DOI:10.11918/202511098

分层强化学习驱动无人作战飞机空战自主决策

王文飞[1,2,3],茹 乐[1,2,3],吕茂隆[2,4],莫 雳[5]

(1.空军工程大学 装备管理与无人机工程学院,西安 710051;2.无人飞行器技术全国重点实验室(西北工业大学),西安 710051;3.空军工程大学 陕西高校青年创新团队,西安 710051;4.空军工程大学 空管领航学院,西安 710051;5.北京理工大学 空天科学与技术学院,北京 100081)

摘 要:为应对未来智能化空战对无人机自主决策能力提出的迫切需求,解决在复杂高动态环境下传统方法决策效率低、策略鲁棒性差的问题,提出一种结合分层强化学习与两阶段训练范式的无人机空战自主决策新框架。首先,设计了一种分层决策架构,将端到端的庞大决策空间进行策略性解耦:上层利用深度Q网络(DQN)学习高层战术决策,将原始连续动作空间压缩为有限的离散战术指令集,有效降低了学习复杂度;下层通过多模式控制和状态驱动机制,将高层战术意图实时精确映射为具体的飞控指令序列,确保决策的可执行性与物理安全性。其次,提出并实现了一种"专家模仿-自博弈进化"的两阶段训练方法:第1阶段,构建了基于SECA规则架构的专家对手,智能体通过与其对抗进行模仿学习,获得高质量的初始策略和先验知识;第2阶段,智能体转入虚拟自博弈环境,通过与其历史最优版本持续对抗,驱动策略不断突破初始专家知识的局限。最后,搭建高保真空战仿真环境对本文方法性能进行实验验证。研究结果表明,与传统DQN及其他基准算法相比,所提方法可以有效提升智能体学习训练的奖励收敛性能和最终对抗胜率,有效提升无人机智能体空战对抗水平。

关键词:深度强化学习;无人机;自主空战决策;专家对抗;自博弈

中图分类号:V249 **文献标志码**:A **文章编号**:0367-6234(2026)08-0125-16

Autonomous decision-making of unmanned combat aircraft in air combat driven by hierarchical reinforcement learning

Wang Wenfei[1,2,3], Ru Le[1,2,3], Lü Maolong[2,4], Mo Li[5]

(1. School of Equipment Management and Unmanned Aerial Systems Engineering, Air Force Engineering University, Xi'an 710051, China; 2. National Key Laboratory of Unmanned Aerial Vehicle Technology (Northwestern Polytechnical University), Xi'an 710051, China; 3. The Youth Innovation Team of Shaanxi Universities, Air Force Engineering University, Xi'an 710051, China; 4. School of Air Traffic Control and Navigation, Air Force Engineering University, Xi'an 710051, China; 5. School of Aerospace Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: To address the urgent demand for the autonomous decision-making capability of unmanned aerial vehicles in future intelligent air combat and to solve the bottlenecks of low decision-making efficiency and poor strategy robustness of traditional methods in complex and highly dynamic environments, this paper proposed a novel framework for the autonomous decision-making of unmanned aerial vehicles in air combat combining hierarchical reinforcement learning and a two-phase training paradigm. Firstly, this paper designed a hierarchical decision-making architecture to strategically decouple the vast end-to-end decision space. The upper layer utilized a deep Q-network (DQN) to learn high-level tactical decisions, compressing the original continuous action space into a finite set of discrete tactical commands, which effectively reduced the learning complexity; the lower layer mapped the high-level tactical intentions into specific flight control command sequences in real time through multi-mode control and state-driven mechanisms, which ensured the executability and physical safety of the decisions. Secondly, a two-phase training method of "expert imitation and self-game evolution" was proposed and implemented. In the first phase, an expert opponent based on the SECA rule architecture was constructed, and the agent obtained high-quality initial policies and prior knowledge through imitation learning by fighting against it; in the second phase, the agent transitioned into a virtual self-game environment to continuously fight against its historically optimal versions, which drove the policy to continuously break through the limitations of initial expert knowledge. Finally, a high-fidelity air combat simulation environment was built to experimentally verify the performance of the proposed

收稿日期:2025-11-29;**录用日期**:2026-01-16;**网络首发日期**:2026-06-01

网络首发地址:https://link.cnki.net/urlid/23.1235.t.20260601.1242.002

基金项目:国家自然科学基金(62303489,GKJJ24050502,62350048,T2121003);中国博士后科学基金面上基金(2022M723877);中国博士后科学基金(2023T160790);中国博士后科学基金国际交流计划引进项目(YJ20220347);陕西省青年人才托举计划项目(20220101);陕西省自然科学基础研究计划项目(2024JC-YBQN-0668,2025JC-QYCX-052)

作者简介:王文飞(1998—),男,博士研究生;茹 乐(1978—),男,教授,博士生导师;吕茂隆(1991—),男,副教授,博士生导师

通信作者:吕茂隆,maolonglv@163.com

method. The research results indicate that, compared with the traditional DQN and other baseline algorithms, the proposed method effectively improves the reward convergence performance during the agent learning and training and the final confrontation winning rate and effectively enhances the air combat confrontation level of the unmanned aerial vehicle agent.

Keywords: deep reinforcement learning; unmanned aerial vehicle; autonomous decision-making in air combat; expert confrontation; self-game

随着无人机(unmanned aerial vehicle, UAV)技术的不断发展以及现代空战形态的深刻演变,空中作战呈现出作战节奏快速、信息高度饱和、对抗复杂动态剧烈等特点,单纯依靠传统的人为操控飞行器作战模式已难以满足现代空战需求[1-3]。在此背景下,开展无人机自主空战机动决策研究,实现无人机自主智能感知态势并规划决策攻防机动,已成为当前占据空战决策技术制高点的迫切需求,也是未来智能化战争发展的必然趋势[4-6]。

对于自主空战机动决策,当前常用方法包括基于博弈论方法、基于优化理论方法,以及基于人工智能方法。基于博弈论方法主要包括微分对策、矩阵对策,以及影响图等方法。车竞等[7]针对双机空战机动决策问题,基于基本机动动作和攻击与防御准则构建双方空战支付矩阵,并基于矩阵博弈方法进行仿真求解。刘昊天等[8]利用矩阵对策法,结合对抗双方态势信息建立支付博弈模型,并基于对局迭代方法求解策略。周思羽等[9]将随机决策理论与影响图方法相结合,基于 5 种典型机动准则设计以在特定决策时刻给出决策方案。基于优化理论方法主要包括动态规划方法、遗传算法,以及贝叶斯推理方法等。Mcgrew 等[10]针对单机空战机动决策问题,基于近似动态规划方法来求解二维空间中的机动策略,具有良好的实时决策效果。邓可等[11]将矩阵博弈与遗传算法相结合,首先基于矩阵对策求解空战初始策略范围,然后利用遗传算法来求解最优空战机动策略,一定程度上提升了空战决策效率与准确性。孟光磊[12]等针对单机空战决策问题,分析提取影响空战对抗态势的典型因素,基于因素之间的因果关系构建基于动态贝叶斯的空战决策网络,仿真验证了该方法的有效性。基于博弈论方法求解空战决策问题的优势在于其数学建模明确清晰,求解简单直观,但是其实时性效果较差且容易出现计算量激增的问题;基于优化理论方法的优势在于其方法适应性较强,易扩展,但是其实时性效果也较差且容易陷入局部最优;基于人工智能方法的学习能力较强,且具有良好的决策实时性,可以有效提升空战决策自主智能化水平,目前基于强化学习的智能空战决策方法正在逐渐受到国内外相关研究人员的广泛关注与研究。

强化学习方法主要分为基于价值的[13-15]和基于策略[16-17]的强化学习。基于价值的强化学习主要是通过学习动作的价值来决定后续的动作执行,而基于策略的强化学习则是直接学习智能体的不同策略,如基于策略梯度的方法。Li 等[18]针对单机空战机动决策问题,结合专家经验来改进强化学习决策方法,构建专家经验重播缓冲区并结合导弹攻击区构建,仿真结果显示该方法具有良好的决策适应性。廉云霄等[19]针对多机协同空战决策问题,基于图注意力机制来融合处理多源观测信息,并在此基础上利用改进的时空信息融合方法来聚合战机和时序信息,最后基于强化学习方法进行求解,仿真结果显示该决策方法使得多机群体具有良好的协同性。倪浩等[20]针对无人机智能空战决策问题,首先利用自博弈方法来获得有效的空战交互数据,其次基于域随机化方法来提升决策鲁棒性,仿真结果显示该决策方法显著提升了无人机空战决策的泛化性和鲁棒性。

基于强化学习方法的空战决策虽然可以赋予战机智能体更加自主智能的机动决策能力,但是现有的空战决策端到端模型的动作空间巨大,导致策略难以收敛且缺乏可解释性。其次,仅依赖简单的对抗学习环境来进行策略训练,所得到的智能体能力难以确保可以应对复杂多变的空战环境且易陷入局部最优[21]。针对上述问题,本文首先提出设计一种分层强化学习框架以将决策过程解耦,该框架的核心创新在于将端到端的决策过程策略性解耦为高层战术学习与底层动作执行,上层利用深度 Q 网络

(deep Q-network,DQN)学习抽象的战术意图,从而将巨大的原始动作空间压缩为有限的离散战术指令集,从根本上缓解了策略难以收敛的问题。底层则通过基于多模式控制和有限状态机的驱动机制来将高层战术实时转化成具体的机动动作指令组合。这种分层设计不仅大幅降低了学习复杂度,还增强了决策过程的模块化和可解释性。其次,在训练方法层面,研究设计专家模仿-自博弈训练进化的两阶段训练机制,以解决稀疏奖励环境下的学习效率和策略固化问题。第 1 阶段,构建了基于 SECA 规则架构的专家对手,智能体通过与其对抗进行模仿学习,获得高质量的初始策略和先验知识,以有效克服强化学习"冷启动"难题;第 2 阶段,智能体转入虚拟自博弈环境,通过与其历史最优版本持续对抗,驱动策略不断突破初始专家知识的局限,探索更优和鲁棒性更强的战术策略。

1　问题描述

本文首先对红蓝双方空战场景以及无人机动力学模型进行定义描述,在此基础上,给出空战决策马尔可夫决策模型,主要包含状态空间、动作空间、奖励函数机制,以及状态转移函数的定义设计,为后续的空战机动决策算法设计奠定模型基础。

1.1　空战无人机动力学模型

无人机自主空战主要指对抗双方通过机动占位来获取有利的攻击态势,并适时发射导弹攻击目标的作战模式[22-23]。无人机空战过程主要为:在目标搜索与跟踪阶段,无人机重点在于搜索目标,在捕获目标后开始跟踪,以寻找最佳攻击时机。在捕捉到目标的基础上,无人机需要根据实际情况及时动态调整机动策略来确保处于最佳攻击位置和导弹发射时机,并伺机发射导弹攻击目标。在完成导弹打击之后,无人机需要根据任务完成情况迅速分析是否需要重新进入空战攻击阶段或撤出战场。无人机作为空战对抗的主体,其机动行为决策对空战结果具有至关重要的影响,本文主要采用六自由度模型[1]来建立无人机运动与动力学模型,无人机的运动学方程为

$$\begin{cases}\dot{x}=v\cos\gamma\cos\chi\\ \dot{y}=v\cos\gamma\sin\chi\\ \dot{z}=-v\sin\gamma\end{cases}\tag{1}$$

式中:(x,y,z)为无人机的空间位置,γ、χ 分别为无人机速度向量 v 相对于地面坐标系的航迹偏角和航迹倾角。

无人机的动力学方程为

$$\begin{cases}m\dot{v}=F-Z-mg\sin\gamma\\ v\dot{\gamma}=n\cos\varphi-\cos\gamma\\ v\dot{\chi}\cos\gamma=ng\sin\varphi\end{cases}\tag{2}$$

式中:F 为发动机推力,Z 为阻力,m 为无人机质量,g 为重力加速度,n 为过载,φ 为无人机滚转角。无人机的运动控制量主要为过载、滚转角速率和油门指令,针对不同的空战态势通过生成不同的控制指令组合来适时做出不同的机动动作以占据有利的空战对抗态势。在底层控制器生成连续控制量时,根据当前飞行状态动态调整控制指令上下限,即

$$n_i=\mathrm{clip}(n_i^{\mathrm{cmd}},-3,9)\tag{3}$$

式中:n_i^{cmd} 为控制器当前计算所得到的过载值,$\mathrm{clip}(\cdot)$为截断函数。同时,飞机的滚转角速率限制为 $\dot{\phi}_i=\mathrm{clip}(\dot{\phi}_i^{\mathrm{cmd}},-4,4)$, rad/s。

飞机在机动过程中的安全迎角保护如下:

$$n_i^s=\begin{cases}n_i, & |\alpha|\leqslant 25^\circ\\ n_i\cdot\exp\left(-\dfrac{|\alpha|-25}{5}\right), & 25^\circ<|\alpha|\leqslant 35^\circ\\ 0, & |\alpha|>35^\circ\end{cases}\tag{4}$$

式中:n_i 为经物理限制后的法向过载指令,α 为当前攻角。

飞机在战术机动选择时,考虑当前状态下的最小转弯半径限制为

$$R_{\min}=\frac{V^2}{g\sqrt{n_{\max}^2-1}}\tag{5}$$

式中:V 为飞机速度,g 为重力加速度,$n_{\max}$为当前最大可用过载。在位置追踪、角度追踪等追踪类机动过程中,若所需转弯半径小于 $R_{\min}$,则自动切换为其他能量管理机动模式,以避免执行不可行的机动指令。

基于深度强化学习的智能空战机动决策学习的整体研究框架如图 1 所示。

1.2　空战决策马尔可夫模型

本文将空战机动决策问题建模为一个马尔可夫决策过程(Markov decision process, MDP),其核心要素定义如下。

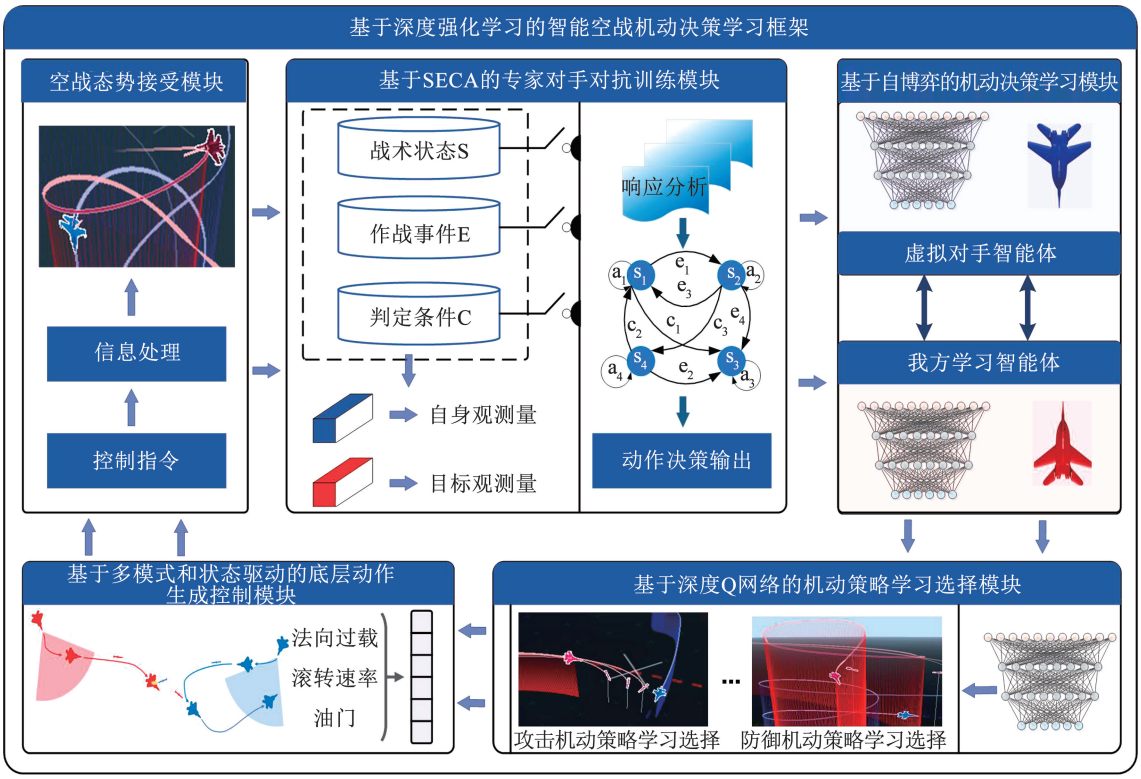


图 1 自主智能空战机动决策学习整体框架

Fig. 1 Overall framework of autonomous intelligent air combat maneuver decision-making learning

1.2.1 状态空间

状态向量在时刻 t 定义为

$$\boldsymbol{s}_t=[x_{\mathrm{b}},y_{\mathrm{b}},z_{\mathrm{b}},v_{\mathrm{b}},\psi_{\mathrm{b}},x_{\mathrm{r}},y_{\mathrm{r}},z_{\mathrm{r}},v_{\mathrm{r}},\psi_{\mathrm{r}}]^{\mathrm{T}} \tag{6}$$

式中：$(x_{\mathrm{b}},y_{\mathrm{b}},z_{\mathrm{b}})$、$(x_{\mathrm{r}},y_{\mathrm{r}},z_{\mathrm{r}})$分别为蓝方和红方在北东地(North east down, NED)坐标系下的位置坐标；v_{b}、v_{r} 分别为蓝方和红方战机的地面速度；ψ_{b}、ψ_{r} 分别为蓝方和红方的进攻角(即其速度矢量与目标视线之间的夹角)。

基于空战状态空间，红、蓝双方空战对抗之间的相对态势计算如下：

1)红、蓝双方相对位置矢量为

$$\boldsymbol{r}=[x_{\mathrm{r}}-x_{\mathrm{b}},y_{\mathrm{r}}-y_{\mathrm{b}},z_{\mathrm{r}}-z_{\mathrm{b}}]^{\mathrm{T}} \tag{7}$$

2)红、蓝双方之间的间隔距离为 $d=\|r\|_2$，红方的进攻角为

$$\psi_{\mathrm{r}}=\arccos\left(\frac{v_{\mathrm{r}}\cdot r}{\|v_{\mathrm{r}}\|\cdot\|r\|}\right) \tag{8}$$

1.2.2 动作空间

动作空间为离散集合，表示不同的基础空战机动模式为

$$a_t\in A=\{0,1,2,3,4,5,6,7,8,9,10\} \tag{9}$$

式中，每个动作 a_t 对应一种预定义的战术机动[24]。其中：0 为加速，1 为定航迹角爬升，2 为位置追踪，3 为角度追踪，4 为高 YoYo 机动，5 为低 YoYo 机动，6 为半滚倒挂，7 为大回环，8 为赫伯斯特机动，9 为 F-pole 机动，10 为眼镜蛇机动。这些机动由底层控制器解析为具体的过载、滚转指令执行。

1.2.3 奖励函数

奖励函数 $R(s_t,a_t,s_{t+1})$ 旨在引导智能体学习优势空战态势，主要由环境态势奖励和终局奖励构成。

1)环境态势奖励

环境态势奖励 R_{e} 为

$$R_{\mathrm{e}}=w_aR_a+w_dR_d+w_vR_v+w_hR_h \tag{10}$$

式中：R_a 为角度优势奖励，R_d 为距离优势奖励，R_v 为速度优势奖励，R_h 为高度优势奖励，w_a、w_d、w_v、w_h 分别为各个优势奖励的权重。

①角度优势奖励 R_a 为

$$R_a=1-\frac{|\psi_{\mathrm{r}}|+|\phi_{\mathrm{r}}|}{2\pi} \tag{11}$$

式中，ϕ_{r} 为红方的逃离角，此项奖励机制鼓励战机指向敌机并规避敌机的机头指向。

②距离优势奖励 R_d 为

$$R_d=\exp\left(-\frac{|d-d_{\mathrm{s}}|}{d_{\mathrm{o}}}\right) \tag{12}$$

式中：d_{s} 为对比标准攻击距离，本文将其设置为 10 km；d_{o} 为距离尺度因子，本文将其设置为 100 km。

③高度优势奖励 R_h 为

$$R_h=\begin{cases}\dfrac{z_r}{z_b}-0.3, & z_r<z_b\\ \exp\left(-\dfrac{|z_r-h_s|}{h_s}\right), & \text{其他}\end{cases}\tag{13}$$

式中,h_s 为对比标准空战高度,本文将其设置为 2 000 m。

④速度优势奖励 R_v 为

$$R_v=\begin{cases}\exp\left(\dfrac{v_s-v_r}{v_r}\right), & v_r>v_s\\ 1.5, & 1.2v_b<v_r\leqslant v_s\\ \dfrac{v_r}{v_b}-0.5, & 0.3<v_r\leqslant 1.2v_b\\ 0.5, & \text{其他}\end{cases}\tag{14}$$

式中,v_s 为对比标准空战速度,本文将其设置为1 *Ma*。

2)空战终局奖励

空战终局奖励 R_t 为

$$R_t=\begin{cases}+150, & \text{我方存活且敌方被击杀}\\ -150, & \text{敌方存活且我方被击杀}\\ +30, & \text{双方同归于尽}\\ -30, & \text{双方均存活}\end{cases}\tag{15}$$

3)总体空战奖励

总体空战奖励 R_s 为

$$R_s=w_eR_e+w_tR_t\tag{16}$$

式中,w_e、w_t 分别为环境态势奖励和空战终局奖励权重。若强调空战过程中的态势控制能力(如持续占据优势位置),则应提高 w_e,使智能体更关注实时态势变化;若强调最终胜负结果(如尽快击落敌机),则应提高 w_t,鼓励智能体以达成击落或避免被击落为首要目标。在奖励设置中,需保证 w_eR_e 和 w_tR_t 在数值尺度上相匹配,避免某一项奖励主导整个奖励信号,导致策略偏向单一目标。在实际训练中,通过进行多组权重实验仿真试验的对比分析,本文设定 $w_e=0.7$,$w_t=0.3$,该权重设定本质上是对过程奖励与结果奖励的权衡,旨在较好地平衡对抗胜率与态势控制效果。

1.2.4　状态转移函数

状态转移函数由红、蓝双方空战环境决定:

$$s_{t+1}\sim P(s_{t+1}|s_t,a_t)\tag{17}$$

在红、蓝双方空战对抗过程中,红方接收空战态势信息 s_t,并执行动作 a_t,空战对抗环境发生变化,状态发生迁移 s_{t+1}。自主机动决策的目标是基于当前所观察到的态势信息,通过决策映射函数机制输出每架飞机在时刻 t 应该执行的动作。

2　基于分层强化学习的空战机动决策

本文主要研究构建分层强化学习框架,将智能体的决策过程进行宏观战术决策与微观动作执行解耦,首先研究构建了上层的基于深度 Q 网络的机动决策训练与策略学习方法来输出宏观战术机动策略,随后聚焦在底层研究基于多模式和状态驱动的连续动作生成控制方法以生成有限动作指令组合来控制动作输出。

2.1　基于深度 Q 网络的机动决策训练与策略学习

本文主要采用深度 Q 网络算法来研究自主空战机动决策问题,其主要采用双网络结构,包含评估网络 $Q(s,a;\theta)$ 和目标网络 $Q(s,a;\theta^-)$,网络结构为多层感知机,即

$$\begin{cases}h=\mathrm{ReLU}(\boldsymbol{W}_1s+\boldsymbol{b}_1)\\ Q=\boldsymbol{W}_2h+\boldsymbol{b}_2\end{cases}\tag{18}$$

式中:$\boldsymbol{W}_1\in\mathbf{R}^{50\times 10}$,$\boldsymbol{b}_1\in\mathbf{R}^{50}$,$\boldsymbol{W}_2\in\mathbf{R}^{11\times 50}$,$\boldsymbol{b}_2\in\mathbf{R}^{11}$,$\theta=\{\boldsymbol{W}_1,\boldsymbol{b}_1,\boldsymbol{W}_2,\boldsymbol{b}_2\}$。

以学习最优动作价值函数 Q^* 为决策目标,该函数需满足贝尔曼最优方程为

$$Q^*(s,a)=E_{s'\sim P}[R(s,a,s')+\gamma\max_{a'}Q^*(s',a')]\tag{19}$$

式中,$\gamma=0.99$ 为折扣因子,用于权衡当前与未来奖励。

在学习训练过程中,使用回放缓冲区 D 存储智能体的经验元组 $e_t=(s_t,a_t,r_t,s_{t+1})$,训练时从中均匀采样一个小批次的经验 $B\sim\mathrm{Uniform}(D)$ 以打破数据之间的相关性。通过最小化均方时序差分误差来训练评估网络为

$$L(\theta)=E_{(s,a,r,s')\in B}[(y-Q(s,a;\theta))^2]\tag{20}$$

其中目标值 y 由目标网络计算:

$$y=r+\gamma\max_{a'}Q(s',a';\theta^-)$$

使用 Adam 优化器以学习率 $\alpha=0.01$ 更新参数 θ,每经过 150 次训练步,将评估网络的参数复制给目标网络 $\theta^-\leftarrow\theta$ 以稳定训练过程。采用 ε-贪婪策略来选择动作,以平衡探索与利用,即

$$a_t=\begin{cases}\text{随机从 } A \text{ 中选择}, & \text{概率为 } \varepsilon\\ \underset{a}{\mathrm{argmax}}\,Q(s_t,a;\theta), & \text{概率为 } 1-\varepsilon\end{cases}\tag{21}$$

在训练过程中,探索率从初始值 0.9 线性衰减。通过上述 MDP 建模与 DQN 训练框架,智能体能够通过与仿真环境的交互,自主学习在复杂空战态势下选择最优战术机动的策略。高层策略 π_{high} 是一个

参数化的策略,其输入为状态 s_t,输出一个离散的机动指令 m_t 为

$$m_t = \pi_{\text{high}}(s_t;\theta) \in A \tag{22}$$

式中:A 为离散的机动指令集合,每个指令对应一种空战基本战术动作;θ 为 DQN 网络的参数,该层级的目的是根据当前态势选择最有利的战术机动类型。

2.2 基于多模式和状态驱动的底层动作生成控制

底层控制器是一个确定性函数 F,它将高层的离散机动指令 m_t 和当前飞机的连续状态 x_t 以及目标敌机的连续状态 x_{target} 映射为具体的动力学模型的连续控制量 u_t,即

$$\boldsymbol{u}_t = F(m_t, x_t, x_{\text{target}}) \tag{23}$$

式中,$\boldsymbol{u}_t = [n_z, \dot{\phi}_{\text{cmd}}, T_{\text{h}}]^{\text{T}}$ 为生成的控制向量(其中:n_z 为法向过载指令,用于控制飞机爬升或俯冲;$\dot{\varphi}_{\text{cmd}}$ 为滚转角速率指令,用于控制飞机滚转与转弯;T_{h} 为油门指令,通常被设置为固定值)。

底层控制器可以建模为一个六元组(M, X, U, G, F, τ)。其中:M 为输入机动指令集合,即高层动作空间;X 为系统状态,包括本机状态 x_t 和目标状态 x_{target};U 为输出控制指令空间;G:$M \times X \rightarrow U$ 为基本动作生成器,为大多数机动指令提供连续控制律;$M \times X \times \tau \rightarrow U \times \tau$ 为序列动作生成器,处理需要分阶段执行的复杂机动(τ 为内部状态(标签)集合)。

2.2.1 基本动作生成器 G 的详细控制律

对于非序列机动,控制器 G 直接根据指令和当前状态计算 u_t:

1)模式 1&2:能量管理(加速/爬升)

①模式 1(加速):

$$\begin{cases} n_z = \cos(\gamma) \\ \dot{\varphi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) = K_\varphi(0 - \phi) \\ T_t = 100 \end{cases} \tag{24}$$

式中:$n_z = \cos(\gamma)$ 主要是为了平衡重力以保持高度,$\dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) = K_\phi(0 - \phi)$ 是为了改平机翼。

②模式 2(定航迹角爬升):

$$\begin{cases} n_z = \dfrac{V_g}{g \cdot t_c}\sin(\gamma_{\text{des}} - \gamma) + \cos(\gamma) \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) = K_\phi(0 - \phi) \\ T_t = 100 \end{cases} \tag{25}$$

式中:$\gamma_{\text{des}} = 25°$ 为期望航迹角,t_c 为控制时间常数,设定为 1 s,以调节响应速度。

2)模式 3&6:追踪(位置/姿态)

①模式 3(位置追踪–比例导引):相对位置与视线角为

$$\begin{cases} r = p_t - p_r \\ \lambda_y = \arctan 2(r_{\text{east}}, r_{\text{north}}) \\ \lambda_p = \arctan 2(-r_{\text{down}}, \sqrt{r_{\text{north}}^2 + r_{\text{east}}^2}) \end{cases} \tag{26}$$

在弹道系计算加速度指令:

$$a_{\text{cmd}}^{\text{traj}} = \begin{bmatrix} 0 \\ 2V_g\sin(\lambda_y - \chi) \\ -2V_g\sin(\lambda_p - \gamma) \end{bmatrix} \tag{27}$$

转换至速度系并补偿重力:

$$\begin{cases} a_{\text{cmd}}^{\text{vel}} = R_x(\phi)^{\text{T}} a_{\text{cmd}}^{\text{traj}} \\ a_{\text{total}}^{\text{vel}} = a_{\text{cmd}}^{\text{vel}} - [0,0,g]^{\text{T}} \end{cases} \tag{28}$$

解算最终控制指令:

$$\begin{cases} n_z = -\dfrac{a_{\text{total}}^{\text{vel}}[2]}{g} \\ \phi_{\text{des}} = \arctan 2(a_{\text{total}}^{\text{vel}}[1], -a_{\text{total}}^{\text{vel}}[2]) \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) \\ T_t = 100 \end{cases} \tag{29}$$

②模式 6(姿态追踪):针对目标姿态追踪机动模式,首先计算目标在机体系下的方位:

$$r^{\text{body}} = R_{\text{body}}^{\text{ned}}(p_t - p_r) \tag{30}$$

在得到目标的确切方位信息之后,计算姿态误差角:

$$\begin{cases} \varepsilon_y = \arctan 2(r_y^{\text{body}}, r_x^{\text{body}}) \\ \varepsilon_p = \arctan 2(-r_z^{\text{body}}, r_x^{\text{body}}) \end{cases} \tag{31}$$

结合目标姿态追踪的姿态误差角,计算生成角速率指令并转换如下:

$$\begin{cases} \dot{\chi}_{\text{cmd}} = K_{\text{p},y}\varepsilon_y \\ \dot{\gamma}_{\text{cmd}} = K_{\text{p},p}\varepsilon_p \\ n_z = \dfrac{V_g}{g}\dot{\gamma}_{\text{cmd}} + \cos(\gamma) \\ \dot{\phi}_{\text{cmd}} = K_\varphi\dot{\chi}_{\text{cmd}} \\ T_t = 100 \end{cases} \tag{32}$$

3)模式 4&5:YoYo 机动(高/低),YoYo 机动通过追踪一个虚拟点来实现。

虚拟点位置为

$$H_c = \text{sign} \cdot \eta \cdot \frac{V_g^2}{2g} \tag{33}$$

式中:sign = +1 表示高 YoYo,sign = −1 表示低

YoYo;η 为增益系数;$p_{\text{virtual}} = p_t + [0,0,H_c]^{\mathrm{T}}$ 作为新模式 3 的目标,调用位置追踪控制律。

2.2.2　序列动作生成器 F 的详细控制律

对于序列机动(模式 7 ~ 11),控制器 F 维护一个内部状态 $\text{tag} = [\text{tag}_0, \text{tag}_1] \in B^2$ 来管理机动执行阶段。

1)模式 7:半滚倒挂(Split-S)

①阶段 1($\text{tag}_0 = \text{True}$):滚转至倒飞为

$$\begin{cases} \phi_{\text{des}} = 180° \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) \\ n_z = 2.0 \end{cases} \tag{34}$$

式中,$n_z = 2$ 表示轻微拉起,辅助滚转。若 $|\phi| > 160°$,则 $\text{tag}_0 = \text{False}$。

②阶段 2($\text{tag}_0 = \text{False}, \text{tag}_1 = \text{True}$):倒飞状态下拉为

$$\begin{cases} \phi_{\text{des}} = 180° \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) \\ n_z = 9.0 \end{cases} \tag{35}$$

式中,$n_z = 9$ 表示大过载,向下机动。若 $\theta > 0°$,则 $\text{tag}_1 = \text{False}$。

③阶段 3($\text{tag}_0 = \text{False}, \text{tag}_1 = \text{False}$):

$$\begin{cases} \phi_{\text{des}} = 0° \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) \\ n_z = \cos(\gamma) \end{cases} \tag{36}$$

2)模式 8:大回环

①阶段 1:拉起进入筋斗

$$\begin{cases} n_z = 6.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \end{cases} \tag{37}$$

式中,$\dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi)$ 主要是为了保持机翼水平。

②阶段 2:在顶点附近滚转 180 °,即

$$\begin{cases} n_z = 2.0 \\ \phi_{\text{des}} = 180° \\ \dot{\phi}_{\text{cmd}} = K_\phi(\phi_{\text{des}} - \phi) \end{cases} \tag{38}$$

③阶段 3:改平并进入平飞为

$$\begin{cases} n_z = \cos(\gamma) \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \end{cases} \tag{39}$$

3)模式 9(赫伯斯特机动)。赫伯斯特机动是一种大迎角、快速机头指向变化的超机动动作,适用于空战对抗中快速改变机头指向以抢占攻击角度,主要将其分为 3 个阶段。

①阶段 1:快速拉起进入大迎角,当迎角 $\alpha > 30°$ 时,则进入下一阶段,具体的控制律为

$$\begin{cases} n_z = 6.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \end{cases} \tag{40}$$

②阶段 2:机头快速偏转,当偏航角变化率 $\dot{\psi} > 30$ °/s 且持续 1 s 之后则进入下一阶段,控制律为

$$\begin{cases} n_z = 2.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(180° - \phi) \end{cases} \tag{41}$$

③阶段 3:改出并恢复稳定飞行状态,使得飞机的迎角 $\alpha < 10°$ 且滚转角 $|\phi| < 20°$,控制律为

$$\begin{cases} n_z = \cos(\gamma) \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \\ T_t = 100 \end{cases} \tag{42}$$

4)模式 10(F-pole 机动)。F-pole 机动是一种在导弹攻击中使用的规避与占位复合机动,通过侧向机动维持对敌角度优势的同时规避敌方武器攻击。

①阶段 1:飞机侧向滚转并拉起以快速建立侧向位移,使飞机滚转角达到约 55° ~ 65°范围,控制律为

$$\begin{cases} n_z = 4.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(60° - \phi) \end{cases} \tag{43}$$

②阶段 2:飞机保持侧滑,在维持侧向位移的同时调整机头指向目标,在此过程中,飞机的滚转角速率与前一阶段保持一致,同时保持飞机的侧滑角约为 5°左右,在视线角变化率 $\dot{\lambda} < 1$°/s 时则进入阶段 3。

③阶段 3:飞机改平并恢复对目标的稳定追踪,此时飞机的控制律和赫伯斯特机动的阶段 3 的控制律保持一致。

5)模式 11(眼镜蛇机动)。眼镜蛇机动是一种典型的过失速机动,通过快速拉起至大迎角后迅速改出,用于快速减速并迫使后方敌机冲前。

①阶段 1:飞机快速拉起至大迎角,使飞机迎角迅速增加至约 70° ~ 90°范围,控制律为

$$\begin{cases} n_z = 8.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \\ T_t = 30 \end{cases} \tag{44}$$

②阶段 2:飞机保持大迎角状态,并迅速减速,具体的控制律为

$$\begin{cases} n_z = 0.5 \\ \dot{\phi}_{\text{cmd}} = K_\phi(0 - \phi) \\ T_t = 10 \end{cases} \tag{45}$$

③阶段 3:当飞机的空速下降至原速度的 60% 左右或持续 1 s 之后,快速改出并恢复小迎角飞行状态,控制律为

$$\begin{cases} n_z = -3.0 \\ \dot{\phi}_{\text{cmd}} = K_\phi(0-\phi) \\ T_t = 100 \end{cases} \tag{46}$$

底层控制器的总入口函数如下:

$$u_t, \text{tag}_{t+1} = F(a_t, x_r, x_t, \text{tag}_t) = \begin{cases} F(a_t, x_r, x_t, \text{tag}_t), & a_t \in \{7,8,9,10,11\} \\ (G(a_t, x_r, x_t), \text{tag}_t), & \text{其他} \end{cases} \tag{47}$$

通过这种分层与分模块的设计,驱使高层 DQN 在简化的抽象动作空间中高效学习战术决策,而底层动作控制则确保这些战术决策能够被安全、可靠且符合物理规律地执行。这两部分通过明确的接口 a_t 和 u_t 进行耦合,共同构成了一个完整、高效的空战智能体。

3 基于预训练规则对手的自博弈强化学习训练

为实现红方 DQN 智能体的战术能力从“基础认知”到“突破优化”的递进,训练过程分为“规则对手对抗训练”与“自博弈强化学习”两个紧密衔接的阶段。首先,以基于 SECA 架构构建的蓝方规则对手为训练对象,在对抗训练阶段,智能体通过 ε-贪心策略选择动作并存储经验至预训练经验池,同步优化 DQN 网络参数,最终使智能体形成对空战环境的基础认知。在此基础上,进入自博弈强化学习阶段,以预训练成果为初始化,构建动态对抗 MDP,采用“当前智能体 vs 历史最优智能体”的对抗模式,通过迭代流程实现策略优化,促使智能体更加高效地学习空战机动策略。

3.1 基于 SECA 规则架构的专家对手对抗训练

基于专家系统的状态-事件-条件-动作(state-event-condition-action, SECA)分层决策框架,将蓝方规则型对手的决策逻辑形式化建模为六元组模型,统一规则的抽象描述、可视化表达与机器部署格式。蓝方专家规则对手的 SECA 决策模型 $M_B = (\text{ID}_B, Q_B, E_B, C_B, A_B, \Pi_B)$,其中:$\text{ID}_B$ 为蓝方的唯一标识号;$Q_B = \{q_1, q_2, \cdots, q_8\}$ 为有限战术状态集合;$E_B = \{e_1, e_2, \cdots, e_6\}$ 为触发状态转换的事件集合,基于战场态势变化生成;$C_B = \{c_1, c_2, \cdots, c_5\}$ 为状态转换的条件集合,量化事件触发后的判定标准;$A_B = \{a_1, a_2, \cdots, a_8\}$ 为蓝方可执行的动作集合;$\Pi_B: Q_B \times E_B \times C_B \rightarrow Q_B \times A_B$ 为状态转换函数,满足 $(a_b, q'_b) = \Pi_B(q_b, e_b, c_b)$,表示当前状态 q_b 下,触发事件 e_b 且满足条件 c_b 时,蓝方执行动作 a_b 并转换至新状态 q'_b。蓝方战术状态集合、作战事件集合见表 1、2。

表 1 空战战术状态

Tab. 1 Tactical states of air combat

状态	描述	对应作战阶段
q_1	能量优势状态	区域巡航飞行,预先积累能量
q_2	远距机动状态	双方间隔较远,主动搜索目标并占位
q_3	机动防御状态	触发危险告警,进入规避防御阶段
q_4	攻击交互状态	达成攻击条件,进入作战攻击阶段

表 2 空战作战事件

Tab. 2 Air combat events

事件	描述	触发源
e_1	高度/速度过低	自身状态监测
e_2	敌我距离中远程	态势感知
e_3	敌我距离近距离	态势感知
e_4	方位角过大	角度计算
e_5	被敌方咬尾	敌我相对姿态
e_6	目标摧毁	战场状态更新

从表 2 可以看出,事件 e_1 中的高度过低预警值设定为 3 km,速度过低预警值设定为 80 m/s;事件 e_2、e_3 中的敌我间距中远程与近距离的界定值设定为 5 km;事件 e_4 中的方位角过大判定条件设定为天线瞄准角 $\alpha > 60°$;事件 e_5 中的被敌方咬尾的判定条件设定为天线瞄准角 $\alpha > 120°$ 且目标进入角 $\alpha_e > 150°$。

蓝方判定条件集合见表 3。

表 3 空战判定条件

Tab. 3 Air combat judgment conditions

条件	描述	数学表达式
c_1	自身能量低于敌方 60%	$\Delta E < 0.6\Delta E_T$
c_2	自身速度高于最优速度	$v > v_{\text{opt}} = 260$ m/s
c_3	自身速度低于最优速度	$v \leqslant v_{\text{opt}} = 260$ m/s
c_4	敌方位于自身下方	$\Delta z = z_{\text{own}} - z_{\text{tar}} > 0$
c_5	敌方位于自身上方	$\Delta z = z_{\text{own}} - z_{\text{tar}} < 0$

蓝方机动动作集合包括:a_1 为加速,a_2 为爬升,a_3 为位置追踪,a_4 为角度追踪,a_5 为高 YoYo 机动,a_6 为低 YoYo 机动,a_7 为半滚倒挂,a_8 为大回环,各

类动作含义及生成控制方式同前述章节。

基于表 1 ~3 的核心元素定义,蓝方规则对手的状态转换规则通过 SECA 规则描述语言形式化,关键规则示例见表 4。

表 4　空战规则

Tab. 4　Air combat rules

状态 q_b	事件 e_b	条件 c_b	动作 a_b	继任状态 q_b'
q_1	$e_1 \wedge e_2$	c_3	a_1	q_1
q_1	e_2	$\neg c_1$	a_2	q_2
q_1	e_5	c_5	a_6	q_3
q_1	e_5	c_4	a_7	q_3
q_1	e_3	$\neg c_1$	a_3	q_4
q_2	e_1	c_1	a_1	q_1
q_2	e_3	$\neg c_1$	a_3	q_4
q_2	e_4	c_2	a_4	q_2
q_2	e_4	c_3	a_5	q_2
q_2	$\neg e_4$		a_2	q_2
q_3	$\neg e_5$	c_1	a_1	q_1
q_3	$\neg e_4 \wedge \neg e_6$	$\neg c_1$	a_2	q_4
q_3	$e_5 \wedge \neg e_6$	c_5	a_6	q_3
q_3	$e_5 \wedge \neg e_6$	c_4	a_7	q_3
q_4	e_5	c_5	a_6	q_3
q_4	e_5	c_4	a_7	q_3
q_4	$e_3 \wedge \neg e_5$	c_1	a_3	q_4
q_4	$e_2 \wedge \neg e_6$		a_2	q_4
q_4	e_6		a_1	q_2

表 4 将表 1 ~3 中的所有状态集合、事件集合,以及条件集合按照 SECA 决策模型的决策链条进行串联结合以形成具体的机动规则,如表 4 中的规则链 $q_2 \rightarrow e_3 \rightarrow \neg c_1 \rightarrow a_3 \rightarrow q_4$ 所示,该规则表示蓝方初始处于远距机动状态,在触发敌我双方间距逐渐缩小至近距范围内事件时,满足机动能量较高的条件下开始执行位置追踪机动动作以寻求在尽可能短的时间内来捕捉到合适的攻角,并迁移至攻击交互状态来打击对方。图 2 清晰地展示了蓝方状态机的决策逻辑流程,每个节点代表一个状态,连线上的标签表示事件→条件→动作,最终指向继任状态。通过这种形式化的描述和状态迁移图,蓝方规则策略的逻辑变得清晰、严谨且易于理解和复现。它展示了如何根据空战态势(事件和条件)在不同的战术状态之间进行切换,并为每个状态选择合适的战术动作。

基于上述蓝方专家对手规则建模,使红方智能体与其进行对抗训练,其核心是让红方 DQN 智能体通过与蓝方 SECA 规则对手交互,学习战场态势感知、动作价值评估,以及最优决策生成,具体过程形式化如图 2 所示。

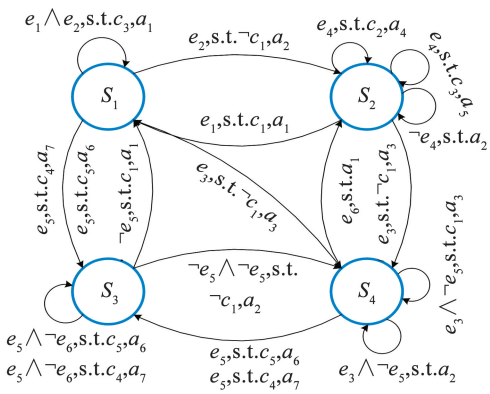


图 2　状态迁移图

Fig. 2　State transition diagram

定义对抗训练的马尔可夫决策过程(MDP) $M = (S, A_R, A_B, P, R, \gamma)$。其中:状态空间 S 包括红方状态 $s_R(t)$、蓝方状态 $s_B(t)$ 与环境状态。红方状态 $s_R(t)$ 与前述章节中的状态空间保持一致,蓝方状态 $s_B(t)$ 与 $s_R(t)$ 结构一致,新增 SECA 状态 $q_B(t) \in Q_B$,环境状态 $s_E(t) = [d(t), \alpha(t), \alpha_e(t), \Delta E(t)]$,即包含敌我距离、方位角、敌方方位角和能量差;动作空间 A_R 为红方 DQN 智能体的控制指令集合,$a_R(t) = [n_{zR}(t), \dot{\phi}_R(t), \lambda_{mR}(t)]$,即包含法向过载、滚转角速率和导弹发射指令等,动作空间 A_B 为蓝方 SECA 规则对手的动作集合 $a_B(t) = \prod_B(q_B(t), e_B(t), c_B(t))$;状态转移概率 $P(s'(t) \mid s(t), a_R(t), a_B(t))$ 由红、蓝双方对抗交互而决定,奖励函数与前述章节中所述奖励机制保持一致。对抗训练的数据流通过时序采样与经验回放池存储实现,具体过程见算法 1。

算法 1:专家规则对手对抗训练学习算法

1: 初始化:红方 DQN 网络参数 θ,目标网络参数 θ^-,蓝方 SECA 状态 $q_B^0 = q_1$,生成初始状态 s_0

2: **for** $t = 0, 1, \cdots, T-1$:

3: $a_R(t) = \underset{a \in A_R}{\operatorname{argmax}} Q(s(t), a; \theta)$(贪心概率 $1-\varepsilon$)或随机选择动作(概率 ε)

4: $e_B(t) = \rho(s(t), q_B(t))$(事件提取),$c_B(t) = \zeta(s(t), q_B(t))$(条件判定),$a_B(t) = \prod_B(q_B(t), e_B(t), c_B(t))$

5: 执行红、蓝双方动作并转移至新状态 $s'(t) = f(s(t), a_R(t), a_B(t))$,计算奖励 $R(t)$

6: 存储状态、动作和奖励经验元组 $(s(t), a_R(t), R(t), s'(t), \text{done}(t))$ 至经验回放池 D

7: 网络更新:每 K 步从 D 中随机采样批量经验 B,最小化时序差分误差:$L(\theta) = E_{(s,a,R,s',\text{done}) \in B}[(Q(s,a;\theta) - y)^2]$

8: 目标网络同步:每 L 步更新 $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$

3.2 基于虚拟自博弈的空战机动决策强化学习

结合基于 SECA 规则架构的专家对手对抗训练,对抗训练阶段的核心目标是让智能体预先熟悉对抗环境,而后续的自博弈是为了突破固有专家知识经验,两者通过初始化迁移与对手迭代实现衔接。将自博弈过程建模为动态对抗马尔可夫决策过程(dynamic adversarial MDP, DAMDP),形式化表述为

$$M_{\mathrm{SP}}=(S,A,P_{\mathrm{SP}},R_{\mathrm{SP}},\gamma,\boldsymbol{\Theta}) \tag{48}$$

式中:S 为状态空间,与对抗训练阶段一致,包含红方状态、蓝方状态与环境状态;A 为动作空间,双方采用统一的动作空间 $a(t)=[n_z(t),\dot{\phi}(t),\lambda_m(t)]$;$P_{\mathrm{SP}}$为自博弈状态转移概率,由双方智能体的动作共同决定 $P_{\mathrm{SP}}(s'(t)\mid s(t),a_1(t),a_2(t))$,其中 a_1 为当前智能体动作,a_2 为历史最优智能体动作;R_{SP}为自博弈奖励函数,与前述章节中的奖励函数机制保持一致;γ 为折扣因子,平衡即时奖励与长期收益;$\boldsymbol{\Theta}$ 为策略池,存储自博弈迭代过程中所有历史智能体的网络参数 $\boldsymbol{\Theta}=\{\theta_0,\theta_1,\cdots,\theta_k\}$,$\theta_k$ 为第 k 轮迭代的智能体参数。

策略池存储自博弈迭代过程中所有有效策略网络参数,其动态更新为

$$\boldsymbol{\Theta}^{k+1}=\begin{cases}\boldsymbol{\Theta}^k\cup\{\theta_k\}, & W(k)>W_{\mathrm{b}}+\Delta\omega\\ \boldsymbol{\Theta}^k, & \text{其他}\end{cases} \tag{49}$$

式中:$\boldsymbol{\Theta}^k$ 为第 k 轮迭代后的策略池,$W(k)$为第 k 轮当前智能体(参数 θ_k)对抗历史最优智能体(θ_{b})的胜率,W_{b} 为当前历史最优胜率,$\Delta\omega=0.05$ 为胜率提升阈值。

从更新后的策略池中选择胜率最高的策略作为新的历史最优,即

$$\theta_{\mathrm{b}}^{k+1}=\underset{\theta_i\in\boldsymbol{\Theta}^{k+1}}{\operatorname{argmax}}\,W(k) \tag{50}$$

同时更新历史最优胜率:$W_{\mathrm{b}}^{k+1}=\max\limits_{\theta_i\in\boldsymbol{\Theta}^{k+1}}W(k)$。

自博弈的初始化依赖对抗训练阶段的输出为 $\theta_0=\theta_{\mathrm{pre}}$,其中 θ_0 为初始智能体参数,θ_{pre}为对抗训练后红方 DQN 的网络参数。自博弈训练阶段初始经验池为 $D_0=D_{\mathrm{pre}}$,其中 D_0 为初始经验池,D_{pre}为对抗训练阶段存储的经验元组集合。在迭代对抗过程中,随机生成对抗初始态势,当前智能体采用 ε-贪心策略选择动作为

$$a_1(t)=\begin{cases}\underset{a\in A}{\operatorname{argmax}}\,Q(s(t),a;\theta_k), & \text{概率为 } 1-\varepsilon\\ \operatorname{random}(a\in A), & \text{概率为 } \varepsilon\end{cases} \tag{51}$$

历史最优智能体采用纯贪心策略选择动作以对抗压力稳定为

$$a_2(t)=\underset{a\in A}{\operatorname{argmax}}\,Q(s(t),a;\theta_{\mathrm{b}}) \tag{52}$$

红、蓝双方执行动作并更新状态,计算自博弈奖励并判定终止条件,存储经验元组至当前轮经验池。策略更新过程为:从当前轮经验池中随机采样批量经验 $B=\{(s_i,a_i,R_{\mathrm{SP}},s_i',\mathrm{done}_i)\}$,目标值计算 y_i 为

$$y_i=R_{\mathrm{SP},i}+\gamma(1-\mathrm{done}_i)\cdot Q(s_i',Q(s_i',a';\theta_k);\theta_k^-) \tag{53}$$

式中,θ_k^- 为当前轮目标网络参数,与当前网络参数的关系为软更新。

采用时序差分(temporal difference, TD)损失函数,同时增加策略正则化项避免策略与预训练基础偏差过大,即

$$L(\theta_k)=E_{(s,a,R,s',\mathrm{done})\in B}[(Q(s,a;\theta_k)-y)^2)+\lambda\cdot\sum_i(\theta_{k,i}-\theta_{\mathrm{pre},i})^2 \tag{54}$$

式中 λ 为正则化系数。

采用 Adam 优化器最小化损失函数,更新当前网络参数为

$$\theta_k\leftarrow\theta_k-\alpha\cdot\nabla_{\theta_k}L(\theta_k) \tag{55}$$

式中:α 为学习率,$\nabla_{\theta_k}L(\theta_k)$为损失函数对 θ_k 的梯度。

4 仿真实验与结果分析

4.1 实验参数设置

本文主要针对红、蓝双方空战对抗仿真实验进行仿真参数设置,实验基于 Python3.7 和 PyTorch 框架搭建空战仿真平台,并基于名为“SNAKE”的开源软件进行空战对抗仿真实验。

本次实验采用 DQN 算法进行空战机动决策学习训练,相关的网络参数、训练参数见表 5、6。

表 5 相关网络参数说明

Tab. 5 Description of relevant network parameters

参数名称	数值	说明
输入层维度	10	我方状态维度
隐藏层结构	[64, 64]	两层全连接网络,各 64 神经元
输出层维度	11	对应 11 种机动动作
激活函数	ReLU(隐藏层)、线性(输出层)	非线性拟合与动作价值输出
学习率	1×10^{-4}	Adam 学习率
目标网络软更新系数	0.001	目标网络缓慢跟随当前网络

表 6　相关训练参数说明

Tab. 6　Description of relevant training parameters

参数名称	数值	说明
经验池容量	1×10^6	存储历史经验的最大容量
批量采样大小	32	每次更新的样本数量
折扣因子	0.99	平衡即时与长期奖励
初始探索率	0.9	ε-贪心策略初始值
探索率衰减系数	0.995	每轮迭代的探索率衰减比例

在红、蓝双方的空战对抗场景中,双方空战战场环境参数见表 7。

表 7　对抗场景初始态势参数设置

Tab. 7　Initial situation parameter settings of confrontation scenario

参数名称	数值
初始间隔范围/m	[1 500,50 000]
初始方位角范围/(°)	[30,120]
仿真步长/s	0.1
初始速度范围/*Ma*	[0.5,1.5]
初始导弹数量/枚	4
导弹发射间隔/s	2

4.2　决策方法性能对比实验

为更好地验证本文所提决策方法的有效性和优越性,将本文提出的改进决策方法 IDQN 与原始 DQN 方法和 PPO 基线算法进行对比,对比的指标主要包含决策方法在智能体学习训练过程中的累积奖励的变化情况以及对抗胜率比例的变化情况,在每一轮的对抗训练中,红、蓝双方的初始态势随机初始化,具体的仿真情况如图 3、4 所示。

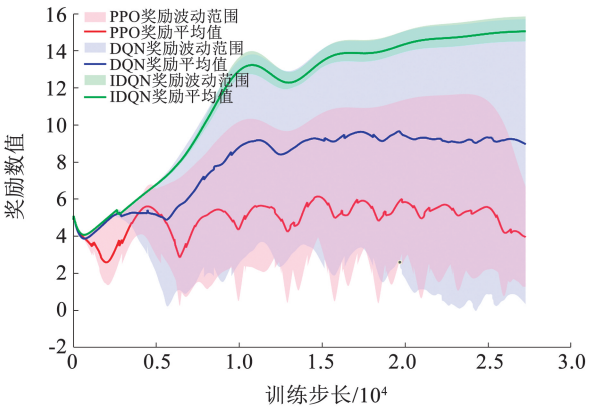


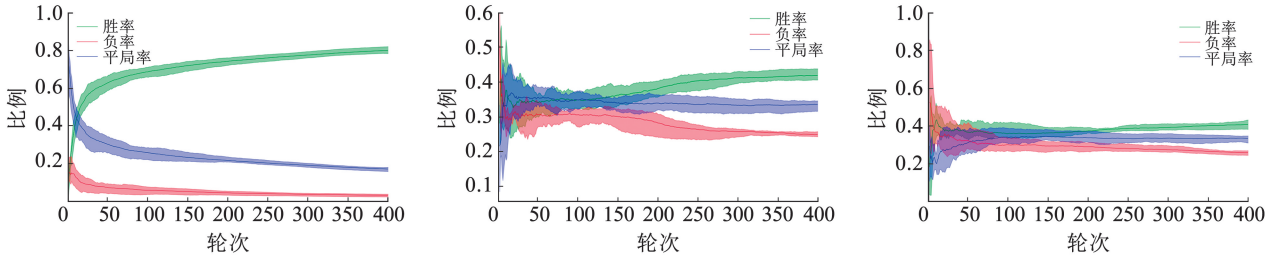
图 3　不同方法奖励变化情况对比

Fig. 3　Comparison of reward changes of different methods

图 3 展示了 3 种不同决策方法在智能体学习训练过程中的奖励累积变化情况,所有方法的学习训练共进行了 5 个线程的仿真试验,图中的变化曲线为 5 个线程实验数据的平均值变化曲线,阴影区域为所有线程实验数据的最大值和最小值组成,同时也对数据做了平滑处理以更好地观察奖励数据变化趋势。

从图 3 可以看出,原始 DQN 方法的训练奖励虽然最终达到稳定的收敛水平,但是本文提出的改进 DQN 决策方法展现出更快的收敛速度并最终收敛到更高的奖励水平且其波动变化范围更小,其奖励曲线从一开始就处于更高水平,这证明了专家预填充提供了高质量的初始数据,克服了 DQN 的“冷启动”问题并保证了学习训练过程的整体稳定性。其次本文决策方法的奖励变化曲线平滑、快速上升至一个明显更高的平台,且波动范围最小。这证明了“分层架构”通过缩减决策空间,降低了学习难度,使策略优化过程更加高效和稳定。标准 PPO 的奖励变化曲线变化幅度波动性较大且最终收敛水平较低,这也间接表现出在本文设定的高动态空战对抗决策仿真实验中,在线策略的 PPO 算法在样本效率和训练稳定性上存在不足,而离线学习与精心设计的数据复用机制可以在一定程度上加快提升智能体的学习性能。

图 4 中展示了本文提出的改进 DQN 算法、原始 DQN 算法,以及 PPO 算法各自在智能体训练过程中的对抗胜率的变化情况。从图 4 可以看出,IDQN 的对抗胜率可以达到更高的数值约为 0.80,且其负率最终降至约为 0.05,这表明了本文提出的决策方法可以使得智能体具有更强的对抗能力。相比之下,基准 PPO 方法胜率曲线增长缓慢且伴随反复,胜率、负率、平局率三者长期交织,表明智能体始终在探索基础战术,未能形成稳定、高胜率的优势策略。DQN 方法的胜率、负率、平局率三者在前期也存在交织变化现象,虽然最终胜率变化高于负率和平局率,但是增长的幅度不是很明显。本文所提的决策方法在训练早期,胜率即快速脱离随机水平并持续上升,负率相应迅速下降。这对应了智能体通过模仿专家,迅速掌握了基本的攻防转换和战术决策能力。中期以后,胜率曲线以更稳定的斜率增长,平局率开始显著下降,这表明智能体通过自博弈训练机制开始学习到更加稳定且有效的空战对抗策略。最终,智能体的胜率稳定在约 0.80 的高位水平,显著高于传统 DQN 的 0.43 和基准 PPO 的 0.40。对抗胜率的变化情况对比实验表明本文方法可以有效实现智能体性能的稳健提升,并极大提升智能体的空战对抗能力。



(a) IDQN对抗胜率变化　(b) DQN对抗胜率变化　(c) PPO对抗胜率变化

图 4　不同方法对抗胜率变化

Fig. 4　Variation of confrontation winning rate of different methods

为更好地对比本文决策方法的决策性能，将分别采用本文决策方法、传统 DQN 方法，以及基准 PPO 算法所控制的智能体分别与一个共同的对手进行对抗，并在每次仿真对抗回合中都采用随机的初始态势参数设置。仿真对比的重点主要包括两个指标：效率成本比（越低越好，反映消耗的导弹数量与消灭的目标数量之比）以及每个决策步骤的 CPU 时间（越低越好，表示决策速度越快）、对抗胜率。经过 5 次仿真试验取平均值以减少随机性，详细对比结果见表 8。

表 8　空战决策指标对比

Tab. 8　Comparison of air combat decision-making indicators

性能指标	效率成本比	单步决策耗时/ms	对抗胜率/%
本文方法	0.36	1.23	78.5
原始 DQN	0.42	1.19	65.2
PPO	0.48	1.35	58.7

从表 8 可以看出，效率成本比综合反映了决策方法所控制的智能体利用导弹击杀对手的情况优劣，可以看出本文方法的效率成本比平均为 0.36，这反映出了本文决策方法所控制的智能体可以使用更少的导弹数量来击杀对手。单步决策耗时表示从状态输入到动作输出的计算时间，反映出了算法的实时性，本文决策方法的平均决策耗时为 1.23 ms，少于基准 PPO 算法耗时，但多于原始 DQN 算法耗时，这主要是由于本文决策方法所采用的分层决策机制和两阶段训练机制在一定程度上增加了决策时间，但是在对抗胜率指标的效果衡量中可以看出，本文决策方法的平均对抗胜率达到约 78.5%，高于原始 DQN 算法和基准 PPO 算法的平均对抗胜率，这也进一步验证了本文决策方法可以使得飞机智能体在满足空战高动态环境下的实时性要求的情况下具有更强的空战对抗能力。

4.3　决策方法对比消融实验

为进一步验证本文所提方法决策机制的有效性，将本文提出的改进决策方法 IDQN 与两种变体进行对比：ECM（expert confrontation method，移除自博弈，全程与专家对手对抗）和 SGM（self-game method，移除专家引导，全程自博弈训练）。对比的指标主要包含决策方法在智能体学习训练过程中的累积奖励的变化情况以及对抗胜率比例的变化情况，在每一轮的对抗训练中，红方智能体的对抗初始态势在攻角角度劣势的变化范围[90°,120°]内随机初始化以更进一步检验智能体的对抗性能，具体仿真如图 5、6 所示。

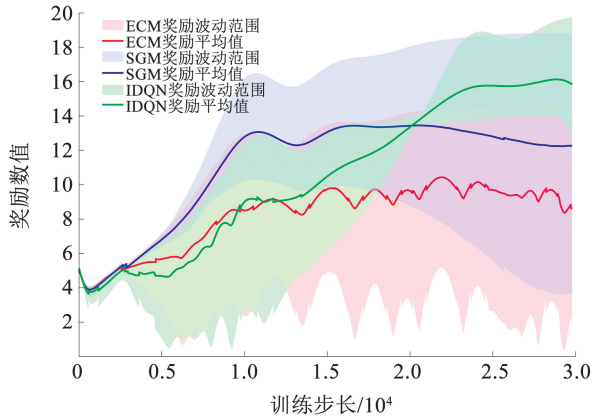
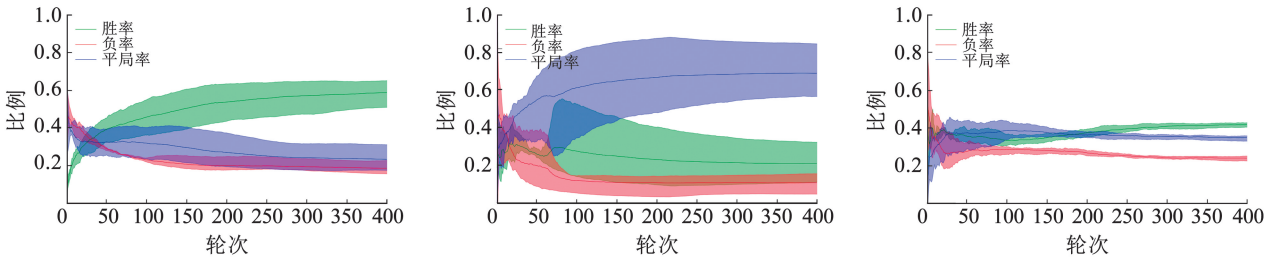


图 5　不同训练机制奖励变化情况对比

Fig. 5　Comparison of reward changes of different training mechanisms

图 5 展示了 3 种不同方法机制的奖励变化情况。从图 5 可以看出，本文的决策方法虽然收敛速度较慢，但最终所达成的奖励收敛水平最高，而 ECM 方法受限于和固定专家对手的对抗训练，因而其容易陷入局部最优，ECM 训练机制初期收敛较快，但很快陷入平稳变化阶段，其最终达成的奖励稳定收敛水平最低，这正是其策略被固定专家局限的体现。相比之下，自博弈方法奖励变化速率较快，这

可能是由于另外两种方法中都结合了专家对手的对抗训练机制,因而需要更多的训练步长来达到收敛水平,SGM 训练机制在整体训练过程中通过自我探索能够有效提升智能体奖励收敛水平且高于 ECM 训练机制,这显示出了该方法的探索训练的有效性,但是同样其最终奖励收敛水平也低于本文决策方法,且其中后期的变化波动幅度也大于本文的决策方法,这表明本文决策方法可以综合利用专家知识和有效探索的优点,在保证平稳变化提升的同时来达到更高的奖励收敛水平。



(a) IDQN对抗胜率变化　　(b) ECM对抗胜率变化　　(c) SGM对抗胜率变化

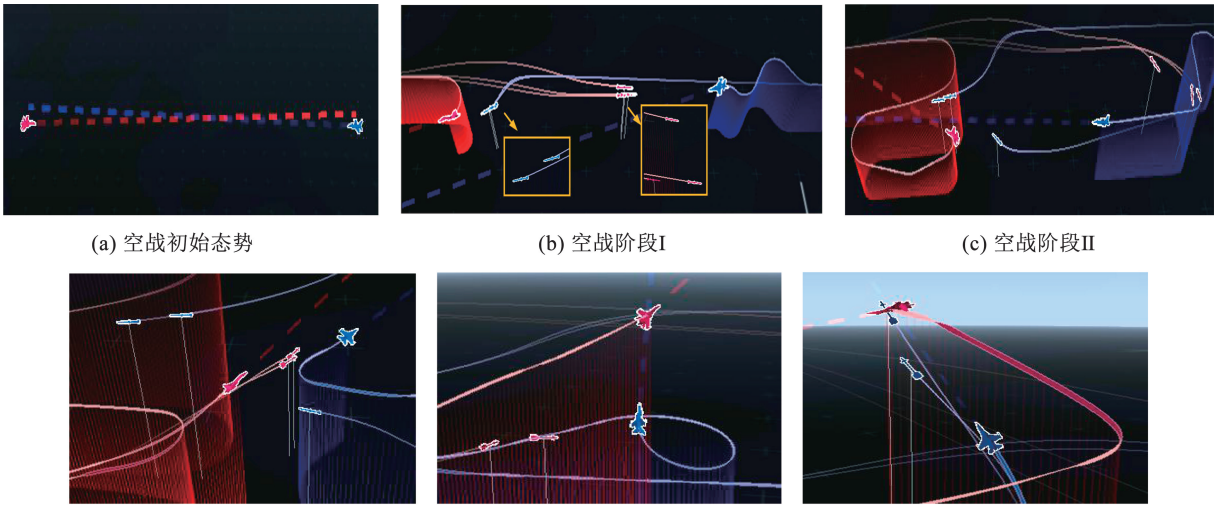
图 6　不同训练机制对抗胜率变化情况对比

Fig. 6　Comparison of confrontation winning rate changes of different training mechanisms

图 6 展示了 3 种不同决策方法机制的对抗胜率变化情况。从图 6 可以看出,ECM 方法由于直接全程和专家对手进行对抗训练,因而其对抗胜率一直处于较低水平(胜率被限制在约 0. 20 的较低水平),而其平局率占比最大,这可能是由于专家对手较为强大,因而其对抗情况一直处于劣势水平,这表明仅向固定专家学习,智能体最多只能达到与专家性能相近的水平,无法激发智能体学习到较好的应对策略。相比之下,SGM 方法的对抗胜率变化情况较好(最终胜率约 0. 40),虽然其胜率优于 ECM 训练机制,但是其胜率、负率,以及平局率三者变化幅度较为接近,策略波动大、不稳定,这反映了从零开始的自博弈探索效率低下,且容易陷入策略循环。本文提出的决策方法的对抗胜率最终达到最高水平约为 0. 60,显著高于 ECM 和 SGM 训练机制,且其平局率和负率也降至较小范围内,这有效证明“两阶段训练”是一个有机整体:第 1 阶段(ECM)提供了高质量的策略起点和搜索先验,使第 2 阶段的自我博弈能够在一个更高的基础上,更高效地探索超越已知专家知识的策略空间,从而学得既稳健又强力的策略。

4.4　空战场景策略仿真实验

为突出验证本文所提出决策方法的机动策略的有效性,开展红、蓝双方空战机动决策场景仿真实验,空战双方的初始态势依据表 7 中的相关指标随机初始化,其中蓝方智能体由本文所提出的决策方法进行训练控制,红方智能体由原始 DQN 方法控制训练,具体的仿真场景如图 7 所示。



(a) 空战初始态势　　(b) 空战阶段I　　(c) 空战阶段II

(d) 空战阶段III　　(e) 空战阶段IV　　(f) 空战阶段V

图 7　红、蓝双方空战对抗场景

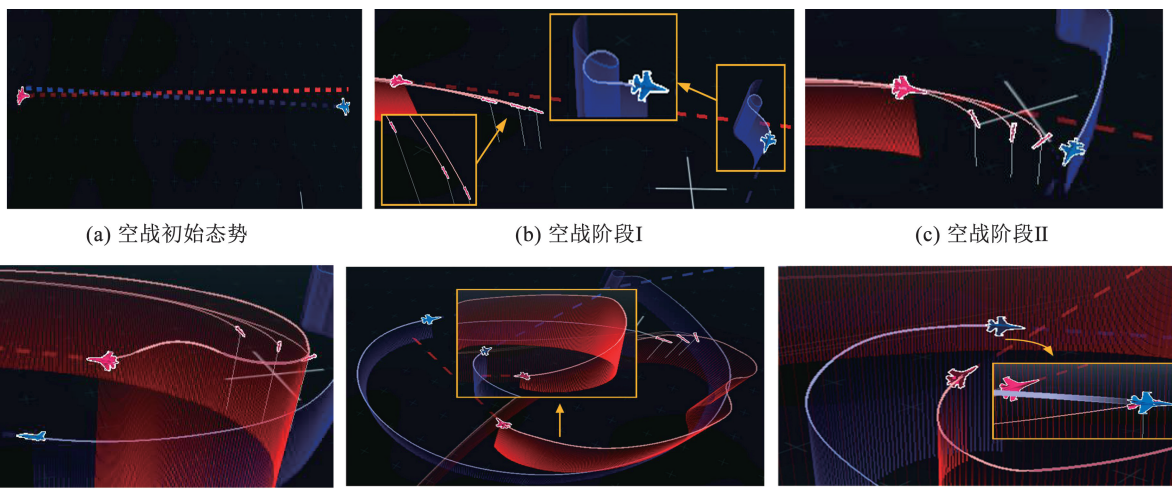
Fig. 7　Air combat confrontation scenario between red and blue sides

从图 7(a)可以看出,在初始状态下,红、蓝双方战机相互对峙并将机头指向对方以试图寻找攻击时机。从图 7(b)可以看出,红、蓝双方在相互接近的过程中,蓝方战机预先执行 F-pole 机动来迷惑对方的同时寻找合适的攻击角度,此时红、蓝双方互发导弹攻击,而红方此时开始施展置尾降高机动来规避导弹攻击。从图 7(c)可以看出,蓝方通过 F-pole 机动预先消耗了红方导弹能量,最后凭借偏置规避机动成功躲避红方导弹攻击,而红方此时也成功规避了蓝方导弹攻击,但此时蓝方机头改变幅度较小,因此先于红方捕获到合适的攻击角度并发射导弹攻击,而红方此时也再次执行 U 形转弯机动来规避导弹攻击。

从图 7(d)可以看出,红方在成功规避蓝方的第 2 波导弹攻击之后,开始再次向蓝方发射导弹攻击,而蓝方此时也执行 U 形转弯机动来规避导弹攻击。从图 7(e)可以看出,在前期空战交互之后,红、蓝双方均没有有效击杀对方,而红方此时开始爬升以试图占据高度优势,而蓝方此时试图借助赫伯斯特机动来爬升的同时改变机头指向来寻找合适的攻击角度,此后蓝方战机通过眼镜蛇机动来适当减速保持对红方的角度追踪并最终成功击杀红方(图 7(f))。

为进一步验证本文决策方法的有效性,开展了红、蓝双方空战对抗以及关键指标分析仿真实验,其中:红方为我方战机,蓝方为对手战机。具体仿真实验情况如图 8 所示。从图 8(a)可以看出,红、蓝双方的初始态势为双机相互对峙,此时红、蓝双方需要通过合理有效的机动决策来形成有利的攻击态势以达成有效击杀对方的作战目的。从图 8(b)可以看出,蓝方在初始时刻开始执行小范围的回转盘旋机动(图 8(b)中右侧黄色框所示),以迷惑红方的同时来适时寻找合适的机动角度,而红方此时在逐渐逼近蓝方的同时先后发射 3 枚导弹(图 8(b)中左侧黄色框所示),以试图从远距离来击杀蓝方。

从图 8(c)可以看出,蓝方在面对红方的导弹攻击时,顺势从回转机动中撤出并向一侧进行偏置规避机动来躲避红方的多枚导弹威胁,最后蓝方成功规避掉红方的导弹攻击。从图 8(d)可以看出,在蓝方成功规避红方的导弹攻击之后,蓝方由于施展了偏置规避机动,因而红方此时成功咬住蓝方尾部,并同时开始爬升以占据高度优势来寻求合适的攻击角度,而蓝方开始向前加速以试图摆脱红方。从图 8(e)可以看出,蓝方在加速摆脱的同时开始执行大回环机动来拖曳红方,而红方此时开始从一定高度向下俯冲以换取机动能量来试图捕捉有利攻角(图 8(e)中黄色框所示),同时蓝方加大回环机动幅度来干扰红方并成功破坏了红方的进攻态势。从图 8(f)可以看出,红方在俯冲未达成有效的攻击角度之后,利用蓝方的回转机动成功绕至其右侧后方并开始爬升来逼近蓝方尾部,而蓝方此时试图加速脱离红方的追击,但是红方此时已成功拉近与蓝方的距离并达成导弹发射角度条件,最终成功发射导弹击杀蓝方(图 8(f)中黄色框所示)。



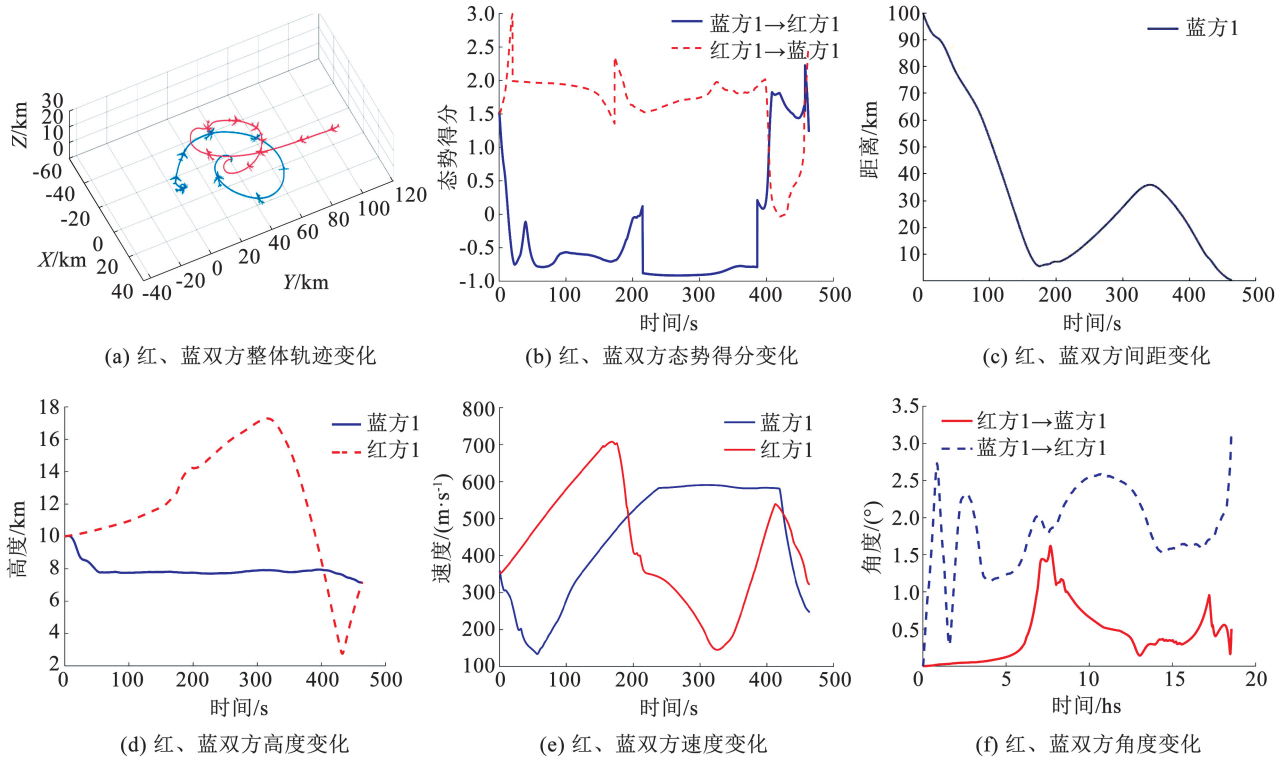
(a) 空战初始态势 (b) 空战阶段I (c) 空战阶段II

(d) 空战阶段III (e) 空战阶段IV (f) 空战阶段V

图 8 红、蓝双方空战对抗场景

Fig. 8 Air combat confrontation scenario between red and blue sides

图 9 展示了红、蓝双方空战过程中的整体机动轨迹和相关指标变化情况。图 9(a)对应图 8 所示的整体空战轨迹变化情况，从图 9(a)可以看出，红、蓝双方在相互接近的过程中，整体上呈现出相互缠绕交互的动态对抗态势，与图 8 所描述的各个阶段的空战对抗过程保持一致。从图 9(b)可以看出，红方的态势得分在整体上高于蓝方，这表示红方的机动决策更加合理有效，以更好地营造有利的攻击态势。从图 9(c)~9(f)可以看出，红方的高度优势在大部分时间都高于蓝方，且红方的攻角角度一直小于蓝方，这表示红方能够更好地瞄准蓝方。红方在前期的速度提升主要在于追击蓝方以更好地咬住蓝方的侧翼薄弱区域，而后期的速度降低主要是为了更好地扭转机头来改变对抗态势，最后的速度上升是为了确保完成击杀任务。综合图 9 的指标变化情况可以看出，随着前期红、蓝双方间隔距离的不断缩近，红方一直占据着高度优势，同时红、蓝双方的速度交织变化情况展示出了前期红、蓝双方追击过程的剧烈性，且对抗全程红方一直占据着攻角优势，这也表明红方一直在蓝方的后方及侧翼范围内来寻找合适的攻击角度。在空战对抗后期，蓝方在一定时间内拉开了与红方的间隔距离，但红方合理利用高度与速度能量之间的转化来再度拉近与蓝方的间隔距离，并在最后成功达成有利的攻击角度条件，这进一步验证了本文所提出的决策方法的机动决策有效性，有力提升了无人机智能体的空战机动能力。



(a) 红、蓝双方整体轨迹变化　(b) 红、蓝双方态势得分变化　(c) 红、蓝双方间距变化

(d) 红、蓝双方高度变化　(e) 红、蓝双方速度变化　(f) 红、蓝双方角度变化

图 9　红、蓝双方空战对抗相关指标对比

Fig. 9　Comparison of relevant indicators of air combat confrontation between red and blue sides

5　结　论

1)设计了一种分层决策架构，通过上层深度 Q 网络(DQN)学习宏观战术指令，下层采用多模式控制与状态驱动机制将指令转化为连续动作，有效解耦了决策过程，降低了学习复杂度与计算负担。

2)提出“专家对抗—自博弈进化”两阶段训练范式。首先利用基于 SECA 规则构建的专家对手为智能体提供高质量初始策略。随后通过虚拟自博弈使其持续与历史最优策略对抗，推动策略突破专家知识局限，提升了策略的适应性与鲁棒性。

3)仿真结果表明，相比传统 DQN 与 PPO 等基准方法，本文方法在奖励收敛速度、最终收敛水平，以及稳定性方面均显著提升，消融实验进一步证明了两阶段训练机制的有效性，表明本文方法在提升决策实时性的同时，显著增强了无人机在动态空战环境中的对抗能力。

参考文献

[1]李永丰，史静平，章卫国，等. 深度强化学习的无人作战飞机空战机动决策[J]. 哈尔滨工业大学学报，2021，53(12)：33
Li Yongfeng，Shi Jingping，Zhang Weiguo，et al. Maneuver decision of UCAV in air combat based on deep reinforcement learning[J].

Journal of Harbin Institute of Technology, 2021, 53(12): 33. DOI: 10.11918/202005108

[2] 赵克新, 黄长强, 魏政磊, 等. 改进决策树的无人机空战态势估计[J]. 哈尔滨工业大学学报, 2019, 51(4): 66
Zhao Kexin, Huang Changqiang, Wei Zhenglei, et al. Situation assessment for unmanned aerial vehicle air combat based on anti-reasoning rules decision tree [J]. Journal of Harbin Institute of Technology, 2019, 51(4): 66. DOI: 10.11918/j.issn.0367-6234.201801006

[3] 潘耀宗, 张健, 杨海涛, 等. 战机自主作战机动双网络智能决策方法[J]. 哈尔滨工业大学学报, 2019, 51(11): 144
Pan Yaozong, Zhang Jian, Yang Haitao, et al. Dual network intelligent decision method for fighter autonomous combat maneuver [J]. Journal of Harbin Institute of Technology, 2019, 51(11): 144. DOI: 10.11918/j.issn.0367-6234.201811083

[4] Li Weihua, Shi Jingping, Wu Yunyan, et al. A multi-UCAV cooperative occupation method based on weapon engagement zones for beyond-visual-range air combat [J]. Defence Technology, 2022, 18(6): 1006. DOI: 10.1016/j.dt.2021.04.009

[5] Han Haoran, Cheng Jian, Lv Maolong, et al. Augmenting the robustness of tactical maneuver decision-making in unmanned aerial combat vehicles during dogfights via prioritized population play with diversified partners [J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(5): 12892. DOI: 10.1109/TAES.2025.3578294

[6] Wang Wenfei, Ru Le, Lü Maolong, et al. Exploring hierarchical hybrid autonomous maneuvering decision-making architecture in beyond visual range air combat[J]. IEEE Transactions on Vehicular Technology, 2025, 74(10): 15491. DOI: 10.1109/TVT.2025.3568377

[7] 车竞, 钱炜祺, 和争春. 基于矩阵博弈的两机攻防对抗空战仿真[J]. 飞行力学, 2015, 33(2): 173
Che Jing, Qian Weiqi, He Zhengchun. Attack-defense confrontation simulation of air combat based on game-matrix approach[J]. Flight Dynamics, 2015, 33(2): 173. DOI: 10.13645/j.cnki.f.d.20150106.001

[8] 刘昊天, 王玉惠, 陈谋, 等. 基于对局迭代的无人机空战博弈研究[J]. 电光与控制, 2022, 29(2): 1
Liu Haotian, Wang Yuhui, Chen Mou, et al. UAV air combat game based on iteration method[J]. Electronics Optics & Control, 2022, 29(2): 1. DOI: 10.3969/j.issn.1671-637X.2022.02.001

[9] 周思羽, 吴文海, 孔繁峨, 等. 基于随机决策准则的改进多级影响图机动决策方法[J]. 北京理工大学学报, 2013, 33(3): 296
Zhou Siyu, Wu Wenhai, Kong Faner, et al. Improved multistage influence diagram maneuvering decision method based on stochastic decision criterions [J]. Transactions of Beijing Institute of Technology, 2013, 33(3): 296. DOI: 10.15918/j.tbit1001-0645.2013.03.012

[10] 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. DOI: 10.2514/1.46815

[11] 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]. 火力指挥与控制, 2019, 44(12): 61
Deng Ke, Peng Xuanqi, Zhou Deyun. Study on air combat decision method of UAV based on matrix game and genetic algorithm[J]. Fire Control & Command Control, 2019, 44(12): 61. DOI: 10.3969/j.issn.1002-0640.2019.12.013

[12] 孟光磊, 罗元强, 梁宵, 等. 基于动态贝叶斯网络的空战决策方法[J]. 指挥控制与仿真, 2017, 39(3): 49
Meng Guanglei, Luo Yuanqiang, Liang Xiao, et al. Air combat decision-making method based on dynamic Bayesian network[J]. Command Control & Simulation, 2017, 39(3): 49. DOI: 10.3969/j.Issn.1673-3819.2017.03.011

[13] Shi Haobin, Xu Meng. A multiple-attribute decision-making approach to reinforcement learning [J]. IEEE Transactions on Cognitive and Developmental Systems, 2020, 12(4): 695. DOI: 10.1109/TCDS.2019.2924724

[14] Xu Xin, Zuo Lei, Li Xin, et al. A reinforcement learning approach to autonomous decision making of intelligent vehicles on highways [J]. IEEE Transactions on Systems, Man and Cybernetics: Systems, 2020, 50(10): 3884. DOI: 10.1109/TSMC.2018.2870983

[15] 吕茂隆, 丁晨博, 韩浩然, 等. 基于深度强化学习的无人机自主感知-规划-控制策略[J]. 自动化学报, 2025, 51(6): 1305
Lü Maolong, Ding Chenbo, Han Haoran, et al. Autonomous perception-planning-control strategy based on deep reinforcement learning for unmanned aerial vehicles[J]. Acta Automatica Sinica, 2025, 51(6): 1305. DOI: 10.16383/j.aas.c240639

[16] Yu Xiaoming, Wu Wenjun, Liao Xingchuang, et al. Dynamic stock-decision ensemble strategy based on deep reinforcement learning[J]. Applied Intelligence, 2023, 53(2): 2452. DOI: 10.1007/s10489-022-03606-0

[17] Ma Xiaoteng, Chen Junyao, Xia Li, et al. DSAC: distributional soft actor-critic for risk-sensitive reinforcement learning[J]. Journal of Artificial Intelligence Research, 2025, 83: 2. DOI: 10.1613/jair.1.17526

[18] Li Bo, Bai Shuangxia, Liang Shiyang, 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. DOI: 10.1049/cit2.12195

[19] 廉云霄, 李霓, 谢锋, 等. 基于时空信息融合的多机协同空战决策方法[J]. 航空学报, 2026, 47(6): 296
Lian Yunxiao, Li Ni, Xie Feng, et al. A multi-UAV cooperative air combat decision-making method based on spatial-temporal information fusion[J]. Acta Aeronautica et Astronautica Sinica, 2026, 47(6): 296. DOI: 10.7527/S1000-6893.2025.32633

[20] 倪浩, 章胜, 刘福炜, 等. 基于域随机化增强 EfficientZero 的无人机空战智能决策[J]. 控制与决策, 2025, 40(11): 3273
Ni Hao, Zhang Sheng, Liu Fuwei, et al. UAV air combat intelligent decision-making based on domain randomization enhanced EfficientZero[J]. Control and Decision-Making, 2025, 40(11): 3273. DOI: 10.13195/j.kzyjc.2024.1443

[21] Wang Wenfei, Ru Le, Lü Maolong, et al. Dynamic and adaptive learning for autonomous decision-making in beyond visual range air combat [J]. Aerospace Science and Technology, 2025, 163: 110327. DOI: 10.1016/j.ast.2025.110327

[22] Hou Yueqi, Liang Xiaolong, Zhang Jiaqiang, et al. Hierarchical decision-making framework for multiple UCAVs autonomous confrontation[J]. IEEE Transactions on Vehicular Technology, 2023, 72(11): 13953. DOI: 10.1109/TVT.2023.3285223

[23] Yang Bingwei, Mo Li, Lü Maolong, et al. Exploring critical decision schemes and tactical parameters in typical beyond visual range air combat[J]. IEEE Transactions on Vehicular Technology, 2025, 74(1): 348. DOI: 10.1109/TVT.2024.3458076

[24] 史桐雨, 王昊, 王酉琨, 等. 无人作战飞机超视距空战博弈对抗决策仿真[J]. 现代防御技术, 2026, 54(3): 93
Shi Tongyu, Wang Hao, Wang Youkun, et al. Simulation of game-theoretic decision-making for beyond-visual-range combat with UCAVs[J]. Modern Defence Technology, 2026, 54(3): 93. DOI: 10.3969/j.issn.1009-086X.2026.03.009

(编辑 张 红)