http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2026.0377

基于自适应加权优先虚拟自博弈的多机协同空战决策算法

周尧明[1,*]，韩韬[1]，姜雨菲[1]，杨政[2]，田艺枫[2]

(1. 北京航空航天大学 航空科学与工程学院，北京 100191；2. 中国电子科技集团公司第二十九研究所，成都 610036)

摘 要：针对多机协同超视距(BVR)空战自博弈训练中收敛速度慢、策略循环振荡的问题，提出了一种结合优先虚拟自博弈(PFSP)与自适应胜局加权机制的多机协同空战决策算法。在多智能体强化学习框架下构建4v4超视距空战仿真对抗环境；引入Elo机制动态评估策略能力，实现历史对手策略的非均匀采样，打破策略循环；同时，设计基于胜率的损失函数加权机制，在初期高价值样本稀缺情况下，引导梯度更新方向。实验结果表明，所提算法在第200个训练回合便逼近平衡状态，较传统FSP算法的收敛效率提升约55%，且中后期的博弈振荡幅值相对较小。进一步采用消融实验证实了损失函数加权机制的加速收敛效果。采用PFSP算法的自博弈智能体策略在面对专家对手时，能够取得70.3%的胜率，展现出强大的空战决策能力。

关 键 词：空战决策；自博弈；强化学习；胜局加权；多机协同
中图分类号：V279
文献标志码：A **文章编号**：1001-5965(2026)10-3580-12

制空权是现代信息化战争中的关键，掌握制空权在一定程度上意味着获得了战争的主动权[1]，随着智能技术的日趋成熟，空战正朝着分布化、协同化、无人化、智能化的方向发展，与传统空战相比，未来空战的战争形态和面貌正在发生深刻变化。其中，多架高性能战斗机可在数据链的支持下协调配合，进行超视距(beyond-visual-range, BVR)空战。超视距空战也被称为中远距空战，是指基于中远距空空导弹，交战双方飞行员在目视观察范围以外进行的空战，与近距离空战相比，主要区别在于作战距离和作战方式[2]。随着战场环境的复杂化，飞行员面临态势感知压力激增与决策窗口缩短的挑战，因此，将无人机与先进的自主智能决策技术相结合具有重要意义。

近年来，深度强化学习(deep reinforcement learning, DRL)凭借其在处理高维连续状态空间和非线性决策问题上的卓越表现，在空战决策领域得到了广泛应用[3-5]。相比于传统的专家经验决策算法，DRL能够通过与环境的交互反馈获得奖励，不断自主优化策略。当前，较为常用的DRL算法包括深度Q网络(deep Q-network, DQN)、近端策略优化(proximal policy optimization, PPO)、深度确定性策略梯度(deep deterministic policy gradient, DDPG)等。

Li等[6]针对信息缺失问题，将Transformer网络引入DDPG算法的Actor-Critic框架，有效提升了部分可观测环境下的决策胜率与稳定性。Yang等[7]提出了耦合奖励机制，通过将角度与距离奖励相乘而非传统加权相加，使二者相互制约，在动态场景中展现出更优的收敛速度与鲁棒性。Wang等[8]提出将双向长短期记忆网络(bidirectional long short-term memory, BiLSTM)与多头自注意力机制融合

收稿日期：2026-06-09；**录用日期**：2026-07-17；**网络出版时间**：2026-08-06 14:42
网络出版地址：link.cnki.net/urlid/11.2625.V.20260806.1112.002
*通信作者. E-mail：zhouyaoming@buaa.edu.cn

引用格式：周尧明，韩韬，姜雨菲，等. 基于自适应加权优先虚拟自博弈的多机协同空战决策算法[J]. 北京航空航天大学学报，2026，52(10)：3580-3591. Zhou Y M，Han T，Jiang Y F，et al. A multi-aircraft cooperative air combat decision-making algorithm based on an adaptive weighted priority virtual self-play theory approach[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(10)：3580-3591 (in Chinese).

以增强态势感知, 实现了在复杂对抗中更高的平均回报与胜率。Mei 等[9] 提出一种双层分层决策框架, 利用门控循环单元(gated recurrent unit, GRU)的网络记忆能力补偿部分观测信息, 并结合自博弈课程学习, 显著提升了信息缺失环境下的决策鲁棒性与胜率。Zhang 等[10] 构建了融合规则、子目标与动态态势评估的复合奖励范式, 有效平衡了奖励信号的引导, 提升了训练效率与最终性能。Zhu 等[11] 通过引入中断机制与分层动作空间, 使智能体能够快速响应突发威胁并组合复杂机动, 结合高保真分布式仿真, 实现了对专家规则系统的战术超越。Gao 等[12] 通过改进 TD3 算法并引入场景迁移训练, 大幅提升了训练效率, 所生成的近距机动策略与飞行员经验高度吻合。Shi 等[13] 提出了融合态势感知、决策与控制的完整框架, 其创新的优先航向轮询随机观测权重(priority heading polling random observation weight, PHP-ROW)方法在六自由度模型上有效解决了超视距博弈中机动性与稳定性的平衡难题。

自博弈机制可以与 DRL 结合, 在不需要人类先验知识设计专家对手的条件下, 通过“左右互搏”的方式不断进化, 从而自发涌现出超越人类预设逻辑的高级战术[14]。Li 等[15] 将并行自博弈与软演员-评论家(soft Actor-Critic, SAC)算法结合, 通过多环境中的样本共享和策略共享提升了无人机机动决策的训练效率与泛化能力; Kong 等[16] 面向双机编队近距空战构建分层决策架构, 并利用竞争式自博弈联合训练主策略与子策略, 增强了策略的对称性与鲁棒性。在复杂观测与高保真验证方面, Tasbas 等[17] 将竞争式自博弈与状态堆叠结合, 提升了智能体在传感器噪声条件下的胜率与稳健性; 单圣哲等[14] 采用基于群落的自博弈训练连续动作空战智能体, 使其相对单智能体训练模型取得 70% 以上胜率, 并涌现出类似人类“单环/双环”的战术。

尽管上述研究验证了自博弈在生成强对手、促进战术涌现和提升策略鲁棒性方面的潜力, 现有工作仍主要集中于近距一对一或双机编队场景, 对多机协同超视距空战关注不足。将自博弈机制直接应用于多机超视距空战环境时, 仍面临以下 2 个问题: ①对手策略持续变化使训练环境呈现非平稳性, 超视距空战策略库中往往存在类似“剪刀-石头-布”的循环克制关系, 导致传统自博弈在采样对战对手时容易陷入性能振荡, 无法保证策略能力持续进化, 甚至出现灾难性遗忘; ②自博弈要求在无人类专家作为对手指导下自发进化, 导致在训练初期智能体成长极为缓慢, 高价值胜利样本较为稀缺, 易被大量的无意义经验样本所淹没, 最终使得收敛缓慢甚至停滞。

针对上述问题, 本文在多智能体 PPO (multi-agent PPO, MAPPO)算法的基础上, 引入优先虚拟自博弈(priority fictitious self-play, PFSP)机制, 并设计了优势经验损失加权的算法改进, 建立了无人机自博弈决策模型。本文主要创新如下:

1) 基于 Elo 机制的 PFSP。本文摒弃了传统的均匀采样策略, 在策略池中引入 Elo 评分系统与优先选择机制, 通过动态评估历史策略的绝对实力, PFSP 能够在“探索”与“利用”之间实现自适应平衡, 打破策略循环, 保证模型能力的持续进化。

2) 自适应胜率优势经验损失加权的 MAPPO 算法。本文提出了一种新型的梯度加权策略, 根据当前对局胜率动态调整不同样本在损失函数中的权重, 该机制能够显著增强高价值经验对策略更新的指引作用, 大幅提升样本利用率并加速收敛。

1　超视距空战环境建模

为真实刻画超视距空战中飞机的机动特性, 本文采用六自由度刚体动力学模型对飞机进行建模。该模型能够同时描述飞机在三维空间中的平动与转动行为, 相较于三自由度飞行运动, 六自由度这种更高维度的机动模型能够更准确地模拟无人机的姿态变化, 更接近无人机的真实飞行[18]。

1.1　六自由度无人机机动模型

本文动作控制采用标准飞行控制分层结构, 上层模块将期望高度、期望速度、期望航向角转换为姿态角和推力指令, 下层控制模块再将姿态角与推力转换为舵面偏角与油门, 实现了飞行器稳定控制。本文假设无人机为左右对称的理想刚体, 实验中所处空域的重力加速度、空气密度和无人机自身质量始终保持不变。图 1 为飞机模型示例。

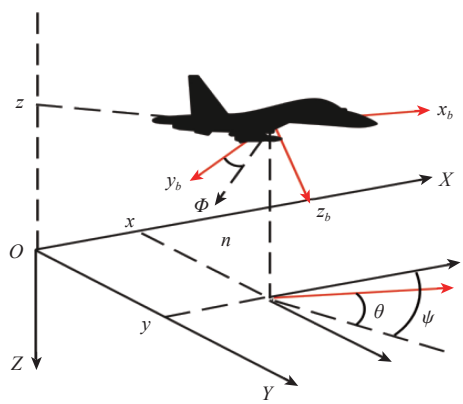


图 1　飞机模型

Fig. 1　Aircraft model

六自由度飞机的状态量可由如下非线性运动方程控制：

$$\begin{cases}\dot{x}=V\cos\gamma_{\mathrm{f}}\cos\chi\\ \dot{y}=V\cos\gamma_{\mathrm{f}}\sin\chi\\ \dot{z}=-V\sin\gamma_{\mathrm{f}}\\ \dot{\Phi}=p+\tan\theta(u\cos\Phi+q\sin\Phi)\\ \dot{\theta}=q\cos\Phi-u\sin\Phi\\ \dot{\psi}=(u\cos\Phi+q\sin\Phi)/\cos\theta\\ \dot{V}=\dfrac{F_T\cos\alpha-F_D}{m}-g\sin\gamma_{\mathrm{f}}\\ \dot{\gamma}_{\mathrm{f}}=\dfrac{(L+F_T\sin\alpha)\cos u}{mV}-\dfrac{g}{V}\cos\gamma_{\mathrm{f}}\end{cases}\tag{1}$$

式中：V为飞机的速度；$[x,y,z]$为飞机在三维空间中的坐标；α为飞机的迎角；γ_{f}为飞机的航迹倾角；χ为飞机的航迹偏角；$[\Phi,\theta,\psi]$分别表示飞机的滚转角、俯仰角和偏航角；$[p,q,u]$分别表示滚转角速度、俯仰角速度、偏航角速度；F_T为推力；F_D为阻力；g为重力加速度；m为飞行器质量；L为升力。

1.2 机载雷达建模

为描述超视距空战中无人机对目标的搜索与跟踪过程，本文建立机载雷达探测模型。以一架我方战机和一架敌方战机为例，位置坐标分别为$\boldsymbol{P}_A=(x_A,y_A,z_A)$和$\boldsymbol{P}_B=(x_B,y_B,z_B)$，可以得到敌方战机$B$相对于我方战机$A$的相对位置$\boldsymbol{P}_{AB}=(x_B-x_A,y_B-y_A,z_B-z_A)$。

1）雷达距离探测约束。由此可以得出二者之间的距离为

$$d_{AB}=\|\boldsymbol{P}_{AB}\|=\sqrt{(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2}\tag{2}$$

本文中，机载火控雷达采用基本雷达方程进行建模：

$$d_{\max}=\left[\frac{P_{\mathrm{tx}}G^2\xi}{(4\pi)^3P_{\min}}\right]^{1/4}\tag{3}$$

式中：P_{tx}为雷达发射功率；G为雷达天线的增益；ξ为目标雷达截面积；$P_{\min}$为雷达最小可检测信号；$d_{\max}$为雷达对于雷达截面积为ξ的目标最大作用距离。

则得到目标满足距离探测条件：

$$d_{AB}\leqslant d_{\max}\tag{4}$$

2）雷达方位角探测约束。设本机速度方向单位矢量为

$$\boldsymbol{e}_v=\frac{\boldsymbol{V}_A}{\|\boldsymbol{V}_A\|}\tag{5}$$

目标相对方向单位矢量为

$$\boldsymbol{e}_r=\frac{\boldsymbol{P}_{AB}}{\|\boldsymbol{P}_{AB}\|}\tag{6}$$

可以得到目标相对方位角为

$$\lambda_{AB}=\arccos(\boldsymbol{e}_v\boldsymbol{e}_r)\tag{7}$$

则得到目标满足雷达方位角探测条件：

$$|\lambda_{AB}|\leqslant\lambda_{\max}\tag{8}$$

式中：$\lambda_{\max}$为雷达水平搜索半视场角。

3）雷达俯仰角探测约束。目标相对俯仰角定义为

$$\theta_{AB}=\arctan\left(\frac{z_B-z_A}{\sqrt{(x_B-x_A)^2+(y_B-y_A)^2}}\right)\tag{9}$$

设雷达俯仰搜索半视场角为$\theta_{\max}$，则得到目标满足雷达俯仰角探测条件：

$$|\theta_{AB}|\leqslant\theta_{\max}\tag{10}$$

综合距离、方位和俯仰约束，目标探测判定函数定义为

$$I_{AB}=\begin{cases}1 & d_{AB}\leqslant d_{\max},|\lambda_{AB}|\leqslant\lambda_{\max},|\theta_{AB}|\leqslant\theta_{\max}\\ 0 & \text{其他}\end{cases}\tag{11}$$

$I_{AB}=1$表示成功探测目标，$I_{AB}=0$表示不能探测到目标。当目标满足探测条件时，本机可获得目标的距离、方位及速度等局部信息，并更新目标态势信息；否则，目标信息置为空。

1.3 导弹交战规则模型

1.3.1 导弹运动学建模

本文选用AIM-120D空空导弹并建立运动学模型。建立以导弹重心O_M为原点的弹体坐标系$O_Mx_My_Mz_M$。其中，O_Mx_M位于导弹对称平面内，平行于弹身轴线，指向弹头方向；O_Mz_M轴位于导弹的对称面内，垂直于O_Mx_M轴且方向向下；O_My_M轴垂直于战机的对称面，遵循右手定则方向向右。

导弹的运动学模型可表示为

$$\begin{cases}\dot{x}_M=V_M\cos\sigma\cos\varepsilon\\ \dot{y}_M=V_M\cos\sigma\sin\varepsilon\\ \dot{z}_M=-V_M\sin\sigma\end{cases}\tag{12}$$

式中：(x_M,y_M,z_M)为导弹的空间坐标；V_M为导弹的速度大小；σ和ε分别为导弹的弹道倾角和弹道偏角，其中，弹道倾角σ为弹体轴O_Mx_M与水平面$O_gx_gy_g$之间的夹角，规定弹头上扬时为正，弹道偏角ε为弹体轴O_Mx_M在水平面$O_gx_gy_g$上的投影与O_gx_g轴之间的夹角，规定导弹右偏航时为正。

1.3.2 基于专家规则的导弹发射决策模型

在超视距空战仿真建模中，导弹发射不仅依赖于底层的导弹动力学包线，更受到战术准则、战场态势感知及火力资源配置等多重约束。本文基于专家系统框架，构建了一套多维度的导弹发射逻辑模型。该模型在无人机进入攻击阶段后，通过对感知目标列表进行条件过滤与动态分配，最终输出开火指令及导弹协同参数。

1) 传感器跟踪精度约束。导弹的火控解算高度依赖于机载传感器(雷达/红外搜索与跟踪(infrared search and track, IRST)系统)对目标的跟踪精度。设定目标的距离误差标准差为σ_{R}, 方位角误差标准差为σ_{Az}, 俯仰角误差标准差为σ_{El}。只有当$(\sigma_{\mathrm{R}} \leqslant \sigma_{\mathrm{R0}}) \cap (\sigma_{\mathrm{Az}} \leqslant \sigma_{\mathrm{Az0}}) \cap (\sigma_{\mathrm{El}} \leqslant \sigma_{\mathrm{El0}})$时, 才允许发射导弹, 其中, σ_{R0}、σ_{Az0}、σ_{El0}分别为对应导弹允许的最大误差阈值。

2) 离轴角发射约束。为确保导弹发射后中继数据链的稳定传输, 对发射时刻的离轴角λ进行约束, 离轴角定义为我方载机机体纵轴(机头指向)与弹目视线之间的夹角。根据专家经验, 设置最大离轴角为30°。当$\lambda \leqslant 30°$时, 认为敌机在机载火控雷达的照射扇区或导弹导引头的最大搜索视场内, 允许导弹发射。

3) 防误伤机制。当敌我机群发生近距离混战时, 主动雷达弹易发生误伤。遍历我方无人机的位置并计算目标与最近我方战机的距离, 如果距离过近, 则为避免误伤友机, 禁止发射。

4) 飞机姿态与发射时间间隔。导弹的超视距有效射程高度依赖于其发射时的初始动力学条件, 当载机处于过大负俯仰角状态下发射时, 导弹弹道将迅速指向低空高密度大气区域, 导致空气动力阻力急剧增加, 动能过快耗散, 进而压缩导弹的动力学杀伤包线。定义发射最低俯仰角$\theta_{\min} = -5°$, 以保证导弹获得平飞或跃升初始姿态。

此外, 对最小发射时间间隔做出限制, 预定2枚导弹之间的最小时间间隔为0.5 s, 以符合实际空战场景。

2　基于优先虚拟自博弈的无人机集群博弈对抗算法

在多无人机集群超视距对抗任务中, 智能体之间的博弈关系呈现出典型的强对抗性与非平稳性特征。随着策略的持续更新, 环境动态分布不断变化, 使得强化学习过程面临显著的训练不稳定问题。此外, 超视距空战奖励稀疏, 加剧了策略收敛的难度。本文以4v4超视距空战为例, 设计了基于Elo评分的PFSP机制, 并引入胜率驱动的自适应损失加权算法, 引导梯度更新方向。

2.1　基于 Elo 评分的优先虚拟自博弈机制

在多智能体强化学习中, 传统的虚拟自博弈(fictitious self-play, FSP)通常采用均匀采样的方式从历史策略池中抽取对手进行训练。然而, 这种方式忽略了不同历史策略对当前智能体能力提升的差异性, 容易导致训练效率低下。为解决这一问题, 本文引入了PFSP机制。该机制的核心思想是: 通过对策略池中的历史策略进行能力评估, 并基于评估结果进行非均匀采样, 从而优先选择对当前策略具有较高学习价值的对手参与训练。为量化策略库中各策略的真实水平并计算抽取概率, 本文引入了动态Elo等级分系统。

2.1.1　Elo 评分机制

Elo等级分机制的核心原理是: 为每个参与博弈的策略赋予一个初始分数, 并通过与其他策略的实际对弈结果(胜、负、平)来动态调整该分数。分数的变动幅度不仅取决于实际对局结果, 还取决于双方对弈前的实力差距。本文将策略池中策略的初始Elo评分统一设置为1 200分。对于训练过程中达到入池条件的新策略, 其Elo分数由其在进入策略池前的对抗评估过程决定。具体而言, 候选策略在满足入池条件之前, 需要与策略池中的已有策略进行多轮博弈, 并根据每轮对抗结果按照Elo更新公式持续调整自身评分。当候选策略经过评估达到预设入池标准后, 将其当前累计更新后的Elo评分作为加入策略池时的初始评分。该设计能够保证新加入策略的评分与其实际训练表现保持一致, 其入池的Elo评分能够反映该策略相对于已有策略的实际表现。

设当前策略与对手策略的Elo评分分别为E_i和E_j, 则当前策略的期望胜率为

$$P = \frac{1}{1 + 10^{(E_j - E_i)/400}} \tag{13}$$

根据对战结果, 对Elo评分进行如下更新:

$$E_{i'} = E_i + k_{\mathrm{Elo}}(s_i - P) \tag{14}$$

式中: s_i表示实际对战结果, 胜出时$s_i = 1$, 平手时$s_i = 0.5$, 失败时$s_i = 0$; k_{Elo}为更新系数, 用于控制评分变化幅度。在训练初期, 为提高评分更新的敏感性, 采用较大的k_{Elo}值; 随着训练进行, 逐步减小k_{Elo}, 以提升评分稳定性。具体而言, 本文采用随训练轮次递减的动态更新策略, 使Elo评分能够兼顾收敛速度与稳定性。

2.1.2　多局对抗下的 Elo 更新

由于单次对抗结果具有较大的随机性, 在实际训练中, 每轮策略评估由多局对战构成。为降低偶然性对评分的影响, 本文采用多局累积更新的方式对Elo评分进行修正, 即在多次对战结果基础上迭代更新评分, 从而获得更加稳定的策略能力估计。

$$\begin{cases} E_1 = E_0 + k_{\mathrm{Elo}}(s_0 - P_0) \\ E_2 = E_1 + k_{\mathrm{Elo}}(s_1 - P_1) \\ \quad \vdots \\ E_n = E_{n-1} + k_{\mathrm{Elo}}(s_{n-1} - P_{n-1}) \end{cases} \tag{15}$$

该机制能够有效平滑单局对战中的随机波动，使得Elo评分更具鲁棒性。

2.1.3 基于Elo的优先对手采样机制

在获取了策略池中所有对手的Elo评分后，PFSP机制需要依据分数制定非均匀的对手采样分布。本文采用基于Softmax函数的概率选取方法，使得得分越高的强大对手被抽取的概率越高。策略池中第i个策略被选为对手的概率P_i^{opp}计算如下：

$$P_i^{\text{opp}} = \mathrm{e}^{(E_i/\tau)} \Big/ \sum_j \mathrm{e}^{(E_j/\tau)} \tag{16}$$

式中：τ为温度系数，用于调节抽取策略的探索与利用平衡。当τ取值较大时，指数分布趋于平缓，各个策略被抽取的概率趋于均等，系统倾向于进行全局探索；当τ取值较小时，系统呈现出“赢者通吃”的贪婪特性，几乎只采样当前策略池中Elo得分最高的精英策略。

综合考虑训练的多样性与难度进阶，本文在实验中将温度系数设置为$\tau = 300$，以确保智能体在主要挑战高水平对手的同时，仍保留与历史不同水平策略交手的概率，有效克服神经网络的灾难性遗忘。

2.2 基于自适应胜局信号的策略损失加权机制的MAPPO算法

多智能体空战对抗任务具有高维连续状态空间、强对抗性及多智能体耦合决策等特点。在该类场景中，智能体不仅需要完成个体机动控制（如速度、高度与航向调节），还需在编队协同与敌我对抗中进行策略博弈，使得环境呈现出显著的非平稳性与部分可观测性。

为了形式化描述该过程，本文将多智能体空战对抗任务建模为一个去中心化部分可观测马尔可夫决策过程。该过程可以由一个元组$\langle S, A, O, R, p_{\text{tr}}, n, \gamma \rangle$表示。其中，$A$为各智能体共享的动作空间；$o_t^i = O(s_t; i)$表示智能体$i$在时刻$t$对于环境状态$s_t \in S$的局部观测；$r_t = R(s_t, a_t)$表示各智能体共享的奖励函数；$p_{\text{tr}}$为状态转移方程，描述了在联合动作$a_t$下环境状态的变化；$\gamma \in [0,1)$为折扣因子；$n$为智能体数量。

针对上述问题，本文采用MAPPO算法作为基础强化学习框架。MAPPO算法基于集中训练、分散执行范式，在训练阶段利用全局信息构建价值函数，在执行阶段仅依赖局部观测进行决策，兼顾了多智能体协同建模能力与实际部署可行性。

然而，在多智能体空战对抗任务中，强化学习仍面临显著的稀疏奖励问题。在训练早期阶段，由于智能体策略尚未收敛，胜局出现频率较低，导致：①有效策略样本（胜局经验）占比较小；②胜局经验在梯度更新中被大量无效样本稀释；③PFSP机制导致强化学习要同时针对策略池中多个策略进行更新，学习过程收敛缓慢。

传统方法通常通过建立经验回放池或奖励重塑缓解该问题，但对于基于on-policy的MAPPO算法，直接使用经验回放池会破坏策略分布一致性。因此，本文引入一种基于胜局信号的策略损失加权机制，在不改变采样分布的前提下，提高关键样本对策略更新的贡献。

2.2.1 MAPPO算法框架

MAPPO算法采用Actor-Critic结构，其核心由分散执行的策略网络与集中式价值网络构成。

每个智能体基于自身局部观测s_i输出动作分布$a_i \sim \pi_\vartheta(a_i|s_i)$。

在训练阶段，引入全局状态S构建集中式价值函数$V_\phi(S)$，该结构能够融合敌我双方所有智能体的状态信息，从而提高价值估计精度。

在MAPPO算法中，策略更新基于如下裁剪目标函数：

$$L^{\text{policy}}(\vartheta) = -E_t\left[\min\left(\rho_t(\vartheta)A_t, \text{clip}(\rho_t(\vartheta), 1-\epsilon, 1+\epsilon)A_t\right)\right] \tag{17}$$

式中：$\rho_t(\vartheta) = \pi_\vartheta(a_t|s_t)/\pi_{\vartheta\text{old}}(a_t|s_t)$为重要性采样比，$\vartheta$为神经网络策略；$A_t$为优势函数；$\epsilon$为裁剪系数。该目标函数本质上通过优势函数$A_t$决定策略更新方向，通过裁剪机制限制更新幅度。

2.2.2 自适应胜局加权机制

为提升胜局样本在策略更新中的影响力，本文引入样本权重w_t对策略损失进行加权，如下：

$$L_{\text{weight}}^{\text{policy}}(\vartheta) = -E_t[w_t \cdot \min(r_t(\vartheta)A_t, \text{clip}(r_t(\vartheta), 1-\epsilon, 1+\epsilon)A_t)] \tag{18}$$

式中：权重w_t根据样本t所属对局结果自适应确定，若$t \in$胜局，则$w_t = w_{\text{win}}(w_{\text{win}} \geqslant 1)$，否则，$w_t = 1$。

该加权机制等价于在梯度层面对胜局样本进行放大：

$$\nabla_\vartheta L_{\text{weight}}^{\text{policy}}(\vartheta) \propto w_t A_t \nabla_\vartheta \ln \pi_\vartheta(a_t|s_t) \tag{19}$$

需要指出的是，本文的胜局损失加权机制并未改变PPO算法中重要性采样比值及clip约束过程，而是在裁剪目标函数计算完成后对不同样本的损失贡献进行调整。PPO算法中的clip机制主要限制新旧策略概率比值的变化范围，以避免单次策略更新幅度过大。而本文引入的权重w_t仅作用于样本损失尺度，对梯度贡献进行重新分配，不影响$r_t(\vartheta)$的计算及裁剪边界。因此，本文提出的自适应胜局加权机制在保证on-policy特性的前提下，增强了胜局经验对策略更新的权重。

2.2.3　动态权重调节策略

为避免固定权重导致过拟合或训练不稳定, 本文进一步引入基于胜率的自适应权重调节机制。设当前训练阶段的胜率为r_{current}, 目标胜率为r_{target}, 定义比值为

$$\eta = \frac{r_{\text{current}}}{r_{\text{target}}} \tag{20}$$

则胜局样本权重w_t定义为

$$w_t = \begin{cases} 1 & r_{\text{current}} \geqslant r_{\text{target}} \\ 1 + (1-\eta)(w_{\max} - 1) & r_{\text{current}} < r_{\text{target}} \end{cases} \tag{21}$$

式中: $w_{\max}$为策略损失加权机制的最大权重, 本文中$w_{\max}$设置为 1.5, 将单个胜局样本的最大梯度放大限制在 50%, 以平衡胜局信号增强与策略更新稳定性。通过引入动态策略损失加权机制, 在低胜率的训练前期放大了胜局经验的策略损失权重, 在高胜率阶段回归标准策略避免过拟合, 最终提高了训练效率和策略收敛速度。

2.2.4　加权机制对策略更新的影响分析

1) 策略梯度估计。设 PPO 算法裁剪目标中的单样本目标项为

$$f_t(\vartheta) = \min\left(\rho_t(\vartheta)A_t, \text{clip}(\rho_t(\vartheta), 1-\epsilon, 1+\epsilon)A_t\right) \tag{22}$$

对于包含B_{batch}个样本的训练批次, 标准 MAPPO 算法的策略梯度估计为

$$\hat{\boldsymbol{g}} = \frac{1}{B_{\text{batch}}} \sum_{t=1}^{B_{\text{batch}}} \nabla_\vartheta f_t(\vartheta) \tag{23}$$

引入胜局权重后, 策略梯度估计变为

$$\hat{\boldsymbol{g}}_w = \frac{1}{B_{\text{batch}}} \sum_{t=1}^{B_{\text{batch}}} w_t \nabla_\vartheta f_t(\vartheta) \tag{24}$$

在每次参数更新过程中, 样本权重由对局结果和更新前统计得到的胜率确定, 因此, 可视为与当前梯度计算相独立的常数, 不产生额外的权重求导项。从梯度期望角度, 有

$$E[\hat{\boldsymbol{g}}_w] = E[\hat{\boldsymbol{g}}] + E[(w_t - 1)\nabla_\vartheta f_t(\vartheta)] \tag{25}$$

第 2 项表示胜局经验产生的附加更新方向。因此, 该方法仍使用当前策略采集的轨迹数据, 保持了数据采集的 on-policy 属性, 但改变了原始 PPO 算法目标中不同轨迹的相对贡献, 从而形成面向胜局经验的重加权策略梯度估计。

2) 训练稳定性分析。自适应加权机制在增强有效信号的同时, 也可能放大胜局样本梯度, 从而增大训练早期的梯度波动。令

$$\boldsymbol{g}_t = \nabla_\vartheta f_t(\vartheta) \tag{26}$$

在近似独立采样条件下, 加权梯度估计的方差满足:

$$\text{Var}(\hat{\boldsymbol{g}}_w) = \frac{1}{B_{\text{batch}}} \text{Var}(w_t \boldsymbol{g}_t) \tag{27}$$

由于$1 \leqslant w_t \leqslant w_{\max}$, 其二阶矩满足:

$$E[\|w_t \boldsymbol{g}_t\|^2] \leqslant w_{\max}^2 E[\|\boldsymbol{g}_t\|^2] \tag{28}$$

由此分析可知, 过大的$w_{\max}$可能导致梯度方差和单次策略更新幅度增加。但本文通过以下 3 种方式控制该影响: ①将权重限制在有限区间$[1, w_{\max}]$, 避免少量胜局样本主导整个批次更新; ②PPO 算法裁剪项限制新旧策略概率比, 降低单次更新过大导致策略退化的风险; ③随着当前策略胜率接近目标胜率, 胜局权重逐步衰减至 1, 使算法在训练中后期回归标准 MAPPO 算法更新。因此, 该机制在可控的方差增幅下提高稀缺胜局信号的有效性, 并通过动态调节避免持续放大带来的训练振荡。

3) 探索能力分析。从探索与利用关系来看, 胜局加权提高了已获得成功回报的动作序列对策略更新的影响, 因此, 会使算法在训练初期更加倾向于利用已发现的有效战术。但当胜率提高后, 胜局权重逐渐下降至 1, 避免算法长期偏向少数早期成功策略, 同时 PFSP 机制通过从历史策略池中采样不同能力水平的对手, 使智能体持续面对多样化对抗策略, 从对手分布层面维持策略探索。因此, 动态胜局加权主要在训练初期增强有效经验利用, 随着胜局权重逐渐下降及多样化对手出现, 探索能力逐渐上升。

整体 PFSP 的伪代码如下。

输入: 初始策略π_0; 空策略池$M = \{\pi_0\}$; 初始 Elo 评分$E_0 = 1\,200$; 最大训练轮数T; 每轮训练的步数N; 温度系数$\tau = 300$; 目标胜率r_{target}; 最大损失权重$w_{\max}$。

输出: 训练收敛后的最优策略π^*。

1.初始化, 设置当前训练智能体$\pi_{\text{curr}} \leftarrow \pi_0$。

2. For 训练轮次$t = 1, 2, \cdots, T$ do

3.优先对手采样(PFSP)。计算策略池M中各个历史策略π_i的采样概率: $P_i^{\text{opp}} = \mathrm{e}^{(E_i/\tau)} \Big/ \sum_j \mathrm{e}^{(E_j/\tau)}$, 根据分布随机抽取对手策略$\pi_{\text{opp}}$。

4.多局对抗评估, 初始化本局累计得分E_0。

For 步数$n = 1, 2, \cdots, N$ do:

在 4v4 超视距空战环境中运行π_{curr}与π_{opp}的对局, 获取对局结果$s_n \in \{1, 0.5, 0\}$(胜、平、负), 收集训练轨迹数据D_n用于策略强化学习更新。

End For

5.自适应胜局信号加权更新。计算当前胜率$r_{\text{current}} = N_{\text{win}}/N$, 计算胜率比$\eta = r_{\text{current}}/r_{\text{target}}$。

计算本次训练中胜局样本权重：$w_t=\begin{cases}1 & r_{\text{current}}\geqslant r_{\text{target}}\\ 1+(1-\eta)(w_{\max}-1) & r_{\text{current}}<r_{\text{target}}\end{cases}$。利用加权目标函数更新策略参数$\vartheta$：$L_{\text{weight}}^{\text{policy}}(\vartheta)=-E_t\left[w_t\cdot\min(\rho_t(\vartheta)A_t,\ \text{clip}(\rho_t(\vartheta),1-\epsilon,1+\epsilon)A_t)\right]$。

6.动态Elo评分更新。计算当前策略相对于对手的期望胜率：$P=\dfrac{1}{1+10^{(E_{\text{opp}}-E_{\text{curr}})/400}}$。根据训练进程调整更新系数$k_{\text{Elo}}$。

For $n=1,2,\cdots,N$:

$E_{\text{curr}}\leftarrow E_{\text{curr}}+k_{\text{Elo}}(s_i-P)$

End For

If 当前胜率$r_{\text{current}}\geqslant$目标胜率$r_{\text{target}}$ And 入池间隔$G_{\text{current}}\geqslant$目标入池间隔$G_{\text{target}}$, then:

将当前策略π_{curr}加入策略池M，并记录当前的E_{curr}

End For

Return 最优策略π^*

整体的代码框架如图2所示。

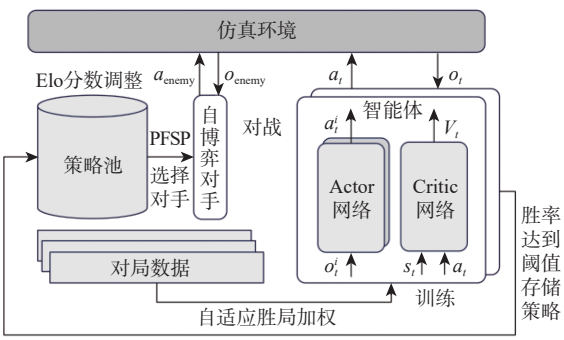


图2 代码框架

Fig. 2 Code framework

3 状态空间与奖励函数设置

3.1 状态空间

根据多机协同超视距空战集中式训练分布式执行架构的特点，分别设计了全局状态空间和局部状态空间，全局状态空间作为集中式训练过程中价值函数的输入，而各智能体的决策只能基于各自的局部观测信息实现。以4v4协同超视距空战为例，我方为蓝方，包括蓝1号机、蓝2号机、蓝3号机、蓝4号机；敌方为红方，包括红1号机、红2号机、红3号机、红4号机。在自博弈情况下，红蓝双方均视为智能体。

3.1.1 全局状态空间设计

本文全局状态空间S设计为8×424大小，每个维度是针对一架飞机的全局状态空间。以蓝1号机为例，其全局状态空间S_1如表1所示。

3.1.2 局部状态空间设计

局部状态空间o设计为8×236大小，每个维度也是针对一架飞机的局部状态空间。以蓝1号机为例，其局部状态空间o_1如表2所示。

3.2 动作空间

本文设计智能体i的动作空间为$a_i=\{h_i,v_i,\psi_i\}$，其中，h_i为期望高度，v_i为期望速度大小，ψ_i为期望航向角。h_i取值为3 000,4 000,5 000,6 000,7 000,8 000,9 000,10 000,11 000,12 000,13 000 m，v_i取值为200,250,300,350,400 m/s，ψ_i取值为0°,45°,90°,135°,180°,225°,270°,315°。动作空间中的期望动作虽为离散值，但战机通过上述期望高度、期望速度、期望角度组成动作对，可以实现在仿真作战环境的自由机动。这是因为战机跟踪期望动作需要时间，因此，即使上层动作指令离散，实际飞行轨迹仍具有连续性和较强的机动灵活性。

3.3 奖励函数

奖励设计是空战中至关重要的一环。本文设计了较为细致的结果奖励和状态奖励，更好地引导无人机根据实际态势进行正确的空战决策，提升无人机作战能力。

3.3.1 状态奖励

状态奖励的目的是缓解只有事件奖励时存在

表1 全局状态空间

Table 1 Global state space

状态分量	描述	维数
本机(蓝1号机)	位置信息，包括x坐标、y坐标、z坐标；	3
	角度信息，包括航向角、俯仰角、滚转角；	3
	速度信息，包括速度大小、北向速度分量、地向速度分量、东向速度分量；	4
	加速度信息，包括北向加速度分量、地向加速度分量、东向加速度分量	3
友机(蓝2号机，蓝3号机，蓝4号机)	与本机类似，但状态信息是相对于本机的相对数据	3×13
敌机(红1号机，红2号机，红3号机，红4号机)	与本机类似，但状态信息是相对于本机的相对数据	4×13
蓝方导弹信息(预留16枚导弹)	相对位置信息(x坐标、y坐标、z坐标)、相对速度信息(速度大小、北向速度分量、地向速度分量、东向速度分量)	16×7
红方导弹信息(预留16枚导弹)	与蓝方导弹信息类似	16×7
导弹告警信息(预留16枚导弹)	方位角、俯仰角(相对值)	16×2
导弹制导信息(预留16枚导弹)	标识符、方位角、俯仰角、制导步数	16×4

表 2　局部状态空间

Table 2　Local state space

状态分量	描述	维数
本机(蓝1号机)	位置信息, 包括x坐标、y坐标、z坐标;	3
	角度信息, 包括航向角、俯仰角、滚转角;	3
	速度信息, 包括速度大小;	1
	加速度信息, 包括北向加速度分量、地向加速度分量、东向加速度分量	3
友机(蓝2号机, 蓝3号机, 蓝4号机)	与本机类似, 但状态信息是相对于本机的相对数据	3×10
敌机(红1号机, 红2号机, 红3号机, 红4号机)	只包括位置信息(x坐标、y坐标、z坐标)、航向角、速度	4×5
蓝方导弹信息(预留8枚导弹)	只包括位置信息(x坐标、y坐标、z坐标)、航向角、速度	8×5
红方导弹信息(预留8枚导弹)	与蓝方导弹信息类似	8×5
导弹告警信息(预留16枚导弹)	方位角、俯仰角(相对值)	16×2
导弹制导信息(预留16枚导弹)	标识符、方位角、俯仰角、制导步数	16×4

的奖励稀疏问题, 引导无人机决策学到进攻、躲弹等决策, 并更多地产生与目标对抗的经验。

1) 探测敌机奖励。为增强智能体的态势感知能力与主动搜索能力, 本文设计了敌方目标探测奖励项。当敌方目标进入己方传感器(如机载雷达)的有效探测范围时, 给予正奖励, 从而引导智能体学习主动占据有利观测位置, 并与敌机产生更多的对抗经验。

具体而言, 设智能体的位置向量$\boldsymbol{P}_i=(x_i,y_i,z_i)$, 敌机的位置向量$\boldsymbol{P}_t=(x_t,y_t,z_t)$, 计算敌机相对我方智能体的方位角$\lambda_{xoy}$和俯仰角$\theta_{xoz}$。当满足以下条件时: $\lambda_{xoy}\in\left[-45°,45°\right]$, $\theta_{xoz}\in[-180°,180°]$, 智能体将得到探测正奖励$r_{\text{detect}}=0.000\,1$。

2) 敌方导弹接近负奖励。为提升智能体在空战环境中的生存能力, 本文在奖励函数中引入了导弹威胁感知惩罚项, 用于表征来袭导弹对自身平台的威胁程度。当敌方导弹进入一定攻击条件范围内时, 对智能体施加负奖励, 以引导其学习规避导弹的机动策略。

具体而言, 设智能体的位置向量$\boldsymbol{P}_i=(x_i,y_i,z_i)$, 导弹的位置向量$\boldsymbol{P}_m=(x_m,y_m,z_m)$, 二者之间的相对位置向量$\boldsymbol{d}=\boldsymbol{P}_i-\boldsymbol{P}_m$, 对应距离$D=\|\boldsymbol{d}\|$, 同时根据导弹的姿态信息计算其速度方向向量(即导弹航向)$\boldsymbol{v}_m$, 进一步定义导弹攻击角为导弹航向与目标方向之间的夹角$\eta_m=\arccos\left(\dfrac{\boldsymbol{d}\cdot\boldsymbol{v}_m}{\|\boldsymbol{d}\|\cdot\|\boldsymbol{v}_m\|}\right)$。当同时满足以下条件时: $|\eta_m|<30°$, $D<18\,\text{km}$, 在该条件下, 智能体将得到一个固定的负奖励: $r_{\text{missile_threat}}=-0.003$。

3) 高度奖励。为引导智能体在不同作战距离下占据合理的高度区间, 从而获得潜在的能量优势与战术机动空间, 本文设计了一种基于敌我距离的自适应高度奖励机制。

首先, 在所有敌方目标中选取与己方距离最近的目标作为主要参考对象, 并计算智能体与其三维距离$d_{\min}=\min\limits_i d_i$。

其次, 为使智能体在不同交战阶段采取不同的高度策略, 定义期望高度h^*并设置距离分段函数:

$$h^*=\begin{cases}h_{\min} & d_{\min}\leqslant d_{\min}^{\text{low}}\\ h_{\max} & d_{\min}\geqslant d_{\min}^{\text{high}}\\ h_{\min}+\dfrac{h_{\max}-h_{\min}}{d_{\min}^{\text{high}}-d_{\min}^{\text{low}}}\left(d_{\min}-d_{\min}^{\text{low}}\right) & \text{其他}\end{cases}\tag{29}$$

式中: $d_{\min}^{\text{low}}=60\,\text{km}$, $d_{\min}^{\text{high}}=100\,\text{km}$, $h_{\min}=6\,\text{km}$, $h_{\max}=10\,\text{km}$。由此可计算距离不同敌人时的期望高度h^*。

定义当前智能体的高度为h_i, 则高度偏差$\Delta h=|h_i-h^*|$。在容差范围: $\Delta h\leqslant\delta_h$, 则给予线性衰减奖励:

$$r_{\text{height}}=\lambda_{\text{height}}(1-\Delta h/\delta_h)\tag{30}$$

否则为 0。其中, 高度容差$\delta_h=2\,\text{km}$, 高度奖励系数$\lambda_{\text{height}}=0.000\,1$。由此可以确定智能体获得的高度正奖励$r_{\text{height}}$。

4) 制导奖励。为提升智能体对己方导弹的态势感知能力与持续制导能力, 本文设计了一种导弹制导奖励机制。

具体而言, 设智能体的位置向量$\boldsymbol{P}_i=(x_i,y_i,z_i)$, 我方已发射导弹的位置向量$\boldsymbol{P}_m=(x_m,y_m,z_m)$, 计算每枚导弹相对我方智能体的方位角$\lambda_{xoy}$和俯仰角$\theta_{xoz}$。当满足以下条件时: $\lambda_{xoy}\in\left[-60°,60°\right]$, $\theta_{xoz}\in\left[-60°,60°\right]$, 导弹位于该前向扇区内时, 认为其处于可制导状态, 智能体将得到探测正奖励$r_{\text{ego}}=0.001\,667$。

5) 置尾逃逸奖励。为提高智能体在空战中的生存能力, 引入尾部逃逸奖励机制。当雷达探测到敌方导弹接近时, 通过计算得到敌方导弹相对飞机的方位角λ_{xoy}, 若导弹位于飞机后方, 即$|\lambda_{xoy}|>90°$,

则认为飞机处于置尾逃逸状态，智能体将得到置尾逃逸正奖励$r_{\text{tail_escape}}=0.005$。

3.3.2 事件奖励

事件奖励通过给超视距空战中涉及的标志性事件单独设计奖励，从而引导无人机在攻击目标的同时，能够躲避导弹，生存下来。主要事件和对应的奖励如表3所示。

表3 事件奖励

Table 3 Event rewards

事件	奖励值
我方飞机发射一枚导弹r_{shoot}	+2
击落一架敌机r_{kill}	+10
我方一架飞机被击落r_{out}	−10
导弹告警消失且我方飞机存活$r_{\text{threat_gone}}$	+0.25

事件奖励和状态奖励构成了总体的奖励函数，每个时间步计算的奖励函数如下：

$$r=r_{\text{detect}}+r_{\text{missile_threat}}+r_{\text{height}}+r_{\text{ego}}+r_{\text{tail_escape}}+r_{\text{shoot}}+r_{\text{kill}}+r_{\text{out}}+r_{\text{threat_gone}} \tag{31}$$

本文奖励函数融合了目标探测、导弹规避、能量管理和持续制导等超视距空战领域先验知识，旨在缓解训练初期奖励稀疏的问题，加快智能体策略的形成。但奖励设计仍具有一定的人工经验依赖，未来将进一步研究更加稀疏和自主的奖励学习方法，以提升算法的探索能力及跨场景泛化性能。

4 仿真结果

为验证本文提出的多机协同超视距空战自博弈决策框架的有效性，在4v4对抗场景下开展仿真实验。首先，通过对比PFSP机制与标准FSP机制，验证优先采样策略在提升博弈稳定性方面的优越性；其次，通过对自适应胜率损失加权机制进行消融实验，量化分析该机制对算法收敛效率的贡献。此外，还引入基于规则的专家对手，旨在评估纯自博弈训练智能体的泛化对抗能力，并对仿真对局中自主涌现的典型战术机动进行深入剖析。

4.1 仿真环境设置

本文在仿真实验环境中设置红蓝双方，我方为蓝机，敌方为红机。自博弈红蓝双方神经网络初始设置均采用随机初始化参数的空白策略，设置多进程并行实验进程组数为4组，实验中每次对战最大时长为1 200 s，智能体控制间隔为0.4 s。对战双方初始距离为140 km，飞机初始高度为9 000 m，初始速度为300 m/s，蓝方飞机初始航向为90°，红方飞机初始航向为270°。在单场对局中，任意一方4架飞机全部阵亡，或者达到最大仿真对战时长，本局对战结束。仿真实验中，敌我双方位置轮流调换，消除地理位置偏差。

此外，多机协同超视距空战自博弈算法的相关参数如表4所示。

表4 算法参数

Table 4 Algorithm parameters

参数分类	参数	取值
神经网络参数	隐藏层神经元数量	256
	隐藏层数量	2
	Actor网络学习率	0.000 5
	Critic 网络学习率	0.000 5
	PPO裁剪参数	0.2
	折扣因子	0.995
	批数量	1
自博弈参数	策略池容量	5
	策略入库胜率阈值	0.6
	策略入库间隔阈值	5
	Elo初始评分	1 200
改进机制参数	策略损失加权机制的最大权重	1.5

4.2 对比实验与消融实验

设计仿真4v4超视距空战。分别进行PFSP自博弈算法与FSP自博弈算法对比实验，以及增加胜局加权机制的PFSP算法与仅PFSP算法的消融实验。以训练时自博弈对局胜利作为评价标准，分析其收敛情况。在对称自博弈演化框架下，博弈双方策略的均衡状态（即逼近平衡状态）表现为胜率曲线在50%基准线附近的收敛与微幅振荡，如图3所示。

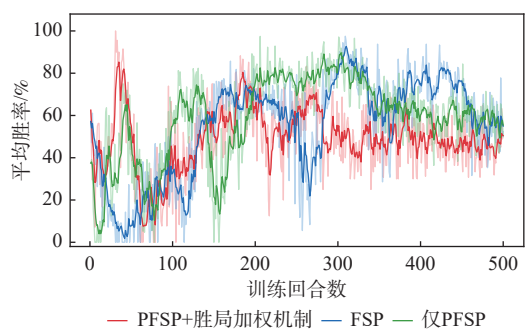


图3 算法对比实验胜率

Fig. 3 Win rates in algorithm comparative experiments

4.2.1 与基准FSP算法的对比实验

首先，从收敛效率纬度，本文提出的“PFSP+胜局加权机制”表现出显著的加速收敛优势。在经历训练初期（前150个训练回合）因策略快速试错引起的剧烈激战期后，红线在第200个训练回合数附近便率先达到50%胜率，随后经历一段时间的振荡后进入稳定期，在50%胜率上下微幅波动。相比之

下, 基准 FSP 算法由于缺乏对手针对性筛选与有效的梯度反馈, 其策略更新路径呈现明显的随机性, 直到第 450 个回合附近才初步具备收敛到均衡线的迹象。实验结果表明, 本文算法使算法逼近均衡状态所需的训练回合数缩短了约 55%, 显著降低了多智能体博弈训练的计算成本。

其次, 在博弈稳定性纬度, 标准 FSP 算法在整个训练中后期(200～450 回合)表现出严重的“策略循环振荡”现象, 其平均胜率在 40%～90% 之间产生较大幅度的周期性起伏。这表明在多机协同超视距空战的复杂高动态环境中, 传统 FSP 算法易受到策略库中部分过时或劣质策略的干扰, 导致智能体陷入由于对手改变而频繁推倒重建的策略振荡。而本文算法在 250 个训练回合数后, 其曲线不仅拟合 50% 的平衡状态线, 且其振荡幅值较小, 展现出较好的策略稳定性。

4.2.2　与仅使用 PFSP 算法的消融实验

如图 3 中绿线(仅 PFSP 机制)所示, 当移除胜局加权机制后, 智能体虽然在早期能够凭借优先自博弈提供的合理对手, 避免进入策略循环振荡, 但在训练中期(200～300 回合)由于缺乏合适的梯度指引, 模型总体学习速率较低, 直至 350 个回合后才逐步稳定向均衡线靠拢。对比红线(本文算法)可知, 胜局加权机制大幅度加速了收敛速率, 引导智能体更早地逼近平衡状态。

4.3　对战专家对手实验

为了验证这组收敛后的本文算法模型在面对未知的外部对手时的真实作战能力, 进一步引入基于规则的专家对手进行泛化对抗评估。该专家对手具备“进攻-防御”双重战术态势切换能力, 其核心设计理念为“先发制敌、射后俯冲制导后防御”。具体而言, 专家对手在远距阶段保持雷达常开探测和加速接敌, 当敌机进入杀伤包线后, 优先爬升抢占有利态势, 并实施先手打击; 导弹发射后, 根据导弹制导需求朝向敌方目标飞行, 并通过俯冲、偏置制导提高自身生存能力。当接收到导弹告警时, 策略优先切换至防御机动, 通过转弯背向导弹方向降低高度躲避导弹。

选择训练收敛时的智能体分别对战专家对手。以 3 种算法达到 500 回合时的智能体策略, 分别与专家对手对战 300 场, 得到胜率效果如表 5 所示。

分析可知, 本文提出的结合胜局加权机制的 PFSP 算法在各项空战指标中相对于 FSP 算法和仅 PFSP 算法均有较大领先优势。在与专家对手的 300 场对抗中, 本文算法最终取得了 70.3% 的平均胜率, 大幅度领先 FSP 算法取得的 42.7% 平均胜率和仅 PFSP 算法的 53.3% 平均胜率。同时, 本文算法的平均战损比达到了较为优秀的 1.8:1, 场均我方存活数为 1.9 架, 是 FSP 算法的 2.4 倍。战损比和我方场均存活数表明, 结合胜局加权机制的 PFSP 算法训练出了具有较高决策能力的智能体, 能够在优先保证自身生存的前提下, 通过高超的战术规划击败对手。

表 5　对战专家对手胜率

Table 5　Win rates against expert opponents

决策算法	对战专家胜率/%	战损比(敌: 我)	我方场均存活数/架	平均对局时长/s
PFSP+胜局加权机制	70.3	1.8:1	1.9	455.1
FSP	42.7	0.9:1	0.8	654.2
仅PFSP	53.3	1.2:1	1.3	545.5

4.4　战术分析

以强化学习自博弈对战专家对手为例, 展示多机协同超视距空战决策算法涌现的空战战术。蓝方采用自博弈收敛时的智能体策略, 红方为专家对手, 对战双方初始距离为 140 km, 飞机初始高度为 9 000 m, 初始速度为 300 m/s, 蓝方飞机初始航向为 90°, 红方飞机初始航向为 270°。在单场对局中, 任意一方 4 架飞机全部阵亡, 或者达到最大仿真对战时长, 本局对战结束。

空战场景中红蓝双方对抗三维轨迹如图 4 和图 5 所示。

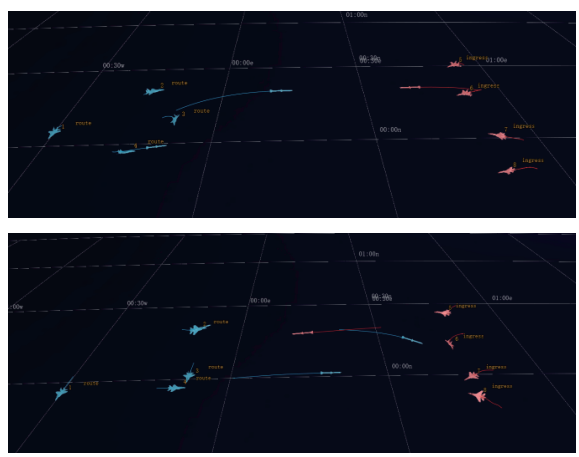


图 4　对战初期仿真表现

Fig. 4　Simulation performance in the initial phase of combat

在对战初期, 蓝 3 号机先视先射发射导弹, 蓝 2 号机和蓝 4 号机采用钳形攻势, 背对飞行, 拉开队内间距。蓝 3 号机完成导弹发射后, 俯冲侧向对敌, 采取防御机动, 此时重新转向敌方的蓝 2 号机和蓝 4 号机对蓝 3 号机发射的导弹进行接力制导。

进一步, 蓝 2 号机和蓝 4 号机同时爬升, 发射高抛弹使用火力压制敌方飞机。第 1 波态势弹有效压制敌方飞机的能量, 敌方红 1 号机和红 3 号机



图5 对战中后期仿真表现

Fig. 5 Simulation performance in mid-to-late phases of combat

躲避导弹后，在低空重新进攻，蓝2号机和蓝4号机继续发射导弹进行第2波进攻。第2波进攻距离近，且蓝方占据高度优势，成功命中敌机。进一步对红2号机进行持续压制，通过发射导弹逼迫红方飞机侧向面对己方，从而使得红方飞机导弹失去制导，无法命中我方飞机，最终全歼敌方飞机。

5 结 论

本文针对多机协同超视距空战场景下自博弈训练较慢、策略循环振荡问题，提出了结合胜局加权机制的PFSP算法，通过理论建模和仿真验证，得出如下结论：

1）结合胜局加权算法的PFSP在训练收敛速率、收敛稳定性方面均显著优于基准FSP算法，收敛效率提升约55%。

2）使用本文算法训练的智能体对战专家对手时，能够取得70.3%的胜率，显著优于基准算法42.7%的胜率，展现出强大的空战决策能力。

3）仿真战术分析表明本文所提出的自博弈算法具有优异的表现效果，涌现出了钳形攻势、侧翼包抄等符合人类认知的专家级作战战术。

需要指出的是，本文采用离散化的期望高度、期望速度和期望航向动作，能够有效压缩动作搜索空间，降低多机空战决策中的探索难度，但也可能限制智能体对飞行控制量的精细调节能力，从而损失部分精细机动潜力。未来将进一步研究连续动作空间建模方法，以提升智能体在复杂空战场景下的精细机动与控制能力。

参考文献（References）

[1] 樊会涛, 闫俊. 空战体系的演变及发展趋势[J]. 航空学报, 2022, 43(10): 527397.
Fan H T, Yan J. Evolution and development trend of air combat system[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(10): 527397(in Chinese).

[2] 孙智孝, 杨晟琦, 朴海音, 等. 未来智能空战发展综述[J]. 航空学报, 2021, 42(8): 525799.
Sun Z X, Yang S Q, Piao H Y, et al. A survey of air combat artificial intelligence[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 525799(in Chinese).

[3] 罗越群, 丁达理, 谭目来, 等. 无人作战飞机自主机动决策方法综述[J]. 航空学报, 2025, 46(7): 130884.
Luo Y Q, Ding D L, Tan M L, et al. A review of autonomous maneuver decision methods for unmanned combat aerial vehicle[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(7): 130884(in Chinese).

[4] 李波, 黄晶益, 万开方, 等. 基于深度强化学习的无人机系统应用研究综述[J]. 战术导弹技术, 2023(1): 58-68.
Li B, Huang J Y, Wan K F, et al. A review of research on the application of UAV system based on deep reinforcement learning[J]. Tactical Missile Technology, 2023(1): 58-68(in Chinese).

[5] 方思雨, 张振华, 牛经龙, 等. 多无人机空战机动决策中的深度强化学习: 实践中的关键技术综述与未来展望[J]. 航空兵器, 2025(2): 19-33.
Fang S Y, Zhang Z H, Niu J L, et al. Deep reinforcement learning in multi-UAV air combat maneuver decision-making: a review of key techniques in practice and future prospects[J]. Aero Weaponry, 2025(2): 19-33(in Chinese).

[6] Li W T, Fang F, Peng D L, et al. An intelligent maneuver decision-making approach for air combat based on deep reinforcement learning and transformer networks[J]. Entropy, 2024, 26(12): 1036.

[7] Yang J, Wang L P, Han J L, et al. An air combat maneuver decision-making approach using coupled reward in deep reinforcement learning[J]. Complex & Intelligent Systems, 2025, 11(8): 364.

[8] Wang H M, Zhou Z F, Jiang J Z, et al. Autonomous air combat maneuver decision-making based on PPO-BWDA[J]. IEEE Access, 2024, 12: 119116-119132.

[9] Mei J R, Li G, Huang H S. Deep reinforcement-learning-based air-combat-maneuver generation framework[J]. Mathematics, 2024, 12(19): 3020.

[10] Zhang X, Dong W H, Zhang P, et al. Research on the reward design method for deep reinforcement learning in WVR air combat[J]. IEEE Access, 2024, 12: 182693-182707.

[11] Zhu J Y, Kuang M C, Zhou W Q, et al. Mastering air combat game with deep reinforcement learning[J]. Defence Technology, 2024, 34: 295-312.

[12] Gao X Z, Zhang Y, Wang B L, et al. The optimal strategies of maneuver decision in air combat of UCAV based on the improved TD3

algorithm[J]. Drones, 2024, 8(9): 501.

[13] Shi G Q, Cao Y, Wang D H, et al. Towards intelligent unmanned adversarial games: a reinforcement learning framework with the PHP-ROW method[J]. Drones, 2025, 9(5): 331.

[14] 单圣哲, 张伟伟. 基于自博弈深度强化学习的空战智能决策方法[J]. 航空学报, 2024, 45(4): 328723.
Shan S Z, Zhang W W. Air combat intelligent decision-making method based on self-play and deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(4): 328723(in Chinese).

[15] Li B, Huang J Y, Bai S X, et al. Autonomous air combat decision-making of UAV based on parallel self-play reinforcement learning[J]. CAAI Transactions on Intelligence Technology, 2023, 8(1): 64-81.

[16] Kong W R, Zhou D Y, Zhou Y, et al. Hierarchical reinforcement learning from competitive self-play for dual-aircraft formation air combat[J]. Journal of Computational Design and Engineering, 2023, 10(2): 830-859.

[17] Tasbas A S, Sahin S O, Üre N K. Reinforcement learning based self-play and state stacking techniques for noisy air combat environment: AIAA 2023-1077[R]. Reston: AIAA, 2023.

[18] Yang M C, Shan S Z, Zhang W W. Decision-making and confrontation in close-range air combat based on reinforcement learning[J]. Chinese Journal of Aeronautics, 2025, 38(9): 103526.

A multi-aircraft cooperative air combat decision-making algorithm based on an adaptive weighted priority virtual self-play theory approach

ZHOU Yaoming[1,*], HAN Tao[1], JIANG Yufei[1], YANG Zheng[2], TIAN Yifeng[2]

(1. School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China;
2. The 29th Research Institute, China Electronics Technology Group Corporation, Chengdu 610036, China)

Abstract: This study suggests an approach that combines prioritized fictitious self-play (PFSP) with an adaptive winning-game weighting mechanism to address the delayed convergence and policy oscillation in self-play training for multi-aircraft cooperative beyond-visual-range (BVR) air combat. Firstly, a 4v4 BVR air combat simulation environment is established under a multi-agent reinforcement learning framework. Second, in order to break the policy cycle, the Elo method is developed to conduct non-uniform sampling of past opponents and dynamically assess policy capabilities. Meanwhile, a win-rate-based loss function weighting mechanism is designed to guide gradient updates when early high-value samples are scarce. According to experiments, the suggested approach greatly lessens later-stage oscillation and converges to the Nash equilibrium by the 200th episode, increasing convergence efficiency by 55% over conventional FSP. The weighting mechanism's accelerating impact is further supported by ablation investigations. Lastly, agents educated using PFSP have strong air combat decision-making ability, winning 70.3% of the time against skilled opponents.

Keywords: air combat decision-making; self-play; reinforcement learning; winning-game weighting; multi-aircraft cooperation

Received: 2026-06-09; **Accepted**: 2026-07-17; **Published Online**: 2026-08-06 14:42
URL: link.cnki.net/urlid/11.2625.V.20260806.1112.002

* **Corresponding author**. E-mail: zhouyaoming@buaa.edu.cn