arXiv:2301.05980v1 [cs.RO] 14 Jan 2023

基于深度强化学习的路径规划：利用距离传感器作为观测的工业机器人路径规划

Teham Bhuiyan[1]*、Linh Kästner[1],Yifan Hu[1],Benno Kutschank[1] 和 Jens Lambrecht[1]

摘要——工业机器人凭借其高效执行装配或焊接等重复性任务的能力，广泛应用于各种制造环境。这些应用中一个常见的问题是，如何在不与障碍物或其他机械臂发生碰撞的情况下到达目的地。最常用的基于采样的路径规划方法（如 RRT）通常需要较长的计算时间，尤其是在复杂环境中。此外，使用这些方法时必须预先了解环境。在新环境中采用这些方法时，需要投入大量工程精力来设置超参数，耗时又费钱。另一方面，深度强化学习在应对未知环境、泛化至新的问题实例以及高效解决运动规划问题方面取得了显著成果。因此，本文提出一种基于深度强化学习的机器人机械臂运动规划器。我们通过多项实验，将模型与最先进的基于采样的规划器进行了比较。结果表明，我们的规划器在路径长度和执行时间方面表现更优。

I. 引言

工业机器人广泛应用于制造流程中。尽管机器人通常能够自动执行特定任务，但仍需对其进行手动编程。基于采样的规划方法虽然能够很好地应对简单环境，但由于计算复杂度高且动态变化难以预料，在复杂且未知的场景中实现可靠的运动规划仍是一项重大挑战。尤其是在每个新环境中，工业机器人都需要借助离线编程等方法重新编程。这些方法要求熟练工人手动设置各种超参数，既费时又费钱，也需要投入繁重的工程工作。此外，RRT 等经典方法只能用于已知环境。另一方面，深度强化学习（DRL）作为一种端到端学习方法应运而生，能够直接将传感器数据映射为机器人动作，并在习得复杂行为规则、提高对噪声的鲁棒性以及泛化到新的问题实例方面展现出良好前景 [1], [2], [3]。本文提出一种基于深度强化学习的方法，用于自动化机械臂的运动规划。具体而言，我们提出了一套训练流程，其中包含虚拟激光观测，以提高训练效率；并在逼真的三维仿真中，针对多种面向工业的任务对该方法进行评估。

本研究的主要贡献如下：

[1] Linh Kästner、Yifan Hu、Benno Kutschank 和 Jens Lambrecht 任职于德国柏林柏林工业大学电气工程与计算机科学学院工业级网络与云计算教席。teham@tu-berlin.de

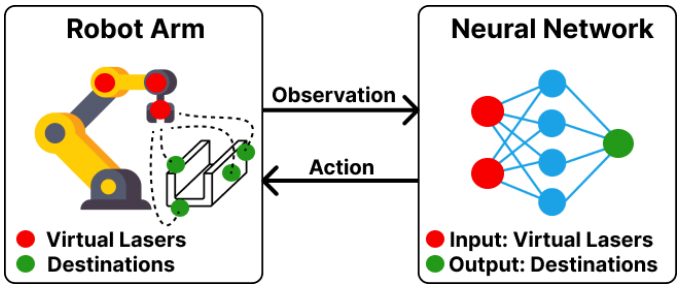


图 1：本文提出了一种基于深度强化学习的运动规划方法，并使用虚拟激光雷达作为观测。给定目标和虚拟传感器观测后，训练好的智能体能够生成一条通往目标的无碰撞路径。

- 在 3D 模拟器 PyBullet 中实现真实的工业机器人任务
- 提出一种使用激光扫描作为观测的高效三维运动规划算法
- 在复杂场景中，将该方法与经典基线方法进行了广泛评估

本文结构安排如下。第 2 节介绍相关工作，随后在第 3 节阐述方法。接着，第 4 节展示结果与评估。最后，第 6 节给出结论与展望。

II. 相关工作

机器人机械臂的无碰撞路径规划算法已在各类研究文献中得到广泛研究。人工势场和基于采样的算法等传统方法，是目前工业机器人机械臂中最常用的方法之一 [4] [5][6]。近期的研究包括 Xinyu 等人的工作 [7]。研究人员提出了一种基于 RRT 和人工势场的随机化算法 P-RT*-connect。该算法通过使用 RRT* 从起始节点和目标节点分别扩展两棵路径树来寻找运动路径。Wang 等人还提出了针对 RRT 规划器的一种近期优化方法[8]。该方法是对双向 RRT 算法（Bi-RRT，也称为 RRT-ConCon）的改进版本，通过增加一棵从目标构型开始扩展的 RRT 树来改进标准 RRT 算法，并以相同速度扩展两棵树。Salehian 等人 [9] 提出了一种统一的协调多臂运动规划框架，在自碰撞规避约束下实时提供集中式逆运动学求解器。Zhang 等人[10] 提出了一种适用于动态环境中机器人机械臂的实时运动学控制策略。尽管基于采样的方法能够实现无碰撞规划，但它们

仍然速度较慢、效率低下，在复杂环境中尤其如此。大多数基于采样的算法都假设已掌握环境的完整信息，但在某些场景中未必能够获得这些信息。深度强化学习已成为一种端到端方法，有望在未知环境中学习复杂行为。在机器人导航领域，基于深度强化学习的方法已展现出良好的应用前景 [11]⁊[12], [1], [13]。Wen 等人提出使用深度确定性策略梯度（DDPG）规划机械臂轨迹，仅依靠深度强化学习实现避障[14] 。Bianca Sangiovanni 等人 [15]提出一种混合控制方法，以实现仿人机器人操作臂的全身碰撞规避。他们将经典运动规划算法与深度强化学习（DRL）相结合，在操作空间中执行避障并完成到达任务。Faust 等人 [16] 和 Chiang 等人 [2]的类似研究，将基于深度强化学习的运动规划与 RRT、概率路线图等经典方法相结合，用于地面机器人长距离运动规划。近来，Kästner 等人 [11]、Dugas 等人 [17], 和 Guldenring 等人 [18] 的研究表明，深度强化学习方法在未知动态环境中快速避障方面表现出优势。多项研究也将基于深度强化学习的方法用于固定机器人的运动规划和碰撞规避。Kamali 和 Bonev [19] 提出一种基于深度强化学习的方法，用于解决远程操控应用中的机械臂运动规划问题。他们将人手运动实时映射到机械臂上，同时避免碰撞。该方法的缺点是障碍物必须是静态且预先定义的。Zhong 等人 [20] 提出了一种基于深度强化学习的焊接机械臂路径规划器。他们通过三个特定焊接任务证明，基于深度强化学习的模型能够取得与基于采样的方法相当的结果。Meyesa 等人 [21] 提出了一种基于强化学习和 Q学习的方法，使智能体能够控制六轴工业机器人使用摄像机玩套环游戏。Zhu 等人 [22] 提出了一种新颖的碰撞规避框架，该框架使用深度摄像机，使机器人能够在非结构化的复杂环境中与人类操作员协同工作。值得注意的是，上述研究均依赖传感器感知环境并据此采取行动。然而，对于焊接等工业任务，使用静态传感器并非最优选择，因为机器人可能会移出传感器的探测范围，或者传感器可能无法检测狭窄空间内的区域。因此，我们提出一种方案，将虚拟激光传感器策略性地布置在机械臂的不同部位。这样，传感器便可随机械臂一同移动，使机器人能够在复杂环境中执行无碰撞运动规划。

III. 方法

本章介绍所提出框架的方法。系统设计如图2所示。我们训练一个深度强化学习智能体，仅使用虚拟激光扫描观测来为工业机器人进行无碰撞运动规划

。训练和测试均在 Bullet3 [23] 仿真器中进行。

A. 机器人运动学

机器人控制算法依靠运动学来计算精确的指令。对于由刚性连杆和关节组成的机器人，常见的建模方法是 Denavit-Hartenberg（D-H）表示法。该方法使用齐次变换矩阵，这些矩阵是机器人构型 $\mathbf{q} = [q_1, q2, \ldots, q_n]$ 的函数，用于描述两个相邻连杆的坐标系之间的关系。将给定机器人的这些矩阵依次相乘，可得到从机器人基座连杆到末端执行器的变换，而末端执行器在大多数应用中就是控制目标。对于六自由度机器人，从基座到末端执行器的变换为

$$ {}^{0}_{\mathrm{E}}\mathbf{T}(\mathbf{q}) = (\prod_{i=1}^{6} {}^{i-1}_{i}\mathbf{T}(q_i)) \cdot {}^{6}_{\mathrm{E}}\mathbf{T}, \tag{1} $$

由单个矩阵 ${}^{i-1}_{i}\mathbf{T}(q_i)$ 定义，其表达式为

$$ {}^{i-1}_{i}\mathbf{T}(q_i) = \begin{bmatrix} \cos(\theta_i) & \sin(\theta_i) & 0 & 0 \\ -\sin(\theta_i)\cos(\alpha_i) & \cos(\theta_i)\cos(\alpha_i) & \sin(\alpha_i) & 0 \\ \sin(\theta_i)\sin(\alpha_i) & \sin(\alpha_i) & \cos(\alpha_i) & 0 \\ a_i\cos(\theta_i) & a_i\sin(\theta_i) & d_i & 1 \end{bmatrix}^{t}. \tag{2} $$

方程（1）可理解为机器人的正运动学：对于给定构型 $\mathbf{q}$，变换 ${}^{0}\mathbf{ET(q)}$ 给出了机器人基座坐标系中末端执行器的姿态（左上角的 3×3 子矩阵）和位置（最右列的前三个元素）。然而，机器人领域的大多数任务是在工作空间中定义的，而非构型空间中。因此需要逆运动学，将工作空间中定义的轨迹映射到构型空间，随后用于生成控制输入。一般而言，在逆运动学中，期望的工作空间位移 $\dot{\mathbf{x}}$ 与一个初始未知的构型空间位移 $\dot{\mathbf{q}}$相关，其关系为

$$ \dot{\mathbf{x}} = \mathbf{J}(\mathbf{q}) \cdot \dot{\mathbf{q}}, \tag{3} $$

其中 $\mathbf{J(q)}$ 是机器人的雅可比矩阵。由于精确解通常不切实际，人们提出了多种数值方法来求解该方程中的 $\dot{\mathbf{q}}$ 。本研究采用的方法是 Wampler[24] 提出的阻尼最小二乘法。该方法通过以下方式近似求解方程 (3)：

$$ \dot{\mathbf{q}} = \mathbf{J}(\mathbf{q})^T \cdot (\mathbf{J}(\mathbf{q}) \cdot \mathbf{J}(\mathbf{q})^T + \lambda^2 \cdot \mathbf{I})^{-1} \cdot \dot{\mathbf{x}}, \tag{4} $$

其中， λ 是阻尼因子，用于在机器人移动到接近奇异位形的构型时降低 $\dot{\mathbf{q}}$ 。这样即可采用算法1所示的标准逆运动学方法。

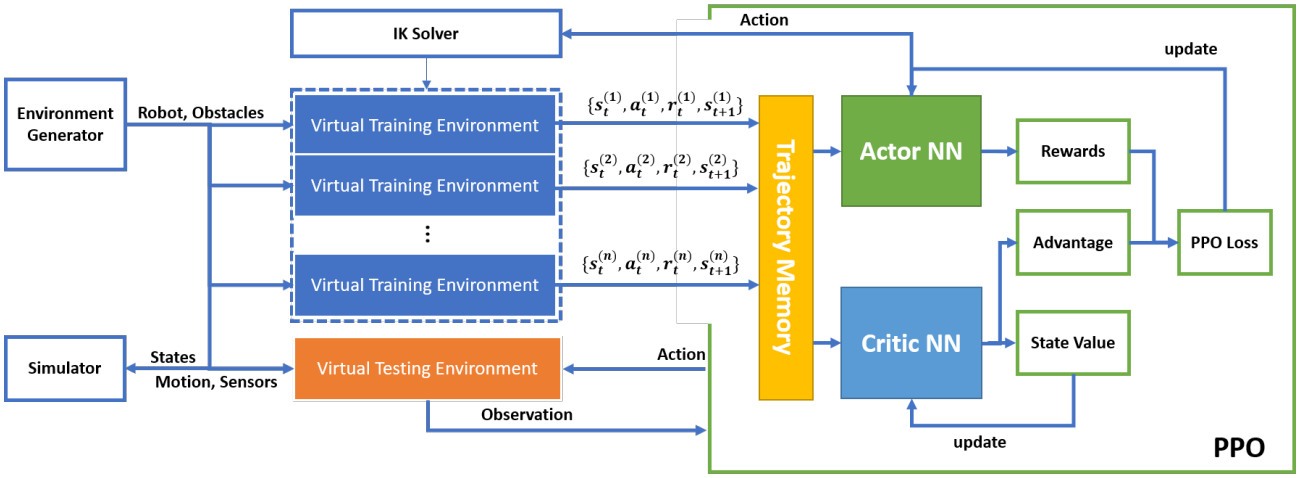


图2：我们提出的方法的系统设计。对于观测，我们使用三个虚拟激光传感器（两个位于末端执行器上，一个位于腕部上方，如图3所示）。我们采用演员-评论家方法，由智能体输出连续值，以预测机器人当前位置姿态的变化。机器人的运动通过逆运动学（IK）求解器实现。该过程会重复进行，直到机器人到达目标。

算法1： 阻尼最小二乘逆运动学

运动学

输入：期望的工作空间位姿$\mathbf{x}_d$，积分步长 Δ_{int}，
微分步长 Δ_{dif}，精度阈值 ε

输出：期望的构型空间位姿$\mathbf{q}_d$

$\mathbf{q}_{\text{it}} \leftarrow \mathbf{q}_c$

$\mathbf{x}_c \leftarrow$ 通过正运动学从当前工作空间位姿从 $\mathbf{q}_{\text{it}}$

$\mathbf{x}_e \leftarrow \mathbf{x}_c - \mathbf{x}_d$

做

$\dot{\mathbf{x}}_e \leftarrow \frac{\mathbf{x}_e}{\Delta_{\text{dif}}}$

$\dot{\mathbf{q}} \leftarrow \mathbf{J}(\mathbf{q}_{\text{it}})^T \cdot (\mathbf{J}(\mathbf{q}_{\text{it}}) \cdot \mathbf{J}(\mathbf{q}_{\text{it}})^T + \lambda^2 \cdot \mathbf{I})^{-1} \cdot \dot{\mathbf{x}}_e$

$\mathbf{q}_{\text{it}} \leftarrow \mathbf{q}_{\text{it}} + \dot{\mathbf{q}} \cdot \Delta_{\text{int}}$

$\mathbf{x}_c \leftarrow$ 通过正运动学求当前工作空间位姿运动学，来自$\mathbf{q}_{\text{it}}$

$\mathbf{x}_e \leftarrow \mathbf{x}_c - \mathbf{x}_d$

当 $||\mathbf{x}_e|| > \varepsilon$

$\mathbf{q}_d \leftarrow \mathbf{q}_{\text{it}}$

返回 $\mathbf{q}_d$

B. 观测空间

观测空间由O_{robot} 和O^{rays}组成。O_{robot} 是一个由以下内容构成的13维向量：

- 6当前关节角，单位为弧度（适用于 6 自由度机械臂）。
- 3 标量值，表示当前笛卡尔坐标位置与目标位置之间的差值
- 3 描述当前末端执行器欧拉角位姿的标量值
- 1 表示当前位置到目标位置距离的标量值

虚拟传感器放置在末端执行器的尖端、末端执行器的外表面以及腕部顶部。图3展示了两种不同机器人模型的射线排列。虚拟激光器提供的观测O^{rays} 是一个129维向量，由这129条射线的命中比例构成。根据具体的末端执行器或机器人模型，可能需要重新配置射线的确切数量和角度。环境提供给智能体的观测是O_{robot}和O^{rays}的拼接。

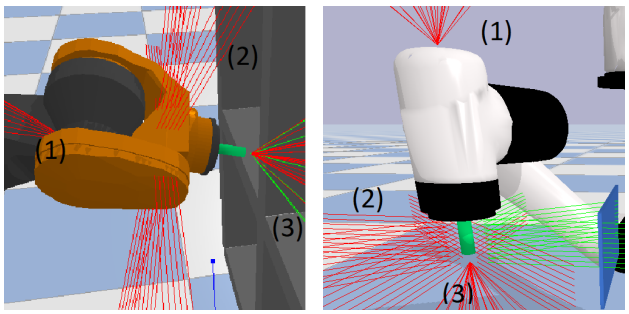


图3：两种不同机器人模型上的虚拟激光器（左：Kuka，右：UR5）。(1)–(3)中的数字表示各传感器发出的射线。绿色射线表示是否检测到障碍物。每个传感器的探测范围为 0.4 m

对于这两种机器人模型，传感器的布置如下：

- 腕部传感器由 24 条射线组成。每条射线在垂直方向上以 $\phi = 30°$ 的角间距排列，极角为 $\vartheta = 40°$。
- 末端执行器的环绕表面上共布置了 80 条射线。每 10 条射线在水平方向上以 $\phi = 45°$ 的间距排列。
- 25 射线布置在末端执行器的尖端，极角为 $\vartheta = 20°$，并沿水平方向间隔

$\phi = 30°$。

1) 奖励系统：奖励系统需要经过精心设计，因为它对智能体的行为起着至关重要的作用。我们的奖励函数 $r(s_t,a_t)$ 定义如下：

$$r(s_t,a_t) = r^t_{suc} + r^t_c + r^t_d + r^t_{shake} + r^t_o + r^t_e \tag{5}$$

r^t_{suc} 是成功奖励。当d_t（到目标的距离）小于距离阈值d_{Th}时，视为成功：

$$r^t_{suc} = \begin{cases} 10, & d_t < d_{Th} \\ 0, & otherwise. \end{cases} \tag{6}$$

r^t_c 是碰撞惩罚：

$$r^t_c = \begin{cases} 0, & otherwise \\ -10, & if\ robot\ collides. \end{cases} \tag{7}$$

r^t_d 是距离惩罚：

$$r^t_d = -0.01 \cdot d_t \tag{8}$$

- r_{to} 是越界工作空间惩罚：

$$r^t_c = \begin{cases} 0, & otherwise \\ -10, & if\ x_t < x_{min}\ or\ x_t > x_{max} \\ & or\ y_t < y_{min}\ or\ y_t > y_{max} \\ & or\ z_t < z_min\ or\ z_t > z_{max}. \end{cases} \tag{9}$$

- r_{te} 是单个回合步数耗尽时的惩罚。它可以防止机器人通过避免碰撞而放弃空间探索：

$$r^t_e = \begin{cases} 0, & otherwise \\ -5, & step > step_{max}. \end{cases} \tag{10}$$

r^t_{shake} 是不平滑运动的惩罚，其中n_s表示机器人在过去 10 步中来回移动的次数。n_s 的计算方法见算法 2。

$$r^t_{shake} = -0.005 \cdot n_s \tag{11}$$

2) 神经网络架构与智能体设计：神经网络架构如图 4 所示。它由两个网络组成，一个用于价值函数，另一个用于策略函数。网络的输入包括来自三次虚拟激光雷达扫描的共 129 条射线，以及目标位置和机器人的当前位置姿态。我们采用连续动作状态形式，以提高动作的灵活性和平滑性 [16]。动作空间A 定义如下：

$$A = \{\Delta x, \Delta y, \Delta z, \Delta\theta_y, \Delta\theta_p, \Delta\theta_r\} \tag{12}$$

$$\Delta x\ \Delta y\ \Delta z \in [-0.005, 0.005]\ m \tag{13}$$

$$\Delta\theta_y, \Delta\theta_p, \Delta\theta_r \in [-0.005, 0.005]\ rad \tag{14}$$

算法 2： 过去不均匀运动的次数 10 步

输入： 用于记录过去距离的队列Q_d，列表 L_m 用于历史方向记录
输出： 不均匀运动次数n_s

```
n_s=0 ;
如果 len(Q_d) > 10 则
    Q_d.从左侧弹出0
end
问_d.追加(d_t);
对于 len(Q_d) − 1，执行
    L_m.append(0) 如果 Q_d[i+1] − Q_d[i] ≥ 0 否则
        L_m.append(1)
end
对于 len(L_m) − 1，执行
    如果 L_m[j] ≠ L_m[j+1]，则
        n_s += 1
    end
end
返回 n_s
```

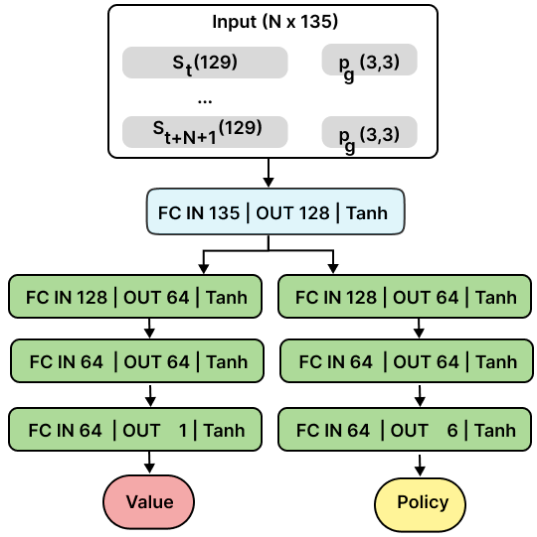


图 4：我们将 N 个智能体的激光雷达扫描观测作为输入传入网络。对于演员网络和评论家网络，我们使用相同的头部。

3) 训练算法：我们使用第3节所述的 PPO [25] 算法训练我们的深度强化学习智能体。PPO 内部采用优势演员-评论家结构，其中演员提供策略，评论家提供价值函数。每个网络都包含策略网络和价值网络。我们选择了一种模型，其中演员与评论家共享一个特定大小的特征提取器，用于从输入中提取特征，以节省计算时间。如图2和图4所示，通过同时模拟多个智能体和环境，实现并行训练。我们还通过简化激光数据并设置增强目标，进一步加快了训练。Beik 等人 [26] 提出了一种自适应策略，用于调整目标的可达性。训练一个“比实际更大”的目标可以简化训练过程。增强目标尺寸会根据训练成果进行调整——

算法 3： 近端策略优化

输入： 初始策略参数 π_0，初始价值函数参数 ϕ_0，裁剪阈值 ε

对 $N1$ **执行**

  重置环境并获取初始状态s_0

  对 $N2$ **执行**

    对于T **执行**

      通过运行策略 π_θ 选择动作a_t k

      在环境中执行a_t

      获取奖励r_t 和下一状态 s_{t+1}

      将状态转移 $(s_t,\ a_t,\ r_t,\ s_{t+1})$ 收集到轨迹集合 $\mathfrak{D}_k$

    end

    计算后续奖励 $\hat{R}_t$

      ˆ 基于当前$A^{\pi_{k_t}}$ 估计优势

    价值函数$V\phi_k$

  end

  最大化PPO-Clip目标：

$$\theta_{k+1} = arg\max_\theta \mathfrak{L}^{CLIP}_{\theta_k}(\theta)$$

  更新价值函数：

$$\phi_{k+1} = arg\min_\phi \frac{1}{\mathfrak{D}_k T}\sum_{\tau\in\mathfrak{D}_k}\sum_{t=0}^{T}(V\phi(s_t) - \hat{R}_t)^2$$

end

如果智能体持续无法到达目标，则增大目标大小；否则，减小目标大小。

$$\tilde{\rho}(e) = \begin{cases} \rho,\ e < e_\zeta \\ \tilde{\rho}(e-1) + \delta^+,\ \eta(k,e) < P_\zeta, e \geq e_\zeta, \tilde{\rho}(e) < \tilde{\rho}_{max} \\ \tilde{\rho}(e-1) - \delta^-,\ \eta(k,e) \geq P_\zeta, e \geq e_\zeta, \tilde{\rho}(e) > \tilde{\rho}_{min} \\ \rho,\ \eta(k,e) = 1, e \geq e_\zeta, \tilde{\rho}(e) = \rho \\ \tilde{\rho}(e-1),\quad otherwise \end{cases} \tag{15}$$

公式 15 给出了增强目标尺寸 $\tilde{\rho}(e)$ 的定义。这里，$\tilde{\rho}(e)$ 表示第 e 个回合时的增强目标尺寸，e_ζ 是尺寸保持不变的回合编号。目标尺寸的增量和减量分别由 δ^+ 和 δ^- 表示。成功阈值由 P_ζ 描述，而 $\eta(k,e)$ 表示在过去 k 个回合中，成功到达（增强）目标的次数。

$$\eta(k,e) = \frac{1}{k}\sum_{e'=e-k}^{e} S_{e'} \tag{16}$$

S_e 表示智能体在第 e 个回合结束时的成功/失败状态：

$$S_e = \begin{cases} 1, & if\ ||P_t - P_{target}|| \leq \tilde{\rho}(e) \\ 0, & otherwise \end{cases} \tag{17}$$

增强目标尺寸所用的参数见表 II。

4) 训练设置：智能体在随机化环境中进行训练。每个回合结束后，都会随机生成静态和动态障碍物。障碍物的数量会根据智能体的成功率增加或减少。训练过程通过使用图形处理器进行了优化，并在一个 NVIDIA RTX 3090 图形处理器。训练约耗时 6 小时后收敛。图 5 展示了两台机器人各自的训练过程

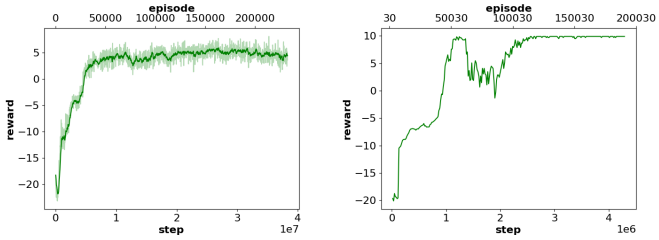


图 5：两台机器人的训练过程学习曲线（左：UR5，右：Kuka）。

模型上进行训练。超参数列于表 I 中。

在下一章中，我们将介绍实验、结果和评估内容。其中包括将两种基于采样的规划器——RRT 和一种名为 NC-RRT 的新型 RRT——与我们提出的深度强化学习规划器 [8] 进行比较。我们在两种不同的机器人（UR5 和 Kuka KR 16 ）上进行了实验，共创建了七种难度逐渐增加的测试场景。对于每种规划器，我们都在同一场景下进行了 30 次测试。图 6 展示了三组实验，其中包含不同类型的金属板障碍物。UR5 必须在移动至目标的过程中避开这些金属板。每组实验下方均展示了相应的轨迹。

UR5 实验旨在展示我们模型的一般路径规划能力，而 Kuka 实验则侧重于工业应用。在焊接、喷漆或装配等任务中，机械臂需要到达复杂且狭窄的环境中的特定位置。图 9 展示了该实验的仿真环境，其中包括机器人（Kuka KR 16）和一个用作障碍物的货架状结构。任务是依次穿过编号为 1 至 4 的每个空间。

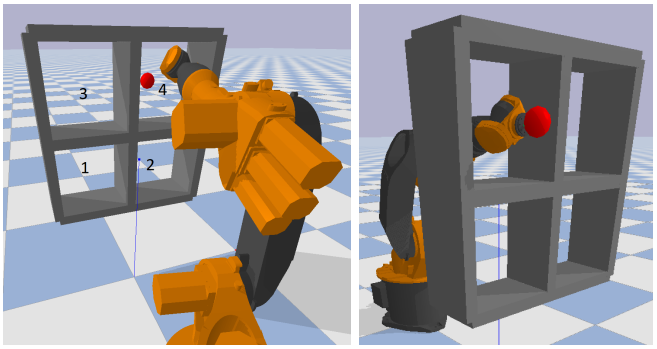


图 9：左侧和右侧的图示从不同角度展示了同一实验

图 7 展示了 Kuka 实验的定性结果。末端执行器的初始位置以蓝色球体标示，目标位置则以红色星形标示。粗线表示相应规划器的平均轨迹。

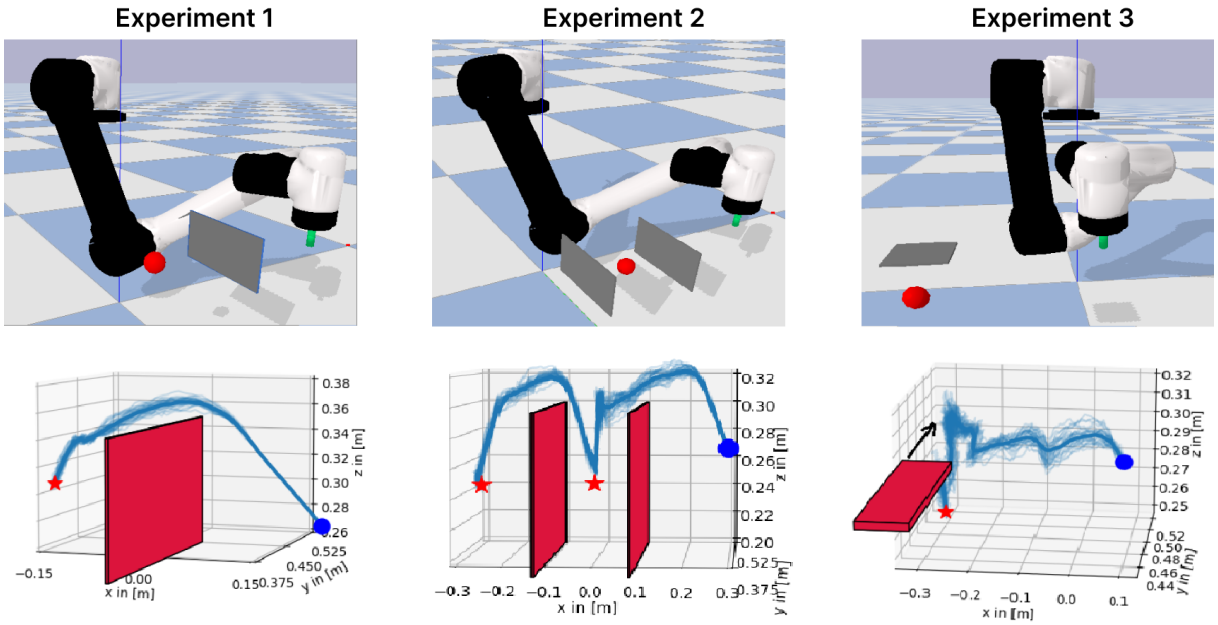


图6：初始位置和目标位置分别以蓝色圆圈和红色星形表示。粗线轨迹表示所有单次运行的平均路径。实验3包含一个移动障碍物（箭头表示其运动方向）。图中仅展示我们的深度强化学习规划器，因为基于采样的规划器变化较大，且其轨迹长得多、占用空间过大。

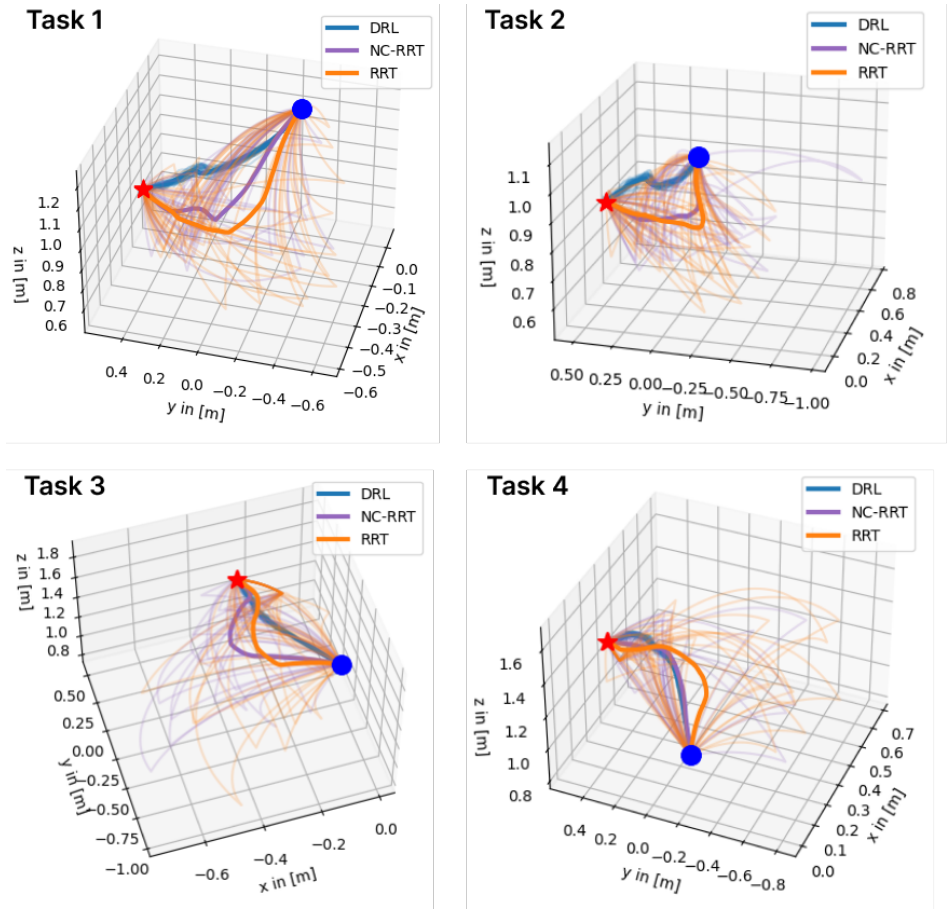
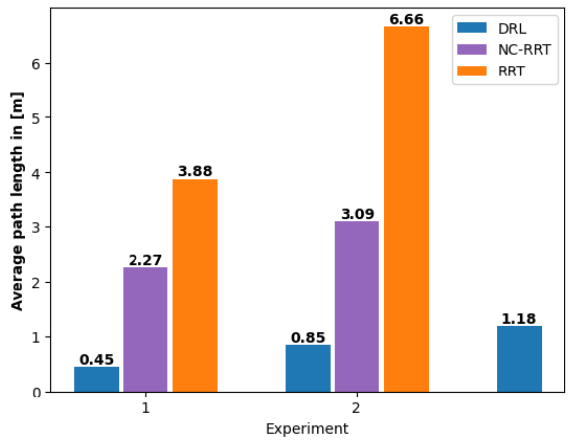
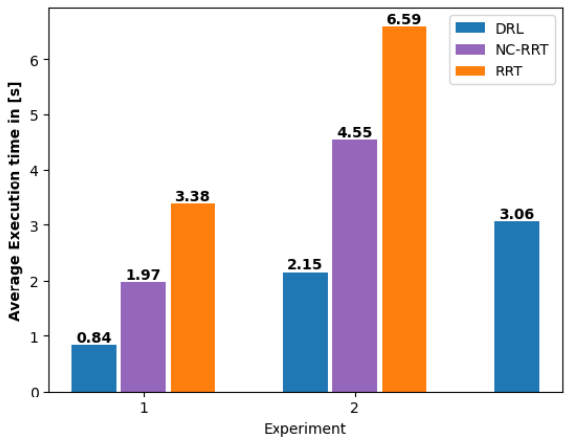


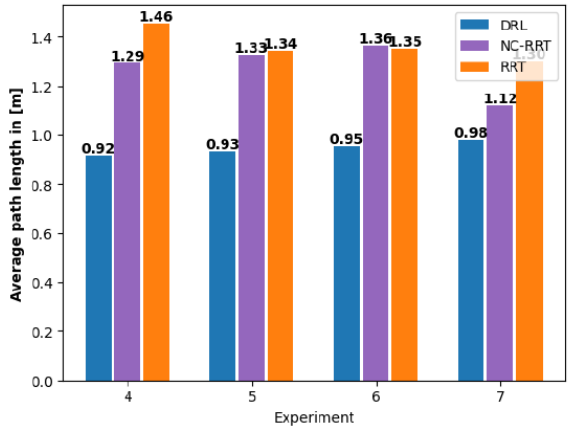
图7：Kuka实验的轨迹



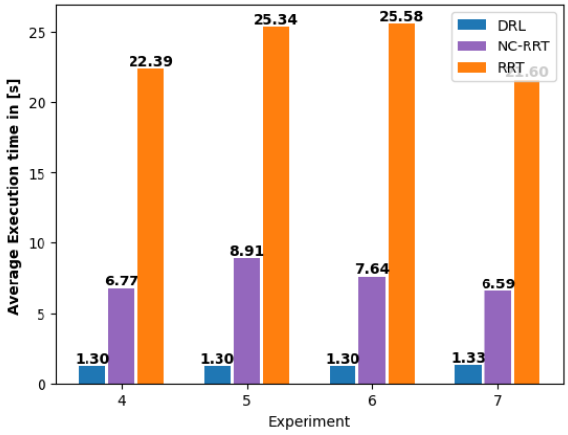
(a) Average path length for the UR5 task



(b) Average execution time for the UR5 task



(c) Average path length for the Kuka task



(d) Average execution time for the Kuka task

图8：所有实验的定量结果

IV. 讨论

A. UR5 轨迹

图6展示了深度强化学习规划器的三个仿真实验及各自的定性结果。未列出基于采样的规划器，因为其轨迹在每次运行中都存在明显偏差。由于这些轨迹混乱且缺乏规律，因此进行比较并无意义。实验1包含一个代表金属板的静态障碍物。轨迹显示，机器人末端执行器会抬高以避开板。在实验2中，增加了另一个目标和一块板，迫使机器人依次避开两个障碍物。实验3旨在测试我们模型的响应能力，其中包含一个移动障碍物（v=0.2m/s），该障碍物被设置为阻挡末端执行器到达目标位置。因此，智能体抬高末端执行器，使其靠近目标位置，并在障碍物移开后到达目标。在每个实验中，各次运行的轨迹都很相似，表明深度强化学习规划器具有很高的一致性。

B. Kuka 轨迹

图7展示了所有规划器在各项任务中的定性轨迹。可以明显看出，基于采样的规划器的轨迹在不同运行之间存在很大偏差。与此同时，我们的深度强化学习模型生成的轨迹似乎更加一致，因为各条独立轨迹与平均轨迹几乎无法区分。此外，基于采样的规划器生成的轨迹表明其路径长得多。

C. 定量评估

图8展示了三种规划器在所有实验中的定量评估结果。我们针对每种规划器，在7种不同的实验中总共进行了30次测试。我们通过计算行驶距离的平均值以及到达目标所需时间的平均值，来评估路径规划的效率。所有规划器均在未发生任何碰撞的情况下完成了实验。

1) UR5： 图8 (a) 展示了实验1-3的平均轨迹。与基于采样的规划器相比，深度强化学习模型生成的轨迹明显更短。轨迹更短意味着执行时间更快，如

图8 (b) 所示。需要注意的是，基于采样的规划器执行时间较长，不仅是因为轨迹较长，也因为计算耗时较高。在实验1中，我们的深度强化学习规划器完成任务的速度几乎是 NC-RRT 的两倍、RRT 的四倍。实验2结果相似：深度强化学习规划器的速度再次达到 NC-RRT 的两倍、RRT 的三倍。在实验3中，由于规划时间过长，经典方法无法应对动态障碍物，最终全部失败。因此，只有深度强化学习方法能够接受评估。

2) Kuka：货架障碍物是所有实验中最复杂的环境，因为其中包含多个狭窄隧道，增加了发生碰撞的可能性。与实验1和2类似，深度强化学习方法在所有实例中均优于RRT 规划器。图8（c）和（d）表明深度强化学习规划器具有一致性：实验4至7中的路径长度和时间基本相同，而RRT 规划器则表现出轻微差异。特别是，深度强化学习规划器的速度约为NC-RRT的6倍，最高可达RRT的19倍。

V. 结论

本文提出了一种基于深度强化学习的工业机器人运动规划模型。结果表明，在所有实验中，我们的模型均显著优于两种基于采样的规划器。在简单环境（实验1至2）中，我们的模型速度几乎是NC-RRT的两倍；而在更复杂的环境（实验4至7）中，NC-RRT的性能几乎落后六倍。由此可知，与基于采样的方法相比，我们提出的规划器对环境复杂性的敏感度更低。此外，借助激光传感器，我们的规划器能够对意外情况作出反应，例如物体因使用不当而被错误放置。在未来的工作中，我们计划通过加入路径规划之外的其他功能来扩展该模型。例如，在焊接任务中，机器人不仅需要移动到目标位置，还必须通过其末端执行器（焊枪）保持所需位姿。传统上，每个焊点都需要由熟练工人或第三方工具手动设计。我们将尝试开发一种用于路径规划和位姿估计的端到端模型。此外，我们还将加入其他类型的传感器（例如RGB相机），进一步改进该模型。在协作环境中，使用基于传感器的规划器至关重要，因为机器人必须能够在避开人类或其他机器人等动态障碍物的同时执行任务。另一个尚未解决的挑战是仿真到现实的鸿沟。尽管我们在仿真中使用虚拟传感器取得了令人鼓舞的结果，但这些方法仍需在真实机器人上进行验证。然而，在离线编程中，我们的深度强化学习规划器可能是经典运动规划方法的一种更高效替代方案。

附录

代码可在此处获取：https://github.com/ignc-research/IR-DRL

超参数	**值**
折扣因子	0.99
批量大小	256
学习率	0.0003
裁剪范围	0.2
单个回合的最大步数	1024
损失计算的熵系数	0.0
损失计算的价值函数系数	0.5
最大梯度	0.5
权衡因子	0.95

表 I：训练超参数

$\tilde{\rho}_{max}$	$\tilde{\rho}_{min}$	δ^+	δ^-	e_ζ	P_ζ	k
0.1	0.01	0.001	0.01	1e3	0.9	50

表 II：增强目标尺寸的参数

	执行时间（ [秒]）			**路径长度（ [米]）**		
	规划器					
实验	DRL	RRT	NC-RRT	DRL	RRT	NC-RRT
1	0.84	3.38	1.97	0.45	3.88	2.27
2	2.15	6.59	4.54	0.85	6.65	3.08
3	3.05	/	/	1.18	/	/
4	1.29	22.38	6.76	0.91	1.45	1.29
5	1.29	25.34	6.36	0.93	1.34	1.32
6	1.30	25.58	5.46	0.95	1.35	1.36
7	1.33	21.59	4.70	0.98	1.30	1.12

表 III：所有实验 30 次运行的平均值。最佳和最差表现分别以灰色和红色标示。

参考文献

[1] Everett, Michael, Chen, Y. Fan, How, and J. P，《利用深度强化学习在动态环境中的决策智能体之间进行运动规划》，载于*2018年IEEE/RSJ*智能机器人与系统国际会议*(IROS)*。IEEE，2018年，第3052–3059页。[2] Chiang, H.-T. Lewis, Faust, Aleksandra, Fiser, Marek, Francis, and Anthony，《利用autorl端到端学习导航行为》，*IEEE*机器人与自动化快报，第4卷，2014，2019 2，2007 no. pp. – 。[3] Kästner, Linh, Buiyan, Teham, Jiao, Lei, Le, T. Anh, Zhao, Xinlin, Shen, Zhengcheng, Lambrecht, and Jens，《Arena-rosnav：将基于深度强化学习的避障应用于传统自主导航系统》，载于*2021年IEEE/RSJ*智能机器人与系统国际会议*(IROS)*。IEEE，第6456–6463页。[4] L. Wu 和 Y. Hori，《基于八叉树模型的机器人机械臂实时无碰撞路径规划》，第九届*IEEE*高级运动控制国际研讨会，2006年。[5] T. Zhang, M. Zhang 和 Y. Zou，《变化环境中机械臂的实时路径规划》，*International Journal of Control and AutomationandSystems*，2010年。[6] Wei 和 Ren，《基于改进RRT算法的机器人机械臂自主避障动态路径规划方法》，传感器，2018年。[7] W. XINYU, L. XIAOJUAN, G. YONG, S. JIADONG 和 W. RUI，《双向势场引导RRT*运动规划》，*IEEE Access, vol. 7, pp. 95046-95057, 2019, doi: 10.1109/ACCESS.2019.2928846.*，2019年。[8] X. Wang, X. Luo, B. Han, Y. Chen, G. Liang 和 K. Zheng，《基于改进快速探索随机树算法的机器人无碰撞路径规划方法》，*Applied Sciences*，2020年。[9] S. S. M. Salehian, N. Figueroa 和 A. Billard，《协调多臂运动规划的统一框架》，国际机器人研究杂志 *2018, Vol. 37(10) 1205–1232*，2018年。[10] H. Zhang, H. Jin, Z. Liu, Y. Liu, Y. Zhu 和 J. Zhao，《时变环境中冗余机械臂的实时运动学控制：多动态障碍物避障与移动物体快速跟踪》，*IEEE*工业信息学汇刊，2020年。

[11] Kästner, Linh, Bhuiyan, Teham, Le, T. Anh, Treis, Elias, Cox, Johannes, Meinardus, Boris, Kmiecik, Jacek, Carstens, Reyk, Pichel, Duc, Fatloun, Bassel 等，《Arena-bench：高度动态环境中避障方法的基准测试套件》，*IEEE*机器人与自动化快报，第7卷，第4期，第9477–9484页，2022年。[12] Dugas, Daniel, Nieto, Juan, Siegwart, Roland, Chung 和 J. Jen，《Navrep：用于动态人类环境中机器人导航强化学习的无监督表征》，载于*2021*年*IEEE*机器人与自动化国际会议*(ICRA)*。IEEE，2021年，第7829–7835页。[13] Kastner, Linh, Cox, Johannes, Buiyan 和 Lambrecht，《All-in-one：结合先进导航规划器的基于深度强化学习的控制切换器》，载于*2022*年机器人与自动化国际会议*(ICRA)*。IEEE，2022年，第2861–2867页。[14] S. Wen, J. Chen, S. Wang 和 H. Zhang，《基于深度确定性策略梯度的人形机械臂路径规划》，机器人与仿生学国际会议，2018年。[15] B. Sangiovanni, G. P. Incremona, M. P. Antonella Ferrara 和 A. Ferrara，《通过深度强化学习实现带避障功能的自配置机器人路径规划》，第*17*届计算机与机器人视觉会议*(CRV)*，2020年。[16] Faust, Aleksandra, Oslund, Kenneth, Ramirez, Oscar, Francis, Anthony, Tapia, Lydia, Fiser, Marek, Davidson 和 James，《PRM-RL：结合强化学习与基于采样的规划完成远距离机器人导航任务》，载于*2018*年*IEEE*机器人与自动化国际会议*(ICRA)*。IEEE，2018年，第5113–5120页。[17] Dugas, Daniel, Nieto, Juan, Siegwart, Roland, Chung 和 J. Jen，《Navrep：用于动态人类环境中机器人导航强化学习的无监督表征》，*arXiv*预印本*arXiv:2012.04406*，2020年。[18] Güldenring, Ronja, Görner, Michael, Hendrich, Norman, Jacobsen, N. Jul, Zhang 和 Jianwei，《用于室内环境人类感知导航的局部规划器学习》，2020年。[19] K. Kamali, I. A. Bonev 和 C. Desrosiers，《利用动态目标深度强化学习进行机器人遥操作的实时运动规划》，第*17*届计算机与机器人视觉会议*(CRV)*，2020 年。[20] Zhong, Wang, Lin, Xu 和 Cheng，《通过深度强化学习与逆运动学混合算法实现焊接机械臂的无碰撞路径规划》，*Complex& IntelligentSystems*，2021年。[21] R. C. Luo 和 C.-W. Kuo，《用于制造自动化工业信息物理系统的智能七自由度机器人：动态避障与三维物体识别》，*IEEE*会议录，第*104*卷，第*5*期，第*1102-1113*页，2016年。[22] X. Zhu, Y. Liang, H. Sun, X. Wang 和 B. Ren，《使用深度强化学习的机器人避障系统》，智能机器人与应用国际期刊，2021年。[23] “Bullet3”，https://github.com/bulletphysics/bullet3，访问日期：2022-08-31。[24] Wampler 和 C. W.，《基于向量公式和阻尼最小二乘法的机械臂逆运动学解》，*IEEE*系统、人类与控制论汇刊，第16卷，第1期，第93–101页，1986年。[25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford 和 O. Klimov，《近端策略优化算法》，*arxiv*预印本：*1707.06347*，2017年。[26] B. Mohammadi, Hadi, Zamani, M. Ali, Kerzel, Matthias, Wermter 和 Stefan，《用于伸手抓取任务的混合现实深度强化学习》，载于人工神经网络国际会议。Springer，2019年，第611–623页。