文章编号:1671 - 4598(2025)10 - 0165 - 09 DOI:10. 16526/j. cnki. 11-4762/tp. 2025. 10. 022 中图分类号:TP393 文献标识码:A

基于 PPO 算法的一对一空战格斗决策方法

周琪栋，江志东，霍立平，赵冬梅
（海军航空大学青岛校区，山东 青岛 266041）

摘要：空战格斗具有作战要素多、态势变化快和作战氛围紧张等特点，其决策方法是人工智能领域的热点研究课题；目前关于近距空战算法的研究大都在简化的低精度场景或现有仿真平台中进行，受实际问题的复杂性和仿真效能的限制大多简化了空战决策模型，降低了研究结果的参考价值；针对此问题，基于 Unity3D 搭建了满足研究需求的可视化空战平台并设计了飞机的机动动作集，根据空空格斗时的敌我态势特点定义了态势评估函数和奖励函数，在此基础上构建了基于近端策略优化算法的一对一空战格斗决策框架；实验结果表明，决策模型能够驱动智能体根据战场态势进行灵活的机动决策，具备较强的自主决策的能力，验证了方法的有效性。

关键词：人工智能；空战格斗；强化学习；近端策略优化；空战决策

1V1 Close-range Air Combat Maneuvering Decision-Making Method Based on the PPO Algorithm

ZHOU Qidong，JIANG Zhidong，HUO Liping，ZHAO Dongmei
（Qingdao Branch，Navy Aviation University，Qingdao 266041，China）

Abstract：Abstract：The close-range air combat has the characteristics of multiple combat elements，rapid situational changes，and tense combat atmosphere，it's decision-making method is a hot research topic in the field of artificial intelligence. At present，research on close-range air combat algorithms is mostly conducted in simplified low precision scenarios or existing simulation systems，due to the complexity of practical problems and limitations in simulation effectiveness，most simply air combat decision models are simplified，which reduces the reference value of research results. In response to this issue，a visual air combat platform based on the Unity3D is built to meet research requirements and design a aircraft maneuvering action set. Based on the characteristics of the enemy-friendly situation during the close-range air combat，the situation evaluation and reward functions are defined. On this basis，a one-on-one close-range air combat decision-making framework based on the proximal policy optimization (PPO) algorithm is constructed. Experimental results show that the decision model can drive the intelligent agent to make a flexible maneuvering decision based on a battlefield situation，and has a strong autonomous decision-making ability，which verifies the effectiveness of the method.

Keywords：artificial intelligence；close-range air combat；reinforcement learning；PPO；air-combat decision

0 引言

空战智能决策是模拟作战飞行员在各种空战态势下对飞行器操纵的决策，是智能作战飞行器的“灵魂”和“大脑”[1-2]。近距空战中，飞行器需要做大量战术机动以规避敌机并构成武器发射条件，因此智能决策是近距空战需要研究的关键问题[3]。

深度强化学习等智能算法对于需要快速处理海量高维信息的空战决策来说有其独特优势，以其为核心控制算法的空战智能决策逐渐被广泛应用，大致可分为：基于值函数的方法、基于确定性策略梯度方法及基于随机性策略梯度的方法等[4]。

基于值函数的方法主要是将 Q-learning[5]、深度 Q 网络（DQN，deep q-learning network）[6]及其变种算法

收稿日期:2024 - 09 - 21； **修回日期**:2024 - 10 - 31。
基金项目:中国人民解放军海军航空大学基金(H3202204022)。
作者简介:周琪栋(1989 -),男,硕士,助教。
通讯作者:江志东(1985 -),男,博士,讲师。
引用格式:周琪栋,江志东,霍立平,等. 基于 PPO 算法的一对一空战格斗决策方法[J]. 计算机测量与控制,2025,33(10):165 - 173.

应用在智能决策中。文献［7］基于 DQN 算法框架，对奖励函数及超参数设置进行了优化研究，训练智能体学习较优的机动策略。文献［8］通过加入启发式因子和双 Q 表交替学习机制，改进了传统 Q-learning 算法，并设计了动态栅格规划环境，构建了侧向机动决策算法。基于确定性策略梯度方法的研究主要集中在深度确定性策略梯度（DDPG，deep deterministic policy gradient）算法[9]及其变种算法的应用上。文献［10］基于 DDPG 算法，增加飞行高度上下限、飞行过载以及飞行速度上下限，通过全连接的载机速度控制网络与环境奖励网络。文献［11］将 DQN 与 DDPG 算法相结合，使用 DQN 生成动作库指令，通过概率神经网络预测目标机动指令，将预测结果使用在 DDPG 上，使无人机能够自主进行决策。文献［12］基于双延迟确定性策略梯度算法（TD3，twin delayed deep deterministic policy gradient），采用奖励函数与样本优先度排序方法，对智能体进行仿真博弈训练。文献［13］通过结合成熟的飞行控制技术，发展出基于航迹导引指令的空战机动决策与控制方案，并且设计了机动决策与飞行控制分离的硬件架构进行策略分析。基于随机性策略梯度方法的研究主要集中在近端策略优化（PPO，proximal policy optimization）算法[14]及其变种算法的应用上。将长短时记忆网络（LSTM）与 PPO 结合[15-16]增强对样本数据的学习能力，对状态进行特征提取和态势感知，加入随机噪声提高智能体对未知状态空间的探索能力。对 PPO 的奖励函数与优势函数进行重塑和优化[17-18]，可进一步改进 PPO 算法的性能，加快收敛。文献［19］在 PPO 算法基础上，将动作空间进行分层，采用自回归结构，使用多维离散的动作空间。文献［20］提出了机动决策分层框架下的基于双重奖励的 PPO 优化算法，将空战任务分为决策与控制两个子问题，决策层使用 PPO 算法，控制层使用比例积分微分算法，将高层决策转换并输出原始控制指令。文献［21－23］分别将门控循环单元融合前序态势信息、自博弈、option 等引入 PPO 算法，合理设计奖励函数进行空战决策。

在上述智能空战模型使用的诸多算法中，PPO 算法是一种基于策略的强化学习算法，利用旧策略的决策轨迹进行训练，同时通过限制策略更新的幅度，使得学习过程更加稳定。与其他算法相比，PPO 算法具备以下优势：1）通过限制新旧策略的更新步长，放缓策略变化率，解决了策略梯度算法普遍存在的步长选择困难问题；2）通过算法参数的更新方式，确保在训练过程中值函数单调上升；3）通过重要性采样原理实现策略的离线更新，提高数据的利用率。因此本文以 PPO 算法为基础，合理设计算法超参数和奖励函数，构建智能体训练框架。

在仿真实验方面，目前深度强化学习算法在智能空战决策应用中的仿真大部分都是在高度简化的低精度二维或三维场景中进行的，如 MaCA 平台、Matlab 平台、Simulink 平台等，与实际的空战场景相距较大。有些研究使用现有的仿真系统如墨子联合作战推演系统[24]、WUKONG 空战环境[25]等，受实际问题的复杂性和仿真效能的限制大多简化空战决策过程，降低了研究结果的参考价值。因此本文基于 Unity3D 自主搭建贴合实战仿真需求的一对一近距空战场景作为实验环境。

综上所述，本文针对单机近距空战决策问题，首先基于 Unity3D 搭建满足课题研究需求的可视化空战平台，定义适用于空战格斗的机动动作集，构建基于 PPO 算法的一对一近距空战格斗决策模型，同时针对近距空战问题，设计并优化了态势评估函数和奖励函数，引导智能体在与高水平专家系统对抗中快速收敛。通过对训练过程和对战数据进行评估和分析，验证了态势评估函数和奖励函数的有效性，同时可以看出经过训练的智能体具备较强的一对一空战格斗决策能力。

1 空战格斗仿真平台构建

1.1 空战环境设定

空战场景基于 Unity3D 搭建，设定博弈双方驾驶性能完全相同的两架飞机在相同空战环境下进行一对一空战格斗。

空战仿真坐标系以设定的地面中心点为坐标原点（0，0，0），以米为单位。其中，作战场景东西方向范围200 km，用 x［－100 000，1 000 000］表示；垂直方向限高 20 km，用 y［0，20 000］表示；南北方向范围 200 km，用 z［－100 000，1 000 000］表示。为提高深度强化学习的泛化性，两架飞机的初始位置设定为在一定范围内随机生成。为避免飞机在训练初期出界，初始作战位置设为场景中部，即两机初始坐标（Px、Py、Pz）定义为：$Px\in$［－5 000，5 000］，$Py\in$［7 000，10 000］，$Pz\in$［－5 000，5 000］。

空战场景设定两机武器均为有效射程 1.8 km 的航炮，满足发射条件时自动开火，即若敌机出现在我机机头延长线夹角 1°、距离 1.8 km 范围内，则视为被击中，被击中将产生 20 点/秒血量值损失速度，两机均设定初始 100 点血量值，即目标在飞机朝向射程范围内 5 s 被击落。空战格斗场景如图 1 所示。

1.2 飞机状态空间设计

飞机在 t 时刻的状态表示为 $\boldsymbol{S}_t$，是一个包含 14 个维度的信息的向量：$\boldsymbol{S}_t=\{\boldsymbol{X}_{\text{Position}}, \boldsymbol{Y}_{\text{Position}}, \boldsymbol{Z}_{\text{Position}}, \boldsymbol{X}_{\text{eulerAngles}}, \boldsymbol{Y}_{\text{eulerAngles}}, \boldsymbol{Z}_{\text{eulerAngles}}, \boldsymbol{X}_{\text{attitude}}, \boldsymbol{Y}_{\text{attitude}}, \boldsymbol{Z}_{\text{attitude}}, speed, Acc, Yaw, Pitch, blood\}$，其中各分量含义描

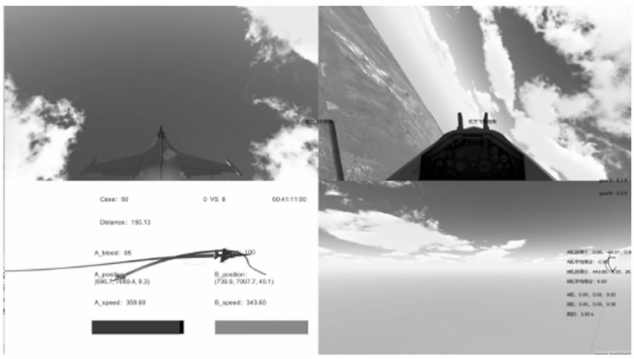

图 1　空战仿真场景

述如下：

1）$\boldsymbol{X}_{\text{Position}}$，$\boldsymbol{Y}_{\text{Position}}$，$\boldsymbol{Z}_{\text{Position}}$：表示 t 时刻飞机在世界坐标系中的三维坐标；

2）$\boldsymbol{X}_{\text{eulerAngles}}$，$\boldsymbol{Y}_{\text{eulerAngles}}$，$\boldsymbol{Z}_{\text{eulerAngles}}$：表示 t 时刻飞机坐标系相对于世界坐标系的欧拉角，该角度反映了此刻飞机在世界坐标系中的飞行姿态；

3）$\boldsymbol{X}_{\text{attitude}}$，$\boldsymbol{Y}_{\text{attitude}}$，$\boldsymbol{Z}_{\text{attitude}}$：表示 t 时刻飞机在世界坐标系下的机头朝向向量；

4）$speed$：表示 t 时刻飞机的飞行速度；

5）Acc，Yaw，$Pitch$：分别表示 t 时刻飞机的加速度、偏航角速度和俯仰角速度；

6）$blood$：表示 t 时刻飞机在本局格斗中的剩余血量。

1.3　飞机动作空间设计

飞机有一个建立在自身形态基础上的航迹坐标系，如图 2 所示。其中 Z 轴指向飞机速度方向，Y 轴垂直于 Z 轴指向飞机机身上方，X 轴垂直于 Y 轴和 Z 轴指向飞机机身右侧。

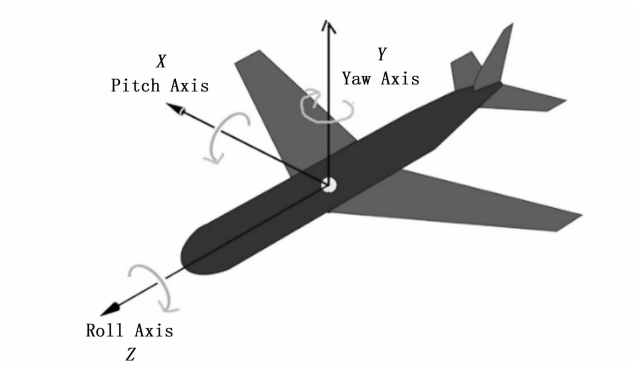

图 2　飞机坐标系

飞机在仿真环境中具有 3 种独立的动作，分别为加速度、偏航角速度和俯仰角速度，可以同时作用于自身坐标系。考虑到飞机机动能力，将 3 个动作维度的取值范围定义如下：

1）加速度 $Acc\in[-30,90]$（m/s^2），表示此刻飞机以区间内某一个加速度朝着自身坐标系 Z 轴方向加速（或减速）向前机动。飞机的速度取值范围参考了国内外先进战机最大飞行速度（F-22 飞机 2.25 马赫、歼-20 飞机 2.5 马赫）以及战斗机失速速度（300 km/h），设为 $speed\in[100,800]$（m/s）；

2）偏航角速度 $Yaw\in[-60,60]$（°/s）：表示此刻飞机以区间内某一个角速度绕自身坐标系 Y 轴旋转；

3）俯仰角速度 $Pitch\in[-30,30]$（°/s）：，表示飞机以区间内某一个角速度绕自身坐标系 X 轴旋转。

1.4　空战态势评估

在空战自主决策中，态势评估值最优的方案会更大概率被使用，因此要保证态势评估的合理性和有效性。本算法中将态势评估划分为两个部分，第一部分为角度优势评估，第二部分为距离优势评估，最后将两部分评估值结合。

1.4.1　角度优势评估

角度优势评估值与敌我双方飞机飞行朝向以及相对位置相关，如图 3 所示。其中，O-XYZ 为我机航迹坐标系，p 为目标的方位角，q 为目标的进入角，V_R 为目标的速度，$speed$ 为我方的速度，R 为两机的相对距离。

t 时刻飞机的角度优势函数 T_a 定义为：

$$T_a=\frac{|p|+|q|}{360}\quad T_a\in[0,1]$$

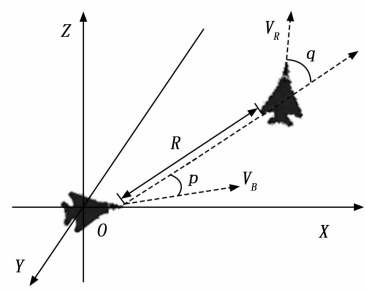

图 3　敌我双方占位态势关系

显然，角度函数 T_a 越小则优势越大，当 $T_a=0$ 时，意味着我机和敌机在同一直线上飞行，我机航炮正对敌机尾部，我机优势最大；相反，$T_a=1$ 时，我机与敌机在同一直线上飞行并且敌机航炮正对我机尾部，我机优势最小。

1.4.2　距离优势评估

敌我飞机之间的距离在不同的角度下评估出的优势截然不同，因此在衡量距离优势时要充分考虑角度优势。t 时刻飞机的距离优势函数 T_r 分析及定义如下。其中，R_m 为航炮射程，R 为两机距离，T_r 越小说明此时的距离优势越大。

1）当我方角度处于优势（$T_a<0.5$），且敌机在射程内（$R\leqslant R_m$）时，两机距离越小则命中目标的概率越

大。该态势下我方距离优势大，取值范围设为［0，1］。距离优势 T_r 定义为：

$$T_r = 1 - \frac{R_m - R}{R_m},\ T_r \in [0,1]$$

2）当我方角度处于优势（$T_a < 0.5$），且敌机在射程外（$R > R_m$）时，两机距离越小则将来命中目标的概率越大。该态势下我方距离优势较大，取值范围设为［1，2］。距离优势 T_r 定义为：

$$T_r = 1 - \frac{R_m - R}{R},\ T_r \in [1,2]$$

3）当我方角度处于均势（$T_a = 0.5$）时，两机距离对距离优势影响小。该态势下我方距离优势 T_r 取距离优劣势的中间值：

$$T_r = 2$$

4）当我方角度处于劣势（$T_a > 0.5$），且在敌机射程外时（$R > R_m$），两机距离越大则被敌机命中的概率越低。该态势下我方距离优势较小，取值范围设为［2，3］。距离优势 T_r 定义为：

$$T_r = 3 + \frac{R_m - R}{R},\ T_r \in [2,3]$$

5）当我方角度处于劣势（$T_a > 0.5$），且在敌机射程内（$R \leqslant R_m$）时，两机距离越大则越容易摆脱敌机攻击。该态势下我方距离优势小，取值范围设为［3，4］。距离优势 T_r 定义为：

$$T_r = 3 + \frac{R_m - R}{R_m},\ T_r \in [3,4]$$

1.4.3 综合优势评估

t 时刻飞机的综合优势应结合飞机的速度优势与距离优势，综合优势评估函数 T 定义为 T_a 与 T_r 的乘积：

$$T = T_a \times T_r,\ T \in [0,4]$$

2 基于 PPO 算法的空战决策实现

2.1 PPO 算法介绍

PPO 算法是一种经典的深度强化学习方法，具有高稳定性、强适用性等特点，是基于 A2C（Advantage Actor Critic）算法和区域策略优化（TRPO，trust region policy optimization）的改进。

A2C 是一种基于策略梯度的强化学习算法，策略梯度项为：

$$\triangle_\theta J(\theta) \sim \left[\sum_{t=0}^{T-1} \log P_{\pi_\theta}(a_t \mid s_t)\right] A(s_t, a_t)$$

式中，π_θ 为策略，$J(\theta)$ 为策略的性能度量，$\triangle_\theta J(\theta)$ 是 $J(\theta)$ 对 θ 的偏微分，$A(s_t, a_t)$ 为优势函数，公式为：

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$$

式中，$A(s_t, a_t)$ 为在状态 s_t 下采取动作 a_t 的平均相对优势。根据贝尔曼方程，可以从价值估计中推出优势函数的估计：

$$A(s_t, a_t) = r_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

式中，γ 为折扣因子，r_{t+1} 表示 $t+1$ 时刻的奖励。

TRPO 算法利用旧策略的动作、状态和优势来训练当前策略。为了提高样本利用率，通过多次重复使用同一批样本来训练当前策略。计算目标函数 $f(x)$ 的期望值的重要性采样方法为：

$$E_{x\sim p(x)}[f(x)] = E_{x\sim q(x)}\left[\frac{px}{qx} f(x)\right]$$

为了解决梯度更新稳定性问题，TRPO 算法使用下一次梯度变化迭代中可采取的最大步长半径来约束梯度更新的幅度，同时根据上一次梯度更新后的性能情况来拓展或缩小边界。PPO 算法将约束集成到目标函数中，目标函数定义为：

$$L^{\mathrm{CLIP}}(\theta) = E_t\{min[r_t(\theta)\ A_t, clip_\varepsilon[r_t(\theta)]A_t]\}$$

$$clip_\varepsilon(x) = clip(x, 1-\varepsilon, 1+\varepsilon)$$

式中，$r_t(\theta) = \pi_\theta(a_t \mid s_t / \pi_(\theta_o ld)(a_t \mid s_t)$ 表示重要性采样权重。

PPO 算法流程描述如下：

Function
初始化策略参数
for episode = 1,2,… do
for actor = 1,2,…,N do
使用策略 $\pi_{\theta_{old}}$ 在环境中连续运行 T 次
计算优势函数 $A_1(s_t, a_t), A_2(s_t, a_t), \cdots, A_T(s_t, a_t)$
end for
使用 minibatch 样本将目标函数 $L^{dip}(\theta)$ 更新 K 次
更新策略网络参数 $\theta_{\mathrm{old}} = \theta$
end for

2.2 算法与空战平台通信

通信模块用于将强化学习算法的输出（智能体作动）传送给空战平台，进行智能体控制，同时从空战平台接收飞机以及环境的状态信息。通信模块分空战平台和算法两个部分。

空战平台的通信模块基于 C++实现 TCP 连接，客户端模型开启了分别负责数据的接收与发送的两个进程，分别在两个端口进行数据收发；相应地，强化学习算法也开启两个进程，分别连接空战平台的两个收发端口，负责数据的接收与发送。

强化学习算法的通信模块使用 Python 编写。算法从空战平台接收的信息包括：对战轮数、帧数、两飞机编号、两飞机位置、两飞机姿态、两飞机机头朝向、两飞机速度、两飞机血量、两飞机加速度、两飞机偏航角速度、两飞机俯仰角速度。

2.3 数据预处理

复杂空战问题中状态量维度较大，因此需通过对状态空间的预处理，将空战过程的机理与演进流程进行筛

选和分析，提取复杂空战环境下取胜的关键影响因素，从而让算法更好地收敛，同时得到更好的飞机控制效果。预处理完成的工作如下：

1）计算敌机相对于智能体在 3 个维度上的角度；

2）计算两飞机之间的距离；

3）计算敌机和智能体机头的夹角；

4）计算智能体和敌机机头的夹角；

5）计算敌机相对于智能体的相对位置。

此外，为了更好地引导智能体学习，还需要一些态势信息，包括飞机血量、飞机位置、飞机姿态、飞机加速度、飞机偏航角速度和飞机俯仰角速度，并做归一化处理。

2.4　奖励函数设计

合理的奖励函数设计可以正确引导智能体学习方向。由于空战问题的智能体状态空间较大，且一局对战的时间比较长，因此仅用每局结束后的胜负作为奖励，算法将无法收敛。因此，在算法训练时，智能体与环境的每次交互均需要设计奖励值，用于评价执行当前动作的优劣，以解决稀疏奖励问题。

每次算法发送作动，得到一个新的状态都需要计算一个奖励值，获得的奖励值越高则表示此次作动决策越优。首先，如果两飞机某一次交互是一局对战中的最后一次交互，若智能体血量大于敌机血量，则获得最高奖励值 1；若智能体血量小于敌机血量，则获得最低奖励值−1；若两机血量相同则奖励值为 0。其次，如果某一次交互不是一局对战中的最后一次交互，需综合考虑交互前后两飞机的血量变化量、角度变化量、距离变化量、飞机速度和两飞机绝对距离进行奖励函数设计，为了让强化学习算法更好地收敛，每部分奖励值做归一化处理，最后将各部分奖励值加权处理。

1）血量变化奖励。血量变化是反映战斗态势变化的最直接方式。由于被航炮命中每秒损失 20 血量，因此算法每次与空战平台的交互时间 200 ms（参照飞行员 APM 平均在 300～600 之间，即飞行员决策控制率为 100～200 ms 设定）内每架飞机的血量最多变化 4 点，利用该值做归一化处理。血量奖励 R_b 计算方法如下，其中 B_1 和 B_2 分别为智能体血量变化和敌机血量变化：

$$R_b = \frac{(B_1 - B_2)}{4}$$

2）角度奖励。血量扣除条件为敌机在我方机头朝向夹角为 1°内，所以智能体朝向与敌机方向的夹角越小则越有优势，反之亦然。因此，角度奖励设计需综合考虑智能体的朝向与敌机的角度奖励（A_{angle}），以及敌机的朝向与智能体的角度惩罚（B_{angle}），即同时考虑我方打中敌机的优势与敌机打中我方的劣势，计算方法如下。其中，A_{pos} 和 B_{pos} 分别表示智能体与敌机位置，A_{dir} 和 B_{dir} 分别为智能体与敌机朝向：

$$A_{\text{angle}} = |\arccos\left[A_{\text{dir}} \times \frac{(A_{\text{pos}} - B_{\text{pos}})}{(\|A_{\text{pos}} - B_{\text{pos}}\|)}\right]$$

$$B_{\text{angle}} = |\arccos\left[B_{\text{dir}} \times \frac{(B_{\text{pos}} - A_{\text{pos}})}{(\|B_{\text{pos}} - A_{\text{pos}}\|)}\right]$$

3）相对距离奖励。智能体作动过程中距离敌方飞机越近则打中敌机的概率越大，因此在两飞机相对距离变化时给予相对距离奖励（R_d），若作动前后两飞机相对距离变小则获得正奖励，否则给予负奖励。飞机最大速度是 0.8 km/s，在算法每次与空战平台的交互时间 200 ms内每架飞机最多可行进 0.16 km，即两飞机相对距离最多可变换 0.32 km，利用该值进行归一化处理，相对距离奖励计算方法如下，其中和分别为作动前和作动后两机的相对距离：

$$R_d = \frac{R' - R}{0.32}$$

4）绝对距离奖励。在距离奖励层面，除了要考虑飞机两帧之间的距离变化量，还应该考虑两飞机的绝对距离（dist），两机绝对距离越近奖励越大。为了保证该部分奖励值范围在−1～1 之间，利用指数函数设计分段函数，函数公式如下，其中 F_m 为分段函数边界，程序中设置为 10 km：

$$R_a = \begin{cases} 1 - \dfrac{dist}{F_m}, & dist \leqslant F_m \\ e^{-[(\text{dist}/F_m)-1]} - 1, & ist > F_m \end{cases}$$

5）在仿真中发现，若智能体速度太快会出现严重的超调状况，导致智能体绕着敌方飞机高速转圈，对射击动作产生不利影响。因此当飞机速度（S）大于 150 m/s 时，给予一定的负奖励（R_S），让智能体学到降速飞行的策略，速度奖励计算如下：

$$R_S = \left[0.5 - \left(\frac{S - 150}{650}\right)\right] \times 2$$

将各部分奖励值加权处理，奖励值（reward）计算公式如下：

$$reward = \begin{cases} 0.5 \times R_a + 0.3 \times R_d + 0.1 \times (A_{\text{angle}} - B_{\text{angle}}) + 0.1 \times P_S, & dist > Fm \\ 0.6 \times dist + 0.1 \times R_d + 0.3 \times (A_{\text{angle}} - B_{\text{angle}}) + 0.1 \times P_S, & 1.8 \leqslant dist \leqslant Fm \\ 0.2 \times R_b + 0.7 \times (A_{\text{angle}} - B_{\text{angle}}) + 0.1 \times R_S, & dist < 1.8 \end{cases}$$

2.5　模型训练流程

空战自主决策模型训练包含 PPO 算法模块和空战平台模块，模型训练流程如图 4 所示。

1）初始化算法超参数。算法的主要超参数如表 1

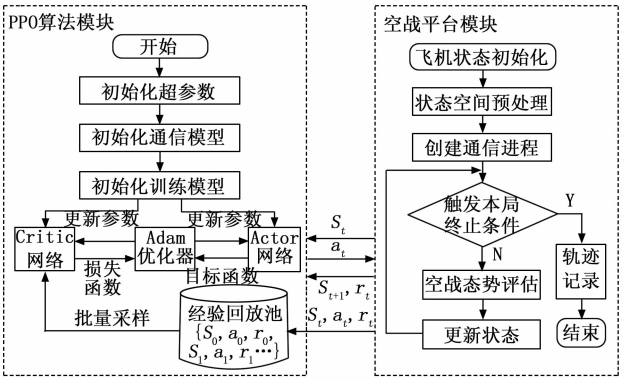

图 4 模型训练流程图

所示。

表 1 算法超参数设置

参数名称	参数值
策略网络学习率	0.000 2
价值网络学习率	0.000 4
折扣因子	0.95
裁剪系数	0.2
每局步数上限	600
动作更新步数	20
价值更新步数	20
样本重复训练次数	5
批次大小	32
经验缓冲区容量	8 192

2）初始化平台通信模型。创建服务器接收模型和服务器发送模型，用于算法与空战平台之间的通信。

3）初始化算法训练模型。训练模型包括 Actor 和 Critic，其中，Actor 为动作网络模型，它接收来自空战平台模块的环境状态 S_t，根据策略网络采样输出决策动作 a_t 发回空战平台，由智能体执行动作 a_t，同时根据空战态势评估得到奖励 r_t 并进入下个状态 S_{t+1}。Critic 为价值网络模型，它的输入是智能体状态，输出为价值预测，Critic 网络的训练方向就是让预测值尽量接近样本值。

4）开始训练。在智能体与敌机对战过程中，将 Actor 与空战平台的交互中收集到的大量样本集合 $\{S_t, a_t, r_t, S_{t+1},\cdots\}$ 存入经验缓冲池。从经验池中批量采集数据用于神经网络训练，分别计算 Actor 和 Critic 网络的目标函数和损失函数，并利用 Adam 优化器对网络参数进行更新，实现模型的梯度上升。利用不断更新的网络继续与空战平台交互，直至触发本局终止条件，即一方被击落、出界或撞机平局时，本局训练结束。

3 仿真实验与效果评估

在构建好的一对一空战格斗仿真平台下进行模型训练，决策模型的目标是用算法训练和控制飞机 A，将基于复杂数学模型决策的飞机 B 击败。

3.1 训练结果分析

奖励值是评估智能体学习效果的重要指标。在训练过程中，智能体所操纵的飞机 A 的奖励值随训练局数的变化情况如图 5 所示。从奖励值的变化情况可以看出，通过 PPO 算法数百局的训练，智能体所获取的奖励值不断提升并趋于稳定。

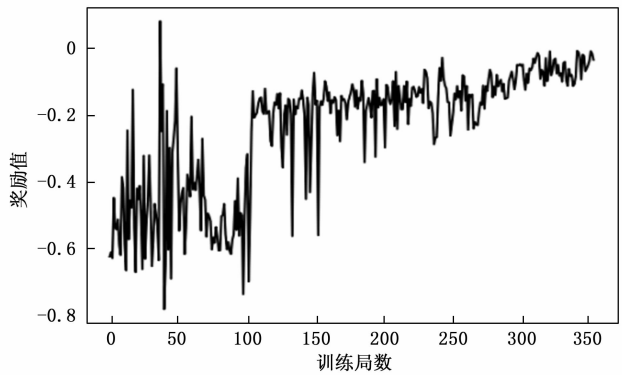

图 5 奖励值随训练局数变化

在 350 局的训练过程中，整体对战情况为飞机 A 胜利 62 局，失败 151 局，打平 137 局，战绩结果如图 6 所示，各阶段获胜及胜率情况如表 2 所示。

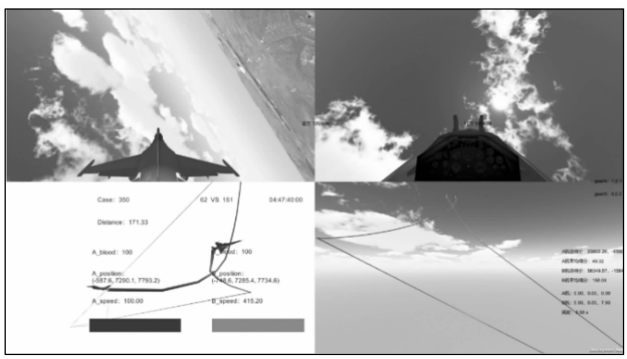

图 6 算法模型 350 局战绩结果

表 2 训练各阶段智能体胜率

训练阶段	胜利局数	失败局数	平局数	胜率（去除平局）/%
初期(0～100 局)	1	85	14	1.16
中期(101～200 局)	9	36	55	13.33
后期(201～350 局)	52	30	68	63.41

3.2 各阶段训练情况

3.2.1 训练初期

在训练的前 100 局，奖励值维持在较低范围，且波动大，说明智能体此时在学习难度大的环境中较难获得正奖励，在探索有效作战策略的过程中常处于较劣的空战态势下。

训练第 1 局两机的作战飞行轨迹如图 7 所示，其中

实心点是两机随机生成的初始位置。从作战轨迹可以看出，智能体的飞行轨迹几乎是直线，在短时间内就被敌机击败，表明此时智能体以高机动速度逃逸为主，尚不具备基本的躲避攻击能力。

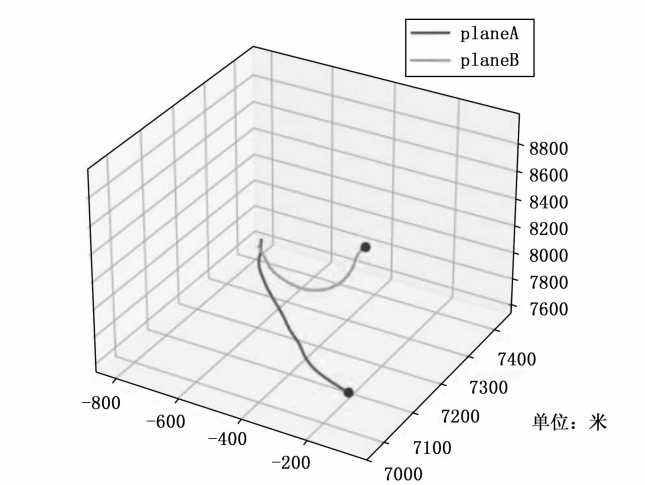

图 7　训练第 1 局两机作战轨迹

训练第 50 局两机的作战飞行轨迹如图 8 所示。从作战轨迹和对战录像中可以看出，智能体在被敌机尾后追击过程中以直飞为主，能够通过调整速度、俯仰等躲避攻击，实现较长时间的逃逸。这表明此时智能体已经学会了基本的躲避攻击技巧。

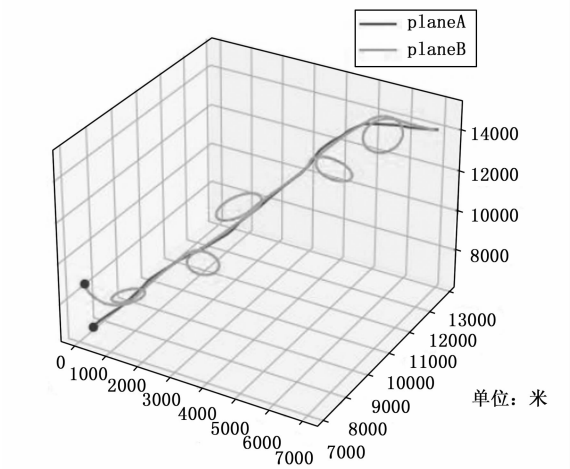

图 8　训练第 50 局两机作战轨迹

3.2.2　训练中期

在训练的第 101～200 局，智能体所获得的平均奖励有了较大提升，但波动较大，尚未收敛。该阶段智能体胜率为 13.33%，说明此时智能体已经探索到制胜策略。

训练第 200 局，两机的作战飞行轨迹如图 9 所示。通过智能体的作战轨迹分析，智能体目前学习到较复杂的机动策略，可通过转弯、爬升等动作有效规避敌方的尾后攻击；从对战录像分析，本阶段智能体仍以机动逃逸为主，在盘旋机动中与敌机产生较多的碰撞平局。从获胜局录像分析，智能体偶然探索到一种诱敌出界策略，但尚未稳定掌握。

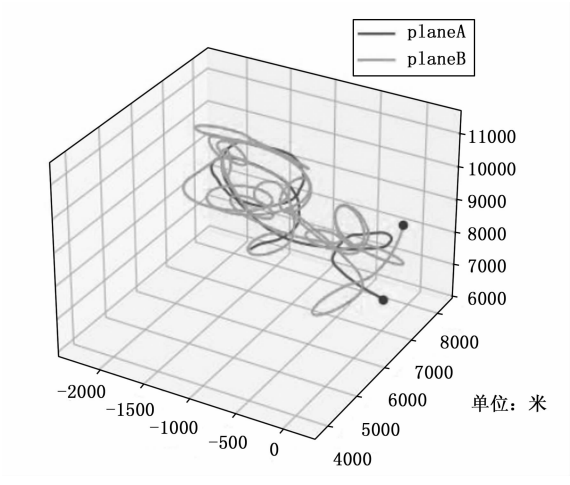

图 9　训练第 200 局两机作战轨迹

3.2.3　训练后期

在训练的第 201～350 局，智能体所获得的平均奖励值较训练中期有小幅提升，奖励值波动较小，对战胜率为 63.41%。此时智能体已经学习到稳定有效的作战策略，训练趋于收敛。

训练第 350 局，两机的作战飞行轨迹如图 10 所示。从该阶段作战飞行轨迹分析，智能体做出较多爬升和俯冲的机动；通过对战录像分析，智能体在敌尾后追击过程中频繁向地面俯冲和拉起，诱骗敌机在追击过程中因撞地而落败。这表明智能体已经学习到稳定的诱敌出界策略。

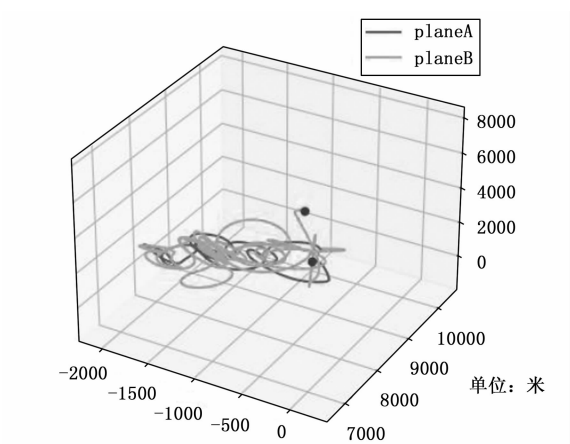

图 10　训练第 350 局两机作战轨迹

3.3　算法横向对比

QR-DQN 算法是深度 Q-网络（DQN）的改进算法，通过使用分布式的估计方法使其在处理值函数不确

定性时更强大，提供更稳健的学习策略。在相同条件下，使用 QR-DQN 算法进行数百局训练，奖励值随局数变化情况如图 11 所示。可以看出，说明 QR-DQN 算法在数百局的训练中没有收敛的趋势。

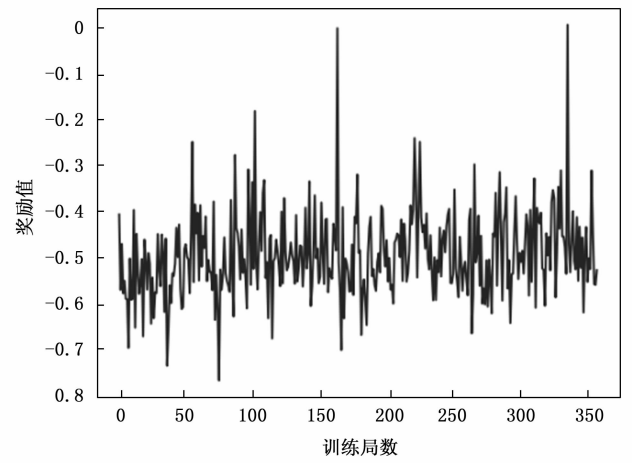

图 11 QR-DQN 奖励值随训练局数变化

在训练过程中，QR-DQN 算法取得的平均奖励值随局数变化情况曲线如图 12（a）所示，PPO 算法平均奖励值变化如图 12（b）所示。从变化趋势来看，QR-DQN 在前 200 局学习速度明显低于 200 局之后。QR-DQN 算法在 350 局训练的平均奖励值为－0.48，低于 PPO 算法的－0.22。

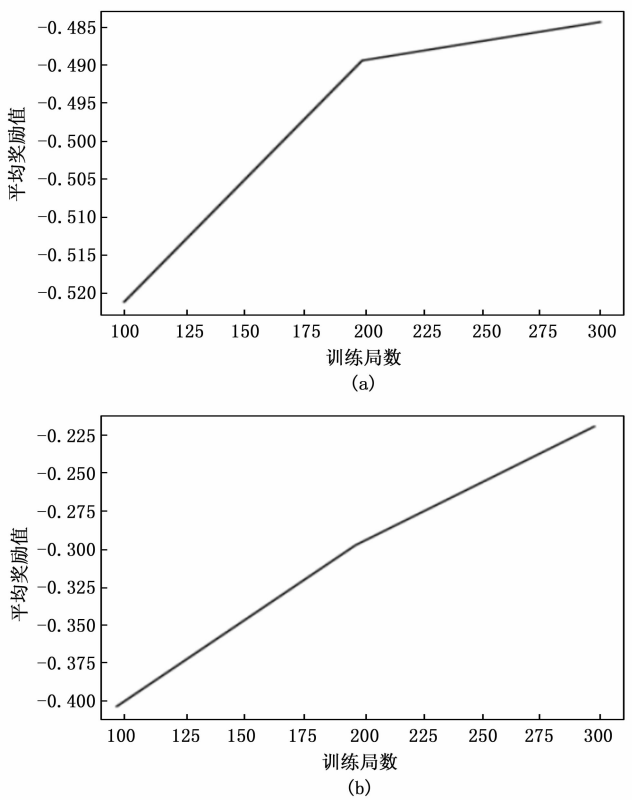

图 12 算法平均奖励值随局数变化

QR-DQN 算法的第 100 局和第 350 局训练作战飞行轨迹分别如图 13 所示。从智能体飞行轨迹分析，在训练初期智能体就学会了转弯、俯冲等战术动作来躲避敌机攻击，而经过 350 局训练后智能体能够做出转弯、爬升、俯冲、横滚、筋斗等复杂的战术机动动作，可以做到长时间逃逸。

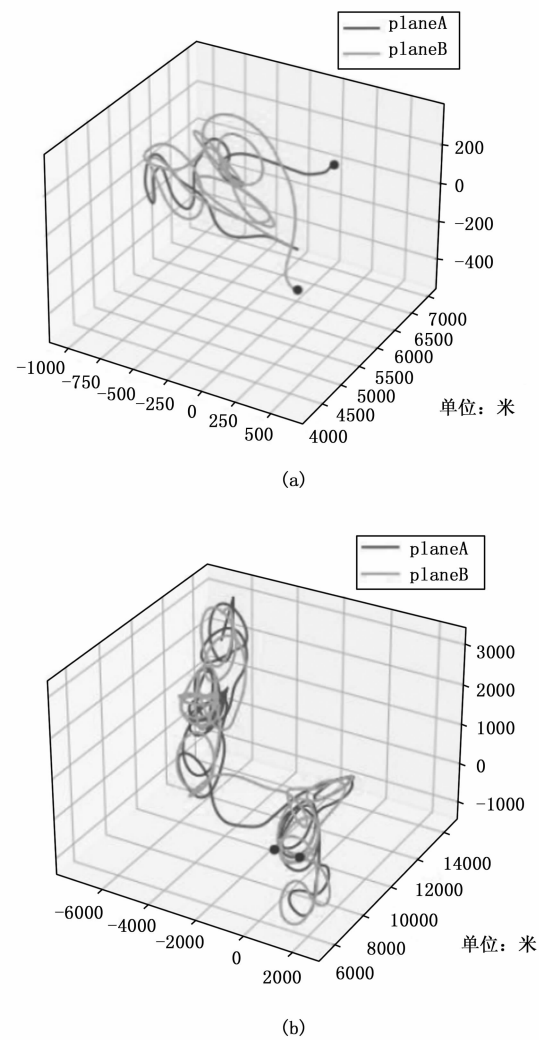

图 13 QR-DQN 训练中两机作战轨迹

进一步分析智能体作战策略，将 QR-DQN 模型进行数十局推理验证，如图 14 所示，智能体在 52 局博弈中共失败 6 局、打平 46 局。从作战过程分析，智能体在数十局中均以中低速、多战术动作逃逸为作战策略，没有学习到对敌机进行攻击及诱敌出界的获胜策略。智能体在与敌盘旋过程中有较大概率造成碰撞平局。

综上所述，与 QR-DQN 算法相比，一方面本文实现的 PPO 算法在训练过程中可以快速收敛，另一方面算法能够引导智能体学习到稳定有效的制胜策略。

4 结束语

本文针对复杂状态空间和动作空间的一对一近距空

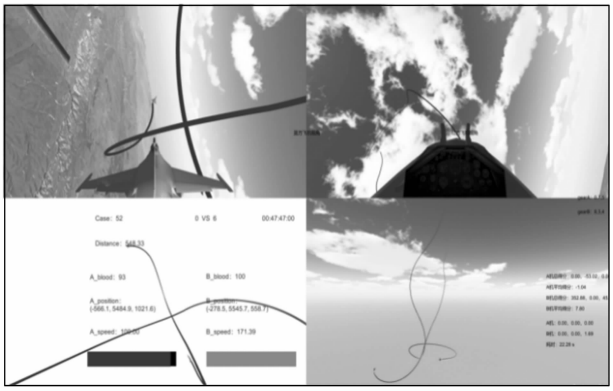

图 14　QR-DQN 算法推理情况

战决策问题，在基于 Unity3D 自主搭建的空战平台中设计和实现了空战自主决策框架，并在可视化仿真平台推理智能体决策模型、分析作战数据。仿真结果表明，本文实现的强化学习算法解决了复杂空战环境下自主决策的收敛困难问题和空战策略有效性问题，具备一定的工程应用参考价值。基于本文的研究，后续可以进一步考虑超视距、多智能体协同等空战问题，以适应更广泛的空战场景。

参考文献：

[1] 傅　莉，李　伟. 战机空战决策方法及分析 [J]. 沈阳航空航天大学学报，2013，30 (6)：48 - 52.

[2] 宋遐淦. 不确定环境下智能空战优化决策算法研究 [D]. 南京：南京航空航天大学，2017.

[3] 董一群，艾剑良. 自主空战技术中的机动决策：进展与展望 [J]. 航空学报，2020，41 (s2)：4 - 12.

[4] 陈　浩，黄　健，刘　权，等. 自主空战机动决策技术研究进展与展望 [J]. 控制理论与应用，2023，40 (12)：2104 - 2129.

[5] VAN HASSELT H，GUEZ A，SILVER D. Deep reinforcement learning with double q-learning [C] // Menlo Park：AAAI，2016：2094 - 2100.

[6] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning [J]. Nature，2015，518 (7540)：529 - 533.

[7] 张婷玉，孙明玮，王永帅，等. 基于深度 Q 网络的近距空战智能机动决策研究 [J]. 航空兵器，2023，30 (3)：41 - 48.

[8] 姚培源，魏潇龙，俞利新，等. 基于 Q-learning 算法的无人机空战机动决策研究 [J]. 电光与控制，2023，30 (5)：16 - 22.

[9] LILLICRAP T P，HUNT J J，PRITZEL A，et al. Continuous control with deep reinforcement learning [J]. ArXiv Preprint，2016，Arxiv：1509.02971.

[10] 贺宝记，白林亭，文鹏程. 基于态势评估及 DDPG 算法的一对一空战格斗控制方法 [J]. 航空工程进展，2024，15 (2)：179 - 187.

[11] 李永丰，吕永玺，史静平，等. 深度确定性策略梯度和预测相结合的无人机空战决策研究 [J]. 西北工业大学学报，2023，41 (1)：56 - 64.

[12] 周　攀，黄江涛，章　胜，等. 基于深度强化学习的智能空战决策与仿真 [J]. 航空学报，2023，44 (4)：99 - 112.

[13] 章　胜，周　攀，何　扬，等. 基于深度强化学习的空战机动决策试验 [J]. 航空学报，2023，44 (10)：122 - 135.

[14] SCHULMAN J，WOLSKI F，DHARIWAL P，et al. Proximal policy optimization algorithms [J]. ArXiv Preprint，2017，Arxiv：1707.06347.

[15] 丁　维，王　渊，丁达理，等. 基于 LSTM-PPO 算法的无人作战飞机近距空战机动决策 [J]. 空军工程大学学报 (自然科学版)，2022，23 (3)：19 - 25.

[16] 丁云龙，匡敏驰，朱纪洪，等. 基于 LSTM-PPO 算法的多机空战智能决策及目标分配 [J]. 工程科学学报，2024，46 (7)：1179 - 1186.

[17] 邱　妍，赵宝奇，邹　杰，等. 基于 PPO 算法的无人机近距空战自主引导方法 [J]. 电光与控制，2023，30 (1)：8 - 14.

[18] 钱殿伟，齐红敏，刘　振，等. 基于改进近端策略优化的空战自主决策研究 [J]. 系统仿真学报，2024，36 (9)：2208 - 2218.

[19] 李佐龙，朱纪洪，匡敏驰，等. 基于混合动作的空战分层强化学习算法 [J]. 航空学报，2024，45 (19)：530053.

[20] 张祥瑞，谭　泰，李　辉，等. 基于深度强化学习的无人机空战机动决策方法 [J/OL]. 计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0069621.

[21] 田成滨，李　辉，陈希亮，等. 基于深度强化学习的抗感知误差空战机动决策 [J/OL]. 工程科学与技术. https://doi.org/10.15961/j.jsuese.202300259.

[22] 单圣哲，张伟伟. 基于自博弈深度强化学习的空战智能决策方法 [J]. 航空学报，2024，45 (4)：206 - 218.

[23] 吴宜珈，赖　俊，陈希亮，等. 强化学习算法在超视距空战辅助决策上的应用研究 [J]. 航空兵器，2021，28 (2)：55 - 61.

[24] Huashu Defense. Mozi joint operations deduction system [EB/OL]. (2020 - 08 - 01) [2021 - 03 - 30]. http://www.hs-defense.com/col.jsp? id=114.

[25] SHIN H，LEE J，KIM H，et al. An autonomous aerial combat framework for two-on-two engagements based on basic fighter maneuvers [J]. Aerospace Science and Technology，2018，72 (1)：305 - 315.