基于深度强化学习反无人机拦截器机动控制算法

刘欣皓，刘颜歌，马思齐，周永恒，石永仪

（西安电子科技大学）

负责人：吴建设，联系方式：XXXX，邮箱：XXXX

2024.10

摘 要

随着无人机技术的快速发展，其在军事领域的应用日益广泛，对地防空系统构成了新的挑战。为此，构建一个智能化反无人机作战体系变得尤为重要。本文提出了一种基于深度强化学习(DRL)的无人机智能拦截系统，采用了在连续动作上有较好性能的Twin Delayed Deep Deterministic policy gradient (TD3) 算法来优化智能体的策略。该系统能够自动跟踪并拦截敌方无人机，以达到高精度和快速响应的目的。

针对有过载及速度约束的无人机拦截问题，在本方案使用连续动作进行控制，根据偏角和距离等条件设计了观测，并综合考虑速度、终止条件设计了奖励，将智能体与环境之间的交互建模为马尔可夫决策过程（MDP）。仿真结果表明，在低速环境下，本方案实现了对两组100个目标100%的拦截率，并取得了约37秒的平均拦截时间和18米的平均拦截距离；在高速环境下，对于两组12个目标同样达到了100%的拦截率，取得了约33秒的平均拦截时间和49米的平均拦截距离。

本方案提出的基于深度强化学习的无人机智能拦截算法有一定应用前景，不仅为智能化反无人机作战提供了技术支持，也为其他需要高效决策和精准控制的应用场景提出了思路。

关 键 词：深度强化学习，反无人机，TD3算法，智能拦截系统

目录

**[基于深度强化学习反无人机拦截器机动控制算法 1](#_Toc24850)**

**[摘 要 2](#_Toc22691)**

**[1.概述 4](#_Toc20415)**

[1.1.研究背景和需求分析 4](#_Toc17010)

[1.2.研究现状 4](#_Toc19956)

**[2.方案描述 5](#_Toc32142)**

[2.1.强化学习基本定义 5](#_Toc16782)

[2.2 深度强化学习概述 7](#_Toc24980)

[2.3 TD3算法 7](#_Toc20514)

[2.4 智能无人机拦截算法框架 8](#_Toc8324)

[2.5 仿真环境设置 9](#_Toc874)

**[3.仿真验证情况 13](#_Toc9736)**

[3.1 低速环境仿真 13](#_Toc17904)

[3.2 高速环境仿真 15](#_Toc20957)

**[4.创新性分析 16](#_Toc8793)**

**[5.应用前景分析 16](#_Toc28297)**

**[6.总结和展望 17](#_Toc16830)**

**[参考文献 17](#_Toc3782)**

# 1.概述

## 1.1.研究背景和需求分析

无人机因其无需人类驾驶员操控，具备低成本、长续航能力以及良好的隐身特性等优势，在近年来随着计算机技术、通信技术和传感器技术的迅猛发展，已在军事领域中找到了广泛的应用场景[1]。尤其在执行对地防空系统作战及空对空中任务时，无人机展现出了极高的威胁性和强大的打击效能。因此，构建一套智能化反无人机作战体系对于加强国防安全而言至关重要。

本方案旨在探索并实现一种能够自主决策的拦截系统，使其自动跟踪敌方无人机目标并完成拦截任务，最终达到尽可能高的拦截精度和尽可能短的拦截时间。通过这样的技术创新，我们期望能够显著提升反无人机作战效率，为维护国家安全提供强有力的技术支持。

## 1.2.研究现状

针对无人机或拦截器的机动决策问题，当前的研究方法主要包括传统控制技术、启发式优化算法以及人工智能算法。

传统的控制方法如PID控制器，是一种广泛应用于反馈控制系统中的经典方法。通过精细调节比例、积分和微分三个关键参数，PID控制器能够有效管理无人机的飞行姿态及稳定性。然而，这类控制方法往往依赖于丰富的专家知识进行参数调优，并且在复杂多变的任务环境中可能难以达到最佳性能。

启发式优化算法主要包括遗传算法、粒子群算法等，提供了一种探索最优解的新途径。邓可等人[2]采用遗传算法对无人机的行动策略进行了优化设计，展示了该类方法在搜索全局最优解方面的潜力。不过，此类算法通常伴随着较高的计算成本与较长的求解周期，限制了它们在实时性要求高的应用场景中的适用范围。

人工智能算法主要包括深度强化学习、专家系统等，在解决机动决策问题上展现了巨大潜力。王昱等人[3]利用深度强化学习算法DDQN[4]研究了无人机在空战场景中的自主决策能力，取得了显著成果。相较于传统手段，该类算法较为先进，控制决策时只涉及推理，求解速度较快；但同时也面临着数据需求量大、训练过程耗时长以及可解释性较低等问题。

# 2.方案描述

## 2.1.强化学习基本定义

强化学习是一个多领域交叉的学科，包括了计算机科学、最优控制等学科。为了描述智能体(agent)和环境(environment)的交互过程，有如下概念的定义：

## 2.1.1 动作、观测与状态

状态(state)用于描述智能体某一时刻所处环境的全部信息，可以是离散的也可以是连续的，比如在无人机控制环境中，状态可以是敌我无人机的位置，速度，过载等信息。环境中所有可能状态的集合称为状态空间(state space)，用表示。

观测(observation)则指的是智能体能从环境中感知到的部分信息，是环境某一状态的局部表现。比如在控制无人机时，观测可以包含敌方相对我方的偏角，但不包含双方的坐标信息。在每个状态，智能体根据所接收到的观测，可以选择执行不同的动作(action)，所有可能的动作集合称为动作空间(action space)，用表示。

从一个状态过渡到另一个状态的过程称为状态转移(state transition)。用表示从状态s执行动作到达状态s'的概率。对于无人机拦截的环境，由于无人机的运动轨迹由动力学方程决定，当时，状态转移是确定性的。

### 2.1.2 策略和奖励

策略(policy)决定了智能体在给定状态下采取何种动作，用表示，代表在状态s下选择动作a的概率。强化学习中智能体的目标便是学习一个最优策略，通常用累积奖励来衡量策略的优劣。

根据策略执行动作后，智能体会得到一个奖励(reward)，用表示，代表在状态s选取动作a时得到的奖励。奖励是智能体在环境中得到的即时反馈信号，其设置的合理性对智能体的学习过程至关重要。正奖励激励智能体采取相应的动作，负奖励则相反。

智能体根据策略执行动作会生成一条轨迹(trajectory)，轨迹是一条状态-动作-回报链。轨迹上所有奖励的和称为回报(return)[5]，用G表示。回报由即时奖励和未来奖励组成，由于轨迹可能是无限长的，回报也可能无限大，为此需引入折扣因子 (discount rate)。折扣回报的公式如式(2- 1)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 1) |

其中折扣因子，用于平衡即时奖励和未来奖励的重要性，使智能体短视或者远视。

当智能体与环境的交互到达设定的时间上限或进入终止状态，此时得到的一条轨迹称为一集(episode)。完成一集的交互后，智能体会回到初始状态进行下一集的交互。

### 2.1.3 马尔可夫决策过程

马尔可夫决策过程(Markov decision processes , MDP)用于描述随机动态系统，模拟环境中智能体的随机性策略与回报。MDP是状态空间S，动作空间A，状态转移概率P，奖励期望R的四元组(S,A,P,R)。即智能体在状态s，执行动作a，得到奖励r并根据状态转移概率进入新状态s'的决策过程。

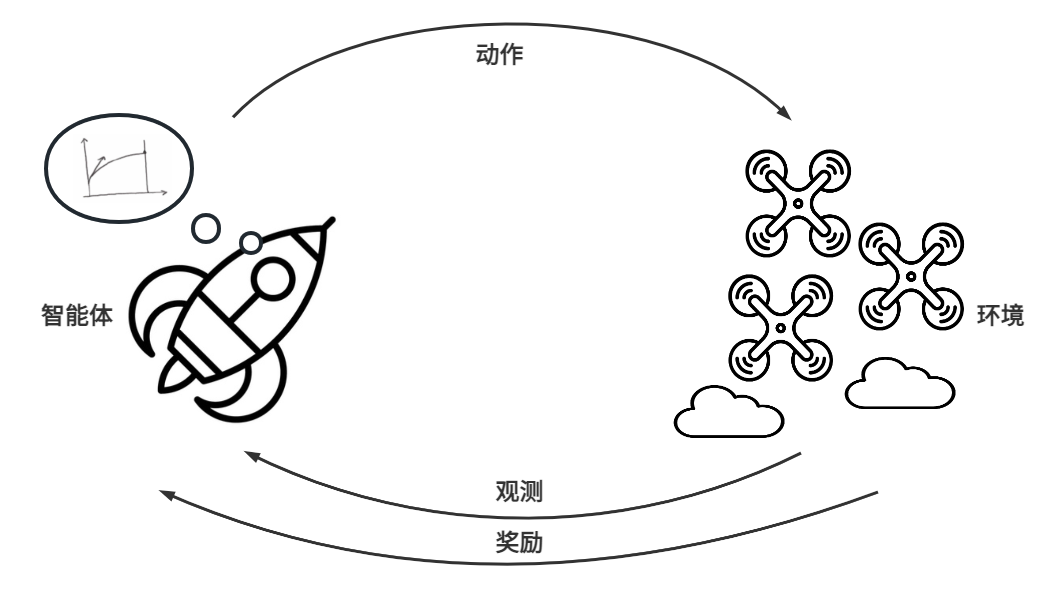


图2.1 强化学习中智能体和环境的交互过程

MDP具有马尔可夫性质，即未来状态仅依赖于当前状态和当前采取的动作，而不受之前的历史影响，用公式表示如(2- 2)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 2) |

## 2.2 深度强化学习概述

传统的强化学习算法使用表格的形式储存状态值以及策略，在应对复杂任务时有很大局限性。深度强化学习(Deep Reinforcement Learning , DRL)用神经网络来拟合函数，分为值函数近似(Value Function Approximation)和策略函数近似(Policy Function Approximation)，通常用表示近似的值函数，用表示近似的策略函数，其中和是神经网络的参数。在将值函数或策略参数化后，深度强化学习算法通过梯度下降法来优化参数以得到最优策略。

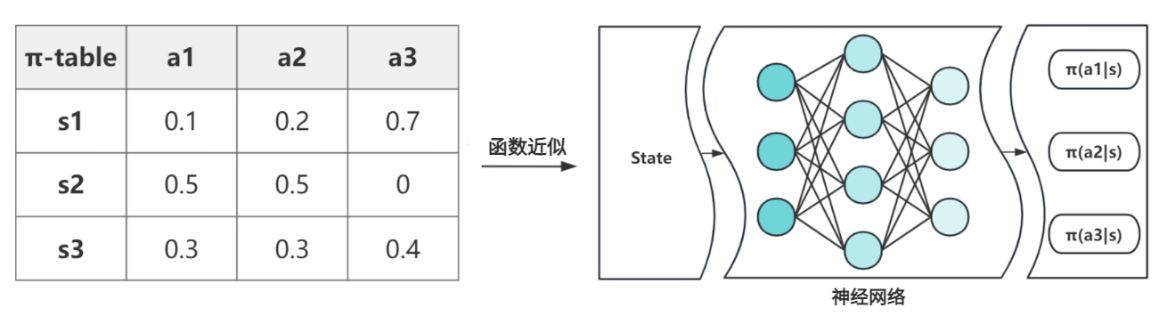


图2.2 DRL中的函数近似

在数据的使用上，传统强化学习算法如SARSA得到一个经验后会立即更新策略并用新策略产生新的经验。而在深度强化学习中，通常使用经验回放(Experience Replay)等技巧来重复利用经验，提高数据的利用率。

## 2.3 TD3算法

Twin Delayed Deep Deterministic policy gradient（TD3）[6]是由Scott Fujimoto等人于2018年在Deep Deterministic Policy Gradient（DDPG）[7]算法上改进得到的一种用于解决连续控制问题的异策略深度强化学习算法。相较于DDPG，TD3加入了3个技巧：

1.双重网络:采用了两个Critic网络，计算目标值时取二者中的较小值，从而抑制Q值过估计的问题。

2.目标策略平滑：更新策略时为目标动作添加噪声，使得策略更难利用到Q值过估计的动作。

3.延迟更新：策略网络更新频率比价值网络更低。由于策略网络需要价值网络输出的Q值来更新，价值网络更新稳定后再更新策略网络能使得学习过程更稳定。

TD3算法的伪代码如下:

|  |
| --- |
| **Algorithm 1** TD3 |
| 初始化价值网络和，策略网络  初始化目标网络  初始化经验回放池B  **for** to **do**  选取添加噪声后的动作，，得到经验储存到B中  从B中采样N个经验  计算  计算  更新价值网络参数  **if** **then**  用确定性策略梯度更新：    更新目标网络：    **end if**  **end for** |

## 2.4 智能无人机拦截算法框架

本方案采用深度强化学习中的TD3算法来训练智能体，让智能体调节过载以控制拦截器轨迹。由于智能体控制的是单个拦截器，在面对敌方无人机群时，对于进入射程的n个目标，使用n个相同的智能体控制n个拦截器进行拦截。

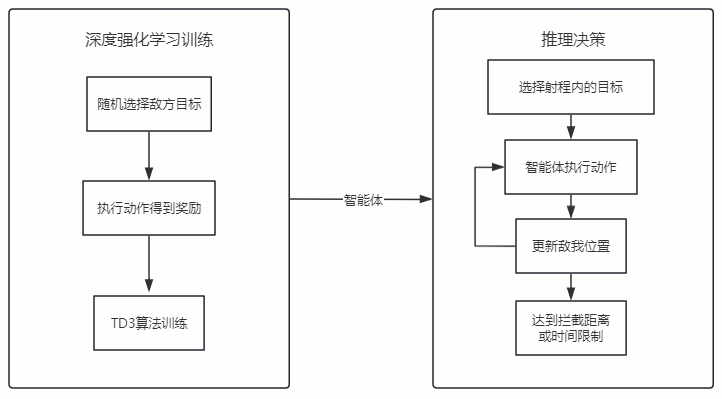


图2.3 智能无人机拦截算法框架

如图所示，算法分为训练和推理两个部分。训练阶段随机选择敌方目标，让智能体与其交互得到奖励信号，使用TD3算法来优化策略；训练完成后，由智能体按照学到的策略控制过载来实现追踪打击处于射程范围内的敌方无人机，根据动力学方程和目标数据来实时更新敌我位置，直到目标被拦截或者达到时间限制后我方拦截器速度降为0。

## 2.5 仿真环境设置

### 2.5.1 拦截器运动模型

敌方无人机的位置已经由数据给出，我方拦截器则需要对运动模型进行建模来模拟其飞行过程。

低速拦截器速度的约束条件如下：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 3) |

低速拦截器过载的约束条件如下，和的约束相同：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 4) |

高速拦截器速度的约束条件如式(2- 5)：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 5) |

高速拦截器过载的约束条件如下，和的约束相同：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 6) |

在设计拦截敌方无人机的拦截器运动模型时，将我方无人机看作质点，忽略空气阻力，高度变化对重力加速度的影响以及飞行过程中油耗对自重的影响。在航迹坐标系内，拦截器的三自由度动力学方程如(2- 7)所示：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 7) |

其中，是拦截器沿航迹坐标系X轴(即飞行方向)的速度，满足约束条件；是拦截器的实时俯仰角(即飞行方向和地面坐标系XOY面的夹角)；是拦截器的实时偏航角(即飞行方向在地面坐标系XOY面的投影与地面坐标系X轴的夹角)；代表重力加速度；代表侧向过载，代表纵向过载，满足约束条件。

再根据如下质心运动学方程，可以计算出拦截器在地面坐标系下的坐标：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 8) |

以上微分方程的求解通过matlab中的ode45函数实现。需要注意的是，由于起始时刻，会导致方程中出现除以0的数值错误，因此模拟时将起始时刻定为0.001秒而不是0秒，从而避免这一问题。

### 2.5.1 动作空间设置

智能体通过调节和来控制拦截器轨迹，因此动作应该决定过载的大小。由于过载的最大值约束随时间变化，不能直接将过载的值定义为动作，故定义动作为二维行向量,。其中和在-1到1的连续区间上取值，作为当前时刻的最大过载系数。设t时刻过载的最大值为，则t时刻过载为和。智能体每0.1秒选择并执行一次动作。

由于拦截器在起始阶段的速度会迅速增加到到最大值，为了避免拦截器早期就严重偏离目标，设定拦截器的初始飞行方向朝向目标的起始位置，且在50个目标的低速环境中，起始的2秒内无论智能体选择什么动作，和分别固定为0和，这样可以确保拦截器保持稳定的飞行方向朝向目标的起始位置。在6个目标高速环境中则是在起始3秒内和分别固定为0和。

### 2.5.2 观测空间设置

为了使智能体能够有效地做出决策并跟踪目标，观测需要提供简洁而有用的信息。本方案使用了拦截器速度方向与敌方相对我方的方位角等信息。用表示目标与拦截器连线与地面坐标系X轴逆时针方向的夹角，是拦截器的实时偏航角，令，表示拦截器速度与敌我连线在水平方向上的偏角，其取值范围为[-180°, -180°]。

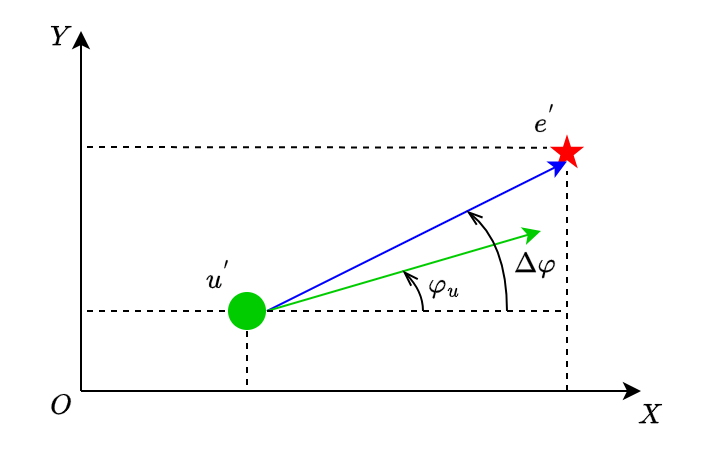


图2.4 XOY面拦截示意图

拦截器和目标在地面坐标系XOY平面的投影点分别是和，以和的连线为轴，平行于轴向上，作为轴。在平面内，目标和拦截器的连线与轴逆时针方向形成的夹角为，是拦截器的实时俯仰角，令，表示拦截器速度与敌我连线在竖直方向上的偏角，其取值为[-180°, -180°]。

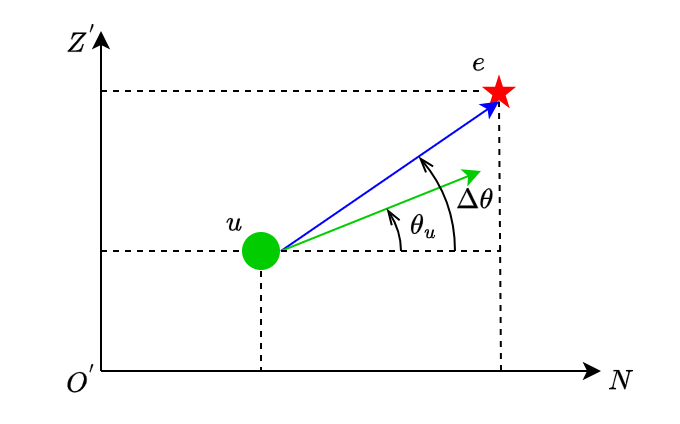


图2.5 面拦截示意图

用表示拦截器与目标的欧式距离，表示拦截器当前速度的大小，最终观测空间定义为4维行向量,。由于过载的约束条件和速度的约束条件成正比，相当于同时包含了二者的信息。注意到观测中各值的大小差异很大，需要进行归一化处理，然而，min-max归一化或者Z-score标准化等方法不能保证和的正负信息保持不变，故将各值除以其取值的最大值来实现归一化，即观测归一化后变为,。

### 2.5.3 奖励函数设置

奖励的设置将影响智能体的行为模式，为了使智能体达到对目标的跟踪效果，在50个低速拦截器的实验中设置了如下的奖励函数，每步起始奖励为0，表示奖励的变化量：

非终止状态：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 9) |

终止状态：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 10) |

其中是当前时刻拦截器速度方向和敌我连线方向夹角的余弦值，为上一时刻该夹角的余弦值，余弦值越大，说明夹角越小；和则是当前时刻和上一时刻敌我的欧式距离。当夹角和距离都减小时给予+1的奖励，否则给予-1的惩罚。并且，若夹角保持较小，即时给予额外+2的奖励。

结算完以上两个每步都会固定获得的奖励后，对当前的奖励(也可能为负表示惩罚)按照速度大小进行增幅，即速度大时没有偏离目标的奖励应该更大，偏离目标的惩罚也应该更大[8]，增幅公式如式(2- 11)下：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 11) |

当敌我距离小于期望距离时，再额外给予一个很大的奖励并进入终止状态(该奖励及终止状态的奖励不进行增幅)，并且下轮训练时会减小0.2，逐步从50减到20为止，相当于奖励和得到奖励的难度都有所提高。到达终止状态时会根据最后时刻的敌我距离给予奖励或惩罚，若大于起始最大距离的一半为惩罚，否则为奖励。

对于6个高速拦截器的实验，奖励函数在数值上有些许变化：

非终止状态：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 12) |

终止状态：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 13) |

其中每次减少1，逐步从300减到100。

奖励的增幅公式修改为：

|  |  |  |
| --- | --- | --- |
|  |  | (2- 14) |

### 2.5.4 策略网络和价值网络结构

TD3算法中有一个策略网络和两个价值网络，本方案中高速环境和低速环境均采用同样的网络结构，且两个价值网络的结构相同。

策略网络接收观测作为输入，经过全连接层1输出224个隐藏状态，再经过全连接层2输出224个隐藏状态，全连接层间使用relu激活函数；最后由全连接层3输出2维向量代表动作的概率分布，由tanh激活函数限制在-1到1之间。

价值网络的输入有两条路径，全连接层1接收观测作为输入，全连接层2接收动作作为输入，分别5输出224个隐藏状态，拼接成448维的向量输入到全连接层3并输出224个隐藏状态，再由全连接层4输出单个值代表预测的Q值，全连接层间使用relu激活函数。

### 2.5.5 深度强化学习超参数

在完成动作、观测、奖励以及网络结构的定义后，就可以使用TD3算法对智能体进行训练。

对于低速环境，设定41秒后进入终止状态，每次到达终止状态后随机选择下一个目标进行训练，一共训练200轮，批次大小为20，折扣因子为0.99，目标更新平滑因子为0.003，动作噪声的上下限为±0.3，策略网络和价值网络的学习率都为0.0002。

对于高速环境，设定34秒后进入终止状态，每次到达终止状态后随机选择下一个目标进行训练，一共训练1000轮，批次大小为30，折扣因子为0.995，目标更新平滑因子为0.002，动作噪声的上下限为±0.2，策略网络学习率为0.0001，价值网络为0.0002。

训练过程中的日志如下所示，平均奖励呈现上升趋势。

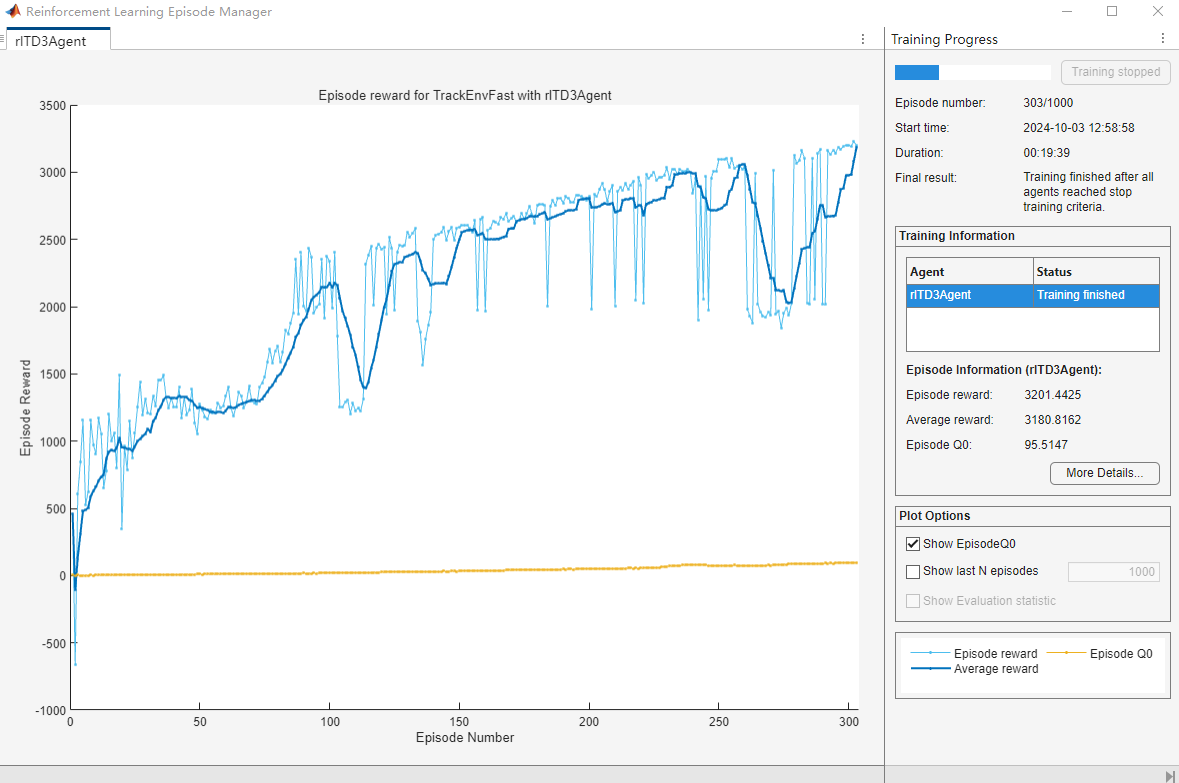


图2.6 使用TD3算法对智能体进行训练的日志

# 3.仿真验证情况

## 3.1 低速环境仿真

我方拦截器全部位于地面坐标系坐标原点，在t=0时刻所有50个目标已进入射程，故将50个拦截器交给50个相同的智能体，控制轨迹打击所有目标。

智能体仅由“wrj50jd.dat”这组数据训练而成。设定小于20米为拦截成功，拦截率、平均拦截时间(敌方目标从起飞到被拦截所耗的时间)和平均拦截距离(拦截成功时敌我距离)如表3.1所示：

表3.1 低速目标的拦截结果

|  |  |  |  |
| --- | --- | --- | --- |
| 目标数据 | 拦截率 | 平均拦截时间/s | 平均拦截距离/m |
| wrj50jd.dat | 100% | 37.14 | 17.79 |
| wrj50pf.dat | 100% | 37.16 | 18.32 |

其中几个目标的拦截轨迹如下图所示：

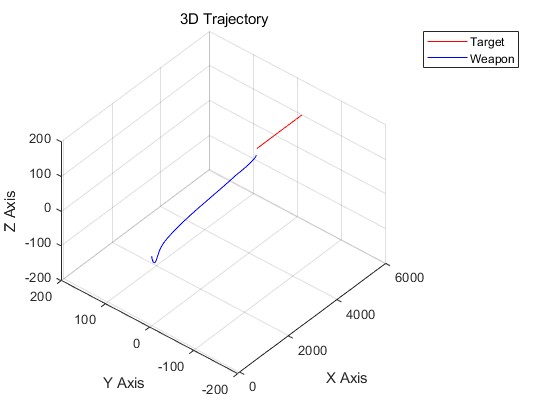


图3.1 智能体对低速平飞目标进行拦截结果图

智能体对该低速平飞目标进行拦截，拦截距离为18.35米。

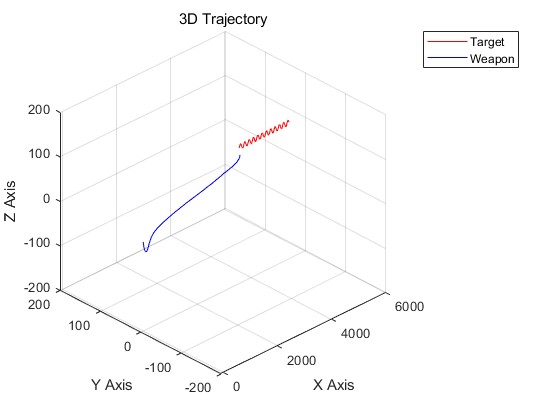


图3.2 智能体对低速平飞加机动目标进行拦截结果图

智能体对该低速平飞加机动目标进行拦截，拦截距离为16.78米。

## 3.2 高速环境仿真

我方拦截器全部位于地面坐标系坐标原点，在t=0时刻所有6个目标已进入射程，故将6个拦截器交给6个相同的智能体，控制轨迹打击所有目标。

智能体由“wrj6jd.dat”和“wrj6pf.dat”这两组数据训练而成。设定小于100米为拦截成功，拦截率、平均拦截时间和平均拦截距离如下表所示：

表3.2 高速目标的拦截结果

|  |  |  |  |
| --- | --- | --- | --- |
| 目标数据 | 拦截率 | 平均拦截时间/s | 平均拦截距离/m |
| wrj6jd.dat | 100% | 32.95 | 50.94 |
| wrj6pf.dat | 100% | 32.95 | 47.03 |

其中几个目标的拦截轨迹如下图所示：

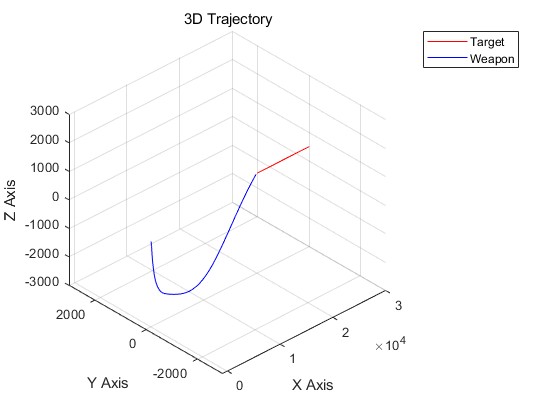


图3.3 智能体对高速平飞目标进行拦截结果图

智能体对该高速平飞目标进行拦截，拦截距离为37.92米。

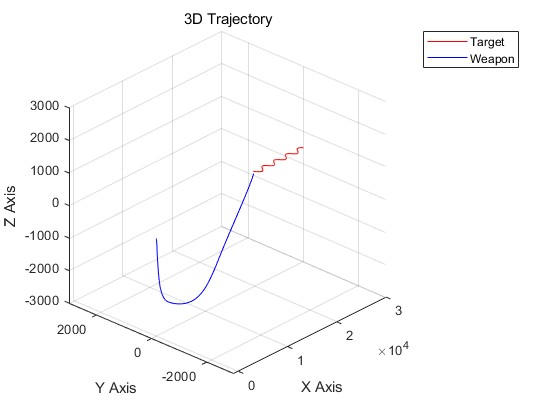


图3.4 智能体对高速平飞加机动目标进行拦截结果图

智能体对该高速平飞目标进行拦截，拦截距离为55.69米。

# 4.创新性分析

本方案使用深度强化学习解决智能体控制拦截器追踪目标的问题，相较于传统的PID控制，DRL能够在与环境的交互过程中，根据奖励信号而非专家经验自动调整其行为策略，从而在复杂多变的环境中展现出色的表现。低速环境的实验也表明，训练好的智能体具有较好的泛化性能，可以自动控制拦截器飞行轨迹，应对不同来袭情况的敌方无人机群。

在观测的设计，本方案选取了水平差角、垂直差角、距离和速度四个信息，简洁有效地描述了拦截器和敌方目标的关系，并且不需要特征工程，使用神经网络自动从原始观测提取特征

在奖励的设置上，本方案考虑了多种情况下的奖励，以引导智能体学习跟踪的策略。非终止状态下，同时结合角度和距离进行奖惩，并根据速度进行增幅；达到终止状态时，根据拦截距离给出奖惩，并缩短下一轮的预期拦截距离，以鼓励拦截器不断接近目标。

# 5.应用前景分析

本方案的拦截器是个抽象的理想质点模型，在实际应用中可以是无人机、导弹等多种飞行器或汽车之类的载具。

在军事与国防应用方面，可以用类似的方法对导弹的航迹进行模拟，以达到更好的跟踪和性能。在航空航天领域，本方案可以应用于飞行器路径规划的研究之中，通过对环境添加更多约束条件并改变奖励函数，可实现路径规划及避障等功能。

在汽车自动驾驶领域，本方案也起到了一定启发作用。虽然自动驾驶不需要控制z轴，且与飞行控制有相似的逻辑，但往往有更多的约束和规则限制，可将基于深度强化学习的控制和模仿学习[9]、专家系统相结合，实现更智能的驾驶。

在工厂或仓库等需要机器人跟踪物体或追踪路径移动的环境，本方案可以帮助开发者模拟测试机器人的导航逻辑，确保它们能够在复杂的工业环境中可靠地运作。

# 6.总结和展望

近年来，深度强化学习技术得到快速发展，被广泛地应用于各类智能决策问题。本方案对三维环境中的无人机分配及拦截问题进行了研究，将拦截器和目标的交互过程建模为马尔可夫决策过程，使用深度强化学习算法优化智能体的策略。

方案首先对飞行模型进行建模，利用动力学方程对拦截器的飞行轨迹进行仿真，实时更新敌我位置。然后定义了动作、观测以及奖励，引导智能体学习追踪策略。最终由TD3算法训练好的智能体能成功追踪目标，并取得较高的拦截精度和较短的拦截时间。

本方案虽在模拟环境中取得了较好的性能，但在实际应用中仍存在一些问题需要改进。方案中没有考虑拦截器之间飞行轨迹的相互干扰和影响。并且针对高速环境和低速环境各训练了一个智能体，能否统一多种情况，提高模型的泛化性能，使得单个智能体也能面对各种场景设置是一大挑战。其次，在一些规则较复杂的环境中，如何将专家经验和规则融入决策，实现人机共同决策的混合智能系统，也是亟待解决的问题。

参考文献

[1]孙韬, 王俊达, 王睦深, 等. 集群式侦察探测防御系统[J]. 现代防御技术, 2020, 48(03): 92-98.

[2]邓可, 彭宣淇, 周德云. 基于矩阵对策与遗传算法的无人机空战决策[J]. 火力与指挥控制, 2019, 44(12): 61-66,71.

[3]王昱, 任田君, 范子琳. 基于引导 Minimax-DDQN 的无人机空战机动决策[J/OL]. 计算机应用: 1-10.

[4]Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[5]Zhao S. Mathematical Foundations of Reinforcement Learning[M]. Berlin, Germany:Springer,2024.

[6]Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International conference on machine learning. PMLR, 2018: 1587-1596.

[7]Lillicrap T P. Continuous control with deep reinforcement learning[J].arXiv preprint arXiv:1509.02971, 2015.

[8]Bakos S, Davoudi H. Mitigating Cowardice for Reinforcement Learning Agents in Combat Scenarios[C]//2022 IEEE Conference on Games (CoG). IEEE, 2022: 377-384.

[9]Hussein A, Gaber M M, Elyan E, etal. Imitation learning: A survey of learning methods[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-35.