arXiv:2305.00929v1 [cs.RO] 1 May 2023

从人类示范和实时不确定性感知干预中学习飞行控制系统⋆

Prashant Ganesh * **J. Humberto Ramos** * **Vinicius G. Goecks** **
Jared Paquet * **Matthew Longmire** * **Nicholas R. Waytowich** **
Kevin Brink ***

*佛罗里达大学机械与航空航天工程系，*REEF*，*Shalimar*，佛罗里达州 *32579*，美国（电子邮箱：*prashant.ganesh@ufl.edu,jramoszuniga@ufl.edu, jaredpaquet@ufl.edu, m.longmire@ufl.edu*）

**美国陆军研究实验室人类研究与工程局，阿伯丁试验场，马里兰州 *21005*，美国（电子邮箱：*vinicius.goecks@gmail.com, nicholas.r.waytowich.civ@army.mil*）

***空军研究实验室弹药局，埃格林空军基地，佛罗里达州 *32542*，美国（电子邮箱：*kevin.brink@us.af.mil*）

摘要： 本文介绍了一种学习飞行控制系统的方法，该方法在学习过程中利用人类示范和干预，并考虑学习得到的模型中的估计不确定性。所提出的方法利用人类示范通过模仿学习训练初始模型，随后通过实时人工干预反复提升其性能。干预旨在纠正不良行为，并使模型适应任务动力学的变化。通过蒙特卡洛丢弃实时估计学习得到的模型的不确定性；当该不确定性超过预定义阈值时，通过视听信号提示人类监督者进行干预。该方法在固定平台和移动平台上的自主四旋翼着陆任务中得到验证。结果表明，借助该算法，人类可以通过演示专家轨迹，快速教会无人飞行器执行飞行任务；当学习得到的控制器出现不良机动、任务发生变化和/或模型不确定性超过阈值时，人类还可以通过干预来调整学习得到的模型。

关键词： 人机系统、自主机器人系统、智能控制器、知识建模、人在回路机器学习、反馈学习、实验验证。

1. 引言

对于自主或智能系统所需的许多能力而言，解析控制律的设计与实现可能难以完成，或面临极大的挑战。对于极其难以建模的高度复杂系统，这一点尤为明显。在这些情况下，基于机器学习的方法可以作为一种有用的替代方案，无需设计经典控制器，也能实现复杂的能力并捕捉复杂的动态行为。尽管机器学习技术往往缺乏解析保证，但只需相对少量的第一性原理设计与调参，就能生成高效的控制器（也称为策略），以执行具有挑战性的任务。此前的研究已证明，这类技术可以应用于具身硬件平台 Ng 等人（2006）； Ross 等人（2013）； Giusti 等人（2015）； Zhang 等人（2016）； Bojarski 等（2016）； Gandhi 等人（2017）； Goecks 等人（2018），以及模拟环境中的 Gu 等人（2017）； Nair 等（2018），和实体机器人上的 OpenAI 等（2018）； Tobin 等人（2018）； Thomas 等人（2018），包括飞行器上的 Blukis 等人（2018）； Anderson 等人（2018）； Mei 等人（2016）； Misra 等人（2017）； Tai 等人（2017）。

这些端到端数据驱动方法通常使用深度神经网络来表示控制策略，这需要大量数据，因此需要较长的训练时间才能实现预期的系统行为。此外，在达到预期行为之前，表现不佳的策略可能会导致灾难性动作，而这在使用物理系统时可能无法接受。学习得到的行为也可能很脆弱，无法适应环境或受控对象的细微变化，从而导致整体表现不佳。为应对这些挑战，本文提出一种人在回路中的方法，以在线方式学习并调整控制策略。具体而言，该方法从人类示范任务中学习控制策略，并

⋆本研究由美国陆军研究实验室资助，并依据编号为 W911NF-18-2-0134 的合作协议开展。本文所含观点和结论均为作者观点，不应被解读为代表美国陆军研究实验室或美国政府明示或暗示的官方策略。

在学习得到的控制器执行任务时，根据人工干预对其进行调整。由于该方法在线进行模型的学习和调整，因此学习速度更快，经过物理验证，且通常所需的训练数据更少。

此前已有研究探讨了如何从人类示范和人工干预中学习 Ross 等人 (2013)，以重新标注不良行为，并主动介入以避免灾难性故障 Hilleli 和 El-Yaniv(2018)，以及从这些干预中学习 Goecks 等人 (2019a)；通过稀疏评价反馈间接引导学习 Warnell 等人 (2018)，并将这些人类反馈方式与强化学习相结合 Way- towich 等人 (2018)；Goecks 等人 (2019b)，甚至在人类示范和干预由熟练与非熟练用户共同提供、且用户需要判断何时进行干预的情况下也是如此 Wu 等人(2022)。然而，在任务复杂度发生变化且实时人工干预依据模型不确定性的情形下，尚未应用人在回路机器学习框架。

本文采用由不确定性信息引导的实时人类示范和干预，训练基于深度神经网络的四旋翼飞行控制器，使其能够在静态和动态着陆平台上着陆。首先，仅利用人类示范训练四旋翼飞行器在静态平台上着陆，并将其与执行相同任务的经典控制器进行比较。随后表明，当模型不确定性超过预定义阈值时，利用实时干预提供的额外数据纠正系统行为并重新训练模型，可以提升模型性能。神经网络不确定性通过蒙特卡洛丢弃计算 Gal 和 Ghahramani (2016)，在监督车辆行为时，则通过视听提示通知人类进行干预。最后，任务变为在移动平台上着陆；实验表明，少量干预就足以使模型实时适应任务变化，而经典控制器则需要离线重新设计。我们展示了，借助这一人在回路机器学习框架，可以增强类人控制策略，只需极少的训练时间便能取得成功，且无需进一步手动调参。所有所述机器学习程序均在车辆上运行。

本工作的主要贡献如下：

- 新颖的人类参与式机器学习框架，利用估计的模型不确定性进行有针对性的训练，以改进任务执行；
- 通过实验将不确定性感知的人类参与式机器学习技术与经典控制器进行比较，展示了该技术在处理动态任务时的优势；
- 对一种新颖的人类参与式机器学习技术进行了广泛的硬件验证，该技术能够实时学习任务 Goecks 等人（2019a）。

2. 背景

在本研究中，我们使用深度神经网络来表示学习控制器。尽管深度神经网络技术已取得巨大成功，但通常需要大量数据和漫长的训练时间才能收敛到良好的解。人在回路机器学习是一类利用通过各种形式的人机交互收集的数据来增强训练过程的技术。

2.1 从人类示范中学习

示范学习 （LfD）利用人类示范或任务示例进行学习 Argall 等人 （2009年）； Osa 等人 （2018年）。在 LfD 中，策略 π_θ 是一个由 θ 参数化的函数，将状态 $\boldsymbol{s}$ 映射为动作 $\boldsymbol{a}$。

$$\pi_\theta : S \rightarrow A \tag{1}$$

在控制理论的记法中，$\boldsymbol{s}$ 类似于系统状态向量 $\boldsymbol{x}$，而 $\boldsymbol{a}$ 对应于控制向量$\boldsymbol{u}$，如图 1a-b 所示。对于部分可观测的情况，则使用观测$\boldsymbol{o}$ 代替状态$\boldsymbol{s}$；它对应于控制理论中用于没有全状态反馈的系统的测量向量$\boldsymbol{y}$ 。

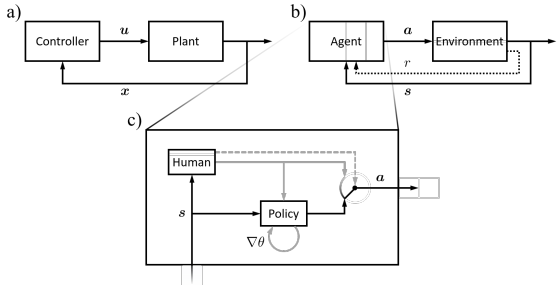


图 1. a）典型反馈控制系统和 b）具有奖励信号（虚线）的类比学习智能体系统的框图。c）LfD 和 LfI 的智能体模块的插图，其中人类触发器（虚线）可在当前策略输出的动作（黑色）与人类自身输出的动作（灰色）之间切换，同时更新策略。

策略 π_θ 经过训练，可在任务示范过程中访问的一组 $\mathcal{D}$ 状态和动作上进行泛化；该示范持续T 个时间步，即状态和动作轨迹随时间变化的历史记录：

$$\mathcal{D} = \{\boldsymbol{a}_1, \boldsymbol{s}_1, \boldsymbol{a}_2, \boldsymbol{s}_2, ..., \boldsymbol{a}_T, \boldsymbol{s}_T\}. \tag{2}$$

该示范可由解析推导的控制器、预训练策略或人类监督者执行。对于人类示范，人类根据观察到的状态提供动作，这些动作可表示为其内部策略 π_{sup}的响应，从而生成 $\mathcal{D}$。策略 π_θ 的训练属于标准监督学习问题，旨在估计人类的内部策略 $\hat{}\pi_{sup}$；其中，参数θ 经过训练以最小化损失函数，例如均方误差，如式 (3) 所示。

$$\hat{\pi}_{sup} = \underset{\pi_\theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} ||\pi_\theta(\boldsymbol{s}_t) - \boldsymbol{a}_t||^2 \tag{3}$$

近年来，在自动驾驶汽车等领域应用模仿学习来训练自主系统，已取得多项实证成果 Bojarski 等

（2016年）以及模拟环境中的Goecks 等人 （2018年）和真实无人机飞行中的Giusti 等人 （2015年）。

2.2 从人类干预中学习

在 从干预中学习 (LfI) 中，人类充当监督者，观察智能体执行任务，并在必要时进行干预，即覆盖智能体的动作，以避免可能导致灾难性状态的不良行为或不安全行为。这项技术被用于在安全强化学习框架下训练智能体，使其不会发生灾难性故障 Saunders 等人 (2017年)，其中人类监督者会监控学习中的智能体，并在认为智能体的动作可能造成灾难性后果时提供输入，阻止其执行这些动作。随后，研究人员利用这些数据训练一个函数，使其能够自动执行阻止操作。Hilleli 和 El-Yaniv (2018年)开展的类似研究则利用人机交互训练分类器，以检测不安全状态。

2.3 从演示和干预中学习

结合从演示中学习和从干预中学习已成功应用于 Goecks 等人 （2019a），用于训练全连接深度神经网络，使无人机能够在高保真无人机模拟器中降落到目标着陆垫上；这两种方法也应用于自动驾驶任务 Kelly 等人（2019年），并在模拟和真实测试车辆中进行了验证。先前的两项研究均表明，这些学习技术相结合的效果优于单独使用任一技术，且数据效率更高；这也是本研究的主要灵感来源。

然而，提供干预的用户只能依靠自己的直觉来判断何时纠正智能体的行为，这进一步加重了飞行员原本就承受的认知负担。在本研究中，我们通过实时计算模型的不确定性并将其显示给人类，作为何时进行干预的指示，从而减轻了这一负担。

3. 方法

所提出的方法首先从专家人类示范者处收集一组任务示范；在本例中，任务是让四旋翼飞行器降落在固定平台上。策略 π_θ 最初根据公式 3，使用这组任务示范进行训练。该策略的输入包括四旋翼飞行器与着陆垫之间的真实相对 x 和 y 位置、以欧拉角表示的飞行器真实全局偏航角，以及通过开源 REEF 状态估计器 Ramos 等人(2019年)计算出的飞行器估计位置（高度）。策略的输出是由人类示范者通过操纵杆输入生成的速度。这些参考速度随后被发送至 REEF 自适应控制器，由其输出扭矩指令，以达到期望速度。该自适应控制器的作用是应对飞行器动力学的细微变化，例如配平或电池配置的变化，并提升飞行性能。

策略通过这组初始任务示范训练完成后，将控制权交给智能体，智能体执行π_θ ，同时人类监督智能体的动作。在智能体执行任务时，我们计算神经网络使用蒙特卡洛丢弃的不确定性Gal 和 Ghahramani (2016年），这是一种在推理过程中使用丢弃层来生成多次随机前向传播，然后对这些传播结果进行汇总，以计算神经网络模型的预测均值和预测不确定性的技术。该方法的主要优势在于，只要能够在推理过程中启用丢弃层，就可以应用于已经使用丢弃层训练的现有神经网络模型。

估计不确定性数值会转换为视听信号，在智能体于高不确定性区域导航时提示人类监督者进行干预。每当智能体表现出不良行为时，人类也会介入，提供纠正操作，并在认为合适时将控制权交还给智能体。随后，智能体通过将干预期间记录的状态和动作添加到原始训练数据集中来学习这次干预，并对 π_θ 进行微调。智能体随后执行新的策略 π_θ；该策略会在收集到干预数据后更新。人类用户继续监督，并根据需要进行干预，通过操纵杆上的按钮，在自己与当前智能体策略之间切换飞行器的控制权。

人类用户可以根据需要灵活切换：既可以提供演示，即覆盖整个着陆机动、不中断的状态-动作轨迹；也可以提供干预，即针对着陆机动中某些片段施加的纠正性状态-动作轨迹，如图 1c 所示。结合演示和干预后，该框架能够超越仅从人类示范中学习得到的策略，而且所需的训练数据更少Goecks 等人(2019a）。这部分归因于从干预中学习阶段对状态空间进行了更多样的探索，而从演示中学习阶段的探索则未能达到这一程度，并因此导致失败。在这些失败状态下进行有针对性的干预，可以让智能体在策略最薄弱的区域高效地更新策略。智能体熟练掌握在固定平台上着陆后，任务会通过逐步增加着陆垫的运动而变得更加复杂。一旦平台开始移动，就会收集更多人工干预，以便更新策略 π_θ ，使其实时适应这一新任务。

3.1 实现细节

本文所采用的学习框架总结于算法 1 中。主程序初始化智能体的策略 π_θ、人工数据集 $\mathcal{D}$，以及一个次级策略更新例程，用于利用新的演示或干预数据更新策略。智能体的策略 π_θ 是一个全连接的三隐层神经网络，三层分别包含 130、72 和 40 个神经元。所有层均为全连接层，并采用双曲正切函数作为激活函数。前两层使用Srivastava 等人(2014)提出的丢弃层，丢弃概率为 25%，以防止过拟合并计算不确定性。网络通过从正态分布中采样权重进行随机初始化。

主循环主要由读取当前环境状态，以及执行由智能体当前策略或人类监督者提供的动作构成。当人类监督者按住操纵杆上的触发器并输入飞行指令时，这些

算法 1 从人类示范中学习与不确定性感知干预

过程 MAIN
  初始化 策略 π_θ，数据集 $\mathcal{D}$
  当 True**时**
    记录状态 $\boldsymbol{s}$
    如果人类正在交互**则**
      记录人类动作$\boldsymbol{a}$
      添加状态-动作对$\boldsymbol{s}$，$\boldsymbol{a}$ 到 $\mathcal{D}$
    else
      从策略 $\pi_\theta(\boldsymbol{s})$ 中采样多个动作$\boldsymbol{a}$
      计算以下内容的均值和标准差这些动作
      将均值作为选定动作 $\boldsymbol{a}$
      将标准差用作模型不确定性
    执行动作 $\boldsymbol{a}$
    如果 任务已完成 **则**
      重置任务并读取初始状态 $\boldsymbol{s}_0$
      初始化 过程 POLICY UPDATE
      返回 初始状态 $\boldsymbol{s}_0$
过程 POLICY UPDATE
  初始化 训练轮数 n_{max}，批量大小 N
  当 True **时**
    加载人工数据集 $\mathcal{D}$
    如果 $\mathcal{D}$ 中有新样本**则**
      当 $n < n$ **时**$_{max}$ **do**
        从以下数据中采样 N 个状态-动作对（$\boldsymbol{s}$，$\boldsymbol{a}$）：专家数据集 $\mathcal{D}$
        从策略 $\pi_\theta(\boldsymbol{s})$ 中采样动作 $\hat{}\boldsymbol{a}$
        计算损失 l
        递增训练轮数 n
        对 π_θ执行梯度下降更新

动作由智能体执行，并与当前状态一同存储在数据集 $\mathcal{D}$ 中。该操纵杆触发器用于在人类和当前策略之间切换控制权。如果人类未按下该操纵杆触发器，则根据当前状态 $\boldsymbol{s}$，从当前策略 π_θ 中采样多个动作 $\boldsymbol{a}$，并由智能体执行其均值。动作的标准差用作模型不确定性的代理指标，以通过人工干预指导数据收集。

在辅助策略更新 流程中，初始化最大训练轮数 n_{max} 和批量大小 N。当人类用户提供示范数据或干预数据时，该流程会更新策略。流程加载当前数据集 $\mathcal{D}$，并根据样本数量检查是否有新增数据。如果发现新样本，此循环会从 $\mathcal{D}$中采样一个包含 $N = 64$ 个状态-动作对的批次。给定这些状态$\boldsymbol{s}$，从当前策略 π_θ中采样对应的动作 $\hat{}\boldsymbol{a}$，并计算均方误差损失 l：

$$l = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{a}}_i - \boldsymbol{a}_i)^2. \tag{4}$$

如果优化轮次低于最大值 $n_{max} = 30$，则使用常用的均方根传播 (RMSProp) 优化器，以 1e-4 的学习率优化策略 π_θ。

4. 硬件实验结果

本节介绍用于在硬件平台上执行所提框架的实验装置，并展示实验结果。

4.1 实验装置

人类操作员通过 Logitech 游戏手柄控制器，向机器学习模型提供示范和干预，输入机体级（前向和侧向）速度、偏航角速度及高度指令。这些指令随后输入速度和高度控制器，再由该控制器向运行 ROSFlight 的飞行控制器发送姿态和推力指令Jackson 等人(2016年）。在此着陆任务中，我们使用 REEF PID 和 REEF 自适应控制器 Ramos 等人(2019年），以补偿多旋翼配平设置的细微变化。REEF 控制器是由佛罗里达大学研究与教育工程设施（REEF）开发并维护的开源飞行控制器。图 2 展示了整体实验系统的框图。有关飞行控制器的详细实现，请参见 GitHub 页面 [1]。

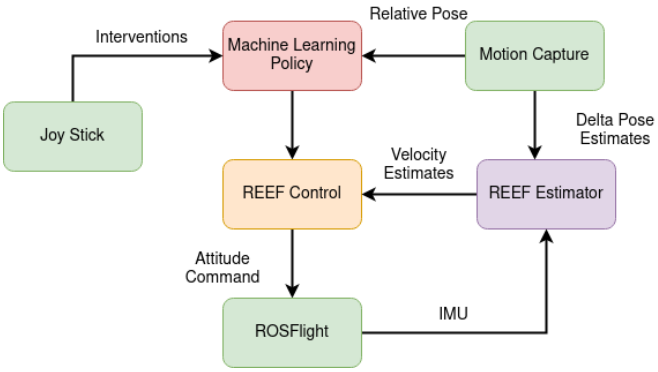


图 2. 展示用于演示、干预以及在四旋翼飞行器控制框架中部署已学习策略 π_θ 的整体系统框图。

神经网络的输入是四旋翼飞行器在相机坐标系中相对于目标的位置。该测量数据由动作捕捉系统获取，该系统会跟踪四旋翼飞行器和着陆垫的位置。使用Ganesh 等人(2021年)进行外参标定，从而可将动作捕捉测量结果作为机载测量的替代，例如使用 ArUco 标记。除相对位置外，神经网络还会接收四旋翼飞行器的航向和高度，分别由动作捕捉系统和 REEF 状态估计器获取。记录着陆示范后，网络便会使用图 2 所示的相同框架进行训练和部署（也称为展开运行）。

在每次展开运行中，系统会引导人类监督者根据当前模型不确定性的视听呈现（图 3）决定是否进行干预。

[1] REEF 自适应控制器：https://github.com/uf-reef-avl/reef_adaptive_control

当估计不确定性超过用户定义的阈值时，系统会逐步点亮相应数量的指示灯，以反映模型不确定性的程度。人类监督者可根据这些信号了解干预的紧迫程度。此外，当不确定性超过最大阈值时，蜂鸣器会发出提示音，表明模型不可靠，需要人类进行干预。所有不确定性阈值均根据演示飞行中获得的估计不确定性，通过启发式方法设定。该不确定性通过神经网络的变化性进行估计，并以蒙特卡洛丢弃所产生的标准差加以量化。人类监督者进行干预时，系统会记录每次干预，并将其用于微调智能体的行为。

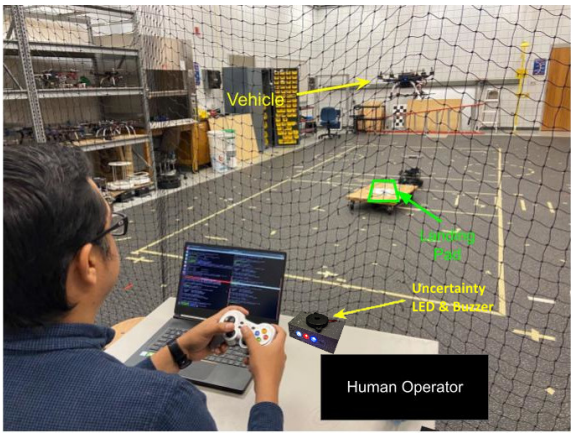


图3. 人类操作员介入，纠正四旋翼飞行器不理想的动作，以提升着陆任务的性能。图中还可以看到由地面机器人牵引的着陆垫。

为了证明所提出的框架能够轻松应对环境或动力学的变化，着陆任务先在固定着陆垫上执行，随后在移动着陆垫上执行。如图4所示，实验装置由一台地面机器人和一辆拖曳着陆垫的推车组成。着陆垫的线速度约为0.25米/秒。

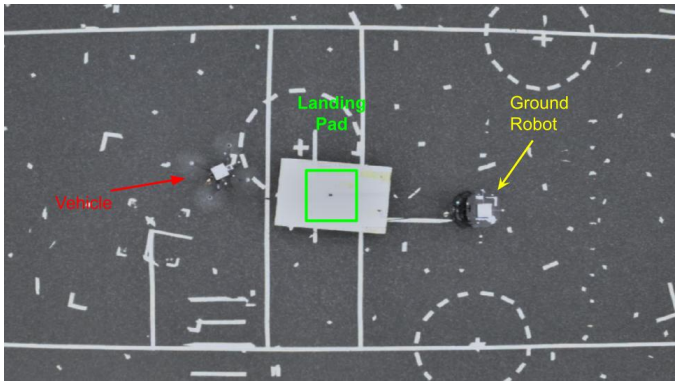


图4. 实验室空间中由上方摄像机拍摄的典型实验。图中显示四旋翼飞行器正在跟随由地面机器人牵引的着陆垫。使用 ArUco标记标示着陆垫中心。

4.2 硬件实验结果

硬件实验结果将在以下小节中介绍：演示与部署，使用人类示范训练网络，使其模仿在静态着陆垫上的着陆任务；不确定性感知干预，通过干预降低网络不确定性并微调其性能；以及在移动平台上着陆，通过干预使学习到的模型适应环境变化。在本例中，需要使学习到的着陆策略适应移动着陆垫。

示范与展开运行 为了执行着陆任务，网络使用了由10个人类示范组成的训练集。每次示范中，四旋翼飞行器都从不同的初始位置开始学习实验，以丰富训练数据。在示范——

示范数据、人类输入（即速度、高度和偏航角速度）以及网络输入均会被保存。随后，网络经过30个训练轮数的训练，并从学习集起飞位置以外的初始位置中均匀选取位置进行部署，以执行着陆任务。图5展示了四旋翼飞行器从6个不同起飞位置出发时的平均着陆位置。图中还显示了人类示范和经典控制器的平均着陆位置。从图中还可以看出，仅使用示范训练的网络能够模仿人类飞行员，并成功使四旋翼飞行器降落在着陆垫内，位置接近中心，但偏离了示范平均着陆点。

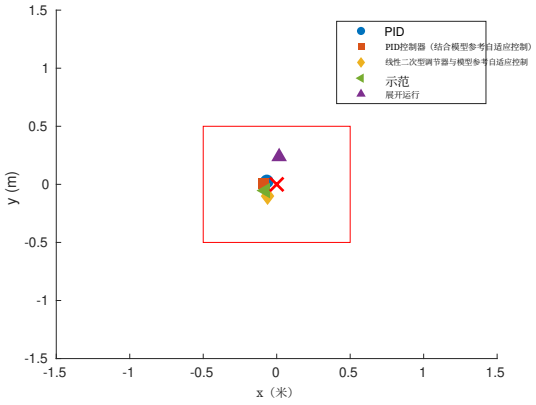


图5. 图中显示了四旋翼飞行器在不同控制策略下的平均着陆位置，并以方形着陆区域作为参照，该区域代表四旋翼无人机底座的尺寸（约为50厘米×50厘米）。着陆垫中心以十字标示。示范的平均着陆位置与经典控制器的结果相当，平均径向误差为12.73厘米。由于模型不确定性，初始展开运行的平均着陆位置相较于示范结果有所偏移。

仅使用示范训练时，网络生成的10条着陆轨迹如图6所示。着陆轨迹根据所获得的模型不确定性以不同深浅显示。数值越大，表示模型不确定性越高。具体而言，颜色条表示各控制轴（前向和

横向速度以及高度）上的方差之和。着陆任务从大致相同的位置开始时，虽然遵循相似的轨迹，但不确定性指标确实捕捉到了输出变异性。这一点在下降阶段尤为明显：由于地面效应以及示范中的最终修正指令，相对位置结果的变化更大。

不确定性感知干预 由于学习过程是实时进行的，因此也可以实时、定期并按需决定是否进行干预。为了判断是否需要干预，人类操作员将控制权交给智能体，并观察生成的飞行路径是否合适，以及网络不确定性是否可能导致不安全或不规则飞行。为估计网络不确定性，我们使用由蒙特卡洛丢弃产生的一组网络输出的标准差。具体而言，将当前输入通过网络进行 1000 次前向传播，每次随机丢弃一组网络神经元。这样，每个输入会产生 1000 个输出；我们将这 1000 个输出的样本标准差视为模型的估计不确定性。这项计算在线进行，以提示飞行员何时应进行干预。如前所述，当模型不确定性较高并超过上述阈值时，蜂鸣器和 LED 灯会向飞行员发出通知，如图 3 所示。请注意，如果控制器执行了意料之外的飞行动作，人类也可以进行干预，并通过手动遥控指令直接控制四旋翼力矩，从而覆盖原有控制输入。

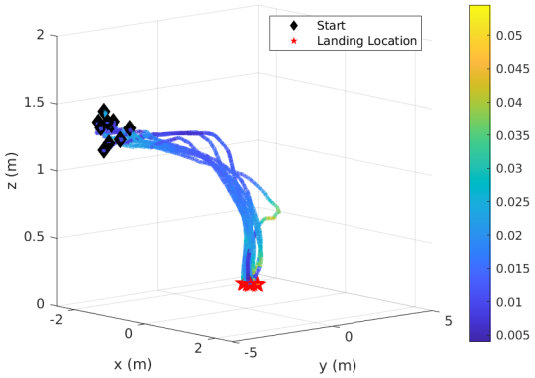


图6. 四旋翼飞行器从同一起始位置执行着陆任务时的轨迹图。轨迹根据使用10次示范训练的模型所估计的不确定性进行着色；该模型与前文介绍的、采用多个起飞位置的静止平台着陆模仿模型类似。

当模型从 4 个不同的起始位置展开运行时，系统会实施干预。每次展开运行中，飞行员根据代表模型不确定性的视听提示决定何时进行干预。图 7 展示了将不确定性感知干预纳入模型后，着陆任务的着陆轨迹。图 8 展示了这些展开运行的平均着陆位置。

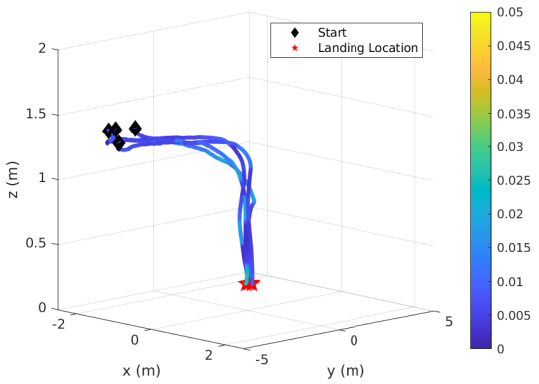


图7. 微调后四旋翼飞行器的轨迹

具有不确定性感知的人类干预的策略。此次干预改善了不确定性和最终着陆位置。

从图 7 和图 8 可以看出，干预后四旋翼飞行器的轨迹变化较小，整体上也能更好地复现示范轨迹。表 1 使用平均着陆位置误差以及笛卡尔方向和径向方向上的偏差，定量呈现了干预后着陆任务的改进情况。干预训练策略的着陆精度与经典控制器和示范相当，但精确度更高。值得注意的是，模型不确定性也得到了改善：与仅使用示范进行训练的表现相比，模型不确定性降低了 5 倍。此外，请注意，唯一新增的信息来自视听设备发出不确定性超过阈值信号时所实施的干预。

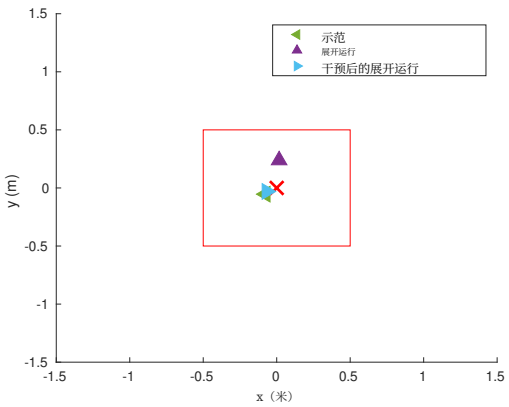


图 8. 学得的控制策略下四旋翼飞行器的平均着陆位置，突出显示干预的效果。干预改善了策略的展开运行，使平均着陆位置与初始示范中的着陆位置相仿。

在移动平台上着陆所提出的框架允许引入人工干预，使模型行为能够适应相关但显著不同的任务。为展示这一概念，我们通过从人工干预中学习，将用于在静止目标上着陆的模型调整为在移动目标上着陆。

表 1. 每种控制策略下着陆位置的均值和偏差。

PID	x（厘米）	y（厘米）	r（厘米）
μ	-6.73	2.57	17.99
σ	15.35	11.39	6.27
MRAC	x（厘米）	y（厘米）	r（厘米）
μ	-9.03	0.73	11.74
σ	10.85	5.34	8.92
LQR	x（厘米）	y（厘米）	r（厘米）
μ	-6.23	-10.08	15.08
σ	5.91	12.00	8.64
演示。	x（厘米）	y（厘米）	r（厘米）
μ	-7.61	-5.34	12.73
σ	5.53	8.48	4.30
展开运行	x（厘米）	y（厘米）	r（厘米）
μ	1.85	23.83	26.82
σ	3.51	18.56	13.35
R.A.I.	x（厘米）	y（厘米）	r（厘米）
μ	-6.82	-3.02	5.26
σ	6.28	8.83	1.10

图 9 展示了在移动目标上着陆时，策略在干预前后的表现，以及经典控制器的表现。所有经典控制器以及仅在静止目标上训练的策略 π_θ ，都未能降落到动态平台上。

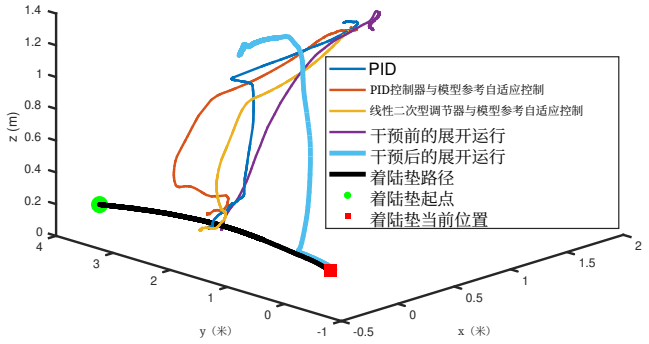


图 9. 多种控制策略下尝试在移动目标上着陆。只有经过干预后的网络策略能够降落到目标上。绿色圆圈表示着陆垫在轨迹起点处的移动方向，红色方块表示轨迹终点。

为调整模型行为，策略 π_θ 在经过三次人工干预后重新训练，以纳入移动着陆垫的信息。网络使用这三次额外干预训练了 10 轮。如图 9 中蓝色曲线所示，改进后的策略成功学会了动态着陆平台的新动力学特性，并能够在移动目标上着陆。

请注意，在移动平台上着陆只是任务／环境发生变化的一个例子。该框架同样可用于执行其他相对复杂的任务，例如绕障飞行、执行特定的着陆机动（如螺旋着陆），甚至无需严谨的模型即可穿窗飞行。

5. 结论

在这项工作中，我们提出了一种新颖的人在回路机器学习框架，该框架利用不确定性估计，对基于深度神经网络的控制器进行针对性训练，从而提升任务执行效果。我们提出的方法成功地将从人类示范中学习与不确定性感知干预相结合，教会一架小型无人飞行器在固定着陆垫上着陆；此外，通过利用不确定性感知干预增强已学习的任务，还教会它在移动着陆平台上着陆。我们展示了硬件验证实验，将不确定性感知的人在回路机器学习技术与经典控制器进行了比较。结果表明，在处理动态任务时，除基于人类示范的学习外，不确定性感知干预也优于经典方法。通过人工干预实现的深度神经网络适应能力，有望使机器人能够适应变化环境，并在线学习与人类目标一致的新任务。

参考文献

Anderson, P., Wu, Q., Teney,D., Bruce, J., Johnson,M.,S¨underhauf, N., Reid, I., Gould, S. 和 van den Hengel, A.（2018）。视觉与语言导航：在真实环境中解读以视觉为基础的导航指令。载于 *IEEE*计算机视觉与模式识别会议（*CVPR*）。Argall, B.D., Chernova, S., Veloso, M. 和 Browning, B.（2009）。机器人示范学习综述。《机器人与自主系统》，57(5)，469–483。Blukis, V., Brukhim, N., Bennett, A., Knepper, R.A. 和 Artzi, Y.（2018）。通过模仿学习在模拟四旋翼无人机上遵循高级导航指令。载于《机器人：科学与系统会议论文集》。Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J. 和 Zieba, K.（2016）。端到端自动驾驶汽车学习。1–9。URLhttp://arxiv.org/abs/1604.07316。Gal, Y. 和 Ghahramani, Z.（2016）。将Dropout作为贝叶斯近似：在深度学习中表示模型不确定性。载于 M.F. Balcan 和 K.Q. Weinberger（编），《第*33*届国际机器学习会议论文集》，第48卷，收录于《机器学习研究论文集》，1050–1059。PMLR，纽约，纽约州，美国。URL https://proceedings.mlr.press/v48/gal16.html。Gandhi, D., Pinto, L. 和 Gupta, A.（2017）。在坠机中学习飞行。载于 *2017*年*IEEE/RSJ*国际智能机器人与系统会议（*IROS*），3948–3955。IEEE。Ganesh, P., Volle, K., Buzaud, P., Brink, K. 和 Willis, A.（2021）。相机与动作捕捉系统的外参标定。载于 *2021*年东南会议，01–08。IEEE。Giusti, A., Guzzi, J., Cire, D.C., He, F.l., Rodr´ıguez, J.P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Caro, G.D., Scaramuzza, D. 和 Gambardella, L.M.（2015）。一种用于移动机器人森林小径视觉感知的机器学习方法。3766(c)，1–7。doi:10.1109/LRA.2015.2509024。

Goecks, V.G., Gremillion, G.M., Lehman, H.C. 和 Noth- wang, W.D.（2018）。一种基于任务示范、用于学习人类意图并塑造机器人行为的网络-人类协作方法。载于 *2018*年国际联合神经网络会议（*IJCNN*），1–7。doi:10.1109/ IJCNN.2018.8489595。Goecks, V.G., Gre million, G.M., Lawhern, V.J., Valasek, J. 和 Waytowich, N.R.（2019a）。在实时环境中高效结合人类示范与干预，以实现自主系统的安全训练。载于*AAAI*人工智能会议论文集，第33卷，2462–2470。Goecks, V.G., Gremillion, G.M., Lawhern, V.J., Valasek, J. 和 Waytowich, N.R.（2019b）。结合行为克隆与强化学习，以提升稠密和稀疏奖励环境中的性能。*arXiv*预印本 *arXiv:1910.04281*。Gu, S., Holly, E., Lillicrap, T. 和 Levine, S.（2017）。通过异步离策略更新实现机器人操作的深度强化学习。载于 *2017*年*IEEE*国际机器人与自动化会议（*ICRA*），3389–3396。doi:10.1109/ICRA.2017.7989385。Hilleli, B. 和 El-Yaniv, R.（2018）。迈向无需模拟器的深度强化学习：一个自主转向示例。*AAAI*人工智能会议，1471–1478。
U
R
L
https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16929。Jackson, J., Ellingson, G. 和 McLain, T.（2016）。Rosflight：一种轻量、低成本的微型无人机研发工具。载于*2016*年国际无人机系统会议（*ICUAS*），758–762。IEEE。Kelly, M., Sidrane, C., Driggs-Campbell, K. 和 Kochen- derfer, M.J.（2019）。HG-DAGGER：与人类专家开展交互式模仿学习。载于*2019*年国际机器人与自动化会议（*ICRA*），8077– 8083。doi:10.1109/ICRA.2019.8793698。Mei, H., Bansal, M. 和 Walter, M.（2016）。倾听、关注并行走：将导航指令神经映射为动作序列。载于 *AAAI*人工智能会议。
U
R
L
h
t
t
p
s://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12522。Misra, D., Langford, J. 和 Artzi, Y.（2017）。利用强化学习将指令和视觉观测映射为动作。载于*2017*年自然语言处理实证方法会议论文集，1004–1015。计算语言学协会。doi:10.18653/v1/D17-1106。URLhttps://www.aclweb.org/anthology/D17-1106。Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W. 和 Abbeel, P.（2018）。利用示范克服强化学习中的探索问题。载于 *2018*年*IEEE*国际机器人与自动化会议（*ICRA*），6292–6299。doi:10.1109/ICRA.2018.8463162。Ng, A.Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E. 和 Liang, E.（2006）。通过强化学习实现自主倒飞直升机飞行。载于《实验机器人学*IX*》，363–372。施普林格。OpenAI, Andrychowicz, M., Baker, B., Chociej, M., J´ozefowicz, R., McGrew, B., Pachocki, J.W., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., Schnei- der, J., Sidor, S., Tobin, J., Welinder, P., Weng, L. 和 Zaremba, W.（2018）。灵巧手内操作学习。*CoRR*，abs/1808.00177。URLhttp://arxiv.org/abs/1808.00177。Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P. 和 Peters, J.（2018）。从算法视角看模仿学习。*ArXiv*，abs/1811.06711。Ramos, J.H., Ganesh, P., Warke, W., Volle, K. 和 Brink, K.（2019）。REEF 状态估计器：一种简化的多旋翼飞行器开源估计器和控制器。载于*2019 IEEE*全国航空航天与电子会议（*NAECON*），606–613。doi:10.1109/NAECON46414.2019.9058099。Ross, S., Melik-Barkhudarov, N., Shankar, K.S., Wendel, A., Dey, D., Bagnell, J.A. 和 Hebert, M.（2013）。在杂乱的自然环境中学习单目反应式无人机控制。载于*2013 IEEE*国际机器人与自动化会议，1765–1772。doi:10.1109/ICRA.2013.6630809。Saunders, W., Sastry, G., Stuhlm¨uller, A. 和 Evans, O.（2017）。无差错试验：通过人工干预实现安全强化学习。*CoRR*，abs/1707.05173。网址http://arxiv.org/abs/1707.05173。Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. 和 Salakhutdinov, R.（2014）。Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》，15(1)，1929–1958。Tai, L., Paolo, G. 和 Liu, M.（2017）。虚实迁移深度强化学习：用于无地图导航的移动机器人连续控制。载于*2017*年 *IEEE/RSJ*国际智能机器人与系统会议（*IROS*），31–36。doi:10.1109/IROS.2017.8202134。Thomas, G., Chien, M., Tamar, A., Ojea, J.A. 和 Abbeel, P.（2018）。从 CAD 学习机器人装配。载于*2018*年 *IEEE*国际机器人与自动化会议（*ICRA*），1–9。doi:10.1109/ICRA.2018. 8460696。Tobin, J., Biewald, L., Duan, R., Andrychowicz, M., Handa, A., Kumar, V., McGrew, B., Ray, A., Schnei-der, J., Welinder, P., Zaremba, W. 和 Abbeel, P.（2018）。用于机器人抓取的域随机化与生成模型。载于*2018*年 *IEEE/RSJ*国际智能机器人与系统会议（*IROS*），3482–3489。doi:10.1109/IROS.2018.8593933。Warnell, G., Waytowich, N., Lawhern, V. 和 Stone, P.（2018）。Deep TAMER：在高维状态空间中进行交互式智能体塑造。*AAAI*人工智能会议，1545–1553。网址https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16200。Waytowich, N.R., Goecks, V.G. 和 Lawhern, V.J.（2018）。通过人机交互实现自主系统的循环学习。*CoRR*，abs/1808.09572v1。网址https://arxiv.org/abs/1808.09572v1。Wu, J., Huang, Z., Hu, Z. 和 Lv, C.（2022）。迈向人机协同人工智能：通过实时人工引导增强自动驾驶的深度强化学习。《*Engineering*》。doi:https://doi.org/10.1016/j.eng. 2022.05.017。网址
h
t
t
p
s://
www.sciencedirect.com/science/article/pii/S2095809922004878。Zhang, T., Kahn, G., Levine, S. 和 Abbeel, P.（2016）。利用模型预测控制引导的策略搜索学习自主飞行器的深度控制策略。载于*2016*年 *IEEE*国际机器人与自动化会议（*ICRA*），528–535。IEEE。