arXiv:2305.02554v1 [cs.RO] 4 May 2023

学习可泛化的枢转技能

Xiang Zhang[1], Siddarth Jain[2], Baichuan Huang[3],Masayoshi Tomizuka[1],和 Diego Romeres[2]

摘要—**利用机器人系统对物体进行枢转操作是一项具有挑战性的技能，因为作用于系统的外力主要由接触交互产生。当要求相同的技能在不同物体之间实现泛化时，其复杂性会进一步增加。本文提出了一种用于学习具有鲁棒性和泛化能力的枢转技能的框架，该框架包含三个步骤。首先，我们利用强化学习（RL）在一个“单一”物体上学习枢转策略。随后，通过监督学习获得该物体的特征空间，以编码任意物体的运动学属性。最后，为了使单一策略适用于多个物体，我们基于物体特征学习数据驱动投影，以调整新枢转任务的状态与动作空间。所提出的方法完全在仿真环境中训练。该方法仅需要物体的一张深度图像，即可实现向真实世界物体的零样本迁移。我们验证了该方法在仿真到现实迁移方面的鲁棒性，以及对多个物体的泛化能力。**

I. 引言

枢转等桌面操作技能通常用于重新调整物体的方向，从而为其他操作技能创造前置条件。例如，桌上的书可能太大，机器人无法抓取；销钉也可能方向不对，无法执行插入任务。然而，通过枢转运动重新调整书和销钉的方向，就能为成功抓取创造条件。图 1(a) 展示了物体位于两个外部表面之间时的枢转设置，机器人需要利用与这些表面的交互来枢转物体。枢转的一大难点在于，机器人必须维持物体与夹爪之间以及物体与表面之间的接触。此外，多个物体在运动学和惯性属性上的差异，还会带来不稳定、滑动和滚动等问题，增加操作的复杂性。

本文提出了一种框架，仅凭仿真经验即可学习能够泛化至真实世界物体的枢转机器人技能。具体而言，我们希望基于物体深度图像，将针对一个物体学习到的枢转策略迁移到多个物体上。如图 1(b) 所示，所提出的框架由三个部分组成。首先，使用强化学习学习枢转策略，以枢转一个特定物体，我们将其称为
这个“单一”物体。然后，为了从高维物体深度图像中提取低维物体信息，我们对仿真中采集的数据集采用监督学习，通过预测物体类别和尺寸来学习特征空间。最后，学习特定物体的状态与动作投影，以使单一策略适用于一个

[1]加州大学伯克利分校机械系统控制实验室，美国加利福尼亚州伯克利。{xiang zhang 98, tomizuka}@berkeley.edu[2]三菱电机研究实验室（MERL），美国马萨诸塞州剑桥 {sjain,romeres@merl}.com [3]罗格斯大学计算机科学系，美国新泽西州皮斯卡塔韦。baichuan.huang@rutgers.edu

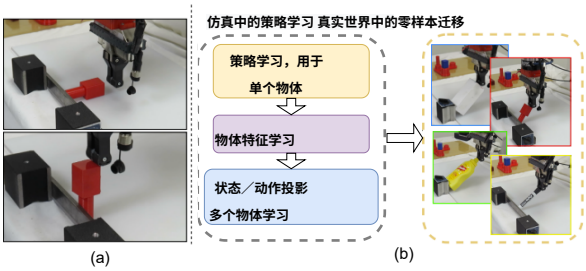


图 1：a）枢转任务设置。b）在仿真中进行策略学习，并零样本迁移至真实世界

通过调整状态与动作空间来适应新物体。这些投影是由物体特征得到的线性变换，并通过基于策略梯度的方法进行学习。直观来说，状态投影会调整新物体的状态，使其与单一物体相似。相应地，针对新物体的动作投影会改变策略输出，以提升策略性能。所提出的方法完全在仿真中训练，并通过零样本迁移应用于一系列真实世界的枢转任务。

总而言之，我们的工作作出了以下贡献：

- 提出了一个强化学习框架，仅通过仿真训练，即可学习适用于真实世界物体的泛化枢转技能。
- 提出一种基于策略梯度的方法，通过学习状态／动作变换，实现对未见物体的泛化。
- 在仿真和真实世界实验中对所提出的方法进行了全面评估，并取得了令人鼓舞的成功率。

II. 相关工作

A. 机器人接触丰富型操作

机器人接触丰富型操作的方法主要分为两类：基于模型的方法和无模型方法。基于模型的方法通过对接触动力学进行建模，设计开环和反馈控制策略。然而，由于接触过程中存在不同的接触模式，接触动力学具有混合特性。研究人员要么将混合接触动力学建模为互补约束 [1]–[3]，要么直接构建混合整数规划问题 [4]，以求得最优控制律。近期，一些方法 [5], [6] 采用分层框架，分别求解离散接触模式和机器人运动，以缩短计算时间。然而，基于模型的方法需要对初始条件和接触对作出严格假设，还需要进行系统辨识以获取物体的物理参数。

与基于模型的方法不同，无模型方法省去了复杂的接触动力学建模，直接学习策略来完成操作任务。近期，研究人员引入强化学习（RL）方法，通过优化设计的奖励来学习操作技能。相关实例包括机器人装配任务 [7]–[9]、双手操作 [10]和桌面操作 [11]。具体而言，[12] 的作者利用强化学习研究了枢转任务。然而，他们的枢转策略是在训练过程中针对单个物体学习得到的，只适用于尺寸相近的物体，因而限制了其应用。

B. 学习可泛化的机器人技能

泛化是机器人技能学习中一个重要的关注点，因为我们关注的是能够适应环境和任务条件变化的鲁棒机器人技能。事实上，不同物体具有不同的几何形状、物理属性和接触动力学。这些差异中的每一项都会改变机器人与物体的交互方式，使技能的泛化变得困难。研究人员提出了多种方法，用于学习能够适应不同任务设置或快速适应新任务的鲁棒技能。领域随机化 [13] 可用于提升鲁棒性，迫使学习到的策略从状态中提取有用信息。此外，模型无关元学习（MAML） [14]和在线模型学习 [15] 能够在少数几次试验内，使动力学模型或策略适应新任务。

另一种方法是将任务信息编码到策略中。因此，学习到的策略所生成的机器人动作会以任务为条件。这样，机器人便能根据不同任务调整技能，从而实现泛化。通常，可以从机器人轨迹、图像或独热编码中推断任务信息。这种方法在装配 [8]、抓取与放置 [16] 以及腿式运动 [17], [18]任务中也已被证明是成功的。

研究人员还将技能泛化视为一个领域适应问题。对于不同任务，状态与动作空间可能会根据任务设置（例如，机器人类型、物体形状、目标状态等）发生变化，因此需要分析如何将已学习的策略迁移到其他任务。一种方法是在任务之间寻找一个不受任务设置影响的共享潜在空间。在不变的潜在空间中学会技能后，便可借助任务特定映射 [19], [20] 将其迁移到不同的任务设置中。其他方法则直接对状态空间或动作空间进行映射，从而将技能从源领域迁移到测试领域 [21]–[22]。然而，对于以往的方法，状态空间或动作空间映射要么是通过分析空间之间的差异手动获得，要么是通过点云配准算法获得。因此，这些方法会利用先验人类知识来找出正确的映射。我们提出的方法通过最大化轨迹回报，自动发现状态空间或动作空间中的潜在映射。

III. 问题表述

在这项工作中，我们提出了一种框架，用于学习在结构化环境中枢转真实世界物体的机器人技能，并通过从仿真到真实世界的零样本迁移学习实现这一目标。我们的主要重点是将学到的枢转技能泛化到任意未见物体。

考虑图1(a)所示的环境，其中刚性物体 o 静止放置在类似桌面的平面上，并且靠近第二个表面，该表面称为墙面，与桌面垂直。物体 o 可以通过机器人机械臂的任意末端执行器进行操作，该末端执行器能够与物体建立面接触，例如夹爪的手指。目标是学习一种策略 $\pi(\cdot|s)$，其中 s 表示系统状态，该策略能够首先使夹爪与 o接触，然后建立 o 与墙面之间的接触，最后利用环境接触（夹爪-物体、物体-桌面和物体-墙面）将物体枢转至直立位置。

需要注意的是，我们并未固定环境中任何元素的初始条件。具体而言，我们不要求墙面-物体接触和夹爪-物体接触预先存在，也不假设已知墙面和物体的精确位置与方向。

我们试图在这一场景中回答的问题是：1）如何学习一种能够从仿真迁移到真实世界的零样本枢转策略？2）同一策略能否泛化到不同物体？

IV. 提出的方法

本节将介绍我们为解决第III节所述问题而提出的方法。其基本思路是：不同物体的枢转操作可以通过对在一个主要物体上学习到的策略进行变换来计算，而不必每次都从头开始学习。我们将该方法分解为三个主要步骤。首先，在仿真中针对任意物体学习一种枢转策略，该物体称为“单一”物体。其次，学习一个物体特征的潜在空间，以基于仿真中生成的合成深度图像表示物体的形状。最后，训练两个神经网络，通过调整状态与动作空间，使单一策略适应新物体。整体框架如图2所示，具体步骤将在下节中详细介绍。

A. 单元物体枢转的强化学习

本节介绍用于学习单一物体枢转策略的强化学习框架。强化学习框架在数学上定义为一个 MDP: $= \{S, A, R, T, \gamma\}$，它是一个元组，其中 S 是状态集合，A 是动作集合，R 是为每个状态/动作对分配实数值的奖励函数，T 是状态转移函数，而 γ 是折扣因子。

仿真环境。单一物体枢转策略$\pi^u_\theta(a|s)$由 θ 参数化，并在 Mujoco [23] 仿真环境中学习，如图3(a)所示。仿真包含机器人夹爪和单一物体，后者是一个$9 \times 9 \times 3\ cm^3$ 盒子。尺寸是任意选定的，不会影响算法。此外，我们在所定义的世界坐标系原点处放置了一面刚性墙面，作为外部表面。

训练 $\pi^u_\theta(a|s)$ 的详细信息如下：

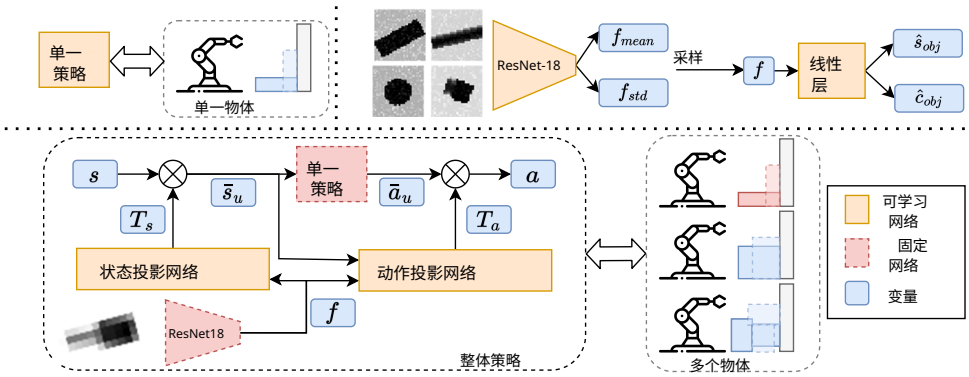


图 2：仿真中的枢转策略学习分为三个步骤：a）在单一物体上学习枢转策略。b）通过预测物体类别和尺寸进行物体特征学习。c）针对多个物体的枢转进行状态/动作投影网络学习

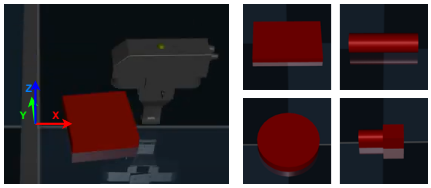


图 3：Mujoco 仿真：a）在单一物体上训练，b）四类物体（盒子、圆形物体、圆柱体、销钉）

状态空间 由三个部分定义：物体位姿 s_o、夹爪位姿 s_g，以及由机器人机械臂腕部、夹爪正上方的力/力矩传感器测得的外力 s_F。物体和夹爪位姿包括笛卡尔位置（X、Y、Z 坐标轴）和四元数方向 s_o，$s_g \in \mathbb{R}^7$ 而 s_F 包含沿 X、Y 和 Z 坐标轴测得的力 $s_F \in \mathbb{R}^3$。因此，状态 $s := [s_o, s_g, s_F] \in \mathbb{R}^{17}$。仿真中机器人施加的最大力在每个坐标轴上均为 $\pm 10N$，并且s_F 被归一化至 $\pm 1N$。

动作空间 由机器人夹爪在 X、Y 和 Z 坐标轴上的线速度以及俯仰方向的角速度定义，$a = [a_x, a_y, a_z, a_\rho] \in \mathbb{R}^4$。动作受限于设置为 25 mm的移动阈值。如果夹爪在训练期间移动超过此限制，机器人将停止并继续执行下一个动作。

奖励函数 是当前物体旋转矩阵 R 与目标旋转矩阵 R^{goal} 之间的距离，其定义为：

$$r = \frac{\pi}{2} - d, \text{ with } d = \arccos\left(0.5(\mathrm{Tr}(R^{goal}R^T) - 1)\right) \quad (1)$$

其中 $Tr(\cdot)$ 表示矩阵的迹。将 $\frac{\pi}{2}$ 添加到奖励中，使初始奖励接近 0。该奖励鼓励机器人将物体枢转至目标方向 R^{goal}，该方向设定为物体与地面垂直时的方向。

领域随机化 用于通过三种噪声提高枢转策略的鲁棒性：

1) 墙壁位置的不确定性。世界坐标系的原点设在墙面上，并假设墙面的准确位置未知。由于物体和夹爪的位置都是相对于墙面测量的，因此我们通过在物体和夹爪的位置上添加标准差为 $2\ cm$ 的零均值高斯噪声来模拟这种不确定性。

2）力测量噪声。为模拟力/力矩传感器测量中的噪声，在仿真中向测得的力添加标准差为 $0.5\ N$ 的零均值高斯噪声。

3）初始位姿分布。在训练过程中，我们对物体的初始位姿进行随机化。具体来说，将物体放置在与墙面相隔一定偏移量的位置，并使其朝向一个随机角度。初始偏移量 Δ_x 在范围 $[0, 5\ cm]$ 内均匀采样。初始旋转角度的限值定义为 $\pm\arctan(\Delta_x/4.5)$，并从该范围内采样初始旋转角度，以避免出现不可行的初始位姿。

B. 物体特征的表征学习

目标是使枢转单一策略能够适用于多个物体。实现这一目标需要获取新物体的运动学信息。我们采用表征学习，根据物体的俯视深度图像学习其低维特征空间。RGB 纹理对于此任务并非必需；我们也不依赖物理测量，因为这会要求用户额外进行工程操作。

我们首先在仿真中构建一个深度图像数据集，以学习这样的特征空间。如图 3(b) 所示，我们从四种物体类别 c_{obj}生成了该数据集：长方体、圆形物体、圆柱体和销钉。每个类别随机生成 100 个不同尺寸的物体。生成的数据集由以下内容组成 $\mathcal{D} := \{\mathcal{I}_i, c_{obj,i}, s_{obj,i}\}_i^{400}$每个物体的深度图像，以及 $s_{obj,\ i} = [l_x, l_y, l_z]$ 物体的尺寸（作为标签记录）。然后，通过对每个原始数据点进行随机平移、旋转和添加高斯噪声，将 $\mathcal{D}$ 扩增十倍，得到 4000 个数据点。

图2(b)展示了所提出的网络 $F(f|\mathcal{I})$，用于基于物体深度图像$\mathcal{I}$学习物体特征 f 。首先，标准 ResNet18 架构 [24] 处理物体深度图像 $\mathcal{I}$，并输出物体特征的均值和标准差 f_{mean} 与 f_{std}。其次，与变分自编码器 [25], 类似，我们使用重参数化技巧对物体特征f进行采样。最后，另一个线性层输出预测的物体

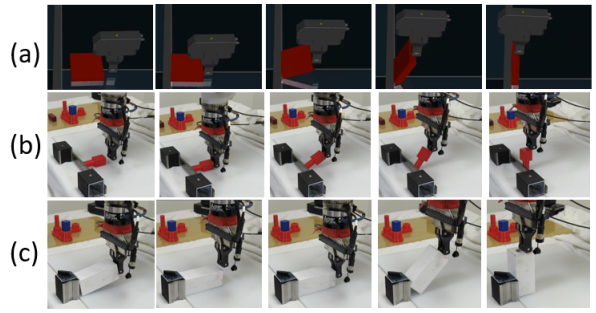


图 4：以下场景的截图：a）在仿真中枢转单一物体；b）在真实世界中枢转销钉；c）恢复行为。

尺寸 $\hat{s}_{obj}$ 以及物体类别的 logits $\hat{c}_{obj}$。用于训练该网络的损失函数设计如下：

$$L = L_{shape} + L_{class} + \beta L_{KL} = \|s_{obj} - \hat{s}_{obj}\|^2 + L_{CE}(c_{obj}, \hat{c}_{obj}) + \beta D_{KL}(N(f_{mean}, f_{std}), N(0,1))$$

其中，L_{CE} 是交叉熵损失，β 是 KL 散度损失的权重。损失函数的前两项用于通过监督学习预测物体尺寸和类别。KL 散度损失用于约束所学习的特征空间，并缓解 [25]中提到的过拟合。

C. 状态与动作投影网络

IV-A 节中采用的标准领域随机化技术无法泛化到运动学属性差异显著的物体；实现这种程度的泛化是我们的主要目标之一。相同的马尔可夫决策过程（MDP）描述了不同的物体，但单一物体会接收未见过的状态作为输入，并输出不适用于多个物体的动作。然而，如图 5(b) 所示，不同物体在枢转过程中的轨迹形状相似，并且可能可以通过在恰当的空间中进行简单变换来描述。

受这一直觉启发，我们提出学习基于物体的变换，以调整状态与动作空间，并将单一策略泛化到新物体上，而不是从头开始学习。具体而言，为了简便起见，我们选择使用线性变换，并将这些变换称为状态和动作投影网络，如图2(c)所示。状态投影网络：$T_s = \rho_\phi(f)$由 ϕ 参数化，以物体 f 的特征作为输入，输出一个 T_s 状态维度的对角矩阵。该输出被用作线性算子，将物体状态 s 投影到类似于单一状态 $\bar{s}_u$的空间：$\bar{s}_u = T_s s$。投影后的状态$\bar{s}_u$ 被输入单一策略 $\bar{a}_u = \pi^u_\theta(\bar{a}_u|\bar{s}_u)$，以生成枢转动作。然而，$\bar{a}_u$ 需要经过变换才能作用于原始物体。因此，我们训练动作变换网络：$T_a = \rho_\psi(f,\ \bar{s}_u)$；该网络给定 f，s_u后，输出一个 T_a 动作维度的对角矩阵。利用该矩阵计算枢转动作：$a = T_a \bar{a}_u$。

整体动作推断过程可概括为：

$$a = T_a \bar{a}_u = T_a \pi^u_\theta(\bar{a}_u|\bar{s}_u) = T_a \pi^u_\theta(\bar{a}_u|T_s s) \quad (2)$$

因此，整体的泛化枢转策略由以下部分组成：单一策略、物体特征提取网络以及状态/动作投影网络。前两者已经训练完成，且其权重已冻结。只需训练状态/动作投影网络，即可使枢转策略适应不同物体。具体而言，我们采用策略梯度方法学习状态/动作投影网络，以最大化对不同物体进行枢转的轨迹回报。假设我们收集到一个特征为 f 的物体的枢转轨迹 τ =(s_1, a_1, T_s, T_a, f, r_1, ..., s_T, a_T, T_s, T_a, f, r_T)，则状态/动作变换网络的优势函数分别为：

$$\hat{A}_s = \frac{1}{T}\sum_{i=0}^{T} \gamma^i r_i;\ \hat{A}_{at} = \sum_{i=t}^{T} \gamma^i r_i \quad (3)$$

其中，r_i 是时间 i 的奖励，γ 是折扣因子。然后，对状态和动作投影网络应用更简单版本的 PPO [26] 更新：

$$L_s(f, T_s, \phi, \phi_{old}) = \min\left(\frac{\rho_\phi(f)}{\rho_{\phi_{old}}(f)}, \text{clip}\left(\frac{\rho_\phi(f)}{\rho_{\phi_{old}}(f)}, 1+\epsilon_s, 1-\epsilon_s\right)\right) A_s \quad (4)$$

$$L_a(f, s_u, T_a, \psi, \psi_{old}) = \min\left(\frac{\rho_\psi(f, s_u)}{\rho_{\psi_{old}}(f, s_u)}, \text{clip}\left(\frac{\rho_\psi(f, s_u)}{\rho_{\psi_{old}}(f, s_u)}, 1+\epsilon_a, 1-\epsilon_a\right)\right) A_a \quad (5)$$

其中，$\frac{\cdot \rho_1(\ )}{\rho_2(\cdot)}$ 是两个投影的似然比，ϵ_s, ϵ_a 是更新的裁剪因子。两个投影网络的权重按以下方式更新：

$$\phi_{k+1} = \underset{\phi}{\text{argmax}} \underset{(f,T_s)\sim\pi_{\phi_k}}{\mathbb{E}} L_s(f, T_s, \phi, \phi_k) \quad (6)$$

$$\psi_{k+1} = \underset{\psi}{\text{argmax}} \underset{(f,s_u,T_s)\sim\pi_{\psi_k}}{\mathbb{E}} L_a(f, s_u, T_a, \psi, \psi_k) \quad (7)$$

本质上，我们最大化具有更高后续奖励的状态或动作转换的似然。我们所提出方法的详细信息总结于算法 1 中。

算法 1： 学习状态和动作投影

初始化状态和动作投影网络
  $\rho_\phi(f)$ $\rho_\psi(f, s_u)$，并使用随机权重 ϕ, ψ
初始化单一策略 $\pi^u_\theta(\bar{a}_u|\bar{s}_u)$ 和特征
  使用预训练权重的提取网络 $F(f|I)$。
对于 i = *0*, *1*、*2*、.....直至收敛 **执行**
  对于迭代 k = *1* 至 K **执行**
    随机采样一个物体 o_k 及其图像 $\mathcal{I}_k$
    推断物体特征 $f_k \sim F(f|\mathcal{I}_k)$
    使用（2）推断动作以收集一条轨迹
  end
  计算状态和动作的优势值
    使用 (3) 进行变换
  对迭代 m = *1* 至 M **执行**
    更新状态/动作投影网络(6),(7)
  end
end

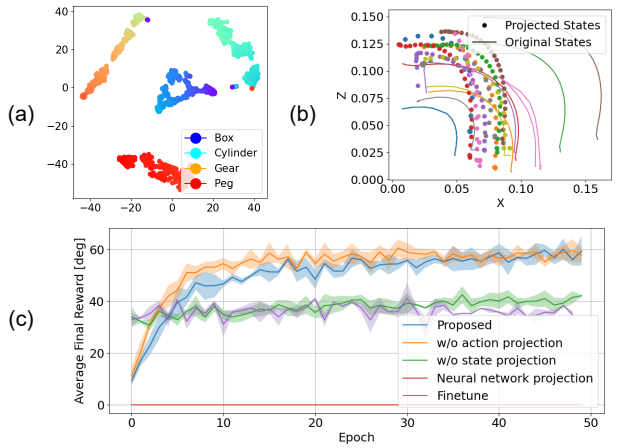


图 5：a）学习到的物体特征的 t-SNE 可视化，同一类别内的颜色差异表示不同的物体尺寸；b）不同物体在投影前后的枢转轨迹；c）策略适应的学习曲线。

V. 实验

A. 仿真：训练与验证实验

单一物体枢转的强化学习：单一物体的枢转策略 $\pi_\theta^y(a|s)$ 使用 SAC [27] 算法进行训练，训练所依据的马尔可夫决策过程（MDP）如第 IV-A 节所述，算法采用 RLkit [28] 中的实现。策略网络和 Q 函数网络分别采用具有 128 个和 256 个单元的两层 ReLU 网络进行参数化。批量大小为 1024，Q 函数和策略的学习率分别为 $5e-3$ 和 $3e-4$。图 4(a) 展示了策略训练完成后，单一物体在仿真中进行枢转的一系列连续快照。夹爪首先将物体推向墙面，使物体与墙面建立接触。随后，机器人在将物体推向墙面的同时向上移动，从而开始枢转。当物体旋转到一定角度后，我们注意到机器人施加了向下作用力，以保持与物体的接触。这是一种能够稳定物体并防止物体掉落的可靠方式。最后，物体被翻转起来，并直立在桌面上。

物体特征表示学习：编码神经网络采用第 IV-B 节所述的数据集 $\mathcal{D}$ 以监督方式进行训练，并使用 t-SNE 方法 [29]评估计算得到的物体特征空间。如图 5(a) 所示，物体特征在学习到的特征空间中聚为四组，分别表示四种物体类别。此外，在每个类别内部，该特征空间还能区分不同物体的尺寸信息，这表明学习到的特征空间可用于下游任务。

状态/动作投影网络训练： 单一策略和物体特征学习完成后，在相同的仿真环境中（见第 IV 节）训练状态/动作投影网络，方法是随机选择 40 个大小各异的物体，每类选择 10 个 c_{obj}。两个投影网络分别是具有 16 个和 35 个单元的两层 ReLU 网络；算法的每个训练轮次会选取 200 条轨迹组成批次，用于更新网络。由于状态中的四元数元素彼此耦合，且会受到变换影响，因此对于状态投影网络，我们仅对物体和夹爪的位置应用相同的投影，即 $T_s \in \mathbb{R}^{3\times3}$ 和 $T_a \in \mathbb{R}^{4\times4}$保持不变。

多物体泛化性能：所提出的方法在仿真中进行评估，并与三项消融研究和两种基线方法进行比较：

1)神经网络投影：训练两个神经网络 $\bar{s}_u = \rho_s(f, s)$、$a = \rho_a(f, \bar{s}_u)$，以替代状态/动作投影网络中的线性变换；2) 无状态：我们的方法不使用状态投影网络；3) 无动作：我们的方法不使用动作投影网络；4) 微调：在不使用状态和动作投影网络的情况下，使用 PPO 对单一策略进行微调；

5) *PEARL*：训练一种最先进的元强化学习方法 PEARL [18]。方法 1) 至 4) 按照算法 1 进行训练，但更新的神经网络不同。PEARL 从头开始学习。

如图 5(c) 所示，所提出的方法和消融实验 无动作 均优于其他所有方法。神经网络投影 无法使单一策略适应多个物体，这可能是因为它无法利用线性变换的结构，并且可能需要更多数据才能学会该任务。我们还注意到，状态投影网络对适应的帮助最大，且收敛速度比所提出的方法更快。原因是所提出方法的动作投影在初始阶段并不完美，因此训练速度较慢。然而，如表 I 所示，所提出的方法取得了高于所有基线方法的成功率，这表明动作投影有助于根据物体特征调整策略。在我们的实验中，PEARL 可以接近物体，但无法学会对多个物体进行枢转。我们得出结论，物体之间差异过大，使 PEARL 难以从头开始学习适用于所有物体的策略。由于仿真中的物体是随机生成的，有些物体相对于自身尺寸而言，可供站立的表面较小，这使其难以枢转，因此我们的方法会失败。

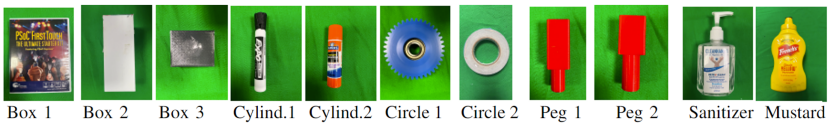
Ours	无动作	无状态	微调	NN	PEARL
30/40	27/40	19/40	15/40	0/40	0/40

表 I：在训练环境中对 40 个物体进行 3 次随机运行的平均成功率。

我们进一步通过绘制两个最具代表性的轴X，Z上不同物体在投影前后的枢转轨迹，分析状态投影网络的效果。如图 5(b) 所示，尽管原始轨迹（实线）差异很大，投影网络仍能使枢转轨迹趋于一致，从而使单一策略能够适应新物体。

B. 真实世界实验

真实机器人设置与视觉反馈：我们使用一台六自由度三菱 Assista RV-5AS-D 协作机械臂，配有 WSG32 夹爪，如图 7(a) 所示。机器人



	Box 1	Box 2	Box 3	Cylind.1	Cylind.2	Circle 1	Circle 2	Peg 1	Peg 2	Sanitizer	Mustard
April-Tag	10/10	10/10	10/10	9/10	9/10	10/10	10/10	9/10	10/10	10/10	10/10
Vision tracking	10/10	7/10	10/10	10/10	5/10	10/10	10/10	10/10	10/10	10/10	7/10

图 6：两种视觉系统的测试物体及成功率

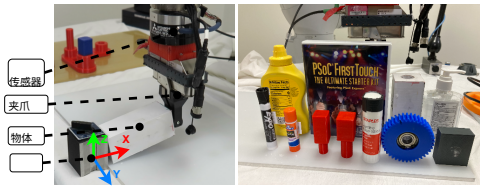


图 7： a) 真实世界中的设置。b) 真实世界中的物体

世界。

以阻抗控制模式运行，刚度设为 $12N/mm$。外力由腕部安装的力/扭矩传感器测量。对于状态估计，我们比较了两种系统，以从 RGB-D 相机（Intel Realsense D435）获取物体位姿。第一种系统使用物体上的 AprilTag [30] 进行跟踪。该系统可提供高精度的状态信息，但需要在物体上贴附标签。第二种系统由基于视觉的 6D 位姿估计组成，包含掩膜和深度特征提取 [31], [32] 流程，并结合位姿跟踪。为了利用 RGB-D图像快速跟踪运动中的新物体，我们基于 BundleTrack [34]引入了多项改进，以增强位姿跟踪 [33] 。该方法不需要 CAD模型；由于传感器噪声，状态估计可能会更加嘈杂。我们使用这两种系统进行枢转操作测试，以评估所提方法的鲁棒性。

物体数据集： 图 7(b) 展示了我们在真实世界实验中使用的物体。我们选取了九种物体，在仿真中可归为四个物体类别，用于测试所提出方法的仿真到现实迁移性能。此外，我们还在两个形状不规则的物体（洗手液和芥末酱瓶）上测试了泛化能力。请注意，这些物体均未在仿真中出现过；同样，摄像头拍摄的真实物体深度图像也未在仿真中出现过，并用于推断物体特征。

枢转技能的仿真到现实迁移与泛化： 所提出的方法通过对各类物体进行零样本迁移学习来评估。当物体到达稳定直立位置时，即判定实验成功，成功率如图 6 所示。首先，我们分析使用 AprilTag系统估计状态的情况。尽管这些物体的形状和尺寸差异很大，即 $l_x = [6, 18.5]$, $l_y = [1, 15]$, $l_z = [1, 5][cm]$, 我们的方法在几乎所有物体上的成功率都达到 100%，展现了直接仿真到现实迁移的能力，以及对多种物体的泛化能力。失败情况全部发生在底部为圆柱体的物体上，即圆柱体 1、2 和销钉 2，因为它们的形状容易产生滚动行为，而且用于站立的底面较小。然而，该策略是在大多数情况下，该策略都能抵御这些困难，且仅失败过一次。图 4(b)-(c) 通过一系列快照展示了成功的枢转实验。值得注意的是，如图 4(c) 所示，学习到的策略能够在失败后恢复。第一次枢转尝试可能会失败，因为盒子是我们所用物体中最重的（$450[g]$ 而其他物体的重量介于 $[0, 200][g]$），这可能会加剧推动时接触处的滑移效应。不过，学习到的策略会移动到物体后方，并在第二次尝试时成功完成枢转，展现出出色的鲁棒性。

接下来，我们测试了对两个分布外物体——“洗手液”和“芥末酱瓶”的泛化能力。这两种物体的形状复杂、非凸，且接触表面不规则，并非平面。此外，训练过程中没有使用任何形状相似的物体。然而，我们的方法在这两种物体上均取得了成功，成功率达到 100%，展现了良好的泛化能力。

最后，我们评估了使用视觉跟踪系统时的性能，该系统返回的状态估计噪声更大。如图 6 所示，对于大多数物体，我们仍然取得了 100% 的成功率。失败案例是由状态中的噪声过大造成的：盒子 2 上是因为物体没有纹理，圆柱体 2 上是因为物体非常小，而芥末酱瓶则是在接近目标时跟踪丢失。

VI. 结论

我们提出了一种框架，用于学习枢转真实世界物体的机器人技能。该方法仅在仿真中训练，只需一张被操作物体的深度图像，即可迁移到真实世界任务。此外，同一策略能够泛化到多个真实世界物体的枢转操作。核心思想是先为一个“单一”物体学习鲁棒的强化学习策略，再学习两个投影网络，以调整输入该策略的状态以及该策略输出的动作。通过物体的俯视深度图像，我们学习物体特征空间，以编码尺寸和形状等运动学属性。真实世界实验表明，该方法成功实现了跨越仿真到真实世界差距的零样本迁移，并泛化到多个物体。

所提出的方法根据物体的视觉特征调整状态／动作空间。不过，也可以考虑摩擦和惯性属性等其他特征，通过额外的传感器输入或系统辨识程序，将我们的方法推广到更广泛的任务类别。此外，我们对状态和动作应用线性变换，以构建结构更明确的模型并提高数据效率。尽管如此，更复杂的操作任务仍需学习依赖状态的非线性变换。

参考文献

[1] Y. Shirai、D. K. Jha、A. U. Raghunathan 和 D. Romeres，“鲁棒枢转：利用双层优化实现摩擦稳定性”，载于 2022 年国际机器人与自动化会议（ICRA），2022 年，第 992–998 页。[2] S. Jin、D. Romeres、A. Ragunathan、D. K. Jha 和 M. Tomizuka，“可变形物体操作的轨迹优化：皮带传动单元的装配”，载于 2021 年 IEEE 国际机器人与自动化会议（ICRA），2021 年，第 10 002–10 008 页。[3] A. U. Raghunathan、D. K. Jha 和 D. Romeres，“Pyrobocop：用于操作的 Python 机器人控制与优化软件包”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 985–991 页。[4] Y. Shirai、D. K. Jha、A. Raghunathan 和 D. Romeres，“接触丰富型系统中的机会约束优化，实现鲁棒操作”，arXiv 预印本 arXiv:2203.02616，2022 年。[5] F. R. Hogan、E. R. Grau 和 A. Rodriguez，“基于凸混合 MPC 的响应式平面操作”，载于 2018 年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，2018 年，第 247–253 页。[6] X. Cheng、E. Huang、Y. Hou 和 M. T. Mason，“用于三维准动态灵巧操作的接触模式引导运动规划”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 2730–2736 页。[7] X. Zhang、S. Jin、C. Wang、X. Zhu 和 M. Tomizuka，“通过离散-连续混合动作空间学习机器人装配任务的插入基元”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 9881–9887 页。[8] T. Z. Zhao、J. Luo、O. Sushkov、R. Pevceviciute、N. Heess、J. Scholz、S. Schaal 和 S. Levine，“面向工业插入任务的离线元强化学习”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 6386–6393 页。[9] X. Zhang、L. Sun、Z. Kuang 和 M. Tomizuka，“通过逆强化学习为力相关任务学习可变阻抗控制”，《IEEE 机器人与自动化快报》，第 6 卷，第 2 期，第 2225–2232 页，2021 年。[10] R. Chitnis、S. Tulsiani、S. Gupta 和 A. Gupta，“利用学习得到的任务模式实现高效双臂操作”，载于 2020 年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，2020 年，第 1149–1155 页。[11] S. Nasiriany、H. Liu 和 Y. Zhu，“利用行为基元增强强化学习，以应对多种操作任务”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 7477–7484 页。[12] W. Zhou 和 D. Held，“借助涌现的外在灵巧性学习抓取不可抓取物体”，载于 ICRA 2022 研讨会：接触丰富型操作的强化学习，2022 年。 [在线]。链接：https://openreview.net/forum?id=Zrp4wpa9lqh [13] M. Laskin、K. Lee、A. Stooke、L. Pinto、P. Abbeel 和 A. Srinivas，“利用增强数据进行强化学习”，《神经信息处理系统进展》，第 33 卷，第 19 884–19 895 页，2020 年。[14] C. Finn、P. Abbeel 和 S. Levine，“用于快速适应深度网络的模型无关元学习”，载于国际机器学习会议。PMLR，2017 年，第 1126–1135 页。[15] C. Wang、Y. Zhang、X. Zhang、Z. Wu、X. Zhu、S. Jin、T. Tang 和 M. Tomizuka，“利用图神经网络对电缆操作中的形变模型进行离线-在线学习”，《IEEE 机器人与自动化快报》，第 7 卷，第 2 期，第 5544–5551 页，2022 年。[16] K. Hausman、J. T. Springenberg、Z. Wang、N. Heess 和 M. Riedmiller，“学习可迁移机器人技能的嵌入空间”，载于国际学习表征会议，2018 年。[17] T. Li、N. Lambert、R. Calandra、F. Meier 和 A. Rai，“通过分层强化学习学习可泛化的运动技能”，载于 2020 年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，2020 年，第 413–419 页。[18] K. Rakelly、A. Zhou、C. Finn、S. Levine 和 D. Quillen，“通过概率上下文变量实现高效的离策略元强化学习”，载于国际机器学习会议。PMLR，2019 年，第 5331–5340 页。[19] Z.-H. Yin、L. Sun、H. Ma、M. Tomizuka 和 W.-J. Li，“利用不变表征实现跨域机器人模仿”，载于 2022 年国际机器人与自动化会议（ICRA）。IEEE，2022 年，第 455–461 页。[20] K. Kim、Y. Gu、J. Song、S. Zhao 和 S. Ermon，“领域自适应模仿学习”，载于国际机器学习会议。PMLR，2020 年，第 5286–5295 页。

[21] M. E. Taylor、P. Stone 和 Y. Liu，“通过任务间映射进行迁移学习以实现时序差分学习。”机器学习研究杂志，第 8 卷，第 9 期，2007 年。[22] T. Tang、C. Liu、W. Chen 和 M. Tomizuka，“通过切空间映射和非刚性配准对可变形物体进行机器人操作”，载于*2016*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），2016 年，第 2689–2696 页。[23] E. Todorov、T. Erez 和 Y. Tassa，“Mujoco：用于基于模型的控制的物理引擎”，载于*2012*年*IEEE/RSJ*智能机器人与系统国际会议IEEE，2012 年，第 5026–5033 页。[24] K. He 、 X. Zhang 、 S. Ren 和 J. Sun，“ 用于图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议论文集，2016 年，第 770–778 页。[25] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，*arXiv* 预印本 *arXiv:1312.6114*，2013 年。[26] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，“近端策略优化算法”，*arXiv* 预印本 *arXiv:1707.06347*，2017 年。[27] T. Haarnoja、A. Zhou、K. Hartikainen、G. Tucker、S. Ha、J. Tan、V. Kumar、H. Zhu、A. Gupta、P. Abbeel 等，“软演员-评论家算法及其应用”，*arXiv* 预印本 *arXiv:1812.05905*，2018 年。[28] Rail-Berkeley，“Rail-berkeley/rlkit：强化学习算法集合。”[在线]。可用地址：https://github.com/rail-berkeley/rlkit [29] L. Van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据。”机器学习研究杂志，第 9 卷，第 11 期，2008 年。[30] E. Olson，“AprilTag：一种稳健且灵活的视觉基准标记系统”，载于*2011*年*IEEE*机器人与自动化国际会议。IEEE，2011 年，第 3400–3407 页。[31] C. Mayer、M. Danelljan、G. Bhat、M. Paul、D. P. Paudel、F. Yu 和 L. Van Gool，“转换用于跟踪的模型预测”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022 年，第 8731–8740 页。[32] J. Revaud、P. Weinzaepfel、C. De Souza、N. Pion、G. Csurka、Y. Cabon 和 M. Humenberger，“R2D2：可重复且可靠的检测器和描述子”，*arXiv* 预印本 *arXiv:1906.06195*，2019 年。[33] B. Huang、J. Yu 和 S. Jain，“ EARL：用于主动位姿估计动态抓取的眼在手上强化学习器”，2023 年。[34] B. Wen 和 K. Bekris，“BundleTrack：无需实例级或类别级三维模型的新型物体六维位姿跟踪”，载于*2021*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2021 年，第 8067–8074 页。