深度强化学习驱动的超视距空战自主决策方法

吕茂隆[1,2] 王金河[3] 韩浩然[4] 丁晨博[3] 万路军[1]

摘 要 随着机载传感器和中远距空空导弹技术的快速发展, 超视距空战已经成为现代空战的主流形式. 在这种复杂多变的作战环境中, 开发能够实时掌握战场态势并制定合理机动决策的智能化技术, 已成为军事技术研究领域的热点问题. 首先, 构建一个涵盖飞机六自由度动力学模型、导弹制导系统模型和雷达传感器系统的高保真仿真环境; 接着, 融合模仿学习和自博弈方法, 提出基于对手学习的空战决策框架, 以解决深度强化学习在空战中适应性和泛化性差的缺点, 提升智能体在复杂多变战场环境中快速适应和策略优化的能力; 最后, 构建 10 种具有显著战术差异性的专家系统, 在高保真空战仿真平台中与智能体进行博弈对抗. 结果表明, 在收敛速度和胜率等关键指标上, 所提出的空战决策框架优于传统深度强化学习决策策略, 有效性和泛化性强, 可为复杂超视距空战态势下快速生成可靠策略提供技术支持.

关键词 深度强化学习; 对手学习; 超视距空战; 智能控制

引用格式 吕茂隆, 王金河, 韩浩然, 丁晨博, 万路军. 深度强化学习驱动的超视距空战自主决策方法. 自动化学报, 2026, **52**(3): 510−524

DOI 10.16383/j.aas.c250334 **CSTR** 32138.14.j.aas.c250334

An Autonomous Decision-making Method for Beyond Visual Range Air Combat Driven by Deep Reinforcement Learning

LV Mao-Long[1,2] WANG Jin-He[3] HAN Hao-Ran[4] DING Chen-Bo[3] WAN Lu-Jun[1]

Abstract With the rapid development of airborne sensor technologies and medium-to-long-range air-to-air missile technologies, beyond visual range air combat has become the dominant form of modern air warfare. In such a complex and dynamic operational environment, the development of intelligent technologies capable of real-time battlefield situation awareness and rational maneuver decision-making has emerged as a research hotspot in the field of military technology. First, a high-fidelity simulation environment is constructed, encompassing a six-degree-of-freedom aircraft dynamics model, a missile guidance system model, and a radar sensor system. Subsequently, integrating imitation learning and self-play methods, an opponent-learning-based air combat decision-making framework is proposed to address the poor adaptability and generalization of deep reinforcement learning in aerial combat, thereby enhancing the agent′s ability to rapidly adapt and optimize strategies in complex and variable battlefield environments. Finally, ten expert systems with significant tactical differences are developed to engage in game-based confrontations with the agent within the high-fidelity air combat simulation platform. The results demonstrate that the proposed decision-making framework significantly outperform traditional deep reinforcement learning strategies in key metrics such as convergence speed and winning rate, exhibiting strong effectiveness and generalization. This work can provide technical support for the rapid generation of reliable strategies in complex beyond visual range air combat scenarios.

Keywords deep reinforcement learning; opponent learning; beyond visual range air combat; intelligent control

Citation Lv Mao-Long, Wang Jin-He, Han Hao-Ran, Ding Chen-Bo, Wan Lu-Jun. An autonomous decision-making method for beyond visual range air combat driven by deep reinforcement learning. *Acta Automatica Sinica*, 2026, **52**(3): 510−524

收稿日期 2025-07-21 录用日期 2025-12-03
Manuscript received July 21, 2025; accepted December 3, 2025
国家自然科学基金 (62303489, GKJJ24050502), 博士后面上基金 (2022M723877), 博士后特别资助 (2023T160790), 中国博士后国际交流引进计划 (YJ20220347), 军事科技领域青年人才托举工程 (2022-JCJQ-QT-018), 陕西省自然科学基础研究计划重点项目 (2025JC-QYCX-052) 资助
Supported by National Natural Science Foundation of China (62303489, GKJJ24050502), Postdoctoral General Fund (2022M723877), Special Postdoctoral Funding (2023T160790), China Postdoctoral International Exchange and Introduction Program (YJ20220347), Youth Talent Support Program for Military Science and Technology (2022-JCJQ-QT-018), and Key Project of the Natural Science Basic Research Program of Shaanxi Province (2025JC-QYCX-052)
本文责任编委 陈谋
Recommended by Associate Editor CHEN Mou

1. 空军工程大学空管领航学院 西安 710051 2. 空军工程大学无人飞行器技术全国重点实验室 西安 710051 3. 空军工程大学研究生院 西安 710051 4. 电子科技大学信息与通信工程学院 成都 611731
1. Air Traffic Control and Navigation College, Air Force Engineering University, Xi′an 710051 2. National Key Laboratory of Unmanned Aerial Vehicle Technology, Air Force Engineering University, Xi′an 710051 3. Graduate School, Air Force Engineering University, Xi′an 710051 4. School of Information and Communication Engineering, University of Electronic Science and Technology, Chengdu 611731

超视距空战 (beyond visual range air combat, BVR) 作为依托远程探测与制导武器实现视距外作战的形态, 其核心优势在于通过先发制人的远程打击能力显著压缩敌方反应时间, 并降低自身暴露风险, 从而重塑现代空战战术博弈格局. 2025 年 5 月 7 日印巴冲突中, 巴方 J-10CE 战机在 150 km 外发射 PL-15E 导弹击落多架印机[1], 凸显体系化支撑下超视距空战打击的核心地位. 在 2023 年 9 月美国 DARPA 的“空战进化”项目中, 深度强化学习 (deep reinforcement learning, DRL) 控制的 X-62A 与有人驾驶的 F-16 进行首次空战对抗测试, 使智能化空战形态成为该领域研究热点. DRL 能够使战斗机在动态对抗中自主优化战术, 突破传统人工决策反应局限, 形成“体系赋能装备、智能提升决策”的双重驱动, 大幅压缩“观察、判断、决策、行动”循环周期, 推动超视距空战成为未来空天战场的制胜核心[2].

目前, 传统超视距空战机动决策方法主要包括微分对策法[3]、影响图法[4] 和专家系统法[5]. 微分对策法通过设定战机位置、油门等状态与控制变量, 实现对连续状态演变及相互作用的描述; 影响图法通过整合空战决策者与专家见解, 以图形化方式实现解决空战不确定性问题; 专家系统法通过构建空战知识库等框架, 实现模拟人类飞行员空战决策过程. 值得注意的是, 微分对策法求解过程复杂、效率低, 难以快速响应空战中动态多变的对抗态势; 影响图法求解涉及复杂概率计算与优化, 难以在空战实时态势评估与决策中应用; 专家系统法过度依赖专家经验, 在空战未知复杂场景下适应性和灵活性差. DRL[6−7] 作为结合深度学习与强化学习的人工智能关键分支, 突破传统方法的局限, 具有自主学习和长期决策优化的优点, 已成功应用于空战自主策略控制[2, 8].

具体而言, 针对构建高保真空战对抗仿真场景问题, 文献 [9−11] 采用三自由度模型, 通过描述无人机的质心运动及相关动力学特性, 为仿真提供空战实验平台, 但三自由度模型未考虑目标与载机之间相对姿态变化对空战动态特性的影响, 导致对复杂空战态势演变的刻画能力不足, 难以适配高机动场景下的动态特性分析需求. 针对超视距空战中动作空间维度过高的问题, 文献 [12−15] 采用离散动作空间, 通过将高维连续动作空间转化为有限的离散动作集合, 实现降低动作空间复杂度并提升决策效率, 但在面对复杂多变的空战环境时, 缺乏根据实时态势动态调整动作空间的能力, 难以适应战场环境的非线性变化. 针对复杂动态环境下智能体需精确控制的问题, 文献 [9−10, 16] 采用连续动作空间, 通过直接操控连续控制量, 实现动态环境下的灵活决策, 但随着空战规模扩大与持续时间延长, 高维度连续动作空间将导致智能体的计算复杂度与资源消耗激增, 严重制约算法在实际应用中的实时性与可扩展性. 针对空战智能决策中奖励信号与策略优化耦合关联的问题, 文献 [13, 17−18] 采用稠密奖励函数, 通过高频、精细化的奖励反馈, 实现策略快速收敛, 但改进后的奖励函数设计复杂, 需要精确调整各因素的权重, 在实际应用中难以平衡.

综上, 当前关于 DRL 在空战自主决策方面的应用主要存在以下 3 个问题: 1) 在仿真环境构建时, 未考虑飞机的六自由度特性, 限制了无人机在复杂空战场景中的机动灵活性和决策精度; 2) 在设计奖励函数时, 没有充分考虑空战态势对奖励函数设计的影响, 降低了智能体训练效率; 3) 在设计动作空间时, 只采用离散动作空间或连续动作空间, 忽略单一类型动作空间的局限性, 即离散型缺失精细度、连续型耗费计算资源, 不能够全面真实地描述空战中的复杂决策过程.

为解决以上问题, 本文在设计基于深度 Q 网络 (deep Q network, DQN) 的深度强化学习方法基础上, 融合模仿学习和自博弈方法, 提出一种基于对手学习的超视距空战决策策略. 首先, 将空战机动问题转化为对载机、导弹制导系统及机载火控雷达系统的数学建模问题, 为后续决策方法设计提供仿真平台支持; 其次, 设计基于 DQN 的深度强化学习决策方法, 解决基于部分可观测马尔科夫决策过程 (partially observable Markov decision process, POMDP) 的空战决策问题; 然后, 综合模仿学习和自博弈的优势, 提出一种基于对手学习的决策方法, 以弥补传统深度强化学习方法对动态对抗环境适应性和泛化性差的缺点; 最后, 融入空战时间线决策逻辑设计专家系统, 通过全因子对抗仿真实验, 验证对手学习方法在超视距空战博弈决策研究中的有效性和泛化性.

1　超视距空战仿真环境构建

1.1　飞机动力学模型建模

本文建立的六自由度动力学模型[19] 涵盖飞机角速度、姿态角、位置以及速度等核心参数, 完整刻画飞机在空中的复杂运动状态, 包括描述机身旋转特性的转动运动方程和反映空间位置变化的平动运动方程. 定义飞机绕机体坐标系 OX, OY, OZ 轴的角速度分量为 p, q, k, 转动运动方程可表示为飞机在机体坐标系中的角运动方程, 即

$$\begin{cases}\dot{p}=\dfrac{I_y-I_z}{I_x}qk+\dfrac{M_x}{I_x}\\ \dot{q}=\dfrac{I_z-I_x}{I_y}kp+\dfrac{M_y}{I_y}\\ \dot{k}=\dfrac{I_x-I_y}{I_z}pq+\dfrac{M_z}{I_z}\end{cases}\tag{1}$$

式中, M_x、M_y、M_z 分别是绕机体坐标系OX、OY、OZ 轴的力矩, I_x、I_y、I_z 分别是飞机绕机体坐标系 OX、OY、OZ 轴的转动惯量. 由式 (1) 可得姿态角变化方程:

$$\begin{cases}\dot{\phi}=p+(q\sin\phi+k\cos\phi)\tan\vartheta\\ \dot{\vartheta}=q\cos\phi-k\sin\phi\\ \dot{\psi}=\dfrac{q\sin\phi+k\cos\phi}{\cos\vartheta}\end{cases}\tag{2}$$

式中, ϕ 是滚转角, ϑ 是俯仰角, ψ 是偏航角. 定义飞机质量为 m, 重力加速度为 g, 飞机平动运动方程可描述为在机体坐标系中的线运动方程:

$$\begin{cases}\dot{u}=\dfrac{1}{m}(T\cos\alpha\cos\beta-B)-g\sin\vartheta+kv-qw\\ \dot{v}=\dfrac{1}{m}(T\cos\alpha\sin\beta-Y)+g\cos\vartheta\sin\psi+pw-ku\\ \dot{w}=\dfrac{1}{m}(T\sin\alpha-F)+g\cos\vartheta\cos\psi+qu-pv\end{cases}\tag{3}$$

式中, u、v、w 分别是飞机在机体坐标系 OX、OY、OZ 轴上的速度分量, T 是发动机推力, B 是阻力, Y 是侧向力, F 是升力, α 是攻角, β 是侧滑角. 将式 (3) 转换至大地坐标系, 可得飞机的位置变化方程:

$$\begin{cases}\dot{x}=u\cos\vartheta\cos\psi+v\cos\vartheta\sin\psi-w\sin\vartheta\\ \dot{y}=-u\sin\psi+v\cos\psi\\ \dot{z}=u\sin\vartheta\cos\psi+v\sin\vartheta\sin\psi+w\cos\vartheta\end{cases}\tag{4}$$

式中, x、y、z分别是飞机在大地坐标系中的位置坐标.

1.2 导弹制导系统建模

本文研究的空空导弹为主动雷达制导导弹, 其制导过程包括中制导和末制导两个阶段. 在中制导阶段, 导弹主要依靠载机对目标的持续照射来获取目标位置和速度信息. 因此在建模时, 主要关注导弹的末制导阶段. 在末制导阶段, 导弹使用的制导方法是比例导引法. 如图 1 所示, 比例导引法的核心思想是通过调整导弹的速度矢量, 使其转动角速度与目标视线的转动角速度成比例, 从而实现对目标精确拦截. 本文选取目标和导弹相对运动轨迹的平面为攻击平面, 得到以下制导方程[20]:

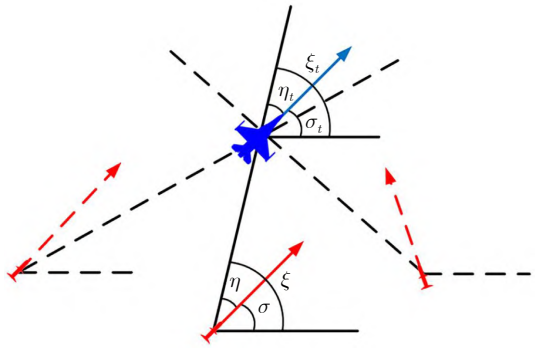


图 1 比例导引法

Fig. 1 Method of proportional navigation guidance

$$\begin{cases}\dfrac{\mathrm{d}\rho_m}{\mathrm{d}t}=V_t\cos\eta_t-V\cos\eta\\ \rho_m\dfrac{\mathrm{d}\xi}{\mathrm{d}\rho_m}=V\sin\eta-V_t\sin\eta_t\\ \xi=\chi+\eta\\ \xi_t=\chi_t+\eta_t\\ \dfrac{\mathrm{d}\chi}{\mathrm{d}t}=K\dfrac{\mathrm{d}\xi}{\mathrm{d}t}\end{cases}\tag{5}$$

式中, ρ_m 是导弹相对目标的距离, V、V_t 分别是导弹、目标的速度矢量, ξ、ξ_t 分别是导弹、目标线与基准线之间的夹角, χ、χ_t 分别为导弹、目标速度矢量与基准线之间的夹角, η、η_t 分别为导弹、目标速度矢量与目标线之间的夹角, K 是比例导引系数. 与文献 [21] 类似, 本文设定导弹与目标相对距离为 25 km 时, 导弹进入末制导, 当相对距离在 15 ~ 25 km 时, 取 $K=3$; 当相对距离在 10 ~ 15 km 时, 取 $K=4$; 当相对距离小于 10 km 时, 取 $K=5$.

1.3 雷达传感器建模

雷达传感器模块包含本次实验中的所有空中态势信息, 其属性和方法涉及雷达和传感器的多个方面, 包括预警探测系统、机载探测雷达系统、机载火控雷达系统、雷达告警系统. 与文献 [21] 类似, 预警探测系统以预警机布设位置为圆心, 划定半径为 300 km 的圆形区域作为其有效作用范围. 对于机载探测雷达系统和机载火控雷达系统, 由于二者均为单脉冲雷达, 其发射天线增益 G_t 与接收天线增益 G_r 相等, 均记作 G. 由此可推导出其最大作用距离 $d_{\max}$ 的计算式为

$$d_{\max}=\left(\frac{P_tG^2\lambda^2\sigma}{(4\pi)^3P_{\min}}\right)^{\frac{1}{4}}\tag{6}$$

对于雷达告警系统, 主要功能是检测来袭导弹的位置, 其最大作用距离 $d_{\max}$ 的计算方程为

$$d_{\max}=\left(\frac{P_tG_tG_r\lambda^2}{(4\pi)^2P_{\min}}\right)^{\frac{1}{2}} \tag{7}$$

式 (6) 和式 (7) 中, P_t 是雷达的发射功率, λ 是信号波长, σ 是目标的雷达反射截面积, $P_{\min}$ 是最小检测阈值. 为排除不同机型和武器参数对实验结果的影响, 本文设定空战双方为同代且性能相同的制空型飞机, 雷达反射截面积均为 4 m^2, 其中红、蓝双方型号均为 F-16, 装备性能参数来自文献 [21], 双方机载探测雷达系统、机载火控雷达系统和雷达告警系统的具体参数设计如表 1 所示.

表 1　雷达传感器参数

Table 1　Parameters of radar sensors

参数 (单位)	机载探测雷达系统	机载火控雷达系统	雷达告警系统
飞机发射功率 (kW)	30	22	—
导弹发射功率 (kW)	—	—	1
双方雷达反射截面积 (m^2)	4	4	—
发射天线增益 (dBi)	42	38	20
接收天线增益 (dBi)	42	38	30
信号波长 (m)	0.037	0.032	0.024
方位角 (°)	[−120, 120]	[−60, 60]	[−180, 180]
俯仰角 (°)	[−60, 60]	[−15, 15]	[−90, 90]

空战环境下, 在一定高度范围内, 由于高空视距扩展及地面杂波干扰减弱的影响, 雷达探测能力随飞行高度升高呈增强趋势, 如图 2 所示. 本文中, 当飞行高度从海平面提升至 8 000 m 时, 机载雷达信号的最小检测阈值降低至−100 dBm, 导弹末制导雷达信号的最小检测阈值降低至−30 dBm. 基于该作战条件, 通过式 (6) 和式 (7) 计算得到机载探测雷达系统、机载火控雷达系统、雷达告警系统的有效作用距离分别是 120 km、65 km 和 80 km. 值得注意的是, 雷达告警系统对来袭导弹末制导雷达信号的探测距离, 优于该末制导雷达对载机的探测距离, 这使载机能够在来袭末制导雷达开机时截获威胁信号[21].

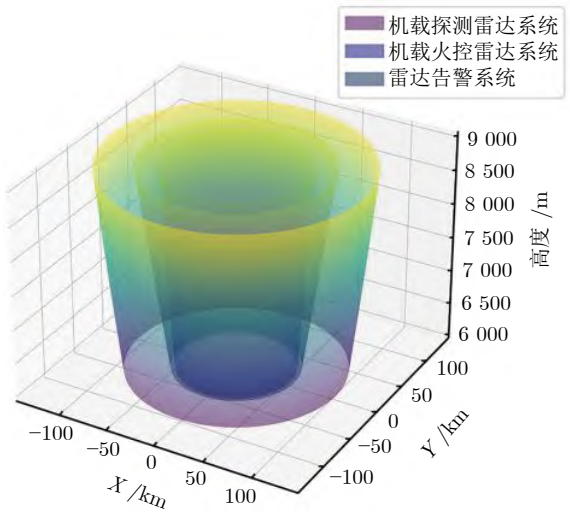


图 2　雷达传感器作用距离

Fig. 2　The operating range of radar sensors

2　基于 DQN 的超视距空战决策架构设计

2.1　基于 DQN 的空战决策方法描述

2.1.1　问题描述

在超视距空战环境中, 战斗机需要在复杂且动态变化战场态势中做出快速而精准的决策. 然而, 超视距空战场景具有高度不确定性, 战斗机难以完全获取对手的详细信息和战术意图, 同时作战环境中的电磁干扰、气象条件以及目标机动等因素, 进一步增加了决策的复杂性. 对于超视距空战中复杂不确定性决策问题, 与文献 [14] 保持一致, 本文采用 POMDP 进行决策过程建模. 在这种模型中, 战斗机根据自身局部观测信息独立地选择动作, 并最大化整体作战效能.

POMDP 模型由七元组 $\langle S, A, \Theta, P, R, \Omega, \gamma\rangle$ 定义. 其中, S 表示全局状态空间, 包括战斗机的速度、位置、角度; A 表示战斗机的动作空间, 包括离散战术动作和连续机动动作; Θ 表示观测空间, 战斗机通过观测函数获取观测值 $\theta\in\Theta$; P 表示状态转移概率, 描述为 $P(s'\mid s, a)$, 表示在当前状态 s 下执行动作 a 后, 系统转移到状态 s' 的概率; R 表示奖励函数, 描述为 $R(s, a)$, 表示战斗机在状态 s 下执行动作 a 后获得的即时奖励, 包括关键事件奖励和状态奖励; Ω 表示观测函数, 描述为 $\Omega(\theta\mid s, a)$, 表示在状态 s 下执行动作后, 观测到观测值 θ 的概率; γ 表示折扣因子. 本文设计超视距空战决策架构时, 将战斗机视为具备环境感知、自主决策能力的智能体, 并对空战问题中的作战流程与对抗机制进行建模仿真, 从而可以将深度强化学习方法应用于该问题中. 智能体通过与环境的交互, 获取局部观测信息 $\theta\in\Theta$, 并根据深度强化学习算法选择最佳动作 $a\in A$. 在执行动作后, 智能体获得奖励 $R(s, a)$, 并转移到新的状态 s'. 通过不断学习和优化, 智能体能够逐步逼近最优决策策略, 从而提高作战效能.

2.1.2　DQN 算法描述

DQN[13] 结合深度学习和 Q-学习, 是一种典型的无模型深度强化学习算法, 其更新计算式为

$$Q(s_t,\ a_t) = Q(s_t,\ a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},\ a) - Q(s_t,\ a_t)\right] \quad (8)$$

式中, α 代表学习率, γ 是折扣因子, r 为即时奖励, s_t 表示 t 时刻状态, s_{t+1} 表示 $t+1$ 时刻状态.

DQN 算法流程如下: 首先, 初始化深度 Q 网络 $Q(s,\ a;\theta)$ 和目标网络 $Q(s,\ a;\theta^-)$ 参数, 以及经验回放缓冲区 N, 并选择一个初始状态 s. 然后, 在每个时间步 t, 根据当前的 ε-贪心策略选择动作 a, 执行该动作后观察奖励 r 和新状态 s', 将经验 $(s,\ a,\ r,\ s')$ 存储到 N 中. 接着, 从 N 中随机抽取一个批次的样本 $\{(s_i,\ a_i,\ r_i,\ s_i')\}$, 计算目标值 y_i, 如果是终止状态 s_i', 则 $y_i = r_i$; 否则, $y_i = r_i + \gamma \max a' Q(s',\ a';\theta^-)$. 最后, 每间隔固定步数调用均方误差损失函数:

$$MSE_1 = \frac{1}{N} \times \sum (y_i - Q(s_i,\ a_i,\ \theta)^2) \quad (9)$$

更新当前网络参数 θ, 并定期更新目标网络参数 θ^- 为 θ. 该训练过程不断迭代, 直至达到最大训练步数或者智能体性能指标达到预定的标准时结束.

2.2 深度强化学习决策方法设计

2.2.1 状态空间设计

在设计深度强化学习状态空间时, 需遵循以下步骤[21]: 任务分解、特征选择与设计、状态抽象化与形式统一、效果评估. 这些步骤确保状态空间在不同任务场景下保持一致性, 从而增强其泛化能力.

依据以上步骤, 超视距空战划分为四个子阶段: 搜索与识别、跟踪与截击、参战与脱离、回转评估. 这四个子阶段的关键特征包括: 双方战机的位置、速度和航向, 双方之间的相对距离和角度, 以及导弹的相对位置和速度. 基于这些特征, 本文设计一个结构化状态空间, 以满足空战场景中决策需求.

值得注意的是, 在空战场景中, 双方战机均可作为载机或目标机. 载机是指发射导弹的飞机, 而目标机则是被载机攻击的对象. 这种角色划分是基于任务阶段和战术需求动态确定, 而非固定不变的. 载机的状态信息可表示为

$$S_i = [L_i,\ v_i,\ T_i,\ O_i] \quad (10)$$

式中, L_i 是载机的三维坐标, v_i 是载机速度矢量, T_i 是载机天线偏置角, O_i 是载机尾后角. 目标机的状态信息可表示为

$$S_j = [L_j,\ v_j,\ T_j,\ O_j] \quad (11)$$

式中, L_j 是目标机的三维坐标, v_j 是目标机速度矢量, T_j 是目标机天线偏置角, O_j 是目标机尾后角.

在空战过程中, 载机状态空间的构建不仅需要准确获取目标机的位置和速度信息, 以伺机攻击, 还需要感知来袭导弹的位置数据, 以及时防御. 本文中, 规定每架飞机可发射两枚空空导弹, 因此状态空间 S 设计为

$$S = [S_i,\ S_j,\ D_1,\ D_2] \quad (12)$$

式中, D_1 是载机与第 1 枚来袭导弹相对位置矢量, D_2 是载机与第 2 枚来袭导弹相对位置矢量.

2.2.2 动作空间设计

本文设计一个两级结构的参数化动作空间[22−23], 该结构包括战术指令层和机动参数层, 如图 3 所示. 该设计旨在为智能体提供一个灵活而精确的动作选择机制, 使其能够在复杂的战术环境中做出有效的决策. 智能体首先从一组离散动作 $\{A_1,\ A_2,\ \cdots,\ A_K\}$ 中选择一个动作 A_i, 然后为该动作指定维度的连续动作参数 $\{t_1^i,\ t_2^i,\ \cdots,\ t_{n_i}^i\}$, 于是动作空间可以形式化表示为

$$A = \bigcup_{A_i \in \{A_1,\ A_2,\ \cdots,\ A_K\}} \left\{(A_i,\ t_1^i,\ t_2^i,\ \cdots,\ t_{n_i}^i)\right\} \quad (13)$$

战术指令动作集参照第 2.2.1 节中状态空间的四种子阶段, 可划分为五种战术指令: 搜索、锁定、攻击、规避、脱离, 其中每一种战术指令可分解为多个机动参数. 机动参数包括四种基本参数: 速度、滚转角、俯仰角、偏航角, 以及四种专项动作参数: 机载探测雷达参数、机载火控雷达参数、雷达告警参数、导弹发射参数. 战术指令层动作集设计如表 2 所示.

在机动参数层, 四种基本参数和四种专项动作

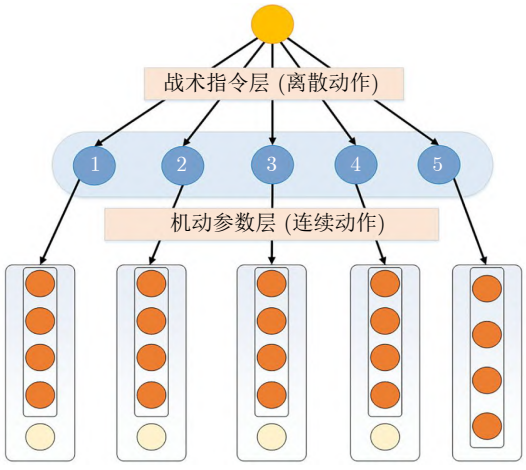


图 3 混合动作空间示意图

Fig. 3 Illustration of a hybrid action space

表 2　战术指令层动作集

Table 2　Tactical command layer action set

序号	战术指令	机动参数
1	搜索	四种基本参数、机载探测雷达参数
2	锁定	四种基本参数、机载火控雷达参数
3	攻击	四种基本参数、导弹发射参数
4	规避	四种基本参数、雷达告警参数
5	脱离	四种基本参数

参数均为连续动作参数. 智能决策模型通过调整速度、角度的大小对载机进行控制, 通过调整机载雷达参数、雷达告警参数及导弹相关参数, 来控制武器状态, 即 $a_t = [a_v, a_\phi, a_\theta, a_\psi, a_{r_1}, a_{r_2}, a_{r_3}, a_m]$, 其中速度 a_v 取值范围为 [0.6 Ma, 2.0 Ma]; 滚转角 a_ϕ 取值范围为 [−360°, 360°]; 俯仰角 a_θ 取值范围为 [−20°, 30°]; 偏航角 a_ψ 取值范围为 [−60°, 60°]; 机载探测雷达参数 a_{r_1}、机载火控雷达参数 a_{r_2}、雷达告警参数 a_{r_3} 取值范围均为 [0, 1], 满足:

$$f(a_{r_k}) = \begin{cases} \text{雷达关机}, & a_{r_k} \in [0,\ 0.5) \\ \text{雷达开机}, & a_{r_k} \in [0.5,\ 1] \end{cases} \tag{14}$$

式中, $k = 1, 2, 3$. 导弹发射指令 a_m 取值范围为 [0, 3), 并引入 int 函数, 表示某时刻执行攻击指令时, 载机发射第 $[a_m]$ 枚导弹. 这种参数化动作空间的设计不仅提供了战术层面的灵活性, 还通过连续参数的调整实现了动作执行的精确控制.

2.2.3　奖励函数设计

本文设计一种混合奖励函数[16, 24], 用于指导智能体在空战环境中的决策. 奖励函数由关键事件奖励和状态奖励组成, 旨在引导智能体采取行动以规避敌方导弹、击落目标, 如式 (15) 所示. 其中, 关键事件奖励以文献 [16] 为基础, 依据超视距空战规则、空战态势转变以及空战战术选择进行设定, 如表 3 所示.

$$\begin{cases} R_{\text{tr}} = R_{\text{kr}} + R_{\text{sr}} \\ R_{\text{kr}} = R_{\text{win}} + R_{\text{draw}} + R_{\text{fail}} + R_{\text{lock}} + \\ \qquad R_{\text{locked}} + R_{\text{launch}} \\ R_{\text{sr}} = R_{\text{df}} + R_{\text{th}} + R_{\text{ad}} + R_{\text{ts}} \end{cases} \tag{15}$$

式中, R_{tr} 为总奖励函数; R_{kr} 为关键事件奖励, 旨在驱动智能体规避敌导弹并击落目标, 确保其在关键事件上获得显著反馈, 包括六个常数奖惩部分: 击落奖励 R_{win}、平局惩罚 R_{draw}、被击落惩罚 R_{fail}、锁定奖励 R_{lock}、被锁定惩罚 R_{locked}、发射导弹惩罚 R_{launch}; R_{sr} 为状态奖励, 旨在驱动智能体综合态势并寻找发射阵位, 在安全飞行的前提下, 不断采取有利战术动作, 包括四个奖惩函数部分: 危险飞行惩罚 R_{df}、威胁惩罚 R_{th}、优势奖励 R_{ad}、时间步惩罚 R_{ts}, 具体函数设计如下:

表 3　奖励事件及取值设计

Table 3　Reward event and value design

类型	名称	取值
关键事件奖励	击落	50
	平局	−20
	被击落	−50
	锁定	10
	被锁定	−10
	发射导弹	−5
状态奖励	危险飞行	R_{df}
	威胁	R_{th}
	优势	R_{ad}
	时间步	R_{ts}

1) 在超视距空战中, 空战双方为保持空战优势, 高度 H_B 必须控制为 [6 000 m, 9 000 m]. 若偏离此范围则给予惩罚, 惩罚力度与偏离程度成正比, 危险飞行惩罚 R_{df} 为

$$R_{\text{df}} = \begin{cases} -0.003(6\,000 - H_B), & H_B < 6\,000 \text{ m} \\ -0.002(H_B - 9\,000), & H_B > 9\,000 \text{ m} \\ 0, & \text{其他} \end{cases} \tag{16}$$

2) 优势奖励用于反映载机相对于目标机的作战优势, 由能量优势和角度优势组成. 本文定义 V_O、V_E 分别是载机、目标机的速度, H_O、H_E 分别是载机、目标机的高度, 当 $0.6\text{ Ma} \leq V_E < V_O \leq 2.0\text{ Ma}$ 且 $6\,000\text{ m} \leq H_E < H_O \leq 9\,000\text{ m}$ 时, 能量优势为包括速度和高度的函数, 以反映载机飞机相对于目标机的能量状态:

$$R_{\text{energy}} = 0.1\frac{V_O}{V_E}\text{e}^{\frac{H_O - H_E}{1\,500}} \tag{17}$$

值得注意的是, 当不满足上述条件时, 能量状态 R_{energy} 为 0. 此外, 角度优势需要考虑载机对目标机的攻击角度 θ_{attack}:

$$R_{\text{angle}} = \begin{cases} 0.8\text{e}^{-\frac{(|\theta_{\text{attack}}| - 30)^2}{100}}, & 0° \leq |\theta_{\text{attack}}| < 60° \\ 0, & \text{其他} \end{cases} \tag{18}$$

将能量优势和角度优势结合起来, 得到优势奖励 R_{ad}:

$$R_{\text{ad}} = R_{\text{energy}} + R_{\text{angle}} \tag{19}$$

3) 威胁惩罚 R_{th} 利用来袭导弹与飞机之间距

离 d, 反映智能体在空战环境中受威胁程度:

$$R_{\text{th}}=\begin{cases}-15\cdot 2^{-\frac{d}{5}+1}, & 0\ \text{km}<d<25\ \text{km}\\ 0, & \text{其他}\end{cases}\tag{20}$$

4) 时间惩罚旨在激励智能体尽快完成任务, 避免长时间无效战斗. 惩罚与空战时间 t (单位: s) 呈线性关系:

$$R_{\text{ts}}=-0.01t\tag{21}$$

3 基于对手学习的决策方法研究

3.1 对手学习方法描述

3.1.1 模仿学习理论分析

行为克隆 (behavioral cloning, BC) 是模仿学习[25] 中一种简单而直接的方法, 其核心在于通过专家提供的状态–动作对, 在特定状态下, 训练智能体输出与专家相同或相似的动作, 训练过程主要包括以下四个步骤:

1) 数据收集. 从专家处获取一系列状态和对应的动作, 形成训练数据集.

2) 数据预处理. 对收集到的数据进行清洗和标准化, 去除噪声和异常值, 然后对状态和动作数据进行归一化处理, 以便更好地适应模型训练.

3) 模型训练. 使用监督学习算法, 将专家的状态–动作对作为训练样本, 训练智能体的策略网络.

4) 策略执行. 在测试阶段, 智能体根据训练好的策略网络输出动作.

假设专家的策略为 $\pi_E(a|s)$, BC 的目标是最小化智能体策略 $\pi(a|s)$ 与专家策略之间的差异, 并使用均方误差作为损失函数, 即

$$MSE_2=\text{E}_{(s,\,a)\sim D}\left[(\pi(a|s;\theta)-\pi_E(a|s))^2\right]\tag{22}$$

式中, D 是从专家处收集的数据集, θ 是策略网络的参数.

3.1.2 自博弈理论分析

在深度强化学习中, 自博弈[16] 可以视为一种特殊的训练机制, 通过让智能体与自身博弈, 智能体能够不断探索环境并优化策略. 自博弈的关键机制包括以下三个方面:

1) 自我对抗. 智能体通过与自身历史策略或同策略衍生的虚拟对手进行空战对抗, 在博弈过程中发现自身策略缺陷并针对性改进, 以克服依赖外部固定对手的局限性.

2) 策略迭代. 依托自我对抗产生的历史经验, 智能体通过深度强化学习算法持续迭代策略参数, 逐步提升在复杂对抗环境中的决策表现能力.

3) 动态平衡. 自博弈通过动态策略更新, 确保探索新策略与利用最优策略的平衡, 以发现更优解并实现最大回报.

假设初始化智能体的策略为 $\pi(\theta)$, 自博弈的目标是通过与自身博弈, 逐步优化策略 π, 使其在长期对抗中表现最优. 自博弈方法优化过程可以通过下式描述:

$$\pi_{\text{new}}=\arg\max_{p}\text{E}_{(s,\,a)\sim p}\Big[R(s,\,a)\Big]\tag{23}$$

式中, $R(s,\,a)$ 是智能体在状态 s 下采取动作 a 的奖励函数. 通过不断更新策略 π, 智能体能够逐步提高其在对抗环境中的表现.

3.2 专家知识规则驱动的对手策略设计

规则驱动的空战专家系统[26] 由三大核心要素构成: 机动行为库、战术知识库以及推理机. 值得注意的是, 在设计机动行为库时严格遵守超视距空战时间线的交战规则, 如表 4 所示, 其中每个距离的典型战术机动如图 4 所示.

表 4 超视距空战时间线

Table 4 Timeline of BVR

名称	符号	描述	关键距离 (km)
最小中断距离	MAR[1]	执行 Short Skate 机动	30
最小回转距离	MOR[2]	执行 Skate 机动	45
分配距离	TR[3]	进行目标分配	65
前出距离	CR[4]	前出发起攻击	80
画面距离	PR[5]	获取战场态势信息	120

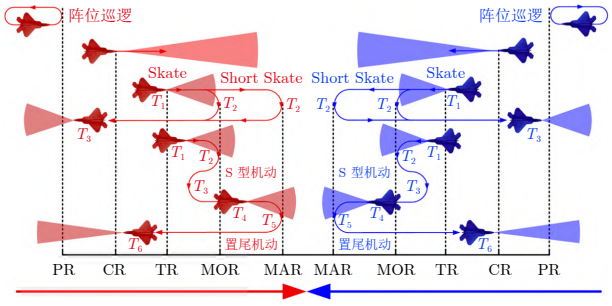


图 4 超视距空战时间线示意图

Fig. 4 A timeline diagram of BVR

规则驱动的决策框架涵盖从态势感知到决策输出的全流程模块设计, 该决策框架采用生成式规则表示方法, 每一条规则由四个核心要素组成: 态势 S、事件 E、条件 C 和动作 A. 其中, 态势 S 指敌对双方所处的实时空战相对位置状态集合, 确保决策与当前空战环境相匹配, 如表 5 所示; 事件 E 指需

[1] MAR: Minimum abort range; [2] MOR: Minimum out range;
[3] TR: Target range; [4] CR: Commit range; [5] PR: Picture range.

要专家系统响应的关键空战战术节点, 通过从连续的态势数据流中提取“决策触发点”, 使规则从“静态判断”转向“动态响应”, 如表 6 所示; 条件 C 是规则生效的约束边界, 主要依据表 4 中时间线来判断“是否执行动作”, 如表 7 所示; 动作 A 是专家系统最终输出的空战战术指令, 如表 8 所示. 此外, 每条规则包含预期状态转移, 用于明确智能体执行动作 A 后将达到的新态势 S'. 战术决策过程中, 专家系统首先对战术事件 E 进行识别, 随后由推理机对战斗机的作战条件进行评估, 涵盖剩余弹药、与敌相对位置等关键维度, 以确保系统能够精准把握飞机在当前战场环境下的作战效能. 在此基础上, 规则匹配模块从知识库中动态检索战术规则, 结合实时环境态势信息进行多维度条件匹配, 最终生成最优战术决策, 如表 9 所示. 值得注意的是, 规则驱动的空战专家系统的规则调整无优先级, 因此只具备有限的战术调整能力.

为进一步阐述表 9 中规则集在不同战场态势下的转换过程, 本文构建战术规则状态转换图, 如图 5 所示. 图中节点代表表 5 中的典型态势 S, 节点之间连线表示由典型事件 E 和典型条件 C 触发

表 5　基于规则的超视距空战决策框架典型态势
Table 5　Typical situations of rule-based BVR decision-making framework

态势 S	描述	战术分类
s_1	在 PR 外前出	基础状态
s_2	在 TR ∼ PR 之间前出	进攻性
s_3	在 MAR ∼ TR 之间前出	进攻性+防御性
s_4	返航	基础状态
s_5	任务失败	基础状态

表 6　基于规则的超视距空战决策框架典型事件
Table 6　Typical events of rule-based BVR decision-making framework

事件 E	描述	战术分类
e_1	目标雷达开机	基础状态
e_2	目标进入武器攻击区	进攻性
e_3	锁定目标	进攻性
e_4	被目标锁定	防御性
e_5	击杀目标	基础状态
e_6	被目标击杀	基础状态
e_7	目标逃逸	基础状态

表 7　基于规则的超视距空战决策框架典型条件
Table 7　Typical conditions of rule-based BVR decision-making framework

条件 C	描述	战术分类
c_1	在 MAR ∼ TR 之间	进攻性
c_2	导弹剩余数量大于 0	进攻性
c_3	在 MOR 外被攻击	进攻性
c_4	在 MAR 内被攻击	防御性

表 8　基于规则的超视距空战决策框架基础机动动作
Table 8　Basic maneuvers of rule-based BVR decision-making framework

动作 A	动作名称	描述	战术分类
a_1	匀速前飞	适用于巡航和搜索阶段	基础状态
a_2	爬升	提升高度以获得高度优势	基础状态
a_3	俯冲	降低高度以获得速度优势	基础状态
a_4	置尾机动	180° 转弯机动	防御性
a_5	Skate 机动	在 MOR 外发射后脱离	进攻性
a_6	Short Skate 机动	在 MAR 外发射后脱离	进攻性
a_7	发射	发射导弹	进攻性

表 9　基于规则的超视距空战决策框架规则集
Table 9　Rule set of rule-based BVR decision-making framework

规则编号	当前状态 S	触发事件 E	满足条件 C	执行动作 A	下一状态 S'	战术类别
1	s_1	—	—	a_1	s_1	基础状态
2	s_1	—	—	a_1	s_2	基础状态
3	s_1	e_7	—	—	s_4	基础状态
4	s_2	e_7	—	—	s_4	基础状态
5	s_2	e_1	—	a_1	s_2	进攻性
6	s_2	e_3	—	a_1	s_3	进攻性
7	s_3	e_2	c_1	a_1	s_3	进攻性
8	s_3	e_2	c_1	a_2	s_3	进攻性
9	s_3	e_3	$c_1 \cap c_2$	a_7	s_3	进攻性
10	s_3	e_3	$\neg c_2$	a_7	s_4	防御性
11	s_3	e_5	—	a_4	s_4	基础状态
12	s_3	e_4	c_3	a_5	s_3	防御性+进攻性
13	s_3	e_4	c_4	a_6	s_3	防御性+进攻性
14	s_3	e_4	c_3	a_3	s_3	防御性
15	s_3	e_4	c_4	a_4	s_3	防御性
16	s_3	e_6	—	—	s_5	基础状态

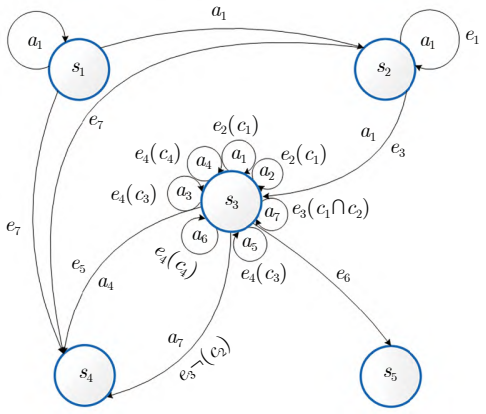


图 5　战术规则状态转换图
Fig. 5　Transition diagram of tactical rule state

的典型动作 A, 通过执行典型动作 A 进入新的态势 S'. 例如, 当载机到达态势 s_2 时, 触发典型事件 e_1, 载机开始对目标机进行搜索. 之后载机根据与目标的距离, 建立阵位进行巡逻, 维持原态势 s_2; 或者继续前出, 进入新的典型态势 s_3. $e_2(c_2)$ 表示 e_2 s.t. c_2, 其余类推.

3.3 基于对手学习的空战策略设计

3.3.1 近端策略优化算法描述

近端策略优化 (proximal policy optimization, PPO) 算法[27] 的核心思想是基于策略更新的"近端"约束, 即限制每次策略更新幅度, 避免新旧策略之间差异过大. 具体而言, PPO 通过引入新旧策略比率项:

$$r_t(\theta)=\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \tag{24}$$

将策略更新的目标函数转化为对旧策略的局部近似优化.

PPO 算法的实现流程分为数据收集、优势估计和策略优化三个阶段. 首先, 算法使用旧策略 π_{old} 与环境交互, 收集状态–动作–回报序列. 其次, 通过广义优势估计计算各时间步的优势值 A^{GAE}, 其表达式为

$$\begin{cases} A^{\text{GAE}}=\displaystyle\sum_{t=0}^{T}(\gamma\lambda_{\text{GAE}})^t\delta_{t+l} \\ \delta_t=r_t+\gamma Q(s_{t+1})-Q(s_t) \end{cases} \tag{25}$$

式中, γ 为折扣因子, λ_{GAE} 用于权衡优势估计的偏差与方差. 最后, 算法对收集的样本进行多轮次的小批量随机梯度上升优化, 更新策略参数 θ 和值函数参数. 每一轮迭代完成后, 旧策略参数 θ_{old} 被更新为当前策略参数 θ, 以开启下一轮数据收集.

3.3.2 对手学习决策框架设计

在第 2.2 节深度强化学习决策方法设计的基础上, 本文设计对手学习决策框架[21], 通过阶段性对抗训练流程, 旨在解决超视距空战中智能体面临的复杂动态决策问题. 如图 6 所示, 决策框架主要包括四个模块, 分别为专家系统模块、模仿学习模块、自博弈模块、对抗测试模块. 其中, 专家系统模块储存空战领域的专家知识和经验, 输出丰富的数据集; 模仿学习模块接收并处理数据集, 输出初始化的经验池; 自博弈模块接收经验池中的数据进行自我对抗训练, 输出优化的智能策略; 对抗测试模块提供具有战术差异性的空战专家系统, 在高保真的对抗环境中测试智能化策略, 具体过程如算法 1 所示.

算法 1. 对手学习决策算法

输入. 专家知识经验, 学习率 l_r, 模仿学习迭代回合数 P, 环境 E, 初始策略参数 θ, 自博弈训练回合数 N, 更新间隔 U.

1. 用专家系统模块生成式 (26) 中数据集 D_{expert}
2. 进入模仿学习模块, 初始化策略 $\pi(\theta)$, 经验池 H
3. 对于每一轮训练 episode $=1,2,\cdots,P$, 执行以下代码:
4. 从 D_{expert} 中随机采样生成样本 B
5. 初始化损失 $L=0$
6. 判定 B 未完全遍历时, 执行以下代码:
7. 预测动作 $a'\leftarrow\pi(a|s;\theta)$
8. 计算损失 $L\leftarrow L+||a'-a||^2$
9. 循环结束
10. 使用式 (27) 将损失 L 平均化处理
11. 计算损失函数关于 θ 的梯度 $\nabla_\theta L$
12. 用式 (28) 梯度下降法更新策略, 并更新式 (29) 中经验池 H
13. 循环结束
14. 进入自博弈模块, 初始化策略 $\pi(\theta)$, 并使对手策略 $\pi_{\text{opponent}}(\theta_{\text{opponent}})=\pi(\theta)$
15. 对于每一轮训练 episode $=1,2,\cdots,N$, 执行以下代码:
16. 初始化环境 E 为 E_0
17. 判定 E_0 未达到终止状态时, 执行以下代码:
18. 当判定满足条件 episode mod $2=1$ 时:
19. 生成动作 $a_t\leftarrow\pi_{\text{opponent}}(E_0;\theta_{\text{opponent}})$
20. 否则, 执行以下代码:
21. 生成动作 $a_t\leftarrow\pi(E_0;\theta)$
22. 条件结束
23. 生成新状态、奖励、结束标志 E_1, r_t, done
24. 更新经验池 H.add(E_0, a_t, r_t, E_1, done)
25. 更新环境状态 $E_0\leftarrow E_1$, 进入下一状态继续交互
26. 当前回合的环境交互循环结束
27. 从 H 中采样生成样本 B
28. 用式 (31) 更新智能体策略参数 θ
29. 当判定满足条件 episode mod $U=0$ 时:
更新对手策略参数 $\theta_{\text{opponent}}\leftarrow\theta$
30. 条件结束
31. 循环结束
32. 生成最终策略参数 θ^*, 并在对抗测试模块进行评估

1) 专家系统模块首先将空战领域专家知识和经验转化为规则库, 接着在模拟环境中仿真生成大量状态–动作数据, 为模仿学习提供可靠的数据集. 本文定义专家系统数据集为

$$D_{\text{expert}}=\left\{(s_i,\ a_i)\right\}_{i=1}^{N} \tag{26}$$

式中, s_i 表示状态, a_i 表示动作, N 为数据总数.

2) 模仿学习模块首先对数据集进行预处理, 对

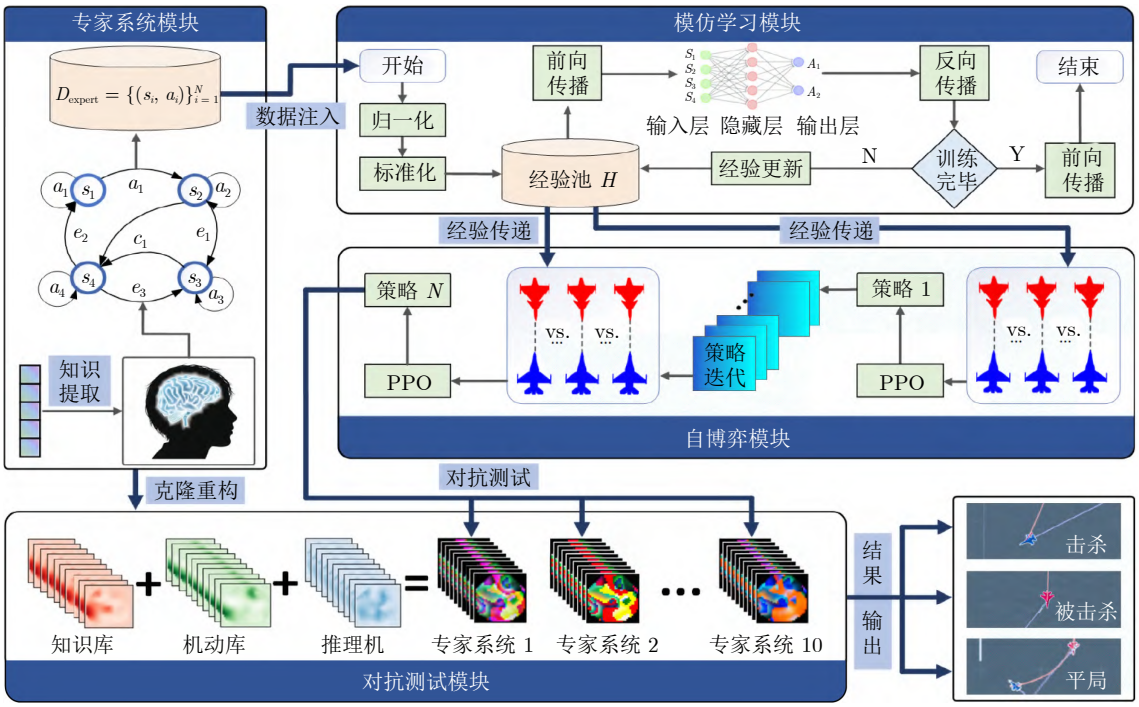


图 6　对手学习决策框架图

Fig. 6　Framework diagram of adversarial learning decision

每个 $(s,\ a)$ 的特征进行线性归一化, 使特征值域映射到 $[0,\ 1]$ 区间. 接着, 针对不同量纲的特征, 该模块采用 Z-Score 标准化方法, 使其转化为服从 N(0, 1) 分布的数据. 其次, 设计神经网络架构作为策略网络 π_θ, 该网络以状态 s_i 作为输入, 以动作 a_i 作为输出. 为防止过拟合, 在隐藏层之间添加 Dropout 层, 设置参数为 Dropout(0.2), 以随机丢弃部分神经元的输出. 然后, 定义均方误差函数损失, 来衡量策略网络输出的动作与专家动作之间的差异:

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}||\pi_\theta(s_i)-a_i^*||^2 \tag{27}$$

式中, N 是数据样本数量, $\pi_\theta(s_i)$ 是策略网络在状态 s_i 下的输出动作, a_i^* 是专家在状态 s_i 下执行的动作. 最后, 通过反向传播算法计算 $L(\theta)$ 关于网络参数 θ 的梯度 $\nabla_\theta L(\theta)$, 并更新网络参数 θ:

$$\theta\leftarrow\theta-l_r\nabla_\theta L(\theta) \tag{28}$$

式中, l_r 是学习率. 该模块训练完毕后得到经验池 H:

$$H=\left\{(s_i,\ a_i,\ r_i,\ s_{i+1})\right\}_{i=1}^{N} \tag{29}$$

式中, r_i 为状态 s_i 下执行动作 a_i 后获得的奖励, s_{i+1} 为下一个状态.

3) 自博弈模块采用 Actor-Critic 架构[28−29], 在策略更新时, 通过重要性采样方法对数据进行采样, 具体过程为: 首先, 从与自身对抗生成的经验池 H 中随机选择 M 个样本 $\{(s_i,\ a_i,\ r_i,\ s_{i+1})\}_{i=1}^{M}$, 接着计算样本 j 重要性权重:

$$\rho_j=\frac{\pi_{\rm new}(a_j|s_j)}{\pi_{\rm old}(a_j|s_j)} \tag{30}$$

式中, $\pi_{\rm new}$ 和 $\pi_{\rm old}$ 分别为更新后策略和旧策略. 然后, 使用 PPO 算法中损失函数更新智能体策略:

$$L^{\rm PPO}=\hat{E}_j\Big[\min(\rho_j A^{\pi_{\rm old}}(s_j,\ a_j),\ \mathrm{clip}(\rho_j,\ 1-\varepsilon,\ 1+\varepsilon)A^{\pi_{\rm old}}(s_j,\ a_j))\Big] \tag{31}$$

式中, ε 为剪切参数; $A^{\pi_{\rm old}}$ 为优势函数, 由 Critic 网络进行估计. 值得注意的是, 每间隔 50 轮, 该模块从经验池 H 中引入经验数据, 驱动生成新的策略, 避免策略在固定对手上过拟合, 从而提升训练效率并增强策略的泛化能力. 此外, 当智能体与采取相同策略的对手智能体进行自博弈对抗时, 将冻结对手智能体, 以增强训练环境的稳定性.

4) 本文基于 3.2 节中的专家系统框架, 并结合表 4 中关键距离, 从雷达探测、导弹发射、能量机动三个角度, 在对抗测试模块中构建 10 种具有战术差异性的空战专家系统, 用于对智能化方法进行全因子对抗测试, 以评估智能化方法的性能, 具体如表 10 所示.

表 10 10 种空战专家系统设计
Table 10 Ten designs of expert system for air combat

类型	名称	功能特点
偏好攻击型	对手 1	首轮双弹打击
	对手 2	高速接近突袭
	对手 3	保持高位压制
攻防均衡型	对手 4	能量–射程均衡
	对手 5	雷达扫描–锁定优化
	对手 6	高度–速度攻防转换
	对手 7	多弹压制–复合规避
偏好防御型	对手 8	TR 巡逻规避
	对手 9	MOR 巡逻规避
	对手 10	持续转冷规避

偏好攻击型对手的核心特点是: 导弹进入末制导前, 载机保持稳定飞行不实施机动, 以提升导弹命中率, 具体包括 3 种类型. 首轮双弹打击系统牺牲防御机动性, 当目标进入导弹最大攻击包线时, 立即发射两枚导弹; 高速接近突袭系统以对手 1.2 倍速度接近; 保持高位压制系统要求载机持续爬升至目标机上空 1 000 m 处, 利用重力加速俯冲发射导弹. 攻防均衡型对手的核心特点是: 兼顾进攻效能与防御安全, 通过动态调整策略伺机击杀目标, 具体包括 4 种类型. 能量–射程均衡系统是指当目标处于载机导弹最大攻击包线时, 载机立即发射, 当载机处于目标导弹不可逃逸区时, 执行“径向+俯冲”机动脱离; 雷达扫描–锁定优化系统间歇性切换雷达扫描模式, 防止目标预测锁定节奏, 当被目标机雷达持续锁定时, 立即进行偏置机动; 高度–速度攻防转换系统攻击时, 爬升用动能换势能, 防御时, 俯冲将势能转为动能; 多弹压制–复合规避系统要求载机攻击时, 第一枚导弹发射后, 若目标规避则 3 s 内补射第二枚封堵路径, 防御时, 执行“S 型+置尾”复合机动. 偏好防御型对手的核心特点是: 优先执行防御机动, 力求保证载机安全, 并在防御中寻找反攻时机, 具体包括 3 种类型. TR 巡逻规避系统要求载机在目标 TR 距离外保持巡逻阵位, 不主动出击; MOR 巡逻规避系统要求载机在目标 MOR 距离外保持巡逻阵位, 不主动出击; 持续转冷规避系统要求载机保持在目标导弹不可逃逸区外, 并持续向目标冷边飞行, 到达冷边时反攻.

4 超视距空战仿真实验

4.1 想定描述

在构建高保真超视距空战场景基础上, 为系统研究单机状态下的超视距空战问题, 本文设计以下涵盖典型空战阶段的作战任务想定.

红方和蓝方在指定任务区内展开对抗, 目的是夺取该区域的制空权, 双方使用机动性能相同的战机, 且机载传感器和武器系统的性能指标一致, 具体型号及装备性能参数见第 1.3 节. 在任务初始阶段, 红蓝双方在任务区内处于“空中待命”状态, 红方由智能化机动决策算法操控, 蓝方由基于规则的专家系统操控, 双方各挂载 2 枚中距空空导弹. 双方接收战斗指令后, 开始执行夺取“制空权”任务; 在雷达未探测到目标时, 双方战机在任务区内进行警戒巡逻; 当雷达探测到目标后, 双方立即进入战术对抗阶段, 且战斗在一方战机被击落或达到最大仿真时长时终止, 并根据双方的生存状态和作战效能判定胜负. 该过程包含三个关键子阶段:

1) 跟踪阶段: 载机根据雷达提供的目标方位、距离和高度等信息, 结合目标性能参数和当前战场态势, 预测出最佳拦截位置.

2) 攻击阶段: 在目标进入导弹有效攻击区后, 载机将迅速调整飞行姿态和航向, 以获得最大能量优势和角度优势, 伺机发起攻击.

3) 规避阶段: 当目标攻击载机时, 载机将执行一系列机动动作, 如 Crank 机动或偏置机动, 以迅速改变航向和飞行姿态, 拉开与来袭导弹距离.

4.2 实验设计

本文中, CPU 采用 AMD Ryzen7 5700U, Radeon Graphics, 仿真环境与文献 [21] 保持一致, 开发工具为 Pycharm 2024.3.3, 基于 Python 语言实现. 此外, 本文采用 Tacview 仿真分析工具, 以可视化飞行数据.

在仿真实验中, 通过构建载机底层控制指令模型、导弹制导律模型与雷达态势更新模型, 建立了完整的超视距空战仿真环境, 如图 7 所示. 图 7 中, ATA (aspect angle) 表示目标进入角, AOT (angle-off-tail) 表示目标偏离角. 该仿真环境支持红蓝双方智能决策算法的对抗验证, 为战术决策系统的性能评估提供了实验平台, 训练参数设置如下: 学习

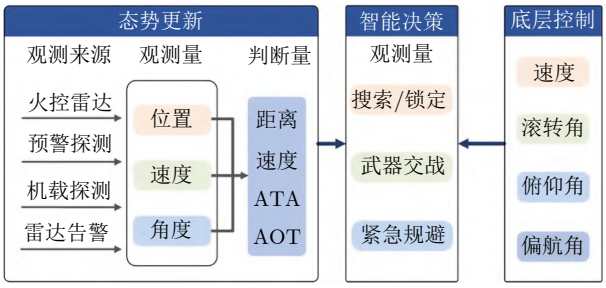


图 7 仿真环境实验框架
Fig. 7 Experiment framework of simulation environment

率为 0.000 1, 折扣因子为 0.9, 模仿学习模块迭代回合数为 200 轮, 自博弈模块训练回合数为 500 轮, 且每 1 000 个时间步取一个记录点, 以确保智能体有足够的时间学习策略, 同时平衡学习速度和稳定性.

仿真实验分为两个阶段: 第 1 阶段为训练结果分析阶段, 其核心目的是, 通过将基于对手学习的方法与基于 DQN 的深度强化学习方法进行对比实验, 分析奖励值曲线趋势、胜败率和平局率等指标, 验证对手学习方法的有效性; 第 2 阶段为测试效果分析阶段, 其核心目的是验证对手学习方法能在面对未知的专家系统争夺制空权时, 是否具备一定的泛化能力.

值得注意的是, 本文选取文献 [26] 中的专家系统作为实验验证的对手原型, 主要基于以下两方面原因: 1) 该系统当前已在空战领域形成成熟应用框架, 采用 SECA 规则体系, 通过解析空战态势、事件触发条件及规则匹配机制, 可为战机提供结构化机动决策; 2) 该系统的规则库具有高度可解释性, 便于与深度强化学习模型的决策过程进行对比分析.

4.3　实验结果与性能分析

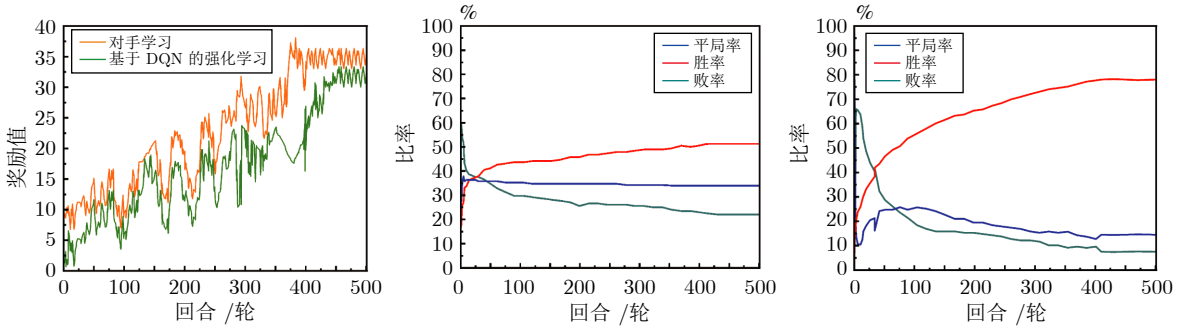
4.3.1　训练结果分析阶段

如图 8(a) 所示, 在与第 3.2 节专家系统对抗训练过程中, 对手学习和基于 DQN 的深度强化学习方法呈现出不同的奖励值变化趋势, 图 8(a) 中横轴表示训练轮数, 纵轴表示奖励值. 在训练初期, 两种方法的奖励值均呈现上升趋势, 表明两者通过学习持续更新策略, 其中对手学习方法的奖励值增速明显快于基于 DQN 的深度强化学习方法, 这表明模仿学习在对手学习方法的早期训练中发挥重要作用; 经过 150 轮训练迭代后, 对手学习方法的奖励值继续保持平稳增长, 而基于 DQN 的深度强化学习方法则呈现阶梯式上升的特征, 并伴随明显波动, 其奖励值最大波动幅度达 12.7%, 反映出基于 DQN 的深度强化学习方法在策略更新过程中存在不稳定性; 经过 400 轮训练迭代后, 对手学习方法的奖励值峰值突破并最终收敛于 35, 基于 DQN 的深度强化学习方法则在奖励值继续增长后收敛, 最终未突破 35. 这一结果不仅验证了对手学习在策略收敛性方面的优势, 更揭示了其在探索–利用平衡机制上具备优越性.

本文引入监测机制, 在每轮训练结束后记录当前博弈状态, 最终生成如图 8(b) 和图 8(c) 所示的比率演化曲线. 随训练轮数的增加, 对手学习的胜率呈递增趋势, 并在完成 150 轮训练后快速收敛至 80%, 这一胜率显著高于基于 DQN 的深度强化学习方法. 此外, 当两种方法均达到收敛状态时, 对手学习方法败率更低, 并且其平局率相较于基于 DQN 的深度强化学习方法降低 57%. 这一结果进一步证实, 对手学习方法中模仿学习模块能有效识别对手策略缺陷. 综上所述, 在策略优化效率和收敛速度方面, 对手学习方法均优于传统深度强化学习方法.

4.3.2　测试效果阶段

测试效果阶段主要是为系统验证对手学习算法在以下两个维度的泛化能力: 1) 对极端战术倾向的鲁棒性; 2) 对不同战术风格的适应性. 该阶段采用对照实验设计, 通过构建涵盖“攻击型–均衡型–防御型”的战术风格连续谱系, 使两种智能化方法分别与 10 种专家系统进行组合对抗, 每组组合进行 120 轮独立重复实验, 共计获得 2 400 组对抗数据, 统计的胜率结果如图 9 所示. 由图可知, 对于对手 3、4、5、6、8、10, 对手学习方法都能保持超过 70% 的较高胜率; 对于对手 1、2、7、9, 对手学习方



(a) 奖励值对比图
(a) The comparison chart of rewards

(b) 深度强化学习决策下的比率
(b) The rates under deep reinforcement learning decision-making

(c) 对手学习决策下的比率
(c) The rates under adversarial learning decision-making

图 8　训练结果对比图

Fig. 8　The comparison chart of training results

法也能维持较高胜率. 在特定场景下, 虽然基于DQN 的深度强化学习方法胜率超过对手学习方法,

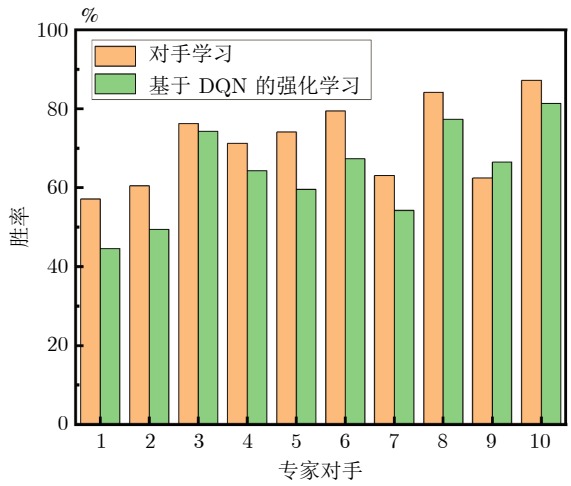


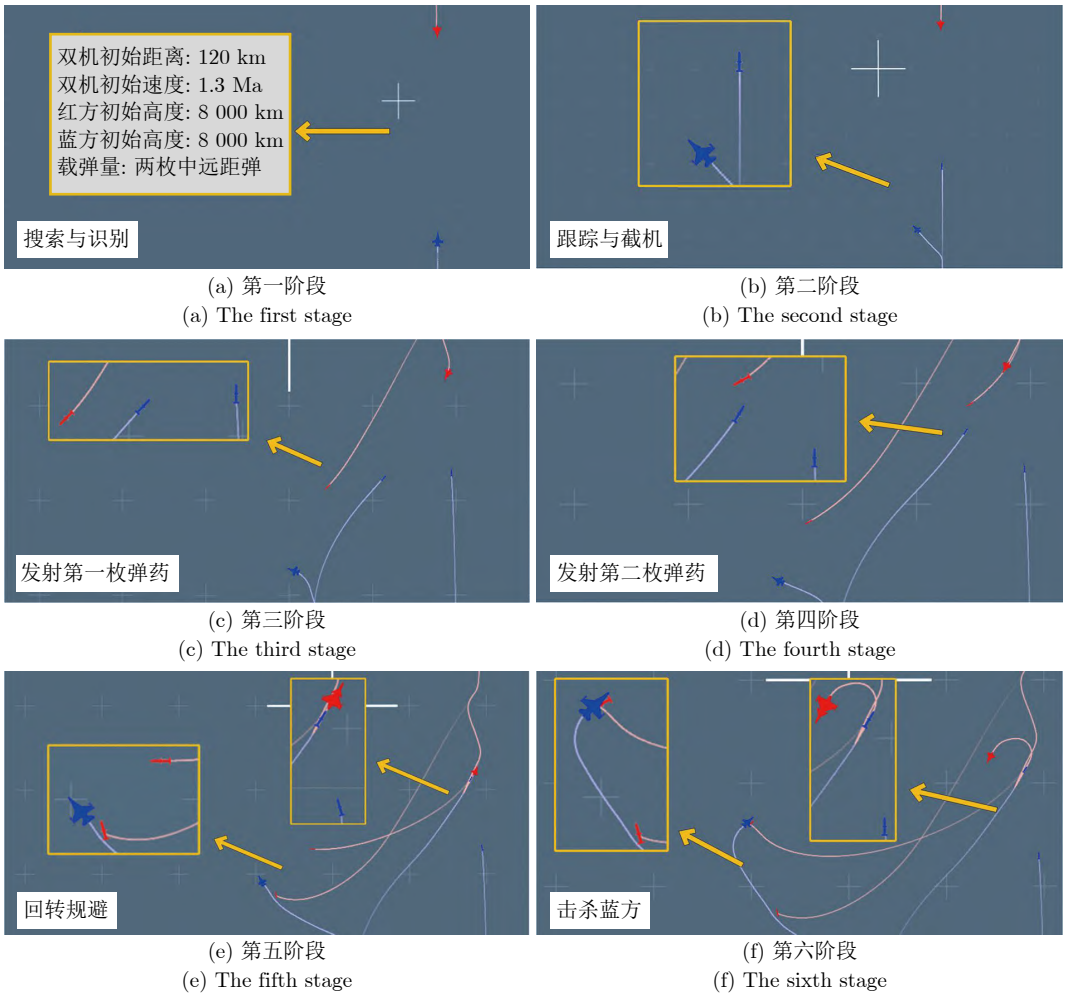
图 9 对抗胜率图

Fig. 9 Chart of win rates in confrontation

如在与对手 9 的对抗中, 但值得注意的是, 对手学习方法在对抗对手 9 时仍能保持 63% 的基准胜率. 综上所述, 对手学习方法在对抗均衡型对手策略时维持较高且稳定的胜率, 同时在应对偏好防御型和偏好攻击型, 即具有极端战术倾向的对手时, 仍展现出良好的战术适应性, 验证了对手学习方法在“攻击型–均衡型–防御型”战术连续谱系中具备良好的泛化能力.

为展示此次实验超视距空战具体对抗过程, 本文选取对手学习抗击“攻防均衡”型对手 7 的过程中一次案例, 并结合表 4 对双方执行的关键决策进行分析, 如图 10 所示.

在超视距空战初始阶段, 红方和蓝方战斗机在高空以 1.3 Ma 的速度飞行, 双方高度均为 8000 km, 且均处于对手的 PR 处. 双方开启机载雷达, 对周边空域进行扫描, 通过敌我识别系统搜索识别对手. 蓝方先于红方识别目标, 执行“先敌发现, 先敌攻击”战术. 在距离红机 TR 处, 蓝方锁定红方, 此时



(a) 第一阶段 (a) The first stage

(b) 第二阶段 (b) The second stage

(c) 第三阶段 (c) The third stage

(d) 第四阶段 (d) The fourth stage

(e) 第五阶段 (e) The fifth stage

(f) 第六阶段 (f) The sixth stage

图 10 基于对手学习的对抗过程

Fig. 10 The adversarial process based on adversarial learning

判定蓝方有剩余弹药, 于是发射第一枚导弹进行攻击. 此时, 红方与蓝方的距离远大于 MAR, 红方继续前出, 并伺机攻击. 在成功锁定蓝方后, 红方进入攻击环节. 在双机距离 55 km 时, 蓝方再次锁定红方, 发射第二枚导弹, 此时红方选择一枚中程空空导弹进行发射. 在红方前出至 MOR 处, 红方根据雷达提供的目标位置信息, 立刻发射第二枚导弹. 红方在 MAR 前对第二枚导弹完成制导并立刻回转规避, 获得规避导弹奖励. 当蓝方感知到被红方导弹锁定, 且准备执行紧急置尾机动时, 为时已晚, 最终被红方导弹击杀.

上述结果表明, 在复杂动态环境中, 面对同一专家系统, 基于对手学习的超视距空战决策方法比基于 DQN 的方法表现更加出色, 验证了本文提出的对手学习决策框架的有效性和泛化性.

5 结束语

针对超视距空战博弈问题, 本文在高保真空战仿真环境中, 构建基于 DQN 的深度强化学习空战决策架构, 并对此架构进行优化, 形成基于对手学习的空战决策框架. 该框架针对超视距空战中智能体面临的复杂动态决策问题, 融合模仿学习与自博弈机制, 采用 PPO 算法, 针对性弥补了传统深度强化学习对动态对抗环境收敛慢和泛化性差的缺点.

此外, 本文引入专家系统进行对抗训练测试, 首先构建一种专家知识规则驱动的对手策略, 分别对两种空战决策架构进行对抗训练, 接着对该专家系统进行重构, 建立“攻击型–均衡型–防御型”的战术风格连续谱系, 分别对两种空战决策架构进行对抗测试. 训练结果分析阶段表明, 对手学习方法在收敛速度和胜率上均优于传统的深度强化学习方法; 测试效果阶段表明, 对手学习方法在多场景对抗中能维持较高且稳定的胜率, 能适应性地对抗多种战术倾向的对手, 具备泛化性.

参考文献

1 Wu Yan, Ma Jun. The India-Pakistan conflict reflects the characteristics of future warfare [Online], available: https://mil.huanqiu.com/article/4MdrTXqZu4e, May 12, 2025
(武彦, 马俊. 印巴冲突折射未来战争特征 [Online], available: https://mil.huanqiu.com/article/4MdrTXqZu4e, 2025-05-12)

2 Sun Zhi-Xiao, Yang Sheng-Qi, Piao Hai-Yin, Bai Cheng-Chao, Ge Jun. A survey of air combat artificial intelligence. *Acta Aeronautica et Astronautica Sinica*, 2021, **42**(8): Article No. 525799
(孙智孝, 杨晟琦, 朴海音, 白成超, 葛俊. 未来智能空战发展综述. 航空学报, 2021, **42**(8): Article No. 525799)

3 Ren Hao, Ma Ya-Jie, Jiang Bin, Liu Cheng-Rui. Fault-tolerant control for spacecraft formation with communication faults based on zero-sum differential game. *Acta Automatica Sinica*, 2025, **51**(1): 174−185
(任好, 马亚杰, 姜斌, 刘成瑞. 基于零和微分博弈的航天器编队通信链路故障容错控制. 自动化学报, 2025, **51**(1): 174−185)

4 Herrala O, Terho T, Oliveira F. Risk-averse decision strategies for influence diagrams using rooted junction trees. *Operations Research Letters*, 2025, **61**: Article No. 107308

5 Shi Tong-Yu, Wang Hao, Wang You-Kun, Lv Mao-Long. Simulation of game-theoretic decision-making for beyond-visual-range combat with UCAVs [Online], available: http://link.cnki.net/urlid/11.3019.tj.20250828.1438.008, December 3, 2025
(史桐雨, 王昊, 王酉琨, 吕茂隆. 无人作战飞机超视距空战博弈对抗决策仿真 [Online], available: http://link.cnki.net/urlid/11.3019.tj.20250828.1438.008, 2025-12-03)

6 Lv Mao-Long, Ding Chen-Bo, Han Hao-Ran, Duan Hai-Bin. Autonomous perception-planning-control strategy based on deep reinforcement learning for unmanned aerial vehicles. *Acta Automatica Sinica*, 2025, **51**(6): 1305−1319
(吕茂隆, 丁晨博, 韩浩然, 段海滨. 基于深度强化学习的无人机自主感知–规划–控制策略. 自动化学报, 2025, **51**(6): 1305−1319)

7 Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623
(施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623)

8 Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, Huang Ting-Wen, Yang Chun-Hua, Gui Wei-Hua. Survey on multi-agent reinforcement learning for control and decision-making. *Acta Automatica Sinica*, 2025, **51**(3): 510−539
(罗彪, 胡天萌, 周育豪, 黄廷文, 阳春华, 桂卫华. 多智能体强化学习控制与决策研究综述. 自动化学报, 2025, **51**(3): 510−539)

9 Guo Wan-Chun, Xie Wu-Jie, Yin Hui, Dong Wen-Han. Research on UAV anti-pursing maneuvering decision based on improved twin delayed deep deterministic policy gradient method. *Journal of Air Force Engineering University (Natural Science Edition)*, 2021, **22**(4): 15−21
(郭万春, 解武杰, 尹晖, 董文瀚. 基于改进双延迟深度确定性策略梯度法的无人机反追击机动决策. 空军工程大学学报 (自然科学版), 2021, **22**(4): 15−21)

10 Sun Shi-Bin, Wang Qing-Ling. Large scale UAV cluster adversarial game based on improved multi-agent reinforcement learning. *Journal of Naval Aviation University*, 2025, **40**(4): 528−538
(孙世彬, 王庆领. 基于改进多智能体强化学习的大规模无人机集群博弈对抗. 海军航空大学学报, 2025, **40**(4): 528−538)

11 Ou Yang, Xu Yang, Zhang Jin-Peng, Luo De-Lin. UAV air combat dueling and double deep reinforcement learning maneuver adversarial decision making. *Journal of Xiamen University (Natural Science)*, 2022, **61**(6): 975−985
(欧洋, 徐扬, 张金鹏, 罗德林. 无人机空战的竞争与双重深度强化学习机动对抗决策. 厦门大学学报 (自然科学版), 2022, **61**(6): 975−985)

12 Wu Yi-Jia, Lai Jun, Chen Xi-Liang, Cao Lei, Xu Peng. Research on the application of reinforcement learning algorithm in decision support of beyond-visual-range air combat. *Aero Weaponry*, 2021, **28**(2): 55−61
(吴宜珈, 赖俊, 陈希亮, 曹雷, 徐鹏. 强化学习算法在超视距空战辅助决策上的应用研究. 航空兵器, 2021, **28**(2): 55−61)

13 Kong Wei-Ren, Zhou De-Yun, Zhao Yi-Yang, Yang Wan-Sha. Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play. *Control Theory and Applications*, 2022, **39**(2): 352−362
(孔维仁, 周德云, 赵艺阳, 杨婉莎. 基于深度强化学习与自学习的多无人机近距空战机动策略生成算法. 控制理论与应用, 2022, **39**(2): 352−362)

14 Wang Yi-Song, Zhao Ming-Hui, Zhang Xue-Bo. ASM^2: Multi-agent multi-opponent game algorithm for joint sea-air scenarios. *Control Theory and Applications*, 2025, **42**(7): 1275−1284
(王臆淞, 赵铭慧, 张雪波. ASM^2: 面向海空联合场景的多对手多智能体博弈算法. 控制理论与应用, 2025, **42**(7): 1275−1284)

15 Piao H Y, Sun Z X, Meng G L, Chen H C, Qu B H, Lang K J, et al. Beyond-visual-range air combat tactics auto-generation by reinforcement learning. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Glasgow, UK:

IEEE, 2020. 1–8

16 Shan Sheng-Zhe, Zhang Wei-Wei. Air combat intelligent decision-making method based on self-play and deep reinforcement learning. *Acta Aeronautica et Astronautica Sinica*, 2024, **45**(4): Article No. 328723
(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. 航空学报, 2024, **45**(4): Article No. 328723)

17 Zhou Pan, Huang Jiang-Tao, Zhang Sheng, Liu Gang, Shu Bo-Wen, Tang Ji-Gang. Intelligent air combat decision making and simulation based on deep reinforcement learning. *Acta Aeronautica et Astronautica Sinica*, 2023, **44**(4): Article No. 126731
(周攀, 黄江涛, 章胜, 刘刚, 舒博文, 唐骥罡. 基于深度强化学习的智能空战决策与仿真. 航空学报, 2023, **44**(4): Article No. 126731)

18 Li Yin-Tong, Han Tong, Sun Chu, Wei Zheng-Lei. An optimization method of air combat situation assessment function based on inverse reinforcement learning. *Fire Control and Command Control*, 2019, **44**(8): 101–106
(李银通, 韩统, 孙楚, 魏政磊. 基于逆强化学习的空战态势评估函数优化方法. 火力与指挥控制, 2019, **44**(8): 101–106)

19 Shi Y Y, Li J, Lv M L, Wang N, Zhang B Y. Distributed consensus control for 6-DOF fixed-wing multi-UAVs in asynchronously switching topologies. *IEEE Transactions on Vehicular Technology*, 2025, **74**(4): 5649–5663

20 Liang Yu-Feng, Zhao Jing-Chao, Liu Wang-Kui, Wang Lei, Wang Shi-Peng, Ruan Shi-Long. Air combat guidance method based on top rolling optimization and bottom tracking. *Systems Engineering and Electronics*, 2023, **45**(9): 2866–2872
(梁玉峰, 赵景朝, 刘旺魁, 王雷, 王世鹏, 阮仕龙. 基于顶层滚动优化和底层跟踪的空战导引方法. 系统工程与电子技术, 2023, **45**(9): 2866–2872)

21 Wang W F, Ru L, Lv M L, Hou Y Q, Yin H. Exploring hierarchical hybrid autonomous maneuvering decision-making architecture in beyond visual range air combat. *IEEE Transactions on Vehicular Technology*, 2025, **74**(10): 15491–15506

22 Wang W F, Ru L, Lv M L, Mo L. Dynamic and adaptive learning for autonomous decision-making in beyond visual range air combat. *Aerospace Science and Technology*, 2025, **163**: Article No. 110327

23 Xu Y H, Wei Y R, Jiang K Y, Chen L, Wang D, Deng H B. Action decoupled SAC reinforcement learning with discrete-continuous hybrid action spaces. *Neurocomputing*, 2023, **537**: 141–151

24 Han H R, Cheng J, Lv M L, Duan H B. Augmenting the robustness of tactical maneuver decision-making in unmanned aerial combat vehicles during dogfights via prioritized population play with diversified partners. *IEEE Transactions on Aerospace and Electronic Systems*, 2025, **61**(5): 12892–12907

25 Jiang C R, Wang H, Ai J L. Autonomous maneuver decision-making algorithm for UCAV based on generative adversarial imitation learning. *Aerospace Science and Technology*, 2025, **164**: Article No. 110313

26 Hou Y Q, Liang X L, Zhang J Q, Lv M L, Yang A W. Hierarchical decision-making framework for multiple UCAVs autonomous confrontation. *IEEE Transactions on Vehicular Technology*, 2023, **72**(11): 13953–13968

27 Chen C, Song T, Mo L, Lv M L, Lin D F. Autonomous dogfight decision-making for air combat based on reinforcement learning with automatic opponent sampling. *Aerospace*, 2025, **12**(3): Article No. 265

28 Chen Can, Mo Li, Zheng Duo, Cheng Zi-Heng, Lin De-Fu. Cooperative attack-defense game of multiple UAVs with asymmetric maneuverability. *Acta Aeronautica et Astronautica Sinica*, 2020, **41**(12): Article No. 324152
(陈灿, 莫雳, 郑多, 程子恒, 林德福. 非对称机动能力多无人机智能协同攻防对抗. 航空学报, 2020, **41**(12): Article No. 324152)

29 Chen C, Song T, Mo L, Lv M L, Yu Y N. Scalable cooperative decision-making in multi-UAV confrontations: An attention-based multiagent actor-critic approach. *IEEE Transactions on Aerospace and Electronic Systems*, 2025, **61**(6): 15195–15209

吕茂隆 国家级青年人才, 空军工程大学副教授, 荷兰代尔夫特理工大学博士. 主要研究方向为集群无人机协同打击, 有人–无人协同空战, 智能空战. 本文通信作者.
E-mail: maolonglv@163.com
(**LV Mao-Long** National-Level Young Talent, associate professor at Air Force Engineering University, Ph.D. at Delft University of Technology, the Netherlands. His research interests include cooperative strike by swarming unmanned aerial vehicles, manned-unmanned collaborative air combat, and intelligent air combat. Corresponding author of this paper.)

王金河 空军工程大学硕士研究生. 主要研究方向为智能空战.
E-mail: goldenriver2025@163.com
(**WANG Jin-He** Master student at Air Force Engineering University. His main research interest is intelligent air combat.)

韩浩然 电子科技大学信息与通信工程学院博士研究生. 主要研究方向为强化学习技术与应用.
E-mail: hanadam@163.com
(**HAN Hao-Ran** Ph.D. candidate at the School of Information and Communication Engineering, University of Electronic Science and Technology. His research interests include reinforcement learning techniques and applications.)

丁晨博 空军工程大学博士研究生. 主要研究方向为有人–无人协同空战.
E-mail: chenbo_ding2024@163.com
(**DING Chen-Bo** Ph.D. candidate at Air Force Engineering University. His main research interest is manned-unmanned collaborative air combat.)

万路军 空军工程大学副教授. 主要研究方向为智能空域管理, 空域冲突检测与冲突解除, 空间网格管理.
E-mail: pandawlj@126.com
(**WAN Lu-Jun** Associate professor at Air Force Engineering University. His research interests include intelligent airspace management, airspace conflict detection and deconfliction, and spatial grid management.)