关系恢复物体的位姿。文献[40]对每个像数预测其三维坐标，随后将 2D 与 3D 对应恢复物体的 6D 位姿。文献[41] [42]通过构建 2D 与 3D 的面片构建坐标对应关系图恢复物体的 6D 位姿。文献[47]通过图像进行特征编码和 3D 模型进行特征编码，构建密集的 2D 与 3D 对应关系恢复物体的 6D 位姿。然而，坐标对应法往往是稠密的对应关系，虽然对遮挡具有较强的鲁棒性，但较大的输出空间耗费了更多的训练时间。

其他方式。除了以上几种方式外，还存在一些方法，如直接法和迭代优化法。直接法旨在通过四元数或者变换角度回归物体的 6D 位姿[18] [43]，但是该种方式受图像质量影响较大，并且旋转空间的非线性导致预测效果很差。迭代优化法[44] [45]旨在利用上一轮预测的值与真值不断地迭代优化真值。

Table 2. Performance of RGB-based pose estimation methods on mainstream datasets
表 2. 基于 RGB 位姿估计方法在主流数据集上的表现

数据集	方法	年份	类型	输入	2D 投影	ADD(-S)	5°5 cm
LINEMOD	文献[25]	2017	直接法	RGB	70.2	62.7	19.4
	文献[33]	2017	关键点法	RGB	89.3	62.7	69.0
	文献[34]	2018	关键点法	RGB	90.37	55.95	-
	文献[36]	2019	关键点法	RGB	99.00	86.27	-
	文献[37]	2020	关键点法	RGB	-	91.30	-
	文献[39]	2019	密集匹配	RGB	98.10	89.86	94.31
	文献[40]	2019	密集匹配	RGB	-	72.4	-
	文献[41]	2020	密集匹配	RGB	-	82.98	-
Occlusion LINEMOD	文献[18]	2017	直接法	RGB	17.2	24.9	-
	文献[34]	2018	关键点法	RGB	6.16	6.42	-
	文献[36]	2019	关键点法	RGB	61.06	40.77	-
	文献[37]	2020	关键点法	RGB	-	47.5	-
	文献[35]	2018	关键点法	RGB	60.9	30.4	-
	文献[40]	2019	密集匹配	RGB	-	32.0	-
	文献[41]	2020	密集匹配	RGB	-	32.79	-
	文献[44]	2018	迭代优化	RGB	56.6	55.5	30.9
YCB	文献[18]	2017	直接法	RGB	3.7	21.3	61.3
	文献[36]	2019	关键点法	RGB	47.4	-	73.4
	文献[35]	2018	关键点法	RGB	39.4	-	72.8
	文献[44]	2018	迭代优化	RGB	-	-	81.9
	文献[45]	2020	迭代优化	RGB	-	89.8	-

4.3. 基于 RGB-D 的方法

基于 RGBD 的方法相较其他两种方式有着较好鲁棒的效果。因为，对于图像数据而言，受光照、拥挤和截断等影响，而对于点云数据而言，受点云缺失和弥散等因素的影响。而将两种数据结合在一起一定程度上达到信息弥补的作用。由于 RGBD 方法有着大量的数据输入，因此主要通过关键点检测的方式进行 6D 位姿估计。如图 3 所示，展示了网络整体执行效果，而各种网络在主流数据集的评分结果放置于表 3。

基于关键点的方法。对于以 RGB 为输入的关键点方法，由于透视投影限制了推测的 2D 关键点与 3D 关键点间的几何关系，从而使网络减少了一定的性能。如文献[36]中，通过霍夫投票方式获取物体的 2D 关键点，随后通过 PnP 算法恢复物体的位姿。但是，几何关系的透视投影对于对称或者重叠区域的点集来说有着一定的模糊性。文献[46]将其推广到三维方法中。具体来说，通过对 RGB 和点云特征的提取，将两种信息的特征进行混合，分别通过中心点检测、语义分割和关键点检测模块进行特征分类。而关键点检测同样通过霍夫投票的方式获取最高置信度的三维点。通过预测的关键点和模型关键点恢复物体的 6D 位姿。文献[47]通过结合颜色和物体的几何信息分割并预测物体的遮掩区域，并通过逐点的特征回归三维点向量。最终借助各个区域的几何约束恢复物体的 6D 位姿。文献[48]在文献[46]的工作中指出，点云数据应该与 RGB 数据相互融合的，为此提出了双流融合网络获取具有融合点云和像数数据。并在此基础上提出了 SIFT-FPS 关键检测算法，克服了关键点出现在图像之外的情况发生。文献[49]提出了一种径向的投票方案，对遮挡和拥挤场景具有一定的鲁棒性。CNN 被训练来估计 RGB 图和深度图对应像素点与 3D 点之间的距离关系，并且在目标对象中定义一组 3 个分散的关键点框架。根据网络的推断，以每个 3D 点为中心，生成一个半径等于估计距离的球体。球体的表面投票增加一个 3D 空间的累加器，累计的峰值指示关键点位置。与之前方法相比，该种方法只需要预测 3 个关键点，所以检测的速度较快。

其他方法。除了上述方法外，还有通过其他方式来预测物体的 6D 位姿。文献[50] [53] [54]通过迭代优化的方式进行优化位姿。通过 RGB 特征和点云特征提取器提取各种通道的特征，并将其特征融合在一起，迭代优化物体的位姿。文献[51]利用空间感知来预测目标的位姿，通过联合多个对象的位姿解决遮挡和杂乱场景中信息紊乱的问题，提高了位姿估计的准确性。文献[52]通过 YOLOv4 和 PointRCNN 识别图像和点云中的目标，在视觉中定位方面，利用 RANSAC 剔除的非线性优化方法求解相机位姿。

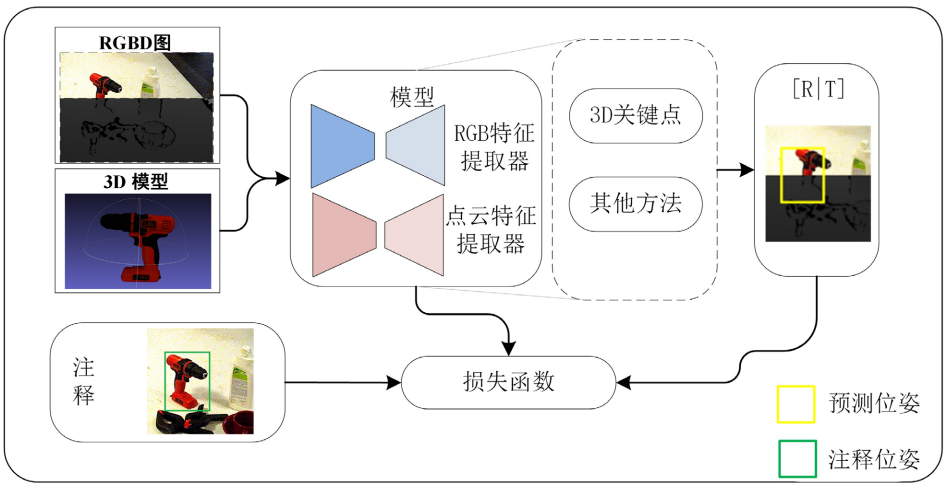

Figure 3. Image examples of commonly used datasets
图 3. 常用数据集的图像示例

Table 3. Performance of RGBD-based pose estimation methods on mainstream datasets
表 3. 基于 RGBD 位姿估计方法在主流数据集上的表现

数据集	方法	类型	输入	ADD(-S)
LINEMOD	文献[46]	关键点法	RGB-D	99.4
	文献[48]	关键点法	RGB-D	99.7
	文献[47]	关键点法	RGB-D	98.4
	文献[49]	关键点法	RGB-D	99.7
	文献[50]	其他	RGB-D	94.3
	文献[53]	其他	RGB-D	98.7
	文献[54]	关键点法	RGB-D	98.9
Occlusion LINEMOD	文献[46]	关键点法	RGB-D	63.2
	文献[48]	关键点法	RGB-D	66.2
	文献[54]	关键点法	RGB-D	65.4
YCB	文献[46]	关键点法	RGB-D	95.5
	文献[48]	关键点法	RGB-D	97.0
	文献[49]	关键点法	RGB-D	97.0
	文献[50]	其他	RGB-D	96.8
	文献[51]	其他	RGB-D	95.7
	文献[53]	其他	RGB-D	92.4

5. 未来可能的工作和挑战分析

5.1. 基于 D 方法的发展方向

1) 对于点云预处理阶段，可以构建因素弥补型模型，它可以有效避免干扰因素而造成的点云空洞、缺失等情况。而补齐后的完整点云模型可以进一步提升网络模型识别的精度。

2) 点云特征提取阶段，可以借鉴于其他领域的成就。如自然语言处理方面的 Tansformer 模型，它可以通过遮掩策略迫使网络学习到被遮掩到的信息，进一步提高模型学习语句的高维语义能力。而对于点云缺失、弥散等情况而言可以通过遮掩策略较好的解决此问题。

3) 对于点云的回归模型中，可以将 2D 相关思想提升到 3D 中。一般通过物体的点云还原物体的 6D 位姿，而很多优异的 2D 特征提取器可以克服点云缺失、弥散缺失的特征，将检测较为鲁棒的 2D 特征检测器，还原到 3D 中以此来对 6D 位姿的估计。

5.2. 基于 RGB 方法的发展方向

1) 通过 RGB 方式采集的数据一般受遮挡、拥挤和曝光等因素的影响。因此，可以针对特定场景下的挑战进行逐个解决这些因素的影响。

2) 通过 RGB 方式采集的数据较为鲁棒的网络常应用于室内场景中，对于室外场景中还需要采集更多的数据集进行实验，而无人驾驶中需要进一步克服的是较大视点的变化。

3) 通过RGB方式通常需要较大样本的数据集，通过小样本学习的方法克服数据集样本过大的缺点，但是对于多变场景是有所欠缺的，对于固定工位场景有着较好的表现。

4) 利用 2D 场景重建技术，将 2D 场景还原近似 3D 场景，求解近似 3D 场景与 3D 模型直接的关系还原 6D 位姿。

5.3. 基于 RGB-D 方法的发展方向

1) 将位姿估计的 2D 方法扩展到 3D 中，通过结合点云数据提高网络的精度。

2) 对于数据特征提取模块往往很少研究 2D 与 3D 数据之间的共性，而 2D 与 3D 之间数据关联性已经被证明很重要。

3) 建立小样本的 6D 位姿估计数据集，通过 Gan，Diffusion Model 等方法进一步扩充数据集量，达到小样本进校大数据的学习方式。

4) 基于 RGB-D 的方法相较基于 RGB 的方法对位姿估计的准确率得到提高，但是消耗的时间变的更多。因此，在网络的设计过程中，如何减少消耗的时间可以作为网络设计中的一个亮点。

6. 总结

本文详细概述了 6D 位姿在室内场景的最新研究现状。本文的介绍包括：评估指标、数据集、6D 位姿估计方法和各个方向发展的预期等内容。其中，对于 6D 位姿估计方法，我们根据数据的输入不同划分为以 D、RGB 和 RGBD 为输入的模型。对这几类的研究方法进行了总结，发现通过关键点的方法目前研究展开得过多，主要因为其速度快并且特征选点较为鲁棒，对于一般非光照变化过大的场景来说网络表现效果较好。而其他一些方法，主要针对场景不同有着各自的需要克服和解决的问题，如无纹理、拥挤和形状复杂等情况。最后，结合各个方法的优缺点在未来可能的工作和挑战中指出各类方法未来的发展方向，希望给予相关研究人员一定的帮助。

参考文献

[1] Kiran, B.R., Sobh, I., Talpaert, V., *et al*. (2021) Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 4909-4926. https://doi.org/10.1109/TITS.2021.3054625

[2] Zhu, Z. and Zhao, H. (2021) A Survey of Deep RL and IL for Autonomous Driving Policy Learning.

[3] Bousmalis, K., Irpan, A., Wohlhart, P., *et al*. (2018) Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping. *Proceedings of the IEEE International Conference on Robotics and Automation*, Brisbane, 21-25 May 2018, 4243-4250. https://doi.org/10.1109/ICRA.2018.8460875

[4] James, S., Wohlhart, P., Kalakrishnan, M., *et al*. (2019) Sim-to-Real via Sim-to-Sim: Data-Efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 12627-12637. https://doi.org/10.1109/CVPR.2019.01291

[5] Fang, W., Zheng, L. and Wu, X. (2017) Multi-Sensor Based Real-Time 6-DoF Pose Tracking for Wearable Augmented Reality. *Computers in Industry*, **92**, 91-103. https://doi.org/10.1016/j.compind.2017.06.002

[6] Strobl, K.H., Mair, E., Bodenmuller, T., *et al*. (2018) Portable 3-D Modeling Using Visual Pose Tracking. *Computers in Industry*, **99**, 53-68. https://doi.org/10.1016/j.compind.2018.03.009

[7] Hodaň, T., Zabulis, X., Lourakis, M., *et al*. (2015) Detection and Fine 3D Pose Estimation of Texture-Less Objects in RGB-D Images. *IEEE International Conference on Intelligent Robots and Systems*, Hamburg, 28 September-3 October 2015, 4421-4428. https://doi.org/10.1109/IROS.2015.7354005

[8] Hinterstoisser, S., Cagniart, C., Ilic, S., *et al*. (2012) Gradient Response Maps for Real-Time Detection of Textureless Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 876-888. https://doi.org/10.1109/TPAMI.2011.206

[9] Chen, J., Lei, B., Song, Q., *et al*. (2020) A Hierarchical Graph Network for 3D Object Detection on Point Clouds. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 392-401. https://doi.org/10.1109/CVPR42600.2020.00047

[10] Shi, W. and Rajkumar, R. (2020) Point-gnn: Graph Neural Network for 3D Object Detection in a Point Cloud. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June

2020, 1711-1719. https://doi.org/10.1109/CVPR42600.2020.00178

[11] Zhu, C., Chen, F., Ahmed, U., *et al*. (2021) Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 8782-8791. https://doi.org/10.1109/CVPR46437.2021.00867

[12] Yang, C., Wu, Z., Zhou, B., *et al*. (2021) Instance Localization for Self-Supervised Detection Pretraining. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 3987-3996. https://doi.org/10.1109/CVPR46437.2021.00398

[13] Sahin, C. and Kim, T.K. (2018) Recovering 6D Object Pose: A Review and Multi-Modal Analysis. *Proceedings of the European Conference on Computer Vision*, Munich, 8-14 September 2018, 15-31. https://doi.org/10.1007/978-3-030-11024-6_2

[14] Patil, A.V. and Rabha, P. (2018) A Survey on Joint Object Detection and Pose Estimation Using Monocular Vision. *MATEC Web of Conferences*, **277**, Article No. 02029. https://doi.org/10.1051/matecconf/201927702029

[15] Kleeberger, K., Bormann, R., Kraus, W., *et al*. (2020) A Survey on Learning-Based Robotic Grasping. *Current Robotics Reports*, **1**, 239-249. https://doi.org/10.1007/s43154-020-00021-6

[16] Sahin, C., Garcia-Hernando, G., Sock, J., *et al*. (2020) A Review on Object Pose Recovery: From 3d Bounding Box Detectors to Full 6d Pose Estimators. *Image and Vision Computing*, **96**, Article ID: 103898. https://doi.org/10.1016/j.imavis.2020.103898

[17] Hinterstoisser, S., Lepetit, V., Ilic, S., *et al*. (2012) Model Based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes. In: *Asian Conference on Computer Vision*, Springer, Berlin, 548-562.

[18] He, Y., Wang, Y., Fan, H., *et al*. (2022) FS6D: Few-Shot 6D Pose Estimation of Novel Objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seoul, 27 October-2 November 2022, 6814-6824.

[19] Di, Y., Manhardt, F., Wang, G., *et al*. (2021) So-Pose: Exploiting Self-Occlusion for Direct 6d Pose Estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2021, 12396-12405. https://doi.org/10.1007/978-3-319-49409-8_52

[20] Shotton, J., Glocker, B., Zach, C., *et al*. (2013) Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 2930-2937. https://doi.org/10.1109/CVPR.2013.377

[21] Hodaň, T., Matas, J. and Obdržálek, Š. (2016) On Evaluation of 6D Object Pose Estimation. In: Hua, G. and Jégou, H., Eds., *European Conference on Computer Vision*, Springer, Cham, 606-619. https://doi.org/10.1007/978-3-319-49409-8_52

[22] Brachmann, E., Krull, A., Michel, F., *et al*. (2014) Learning 6d Object Pose Estimation Using 3d Object Coordinates. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *European Conference on Computer Vision*, Springer, Cham, 536-551. https://doi.org/10.1007/978-3-319-10605-2_35

[23] Singh, A., Sha, J., Narayan, K.S., *et al*. (2014) BigBIRD: A Large-Scale 3D Database of Object Instances. *Proceedings of the IEEE International Conference on Robotics and Automation*, Hong Kong, 31 May-7 June 2014, 509-516. https://doi.org/10.1109/ICRA.2014.6906903

[24] Calli, B., Singh, A., Walsman, A., *et al*. (2015) The ycb Object and Model Set: Towards Common Benchmarks for Manipulation Research. *Proceedings of the IEEE International Conference on Robotics and Automation*, Seattle, 26-30 May 2015, 510-517. https://doi.org/10.1109/ICAR.2015.7251504

[25] Xiang, Y., Schmidt, T., Narayanan, V., *et al*. (2017) Posecnn: A Convolutional Neural Network for 6d Object Pose Estimation in Cluttered Scenes. https://doi.org/10.15607/RSS.2018.XIV.019

[26] Wang, H., Sridhar, S., Huang, J., *et al*. (2019) Normalized Object Coordinate Space for Category-Level 6d Object Pose and Size Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 2642-2651. https://doi.org/10.1109/CVPR.2019.00275

[27] Hodan, T., Haluza, P., Obdržálek, Š., *et al*. (2017) T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Santa Rosa, 24-31 March 2017, 880-888. https://doi.org/10.1109/WACV.2017.103

[28] Doumanoglou, A., Kouskouridas, R., Malassiotis, S., *et al*. (2016) Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3583-3592. https://doi.org/10.1109/CVPR.2016.390

[29] Brégier, R., Devernay, F., Leyrit, L., *et al*. (2017) Symmetry Aware Evaluation of 3d Object Detection and Pose Estimation in Scenes of Many Parts in Bulk. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, Venice, 22-29 October 2017, 2209-2218. https://doi.org/10.1109/ICCVW.2017.258

[30] Kleeberger, K., Landgraf, C. and Huber, M.F. (2019) Large-Scale 6d Object Pose Estimation Dataset for Industrial

Bin-Picking. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, Macau, 3-8 November 2019, 2573-2578. https://doi.org/10.1109/IROS40897.2019.8967594

[31] Gao, G., Lauri, M., Hu, X., *et al*. (2021) CloudAAE: Learning 6D Object Pose Regression with On-Line Data Synthesis on Point Clouds. 2021 *IEEE International Conference on Robotics and Automation* (*ICRA*), Xi'an, 30 May-5 June 2021, 11081-11087. https://doi.org/10.1109/ICRA48506.2021.9561475

[32] Hagelskjær, F. and Buch, A.G. (2020) PointVoteNet: Accurate Object Detection and 6 DOF Pose Estimation in Point Clouds. *Proceedings of the IEEE International Conference on Image Processing*, Macau, 1-4 December 2020, 2641-2645. https://doi.org/10.1109/ICIP40778.2020.9191119

[33] Rad, M. and Lepetit, V. (2017) Bb8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3d Poses of Challenging Objects without Using Depth. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 3828-3836. https://doi.org/10.1109/ICCV.2017.413

[34] Tekin, B., Sinha, S.N. and Fua, P. (2018) Real-Time Seamless Single Shot 6d Object Pose Prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 292-301. https://doi.org/10.1109/CVPR.2018.00038

[35] Oberweger, M., Rad, M. and Lepetit, V. (2018) Making Deep Heatmaps Robust to Partial Occlusions for 3d Object Pose Estimation. *Proceedings of the European Conference on Computer Vision*, Munich, 8-14 September 2018, 119-134. https://doi.org/10.1007/978-3-030-01267-0_8

[36] Peng, S., Liu, Y., Huang, Q., *et al*. (2019) Pvnet: Pixel-Wise Voting Network for 6d of Pose Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 4561-4570. https://doi.org/10.1109/CVPR.2019.00469

[37] Song, C., Song, J. and Huang, Q. (2020) Hybridpose: 6d Object Pose Estimation under Hybrid Representations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 431-440. https://doi.org/10.1109/CVPR42600.2020.00051

[38] Brachmann, E., Michel, F., Krull, A., *et al*. (2016) Uncertainty-Driven 6d Pose Estimation of Objects and Scenes from a Single RGB Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3364-3372. https://doi.org/10.1109/CVPR.2016.366

[39] Li, Z., Wang, G. and Ji, X. (2019) Cdpn: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6d of Object Pose Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 7678-7687. https://doi.org/10.1109/ICCV.2019.00777

[40] Park, K., Patten, T. and Vincze, M. (2019) Pix2pose: Pixel-Wise Coordinate Regression of Objects for 6d Pose Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 7668-7677. https://doi.org/10.1109/ICCV.2019.00776

[41] Zakharov, S., Shugurov, I. and Ilic, S. (2019) Dpod: 6d Pose Object Detector and Refiner. *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 1941-1950. https://doi.org/10.1109/ICCV.2019.00203

[42] Hodan, T., Barath, D. and Matas, J. (2020) Epos: Estimating 6d Pose of Objects with Symmetries. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 11703-11712. https://doi.org/10.1109/CVPR42600.2020.01172

[43] Park, K., Mousavian, A., Xiang, Y., *et al*. (2020) Latentfusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 10710-10719. https://doi.org/10.1109/CVPR42600.2020.01072

[44] Li, Y., Wang, G., Ji, X., *et al*. (2018) Deepim: Deep Iterative Matching for 6d Pose Estimation. *Proceedings of the European Conference on Computer Vision*, Munich, 8-14 September 2018, 683-698. https://doi.org/10.1007/978-3-030-01231-1_42

[45] Labbé, Y., Carpentier, J., Aubry, M., *et al*. (2020) Cosypose: Consistent Multi-View Multi-Object 6d Pose Estimation. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.-M., Eds., *Proceedings of the European Conference on Computer Vision*, Springer, Cham, 574-591. https://doi.org/10.1007/978-3-030-58520-4_34

[46] He, Y., Sun, W., Huang, H., *et al*. (2020) Pvn3d: A Deep Point-Wise 3d Keypoints Voting Network for 6d of Pose Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 11632-11641. https://doi.org/10.1109/CVPR42600.2020.01165

[47] Chen, W., Duan, J., Basevi, H., *et al*. (2020) PointPoseNet: Point Pose Network for Robust 6D Object Pose Estimation. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Snowmass Village, 1-5 March 2020, 2824-2833. https://doi.org/10.1109/WACV45572.2020.9093272

[48] He, Y., Huang, H., Fan, H., *et al*. (2021) FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25

June 2021, 3003-3013. https://doi.org/10.1109/CVPR46437.2021.00302

[49] Wu, Y., Zand, M., Etemad, A., *et al*. (2021) Vote from the Center: 6 DoF Pose Estimation in RGB-D Images by Radial Keypoint Voting. 17*th European Conference*, Tel Aviv, 23-27 October 2022, 335-352. https://doi.org/10.1007/978-3-031-20080-9_20

[50] Wang, C., Xu, D., Zhu, Y., *et al*. (2019) Densefusion: 6d Object Pose Estimation by Iterative Dense Fusion. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3343-3352. https://doi.org/10.1109/CVPR.2019.00346

[51] Wada, K., Sucar, E., James, S., *et al*. (2020) Morefusion: Multi-Object Reasoning for 6d Pose Estimation from Volumetric Fusion. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 14540-14549. https://doi.org/10.1109/CVPR42600.2020.01455

[52] 张磊, 徐孝彬, 曹晨飞, 等. 基于动态特征剔除图像与点云融合的机器人位姿估计方法[J/OL]. 中国激光, 2022, 49(6): 126-137.

[53] Chen, W., Jia, X., Chang, H.J., *et al*. (2020) G2l-net: Global to Local Network for Real-Time 6d Pose Estimation with Embedding Vector Features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 4233-4242. https://doi.org/10.1109/CVPR42600.2020.00429

[54] Hua, W., Zhou, Z., Wu, J., *et al*. (2021) Rede: End-to-End Object 6d Pose Robust Estimation Using Differentiable Outliers Elimination. *IEEE Robotics and Automation Letters*, **6**, 2886-2893. https://doi.org/10.1109/LRA.2021.3062304