文章编号:1674-8190(XXXX)XX-001-16

面向智能空战的深度强化学习技术综述

李霓[1],廉云霄[1],周攀[1],谢锋[2],汤志荔[1],周浩然[3],陈军[4]
(1.西北工业大学 航空学院,西安 710072)
(2.航空工业成都飞机设计研究所,成都 610041)
(3.西北工业大学 网络空间安全学院,西安 710129)
(4.西北工业大学 电子信息学院,西安 710129)

摘 要: 目前主要航空大国及相关研究机构将着力点聚焦于智能空战关键技术的探索与研究,而深度强化学习结合了深度学习的感知能力与强化学习的决策能力,在空战能力涌现方面表现出巨大优势。本文结合智能空战发展的迫切需求,在分析和归纳深度强化学习技术领域主流算法的基础上,探讨了其与空战领域的结合点;从算法实现角度指明深度强化学习在空战中的关键技术;通过梳理当前空战领域前沿技术成果,得出未来深度强化学习研究将由单机空战向集群空战发展这一趋势,进而提出了其面临的挑战,为智能空战中智能算法的发展提供借鉴和指导。

关键词: 智能空战;深度强化学习;作战飞机;关键技术;发展趋势
中图分类号: V24 **文献标识码:** A
DOI: 10.16615/j.cnki.1674-8190.XXXX.XX.01

A survey of deep reinforcement learning technologies for intelligent air combat

LI Ni[1], LIAN Yunxiao[1], ZHOU Pan[1], XIE Feng[2], TANG Zhili[1], ZHOU Haoran[3], CHEN Jun[4]
(1.School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China)
(2.AVIC Chengdu Aircraft Design and Research Institute, Chengdu 610041, China)
(3.School of Cyberspace Security, Northwestern Polytechnical University, Xi'an 710129, China)
(4.School of Electronic Information, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Major aviation nations and related research institutions are focusing on exploration and research of key technologies for intelligent air combat. Deep reinforcement learning combines the perceptual ability of deep learning with the decision-making ability of reinforcement learning, demonstrating significant advantages in the emergence of air combat capabilities. Based on the urgent needs of intelligent air combat development, the points of integration with the air combat field are explored by analyzing and summarizing the mainstream algorithms in the field of deep reinforcement learning. From the perspective of algorithm implementation, the key technologies of deep reinforcement learning in air combat are pointed out. By sorting out the current cutting-edge technological achievements in the field of air combat, it is concluded that the future research on deep reinforcement learning will develop from single-to-single air combat to cluster air combat. The challenges algorithm faces are proposed, which can provide the reference and guidance for the development of intelligent algorithms in intelligent air combat.

Key words: intelligent air combat; deep reinforcement learning; combat aircrafts; key technologies; development trend

收稿日期: 2024-03-12; **修回日期:** 2024-03-28
基金项目: 国家自然科学基金(52372398, 61305133)
通信作者: 陈军(1986—),男,博士,副教授。E-mail: junchen@nwpu.edu.cn
引用格式: 李霓,廉云霄,周攀,等. 面向智能空战的深度强化学习技术综述[J]. 航空工程进展, XXXX, XX(XX): 1-16.
LI Ni, LIAN Yunxiao, ZHOU Pan, et al. A survey of deep reinforcement learning technologies for intelligent air combat[J]. Advances in Aeronautical Science and Engineering, XXXX, XX(XX): 1-16. (in Chinese)

0 引 言

空战作为夺取制空权的一个重要手段，是形成并扩大对敌作战优势中最重要的环节之一。随着作战飞机朝着无人化、智能化与集群化方向发展[1]，空战对抗变得愈发激烈，空中战场逐步进入智能空战时代。智能空战是指以人工智能的手段解决复杂空战对抗环境下作战飞机的决策和控制问题。

自20世纪60年代以来，国外已经针对智能空战开展了广泛的探索与研究，基于专家系统的自适应机动逻辑软件是智能空战技术的首次系统性尝试[2-3]。20世纪90年代，NASA兰利研究中心开发了战术研究与评估系统[4]，实现了空战规则自动生成。2010年，麻省理工学院基于近似动态规划方法开发的空战系统已经具有在环境中探索学习的思想[5]。2016年，辛辛那提大学基于遗传模糊树理论开发了智能体Alpha AI，在与著名飞行教官基恩·李的空战模拟中大获全胜，展现了启发式智能空战算法的威力[6]。2019年，美国国防部高级研究项目局（DARPA）举办的Alpha Dog-fight空战比赛中，由苍鹭系统公司开发的空战决策智能体以5:0的成绩击败F-16飞行教官，彰显了其在空战能力涌现方面的巨大优势[7]。2023年，美国空军研究实验室完成了隐形XQ-58A“女武神”无人机的试飞，标志着美国已经初步具备人工智能自主执行空战任务的能力[8]。中国科研机构，如空军工程大学、北京航空航天大学、清华大学、西北工业大学、国防科技大学、南京航空航天大学、复旦大学、空天技术研究所等[9-16]，近年来也加快了对智能空战技术研究的步伐，并取得了丰富的科研成果。

机动决策是空战过程的核心，是指战机根据感知到的战场态势输出合适的机动指令。深度强化学习结合了深度学习的感知记忆和强化学习的探索学习，是一种具备自主决策能力的人工智能技术。基于深度强化学习的机动决策致力于完成不同战场态势到机动指令的映射，通过智能体不断地在环境中探索与学习，从而掌握一种符合人类期望甚至超越人类能力的决策策略[17]，因此在空战机动决策方面表现出巨大优势。同时，智能空战领域亟待发展的有人机智能辅助决策系统、无人机智能自主决策系统和可供飞行员训练的智能对抗模拟系统为空战智能体提供了广阔的应用前景。然而，复杂空战环境对智能算法决策速度、可靠性、决策效果的要求又使得深度强化学习面临巨大的挑战。因此，如何进一步深入研究深度强化学习算法以提高智能体的空战能力是一个值得探讨的问题。

基于上述分析，本文首先从深度强化学习技术领域出发，梳理了深度强化学习领域的主流算法，在此基础上重点分析了其与空战领域的结合点；然后从算法实现的角度对空战中深度强化学习的各个环节进行分析和评价；在此基础上，通过归纳当前空战领域前沿技术成果，提出未来深度强化学习技术在空战中所面临的挑战，以期为面向智能空战的深度强化学习技术发展提供借鉴和指导。

1 深度强化学习技术概述

1.1 深度强化学习基础

深度强化学习的基础来源于深度学习和强化学习。强化学习是一种辅助智能体在与环境的多次交互过程中形成决策策略的方法，其基本思想是通过最大化从环境获得的回报期望值，从而使智能体学习到最优策略。强化学习问题可以通过马尔可夫决策过程（Markov Decision Process，简称MDP）进行描述，用元组$\langle S, A, R, P, \gamma \rangle$表示。其中，$s_t \in S$和$a_t \in A$分别为智能体在$t$时刻的状态和采取的动作，$S$为状态空间，$A$为动作空间；$R$为奖励函数，$r_t$表示智能体执行动作$a_t$后从环境中获得的奖励；$P$为状态转移函数，$p[s_{t+1}|s_t, a_t]$表示在执行动作$a_t$后，环境由$s_t$转移到$s_{t+1}$的概率；$\gamma$为折扣因子，表示当前时刻奖励与未来奖励的权重。定义强化学习的回报函数为

$$U_t = \sum_{i=t}^{n} \gamma^{i-t} r_i \tag{1}$$

强化学习通过动作价值函数$Q_\pi(s_t, a_t)$实现对当前状态和策略下所执行动作的评价，表示未来回报的期望值。

$$Q_\pi(s_t, a_t) = E_{S_{t+1}, A_{t+1}, \cdots, S_n, A_n}[U_t | s_t, a_t] \tag{2}$$

为使智能体学习到最优策略π^*，最优动作价值函数表示为

$$Q^*(s_t, a_t) = \max_{\pi} Q_\pi(s_t, a_t) \tag{3}$$

由此可以推导出最优贝尔曼方程（Optimal

Bellman Equation)为

$$Q^*(s_t,a_t)=E_{S_{t+1}}\left[r_t+\gamma\max_A Q^*(s_{t+1},A)\right] \quad (4)$$

该方程是在已知t时刻真实信息情况下对最优动作价值函数的更精确表示。以Q学习为例,用表格近似最优动作价值函数$Q(s_t,a_t;\theta)\approx Q^*(s_t,a_t)$,其中$\theta$为表格参数;根据最优贝尔曼方程得到更精确的目标值$\hat{y}_t=r_t+\gamma\max_A Q^*(s_{t+1},A)$,由此定义损失函数对表格参数进行梯度更新,损失函数表示为

$$L(\theta)=f\left[Q(s_t,a_t;\theta),\hat{y}_t\right] \quad (5)$$

参数的梯度更新公式表示为

$$\theta_{new}=\theta_{old}-\alpha\cdot\delta_t\cdot\nabla_\omega L(\theta) \quad (6)$$

式中:α为学习率;δ_t为步长;$\nabla_\omega L(\theta)$为损失函数对参数θ的梯度。

为了评价当前局势,强化学习定义了状态价值函数$V_\pi(s_t)$,以表征当前策略下智能体所处状态的好坏。

$$V_\pi(s_t)=E_{A_t\sim\pi}\left[Q_\pi(s_t,a_t)\right] \quad (7)$$

式中:$E_{A_t\sim\pi}$为对动作A_t的期望。

状态价值函数依赖于状态s_t和策略π,因此对状态$S\sim d(\cdot)$求期望得到目标函数$J(\theta)$,其只依赖于策略π,可以实现对当前策略的评价。

$$J(\theta)=E_{S\sim d(\cdot)}\left[V_\pi(s_t)\right] \quad (8)$$

传统表格型强化学习方法只能处理状态空间和动作空间离散且维度较低的任务,然而实际中大多数任务具有连续且高维的状态空间和动作空间,解决该问题的一种有效手段是利用深度学习强大的网络映射能力。深度学习的概念来源于神经网络,神经网络是由大量神经元堆叠而成的一种网络结构,每一个神经元都可以看作一种简单非线性函数,通过多层叠加和参数调整,使神经网络可以拟合各种复杂函数。深度学习经过多年的发展,在传统多层神经网络的基础上逐渐演变出卷积网络[18]、循环网络[19]、生成对抗网络[20]、图神经网络[21]和Transformer网络[22],这些网络应用在不同领域表现出良好的效果。与此同时,强化学习欠缺的策略拟合能力为深度学习的嵌入提供良好的平台,由此发展而来的深度强化学习具备更强的决策能力。

1.2　深度强化学习的发展

深度强化学习的开创性工作是Mnih等[23]结合卷积神经网络与Q学习提出的深度Q网络(Deep Q Network,简称DQN),该模型被用于处理基于视觉感知的控制任务[24]。DQN以神经网络代替状态-Q值的表格映射(又称值网络),使其在面对连续状态任务时表现出良好的效果。然而,在使用贝尔曼方程求目标Q值时,以值网络求取目标值会引起“自举”,同时最大化Q值会带来高估的问题,从而影响算法的准确性。Van Hasselt等[25]基于此提出一种双Q学习的深度Q网络(Double DQN),通过增加目标网络求取较为稳定的目标Q值,从而缓解了DQN的高估问题。为提高DQN算法的收敛速度,Wang等[26]提出一种对决深度Q网络(Dueling DQN),通过定义优势函数来评价当前状态下动作的相对好坏,并以神经网络分别拟合优势函数和状态价值函数实现对Q值的加权求解,提高了算法对经验的利用效率。DQN算法通过设置经验回放机制均匀采样以消除样本的相关性,但忽略了样本的重要程度。Schaul等[27]提出一种比例优先级的非均匀采样方法,通过提高高价值样本的采样概率加快算法的训练速度。深度强化学习通过全连接网络实现高维特征到低维特征的映射,Hausknecht等[28]将长短时记忆网络代替全连接网络,使深度强化学习算法具备时间轴记忆功能,在处理时间序列样本时表现出更好的性能。

以DQN为代表的值网络结构通过Q值选取动作,难以应对复杂动作空间。相比之下,策略网络结构通过神经网络直接输出动作结果,在面对复杂动作空间时表现出巨大优势。基于策略的深度强化学习以式(8)为目标函数,通过策略梯度上升更新网络参数θ。由于在求取策略梯度过程中需要已知状态S的概率密度函数$d(\cdot)$并求期望,实际过程中难以执行,因此可以通过蒙特卡洛方法近似目标函数期望值。Williams等[29]提出一种REINFORCE算法,通过结合实际观测到的回报值和对策略网络求取的梯度,从而实现策略网络的参数更新。REINFORCE算法的缺陷在于近似动作价值函数过程中方差较大,为此Schulman等[30-31]提出一种置信域策略优化(Trust Region Policy Optimization,简称TRPO)方法,基于广义

优势函数的基线算法保证在缩小方差的同时偏差较小，同时通过限制新旧策略的散度差异确定步长参数范围。

REINFORCE算法利用实际观测到的回报值来近似动作价值函数，相比之下采用神经网络近似动作价值函数的方法得到更广泛的应用，称为行动者-评论者（Actor-Critic，简称AC）框架。AC框架结合了值网络和策略网络的优势，策略网络负责做出动作，值网络负责评价动作的好坏。Lillicrap等[32]基于AC框架提出了深度确定性策略梯度（Deep Deterministic Policy Gradient，简称DDPG）算法，策略网络输出确定性动作，值网络拟合动作价值函数，通过策略梯度算法对策略网络进行参数更新。DDPG存在与DQN类似的高估和自举问题，由此衍生了双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，简称TD3）算法[33]，通过引入目标网络解决自举的问题，同时设计两套值网络减小由最大化带来的高估。此外，TD3算法通过在目标策略网络中加入截断正态分布噪声和降低策略网络更新频率从而进一步提高了算法的性能。TD3算法在面对连续控制问题时表现出良好的效果，但其它的输出动作为确定性的，SAC（Soft Actor-Critic）算法[34]，它具有和TD3相似的网络结构。SAC的显著特点是采用了熵正则的策略梯度算法，在致力于获得最大回报的同时使动作熵最大，从而保证SAC算法的策略随机性。TD3算法和SAC算法在应用过程中存在较多需要设定的超参数，Schulman等[35]对TRPO算法进行改进与简化进而提出一种近端策略优化（Proximal Policy Optimization，简称PPO）算法，通过设计一种Clip机制限制新旧策略差异，在策略更新过程中算法自动进行超参数的调整，使得算法在复杂性和表现效果之间取得有利的平衡。典型的单智能体深度强化学习算法分类及特点如表1所示。

表1 单智能体深度强化学习算法

Table 1 Single agent deep reinforcement learning algorithm

序号	分类	算法名称	特点
1	值网络	DQN	以神经网络代替Q表格，使得状态空间由离散变为连续
2		Double DQN	引入目标网络，缓解高估问题
3		Dueling DQN	定义对决网络评价动作好坏，提高经验利用效率
4	策略网络	REINFORCE	神经网络输出动作结果，可以处理连续动作空间
5	AC框架	TRPO	引入广义优势函数和限制新旧策略差异，提高算法稳定性
6		TD3	引入目标网络缓解高估问题，降低策略网络更新频率提高算法的稳定性
7		SAC	采用熵正则的策略梯度算法，保证策略的随机性
8		PPO	设计Clip机制限制新旧策略差异，提高算法稳定性

由于在某些任务中面对的智能体不只一个，如机器人协作，由此在上述单智能体深度强化学习（简称单智能体）方法的基础上发展出一种多智能体深度强化学习（简称多智能体）方法。区别于单智能体任务，多智能体的环境动态性取决于所有智能体的联合动作，环境中每个智能体都面临由于其他智能体策略改变而导致的环境不平稳问题，此外，不断变化的智能体维度及智能体之间的协作和对抗也使得多智能体环境更为复杂。当前多智能体方法仍处于蓬勃发展阶段，诸如大搜索空间、部分观测、环境非平稳和稀疏奖励等深度强化学习基本问题仍待研究和解决[36]。

1.3 深度强化学习在智能空战中的应用

随着武器装备技术的发展，现代空战已经从传统的视距内（With Visual Range，简称WVR）空战扩展到超视距（Beyond Visual Range，简称BVR）空战，智能化技术的突破也使得作战飞机从有人驾驶到有人—无人协同，并最终向无人自主作战方向发展。具备自主作战能力的无人机需要完成集观测—判断—决策—行动为一体的全周期任务，即OODA（Observation，Orientation，Decision，简称Action）环[37-38]（如图1所示）。其中，观测指作战飞机通过空天地（机载传感器、预警机、卫星、地面雷达等）的手段收集空战信息；判断指

作战飞机根据空战信息进行态势评估;决策指作战飞机根据判断结果完成目标和机动动作选取;行动指作战飞机执行机动指令和完成武器发射。在对战过程中,能率先完成OODA环的一方将获取空战优势[39]。

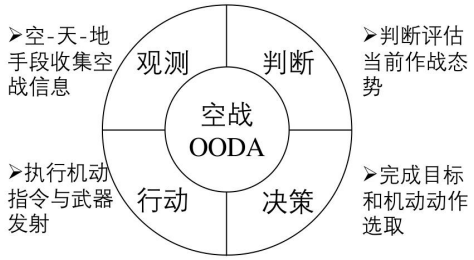


图1　空战OODA环

Fig. 1　OODA for air combat

由深度强化学习构建的策略网络可以实现观测状态到机动动作的端到端映射,通过试错学习机制实现对映射过程的训练。结合空战的OODA及马尔可夫决策过程,将单智能体的空战机动决策通过MDP进行描述为:假设在t时刻空战环境的状态信息为S_t,作战飞机依据自身策略π执行动作A_t,由状态转移函数P产生$t+1$时刻环境状态信息S_{t+1},同时环境根据新的状态S_{t+1}给予作战飞机奖励R_t。由$\langle S_t, A_t, R_t, S_{t+1}\rangle$构成的四元组可以作为经验对智能体进行训练。在考虑单智能体的局部观测问题时,智能体从全局状态信息S_t中观测到部分信息O_t,此时经验元组变为$\langle S_t, O_t, A_t, R_t, S_{t+1}\rangle$,智能体所执行的动作$A_t$依赖于信息$O_t$,而$S_t$、$R_t$和$S_{t+1}$用于对$O_t \rightarrow A_t$的策略网络映射过程进行评价与训练。对于具有$N$个智能体的空战机动决策问题,可以定义状态集合$\{S_t^i\}$和联合动作$\{A_t^i\}$,其中$i=1,2,\cdots,N$,每个智能体依据自身观测$S_t^i$和策略$\pi^i$生成动作$A_t^i$,由状态转移函数$P$产生下一时刻的状态集合$\{S_{t+1}^i\}$。在这个过程中,根据多智能体的合作模式,选择给予个体奖励或集体奖励R,由此构成的经验元组$\langle \{S_t^i\}, \{A_t^i\}, R_t, \{S_{t+1}^i\}\rangle$完成智能体的训练。

将深度强化学习应用到智能空战主要解决作战飞机的机动决策问题,它的思想来源于更早的博弈理论和对策方法[40-43],根据对手的行动从我方机动动作库中选取一个最佳动作。空战机动决策关注于与对手的动态博弈,博弈的目标是在对抗双方不断变换位置中占据和保持位置优势,为快速满足武器发射创造条件。通常将纳什均衡作为优化双方机动策略的目标,在纳什均衡条件下,两方的收支达到平衡,在不改变对手策略的情况下自身策略达到最佳。以DQN及其改进算法为代表,旨在使作战飞机自主完成战术指令或机动动作的选取。随着深度强化学习技术的发展,基于Actor-Critic框架的深度强化学习算法被广泛用于解决空战自主决策问题(如图2所示),可以根据科研者的需求实现连续动作或离散动作的输出,且算法的稳定性和收敛速度得到明显提高,大幅提高了作战飞机的决策能力。

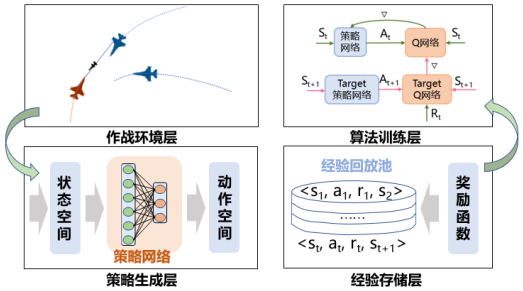


图2　面向空战的深度强化学习框架

Fig. 2　Deep reinforcement learning framework for air combat

2　空战自主决策关键技术

在应用深度强化学习实现空战自主决策过程中,需要结合空战背景解决其中的关键问题。本章从空战中深度强化学习算法实现角度出发,对其中的关键技术进行分析和讨论,包括交互环境、状态空间、动作空间、奖励函数和采样训练方法。

2.1　交互环境

深度强化学习的本质是一种探索和试错学习,在学习过程中依赖于和环境的交互产生经验片段,以此作为自身训练的数据集。交互环境有两个基本作用:状态更新和奖励反馈。状态更新接收来自算法的动作指令,对空战状态信息进行更新。奖励反馈根据状态信息和设计的奖励函数给予算法奖励值,以实现算法的参数训练。在空战研究中,作战飞机的数学模型和奖励函数可以构建出最基本的交互环境,通过作战飞机数学模型可以实例化敌我战机实现状态更新,通过奖励函数可以实现奖励反馈。深度强化学习算法环境交互示意图如图3所示。

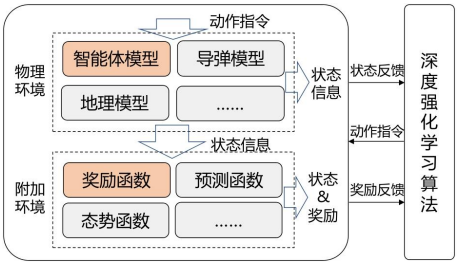


图 3 环境交互原理

Fig. 3 Environmental interaction principle

在考虑构建更贴合实际的交互环境时，一方面需要建立更为精确的飞机模型，如从三自由度质点运动模型扩展到六自由度非线性运动模型，并考虑飞机的外形结构参数和气动参数等[16]。另一方面需要考虑武器和航电系统等，如导弹和机炮的弹道模型、杀伤范围模型，雷达的工作模式和探测范围模型[12]。从现有空战研究中看，各大科研院所多采用自我研发的空战仿真平台或开源空战平台作为算法交互环境，如基于 Unity 3D[11,44]引擎开发的仿真环境，开源空战平台 Harfang 3D[45]，JSBSim[46]等（仿真界面如图 4 所示）。研究者可以考虑基于这些平台进行算法验证和可视化分析，或丰富和改进模型环境以开展更为深入的算法研究。

（a）Unity 3D

（b）Harfang 3D

图 4 空战仿真平台

Fig. 4 Air combat simulation platform

2.2 状态空间

研究空战机动决策问题的关键是要解决作战飞机在不同状态下如何生成最佳的机动动作，状态信息作为决策的前提条件，在这个过程中扮演着至关重要的作用。状态信息是指空战中飞机通过航电设备探测或数据链通信得到的我方和敌方飞行状态信息。状态信息是智能体进行态势评估和动作生成的依据，文献[47]总结了影响空战胜负的状态信息，并将其分为实体信息、目标信息和威胁信息。实体和目标信息分别用于描述我方和敌方飞机的状态，包括位置、姿态、速度等，威胁信息描述我方告警系统感知到的威胁信息，如来袭导弹信息。对于深度强化学习算法而言，提前设计好状态空间作为策略网络输入以实现端到端映射，将大量的状态信息作为状态空间子集时，智能体的探索范围巨大，算法训练久且学习困难。设计的状态信息过少又将导致智能体观测到的信息不全面，难以应对复杂的战场环境。

根据当前空战文献，智能体的状态空间可以从自身状态信息和态势信息两个方向进行设计和组合（如图 5 所示）。自身状态信息指空战中飞机自身的飞行状态参数，如位置、姿态、速度等，这些信息用于保证飞机稳定飞行且不与地理环境发生冲突。态势信息指飞机自身与目标飞机之间的相对信息，如距离、相对速度、相对角度等，态势信息是空战智能体机动决策的重要依据。此外，为了应对更加复杂的空战环境，对目标战机飞行轨迹进行预测得到的预测信息也可以作为态势信息的一部分。在开展集群空战状态空间设计时，由于双方战机数量增加，态势信息也将成倍的增长，除了要考虑敌我之间的相对态势，也要引入临近友方的相对态势，从而涌现出我方的“协同”策略和避免敌方的“围剿”策略。

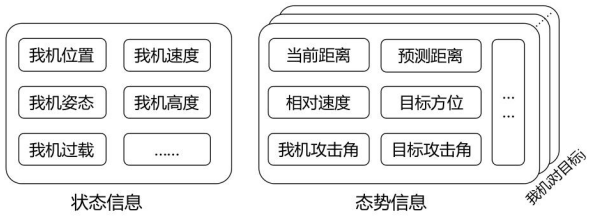


图 5 状态空间设计

Fig. 5 State space design

2.3　动作空间

动作空间作为深度强化学习算法的输出集合,其动作结果作为指令由作战飞机执行。根据所执行的动作类型的不同,将动作空间分为战术策略集、机动动作集和姿态调控集三种(如表2所示)。战术策略集由多个战术策略组成(如盘旋、筋斗等),每一个战术策略都是一个时间序列的动作集合,由空战专家根据经验总结而成,作战飞机在实际空战中根据面临场景的不同从策略集合中选取一个最佳策略执行。机动动作集来源于美国航空航天局(NASA)规定的7种经典机动动作(分别是匀速直飞、加速直飞、减速直飞、左转、右转、上升、下降),在早期采用对策的方法研究空战机动决策时常以该集合作为作战飞机动作库。为了进一步细化机动动作以提高飞机的机动能力,研究者们将飞行方向和飞行加速度进一步细化和组合,从而扩大了动作集的数量。姿态调控集是根据作战飞机数学模型和控制回路模型而建立的姿态指令集合(如迎角、滚转角、舵偏角等),飞机根据集合中的姿态指令和自身的姿态控制回路调整飞行姿态,进而完成相应的机动。

表2　作战飞机动作空间

Table 2　Actor space for combat aircraft

序号	动作类型	内容	特点	参考文献
1	战术策略集	直飞、追踪、盘旋、筋斗、攻击、躲避	离散动作	[48]
2		进攻、防御、追踪射击、撤离		[49]
3	机动动作集	匀速直飞、加速直飞、减速直飞、左转、右转、上升、下降、加/减速左转、加/减速右转	离散动作	[50-52]
4	姿态调控集	法向过载、推力、速度滚转角	连续动作	[53]
5		迎角、滚转角、油门		[16]

深度强化学习基于Q网络或策略网络完成机动指令的生成,其既可以输出离散动作(战术策略集和机动动作集),也可以输出连续动作(姿态调控集)。对于战术策略集而言,其策略元素依据专家经验设计,具有较好的机动稳定性和空战适用性,在近距离空战格斗中能够表现出人类期望的机动效果。然而,固化的战术策略使得无人机行动刻板,难以表现出超越人类的空战能力。机动动作集使得作战飞机机动更为灵活,一方面其可以依靠自动驾驶仪或稳定控制回路使得飞机控制更加稳定,另一方面通过细化组合航向和加速度来变更动作空间的复杂度,从而使科研者在机动灵活性和算法运行效率之间寻求平衡。姿态调控集直接输出连续姿态指令到作战飞机,飞机的机动性和灵活性得到最大发挥,作战能力也大大加强,但对深度强化学习算法而言其控制难度也相应增大,算法训练周期较长。

此外,动作空间的设计也应当符合实际任务需求。如在空战任务中需要考虑规避敌方导弹攻击时,基于专家经验设计一种摆脱导弹追踪的战术策略(例如置尾逃逸与爬升俯冲交替)作为动作空间子集往往会取得令人满意的效果。执行近距离空战任务时需要作战飞机快速变换位置和姿态,因此采用姿态调控集可以充分发挥作战飞机的机动性能。在执行位置姿态不急剧变化的空战任务时,采用机动动作集可以较好的平衡飞机灵活性、控制稳定性和算法训练难度之间的关系。在执行中远距离空战任务时,导弹的高机动能力对飞机造成很大威胁,仅执行简单的机动动作集难以规避导弹,且飞行轨迹容易被预测和追踪。

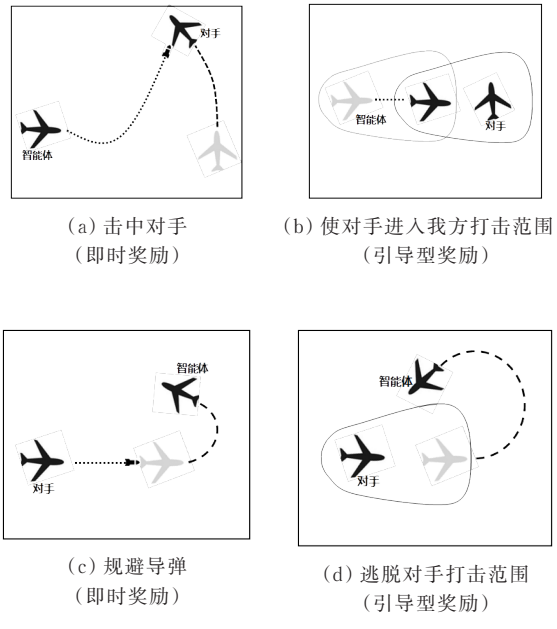
2.4　奖励函数

奖励是环境对智能体执行某种动作后根据新的状态以一定规则给予智能体的反馈,用于指导智能体的策略生成。在空战场景中,奖励函数的目的是驱使作战飞机机动以占据优势位置或击败对手。奖励函数的设计应以所假定空战的制胜准则为标准,空战制胜准则经历了角度准则、能量准则和体系准则的发展过程[39,54-55]。角度准则趋向于有利的空间占位,通过尾追占位为我方战机武器发射创造有利的攻击条件。能量准则致力于使战机拥有最佳的能量优势,包括最佳飞行速度和飞行高度。体系准则的核心是OODA循环,其制胜关键是缩短OODA闭环时间,达到先敌发现和先敌攻击的目的。在体系准则下,奖励函数的设计应以尽可能快地满足武器发射条件为目标。

空空导弹和机载航电技术的发展将传统视距内空战扩展到超视距空战,尽管具备先敌发现和先敌打击优势的超视距空战是未来发展趋势,但现有文献表明,视距内空战在未来空战中仍有至

关重要的作用。根据空战范围的不同，奖励函数的设计也略有差异。目前视距内空战的研究[50]主要沿用传统的角度准则，受限于近距离空空导弹的打击能力，作战飞机的机动性优势更加明显，奖励函数的设计侧重于作战飞机的优势占位，包括角度优势、高度优势、距离优势等。而在现有超视距空战研究[52,56-57]中，空空导弹的全向和远距离打击能力弱化了一定的角度优势和高度优势，其奖励函数的设计与空空导弹攻击范围密切相关，奖励函数的设计更加迎合导弹的杀伤模型。

奖励函数对深度强化学习算法的收敛性有很大影响。由于空战环境的状态空间巨大，沿用传统回合制奖励机制会导致奖励稀疏，智能体学习困难。现有文献多采取细化奖励的方式对空战中的奖励函数进行改进。Li W H等[56]设计了一种基于事件的奖励函数塑造方法细化了原始的回合制奖励。Zhang J D等[58]、Hu J W[59]等、殷宇维等[60]基于态势函数设计了连续型奖励，意图对空战智能体进行引导。本文据此提出一种即时奖励和引导型奖励联合的奖励函数设计方法，其中即时奖励对空战中典型事件给予较大的奖罚值，如导弹命中、战机损毁等，引导型奖励则根据当前作战态势给予智能体较小的连续性奖励，驱使智能体机动以不断提高作战态势。奖励函数设计思路如图6所示。



(a) 击中对手
(即时奖励)

(b) 使对手进入我方打击范围
(引导型奖励)

(c) 规避导弹
(即时奖励)

(d) 逃脱对手打击范围
(引导型奖励)

图6 奖励函数设计
Fig. 6 Design of prize function

2.5 采样训练方法

设计一个良好的样本采样方法可以提高算法训练的效率。在1.2节中提到，DQN算法利用均匀采样机制训练网络，它是一种异策(Off-policy)学习方法，该方法将智能体与环境的交互经验存储到经验回放缓存(又称经验池)中，在训练智能体时以均匀概率采样从经验回放缓存中提取数据，这种方法可以打破交互经验序列的相关性，提高数据的利用效率。研究发现，存放在经验池中的数据重要性不同，以一定机制对样本的重要程度进行排序并对高重要性样本优先采样可以提高算法的收敛速度和学习效率。HU D Y等[52]提出了一种样本质量评估方法，其认为空战中的高态势和低态势样本对空战胜负起关键性作用，因此文中通过态势函数计算出样本的态势值，并给予高/低态势样本以高质量，从而提高智能体对高质量样本的学习概率；周攀等[16]设计了一种样本筛选策略，以样本的动作价值作为评判依据，当多个样本的状态较为接近时舍弃动作价值低的样本，从而提高经验池的样本质量；殷宇维[61]认为网络训练的目标是降低损失值，因此文中将TD误差作为样本价值的判断依据，对TD误差值较大的样本设计更高的采样概率从而加快算法的收敛速度。

综合上述分析，在空战智能体训练过程中，若深度强化学习算法训练缓慢，或智能体在某种场合机动决策效果不佳，可以设计采用样本优先采样机制提高算法的收敛速度，有针对性的对表现不佳的样本优先采样，从而改善和提高智能空战中深度强化学习的整体效果。

3 当前研究及挑战

本文讨论了面向空战的深度强化学习算法实现需要解决的关键技术，旨在帮助读者快速搭建深度强化学习算法，在本节为读者梳理当前领域前沿技术成果，并提出下一步智能空战算法研究所面临的挑战。从现有文献中看(如图7所示)，针对单机的深度强化学习空战算法已经有较多的研究，这些研究主要从探讨深度强化学习算法的可实现性、对算法进行优化以提高空战智能体作战性能的角度开展仿真和分析，例如对比不同机动决策算法研究[62-68]、深度强化学习算法分层策略研究[44,56]、优化采样训练方法研究[69-72]、奖励函数改

进措施研究[73-74]等。现有研究中普遍采用的深度强化学习算法包括 TD3、PPO、SAC 及其改进版本,这些算法同样是当前深度强化学习领域的先进算法,在迁移到空战领域后仍表现出不俗的效果。在单机对抗任务的基础上,学者们考虑了更加复杂的空战任务环境以进一步提高智能体的决策能力,例如对目标敌机机动预测[75-78]、传感器干扰下的算法鲁棒性研究[74,76]以及设计态势评估函数[79-83]。在集群空战方面现有研究相对较少,原因在于集群空战相对更加复杂,且仍有一些关键技术尚待攻克。当前集群空战的研究重点包括机动决策算法研究[84-85]、应对高维状态空间研究[86-88]、设计分层的机动策略研究[89]等。

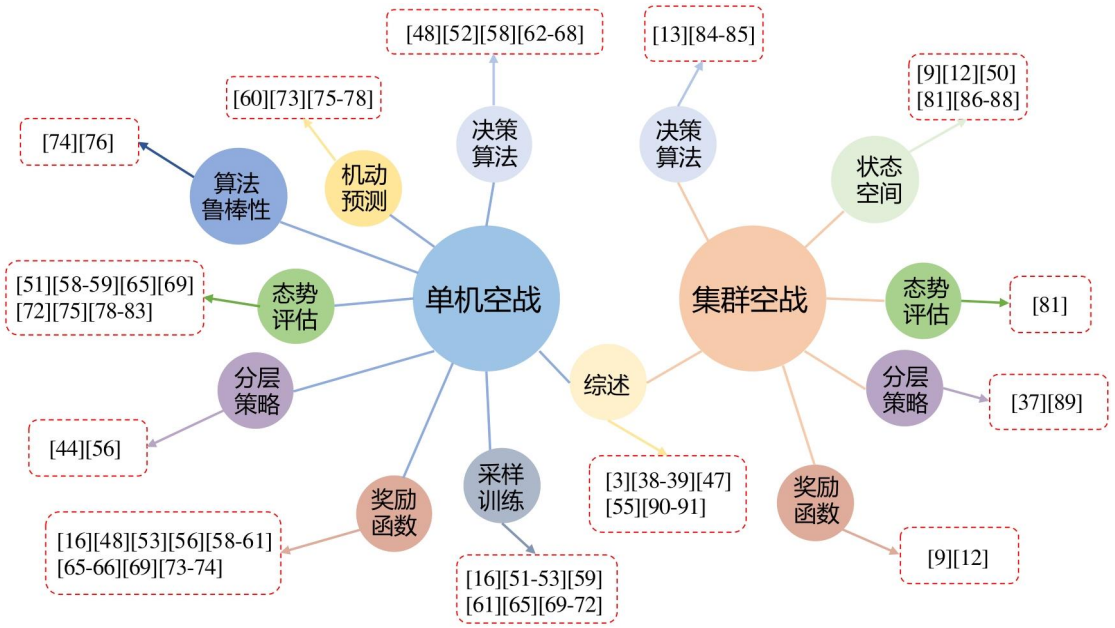


图 7　智能空战研究

Fig. 7　Research on intelligent air combat

从空战发展趋势[90-91]来看,由于单机作战能力有限,智能化与集群化的空战是未来研究重点,深度强化学习技术已经逐步由单智能体向多智能体方向发展[36],现有的深度强化学习决策算法研究集中在单机空战,集群空战仍面临诸多困难和挑战,因此,未来深度强化学习技术的发展方向在集群空战。

3.1　状态信息不完全

状态信息不完备是指在真实作战环境中敌我双方状态信息存在不透明、不连续和不真实的问题。由于目前深度强化学习技术尚未成熟,且开展真实空战实验成本过高,因此现阶段开展的单机空战深度强化学习算法研究均通过仿真环境实现。在仿真训练过程中,普遍假设敌我双方的状态信息是全局已知的,即对战双方都能随时访问到全局状态信息,供态势评估和策略生成。然而实际空战中对手的状态信息并不能随时获得,一方面受限于机载探测设备的信息收集能力,对手的部分状态信息无法探知;另一方面空战过程中双方快速机动、对手采取躲避探测、电磁干扰等对抗策略,原本可以探测到的信息也可能发生短暂丢失,此时智能体将面临信息缺失的风险;此外,由于传感器测量误差或对手采取欺诈手段,获取到的状态信息也并非完全真实,这对训练好的深度强化学习算法能否生成期望的行动是一个挑战。

在当前空战研究中,已有文献开展目标意图识别和状态预测研究,以提高算法的机动决策能力。如 Zhang J D[58]考虑了实际空战中态势信息的时间连续性和相关性,并以此塑造强化学习的奖励函数;Kong W R[60]提出一种最小化 Q 值预测目标动作的轨迹预测方法;JIA L Y 等[69]设计了三种典型意图来构建深度强化学习算法决策模型,为实际对抗训练中的智能体提供意图引导;李永丰等[75]设计了一种基于概率神经网络的目标机动指

令预测模型;Huang C Q[76]提出基于基本动作要素建立对手的机动预测模型,将对手的当前状态作为惯性属性预测其位置。这些研究为克服不完全状态信息难题提供了有效的解决思路,通过对目标进行意图识别和状态预测,以弥补因状态信息缺失导致决策不可靠问题,同时也能显著提高智能体的机动决策能力。

3.2 状态维度高动态变化

由于单机作战能力有限,实际空战中往往难以出现势均力敌的两架作战飞机进行1对1的空战格斗,未来更有可能的空战模式是集群空战。集群空战是指两架或两架以上的作战飞机协同执行空中拦截、对空打击的任务,在空战过程中更加强调协同占位与协同进攻。在集群空战中,作战飞机往往不是单独行动,其面对的对手数量也是未知的,并且在格斗过程中会不断出现敌我战机伤亡的情况,这将导致深度强化学习算法面临状态维度高动态变化的风险。与此同时,现有的深度强化学习方法通过神经网络完成状态到动作的映射,其输入和输出的维度是提前设计好且固定的,因此难以应对集群空战中的维度高动态变化问题。

从现有的集群空战研究中看,开展多智能体深度强化学习方法研究,或将单机空战算法与目标分配方法相融合是解决集群空战中维度动态变化问题的可行路线。如张耀中[12]建立了一种无人机通信交互模型,通过划定通信范围确定动态变化的友机状态信息;Jiang F L[50]在Transformer网络的基础上引入虚拟对象和Mask机制,虚拟对象用于补足和对齐深度强化学习状态空间的维度,Mask机制可以屏蔽不可用或不重要的状态信息,Transformer网络的多头注意力机制可以实现不同无人机态势信息的自主权重分配;唐文泉等[86]设计了一种引入注意力机制的PPO算法,通过为敌方战机分配注意力权重筛选有用的敌方状态信息。这些文献研究的重点在于改进深度强化学习算法的结构。

目标威胁评估技术是实现目标分配的重要手段,在集群空战中的作用尤为重要。目标威胁评估是根据战场态势信息计算敌方目标对我机的威胁程度,确定目标的威胁顺序,进而有助于我方进行合理决策和机动[79]。当前智能空战研究中尚未有文献将目标威胁评估技术与深度强化学习技术相融合,而传统的目标威胁评估技术研究较为丰富,其包括威胁模型评估法[92]和人工智能评估法[93]。前者通过建立目标的威胁评估指标模型量化其威胁程度,近年来的研究及理论包括模糊综合评价法[94],Dempster-Shafer证据理论[95]等。后者则是通过数据驱动的方式对威胁评估模型进行拟合。由于深度强化学习空战智能体依赖状态信息完成机动决策,因此为解决集群空战中的目标分配问题,将目标威胁评估的结果作为状态信息的一部分输入到策略网络中,或根据威胁评估结果进行多目标状态信息筛选是完成两种技术融合的一种途径。

3.3 空战算法评价与验证

面向空战的深度强化学习算法评定与效果验证依赖于对手策略和仿真平台。由于空战中的智能体是在博弈和对抗过程中探索和学习,因此深度强化学习算法的收敛性与对手的智能程度密切相关。现有文献中常见的对手策略包括:敌机执行基本的战术指令[51](如直线飞行或空中盘旋),敌机采用对策的方法从动作库中选取最优动作[42],为敌机设计智能程度相对较低的AI智能体[70],通过自博弈方法训练对抗双方智能体以达到均衡[96],人机空战对抗[49]。不同的对手策略对算法的训练效果截然不同,如智能程度较低的对手策略有助于快速实现算法收敛,但训练后的智能体作战能力有所欠缺。智能程度较高的对手策略有助于提高智能体的空战水平,但若奖励函数设计存在缺陷,算法不易收敛。因此对手策略如何设计在当前研究中仍是一项具有挑战性的工作。

空战仿真平台是深度强化学习算法训练和验证的基础。由于当前基于深度强化学习的智能空战研究尚浅,各研究文献中的仿真环境和仿真条件存在较大差异,这对空战算法的对比验证和快速发展形成制约,同时对构建的仿真环境真实性提出较大考验。受深度强化学习领域蓬勃发展依赖的开源游戏平台Atari 2600[23]、星际争霸Ⅱ[97]的启发,当前智能空战领域需要一款开源空战仿真对抗平台,通过各研究者嵌入算法进行空战对抗

和算法效果验证,以实现算法的改进和空战智能体决策能力的提升,最终促进深度强化学习智能空战决策技术的快速突破和发展。

3.4　真实环境迁移问题

当前空战中的深度强化学习算法研究尚处于仿真阶段,随着算法日趋成熟,未来智能空战研究需要在真实环境中开展飞行试验和算法验证。从当前的研究进展看,制约开展真实环境研究的因素包括:

1) 神经网络可解释性差带来的信任问题

可解释的算法对于保障机动决策的可靠性至关重要,然而在智能空战研究中尚未出现可解释神经网络的相关研究,但在其它领域,如何理解"黑匣子"模式的神经网络已经得到研究人员的关注并取得初步成果[98-99]。

2) 从仿真到真实环境的算法迁移问题

章胜等[100]开展了真实环境下的空战算法迁移研究,并成功将深度强化学习决策模型应用于人机对抗飞行试验,为空战智能决策技术的工程实现提供了良好的参考。尽管如此,深度强化学习技术的真实环境迁移仍面临诸多考验,包括如何实现算法的快速部署,如何克服由于仿真和真实环境差异导致算法效果不同,如何克服有限的机载算力难以应对大网络模型等。

3) 不同场景下经验数据的有效融合问题

在开展从仿真到真实环境的空战决策算法迁移研究时,需要处理来自真实环境、仿真环境和人类飞行员经验的数据。来自真实试验的数据量较少但可信度高,来自人类飞行员/指战员的经验数据可以保证相关机动符合人类期望,从而提高机动品质[101],对智能体均具有优先学习的价值。来自仿真环境的数据获取成本低,可以充分实现智能体的利用与学习。因此,在未来空战决策研究中如何克服算法面临的经验数据来源不同的问题,以及如何平衡不同数据的运用关系是一项重要的研究内容。

4　结束语

随着深度强化学习技术的发展,可以预想到,智能空战在未来将呈现出新的作战模式。这些空战智能体将超越人类飞行员的生理极限,凭借其快速的决策与高难度的机动,在高动态的战场环境中占据优势。通过深度强化学习,空战智能体可以持续积累经验、优化策略,并与友机进行数据共享和策略协同,从而在复杂多变的战场环境中合力制胜。

在这一背景下,本文结合当前人工智能技术热点,阐述了深度强化学习在智能空战中的巨大发展潜力,详细分析了深度强化学习技术原理,归纳了深度强化学习技术领域的主流算法,探讨了其与空战领域的结合点。针对空战自主决策算法实现需要解决的关键技术,梳理当前智能空战领域的前沿技术成果,根据未来由单机空战向集群空战发展这一趋势,总结下一步深度强化学习技术发展所面临的挑战,以期为智能空战算法的发展提供借鉴。

参 考 文 献

[1] 李霓, 布树辉, 尚柏林, 等. 飞行器智能设计愿景与关键问题[J]. 航空学报, 2021, 42(4): 213-230.
LI Ni, BU Shuhui, SHANG Bolin, et al. Aircraft intelligent design: visions and key technologies[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(4): 213-230. (in Chinese)

[2] BURGIN G H, OWENS A J. An adaptive maneuvering logic computer program for the simulation of one-on-one air-to-air combat: NASA CR-2582, CR-2583[R]. Washington, D. C.: NASA, 1975.

[3] 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): 35-49.
SUN Zhixiao, YANG Shengqi, PIAO Haiyin, et al. A survey of air combat artificial intelligence[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 35-49. (in Chinese)

[4] GOODRICH K, MCMANUS J. Development of a tactical guidance research and evaluation system (TGRES)[C]// Flight Simulation Technologies Conference and Exhibit. Reston: AIAA, 1989: 1-8.

[5] MCGREW J S, HOW J P, WILLIAMS B, et al. Air-combat strategy using approximate dynamic programming[J]. Journal of guidance, control, and dynamics, 2010, 33(5): 1641-1654.

[6] ERNEST N, CARROLL D, SCHUMACHER C, et al. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions[J]. Journal of Defense Management, 2016, 6(1): 2167-0374.

[7] HITCHENS T. DARPA's alpha dogfight tests AI pilot's combat chops [EB/OL]. (2020-08-18) [2023-12-31]. https:// breakingdefense. com/2020/08/darpas-alphadog-

fight-tests-ai-pilots-combat-chops/.

[8] DINARDO G. Artificial intelligence flies XQ-58A Valkyrie drone[EB/OL]. [2024-03-12]. https://www.defensenews.com/unmanned/2023/08/03/artificial-intelligence-flies-xq-58a-valkyrie-drone/.

[9] 左家亮, 杨任农, 张滢, 等. 基于启发式强化学习的空战机动智能决策[J]. 航空学报, 2017, 38(10): 217-230.
ZUO Jialiang, YANG Rennong, ZHANG Ying, et al. Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2017, 38(10): 217-230. (in Chinese)

[10] WANG M L, WANG L X, YUE T, et al. Influence of unmanned combat aerial vehicle agility on short-range aerial combat effectiveness[J]. Aerospace Science and Technology, 2020, 96: 105534.

[11] 周文卿, 朱纪洪, 匡敏驰. 一种基于群体智能的无人空战系统[J]. 中国科学: 信息科学, 2020, 50(03): 363-374.
ZHOU Wenqing, ZHU Jihong, KUANG Minchi. An unmanned air combat system based on swarm intelligence[J]. Scientia Sinica Informationis, 2020, 50(03): 363-374. (in Chinese)

[12] 张耀中, 许佳林, 姚康佳, 等. 基于DDPG算法的无人机集群追击任务[J]. 航空学报, 2020, 41(10): 314-326.
ZHANG Yaozhong, XU Jialin, YAO Kangjia, et al. Pursuit missions for UAV swarms based on DDPG algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(10): 314-326. (in Chinese)

[13] 施伟, 冯旸赫, 程光权, 等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.
SHI Wei, FENG Yanghe, CHENG Guangquan, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623. (in Chinese)

[14] LI S Y, CHEN M, WANG Y H, et al. Air combat decision-making of multiple UCAVs based on constraint strategy games[J]. Defence Technology, 2022, 18(3): 368-383.

[15] 马金毅, 王灿, 薛涛, 等. 空战格斗飞行机动数据库建立及应用[J]. 航空学报, 2023, 44(s1): 39-47.
MA Jinyi, WANG Can, XUE Tao, et al. Development and illustrative applications of an air combat engagement database[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(S1): 39-47. (in Chinese)

[16] 周攀, 黄江涛, 章胜, 等. 基于深度强化学习的智能空战决策与仿真[J]. 航空学报, 2023, 44(4): 99-112.
ZHOU Pan, HUANG Jiangtao, ZHANG Sheng, et al. Intelligent air combat decision making and simulation based on deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(4): 99-112. (in Chinese)

[17] KAUFMANN E, BAUERSFELD L, LOQUERCIO A, et al. Champion-level drone racing using deep reinforcement learning[J]. Nature, 2023, 620: 982-987.

[18] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[19] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[20] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[21] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4-24.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. [2024-0312]. https://blog.csdn.net/chengyq116/article/details/106065576/.

[23] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.

[24] 刘全, 翟建伟, 章宗长等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Quan, ZHAI Jianwei, ZHANG Zongzhang, et al. A survey on deep reinforcement learning[J]. Chinese journal of computers, 2018, 41(1): 1-27. (in Chinese)

[25] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning[C]// 2016 the AAAI Conference on Artificial Intelligence. Palo Alto: Assoc Advancement Artificial Intelligence, 2016: 1-8.

[26] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]// International Conference on Machine Learning. San Diego: Machine Learning Research, 2016: 1995-2003.

[27] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[DB/OL]. [2024-0312]. https://blog.csdn.net/beingstrong/article/details/134562992.

[28] HAUSKNECHT M, STONE P. Deep recurrent Q-learning for partially observable MDPs[C]// AAAI 2015 Fall Symposium. California: AI Access Foundation, 2015: 29-37.

[29] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8: 229-256.

[30] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[EB/OL]. [2024-03-12]. https://www.xueshufan.com/publication/1191599655.

[31] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]// International Conference on Machine Learning. San Diego: Journal Machine Learning Research, 2015: 1889-1897.

[32] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6): A187.

[33] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// International Conference on Machine Learning. San Diego: Machine Learning Research, 2018: 1587-1596.

[34] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]// International Conference on Machine Learning. San Diego: Machine Learning Research, 2017: 1352-1361.

[35] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. [2024-0312]. https://arxiv.org/pdf/1707.06347.pdf.

[36] WONG A, BÄCK T, KONONOVA A V, et al. Deep multiagent reinforcement learning: Challenges and directions [J]. Artificial Intelligence Review, 2023, 56(6): 5023-5056.

[37] 王欢, 周旭, 邓亦敏, 等. 分层决策多机空战对抗方法[J]. 中国科学: 信息科学, 2022, 52(12): 2225-2238.
WANG Huan, ZHOU Xu, DENG Yimin, et al. A hierarchical decision-making method for multi-aircraft air combat confrontation[J]. Scientia Sinica Informationis, 2022, 52(12): 2225-2238. (in Chinese)

[38] 黄长强. 未来空战过程智能化关键技术研究[J]. 航空兵器, 2019, 26(1): 11-19.
HUANG Changqiang. Research on key technology of future air combat process intelligentization[J]. Aero Weaponry, 2019, 26(1): 11-19. (in Chinese)

[39] 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): 296-305.
FAN Huitao, YAN Jun. Evolution and development trend of air combat system[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(10): 296-305. (in Chinese)

[40] AUSTIN F, CARBONE G, FALCO M, et al. Game theory for automated maneuvering during air-to-air combat[J]. Journal of Guidance, Control, and Dynamics, 1990, 13(6): 1143-1149.

[41] CRUZ J B, SIMAAN M A, GACIC A, et al. Game-theoretic modeling and control of a military air operation[J]. IEEE Transactions on Aerospace and Electronic Systems, 2001, 37(4): 1393-1405.

[42] LI S Y, CHEN M, WANG Y H, et al. A fast algorithm to solve large-scale matrix games based on dimensionality reduction and its application in multiple unmanned combat air vehicles attack-defense decision-making[J]. In-formation Sciences, 2022, 594: 305-321.

[43] RAMÍREZ LÓPEZ N, ŻBIKOWSKI R. Effectiveness of autonomous decision making for unmanned combat aerial vehicles in dogfight engagements[J]. Journal of Guidance, control, and Dynamics, 2018, 41(4): 1021-1024.

[44] 吴宜珈, 赖俊, 陈希亮, 等. 强化学习算法在超视距空战辅助决策上的应用研究[J]. 航空兵器, 2021, 28(2): 55-61.
WU Yijia, LAI Jun, CHEN Xiliang, et al. Research on the application of reinforcement learning algorithm in decision support of beyond-visual-range air combat[J]. Aero Weaponry, 2021, 28(2): 55-61. (in Chinese)

[45] ÖZBEK M M, YILDIRIM S, AKSOY M, et al. Harfang3D dogfight sandbox: a reinforcement learning research platform for the customized control tasks of fighter aircrafts [EB/OL]. [2024-0312]. https://arXiv preprint: 2210.07282, 2022.

[46] BERNDT J S. JSBSim: an open source flight dynamics model in C++[C]// AIAA Modeling and Simulation Technologies Conference. Virginia: American Institute of Aeronautics and Astronautics, 2004: 261-287.

[47] 贺嘉璠, 汪慢, 方峰, 等. 深度强化学习技术在智能空战中的运用[J]. 指挥信息系统与技术, 2021, 12(5): 6-13.
HE Jiafan, WANG Man, FANG Feng, et al. Application of deep reinforcement learning technology in intelligent air combat[J]. Command Information System and Technology, 2021, 12(5): 6-13. (in Chinese)

[48] ZHU J Y, KUANG M C, ZHOU W Q, et al. Mastering air combat game with deep reinforcement learning[J/OL]. (2023-5-11)[2024-03-12]. http://onineli-brary.wiley.com/doi/10.20944/preprints202305.0862.v1.

[49] 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.

[50] JIANG F L, XU M Q, LI Y Q, et al. Short-range air combat maneuver decision of UAV swarm based on multi-agent Transformer introducing virtual objects[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106358.

[51] LI Y F, SHI J P, JIANG W, et al. Autonomous maneuver decision-making for a UCAV in short-range aerial combat based on an MS-DDQN algorithm[J]. Defence Technology, 2022, 18(9): 1697-1714.

[52] HU D Y, YANG R N, ZHANG Y, et al. Aerial combat maneuvering policy learning based on confrontation demonstrations and dynamic quality replay[J]. Engineering Applications of Artificial Intelligence, 2022, 111: 104767.

[53] 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): 206-218.
SHAN Shengzhe, ZHANG Weiwei. Air combat intelligent decision-making method based on self-play and deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(4): 206-218. (in Chinese)

[54] CRUMPACKER J B, ROBBINS M J, JENKINS P R. An approximate dynamic programming approach for solving an air combat maneuvering problem[J]. Expert Systems with Applications, 2022, 203: 117448.

[55] 孙聪. 从空战制胜机理演变看未来战斗机发展趋势[J]. 航空学报, 2021, 42(8): 8-20.
SUN Cong. Development trend of future fighter: A review of evolution of winning mechanism in air combat[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 8-20. (in Chinese)

[56] SUN Z X, PIAO H Y, YANG Z, et al. Multi-agent hierarchical policy gradient for air combat tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

[57] LI W H, SHI J P, WU Y Y, 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): 1006-1022.

[58] ZHANG J D, YU Y F, ZHENG L H, et al. Situational continuity-based air combat autonomous maneuvering decision-making[J]. Defence Technology, 2023, 29: 66-79.

[59] HU J W, WANG L H, HU T M, et al. Autonomous maneuver decision making of dual-UAV cooperative air combat based on deep reinforcement learning [J]. Electronics, 2022, 11(3): 467.

[60] KONG W R, ZHOU D Y, YANG Z, et al. Maneuver strategy generation of UCAV for within visual range air combat based on multi-agent reinforcement learning and target position prediction[J]. Applied Sciences, 2020, 10(15): 5198.

[61] 殷宇维, 王凡, 吴奎, 等. 基于改进DDPG的空战行为决策方法[J]. 指挥控制与仿真, 2022, 44(1): 97-102.
YIN Yuwei, WANG Fan, WU Kui, et al. Research on air combat behavior decision-making method based on improved DDPG[J]. Command Control & Simulation, 2022, 44(1): 97-102. (in Chinese)

[62] 吴冯国, 陶伟, 李辉, 等. 基于深度强化学习算法的无人机智能规避决策[J]. 系统工程与电子技术, 2023, 45(6): 1702-1711.
WU Fengguo, TAO Wei, LI Hui, et al. UAV intelligent avoidance decisions based on deep reinforcement learning algorithm[J]. Systems Engineering and Electronics, 2023, 45(6): 1702-1711. (in Chinese)

[63] 李波, 白双霞, 孟波波, 等. 基于SAC算法的无人机自主空战决策算法[J]. 指挥控制与仿真, 2022, 44(5): 24-30.
LI Bo, BAI Shuangxia, MENG Bobo, et al. Autonomous air combat decision-making algorithm of UAVs based on SAC algorithm[J]. Command Control & Simulation, 2022, 44(5): 24-30. (in Chinese)

[64] 郭万春, 解武杰, 尹辉, 等. 基于改进双延迟深度确定性策略梯度法的无人机反追击机动决策[J]. 空军工程大学学报(自然科学版), 2021, 22(4): 15-21.
GUO Wanchun, XIE Wujie, YIN Hui, et al. Research on UAV anti-pursing maneuvering decision based on improved twin delayed deep deterministic policy gradient method[J]. Journal of Air Force Engineering University (Natural Science Edition), 2021, 22(4): 15-21. (in Chinese)

[65] ZHANG H P, WEI Y J, ZHOU H, et al. Maneuver decision-making for autonomous air combat based on FRE-PPO [J]. Applied Sciences, 2022, 12(20): 10230.

[66] HU D Y, YANG R N, ZUO J L, et al. Application of deep reinforcement learning in maneuver planning of beyond-visual-range air combat[J]. IEEE Access, 2021, 9: 32282-32297.

[67] 丁维, 王渊, 丁达理, 等. 基于LSTM-PPO算法的无人作战飞机近距空战机动决策[J]. 空军工程大学学报(自然科学版), 2022, 23(3): 19-25.
DING Wei, WANG Yuan, DING Dali, et al. Maneuvering decision of UCAV in close air combat based on LSTM-PPO algorithm[J]. Journal of Air Force Engineering University (Natural Science Edition), 2022, 23(3): 19-25. (in Chinese)

[68] 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 Shengzhe, YANG Mengchao, ZHANG Weiwei, et al. Continuous decision-making method for autonomous air combat[J]. Advances in Aeronautical Science and Engineering, 2022, 13(5): 47-58. (in Chinese)

[69] JIA L Y, CAI C T, WANG X M, et al. Multi-intent autonomous decision-making for air combat with deep reinforcement learning [J]. Applied Intelligence, 2023, 65: 1-18.

[70] YANG Q M, ZHANG J D, SHI G Q, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning[J]. IEEE Access, 2020, 8: 363-378.

[71] 付宇鹏, 邓向阳, 朱子强, 等. 基于价值滤波的空战机动决策优化方法[J]. 航空学报, 2023, 44(22): 20-33.
FU Yupeng, DENG Xiangyang, ZHU Ziqiang, et al. Value-filter based air-combat maneuvering optimization[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(22): 20-33. (in Chinese)

[72] 祝靖宇, 张宏立, 匡敏驰, 等. 稀疏奖励下基于课程学习的无人机空战仿真[J/OL]. 系统仿真学报[1]祝靖宇,张宏立,匡敏驰,等. 稀疏奖励下基于课程学习的无人机空战仿真[J/OL]. 系统仿真学报:1-15[2024-03-29]. https://doi.org/10.16182/j.issn1004731x.joss.23-0349.
ZHU Jingyu, ZHANG Hongli, KUANG Minchi, et al. Curriculum learning based simulation of UAV air combat under sparse rewards[J/OL]. Journal of System Simulation:1-15[2024-03-29]. https://doi.org/10.16182/j.issn1004731x.joss.23-0349.

[73] 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): 33-41.
LI Yongfeng, SHI Jingping, ZHANG Weiguo, et al. Maneuver decision of UCAV in air combat based on deep rein-

forcement learning[J]. Journal of Harbin Institute of Technology, 2021, 53(12): 33-41. (in Chinese).

[74] KONG W R, ZHOU D Y, YANG Z, et al. UAV autonomous aerial combat maneuver strategy generation with observation error based on state-adversarial deep deterministic policy gradient and inverse reinforcement learning[J]. Electronics, 2020, 9(7): 1121.

[75] 李永丰, 吕永玺, 史静平, 等. 深度确定性策略梯度和预测相结合的无人机空战决策研究[J]. 西北工业大学学报, 2023, 41(1): 56-64.
LI Yongfeng, YongxiLYU, SHI Jingping, et al. UAV′s air combat decision-making based on deep deterministic policy gradient and prediction[J]. Journal of Northwestern Polytechnical University, 2023, 41(1): 56-64. (in Chinese)

[76] HUANG C Q, DONG K S, HUANG H Q, et al. Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization[J]. Journal of Systems Engineering and Electronics, 2018, 29(1): 86-97.

[77] 张百川, 毕文豪, 张安, 等. 基于Transformer模型的空战飞行器轨迹预测误差补偿方法[J]. 航空学报, 2023, 44(9): 291-304.
ZHANG Baichuan, BI Wenhao, ZHANG An, et al. Transformer-based error compensation method for air combat aircraft trajectory prediction[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(9): 291-304. (in Chinese)

[78] 毛梦月, 张安, 周鼎, 等. 基于机动预测的强化学习无人机空中格斗研究[J]. 电光与控制, 2019, 26(2): 5-10,22.
MAO Mengyue, ZHANG An, ZHOU Ding, et al. Reinforcement learning of UCAV air combat based on maneuver prediction [J]. Electronics Optics & Control, 2019, 26(02): 5-10,22. (in Chinese)

[79] CAO Y, KOU Y X, XU A, et al. Target threat assessment in air combat based on improved glowworm swarm optimization and ELM neural network[J]. International Journal of Aerospace Engineering, 2021, 37: 1-19.

[80] 杨少博, 张志伟, 蒋道刚, 等. 智能空战仿真中的目标威胁估计方法研究[J]. 飞行力学, 2020, 38(4): 81-86.
YANG Shaobo, ZHANG Zhiwei, JIANG Daogang, et al. Research on target threat assessment method in intelligent air combat simulation[J]. Flight Dynamics, 2020, 38(4): 81-86. (in Chinese)

[81] 朱星宇, 艾剑良. 多对多无人机空战的智能决策研究[J]. 复旦学报(自然科学版), 2021, 60(4): 410-419.
ZHU Xingyu, AI Jianliang. Research on intelligent decision making of many to many unmanned aerial vehicle air combat [J]. Journal of Fudan University (Natural Science), 2021, 60(4): 410-419. (in Chinese)

[82] 马钧文, 毕文豪, 张安, 等. 基于模糊动态权重的近距空战态势评估方法[J/OL]. 控制与决策: 1-10[2024-03-12]. https://doi.org/10.13195/j.kzyjc.2023.0453.
MA Junwen, BI Wenhao, ZHANG An, et al. Close-range air combat situation assessment based on fuzzy dynamic weight[J/OL]. Control and Decision: 1-10[2024-03-12]. https://doi.org/10.13195/j.kzyjc.2023.0453.

[83] 欧洋, 徐扬, 张金鹏, 等. 无人机空战的竞争与双重深度强化学习机动对抗决策[J]. 厦门大学学报(自然科学版), 2022, 61(6): 975-985.
YangOU, XU Yang, ZHANG Jinpeng, et al. UAV air combat dueling and double deep reinforcement learning maneuver adversarial decision making[J]. Journal of Xiamen University (Natural Science), 2022, 61(6): 975-985. (in Chinese)

[84] XUAN S Z, KE L J. UAV swarm attack-defense confrontation based on multi-agent reinforcement learning[C]// 2020 International Conference on Guidance, Navigation and Control. Berlin: Springer. 2022: 5599-5608.

[85] 周文卿, 朱纪洪, 匡敏驰, 等. 基于预知博弈树的多无人机群智协同空战算法[J]. 中国科学: 技术科学, 2023, 53(2): 187-199.
ZHOU Wenqing, ZHU Jihong, KUANG Minchi, et al. Multi-UAV cooperative swarm algorithm in air combat based on predictive game tree[J]. Scientia Sinica(Technologica), 2023, 53(2): 187-199. (in Chinese)

[86] 唐文泉, 孙莹, 杨奇, 等. 一种面向2V2近距空战的强化学习算法[J]. 战术导弹技术, 2022(1): 120-130.
TANG Wenquan, SUN Ying, YANG Qi, et al. A reinforcement learning algorithm for 2V2 close-range air combat [J]. Tactical Missile Technology, 2022(1): 120-130. (in Chinese)

[87] 孔维仁, 周德云, 赵艺阳, 等. 基于深度强化学习与自学习的多无人机近距空战机动策略生成算法[J]. 控制理论与应用, 2022, 39(2): 352-362.
KONG Weiren, ZHOU Deyun, ZHAO Yiyang, et al. Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play[J]. Control Theory & Applications, 2022, 39(2): 352-362. (in Chinese)

[88] 陈宇轩, 王国强, 罗贺, 等. 基于Actor-Critic算法的多无人机协同空战目标重分配方法[J]. 无线电工程, 2022, 52(7): 1266-1275.
CHEN Yuxuan, WANG Guoqiang, LUO He, et al. Target reassignment method for multi-UAV cooperative air combat based on Actor-Critic algorithm[J]. Radio Engineering, 2022, 52(7): 1266-1275. (in Chinese)

[89] 张建东, 王鼎涵, 杨啟明, 等. 基于分层强化学习的无人机多维空战决策[J/OL]. 兵工学报: 1-6. [2024-03-12]. https://kns.cnki.net/kcms/detail/11.2176.TJ.20221213.1026.002.html.
ZHANG Jiandong, WANG Dinghan, YANG Qiming, et al. Multi-dimensional air combat decision-making of UAV based on HRL[J/OL]. Acta Armamentarii: 1-6. [2024-03-12]. https://kns.cnki.net/kcms/detail/11.2176.TJ.

20221213.1026.002.html.

[90] 谢育星，陆屹，管聪，等. 协同空战与多智能体强化学习下的关键问题[J]. 飞机设计，2023，43(1)：6-10.
XIE Yuxing，LU Yi，GUAN Cong，et al. Key problems in coordinated air combat and multi-agent reinforcement learning[J]. Aircraft Design，2023，43(1)：6-10.(in Chinese)

[91] 梁晓龙，胡利平，张佳强，等. 航空集群自主空战研究进展[J]. 科技导报，2020，38(15)：74-88.
LIANG Xiaolong，HU Liping，ZHANG Jiaqiang，et al. Research status and prospect of aircraft swarms autonomous air combat[J]. Science & Technology Review，2020，38(15)：74-88.(in Chinese)

[92] 王新，杨任农，于洋. 基于TOPSIS的空战效能多指标评估模型[J]. 航空工程进展，2020，11(1)：69-76.
WANG Xin，Yang Rennong，Yu Yang. Multi-index evaluation model based on TOPSIS for air combat efficiency[J]. Advances in Aeronautical Science and Engineering，2020，11(1)：69-76.(in Chinese)

[93] 汪泽辉，方洋旺. 基于贝叶斯网络的空战效能评估方法研究[J]. 航空工程进展，2018，9(1)：35-42.
WANG Zehui，FANG Yangwang. Effectiveness evaluation method of air-combat based on Bayesian networks[J]. Advances in Aeronautical Science and Engineering，2018，9(1)：35-42.(in Chinese)

[94] 刘涛，于楠，王方奕，等. 基于模糊综合评价的无人集群系统效能评估[J]. 指挥信息系统与技术，2023，14(4)：20-25,44.
LIU Tao，YU Nan，WANG Fangyi，et al. Effectiveness evaluation of unmanned cluster system based on fuzzy comprehensive evaluation[J]. Command Information System and Technology，2023，14(4)：20-25,44.(in Chinese)

[95] ZHOU Y，TANG Y C，ZHAO X Z. Situation assessment in air combat considering incomplete frame of discernment in the generalized evidence theory[J]. Scientific Reports，2022，12(1)：22639.

[96] AUSTIN F，CARBONE G，FALCO M，et al. Game theory for automated maneuvering during air-to-air combat[J]. Journal of Guidance，Control，and Dynamics，1990，13(6)：1143-1149.

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

[98] ORTEGA-FERNANDEZ I，SESTELO M，VILLANUEVA N M. Explainable generalized additive neural networks with independent neural network training[J]. Statistics and Computing，2024，34(1)：6.

[99] BLAZEK P J，LIN M M. Explainable neural networks that simulate reasoning[J]. Nature Computational Science，2021，1(9)：607-618.

[100] 章胜，周攀，何扬，等. 基于深度强化学习的空战机动决策试验[J]. 航空学报，2023，44(10)：122-135.
ZHANG Sheng，ZHOU Pan，HE Yang，et al. Air combat maneuver decision-making test based on deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica，2023，44(10)：122-135.(in Chinese)

[101] 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)：4-12.
DONG Y Q，AI J L. Decision making in autonomous air combat：Review and prospects[J]. Acta Aeronautica et Astronautica Sinica，2020，41(s2)：4-12.(in Chinese)

(编辑:丛艳娟)