深圳大学

SHENZHEN UNIVERSITY

计算机前沿技术

研究生创新示范课程研究报告

姓名：肖非凡

学号：2210273065

2022 年 12 月 27 日

Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter[1]

Michel Breyer,Jen Jen Chung

摘要

一般的机器人在混乱中抓取需要综合抓取能力，这些抓取能力适用于以前看不见的物体，并且对物理交互（如与场景中其他物体的碰撞）也很鲁棒。在这项工作中，我们设计并训练了一个网络，该网络根据从车载传感器（如手腕安装的深度相机）收集的 3D 场景信息预测 6 自由度抓取。我们提出的体积抓取网络（VGN）接受场景的截断有符号距离函数（TSDF）表示，并直接输出所查询的 3D 体积中每个体素的预测抓取质量和相关的抓取器方向和开口宽度。我们表明，我们的方法可以在仅 10ms 内规划抓取，并且能够在真实世界的杂波去除实验中清除 92% 的对象，而无需明确的碰撞检查。实时能力为闭环抓取规划提供了可能性，允许机器人处理干扰、从错误中恢复并提供更强的鲁棒性。

关键词：抓取合成；3D 卷积神经网络

1 引言

传统上，机器人操作主要考虑了在严格控制的空间中执行的重复任务。然而，最近人们对将机器人部署到需要更多灵活性的领域产生了浓厚的兴趣。例如，辅助机器人可以通过接管医院获取供应的耗时任务来缓解医务人员的工作量。为了在这种非结构化环境中表现良好，系统必须能够为它可能遇到的大量对象计算抓取，同时处理车载传感器的杂波、遮挡和高维噪声读数。

由于这些挑战，最近在抓取合成方面的研究压倒性地支持直接从传感器数据计划抓取的数据驱动方法，优于手动设计的策略。最近几项著名的工作涉及全 6 自由度（DOF）抓取姿势检测。然而，这些方法通常会导致单个、孤立的对象在放置在场景中时需要额外的碰撞检查，或者它们相当高的计算时间（以秒为单位）使得它们不适合闭环执行，这对于更高级的交互和对动态变化的反应至关重要。

在这项工作中，我们提出了一种新的实时 6 自由度抓取合成方法。我们算法的输入是一个三维体素网格，其中每个单元包含到最近曲面的截断距离。我们训练全卷积网络（FCN）将输入 TSDF 映射到具有相同空间分辨率的体积，其中每个单元包含在体素中心执行的抓取的预测质量、方向和宽度。

总而言之，这项工作的贡献是一种新颖的抓取网络：1) 能够实现实时的 6D 抓取合成。2) 使用完整的 3D 场景信息直接学习无碰撞抓取策略。

2 相关工作

2.1 深度学习在抓取领域的研究现状

美国康奈尔大学 Lenz 等人[2]提出了一种基于 RGBD 图像的级联神经网络模型对目标物体进行抓取规划。整个系统分为候选抓取集采样和抓取评估选择两部分，每一部分分别由一个级联神经网络完成。候选抓取集采样是从图片中所有可能的抓取区域进行评分筛选分数相对较高的少部分，然后再进入第二个级联神经网络进行更加精细的评分和筛选. 如图 3所示

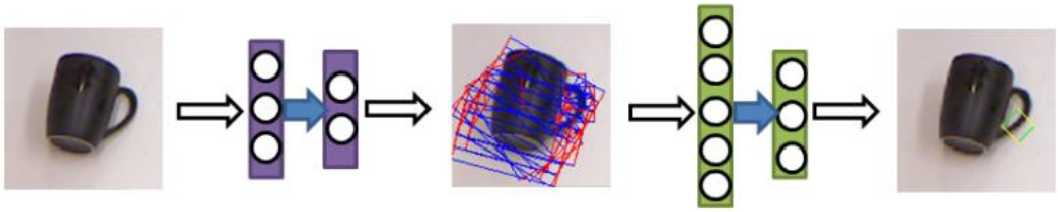

图 1: 双级联神经网络抓取模型

谷歌、剑桥大学以及加州大学伯克利分校成立研究团队提出一种基于 Q 函数的强化学习算法训练真实环境中的机器人进行抓取，共同探讨如何将深度强化学习应用在机器人操作中。美国华盛顿大学 Joseph Redmon 等人提出一种基于卷积神经网络的实时、准确抓取检测算法，可以在不使用标准滑动窗口或者区域建议技术的情况下对可抓取的边界框执行单阶段回归。同时，该网络可以同时执行分类，以便在一个步骤中识别对象并找到一个良好的抓取矩形，对该模型的修改通过使用局部约束的预测机制来预测每个对象的多个抓取，如图 3所示。局部约束模型的表现明显更好，特别是对于可以通过各种方式掌握的物体.

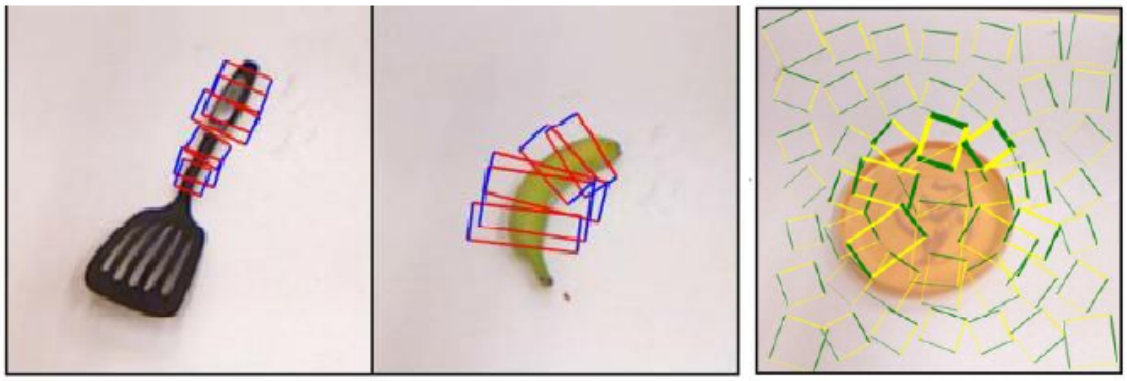

图 2: 华盛顿大学对于具有多抓取物体的抓取检测

3 本文方法

3.1 本文方法概述

此部分对本文将要复现的工作进行概述。

Observations: 根据深度相机捕获的具有已知内在和外在的深度图像，我们将其融合成 TSDF，该 TSDF 是一个 N^3 的体素网格 V，其中每个单元 V_i 包含到最近曲面的截断的有符号距离。TSDF 的空间数据表示结构能够很好的表示空间信息，有助于提高抓取检测的表现，将深度图片转化为 TSDF 体素之后，我们将它作为模型的输入。

Grasp: 我们将 6-DOF 抓取 g 定义为抓取中心位置 $t \in R^3$、抓取的方向 $r \in SO(3)$ 和手指之间的开口宽度 $w \in R$。

关于模型的架构 VGN，本文主要采用的是全连接卷积网络 (FCN)。首先感知模块由 3 个分别有 16，32，64 的 filters 的卷积层组成，将输入的 TSDF 数据映射到 64×5^3 的特征图上。第二部分由三个三卷积层和双线性上采样交织组成，最后输出三头数据。模型架构如图 3所示：

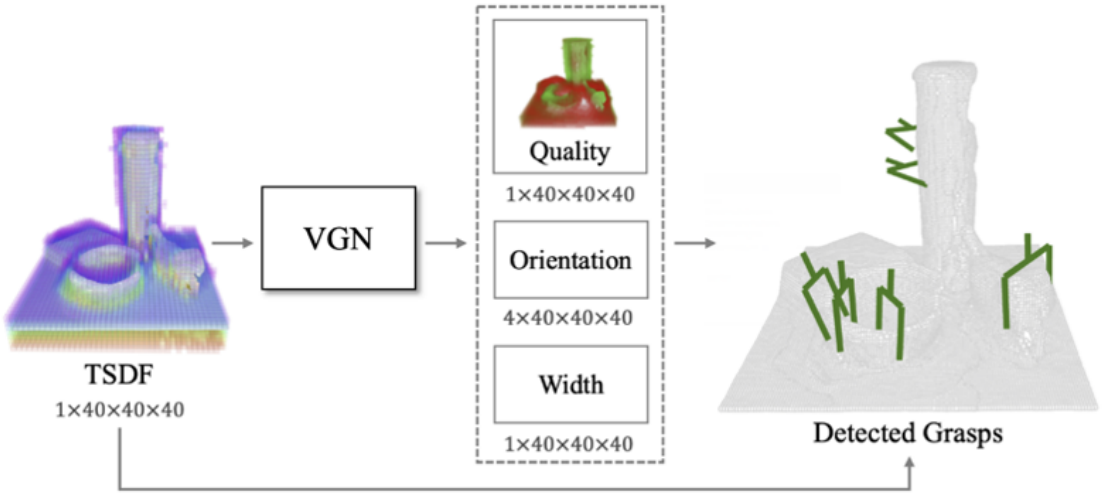

图 3: 方法示意图

3.2 问题定义

我们的目标就是要寻找个一个映射，能够将输入的数据映射为能够表示抓取。VGN 模型输出的抓取由机器人坐标系下夹持器的位置 t，方向 r，以及夹持器的张开宽度 w 来进行表示。输出的每一个姿态都对应着一个标量 $q \in [0,1]$，代表着抓取成功的可能性。

下式为各个参数的转换公式。

$$t = T_{RV}(t)/v, r = T_{RV}(r), w = w/v \tag{1}$$

3.3 损失函数定义

损失函数的定义如下：

$$L(\hat{g}_i, \tilde{g}_i) = L_q(\hat{q}_i, \tilde{q}_i) + q_i(L_r(\hat{r}_i, \tilde{r}_i) + L_w(\hat{w}_i, \tilde{w}_i)) \tag{2}$$

其中，$q_i \in \{0,1\}$ 表示目标抓取 $\hat{g}_i$ 的真值抓取标签, L_q 是预测和真实标签 $\hat{q}_i$ 和 $\tilde{q}_i$ 之间的二元交叉熵,L_w 预测宽度 $\hat{w}_i$ 和目标宽度 $\tilde{w}_i$ 之间的均分误差（MSE）。我们使用内积来计算预测的四元数和目标四元数之间的距离,$L_{quat} = 1 - |\hat{r} \cdot \tilde{r}|$。所以关于旋转的损失函数为

$$L_r(\hat{r}, \tilde{r}) = min(L_{quat}(\hat{r}, \tilde{r}), L_{quat}(\hat{r}, \tilde{r}_\pi)) \tag{3}$$

4 复现细节

4.1 与已有开源代码对比

复现论文的作业提供了源代码，所以本次实验在原作者的基础上进行了代码的重构与改进。在网络架构的实现部分，此部分的代码（即编码器、解码器等）由本人进行了重新编写。并且在论文的 VGN 网络架构的基础上，对于网络进行了改进，增加了全连接卷积网络的层数，增添了残差网络连接，并尝试采用卷积占用网络 (CONet) 进行建模。

伪代码如下所示:

Procedure 1 Volumetric Grasping Network for grasping.

Input: a *TSDF*, an N^3 voxel grid V
Output: a grasp G
for i ***in*** *target frame indices* **do**
$V_{in} = \text{Resnet}(\text{Encoder}(V))$
$V_{out} = \text{Decoder}(V_{in})$ $Quality = \text{Sigmoid}(\text{Conv3d}(V_{out}))$ $Orientation = \text{Normalize}(\text{Conv3d}(F_{out}))$ $Width = \text{Conv3d}(F_{out})$
end

4.2 实验环境搭建

我们在 PyBullet 中模拟实际设置建立了一个模拟环境，允许我们生成抓取试验的大型数据集。我们从不同的来源组装了一组 343 个对象网格，并将它们分为 303 个训练对象和 40 个测试对象。整个实验是用 Python 语言实现的。网络训练和模拟实验是在配备了 GeForce GTX 3090 显卡的计算机上进行的。

4.3 界面分析与使用说明

下图给出了抓取场景 (场景不同位置的抓取质量) 的例子，并在图中预测了较优抓取。我们的第一个观察是，我们的方法能够考虑到场景的上下文并预测无碰撞抓取。其中对象的一部分是可抓住的 (红色)，而对称部分不是 (灰色)，并且抓住这些灰色区域可能会导致与邻近对象的碰撞。这表明，我们的模型能够理解来自自我监督抓取试验的训练的场景信息，并在进行抓取预测时考虑了实际约束。

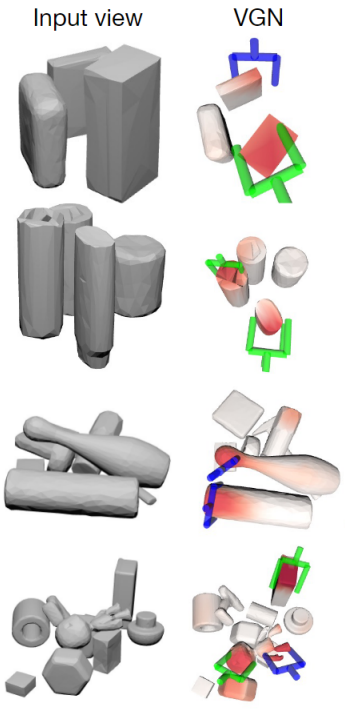

图 4: 操作界面示意

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。

表 1: 实验结果为抓取的成功率，分别采用了 $\varepsilon = 0.95, 0.90, 0.80$ 随机选择其预测抓取质量高于某个阈值的抓取，实验场景为 5 个物体，并进行了 200 轮的模拟训练。

	Blocks	Pile
VGN(ε =0.95)	83.2	47.5
VGN(ε = 0.90)	78.4	52.8
VGN(ε = 0.80)	74.4	41.1

6 总结与展望

本次课程设计的目的是对于该篇论文工作进行一个复现的工作，在实验过程中，我们对于论文本身的思路进行了详细的梳理并进行了代码的撰写，对于一些重要的代码比如模拟数据的生成，模拟抓取图像的生成等我们采用了原作者的代码，而在网络架构方面我们进行了代码的重构。往往在模拟环境之中的数据在现实环境中并不能得到很好的匹配，但由于资源的限制，本次实验并没有像原作者进行真机实验，所以我们只有模拟环境的实验结果。本次实验只考虑了一种类型的夹爪，该体积抓取网络是否能够胜任其他形状的抓手也是我们可以去考虑的事情。

参考文献

[1] BREYE M. Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter[J]. In Conference on Robot Learning, 2020.

[2] JIANG Z. Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations[J]. robotics science and systems, 2021.