基于强化学习的三维仿真智能体路径规划

小组成员 (Group members)：贾欣乐　张方哲　夏特雄　郭兆容　刘姝含

指导老师 (Mentors)：袁春 俞承霖　张洋　郑江茂

X-Institute 深圳零一学院

Abstract 摘要

基于传统算法（如A*，动态规划法）的路径规划对复杂场景和多机交互情景的泛化能力不足，因而本组设计了基于强化学习的无人机三维动态路径规划方法，为无人机路径规划提供更广泛的方法选择，在很多情景中（如灾难救援，多机交互）有不可替代的作用。

Introduction 问题简述

行业现状：低空经济迅速发展，同时无人机市场规模雄厚。但传统算法在大量应用场景中难以承担任务。基于强化学习的三维仿真智能体路径规划能有效降低实地测试的成本，可以为物流、拍摄等应用提供技术支持

技术难题：原有的算法只能应对静态环境下的避障，需要已知地图，不能直接利用感知信息，且泛化能力差；现有强化学习任务环境简单，聚焦于离散动作空间。

项目目标：在此基础上，我们希望利用三维仿真库模拟实现动态环境下的实时避障，通过传感器获取感知信息，进行分析，通过强化学习训练，得到拥有良好避障能力的模型。

项目优势：（1）实时处理，泛化能力强，未来应用场景广泛 （2）可移植性高，便于在无人机上使用　（3）三维仿真，与真实环境更接近

Plan 研究计划

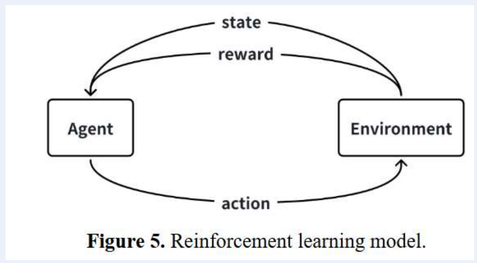
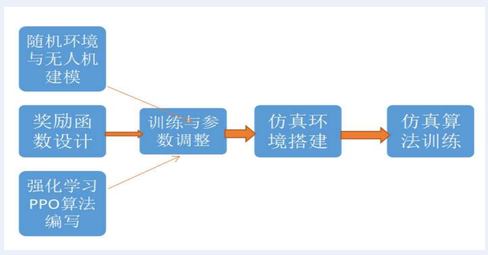


Figure 5. Reinforcement learning model.

强化学习

1.文献阅读和网络调研---》对强化学习的可行性和进化性的了解
2.程序设计和修改--》提升奖励函数的合理性和训练的稳定性
3.训练和仿真--》提升智能体的泛化能力，鲁棒性，并初步实现对仿真环境中物理规律的学习能力
4.未来展望：进行多智能体交互和仿真训练升级，为无人机交通网的现实实现赋能



思维导图

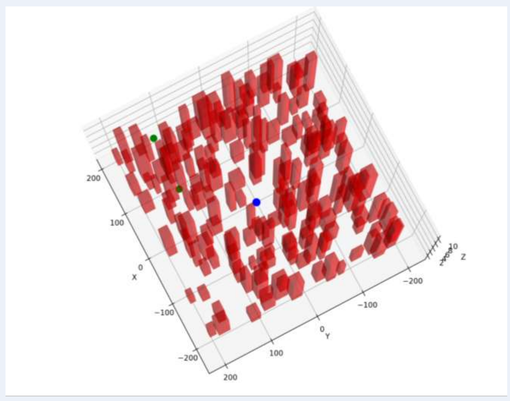
Primary Results 初步研究结果

1.**奖励函数优化设计**：设计了多目标奖励公式，包括**目标导向奖励**（鼓励智能体持续向当前目标靠近），**目标达成奖励**（强化完成任务的动力），**避障奖励**（主动规避危险），**碰撞惩罚**（禁止危险行为）和**边界违规惩罚**（限制活动范围）等

```python
def _calculate_reward(self, action, current_distance):  1 个用法
    # 1. 碰撞惩罚：降低惩罚强度，保留足够警示但避免彻底否定
    collision = not self._is_position_safe(self.position)
    if collision:
        return -5000.0  # 从-2000降至-500，减轻惩罚力度

    # 2. 目标导向奖励：即使远离目标，也给予基础正向奖励（避免持续负反馈）
    distance_reduction = self.prev_distance - current_distance
    goal_base_reward = 50.0  # 基础存活奖励（无论是否靠近目标）
    goal_reward = goal_base_reward + 30.0 * max(distance_reduction, -2.0)  # 降低距离奖励系数至30

    # 3. 避障奖励：提高权重，强化避障优先级
    lidar = self._get_lidar_scan()
    min_obstacle_dist = min(lidar)
    obstacle_reward = 0.0
    # 安全距离越长、持续越久，奖励越高（鼓励主动保持安全）
    if min_obstacle_dist > 20.0:  # 更远的安全距离
        obstacle_reward += 80.0 + 2.0 * self.current_step  # 随步数累积奖励
    elif min_obstacle_dist > 10.0:
        obstacle_reward += 40.0 + 1.0 * self.current_step
    elif min_obstacle_dist > 3.0:
        obstacle_reward += 10.0  # 近距安全但需谨慎
    else:
        obstacle_reward -= 200.0  # 极近障碍物，警告性惩罚（非碰撞）

    # 4. 动作平滑惩罚：保持不变，鼓励稳定移动
    action_penalty = -5.0 * np.sum(np.square(action))
```

奖励函数代码

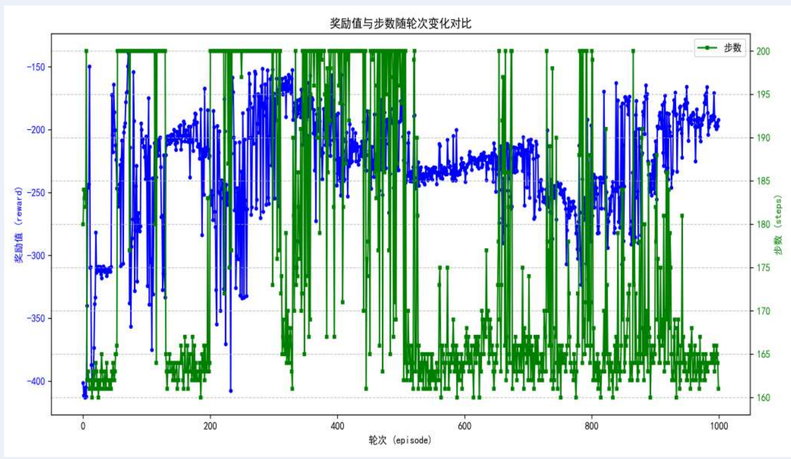
2.**仿真环境搭建**：基于pybullet构建了一个高度可配置的3D无人机配送仿真环境，仿真环境具有提供视觉（3D场景渲染）、物理（碰撞检测）和动力学（风速扰动）等多维度仿真数据，并且允许在零风险条件下测试极端场景（如电池耗尽、强风天气）等优势



仿真环境

3.**采用PPO算法**：其实现简单、训练稳定、样本效率高、在多种环境中表现优异，其采用平滑裁剪（clip）避免奖励突变，可以动态调整奖励系数（如completion_bonus），匹配PPO的阶段性学习特征，并且可以增量式环境变化（如风力）避免剧变导致策略崩溃，动态难度调整机制（目标扰动概率随训练轮次增加）等诸多优势

研究现状：（1）在仿真环境中基于强化学习进行六百万次的训练，得到初期的模型；
（2）在简化仿真环境下，实现了静态和动态环境中的模拟并得到了较好的结果；
（3）基于结果进行了图像生成和数据分析



Reference 参考文献

1.2hang,, & Wang, H. 2022. UAV path plamning based on the improved pp0 algorithm. 2022 internationalConferenceon Unmanned Aircrat systems ICUs, 1253-1260
2.王伟,张强,李华.(2023).基于改进PPO算法的AUV路径规划研究,电光与控制, 30(2),56-62.
3.北京航空航天大学.(2023).基于改进PPO算法的多无人机路径规划方法[Patent].CN117193378B.
4.Li, Y, et al. (2023). Mean-field deep RI for fair and eficient UAV control. IEE Internet of ThingsJournal, 10(5), 4125-4138.