计算机科学
COMPUTER SCIENCE

基于深度学习的刚体位姿估计方法综述

郭楠，李婧源，任曦

引用本文

郭楠，李婧源，任曦.基于深度学习的刚体位姿估计方法综述[J]．计算机科学, 2023, 50(2): 178-189.

GUO Nan, LI Jingyuan, REN Xi. Survey of Rigid Object Pose Estimation Algorithms Based on Deep Learning [J]. Computer Science, 2023, 50(2): 178-189.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200164

基于深度学习的刚体位姿估计方法综述

郭 楠 李婧源 任 曦
东北大学计算机科学与工程学院 沈阳 110167

摘 要 刚体位姿估计旨在获取刚体在相机坐标系下的3D平移信息和3D旋转信息,在自动驾驶、机器人、增强现实等快速发展的领域起着重要作用。现对2017—2021年间的基于深度学习的刚体位姿估计方向具有代表性的研究进行汇总与分析。将刚体位姿估计的方法分为基于坐标、基于关键点和基于模板的方法。将刚体位姿估计任务划分为图像预处理、空间映射或特征匹配、位姿恢复和位姿优化4项子任务,详细介绍每一类方法的子任务实现及其优势和存在的问题。分析刚体位姿估计任务面临的挑战,总结现有解决方案及其优缺点。介绍刚体位姿估计常用的数据集和性能评价指标,并对比分析现有方法在常用数据集上的表现。最后从位姿跟踪、类别级位姿估计等多个角度对未来研究方向进行了展望。

关键词:计算机视觉;刚体目标;位姿估计;位姿优化;深度学习
中图法分类号 TP391

Survey of Rigid Object Pose Estimation Algorithms Based on Deep Learning

GUO Nan, LI Jingyuan and REN Xi
School of Computer Science and Engineering, Northeastern University, Shenyang 110167, China

Abstract Rigid object pose estimation aims to obtain 3D translation and 3D rotation information of the rigid object in the camera coordinate system, which plays an important role in rapidly developing fields such as autonomous driving, robotics and augmented reality. The representative papers on rigid object pose estimation based on deep learning from 2017 to 2021 are summarized and analyzed. The rigid object pose estimation methods are divided into coordinate-based, keypoints-based and template-based methods. The rigid object pose estimation task is divided into four sub-tasks: image preprocessing, spatial mapping or feature matching, pose recovery, and pose optimization. The subtask realization of each method and its advantages and problems are introduced in detail. The challenges of rigid object pose estimation are analyzed, and the existing solutions and their advantages and disadvantages are summarized. Based on the rigid object pose estimation method, the articulated object and deformable object pose estimation are analyzed. The common datasets and performance evaluation indexes of rigid object pose estimation are introduced, and the performance of existing methods on common datasets is compared and analyzed. Finally, the future research directions of pose tracking and class rigid object pose estimation are prospected.

Keywords Computer vision, Rigid object, Pose estimation, Pose optimization, Deep learning

1 引言

刚体位姿估计能够获取相机坐标系下目标刚体的3D平移信息和3D旋转信息,是自动驾驶、机器人和增强现实领域的关键技术。在自动驾驶领域,获取刚体的位姿信息可以精确感知道路上的刚体,从而有效躲避障碍物,提升自动驾驶的安全性。在机器人领域,预测位姿信息可以使机械臂有效抓取、摆放和整理刚体。在增强现实领域,利用刚体的位姿信息可以有效地与虚拟刚体实现交互,从而为工业领域的辅助维修、辅助装配等应用提高仿真的真实性。

早期的刚体位姿估计方法使用基于传感器或多种传感器组合的方法来估计目标刚体的位姿信息,但这种方法会导致系统过于复杂和庞大。之后又有研究提出依据目标刚体的几何特征进行关键点的选取或模板的匹配来实现刚体位姿估计,同时为了提升算法的鲁棒性,利用SIFT[1]等具备良好不变性的特征描述子或利用数学方法[2]来实现刚体位姿估计。但是在现实世界中,刚体种类十分丰富且所处环境过于复杂,导致刚体位姿估计方法在实际应用中的效果显著下降[3-6]。

在实际应用中,刚体位姿估计任务面临的主要挑战包括:1)刚体受自身材质或光线影响,其表面纹理信息较少甚至无纹理信息,导致刚体相关纹理特征难以提取,从而影响位姿信息的获取[7];2)刚体所处环境较为复杂,常被其他刚体遮挡,

到稿日期:2021-12-14 返修日期:2022-03-27
基金项目:国家自然科学基金(52130403);中央高校基本科研业务费专项资金(N2017003)
This work was supported by the National Natural Science Foundation of China(52130403) and Fundamental Research Funds for the Central Universities of Ministry of Education of China(N2017003).
通信作者:郭楠(guonan@mail.neu.edu.cn)

导致提取信息有误，降低了刚体位姿估计的准确率；3）刚体自身具备全局或局部对称性，会出现一幅输入图像对应多个位姿的现象，导致网络训练过程出现歧义[8]。

2017 年 Rad 等[9]通过训练大量 3D 模型数据，更全面地调整相关参数，使位姿估计具有了更高的鲁棒性和精度。因此基于深度学习的刚体位姿估计逐渐成为计算机视觉领域的一个热点问题。近年来，针对刚体位姿估计问题，其已有一些学者展开了持续的研究，其中，Wang 等提出了包括文献[10]和文献[11]在内的有关实例级刚体位姿估计和刚体位姿跟踪的研究。Brachmann 等[12-13]旨在提供通用的位姿估计系统，依据随机森林解决刚体位姿估计问题，其团队考虑了输入类型在实际应用中的影响，将研究方向从基于 RGB-D 的刚体位姿估计转换为基于 RGB 的刚体位姿估计。Hodan 等[14]主要针对刚体位姿估计面临的挑战，包括刚体弱纹理及刚体对称性带来的影响进行了研究，还提出了解决弱纹理刚体位姿估计的 T-LESS 数据集。Hu 等[15-16]主要对基于关键点的刚体位姿估计方法进行了研究，将端到端的实现作为主要方向开展研究工作。Pham 等[17-18]针对刚体位姿估计的子任务进行了研究，于 2019 年提出了对点云信息的处理方法，于 2020 年提出了获取 2D-3D 对应信息的方法。

不同于已发表综述[19-21]，本文对基于深度学习的刚体位姿估计进行分析、总结和讨论，主要贡献如下：

（1）参考文献选取 CVPR，ICCV，ECCV，AAAI 等计算机视觉领域顶级会议论文，着重分析了近 5 年基于深度学习的刚体位姿估计方法，明确了基于深度学习的刚体位姿估计的最新研究进展。

（2）将刚体位姿估计任务按映射和特征关系分为 3 类：基于坐标、基于关键点和基于模板的方法。同时按照不同的实现方案划分了子类，既给出了文献之间在研究深度上的递进关系，又体现了文献之间在研究广度上的关联性。

（3）将刚体位姿估计任务细化为 4 项子任务：图像预处理、空间映射或特征匹配、位姿恢复和位姿优化。对 4 项子任务的实现方式进行了分析和总结，实现过程清晰，且可依据各子任务的实现方案及方案的优缺点，根据实际需求进行方案的选取。

（4）总结了刚体位姿估计面临的 3 方面挑战：刚体弱纹理、刚体具备对称性和刚体被遮挡。并分析了挑战产生的原因和对应的解决方案，可依据实际应用需求考虑需解决的挑战，并选取合适的方案来解决问题。

（5）将刚体位姿估计任务扩展至铰接体和可变形体位姿估计，并进行了分析和介绍，可依据刚体的相关位姿估计方法进行铰接体和可变形体位姿估计方法的改进。

2　刚体位姿估计任务概述

刚体位姿估计的一般流程包括：确定目标在图像中的位置信息，获取 2D 与 3D 的空间映射或输入图像与特征库间的特征匹配，利用映射关系求解刚体位姿信息（3D 平移信息和 3D 旋转信息），对初始的位姿信息进行优化，从而提升刚体位姿估计的准确性。因此本文将刚体位姿估计任务细化为 4 项子任务：图像预处理、空间映射或特征匹配、位姿恢复以及位姿优化。表 1 列出了典型的基于深度学习的刚体位姿估计方法实现。

表 1　近 5 年深度学习刚体位姿估计方法实现

Table 1　Implementation of deep learning rigid object pose estimation method in recent 5 years

类别	文献	训练图像		数据格式		实现过程							位姿优化	处理类别		解决何种挑战			评价指标
						图像预处理			空间映射/特征匹配			位姿恢复							LineMOD
		真实	合成	RGB	RGBD	检测	分割	深度	网络结构	网络输出	映射/匹配关系	平移	旋转		实例	类别	弱纹理	对称	遮挡	ADD(-S)
基于坐标	[22]	√	√	√	—	√	—	—	GAN	3D 坐标	2D-3D	RANSAC+PnP		—	√	—	√	√	√	72.40
	[23]	√	√	√	—	√	—	—	CNN	3D 坐标	2D-3D	CNN	CNN	—	√	—	—	√	√	93.70
	[24]	√	√	√	—	—	√	—	CNN	UV 贴图	2D-3D	RANSAC+PnP		√	√	—	—	—	√	95.15
	[25]	√	—	—	√	√	—	√	CNN	平移信息	2D-3D	CNN	MLP	—	√	—	—	—	√	98.70
	[26]	—	√	—	√	—	√	—	CNN	NOCS 映射	2D-3D	RANSAC Umeyama		—	—	√	√	—	—	—
基于关键点	[27]	√	√	√	—	—	√	—	CNN	向量图	3D-2D	不确定性 PnP		—	√	—	—	—	√	86.27
	[28]	√	—	√	—	√	—	—	CNN	热图	3D-2D	PnP		—	√	—	—	—	√	—
	[29]	√	√	√	—	—	√	√	CNN	2D 坐标	3D-2D	PnP		—	√	—	—	—	√	58.60
	[30]	√	√	—	√	—	√	√	CNN MLP	3D 关键点	3D-2D	最小二乘拟合		—	√	—	—	—	√	99.70
	[31]	√	√	—	√	—	√	√	CNN	3D 关键点	3D-2D	最小二乘拟合		—	√	—	—	—	√	99.40
基于模拟	[32]	√	√	√	—	√	—	—	AE	图像代码	代码-码本	CNN		—	√	—	—	√	√	31.41
	[33]	√	√	√	—	√	—	—	CNN	图像代码	代码-码本	CNN		—	√	—	—	√	√	76.30
	[34]	√	√	√	—	√	√	—	CNN	图像代码	预训练-标记	CNN	CNN	√	√	—	√	√	√	—

图像预处理指处理采集到的信息以明确目标刚体的位置信息，其目标是对采集到的图像信息进行处理，主要通过目标检测或目标分割的相关算法来实现。目标检测可以实现对目标刚体的快速分类和定位，因此常用于网络的初始阶段，以快速削弱由背景以及非目标刚体产生的干扰；目标分割可以实现对目标刚体的像素级分类和定位，能够获取高精确度的目标刚体区域。此外，对于包含深度信息的输入，还需将深度信息转换为点云信息。

空间映射指经目标定位处理后的数据与刚体 3D 模型之间的映射。特征匹配指获取处理后数据的特征与预先定义的特征库之间的匹配。常利用卷积神经网络（Convolutional Neural Networks，CNN）、对抗生成网络（Generative Adversarial Networks，GAN）、自动编码器（Auto-Encoder，AE）、多层感知机（Multi-Layer Perceptron，MLP）等实现。

位姿恢复指对获取的空间映射信息进行处理进而恢复出目标刚体的3D平移信息和3D旋转信息。在位姿恢复的任务中，常使用随机抽样一致(RANdom SAmple Consensus，RANSAC)算法[35]和PnP[36]等算法进行异常映射关系的去除和位姿恢复，利用深度学习模型来实现端到端的位姿恢复。

位姿优化指对初始位姿恢复后的信息进行细化，获取更精确的目标刚体位姿信息。常常通过加入深度信息，利用迭代最近点(Iterative Closest Point，ICP)方法[37]进行点云信息与位姿信息的配准，从而实现位姿优化；利用深度学习模型，如DeepIM模型[38]等来实现位姿优化。

表2列出了基于深度学习的刚体位姿估计子任务不同实现方式的优缺点。

表2 基于深度学习的刚体位姿估计子任务实现方式及优缺点

Table 2 Implementation of rigid body pose estimation subtask based on deep learning and its advantages and disadvantages

任务	实现方式	优点	缺点
图像预处理	目标检测[22,25,28,32-34,39-40]	推理速度快 网络结构较简单	精确度较差
	目标分割[9-10,15,24-25,27,29,31,40-42]	像素级的分类 区域精确度较高	推理速度较慢，注释过程复杂
空间映射	2D-3D[22-24,26,41]	鲁棒性较好	映射复杂，易出现较多错误对应， 推理时间较长
	3D-3D[25,30-31]	准确率较高	需要点云信息，推理时间较长
	3D-2D[9,15,27-28,39]	推理速度较快	映射较简单，鲁棒性较差
位姿恢复	RANSAC和PnP[22,24,41]	实现简单	不可微，无法实现端到端的训练
	深度学习模型[10,23,25,27,32-34,40,42]	可实现端到端的训练 准确率较高	网络结构较复杂
位姿优化	ICP[34,43]	实现较为简单，配准效果较好， 算法收敛性较好	计算开销较大，需要点云数据， 产生错误对应点
	深度学习模型[10,24,42,44-46]	不需点云数据，优化精度较高	推理时间较长，训练参数较多， 需要位姿注释

3 基于坐标的刚体位姿估计方法

基于坐标的刚体位姿估计方法通过对目标刚体三维模型坐标进行标注，然后训练网络对处理后的图像像素进行坐标预测，获取2D-3D的空间映射，最后使用RANSAC和PnP基本算法或训练网络来恢复目标刚体位姿。基于坐标的刚体位姿估计方法的一般流程如图1所示，图1中展示的每一子任务的实现方式选取最常用的方法表示。

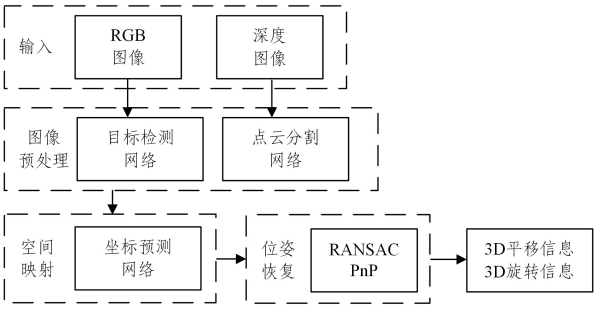

图1 基于坐标的刚体位姿估计方法一般流程

Fig. 1 General process of coordinate-based rigid object pose estimation method

3.1 图像预处理

基于坐标的刚体位姿估计方法将在3.2节阐述，其可通过预测目标刚体的密集3D坐标来提升准确率，因此在对图像进行预处理时，侧重于提升方法的速度，文献[22-23,41,47-48]均使用现有的检测速度较快的单阶段目标检测算法，如利用YOLOv3[49]等对RGB图像进行处理。使用已有算法还可以分离目标检测网络和后续任务的网络，2D目标检测算法的发展，使得可以随时进行目标检测算法的替换，从而提高系统的可扩展性。但利用单阶段目标检测算法的准确率仍有较大提升空间，因此文献[22-23,41]引入了动态放大(Dynamic Zoom-In，DZI)算法进行优化，进一步提升了目标检测算法的准确率。

当后续任务对获取目标刚体区域的精确度要求较高时，利用目标检测相关算法不可实现，故研究者利用目标分割算法来进行处理，同时为了保证系统的快速和轻量，考虑合并目标分割和空间映射网络[24]。其中的UV贴图的实现需获取精确的目标刚体区域，故设计网络来实现目标分割和3D坐标的预测。此外，对于RGB-D类型的输入，图像预处理还包括对深度信息的处理，常利用学习网络在初始阶段将深度信息转换为3D点云信息，进行分割目标刚体点云信息以及对齐操作。

3.2 空间映射

在基于坐标的刚体位姿估计方法中，空间映射主要为2D-3D的映射关系，使用的网络主要结构为CNN。文献[41]中对空间映射的获取可解决后续的对旋转信息的预测，同时为保证系统高效，利用置信图直接回归所有目标刚体坐标进行预测，文献[23]也以此为基础获取映射关系。但直接回归连续坐标易产生较多错误坐标且存在无限连续解空间的问题，因此文献[24]为了解决文献[23,41]存在的问题，提出了通过回归UV贴图来获取映射关系的方法。上述方法虽达到了较高的准确率但鲁棒性较差，故文献[22]提出利用GAN的特性来设计网络，提升了算法的鲁棒性。

为了进一步改进位姿估计的性能，文献[26]引入了归一化刚体坐标空间(Normalized Object Coordinate Space，NOCS)供同类别内的刚体共享，以获取类别级的空间映射关系。文献[25]在此阶段引入深度信息，将3D点云信息作为输入，利用3D点云分割信息来预测局部正则坐标，进而获取映射关系。

3.3 位姿恢复

基于坐标的位姿估计方法易出现较多错误的映射关系，

故文献[22，24，26]考虑利用 RANSAC 算法去除错误的映射关系，并利用 PnP 及其变体算法或 Umeyama[50] 算法恢复位姿信息。但此类方法忽略了平移信息和旋转信息的不同特性，易出现性能不平衡的问题，因此文献[23，25，41]设计了基于解耦的方法，分别预测刚体的平移信息和旋转信息，且文献[23]在文献[41]的基础上提出了 Patch-PnP 算法，解决了文献[41]在恢复旋转信息的过程中存在的不可微问题。为了提高位姿估计的准确率，文献[25]引入了深度信息，提出了点级别的嵌入向量特征并依据文献[51]的方法恢复了目标刚体的平移信息，提升了位姿恢复的准确率。

3.4　基于坐标的刚体位姿估计方法比较

基于坐标的刚体位姿估计方法，在图像预处理后，依据准确的 2D-3D 空间映射预测刚体的位姿信息，降低了由于物体被遮挡而产生的不良影响，在一定程度上提高了遮挡情况下刚体位姿估计的准确性。表 3 列出了典型的基于坐标的刚体位姿估计方法的比较。

表 3　典型的基于坐标的刚体位姿估计方法比较

Table 3　Comparison of typical coordinate-based rigid object pose estimation methods

子分类	文献	特点
基于 GAN 的位姿估计	[22](2019)	可解决位姿估计的多种挑战
基于合成数据的位姿估计	[24](2019) [26](2019)	减少了数据注释的工作量
基于解耦的位姿估计	[41](2019) [25](2020) [23](2021)	依据平移和旋转特性提升位姿恢复准确率
基于端到端的位姿估计	[25](2020) [23](2021)	提高位姿估计的推理速度

4　基于关键点的刚体位姿估计方法

基于关键点的刚体位姿估计方法通过对已知目标刚体的 3D 模型进行 2D 投影并选取关键点进行标注，训练网络对输入图像进行关键点预测，获取 3D-2D 的空间映射，最后使用 PnP 基本算法或训练网络来恢复目标刚体位姿。基于关键点的刚体位姿估计方法实现的一般流程如图 2 所示，所展示的子任务的实现方式选取最常用的方法。

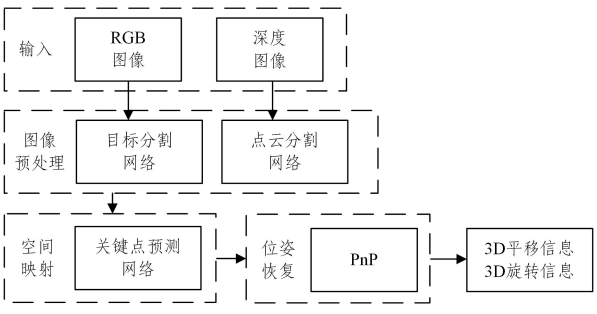

图 2　基于关键点的刚体位姿估计方法一般流程

Fig. 2　General process of keypoints-based rigid object pose estimation method

4.1　图像预处理

基于关键点的刚体位姿估计方法在图像预处理的过程中，文献[9，15，27，30-31]使用准确率更高的目标分割算法进行处理，以便精确预测关键点坐标。文献[9，15]利用分割网络预测对象标签和目标刚体质心，其中文献[9]在文献[15]的基础上进行了改进，设计了两级分割网络，进一步提高了目标刚体分割的准确率。上述方法虽提升了准确率，但在遮挡情况下的鲁棒性仍较差，因此 Peng 等[27] 引入向量图，利用向量图中包含的预测关键点向量来解决遮挡刚体预测不准确的问题，文献[31]对文献[27]进行了扩展，同时基于文献[10]的特征提取算法，引入了深度信息，进一步提升刚体位姿估计的准确率。文献[30]则改进了文献[10]的特征提取算法，增加了双向融合模块，以便更好地共享互补信息，从而更好地学习目标刚体的外观信息和几何信息。虽然上述方法可获得较准确的位姿估计预测结果，但是当对合成数据进行处理时，利用全监督方式进行训练会导致性能显著下降[52]，因此 Yang 等[29] 以弱监督的方式用真实数据及合成数据进行训练，提升了基于合成数据的位姿估计算法的准确率。

为了提高算法的推理速度，文献[28，39]使用目标检测算法进行图像预处理。其中文献[39]基于 YOLO[44] 框架直接预测后续的映射关系，极大地提升了系统的运行速度；文献[28]为了更好地解决刚体被遮挡的问题，利用目标检测算法，快速获取图像补丁信息，提升了图像预处理的效率。

4.2　空间映射

在基于关键点的刚体位姿估计方法中，空间映射主要为 3D-2D 的映射关系，使用的网络主要结构为 CNN。文献[9，15，39]选取 3D 模型的边界角点作为关键点，通过预测目标刚体模型的 3D 边界框角的 2D 投影来获取映射关系。文献[28]为了解决上述方法存在的遮挡情况下预测效果较差的问题，利用小块热图预测 2D 投影，通过解决最优化问题来预测关键点，获取映射关系。然而在定位过程中，边界框角会产生较大的误差，因此文献[27，31]利用霍夫投票和最远点采样(Farthest Point Sampling，FPS)算法生成关键点假设，此类方法生成的关键点均匀地分布于 3D 模型表面，能够减轻杂乱背景的影响，进而提升位姿估计的准确率。

上述方法在准确率上得到了较大的提升，但均利用全监督的方式进行训练，需对大量数据进行注释。为了进一步减轻工作量，Yang 等[29] 利用可见轮廓对齐和双尺度关键点一致性两个自监督函数来保证关键点的可微回归，实现了自监督预测目标刚体的关键点。为了进一步提升位姿估计任务的准确率，文献[54]利用包含关键点、对称向量在内的多种中间件信息来获取映射关系；文献[55]还提出了一种通过匹配 RGB 图和渲染图来获取关键点的方法。

4.3　位姿恢复

基于关键点的刚体位姿估计的预测通常异常值较少，仅产生固定数量的对应(通常为 8 组对应)，因此即使存在空间映射错误也无法进行异常值去除的操作，否则会因映射关系数量过少而无法完成后续的位姿恢复任务，因此不可使用异常值去除操作进行处理。

对于 RGB 类型的输入，仅依据映射关系利用 PnP 及其相关算法进行处理即可实现位姿恢复，在基于关键点的位姿估计中，使用最多的方法为 EPnP 算法。由于对应数量的限制，无法对平移和旋转信息进行解耦处理，会忽略不同关键点

可能具有的不同置信度以及不确定性，因此文献[17]提出了不确定性驱动的 PnP 算法，该算法结合最小二乘优化算法能够直接减少重投影误差，进而恢复刚体的位姿信息。

对于 RGB-D 类型的输入，经常利用最小二乘拟合算法来处理获取的 3D 刚体表面点和处理后的目标刚体点云信息，进而恢复目标刚体的位姿[30-31]。

4.4 基于关键点的刚体位姿估计方法比较

基于关键点的刚体位姿估计方法，在图像预处理后，依据稀疏的 3D-2D 空间映射预测刚体的位姿信息，缩短了关键点坐标预测的时间，在一定程度上能够加快刚体位姿估计的推理速度。表 4 列出了典型的基于关键点的刚体位姿估计方法的比较。

表 4 典型的基于关键点的刚体位姿估计方法的比较

Table 4 Comparison of typical keypoints-based rigid object pose estimation methods

子分类	文献	特点
基于热图的位姿估计	[28](2018)	
基于分割的位姿估计	[9](2017) [15](2019)	提高了遮挡情况下的位姿估计的准确率
基于向量场的位姿估计	[27](2019) [31](2020)	
基于 YOLO 的位姿估计	[39](2018)	极大加快了位姿估计的推理速度
基于自监督和弱监督的位姿估计	[29](2021)	减少了数据注释的工作量

5 基于模板的刚体位姿估计方法

在基于模板的刚体位姿估计方法中，通过建立目标刚体特征数据库，对输入数据进行特征提取，最后与特征库中的特征进行比对，选取匹配度最高的特征作为最优特征来恢复目标刚体的位姿信息。基于模板的刚体位姿估计方法实现如图 3 所示，所展示的子任务实现方式选取最常用的方法。由于基于模板的方法不涉及空间映射关系的获取，因此在位姿恢复子任务中，无法使用 RANSAC 或 PnP 相关算法进行处理，因而基于模板的方法利用 CNN 来进行位姿恢复。

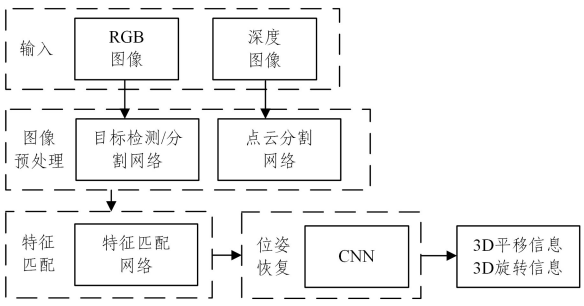

图 3 基于模板的刚体位姿估计方法一般流程

Fig. 3 General process of template-based rigid object pose estimation method

5.1 图像预处理

基于模板的方法依据方法的自身特性来选取处理算法。对于 RGB 类型的输入，主要使用目标检测对输入图像进行处理，其目的是提取 RGB 相关特征并快速实现对目标刚体的裁剪，便于从处理后图像直接回归到位姿信息，如文献[32-33]。基于模板的方法主要从处理后图像直接回归刚体位姿信息，故无须使用目标分割对刚体进行精确定位，而文献[35]使用了目标分割，其目的是解决刚体被遮挡导致的位姿估计准确度下降的问题。

对于 RGB-D 类型的输入，主要使用目标分割方法处理 RGB 图像，其目的是利用图像分割得到的像素级特征匹配深度信息，如文献[10,40,42]，其中文献[10]还考虑如何更充分地利用 RGB 信息和深度信息，提出了一种异构信息融合的方法。在上述以 RGB-D 类型为输入的方法中，常依据文献[56]将深度信息转换为点云信息再进行后续的处理。

5.2 特征匹配

对于 RGB 类型的输入，特征匹配主要指网络预测的特征和目标刚体真实值特征的对应，准确的匹配对后续正确位姿的选取极为重要。对于 RGB 类型的输入，在基于模板的位姿估计方法中，通过训练 CNN 获取输入图像的特征，与已有目标刚体真实值提取的特征制作的数据进行匹配，选取最相似的匹配[35]。文献[32,57]利用图像代码和已知模型码本进行隐式匹配获取了较为准确的匹配结果。依托于文献[26]的思想，文献[57]考虑共享相同特性来关联对象视图，设计了多路径学习的算法，从而实现了类别级的位姿估计任务。

为了提升位姿估计的准确率，文献[10,40,42]引入了深度信息，其中文献[10,42]利用特征融合来更好地处理深度信息，文献[10]将深度信息转换为几何信息来与颜色信息进行融合，但仍存在互补信息利用不充分的问题，故文献[42]提出将 2D 特征和 3D 特征进行融合，解决了此问题。为了实现类别级的位姿估计，文献[40]利用变分自动编码器(Variational Auto-Encoders，VAE)的特性来处理深度信息以预测形状表示，进而获取同类物体的特征匹配。

5.3 位姿恢复

在本节所介绍的方法中，文献[34,40]设计了解耦网络来分别预测平移信息和旋转信息；为了提升位姿估计的准确率，文献[40]引入了深度信息，并依据文献[34]和文献[10]的思想设计网络，成功预测了平移信息和旋转信息；文献[33]为了更好地解决物体对称性产生的歧义性问题，将位姿恢复问题设计为分类问题，进行位姿信息的预测；为了提升位姿估计的准确率，文献[42]使用像素级的特征向量作为位姿恢复网络的输入，训练 MLP 预测位姿信息。

5.4 基于模板的刚体位姿估计方法比较

基于模板的刚体位姿估计方法，在图像预处理后，依据特征匹配预测刚体的位姿信息，在一定程度上能够提高弱纹理刚体位姿估计的准确率。表 5 列出了典型的基于模板的刚体位姿估计方法的比较。

表 5 典型的基于模板的刚体位姿估计方法的比较

Table 5 Comparison of typical template-based rigid object pose estimation methods

子分类	文献	特点
基于图像特征库的位姿估计	[33](2017) [34](2017)	提高了对称物体的刚体位姿估计准确率
基于融合特征的位姿估计	[10](2019) [42](2020)	充分利用互补信息源
基于合成数据和形状表示的位姿估计	[40](2020)	减少了数据注释的工作量，可实现类别级的位姿估计

6　刚体位姿优化方法

位姿优化的目的是提升位姿估计的准确率，现有的位姿优化方法主要针对仅以 RGB 为输入的位姿估计方法，缺少刚体的深度信息导致刚体位姿估计的准确率较低，因此在某些要求刚体位姿估计具备高准确率的场景下，部分方法考虑增加深度信息来进行位姿优化。

常使用的一种方法是将深度信息转换为点云信息，利用 ICP 进行深度点云信息和预测模型点云的配准，基于由 RGB 提供的目标刚体初始位姿进行优化，从而获取更准确的目标刚体最终位姿，例如文献[34，43]均使用了 ICP 进行位姿优化。

常见的方法还有基于 CNN 的刚体位姿优化方法。文献[9]通过最小化预测位姿和真实位姿的关键点的差值来实现位姿优化，但此方法的优化效果仍有较大的提升空间，因此文献[10，42]引入了深度信息，利用其自身特性进行迭代优化，提升了位姿优化的性能。

为了提升位姿优化网络的可扩展性，2018 年 Li 等[38]和 Manhardt 等[44]分别提出了针对 RGB 图像的位姿优化网络，2020 年 LABBÉ 等[45]也对 RGB 图像进行了位姿优化。然而现阶段最常使用的位姿优化算法仍是 Li 等提出的通用 DeepIM 模型，该模型提高了常用的位姿估计算法的准确率，文献[24]在文献[38]的基础上引入了深度信息，增加了对 Z 轴的处理，更好地实现了位姿优化。文献[46]依据文献[38]的思想，利用渲染图来进行位姿优化，通过渲染网络来预测深度信息并采用梯度优化来实现此任务。

7　刚体位姿估计挑战及解决方案

本节主要就刚体位姿估计中刚体弱纹理、刚体对称以及刚体被遮挡这 3 项挑战进行介绍。

刚体弱纹理指刚体表面纹理信息较少或缺失，出现此现象的原因往往是亮度差异过小而无法区分不同信息。对于此类刚体，通常无法提取有效的纹理特征，进而影响图像预处理和空间映射的实现。

刚体对称指刚体局部或全局对称。由于旋转对称对象在某些三维旋转下形状的等效性，某些对象在渲染的视点看起来完全相同。如果在训练过程中将这些作为负例，可能会引入一种歧义，导致训练过程中一个输入图像可能存在多个位姿与之对应[58]，妨碍模型的训练。

刚体被遮挡指目标刚体所处环境较为复杂，易被非目标刚体或环境信息遮挡，从而无法提取足量特征，影响目标刚体图像预处理和空间映射子任务，使位姿估计准确度下降甚至估计失败。

表 6 列出了典型刚体位姿估计方法应对各项挑战的解决方案。在进行刚体位姿估计任务挑战性的分析上，常用数据集是 T-LESS，LineMOD 和 LineMOD Occlusion，常用评价指标是 ADD-S，而文献[47，59]由于主要利用基于重建的方式来进行处理，因此使用的数据集与评价指标不同于其他方法。

表 6　典型刚体位姿估计方法解决方案

Table 6　Typical rigid object pose estimation method solutions

解决挑战	文献	解决方案	数据集	评价指标	性能
弱纹理	[22]	利用 GAN 训练无纹理 3D 模型	T-LESS	$e_{VSD}<0.3$	29.50
	[47]	利用 3D 重建训练无模型物体	Expo	ADD-S＜10%	58.50
	[39]	利用 YOLO 特性对 3D 包围框进行训练	LineMOD	ADD-S＜10%	55.95
对称	[9]	设计分类器对旋转进行分类	T-LESS	ADD-S＜10%	56.10
	[60]	预测多个位姿与真实位姿对应	T-LESS	ADD-S＜10%	56.40
	[33]	考虑不同视点利用分类器预测位姿	LineMOD	ADD-S＜10%	76.30
	[32]	指导预测到最近对称位姿	LineMOD	ADD-S＜10%	72.40
	[59]	设计基于重建的损失函数	Pix3D	EMD	15.82
	[34]	预测与真实模型最近点的偏移量	YCB-Video	ADD-S＜10%	75.90
	[23]	扩展损失函数至对称感知公式	LineMOD	2 cm 2°	62.10
遮挡	[22]	利用 GAN 预测遮挡部分 3D 坐标	Occlusion	ADD-S＜10%	32.00
	[11]	多视图一致性解决位姿跟踪遮挡问题	NOCS-Real275	5 cm 5°	33.30
	[27]	利用向量图预测遮挡部分 3D 坐标	Occlusion	ADD-S＜10%	40.77
	[29]	利用注意力机制预测遮挡部分关键点	Occlusion	ADD-S＜10%	24.90

对于弱纹理刚体的挑战，主要在网络训练过程中处理，利用不需纹理信息的 3D 模型进行训练[22，29]或利用三维重建算法进行训练[47]。对于应对对称性刚体的挑战，可以设计分类器对旋转等信息进行分类来解决，但其工作量较大且推理时间较长[9，32]；此外，还可以设计损失函数在网络训练过程中进行处理，在保证模型推理速度的前提下应对对称性的挑战[22-23，35]。对于刚体被遮挡这一挑战，可通过预测遮挡部分的 3D 坐标来解决，包括基于投票的方法[22]以及基于对抗生成的方法[22]。当方法涉及可微回归时，无法利用基于投票的方法，因此可预测注意力权重，利用基于注意力的平均值来预测刚体被遮挡部分的 3D 坐标，解决刚体被遮挡的问题[27]。若物体处于强遮挡情况下，可考虑增加深度信息或生成点云数据来提高鲁棒性[61]。

8　铰接体与可变形体的位姿估计

铰接体与可变形体的位姿估计任务扩展于刚体的位姿估计任务，相较而言，铰接体和可变形体位姿估计的实现更复杂，难度更高。其中铰接体可被看作由多个刚性部件组合形成，在位姿估计的过程中，除了解决刚体位姿估计问题外，还需对连接节点的属性进行预测，使刚体的自由度更高，刚性部件的链接和约束也会使刚体位姿非独立[62]；对于可变形体而言，方法的实现更趋向于类别级的刚体位姿估计方法的实现，但不同的是，可变形体常利用曲面映射来获取顶点坐标进而恢复刚体位姿信息，在可变形体的位姿估计任务中，因为目标

刚体的尺寸、纹理和位姿信息会随刚体的形状改变而改变，所以其实现难度更大。本节介绍了铰接体和可变形体的位姿估计方法，就二者的实现进行了简要阐述，对所阐述方法的优缺点进行了分析。

Li 等[62]提出了解决类别级铰接刚体的位姿估计问题的方法，该方法仅以深度图像作为输入，基于 NOCS 算法引入了感知关节的规范化坐标空间层次结构，来预测类别级的刚体坐标表示和关节属性，包括关节参数和关节状态；再利用 Umeyama 和非线性最小二乘算法，获取目标刚体的初始位姿和优化后位姿信息。但此方法未考虑环境中的遮挡等问题。

Pavlasek 等[63]于 2020 年提出利用马尔可夫随机场(Markov Random Fields，MRF)来处理铰接体位姿估计的方法，其中，隐藏节点表示目标的位姿信息，图中的边表示部件之间的约束，利用消息传递对图进行推理，在位姿变量之间共享信息，共同给出铰接体的估计状态。该方法设计了一种生成-判别的方法，利用局部热图和目标刚体之间的连接约束来执行 MRF 中的消息传递，从而解决了复杂环境下的铰接体的位姿估计问题。

Chi 等[64]于 2021 年提出了解决类别级的可变形体的位姿估计方法，该方法将可变形体的位姿估计问题转换为标准空间中的形状完成任务，利用同一类别内的服装实例来定义典型形状空间，共享类别级的刚体位姿。该方法将观测到的曲面映射到规范空间中，获取每个顶点正则坐标标签，描述了服装在不同情况下的完整配置。

9 数据集及性能评价指标

在深度学习的方法中，合理利用数据集进行网络训练极为重要。在位姿估计中，训练数据集主要分为合成数据集和真实数据集。真实数据是直接从现实世界环境中测量或收集到的数据，合成数据则是利用计算机人工合成。在位姿估计问题中，合成数据常使用 Unity 等引擎制作合成数据集，其具有以下优点：1)可获取大量不同角度的图像，能够快速扩大数据集；2)可以省去人为进行模型或图像的注释工作，提升了工作效率。合成数据集的缺点在于，应用到真实世界中，会出现域间隙问题，从而影响性能[65]。真实数据在位姿估计问题中的缺点包括：1)可能导致在新环境下泛化能力差；2)获取位姿信息的注释需人为进行，极大地增加了工作量。但真实数据集不存在域间隙问题，应用在真实世界中，与训练时的刚体更加相似，收敛速度更快。综合二者的优点，大部分算法训练时既使用合成数据又使用真实数据。

9.1 常见数据集

表 7 列出了刚体位姿估计中的常用数据集，并介绍了各数据集的注释内容、数据种类、视频数量、处理类别以及数据集中刚体的性质，包括目标刚体被遮挡的情况、目标刚体是否大多为弱纹理刚体、目标刚体是否大多具备对称性。其中实例级数据集的类别数量表示实例刚体的数量，类别级数据集的类别数量表示刚体的种类。

表 7 常用数据集介绍

Table 7 Introduction to common datasets

数据集	注释	类别数量	数据量	处理级别	遮挡情况	弱纹理	对称性
LineMOD[66]	位姿信息 深度信息	15	18273	实例级	一般	√	×
Occlusion[12]	位姿信息 深度信息	8	1214	实例级	严重	√	√
YCB-Video[34]	位姿信息 深度信息 掩膜信息	21	133827	实例级	一般	×	×
T-LESS[14]	位姿信息 深度信息	30	49000	实例级	严重	√	√
HomebrewedDB[67]	位姿信息 深度信息	33	34830	实例级	一般	√	×
RobotP[68]	位姿信息 深度信息 掩膜信息 包围框	8	4200	实例级	一般	×	×
NOCS[26]	位姿信息 深度信息 NOCS 信息	6	30000	类别级	一般	×	×

LineMOD[34]数据集是 Hinterstoisser 等于 2012 年在 ACCV 上提出的，该数据集提供了 15 个对象实例及其对应的无纹理 CAD 模型，包含了 15 个以 RGB-D 类型为输入的真实视频序列，共 18273 帧数据。刚体在该数据集中于不同视角和不同光照条件下捕获，但选取的视角范围有限。注释包括单目标刚体的 6D 位姿信息和深度信息，主要用于评估实例级的刚体位姿估计方法性能，可以了解方法在面临较少遮挡时的效果。在网络训练中，常选取 15%的数据用作训练，85%的数据用作测试，利用 ADD-S 指标对鸡蛋盒和胶水两种对称刚体进行评价，利用 ADD 指标对其余刚体进行评价。

LineMOD Occlusion[12]数据集为 LineMOD 数据集的子集，是 Brachmann 等于 2014 年在 ECCV 上提出的，该数据集提供了从 LineMOD 数据集中选取的 8 个对象实例及其对应的无纹理 CAD 模型，包含了 1214 帧数据。数据选取自 8 个实例刚体中遮挡较多的数据。LineMOD Occlusion 数据集对数据中的所有目标刚体进行注释，注释包括了所有目标刚体的 6D 位姿信息和深度信息，主要用于评估实例级的刚体位姿估计方法的性能，可以了解方法在处理多个刚体、遮挡严重

情况下的效果。在网络训练中，该数据集常被用来对在 LineMOD 数据集训练后的算法进行测试，评价指标与 LineMOD 数据集相同。

YCB-Video[34]数据集是 Xiang 等于 2018 年提出的，该数据集提供了 21 个对象实例及其对应的纹理 CAD 模型，包含了 92 个以 RGB-D 类型为输入的真实视频序列，共 133827 帧数据。数据选取于国外超市中常见的刚体，于不同视角和不同光照条件下捕获，每个数据包含 3～9 个目标刚体。数据集对每帧数据中的所有目标刚体进行注释，注释包括了所有目标刚体的 6D 位姿信息、深度信息以及掩膜信息，部分数据还在真实背景中添加了合成目标刚体，但存在不符合真实世界情况的数据，如悬浮在空中的刚体。该数据集主要用于评估实例级的刚体位姿估计或位姿跟踪方法的性能，可以了解方法在复杂背景下处理多个刚体的效果。在网络训练中，常选取 80 个视频和 80000 个合成图像用于训练，从其余 12 个视频中提取 2949 个关键帧用于测试[69]。

T-LESS[14]数据集是由 Hodan 等于 2017 年提出的，是截至目前公认的最具有挑战性的数据集。该数据集提供了 30 个对象实例及其对应的 CAD 模型，由于刚体本身即为弱纹理刚体，故其模型同样为无纹理 CAD 模型，包含了 20 个以 RGB-D 类型为输入的真实视频序列，共约 49000 帧的数据。数据集中的目标刚体为工业上使用较多的统一颜色的弱纹理刚体，且大多数目标刚体为对称刚体。数据集对每帧数据中的所有目标刚体进行注释，注释包括了所有目标刚体的 6D 位姿信息和深度信息，主要用于评估实例级的目标刚体位姿估计方法的性能，可以了解方法在处理多个刚体、且刚体纹理信息少并具备对称性情况下的效果。在网络训练中，常选取 39000 帧的数据进行训练，这些数据的背景为黑，而选取剩余的约 10000 帧的数据用于测试，测试集中的数据背景较为复杂，包括了不同的光照情况和遮挡等。由于该数据集挑战性较大，故使用该数据集的方法较少。

HomebrewedDB[67]数据集是由 Kaskman 等于 2019 年提出的，该数据集提供了 33 个对象实例及其对应的高精度的基于重建的 CAD 模型，包含了 13 个以 RGB-D 类型为输入的视频序列，共 34830 帧数据。数据包括了纹理和无纹理的刚体，同时每个视频序列还由两个不同的 RGB-D 传感器拍摄，每个刚体处于不同光照条件、不同遮挡的情况下。该数据集主要用于评估实例级的刚体位姿估计方法的性能，可以了解到方法对弱纹理刚体等情况的处理效果。在该数据集中，由于选取了两个传感器进行目标刚体的捕获，故获取的图像分辨率差距较大。

RobotP[68]数据集是由 Yuan 等于 2018 年提出的，该数据集提供了 8 个对象实例及其对应的基于重建的 CAD 模型，包含了 4200 个数据。数据集对所有目标刚体进行注释，注释包含了所有目标刚体的 6D 位姿信息、深度信息、包围盒和掩膜信息，主要用于评估实例级的目标刚体位姿估计方法的性能。在网络训练中，选取 3200 个合成的数据用于训练，其余的数据用于测试。不同于上述数据集，该数据集将目标刚体均匀分布于背景中，防止目标刚体始终处于图像中心。

NOCS[26]数据集是由 Wang 等于 2019 年提出的，该数据集提供了 6 个类别的刚体及其对应的合成 CAD 模型。数据集提供了 300000 张利用 Unity 实现的合成图像和 8000 张从现实世界获取的真实图像。注释包含了目标刚体类的 6D 位姿信息、深度信息和归一化坐标映射。该数据集主要用于评估类别级的刚体位姿估计或位姿跟踪方法的性能。在网络训练中，使用 6 个类别和 3 个实例进行训练；使用 6 个类别进行测试，每个类别使用唯一的实例进行测试。

9.2　性能评价指标

平均点距离(ADD)是 Hintertoisser 等定义的指标，用于计算预测位姿和真实位姿的两个 3D 模型之间对应点的距离的平均值，是位姿估计中最常用的评价指标，该指标对平移和旋转的误差取决于刚体的大小和形状[13]。其中，3D 模型记作 M，x 表示组成模型的 3D 点，真实目标刚体位姿分别用 R 和 T 表示旋转量和平移量，预测目标刚体位姿分别用 $\hat{R}$ 和 $\hat{T}$ 表示旋转量和平移量，则：

$$\omega_{\mathrm{ADD}}=\underset{x\in M}{\operatorname{avg}}\left\|(Rx+T)-(\hat{R}x+\hat{T})\right\| \tag{1}$$

平均最近点距离(ADD-S)用于整体评估对称对象和非对称对象，计算真实 3D 模型点到预测目标刚体模型上最近点的平均距离。其中 3D 模型记作 M，x_1 表示组成真实位姿的 3D 点，x_2 表示组成预测位姿的 3D 点，真实目标刚体位姿分别用 R 和 T 表示旋转量和平移量，预测目标刚体位姿分别用 $\hat{R}$ 和 $\hat{T}$ 表示旋转量和平移量，则：

$$\omega_{\mathrm{ADD\text{-}S}}=\underset{x_1\in M}{\operatorname{avg}}\ \min_{x_2\in M}\left\|(Rx+T)-(\hat{R}x+\hat{T})\right\| \tag{2}$$

对于 LineMOD 和 LineMOD Occlusion 数据集，常将 ADD 度量用于处理非对称对象，ADD-S 度量用于处理对称对象，将此指标定义为 ADD(-S)指标，通常将 ADD(-S)指标的阈值设置为模型直径的 10%，即当度量指标小于模型直径的 10%，则认为成功预测，否则预测失败。

曲线区域面积(AUC)为 ADD-S 或 ADD(-S) 指标形成的曲线包围的区域面积，常应用于 YCB-Video 数据集中。对于 YCB-Video 数据集，通常认为 ADD-S 或 ADD(-S)小于 2cm 时为正确，对应的 AUC 的横轴的最大阈值为 10cm。

上述评价指标较为常用，也有部分评价指标使用较少，如 2D 投影指标，该度量以目标刚体的三维模型为基础，将模型的顶点投影到真实和预测位姿的图像平面上。一般情况下，当刚体全部顶点的平均 2D 投影误差小于 5 个像素时，常认为预测位姿正确。对应顶点的投影距离确定如下：

$$\omega_{\mathrm{2DProj}}=\frac{1}{|V|}\sum_{v\in V}\left\|\hat{K}\hat{R}v-KRv\right\|_2 \tag{3}$$

其中，V 是所有目标刚体模型顶点的集合，$\boldsymbol{K}$ 是相机矩阵。

n cm $n°$指标对 ADD 指标进行了改进，明确定义了公差，分别对平移信息和旋转信息进行处理。一般情况下，当预测平移信息与真实平移信息差值小于阈值，预测旋转信息与真实旋转信息插值也小于阈值时，则认为预测的位姿正确。

表 8、表 9 分别列出了经典的刚体位姿估计方法在 LineMOD 数据集、LineMOD Occlusion 数据集以及 YCB-Video 数据集上的表现效果，其中 * 表示使用了位姿优化技术进行处理后的效果。部分方法提供了推理时间(ms)。为了观测其

运行速度，本文将运行时间统一用处理图像的平均帧率即FPS来计算，同时还需注意刚体的位姿估计评价较为复杂，例如还需考虑方法的学习方式、解决的挑战数目、不同挑战的难度不同等因素，表中结果仅代表性能的一部分，具体方法的性能需要根据综合结果来判断。

表8 各方法在LineMOD和LineMOD Occlusion数据集上的性能指标

Table 8 Performance indicators on LineMOD and LineMOD Occlusion datasets of each method

方法	LineMOD		Occlusion	
	ADD(-S)/%	FPS	ADD(-S)/%	FPS
Brachmann et al.(2016)[13]	32.30	2	—	—
Brachmann et al.(2017)[43]	—	—	76.70	—
PoseCNN(2017)[34]	—	—	24.90	4
BB8(2017)[9]	43.60	4	62.70	3
SSD-6D(2017)[33]	76.30	12	—	—
Heatmaps(2018)[28]	—	—	25.80	4
YOLO-6D(2018)[39]	55.95	**50**	6.42	—
CDPN(2019)[41]	89.86	33	—	—
DPOD(2019)[24]	82.98	40	32.79	—
DenseFusion(2019)[10]	86.20	20	—	—
PVNet(2019)[27]	86.27	25	40.77	—
Hu 等(2019)[15]	—	—	27.00	20
Pix2Pose(2019)[22]	72.40	8～10	32.00	—
G2L-Net(2020)[25]	98.70	23	—	—
Hu 等(2020)[16]	—	—	43.30	**45**
HybridPose(2020)[54]	94.50	30	**79.20**	—
AAE(2020)[32]	31.41	42	—	—
PVN3D(2020)[31]	**99.40**	5	—	—
GDR-Net(2021)[23]	93.70	45	62.20	28
FS-Net(2021)[70]	97.60	20	—	—
SD-Pose(2021)[48]	94.30	—	—	—
SSD-6D*(2017)[33]	90.90	10	—	—
Heatmaps*(2018)[28]	—	—	30.40	—
DPOD*(2019)[24]	95.15	33	47.25	—
DenseFusion*(2019)[10]	94.30	16	—	—

表9 各方法在YCB-Video数据集上的性能指标

Table 9 Performance indicators on YCB-Video dataset of each method

方法	ADD(-S)/%	ADD-S/%	ADD(-S) AUC/%	ADD-S AUC/%	FPS
PoseCNN(2017)[34]	—	—	61.3	75.9	5.0
DeepIM(2018)[38]	—	—	81.9	88.1	—
Heatmaps(2018)[28]	33.6	53.1	—	72.8	—
PVNet(2019)[27]	—	—	73.4	—	—
Hu 等(2019)[15]	39.0	—	—	—	—
DenseFusion(2019)[10]	—	95.3	—	91.2	20.0
G2L-Net(2020)[25]	—	—	—	92.4	**21.0**
Hu 等(2020)[16]	53.9	—	—	—	—
PVN3D(2020)[31]	—	—	91.8	95.5	—
MoreFusion(2020)[42]	—	—	91.0	95.7	—
GDR-Net(2021)[23]	**60.1**	—	84.4	91.6	—
PoseRBPF(2021)[71]	—	—	—	75.4	11.5
PoseCNN*(2017)[34]	—	93.2	—	93.0	0.1
DenseFusion*(2019)[10]	—	**96.8**	—	93.1	16.0
CosyPose*(2020)[59]	—	—	84.5	89.8	—
PVN3D*(2020)[31]	—	—	**92.3**	**96.1**	—

结束语 基于深度学习的位姿估计方法因其强大的学习能力和在处理刚体弱纹理、遮挡等方面的优越性而成为了计算机视觉领域内热门的研究领域。本文介绍了基于深度学习的刚体位姿估计方法的发展，依据实现方式将其分为基于坐标、基于关键点和基于模板的位姿估计方法；详细介绍了位姿估计方法的实现过程、面临的挑战以及相应的解决办法，分析了不同类别方法的优缺点和适用场景，对比分析了常用数据集以及相关算法在常用数据集上的实验结果；最后分析了基于深度学习的位姿估计方法的未来发展趋势。计算机视觉在工业生产、服务、军事等领域的应用日益广泛，位姿估计的应用也随之大幅增加。尽管近年来基于深度学习的位姿估计已经取得了较大进步，但仍有很大的发展空间。

(1)位姿跟踪。现有的位姿跟踪方法较少，且位姿估计虽在图像中已具有较好表现，但在视频中，会出现跟踪位姿不断抖动，在实际应用中会导致效率较低，尤其是在辅助装配领域，无法将其作为静物，会产生较大干扰。

(2)弱监督位姿估计。其根本目的是减少人工注释的工作量，使用少量注释图像来估计非注释的目标刚体位姿。因此，设计弱监督的位姿估计是未来位姿估计研究的一个重要问题。

(3)无监督位姿估计。研究自动标注技术可以极大地提升工作效率，节省大量时间且无须考虑数据不充分的问题。对于位姿估计中此类注释难以标注的问题，无监督位姿估计无疑是未来的一个研究方向。

(4)基于GAN的位姿估计。基于GAN的位姿估计可利用真实场景和模拟数据进行位姿估计，Pix2Pose方法虽适用于GAN，但效果欠佳，故设计GAN来进行位姿估计，有助于获取更好的鲁棒性，应对位姿估计面临的挑战。

(5)类别级位姿估计。现有的位姿估计绝大部分仍为实例级的位姿估计方法，对于类别级的位姿估计，现有方法均利用共享类内空间来解决，但此类方法对于类内形状变换较大的刚体效果极差，故类别级的位姿估计是未来研究一个可考虑的方向。

(6)端到端的位姿估计。现有的效果较好的位姿估计方法大多利用RANSAC和PnP对映射关系进行处理，而此类方法在此阶段不可微，影响位姿估计的训练速度，故可以考虑设计网络来实现位姿估计。

(7)可变形体和铰接体的位姿估计。现有方法绝大部分为刚体的位姿估计方法，但在实际应用中，还存在较多可变形体和铰接体需进行位姿估计，故对可变形体和铰接体的位姿估计研究是未来研究中的一项工作。

(8)基于点云数据的位姿估计。由于激光传感器的费用过高，实际应用较少，故本文未对由其获取的基于点云数据输入的位姿估计方法进行讨论，未来随着激光传感器的发展，其费用会降低，对基于点云数据的位姿估计方法的研究也会增多。

(9)基于强化学习的位姿估计。强化学习可以应用在位姿估计的多个步骤中，如训练智能体进行点云配准可以加快推理速度[72]，使用强化学习进行位姿优化可以不依赖真实6D姿态数据[73]，减少数据标注的工作等。与基于深度学习的位姿估计相比，引入强化学习可以降低对真实6D姿态数据的依赖[74]，实现快速推理[75]等，进一步提升模型的鲁棒性与性能，因此基于强化学习的位姿估计任务也是未来研究中的一项工作。

(10)轻量级的位姿估计。在硬件资源有限的设备上，现有的位姿估计算法不能够同时满足实时性和准确性[76]，故未来有必要设计轻量级的位姿估计网络以应用到实际工程项目中。

基于深度学习的位姿估计方法的研究还需进一步深入，希望未来位姿估计可以为机器人领域和泛虚拟现实领域做出更多的贡献。

参 考 文 献

[1] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the IEEE International Conference on Computer Vision. Kerkyra: IEEE, 1999: 1150-1157.

[2] BRÉGIER R, DEVERNAY F, LEYRIT L, et al. Defining the Pose of any 3D Rigid Object and an Associated Distance[J]. International Journal of Computer Vision, 2018, 126(6): 571-596.

[3] WOHLHART P, LEPETIT V. Learning Descriptors for Object Recognition and 3D Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3109-3118.

[4] COLLET A, BERENSON D, SRINIVASA S S, et al. Object Recognition and Full Pose Registration from a Single Image for Robotic Manipulation[C]// IEEE International Conference on Robotics & Automation. Kobe: IEEE, 2009: 48-55.

[5] DETRY R, PUGEAULT N, PIATER J H. A Probabilistic Framework for 3D Visual Object Representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1790-1803.

[6] GU C, REN X. Discriminative Mixture-of-Templates for Viewpoint Classification[C]// European Conference on Computer Vision. Berlin: Springer, 2010: 408-421.

[7] SHI Y, HUANG J, XU X, et al. StablePose: Learning 6D Object Poses from Geometrically Stable Patches[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 15222-15231.

[8] CORONA E, KUNDU K, FIDLER S. Pose Estimation for Objects with Rotational Symmetry[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018: 7215-7222.

[9] RAD M, LEPETIT V. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3828-3836.

[10] WANG C, XU D, ZHU Y, et al. Densefusion: 6D Object Pose Estimation by Iterative Dense Fusion[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3343-3352.

[11] WANG C, MARTÍN-MARTÍN R, XU D, et al. 6-PACK: Category-Level 6D Pose Tracker with Anchor-Based Keypoints [C]// 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris: IEEE, 2020: 10059-10066.

[12] BRACHMANN E, KRULL A, MICHEL F, et al. Learning 6D Object Pose Estimation Using 3D Object Coordinates[C]// European Conference on Computer Vision. Cham: Springer, 2014: 536-551.

[13] BRACHMANN E, MICHEL F, KRULL A, et al. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3364-3372.

[14] HODAN T, HALUZA P, OBDRŽÁLEK Š, et al. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects [C]// 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa: IEEE, 2017: 880-888.

[15] HU Y, HUGONOT J, FUA P, et al. Segmentation-Driven 6D Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3385-3394.

[16] HU Y, FUA P, WANG W, et al. Single-Stage 6D Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2930-2939.

[17] PHAM Q H, NGUYEN T, HUA B S, et al. Jsis3D: Joint Semantic-Instance Segmentation of 3D Point Clouds with Multi-Task Pointwise Networks and Multi-Value Conditional Random Fields[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8827-8836.

[18] PHAM Q H, UY M A, HUA B S, et al. LCD: Learned Cross-Domain Descriptors for 2D-3D Matching[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 11856-11864.

[19] SAHIN C, GARCIA-HERNANDO G, SOCK J, et al. A Review on Object Pose Recovery: From 3D Bounding Box Detectors to Full 6D Pose Estimators[J]. Image and Vision Computing, 2020, 96: 103898.

[20] DU G, WANG K, LIAN S, et al. Vision-Based Robotic Grasping from Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review[J]. Artificial Intelligence Review, 2021, 54(3): 1677-1734.

[21] YANG B Y, DU X P, WAN Z Q, et al. A Review of Attitude Estimation Methods for Rigid Object in Single Image[J]. Journal of Image and Graphics, 2021, 26(2): 334-354.

[22] PARK K, PATTEN T, VINCZE M. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 7668-7677.

[23] WANG G, MANHARDT F, TOMBARI F, et al. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 16611-16621.

[24] ZAKHAROV S, SHUGUROV I, ILIC S. DPOD: 6D Pose Object Detector and Refiner[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 1941-1950.

[25] CHEN W, JIA X, CHANG H J, et al. G2L-Net: Global to Local

Network for Real-Time 6D Pose Estimation With Embedding Vector Features[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4233-4242.

[26] WANG H, SRIDHAR S, HUANG J, et al. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2642-2651.

[27] PENG S, LIU Y, HUANG Q, et al. PVNet: Pixel-Wise Voting Network for 6D of Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4561-4570.

[28] OBERWEGER M, RAD M, LEPETIT V. Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation[C]// European Conference on Computer Vision. Munich: Springer, 2018: 119-134.

[29] YANG Z, YU X, YANG Y. DSC-PoseNet: Learning 6D of Object Pose Estimation via Dual-Scale Consistency[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3907-3916.

[30] HE Y, HUANG H, FAN H, et al. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3003-3013.

[31] HE Y, SUN W, HUANG H, et al. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6Dof Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11632-11641.

[32] SUNDERMEYER M, MARTON Z C, DURNER M, et al. Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection[J]. International Journal of Computer Vision, 2020, 128(3): 714-729.

[33] KEHL W, MANHARDT F, TOMBARI F, et al. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1521-1529.

[34] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes[J]. arXiv: 1711.00199, 2017.

[35] FISCHLER M A, BOLLES R C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.

[36] LEPETIT V, MORENO-NOGUER F, FUA P. Epnp: An Accurate O(N) Solution to the PnP Problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.

[37] MAKAY B P. A Method for Registration of 3D Shape[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14: 239-256.

[38] LI Y, WANG G, JI X, et al. DeepIM: Deep Iterative Matching for 6D Pose Estimation[C]// European Conference on Computer Vision. Munich: Springer, 2018: 683-698.

[39] TEKIN B, SINHA S N, FUA P. Real-time Seamless Single Shot 6D Object Pose Prediction[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 292-301.

[40] CHEN D, LI J, WANG Z, et al. Learning Canonical Shape Space for Category-Level 6D Object Pose and Size Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11973-11982.

[41] LI Z, WANG G, JI X. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-D of Object Pose Estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 7678-7687.

[42] WADA K, SUCAR E, JAMES S, et al. MoreFusion: Multi-Object Reasoning for 6D Pose Estimation from Volumetric Fusion [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 14540-14549.

[43] MICHEL F, KIRILLOV A, BRACHMANN E, et al. Global Hypothesis Generation for 6D Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 462-471.

[44] MANHARDT F, KEHL W, NAVAB N, et al. Deep Model-Based 6D Pose Refinement in RGB[C]// European Conference on Computer Vision. Munich: Springer, 2018: 800-815.

[45] LABBÉ Y, CARPENTIER J, AUBRY M, et al. CosyPose: Consistent Multi-View Multi-Object 6D Pose Estimation[C]// European Conference on Computer Vision. Cham: Springer, 2020: 574-591.

[46] PARK K, MOUSAVIAN A, XIANG Y, et al. LatentFusion: End-To-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10710-10719.

[47] CAI M, REID I. Reconstruct Locally, Localize Globally: A Model Free Method for Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3153-3163.

[48] LI Z, HU Y, SALZMANN M, et al. SD-Pose: Semantic Decomposition for Cross-Domain 6D Object Pose Estimation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021: 2020-2028.

[49] REDMON J, FARHADI A. Yolov3: An Incremental Improvement[J]. arXiv: 1804.02767, 2018.

[50] UMEYAMA S. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991, 13(4): 376-380.

[51] QI C R, SU H, MO K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 652-660.

[52] HODAŇ T, VINEET V, GAL R, et al. Photorealistic Image Synthesis for Object Instance Detection[C]// 2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019: 66-70.

[53] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger

[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7263-7271.

[54] SONG C, SONG J, HUANG Q. HybridPose: 6D Object Pose Estimation under Hybrid Representations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 431-440.

[55] GEORGAKIS G, KARANAM S, WU Z, et al. Learning Local RGB-to-CAD Correspondences for Object Pose Estimation [C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 8967-8976.

[56] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[J]. arXiv: 1706.02413, 2017.

[57] SUNDERMEYER M, DURNER M, PUANG E Y, et al. Multi-path Learning for Object Pose Estimation Across Domains[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 13916-13925.

[58] PITTERI G, RAMAMONJISOA M, ILIC S, et al. On Object Symmetries and 6D Pose Estimation from Images[C]// 2019 International Conference on 3D Vision(3DV). Quebec City: IEEE, 2019: 614-622.

[59] NAVANEET K L, MATHEW A, KASHYAP S, et al. From Image Collections to Point Clouds with Self-Supervised Shape and Pose Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1132-1140.

[60] MANHARDT F, ARROYO D M, RUPPRECHT C, et al. Explaining the Ambiguity of Object Detection and 6D Pose from Visual Data[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 6841-6850.

[61] LI S F, SHI Z L, ZHUANG C G. Deep Learning-Based 6D Object Pose Estimation Method from Point Clouds[J]. Computer Engineering, 2021, 47(8): 216-223.

[62] LI X, WANG H, YI L, et al. Category-Level Articulated Object Pose Estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 3706-3715.

[63] PAVLASEK J, LEWIS S, DESINGH K, et al. Parts-Based Articulated Object Localization in Clutter Using Belief Propagation [C]// 2020 IEEE International Conference on Intelligent Robots and Systems(IROS). Las Vegas: IEEE, 2020: 10595-10602.

[64] CHI C, SONG S. GarmentNets: Category-Level Pose Estimation for Garments via Canonical Space Shape Completion[J]. arXiv: 2104.05177, 2021.

[65] WANG G, MANHARDT F, SHAO J, et al. Self6D: Self-Supervised Monocular 6D Object Pose Estimation[C]// European Conference on Computer Vision. Cham: Springer, 2020: 108-125.

[66] 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]// Asian Conference on Computer Vision. Berlin: Springer, 2012: 548-562.

[67] KASKMAN R, ZAKHAROV S, SHUGUROV I, et al. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. Seoul: IEEE, 2019: 2767-2776.

[68] YUAN H, HOOGENKAMP T, VELTKAMP R C. RobotP: A Benchmark Dataset for 6D Object Pose Estimation[J]. Sensors, 2021, 21(4): 1299.

[69] LI C, BAI J, HAGER G D. A Unified Framework for Multi-View Multi-Class Object Pose Estimation[C]// European Conference on Computer Vision. Munich: Springer, 2018: 254-269.

[70] CHEN W, JIA X, CHANG H J, et al. FS-Net: Fast Shape-based Network for Category-Level 6D Object Pose Estimation with Decoupled Rotation Mechanism[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1581-1590.

[71] DENG X, MOUSAVIAN A, XIANG Y, et al. PoseRBPF: A Rao-Blackwellized Particle Filter for 6-D Object Pose Tracking [J]. IEEE Transactions on Robotics, 2021, 37: 1328-1342.

[72] BAUER D, PATTEN T, VINCZE M. ReAgent: Point Cloud Registration Using Imitation and Reinforcement Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 14586-14594.

[73] SHAO J, JIANG Y, WANG G, et al. PFRL: Pose-free Reinforcement Learning for 6D Pose Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11454-11463.

[74] SOCK J, GARCIA-HERNANDO G, KIM T K. Active 6D Multi-Object Pose Estimation in Cluttered Scenarios with Deep Reinforcement Learning[C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Las Vegas: IEEE, 2020: 10564-10571.

[75] KRULL A, BRACHMANN E, NOWOZIN S, et al. PoseAgent: Budget-constrained 6D Object Pose Estimation via Reinforcement Learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6702-6710.

[76] JIANG M, CHEN Y, ZHOU Q h, et al. Lightweight Pose Estimation Network for Non-Cooperative Target Acquisition[J]. Computer Engineering, 2022, 48(6): 235-242.

GUO Nan, born in 1977, Ph.D, associate professor. Her main research interests include computer vision, virtual reality, security and privacy.

（责任编辑：何杨）