深度强化学习在无人机智能控制中的应用研究

候 磊[1]，贾贝熙[1*]，杜子亮[1]，张 鹏[1]，王田宇[2]
（1. 中国航空系统工程研究所，北京 100029；2. 北京理工大学，北京 100081）

摘 要：随着无人机在军事与民用领域的广泛应用，智能技术在无人机控制领域的相关研究和应用已经成为该领域的研究热点。深度强化学习能解决无人机复杂控制问题，实现无人机端到端的决策控制，同时也为无人机智能化应用带来了新的机遇和挑战。鉴于此，对深度强化学习在无人机智能控制中的应用进行综述。介绍了深度强化学习的基本原理与常见算法；概述了深度强化学习在无人机姿态控制、飞行控制、目标搜索跟踪、集群协同控制、空战决策控制等领域的应用；指出了深度强化学习在无人机控制应用中存在的问题和挑战，讨论了可能的解决方法；对深度强化学习技术在无人机智能控制中的研究进行总结与展望，旨在为无人机系统向自动化、自主化、智能化方向的发展提供参考。

关键词：智能无人机；深度强化学习；无人机姿态控制；目标搜索跟踪；无人机集群；空战决策控制；虚实迁移

中图分类号：V279 **文献标识码：**A **文章编号：**1009-1300(2024)06-0107-11

DOI：10. 16358/j. issn. 1009-1300. 20240016

Application research of deep reinforcement learning in intelligent control of unmanned aerial vehicle

Hou Lei[1], Jia Beixi[1*], Du Ziliang[1], Zhang Peng[1], Wang Tianyu[2]
(1. Aviation System Engineering Institute of China, Beijing 100029, China;
2. Beijing Institute of Technology, Beijing 100081, China)

Abstract: With the wide application of unmanned aerial vehicles (UAVs) in military and civil domains, the research and application of intelligent technology in the domain of UAV control has become the focus in the related field. Deep reinforcement learning(DRL) can solve complex control problems of UAVs and realize end-to-end decision-making control of UAVs. At the same time, it also brings new opportunities and challenges to the intelligent application of UAVs. In view of these, the application of deep reinforcement learning in the intelligent control of UAVs is reviewed. The basic principles and common algorithms of deep reinforcement learning are introduced, and the application of DRL in the fields of UAV attitude control, flight control, target

收稿日期：2024-02-01；**修回日期：**2024-05-30

作者简介：候磊，工程师。

通讯作者：贾贝熙，工程师。

引用格式：候磊，贾贝熙，杜子亮，等. 深度强化学习在无人机智能控制中的应用研究[J]. 战术导弹技术，2024（6）：107-117.（Hou Lei，Jia Beixi，Du Ziliang，et al. Application research of deep reinforcement learning in intelligent control of unmanned aerial vehicle [J]. Tactical Missile Technology，2024（6）：107-117.）

searching and tracking, cluster cooperative control and air combat decision control is outlined. The problems and challenges existing in the application of DRL in UAV control are pointed out, and the possible solutions are discussed. The summary and prospect of the research on deep reinforcement learning technology in the intelligent control of UAVs is given, so as to provide reference for the development of UAV systems towards automation, autonomy and intelligence.

Key words: intelligent UAV; deep reinforcement learning; UAV attitude control; target searching and tracking; UAV swarm; air battle decision control; sim to real

1 引 言

无人驾驶飞行器（Unmanned Aerial Vehicle，UAV），具备机动性强、成本低廉、操作方便简单、便携程度高等优点，近年来得到了快速发展和资源支持。

传统的无人机控制方法通常依赖于预先编程的通用规则和人工设计的控制策略，这限制了无人机在复杂环境中的适应性和灵活性。随着任务复杂度和难度增加，全程人工干预的无人机决策控制方法变得不再适用，需求转向智能化无人机自主决策与任务分配[1]。通过经验驱动进行自主学习最基础的原理框架是强化学习（Reinforcement Learning，RL）。传统强化学习存在可扩展性不足的问题，且主要适用于较低维度的问题。近年来，深度学习的兴起以及深度神经网络强大的函数逼近和表征学习特性为强化学习提供了新的工具，克服了以往的限制。深度学习最重要的特性之一是其能够自动学习高维数据（如图像、文本和音频）的紧凑低维表示（特征）。这使得强化学习能够扩展到具有高维状态和行动空间的环境中，解决以往难以解决的决策问题，深度学习与强化学习的融合技术—深度强化学习（Deep Reinforcement Learning，DRL）应运而生。深度强化学习中智能体与环境的学习交互过程的一般步骤包括：智能体与环境交互生成训练样本，随后，使用神经网络拟合强化学习中的某些模型（如状态转移函数、值函数、策略函数等），智能体根据拟合得到的模型改进其策略。

深度强化学习与无人机的结合是实现无人机自动化、自主化、智能化的重要核心技术之一，相关算法已广泛应用于无人机控制。为梳理深度强化学习在无人机控制的发展脉络，本文总结了其在无人机姿态控制、集群控制等领域的应用。同时，讨论了基于深度强化学习在无人机领域所面临的问题、机遇与挑战，以期为读者提供参考。

2 深度强化学习在无人机控制中的应用

图 1 展示了典型的深度强化学习中智能体与环境的学习交互过程。依据神经网络在强化学习中应用差异，可以将深度强化学习分为两类：无模型（Model-free）方法和基于模型的（Model-based）方法。其中，无模型方法可以进一步分为：基于值（Value-based）、基于策略（Policy-based）、行动-评论（Actor-Critic）。算法分类关系如图 2 所示[2]。基于模型的深度强化学习算法主要源自最优化控制领域。其一般思路为：首先通过正态随机过程或贝叶斯网络等工具对研究对象进行系统建模，然后通过最优控制方法或者机器学习相关算法对模型进行求解。无模型深度强化学习算法源于机器学习领域，本质是一种基于数据驱动的方法。算法通过大量采样收集状态—动作—奖励对数据，通过累积最大折扣奖励优化评估函数，从而优化动作策略。

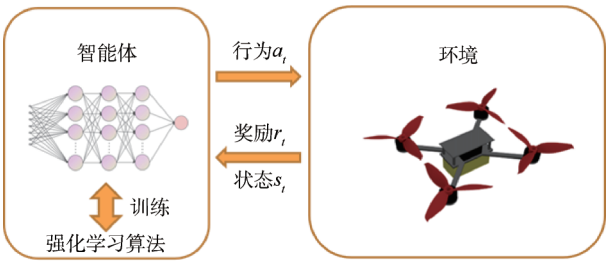

图 1 深度强化学习中智能体与环境的交互过程

Fig. 1 Interaction between agent and its environment in deep reinforcement learning

无人机控制领域面临着环境时变、动力学复杂、对干扰、传感器噪声和未建模动力学模型敏感等问题，经典控制方法的控制效果往往不尽如人意。

研究人员因此引入强化学习方法实现无人机控制。强化学习算法无需环境精确模型，仅通过与环境的交互即可学习控制策略。强化学习在无人机控制中取得了一些成果，但对更高精度和更好鲁棒性的需求促使设计深度强化学习控制方法。本节从不同应用任务类型系统梳理了深度强化学习算法在无人机控制领域中的应用。

深度强化学习在无人机控制中的应用分类如图 3 所示。姿态控制是指控制无人机的俯仰角、滚转角和偏航角，确保无人机的稳定飞行。飞行控制主要包含无人机速度、高度、航向的控制。目标搜索与跟踪是指使用深度强化学习算法通过无人机传感器信息实现对区域内目标高效搜索和目标跟踪。集群控制是指进行无人机之间的无线互联和信息共享，通过多智体强化学习算法实现合作执行任务。空战决策控制是指在空战中做出最优的决策，以提高无人机的生存能力和作战能力。与传统方法相比，深度强化学习算法能够有效地解决复杂的控制问题，如非线性控制、鲁棒控制和最优控制等。无人机控制正在向自主控制的方向发展，而深度强化学习能够为无人机实现自主控制提供强大的算法基础。

2.1　无人机姿态控制与飞行控制

无人机自主飞行控制通常依赖于两个控制环路，内环主要负责控制姿态和速度以执行机动，外环则负责轨迹规划和通信需求，同时也可用于无人机集群管理。在单无人机控制问题中，首先需要关注姿态控制问题。无人机的飞行姿态由许

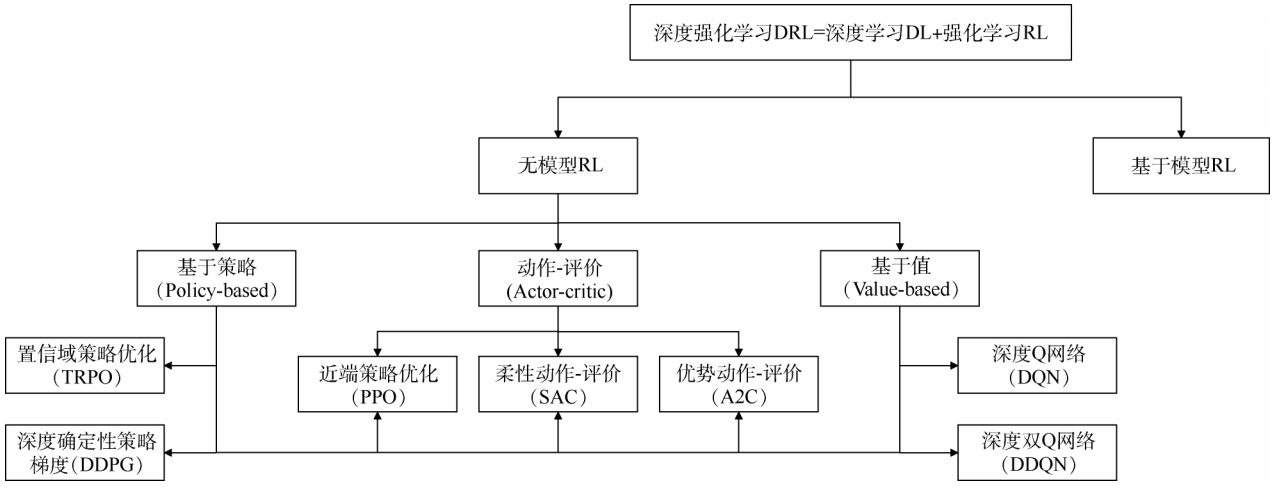

图2　深度强化学习的算法分类

Fig. 2　Taxonomy of deep reinforcement learning algorithms

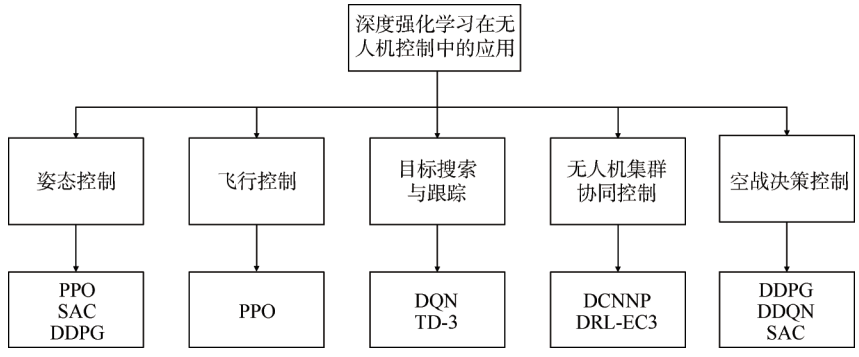

图3　深度强化学习在无人机控制应用中的算法分类

Fig. 3　Taxonomy of deep reinforcement learning algorithms in the context of UAV control applications

多参数同时决定，但在严重的干扰条件下，由于各种非线性因素，当无人机的姿态和空速偏离稳定的水平状态时，由内部控制环路提供的无人机姿态低空稳定功能就会变得越来越困难。在上述情况下，一种可行的解决方案是通过让无人机进行学习训练使无人机能够在湍流或其他极端环境中保持稳定的姿态，实现完全自主飞行。

Koch等[3]对四旋翼无人机的内环姿态角速度控制问题进行了深入研究，利用深度确定性策略梯度（DDPG）、信任区域策略优化（TRPO）和近端策略优化（PPO）三种深度强化学习算法，构建了神经网络控制器，并将其性能与广泛应用的比例-积分-微分（PID）控制器进行了比较。仿真结果表明，基于PPO算法的控制器在上升时间、超调、稳定性等测试评价指标中表现最佳。随后，将得到的PPO控制器迁移到了飞控板上，并进行了上机试验，验证了该控制器的有效性。虽然控制器是基于离散控制任务训练得到的，但研究表明，该方法也可有效应用于连续任务。

Bohn等[4]采用了PPO算法对固定翼无人机进行控制，在考虑动力学非线性以及横向和纵向控制耦合的情况下，采用了PPO算法进行控制器设计，以在各种干扰和不同初始条件下保证飞行稳定性，实现所需的飞行速度、横滚和俯仰角。所提算法能够让无人机的姿态稳定在给定的姿态参考值上，控制器也能很好地适应风和湍流等未知干扰。所提出的姿态控制方案中，奖励函数取决于当前状态与期望状态之间的距离。仿真试验验证了其有效性。

张镭等[5]研究了基于模糊PID和深度强化学习的四旋翼无人机姿态控制，该研究结合了模糊控制的鲁棒性和深度强化学习的自适应性，提出了一种新的控制策略。实验结果表明，该策略能够提高无人机在面对外部干扰时的姿态控制精度和稳定性。Qiu等[6]探讨了基于深度强化学习的移动质量驱动无人机（Moving Mass-Actuated Unmanned Aerial Vehicle，MAUAV）的姿态控制问题，提出一种基于端到端的直接将状态映射到执行器所需偏转量的姿态控制器，解决了由于MAUAV动力学的强非线性和耦合性带来的姿态控制器设计挑战。王伟等[7]设计了一种基于参考模型和DDPG算法的四旋翼无人机姿态控制器，与传统的PID控制器相比，这种基于DDPG和参考模型的控制器在姿态信号跟踪能力和抗干扰性能方面都有显著提升，并显示了强大的鲁棒性。

上述研究都使用PPO算法解决了固定翼无人机和四旋翼无人机的内环姿态控制问题。Xu等[8]制作了一架X布局的无人机模型，该无人机是一种将旋翼和固定翼概念结合起来的特殊无人机，具有垂直起降（VTOL）、长航时、能效高等特点。在仿真环境构建了该飞行器的自动控制系统，该系统由13个状态变量组成；其中四个状态变量是卷积积分误差项，并把实际的物理参数输入到仿真器中并使用PPO算法进行仿真训练，PPO算法在扰动和操作中提供对该复杂耦合系统的纵向和横向控制，训练完成后的控制器能够顺利地控制无人机的悬停、滑翔、降落。该文献提出了一种可以应用于多种无人机的通用控制方案（图4）。

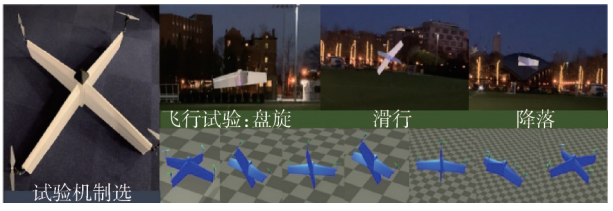

图4 混合布局无人机PPO飞行控制器试飞验证

Fig. 4 Flight test validation of PPO flight controller for hybrid-layout UAV

针对固定翼无人机在复杂环境下的飞行姿态控制问题，孔飞等[9]针对干扰下输入受限的固定翼无人机系统，提出了一种基于行动评论机制的强化学习算法，其中，利用行动者模块求解最小化策略性能函数的控制律，评论家模块实现非线性性能函数的逼近，提高了控制的稳定性，减少了响应时间。为对环境干扰和自身故障有一定的鲁棒性和自适应能力，余自权等[10]提出了一种基于强化学习的自适应容错协同控制算法，引入分数阶微积分算子，给出了强化学习分数阶自适应容错协同控制器。

DDPG、TRPO、PPO等多种深度强化学习算

法被用于无人机姿态、速度等参数的控制，展现出了强大的潜力和价值，已成为无人机姿态控制与飞行控制领域的重要方法，PPO 算法因其性能稳定、适用性强被广泛应用。未来的研究可以在复杂构型无人机姿态控制算法设计、飞行控制算法泛化等方面进行更深入的探索。

2.2　无人机目标搜索与跟踪控制

传统的无人机目标搜索与跟踪方法通常依赖于手工设计的规则和策略，难以适应复杂多变的环境和任务需求。深度强化学习在无人机目标搜索与跟踪控制方面具有广泛的应用潜力并进一步提升系统的决策能力和自主性。通过与传感器数据的实时融合，深度强化学习模型能够根据当前环境状态和目标特征做出准确的决策，并实时调整无人机的行为。这使得无人机能够更加高效地搜索和跟踪移动目标，同时具备避障和路径规划等高级功能。如图 5 所示，无人机群能够协同工作，实现对地面单位的搜索与跟踪。

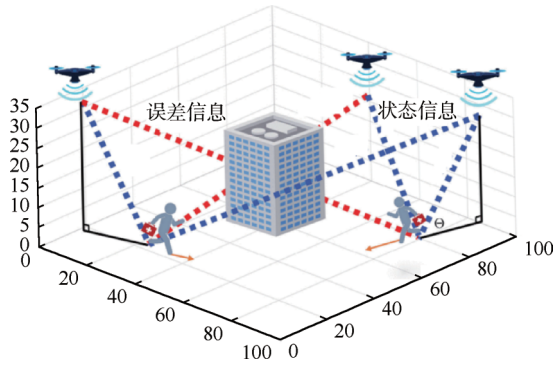

图 5　无人机集群对地面人员的搜索跟踪

Fig. 5　UAV swarm searching and tracking human on ground

针对无人机目标搜索的覆盖路径规划问题，何金等[11]设计了鼓励无人机探索的奖励函数，提出了 PF-DQN 算法，利用 PF 估计无人机的位置和环境信息，将环境信息作为 DQN 的输入，通过 DQN 学习最优的控制策略，用于解决无人机在未知环境中的路径规划问题。针对量战场目标搜寻中存在的目标存活时间短、搜索实时性要求高等难点，杨清清等[12]提出一种基于深度强化学习的海战场目标搜寻规划方法，通过改进 DQN 训练效率较低，且分部信息丢失的缺陷，显著提高了无人机搜寻成功率。牟治宇等[13]基于分层强化学习提出了 option-DQN 算法，并将该算法用于无人机数据收集路径规划任务中，仿真实验验证了算法的有效性和鲁棒性。实验结果表明，该算法能够在未知环境中规划出安全有效的路径，提高了无人机数据收集任务的路径规划效率。针对小型无人机在室内封闭环境下执行目标搜索任务时面临的复杂性和未知性等挑战，赖俊等[14]提出了一种基于空间位置标注的好奇心驱动深度强化学习方法。这种方法利用无人机的访问次数作为好奇心的度量，以鼓励无人机不断探索新的领域。通过这种方式，提高了无人机的感知能力和响应能力，使其能够更快地找到最优搜索策略，从而提升搜索效率和准确率。该方法在室内封闭环境下执行目标搜索任务时表现良好，提高了无人机的搜索效率和准确率。

针对在无人机跟踪任务中，目标的尺度经常发生变化，但是传统的跟踪算法难以适应跟踪目标尺度变化的问题，Zhang 等[15]提出了一种从粗到细的深度强化学习方案，粗跟踪器主导整个边界框，细跟踪器则专注于细化每个边界，二者通过共享具有端到端强化学习架构的感知网络进行联合训练，这种算法优于现有的跟踪算法，在处理无人机跟踪中的长宽比变化时能显著提高精度。为了实现感知-控制端到端无人机目标跟踪，杨兴昊等[16]将无人机所拍摄图像作为卷积神经网络的输入，通过策略网络控制多旋翼无人机电机转速，实现端到端的目标跟踪。为实现无人机对快速机动目标的自主、准确跟踪，Li 等[17]提出了 Meta-TD3 方法。Meta-TD3 方法将深度强化学习与元学习相结合，使无人机能够在目标运动不确定的环境中快速跟踪目标。Meta-TD3 方法的特点包括：将深度强化学习与元学习相结合，使无人机能够更快地适应新的目标运动模式；只需要几步训练，Meta-TD3 方法就能使无人机保持更好的跟踪效果。实验结果表明，与目前最先进的算法相比，Meta-TD3 方法在收敛值和收敛速率方面都有很大提高。李琳等[18]使用卡尔曼滤波用于目标状态估计，DDQN 用于训练无人机在跟踪过程中做出最优决策，仿真结果表明，该方法在目标机动和遮

挡等场景能够有效跟踪机动目标。与传统方法相比，该方法具有更高的跟踪精度和鲁棒性。沈遂欣[19]建立了无人机目标跟踪任务的模型，并使用四种深度强化学习算法进行训练，通过比较在线训练时的指标和离线执行的结果，发现近端策略优化和双决斗深度Q学习算法的目标跟踪性能最佳。黄嘉等[20]提出了一种基于深度确定性策略梯度（DDPG）的无人机目标跟踪算法，该算法通过引入注意力机制来增强无人机对目标的视觉注意力，从而提高了跟踪的准确性和鲁棒性。李明等[21]提出了一种结合深度学习和强化学习的无人机目标跟踪框架，该框架利用深度学习进行特征提取，并通过强化学习优化跟踪策略，实现了对复杂环境下目标的高效跟踪。周立新等[22]研究了一种基于深度强化学习和注意力机制的无人机目标跟踪算法，该算法通过注意力机制来动态调整无人机对目标的关注程度，从而提高了跟踪的精确度和响应速度。

深度强化学习在无人机目标搜索跟踪领域展现出巨大潜力和实际效果。这些研究不仅提高了无人机的自主性和决策能力，也为未来无人机技术的发展提供了新的方向和思路。研究人员采用DQN、DDPG等经典深度强化学习算法大幅提升了传统数学算法、优化算法的性能，随着注意力机制、大模型架构等新技术的引进，可以预见深度强化学习在无人机多目标跟踪、无人机未知区域自主搜索等复杂任务中的应用将更加广泛和深入。

2.3 无人机集群协同控制

鉴于单架无人机在续航时间、探测范围、任务执行能力等方面受到的限制，难以满足冗余任务执行或应对高机动性能目标。通过组建无人机集群，利用无人机之间的无线互联和信息共享，合作执行任务，包括包围和压制敌方机动目标等，可以有效提高任务完成率。无人机集群协同控制框架结构如图6所示。

使用深度强化学习进行无人机集群协同控制的优势在于其自适应性、分布式决策、端到端控制和适应复杂环境和任务的能力[23]。在无人机集群编队控制领域，Tožička等[24]提出了一种使用

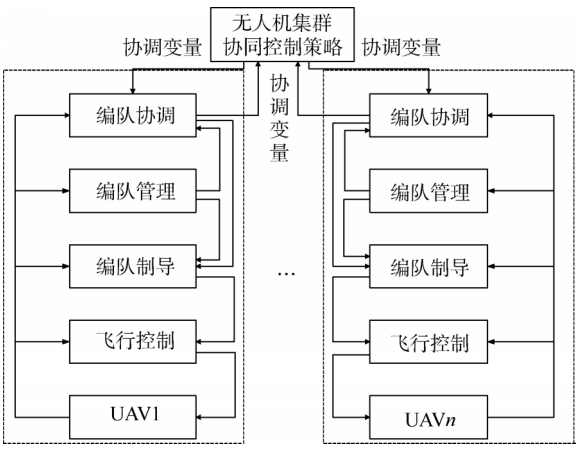

图6 无人机集群协同控制策略

Fig. 6 Collaborative control strategy for UAV swarm

深度卷积神经网络控制策略（DCNNP）来控制五架无人机集群编队的方法。该无人机集群的任务是自主排列成特定形状，并保护指定区域免受潜在威胁。采用了多智能体强化学习（MARL）来解决对群体中无人机状态变量的信息缺乏问题，并通过仿真试验验证了在合作和竞争场景中DCNNP机队控制方法具有良好性能。

在多无人机集群执行任务时，任务分配是一个关键问题。传统的基于贪婪算法或博弈论的分配方法往往效率低下，难以满足实时性和合理性的要求。为解决这一问题，费陈等[25]提出了一种任务分层框架，将目标进行分类，形成任务簇，并映射到无人机编队中。在此基础上，应用MADDPG算法将任务簇内的目标与编队中的无人机进行合理配对。与不分层方法相比，该方法提高了打击完成度和打击效率。这是因为，任务分层框架可以将复杂的任务分解成更小的、更易于管理的任务簇，从而降低了任务分配的难度。此外，深度强化学习算法可以使无人机自主学习和优化任务分配策略，从而提高任务分配的效率和合理性。针对环境变化下任务分配的实时性的问题，刘敬蜀等[26]提出了基于聚类和强化学习的无人机群协同侦察任务规划，该方法将任务区域进行划分，并每个区域的目标归为一个中心目标，从而使得聚类结果更鲁棒的同时有效降低任务的量级，提高任务分配的实时性。

在无人机协同执行任务时，稳定的通信条件

非常重要，如图7所示，无人机需要相互通信实现任务的分配和机动决策，而良好的路由协议能为其在通信条件恶劣场景下的可靠传输提供保障。针对无人机通信中存在的高移动性和节点异常问题，张雅楠等[27]提出了一种基于深度强化学习的无人机可信地理位置路由协议。引入可信第三方提供节点的信任度，使用理论与真实的时延偏差和丢包率作为信任度的评估因子，降低了端到端的延迟并提高了包递交率。稳定良好的通信条件非常重要，但是通信受限环境下的无人机协同策略也必须考虑。针对通信受限环境下的无人机协同决策问题，程进等[28]提出了一种基于动态层级网络通信架构的通信强化学习协同策略。该策略能够有效减少无人机集群之间的通信次数，并准确传递决策所需的信息，从而获得更优的协同策略。

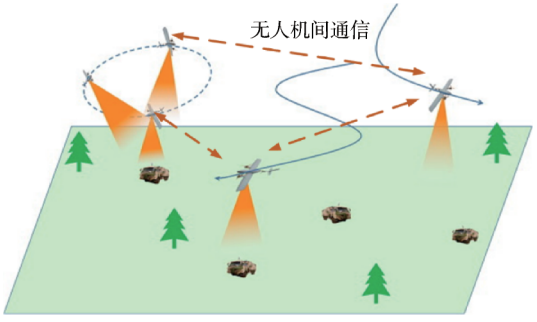

图7　无人机集群在执行任务时的机间通信

Fig. 7　Inter-UAV communication during mission execution for UAV swarm

深度强化学习在无人机集群编队、任务分配、协同通信等方面的应用，极大地提高了无人机集群的自主性和协同性，使无人机集群能够更好地完成复杂的任务。

2.4　无人机空战决策控制

无人机执行侦察、攻击等任务时需要精确、快速且智能的决策控制。基于专家系统、遗传算法、贝叶斯方法的传统空战决策控制方法依赖专家知识，无法自主学习，且决策相对固定，对新环境的适应性较差[29]。随着未来战场规模的扩大和形态的多样化，战场信息量预计将呈现出爆炸性的增长，这使得传统的决策控制方法在处理复杂、动态和不确定的战场环境中面临巨大挑战。在这种情况下，深度强化学习算法的应用显得尤为必要。深度强化学习算法能够通过自我学习和环境交互，自主地从大量复杂、高维度的战场信息中提取有用的特征，并据此做出有效的战术决策。这种方法不仅能够有效地处理信息爆炸问题，而且能够提高无人机的自主决策能力，使其在复杂的战场环境中实现更高效、更智能的空战决策控制。

张强等[30]利用Q网络解决超视距空战中无人机机动决策问题，应用纳什均衡策略选择机动动作。仿真试验结果表明，所提出的方法在超视距空战机动决策任务中具有较好的性能。何金等[31]提出了一种基于马尔可夫决策过程（MDP）的隐蔽接敌深度双Q学习（DDQN）算法，有效地生成了隐蔽接敌策略。张宏鹏等[32]提出了一种使用Q网络进行决策的方法。他们通过构建各种机动动作和模拟来获取样本，训练深度神经网络来预测空战局面，并使用博弈论中的纳什均衡策略来选择最优的机动动作，提高了决策的全局最优性。Li等[33]提出了一种基于多网络深度确定性策略梯度（MN-DDPG）算法和迁移学习技术的无人机机动目标跟踪的方法用于无人机空战决策，迁移学习技术的引入提高了模型的泛化能力。毛梦月等[34]将DDPG算法引入无人机格斗过程中，建立无人机机动模型，并以概率神经网络作为敌方预测单元，在战场完全感知的条件下实现了双方一对一的对抗，该方法能够更准确地预测敌机的机动动作，并选择出更优的决策。Bai等[35]验证了基于双延迟深度确定性策略梯度（TD3）的决策算法比DDPG算法具有更快的收敛速度和更优的决策能力，更适合解决空战决策问题。针对无人机空战决策中的实时性问题，杨霄等[36]提出了一种融合微分对策的强化学习方法。该方法将强化学习的智能决策性与微分对策的精确机动性相结合，以实现战术决策迅速转化为无人机的机动决策。为了增强无人机空战控制决策的泛化性能，Li等[37]提出了一种基于并行自我博弈的PSP-SAC算法。相较于独立训练，该算法通过样本共享和策略共享，在多种战斗环境中有效提高模型的泛化

能力。

未来无人机作战的趋势将朝向多机协同作战的模式发展。多机协同空战涉及多架不同类型的作战飞机之间的相互配合，共同完成对空作战任务。这种作战方式包括协同机动、协同打击和火力掩护等关键环节，是现代海、陆、空、天、电一体化作战模式在空战中的具体实践。因此，提升多机协同的效率对于控制战场制空权、提高对空作战任务的成功率以及减少战斗损失具有至关重要的意义。各国也越来越重视如何通过协同空战来提高无人机群的整体作战效能。相比于单机空战决策，多机协同问题涉及的实体更多，决策空间更广，复杂程度也更高[38]。在这种背景下，深度强化学习技术为实现无人机的自主空战决策提供了可能的解决路径。针对目前决策算法在满足战场实时性决策要求、场景简单且泛化性能差方面存在的共性问题，施伟等[39]提出了一种“集中式训练—分布式执行”多机协同空战决策流程框架。该框架具备实时决策的能力，无需对空战环境和战机飞行动力学进行建模，并对专家经验的需求较小。此外，研究团队提出了四种算法改进机制，有效提高了模型训练的效率和稳定性，为使用深度强化学习算法解决多机协同空战决策问题提供了技术途径。当无人机数量及协同决策内容增加时，多智能体强化学习算法存在训练不易收敛，协同决策水平难以得到显著提升的问题，张磊等[40]提出了一种选择性经验存储策略的多智能体深度确定性策略梯度算法，该算法对经验进行选择，缓解了奖励稀疏的问题，提高了任务完成率。针对无人机空战环境信息复杂、对抗性强所导致的敌机机动策略难以预测，以及作战胜率不高的问题，王昱等[41]设计了一种引导Minimax-DDQN算法，提高了任务完成的成功率，在高对抗的作战环境中具有更强的决策能力，适应性更好。赵琳等[42]将无人机群作战视为公共物品博弈，利用多智能体深度确定性策略梯度（MADDPG）算法求解辅助无人机集群最合理的作战决策，从而以最小的损耗代价实现集群作战胜利。

3 关键问题与讨论

3.1 虚实迁移问题

在无人机飞行控制、自动驾驶等领域，在现实场景中收集数据样本效率低和收集成本高，学者常常利用模拟环境来训练不同的深度强化学习模型。然而，一旦将模型移植到真实无人机中，模拟世界与真实世界之间的“虚拟—现实鸿沟”会降低策略的性能。“虚拟—现实鸿沟”主要有以下几个来源：

（1）环境建模差异。物理仿真器完全复刻真实世界的环境，地形、摩擦等物理特性难以仿真。

（2）感知差异。现实世界中机器通过传感器感知信息，往往存在噪音、光线明暗等因素，会对机器进行环境感知造成影响。

（3）动力学建模差异。仿真建模无法精确地刻画真实机器的动力学、电机模型等方面的特性。

（4）控制差异。受通信传输和机械传动的影响，从机器发出指令到开始执行指令之间存在延时，且控制信号存在噪声，这些都会对控制造成影响[43]。

如何在真实世界环境中通过转移知识和相应调整策略来利用基于仿真的训练，是解决虚实迁移的关键。常见的方法有：（1）建立一个逼真的模拟器，或拥有足够多的模拟经验，以便在现实环境中直接应用模型。这种策略通常被称为“零点转移”或“直接转移”[44]。建立真实世界精确模型的系统识别和领域随机化技术，也可视为一次转移。（2）领域随机化[45]，通过高度随机化模拟，以覆盖真实世界数据的真实分布，Koch等[3]通过域随机化模拟传感器噪声，其对后续无人机控制策略的虚实迁移起到了正向效果。（3）域适应，尝试使源域和目标域两个特征空间统一起来，将仿真环境中训练得到的策略在现实环境中进行再适应。

3.2 算法设计问题

在深度强化学习的算法设计中，目前仍存在许多局限，这些局限阻碍了深度强化学习在无人机领域的广泛应用：（1）奖励函数设计困难：部分问题的奖励函数设计困难。例如，在无人机编队机动控制问题中，每一架无人机的动作都受偏

航、俯仰、滚动和推力等因素的影响。但是，由于无人机在执行任务过程中很难设定中间每步的奖励，只能使用全局奖励。这引发了奖励稀疏且滞后的问题，导致训练困难。（2）训练困难：基于深度强化学习的端到端方法可以对具有相同分布特性的所有问题实例进行求解。但是，现有的深度强化学习模型通常需要消耗大量的时间进行训练。当面对需要即时决策，而战场态势信息变化超出训练模型的预期设定等问题时，很难在短时间内完成模型的训练，严重时可能贻误战机。（3）多智能体协作不足：基于深度强化学习算法所设计的多智能体协同模型很少考虑多智能体间的沟通协作[46]。而真实的多无人机协同环境往往要求具有不同属性特征的作战主体协同配合，仅仅依靠单个主体很难完成目标任务。

3.3　数据生产问题

深度强化学习算法需要大量的数据来学习和优化策略，因此数据生产至关重要[47-48]。数据生产技术可以分为两类：主动数据生产和被动数据生产。主动数据生产技术通过设计特定的数据收集策略来生成数据。例如，可以考虑结合数字孪生技术来解决数据生产问题。无人机数字孪生，是充分利用无人机物理模型、传感器更新、运行历史等数据，创造一个无人机数字孪生体，是对无人机实体对象的动态仿真[49]。研究者们可以将深度强化学习算法应用到无人机数字孪生体上，对孪生体进行实验。这样可以减少对真实无人机的使用，节约成本，提高安全性，并提高数据生产的效率和质量。被动数据生产技术是指利用现有的数据来生成新的数据，例如，可以使用生成对抗网络（GAN）来生成无人机雾霾天气的航拍图像数据。

在强化学习的探索阶段，可以使用数据生产技术来生成新的数据，以帮助算法更好地探索环境[50]。在训练阶段，可以使用数据生产技术来生成高质量的数据，以提高算法的训练效率和性能。数据生产技术可以极大提高深度强化学习算法的性能和效率。

4　结束语

学者们对深度强化学习在无人机控制中的应用正进行着大量研究，但仍然存在一些待解决的问题。其中之一是算法泛化能力不足，即将训练得到的模型应用于新环境时效果不尽如人意；奖励函数对任务环境依赖性较强，如果不契合，则很难获得最优策略；此外，许多任务场景如空战决策，对决策的实时性要求很高，需要很快作出应对，算法的收敛速度和性能也需要提高。虚实迁移问题的存在也使得深度强化学习算法在无人机上的应用挑战重重，无模型强化学习算法在虚实迁移上所面对的挑战巨大，而基于模型的强化学习可能会成为解决虚实迁移的关键。

深度强化学习为解决无人机智能化问题开辟了一条崭新的道路。本文首先介绍了深度强化学习的相关基础理论，随后详细介绍了深度强化学习在无人机智能控制领域的最新进展，并对存在的关键问题进行了讨论。深度强化学习相关应用将进一步拓宽无人机的使用场景，并对其产生深远影响。

[参　考　文　献]

[1] 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.

[2] 郭宪，宋俊潇，方勇纯. 深入浅出强化学习[M]. 北京：电子工业出版社，2020.

[3] Koch W，Mancuso R，West R，et al. Reinforcement learning for UAV attitude control[J]. ACM Transactions on Cyber-Physical Systems，2019，3(2)：1-21.

[4] Bohn E，Coates E M，Moe S，et al. Deep reinforcement learning attitude control of fixed-wing UAVs using proximal policy optimization[C]. 2019 International Conference on Unmanned Aircraft Systems (ICUAS)，Atlanta，GA，USA，2019.

[5] 张镭，李浩. 基于模糊PID和深度强化学习的四旋翼无人机姿态控制研究[J]. 计算机仿真，2018，35(10)：43-47.

[6] Qiu X，Gao C，Wang K，et al. Attitude control of a moving mass-actuated UAV based on deep reinforcement learning[J]. Journal of Aerospace Engineering，

2022，35（2）：4021133.
[7] 王伟，吴昊，刘鸿勋，等. 基于深度强化学习的无人机姿态控制器设计[J]. 科学技术与工程，2023（34）：14888-14895.
[8] Xu J，Du T，Foshey M，et al. Learning to fly：Computational controller design for hybrid UAVs with reinforcement learning [J]. ACM Transactions on Graphics (TOG)，2019，38（4）：1-12.
[9] 孔飞，赵振根，程磊，等. 输入受限及干扰下固定翼无人机强化学习控制[J]. 电光与控制，2024，31（2）：21-28.
[10] 余自权，程月华，张友民，等. 风扰和故障条件下集群无人机强化学习自适应容错协同控制[J]. 厦门大学学报（自然科学版），2022（6）：943-953.
[11] 何金，丁勇，杨勇，等. 未知环境下基于PF-DQN的无人机路径规划[J]. 兵工自动化，2020，39（9）：15-21.
[12] 杨清清，高盈盈，郭玮，等. 基于深度强化学习的海战场目标搜寻路径规划[J]. 系统工程与电子技术，2022（11）：3486-3495.
[13] 牟治宇，张煜，范典，等. 基于深度强化学习的无人机数据收集和路径规划研究[J]. 物联网学报，2020，4（3）：42-51.
[14] 赖俊，饶瑞. 深度强化学习在室内无人机目标搜索中的应用[J]. 计算机工程与应用，2020（17）：156-160.
[15] Zhang W，Song K，Rong X，et al. Coarse-to-fine UAV target tracking with deep reinforcement learning [J]. IEEE Transactions on Automation Science and Engineering，2018，16（4）：1522-1530.
[16] 杨兴昊，宋建梅，佘浩平，等. 基于深度强化学习的无人机空中目标自主跟踪[J]. 兵工学报，2022，43（12）：2551-2560.
[17] Li B，Gan Z，Chen D，et al. UAV maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning[J]. Remote Sensing，2020，12（22）：1-20.
[18] 李琳，张修社，韩春雷，等. 基于卡尔曼滤波和DDQN算法的无人机机动目标跟踪[J]. 战术导弹技术，2022（2）：98-104.
[19] 沈遂欣. 基于深度强化学习的无人机目标跟踪研究[J]. 电子技术（上海），2022（51）：5-8.
[20] 黄嘉，王玉平，刘志勇，等. 基于注意力机制的DDPG无人机目标跟踪算法[J]. 计算机工程与应用，2020，56（12）：1-7.
[21] 李明，张立新，周晓明，等. 结合深度学习和强化学习的无人机目标跟踪框架[J]. 控制理论与应用，2022，39（2）：285-294.
[22] 周立新，陈勇，张凯，等. 基于深度强化学习和注意力机制的无人机目标跟踪算法[J]. 电子技术，2022，45（2）：105-110.
[23] 甄岩，袁健全，池庆玺，等. 深度强化学习方法在飞行器控制中的应用研究[J]. 战术导弹技术，2020（4）：112-118.
[24] Tožička J，Szulyovszky B，de Chambrier G，et al. Application of deep reinforcement learning to UAV fleet control [C]. Intelligent Systems and Applications：Proceedings of the 2018 Intelligent Systems Conference (IntelliSys)，London，UK，2018.
[25] 费陈，郑晗，赵亮. 基于强化学习的无人机智能任务分配方法[J]. 弹箭与制导学报，2022（6）：61-67.
[26] 刘敬蜀，吴嘉琪，刘旭波. 基于聚类和强化学习的多无人机协同侦察任务规划[J]. 中国电子科学研究院学报，2023（1）：21-25+55.
[27] 张雅楠，仇洪冰. 基于深度强化学习的无人机可信地理位置路由协议[J]. 电子与信息学报，2022（12）：4211-4217.
[28] 程进，胡寒栋，江业帆，等. 基于强化学习的通信受限环境多无人机协同策略[J]. 无人系统技术，2022（5）：12-20.
[29] 陈浩，黄健，刘权，等. 自主空战机动决策技术研究进展与展望[J]. 控制理论与应用，2023，40（12）：2104-2129.
[30] 张强，杨任农，俞利新，等. 基于Q-Network强化学习的超视距空战机动决策[J]. 空军工程大学学报：自然科学版，2018，19（6）：7-8.
[31] 何金，丁勇，高振龙. 基于Double Deep Q-Network的无人机隐蔽接敌策略[J]. 电光与控制，2020，27（7）：6-10.
[32] 张宏鹏，黄长强，轩永波，等. 基于深度神经网络的无人作战飞机自主空战机动决策[J]. 兵工学报，2020，41（8）：10-14.
[33] Li B，Yang Z P，Chen D Q，et al. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning[J]. Defence Technology，2021，17（2）：10-13.
[34] 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.
[35] Bai S X，Song S M，Liang S Y，et al. UAV maneuvering decision-making algorithm based on twin delayed

deep deterministic policy gradient algorithm[J]. Journal of Artificial Intelligence and Technology, 2022, 2 (1): 16-22.

[36] 杨霄, 李晓婷, 赵彦东, 等. 基于深度强化学习与微分对策的无人机空战决策研究[J]. 火力与指挥控制, 2021, 46 (5): 71-75.

[37] Li B, Huang J Y, Bai S X, et al. Autonomous air combat decision-making of UAV based on parallel selfplay reinforcement learning[J]. CAAI Transactions on Intelligence Technology, 2022, 1: 1-18.

[38] 李卿莹. 协同空战技术发展概况及作战模式[J]. 科技与创新, 2020 (7): 124-126

[39] 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.

[40] 张磊, 李姜, 侯进永, 等. 基于改进强化学习的多无人机协同对抗算法研究[J]. 兵器装备工程学报, 2023 (5): 230-238.

[41] 王昱, 任田君, 范子琳. 基于引导Minimax-DDQN的无人机空战机动决策[J]. 计算机应用, 2023, 43 (8): 2636-2643.

[42] 赵琳, 吕科, 郭靖, 等. 基于深度强化学习的无人机集群协同作战决策方法[J]. 计算机应用, 2023, 43 (11): 3641-3646.

[43] René T, Hugo C, Timothé L, et al. Continual reinforcement learning deployed in real-life using policy distillation and sim2real transfer [EB/OL]. 2019-06-11. https: //doi.org/10.48550/arXiv.1906.04452.

[44] Fabio M, Christian E, Michael G, et al. Bayesian domain randomization for sim-to-real transfer[J]. IEEE Robotics and Automation Letters, 2021, 6 (2): 911-918.

[45] Arndt K, Hazara M, Ghadirzadeh A, et al. Meta reinforcement learning for sim-to-real domain adaptation [P]. Compiler: Germany, 10. 48550, 2019-09-16.

[46] Zhao W S, Pe J, Li Q Q, et al. Towards closing the sim-to-real gap in collaborative multi-robot deep reinforcement learning[C]. 5th ICRAE, Singapore, 2020.

[47] Ramya R, Ece K, Debadeepta D, et al. Blind spot detection for safe sim-to-real transfer[J]. Journal of Artificial Intelligence Research, 2020 (67): 191-234.

[48] Gupta A, Eysenbach B, Finn C, et al. Unsupervised meta-learning for reinforcement learning[EB/OL]. 2018. https: //doi. org/10. 48550/arXiv. 1806. 04640

[49] Celiberto Jr L A, Matsuura J P, De Mùntaras R L, et al. Using transfer learning to speed-up reinforcement learning: A cased-based approach [C]. 2010-10-23. Latin American Robotics Symposium And Intelligent Robotics Meeting, IEEE, São Bernardo do Campo, Brazi, 2010.

[50] 俞扬. 离线数据强化学习: 途径与进展[J]. 中国基础科学, 2022, 3: 35-39.