·发展综述·

doi：10.15908/j.cnki.cist.2021.05.002

深度强化学习技术在智能空战中的运用*

贺嘉璠[1] 汪 慢[1] 方 峰[1,2] 李清伟[1] 费爱国[1,3]

（1 信息系统工程重点实验室 南京 210023）
（2 杭州电子科技大学自动化学院 杭州 310018）（3 解放军95899部队 北京 100085）

摘 要：随着指挥控制网络化体系由中心集中式决策控制向边缘分布式决策控制发展，在未来空战场中，以深度强化学习技术为代表的智能决策控制方法在末端自主行动指挥方面将大放异彩。在深度强化学习技术理论成果与应用探索基础上，结合编队智能空战发展的迫切需求与关键问题，分析了该技术的应用潜力与实施价值，提出了解决思路与若干思考，可为空中编队指挥控制智能化发展提供指导。

关键词：深度强化学习；编队智能空战；智能决策；自主控制

中图分类号：TP391 **文献标志码**：A **文章编号**：1674-909X(2021)05-0006-08

Application of Deep Reinforcement Learning Technology in Intelligent Air Combat

HE Jiafan[1] WANG Man[1] FANG Feng[1,2] LI Qingwei[1] FEI Aiguo[1,3]

(1 Science and Technology on Information Systems Engineering Laboratory, Nanjing 210023, China)
(2 School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China) (3 Unit 95899 of PLA, Beijing 100085, China)

Abstract: With the development trend of the command and control network system from the centralized decision-making control to the distributed decision-making control, in the field of future air battle, the intelligent decision-making control methods represented by the deep reinforcement learning technology will show great brilliance on the terminal's autonomous operation command. Based on the theory and application of deep reinforcement learning technology, and combined with the urgent requirements and key problems of the formation intelligent air combat development, the application potential and the implementation value of the technology are analyzed. Some solution ideas are also proposed. The ideas can provide guidance for the intelligent development of the air formation command and control.

Key words: deep reinforcement learning; formation intelligent air combat; intelligent decision; automatic control

0 引 言

夺取制空权是形成并扩大对敌作战优势中最重要的环节之一。随着智能化决策与控制技术和空中武器装备性能的快速发展，空中力量对抗程度愈发激烈，对抗节奏愈发迅速，编队智能空战模式将成为夺取制空权的关键手段。如何在高动态、强实时和非对称的复杂编队空战对抗环境下构建基于智能的

* **基金项目**：装备发展部“十三五”预研课题（31505550302）和江苏省企业院士工作站（BM2018535）资助项目。
收稿日期：2021-05-31
引用格式：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.

指挥决策与控制方案，是亟待解决的问题。

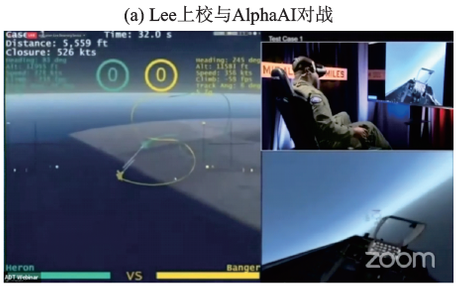
本文从智能空战的研究背景与发展现状出发，剖析了智能空战理论研究与实际应用的若干关键问题。在此基础上，结合新兴的深度强化学习技术方法与问题域概述，分析了应用该技术解决编队智能空战中一系列关键决策问题的可行性，并提出了研究思路、解决方案与落地运用的若干思考。

1　智能空战问题

1.1　外军研究现状

智能空战问题指，如何以人工智能的技术手段解决高动态复杂空战对抗环境下的决策与控制问题。近年来，以美国为代表的世界军事强国，已在空战智能化赋能方面展开了探索性尝试，并取得了初步成果[1-2]。美国空战智能化发展实例如图1所示。

(a) Lee上校与AlphaAI对战



(b) AlphaDogfight比赛直播

图1　美国空战智能化发展实例

2016年，美国辛辛那提大学旗下Psibernetix公司开发了基于遗传模糊树算法的智能飞行员Alpha AI（人工智能），并在空战模拟器中与著名飞行教官基纳·李上校展开了4对2编队超视距对战（图1(a)）。在限制了智能飞行员战机的飞行速度、武器装备和战机作战能力的情况下，Alpha AI仍然大获全胜，展现了人工智能系统在空战问题中的威力[3]。

2019年，美国空军推进Skyborg项目，拟研发可与飞行员一起训练和学习的智能僚机系统[4]。同年5月，美国国防部高级研究项目局（DARPA）宣布启动空战演进（ACE）项目[5]，并于2020年8月全程直播AlphaDogfight近距格斗空战比赛，作为攻克人机协同及提高对自主信任问题的切入点[6]。在AlphaDogfight比赛中，赫伦系统公司的人工智能自主智能体，不仅成功从DARPA选出的8支不同背景的队伍中脱颖而出，而且最终以5:0战绩击败美军武器学校的F-16飞行教官（图1(b)），展示了高级的人工智能算法有能力应用于模拟目视范围内的近距离战斗机空中格斗。值得注意的是，所有参赛团队都基于深度强化学习开发了智能算法，且冠军赫伦系统公司的特色在于未捆绑任何专家系统。赫伦系统公司创建了一个包含有100＋不同智能体的训练环境，不同智能体以不同方式飞行，且具备不同的激励架构和神经网络架构[7]。该试验表明深度强化学习在智能空战方面具有广阔的应用前景。AlphaDogfight给出了一个非常适合智能算法发挥的环境（目标确定、规则简单和环境受限），但赢得该比赛的智能算法在实际复杂作战环境中的价值可能十分有限，尚有进一步学习和提升的空间。

美军ACE计划的第一阶段预计将在2021年结束，其中将包括试验性人工智能驱动系统的飞行测试，形成在小型螺旋桨驱动和喷气式无人飞机上的自主化能力。此后，会有2个阶段的试验，每一阶段均为期16个月，DARPA预期将研发成果转换到更大型的飞行器上测试。ACE项目产生的软件和其他系统，可以帮助改善无人驾驶飞行器的自主操作，并为有人驾驶飞机的机组人员提供新型的智能化协助，并最终预期在2024年左右移植到空军。此外，美军正在寻求推进“忠诚僚机”类项目的开发和演示工作，聚焦多智能体协作和人机协同，在提升无人自主控制软件智能化能力的基础上，应用无人机的低成本和低人员风险优势，实现有人-无人编队智能空战[8]。

1.2　关键问题分析

根据指挥控制的观察—调整—决策—行动（OODA）环概念[9]，编队空战问题本质是综合态势感知、意图预测、战术决策与行动控制的全面性、时效性和精准性的优化问题。

态势感知指，针对先进的机载传感器和数据链获取的态势信息，通过数据融合与算法分析等技术手段感知战场环境、评估威胁程度。意图预测指，根据感知所得的态势中敌方的历史信息集合，分析并预测敌方的战术战法。战术决策指，编队指挥员或长机飞行员通过收集、感知和预测的情报与态势信

息，做出编队级的战术决策。行动控制指，根据编队战术决策分解战术执行动作，并控制空中作战平台进行机动、探测和打击等具体行动。在“人在环”的空战OODA环中，战术决策通常由编队长机飞行员实施，而行动控制由操控每个作战平台的飞行员完成。在未来空战瞬息万变的复杂环境下，编队长机飞行员需要在毫秒级感知、处理数十维态势信息，并做出战术决策指令，同时还需对作战平台进行操控。因此，繁复的工作给其身体与心理带来了巨大的压力与挑战。

综上，在智能空战问题中，战场环境的威胁自主评估与意图预测技术是实现智能化协同作战的前提和条件，战术自主生成与战术自主执行是实现智能化系统作战的核心和过程。因此，通过智能化的技术解决空战过程中的态势全面感知、意图精准预测、战术自主决策与行动自主控制等问题，具有重要意义。

2 深度强化学习算法概述

深度强化学习是一种将深度学习和强化学习技术相结合，使智能体能够从高维空间感知信息，不断试错来训练模型，实现AI自主决策的端到端新技术[10-13]。深度强化学习算法主要包括基于值函数的深度强化学习算法和基于策略的深度强化学习算法等[14]。

2.1 基于值函数的深度强化学习算法

2.1.1 深度Q-网络(DQN)

在DQN算法诞生前，深度学习和强化学习算法在训练数据和学习过程中存在着较大差异，导致深度学习和强化学习难以进行深度融合，无法充分发挥其组合优势。2013年，DeepMind团队的Mnih等[15]提出了DQN算法，该算法为第一个深度强化学习算法，并在2015年得到了进一步完善[16]。DQN算法主要流程是将神经网络与强化学习中的Q-Learning算法相结合，利用神经网络对图像的强大表征能力，将视频帧数据作为强化学习中的状态，并作为神经网络模型的输入；随后，神经网络模型输出每个动作对应的价值，得到将要执行的动作。DQN算法创新之处在于，通过引入目标函数、目标网络及经验回放策略等技术，有效解决了深度学习和强化学习融合过程中存在的挑战，实现了一种端到端的深度强化学习框架。由于深度神经网络具有强大的函数逼近和表示学习特性，使得DQN算法能够解决实际环境中更复杂的决策控制任务，极大扩展了强化学习的应用范围。然而，DQN也有其自身的局限性，主要表现在学习效率较低，且仅能处理离散状态空间中的问题。

2.1.2 基于DQN的改进方法

Hasselt等[17]基于DQN和Double Q-Learning算法提出Double-DQN算法，解决了Q-Learning中过高估计动作值函数的问题。基于回放记忆中经验的重要性不同，Schaul等[18]提出了优先经验回放技术，以便更频繁地回放重要的转移元组，从而更有效地实现学习算法。Wang等[19]提出了一种竞争网络模型取代DQN算法中的网络模型。其核心思想是在神经网络内部将动作值函数分解成状态值函数和动作优势函数，用以解决奖励偏见问题。Raghu等[20]将竞争网络及优先经验回放技术引入DDQN算法，解决了败血症的最佳治疗策略学习问题。其中采用的网络结构结合了DDQN和竞争网络，并命名为竞争双深度Q网络。分布-DQN算法尝试将值函数的分布估计出来，通过使用分布形式表示值函数，智能体可以学习到更丰富的信息，这样可使智能体做出更细致的决策[21]。为了更灵活地提高智能体的探索能力，噪声-DQN在模型参数中加入了一定噪声，通过噪声的随机性改变参数值，进而改变模型的输出值[22]。Rainbow算法基于DQN算法，融合了双层网络、具有优先级的经验存放区技术、竞争网络结构、多步学习、分布式网络和噪声网络特性等技术，形成了针对离散动作空间较完备的优化策略学习算法[23]。

2.2 基于策略的深度强化学习算法

基于值函数求解深度强化学习的方法在实际应用中有一定的局限性，其中最大的不足在于难以高效处理连续动作空间的任务(如DQN算法输出离散状态-动作值函数)。而基于策略梯度的方法，直接通过策略选择动作，有效克服了基于值函数的强化学习求解方法中存在的不足。但由于动作空间的连续性，基于策略梯度的方法，有可能收敛到局部最优解且存在策略评估效率低下的缺点[24]。可分为确定性策略梯度与随机性策略梯度2种深度强化学习算法。

2.2.1 确定性策略梯度

Silver[25]在2014年通过严密的数学推导，证明了确定性策略梯度(DPG)是存在的。基于确定性策略梯度，能够快速有效地求解连续型动作的强化学习任务。其中，每一时间步的动作均为确定性动作值，是无需进行采样的随机策略。2016年，Lillicrap

等[26]将DQN算法采用的深度神经网络的思想应用到连续动作域中，提出一种基于DPG和演员-评论家框架的无模型算法，即深度确定性策略梯度（DDPG）算法。具体地，DDPG采用了经验回放机制，构建了演员、评论家各自单独的策略网络与目标网络，减少了数据间的相关性，增加了算法的稳定性和鲁棒性。DDPG不仅在一系列连续动作空间的任务中表现稳定，而且求得最优解所需的时间步也远远少于DQN。

2.2.2　随机性策略梯度

在非线性优化问题中，梯度的求解相对容易，但合适的优化步长困扰着函数优化的速率。针对该问题，Schulman等[27]提出了可信域策略优化（TRPO）处理随机策略的训练过程，在训练中定义了新策略与旧策略的KL散度（Kullback-Leibler Divergence，又称相对熵），要求状态空间中每个点的KL散度有界限。这样的设计保证了单调改进的策略优化迭代规划，通过引入由KL散度定义的信赖域约束，来选取合适的步长，保证策略的优化总是朝着不变坏的方向进行。针对TRPO的标准解法计算量非常大的问题，Schulman等[28]进一步提出了仅使用一阶优化的近端策略优化（PPO）算法，对代理目标函数简单限定了约束，简化了实现和调参的过程，性能上优于现阶段其他策略梯度算法，表现出同TRPO算法相当的稳定性和可靠性。

3　基于深度强化学习的编队自主行动控制设计与实现

智能空战问题的状态复杂、影响因素交错及作战环境不确定等特性，为深度强化学习技术的运用带来了机遇。如前文所述，智能空战问题是解决以态势全面感知、意图精准预测为前提的战术自主决策下的行动自主控制问题。其中，行动控制包括编队中战机的机动控制、雷达控制和武器控制。基于深度强化学习的智能决策模型可直接实现从态势信息获取到机动动作的映射构建，这种端到端的模型构建方法避免了考虑错综复杂的作战影响因素关系与动态不确定性带来的直接影响[29]。而战机的雷达控制和武器动作控制由于其空间的复杂度不高，可采取固化规则算法降低学习的复杂度，提高智能决策模型的学习效率。

从技术实现角度，编队机动行动自主控制智能体构建的关键因素为编队状态量与动作量的选取、学习训练中回报函数与超参数的设计，基于值函数或策略梯度的学习算法中网络的设计实现等[30]。通过试错对抗，反复学习训练，可得到具有一定能力的编队级智能决策控制体机动决策模型。

3.1　状态空间和动作空间的构造

在深度强化学习理论中，状态空间和动作空间定义的专业性、状态量与动作量划分的精准性直接影响智能决策控制模型的学习性能。针对编队机动决策模型的构造问题，从影响作战胜负的角度出发，状态量大致包含以下3类信息：

1）实体信息：描述我方编队的状态信息，主要包括我方编队中每一架战机的存活状态、空间位置信息（水平位置和垂直高度）、姿态角信息（滚转角、俯仰角和偏航角）、速度信息、加速度信息、剩余载弹量，已发射导弹的信息（是否有效、飞行时间和是否进入主动段）和雷达状态（是否打开和工作模式）等；

2）目标信息：描述我方通过探测或数据链获取的目标信息，主要包括目标与我方编队中某战机的距离信息，目标的速度信息、方位信息、高度信息，目标已被发现/跟踪/打击/丢失的时间信息等；

3）威胁信息：描述我方通过告警系统感知到的威胁信息，主要包括威胁种类（被扫描/被锁定/导弹来袭等），威胁的大致距离信息、方向信息和持续/消失的时间信息等。

从决策方案影响编队机动策略的角度，输出端的机动动作空间中包含的动作量为我方编队中每一架战机的执行动作，主要包括左转、右转、爬升、俯冲、平飞、加速和减速7种基本动作以及左上飞行、左下飞行、右上飞行和右下飞行4种复合动作。综上，状态空间、动作空间及分别包含的状态量与动作量如图2所示。

随着未来大规模无人编队作战体系的成熟与完善，编队中战机数量的增加将使考虑的状态量与动作量的维度成倍增长。因此，探讨空战过程的机理与演进流程，筛选、提取复杂空战环境下影响战斗胜利的关键因素，优化状态空间与动作空间的结构，是一个值得不断深入研究的课题。

3.2　回报值和超参数的整定

回报值通常为判定执行动作有效程度的标量，计算回报值（或描述由动作到状态改变）的映射称为回报函数。在运用深度强化学习技术解决智能空战问题过程中，回报值的确定指根据编队战机动作响应，给予编队级智能决策控制体一个奖励值或惩罚值。回报函数设计得越合理，训练的收敛性越能得

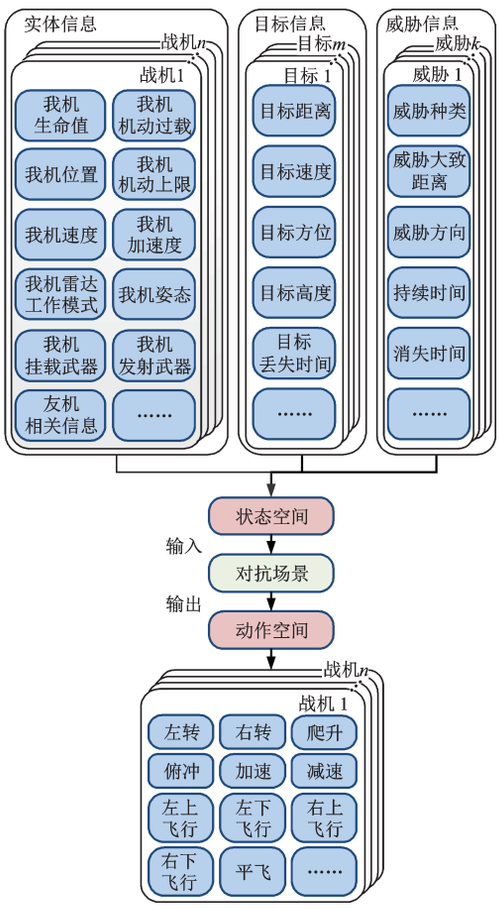


图2 状态空间和动作空间的构造

到保障,则训练所得的编队智能决策控制体性能越优秀。可根据编队空战的战术目标,分别设计引导回报函数、攻击回报函数和防御回报函数等,获得回报值以引导编队级智能决策控制体的学习。其中,引导回报函数可根据编队智能决策控制体的所在位置,给予一个较小的奖励值或惩罚值,引导其进入交战和巡逻等作战状态;攻击回报函数指根据编队智能决策控制体的雷达状态和武器使用状态,对智能体的动作响应(雷达探测目标,雷达锁定目标,发射导弹,导弹处于制导段、进入主动段和命中目标等)分别给予不同的奖励评价;防御回报函数指根据编队智能决策控制体的威胁和规避情况(如被敌雷达锁定、摆脱敌锁定、被敌导弹攻击和摆脱敌导弹等)进行奖励值或惩罚值的设定。此外,回报值的设计也可采用逆强化学习等技术,完善对回报函数的优化,从而提升决策模型的能力。

在深度强化学习技术的运用中,超参数的选取,如每轮空战对抗推演步数、对抗总盘数、学习率、回报折扣因子、记忆区容量大小和最小批容量大小等,也直接决定了训练的有效性与准确性。但同时,超参数的选择也受到算力与资源的约束。随着计算机科学技术的发展,可用计算资源与能力将得到大幅提升,合理的回报值与训练超参数的设计方法将是智能空战问题研究的理论基础之一。

3.3 编队智能决策控制的实现

从空战对抗流程的固有操作序列出发,在智能空战问题中,为了提高学习训练效率,可将深度强化学习算法与固化规则算法的技术有机融合。具体地,机动动作空间的高维性与机动决策的复杂性,导致编队的机动控制难以用精确的建模方法及有效的规划和控制算法实现,而深度强化学习可以利用其端到端的策略生成优势,通过迭代学习形成较可靠的编队机动决策模型。相较之下,雷达控制(包括开关、扫描和锁定)和机载武器控制(发射和制导)具有明确的使用规则,如起落架收起后即可打开雷达、对目标锁定后满足物理约束条件即可进行武器发射等规则,因此,雷达和武器的控制策略可直接采取固化规则的方法进行建模。另一方面,不是所有的时间阶段均需进行导弹发射这一动作的学习,这是因为发弹具有强烈的约束条件。固化发弹规则,将发弹时机作为关键动作量在具有约束的条件下进行学习,可为学习强度“减负”,从而提高学习效率。因此,将时间进程中不同阶段、不同行为的逻辑关系固化为规则算法,与深度强化学习结合,构建合适的学习网络、明确完善的学习机制。

图3给出了以DQN算法与规则算法结合的智能空战指挥决策控制体构建的原理。在双机编队对抗场景下,通过16 000盘(每盘3 500步)的训练获得

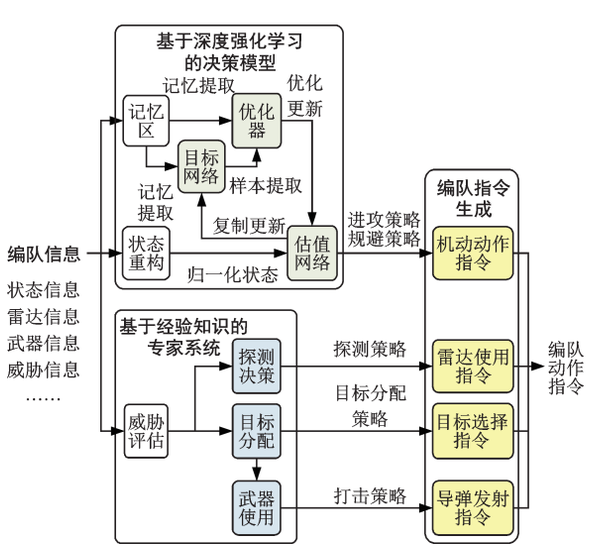
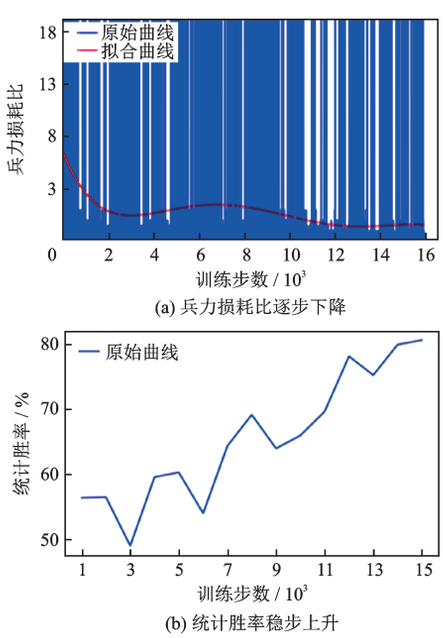


图3 智能空战指挥决策控制体构建原理

试验结果。双机编队智能体完胜对抗训练对手画面如图4所示。由试验结果可知，红方智能体习得目标分配、先敌攻击、双机包夹和车轮战法等多种双机编队战术，与基于经验知识的专家系统进行机机对抗，具有明显优势。图5给出了兵力损耗比和统计胜率试验结果。由图可见，随着训练步数不断增加，对抗模型不断得到优化，所得到的编队智能决策体获胜的概率逐步增高。

图4　双机编队智能体完胜对抗训练对手画面



(a) 兵力损耗比逐步下降

(b) 统计胜率稳步上升

图5　兵力损耗比和统计胜率试验结果

此外，针对智能空战问题，如何实现从混合驱动到自博弈的学习框架均有较高研究价值。因此，在具有细粒度、高逼真的模拟对抗推演环境中，研究基于优秀指战员经验知识的智能空战决策规则，采用深度强化学习解决其中的关键决策点，构造基于联盟智能体群自博弈[31]的学习训练架构（如图6所示），是未来智能空战问题解决的主要思路之一。

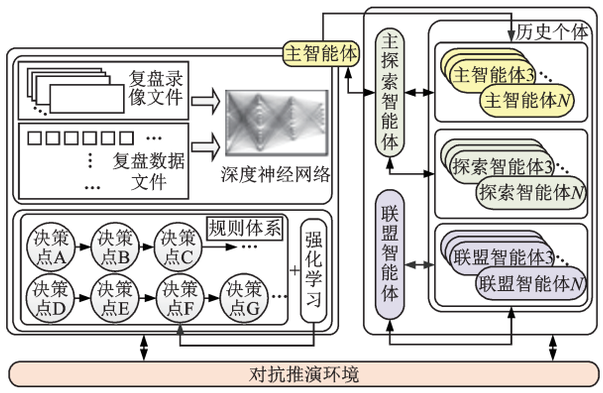


图6　基于联盟智能体群自博弈的学习训练架构

4　展　望

深度强化学习技术为解决智能空战中的决策与控制问题打开了一扇新的大门；同时，如何推动该技术的实际运用与落地实现，也面临着诸多挑战。

1）从单智能体到多智能体的深度强化学习技术

未来空中作战将呈现大编队集群作战趋势，这就要求考虑智能化技术由单智能体向多智能体的深度强化学习技术进行拓展[32-34]。由于智能体数量的增加，状态空间与动作空间存在维数爆炸问题，智能体之间相互影响关系错综复杂，使得多智能体深度强化学习任务相比单智能体深度强化学习任务复杂许多。此外，针对大规模编队下的智能空战问题，构建一个与智能化匹敌的高水平对抗对手，难度也十分巨大。因此，深挖多智能体深度强化学习技术，结合智能空战集群作战场景[35]，研究智能化大规模作战集群的战术决策与行动控制问题，是智能空战未来研究的主要方向之一。

2）从虚拟仿真环境到现实空战环境

考虑到实际环境中的安全因素，目前，在各研究领域中开展的强化学习方法效果验证，均是在仿真环境中先期进行，如智能物流与智慧交通等领域[36-37]。在军事领域，智能算法的仿真验证技术尤为重要，因为战争带来的人力、物力和财力方面的巨大损耗，无法在实际空战中验证所提技术的有效性。因此，借助数字孪生和虚拟现实等高仿真技术，打造空战对抗推演平台是实现智能空战研究的重要环境基础[38]。然而，空战场中的高动态性、强实时性及不确定性，为虚拟环境的打造带来一定难度。此外，验证非平稳环境中算法的迁移泛化能力，也是一个巨大挑战。因此，研究空战实际复杂环境的虚拟化技术，构建高逼真度、细粒度的空战对抗推演平台[39]，

是智能空战技术可行性的基础保障。

3）从单一目标到多目标的深度强化学习技术

未来空中作战必然是体系支援下的空中对抗，因此空战场的战斗任务不再仅是多机编队的超视距空战或近距格斗，而是拓展至体系强对抗下的空-空、空-面作战形式，涉及的作战任务复杂多样，如空中支援、空中掩护、空中突袭和空中巡逻等。因此，研究复杂作战目标情况下的端到端战术决策与执行技术[40]，实现高动态强对抗场景下的空战任务规划、分配和自主执行，是推动智能空战走向未来体系化作战的关键性问题。

5 结束语

本文分析了智能空战问题的基本内涵、外军现状和关键问题，指出了深度强化学习技术在该领域运用的前瞻性与可行性。在此基础上，提出了运用深度强化学习技术解决智能空战问题的基本思路，进一步分析了未来面向大编队、无人化和强对抗应用场景的技术挑战。然而，运用深度强化学习技术解决智能空战问题不仅仅是一个理论问题，更应注重在工程实践中的落地应用，用理论技术指导工程实践，同时用工程实践引导理论技术的发展，形成良性循环，推进空战智能化前进的步伐。

参考文献(References)：

[1] 黄汉桥，白俊强，周欢，等. 智能空战体系下无人协同作战发展现状及关键技术[J]. 导航与控制，2019，18(1)：10-18.

[2] 徐琳，周自力，艾尔，等. 基于大数据的未来智能空战研究[J]. 飞航导弹，2020(10)：41-46.

[3] REILLY M B. Beyond video games：new artificial intelligence beats tactical experts in combat simulation[EB/OL]. (2016-06-27)[2021-04-11]. https://magazine.uc.edu/editors_picks/recent_features/alpha.html.

[4] Air Force Research Laboratory. What is skyborg?[EB/OL]. (2020-02-21)[2021-04-11]. https://afresearchlab.com/technology/vanguards/successstories/skyborg.

[5] MCGLAUN S. DARPA air combat evolution program is on track for live flights this year[EB/OL]. (2021-03-23)[2021-04-10]. https://www.slashgear.com/darpa-air-combat-evolution-program-is-on-track-for-live-flights-this-year-23665043/.

[6] HITCHENS T. DARPA's AlphaDogfight tests AI pilot's combat chops[EB/OL]. (2020-08-18)[2021-04-10]. https://breakingdefense.com/2020/08/darpas-alphadogfight-tests-ai-pilots-combat-chops/.

[7] Heron Systems. Alpha dogfight trials[EB/OL]. (2020-08-18)[2021-04-10]. https://heronsystems.com/work/alpha-dogfight-trials/.

[8] Air Force Technology. Loyal wingman unmanned aircraft[EB/OL]. [2021-04-10]. https://www.airforce-technology.com/projects/loyal-wingman-unmanned-aircraft/.

[9] BOYD J R. A discourse on winning and losin[M]. [S.l.]：Air University Press，1987.

[10] ARULKUMARAN K，DEISENROTH M P，BRUNDAGE M，et al. Deep reinforcement learning：a brief survey[J]. IEEE Signal Processing Magazine，2017，34(6)：26-38.

[11] HENDERSON P，ISLAM R，BACHMAN P，et al. Deep reinforcement learning that matters[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans：AAAI，2018：3207-3214.

[12] 赵冬斌，邵坤，朱圆恒，等. 深度强化学习综述：兼论计算机围棋的发展[J]. 控制理论与应用，2016，33(6)：701-717.

[13] 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.

[14] 刘建伟，高峰，罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报，2019，42(6)：1406-1438.

[15] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing Atari with deep reinforcement learning[EB/OL]. (2013-12-19)[2021-04-10]. https://arxiv.org/abs/1312.5602.

[16] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[17] Van HASSELT H，GUEZ A，SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix：ACM，2016：2094-2100.

[18] SCHAUL T，QUAN J，ANTONOGLOU I，et al. Prioritized experience replay[EB/OL]. (2015-11-18)[2021-04-10]. https://www.arxiv-vanity.com/papers/1511.05952/.

[19] WANG Z Y，SCHAUL T，HESSEL M，et al. Dueling network architectures for deep reinforcement learning[EB/OL]. (2015-11-20)[2021-04-13]. https://arxiv.org/abs/1511.06581.

[20] RAGHU A，KOMOROWSKI M，CELI L A，et al. Continuous state-space models for optimal sepsis treatment：a deep reinforcement learning approach[EB/OL]. (2017-05-23)[2021-04-15]. https://arxiv.org/

pdf/1705.08422.pdf.

[21] BELLEMARE M G,DABNEY W,MUNOS R. A distributional perspective on reinforcement learning[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney:ACM,2017:449-458.

[22] FORTUNATO M,AZAR M G,PIOT B,et al. Noisy networks for exploration [EB/OL]. (2017-06-30)[2021-04-15]. https://arxiv.org/abs/1706.10295v2.

[23] HESSEL M,MODAYIL J,Van HASSELT H,et al. Rainbow: combining improvements in deep reinforcement learning[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI,2018:1-14.

[24] SUTTON R S,BARTO A G. Reinforcement learning: an introduction[M]. Cambridge:MIT Press,1998.

[25] SILVER D,LEVER G,HEESS N,et al. Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing: ACM,2014:387-395.

[26] LILLICRAP T P,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning[EB/OL]. (2015-09-09)[2021-04-15]. https://arxiv.org/abs/1509.02971v6.

[27] SCHULMAN J,LEVINE S,ABBEEl P,et al. Trust region policy optimization[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille:ACM,2015:1889-1897.

[28] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal policy optimization algorithms [EB/OL]. (2017-07-20)[2021-04-15]. https://arxiv.org/abs/1707.06347v2.

[29] 奚之飞,徐安,寇英信,等.多机协同空战机动决策流程[J].系统工程与电子技术,2020,42(2):381-389.

[30] LI Y X. Deep reinforcement learning: an overview[EB/OL]. (2017-01-25)[2021-04-15]. https://arxiv.org/abs/1701.07274v2.

[31] VINYALS O,BABUSCHKIN I,CZARNECKI W M,et al. Grandmaster level in starcraft II using multi-agent reinforcement learning [J]. Nature, 2019, 575(11): 350-354.

[32] 梁星星,冯旸赫,马扬,等.多Agent深度强化学习综述[J].自动化学报,2020,46(12),2537-3557.

[33] PETOSA N,BALCH T. Multiplayer Alphazero[EB/OL]. (2019-10-29)[2021-04-15]. https://arxiv.org/abs/1910.13012.

[34] HU J Z,ZHANG H L,SONG L Y,et al. Cooperative Internet of UAVs: distributed trajectory design by multi-agent deep reinforcement learning [J]. IEEE Transactions on Communications,2020,68(11):6807-6821.

[35] 王肖飞,李冬,陆巍,等.无人机集群战例分析与作战研究[J].舰船电子工程,2020,40(11):16-19.

[36] HAYDARI A,YILMAZ Y. Deep reinforcement learning for intelligent transportation systems: a survey[EB/OL]. (2020-05-02)[2021-04-15]. https://arxiv.org/abs/2005.00935.

[37] KIRAN B R,SOBH I,TALPAERT V,et al. Deep reinforcement learning for autonomous driving: a survey[EB/OL]. (2020-02-02)[2021-04-15]. https://arxiv.org/abs/2002.00444.

[38] HU D Y,ZUO J L,ZHENG W Z,et al. Research on application of LSTM-QDN in intelligent air combat simulation[C]//Proceedings of the 3rd International Conference on Modeling, Simulation and Optimization Technologies and Applications. Beijing: IOPscience, 2020: 1-7.

[39] 金欣.发展智能指挥控制与打造博弈试验平台[J].指挥信息系统与技术,2018,9(5):37-42.

[40] CAO D,HU W H,ZHAO J B,et al. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters[J]. IEEE Transactions on Power Systems,2020,35(5): 4120-4123.

作者简介:

贺嘉璠,男(1990—),工程师,研究方向为机器学习与智能空战。

汪　慢,男(1990—),工程师,研究方向为智能推荐技术与辅助决策技术。

方　峰,男(1991—),讲师,研究方向为空战智能博弈对抗技术。

李清伟,男(1987—),高级工程师,研究方向为智能空战与智能筹划。

费爱国,男(1955—),中国工程院院士,研究方向为军用指挥信息系统与数据链技术,是我国该领域学术带头人之一。

(本文编辑:李素华)