机器人抓取中物体定位算法概述

本文同步于微信公众号：**3D视觉前沿**，欢迎大家关注。

1. 引言

机器人抓取的首要任务，是确定要抓什么，也就是需要定位目标物体在输入数据中的位置。这个过程可以分为三个层次，分别为物体定位但不识别、物体检测、物体实例分割。物体定位但不识别是指获得目标物体的2D/3D范围但是不知道物体的类别；目标检测是指得到目标物体的2D/3D包围盒，同时识别目标物体的类别；目标实例分割提供目标物体所占有的像素或者点级别的区域信息，同时识别目标物体的类别。本文来自论文https://arxiv.org/abs/1905.06658，涉及的论文也都可以在论文中找到，也包含于https://github.com/GeorgeDu/vision-based-robotic-grasping，本文就不引用了。

2. 定位但不识别

当不知道目标物体的类别时，仍然可以采用一些方法获得目标物体的2D/3D区域，进而支撑机器人抓取。当我们知道物体的外轮廓形状时，可以采用拟合形状基元的方法。当我们什么信息都没有时，可以采用显著性物体检测方法，获得潜在可能的待抓取物体区域。

2.1 基于2D RGB图像的方法

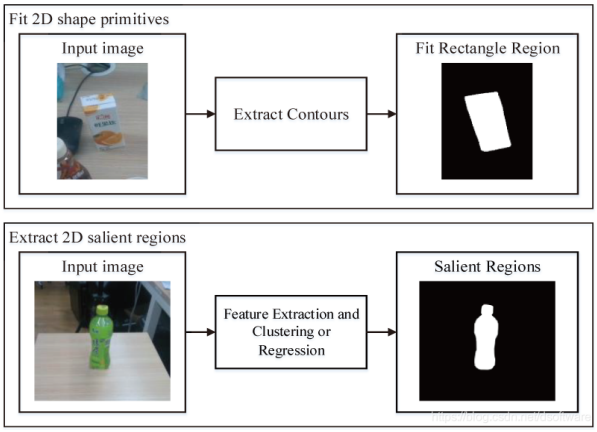

图*1* 基于*2D*图像的定位但不识别方法

a.拟合形状基元： 目标物体的形状可以是一个椭圆、一个多边形或者一个四边形，这些2D形状可以看作是形状基元，通过拟合方法，可以定位到目标物体。该方法的一般步骤为先提取出图像的所有封闭轮廓，其次再用拟合方法得到潜在的可能目标物体，如果存在多个候选，可以使用模板匹配去除干扰。在OpenCV中已经集成了例如拟合椭圆、拟合多边形这样的函数。

b.显著性区域检测： 和特定形状基元相比，显著性区域可以是任意形状。2D显著性区域检测的目的是定位和分割出给定图像中，最符合视觉显著性的区域，这更像一个分割任务。非深度学习的方法主要挖掘低层次的特征表示，或者依据一些经验例如颜色对比、形状先验来得到显著性区域。基于深度学习的方法主要包括基于多层感知机(MLP)的方法，基于全卷积网络(FCN)的方法和基于胶囊网络的方法。

在目前的机器人抓取任务中，该方法仍处在初级阶段。在工业领域，如果待抓取物体形状固定且轮廓清晰，可以采用拟合形状基元的方法。在另外一些机器人抓取任务中，如果背景的颜色信息和目标物体的颜色信息差别较大，也可以去除背景得到目标物体。在Dex-Net2.0中，目标物体放置在绿色背景的桌面上，通过背景颜色分离，可以得到目标物体。

2.2 基于3D点云的方法

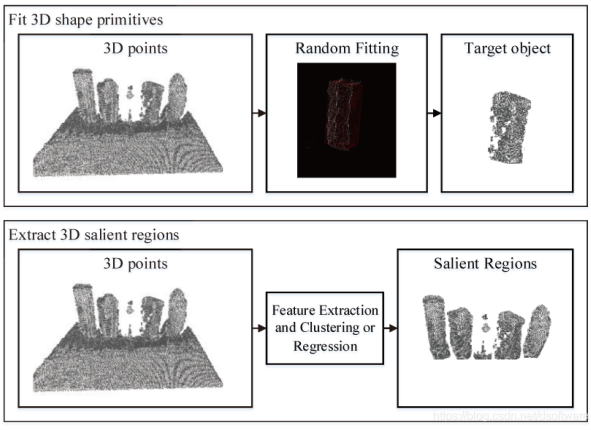

图*2* 基于*3D*点云的定位但不识别方法

a.拟合3D形状基元： 目标物体的形状可以是一个球体、圆柱体或者长方体，这些3D形状可以看作是3D形状基元。存在很多方法能够拟合，包括随机采样点一致(RANSAC)算法，基于霍夫投票(Hough Voting)的方法以及一些其他的聚类算法。针对机器人抓取任务，输入数据是单个视角下不完整的点云，目标是找到一部分点云，使得这些点云能够组合成某一个3D形状基元。一些方法针对输入的点云拟合平面并进行装配组合；一些方法使用长方体拟合室内场景数据；一些方法通过霍夫变换提取点云中的圆柱体，还有一些方法先把背景分割掉，然后对剩下的多个目标物体再拟合形状。

b.3D显著性区域检测： 这类包括基元RGB-D数据的方法和基元3D点云的方法。RGB-D显著性检测方法通过人工设计的或者基于深度学习的方法提取特征并进行融合。基于3D点云的方法主要是提取一个完整物体点云的显著性图谱，而针对机器人抓取任务中是要从3D输入中得到目标物体的3D点区域。一些方法首先去除背景3D点，之后依赖3D特征如曲率等对显著性进行评分，得到显著性区域。

在目前的抓取任务中，该方法被广泛使用，但仍处在初级阶段。通常都在结构化的场景中，利用先验去除背景包含点云(可以利用高度信息，也可以将当前输入和已有的背景3D模型配准去除背景)，之后对于多个目标物体包含的点云进行聚类或者拟合，得到目标物体包含的3D点云。

3. 目标检测

当我们知道要抓取的目标物体的类别时，可以使用2D/3D目标检测以及实例分割算法，获得目标物体的2D/3D包围盒区域或者Mask区域。这里先介绍物体检测算法。目标物体检测是指不仅要定位出目标物体的位置，还要识别出物体的类别，位置通常用2D/3D最小包围盒表示。根据是否生成区域候选，目标检测的方法可以分为两阶段法(Two-stage Method)和单阶段法(One-stage Method)。

3.1 2D目标检测

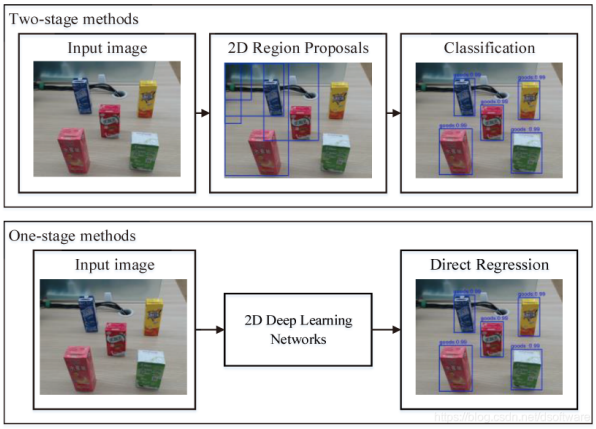

图*3* 基于*2D*图像的目标检测方法

a.两阶段法： 这类方法也称为基于区域候选(Region proposal-based)的方法。大多数传统方法使用滑动窗口策略获得候选包围盒，然后使用包围盒内的特征描述进行识别。大量人工设计的2D描述符，例如SIFT, FAST, SURF, ORB等，常被使用来训练分类器，例如使用神经网络，SVM，Adaboost等。传统方法的缺陷在于需要根据特定物体设定滑动窗口的大小，而且人工设计的特征表征能力不强，不足够支撑一个很强的分类器。

随着深度学习的发展，深度神经网络能够用于计算候选区域，而且能够提取表征能力更强的特征描述符，训练的分类器能力更强(R-CNN)。不仅如此，深度神经网络基于候选区域学习到的特征向量可以直接回归物体的类别，比训练分类器得到了更好的识别效果(Fast R-CNN)。Faster R-CNN进一步提出了候选区域生成网络，允许端到端训练整个目标检测网络。一般而言，两阶段法得到的精度相对较高，但是需要更多的计算资源和计算时间。

b.单阶段法： 这类方法也称为基于回归(Regression-based)的方法。这类方法跳过了候选区域生成步骤，直接在一次估计中预测包围盒以及类别得分。YOLO是代表性单阶段方法，划分网格并同时预测多个包围盒和类别概率。不过由于每个网格只回归两个包围盒，算法不适合小物体。SSD能够为固定集合的锚点包围盒(anchor boxes)预测类别得分和包围盒偏移，比YOLO效果要好。YOLOv2也使用了锚点滑动窗口，效果比YOLO提升。RetinaNet提出了focal loss损失函数用于训练，达到了和两阶段算法相当的精度但速度更快。YOLOv3基于YOLOv2进行了一些优化，达到了更好的效果。此外，存在不使用锚点的单阶段方法，例如FCOS、CornerNet、ExtremetNet、CenterNet等，直接预测单个点处潜在物体的类别概率，以及距离完整包围盒的相对位置。

3.2 3D目标检测

2D目标检测预测的包围盒，能够将物体在2D空间占据的区域完全包含。而3D目标检测旨在寻找到目标物体在3D空间中的完整(amodel)3D包围盒，也即完整的3D物体的最小外接长方体。通常获取的数据是单个视角下的RGB-D数据，通用的3D检测算法都可以使用。但由于目前仅基于RGB进行3D检测与基于3D点云进行3D检测的效果差别较大，本文主要面向机器人抓取领域，因此这里只介绍基于点云的3D检测方法。

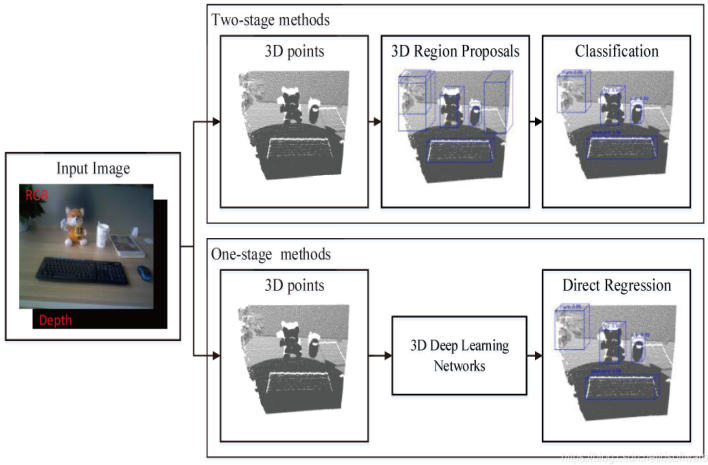

图*4* 基于*3D*点云的目标检测方法

a.两阶段法： 这类方法也称为基于区域候选(Region proposal-based)的方法。传统的3D检测方法主要针对已知形状的物体，3D检测问题就转化为了物体的6D位姿估计问题。此时，可以寻找3D特征点之间的对应，将已有物体和观测数据进行配准，也可以进行基于RANSAC的对齐完成全局配准。然而这些方法要求观测的实例和已有物体模

型几何结构一致，不适用于通用物体3D检测。通用的3D检测任务和2D检测任务类似，也广泛使用3D区域候选。传统方法使用人工设计的3D特征，例如Spin Images, 3D Shape Context, FPFH, CVFH, SHOT等，训练诸如SVM之类的分类器完成3D检测任务，代表方法为Sliding Shapes。

随着深度学习的发展，3D区域候选可以被有效地生成，而且3D包围盒可以被直接回归得到，而不必训练分类器。生成3D区域候选的方法，大致可以分为三类：基于截锥体(frustum-based)的方法，基于全局回归(global regression-based)的方法和基于局部回归(local regression-based)的方法。基于截锥体的方法是指使用成熟的2D目标检测算法来获取物体的3D候选区域，方法最为直接。Frustum PointNets使用2D检测器获得物体的2D区域，对应3D空间是一个截锥体，对里面的点云基于PointeNet进行语义分割，再用MLP回归得到最终的3D包围盒信息；FrustumConvNet对由2D区域得到的3D截锥体进一步划分成多个3D候选区域。基于全局回归的方法根据基于单源或多源输入学习到的特征描述，直接回归得到3D区域候选。Deep Sliding Shapes提出了第一个3D候选区域生成网络，并且结合物体识别网络来回归3D包围盒；MV3D融合多源信息预测3D候选；MMF进行多特征多传感器融合，完成3D检测任务；PartA2使用encoder-decoder网络，对输入的点云预测物体部位的相对位置，预测3D候选包围盒。基于局部回归的方法是指产生逐点的3D区域候选。PointRCNN从输入点云中提取逐点的特征向量，对分割得到每一个前景点，生成3D候选。这些候选经过池化以及正则优化得到最终结果。STD设计了球形锚点生成基于点的候选，并进一步生成稠密的候选包围盒的特征描述，预测最终的包围盒。VoteNet提出深度霍夫投票策略，对采样的每个3D种子点，生成对应的3D投票点，并进一步聚类得到候选包围盒。IMVoteNet结合2D图像，进一步提升了VoteNet的精度。

b.单阶段法： 这类方法也称为基于回归(Regression-based)的方法，通过单个网络直接预测3D包围盒及其类别概率，不需要生成候选3D包围盒以及后处理。VoxelNet将输入点云划分成3D voxels，并且将每个voxel内的点云学习一个统一的特征表示，再用卷积层和候选生成层得到最终的3D包围盒。SECOND相比于VoxelNet，使用稀疏卷积层来划分紧致的体素特征。PointPillars将点云转换成一个稀疏的伪图像，再结合2D卷积网络和检测头预测3D包围盒。以上主要是基于体素的单阶段3D检测器，3DSSD是基于点云的3D检测器，使用了一种融合采样策略，一个候选生成层和一个锚点无关的回归层，实现了精度和速度的平衡。

3D检测能够提供目标物体的大致位置，主要应用在无人车领域，但不足够支撑复杂的抓取。目前的通用3D检测只预测一个平面上的3D包围盒的1个旋转角度，范围在0度-180度(原地旋转180度，包围盒位置朝向不变)，假定物体结构复杂，例如带柄水杯，只给一个中心加旋转角度是不知道如何去抓的。但是3D包围盒能够提供大致的物体抓取位置，而且能够用于机械臂移动避障。

4. 物体实例分割

进行目标检测只能得到矩形框区域，而一些任务需要只属于目标物体的区域，这就需要进行实例分割。物体实例分割是指检测到特定类别物体的逐像素或者逐点的实例，这项任务与目标检测和语义分割极其相关。实例分割算法也分为两阶段法和单阶段法，两阶段法先生成区域候选，单阶段法直接回归。

4.1 2D实例分割

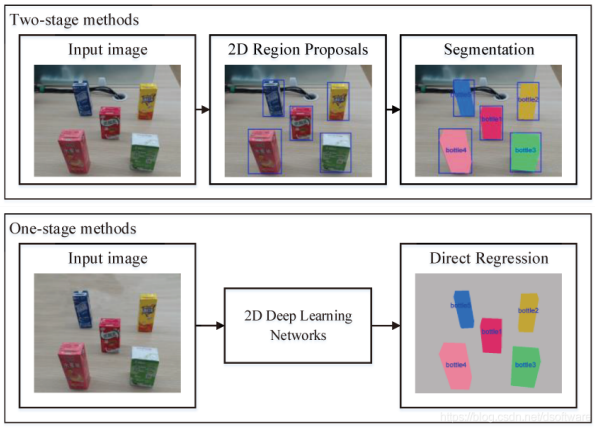

图*5* 基于*2D*图像的实例分割方法

a.两阶段法： 这类方法也称为基于区域候选(Region proposal-based)的方法。成熟的2D检测器常被用来生成包围盒或者区域候选，在他们内部可以进一步计算物体的mask区域。许多方法都基于CNN。SDS使用CNN来识别类别无关的区域候选；MNC通过三个网络进行实例分割；PANet提出路径增强网络提升实例分割效果 ；Mask R-CNN通过增加额外的分支预测物体的mask扩展了Faster R-CNN；HTC对目标检测结果和分割结果进行层叠式优化；PointRend通过迭代细分策略进一步增强了细节的分割效果。

b.单阶段法： 这类方法也称为基于回归(Regression-based)的方法，同时预测分割的mask和存在物体的得分。DeepMask, SharpMask和InstanceFCN为位于中心的物体预测mask；FCIS进行实例级语义分割，能够预测位置敏感的得分进行物体分割和检测；TensorMask在空间域使用结构化的4D张量代表masks，能够预测稠密的masks；YOLACT将实例分割分成两个平行的子任务，即生成一些列原型masks并且预测每个实例mask的系数；YOLACT是第一个实时的单阶段实例分割算法，后续有改进版YOLACT++；PolarMask利用极坐标系，预测实例物体的中心和稠密距离，得到实例物体的轮廓；SOLO引入实例类别的概念，依据每个实例的位置和大小，赋予属于实例物体的每个像素一个类别，将实例分割问题转换成分类问题；CenterMask在目标检测算法FCOS的基础上增加了SAG-Mask分支；BlendMask也基于FCSO，使用了一个渲染模块来预测位置敏感的特征，并且学习每个实例物体的注意力图。

2D实例分割在机器人抓取应用中被广泛使用，如果场景中同一个类别的物体只有一个实例，语义分割也可以。由于输入是RGB-D图像，有了RGB分割结果，可以快速得到对应Depth图像，得到目标物体的3D点云。例如，SegICP使用基于RGB的2D目标分割算法获得只属于目标物体的点，再使用配准方法得到目标物体的6D位姿。很多方法也结合Depth输入共同完成目标分割任务。

4.2 3D实例分割

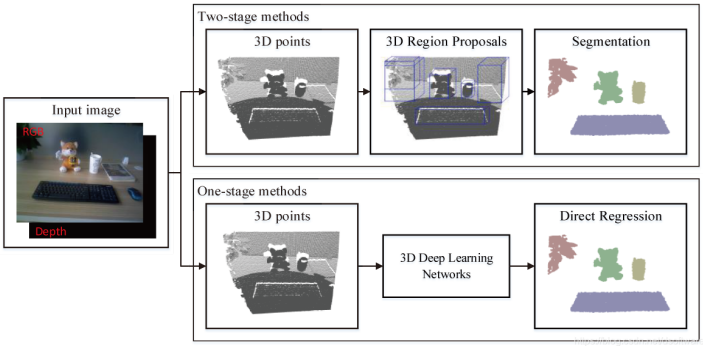

图*6* 基于*3D*点云的实例分割方法

a.两阶段法： 这类方法也称为基于区域候选(Region proposal-based)的方法。一般的方法借助2D/3D检测结果，再对对应3D截锥体或者包围盒区域进行前后景分割得到目标物体的点云。GSPN提出通用形状候选网络生成3D物体的候选区域，进一步使用区域PointNet进行3D物体的实例分割；3D-SIS使用2D和3D融合特征进行3D物体包围盒检测以及语义实例分割。

b.单阶段法： 这类方法也称为基于回归(Regression-based)的方法。很多方法学习如何归类逐点的特征来完成3D实例分割。SGPN提出了相似群候选网络来预测合并后的候选，并给每个候选一个对应的语义类别，也就完成了3D实例分割；MASC使用子空间稀疏卷积预测每个点的语义得分以及邻近点在不同尺度下的紧密关系得分，基于此可以合并得到实例分割结果；ASIS学习具有语义的逐点级别的实例嵌入，之后属于同一个实例的语义特征再进行融合；JSIS3D提出了一个多任务的逐点的网络结构，能够同时预测3D点属于的物体类别，并且将3D点嵌入为更高维的特征向量，进一步聚类得到物体实例；3D-BoNet能够回归所有实例的3D包围盒，同时预测每个实例的逐点的mask；LiDARSeg提出了一种稠密的特征编码框架以及有效的策略解决类别不平衡问题。

3D实例分割在机器人抓取应用中非常重要，如果场景中同一个类别的物体只有一个实例，语义分割也可以。当前常用的做法仍然是利用成熟的2D分割算法，对应到深度图获取3D点云，但是在其分割效果不如直接在3D点云上进行实例分割。随着算法效果和性能的提升，3D实例分割会在未来广泛使用。

5. 总结

给定机器人抓取场景，目前总能找到合适的技术方案来定位目标物体的位置，进而执行后续的物体位姿估计以及机器人抓取位姿估计等任务，但仍存在一些问题。定位但不识别的算法，要求物体在结构化的场景中或者物体与背景具有显著性差异，这些都有益于算法提取出目标物体，这就限制了应用场景；实例级目标检测算法，需要实例级目标物体的大量训练集，而且训练好的检测算法只在训练集上检测精度高；如果需要对新物体进行检测，需要再采集大量的数据进行重新训练，这个过程非常耗时耗力；通用的目标检测算法，泛化能力强，但其精度达不到实例级目标检测算法。实例分割算法也面临同样的问题。对大量训练数据集的需求以及泛化性能差，也是深度学习算法的通用性问题。不过针对限定场景下特定的一些物体，当前算法已经能够得到非常好的满足落地需求的结果。