基于一种改进PPO算法的无人机空战自主机动决策方法研究

张 欣[1]，董文瀚[1]，尹 晖[2]，贺 磊[1]，张 聘[1]，李敦旺[1]
（1.空军工程大学航空工程学院，西安，710038；2.空军工程大学教研保障中心，西安，710051）

摘要 深度强化学习的应用为无人机自主机动决策提供了新的可能。提出一种基于态势评估模型重构与近端策略优化(PPO)算法相结合的无人机自主空战机动决策方法，为一对一近距空战提供了有效策略选择。首先，建立高保真六自由度无人机模型与近距空战攻击模型；其次，基于空战状态划分重构角度、速度、距离和高度态势函数，提出一种描述机动潜力的新型态势评估指标；之后，基于态势函数设计塑形奖励，并与基于规则的稀疏奖励、基于状态转换的子目标奖励共同构成算法奖励函数，增强了强化学习算法的引导能力；最后，设计专家系统作为对手，在高保真空战仿真平台(JSBSim)中对本文工作进行了评估。仿真验证，应用本文方法的智能体在对抗固定机动对手与专家系统对手时算法收敛速度与胜率都得到了有效提升。

关键词 PPO算法；机动潜力；六自由度飞机模型；态势函数；近距空战；专家系统
DOI 10.3969/j.issn.2097-1915.2024.06.010
中图分类号 V279.3 **文献标志码** A **文章编号** 2097-1915(2024)06-0077-10

Research on Autonomous Maneuver Decision Method for Unmanned Aerial Combat Based on an Improved PPO Algorithm

ZHANG Xin[1], DONG Wenhan[1], YIN Hui[2], HE Lei[1], ZHANG Pin[1], LI Dunwang[1]
(1. Aviation Engineering School, Air Force Engineering University, Xi'an 710038, China;
2. Teaching and Research Support Center, Air Force Engineering University, Xi'an 710051, China)

Abstract An application of deep reinforcement learning makes it possible for unmanned aerial vehicles to complete an autonomous maneuver decision-making. This paper proposes an unmanned combat aerial vehicle (UCAV) autonomous air combat maneuver decision-making method based on the reconstruction of situational assessment models in combination with the proximal policy optimization (PPO) algorithm, providing effective strategy choices for 1 vs 1 within visual range (WVR) air combat. In response to the problem of low model fidelity, this paper, firstly, establishes a dynamic model of a six degree of freedom UCAV and defines the attack mode of WVR air combat. And then, in order to improve the adequacy of the situational assessment model in describing air combat, this paper reconstructs the angle, speed, distance, and altitude situational functions based on the division of air combat states, and proposes a new situational function that describes the potential for maneuver. In terms of reward function design, in addition to rule-based sparse rewards, sub target rewards are established based on the transforma-

收稿日期：2024-03-16
基金项目：国家自然科学基金(62303362)
作者简介：张 欣(2000—)，男，山西运城人，硕士生，研究方向为飞行导航、制导与飞行控制。E-mail：m15835924972@163.com
通信作者：董文瀚(1979—)，男，陕西西安人，教授，博士生导师，研究方向为飞行器导航、制导与飞行控制。E-mail：dongwenhan@sina.com

引用格式：张欣，董文瀚，尹晖，等. 基于一种改进PPO算法的无人机空战自主机动决策方法研究[J]. 空军工程大学学报，2024，25(6)：77-86.
ZHANG Xin，DONG Wenhan，YIN Hui，et al. Research on Autonomous Maneuver Decision Method for Unmanned Aerial Combat Based on an Improved PPO Algorithm[J]. Journal of Air Force Engineering University，2024，25(6)：77-86.

tion of air combat states, and shaping reward functions are designed based on situational functions to enhance guidance capabilities. Finally, an expert system is designed to be a competitor to evaluate the work presented in this paper on the high fidelity air combat simulation platform (JSBSim). The simulation verification shows that being confronted with the fixed maneuvering opponents and expert system opponents, the intelligent agent enables to effectively improve the convergence speed and winning rate of the algorithm by using the method proposed in this paper.

Key words PPO algorithm; mobile potential; six degree of freedom aircraft model; situation function; WVR air combat; expert system

随着世界各国智能化进程的推进,未来空战呈现出无人化、自主化、智能化的特点,并逐渐演变为一种机器人代理战争[1]。上世纪 60 年代越南战争开始,无人机就被广泛应用于军事作战,直到最近的俄乌冲突,无人机被证明能够在作战中对战斗人员实现“人机作战”降维打击的效果[2]。未来空战中,各军事团体争相发展无人力量,无人机对战无人机已成趋势,成熟有效的无人机自主空战决策技术成为决定未来战场制空权不可忽视的关键因素。无人机自主空战决策技术,是无人作战飞机(unmanned combat aerial vehicle,UCAV)进行自主空战所必须具备的关键技术,能够使 UCAV 适应动态变化的战场环境,减少地面指挥和操控人员干预,简化信息的传递和处理,消除通信干扰隐患,突破“人在回路”的生理和心理限制,具有十分广泛的发展潜力和应用前景。

世界范围内的众多学者为解决自主决策问题进行了深入研究,涌现出微分对策、影响图、专家系统、深度学习、强化学习等自主决策方法。微分对策法最早见于 Isaacs 等人的著作《Differential Game》[3],其基本思想是基于现代控制理论,将博弈问题作为双边或多边最优控制问题,然后利用最优控制方法来求解[4]。影响图(influence diagram)博弈法结合图论和博弈论,以图的方式描述影响决策的因素,从而计算策略总体效用的概率分布,并按照概率分布给出合理决策信息[5-6]。专家系统根据人的知识和经验设计决策规则,指导 UCAV 进行空战决策,工程化难度低[7-8]。深度学习(deep learning,DL)利用深度人工神经网络(artificial neural network,ANN)强大的非线性拟合能力,隐式表达空战态势和机动决策之间的映射关系。可通过深度神经网络对参数化的战术机动序列进行学习,实现了目标机动识别和最优机动策略求解,具有较强的实时性和鲁棒性[9]。强化学习(reinforcement learning,RL)利用“试错”机制,通过与环境交互来学习策略,使智能体能够通过不断探索,学习决策经验,最终实现空战决策。常用的算法有 Q-Learning[10]、DQN[11]、DDPG[12]、SAC[13]等。传统的空战决策方法要求对复杂多变的空战过程进行精确建模,这通常很难实现。随着决策空间的不断扩大,它也面临着维数爆炸的问题,导致问题求解困难,实时性不高。2016 年,Alpha 飞行员凭借出色的表现击败了一位美国空军上校,以深度强化学习(deep reinforcement learning,DRL)为核心的智能空战领域研究快速发展[14]。以 DRL 为代表的新兴研究方法通过试错机制训练智能体,该机制消除了精确建模,并基于深度神经网络使用数据生成机动决策,以提高解决方案效率[15]。

文献[16]证明了 RL 的自主机动决策的可行性。然而,UCAV 的运动仅在 2D 平面内考虑,对无人作战飞机在三维空间的运动研究较少。文献[17]提出了一种基于深度 Q 学习网络(deep Q-network,DQN)的智能战术决策方法。为了验证该方法的有效性,设计了一个基于 Min-Max 算法的敌情模型。仿真结果表明,DQN 方法的决策性能比 Min-Max 递归方法更快更有效。然而,在神经网络的输入中没有将速度作为特征向量,这可能影响了该方法的收敛速度。文献[18]提出了一种启发式 Q 网络方法来提高强化学习的效率,实现了空中对抗机动策略的自学习,最终帮助战斗机在不同的空中对抗情况下自主做出合理的机动决策。然而,与文献[16]类似,没有考虑高度变化对 UCAV 运动的影响。文献[19]提出了一种基于深度 Q 网络的 UCAV 近程空战自主机动决策模型;然而,在空战态势优势的定义中,速度和角度是耦合的,这可能会干扰具有奖励函数的最优策略的学习过程。文献[20]提出了一种基于状态对抗深度确定性策略梯度算法的机动策略生成算法,但文中的实验结果仅证明了该算法在二维平面上的有效性。

综上,当前关于 DRL 在空战自主决策方面的应用存在以下 2 个不足:一是研究对象保真度低,大多为二维空间中的三自由度无人机;二是没有充分考虑空战态势对奖励函数设计的影响,降低了智能体的训练效率。为解决以上问题,本文提出了一种基于态势评估模型重构与 PPO 算法相结合的空战自主机动决策方法。首先,建立六自由度无人机动力学模型并定义攻击样式,提高了战场环境的保真度。其次,通过空战状态划分实现空战态势的初判断,据

此建立包括基本态势评估与机动潜力态势评估的改进态势评估模型，提高了态势评估模型的评估能力。之后，基于态势评估模型设计塑形奖励，与规则稀疏奖励、状态子目标奖励共同构成总奖励函数，提高了 DRL 算法奖励对于智能体的引导作用。除此之外，模仿飞行员的驾驶方式，构造了基于杆舵控制的连续动作空间，使得智能无人机的机动能力大大提升。同时，还设计了符合空战特性的状态观测空间，为态势评估模型的准确评估提供基础。最后，设计了基于高度优势的专家系统对手。通过一对一近距空战仿真，验证了本文算法的有效性。仿真结果表明，本文提出的机动决策方法能够有效指导高保真六自由度无人机进行空战机动决策，指导 UCAV 自主学习机动策略，并最终实现空战胜利。

1　无人机与空战环境建模

1.1　六自由度飞机模型

现实中的飞机气动布局多样，动力学系统复杂，具有高阶、非线性的特点，为了实现模型研究的通用性，本文对环境与飞机做出适当假设，并在以下假设的基础上建立无人机的数学模型：

1）假设地球是平面的和静止的。

2）假设飞机为具有固定质量分布和恒定质量的刚体，其质量为常数。

3）假设重力加速度为常数，取 $g=9.8\ \mathrm{ms}^2$。

4）假设飞机发动机的安装角为零，则发动机推力指向飞机机头方向。

5）假设空气相对于地球静止。

飞机的运动学方程为：

$$\begin{cases}\dot{\varphi}=p+q\ \sin\varphi\ \tan\theta+r\ \cos\varphi\ \tan\theta\\ \dot{\theta}=q\ \cos\varphi-r\ \sin\varphi\\ \dot{\psi}=(q\ \sin\varphi+r\ \cos\varphi)/\ \cos\theta\end{cases}\tag{1}$$

$$\begin{cases}\dot{x}=u\ \cos\theta\ \cos\psi+v(\sin\theta\ \sin\varphi\ \cos\psi-\ \cos\varphi\ \sin\psi)+\\ w(\sin\theta\ \cos\varphi\ \cos\psi+\ \sin\varphi\ \sin\psi)\\ \dot{y}=u\ \cos\theta\ \sin\psi+v(\sin\theta\ \sin\varphi\ \cos\psi+\ \cos\varphi\ \cos\psi)+\\ w(\sin\theta\ \cos\varphi\ \cos\psi-\ \sin\varphi\ \cos\psi)\\ \dot{h}=u\ \sin\theta-v\ \sin\varphi\ \cos\theta-w\ \cos\varphi\ \cos\theta\end{cases}\tag{2}$$

式中：φ 为飞机滚转角；θ 为飞机俯仰角；ψ 为飞机偏航角；x、y、h 分别为飞机位移在地面坐标系三轴上的分量。

1.2　空战基本几何关系与空战状态划分

在一对一近距空战中，2 架敌对 UCAV 之间的相对位置和相对姿态，构成了一对一近距空战基本几何关系。敌对 2 架 UCAV 的基本几何关系可由相对角度和距离确定，如图 1 所示。

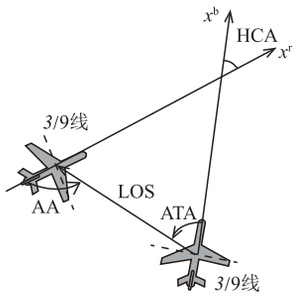

图 1　一对一近距空战基本几何关系

Fig. 1　Basic geometric relationship of 1 vs 1 WVR air combat

以蓝机为例，LOS(line of sight)为 UCAV 之间的观测线，ATA(antenna train angle)为蓝机的天线偏置角，天线偏置角的范围为 ATA∈ [0，π]。AA 为 UCAV 的进入角(aspect angle)，进入角的范围为 AA∈ [0，π]。HCA(heading crossing angle)为航向交叉角，航向交叉角的范围为 HCA∈ [0，π]。3/9 线则是 UCAV 的 3 点钟方位与 9 点钟方位的一条假想连线，可以用于判断一架飞机位于另外一架飞机的前半球还是后半球。

根据空战状态信息与空战基本几何关系，将空战状态分为追击、逃逸、迎面对抗和背身调整。4 个状态的划分如表 1 所示。

表 1　空战状态划分

Tab. 1　Classification of air combat states

几何关系	空战状态
$\mathrm{ATA}<\frac{\pi}{2}\cup\mathrm{AA}<\frac{\pi}{2}$	追击
$\mathrm{ATA}>\frac{\pi}{2}\cup\mathrm{AA}>\frac{\pi}{2}$	逃逸
$\mathrm{ATA}<\frac{\pi}{2}\cup\mathrm{AA}>\frac{\pi}{2}$	迎面对抗
$\mathrm{ATA}\geqslant\frac{\pi}{2}\cup\mathrm{AA}\leqslant\frac{\pi}{2}$	背身调整

这 4 个状态涵盖了空战过程中的所有可能的情况。

1.3　攻击区域建模

无人机的火力打击范围存在一定的限制条件，本文将无人机的火力打击范围简化为一个锥形攻击区域 Z_A，如图 2 中蓝色线条包围区域。攻击区 Z_A 为三维空间子集，表示如下：

$$Z_A=\{P\mid \mathrm{ATA}\leqslant\alpha_{A\max},D_{A\min}\leqslant|\mathrm{LOS}|\leqslant D_{A\max}\}\tag{3}$$

式中：$\alpha_{A\max}$ 为无人机的最大攻击偏角；$|\mathrm{LOS}|$ 为两机之间的距离；$D_{A\min}=152.4$ m 为无人机的最小射程；$D_{A\max}=914.4$ m 为无人机的最大射程；$P=[x,y,z]$ 为坐标点。

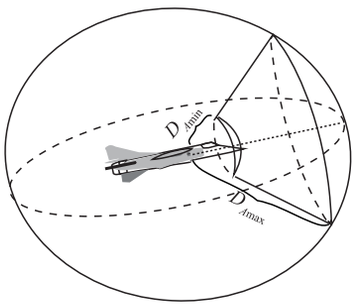

图 2 无人机机载武器攻击范围示意图

Fig. 2 Schematic diagram of the attack range of drone borne weapons

2 态势评估模型重构

2.1 基本态势评估模型建立

2.1.1 角度态势函数

文献[21]中提到一种启发式的角度态势函数，但其态势值变化随角度变化不够平滑，不利于强化学习算法训练的收敛，本文对角度态势函数进行了优化设计，表示为：

$$S_A=\exp(-|\sigma_1 \mathrm{ATA}|-|\sigma_2 \mathrm{AA}|) \tag{4}$$

式中：$\sigma_1>1$、$0<\sigma_2<1$ 为影响系数，用于量化 ATA 与 AA 对角度态势函数的影响。

图 3 的角度态势分布图可以直观看出 ATA 与 AA 对角度态势函数的影响。当敌机航向角不改变时，ATA 通过改变载机姿态可以轻易改变，AA 则需要载机进行长时机动才能够变化，而空战态势变化剧烈，响应速度越快，态势函数值越能反映当前空战的态势，因此借鉴贪婪算法的思想，提高 ATA 对角度态势函数的影响、降低 AA 的影响。

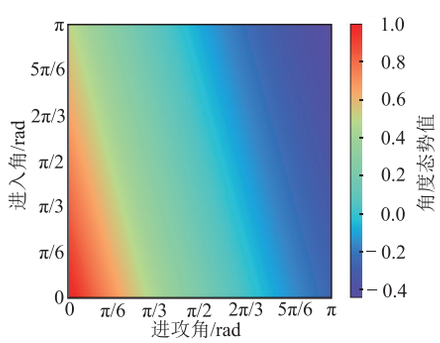

图 3 角度态势分布

Fig. 3 Angle situation distribution

2.1.2 距离态势函数

根据雷达探测范围、武器攻击范围来建立距离态势函数，并根据 1.2 节中的空战状态划分，对距离态势函数进行分类处理。其表示为：

$$S_D=\begin{cases}-1, D_{\mathrm{R}}\leqslant|\mathrm{LOS}|\\ 2\left(\exp\left(-3\dfrac{|\mathrm{LOS}|-D_{A\max}}{D_{\mathrm{R}}-D_{A\max}}\right)-0.5\right), D_{A\max}\leqslant|\mathrm{LOS}|<D_{\mathrm{R}}\\ 1, D_{A\min}\leqslant|\mathrm{LOS}|<D_{A\max}\\ 2\left(\dfrac{|\mathrm{LOS}|}{D_{A\min}}-0.5\right), 0\leqslant|\mathrm{LOS}|<D_{A\min}\end{cases} \tag{5}$$

式中：D_{R} 为雷达范围。距离态势函数的分布情况如图 4 所示。

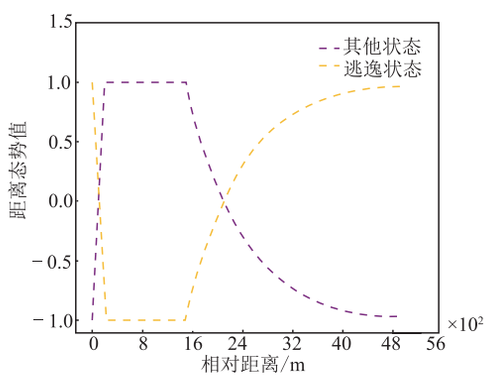

图 4 距离态势分布

Fig. 4 Distance situation distribution

$$S_H=\begin{cases}2\exp\left(0.5\dfrac{\Delta H-2\,000}{2\,000}+\dfrac{H_{\mathrm{b}}-H_{\mathrm{best}}}{H_{\mathrm{best}}}\right)-1,\\ \Delta H<2\,000, H_{\mathrm{b}}<H_{\mathrm{best}}\\ 2\exp\left(5\dfrac{2\,000-\Delta H}{2\,000}+\dfrac{H_{\mathrm{b}}-H_{\mathrm{best}}}{H_{\mathrm{best}}}\right)-1,\\ \Delta H\geqslant 2\,000, H_{\mathrm{b}}<H_{\mathrm{best}}\\ 2\exp\left(0.5\dfrac{\Delta H_{\mathrm{b}}-2\,000}{2\,000}+\dfrac{H_{\mathrm{best}}-H_{\mathrm{b}}}{H_{\mathrm{best}}}\right)-1,\\ \Delta H<2\,000, H_{\mathrm{b}}\geqslant H_{\mathrm{best}}\\ 2\exp\left(5\dfrac{2\,000-\Delta H}{2\,000}+\dfrac{H_{\mathrm{b}}-H_{\mathrm{best}}}{H_{\mathrm{best}}}\right)-1,\\ \Delta H\geqslant 2\,000, H_{\mathrm{b}}\geqslant H_{\mathrm{best}}\end{cases} \tag{6}$$

式中：$\Delta H=H_{\mathrm{b}}-H_{\mathrm{r}}$ 为双方飞机相对高度；H_{b} 为蓝方飞机高度；H_{best} 为最佳作战高度。

2.1.3 高度态势函数

一般情况下，空战中一方高度越高，对另一方的威胁越大。但受制于飞机性能，存在最佳作战高度。因此，根据双方高度差与最佳作战高度构建高度态势函数。

2.1.4 速度态势函数

在空战过程中，速度越大，越容易完成追击，具有更大威胁，但速度的增大不可避免会带来机动能力下降的问题，因此，文献[21]提出根据最佳空战速

度 $V_{F\text{best}}$ 来构造空战速度态势函数。

当 $V_{F\text{best}}>1.5V_T$ 时：

$$S_V=\begin{cases}\exp\left(-\dfrac{V_F-V_{F\text{best}}}{V_{F\text{best}}}\right), V_{F\text{best}}\leqslant V_F\\ 1, 1.5V_T<V_F\leqslant V_{F\text{best}}\\ -0.5+\dfrac{V_F}{V_T}, 0.6V_T<V_F\leqslant V_{F\text{best}}\\ 0.1, V_F<0.6V_T\end{cases}\tag{7}$$

当 $V_{F\text{best}}\leqslant 1.5V_T$ 时：

$$S_V=\begin{cases}\exp\left(-\dfrac{V_F-V_{F\text{best}}}{V_{F\text{best}}}\right), V_{F\text{best}}\leqslant V_F\\ \dfrac{2}{5}\left(\dfrac{V_F}{V_{F\text{best}}}+\dfrac{V_F}{V_T}\right), 0.6V_T<V_F\leqslant V_{F\text{best}}\\ 0.1, V_F<0.6V_T\end{cases}\tag{8}$$

式中：V_F 为己方飞机速度；$V_{F\text{best}}$ 为最佳空战速度；V_T 为敌方飞机速度。

2.2　机动潜力态势评估模型建立

基于本文 1.2 节的空战状态划分，为不同的空战状态构建不同的机动潜力态势函数，在某一状态下，其他状态的机动潜力态势函数值为 0。则该态势可以表示为：

$$S_\eta=S_{\eta_1}+S_{\eta_2}+S_{\eta_3}+S_{\eta_4}\tag{9}$$

式中：$S_{\eta_1}\sim S_{\eta_4}$ 分别为追击、逃逸、迎面对抗、背身调整状态的机动潜力态势函数值。

1）在追击状态下，要求天线偏置角变化率尽可能小，则追击附加态势函数如下所示：

$$S_{\eta_1}=\exp\left(-\xi_1\frac{|\mathrm{ATA}'|}{\mathrm{ATA}'_{\max}}\right)\tag{10}$$

式中：ξ_1 为调整系数；ATA' 为己方飞机天线偏置角变化率，其取值范围 $\mathrm{ATA}'\in\left[-\dfrac{\pi}{2}\ \mathrm{rad/s},\dfrac{\pi}{2}\ \mathrm{rad/s}\right]$；$\mathrm{ATA}'_{\max}$ 为蓝方飞机天线偏置角变化率最大值。

2）在逃逸状态下，要求飞机进入角变化率尽可能大，则逃逸附加态势函数如下所示：

$$S_{\eta_2}=\exp\left(\xi_2\frac{\mathrm{AA}'-\mathrm{AA}'_{\max}}{\mathrm{AA}'_{\max}}\right)\tag{11}$$

式中：ξ_2 为调整系数；AA' 为飞机进入角变化率，其取值范围 $\mathrm{AA}'\in\left[-\dfrac{\pi}{2}\ \mathrm{rad/s},\dfrac{\pi}{2}\ \mathrm{rad/s}\right]$；$\mathrm{AA}'_{\max}$ 为飞机进入角变化率最大值。

3）在迎面对抗状态下，当进入角、天线偏置角与进入角变化率、天线偏置角变化率满足一定关系时，可以合理刻画空战态势。设满足公式 $\mathrm{ATA}+\mathrm{ATA}'\cdot t_{\mathrm{ar}}<\dfrac{\pi}{6}(t_{\mathrm{ar}}\geqslant 0)$ 的最小 t_{ar} 为红方飞机进入攻击状态最短时间，满足公式 $\mathrm{AA}+\mathrm{AA}'\cdot t_{\mathrm{ab}}<\dfrac{\pi}{6}$ $(t_{\mathrm{ab}}\geqslant 0)$ 的最小 t_{ab} 为蓝方飞机进入攻击状态的最短时间，其中 $\dfrac{\pi}{6}$ 为飞机的武器攻击锥圆锥角度，上述 2 个公式表示 t_{ar} 或 t_{ab} 越小，飞机越快建立攻击条件。

$$S_{\eta_3}=\begin{cases}1, t_{\mathrm{ar}}<t_{\mathrm{ab}}\\ 0, t_{\mathrm{ar}}=t_{\mathrm{ab}}\\ -1, t_{\mathrm{ar}}>t_{\mathrm{ab}}\end{cases}\tag{12}$$

4）当飞机进行背身调整时，设满足公式 $\mathrm{ATA}+\mathrm{ATA}'\cdot t_{\mathrm{or}}<\dfrac{\pi}{2}(t_{\mathrm{or}}\geqslant 0)$ 的最小 t_{or} 为红方飞机退出调整状态最短时间，满足公式 $\mathrm{AA}+\mathrm{AA}'\cdot t_{\mathrm{ob}}<\dfrac{\pi}{2}$ $(t_{\mathrm{ob}}\geqslant 0)$ 的最小 t_{ob} 为蓝方飞机退出调整状态的最短时间，其中为 3/9 线状态转换的临界值。上述 2 个公式表示 t_{or} 或 t_{ob} 越小，飞机到达敌方飞机的后半球越快，形成追击态势。

$$S_{\eta_4}=\begin{cases}1, t_{\mathrm{or}}<t_{\mathrm{ob}}\\ 0, t_{\mathrm{or}}=t_{\mathrm{ob}}\\ -1, t_{\mathrm{or}}>t_{\mathrm{ob}}\end{cases}\tag{13}$$

3　PPO 空战决策算法

3.1　PPO 算法

近端策略优化（proximal policy optimization，PPO）算法属于基于策略梯度的 DRL 算法，并且 PPO 的实现采用了演员-评论家架构[22]。在训练过程中，PPO 算法通过梯度上升来最大化目标函数 $L_t^{\mathrm{CLIP}+VF+S}(\theta)$ 来更新神经网络的参数 θ，达到学习最优策略的目的，其性能稳定，能够处理离散和连续动作空间问题，PPO 算法训练流程如图 5 所示。

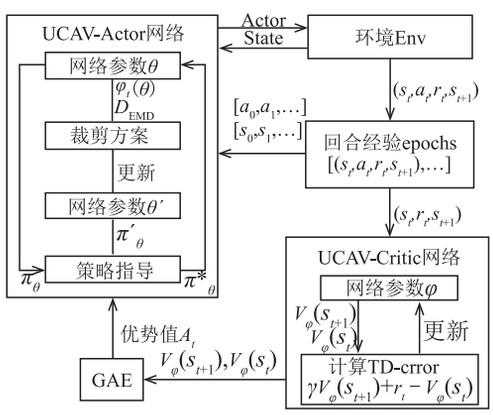

图 5　PPO 算法训练流程

Fig. 5　PPO algorithm training flow chart

本文关于 PPO 算法在空战问题上的整体应用框架主要包含 3 个模块。一对一近距空战格斗训练整体流程如图 6 所示。首先，由初始策略网络进行

动作策略生成，环境交互模块根据生成的动作策略进行环境与状态更新，并判断环境是否达到结束条件，以及对更新后的空战环境进行态势评估；之后由训练器进行奖励函数计算。采样器将过程中的动作、状态、奖励、结束判定信息作为经验存储到经验池内，当达到存储阈值时，算法训练器从经验池内抽取经验进行模型训练与参数更新，更新后的网络再次生成动作策略。

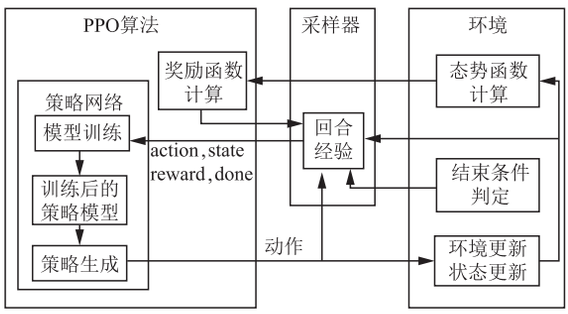

图 6 一对一近距空战格斗训练整体流程

Fig. 6 Overall process of 1 vs 1 close air combat training

3.2 三维近距空战博弈模型建立

3.2.1 状态空间

本文将空战状态信息分为单机状态信息与交互状态信息。其中，单机状态信息包括无人机三轴位置、三轴机动角、速度、法向过载、滚转速率、油门状态等 10 个状态。交互状态信息主要是敌我飞机状态差值，包括：三轴位置差值、速度差值、进攻角、进入角、3/9 线、航向交叉角等 8 个状态。其中 3/9 线状态是由 δ 表示的布尔值变量，$\delta=0$ 表示己方飞机位于敌方飞机的前半球，$\delta=1$ 表示己方飞机位于敌方飞机的后半球。据此建立空战状态空间：

$$S=[s_1,s_2\cdots,s_{18}] \tag{14}$$

式中：$s_1\sim s_{10}$ 为单机状态信息；$s_{11}\sim s_{18}$ 为交互状态信息。

空战状态信息如表 2 所示。

表 2 空战状态信息

Tab. 2 Air combat status information

状态分量	状态信息	状态分量	状态信息
s_1	x	s_{10}	T
s_2	y	s_{11}	e_x
s_3	z	s_{12}	e_y
s_4	v	s_{13}	e_z
s_5	θ	s_{14}	e_v
s_6	φ	s_{15}	ATA
s_7	ψ	s_{16}	AA
s_8	n_z	s_{17}	δ
s_9	$\dot{\varphi}$	s_{18}	HCA

3.2.2 动作空间

为保证飞机机动能力，本文通过操纵驾驶杆与油门杆实现飞行控制，可以得到无人机的连续动作空间，即：

$$A=\{\delta_a,\delta_r,\delta_e,\delta_p\} \tag{15}$$

式中：$\delta_a\in[-80\ \text{N},80\ \text{N}]$ 为滚转杆操纵量；$\delta_r\in[-500\ \text{N},500\ \text{N}]$ 为偏航舵操纵量；$\delta_e\in[-80\ \text{N},160\ \text{N}]$ 为俯仰杆操纵量；$\delta_p\in[0\%,100\%]$ 为油门杆操纵量；其中 $\delta_p=77\%$ 时为加力状态。

3.3 奖励函数设计

本文依据设定的空战规则、划分的空战状态以及设计的动态态势函数分别建立稀疏奖励、子目标奖励以及塑形奖励[21]。构造奖励函数如下所示：

$$R_{\text{total}}=R_{\text{sp}}+R_{\text{sg}}+R_{\text{sh}} \tag{16}$$

式中：R_{sp}(sparse reward)为基于空战规则的稀疏奖励；R_{sg}(subgoal reward)为基于空战状态转换的子目标奖励，R_{sh}(shaping reward)为基于态势函数的塑形奖励。

3.3.1 稀疏奖励

根据空战规则，存在以下 5 种可能的空战结果，并据此设置稀疏奖励，稀疏奖励情况如表 3 所示。

$$R_{\text{sp}}=k_1(R_{\text{sp}}^{+}+R_{\text{sp}}^{-}) \tag{17}$$

表 3 稀疏奖励表

Tab. 3 Sparse reward Tab

空战结果	奖励与惩罚
己方击落敌方	1
敌方超过行动边界	0.5
双方平局	0
己方超过行动边界	−0.5
己方被敌方击落	−1

3.3.2 子目标奖励

根据空战状态转换，构建子目标，并设计基于状态转换的子目标奖励，子目标奖励如图 7 所示。例如当飞机由逃逸状态转到背身调整状态或者迎面对抗状态时，飞机由劣势转为均势，获得 $+b$ 奖励；当飞机由背身调整状态或者迎面对抗状态转到追击状态时，表示由均势转为优势，同样获得 $+b$ 奖励；当状态转换相反时，获得同等大小的惩罚。当飞机保持追击状态时，给予 $+a$ 奖励。其中，$[a,b,c]$ 为常量且满足 $0<a<b<c$。

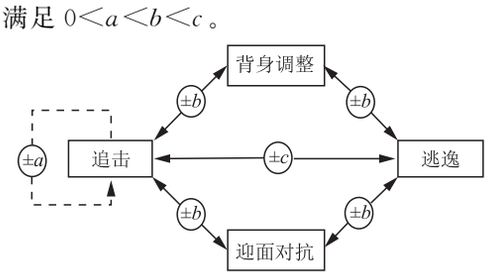

图 7 子目标奖励

Fig. 7 Subtarget reward

3.3.3　塑形奖励

在维持最优策略不变的前提下，根据态势函数构造塑形奖励函数，即：

$$R_{\mathrm{sh}}=k_3\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{total}} \tag{18}$$

式中：$\boldsymbol{\omega}=[\omega_1,\omega_2,\omega_3,\omega_4,\omega_5]^{\mathrm{T}}$；$\boldsymbol{S}_{\mathrm{total}}=[S_A,S_D,S_H,S_V,S_\eta]^{\mathrm{T}}$；$\omega_1\sim\omega_5$ 为塑形奖励的权重；$k_1\sim k_3$ 为 3 种奖励的调整系数。

4　一对一近距空战仿真实验

4.1　仿真配置

本文使用开源飞行动力学模型 JSBsim 软件来模拟战斗机的六自由度模型和地球模型，具体的软硬件环境如表 4 所示。

表 4　仿真软硬件环境

Tab. 4　Simulated software and hardware environment

项目	设置
操作系统	Microsoft Windows10 64 位
处理器	AMD Ryzen5 4 600 H with Radeon Graphics
内存	16.0 GB
开发语言	Python 3.7
可视化软件	Tacview
代码框架	Garage
第三方库	TensorFlow，numpy，scipy，gym，Matplotlib

三维近距空战仿真参数如表 5 所示，使用表中参数进行空战博弈仿真实验。

表 5　三维近距空战仿真参数

Tab. 5　3D WVR air combat simulation parameters

参数	范围
空域 X 坐标范围	(−20 000，20 000)
空域 Y 坐标范围	(−20 000，20 000)
空域 Z 坐标范围	(0，10 000)
单局空战最大步数	1 024
a，b，c	(0.1，0.5，1.0)
$k_1\sim k_3$	(2.0，0.5，2.0)
$\omega_1\sim\omega_5$	(0.4，0.2，0.1，0.1，0.2)
折扣因子	0.99
批处理样本数/训练回合采样步数	Batch size=4 096
数据复用次数	1
估计优势函数剪裁系数	0.2
GAE 调整方差与偏差系数	0.995
Actor 网络结构	18×256×256×4
Actor 网络学习率	2.5×10^{-4}
Critic 网络结构	18×256×256×1
Critic 网络学习率	2.5×10^{-4}

4.2　对抗固定机动仿真实验

首先给对抗双方设定初始态势，仿真实验中视蓝方为己方，则三维空战红蓝双方初始态势如表 6 所示。

表 6　三维空战红蓝双方初始态势

Tab. 6　Initial situation of red and blue parties in 3D air combat

战机	初始坐标/m	速率/(m/s)	航向角/(°)	滚转角/(°)	俯仰角/(°)
蓝方	(1 000.0，−1 000.0，5 000.0)	200.0	0.0	0.0	0.0
红方	(1 000.0，−6 000.0，5 000.0)	200.0	0.0	0.0	0.0

蓝方战机使用 PPO 算法学习生成策略进行机动动作，红方采取爬升机动运动方式。为验证机动潜力态势函数设计的有效性，分别采用普通态势函数（未基于空战状态划分构建且不包含机动潜力的态势函数）、改进态势函数（基于空战状态划分构建但不包含机动潜力的态势函数）与带机动潜力的改进态势函数（基于空战状态划分构建且包含机动潜力的态势函数）进行训练。

在空战仿真中，蓝方智能体的训练奖励变化曲线如图 8 所示，横轴为训练步数，纵轴为训练损失。从图中可以看出，随着训练次数的增加，奖励函数逐渐趋于稳定，证明训练满足了要求。

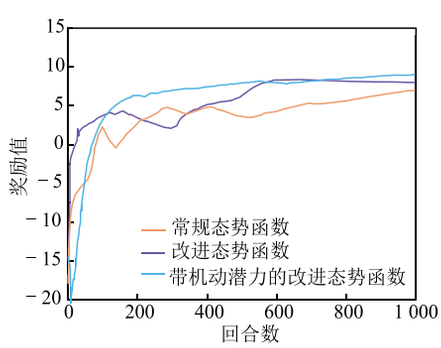

图 8　训练奖励变化曲线

Fig. 8　Training reward variation curve

蓝色曲线表示采用带机动潜力的改进态势函数训练情况，其在第 400 个训练回合后基本收敛；紫色曲线表示采用改进态势函数的训练情况，在接近 200 个训练回合时达到第 1 个奖励峰值，随后经过奖励波动，到第 600 个回合后基本收敛；橙色曲线表示采用普通态势函数的训练情况，其大约在 700 回合基本收敛。

为了测试算法训练效果，在基本收敛后，我们设置了一组胜率测试实验，利用训练后的智能体与本节设计的固定机动对抗 500 场次，记录胜负结果并在图 9 中展示。经计算，3 种方法的胜率分别为 98.60%、97.00%、93.80%。对比运用常规奖励的算法，算法的收敛速度提高了 33.33%，胜率提高了 5.12%。

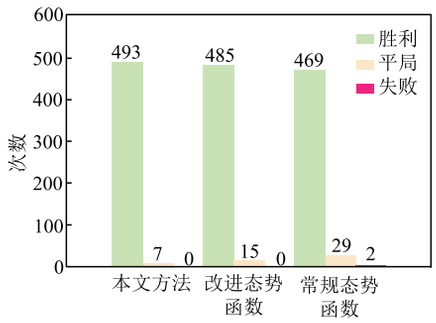

图 9 胜率测试结果统计

Fig. 9 Winning rate test results statistical

训练结束后，记录双方战机的机动数据，并使用 matplotlib 绘图进行轨迹可视化。图 10 展示了采用带机动潜力的改进态势函数的 PPO 算法训练 1 000 回合后的一次红蓝双方战机飞行轨迹。蓝机先通过提升高度获取高度优势，再通过转弯机动拉近距离，最后调整进攻角获取角度优势，直到实现追踪、完成攻击。训练表明采用该算法的蓝方智能体针对红方能够通过自我探索学习到合适的对抗策略，引导战机在该固定机动空战场景中取得胜利。

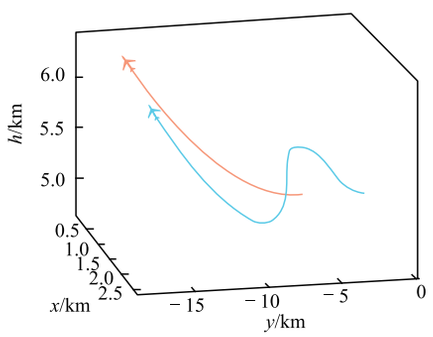

图 10 对抗简单机动飞行轨迹

Fig. 10 Fight against simple maneuvering flight trajectories

图 11 描述了在空战中红蓝双方态势变化情况。由于蓝方始终处于红方飞机的 3/9 线后，其空战态势始终优于红方，分析蓝方态势变化曲线，发现其态势值有 2 次波动变化，在第 1 次波动处，蓝机从提高高度态势向提高距离态势转变，在第 2 次波动，蓝机从提高距离态势向提高角度态势转变，最终实现目标追踪锁定并完成火力打击。

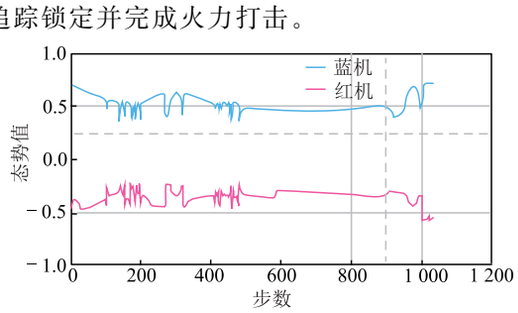

图 11 红蓝双方态势变化曲线

Fig. 11 Red and blue situation change curve

4.3 对抗专家系统仿真实验

设定初始状态，仿真实验中视蓝方为己方，具体参数如表所示。

表 7 三维空战红蓝双方初始态势

Tab. 7 Initial situation of red and blue parties in 3D air combat

	初始坐标/m	速率/(m/s)	航向角/°	滚转角/°	俯仰角/°
蓝方	(1 000.0，−1 000.0，5 000.0)	200.0	0.0	0.0	0.0
红方	(1 000.0，−5 000.0，5 000.0)	200.0	0.0	0.0	0.0

蓝机采取的机动动作生成方式与 4.2 节一致。红方采取专家系统决策生成机动指令进行机动动作，为了保证机动动作的连贯性，设置决策频率为 0.2 Hz，专家系统决策逻辑如图 12 所示。专家系统具体决策逻辑为：无人机先进行获取高度优势，在取得高度优势后，转换高度优势为速度、角度优势进行追踪机动。在机动过程中设置了自保机动保证无人机基本保持在限定的空战空间内。同时，为了提高系统的防御性能，当受到攻击威胁时，会采取防御机动进行攻击规避。

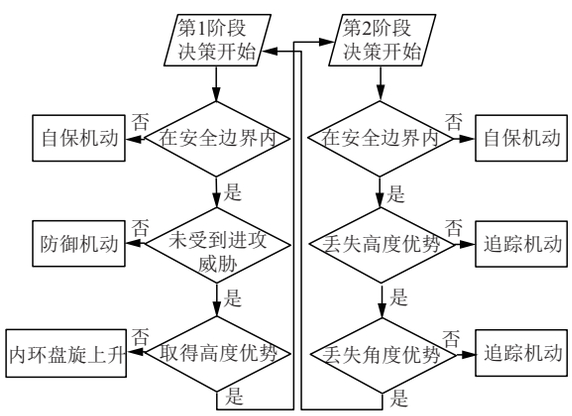

图 12 专家系统决策逻辑示意图

Fig. 12 Schematic diagram of expert system decision logic

红蓝双方对抗训练中，蓝方智能体的训练奖励变化曲线如图 13 所示。从图中可以看出，随着训练次数的增加，奖励函数逐渐趋于稳定，证明训练满足了要求。蓝色曲线在第 450 训练回合后基本收敛；紫色曲线在接近 400 个训练回合时达到第 1 个奖励峰值，随后经过奖励波动，到第 650 回合后基本收敛；橙色曲线大约在 850 回合基本收敛。

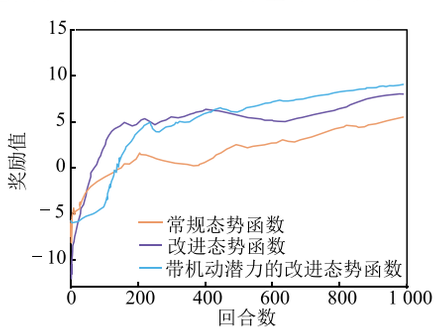

图 13 强化学习奖励变化曲线

Fig. 13 Reinforcement learning reward variation curve

胜率如图 14 所示，3 种方法的胜率分别为 72.80%、60.40%、55.40%，对比运用常规奖励的算法，算法的收敛速度提高了 47.06%，胜率提高了 31.41%。

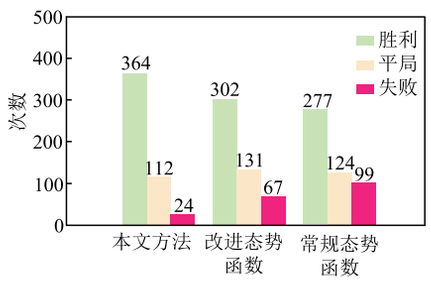

图 14　胜率测试结果统计图

Fig. 14　Winning rate test results statistical chart

在训练结束，记录对抗双方战机的机动数据，并使用 matplotlib 绘图与 Tacview 软件进行轨迹可视化。图 15 展示了训练 1 000 个回合后的红蓝双方战机博弈轨迹。

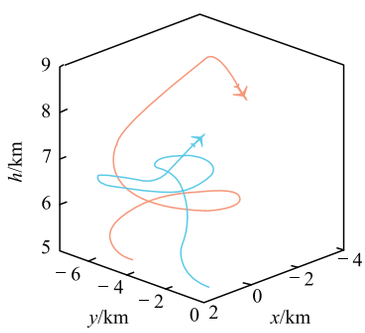

图 15　对抗复杂机动飞行轨迹

Fig. 15　Fight against complex opponents' maneuvering flight trajectories

结合图 16 博弈过程双方态势变化分析对抗过程。在博弈初始阶段，蓝机在速度、高度态势相同，角度态势占优的情况下，提高距离态势，拉近双机距离，便于锁定开火；在蓝方拉近的过程中，红方采取上升盘旋的机动动作应对，通过牺牲自身速度优势，降低了蓝机的高度、角度优势。对抗进行到 300 步后，红方距离态势骤增，双方角度态势呈均势，说明红方通过机动瓦解了蓝方的第 1 次进攻，且红方此时具备了反击机会；此时，蓝机进行了盘旋机动，并凭借高度优势使得红方丧失了进攻条件。在第 2 个阶段，对抗进行到约 1 000 步后，蓝方再次尝试锁定红方，但红方通过右转急拉起规避了蓝方的第 2 次进攻。第 3 个阶段中，蓝方迅速掉头，获取角度优势，拉近距离并提升自身高度，提高距离与高度优势；在取得一定的高度优势后，红方试图俯冲完成反击。在对抗的最终阶段，蓝方实现了对红方飞机的锁定，红方在反击的过程中未能躲避蓝方的第 3 次进攻，最终蓝方获得空战胜利。训练表明使用本文算法的蓝方智能体针对红方能够通过自我探索学习到合适的对抗策略，引导战机再博弈对抗中取得胜利。

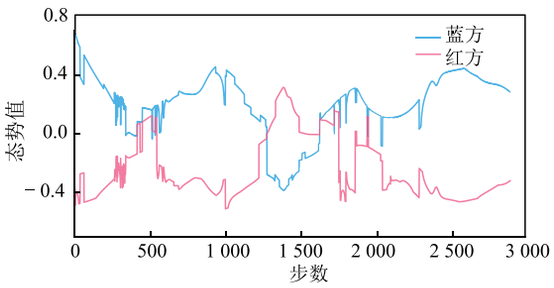

图 16　态势函数变化曲线

Fig. 16　Situation function change curve

5　结语

本文针对 PPO 算法一对一近距空战应用中存在的奖励函数设置困难问题，提出一种基于附加机动潜力态势函数的奖励函数设计方法，建立包含稀疏奖励、子目标奖励与态势塑形奖励的混合奖励结构，取得了良好的实验效果，该方法适用于六自由度高保真飞机模型。在仿真实验中，运用本文方法训练的智能体在对抗专家系统对手时，胜率提升了 31.41%，说明该方法对于复杂对手的适应性较高，能够引导无人战斗机进行机动决策并取得空战胜利。同时，该方法能够提升深度强化学习的训练效率，在对抗固定机动对手时收敛效率提升了 33.33%，对抗专家系统对手时算法收敛速率提升了 47.06%。上述实验结论达到了理论预期，为强化学习方法在智能空战决策的方面应用提供了新的思路。

参考文献

[1]　董康生，黄汉桥，韩博，等. 智能空战决策技术发展分析与展望[C]//第九届中国指挥控制大会论文集. 北京：兵器工业出版社，2021：5.
DONG K S, HUANG H Q, HAN B, et al. Analysis and Prospect of Intelligent Air Combat Decision-making Technology Development[C]//Proceedings of the 9th China Conference on Command and Control. Beijing: National Defense Industry Press, 2021: 5. (in Chinese)

[2]　崔勇平，邢清华. 从俄乌战争看无人机对野战防空的挑战和启示[J]. 航天电子对抗，2022，38(4)：1-3.
CUI Y P, XING Q H. The Challenge and Inspiration of UAVs to Field Air Defense from the Russia-Ukraine War[J]. Aerospace Electronic Warfare, 2022, 38(4): 1-3. (in Chinese)

[3]　BATHER J A, ISAACS R. Differential Games: a Mathematical Theory with Applications to Warfare and Pur-

suit,Control and Optimization[J]. Journal of the Royal Statistical Society SeriesA (General),1966,129(3):474.

[4] WEINTRAUB I E, PACHTER M, GARCIA E. An Introduction to Pursuit-Evasion Differential Games [C]//2020 American Control Conference (ACC). Denver,CO:IEEE,2020:1049-1066.

[5] PENG X Y, ZHAO Y, CAI Y J, et al. Application of Multi-Attribute Decision-Making Method Based on Fuzzy Influence Diagram in Green Supplier Selection [J]. International Journal of Computers Communications & Control,2024,19(2):6111.

[6] WANG X B, XIA X Z, TENG R J, et al. Risk Assessment of Dike Based on Risk Chain Model and Fuzzy Influence Diagram[J]. Water,2022,15(1):108.

[7] QIAN C, ZHANG X, LI L, et al. H3E: Learning Air Combat with a Three-Level Hierarchical Framework Embedding Expert Knowledge [J]. Expert Systems with Applications,2024,245:123084.

[8] LI B, BAI S X, LIANG S Y, et al. Manoeuvre Decision-Making of Unmanned Aerial Vehicles in Air Combat Based on an Expert Actor-Based Soft Actor Critic Algorithm [J]. CAAI Transactions on Intelligence Technology,2023,8(4):1608-1619.

[9] LI G L, WANG Y X, LU C, et al. Multi-UAV Air Combat Weapon-Target Assignment Based on Genetic Algorithm and Deep Learning[C]//2020 Chinese Automation Congress (CAC). Shanghai: IEEE, 2020: 3418-3423.

[10] CHEN Y Y, ZHANG J D, YANG Q M, et al. Design and Verification of UAV Maneuver Decision Simulation System Based on Deep Q-Learning Network [C]//2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). Shenzhen:IEEE,2020:817-823.

[11] HU T M, HU J W, ZHAO C H, et al. Autonomous Decision Making of UAV in Short-Range Air Combat Based on DQN Aided by Expert Knowledge[C]//International Conference on Autonomous Unmanned Systems. Singapore:Springer,2023:1661-1670.

[12] KONG W R, ZHOU D Y, YANG Z, et al. UAV Autonomous Aerial Combat Maneuver Strategy Generation with Observation Error Based on State-Adversarial Deep Deterministic Policy Gradient and Inverse Reinforcement Learning [J]. Electronics, 2020, 9 (7):1121.

[13] 李波,白双霞,孟波波,等. 基于 SAC 算法的无人机自主空战决策算法[J]. 指挥控制与仿真,2022,44(5):24-30.
LI B, BAI S X, MENG B B, et al. Autonomous Air Combat Decision-making Algorithm of UAVs Based on SACalgorithm[J]. Command Control and Simulation. 2022,44(5):24-30. (in Chinese)

[14] HAMBLING D. AI Outguns a Human Fighter Pilot [J]. New Scientist,2020,247(3297):12.

[15] FRANÇOIS-LAVET V, HENDERSON P, ISLAM R, et al. An Introduction to Deep Reinforcement Learning[J]. Foundations and Trends © in Machine Learning,2018,11(3/4):219-354.

[16] MCGREW J S, HOW J P, WILLIAMS B, et al. Air-Combat Strategy Using Approximate Dynamic Programming[J]. Journal of Guidance, Control, and Dynamics,2010,33(5):1641-1654.

[17] LIU P, MA Y F. A Deep Reinforcement Learning Based Intelligent Decision Method for UCAV Air Combat[C]//Asian Simulation Conference. Singapore: Springer, 2017: 274-286.

[18] ZHANG X B, LIU G Q, YANG C J, et al. Research on Air Confrontation Maneuver Decision-Making Method Based on Reinforcement Learning [J]. Electronics, 2018,7(11):279.

[19] YANG Q M, ZHANG J D, SHI G Q, et al. Maneuver Decision of UAV in Short-Range Air Combat Based on Deep Reinforcement Learning[J]. IEEE Access, 2020,8:363-378.

[20] KONG W R, ZHOU D Y, YANG Z, et al. UAV Autonomous Aerial Combat Maneuver Strategy Generation with Observation Error Based on State-Adversarial Deep Deterministic Policy Gradient and Inverse Reinforcement Learning [J]. Electronics, 2020, 9 (7):1121.

[21] 杨爱武,李战武,李宝,等. 基于动态变权重的空战态势评估[J]. 兵工学报,2021,42(7):1553-1563.
YANG A W, LI Z W, LI B, et al. Air Combat Situation Assessment Based on Dynamic Variable Weight [J]. Acta Armamentarii, 2021, 42(7): 1553-1563. (in Chinese)

[22] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-01-01). http://arxiv.org/abs/1707.06347.

(编辑:杜娟)