D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions 复现

何绮宁

摘要

本文是 2022 CVPR 论文《D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions》的复现报告。该论文提出了一个两阶段的动态抓取框架，即抓握阶段以及移动阶段，以单帧静态抓握参考标签以及物体目标位置作为输入，生成抓取物体并移动到指定位置的手部动作序列。对原论文的结果进行了复现。本文对于原文中使用静态抓握标签的设置进行了改进，通过预训练的 Pointnet++，鼓励策略考虑待抓取物体几何信息。但本实现中同样存在抓取不稳定以及自然性不足等缺陷，未来将会针对这两个问题进行进一步的探究。

关键词：手物交互；物体动态抓取；物理动画

1 引言

人类与场景的交互是计算机图形学中一个长久以来的关键问题。而手作为人类与周围环境交互的主要手段，引起了人们对手-物姿态估计，以及静态抓握姿势生成的浓厚兴趣。但对于给定的物体，用灵巧手进行抓握并非是静态的，而是涉及到一个连续的时间过程，包括抓住、拿起以及移动物体。手需要在整个抓握过程中保持稳定，这涉及到了手-物之间的物理相互作用，如碰撞与摩擦力。复杂的动态为这个问题带来了挑战性。

2 相关工作

2.1 人类抓握预测

近年来，手-物交互任务受到了广泛的研究关注。一方面，出现了许多带有手部与物体姿态注释的数据集；另一方面，借此产生了许多直接由 RGB 图像预测出手-物抓握姿态的工作。有人根据手和物体的网格预测抓握姿势[1]，也有工作给定物体，预测其 6DOF 位姿以及对应的手部姿势[2][3]。更直接相关的是生成给定物体的静态抓握的方法，这类方法有时也会加入已知的手部信息。这类方法一般生成手上的接触图[4][5]，或者直接生成手部的关节角配置[6][7]，也有一些工作结合二者，通过接触信息来预测手部姿势[8]。同样，也有工作加入了全身动作条件，给定全身动作以及物体生成合理的局部抓握动作[9]。上述工作集中于静态抓握，并且纯粹是数据驱动的。本文的方法考虑了手-物交互中的动态抓握，并且通过物理模拟来考虑抓握过程中的物理合理性。

2.2 灵巧手操控

以合理地操作物体为目的，产生了许多使用灵巧手操控物体的工作，近年来基于学习的方法也为这方面的工作提供了帮助。一些工作基于专家演示，例如 Rajeswaran 收集专家运动轨迹并模仿它们[10]，Garcia-Hernando 从视频中获取专家演示并使用 RL 进行修正[11]。这种收集专家演示的方法较昂贵，而本文方法只需要每个运动序列的单帧抓取标签，不需要将整个轨迹作为参考，节省成本。Mandikal 提

出了一个专注于学习抓取物体的框架[12]，但只鼓励物体在某个区域被抓住，并未考虑手部姿势的自然性。在他们的后续工作中，额外提出了一个基于手-物交互视频的奖励来解决此问题[13]。而本文解决此问题的方法是通过明确地调节所需的接触点和手的姿势，来产生更自然的抓握。

2.3 物理感知推理

纯粹的数据驱动方法模仿专家演示的手部姿势，可能产生抓取不稳定。许多工作开始引入物理模拟与感知来改善该类问题，例如 Mezghanni 使用物理模拟验证抓取的合理性[14]，Ehsani 通过物理模拟来监督与推理手-物抓握过程中的接触和力[15]，如 2.2 中提到，在一些从视频重建人类抓握姿态的任务中，也增加了基于物理的模块来修正人体姿态的估计。本文方法同样使用物理模拟，但涉及到更细粒度的控制以及抓取物体之后的位移，对手-物之间的物理感知提出了更高的要求。

3 本文方法

3.1 本文方法概述

本文的任务是对于给定的静态抓握标签以及目标物体位置，生成动态的手-物交互序列。为此，本文提出了一个基于 RL、由低级抓取策略以及高级运动生成模块组成的层次框架。以静态抓握标签作为输入，在抓取阶段，只有低级抓取策略起作用，对于指定物体生成一个合理且稳定的手部抓握姿势；在运动合成阶段，加入了用户指定的物体移动位置，抓取策略和运动合成模块同时起作用，生成 action $\boldsymbol{a}$。将 $\boldsymbol{a}$ 传递给 PD 控制器以计算所需力矩 $\boldsymbol{\tau}$，并在物理模拟环境中操纵 MANO 手部模型，更新手部与物体状态，直到完成抓取物体并移动到目标位置的任务。

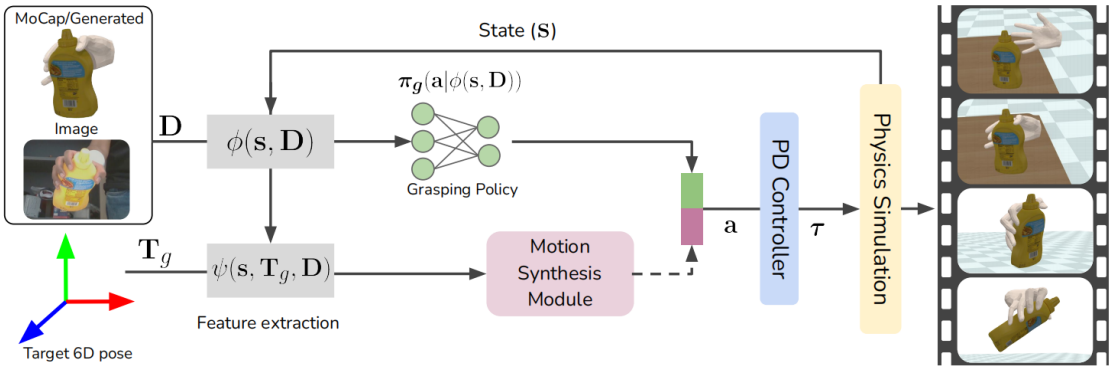

图 1: 方法示意图

3.2 符号定义

在本任务中，输入是一个静态抓握参考标签 $\boldsymbol{D} = (\bar{\boldsymbol{q}}_h, \bar{\boldsymbol{T}}_h, \bar{\boldsymbol{T}}_o)$，其中 T 和 q 分别代表 6D 全局位姿以及 3D 局部姿态，其中 6D 位姿由旋转和平移分量组成 $\boldsymbol{T} = [\boldsymbol{q}|\boldsymbol{t}]$，下标 o 和 h 代表物体与手，即静态抓握标签定义了作为抓取参考的手部与物体位姿。给定静态抓握标签 $\boldsymbol{D}$，目标是抓住物体并将其移动到一个 6D 目标姿态 $\boldsymbol{T}_g$。

3.2.1 State Space

本文将该问题定义为一个马尔科夫决策过程，并使用 RL 进行网络训练。为了训练能够形成稳定合理的抓握的策略，任务状态的表示至关重要。本文将状态表示为 $\boldsymbol{s} = (\boldsymbol{q}_h, \dot{\boldsymbol{q}}_h, \boldsymbol{f}, \boldsymbol{T}_h, \dot{\boldsymbol{T}}_h, \boldsymbol{T}_o, \dot{\boldsymbol{T}}_o)$，$\boldsymbol{q}_h$ 和 $\dot{\boldsymbol{q}}_h$ 分别表示手部关节的角度以及角速度，$\boldsymbol{T}_h$ 和 $\boldsymbol{T}_o$ 分别表示手腕以及物体的 6D 全局位姿，$\dot{\boldsymbol{T}}_h$ 和

$\dot{\boldsymbol{T}}_o$ 为对应的速度；$\boldsymbol{f}$ 为手和物体之间的相互作用力。上述状态参数可以在物理模拟环境中取得。

3.2.2 Action Space

本文在物理模拟中定义了一个动作空间 a 来控制手部模型，表现为一个 51 维向量，其中 6 维表示手部整体的姿态，其余 45 维分别为手部 45 个关节执行器的姿态。本文使用 PD 控制器，以参考关节角度 $\boldsymbol{q}_{ref}$ 作为输入，计算对应的关节扭矩 $\boldsymbol{\tau} = k_p(\boldsymbol{q}_{ref} - \boldsymbol{q}) + k_d\dot{\boldsymbol{q}}$。本方法的参考关节角度不由策略直接得出，而是由当前关节配置 $\boldsymbol{q}_b = \boldsymbol{q}_h$ 与策略输出 a 相加而得：$\boldsymbol{q}_{ref} = \boldsymbol{q}_b + \boldsymbol{a}$。原文发现，与直接预测 $\boldsymbol{q}_{ref}$ 的策略相比，输出偏置值鼓励生成更平滑的手指运动，进而导致更稳定的抓取。

3.3 特征提取层

原文方法并没有将状态 s 直接作为策略 π 的输入，而是应用了一个特征提取层 $\phi(\boldsymbol{s}, \boldsymbol{D})$，以状态 s 和静态抓握标签 $\boldsymbol{D}$ 作为输入，提取出的特征作为策略 π 的输入，即

$$\phi(\boldsymbol{s}, \boldsymbol{D}) = (\boldsymbol{q}_h, \dot{\boldsymbol{q}}_h, \boldsymbol{f}, \tilde{\boldsymbol{T}}_h, \tilde{\boldsymbol{T}}_o, \dot{\tilde{\boldsymbol{T}}}_o, \dot{\tilde{\boldsymbol{T}}}_h, \tilde{x}_o, \tilde{x}_z, \boldsymbol{G})$$

，前三项的符号定义参考前文，$\tilde{\boldsymbol{T}}_o$ 和 $\dot{\tilde{\boldsymbol{T}}}_o$ 代表手腕坐标系下物体的位姿以及对应速度，$\tilde{\boldsymbol{T}}_h$ 和 $\dot{\tilde{\boldsymbol{T}}}_h$ 表示相对于初始手腕姿势的当前手腕姿势与对应速度，$\tilde{x}_o$ 和 $\tilde{x}_z$ 分别表示相对于初始物体位置的当前物体姿势，以及物体当前相对于桌面的距离。最后一项是针对静态抓握参考标签的 $\boldsymbol{G} = [\tilde{\boldsymbol{g}}_x|\tilde{\boldsymbol{g}}_q|\tilde{\boldsymbol{g}}_c]$，分别表示当前手部姿态与参考标签手部姿态的距离差 $\tilde{\boldsymbol{g}}_x$、角度差 $\tilde{\boldsymbol{g}}_q$ 以及是否接触指定点的 one-hot 向量 $\tilde{\boldsymbol{g}}_c$，鼓励手部模型接近参考标签指定的手部姿态，以及到达物体上的指定接触点。

3.4 奖励函数定义

本文方法的奖励函数定义如下：

$$r = w_x r_x + w_q r_q + w_c r_c + w_{reg} r_{reg}$$

包括位置项、角度项、接触项和正则化项，并使用 w_x, w_q, w_c, w_{reg} 等权重因子衡量各奖励之间的重要程度。位置项奖励计算当前每个手部关节位置与目标关节位置之间距离的加权和，即

$$r_x = \sum_{j=1}^{J} w_{x,j} ||\bar{\boldsymbol{x}}_j - \boldsymbol{x}_j||^2$$

；同理，角度项奖励以欧拉角测量当前手部姿态与相应的目标姿态之间的距离，对应的是特征 $\tilde{\boldsymbol{g}}_q$ 的 L2-范数：

$$r_q = ||\tilde{\boldsymbol{g}}_q||$$

；接触项奖励分为两部分，第一部分表示手部模型到达指定接触点的比例，由 one-hot 特征 $\tilde{\boldsymbol{g}}_c$ 表示，第二部分奖励施加在指定接触点上的力，鼓励手在指定位置施加力以抓紧物体。该奖励表示为

$$r_c = \frac{\tilde{\boldsymbol{g}}_c{}^T \boldsymbol{I}_{\boldsymbol{f}>0}}{\tilde{\boldsymbol{g}}_c{}^T \tilde{\boldsymbol{g}}_c} + \min(\tilde{\boldsymbol{g}}_c{}^T \boldsymbol{f}, \lambda m_o)$$

；最后一项奖励为手与物体的线速度和角速度的正则项：

$$r_{reg} = w_{reg,h} ||\dot{\tilde{\boldsymbol{T}}}_h||^2 + w_{reg,o} ||\dot{\tilde{\boldsymbol{T}}}_o||^2$$

4 复现细节

4.1 与已有开源代码对比

本次复现中使用了原作者给出的源代码，并在此基础上进行改进。本文的方法实质上是在物理模拟环境中模仿单帧的静态抓握参考。该方法存在的问题是，无论在训练还是推断中，对于每个抓取都需要抓握的手部姿势作为输入。当遇到一些人类不方便抓取的物体时，便无法获得静态抓握参考，无法指导灵巧手进行合理且自然的交互。

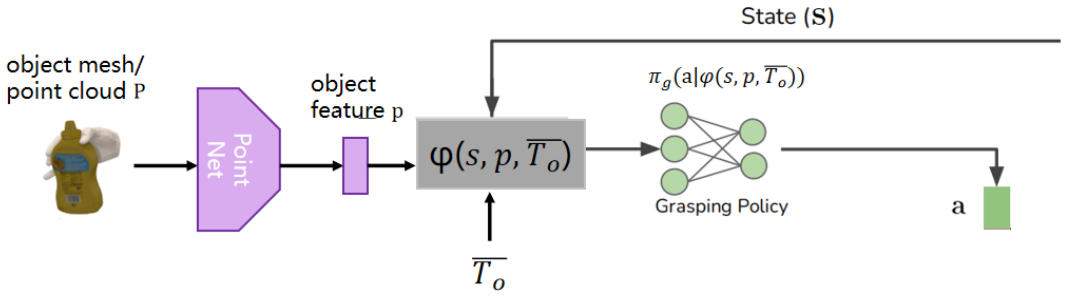

图 2: 特征提取层改进示意图

本次复现对该问题进行了改进。在原文方法中，物体的几何信息载入了物理模拟环境，但并没有作为网络的输入，换而言之，网络并没有考虑到物体本身的信息。针对上面的问题，本次的改进思路是提取出物体的几何特征，用其取代手部参考作为网络输入的一部分，实现对于该类物体几何信息的感知。为了实现该效果，本次复现使用了 pre-trained 的 Pointnet++ 中的 classification 分支，并且使用本文数据集中的物体点云信息对 Pointnet++ 进行微调，将其提取出的 128 维特征代替手部参考作为 D-grasp 中特征提取层的输入。

4.2 实验环境搭建

本次复现的环境为 Ubuntu 20.04，python3.8，使用 Raisim 作为物理模拟环境，在 NVIDIA GeForce RTX 3070 Ti 上进行训练。Raisim 从源码开始构建，安装需要安装的依赖为：eigen 库，cmake>3.10，并进行环境变量的配置。根据 Raisim 官网的指导，进入 RaisimLib 文件夹并进行编译以及安装。为了构建可视化环境，还需要下载 minizip 与 ffmpeg 包。安装上述包后，需要在 Raisim 官网申请激活码，进行可视化环境 RaisimUnity 的激活。

4.3 界面分析与使用说明

在 dgrasp/raisimUnity/linux 路径下可以找到可视化环境 RaisimUnity 的可执行文件，双击打开可得到如 3所示的界面。在保持该界面打开的情况下，运行 runner_motion.py，可以读取预训练模型；点击该界面左上角的“connect”按钮，即可将抓取过程可视化。

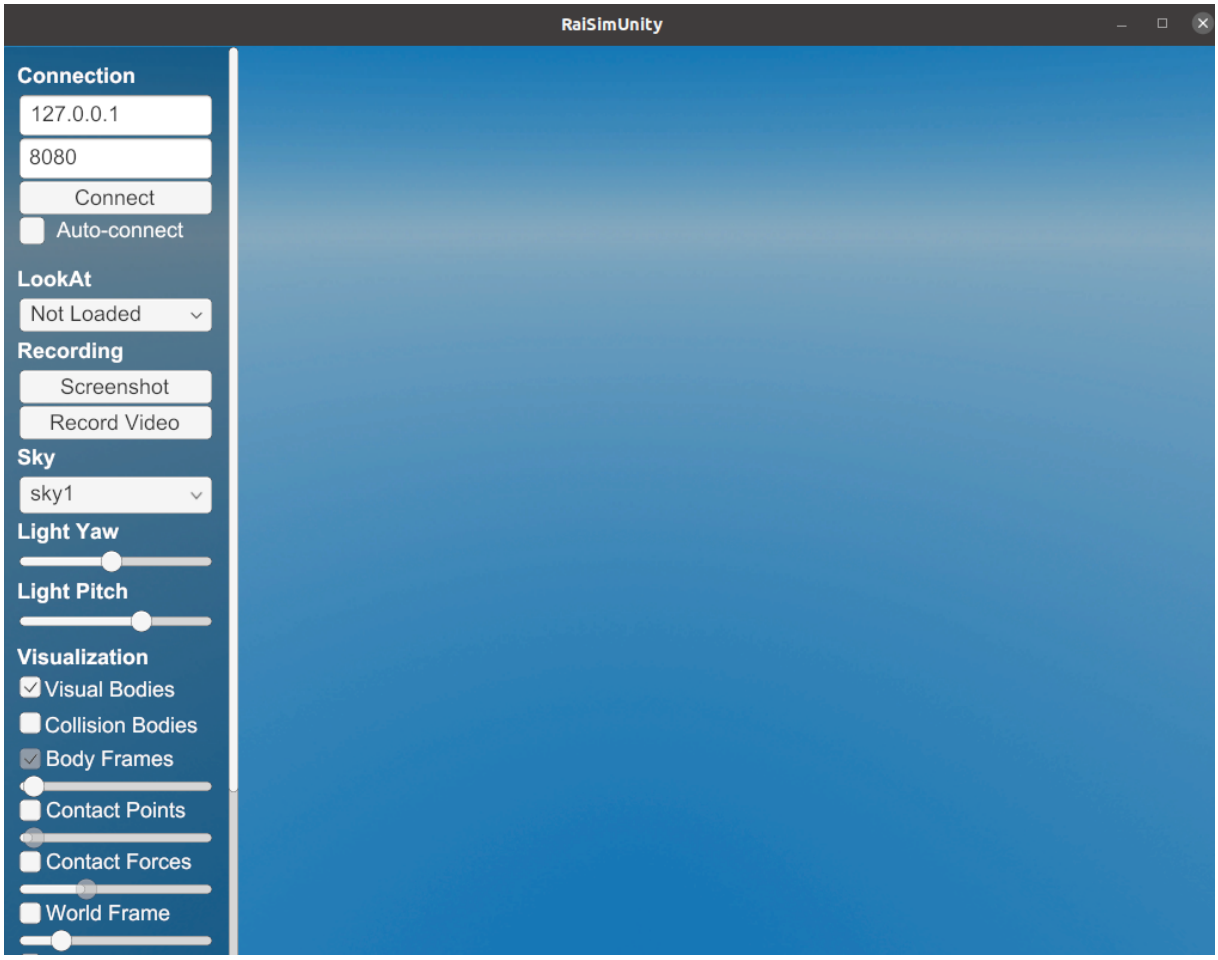

图 3: 操作界面示意

4.4 创新点

本次复现的创新点在于对策略输入的改进，将静态抓握标签的手部姿态信息去除，更改为使用预训练的 Pointnet++ 提取出的物体几何特征。经过改进后，策略能够充分考虑物体的几何信息，不是单纯地拟合抓握标签中的手部姿态。

5 实验结果分析

本次复现使用了原作者提供的代码，完成了 Raisim 的物理模拟环境搭建，并且使用作者提供的预训练模型重现了论文中的结果。在此基础上，新增了用于提取物体几何特征的 Pointnet++ 模块以及自定义的 Dataloader，并且将特征提取层的输入更改为物体几何特征。对于每个单物体策略，训练 3000 epoches 需要 9 ∼ 12 小时。

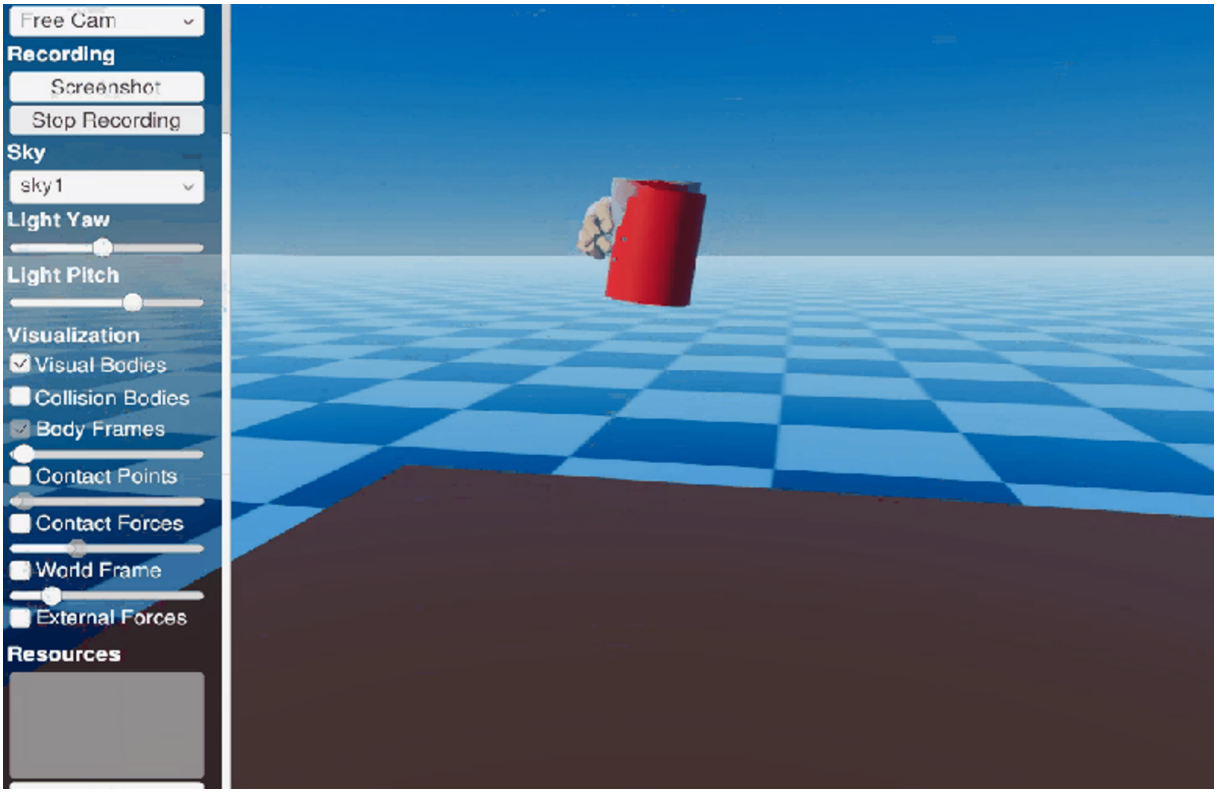

图 4: 实验结果示意-罐头

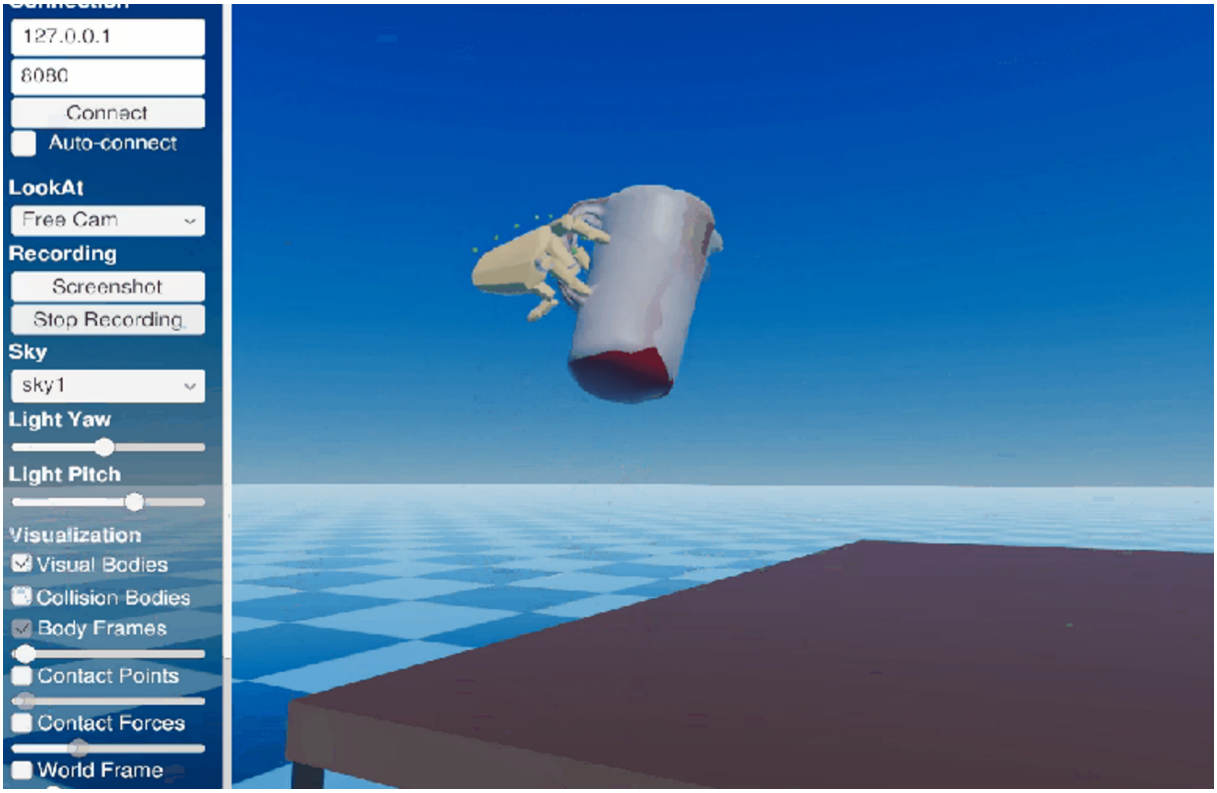

图 5: 实验结果示意-水壶

本次复现的可视化结果如图 4和 5所示。如图所示，以物体几何特征和物体目标位置为特征提取层输入的策略能够准确地抓起物体，保持稳定的抓握姿势，并将物体移动到指定位置。本次复现对抓取的成功率进行了记录。根据原文中消融实验的结果，对比了完全移去特征提取层、将特征提取层的输入更改为物体几何特征、原文的抓握成功率，结果如下表所示：

Models	Success ↑	SimDist [mm/s] ↓	Contact Ratio ↑
w/o ContactRew	0.0	24.18 ± 1.58	0.02
w/o GoalSpace	0.28	14.21 ± 10.50	0.18
w/o FeatLayer	0.47	9.69 ± 10.26	0.21
w/o WristGuidance	0.58	7.88 ± 10.57	0.28
original	0.89	4.83 ± 1.71	0.43
Ours	0.72	5.14 ± 1.61	0.39

本次改进的重点在于如何使策略在推断的时候不需要输入人类抓握姿势作为参考，而是根据物体的几何特征选择合适的抓握方式。从上表可知，通过 3000 次的迭代，本次复现在减少输入的特征的情况下，仍能形成稳定的手-物动态抓握。本次复现的结果中，抓握成功率等参数均优于原文去除了特征提取层等多种设置，也证明了选择物体几何特征作为输入是有效的。

6 总结与展望

本文对 2022 CVPR 论文《D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions》进行了复现。该论文提出了一个两阶段的动态抓取框架，以单帧静态抓握参考标签以及物体目标位置作为输入，生成抓取物体并移动到指定位置的手部动作序列。本次复现工作中使用了作者提供的源代码，进行了环境的配置，成功复现了原论文中的结果。本文对于原文中使用静态抓握标签的设置进行了改进，用待抓取物体的几何特征作为替代。目前实现过程中，策略能够成功考虑物体本身的信息，实现了在推理时不需输入抓握标签作为参考，而是通过考虑 Pointnet++ 提取的物体特征进行抓握。但本实现中同样存在不足，一方面是抓握的稳定性有待提升，在将物体移动到目标位置的过程中有可能跌落；另一方面是抓握的手势自然性不足，存在某根手指姿势怪异的问题。前者可能可以通过增加抓握质量测度来作为奖励的一部分，鼓励形成更加稳定的高质量抓握；后者可以引入新的奖励或损失函数，惩罚不自然的动作；同样，使用更加有代表力的手-物接触特征也是一个可考虑的改进方向。未来将会针对这以上问题进行进一步的探究。

参考文献

[1] HASSON Y, VAROL G, TZIONAS D, et al. Learning Joint Reconstruction of Hands and Manipulated Objects[C]//. 2019: 11799-11808. DOI: 10.1109/CVPR.2019.01208.

[2] HASSON Y, TEKIN B, BOGO F, et al. Leveraging Photometric Consistency over Time for Sparsely Supervised Hand-Object Reconstruction[C]//. 2020.

[3] CAO Z, RADOSAVOVIC I, KANAZAWA A, et al. Reconstructing Hand-Object Interactions in the Wild[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 12397-12406. DOI: 10.1109/ICCV48922.2021.01219.

[4] BRAHMBHATT S, HAM C, KEMP C, et al. ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging[C]//. 2019.

[5] BRAHMBHATT S, TANG C, TWIGG C D, et al. ContactPose: A Dataset of Grasps with Object Contact and Hand Pose[C]//The European Conference on Computer Vision (ECCV). 2020.

[6] KARUNRATANAKUL K, SPURR A, FAN Z, et al. A Skeleton-Driven Neural Occupancy Representation for Articulated Hands[C]//2021 International Conference on 3D Vision (3DV). 2021: 11-21. DOI: 10.1109/3DV53792.2021.00012.

[7] KARUNRATANAKUL K, YANG J, ZHANG Y, et al. Grasping Field: Learning Implicit Representations for Human Grasps[C]//. 2020.

[8] GRADY P, TANG C, TWIGG C D, et al. ContactOpt: Optimizing Contact to Improve Grasps[C]// Conference on Computer Vision and Pattern Recognition (CVPR). 2021.

[9] YE Y, LIU C K. Synthesis of Detailed Hand Manipulations Using Contact Sampling[J/OL]. ACM Trans. Graph., 2012, 31(4). https://doi.org/10.1145/2185520.2185537. DOI: 10.1145/2185520.2185537.

[10] RAJESWARAN A, KUMAR V, GUPTA A, et al. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations[C]//. 2018. DOI: 10.15607/RSS.2018.XIV.049.

[11] GARCIA-HERNANDO G, JOHNS E, KIM T K. Physics-Based Dexterous Manipulations with Estimated Hand Poses and Residual Reinforcement Learning[C/OL]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE Press, 2020: 9561-9568. https://doi.org/10.1109/IROS45743.2020.9340947. DOI: 10.1109/IROS45743.2020.9340947.

[12] MANDIKAL P, GRAUMAN K. Learning Dexterous Grasping with Object-Centric Visual Affordances [C]//2021 IEEE International Conference on Robotics and Automation (ICRA). 2021: 6169-6176. DOI: 10.1109/ICRA48506.2021.9561802.

[13] MANDIKAL P, GRAUMAN K. DexVIP: Learning Dexterous Grasping with Human Hand Pose Priors from Video[J]. ArXiv, 2022, abs/2202.00164.

[14] MEZGHANNI M, BOULKENAFED M, LIEUTIER A, et al. Physically-Aware Generative Network for 3D Shape Modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 9330-9341.

[15] EHSANI K, TULSIANI S, GUPTA S, et al. Use the Force, Luke! Learning to Predict Physical Forces by Simulating Effects[C]//. 2020: 221-230. DOI: 10.1109/CVPR42600.2020.00030.