航 空 学 报
Acta Aeronautica et Astronautica Sinica

Feb. 25 2024 Vol. 45 No. 4
ISSN 1000-6893 CN 11-1929/V

http://hkxb.buaa.edu.cn hkxb@buaa.edu.cn

引用格式：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). doi:10.7527/S1000-6893.2023.28723

基于自博弈深度强化学习的空战智能决策方法

单圣哲[1,2]，张伟伟[1,*]

1. 西北工业大学 航空学院，西安 710072
2. 中国人民解放军93995部队，西安 710306

摘 要：空战是战争走向立体的重要环节，智能空战已经成为国内外军事领域的研究热点和重点，深度强化学习是实现空战智能化的重要技术途径。针对单智能体训练方法难以构建高水平空战对手问题，提出基于自博弈的空战智能体训练方法，搭建研究平台，根据飞行员领域知识合理设计观测、动作与奖励，通过“左右互搏”方式训练空战智能体至收敛，并通过仿真试验验证空战决策模型的有效性。研究结果表明通过自博弈训练，空战智能体战术水平逐步提升，最终对单智能体训练的决策模型构成70%以上胜率，并涌现类似人类“单/双环”战术的空战策略。

关键词：空战；人工智能；深度强化学习；自博弈；智能体

中图分类号：V249 **文献标识码**：A **文章编号**：1000-6893(2024)04-328723-13

随着武器装备的发展，空战的交战样式和作战准则正在发生深刻演变。智能化技术赋能空战将产生革命性效果，杨伟[1]曾论述：空战未来必将进入“智能为王”的时代。2020年，在DARPA组织的人机近距空战Alpha DogFight比赛中，美国苍鹭公司开发的空战决策模型以5∶0的成绩，压倒性地战胜了美空军现役飞行教官[2]。这一事件充分体现了空战智能化将是未来国内外军事领域研究的热点和重点，而空战智能化的核心是空战决策的自主化。

针对空战自主决策问题，早自20世纪60年代，国内外学者就已进行了一系列的理论研究和实践探索，并取得了相应的成果，从理论层次划分，传统方法可分为数学求解方法、机器搜索和数据驱动3大类[3]。但是传统方法只能处理有限维度问题，要在空战场景建模的完整度和算法优化的复杂度之间做权衡，这就导致其决策模型只能应用于简化后的特定空战场景。董一群等[3]总结了近些年国内外研究成果，并指出较为通用和完备的空战决策方法尚未见报道。

自2016年起，深度强化学习(Deep Reinforcement Learning，DRL)方法在智能决策领域取得了一系列巨大成功[4-9]。AlphaGo[4-6]系列算法在围棋领域战胜人类冠军，引发世界关注，AlphaFold v2在蛋白质结构领域取得碾压业界的成果[7]，AlphaTensor在矩阵快速算法领域取得突破成绩[8]。著名学者Silver指出：基于DRL方法通用人工智能(Artificial General Intelligence，AGI)的基础已经具备[9]。

AlphaGo系列算法的成功充分印证了在DRL过程中运用自博弈(Self-Play)训练策略，可以在高维度博弈问题中超越人类顶级决策水平。

收稿日期：2023-03-21；退修日期：2023-06-12；录用日期：2023-08-29；网络出版时间：2023-09-01 17:04
网络出版地址：https://hkxb.buaa.edu.cn/CN/Y2024/V45/I4/328723
基金项目：国防科技重点实验室基金(6142219190302)
*通信作者. E-mail：aeroelastic@nwpu.edu.cn

以此为起点，一系列Self-Play与DRL相结合的方法被相继提出，并广泛应用在诸多复杂博弈领域。其中较为著名的有AlphaStar[10]在StarCraft II游戏中战胜了99.8%人类选手，取得了大师级（Grandmaster）技术段位。StarCraft II游戏的策略搜索空间远超围棋游戏，每一步都有超过10^{29}种可选方案，且具有非完全信息长视决策的特点，而这曾被认为是DRL算法设计领域长期难以克服的难题[11]。另一项较为著名的成果为2019年OpenAI Five[12]在Dota2游戏中以2:0的成绩战胜了世界冠军OG战队。此外，Baker等[13]在hide-and-seek游戏中取得比肩人类的表现水平，Oh等[14]在实时对战游戏Blade&Soul中取得专业级（Pro-Leve）的表现。

DRL方法在高维度决策任务中的优异表现，为解决传统空战决策方法的维度受限问题带来了契机，一系列研究随即展开。2019年，Kurniawan等[15]采用行动者-批评者（Actor-Critic，AC）架构进行空战决策，融合了基于价值和基于策略方法的优点。2019年，Yang等[16-17]采用深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）方法，解决了传统空战决策方法中机动输出的“维度爆炸”问题，实现在连续空间下的动作输出。Piao等[18]采用强化学习方法自演化空战动作，实现战术战法创新。在上述研究思路的基础上，本研究基于AC架构提出了近距空战连续决策的统一框架，主流连续动作空间强化学习算法均可用于近距空战决策[19]。

在上述研究中，空战决策模型的对手往往是“非智能”简易控制程序，训练过程中，智能体战胜执行特定机动的敌机便会收敛，这将极大制约DRL方法高维映射能力的发挥。而基于Self-Play训练策略，可以让空战智能体“左右互搏”，互相扮演自己的高水平对手，不断提高智能体的决策水平，从而解决该问题。目前，Self-Play方法在空战智能决策领域应用尚未见报道。

本研究提出了基于DRL与Self-Play相结合构建空战智能体的基本思路，主要创新点包括：

1）针对空战单智能体训练方法难以构建高水平对手问题，提出了DRL与Self-Play相结合的空战智能训练方法，搭建空战自博弈框架，开发可视化研究平台，为开展相应研究创造基础条件。

2）立足飞行实际，提取空战智能体动作空间和观测空间，从智能体/环境接口角度匹配空战训练任务和连续动作空间强化学习算法，并通过观测特征提取简化训练任务难度。

3）基于空战领域知识，分层次设计多元空战奖励，并合理分配各奖励权重，引导训练收敛并“涌现”空战智能。

1 深度强化学习

强化学习[20]（Reinforcement Learning，RL）是侧重智能体与环境交互的机器学习方法，强调利用奖励刺激产生倾向型行为。深度学习（Deep learning，DL）作为机器学习领域最重要的范式，其强大的高维度映射能力在数据科学、计算机视觉、自然语言处理、生物医药、复杂控制、群体智能等应用领域取得重大突破[21]。2013年Mnih等[22]首次将DL和RL结合，使RL算法具有了高维映射能力，开创了DRL的新范式。一系列新算法不断涌现并取得重大突破，使DRL成为机器学习领域最具潜力的发展方向。

1.1 近端策略优化算法

DRL方法通常分为基于价值的方法、基于策略的方法和基于AC的方法。考虑在空战自博弈过程中，对手的策略不断变化，采用Off-Policy的强化学习方法进行Self-Play训练，可会因为经验池存在过去的决策轨迹，导致整个训练结果的不稳定、不收敛[23]。故本文采用On-Policy的近端策略优化算法[24]（Proximal Policy Optimization，PPO）构建空战决策智能体。

PPO算法具有高鲁棒性、强稳定性等特点，在DRL算法中应用最广泛。PPO使用重要性采样方法，利用旧策略的决策轨迹训练最新策略，同时为防止新旧策略差异的过大，PPO设置优化目标为

$$E_{\pi(\boldsymbol{\theta}')}\Big[\min\Big[r_t(\boldsymbol{\theta})\hat{A}_{\pi(\boldsymbol{\theta}')}(S_t,A_t),\ \mathrm{clip}\big(r_t(\boldsymbol{\theta}),1-\varepsilon,1+\varepsilon\big)\hat{A}_{\pi(\boldsymbol{\theta}')}(S_t,A_t)\Big]\Big] \quad (1)$$

该优化目标从演化动力上限制新旧策略的差异，同时PPO每回合结束就清空经验池，保证策略训练数据都是由本回合训练产生的，从而限

制经验回放导致的数据误差，提高训练的稳定性和收敛性。

PPO算法兼具Off-Policy数据利用效率较高的优点和On-Policy的稳定性，适用于空战智能体的Self-Play训练。

1.2 多智能体与自博弈

在强化学习系统中，若存在一个以上智能体与环境交互，则该系统就转变为多智能体强化学习（Multi-Agent Reinforcement Learning，MARL）系统。一对一空战自博弈训练中，敌我智能体均与环境发生交互，故敌我智能体与空战环境构成MARL系统。

在MARL系统中，每个智能体追求自身累计奖励最大的行为满足了VNM-rational[25]意义下理性参与人要求，博弈论方法将适用于结果的预测和博弈规则的制定。MARL过程通常使用随机博弈[26]（Stochastic Games）和马尔科夫博弈[27]（Markov Games）等数学模型来描述。求解纳什均衡（Nash Equilibrium，NE）通常被认为是MARL任务的概念解（Solution Concept）。

Self-Play方法最早起源于博弈论中的虚拟博弈[28]（Fictitious Play），用以求解双人零和博弈问题的NE。Self-Play应用于DRL领域的最新成果主要包括MuZero[29]、DouZero[30]、δ-Uniform Self-Play[31]、Population Based Training[32]和League Self-Play[10]等方法。

从博弈论角度分析，可将双机近距对抗的强化学习任务认为是完全信息双人零和博弈的MARL任务，Self-Play方法将适用于求解该类问题的NE，本文采用基于群落的自博弈[33]方法，博弈框架如图1所示。

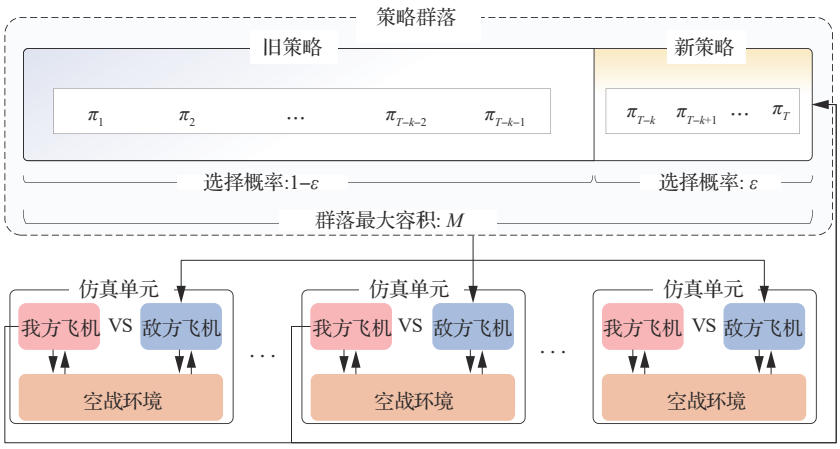


图1 自博弈框架

Fig. 1 Framework of self-Play

该框架采用分布式架构，同时运行多个仿真单元进行训练。空战对手的策略集合为最大人口数量为M的群落，训练中每间隔u个时间步长（steps），便将我方飞机智能体的切片加入群落保存，若此时群落已满，则剔除最早的切片。当我方飞机训练v个steps后，将重新从群落中随机选择智能体切片加载为空战对手，选择对手时以ε的概率选择最新切片作为对手，以$1-\varepsilon$概率均匀选择旧切片作为对手，ε是平衡训练中的探索（Exploration）与利用（Exploitation）关系。为兼顾敌我训练的对称性，每间隔w个steps敌我训练切换，我方加载群落中的智能体切片，敌方飞机进行训练并保存切片。在群落自博弈训练中共有M、u、v、ε、w 5个超参数。

2 空战研究平台

空战对抗是高烈度的三维空间动态，仅依靠数值仿真难以想象空中态势，为方便借鉴飞行员领域知识分析立体态势、评估机动的合理性并做出相应改进，设计如图2所示的技术框架，搭建空战三维可视化研究平台。

该研究平台由空战仿真环境和人工智能训

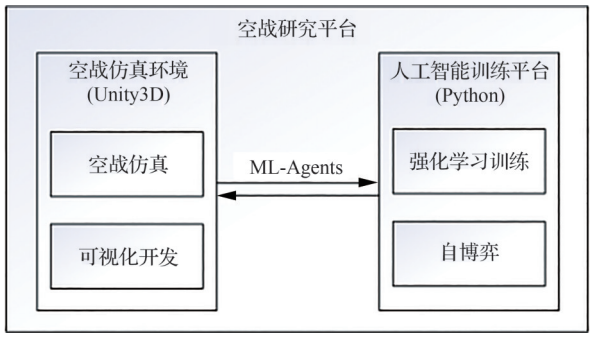


图 2 空战研究平台

Fig. 2 Research platform of air combat

练平台以及两者之间的通信管道构成。空战仿真环境基于Unity3D的物理引擎进行开发,空战决策智能体通过与仿真环境交互不断优化自身策略,同时利用Unity3D场景渲染功能开发多种空战观察视角,以便研究人员从直观角度把握空战动态。人工智能训练平台基于Python语言开发,实现空战决策智能体的强化学习训练及自博弈训练等功能。利用ML-Agents[33]工具包构建Unity3D平台与Python之间的通信管道,将空战仿真环境和人工智能训练平台整合为空战可视化研究平台。

2.1 空战场景设置

尽管超视距空战是未来战争的主要模式之一,但近距空战能力仍为空中作战人员的必备素质,且隐身战机、合理的战术机动等因素都会使超视距空战转入近距空战。无论机载雷达武器设备如何先进,未来空战均不能排除由超视距转入近距作战的可能性,近距空战将仍是未来空战体系中不可或缺的环节。而且近距空战具有高动态、高烈度、高过载的特点,先进战机必须考虑陷入近距作战的风险。对近距空战智能决策的研究仍具有十分重大的意义。

本文将空战的起始态势设置为大进入角态势。若交战双方均未能在超视距达成战果,随即转入近距作战,该态势将成为最有可能的起始态势[34],其成因如图3所示。双方大进入角起始态势下,敌我飞机具有均等的角度和占位优势,由于智能体的策略探索具有随机性,以均势作为起始态势,利于演化出较为复杂的态势场景,实现最终策略的高维性。

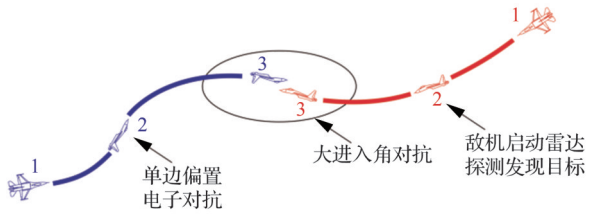


图 3 大进入角起始态势成因

Fig. 3 Genesis of high aspect angle initial situation

设定空战场景以大进入角为起始态势,智能体可以使用雷达、近距格斗导弹及航炮等机载武器设备,同时基于3个维度[19]的连续操纵量形成机动动作,不断探索空战中的机动及设备使用策略。

2.2 飞行运动学模型

沿用文献[19]的思路,空战智能体决策的动作空间中,飞机操纵量为飞机法向过载n_z、推力T与速度滚转角ϕ_a 3个维度的连续量[19]。基于此操纵量可以利用基于四元数的三自由度简化假设,实时求解飞机动态。

飞机的运动学方程为

$$\boldsymbol{V}_{\mathrm{g}}=\begin{bmatrix}\dot{x}_{\mathrm{g}}\\ \dot{y}_{\mathrm{g}}\\ \dot{z}_{\mathrm{g}}\end{bmatrix}=\boldsymbol{L}_{\mathrm{ag}}^{\mathrm{T}}\begin{bmatrix}V\\ 0\\ 0\end{bmatrix} \tag{2}$$

式中:$\boldsymbol{V}_{\mathrm{g}}$为飞机速度向量在地面坐标系下的投影,其3个分量为$\dot{x}_{\mathrm{g}}$、$\dot{y}_{\mathrm{g}}$、$\dot{z}_{\mathrm{g}}$;$V$为飞机空速;$\boldsymbol{L}_{\mathrm{ag}}$为地面坐标系向机体坐标系的投影矩阵。基于四元数的矩阵构建方法参考文献[19]。

飞机的动力学方程为

$$\begin{cases}\dot{e}_0=-0.5\left(\omega_x e_1+\omega_y e_2+\omega_z e_3\right)\\ \dot{e}_1=0.5\left(\omega_x e_0-\omega_y e_3+\omega_z e_2\right)\\ \dot{e}_2=0.5\left(\omega_x e_3+\omega_y e_0-\omega_z e_1\right)\\ \dot{e}_3=0.5\left(-\omega_x e_2+\omega_y e_1+\omega_z e_0\right)\\ \dot{V}=(T-D)/m+2g\left(e_1 e_3-e_0 e_2\right)\end{cases} \tag{3}$$

式中:ω_x、ω_y、ω_z为机体系相对地面系的转动角速度在其机体系下的投影,其求解方法可以参考文献[19];m为飞机质量;D为空气阻力。

2.3 雷达模型

现代近距空战中,机载雷达和离轴发射导弹

被广泛应用，与仅依靠航炮攻击的“狗斗”相比，战斗样式发生了较大变化，因此需要对机载雷达、格斗导弹以及机炮均进行建模，以保证空战仿真更加符合实际情况。

在近距空战中，双方均作大机动飞行，大多数区域搜索、手动截获的空空模式不再适用，雷达须使用空中格斗模式（Air Combat Modes，ACM）进行自动截获。为简化仿真，本文仅对雷达ACM进行建模。

ACM是一种自动截获工作模式，通常包含多个子模式，本文中设定类ACM包含3个子模式：平扫模式（Horizontal Scan，HS）、垂扫模式（Vertical Scan，VS）和定轴模式（Bore Scan，BS），各子模式的设定参数如表1所示。

表1 ACM子模式性能参数

Table 1 Performance parameters of ACM sub-mode

子模式	水平范围/(°)	俯仰范围/(°)	扫描时间/s
HS	−10～10	−15～5	3
VS	−5～5	−10～30	3
BS	−1～1	−1.5～−0.5	0.01

其中，若雷达处于VS或者HS，可以操纵雷达天线进行左偏置或右偏置，从而改变雷达的水平搜索范围，VS状态的水平偏置量为5°，HS状态的水平偏置量为10°。

设定雷达的对敌机最大截获距离为11 km，雷达天线的最大跟踪角速度为60(°)/s，雷达单目标跟踪（Single Target Track，STT）模式的可跟踪范围为水平−50°～50°、俯仰−50°～50°。

若飞机在ACM子模式下截获敌机，雷达将自动转入STT模式；若敌机脱离跟踪，则雷达自动转回ACM模式重新进行截获。

2.4 武器模型

2.4.1 几何态势模型

对空战中双机空战态势进行几何学建模，如图4所示。图中红色飞机为我方飞机，蓝色飞机为敌方飞机，$\boldsymbol{R}$为敌我距离向量，$\boldsymbol{V}_r$为我方的速度向量，$\boldsymbol{V}_b$为敌方的速度向量。从我方飞机视角出发，2个较为重要的几何角度分别为进入角（Aspect Angle，AA）和天线偏置角（Antenna Train Angle，ATA），分别用q和φ表示[19]。

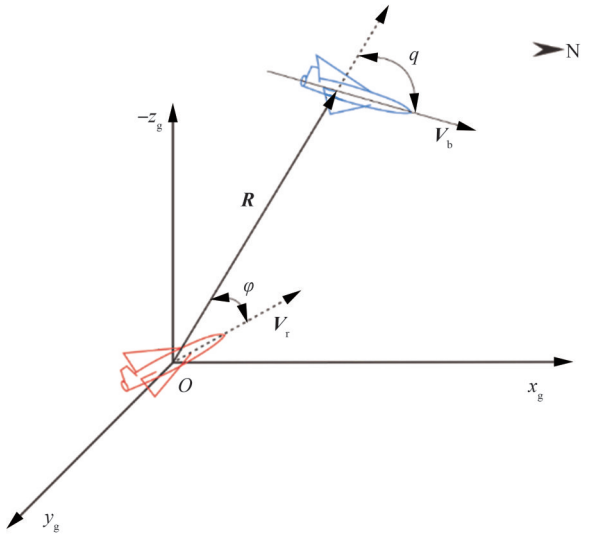


图4 空战几何关系

Fig. 4 Geometry relationship of air combat

2.4.2 格斗导弹模型

以飞行高度H和进入角q为自变量，对格斗导弹红外导引头探测能力建模，得到红外导引头的最大探测距离如图5所示。图5中极坐标角度为飞机方位角q，红外导引头的最大探测距离单位为km。

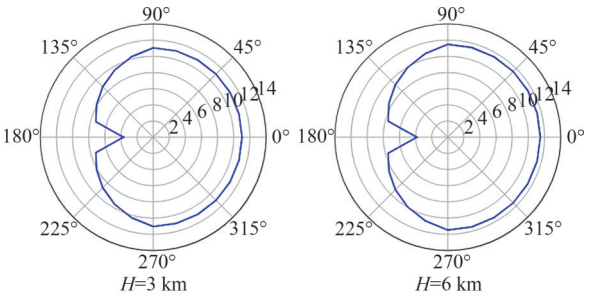


图5 弹红外导引头最大探测距离

Fig. 5 Max range of infrared seeker

参考文献[35]的方法和数据，对导弹的动力学可攻击区进行建模，以敌机速度马赫数为0.9，我机速度马赫数为1.1，交战高度为3 km，敌机机动过载1g为例，不同天线偏置角φ下，导弹动力学的可攻击范围如图6所示。图6中极坐标角度为飞机方位角q，红色包络线为导弹最远可攻击

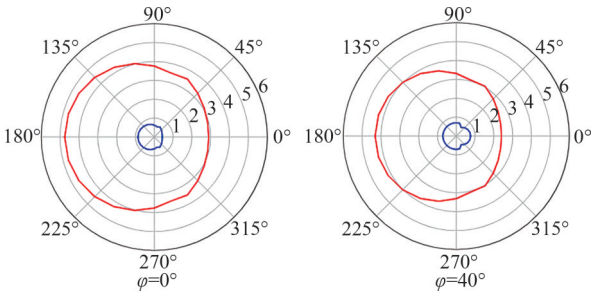


图 6 导弹的动力学可攻击范围

Fig. 6 Attack zone of missile

距离，蓝色包络线为导弹最近可攻击距离，单位为km。

建模雷达与导弹交联关系，区分不同导引头解锁时机，构成不同扫描发射关系，如表2所示。

表 2 导弹的扫描发射方式

Table 2 Scanning and launching method of missile

扫描发射方式	解锁时机	搜索范围
定瞄定发	不解锁	2°锥角
定瞄离发	截获后解锁	2°锥角
定扫离发	截获前解锁	5°锥角
随动离发	雷达截获目标后解锁	水平和垂直±20°方形区域内雷达随动

2.4.3 航空机炮模型

航空机炮在现代空战中的作用有所减弱，但其在特定态势下可对导弹的攻击条件进行必要的扩充，对于现代近距空战仍具有较大的价值。如双机近距快速交会时，瞄准线速率过大会导致无法跟踪截获，此时就可以利用机炮进行快速射击（Snap Shoot）完成击杀[34]。

本文设定航空机炮最大射击距离为900 m，并求解跟踪射击向量$\boldsymbol{L}$，若我方机体纵轴与$\boldsymbol{L}$夹角余弦值大于0.97即认为满足航炮射击条件，若该状态可以持续1 s以上，则认为跟踪一方已实施射击误差的修正，完成航炮击杀。其中跟踪射击向量$\boldsymbol{L}$满足的几何关系如图7所示。图7中$\boldsymbol{V}_{\mathrm{r}}'$和$\boldsymbol{V}_{\mathrm{b}}'$分别为敌我速度向量$\boldsymbol{V}_{\mathrm{r}}$和$\boldsymbol{V}_{\mathrm{b}}$在水平面上的投影，$\boldsymbol{R}'$为敌我距离向量$\boldsymbol{R}$在水平面上的投影，$\Delta H$为炮弹飞行中的下落高度，由自由落体模型进行阻力修正得到。

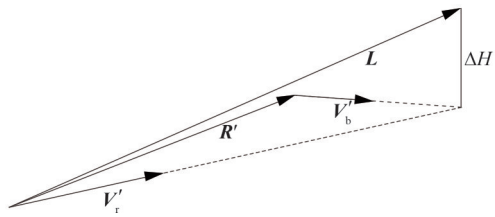


图 7 射击时的几何关系

Fig. 7 Geometry relationship of shooting

3 动作、观测、奖励设计

3.1 动作设计

为使智能体动作更接近现代空战特点，本文在沿用文献[19]中3个飞机机动的操纵量的基础上，增加了雷达子模式选择、天线偏置、导弹导引头解锁3个设备操纵量，共6个动作维度。6个动作空间均为[－1，1]的连续值，设备操纵量对连续空间进行等分来映射实际设备操纵量，为防止决策间隔的动作突变，决策动作和实际操纵量之间做积分处理。

3.2 观测设计

根据飞行训练经验，以大进入角为起始态势，设定交战双方均保持互相目视为基本情况，同时结合雷达等设备获取敌方态势信息，信息获取难度低，故假设双方可完全获取对方态势信息。为了降低智能体训练学习的难度，基于敌我态势信息和我方设备信息提取67维度特征作为观测输入，如表3所示。在特殊观测角度下，若因自身机体遮挡暂时丢失目视，可采用暂时冻结敌方态势方法，经验证仍能保证算法收敛且有效。表3中，数值界限若带“*”则表示空战中的经验界限，而非性能限制的极限范围。如空战中双机高度差通常不大于1 000 m，而极限高度差可达升限。

若考虑电磁干扰等条件，可对观测特征进行删减，经验证，仅提供坐标位置、速度、方位等关键状态量，本研究方法仍然能够有效收敛。

3.3 奖励设计

空战自博弈训练中奖励是稀疏的，只有在决

表 3 观测特征提取

Table 3 Feature extraction of observation

分类	特征名称	数值界限	维度
我机坐标	东西坐标/m	(0,30 000)	1
	南北坐标/m	(0,30 000)	1
	飞行高度/m	(0,12 000)	1
飞行状态	飞行空速/(m·s^{-1})	(0,500)	2
	飞行表速/(m·s^{-1})	(0,500)	2
	马赫数	(0,1.6)	2
	纵向过载	(−1,2)	2
	法向过载	(−4,9)	2
	侧向过载	(−1,1)	2
	转弯角速率/((°)·s^{-1})	(0,50)*	2
	姿态四元数	(−1,1)	8
几何态势	敌我距离向量/m	(−8 000,8 000)*	4
	敌我速度向量/m	(−500,500)*	6
	敌我高度差/m	(−1 000,1 000)*	1
	机炮瞄准系数	(0,1)	2
	水平离轴角/(°)	(−180,180)	2
	离轴角/(°)	(−90,90)	2
	雷达扫描范围/(°)	(−20,20)	8
	导弹扫描范围/(°)	(−20,20)	8
	进入角/(°)	(0,180)	2
	天线偏角/(°)	(0,180)	2
	导弹最大距离/m	(0,8 000)*	2
	导弹最小距离/m	(0,3 000)*	2
	敌我距离标量/m	(0,8 000)*	1
总计			67

出胜负时才有客观的结果奖励发生。智能体在整个探索过程中缺乏引导性奖励,导致训练难以成功。为克服奖励稀疏问题,本文基于飞行员领域知识进行过程奖励设计(Reward Sharping)。

3.3.1 奖励权重设计

过程奖励是基于人为设计的,具有一定主观性,结果奖励虽具有客观性,但只在结束时发生。故权重设计原则为过程奖励权重不得超过折扣后的结果奖励,以保证智能体有足够动力选择结果导向的机动方案,最终智能体策略倾向客观性结果,权重分析如图8所示。

根据经验空战持续时间在4~10 min,若取step=4 s,$\gamma=0.99$,则结果奖励在空战过程中的最小衰减后权重约0.2,过程奖励权重不超过0.2即可满足要求。

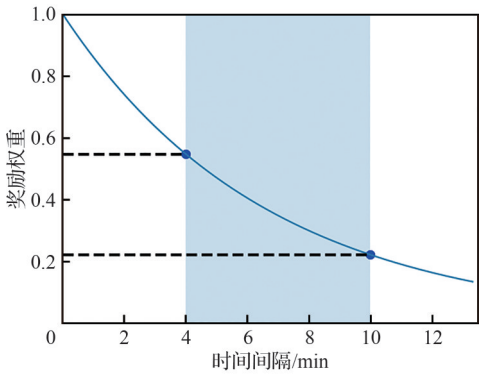


图 8 总体奖励权重分析

Fig. 8 Overview analysis of reward weights

3.3.2 过程奖励设计

过程奖励基于飞行员领域知识设计。根据基本战斗机机动(Basic Fighter Maneuvering, BFM)课程描述,近距空战是平衡角度优势与能量优势的过程,故设计角度优势函数r_a和能量优势函数r_e作为主要过程奖励:

$$\begin{cases} r_a = 1 - \dfrac{\varphi + q}{180} \\ r_e = \dfrac{H_E^r - H_E^b}{H_E^r + H_E^b} \end{cases} \tag{4}$$

式中:φ和q与图4定义相同;H_E^r和H_E^b分别为我机和敌机能量高度,对于飞行速度为V飞行高度为H的飞机,其能量高度为$H_E = H + 0.5V^2/g$。

为引导智能体"勇于"接敌,设计双机距离奖励r_R和双机高度差奖励r_H为

$$\begin{cases} r_R = \dfrac{3\,000 - |\boldsymbol{R}|}{3\,000} \\ r_H = \dfrac{800 - |H_r - H_b|}{800} \end{cases} \tag{5}$$

式中:$\boldsymbol{R}$为敌我距离向量;H_r和H_b为我方飞机和敌方飞机的飞行高度。

3.3.3 边界奖励设计

智能体训练中,可能由于数值超出安全范围发生"撞地""失速""超速"等危险行为导致空战失败。对该类情况若设置条件触发奖励,则会导致

该类奖励稀疏、不连续、无梯度，不利于训练。本文设计“杯型”函数，优化该类奖励，函数构成为

$$p_1=\frac{1}{1+\mathrm{e}^{k(x-b)}}-\frac{1}{1+\mathrm{e}^{k(x-a)}}-1 \tag{6}$$

式中：p_1为惩戒性奖励值；x为触发奖励的数值；a为该数值触发奖励的下边界；b为上边界；k为正实数，代表奖励的突兀程度，该数值越大则奖励触发的越突兀。

以飞行表速为例，其边界奖励如图9所示。该奖励左边界为失速表速，右边界为超速表速，此奖励连续、可微，利于神经网络优化。

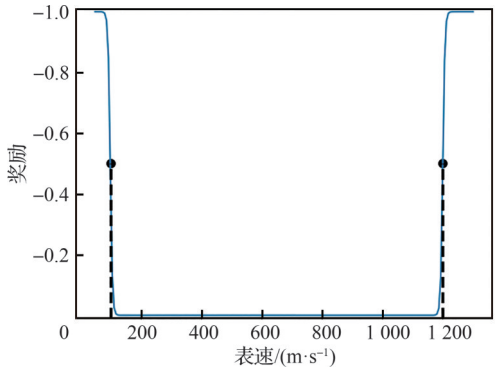


图9 飞行表速边界奖励

Fig. 9 Boundary reward of indicated airspeed

3.3.4 控制区奖励

在近距空战中，有一个重要的区域概念，被称为进攻控制区（Offensive Control Zone，OCZ）。在OCZ内进攻方可以轻松实现对防御方的“咬尾”追踪，在进攻方非失误条件下防御方难以摆脱追踪[34]，区域示意如图10所示。

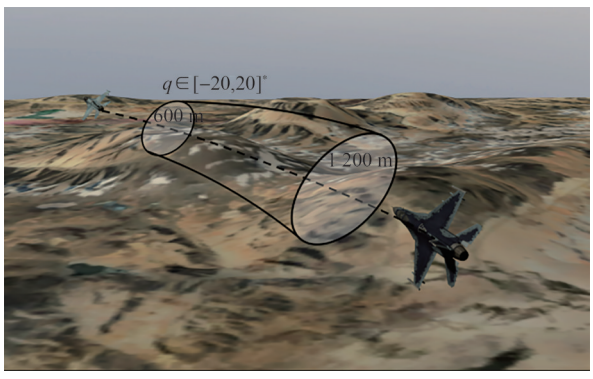


图10 进攻控制区

Fig. 10 Offensive control zone

设计控制区奖励公式为

$$\begin{cases} r_1=-\dfrac{1}{1+\mathrm{e}^{(x-a)/k_1}}-\dfrac{1}{1+\mathrm{e}^{-(x-b)/k_1}}+1 \\ r_2=-\dfrac{1}{1+\mathrm{e}^{(y-c)/k_1}}-\dfrac{1}{1+\mathrm{e}^{-(y-d)/k_1}}+1 \\ r_3=2r_1r_2 \end{cases} \tag{7}$$

式中：r_1、r_2分别为2个维度上的奖励分量；r_3为最终OCZ奖励；x输入为进入角；y输入为敌我距离；(a,b)、(c,d)分别为各维度的奖励触发边界，在该处取值$(0,20)$、$(8,13)$；k_1、k_2为各维度奖励的突兀程度，分别取值1.3、2.3。OCZ奖励曲面如图11所示。

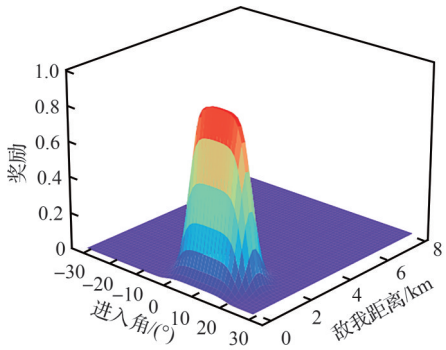


图11 OCZ奖励函数曲面图

Fig. 11 Surface chart of OCZ

3.3.5 最终奖励汇总

将所有奖励汇总，并分配各自权重，可得奖励总体设计，如表4所示。表4中事件奖励不以数值为触发条件，无法使用“杯型”函数优化，其他奖励均已介绍。

从博弈角度来看权重分配，奖励整体符合零和博弈要求。非博弈奖励可使零和博弈转化为常数和博弈，但两者MARL任务等价[36]。虽然诱导双机交战时，利益奖励设置相同，但其权重极小，故总体奖励设置仍可认为符合零和博弈MARL任务要求，Self-Play方法对求解该问题有效。

4 模型训练及验证

4.1 神经网络结构

PPO算法、Actor和Critic网络如图12所示。Actor和Critic网络均采用全连接神经网络

表 4 最终奖励汇总

Table 4 Summary of final reward

奖励类型	博弈分类	奖励名称	权重分配		奖励特性
			我方	敌方	
结果奖励	零和博弈	导弹杀敌	1	−1	稀疏
		机炮杀敌	1	−1	
		敌机撞地	1	−1	
		飞出边界	−1	1	
		相撞/互杀	0	0	
事件奖励	零和博弈	雷达照射	0.05	−0.05	稀疏
		雷达锁定	0.2	−0.2	
		导弹锁敌	0.3	−0.3	
		机炮瞄准	0.5	−0.5	
		达成发射	0.55	−0.55	
过程奖励	零和博弈	角度优势	0.005	−0.005	稠密
		能量优势	0.008	−0.008	
	相同利益	距离奖励	−0.000 1	−0.000 1	
		高度差奖励	−0.000 1	−0.000 1	
边界奖励	零和博弈	控制区	0.1	−0.1	连续
	非博弈	空域坐标	0.8	0.8	
		飞行马赫数	0.8	0.8	
		飞行表速	0.8	0.8	
		双机距离	0.8	0.8	

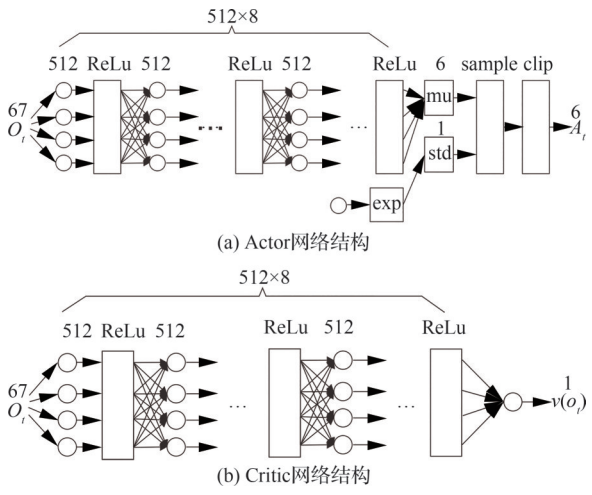


图 12 神经网络结构

Fig. 12 Architecture of neural network

构建，输入层均为67维度的观测O_t，有8个隐藏层，每个隐藏层有512个神经单元，每个隐藏层后是ReLu激活层。Actor输出层为6个维度的连续动作。Critic的输出为1个维度的状态价值评估$v(o)$，用于计算PPO中通用优势估计(Generalized Advantage Estimation，GAE)[37]，GAE采用TD Error(λ)形式。

4.2 自博弈训练

基于群落自博弈方法，设置3组不同超参数组合进行对比试验，如表5所示。

表 5 自博弈超参数

Table 5 Hyper parameters of self-play

名称	超参1	超参2	超参3
群落人口M	100	100	200
博弈概率ε	0.5	0.8	0.5
保存间隔u	2×10^3	2×10^3	4×10^3
重置对手间隔v	100	100	100
训练切换间隔w	2×10^5	2×10^5	4×10^5

表5中各超参数并非完全独立，根据文献[38]的建议，对于一对一训练场景$w=u\times v$，u应与群落人口M和任务难度决定的训练步数成正比。取超参1为baseline，超参2与之控制单一变量ε为对照组，超参3控制M与之对照，其u和w也应与M等比例增大。

自博弈过程中智能体获取奖励不仅取决于自身决策水平，也受到对手水平的影响，以累计奖励评估训练水平不再适用。本文采用文献[39]的方法，利用群落中Elo评级来评估智能体的训练水平及算法的收敛情况。Elo评估机制中，博弈双方分值差与比赛胜率预期唯一对应，其广泛应用于网球联赛、围棋、国际象棋、电子游戏等竞技类比赛中[40]。

Elo胜率预期公式为

$$E_{\mathrm{A}}=\frac{1}{1+10^{(R_{\mathrm{B}}-R_{\mathrm{A}})/400}} \tag{8}$$

式中：A、B为博弈双方；R_{A}、R_{B}分别为A方和B方的Elo分值；E_{A}为A方的胜率预期。

博弈完成后，根据比赛表现，Elo分值更新公式为

$$R'_{\mathrm{A}}=R_{\mathrm{A}}+K(S_{\mathrm{A}}-E_{\mathrm{A}}) \tag{9}$$

式中：R'_{A}为A方更新后的分值；S_{A}为A方博弈表现，博弈获胜其值为1，失败其值为0；K为调整系数，本研究取值为16。

取超参1为baseline，其他为对照，得训练曲

线如图13所示。

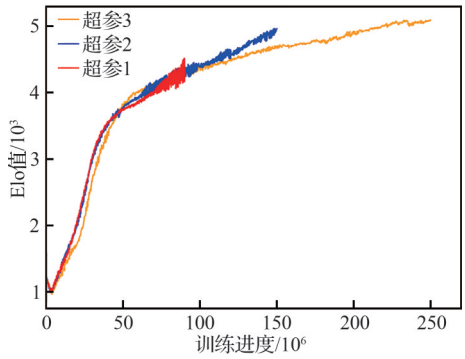


图 13 自博弈训练曲线

Fig. 13 Training curve of self-play

训练中，每个智能体的初始Elo值设定为1 200，经过训练3组超参数对应的Elo值均得到大幅度增长，并最终趋于收敛，3组超参数最终智能体的Elo分值分别为4 397、4 952和5 073。

4.3 结果评估与验证

训练结果评估与验证主要从宏观和微观2个角度开展。

4.3.1 宏观角度评估

基于元博弈思想，统计不同策略切片两两对抗100场空战的胜率，得到博弈矩阵。

分析超参数1，取其训练中50%、75%和100%进度的智能体切片，切片间两两博弈矩阵如图14所示。训练进度越靠后的智能体切片，越能取得更高的胜率。这说明基于群落自博弈训练，智能体可以逐步提升空战决策水平，并最终获得更高水平的空战决策模型。

传统单智能体训练[19]和3组自博弈训练方法，得到的最终智能体两两博弈胜率矩阵如图15所示。由图15可以看出：相对于单智能体训练方法，3组超参数组合下的自博弈训练方法均可以取得70%以上博弈胜率，智能体空战决策水平提升显著；第2组超参数训练模型可以在两两博弈中取得较高胜率，这说明在基于群落的自博弈训练中，提高ϵ比提高M更有可能训练出高水平的空战决策智能体。

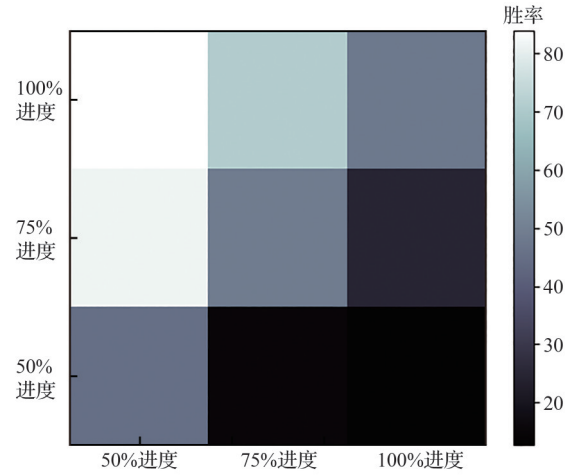


图 14 智能体历史切片博弈矩阵

Fig. 14 Game matrix of agent history slices

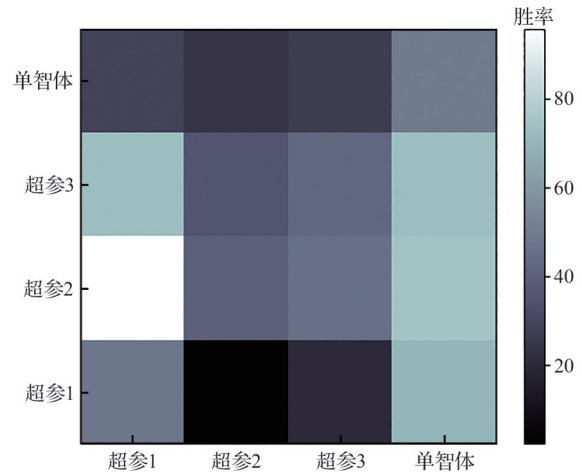


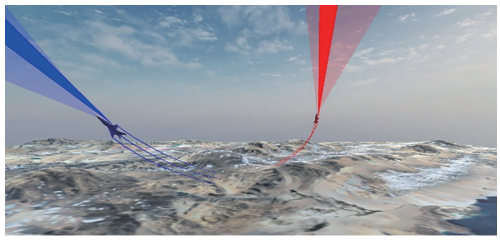
图 15 最终智能体博弈矩阵

Fig. 15 Game matrix of final agents

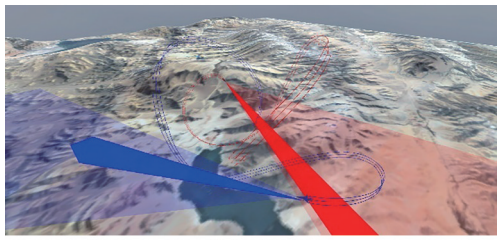
4.3.2 微观角度验证

使用超参2的最终智能体进行对抗演示，并使用可视化平台展示博弈的动态轨迹，利用飞行员经验评估其机动轨迹的合理性。其动态轨迹如图16所示。图16中飞机前方浅色区域为雷达探照范围，深色区域为导弹导引头探照范围，尾后丝带为飞机的尾迹。

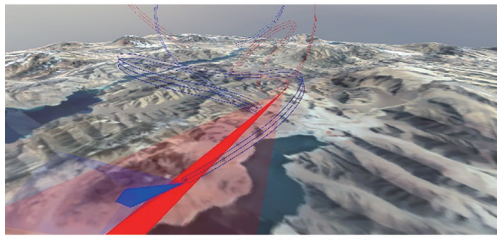
在该场对抗演示中，双方飞机起始为大进入角态势，如图16(a)所示，由于起始速度较大，双方均采用上升转弯机动，形成“双环战”动态。空战时间54 s双方飞机转弯2圈后红方锁定蓝方飞机，但由于蓝方机动角速度较快导致瞄准线速率



(a) 起始态势



(b) 蓝方摆脱锁定



(c) 红方再次锁定并击杀蓝方

图 16 智能体对抗轨迹

Fig. 16 Trails of air combat between agents

较高，未能形成稳定跟踪条件，蓝方成功摆脱锁定，如图16(b)所示。随即红方通过向下机动保持角点速度，形成较大转弯角速度，在114 s成功锁定蓝方，并完成击杀，如图16(c)所示。由于空战持续时间较长，为了清晰展示上述过程的机动轨迹，该空战动态视频已上传至Github[41]。如上，进行多轮对抗轨迹演示，经飞行员经验评估，智能体采取的决策方案均已接近BFM中的经典“单/双环”战术，智能体已基本掌握通用的空战决策能力。在其他的空战动态中，智能体均可以展示出接近人类的战术水平。

5 结论与展望

针对近距空战决策问题，提出基于自博弈和深度强化学习的空战智能体构建方法，搭建了研究平台，并从宏观和微观角度对训练结果进行评估和验证，得出结论如下：

1）由宏观纵向分析可知，采用基于群落的自博弈方法，可以让智能体从零开始不断演化空战战术，逐渐提高空战水平。

2）由宏观横向评估可知，自博弈训练方法相较单智能训练方法，其最终智能体的空战决策水平有较大提升。在自博弈过程中增大与最新智能体博弈的概率，有利于智能体探索更多策略，并提高最终战术水平。

3）从微观角度分析可知，最终训练出的空战决策模型的机动方案已十分接近人类的经典“单/双环战”的战术水平，智能体已基本掌握通用的空战决策能力。

本研究探索基于自博弈的空战训练方法，并发现智能体可掌握系统性战术策略。该方法在快速战术方案生成、虚拟对抗训练、无人空战等领域具有较大应用潜力。

参 考 文 献

[1] 杨伟. 关于未来战斗机发展的若干讨论[J]. 航空学报, 2020, 41(6): 524377.
YANG W. Development of future fighters[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(6): 524377 (in Chinese).

[2] Defense Advanced Research Projects Agency. Alpha dog fight trials go virtual for final event[EB/OL]. (2020-08-07) [2021-03-10]. https://www.darpa.mil/news-events/2020-08-07.

[3] 董一群, 艾剑良. 自主空战技术中的机动决策:进展与展望[J]. 航空学报, 2020, 41(S2): 724264.
DONG Y Q, AI J L. Decision making in autonomous air combat: review and prospects[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(S2): 724264 (in Chinese).

[4] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[5] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge [J]. Nature, 2017, 550(7676): 354-359.

[6] SILVER D, HUBERT T, SCHRITTWIESER J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play [J]. Science, 2018, 362(6419): 1140-1144.

[7] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.

[8] FAWZI A, BALOG M, HUANG A, et al. Discovering faster matrix multiplication algorithms with reinforcement learning[J]. Nature, 2022, 610(7930): 47-53.

[9] SILVER D, SINGH S, PRECUP D, et al. Reward is enough[J]. Artificial Intelligence, 2021, 299: 103535.

[10] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[11] VINYALS O, EWALDS T, BARTUNOV S, et al. StarCraft II: A new challenge for reinforcement learning [DB/OL]. 2017:arXiv preprint:1708.04782.

[12] OpenAI. OpenAI five[EB/OL]. 2018. https://openai.com/research/openai-five.

[13] BAKER B, KANITSCHEIDER I, MARKOV T M, et al. Emergent tool use from multi-agent autocurricula [DB/OL]. arXiv preprint :1909.07528 , 2020.

[14] OH I, RHO S, MOON S, et al. Creating pro-level AI for a real-time fighting game using deep reinforcement learning[J]. IEEE Transactions on Games, 2022, 14(2): 212-220.

[15] KURNIAWAN B, VAMPLEW P, PAPASIMEON M, et al. An empirical study of reward structures for actor-critic reinforcement learning in air combat manoeuvring simulation[C]// Australasian Joint Conference on Artificial Intelligence. Cham: Springer, 2019: 54-65.

[16] YANG Q M, ZHU Y, ZHANG J D, et al. UAV air combat autonomous maneuver decision based on DDPG algorithm[C]// 2019 IEEE 15th International Conference on Control and Automation (ICCA). Piscataway: IEEE Press, 2019: 37-42.

[17] YANG Q M, ZHANG J D, SHI G Q, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning[J]. IEEE Access, 2019, 8: 363-378.

[18] PIAO H Y, SUN Z X, MENG G L, et al. Beyond-visual-range air combat tactics auto-generation by reinforcement learning[C]// 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-8.

[19] 单圣哲, 杨孟超, 张伟伟, 等. 自主空战连续决策方法[J]. 航空工程进展, 2022, 13(5): 47-58.
SHAN S Z, YANG M C, ZHANG W W, et al. Continuous decision-making method for autonomous air combat[J]. Advances in Aeronautical Science and Engineering, 2022, 13(5): 47-58 (in Chinese).

[20] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. 2nd Ed.Cambridge: MIT Press, 2018.

[21] MATHEW A, AMUDHA P, SIVAKUMARI S. Deep learning techniques: an overview[C]//International Conference on Advanced Machine Learning Technologies and Applications. Singapore: Springer, 2021: 599-608.

[22] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[DB/OL]. arXiv preprint: 1312.5602, 2013.

[23] Github. Unity technologies [EB/OL]. (2022-12-14). https://github.com/Unity-Technologies/ml-agents/blob/main/docs/ML-Agents-Overview.md.

[24] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [DB/OL]. arXiv preprint: 1707.06347, 2017.

[25] VON NEUMANN J, MORGENSTERN O. Theory of games and economic behavior: 60th anniversary commemorative edition[M]. Princeton: Princeton University Press, 2007.

[26] SHAPLEY L S. Stochastic games[J]. Proceedings of the National Academy of Sciences of the United States of America, 1953, 39(10): 1095-1100.

[27] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[M]//KAUFMANN M. Machine learning proceedings. Amsterdam: Elsevier, 1994: 157-163.

[28] BROWN G W. Iterative solution of games by fictitious play[J]. Activity Analysis of Production and Allocation, 1951, 13(1): 374-376.

[29] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model [J]. Nature, 2020, 588(7839): 604-609.

[30] ZHA D C, XIE J R, MA W Y, et al. DouZero: Mastering DouDizhu with self-play deep reinforcement learning [DB/OL]. arXiv preprint: 2106.06135, 2021.

[31] BANSAL T, PACHOCKI J, SIDOR S, et al. Emergent complexity via multi-agent competition[DB/OL]. arXiv preprint: 1710.03748, 2017.

[32] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning[J]. Science, 2019, 364(6443): 859-865.

[33] JULIANI A, BERGES V P, VCKAY E, et al. Unity: a general platform for intelligent agentsV[DB/OL]. arXiv preprint: 1809.02627, 2020.

[34] BONANNI P. The art of the kill: A comprehensive guide to modern air combat[M]. Boulder: Spectrum HoloByte, 1993.

[35] 吴文海, 周思羽, 高丽, 等. 基于导弹攻击区的超视距空战态势评估改进[J]. 系统工程与电子技术, 2011, 33(12): 2679-2685.
WU W H, ZHOU S Y, GAO L, et al. Improvements of situation assessment for beyond-visual-range air combat based on missile launching envelope analysis[J]. Systems

Engineering and Electronics, 2011, 33(12): 2679-2685 (in Chinese).
[36] YANG Y D, WANG J. An overview of multi-agent reinforcement learning from game theoretical perspective [DB/OL]. arXiv preprint: 2011.00583v3, 2021.
[37] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation [DB/OL]. arXiv preprint: 1506.02438, 2015.
[38] Unity Technologies. Unity ML-agents toolkit [EB/OL]. (2023-07-10). https://github.com/Unity-Technologies/ml-agents/docs/Training-Configuration-File.md
[39] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning [J]. Science, 2019, 364(6443): 859-865.
[40] Wikipedia. Elo rating system [EB/OL]. 2021. https://en.wikipedia.org.
[41] Github. NWPU-SSZ [EB/OL]. (2023-08-28). https://github.com/NWPU-SSZ/Trajectory-Visualization.

（责任编辑：李丹）

Air combat intelligent decision-making method based on self-play and deep reinforcement learning

SHAN Shengzhe[1,2], ZHANG Weiwei[1,*]

1. School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China
2. 93995 Unit of the Chinese People's Liberation Army, Xi'an 710306, China

Abstract: Air combat is an important element in the three-dimensional nature of war, and intelligent air combat has become a hotspot and focus of research in the military field both domestically and internationally. Deep reinforcement learning is an important technological approach to achieving air combat intelligence. To address the challenge of constructing high-level opponents in single agent training method, a self-play based air combat agent training method is proposed, and a visualization research platform is built to develop a decision-making agent for close-range air combat. The field knowledge of pilots is embedded in the design process of the agent's observation, action, and reward, training the agent to convergence. Simulation experiments show that the air combat tactics of agent gradually improves by self-play training, achieving a win rate of over 70% against the decision making by single agent training and the emerging of the strategies similar to human "single/double loop" tactics.

Keywords: air combat; artificial intelligence; deep reinforcement learning; self-play; agent

Received: 2023-03-21; **Revised**: 2023-06-12; **Accepted**: 2023-08-29; **Published online**: 2023-09-01 17:04
URL: https://hkxb.buaa.edu.cn/CN/Y2024/V45/I4/328723
Foundation item: Science and Technology Foundation of National Defense Key Laboratory (6142219190302)
* **Corresponding author.** E-mail: aeroelastic@nwpu.edu.cn