智能无人系统专辑+无人集群系统深度强化学习控制研究进展

梁鸿涛 [1,3)]，王耀南 [1,3)]✉，华和安 [1,3)]，钟杭 [2,3)]，郑成宏 [1,3)]，曾俊豪 [1,3)]，梁嘉诚 [1,3)]，李政辰 [2,3)]

1) 湖南大学 电气与信息工程学院，湖南 长沙 410082

2) 湖南大学 机器人学院，湖南 长沙 410082

3) 湖南大学 机器人视觉感知与控制技术国家工程研究中心，湖南 长沙 410082

✉ 通信作者，E-mail: yaonan@hnu.edu.cn

摘 要 随着无人集群在物流运输、农业管理、军事行动等场景的试验和应用，其面临的作业环境和任务内容日趋复杂，亟需设计效率更高、泛化能力更强、适应性更好的控制算法。将人工智能引入到无人集群系统控制的研究中，能够大幅提升现有无人集群的能力，完成复杂的作业任务。深度强化学习具有深度学习和强化学习的优点，无人集群系统深度强化学习控制研究受到了国内外科研人员的广泛关注，涌现出许多标志性成果。本文将从原理、特点等方面阐述深度强化学习概念，深入分析深度强化学习的多种典型算法，并讨论无人机集群的各类控制需求，进而介绍深度强化学习在无人机集群控制领域的典型研究成果，最后针对该领域研究成果的落地转化总结了应用前景和面临的挑战。

关键词 无人集群；集群控制；深度强化学习；多智能体；人工智能；集群智能；

分类号 V279; TP242

Research Progress on Deep Reinforcement Learning in Control of Unmanned Swarm System

LIANG Hong-tao[1,3)], *WANG Yao-nan*[1,3)]✉, *HUA He-an*[1,3)], *ZHONG Hang*[2,3)], *ZHENG Cheng-hong*[1,3)], *ZENG Jun-hao*[1,3)], *LIANG Jia-cheng*[1,3)],*LI Zheng-chen*[2,3)]

1) School of Electrical and Information Engineering, Hunan University, Changsha Hunan 410082, China

2) School of Robotics, Hunan University, Changsha Hunan, 410082, China

3) National Engineering Research Center of RVC, Hunan University, Changsha Hunan 410082, China

✉ Corresponding author, E-mail: yaonan@hnu.edu.cn

ABSTRACT As is more common to test and use micro unmanned vehicles, such as unmanned aerial vehicles (UAVs), in scenarios like logistics transportation, agricultural management and military operations, in recent years. It is no longer enough to control a single UAV to accomplish all missions. And with the increasing complexity of operating environments and task requirements, an unmanned swarm requires a series of algorithms with higher efficiency, greater generalization ability and better adaptability. The combination of unmanned swarm with artificial intelligence (AI) is becoming a popular direction to face with the above requirements.

DRL is a machine learning method that combines deep learning (DL) and reinforcement learning (RL). It has the advantages of deep learning and reinforcement learning. With an RL method, an agent has the ability to learn from environment by trial and error, and make decisions which getting higher score autonomously. But when the given

基金项目：中国国家重点研发计划（2022YFB4701800，2021ZD0114503）、湖南省自然科学基金（2023JJ40165）、中国博士后科学基金（2022M721094）、中国国家自然科学基金重大研究计划（92148204）、中国国家自然科学基金（62233011，62027810，61971071，62133005，62173132）

environment is complex, the decision function of agent may be too difficult to be implemented. And then the agent cannot make a right decision. DL method has a strong fitting ability. A suitable deep neural network (DNN) can simulate any linear or nonlinear function. If DL method is used to simulate the decision function in RL, the hybrid method can solve the problem that an agent can't make a right decision in complex environment.
The combination of unmanned swarm and DRL method has been widely concerned and deeply studied. This paper introduces the concept of DRL from the aspects of principle and characteristics. It analyzes a variety of typical algorithms of DRL. Then it discusses the various control requirements of UAV swarm, and focuses on the many achievements of combining deep reinforcement learning and UAV swarm control. Finally, it puts forward viewpoints on the application prospects and challenges for the landing and transformation in the combination field.
The concept of unmanned swarm originated from the study of the behavior of biological groups. Many species of bees, ants, birds, fish and other creatures have complex group behavior. These clusters are a large number of independent individuals in accordance with certain aggregation rules to form a coordinated, orderly group movement mechanism. Similar to biological clusters, in the field of robotics or unmanned aerial vehicles (UAVs), unmanned swarm systems are crowded intelligent systems that are composed of a large number of homogeneous or heterogeneous unmanned equipment to achieve mutual behavior coordination and jointly complete specific tasks, through interactive feedback and incentive response of information.
In practical applications, an unmanned swarm system needs to meet the requirements of open environment, changeable situation, limited resources, and real-time response. It requires the system to have multi-core collaborative capabilities such as distributed collaborative perception, intelligent collaborative decision-making, and robust collaborative control. The distributed intelligent collaborative control method based on deep reinforcement learning can fully meet the control requirements of high intelligence and robustness of unmanned cluster systems.
KEYWORDS Unmanned Swarm; Swarm Control; Deep Reinforcement Learning; Multi-Agent; Artificial Intelligence; Swarm Intelligence;

无人集群的概念起源于对生物群体行为的研究。自然界中生物集群随处可见，许多种类的蜜蜂、蚂蚁、鸟、鱼等生物都有复杂的群体行为，这些集群都是由大量独立的个体按照一定聚集规则形成的有序群体，其群体行为往往表现出分布式、协调性、自组织、环境适应性等特点，而且结构稳定，能够产生超越其中个体的智能。类似于生物学集群，机器人领域的无人集群是由大量同构或异构的无人装备组成，通过信息交互、激励响应等方式相配合，实现集群内行为协同，共同完成超越其中个体能力的特定任务的群体智能系统[1]。

在实际应用中，如无人集群对某区域进行持续搜索巡检，无人集群系统面临环境开放、态势多变、资源受限、响应实时等难题，要求系统具备分布式协同感知、智能协同决策、鲁棒协同控制等重要群体能力[11]。为此，可以通过构建庞大的集群网络，从而实现对外部环境的多方感知与高效的最优决策，进而发挥出集群系统的群体智能优势。无人集群系统由大量小而精的同构或异构无人装备组成，每个个体都具备环境感知、数据分析和临机决策的能力。与独立无人装备系统相比，无人集群系统的一大特点是在行为任务上具备协调一致性。如无人机集群由大量具有自主能力的无人机组成，按照一定结构形式进行三维空间排列，在飞行中可以保持队形稳定，并能够针对外部情况和任务需求变化进行队形的动态调整[19]。通过设计邻近个体之间的分布式交互规则，无人集群相对于单个无人装备也具有更好的鲁棒性和适应性，当集群中部分个体出现故障或失效时，其他个体能够做出补充弥补空缺，从而体现出集群能力的鲁棒性和对动态环境的适应性。无人集群系统通过同构或异构无人平台之间协调互补显著提升了任务执行能力，在大面积区域协同搜索、集群任务配合等许多场景中展现出巨大的应用潜力。

面对无人集群应用场景普遍存在的高动态、不确定等特点，集群系统需要更强大的自主性、鲁棒性和智能性，从而完成任务目标。基于深度强化学习（Deep Reinforcement Learning，DRL）的分布式智能协同控制方法，能够充分满足无人集群系统在复杂环境下对控制算法自主性、鲁棒性和智能性的要求。深度强化学习算法能够使无人机或机器人在无人为干预或少量干预的情况下，通过与环境的交互进行自主学习，根据环境的反馈和奖励信号，不断调整控制策略，优化算法性能。深度强化学习算法能够处理高维、复杂的状态和动作空间，适应各种实际应用中的复杂环境，通过分层表示和高级特征提取，对环境中的重要信息进行抽象和理解，从而生成合适的控制策略来应对多样化的环境。任务环境发生变化时，深度强化学习算法能够通过自主学习和调整策略来适应环境；遇到不可预见的情况或故障时，深度强化学习算法可以协调无人集群中的其他无人机共享信息、提供补充和支持，以保证整个集群的任务完成。深度强化学习是实现无人集群智能控制的重要方法。

1 深度强化学习方法介绍

1.1 深度强化学习原理

深度强化学习是一种结合了深度学习（Deep Learning，DL）和强化学习（Reinforcement Learning，RL）的机器学习方法。它以强化学习为基础，具有自主学习和决策的能力。智能体在一个未知的环境中，发出动作作用于环境，并接收环境对动作的奖励，然后根据奖励不断更新产生动作的策略函数，通过与环境不断交互、反复试错的方式，学习在给定环境中的最优策略，使智能体从环境中获得最大化的累计奖赏。但是当给定的环境愈发复杂，智能体分析决策的逻辑就越复杂，策略函数实现起来就越困难，引入深度学习可以解决这一问题。深度学习利用深度神经网络处理复杂环境产生的大量且高维的奖励参数，提取特征，拟合从奖励到决策的推理过程。如果用深度神经网络代替智能体的策略函数，就能够解决强化学习无法处理复杂环境的难题，形成深度强化学习的机器学习方法。

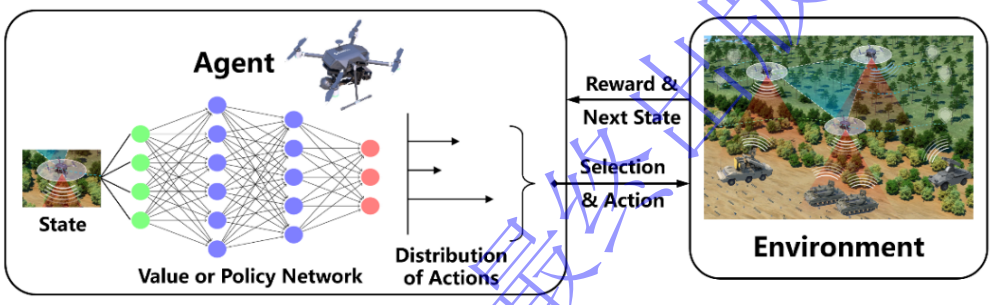


图 1　深度强化学习模型
Fig1　Deep Reinforcement Learning diagram

以无人机作为智能体，其深度强化学习模型如 **Error! Reference source not found.**所示。以强化学习在交互中学习的思想为基础，利用深度神经网络强大的拟合能力逼近难以直接表示的价值或策略函数，赋予智能体适应复杂环境的能力，根据环境奖励优化自身动作策略，得到更高的累计奖励。深度强化学习可以看作强化学习方法与深度神经网络融合的分支，许多概念和原理都与深度学习、强化学习相通。

1.1.1 强化学习

马尔可夫决策过程（Markov Decision Process, MDP）[20]是强化学习的重要基础，MDP 模型如图 2。强化学习的框架建立在智能体-环境系统具有马尔可夫性的基础上，即当前动作后产生的状态、收益等只与当前状态、动作有关，与过去的历史状态无关。在与历史状态有关的场景，如针对棋类游戏的强化学习中，也可以把需要的历史状态封装进当前状态中从而满足马尔可夫性。

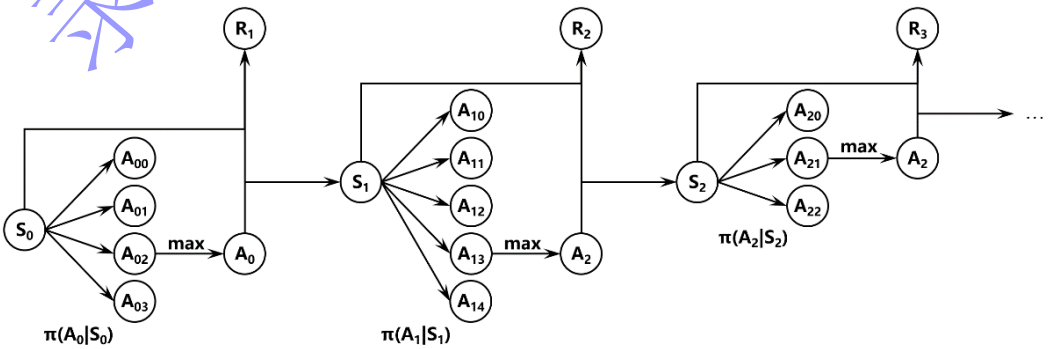


图 2　马尔可夫决策过程模型
Fig2　Model of Markov Decision Process

强化学习作为一种通用的学习框架，只要智能体需要面对多种选择，一般就可以应用。在强化学习模型中，智能体（Agent）处于给定的环境（Environment）但并不预先了解环境信息，智能体根据当前状态（State）和自身策略（Policy）对环境施加动作改变环境，并接收到对应的正向或负向奖励（Reward），同时更新到下一状态，在不断交互和一轮轮尝试中更新策略，寻求全局奖励最大化。依靠智能体通过与环境交互、迭代学习并取得最大期望奖励的特点，基于强化学习的方法在机器人技术[21]、资源分配[26]等多种任务中均取得了众多成果。在无人集群领域，我们有可以构造能够应对复杂环境的智能体，并以集群机器人为载体，实现无人集群系统在复杂现实环境中的智能控制，如图 3 可以构建以无人机集群为载体的多智能体强化学习架构[29]。

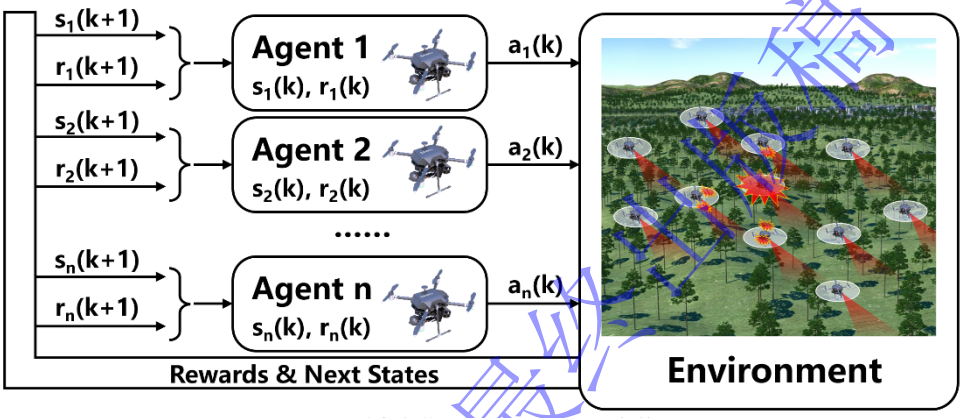


图 3　多智能体强化学习（MARL）架构
Fig3　Architecture of Multi-Agent Reinforcement Learning (MARL)

在强化学习任务中，智能体并不事先了解环境信息，而是通过对环境施加动作、接收奖励从而学习环境、总结策略，因此在训练过程中，智能体需要识别环境特征。环境特征的选取和标记一般由系统的设计者根据自身对任务的了解人工制作，然而人工标记和分类的方法很难处理视觉、语音等复杂的高维感官信息，但是面对实际问题时我们不可避免地要接触和处理这类信息，同时人在处理信息时下意识的忽略和偏见也会影响人工标注的准确性。强化学习效果的优劣很大程度上取决于特征标注的准确性，因此人工标注不再能满足强化学习处理现实问题的需要，环境感知、特征提取的工作应当由智能体自主完成，依据机器标准自成体系地提取和分析环境特征，从而提高智能体感知环境的准确性，智能体依此做出的响应和交互也更加稳定。由于深度学习的高级特征提取和表达能力，引入深度神经网络，智能体能够综合利用视觉、语音等复杂高维感官信息，更好地适应复杂环境，同时进一步提高了智能体对理解环境和做出响应的稳定性。

1.1.2 深度学习

深度学习在复杂数据处理和特征学习方面具有重要地位，是处理高维数据并提取判别性信息的最佳解决方案之一。许多其他算法需要依据领域专家的知识手工完成特征标注，费时费力，而深度学习算法具有从数据中自主提取特征的能力，不需要人总结的知识，并且有效避免了人类经验的不全面和偏见，推动了人工智能的发展。

深度神经网络是深度学习的结构基础，如图 4。深度学习尝试使用有监督或无监督学习算法基于深度神经网络对数据进行高级抽象建模，从而在多个抽象层级归纳和学习。深度学习模型包含多层结构，如自动编码器、受限玻尔兹曼机和卷积层等。在训练期间，原始数据输入到多层网络中，每一层的输出是该层输入的非线性特征变换，并用作网络下一层的输入，然后把最后一层的输出导入限制分类器等其他结构。例如深度神经网络用于图像处理时，在第一层输入图像像素信息，提取并输出图像边缘信息，第二层尝试对边缘进行组合构成一定特征，第三层对特征之间进行组合，并寻找目标模型。深度学习在图像处理中的应用充分展示了它从原始信息中抽象、分离和组合信息特征的分层学习能力，使深度学习成为一种优秀的特征学习方法。

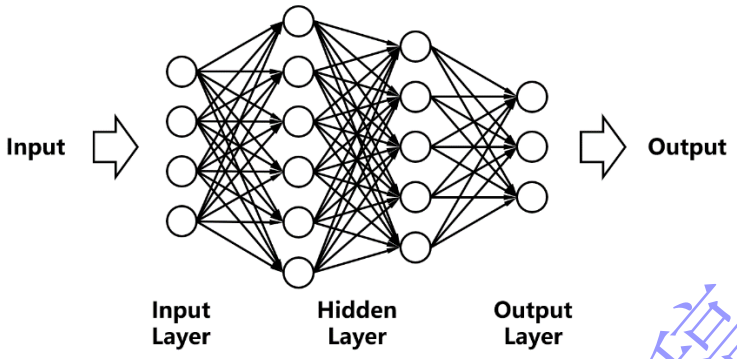


图 4　神经网络，深度学习的结构基础
Fig4　Neural Networks, the structural basis of Deep Learning

能够与强化学习结合的深度神经网络有卷积神经网络（Convolutional Neural Network, CNN）和循环神经网络（Recurrent Neural Network, RNN）等。

1. 卷积神经网络

卷积神经网络是一类有监督的深度特征学习模型，对深度神经网络最早的研究来源于 LeCun Y.于 1989 年利用反向传播网络识别手写字母的研究[30]。随着计算能力的进步，CNN 逐渐在图像处理、语音识别、时序对象识别和检测[31]等领域得到广泛应用。图 5 是一种典型的 CNN 架构。前两层是卷积层和子采样层，卷积层执行卷积操作以创建特征图，它使用局部感受视野和共享权重保证畸变不变性，然后卷积操作的输出通过非线性激活函数输出到子采样层[32]。子采样层执行局部平均或最大池化操作，从而降低进行特征维度，同时保持畸变不变性。卷积层和子采样层的序列可以在特定任务的 CNN 架构中串联使用，最终子采样层的输出送入全连接层，用于分类或识别任务。

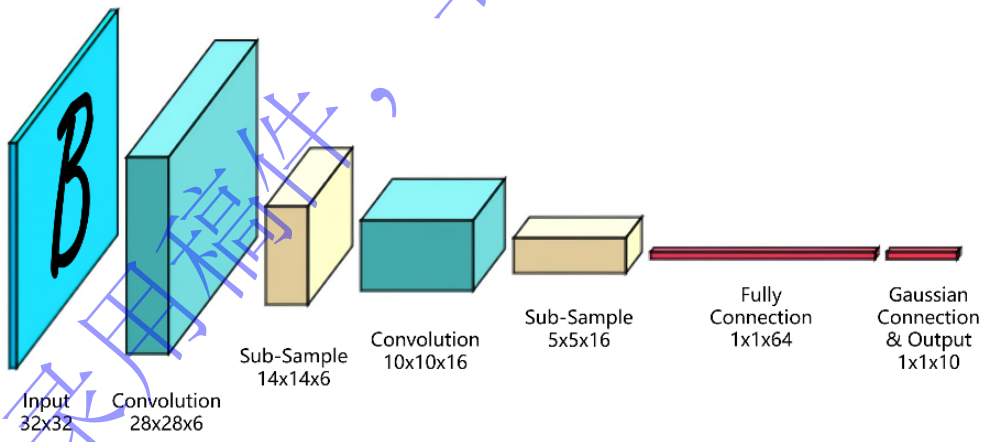


图 5　典型卷积神经网络（CNN）的架构
Fig5　Architecture of a typical convolutional neural network (CNN)

2. 循环神经网络

另一种更适用于时序信息特征提取与学习的算法是循环神经网络。与一般的前馈神经网络（Feedforward Neural Network, FNN）不同的是，循环神经网络具有反馈连接，从而允许网络拥有内部状态，从而网络内部可以设置存储器并保存有关先前输入的信息，使 RNN 能够用于具有时序要求信息的语音识别[33]、自然灾害预测[34]等方面。同时长短期记忆机制（Long Short-Time Memory, LSTM）[35]的引入也解决了 RNN 学习时序数据时可能遇到的梯度消失和梯度爆炸问题，从而弥补了丢失或过度依赖长期事件的缺陷。循环神经网络基本结构如图 6。

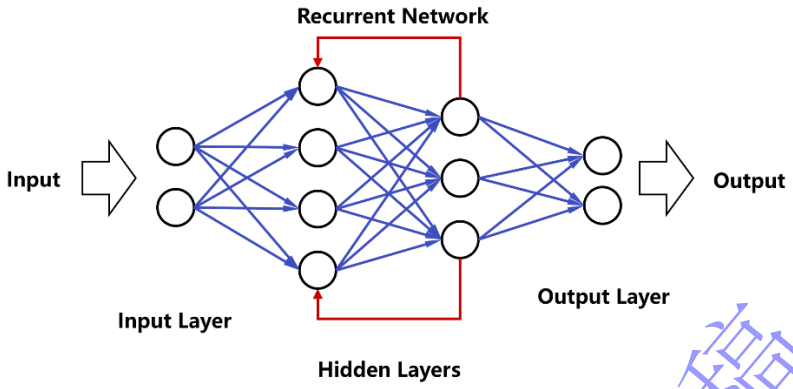


图 6　循环神经网络
Fig6　Recurrent neural network

近年来随着硬件算力的提高，深度学习得到快速发展，深度神经网络展示出强大的抽象和感知能力，在学界获得了巨大的吸引力。深度学习从原始图像、声音等数据中自动提取和归纳特征的能力衍生出众多应用案例，一些新算法和方法的发表也表明深度学习技术可以用于在强化学习问题中准确拟合和表示复杂信息[36]。在强化学习方法中使用深度神经网络，有助于智能体准确感知和把握外部复杂环境信息，催生了多种深度强化学习方法。

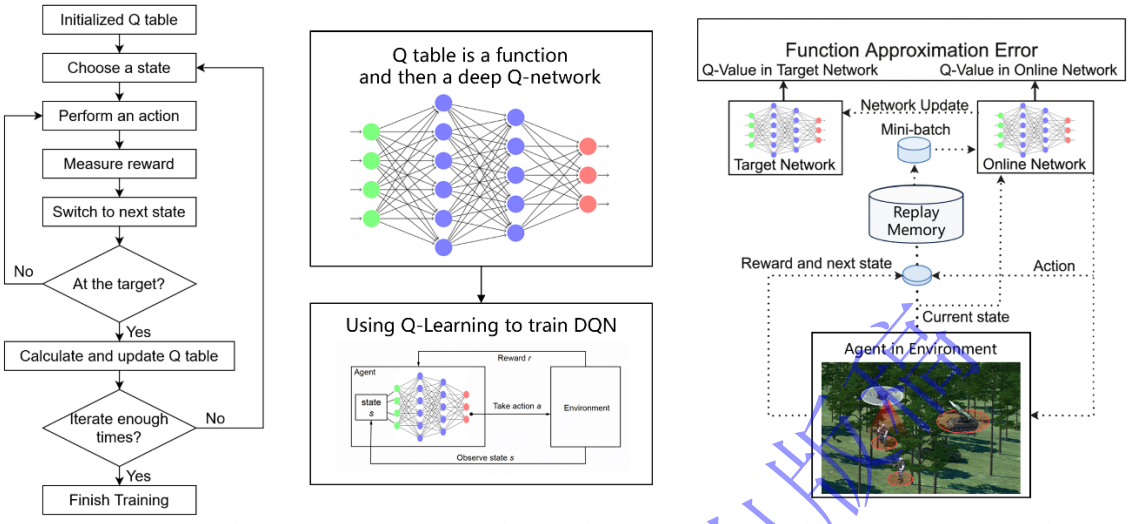
1.2 深度强化学习方法的分类

深度强化学习的根本目的是让智能体学习并生成合适的策略网络，从而在和环境的交互中获得最大全局奖励。围绕这个目的，主要有基于价值（Value-Based）和基于策略（Policy-Based）的两类学习方法。

1.2.1 基于价值的深度强化学习方法

基于价值的 DRL 通过估计智能体动作的价值（Value，同时也是全局奖励的期望）从而做出决策，利用更高的预期价值估计全局奖励期望最大的策略和轨迹，用价值代表策略，用深度神经网络估计价值网络，不断选取最高价值动作，算法思想相较后者更加直接、具体，易于理解，在学界也更先被研究和发表。基于价值的 RL 和 DRL 在发展过程中先后提出三种典型算法结构如图 7。

早在 20 世纪 90 年代，神经网络用于强化学习表示价值网络的研究就有了初步成果，以 TD-Gammon[41]为例，Tesauro 将神经网络与强化学习相结合，在棋类游戏 backgammon 上击败了众多顶尖棋手，但深度强化学习在随后的发展中并不一帆风顺，直到 2013 年深度 Q 网络算法（Deep Q-Learning Network, DQN）[42]问世，并在打砖块游戏 Atari 中表现优秀，以 DQN 算法为代表的基于价值的 DRL 算法迎来快速发展。针对强化学习不稳定和发散等问题，各大研究工作者对 DQN 算法做了许多改进：使用经验回放和目标网络方法，使基于深度神经网络的近似动作值函数趋于稳定；使用端到端的方法，利用卷积神经网络（CNN）把原始图片和游戏得分作为输入，减少模型对领域知识的需求；训练可变的网络，使网络在多种任务中表现有效，并且能够超越人类专业选手。基于 DQN 算法也衍生出许多其他算法和应用，如 Van Hasselt 等[43]提出的 Double DQN 算法，使用 Q Network 和 Target Q Network 两个价值网络分别用于执行动作策略和评估价值，用于解决目标价值过高估计的问题，并在自动驾驶汽车领域得到许多应用[44]；Schaul 等[45]改进了 DQN 对经验均匀采样的策略，提出了优先经验回放策略，利用时序差分误差（Temporal Difference Error, TD Error）对经验的重要性进行衡量，对重要性靠前的经验回放多次，进而提高学习效率。



A)Q-Learning Approach　　B) Deep Q-Network Approach　　C) Double Deep Q-Network Approach

图 7　基于价值的三种强化学习和深度强化学习方法

Fig7　Three value-based RL or DRL Approaches

1.2.2 基于策略的深度强化学习方法

相比基于价值，基于策略的方法在理解上较为抽象，但在逻辑上更加直接。策略学习使用一个神经网络拟合策略函数，称为策略网络（Policy　Network），并使用策略梯度（Policy Gradient）等方法训练策略网络。在使用策略梯度算法的过程中，状态价值函数难以求解的问题引申出 Reinforce 和 Actor-Critic 两种处理方法。

Reinforce 算法[46]在智能体完成一遍学习过程后使用累计奖励作为状态价值的蒙特卡洛近似，从而得到价值的无偏估计。而策另一种方法是构造新的价值网络，使用神经网络拟合价值函数，这样就在强化学习任务中引入了两个神经网络，综合利用基于价值和基于策略的深度强化学习方法，并发展成为双网络架构的 Actor-Critic 算法[47]。确定性策略梯度算法（Deterministic Policy Gradient，DPG）[48]、深度确定性策略梯度算法（Deep Deterministic Policy Gradient，DDPG）[49]和著名的应用案例 AlphaGo[50]都采用了 Actor-Critic 的 DRL 学习方法。以 AlphaGo 为例，为了训练网络，AlphaGo 首先使用基于人类游戏的监督学习，然后使用自我博弈的强化学习，在实战中它还利用蒙特卡洛树搜索简化了估计对手未来下棋策略的复杂度，大大缩短搜索和响应时间。在后继的 AlphaGo　Zero[51]中并没有引入人类棋谱中学习的知识，跳过了初始阶段的监督学习，最终 AlphaGo　Zero 在围棋上完全胜过了初代 AlphaGo，这表明了人为经验不完全适合机器学习，去除人为经验的深度强化学习在性能和稳定性上表现更好。再后来 AlphaZero[52]发布，使人工智能具备了学习和适应多种棋类游戏的能力，并在 Chess、Shogi、Go 多个棋类游戏中在 700 万步训练以内战胜了 Stockfish、Elmo 和 AlphaGo Zero 等专注于单个棋种训练的人工智能模型，体现了 DRL 方法的强泛用性。

2 无人集群的深度强化学习控制

当前无人集群控制领域面临许多难题，随着集群规模的增加，集群系统控制方法的设计难度大幅度上升：集群规模变化会对集群任务和行为产生影响，因此控制方法必须满足集群扩充、缩小的功能要求；高度自治化的集群系统需要支持多样化的任务，因此集群编队等指标发生变化更为频繁；无人集群一般采用分布式控制架构，在集群演化过程中产生的整体变换效果在节点数量、结构、功能和行为目标等方面难以进行预测分析；此外还有单个或部分节点对环境感知不完全的挑战。针对这些难题，许多研究开始尝试将无人集群系统与深度强化学习相结合，着眼于智能体集群在环境中不断试错、求解最大化累计奖励的行为逻辑，为解决无人集群系统复杂、大规模的现实应用难题提出许多处理方法。

2.1 无人集群系统的控制需求

无人集群系统通常可选的逻辑架构有集中式和分布式两种，如图 8 分别构成放射状和网状的两种通信结构。集中式架构无人集群系统的决策和控制中心集中在单一的节点上，系统决策过程较为简单。集中式控制可以更容易地协调集群内布的行为，优化全局性能，而且系统的设计和开发更为简单，便于实施和维护。然而集中式架构有着高度依赖中心节点的缺点，单点故障就可能带来系统崩溃的风险，并且随着系统规模扩大，控制过程的计算和通信负荷增大，对中心节点计算性能要求很高，可能导致系统性能下降。分布式架构的无人集群系统则将决策和控制能力分散到每个节点中，每个节点都能够自主行动，利于系统应对故障和环境变化，而且分布式架构便于实现横向扩展，支持系统规模变化，具有灵活性和鲁棒性。然而分布式架构的系统节点之间协调、通信更加复杂，需要适当的算法和协议来保证系统内合作的稳定，也使得系统设计和调试可能更加困难。

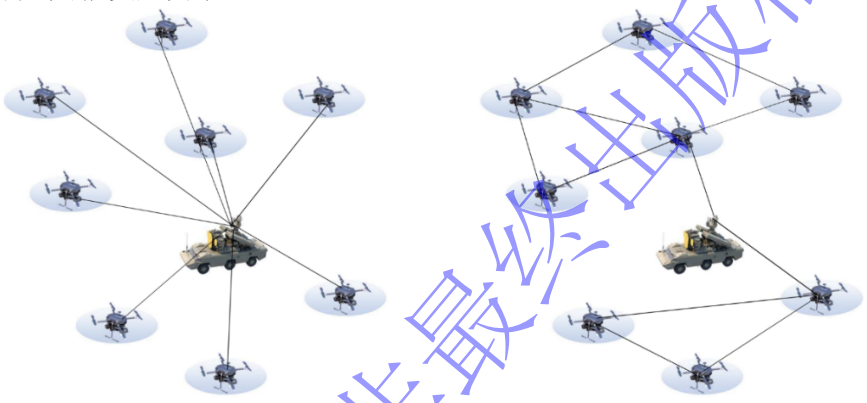


图 8　集中式架构和分布式架构
Fig8　Centralized & Decentralized

无人集群发展越来越注重各无人节点之间的控制、协调、合作、竞争等交互，因此无人集群系统研究更多基于分布式架构，要求各无人节点具有一定智能和自主性，也称作多智能体系统。无人集群系统的控制有着编队控制、跟踪控制、集群避障等多方面的需求。

编队控制（Formation）的目标是驱动多智能体编队维持和变换一定的几何特征，例如节点间的距离、角度、编队形状等，是对无人集群系统进行路径规划的基础，主要在于解决各无人节点之间的协调控制问题。主要分为两类场景：其一，无人节点能够较为准确轻松地获得自身全局位置，此时可以基于位置坐标信息进行系统编队，通常不需要无人节点之间的频繁通信。但这类方法一般需要强大的上位机来满足大量频繁的定位需求，对通信要求较高，而且任务范围不能太大，不能满足大多数集群系统应用场景对大面积的需求，也不能充分发挥系统分布式、弱中心的优势。其二，为了减小对全局位置的需要，发挥分布式架构优势，可以基于无人节点间的几何特征进行系统编队控制，又分为基于距离、基于重心坐标、基于方位的三种基本编队控制方法和他们的混合方法。

跟踪控制（Tracking）的目标是控制无人节点沿给定轨迹、路径移动或维持特定目标在一定视野范围内。轨迹（Trajectory）的几何与时间约束直接符合该节点的运动学和动力学特性，一般是由本节点或模拟节点行驶生成的，可以直接用于实机跟踪，是时刻-位置-姿态三者关系的密集序列或连续函数。无人节点贴近并沿着给定轨迹行驶，而且所用时长趋近参考轨迹的时长，有较为严格的时间要求。路径跟踪控制的效果与轨迹跟踪相近，不同的是路径（Path）定义较为宽泛，可以是路点、折线、样条等。路径跟踪一般对时序要求较为宽松，只关注在部分且离散航点的时刻-位置-姿态关系，在这些点之间可以根据无人节点的动力学约束自主生成和选择轨迹。目标跟踪可以通过设置目标当前位置为航点，并不断更新和切换航点从而达到目的；也可以使用相对位置方法，维持目标在视觉画面特定区域。跟踪目标运动的特点使得跟踪控制具有更高的难度，路径跟踪和轨迹跟踪可以看作目标跟踪一种较为简单的分支。

无人集群的避障控制（Obstacle Avoidance）要求无人集群在执行确切任务的同时能够有效避免节点与节点之间、节点与环境之间的碰撞，面临着外部环境复杂且未知、集群分布拥挤、对通信要求较高等挑战。在线的动态避障是无人集群系统广泛应用于林场搜救、农田管理、货物运送等场景[53]不可或缺的功能。目前在无人车集群领域已经有许多研究工作提出了多种有效的避障办法[56]，然而无人机相比无人车拥有更复杂的三维移动空间和欠驱动的六自由度，为无人机的灵活避障造成了许多困难。而且集群避障任务很难适配集中式算法，强大的集中式算法虽然拥有掌控全局信息的中央服务器[59]，便于进行全局动态规划，但严重依赖低延迟且频繁的通信，这不仅占用大量通信资源，而且与环境动态变化的局部性相冲突，导致系统稳定性较差，采用分布式架构、减少指挥中心对集群底层动作的干涉、赋予无人节点快速自主的避障能力是实现无人集群避障控制的发展趋势。

2.2 基于深度强化学习的无人集群控制方法

近年来无人机技术和应用得到了较快发展，随着人们对无人机技术期望的提升和应用领域的推广，单架无人机执行任务不再能满足全部的需要，无人机集群的发展受到了越来越大的关注。无人集群的需求因为任务场景的复杂化而产生，但集群化的无人技术并没有降低任务环境的复杂程度，反而因为多节点进一步造成了集群-场景系统的复杂，深度强化学习因为面对复杂环境和强大感知和处理能力，逐渐成为解决集群难题的新选择，在无人集群编队控制、跟踪控制、集群避障等领域产生众多研究成果。

2.2.1 编队控制方法

无人机编队飞行是研究集群控制的重要内容，也是降低和拆分集群控制复杂度的可行方法之一。构成稳定编队的无人机集群可以看作一架更大的单体无人机，将集群飞行的问题简化成为单机飞行问题。学界采用了多种拓扑结构研究编队飞行。向心集群结构（Flock Centering）于 1987 年由 Reynolds[61]提出，规定了集群协同的三条原则：避免相碰、速度匹配、尽量相互临近，构建了类似鸟群的无人集群协同飞行典型拓扑结构。La H.M. [62]、Jia Y. [63]、Lee G.T. [64]等人都在自己的研究中使用了这样的拓扑结构。领航集群结构（Leader–Follower Flocking）则是把导航规划任务交由领航机，其余伴飞无人机只须与领航机保持较为固定的相对位置，代表应用有 Quintero S.A. [65]、Hung ShaoMing[66] 、Hao Chen[67]等人的文章。Wang C.等[68]则采用邻居集群（Neighbors Flocking）的拓扑结构，每架无人机只与最邻近的几个节点通信，保持邻居节点之间的相对距离和角度，从而维持编队形制。

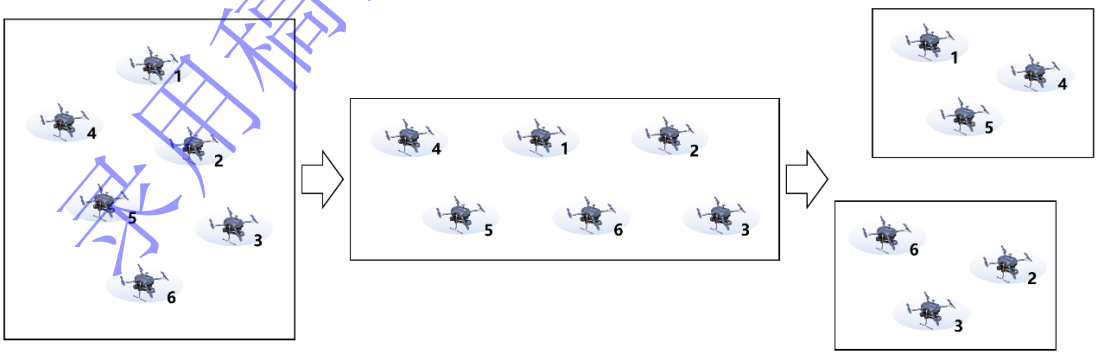


图 9　多架无人机的编队、队形变换、分组等集群任务
Fig9　Swarm tasks of UAVs in formation, transformation, grouping, etc

除了维持编队的基本结构和队形，编队控制还涵盖了编队变换的任务，如图 9 所示。编队变换控制旨在使无人集群能够快速、协调地完成队形变换、分组等操作。队形变换涉及到集群中每个无人机相互之间的位置和姿态的调整，以实现不同的队形。无人集群可以根据任务需求在不同的队形之间切换，如线性队形、V 字队形、菱形队形等，以适应不同的任务环境和场景。队形变换需要通过精确的控制算法和通信协议来确保无人机之间的协同行动和避免碰撞。集群分组则是设计无人机编队的拆分与合并，无人集群通常由多个小组或单元构成，每个组内的无

人机具有相似或相配合的功能，任务变化后需要将无人机重新分组以适应新的需求或应对变化的环境条件。分组管理需要综合考虑多个因素，如通信、协作和资源分配等，以确保集群整体的灵活性和效能。

无人集群编队控制的深度强化学习系统可以采用集中式训练和分布式训练两类方法。集中式强化学习的智能体基于从所有无人机处收集而来的经验，训练公用的集群策略，而在实际飞行中，每架无人机根据其周围的障碍物等本地局部环境信息单独行动。Yan 等人[69]使用近端策略优化算法（Proximal Policy Optimization，PPO）训练集中式公用集群控制策略，其中每架无人机都尽可能靠近集群中心，并根据每架无人机的本地局部环境信息分散执行避障动作。Hung 和 Givigi[70]则采用了领航结构训练无人机导航到目的地，同时避开障碍，并训练了考虑伴飞机与领航机相对位姿的公用策略，完成了无人集群伴飞避障到达目的地的任务。

由于邻居集群结构中邻近节点运动状态不完全同步，导致不同节点的策略需求存在差异，因此针对某个节点集中式训练得到的模型无法不能很好的直接迁移到其他节点上，分布式训练可以克服这一难题[71]。分布式训练中每架无人机都有其对应的智能体负责找到符合无人机自身的最佳集群策略，奖励函数则设计成鼓励无人机维持与其他节点的距离不变且飞行方向一致的形式，也可以根据编队目标定制更丰富的奖励依据。分布式训练方法的问题则是对通信的要求高，单架无人机无法获取完整的集群状态信息，而针对集群优化的强化学习算法需要完整的信息，信息传递失败会导致节点离群或者集群的分裂，这一问题还有待解决。

2.2.2 跟踪控制方法

以无人机集群为多智能体的现实载体，近年来深度强化学习算法在目标跟踪控制，尤其是如图 10 的动态目标跟踪或追捕问题的研究产生了许多成果。符小卫等[72]在二维无约束环境下对多无人机协同追捕策略进行了研究，通过采用多智能体深度强化学习（Multi-Agent Deep Reinforcement Learning，MADRL）方法对追捕无人机进行训练，实现了多无人机对逃逸目标的智能协同追捕。然而上述文献没有考虑实际作战环境中可能存在的障碍物、禁飞区等威胁因素，应用场景局限于二维平面且相对简单，缺乏对复杂任务场景下无人机集群机动策略的研究。基于 DRL 方法的复杂威胁场景单无人机跟踪问题也发表了不少研究结果，如文献[73]针对动态环境中无人机目标跟踪与避障机动控制问题，基于改进 DDPG 算法和迁移学习提出了在线路径规划方法，实现了无人机对地面动目标的自适应跟踪；文献[74]将障碍物约束引入三维作战环境，并采用近端策略优化算法对单无人机自主跟踪与避障的机动策略进行了设计。但单无人机目标跟踪仍然面临着机动避障导致目标丢失、无人机损毁导致跟踪失败等难以避免的容错性问题，这些问题有望通过基于 MADRL 的无人机集群智能协同的方式解决。

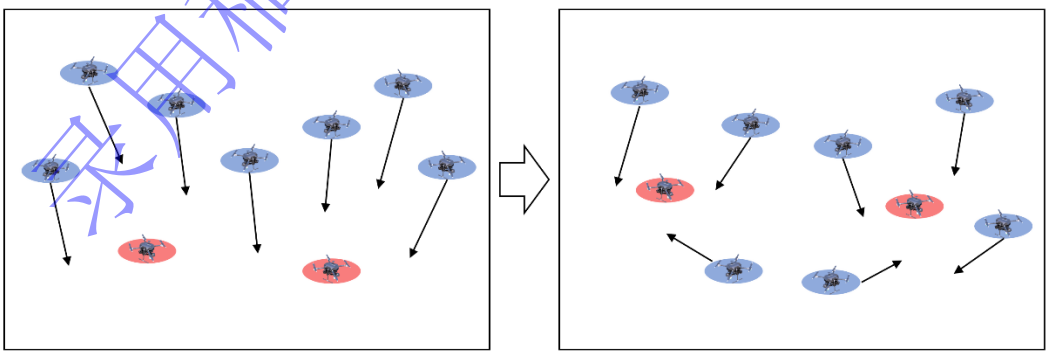


图 10 跟踪控制方法在无人集群协同追捕场景的应用

Fig10 Application of tracking control approaches in UAV swarm cooperative pursuit

综合以上问题，文超等人[75]基于 MADRL 对集群智能协同控制方法展开研究，并落脚于无人机集群协同目标跟踪这一具体问题，提出一种基于解耦型 MADDPG（Decomposed Multi-Agent Deep Deterministic Policy Gradient，DE-MADDPG）的无人机集群自主跟踪与避障机动控制方法，提高了 MADDPG（Multi-Agent Deep Deterministic Policy Gradient）算法的准确性和实时性，满足了复杂空域环境下无人机集群目标自主跟踪任务的需求。面对无障碍或少障碍但

跟踪目标速度更快的特定场景，Cristino de Souza[76]采用双延迟深度确定性策略梯度算法（Twin Delayed Deep Deterministic Policy Gradient Algorithm, TD3）[77]，并融合课程学习训练方法提高训练效率，提出了一种分布式经验共享的 DRL 方法，并在仿真模拟中验证了算法的优秀性能，在无人机定高演示中证明了直接从仿真到实机迁移的可行性。

2.2.3 集群避障方法

分布式架构的集群避障方法主要通过无人机自身的传感器和有限的计算资源控制无人机躲避障碍，而且分布式架构避免了每架无人机与中央服务器频繁的信息交换，节省了大量与中心节点之间的通信资源，降低了对中央服务器计算和调度能力的需求，取而代知的是要求无人机之间的通信，无人机之间需要共享 GPS 位置、预期路径等信息，如高飞团队发表的 Ego-Swarm [78]、McGuire[79]等人无人集群关于探索未知区域的方法。

基于深度强化学习的集群避障将避障问题描述为马尔可夫决策过程，利用传感器输入使无人机直接学习避障策略，而且基于 DRL 的方法不依赖较高的传感器精度，可以通过强大的学习能力实现无地图的在线避障。如 Faust 等人[80]通过为无人机提供全局环境信息来帮助机器人学习避障策略；Leiva 等人[81]使用距离传感器生成的二维点云训练无地图避障策略；Ma 等人[82]提出了一种基于视觉的无人机防撞框架，通过监督学习生成显著性图，然后通过强化学习训练控制策略以避开三维工作空间中的障碍物。

对无人集群深度强化学习避障控制的研究大多对无人机通信有较高要求，也有一部分研究意在降低无人集群避障的通信需求甚至不需要通信[83]，从而提高无人集群在较差通信环境下执行任务的鲁棒性。P. Long 和 T. Fan 等人在文献[84]中通过二维激光扫描和惯性测量的原始传感器测量优化了二维工作空间中完全分散的防撞策略，并在文献[85]对之前的工作做了进一步扩展，在无人机集群的定高任务场景有较好效果，但无法发挥无人机的三维空间运动的优势。H. Huang 等人[86]在前述研究成果的基础上利用 DRL 方法研究了基于视觉的三维工作空间无人集群防撞方法，该方法无需机间通信，且在仿真实验中表现良好。

3 无人集群深度强化学习控制的前景与挑战

近年来无人集群系统领域涌现出一大批已落地应用或在研的优秀成果。例如，无人机编队协同在农业种植业有了新的突破，通过多机协同执行无人机播种、施药治虫等任务，提高大面积农田上的作业效率；在智能仓储系统中，无人机与自动引导小车组成多机种的无人集群系统，相互配合完成货物搬运和清点工作；国防领域中，无人集群相对于单一装备具有不对称优势，有望在协同探测、协同攻击、干扰压制等作战任务中发挥巨大作用，是各国军事科研的重点。大气污染监测、园区无人物流、火灾救援等也是无人集群系统致力发展和创新应用的任务场景。面向复杂场景的无人集群系统具有广阔的应用前景，这要求我们在无人集群协同控制领域开展广泛和持续性的基础理论和关键技术研究。

国际范围内学术界和产业界在复杂环境下无人集群系统自主协同控制相关理论和技术方面取得了一系列突破，并在特定场景下开展了小规模无人集群技术验证试验，但是仍然无法应对高动态、不确定、资源受限等复杂环境带来的技术挑战，深度强化学习有希望成为解决以上诸多困难的有效方法。然而深度强化学习应用于无人集群系统依然面临从虚拟场景转向与现实交互的巨大困难，虚拟的游戏等可以低成本且不间断的反复试错优化模型，一旦面临与实机结合的现实场景，不断试错的做法在金钱成本、时间成本、人力成本等多方面维度都无法实现。当前解决这一问题的思路主要有三种：

1. 构建与现实世界更加逼真的仿真环境

这类方法致力于通过提高交互数据与真实物理世界交互数据的相似性，减小从仿真迁移到真实中的差异。只要虚拟环境足够真实，就可以避免模型落地现实的绝大多数问题。如微软公司开发的 AirSim，专注于空中无人机的强化学习仿真环境，其物理引擎利用了虚幻引擎渲染，模拟程度非常高。

2. 从仿真环境训练迁移到真实环境训练

在构建与真实物理环境接近的仿真环境下，训练成熟一个模型，并保存参数；然后模型部署后，通过与现实世界中的物理环境交互的真实数据，在原来模型的基础上继续训练优化，大大减小了在真实环境中从零训练的成本。

3. 适度引入人类知识减少无用训练

如同上文提到的 AlphaGo，深度强化学习的缺点很大程度上是由于初始的稀疏奖励使得智能体需要探索足够多的状态后，才能获得较好的奖励反馈，在这个初始阶段产生了大量无用的样本。模仿学习让人类首先为智能体构建若干专家示教轨迹，来加速智能体初期的学习速度，避免一些无用的探索。然而模仿学习也存在“人类知识有害”的缺点，在 AlphaGo 与 AlphaGoZero 对弈中，0:100 的胜率就是其中体现，“如何适当地引入人类知识和经验”的问题需要学术界进一步研究。

深度强化学习方法与无人集群控制相结合的领域仍有许多问题有待研究。

4 参考文献

[1] Zhang D N, Cheng Y, Lin Q, et al. Key technologies and development trends of UAV swarm combat. *China New Communications*, 2022, 24(04): 56
(张丹凝, 程岳, 林清等. 无人机集群作战关键技术及发展趋势. 中国新通信, 2022, 24(04): 56)

[2] SHEN Bo, WU Wenliang, YANG Gang, et al. Intelligent evaluation model and method of unmanned cluster system based on swarm OODA. *Journal of Aeronautics*: 2023, 44(14): 263
(沈博, 武文亮, 杨刚等. 基于群体 OODA 的无人集群系统智能评价模型及方法. 航空学报, 2023, 44(14): 263)

[3] Wang L B, Wang M Y, Zhou S Q, et al. Swarm intelligent ensemble tracking control of heterogeneous cluster system considering unknown input. *Science in China: Technical Science*, 2023, 53(02): 291-306
(王林波, 王蒙一, 周思全等. 考虑未知输入的异构集群系统群体智能合围跟踪控制. 中国科学: 技术科学, 2023, 53(02): 291)

[4] Yin G W, Dong D C, You D X, et al. Swarm Intelligence Research: From Bio-inspired Single-population Swarm Intelligence to Human-machine Hybrid Swarm Intelligence. *Machine Intelligence Research*, 2023, 20(1)

[5] Jonathan P, D. O D. Strategies for Scaleable Communication and Coordination in Multi-Agent (UAV) Systems. *Aerospace*, 2022, 9(9)

[6] V. V R. Multi-Agent Logics with Dynamic Accessibly Relations, Projective Unifiers. *Algebra and Logic*, 2022, 61(1)

[7] Yu Z, Zhiyu M, Feifei G, et al. UAV-Enabled Secure Communications by Multi-Agent Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology*, 2020, 69(10)

[8] Difeng H, J. L. V G, Tao W, et al. Multi-agent robotic system (MARS) for UAV-UGV path planning and automatic sensory data collection in cluttered environments. *Building and Environment*, 2022, 221

[9] Zou L Y, Zhang M M, Bai J R, et al. Review of UAV swarm combat modeling and simulation research. *Tactical Missile Technology*, 2021(03): 98
(邹立岩, 张明智, 柏俊汝等. 无人机集群作战建模与仿真研究综述. 战术导弹技术, 2021(03): 98)

[10] Yi S, Huang Q, Yang P F. Function and architecture design of combat simulation system of intelligent unmanned cluster system. *Command Control and Simulation*, 2020, 42 (05): 65
(伊山, 黄谦, 杨鹏飞. 智能无人集群体系作战仿真系统功能与架构设计. 指挥控制与仿真, 2020, 42(05): 65)

[11] Wang Z H, Leng S P, Xiong K. Resource allocation strategy of multi-agent for cooperative sensing of UAV swarm. *Chinese Journal on Internet of Things*, 2023, 7(01): 18
(王志宏, 冷甦鹏, 熊凯. 面向无人机集群协同感知的多智能体资源分配策略. 物联网学报, 2023, 7(01): 18)

[12] Ma Y, Chang T, Zhai M. Rule-Based Unmanned Swarm Collaborative Control Method. Advances in Computer, *Signals and Systems,* 2021, 5(2)
[13] Zhao L, Zhang Y F, Yao M X, et al. Development and prospect of UAV swarm collaborative technology. *Radio Engineering*, 2021, 51(08): 823
(赵林, 张宇飞, 姚明昕等. 无人机集群协同技术发展与展望. 无线电工程, 2021, 51(08): 823)
[14] Tianhao W, Shiqian M, Na X, et al. Secondary Voltage Collaborative Control of Distributed Energy System via Multi-Agent Reinforcement Learning. *Energies*, 2022, 15(19)
[15] Zihua C, Chuanli W, Jingzhao L, et al. Multi-agent collaborative control parameter prediction for intelligent precision loading. *Applied Intelligence*, 2022, 52(14)
[16] Siyu G, Xin W. Neural-Network-Based Collaborative Control for Continuous Unknown Non-linear Systems. *Discrete Dynamics in Nature and Society*, 2021, 2021
[17] Yin H C, Cui H L, Yu Z W, et al. Task allocation method for collaborative perception. *Software Guide*, 2020, 19(04): 14
(尹厚淳, 崔禾磊, 於志文等. 面向协同感知的任务分配方法. 软件导刊, 2020, 19(04): 14)
[18] Huang J C, Zhou D Y. Design and analysis of UAV cooperative combat effectiveness evaluation index system. *Journal of Xi'an Technological University*, 2020, 40(01): 38
(黄吉传, 周德云. 无人机协同作战效能评估指标体系设计与分析. 西安工业大学学报, 2020, 40(01): 38)
[19] Duan H B, Qiu H X, Chen L, et al. Research prospect of UAV autonomous swarming technology. *Science & Technology Review*, 2018, 36(21): 90
(段海滨, 邱华鑫, 陈琳等. 无人机自主集群技术研究展望. 科技导报, 2018, 36(21): 90)
[20] Bellman R. A Markovian decision process. Journal of mathematics and mechanics, 1957: 679-684.
[21] Mohanty S, Elias D S. Control and Coordination of a SWARM of Unmanned Surface Vehicles using Deep Reinforcement Learning in ROS. arXiv preprint arXiv: 2304.08189, 2023
[22] Huang H, Loquercio A, Kumar A, et al. More Than an Arm: Using a Manipulator as a Tail for Enhanced Stability in Legged Locomotion. arXiv preprint arXiv: 2305.01648, 2023
[23] Abouheaf M, Boase D, Gueaieb W, et al. Real-time measurement-driven reinforcement learning control approach for uncertain nonlinear systems. *Engineering Applications of Artificial Intelligence*, 2023, 122: 106029
[24] Herzog A, Rao K, Hausman K, et al. Deep RL at Scale: Sorting Waste in Office Buildings with a Fleet of Mobile Manipulators. arXiv preprint arXiv: 2305.03270, 2023
[25] Miera P, Szolc H, Kryjak T. LiDAR-based drone navigation with reinforcement learning. arXiv preprint arXiv: 2307.14313, 2023
[26] Vengerov D. A reinforcement learning approach to dynamic resource allocation. *Engineering Applications of Artificial Intelligence*, 2007, 20(3): 383
[27] Nguyen K K, Duong T Q, Do-Duy T, et al. 3D UAV trajectory and data collection optimisation via deep reinforcement learning. *IEEE Transactions on Communications*, 2022, 70(4): 2358
[28] Tilahun F D, Abebe A T, Kang C G. Multi-agent reinforcement learning for distributed joint communication and computing resource allocation over cell-free massive MIMO-enabled mobile edge computing network. arXiv preprint arXiv: 2201.09057, 2021
[29] Chincoli M, Liotta A. Self-learning power control in wireless sensor networks. Sensors, 2018, 18(2): 375
[30] LeCun Y, Boser B, Denker S J, et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1989, 1(4)
[31] Lecun Y , Bengio Y . Convolutional Networks for Images, Speech, and Time-Series//*The Handbook of Brain Theory and Neural Networks*.1995
[32] Li Y. Deep reinforcement learning: An overview. arXiv preprint arXiv: 1701.07274, 2017
[33] Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: An overview//*2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013: 8599-8603
[34] Moreno M J, Sánchez M J, Espitia E H. Use of computational intelligence techniques to predict flooding in places adjacent to the Magdalena River. *Heliyon*, 2020, 6(9)

[35] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*, 1997, 9(8): 1735
[36] Mattner J, Lange S, Riedmiller M. Learn to swing up and balance a real pole based on raw visual input data//Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part V 19. Springer Berlin Heidelberg, 2012: 126
[37] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning. CoRR, 2013, abs/1312.5602
[38] Volodymyr M, Koray K, David S, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540)
[39] Böhmer W, Springenberg T J, Boedecker J, et al. Autonomous Learning of State Representations for Control: An Emerging Field Aims to Autonomously Learn State Representations for Reinforcement Learning Agents from Their Real-World Sensor Observations. KI - Künstliche Intelligenz, 2015, 29(4)
[40] Levine S, Finn C, Darrell T, et al. End-to-End Training of Deep Visuomotor Policies. CoRR, 2015, abs/1504.00702
[41] Tesauro G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 1994, 6(2): 215
[42] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning. *CoRR*, 2013, abs/1312.5602
[43] Hasselt V H, Guez A, Silver D. Deep Reinforcement Learning with Double Q-learning. *CoRR*, 2015, abs/1509.06461
[44] Hieu N Q, Hoang D T, Niyato D, et al. Transferable deep reinforcement learning framework for autonomous vehicles with joint radar-data communications. *IEEE Transactions on Communications*, 2022, 70(8): 5164
[45] Schaul T, Quan J, Antonoglou I, et al. Prioritized Experience Replay. *CoRR*, 2015, abs/1511.05952
[46] Andrija P, Mladen N, Miloš J, et al. Fair classification via Monte Carlo policy gradient method. *Engineering Applications of Artificial Intelligence*, 2021, 104
[47] Jia Y, Zhou X Y. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *The Journal of Machine Learning Research*, 2022, 23(1): 12603
[48] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//*International conference on machine learning*. Pmlr, 2014: 387
[49] Lillicrap P T, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. *CoRR*, 2015, abs/1509.02971
[50] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529(7587): 484
[51] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. *Nature*, 2017, 550(7676): 354
[52] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 2018, 362(6419): 1140
[53] Tian Y, Liu K, Ok K, et al. Search and rescue under the forest canopy using multiple UAVs. *The International Journal of Robotics Research*, 2020, 39(10-11): 1201
[54] Ju C, Son H I. Modeling and control of heterogeneous agricultural field robots based on Ramadge–Wonham theory. *IEEE Robotics and Automation Letters*, 2019, 5(1): 48
[55] Chen Z, Alonso-Mora J, Bai X, et al. Integrated task assignment and path planning for capacitated multi-agent pickup and delivery. *IEEE Robotics and Automation Letters*, 2021, 6(3): 5816
[56] Tan Q, Fan T, Pan J, et al. Deepmnavigate: Deep reinforced multi-robot navigation unifying local & global collision avoidance//*2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020: 6952
[57] Gopalakrishnan B, Singh A K, Kaushik M, et al. Prvo: Probabilistic reciprocal velocity obstacle for multi robot navigation under uncertainty//*2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017: 1089

[58] Han R, Chen S, Hao Q. Cooperative multi-robot navigation in dynamic environment with deep reinforcement learning//*2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020: 448
[59] Araki B, Strang J, Pohorecky S, et al. Multi-robot path planning for a swarm of robots that can both fly and drive//*2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017: 5575
[60] Collins L, Ghassemi P, Esfahani E T, et al. Scalable coverage path planning of multi-robot teams for monitoring non-convex areas//*2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021: 7393
[61] Reynolds W C. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 1987, 21(4)
[62] La H M, Sheng W. Flocking control of multiple agents in noisy environments//*2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010: 4964
[63] Jia Y, Du J, Zhang W, et al. Three-Dimensional Leaderless Flocking Control of Large-Scale Small Unmanned Aerial Vehicles. *IFAC PapersOnLine*, 2017, 50(1)
[64] Lee G T, Kim C O. Autonomous control of combat unmanned aerial vehicles to evade surface-to-air missiles using deep reinforcement learning. *IEEE Access*, 2020, 8: 226724
[65] Quintero S A P, Collins G E, Hespanha J P. Flocking with fixed-wing UAVs for distributed sensing: A stochastic optimal control approach//*2013 American Control Conference*. IEEE, 2013: 2025
[66] ShaoMing H, N S G. A Q-Learning Approach to Flocking With UAVs in a Stochastic Environment. *IEEE transactions on cybernetics*, 2017, 47(1)
[67] Hao C, Xiangke W, Lincheng S, et al. Formation flight of fixed-wing UAV swarms: A group-based hierarchical approach. Chinese Journal of Aeronautics, 2021, 34(2): 504
[68] Wang C, Wang J, Zhang X. A deep reinforcement learning approach to flocking and navigation of uavs in large-scale complex environments//*2018 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE, 2018: 1228
[69] Yan P, Bai C, Zheng H, et al. Flocking control of uav swarms with deep reinforcement leaming approach//2020 3rd International Conference on Unmanned Systems (ICUS). IEEE, 2020: 592
[70] Hung S M, Givigi S N. A Q-learning approach to flocking with UAVs in a stochastic environment. *IEEE transactions on cybernetics*, 2016, 47(1): 186
[71] Fadi A, Katarina G. Autonomous Unmanned Aerial Vehicle navigation using Reinforcement Learning: A systematic review. *Engineering Applications of Artificial Intelligence*, 2022, 115
[72] Fu X W, Wang H, Xu Z. Research on Multi-UAV Collaborative Pursuit Strategy Based on DE-MADDPG. *Journal of Aeronautics*, 2022, 43(5): 530
(符小卫, 王辉, 徐哲. 基于 DE-MADDPG 的多无人机协同追捕策略研究. 航空学报, 2022, 43(5): 530)
[73] Li B, Yang Z P, Chen D Q, et al. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning. *Defence Technology*, 2021, 17(2): 457
[74] Hu D X, Dong W H, Xie W J. Proximal strategy optimization of UAV autonomous guided tracking and obstacle avoidance. Journal of Beihang University, 2023, 49(01): 195
(胡多修, 董文瀚, 解武杰. 无人机自主引导跟踪与避障的近端策略优化. 北京航空航天大学学报, 2023, 49(01): 195)
[75] Wen C, Dong W H, Xie W J, et al. Autonomous tracking and obstacle avoidance of UAV swarm based on decoupled MADDPG. *Flight Mechanics*, 2022, 40(06): 24
(文超, 董文瀚, 解武杰等. 基于解耦型 MADDPG 的无人机集群自主跟踪与避障. 飞行力学, 2022, 40(06): 24)
[76] Jr. C S D, Rhys N, Akansel C, et al. Decentralized Multi-Agent Pursuit Using Deep Reinforcement Learning. *IEEE Robotics and Automation Letters*, 2021, 6(3)
[77] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods//International conference on machine learning. PMLR, 2018: 1587

[78] Zhou X, Zhu J, Zhou H, et al. Ego-swarm: A fully autonomous and decentralized quadrotor swarm system in cluttered environments//*2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021: 4101
[79] McGuire K N, De Wagter C, Tuyls K, et al. Minimal navigation solution for a swarm of tiny flying robots to explore an unknown environment. Science Robotics, 2019, 4(35): eaaw9710
[80] Faust A, Oslund K, Ramirez O, et al. Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning//*2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018: 5113
[81] Leiva F, Ruiz-del-Solar J. Robust rl-based map-less local planning: Using 2d point clouds as observations. *IEEE Robotics and Automation Letters*, 2020, 5(4): 5787
[82] Ma Z, Wang C, Niu Y, et al. A saliency-based reinforcement learning approach for a UAV to avoid flying obstacles. Robotics and Autonomous Systems, 2018, 100: 108
[83] Chen Y F, Liu M, Everett M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning//*2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017: 285
[84] Long P, Fan T, Liao X, et al. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning//*2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018: 6252
[85] Fan T, Long P, Liu W, et al. Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios. *The International Journal of Robotics Research*, 2020, 39(7): 856
[86] Huang H, Zhu G, Fan Z, et al. Vision-based Distributed Multi-UAV Collision Avoidance via Deep Reinforcement Learning for Navigation//*2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022: 13745