DOI:10.11784/tdxbz202503008

基于虚实迁移强化学习的机器人按钮操作策略研究

龙晖午[1]，肖聚亮[1]，赵　炜[1]，刘海涛[1]，朱　林[2]，陈　斌[3]
（1. 天津大学机构理论与装备设计教育部重点实验室，天津 300350；
2. 中国空间技术研究院北京卫星制造厂有限公司，北京 100080；3. 西门子中国研究院，北京 100102）

摘　要：具身智能概念的快速发展对智能体与物理世界的交互能力提出了更高要求. 在以机器人为代表的智能载体与环境的交互过程中，主要依靠力反馈信号以决定其动作输出的任务称为力交互任务，例如零件装配、按钮操作和门窗开合等. 针对此类任务交互对象种类繁多、力反馈特性各不相同的挑战，提出了一种具身智能训练方法，基于虚实迁移（sim-to-real）的概念和强化学习方法搭建了机器人高级力交互操作技能学习训练框架，赋予了机器人安全、准确、适应性强大的力交互操作能力. 以经典的机器人力交互场景——按钮操作任务为例：首先，基于域随机化的方法在虚拟环境中构建了大量按钮模型，并从接触刚度的角度划分了机器人与按钮之间的接触阶段；然后，模仿人类在按钮操作过程中的感知方式，结合在线刚度估计算法，在虚拟环境下使用近端策略优化（PPO）算法训练机器人的按钮操作技能；最后，通过 sim-to-real 方法将得到的预训练策略直接部署在真实机器人上，在策略迁移后的实机操作实验中得到了良好的结果. 在针对具有不同力反馈特性的按钮进行操作的泛化能力测试实验中，经上述方法训练得到的策略展现了远优于现有方法的泛化性能.

关键词：机器人；技能学习；力交互；按钮操作；强化学习；虚实迁移
中图分类号：TP242　　**文献标志码**：A　　**文章编号**：0493-2137（2026）04-0361-12

Research on Robot Button Operation Policy Based on Sim-to-Real Reinforcement Learning

Long Huiwu[1]，Xiao Juliang[1]，Zhao Wei[1]，Liu Haitao[1]，Zhu Lin[2]，Chen Bin[3]
（1. Key laboratory of Mechanism Theory and Equipment Design of Ministry of Education，Tianjin University，Tianjin 300350，China；2. Beijing Spacecraft Manufacturing Co.，Ltd.，China Academy of Space Technology，Beijing 100080，China；3. Corporate Technology China，Siemens Ltd.，China，Beijing 100102，China）

Abstract：The rapid development of the concept of embodied intelligence has increased the requirements for the interaction capabilities of agents with the physical world. During the interaction between intelligent carriers，such as robots and the environment，tasks that rely primarily on force feedback signals to determine action outputs are known as force interaction tasks，including component assembly，button operation，and door/window manipulation. To address the challenges posed by the diversity of interaction objects and the varying force feedback characteristics during such tasks，a training method for embodied intelligence was proposed. A training framework for learning advanced robotic force interaction skills was developed based on the sim-to-real concept，as well as reinforcement learning，thereby enabling robots to safely，accurately，and adaptively perform force interaction tasks. Considering the classic robotic force interaction scenario—button operation—as an example，numerous button models were constructed in a virtual environment using domain randomization，and the contact phases between the robot and the button were categorized based on contact stiffness. Thereafter，inspired by human perception of button operation，an online stiffness estimation algorithm was incorporated，and the proximal policy optimization（PPO）algorithm was employed to train the button operation skills of the robot in the virtual environment. Finally，the pretrained policy was

收稿日期：2025-03-05；**修回日期**：2025-03-31.
作者简介：龙晖午（2000—　），男，硕士研究生，13786285162l@163.com.
通信作者：肖聚亮，tianjinxjl@163.com.
基金项目：国家自然科学基金资助项目（52175025，52325501）.
Supported by the National Natural Science Foundation of China（No. 52175025，No. 52325501）.

directly deployed onto a real robot through the sim-to-real method, achieving favorable results in real-world experiments. In generalization tests on buttons with different force feedback characteristics, the policy, which was trained using the proposed method, demonstrated significantly superior generalization performance compared with the existing approaches.

Keywords: robot; skill learning; force interaction; button operation; reinforcement learning; sim-to-real

随着技术的进步,机器人开始承担更多类型的任务,这对其与多变物理世界间的力交互操作能力提出了很高要求. 例如,在使用机器人对电柜、服务器等设备进行维护的研究中[1],存在大量机器人与多种电气开关、按钮的交互问题. 若使用传统的预编程控制方法,首先需要获得各种按钮、开关的精确触发参数(声音、操作行程、按压力),然后依据实时反馈信息判断按纽触发状态. Sukhoy 等[2]在研究机器人自动按压门铃按钮的任务中,通过监测蜂鸣器的发声状态,从声学信息角度来判断门铃按钮是否被成功触发. Wang 等[3]通过在视觉层面检测开关活动部件的位移值来判断开关触发状态,并基于当前的开关状态,选择对应的预设程序以控制机器人进行下一步动作. 上述研究均需要提前获取按钮开关的精确操作参数,然而在非结构化场景中,此类参数难以获取,传统控制方法在实施成本和安全性方面都存在较大的缺陷.

随着学习算法的发展[4-5]和具身智能[6]概念的出现,模仿学习和强化学习(reinforcement learning, RL)等方法快速发展,机器人在非结构化环境中完成力交互操作任务成为可能. Racca 等[7]和 Liu 等[8]基于模仿学习[4],分别从按压力和按压刚度层面对人类的按钮操作技能进行学习,并实现了对同种按钮在不同高度位置等扰动条件下的成功按压,提升了机器人对不同任务条件的适应能力. 另一方面,强化学习方法[5]凭借其对不确定性环境的强大适应能力,发挥了显著作用. 例如,Levine 等[9]提出一种基于深度强化学习的机器人手眼协调抓取学习方法,该方法通过大量抓取数据的训练,可在夹爪磨损、遮挡等不确定的环境因素影响下,针对不同物体进行成功抓取,展现了其训练出的策略的强大适应能力.

强化学习训练过程需要大量交互数据的支撑,而在真实世界收集数据的样本效率低下且成本高昂,故而许多研究者借助仿真环境来训练智能体(agent). 在仿真环境下训练智能体并在真实世界下部署成熟策略的学习方法称为虚实迁移(sim-to-real)[10],该方法不仅能提供近乎无限的潜在数据源,还能降低机器人在训练中的安全风险. 基于 sim-to-real 的学习方法,Matas 等[11]利用深度强化学习算法解决可变形物体的操作问题,通过在仿真环境生成布料模型数据以供虚拟机器人训练的方式,大大减少了训练时间,并在真实世界中驱动机器人成功完成对应操作任务.

在基于 sim-to-real 的强化学习训练流程中,由于仿真器性能限制,导致虚实环境的数据间存在差异. 受此影响,经仿真数据训练的策略在部署到真实机器人上后可能出现性能下降的现象[12],因此仿真数据的质量十分重要. 对于按钮操作任务来说,能否获得可编辑、高仿真度、反馈精确的虚拟按钮模型,直接影响到策略的生成与迁移后的性能. 然而,现有的按钮建模和仿真工作主要集中在对实体按钮的研究[13-14]. 研究者们通过建立按钮仿真的实体平台,利用超声波振动器[15]、线性马达[14]等物理手段来模拟实体按钮的点击感,使操作者获得逼真的按压反馈. 然而,这类方法难以迁移到虚拟按钮模型的仿真问题上,这也是本文的研究动机之一.

虚拟按钮的质量决定策略的有效性,而如何描述机器人与按钮的接触状态并据此进行学习,关系到训练效率和操作策略的性能. 在机器人操作技能学习领域,力交互任务一直是研究的热点. 许多研究通过建立接触模型并依据接触力/力矩的变化[16-17]来划分接触阶段,然后控制机器人执行相应操作. 然而,针对种类繁多的按钮,此类方法难以建立统一的接触模型. 因此,有必要寻找新方法,从更高维度来统一地描述机器人与按钮交互过程中的接触状态,并将其用于按钮操作技能的学习过程中.

针对上述问题,提出了一种基于虚实迁移强化学习的机器人力操作技能学习框架,适用于机器人通用按钮操作技能学习. 为获取高质量的训练数据,开发了一种高仿真度、易编辑的虚拟按钮模型生成方法,支持生成海量数据进行训练,以增强机器人按钮操作技能的鲁棒性和泛化能力. 提出了基于接触刚度的接触状态描述方法,类比人类感知模式,训练基于接触刚度感知的机器人高级按钮操作策略,并在与同类方法的对比实验中体现了优势.

1 机器人力交互按钮操作学习框架

本文设计的面向按钮操作任务的机器人力交互

按钮操作学习框架如图 1 所示，主要分为接触阶段描述与划分、按钮仿真数据生成、强化学习虚拟训练和预训练策略的实机部署共 4 个模块. 该框架专注机器人与按钮的力交互过程，并采用强化学习和 sim-to-real 方法解决非结构化场景中的不确定性.

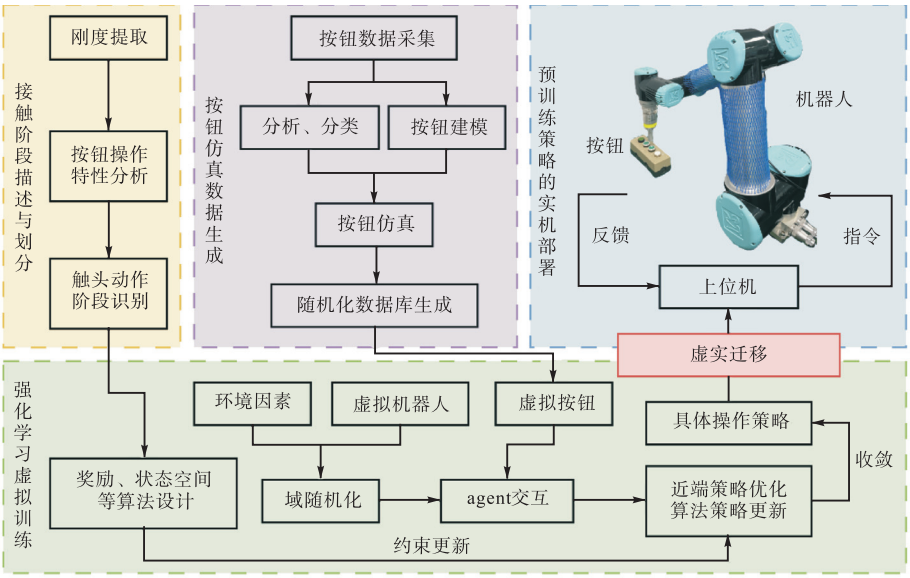


图 1　面向按钮操作任务的机器人力交互按钮操作学习框架

Fig.1　Robot force interaction operation learning framework for button operation tasks

接触阶段描述与划分模块为技能训练提供指导和约束. 人类操作者仅需经过简单的学习，便可操作全新未见的按钮. 受人类的力交互操作技能启发，模仿人在操作时指尖对环境刚度的粗略感知，提出基于接触刚度的接触状态描述和接触阶段划分方法. 该方法可帮助机器人学习理想的操作策略，准确识别触发信号，对按钮进行安全的精确操作.

按钮仿真数据生成模块为技能训练提供数据支持. 首先，依据按钮的力反馈特性对主要电气按钮进行种类划分，并建立几类典型的按钮力反馈模型. 然后，在基于物理引擎的仿真环境中对按钮模型进行仿真，以得到反馈精确的虚拟按钮. 最后，使用基于“控制点”的方法对几类仿真按钮进行处理，随机化生成庞大的虚拟按钮数据库.

强化学习虚拟训练模块在仿真环境中对机器人按钮操作技能进行并行训练. 提出一种类人感知的接触阶段划分方法，并据此设计奖励函数、状态空间、观测空间以约束策略的更新. 然后，对训练场景中的物理参数进行域随机化处理，以生成多种不同的训练环境，并使智能体在不同环境中与数量庞大的虚拟按钮交互以更新按钮操作策略，直至策略收敛.

预训练策略的实机部署模块将训练得到的成熟策略部署到真实机器人上. 在得到固化的按钮操作策略后，将其经过虚实迁移部署到真实环境的上位机中. 预训练策略控制真实机器人对真实按钮进行操作，并通过实时的反馈信息来进行决策，以输出下一个控制循环中的指令.

2　接触状态描述与划分

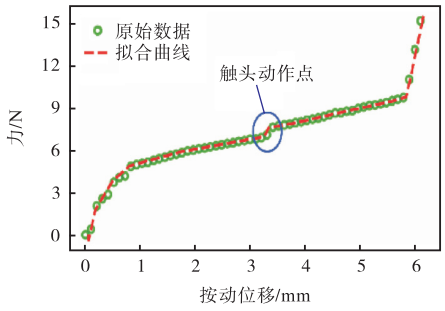
2.1　按钮组成与工作原理

常见电气按钮可分为自复位式和自锁式，内部触点则有常开和常闭之分. 按钮本体主要由按钮帽、按钮外壳、复位机构、触点模块等几个主要部分组成. 在对按钮进行按压时，按钮复位结构内部的弹性元件受力发生弹性形变并提供弹性反力，以实现按钮的回弹和复位功能. 若持续对按钮进行按压操作并使得触点模块的动静触点发生接触/分离的动作，则可实现按钮接通/切断回路的基础功能，并且由于触点间的动作，会产生一个较小的冲击. 若继续向下按压到底，弹性元件形变达到最大，此时再加力按压则可能会损坏按钮. 值得注意的是，在上述按钮操作过程中存在位移、接触力等明显的反馈信号，特别是按钮触点模块发生动作的时刻.

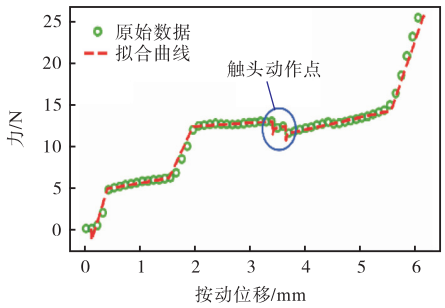
2.2　按钮操作特性分析

本研究针对的按钮型号丰富，按钮的反馈特性、按钮行程与触发力差别巨大. 因此，直接依据接触力、力矩等力学信息或按动行程等几何因素来区分不同按压阶段是不可行的. 为分析按钮操作过程中的反馈特性，本文采集了两种经典电气按钮操作过程中

的反馈数据，并对其进行了分段线性拟合处理，如图 2 所示，其中原始数据经过了稀疏处理.



（a）西门子 SIRIUS ACT 3SU1 按钮



（b）长江电气 LA38-11/203 按钮

图 2 按钮力反馈数据的分段线性拟合

Fig.2 Piecewise linear fitting of button force feedback data

从拟合曲线来看，在按钮触头动作点附近均存在明显的力学信号，这也与人类触发按钮时指尖获取类似“咔哒”感的力觉反馈相符. 这类力学信号代表在按钮触发时机器人与按钮间的接触力发生了跳变. 因此，笔者认为可以通过感知按钮操作过程中力的跳变信号，即按钮刚度的动态变化来识别不同按钮的触发信号并划分接触状态，并在此基础上进行机器人按钮操作技能的学习.

2.3 接触动力学建模与在线刚度估计

机器人操作问题本质上是机器人与环境的力交互问题，一般而言可分为弹性接触和刚性接触两种情况. 在机器人按钮操作任务中，机器人与按钮间的接触属于弹性接触，按钮可等效为弹簧阻尼模型，如图 3 所示. 通过在线辨识按钮模型的动力学参数，即可得到操作过程中按钮的刚度信息.

根据弹性动力学，按钮接触模型为

$$M_{\mathrm{b}}\ddot{x}+C_{\mathrm{b}}\dot{x}+K_{\mathrm{b}}x_{\mathrm{de}}=F \tag{1}$$

式中：M_{b}、C_{b}、K_{b} 分别为按钮的质量、阻尼和刚度参数；$\ddot{x}$、$\dot{x}$、x_{de} 分别为按钮活动部件的加速度、速度和相对按钮表面的位移；F 为接触力. 按钮活动部件质量极小，且机器人按压过程速度变换也不大，因此可将模型中的惯性项忽略，并写成

$$\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\theta}=y \tag{2}$$

式中：$y=F$；$\boldsymbol{\phi}=[\dot{x}\quad x_{\mathrm{de}}]^{\mathrm{T}}$；$\boldsymbol{\theta}=[C_{\mathrm{b}}\quad K_{\mathrm{b}}]^{\mathrm{T}}$.

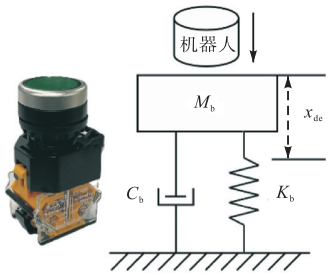


图 3 按钮接触模型示意

Fig.3 Diagram of the button contact model

由于按钮的动力学参数随按压行程动态变化，因此需要在线估计其动力学参数. 本文采用遗忘因子递推最小二乘法（forgetting factor recursive least-squares，FFRLS）来估计按钮的实时动力学参数[18]. 该方法通过调整遗忘因子的值来平衡历史数据和最新数据的影响，从而更好地适应不断变化的数据分布. 假设对长度为 K 的数据集，其表示形式为

$$\boldsymbol{\Phi}_K\boldsymbol{\theta}+\boldsymbol{\delta}=\boldsymbol{Y}_K \tag{3}$$

式中：$\boldsymbol{\Phi}_K=\left[\boldsymbol{\phi}_1^{\mathrm{T}}\quad\boldsymbol{\phi}_2^{\mathrm{T}}\quad\cdots\quad\boldsymbol{\phi}_K^{\mathrm{T}}\right]^{\mathrm{T}}$；$\boldsymbol{\delta}$ 为测量噪声；$\boldsymbol{Y}_K=[y_1\quad y_2\quad\cdots\quad y_K]^{\mathrm{T}}$.

对于 t 时刻的参数估计为

$$\hat{\boldsymbol{\theta}}_t=\hat{\boldsymbol{\theta}}_{t-1}+K_t\left(y_t-\boldsymbol{\phi}_t^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{t-1}\right) \tag{4}$$

$$\boldsymbol{K}_t=\frac{\boldsymbol{P}_{t-1}\boldsymbol{\phi}_t}{\lambda+\boldsymbol{\phi}_t^{\mathrm{T}}\boldsymbol{P}_{t-1}\boldsymbol{\phi}_t} \tag{5}$$

$$\boldsymbol{P}_t=\frac{(\boldsymbol{I}-K_t\boldsymbol{\phi}_t^{\mathrm{T}})\boldsymbol{P}_{t-1}}{\lambda} \tag{6}$$

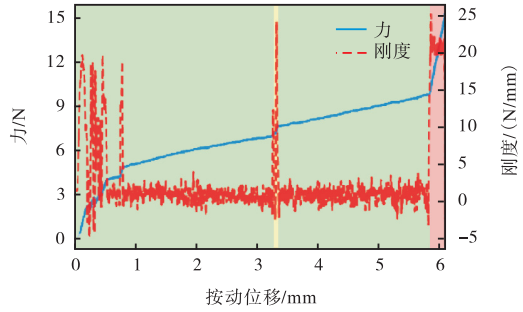
$t=0$ 时，$\boldsymbol{P}_0=10^6\times\boldsymbol{I}$，$K_0=0$，取遗忘因子 $\lambda=0.6$.

2.4 接触阶段划分

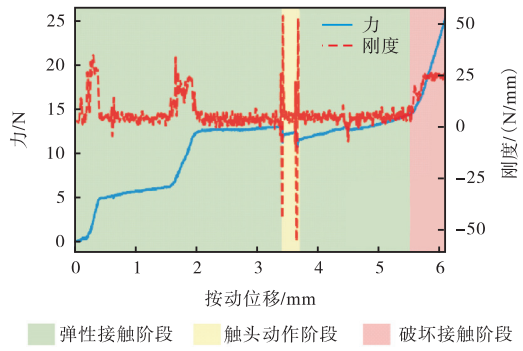
使用第 2.3 节推导得到的带遗忘因子的递推最小二乘法对前文采集的两种按钮的按压数据进行处理，绘制出了图 4 中的按钮接触刚度曲线. 从两种按钮的接触刚度曲线来看，其在按钮触发时具有特殊的力反馈信号，显著区别于触发前、后的接触刚度特征. 基于上述现象，本文将机器人与按钮间的接触状态划分为非接触、弹性接触、触头动作、破坏接触共 4 种接触状态.

如图 4 所示，在机器人按压按钮任务中，当机器人末端刚接触到按钮表面时，进入弹性接触阶段. 此阶段在线估计得到的按钮动力学参数不断变化，并且在开始接触时存在一些冲击，但各段刚度总体是线性

的. 随着按压行程的增加，进入触头动作阶段. 此阶段在按钮内部结构作用下触点接触或分离，产生冲击力并带来了一些非线性因素. 当机器人末端继续下按，进入第 2 个弹性接触阶段，并当按钮行程耗尽后，按钮的非活动部件开始形变，因此在极小的位移内产生巨大的接触力. 此阶段为破坏接触阶段，按钮极易产生不可逆的损坏，应避免进入.



（a）西门子 SIRIUS ACT 3SU1 按钮



（b）长江电气 LA38-11/203 按钮

图 4　按钮刚度特征与接触阶段划分

Fig.4　Stiffness characteristics and contact phase division of buttons

综合对机器人-按钮接触阶段的划分，笔者希望机器人学习到的技能为：通过感知按钮的动力学参数来识别触头动作阶段的反馈信号，并在触发按钮后于第 2 个弹性接触阶段返回初始位置. 这样的按钮操作过程与人类的操作技能相似，可避免进入破坏接触阶段，实现安全、准确的按钮操作.

3　按钮分类与仿真

3.1　按钮分类

依据对按钮操作特性和接触建模的分析，以及文献[13-15]的调研结果，从触头动作阶段的反馈信号特征出发，将按钮分为线性按钮和跳变按钮两类. 线性按钮的反馈力随其按压行程的增长而增长，典型代表如图 4(a)所示. 另一类是图 2(b)所示的跳变按钮，其在被触发时内部簧片机构会瞬间变形，从而出现反馈力突然下降的现象，其典型代表如图 4(b)所示.

仅从触头动作阶段的特性对按钮进行分类的方式并不全面，仍需综合考虑按钮的其他特性以按钮类型进行进一步的划分. 因此，本文从动力学特性和复位方式 2 个角度对按钮类型进行进一步划分，如图 5 所示.

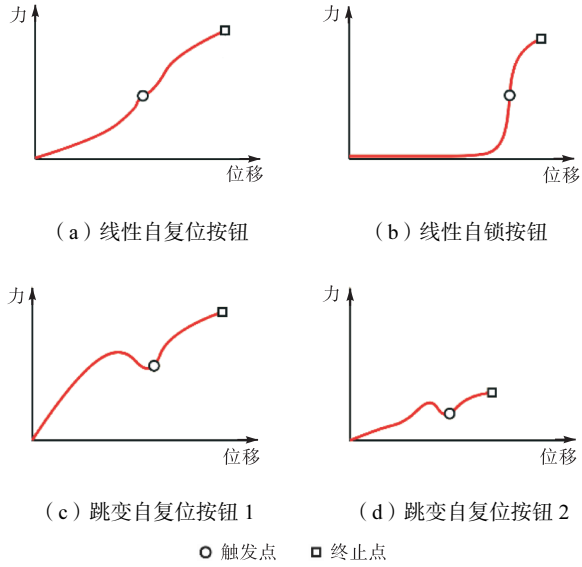


图 5　按钮类型划分

Fig.5　Classification of the buttons

触发信号类型可由按钮被触发时提供的力反馈信号特征区分，主要受按钮内部触发机构的影响，如图 5(a)、(c)所示. 对于按钮的动力学特性，由于各类按钮的设计尺寸以及装配误差等因素的影响，按钮的总体按压行程、触发行程、各段弹簧刚度包括触发力均不同，这些特性反应在按钮的力-位移曲线上便是其转折点位置、高度及各段曲线的斜率不同，如图 5(c)、(d)所示. 复位方式可由按钮被触发后是否可以自动复位来区分，如图 5(a)、(b)所示. 这种按钮分类方法涵盖了大多数类型的按钮，为后续生成全面的虚拟按钮模型数据提供了完善的分类依据.

3.2　按钮仿真数据生成

本文选择 Nvidia 公司的 Isaac Sim 仿真平台作为物理仿真环境，其集成的 PhysX 物理引擎可逼真地模拟刚体接触行为，并支持动态设定虚拟按钮模型的直线关节刚度，可实现对多种按钮模型力反馈特性的精确模拟. 在 Isaac Sim 中创建的虚拟按钮模型及内部直线关节如图 6 所示，其中的绿色网格为其碰撞模型，蓝色方框为按钮直线关节的两侧限位.

为了灵活生成虚拟按钮的力-位移曲线，本文引入文献[13]中的控制点概念，通过对代表按钮反馈特性

的分段线性函数中多个控制点坐标的调整，灵活调整各按压段的斜率、整体曲线形状，以生成多样的按钮模型. 以线性跳变按钮为例，通过调整原始按钮的控制点坐标生成了两种新的虚拟按钮模型，如图 7 所示.

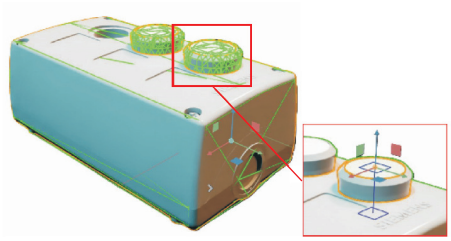


图 6 虚拟按钮模型

Fig.6 Virtual button model

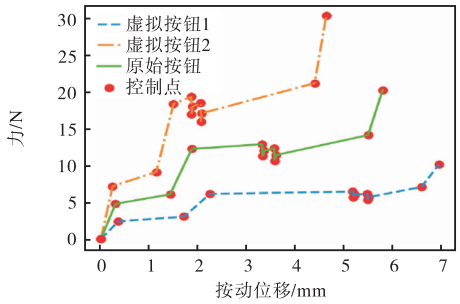


图 7 生成的虚拟按钮反馈曲线

Fig.7 Developed virtual button response curves

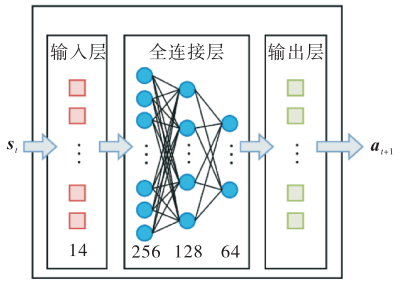
4 机器人按钮操作任务的强化学习训练

经过预训练得到的机器人控制策略可完成如下操作：机器人接近按钮—按压按钮—触发按钮—回退直至离开按钮表面，并可实现对按钮类型及末端执行器与按钮表面初始距离等不确定性因素的适应.

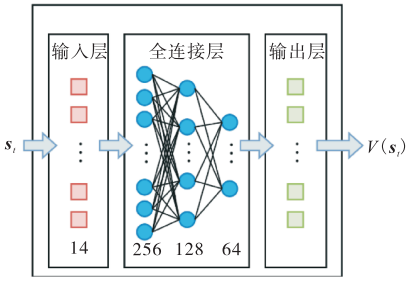
4.1 强化学习训练框架

任务过程可视为有限视界的马尔科夫决策过程(Markov decision process，MDP). MDP 的每个状态迁移均可用元组 $(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})$ 表示. 其中 $\boldsymbol{s}_t$ 为当前状态信息，r_t 和 $\boldsymbol{s}_{t+1}$ 为智能体采取行动 $\boldsymbol{a}_t$ 后从环境得到的奖励和状态反馈信息. RL 通过最大化一个长度为 L 的回合(episode)内的累积奖励 $R_t=\sum_{i=t}^{L}\gamma^{i-1}r(\boldsymbol{s}_t,\boldsymbol{a}_t)$ 来寻找最优策略 π^*，其中 γ 为折扣因子.

本文采用近端策略优化(proximal policy optimization，PPO)算法来进行机器人技能训练. PPO 算法凭借较强的学习稳定性、高计算效率和广泛的适用性，在机器人技能学习领域被大量采用. PPO 算法采用 Actor-Critic 架构，其使用 Actor 网络来选择输出的动作并更新策略，并使用 Critic 网络来评价所选择动作的好坏. 在本文的算法中，这两部分都使用以 elu 为激活函数的多层感知机(multilayer perceptron，MLP)网络，设计的具体网络结构参数如图 8 所示.



(a) Actor策略网络



(b) Critic价值网络

图 8 Actor-Critic 网络结构参数

Fig.8 Parameters of the Actor-Critic network architecture

基于 PPO 算法的机器人按钮操作策略的训练流程如图 9 所示. 机器人在训练回合接收到动作向量 $\boldsymbol{a}_t$ 后，转化为关节命令 $\alpha_{d1}\sim\alpha_{d6}$ 执行并与环境进行交互，交互对象包括随机化和添加噪声后的各种按钮模型. 对每步交互过程中产生的信息进行采集，部分信息经过 FFRLS 算法实时处理得到环境刚度参数，并将全部观测信息输入观测空间得到状态向量 $\boldsymbol{s}_{t+1}$. 将 $\boldsymbol{s}_{t+1}$ 输如 Actor-Critic 网络，输出新的动作 $\boldsymbol{a}_{t+1}$，然后评估状态的价值并更新网络参数.

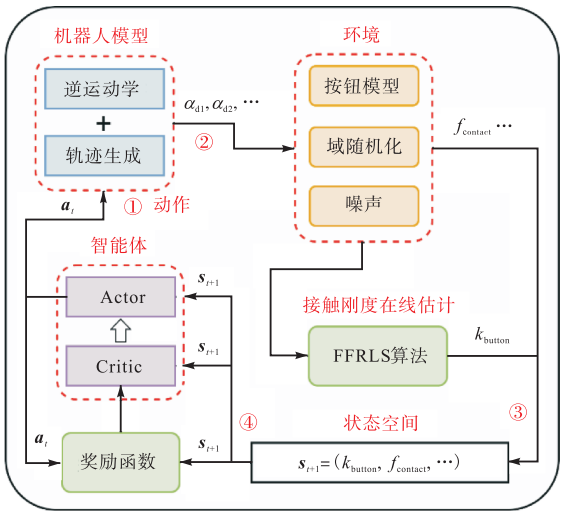


图 9 强化学习训练流程

Fig.9 Reinforcement learning training procedure

4.2　状态空间与动作空间设计

状态空间 $\boldsymbol{S}$ 是一个 14 维的向量，可表示为 $\boldsymbol{S}=\left[l_{\text{progress}}\quad \boldsymbol{J}^{6}_{\text{pos}}\quad \boldsymbol{P}^{3}_{\text{end}}\quad f_{\text{contact}}\quad d_{\text{rela}}\quad k_{\text{button}}\quad b_{\text{break}}\right]$，其中：$l_{\text{progress}}$ 为当前回合的步数进度；$\boldsymbol{J}^{6}_{\text{pos}}$ 为表示机器人关节位置的 6 维向量；$\boldsymbol{P}^{3}_{\text{end}}$ 为表示机器人末端执行器笛卡尔坐标的 3 维向量；f_{contact} 为机器人末端的接触力；b_{break} 表示按钮是否损坏；d_{rela} 为按钮的实时按压行程；k_{button} 为按钮的实时刚度.

动作空间 $\boldsymbol{A}$ 为策略 π 输出的笛卡尔空间下一维增量动作. δ_z 表示 t 时刻机器人末端在按钮表面的法线方向上的增量运动距离. 机器人未接触按钮时，$-0.500\ \text{mm}\leqslant\delta_z\leqslant 0.500\ \text{mm}$；机器人接触按钮后 $-0.025\ \text{mm}\leqslant\delta_z\leqslant 0.025\ \text{mm}$. 将当前帧产生的增量动作指令经逆运动学计算后传输给机器人关节执行.

4.3　奖励函数设计

奖励函数对于塑造智能体的行为非常重要，应能反映任务的最终目标，而不是传授智能体如何实现目标的先验知识. 在上述基础上，奖励函数的形式越简单，越易获得策略上的收敛. 结合任务流程与期望目标，本文设计的奖励函数为

$$r=r_{\text{trigger}}+r_{\text{break}}+r_{\text{complete}} \tag{7}$$

式中：$r_{\text{trigger}}=0.1$，获得条件为按钮被触发；$r_{\text{break}}=-1$，获得条件为机器人按压按钮直至到达破坏接触阶段，按钮视为损坏；$r_{\text{complete}}=1$，在机器人没有破坏按钮的情况下，完成按钮触发行为并脱离接触按钮后可获得.

4.4　域随机化

为了使策略获得适应不同起始按压高度和不同按钮反馈特性的能力，同时也可克服虚实环境之间的物理差异（sim-to-real gap），本文使用了域随机化的方法. 域随机化方法通过随机化训练过程中的物理参数，使预训练策略的适应域同时覆盖虚拟环境与真实世界中不同按钮的任务域，以获得更鲁棒的策略. 本文训练过程中的域随机化参数如表 1 所示.

表 1　域随机化参数

Tab.1　Domain randomization parameters

域随机化参数	说明
按钮类型与反馈特性	第 3.2 节仿真数据库
按钮初始高度/cm	[0,20]
表面材料	随机
材料密度/(kg/m^3)	[0.2,5.0]×初始值
关节摩擦系数	[0.5,2.0]×初始值
力传感器噪声	随机
控制系统时延/ms	[10,30]

在每一次训练开始前，针对上述所有类别的随机因素进行随机抽取并组合成一套随机参数，然后将该套随机参数应用于本次的训练回合中. 本文的域随机化方法主要针对被操作对象即按钮，产生庞大的按钮数据集以供智能体进行交互训练，拟合出通用性更强的按钮操作策略.

4.5　训练结果分析

如图 10 所示，本文使用 PPO 算法在 Isaac Sim 中同时进行 64 组任务场景的模拟训练，每训练回合的长度最大为 1 000 步. 多组场景的并行模拟可极大增加状态-动作对（state-action pair）的采样效率，减少训练时间，保证对环境的充分探索. 图 11 显示了训练过程中每回合的平均奖励，随着训练的进行，智能体获得的平均奖励呈上升趋势，表明智能体在训练中持续学习并向理想策略靠拢. 经过 2 000 个回合（约 818×10^4 训练步骤）的迭代训练后平均奖励曲线趋于稳定，并且训练后期各组智能体的任务完成率均在 90% 以上，表明拟合出的策略性能良好.

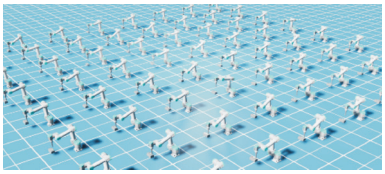


图 10　机器人按钮操作任务训练过程

Fig.10　Training process for robot button operation tasks

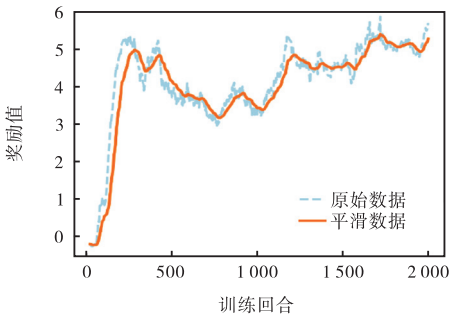


图 11　机器人按钮操作训练的奖励曲线

Fig.11　Reward curves of robot button operation training

5　机器人按钮操作实验

本文首先设计实验以检验预训练策略在模拟环境下的操作性能，然后将预训练策略迁移部署至真实机器人，在真实世界下进行相同的实验，以验证预训练策略的有效性和经虚实迁移后的性能差异. 其次，引入多种力反馈特性不同的按钮作为实验对象，并与 2 种基线方法进行比较，以检验本学习框架所形成策略泛化能力.

5.1　实验准备

本文采用实验室自研的六自由度协作机器人,其负载为 5 kg. 机器人末端配置了一个六维力传感器,其分辨率为 0.025 N. 末端执行器通过法兰与六维力传感器连接. 实验场景如图 12 所示,按钮盒外接信号灯,信号灯亮则代表按钮成功触发. 每次按钮操作成功的标准是显示灯亮(表明机器人准确触发按钮),并且此次操作中按钮行程未进入破坏接触阶段,最后机器人末端执行器离开按钮.

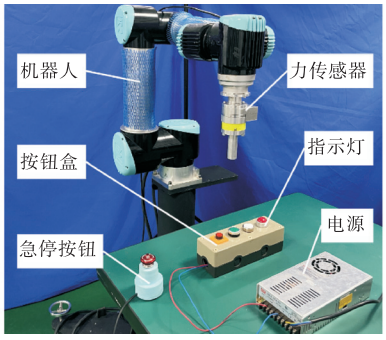


图 12　机器人按钮操作实验平台

Fig.12　Experimental platform of robot button operation

5.2　按钮操作策略的虚实迁移实验

在本实验中,本文同时在虚拟环境和真实世界下进行 2 种经典按钮的操作实验,以检验预训练策略的有效性和 sim-to-real 传输的效果. 在虚实环境中,本文均采用 100 Hz 的控制频率,即完成一次数据接收、策略推理、指令发送和机器人执行的完整控制流程的循环时间为 10 ms. 但由于通信延迟与控制器运动插补功能的最少轨迹点需求,在真实世界下进行实验时,存在 3 个控制循环(30 ms)的延迟. 虚实迁移实验的原理如图 13 所示.

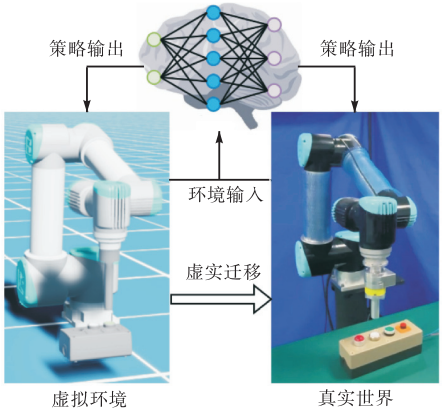


图 13　虚实迁移实验原理

Fig.13　Principle of the sim-to-real experiment

5.2.1　虚拟环境实验

本文通过预先采集的按钮触发特性,建立了 2 种经典的虚拟按钮模型. 在每次实验开始前,将 2 种虚拟按钮的初始位置随机分布到距离桌面高度 0 ~ 10 cm 的区间上,具体高度值对机器人来说是未知的,以模拟实际任务场景. 然后使用预训练策略驱动虚拟机器人与 2 种按钮各进行 50 次按压实验,并统计任务成功率,如表 2 所示. 虚拟机器人对 2 种虚拟按钮的操作成功次数分别为 47 和 49,任务总成功率均为 96%.

表 2　虚实迁移实验结果

Tab.2　Results of sim-to-real experiment

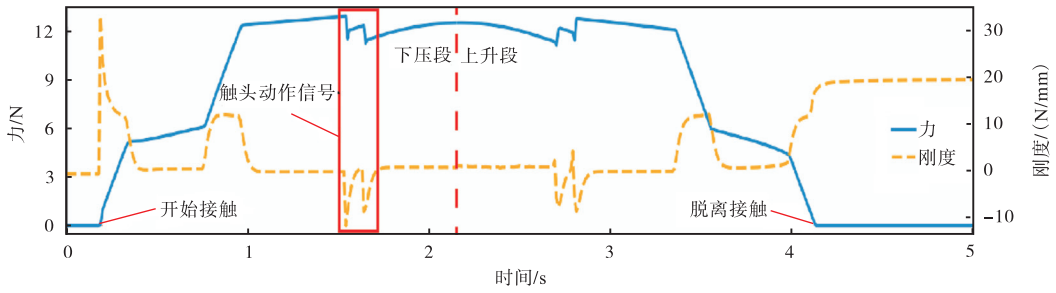
实验场景	成功按压按钮 1 次数	成功按压按钮 2 次数	总成功率/%
虚拟环境	47	49	96
真实世界	39	44	83

2 次成功完成按动任务的力-时间、刚度-时间曲线如图 14(a)、(c)所示. 如图可见,在机器人末端接触按钮的瞬间,由于动冲击产生的较大接触力变化导致在线刚度估计算法产生了较大的环境刚度估计. 在继续向下按压的过程中,机器人侦测到 2 种按钮触头动作阶段的刚度信号后(图 14(a)、(c)红色方框部分),综合力、位移信息来判断按钮的按压阶段和触发状态,认为本次操作已成功触发按钮. 于是在 2.2 s 和 2.6 s 处策略开始输出机器人向上抬起的运动指令,直至机器人离开按钮表面. 实验结果显示,在虚拟环境下该按钮操作策略成功地完成了 2 种不同按钮的操作任务,预训练策略对于 2 种经典按钮的触发信号识别较为准确,并在触发按钮后成功返回初始位置.

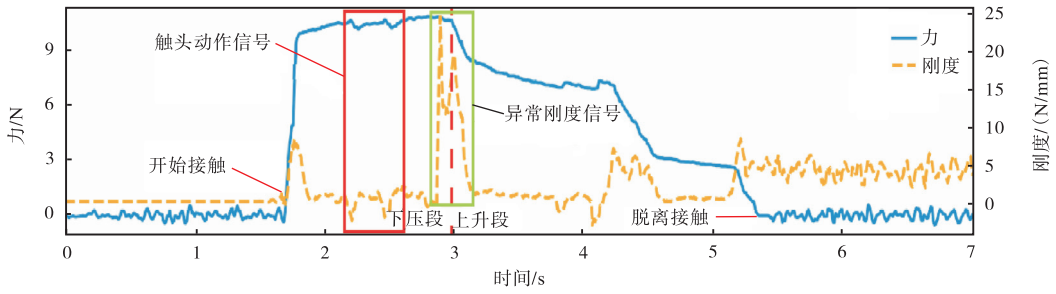
5.2.2　真实世界实验

在虚拟环境下成功验证了按动策略的有效性后,将同一套策略向真实机器人进行迁移,并在两种真实按钮上测试策略效果,实验场景如图 13 所示. 实验结果如表 2 所示,50 次按压实验中真实机器人对 2 种真实按钮的操作成功次数分别为 39 和 44,任务总成功率均为 83%.

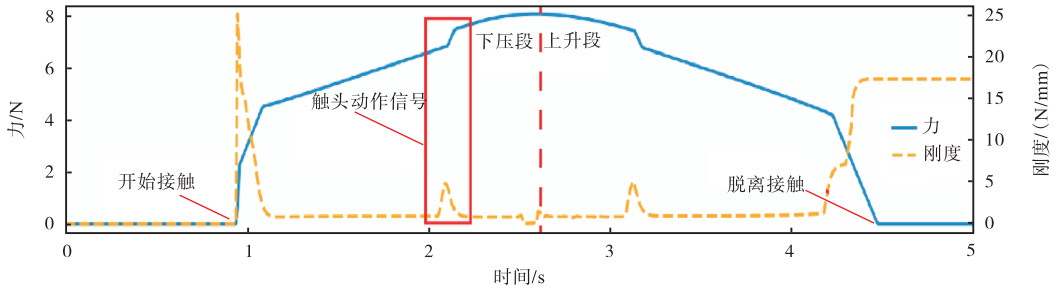
2 次成功完成按动任务的力-时间、刚度-时间曲线如图 14(b)、(d)所示. 与仿真结果类似,真实机器人在初始阶段以较大速度不断向按钮方向进行探索,并且末端执行器与按钮发生接触时侦测到较大的环境刚度. 检测到接触行为的发生后,控制策略减缓机器人按压速度以获得更精确的环境刚度参数. 当机器人继续下压并且识别到按钮触发的刚度信号后(图 14(b)、(d)红色方框),控制策略认为按钮已被成功触发并驱动机器人向上抬起,直至离开按钮表面.



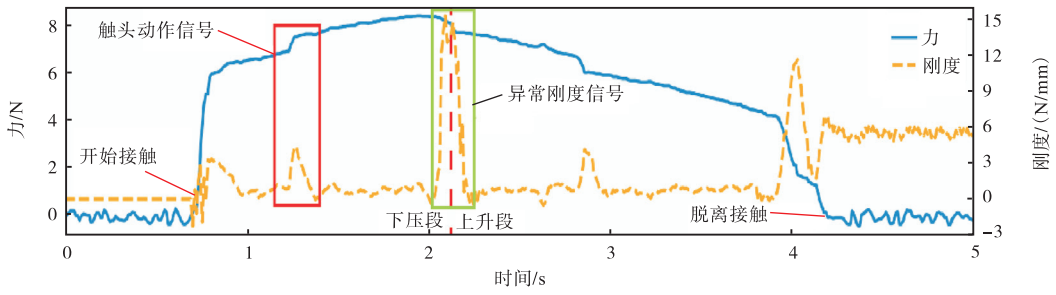
（a）虚拟环境实验的跳变按钮操作过程



（b）真实世界实验的跳变按钮操作过程



（c）虚拟环境实验的线性按钮操作过程



（d）真实世界实验的线性按钮操作过程

图 14　虚实迁移实验的按钮操作信息

Fig.14　Information about pushing the button in sim-to-real experiment

与在真实世界测得的机器人末端抬起时的力-时间曲线要显著低于下压段的力-时间曲线，并且按钮的下压段与抬起段的力-时间曲线并不完全关于返回点对称，这与按钮的结构特性有关[19-20]. 上述这些因素均属于现实环境中的不确定项，对策略的感知和决策有较大干扰，这也是真实世界实验成功率低于虚拟环境的主要原因. 在真实世界实验中，机器人在按压和返回段的交接处还会产生异常的刚度估计（图 14(b)、(d)绿色方框），这是由于在该阶段位移基本不变而接触力在返回阶段出现上述不对称下降的情况引起的. 此时机器人已进入返回阶段，所以该异常刚度值对策略的判断影响不大.

5.2.3　实验结果分析

综合仿真实验与真机实验的效果来看，两者均能以较高的成功率执行不同未知高度下的 2 种按钮的操作任务，证明了本研究提出的基于虚实迁移强化学习的机器人通用按钮操作策略训练框架的有效性. 预训练策略在仿真环境下的效果更好，而在真实机器

人上运行时的操作成功率有一定下降，但仍保持着较高的水平，并可处理一些仿真训练时未出现的干扰因素. 这说明在真实世界的多重非结构化因素的影响下，预训练策略仍保持着一定的决策判断能力，具有一定的鲁棒性. 这也说明了有大量随机化数据参与的 sim-to-real 流程在对抗非结构化环境时在性能、成本和安全性上的优越性.

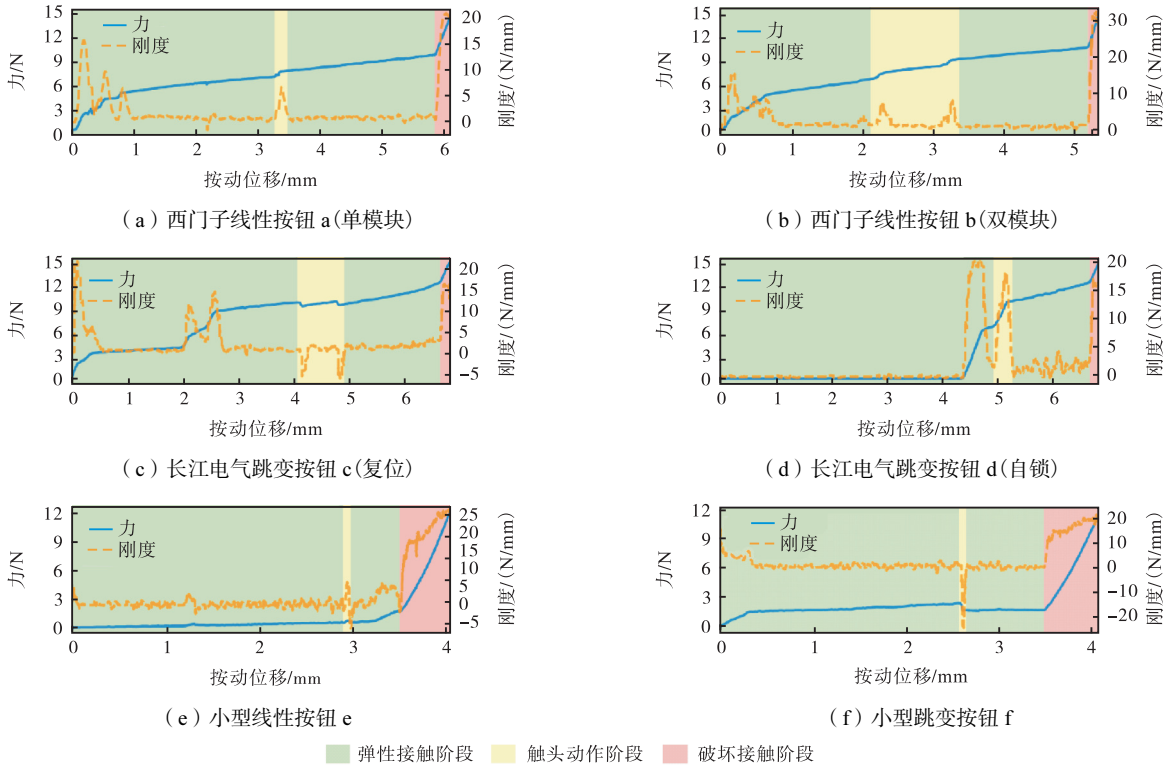
5.3　按钮操作策略的泛化实验

5.3.1　实验对象

为了评估预训练得到的通用按钮操作策略的泛化能力，本研究拟选取多种不同按钮作为操作对象，并将本研究训练出的策略与 2 种经典按钮操作方法[8]进行对比.

本研究依据第 3.1 节提出的按钮分类方法，从不同触发类型、不同动力学参数和是否回弹 3 个方面分类并选取了 6 种按钮作为泛化实验的按钮操作对象，其名称与力-位移反馈特性如图 15 所示.

按钮 a ~ f 具有电气领域按钮的大部分反馈特征，按钮动力学参数差异极大，具体表现在其最大设计按压行程范围为 3.4 ~ 5.8 mm，触发位移为 2.2 ~ 5.2 mm，触发力为 2.5 ~ 10.4 N，触发刚度阈值为 −25 ~ 17 N/mm. 选取的按钮实验对象具有一定代表性，可较好地检验预训练策略在全新按钮样本上的操作能力，体现其泛化性.



(a) 西门子线性按钮 a(单模块)　(b) 西门子线性按钮 b(双模块)

(c) 长江电气跳变按钮 c(复位)　(d) 长江电气跳变按钮 d(自锁)

(e) 小型线性按钮 e　(f) 小型跳变按钮 f

图 15　6种按钮力-位移反馈特性

Fig.15　Force-displacement feedback characteristics of six buttons

5.3.2　基线方法

基线方法主要包括固定力阈值按动方法以及固定刚度阈值按动方法. 固定力阈值按动方法主要思想是在机器人匀速靠近按钮时，检测末端在按钮表面法向上的接触力. 当接触力超过指定按钮的触发力时，控制系统认为已经触发按钮并控制机器人向上抬起直至离开按钮. 固定刚度阈值按动方案是目前已知较先进的机器人按钮操作方法[8]，研究者采用了一种基于模仿学习的刚度阈值学习方法，使机器人从按钮接触刚度层面习得示教者的轨迹、接触力和刚度信息并应用到后续的按钮操作过程中. 主要流程为机器人在按动过程中进行按钮刚度的在线估计，到达预计刚度阈值后认为成功触发按钮，然后生成轨迹退回初始位置.

实验场景如图 16 所示，机器人分别在 3 种策略的控制下，以 100 Hz 的控制频率与 6 种不同按钮进行各 20 次交互. 依据两种基本按钮的力-位移特性(图 4)，将固定力阈值按动方法的触发力阈值设置为 9 N，将固定刚度阈值方法的触发刚度阈值设置为 20 N/mm. 根据指示灯的亮灭情况与操作过程中的力-位移特征曲线记录每次操作的结果，成功则表示本次操作机器人成功触发按钮，并且未达到破坏接触阶段；未触发则表示本次操作机器人未触发按钮便回退；损坏表示本次操作机器人已成功触发按钮，但达

到了按钮的破坏接触阶段.

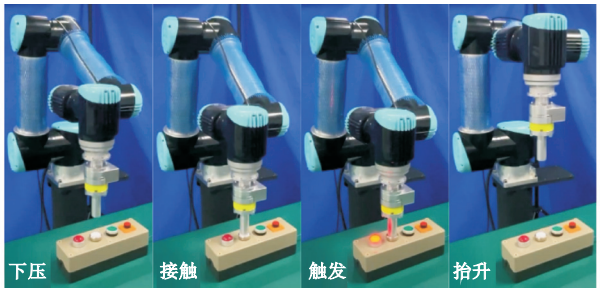


图 16　泛化实验场景

Fig.16　Scenario of the generalization experiment

5.3.3　实验结果分析

各组按钮操作结果如图 17 所示，从图 17(a)可得到 3 种机器人操作策略中固定力阈值策略的总成功率为 33.3%，固定刚度阈值策略的成功率为 0，本文策略成功率为 70.8%. 从泛化操作实验的结果来看，经本文提出的学习框架训练出来的控制策略对不同按钮的适应性远高于其他 2 种方法.

由图 17 可看出，根据固定接触力来判断按钮触发状态的方法对于相近尺寸且力随按动行程线性上升的按钮(按钮 a、b)效果很好；对小尺寸的按钮(按钮 e、f)可实现成功触发但会进入破坏接触阶段；对跳变型的按钮(按钮 c、d)，其反馈力不随按动位移上升而上升，固定力策略在触发按钮前侦测到大于 9 N 的力便回退，无法触发按钮.

固定刚度阈值的操作方法存在两方面问题. 一方面，在机器人末端执行器开始接触按钮表面时会产生冲击载荷，得到较大的接触刚度估计值，导致控制策略认为已触发按钮并控制机器人开始回退. 如图 17(b)所示，在初始刚度较大的跳变按钮(按钮 c、d)的操作过程中，机器人没能成功触发按钮就开始回退. 另一方面，根据固定刚度阈值判断按钮的触发状态往往只能用来识别按钮的破坏接触阶段，如图 17(c)所示. 虽然大部分实验中可成功触发按钮，但无法准确感知按钮的实际触发点. 这对线性按钮或跳变按钮来说均容易带来按钮结构件的永久变形，减少按钮寿命. 甚至对于小型按钮来说，可能一次按压就意味着按钮的损坏，这是不可接受的.

在使用本文策略时，机器人在操作过程中可处理冲击力与振动噪声、触发力与触发位置的不确定性，依据刚度信号准确识别各种按钮的触发状态，实现自主的决策和操作. 如图 17(a)所示，本文策略在除自锁跳变按钮(按钮 d)外的所有未知类型的按钮操作任务中获得了不低于 65% 的操作成功率，证明本文策略的操作性好、稳定性高. 同时，本文方法在成功率、按钮损坏率等指标方面均大幅领先另外 2 种方法，具有安全性和先进性. 最后，经由本文的学习框架训练出的按钮操作策略，未经过真实世界的适应性训练而可直接针对未知操作对象使用，展示了该策略的通用性和良好的泛化能力.

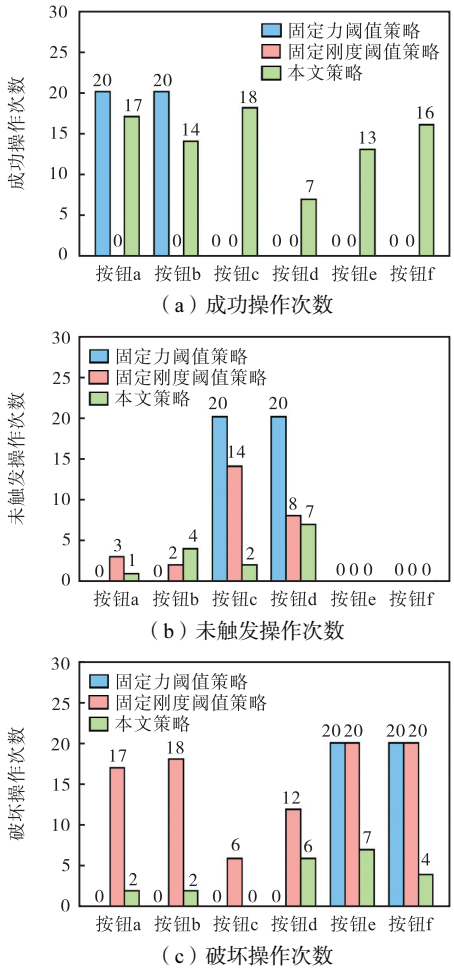


图 17　按钮泛化操作实验结果

Fig.17　Results of generalization experiment for pushing the button

6　结　语

本文提出了一种基于 sim-to-real 方法的机器人力交互操作技能训练框架，并以机器人按钮操作场景为例，训练出了一种通用的按钮操作具身智能系统，可自主决策操作多种不同按钮.

首先，通过划分按钮模型的接触阶段，并模仿人类的按钮操作技能中对接触刚度的感知与对触发信号的分辨来确认学习目标. 然后从 2 类经典按钮模型出发，在虚拟环境中随机生成大量按钮数据并使智能体与之交互，快速安全地拟合出泛化性极强的通用按钮操作策略，并可弥合虚实环境间的差距. 最后，本文通过真实世界机器人与多种未知按钮的交互实

验,验证了在该框架下训练出的按钮操作策略的有效性、鲁棒性、泛化性和先进性.

未来的工作主要集中在优化按钮操作策略的控制性能,例如可加入柔顺控制算法以减少接触初期的冲击力、减少策略误判等情况. 同时,可进一步将该框架推广到更多类似任务的学习过程中,例如学习多段行程开关、旋钮等更广泛类型开关的操作技能,以及学习门窗、抽屉的开合等类似任务.

参考文献:

[1] Cheng M, Xiang D. The design and application of a track-type autonomous inspection robot for electrical distribution room[J]. Robotica, 2020, 38(2): 185-206.

[2] Sukhoy V, Sinapov J, Wu L, et al. Learning to press doorbell buttons[C]//2010 IEEE International Conference on Development and Learning. Ann Arbor, USA, 2010: 132-139.

[3] Wang F, Chen G, Hauser K. Robot button pressing in human environments[C]// 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia, 2018: 7173-7180.

[4] Ravichandar H, Polydoros A S, Chernova S, et al. Recent advances in robot learning from demonstration[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3(1): 297-330.

[5] Brunke L, Greeff M, Hall A W, et al. Safe learning in robotics: From learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5(1): 411-444.

[6] Ren L, Dong J, Liu S, et al. Embodied intelligence toward future smart manufacturing in the era of AI foundation model[J]. IEEE/ASME Transactions on Mechatronics, 2024, 30(4): 2632-2642.

[7] Racca M, Pajarinen J, Montebelli A, et al. Learning in-contact control strategies from demonstration[C]// 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Republic of Korea, 2016: 688-695.

[8] Liu X, Huang P, Liu Z. A novel contact state estimation method for robot manipulation skill learning via environment dynamics and constraints modeling[J]. IEEE Transactions on Automation Science and Engineering, 2022, 19(4): 3903-3913.

[9] Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International Journal of Robotics Research, 2018, 37(4/5): 421-436.

[10] Andrychowicz M, Baker B, Chociej M, et al. Learning dexterous in-hand manipulation[J]. The International Journal of Robotics Research, 2020, 39(1): 3-20.

[11] Matas J, James S, Davison A J. Sim-to-real reinforcement learning for deformable object manipulation[C]// Conference on Robot Learning. Zurich, Switzerland, 2018: 734-743.

[12] Hebecker M, Lambrecht J, Schmitz M. Towards real-world force-sensitive robotic assembly through deep reinforcement learning in simulations[C]//2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). Delft, Netherlands, 2021: 1045-1051.

[13] Liao Y C, Kim S, Lee B, et al. Button simulation and design via FDVV models[C]// Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA, 2020: 1-14.

[14] Kim S, Lee G. Haptic feedback design for a virtual button along force-displacement curves[C]//Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology. St Andrews, UK, 2013: 91-96.

[15] Tashiro K, Shiokawa Y, Aono T, et al. Realization of button click feeling by use of ultrasonic vibration and force feedback[C]// World Haptics 2009—3rd Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. Salt Lake City, USA, 2009: 1-6.

[16] Elguea-Aguinaco Í, Serrano-Muñoz A, Chrysostomou D, et al. A review on reinforcement learning for contact-rich robotic manipulation tasks[J]. Robotics and Computer-Integrated Manufacturing, 2023, 81: 102517.

[17] Jasim I F, Plapper P W. Contact-state monitoring of force-guided robotic assembly tasks using expectation maximization-based Gaussian mixtures models[J]. The International Journal of Advanced Manufacturing Technology, 2014, 73(5/6/7/8): 623-633.

[18] Roveda L, Vicentini F, Tosatti L M. Deformation-tracking impedance control in interaction with uncertain environments[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan, 2013: 1992-1997.

[19] Oulasvirta A, Kim S, Lee B. Neuromechanics of a button press[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal, Canada, 2018: 1-13.

[20] Valverde N, Ribeiro A M R, Henriques E, et al. An engineering perspective on the quality of the automotive push-buttons' haptic feedback in optimal and suboptimal interactions[J]. Journal of Engineering Design, 2019, 30(8/9): 336-367.

(责任编辑: 王晓燕)