汕頭大學
SHANTOU UNIVERSITY

# 基于视觉的深度强化学习分布式避障导航策略学习方法

## Vision-based Deep Reinforcement Learning Distributed Collision Avoidance Navigation Policy Learning Method

答辩人：黄华兴　导师：范衡教授　电子信息

2023/5/19

汕頭大學
SHANTOU UNIVERSITY

# 目录

CONTENTS

汕頭大學
SHANTOU UNIVERSITY

1

# 研究背景

Research Background

# 多机器人避障导航

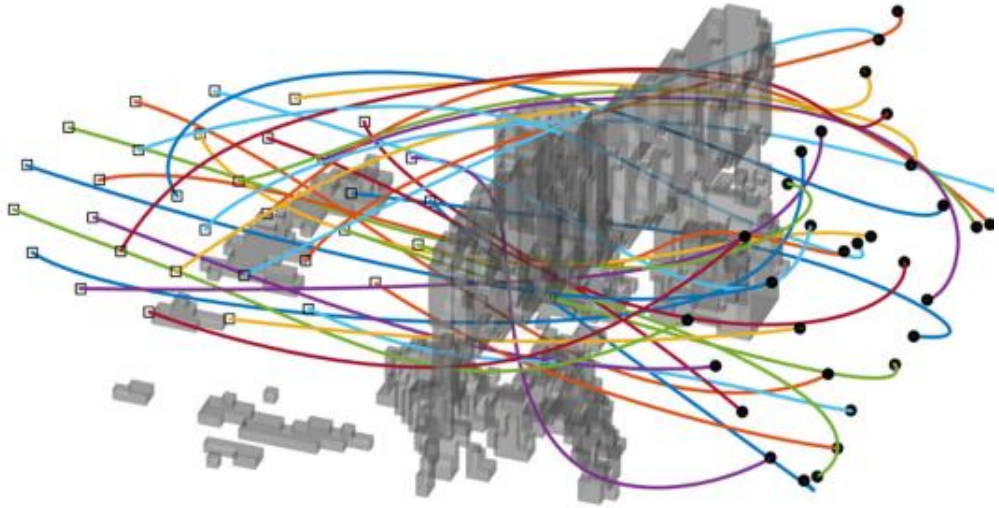
Multi-robot collision avoidance and navigation

- 多机器人避障导航近年来受到机器人学和人工智能的广泛关注，在多机器人搜救、密集人群中导航以及自动化仓储运输等领域有着广泛的应用

- 现实生活中无人机复杂多样的工作空间对多无人机系统的鲁棒性提出了极大的挑战，无人机的任何微小错误都可能对人员或其他财产造成巨大的损害或损失。

# 集中式多机器人避障导航

传统的集中式方法都是由中央控制器来协调所有机器人的运动，根据全局的环境信息和机器人状态，为每个机器人分配合适的目标位置和速度，使得整个系统能够高效地完成任务。



☺

- 能够得到全局最优解
- 保证编队的稳定性和一致性

☹

- 调度的计算成本高
- 严重依赖可靠的同步通信
- 对故障或干扰的容忍度低
- 不适合复杂环境

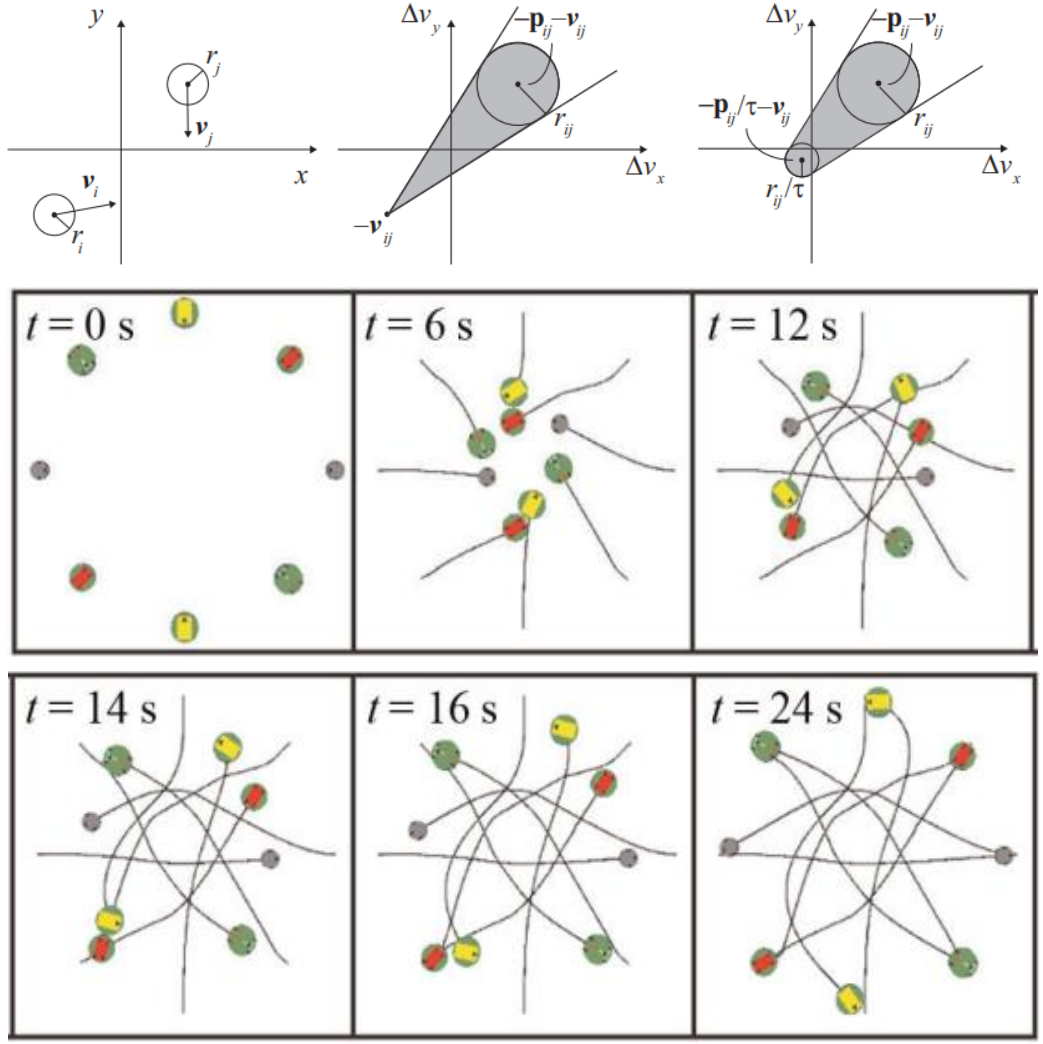
[1] Hönig W, Preiss J A, Kumar T K S, et al. Trajectory planning for quadrotor swarms[J]. IEEE Transactions on Robotics, 2018, 34(4): 856-869.

# 分布式多机器人避障导航

分布式方法是指每个机器人都有自己的控制器和传感器，根据传感器获取的局部的环境信息和机器人状态，为自己决定合适的目标位置和速度，使得整个系统能够协同地完成任务。

目前大多数分布式多机器人避障导航策略都基于Velocity Obstacle (VO)框架，该方法可以保证多个移动机器人在没有通信或中央协调的情况下，通过感知其它机器人的位置和速度，选择合适的自身速度来避免彼此之间的碰撞。

- 假设机器人具有完美的传感器
- 其参数对环境的设置十分敏感




[2] Bareiss D and van den Berg J (2015) Generalized reciprocal collision avoidance. The International Journal of Robotics Research 34(12): 1501–1514.

# 多无人机人避障导航会面临更多挑战

## 环境是大规模的

在这种情况下，基于SLAM的方法将失去效率，因为构建环境地图是不可行的。

## 环境是复杂的

由于基于感知和回避的方法通常设计用于解决稀疏障碍物环境中的导航问题，因此在复杂环境中应用时会失去效率。

## 环境是动态的

显然预先路径规划无法处理这种情况。而强化学习能够适应动态环境，但需要改进以应对大规模复杂环境。

2

# 研究内容

Research Contents

# 问题描述

在一个环境中有 $N$ 架无人机，多无人机的避障导航问题可以描述为部分可观测马尔可夫过程。在每个时间步 $t$ 中，第 $i$ 架无人机$(1 \le i \le N)$ 获得一个观测 $O_i^t$，然后通过策略 $\pi$ 输出一个动作 $a_i^t$，使得无人机从当前位置 $p_i^t$ 朝向目标位置 $g_i^t$ 移动。

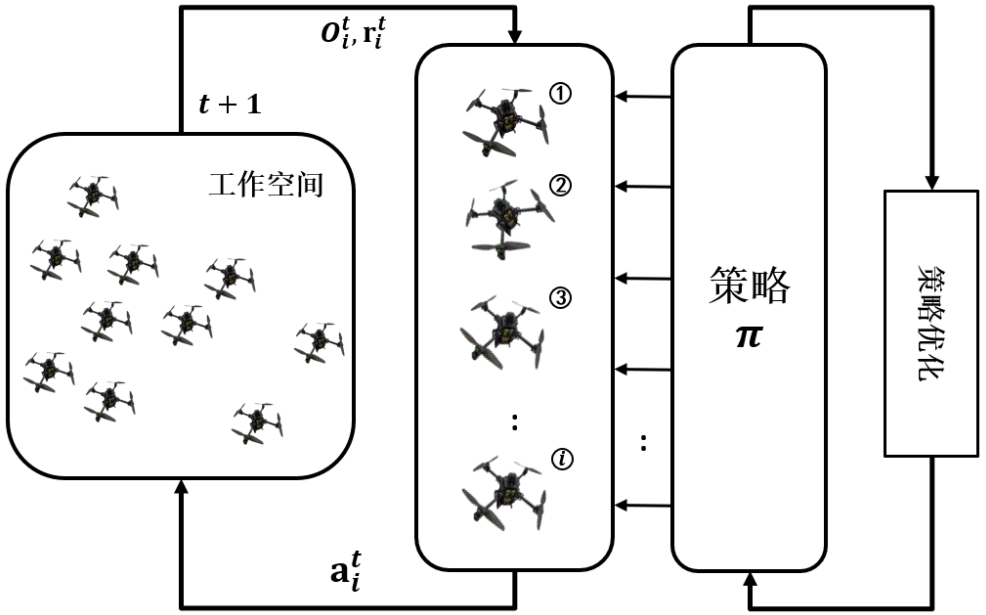



图3-1 本工作方法总体框架示意图

$$O^t = [O_z^t, O_g^t, O_v^t]$$

其中，$O_z^t$ 为视觉信息，$O_g^t$ 为目标点相对无人机的位置，$O_v^t$ 为无人机自身的速度。

$$a^t \sim \pi(a^t | O^t)$$

$$a^t = [v_x^t, v_z^t, w^t]$$

其中，$v^t$是线速度，$w^t$是偏航角的角速度。

# 观测空间

$O_z$：摄像头信息——深度图

深度图包含与视点场景对象表面距离有关信息的图像通道，每个像素值是传感器测出距离物体的实际距离，离摄像头越近越深，越远则越浅。

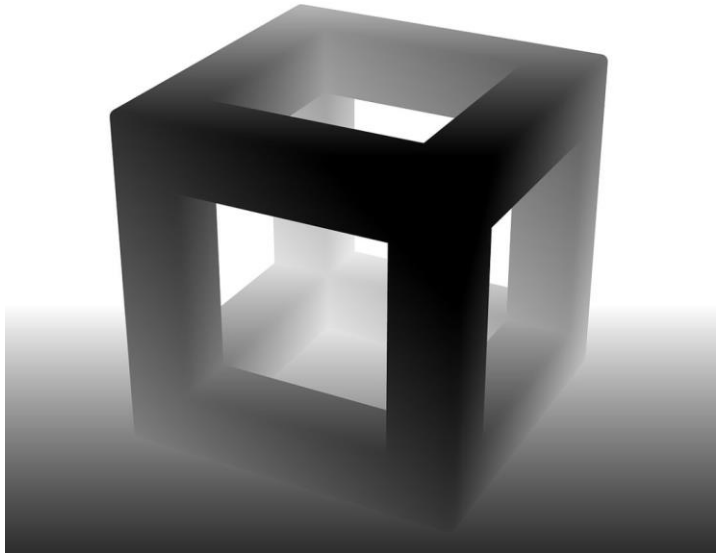


图3-2 深度图示意图

- 提供更多的空间信息，如距离、高度、形状等，有助于无人机识别其他物体。
- 避免视觉上的歧义和干扰，如光照、纹理、颜色等，提高了无人机导航的鲁棒性。
- 可以与其他传感器数据结合，如激光雷达、惯性测量单元等，构建更完善的环境模型和状态表示。

# 观测空间

深度图由每架无人机正前方的深度相机获取的。

为了提高策略学习的效率，本工作还对深度图做了一些处理：

- 使用连续多帧的深度图作为$o_z$，增大状态维度以区分不同状态，以减少过拟合
- 对深度图的像素值设置阈值，当感知到的物体距离太远时设置为白色，以减少图像的冗余信息

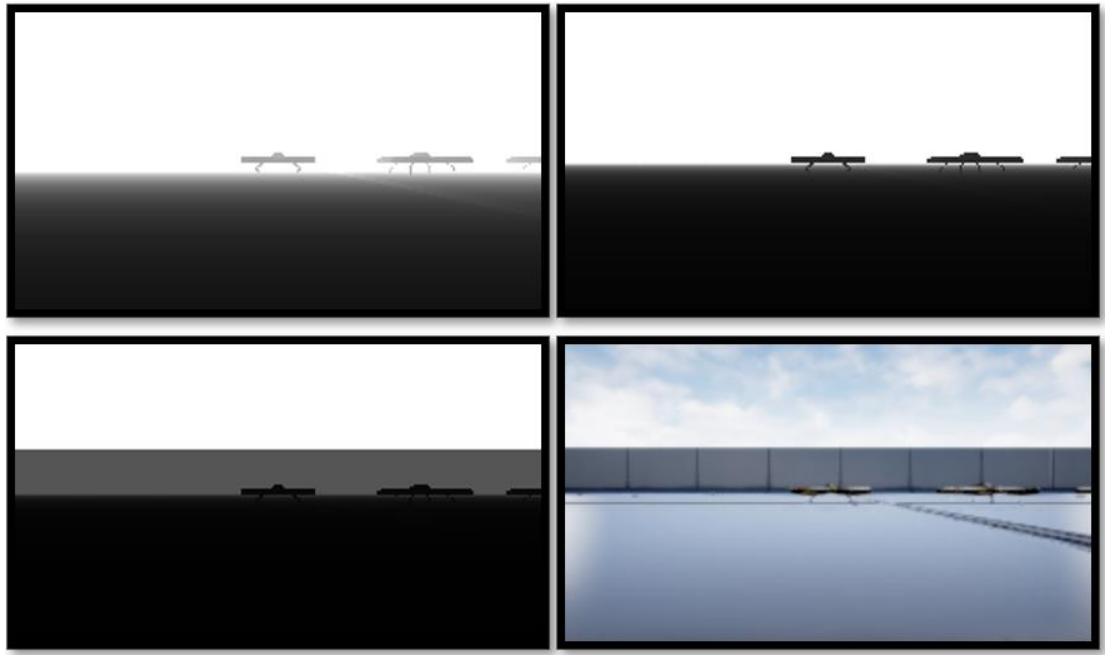


图3-3 不同阈值下的深度图示意图（分别是5m，20m，原始深度图和原始RGB图像）

# 观测空间

$O_g$：目标点相对无人机的位置——机体坐标系

对于四旋翼模型来说，需要两个坐标系：全局坐标系（NED北东地）和机体坐标系。

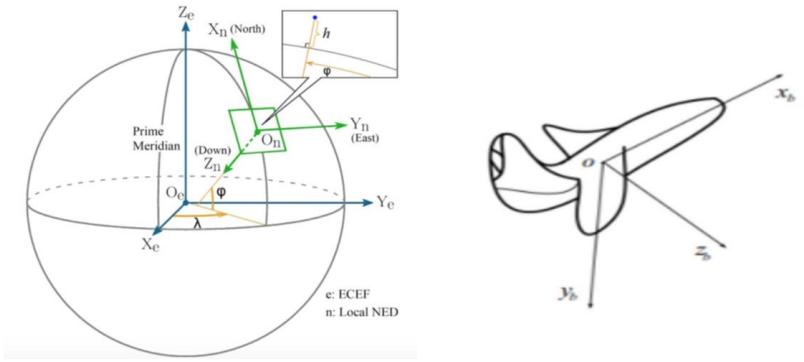



图3-4 NED北东地和机体坐标系示意图

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} R_y = \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix} R_z = \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

全局到机体坐标变换公式为：$P^b = R^{g2b}(P^g - T_b^g)$

其中：

- $P^g$是全局坐标系下的点的坐标，$P^b$是同样的点在机体坐标系下的坐标
- $R^{g2b} = R_x\ R_y\ R_z$，飞机角度为$\theta_x = roll, \theta_y = pitch, \theta_z = yaw$
- $T_b^g$为飞机全局坐标系下的位置

# 观测空间

$\boldsymbol{O_v}$：无人机的当前速度

$$\boldsymbol{O}_{\mathrm{v}}^{t} = [\mathrm{v_x}, \mathrm{v}_{\boldsymbol{z}}, \mathrm{v}_{\boldsymbol{w}}]$$

$\boldsymbol{O_v}$ 包含了两个线速度 $\mathrm{v_x}$ 和 $\mathrm{v}_{\boldsymbol{z}}$（前进速度和爬升速度）和一个角速度 $\mathrm{v}_{\boldsymbol{w}}$（转向速度）。这些速度通常可以使用惯性传感器（如陀螺仪、加速度计）测量得到。

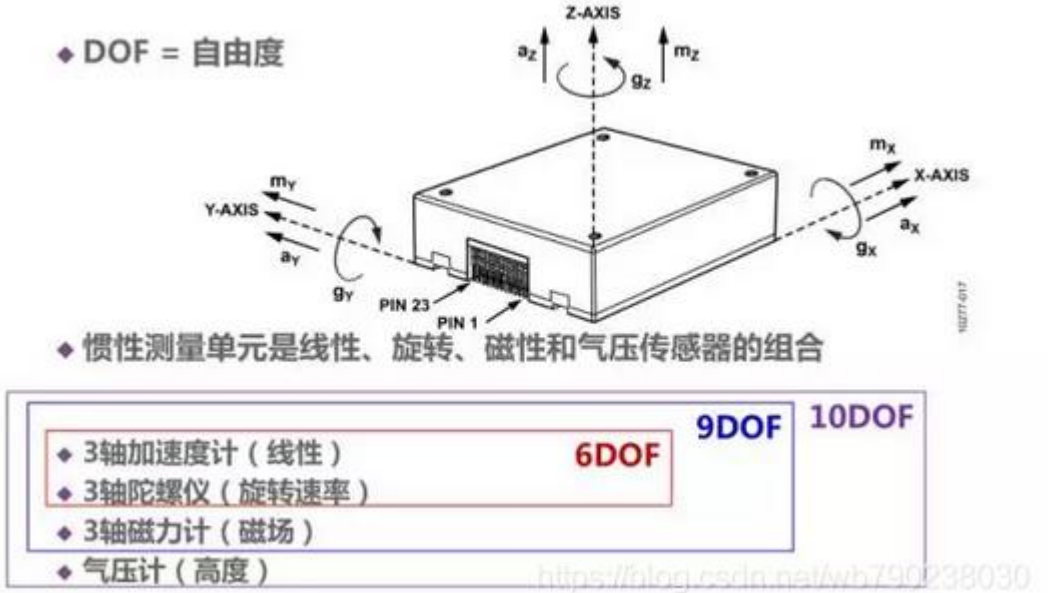



图3-5 惯性测量单元示意图

# 动作空间

$\boldsymbol{a} = [v_x^{cmd}, v_z^{cmd}, v_w^{cmd}]$：无人机速度控制指令

在强化学习中，一般会约束动作空间，因为：

- 动作空间过大会增加强化学习的难度和复杂度
- 动作空间过小会限制机器人的控制能力和灵活性
- 动作空间需要符合物理约束和安全约束

在本文中的工作中，基于以上考量，对动作空间进行了以下的速度约束：

- 前进速度 $v_x^{cmd} \in [0.0, 2.0] m/s$
- 爬升速度 $v_z^{cmd} \in [-0.5, 0.5] m/s$
- 转向速度 $v_w^{cmd} \in [-0.5, 0.5] \mathrm{rad}/s$

注意 $v_x^{cmd} < 0 m/s$是不被允许的！

# 奖励函数

学习目标是在无人机导航过程中避免碰撞，并且最小化所有无人机的到达时间。

奖励函数的设计：

$$r^t = r^t_{\text{goal}} + r^t_{avoid}$$

其中目标奖励$r^t_{\text{goal}}$旨在鼓励每架无人机尽快达到目标点，碰撞惩罚$r^t_{avoid}$是为了避免无人机发生碰撞。

$$r^t_{\text{goal}} = \begin{cases} r_{arrival} & if \|p^t - g^t\| < 0.5 \\ \omega_{\text{goal}}\left(\|p^{t-1} - g^t\| - \|p^t - g^t\|\right) & otherwise \end{cases}$$

$$r^t_{avoid} = \begin{cases} r_{collision} & if \|p^t_i - p^t_j\| < 2R \\ & or \|p^t_i - B_k\| < R \\ \omega_{\text{avoid}} max(r_{safe} - r_{min}, 0) & otherwise \end{cases}$$

其中，实际实验中采用的$r_{arrival} = 50, r_{collision} = -10, \omega_{\text{goal}} = 3, \omega_{\text{avoid}} = -0.05, r_{safe} = 5$。

# 策略学习网络

- 使用自编码器实现图像的降维和策略学习
- 防止actor网络的梯度更新编码器

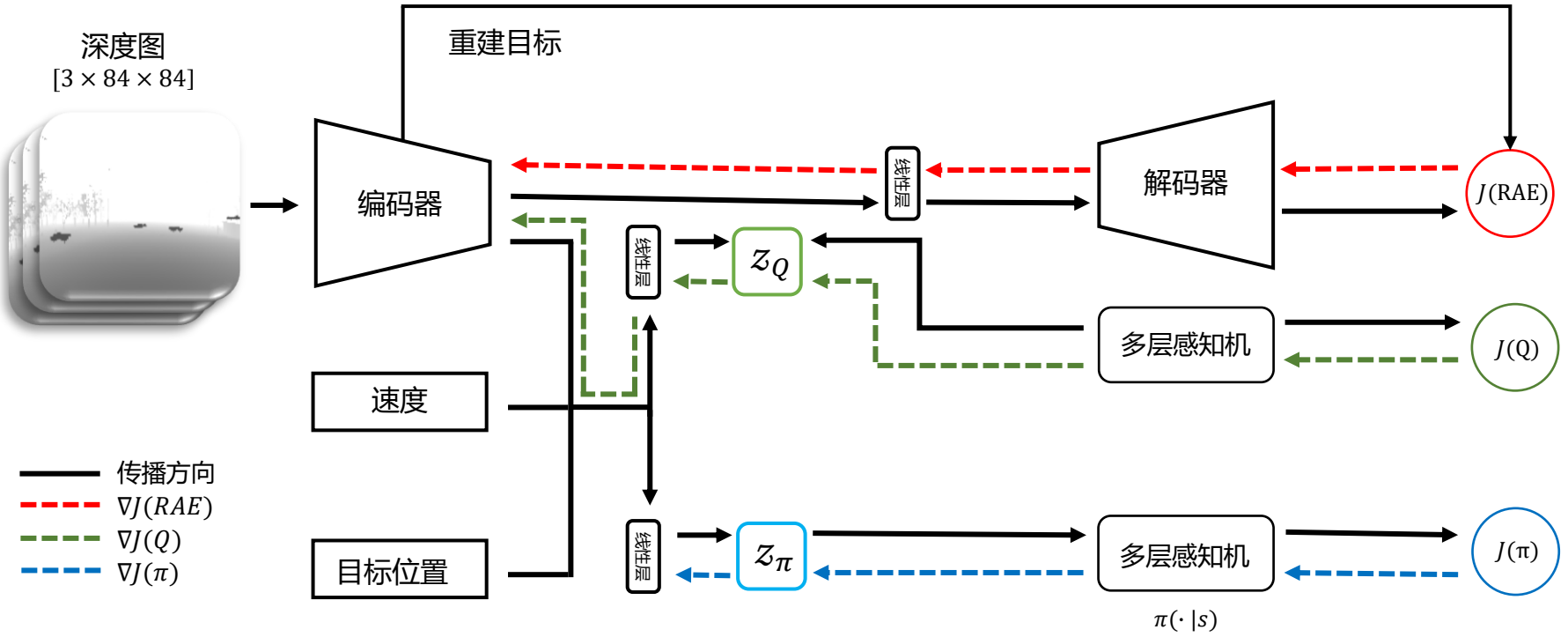



图3-6 策略网络结构示意图

# 策略学习算法

为了提升学习多无人机系统避障导航策略的效果，本文的工作中采用集中式学习，分布式执行的范式：训练过程中使用中央控制器，收集所有智能体的动作、状态和奖励，帮助它们优化策略网络。执行过程中智能体只根据自己的局部观测输出动作，不需要中央控制器的干预。

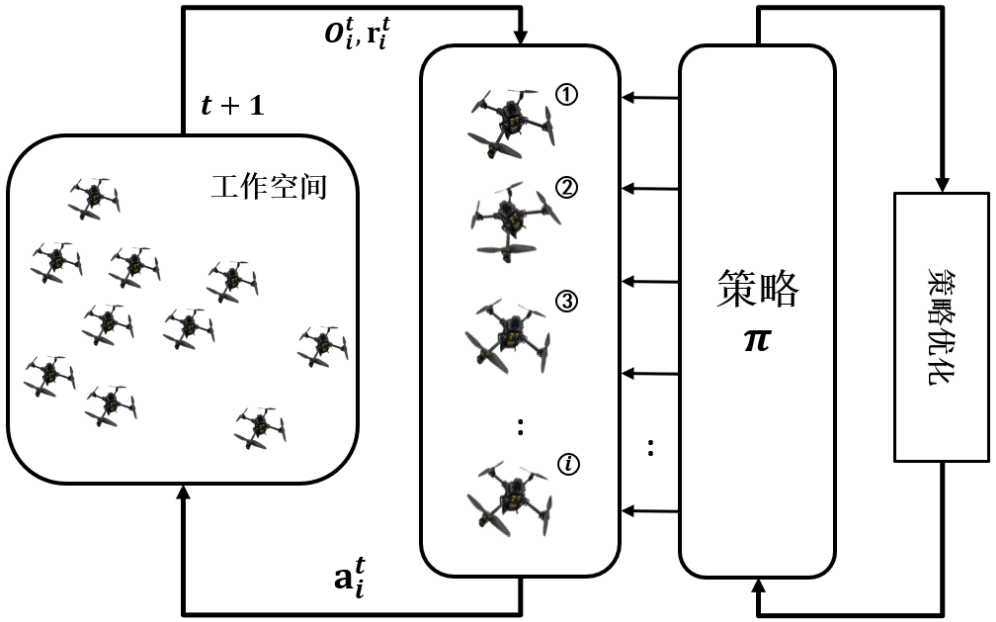



图3-7 策略学习机制示意图

**算法 3-1　多无人机系统避障导航策略学习算法**

初始化编码器，解码器，Actor，Critic 的网络参数和经验回放池 $\mathcal{B}$；
如果 当前回合数 < 最大回合数，重复以下步骤：
　在每个时刻 $t=1, 2, \ldots$ 时：
　　对于第 $i$ 架无人机，其中 $i=1, 2, \ldots, N$：
　　如果 $t <$ 最大时间 $T_{max}$ 或者当前不处于终止状态，重复以下步骤：
　　选择动作 $\boldsymbol{a}_i^t \sim \boldsymbol{\pi}_\theta(\cdot \mid \boldsymbol{o}_i^t)$；
　　得到奖励 $r_i^t$，下一时刻的观测 $\boldsymbol{o}_i^{t+1}$ 和状态终止信号 $d_i^t$；
　　把轨迹 $(\boldsymbol{o}_i^t, \boldsymbol{a}_i^t, r_i^t, \boldsymbol{o}_i^{t+1}, d_i^t)$ 存放到 $\mathcal{B}$ 中；
　如果 $\mathcal{B}$ 中存放的轨迹足够多，以一定更新次数重复以下步骤：
　　随机从 $\mathcal{B}$ 中取出一批轨迹 $(\boldsymbol{o}, \boldsymbol{a}, r, \boldsymbol{o}', d)$；
　　通过 $J(Q)$ 和 $J(\pi)$ 更新 Actor，Critic 中的网络参数；
　　通过 $J(RAE)$ 更新编码器和解码器中的网络参数；

图3-8 策略学习算法流程

3

# 实验结果与分析

Experimental results and analysis

# 实验设置

为了实现鲁棒的多无人机避障导航策略学习，我们在AirSim上构建了多无人机场景。

训练过程中，在每个回合里，在三维空间中会随机生成每架无人机初始位置和目标位置，其中每架无人机的初始位置和目标位置之间有一定的距离约束，保证无人机在初始状态和最终状态不会发生碰撞。

**表4-1 策略学习相关超参数设置**

| 参数 | 值 |
|---|---|
| 经验回放池$\mathcal{B}$容量 | 20000 |
| 批次样本容量 | 128 |
| 折扣因子$\gamma$ | 0.99 |
| 最大回合数 | 200 |
| 优化器 | Adam |
| Critic网络学习率 | $10^{-3}$ |
| Critic目标网络更新频率 | 2 |
| Actor网络学习率 | $10^{-3}$ |
| Actor目标网络更新频率 | 2 |
| 自动编码器学习率 | $10^{-3}$ |

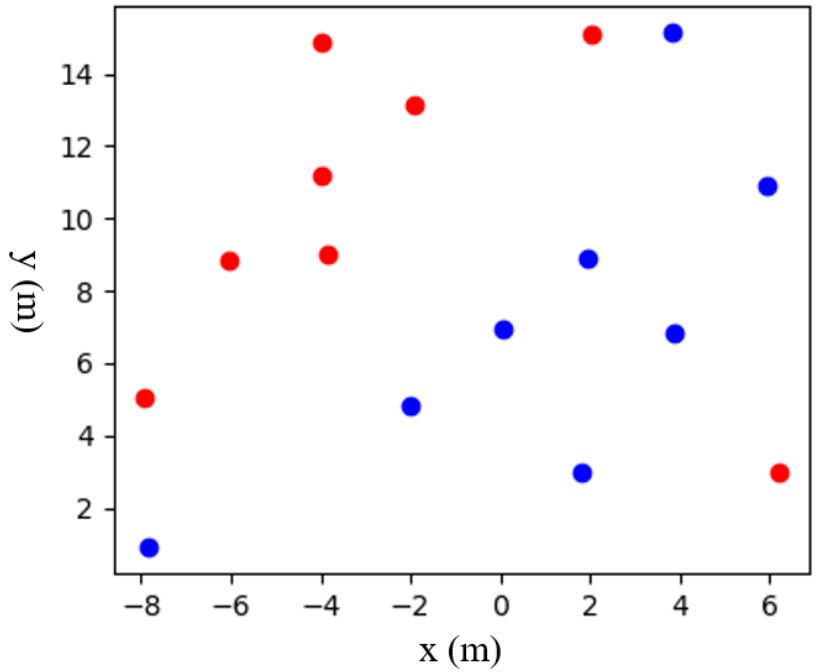



**图4-1训练过程中x-y平面位置生成示意图**

# 实验设置

随机场景：每架无人机的初始位置和目标位置随机生成，场景根据无人机的密集程度分为三种类型，即稀疏场景（约0.04 架/$m^3$）、正常场景（约0.06 架/$m^3$）和密集场景（约0.10 架/$m^3$）。

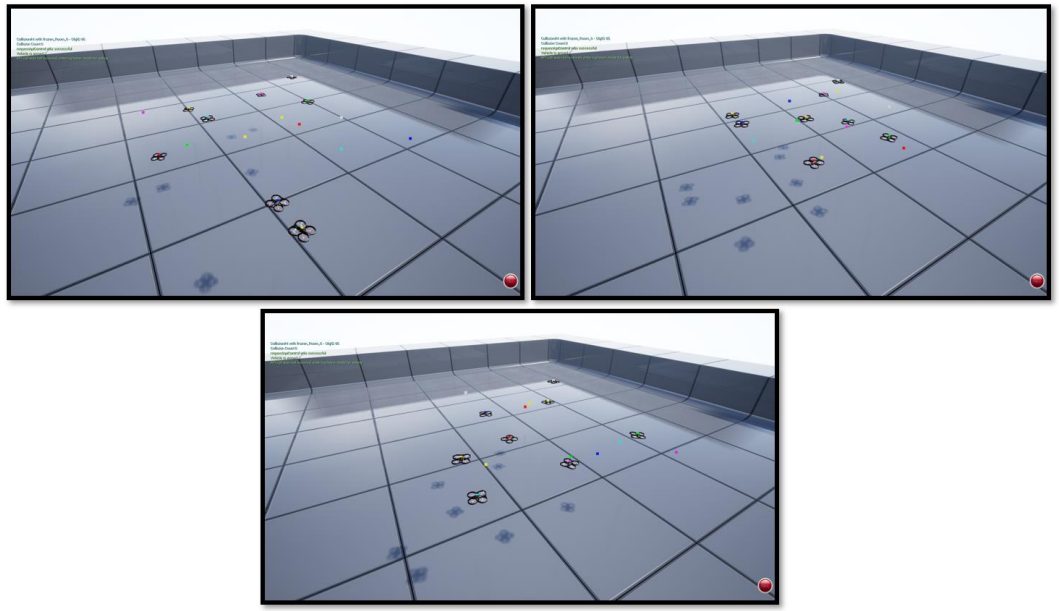


图4-2 随机场景示意图（左：稀疏场景，中：正常场景，右：密集场景）

圆形场景：每台无人机的初始位置都均匀地设置在处于同一高度、同一半径的圆形区域上。而目标位置设置在圆形区域的另一侧，通过增加无人机的数量来调整圆形场景的复杂度。

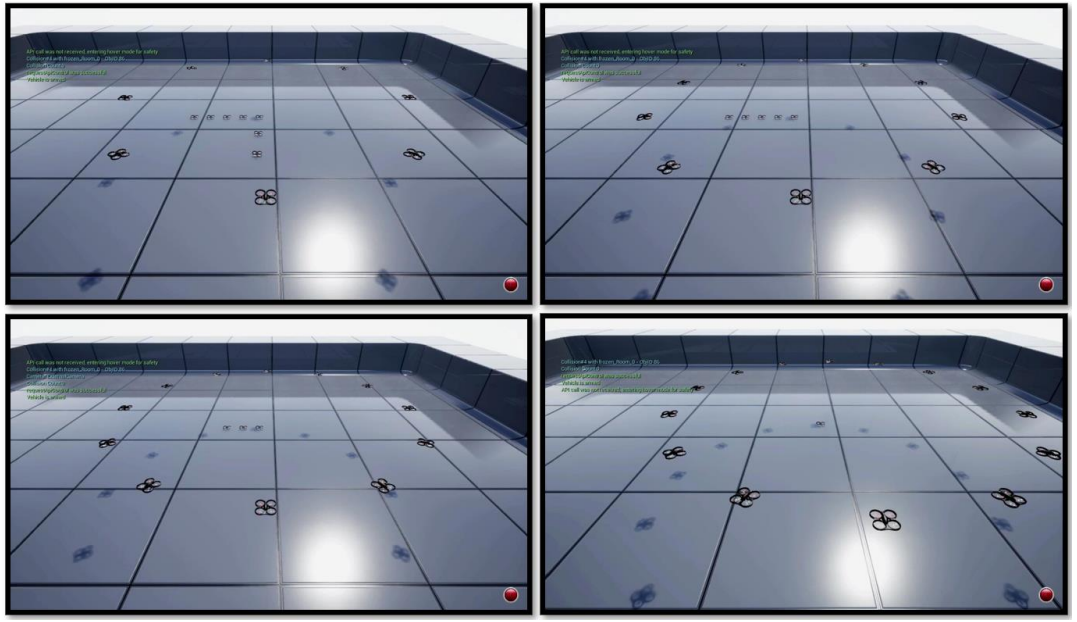


图4-3 圆形场景示意图（半径为12 m，无人机数量为8、10、12、14架）

# 实验设置

我们使用了以下实验指标来衡量不同方法的避障导航性能：

- 成功率：在有限的时间内成功达到目标位置且没有发生任何碰撞的无人机所占的比例。
- 额外距离：与最短路径距离相比，每架无人机所花费的额外移动的距离，单位是米（m）。
- SPL（路径长度加权成功率）：综合考虑了成功率和路径长度。在N架无人机和M轮回合的测试场景中其计算过程如下：

$$\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}S_{i,j}\frac{L_{i,j}}{\max(L_{i,j},l_{i,j})}$$

其中，$L_{i,j}$是第 i 架无人机第 j 轮回合中的最短路径距离，$l_{i,j}$是实际飞行的路径长度，$S_{i,j}$表示是否成功到达目标点，如果成功则值为1，否则值为0。

- 平均速度：每架无人机避障导航过程的平均速度，单位是米每秒（m / s）。

# 实验结果与分析

我们分别在正常密集程度的随机场景上训练了不同的方法，其中对于RL+VAE的方法，我们先是通过采集多张无人机飞行过程中的深度图来训练VAE，之后再进行避障导航策略的训练。

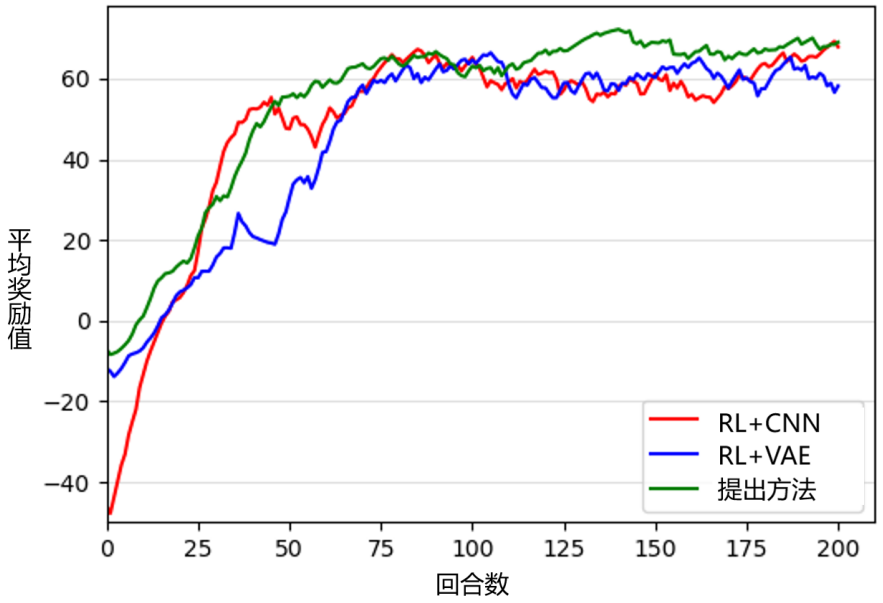



图4-3 训练过程中不同方法平均奖励的比较

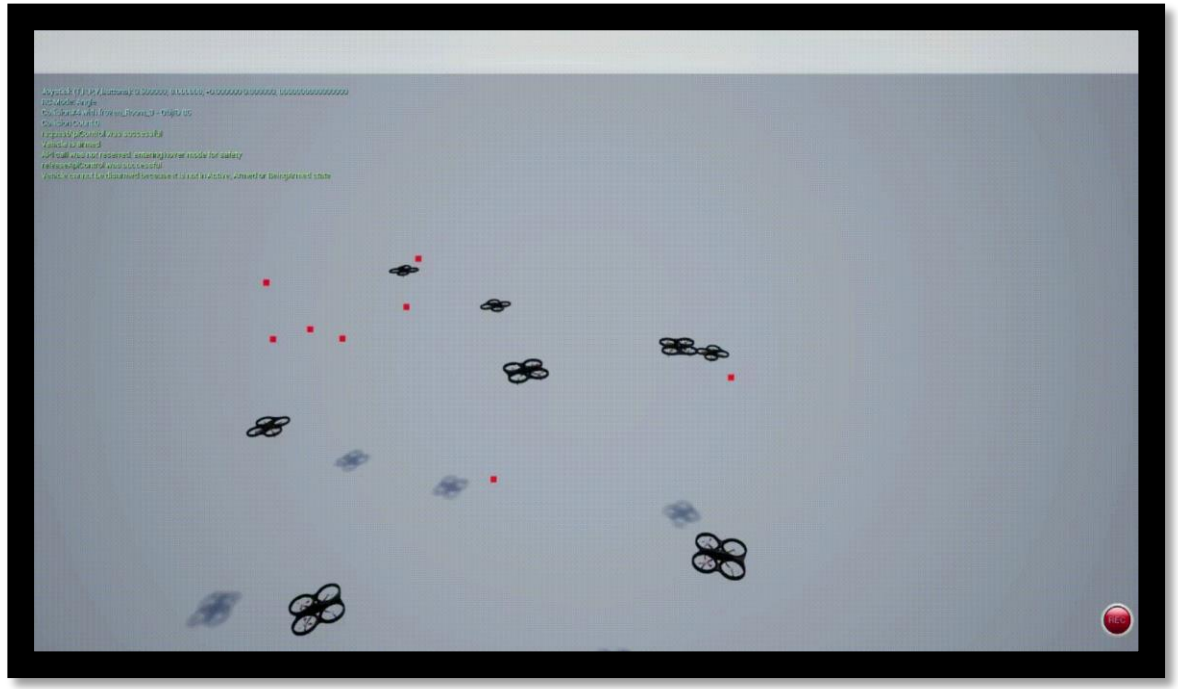


图4-4 训练过程

# 实验结果与分析

测试了1000个回合后不同方法在不同密集程度的随机场景下的性能，结果表明

- 本文提出的方法在性能上表现出极好的一致性。
- 在大多数场景下，本文提出的方法有最高的成功率和SPL。

表4-2不同方法在不同密集程度的随机场景下的性能评估（均值/标准差）

| 场景类型 | 方法 | 成功率 | 额外距离（m） | SPL | 平均速度（m/s） |
|---|---|---|---|---|---|
| 稀疏 | 提出方法 | **0.952** | 1.334/1.202 | **0.852** | **1.066/0.094** |
| | RL+CNN | 0.886 | 1.663/1.133 | 0.774 | 0.773/0.072 |
| | RL+VAE | 0.900 | **0.886/0.536** | 0.833 | 1.063/0.082 |
| 正常 | 提出方法 | **0.934** | 1.784/1.774 | **0.791** | 0.991/0.104 |
| | RL+CNN | 0.829 | 1.499/1.080 | 0.717 | 0.771/0.065 |
| | RL+VAE | 0.851 | **0.777/0.607** | 0.785 | **1.038/0.090** |
| 密集 | 提出方法 | **0.895** | 2.035/2.270 | 0.715 | 0.914/0.130 |
| | RL+CNN | 0.764 | 1.505/1.154 | 0.637 | 0.715/0.070 |
| | RL+VAE | 0.818 | **0.686/0.494** | **0.745** | **0.967/0.099** |

# 实验结果与分析

在圆形场景（半径 = 12m，无人机数量 = 8架）上测试了不同方法，结果表明

- 通过RL+CNN方法训练的避障导航策略无法在圆形场景中适用。
- 本文提出的方法表现出了更好的避障导航性能，除了额外距离稍差于RL+VAE 。


图4-5 通过RL+CNN方法训练的避障导航策略在圆形场景中无人机会在中间点相撞

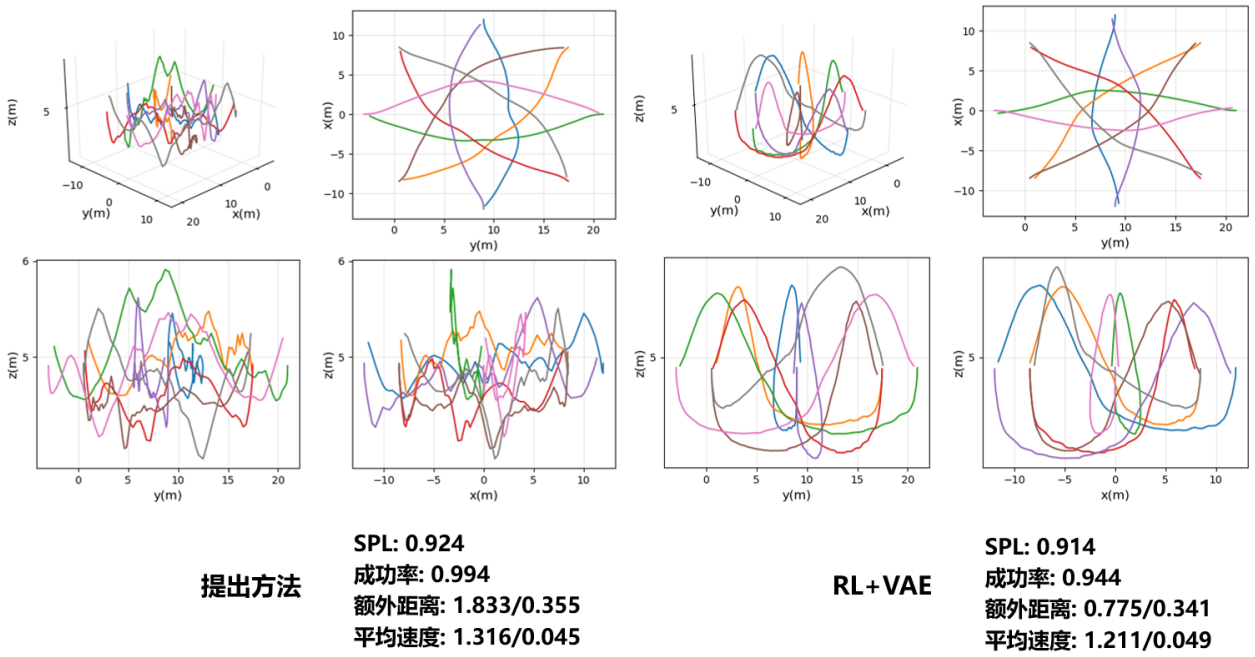



图4-6 在圆形场景下，本文提出的方法和 RL+VAE 的轨迹示意图和性能指标

# 实验结果与分析

在圆形场景（半径 = 12m，无人机数量 = 8, 10, 12, 14架）上测试了不同方法，结果表明随着无人机数量的增加：

- 本文提出的方法的各项性能指标均略有下降，但其中成功率和SPL仍保持在较高水平。
- RL+VAE的成功率和SPL下降明显。

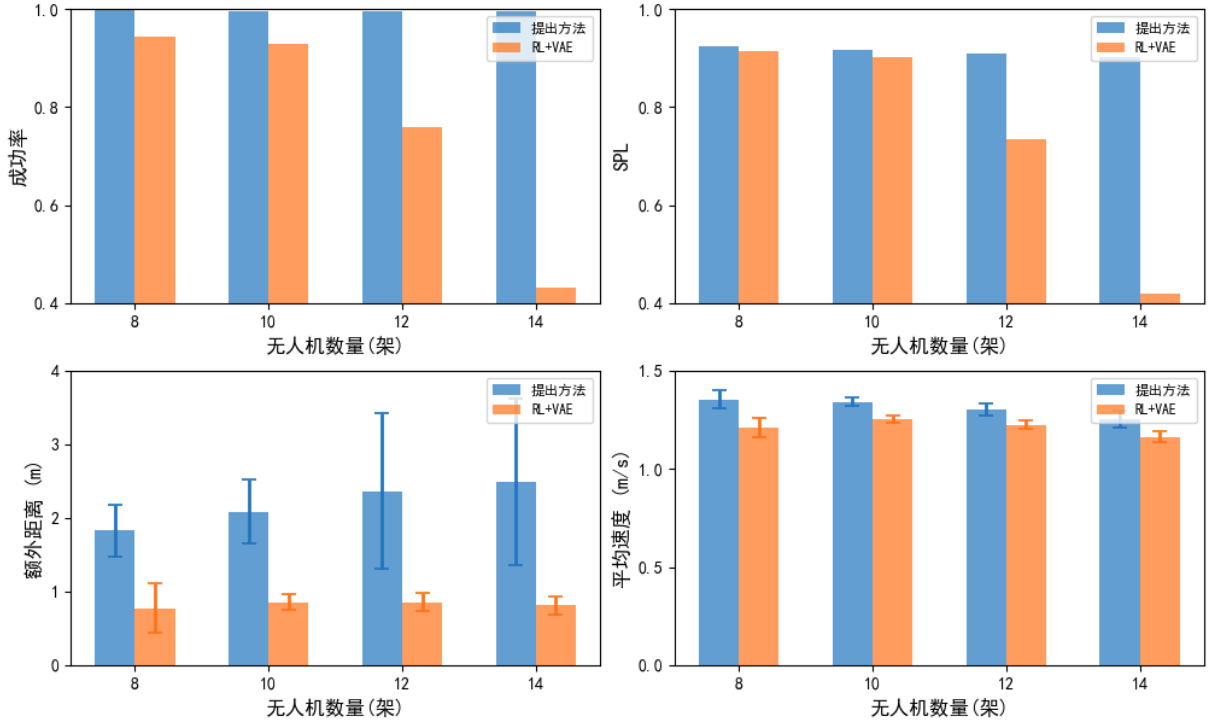



图4-7 在圆形场景不同数量无人机下，本文提出的方法和RL+VAE的性能指标

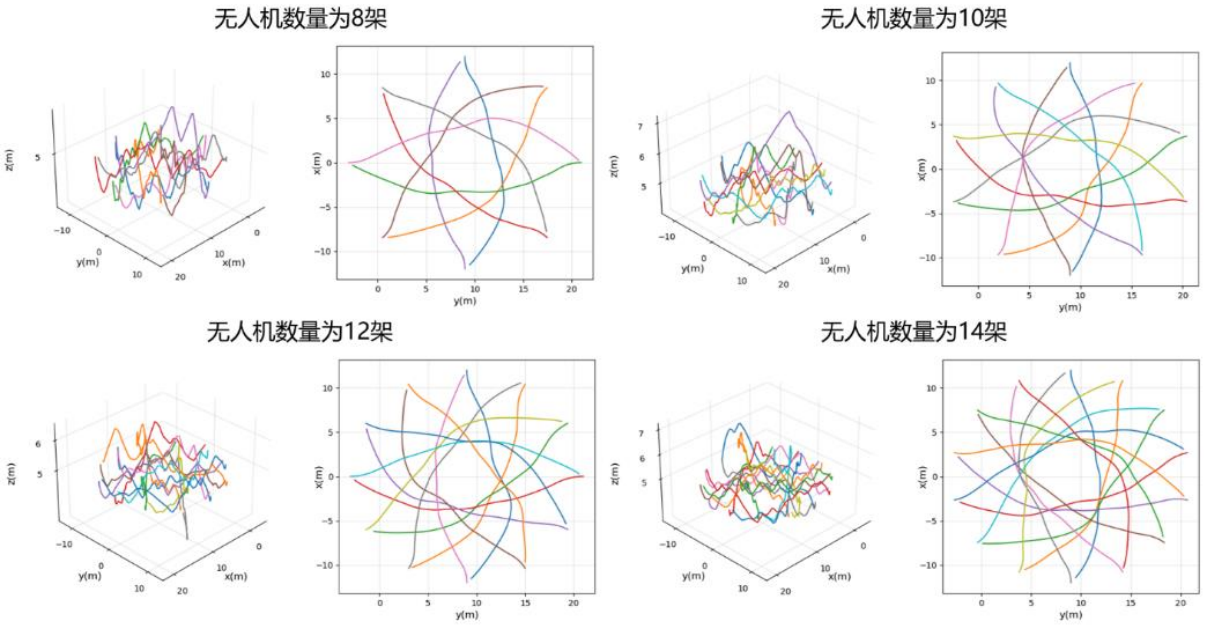
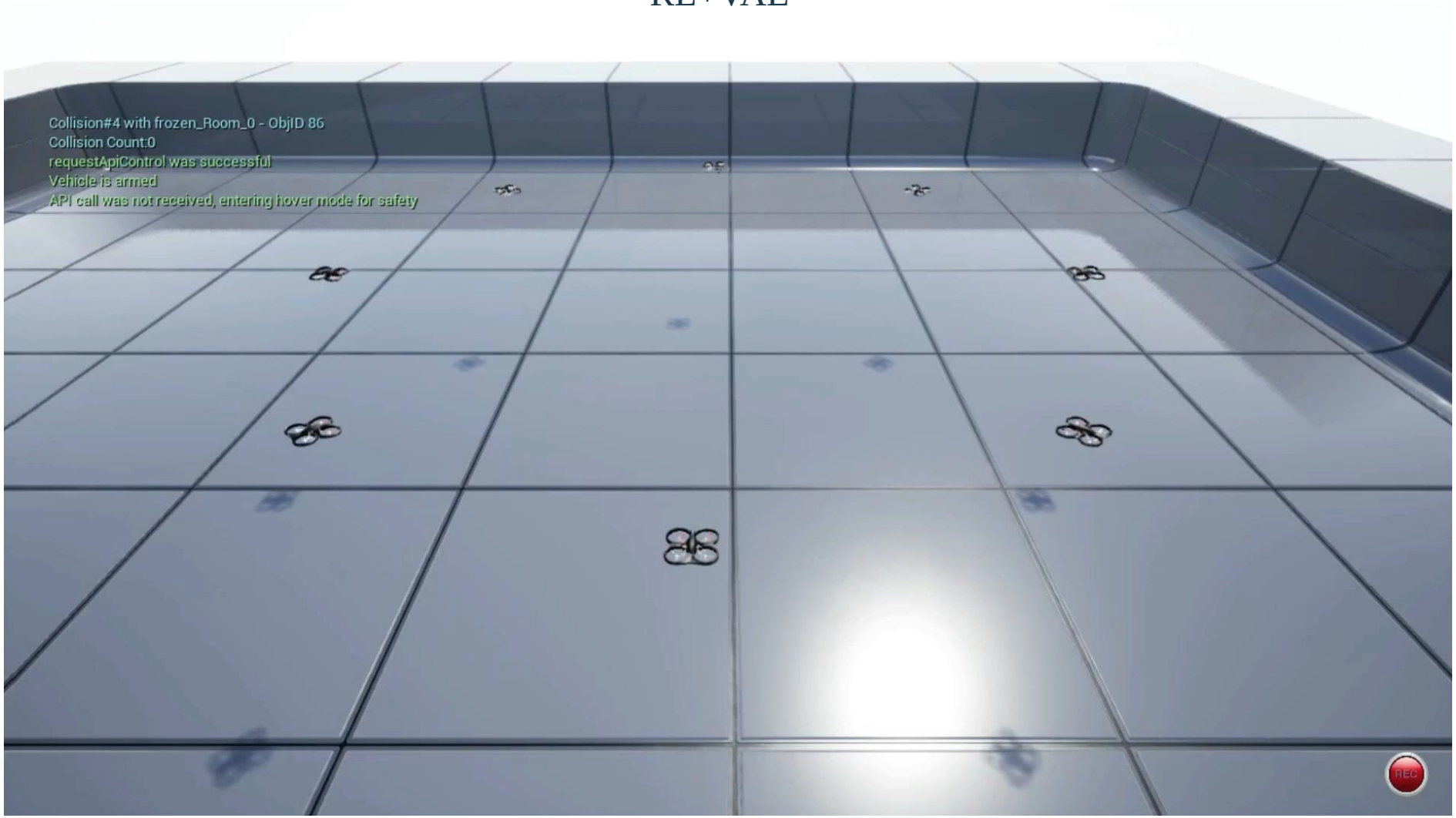



图4-8 在圆形场景不同数量无人机下，本文提出的方法产生的轨迹

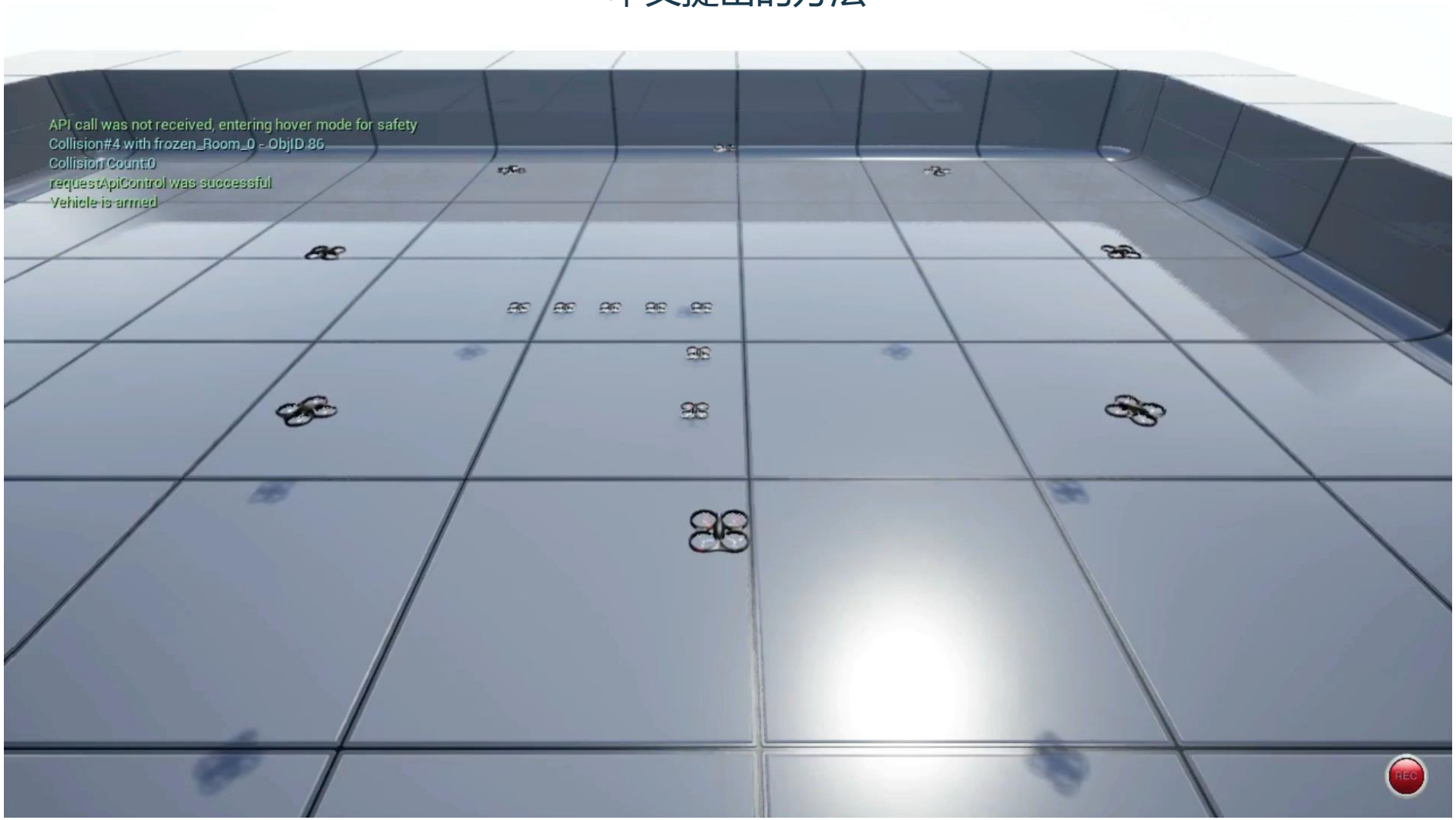
# 结果展示

RL+VAE


Collision#4 with frozen_Room_0 - ObjID 86
Collision Count:0
requestApiControl was successful
Vehicle is armed
API call was not received, entering hover mode for safety
REC

# 结果展示

本文提出的方法



4

# 研究成果

Research Findings

## 总结

- 针对多无人机系统避障导航问题，创新性地提出了一种基于视觉的深度强化学习分布式避障导航策略学习方法，为多无人机协同控制领域提供了一种新的思路和技术。
- 该方法在仿真环境中验证了其在三维工作空间中保证多个无人机的完全自主无碰撞导航的有效性和优越性。并且能够在不同的场景和无人机数量下都保持较好的导航性能。

## 展望

- 目前训练场景不包括任何静态或动态障碍物，我们将在未来的工作中添加它们，以增强无人机对环境的适应能力。
- 探索更多的视觉感知模态，比如部署多个摄像头，光流、单目深度估计等，以增强无人机对环境的感知能力。
- 结合传统的控制策略，比如模型预测控制等，以增强无人机的鲁棒性，为实机实验做准备。

# 攻读硕士学位期间主要的工作成果

## 一、发表的学术论文

[1] Huang H, Zhu G, Fan Z, et al. Vision-based Distributed Multi-UAV Collision Avoidance via Deep Reinforcement Learning for Navigation[C]//2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022: 13745-13752.

[2] C. Wang, D. Wang, M. Gu, H. Huang, Z. Wang, Y. Yuan, X. Zhu, W. Wei, and Z. Fan. Bioinspired Environment Exploration Algorithm in Swarm Based on Levy Flight and Improved Artificial Potential Field[J], Drones, 2022, 6(5): 122.

[3] Cai Y, Zhu G, Huang H, et al. The behavior design of swarm robots based on a simplified gene regulatory network in communication-free environments[C]//International Workshop on Advanced Computational Intelligence and Intelligent Informatics. 2021.

## 二、参与的科研项目

[1] 2022-2025，参与科技创新2030 - “新一代人工智能”重大项目 "因果推理与决策理论模型研究"，项目编号：2021ZD0111502

[2] 2021-2022，参与与汕头市快畅机器人科技有限公司合作开发项目“可编程群体教育机器人研发”

[3] 2020-2022，参与某基础研究项目“共识主动性群体协同机理与环境感知技术研究”

[4] 2018-2020，参与某基础研究项目“基于生物体演化机理的群体智能聚合与涌现研究——群体聚合形态动态转换机理研究”，合同编号：18-163-11-ZT-003-008-02

汕頭大學
SHANTOU UNIVERSITY

# 恳请老师批评指正

THANKS FOR LISTENING

答辩人：黄华兴　导师：范衡教授　电子信息

2023/5/19