报告准备中，请稍候……

智能决策论坛

柯良军

西安交通大学自动化学院

教授

智能感知与决策研究中心

主任

Multi-agent Reinforcement Learning for Large-scale UAV Swarm Cooperative Attack-Defense Confrontation

报告摘要：

This talk considers the problem of Large-scale unmanned aerial vehicle (UAV) swarm attack-defense confrontation in a three-dimensional environment. We propose a new Multi-Agent Reinforcement Learning (MARL) algorithm - Cooperative Deep Deterministic Policy Gradient (CODDPG), which has the following characteristics: It adopts the Actor-Critic framework, and the agents share a common network. In addition, the cooperation between UAVs is considered. To make it suitable for large-scale problems, the Mean Field Theory (MFT) is adopted and the new local state representation and reward allocation method are proposed. Moreover, CODDPG has a new network structure, which considers the executable actions in the current state. Finally, the algorithm uses a framework of centralized training and decentralized execution. In this way, each agent can only rely on its local observation to make decisions. To study the proposed algorithm, a large-scale UAV swarm confrontation platform considering flight constraints and real UAV environment is designed. The experimental results show that CODDPG has better performance than several other mainstream MARL algorithms in the simulation platform.

个人简介：

柯良军，西安交通大学自动化学院教授，博士生导师，IEEE member，西安交通大学智能感知与决策研究中心主任，中国仿真学会智能仿真优化与调度专委会委员，中国自动化学会无人飞行器自主控制专委会委员，IEEE member，《控制理论与应用》编委，出版著作《强化学习》、《蚁群智能优化方法及其应用》。

西安交通大学
XI'AN JIAOTONG UNIVERSITY

基于多智能体深度强化学习的无人机集群对抗研究

柯良军
西安交通大学自动化学院
2020.09.19

西安交通大学
XI'AN JIAOTONG UNIVERSITY

基于多智能体深度强化学习的无人机集群对抗研究

柯良军
西安交通大学自动化学院
2020.09.19

西安交通大学
XIAN JIAOTONG UNIVERSITY

基于多智能体深度强化学习的无人机集群对抗研究

柯良军
西安交通大学自动化学院
2020.09.19

CONTENTS

目录

- 无人机集群是一个由多架无人机相互协作、执行共同任务的统一系统
- 无人机集群有着更加强大的群体协作能力，可以协同完成更复杂的任务和目标

工信部下发《促进新一代人工智能产业发展三年行动计划(2018-2020)》文件中指出，将重点培育包括智能无人机、智能网联汽车、智能服务机器人等在内的八大类人工智能产品，完成一系列人工智能标志性产品的重要突破。

无人机技术已经成为国家战略

1.1 背景和意义

- 无人机集群是一个由多架无人机相互协作、执行共同任务的统一系统
- 无人机集群有着更加强大的群体协作能力，可以协同完成更复杂的任务和目标

搜索救援　　协同侦擦

军事战争　　飞行表演

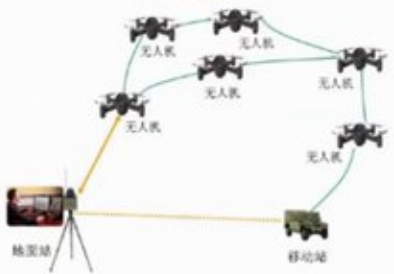
工信部下发《促进新一代人工智能产业发展三年行动计划(2018-2020)》文件中指出，将重点培育包括**智能无人机**、智能网联汽车、智能服务机器人等在内的八大类人工智能产品，完成一系列人工智能标志性产品的重要突破。

无人机技术已经成为**国家战略**，具有重要研究意义

1.2 无人机集群对抗技术

无人机集群对抗是无人机作战的重要模式，是无人机集群的重要形式。它是一群无人机对另一群无人机进行拦截而形成的空中协作式的缠斗，对抗中无人机具有**自组织、自适应**特点和**拟人思维**属性,通过**感知环境**，依据一定的行为规则， 采取攻击、避让、分散、集中、协作、援助等有利策略，使得在整体上涌现出**集群对抗系统的动态特性**。

单无人机系统



群体无人机系统

挑战

- 姿态感知差
- 状态共享滞后，获取信息延迟
- 防撞自主水平差
- 无人机运动相互干扰
- 高动态、未知Markov环境
- 任务分配要求高

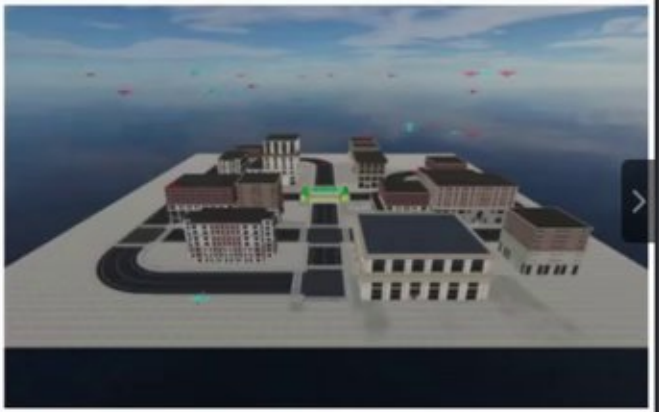
1.2 无人机集群对抗中的关键问题

动力学建模	个体决策对抗行为	集群智能控制方法	集群探测与识别	集群通信技术
• 环境的多元化和不完备、不确定性，复杂的动态随机性 • 利用系统动力学和复杂系统理论建立网络拓扑架构，有利于对其定量和定性分析。	• 个体是动作的发出和执行者，个体与环境交互，促使对抗过程不断演化。 • 集群对抗依赖于个体的对抗规则。	• 集群是一类非线性复杂系统，具有群体智能涌现的复杂性和随机性。 • 集群具有分布式控制的本质特征。单一方法控制往往不够理想，需要考虑集群控制策略。	• 精确地探测和识别是成功的关键。 • 不确定、部分确定条件下，实时在线进行主动感知和目标区分，对于集群的上层决策具有指导意义。	• 既要满足个体的信息交互，同时减少延迟，保证信息的实时性。 • 无人机存在动态加入和退出等状态，通信必须支持无人机数量的变化，

2 问题描述与建模

问题描述

- 建筑群中有两组对抗的无人机群，分别为入侵者(红色)和防御者(蓝色)
  - 双方拥有完全相同的飞行约束
  - 环境中存在随机障碍物，无人机无法提前获知信息
  - 无人机之间相撞或与建筑物相撞则退出环境
  - 至少n架防御者足够接近一架入侵者时，摧毁该入侵者
- 两组无人机拥有各自任务，相互协作和对抗
  - 入侵方目标为进入规定目标点
  - 防御方防守目标，阻止入侵者接近
  - 任何一架入侵者接近目标点，入侵方获胜；所有入侵者被摧毁，判定防御方获胜



三维连续空间

环境约束

所有无人机在飞行时要服从以下约束

1 速度、加速度约束

$$|v_{x,y,z}| \le v_{max_{x,y,z}}$$

$$|a_{x,y,z}| \le a_{max_{x,y,z}}$$

2 高度约束

$$h_{min} \le h \le h_{max}$$

3 最大偏航角(yaw)约束

$$cos(\phi) \le \frac{\alpha_i^T \alpha_{i+1}}{\|\alpha_i\| \bullet \|\alpha_{i+1}\|},$$

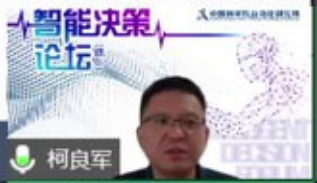
4 障碍物约束

$$l \geqslant R_{safe} + l_{min} + R_{UAV}$$

5 边界约束

建筑群四周是有界的，
无人机不能超出其范围

6 ……



数学模型

数学模型：疆土守卫博弈问题（Territorial defense game problem）
一类微分博弈问题

求解方法：基于博弈论的解析法、群体智能、强化学习

协作与对抗中核心问题：
- 通信问题，即信息交互问题
- 任务生成与分配问题

Xuan, S., Ke, L. Multi-agent reinforcement learning for large-scale UAV swarm co
defense confrontation. *Frontiers of Information Technology & Electronic Engineering*,

环境建模与设计

建模主要流程

1. Unity搭建UAV仿真平台
2. 基于ML-Agent构建Unity RL环境
3. Socket实现Unity与Python通信
4. 基于Gym构建Python RL环境

下一

环境建模与设计

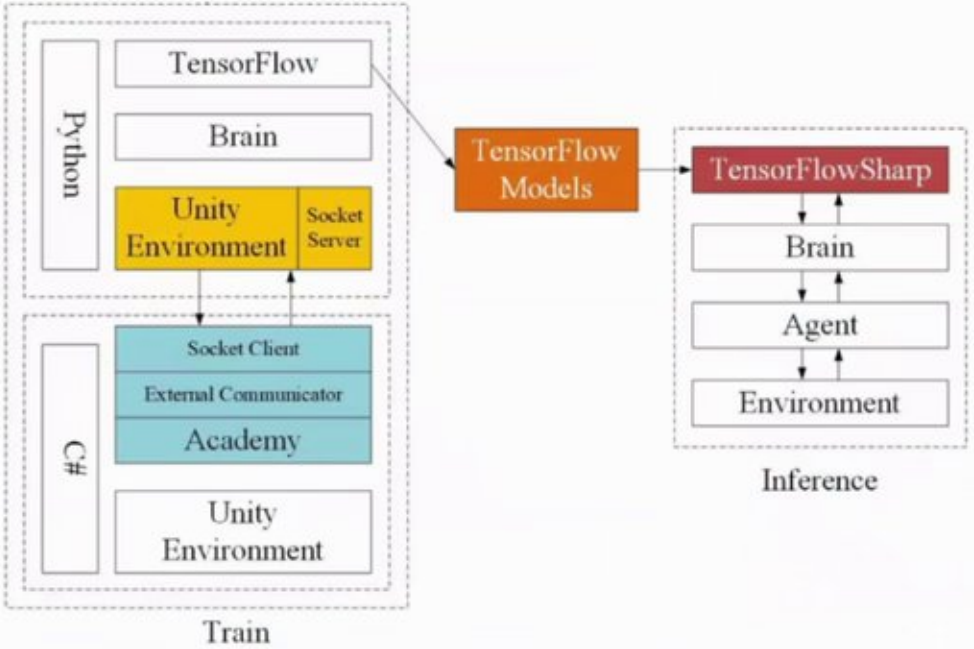
Unity搭建UAV仿真平台

- Unity3D是由Unity Technologies开发的三维视频游戏的多平台的综合型游戏开发工具，是一个全面整合的专业游戏引擎
- 通过Unity可开发符合真实物理环境的三维空间下的UAV集群对抗平台，可实现大规模无人机场景，模拟真实的飞行环境

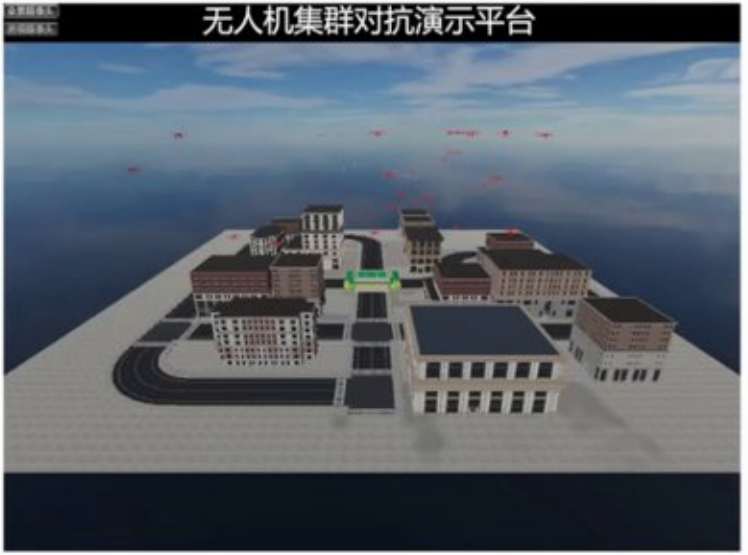
基于ML-Agent构建Unity RL环境

- ML-Agent 是unity专用强化学习组件，依赖Tensorflow 和 Tensorflowsharp 可快速搭建强化学习环境
- 可直接使用Tensorflow进行训练也可修改后将状态信息传递给Python模块
- 可同时并发进行多场仿真实验

环境整体框架



俯视效果

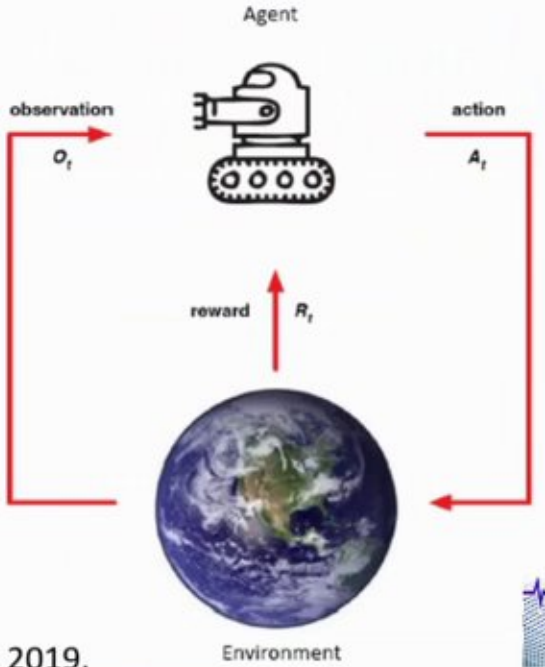


主视效果

3 算法设计

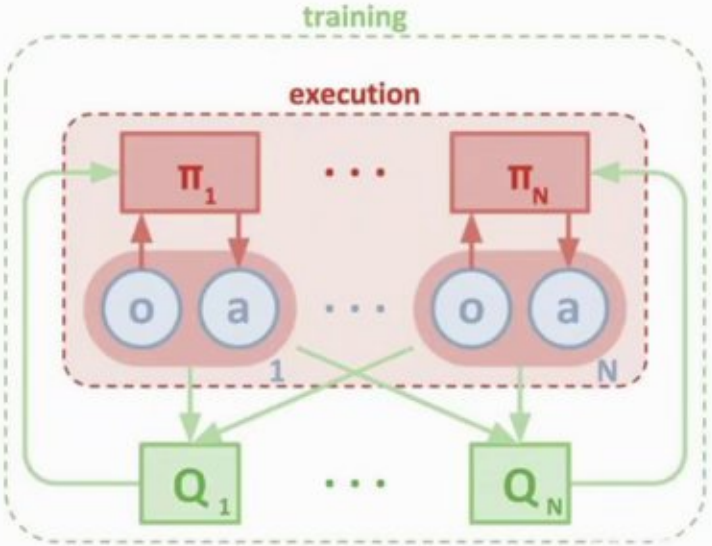
强化学习

- Learning from interaction with the environment
- The agent
  - senses the observations from environment
  - takes actions to deliver to the environment
  - gets reward signal from the environment
- Normally, the environment is stationary



柯良军，王小强. 强化学习，清华大学出版社，2019.

多智能体强化学习



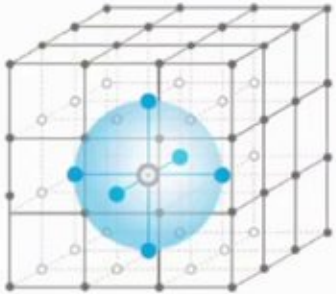
多智能体强化学习算法训练框架

传统RL方法面临的挑战：

- 无人机规模庞大，造成维度灾难，传统的多智能体强化学习算法难以应用
- 多个无人机共同完成同一目标，无人机必须学会组内协作，组间对抗
- 环境存在无人机与无人机、无人机与障碍物的碰撞，必须学习到合理的避障策略

平均场理论MARL

将环境对智能体的作用进行集中处理，以平均作用替代单个作用的加和



$$Q^j(s,\boldsymbol{a}) = \frac{1}{N^j}\sum_{k\in\mathcal{N}(j)} Q^j(s,a^j,a^k)$$

$$\approx Q^j(s,a^j,\bar{a}^j)$$

其中 $\bar{a}^j = \frac{1}{N^j}\sum_k a^k$

Yang, Yaodong, et al. "Mean field multi-agent reinforcement learning." *ICML* (2018).

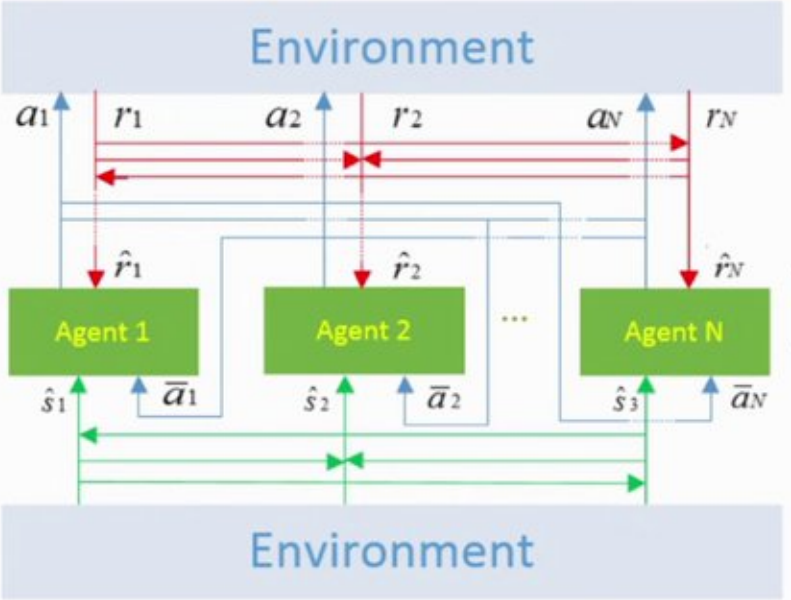
Co-DQL

- 智能体之间的信息交互

state共享 $\hat{s}_k = \langle s_k, \frac{1}{N_k} \sum_{i \in \mathcal{N}(k)} s_i \rangle,$

reward共享 $\hat{r}_k = r_k + \alpha \cdot \sum_{i \in \mathcal{N}(k)} r_i,$

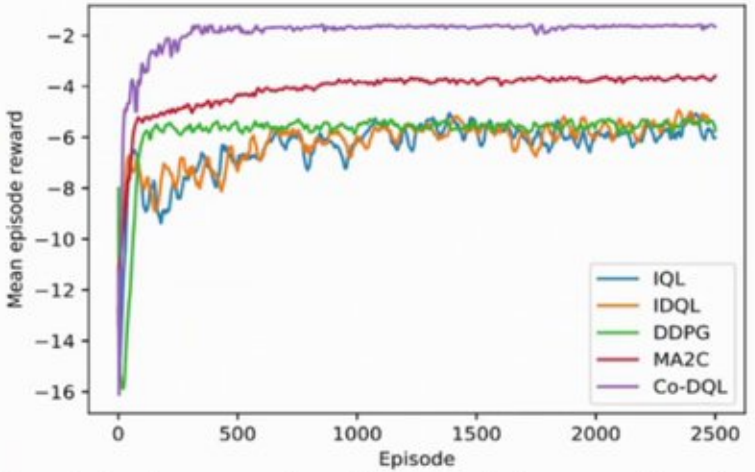
- 为应对过估计问题，采用double Q-learning



Wang, X., Ke, L , et al. 2020. Large-scale traffic signal control using a novel multi-agent reinforcement learning.IEEE transactions on Cybernetics, In press

Co-DQL分析

收敛性证明：在一定条件下，Co-DQL中的Q值以概率1收敛到 Nash Q-value



Wang, X., Ke, L , et al. 2020. Large-scale traffic signal control using a novel multi-agent reinforcement learning. IEEE transactions on Cybernetics, In press

CODDPG

- 关键技术1：智能体State表示
- 难点：传统State表示方法的维度随着智能体数量增加成指数增加，需要寻找一种在不显著增加网络开销的情况下的State表示方法
- 解决方法：基于平均场理论，对局部状态值进行了扩展，引入了其他所有智能体的局部状态的平均值得到了联合状态

$$\widehat{S}_{t,j} = < s_{t,j}, \frac{1}{|d(i)|} \sum_{j \in d(i)} s_{t,j} >$$

- 关键技术2：智能体Reward表示
- 难点：传统方法直接使用环境得到的外部奖励作为自身奖励，难以学习到协作的策略
- 解决方法：基于平均场理论，将其他智能体的奖励的平均值作为外部奖励，环境交互获得的奖励作为内部奖励

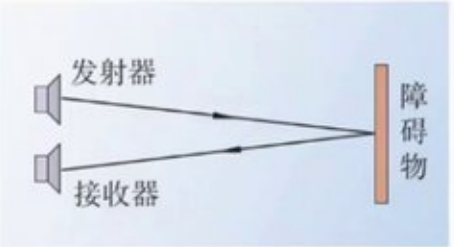
$$\widehat{r}_{t,i} = r_{t,i} + \frac{1}{|d(i)|} \sum_{j \in d(i)} r_{t,j}$$

CODDPG

- 关键技术3：动作表示
- 难点：由于集群对抗中的不确定性，智能体状态价值函数计算中需要将所有智能体的动作作为输入
- 解决方法：使用智能体动作的均值进行平均场近似表示，大大减小了网络的输入，同时降低了干扰因素

$$Q_i^{\mu}(x, a_1, ..., a_n) \approx Q_i^{\mu}(x, a_i, \bar{a}_i)$$

- 关键技术4：避障策略
- 难点：无人机需要考虑与障碍物、其他无人机的碰撞，远距离时无人机无法获取障碍物信息，因此要求无人机检测到障碍物时及时做出反应
- 解决方法：将当前状态下的可执行动作当作网络的输入参与训练，该方法学习到的策略能够合理的避障

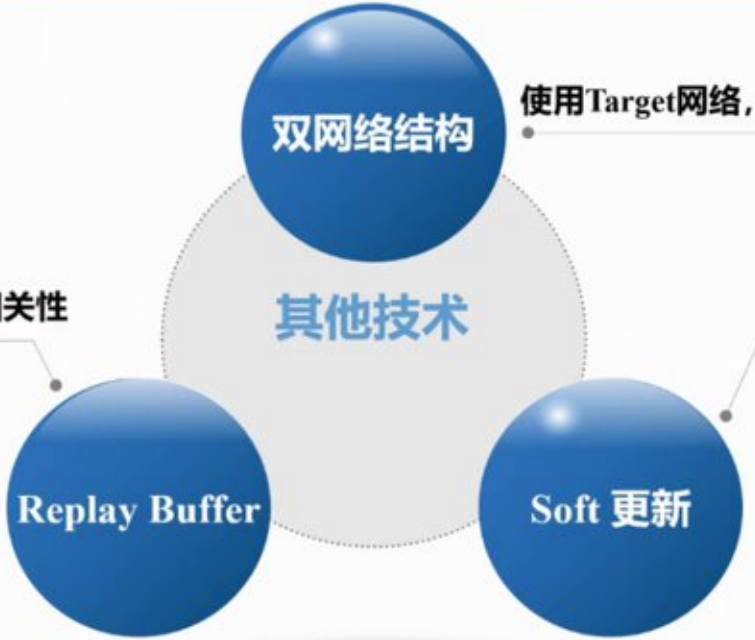


CODDPG

- 关键技术5：通信开销
- 难点：执行环节，无人机需要在极短时间内做出决策，若依赖于全局信息决策，通信会造成极大的开销
- 解决方法：采用集中训练，分散执行策略，在训练环节，无人机相互通信，学习合作策略；在执行环节，无人机仅依赖自身观察决策，不再需要通信

- 关键技术6：网络开销
- 难点：对于大规模场景，为每一个无人机维持一组神经网络会占用极大开销
- 解决方法：将具有相似任务的无人机共享同一组网络以减小网络开销

CODDPG


双网络结构
使用Target网络，避免价值函数的过高估计
其他技术
提高样本利用率，打破样本相关性
增加算法稳定性和鲁棒性
Replay Buffer
Soft 更新


智能决策论坛
柯良军

CODDPG

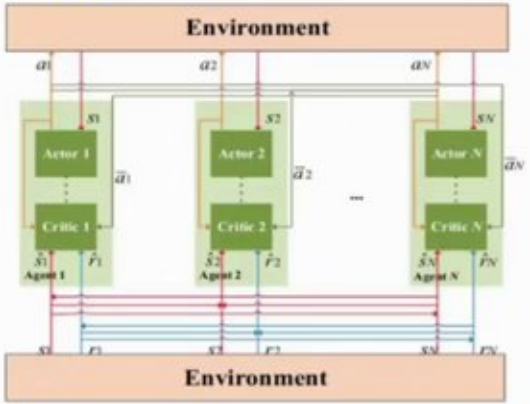


Fig. 2 The framework of CODDPG with n agents. For agent i, s_i and $\hat{s}_i$ represent its local state and joint state, $\hat{r}_i$ is joint reward, a_i and $\overline{a}_i$ represent its own action and average action. The directed lines in red, blue, orange and green respectively represent the information transmission process of state, reward, self action and average action.

Algorithm 1 CODDPG for UAV swarm confrontation for n agent

for episode = 1 to M do **do**
  Initialize simulation environment,initialize s_i and a_i for all $i \in \{1, \cdots, N\}$,
  Initialize evaluation network and target network for Actor network and Critic network.
  Compute $\hat{s}$
  for t = 1 to max-episode-length do **do**
    For each agent i, select action a_i using ϵ-greedy exploration policy
    Execute actions $\boldsymbol{a} = (a_1, \ldots, a_n)$, get reward r and new state s'
    Compute $\overline{\boldsymbol{a}}$, $\hat{r}$ and $\hat{s}'$
    Store $\left\langle \hat{\boldsymbol{s}}, \boldsymbol{a}, \hat{r}, \hat{s}', \overline{\boldsymbol{a}}, \boldsymbol{done} \right\rangle$ in replay buffer $\boldsymbol{D}$
    $s_i \leftarrow s_i'$
    if the length of D > MINI SAMPLE SIZE **then**
      for agent i = 1 to N do **do**
        Sample a random minibatch of S samples $\left\langle \hat{\boldsymbol{s}}, \boldsymbol{a}, \hat{r}, \hat{s}', \overline{\boldsymbol{a}}, \boldsymbol{done} \right\rangle$ from $\boldsymbol{D}$
        Compute $\boldsymbol{y}_i$ by Eq. ??
        Update evaluation Critic and evaluation Actor by Eq (14) and Eq (15)
      end for
      Update target Critic network parameters and target Actor network parameters for each agent i:

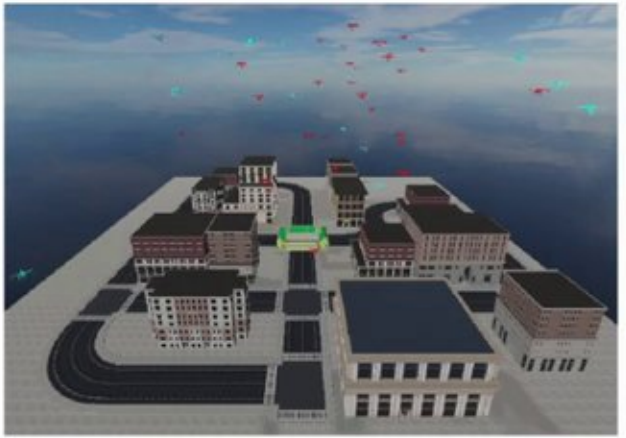
$$\theta' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$$

    end if
    if *done* **then**
      break
    end if
  end for
end for

4 实验与分析

环境设置

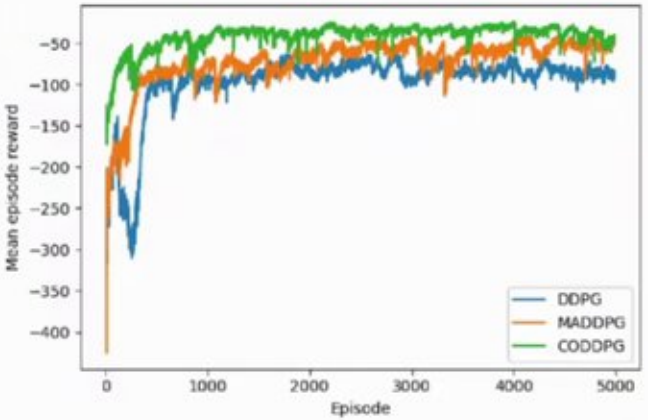
防御者数量	30
入侵者数量	20
感知其他无人机数量	3
摧毁奖励	0.3
入侵奖励	3
入侵者策略	DDPG/人工规则
防御者策略	DDPG/MADDPG/CODDPG



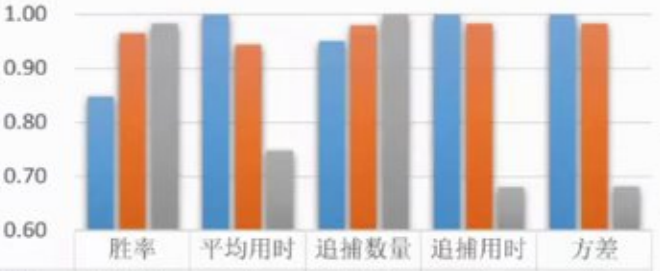
4.1 实验1：DDPG 策略

实验结果

训练曲线



测试结果



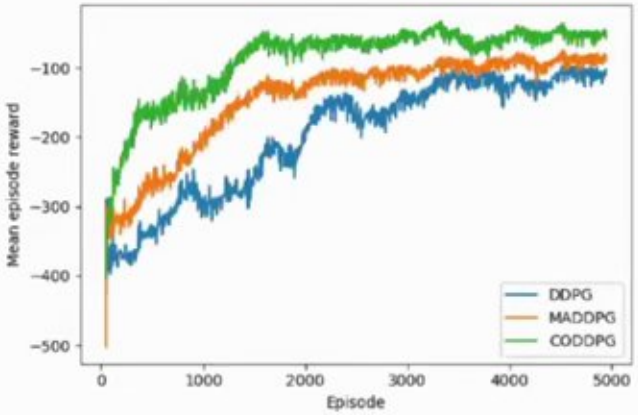
	胜率	平均用时	追捕数量	追捕用时	方差
DDPG	84.74%	1	0.95	1	1
MADDPG	96.52%	0.9439	0.98	0.9824	0.9824
CODDPG	98.28%	0.7472	1	0.6793	0.6814

- **胜率**：摧毁所有无人机的概率，**平均用时**：防御者胜利的平均步长，**追捕数量**：一局中防御者追捕无人机的平均数量，**追捕用时**：追捕一架无人机的平均用时，**平均方差**：平均用时的方差，除胜率外，以上数据经过归一化处理

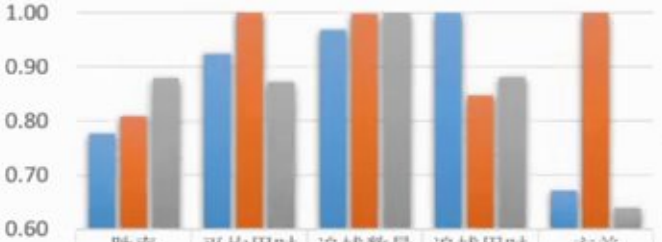
4.1 实验2：人工规则策略

实验结果

训练曲线



测试结果



	胜率	平均用时	追捕数量	追捕用时	方差
DDPG	77.61%	0.9242	0.968	1	0.6705
MADDPG	80.86%	1	0.9977	0.8468	1
CODDPG	87.90%	0.8729	1	0.8819	0.6384

- **胜率**：摧毁所有无人机的概率，**平均用时**：防御者胜利的平均步长，**追捕数量**：一局中防御者追捕无人机的平均数量，**追捕用时**：追捕一架无人机的平均用时，**平均方差**：平均用时的方差，除胜率外，以上数据经过归一化处理

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的未知场景，长期且稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得传统的方法难以学习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

5 总结与展望

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的
稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得
习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

5 总结与展望

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的未知场景，长期且稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得传统的方法难以学习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

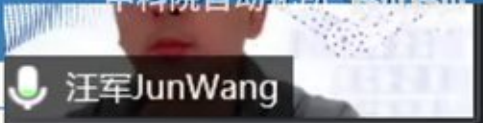
5 总结与展望

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态
稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使
习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

5 总结与展望



1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态
稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使
习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的未知场景，长期且稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得传统的方法难以学习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

5 总结与展望

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。

5 总结与展望

1、本课题主要研究了大规模场景下的无人机集群对抗问题。高动态的
稀疏的奖励、庞大的无人机数量以及复杂的无人机交互等特点，使得
习到好的合作策略。

2、 提出了一种新的多智能体强化学习算法，其引入平均场理论表征无人机间的相互作用，提出了一种新的联合状态和奖励分配形式，并采用集中训练、分散执行的策略，使得无人机能够依赖局部信息实现合作与对抗，该方法理论上可应用于任意规模问题。

3、设计了一个基于Unity和Python的三维无人机集群仿真对抗平台，并用于算法验证，实验结果表明，我们提出的算法能够使得无人机实现智能的合作对抗，表现优于目前主流的强化学习算法，并且可以方便迁移到更大规模的对抗问题中。