孙笠晋,李智. 基于多智能体强化学习的无人机对抗技术研究[J]. 智能计算机与应用,2026,16(4):55-60. DOI:10.20169/j.issn.2095-2163.25120501

基于多智能体强化学习的无人机对抗技术研究

孙笠晋, 李 智
(四川大学 电子信息学院, 成都 610065)

摘 要: 在人工智能和无人机集群技术飞速发展的今天,多无人机对抗成为军事领域的研究热点。多无人机对抗是无人机集群技术的延伸,智能算法控制无人机集群进行空中对抗。在多无人机对抗任务中,本文提出基于注意力机制的 MADDPG 算法在多无人机协同对抗中的应用,研究如何通过点对点信息交换,增强每架无人机的全局态势感知能力,尤其在每个无人机只能观察局部环境的情况下,通过与附近无人机交换信息,实现更好的协作效果。采用注意力机制动态选择其他无人机的关键信息,提升无人机的决策效率与精准度。研究注意力机制如何帮助每架无人机只关注任务相关信息,减少信息处理负担,提高策略优化效果。本文基于多智能体粒子环境搭建了无人机红蓝双方对抗的强化学习环境,红方由八架无人机组成,采用本文提出的强化学习算法,蓝方则使用基于规则的打击策略,实验结果表明,本文所提算法提升了红方胜率 12%。

关键词: 多无人机协同; 注意力机制; MADDPG 算法; 强化学习; 对抗任务
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0055-06

Research on UAV countermeasure technology based on multi-agent reinforcement learning

SUN Lijin, LI Zhi
(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: In today's era of rapid development in artificial intelligence and drone swarm technology, multi-drone combat has become a research hotspot in the military field. Multi-drone combat is an extension of drone swarm technology, where intelligent algorithms control a swarm of drones to engage in aerial confrontations. In multi-drone combat tasks, this paper proposes the application of a MADDPG algorithm based on attention mechanisms for collaborative combat among multiple drones. The study explores how point-to-point information exchange can enhance each drone's global situational awareness, particularly when each drone can only observe the local environment, by exchanging information with nearby drones to achieve better collaboration. The attention mechanism is used to dynamically select key information from other drones, improving decision-making efficiency and accuracy. The research investigates how the attention mechanism helps each drone focus only on task-relevant information, reducing the information-processing burden and enhancing strategy optimization. Based on a multi-agent particle environment, this paper builds a reinforcement learning environment for red-blue drone confrontations. The red team, consisting of eight drones, adopts the proposed reinforcement learning algorithm, while the blue team uses a rule-based attack strategy. Experimental results show that the proposed algorithm effectively increases the red team's win rate by 12%.

Key words: multi-UAV coordination; attention mechanism; MADDPG algorithm; reinforcement learning; adversarial tasks

0 引 言

随着人工智能和无人机技术的迅速发展[1],无人机集群技术在军事领域的应用已经逐渐成为现代战争的重要组成部分。无人机集群通过智能算法控制各个无人机之间的协作,能够有效执行侦察[2]、打击、电子战[3]等多种任务。与传统单架无人机相比,无人机集群能够通过协调合作,优化任务分配和执行效率,进而实现更加高效且灵活的作战能力。然而,随着作战任务的多样化和环境的复杂化,传统的单一无人机作战模式已经无法满足现代军事需求,尤其在面对敌方强烈的对抗和高动态性环境时,

作者简介: 孙笠晋(2001—),男,硕士研究生,主要研究方向:多智能体强化学习。
通信作者: 李 智(1974—),男,博士,教授,主要研究方向:物联网与边缘计算,压缩感知与频谱感知,群体智能。Email:286086522@qq.com。
收稿日期: 2025-12-05

传统控制方法的局限性愈加显著。

近年来,无人机集群对抗技术成为了学术界和军事实践中的研究热点。现有的研究方法大致可以分为两类:一类是基于规则驱动[4]的方法,另一类是基于学习的智能算法方法[5]。规则驱动方法依赖于预设的动作规则和手动调整的策略[6],虽然在一些简单的任务中取得了一定的成功[7],但在复杂、动态、不可预见的战场环境中,这种方法的适应性和灵活性表现不佳。相比之下,强化学习,特别是深度强化学习(Deep Reinforcement Learning, DRL),能够通过智能体与环境的交互进行学习,从而动态优化策略,具备较强的自适应能力,在复杂环境中展现出较高的决策效率和灵活性[8]。然而,当前的强化学习算法在多无人机对抗场景中仍面临着一些挑战,特别是高维状态空间导致的训练不稳定和收敛速度慢等问题,限制了其在实际对抗任务中的应用。

为了解决这些问题,近年来多智能体系统[9-10](MAS)和多智能体深度强化学习[11](MDRL)成为了研究的热点。基于多智能体框架的集中训练[12]、分布执行(CTDE)[13]方法,能够在保证智能体独立决策的同时实现协作,从而提升整体性能。然而,尽管已有基于强化学习的多无人机协作对抗算法取得了一定的成果,这些方法在复杂战斗环境中的表现仍然受限,特别是在强对抗性和多智能体相互作用的场景中,智能体之间的协作效率和学习的稳定性仍然是未解的难题。陈军等[14]提出了一种双层架构的方法,其中底层处理的是 1 对 1 的无人机对抗,而顶层则是多对多的空战对抗模型。该方法通过对 50 架无人机进行对抗仿真测试,探讨了多无人机协同空战中的复杂机动决策问题。Xie 等[15]提出了通过影响图分析多无人机对抗决策的方法,并利用贝叶斯理论对战斗态势进行精确评估。轩书哲等[16]则提出了一种基于近端策略优化的多智能体强化学习算法,采用了“集中式训练与分布式执行”模式,并在一个大规模的真实环境无人机对抗平台上进行了仿真测试。

针对以上挑战,本文提出了一种改进型的基于注意力机制[17]的 MADDPG 算法[18]。近年来注意力机制被广泛应用于目标识别算法中[19],通过引入注意力机制,本文算法能够在多无人机之间动态选择并聚焦于关键的协作信息,从而提升每个无人机的全局协作能力,优化其决策效率与准确性。传统的 MADDPG 算法在训练过程中,智能体通常只能感知到局部信息,难以全面了解全局环境。通过结合注意力机制,智能体能够根据任务相关性动态调整关注点,强化与其他无人机的协作,从而在多无人机协作对抗任务中实现更高效的决策。

本文的研究目标是通过构建一个多无人机对抗的强化学习仿真平台,验证改进型 MADDPG 算法在红蓝对抗场景中的有效性和优势。实验结果表明,基于注意力机制的算法能够显著提高红方无人机群的协作能力,从而提升其胜率,为未来无人机集群智能作战提供理论支持和实践参考。

1 相关理论知识

1.1 MADDPG

多智能体确定性策略梯度算法[20](Multi-Agent Deep Deterministic Policy Gradient, MADDPG)是在 DDPG[21]基础上针对多智能体场景提出的一种集中训练、分布执行(Centralized Training with Decentralized Execution,CTDE)方法。其核心思想是:在执行阶段,每个智能体仅依赖自身的局部观测信息独立做决策;在训练阶段,为每个智能体配置能够访问全局信息的集中式 Critic[22],以缓解多智能体环境的非平稳性问题。

1.1.1 算法框架

假设环境中共有 N 个智能体,第 i 个智能体具有确定性策略为:

$$a_i = \mu_{\theta_i}(o_i) \tag{1}$$

其中,θ_i 表示智能体 i 的局部观测,μ_i 表示其策略网络。执行时,智能体只使用 o_i 输出动作 a_i,而在训练时,第 i 个智能体的 Critic 网络可以访问联合状态(或联合观测)x 以及联合动作 $a = (a_1, a_2, \cdots, a_N)$,从而对当前策略进行全局评估。

1.1.2 策略目标与确定性策略梯度

第 i 个智能体的目标是最大化期望回报,目标函数为:

$$J(\theta_i) = \mathbb{E}_{x, a_i \sim D}[Q_i(x, a_1, a_2, \cdots, a_N)] \tag{2}$$

其中:θ_i 为智能体 i 的策略参数;Q_i 为针对智能体 i 的动作价值函数(由集中式 Critic 近似);D 为经验回放池中的采样数据。

采用确定性策略梯度(Deterministic Policy Gradient)时,第 i 个智能体的策略梯度可写为:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{x \sim D}[\nabla_{\theta_i} \mu_{\theta_i}(o_i) \cdot \nabla_{a_i} Q_i(x, a_1, \cdots, a_N)|_{a_i = \mu_{\theta_i}(o_i)}] \tag{3}$$

其中“·”表示向量内积。也就是说,Critic 对动作 a_i 的梯度会通过链式法则反向传播到策略网络 μ_i,从而实现端到端的策略更新。

1.1.3　集中式 Critic 的训练

第 i 个智能体的 Critic 网络通过最小化 Bellman 残差来学习 Q_i，损失函数为：

$$L_i = \mathbb{E}[(Q_i(x,a_1,\cdots,a_N) - y_i)^2] \quad (4)$$

其中目标值 y_i 定义为：

$$y_i = r_i + \gamma Q_i'(x',a_1',a_2',\cdots,a_N') \quad (5)$$

$$a_j' = \mu_{\theta_j}'(o_j'),\ j = 1,2,\cdots,N \quad (6)$$

其中：r_i 为智能体 i 在当前步获得的奖励；γ 为折扣因子（$0 < \gamma < 1$）；Q_i' 和 μ_{θ_j}' 分别为目标 Critic 网络与目标策略网络，用于提高训练稳定性；x'、o_j' 表示下一时刻的状态（或观测）。

1.1.4　集中训练与分布执行（CTDE）

在 MADDPG 中：分布执行阶段（Decentralized Execution），每个智能体只需根据自己的观测 o_i，通过 μ_i 输出动作 a_i，即 $a_i = \mu_{\theta_i}(o_i)$，不依赖其他智能体的信息，适合在通信受限的无人机对抗场景中部署。

集中训练阶段（Centralized Training），Critic 网络在训练时利用联合状态 x 和联合动作 a，对 $Q_i(x, a)$ 进行评估。这样可以显式建模智能体之间的相互影响，减轻由于其他智能体策略变化造成的环境非平稳性，从而提高学习效率与收敛质量。MADDPG 算法基本结构如图 1 所示。

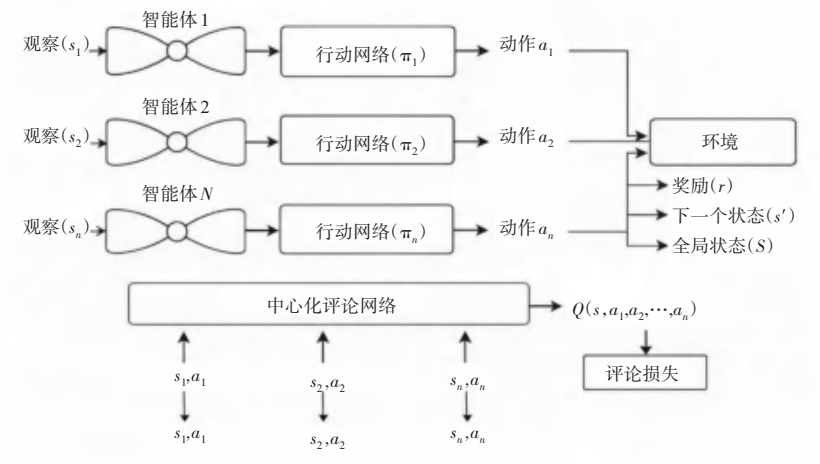

图 1　MADDPG 基本结构

Fig. 1　Basic structure of MADDPG

1.2　注意力机制

注意力机制灵感来源于人类的视觉注意力系统，旨在通过加权选择信息中的关键部分，提高模型对任务的处理效率。其核心思想是根据输入数据的重要性动态调整关注点，从而优化决策过程。

在强化学习的多智能体系统中，每个智能体通常只能感知局部环境信息，这使得其在决策时难以获得全局视野。引入注意力机制后，智能体可以通过为其他智能体的局部观测分配不同的权重，有效增强其全局信息的获取能力，从而优化协作策略。

假设每个智能体 i 的观测为 o_i，而其他智能体的观测为 o_j（其中 $j \neq i$）。注意力机制的核心是计算智能体之间信息的相关性，并基于此分配不同的关注权重。图 2 以经典点积注意力为例，详细展示了该机制的完整计算流程，具体包括查询-键-值向量构建、相关性度量、权重归一化及加权求和四大核心步骤。

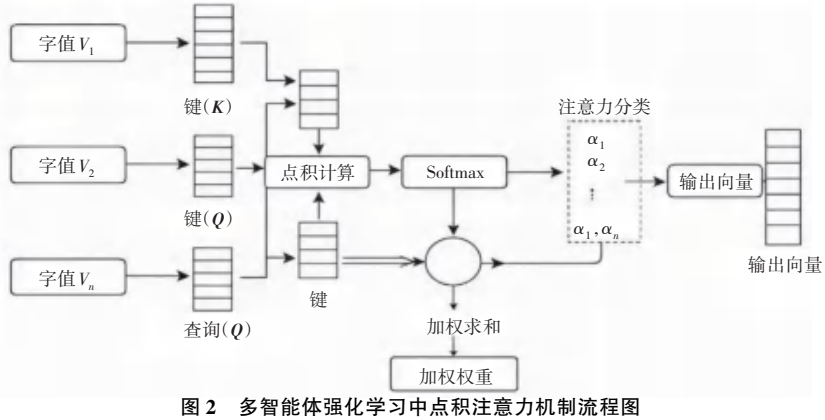

图 2　多智能体强化学习中点积注意力机制流程图

Fig. 2　Flowchart of dot-product attention mechanism in muli-agent reinforcement learning

具体来说,注意力机制通过计算每个智能体与其他智能体之间的关系得分 α_{ij},其公式如下:

$$\alpha_{ij}=\frac{\exp(\text{score}(o_i,o_j))}{\sum_{k\neq i}\exp(\text{score}(o_i,o_k))} \tag{7}$$

其中,$\text{score}(o_i,o_j)$ 表示智能体 i 和 j 之间的相关性得分,可以通过内积或其他距离度量方式来定义。通过计算得分α_{ij},智能体能够选择性地聚焦于与当前任务最相关的其他智能体。在每个时间步,智能体根据这些权重 α_{ij} 进行信息聚合,形成一个加权的观测向量 $\widetilde{\boldsymbol{o}}_i$:

$$\widetilde{\boldsymbol{o}}_i=\sum_{j\neq i}\alpha_{ij}o_j \tag{8}$$

该加权的观测向量 $\widetilde{\boldsymbol{o}}_i$ 将作为智能体的输入,用于策略决策过程。

2 多无人机对抗仿真环境搭建

无人机运动建模:假设红蓝双方无人机处于同一水平面,将对抗场景设置在二维的平面中,只考虑无人机的速度、加速度和航向,无人机的运动方程可以写成如下表达式:

$$\nu_{t+1}=\nu_t+a_t t \tag{9}$$

$$\alpha_{t+1}=\alpha_t+\lambda_t \tag{10}$$

$$\chi_{t+1}=\chi_t+\nu_t\cos(\alpha_t)\,t \tag{11}$$

$$y_{t+1}=y_t+\nu_t\sin(\alpha_t)\,t \tag{12}$$

其中,ν_t 表示 t 时刻的飞行速度;a_t 表示 t 时刻的飞行加速度;α_t 表示 t 时刻的航向;λ_t 表示 t 时刻的转向角;(χ_t,y_t) 表示 t 时刻的坐标。

图 3 构建了该场景下的无人机运动模型坐标系,明确了位置、速度、航向角之间的几何关系,为后续对抗策略的设计提供了直观的物理参考。

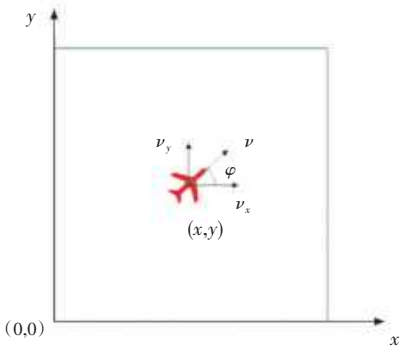

图 3 无人机运动模型

Fig. 3 UAV motion model

场景仿真建模:基于 OpenAI 的多智能体粒子环境搭建一个具有对抗性的多无人机仿真环境,模拟实际任务需求并设置无人机基本运动属性和攻击方式,使无人机能基于全局信息进行任务协作和对抗。如图 4 所示,设定红蓝双方各 8 架无人机。通过对无人机的行为和感知范围进行建模,确保系统能够真实模拟多无人机协作与竞争中的局限性与复杂性。

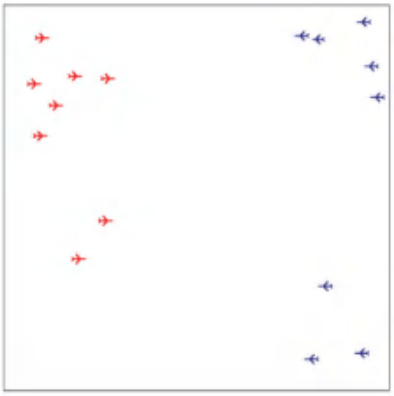

图 4 多无人机对抗场景仿真

Fig. 4 Schematic diagram of multi-UAV confrontation scenario simulation

3 实验与分析

3.1 实验参数及环境配置

在实验中,环境的大小为 10 000 m×10 000 m,红蓝方无人机各有 8 架。每架无人机的状态空间包含了与其他无人机的距离、方位角和存活情况等信息,动作空间由加速度和角速度构成。红方无人机使用基于 MADDPG 算法的深度强化学习进行训练,目标是通过不断交互优化策略,蓝方无人机采用基于规则攻击模式。训练过程中,经验池容量为 10 000,批次大小为 512,训练的总回合数为 20 000,每回合最多进行 500 步。实验的关键参数包括 Actor 和 Critic 的学习率(均为 0.001)、隐藏层维度(64)、折扣因子(0.99)以及目标网络更新速率(0.01)。此外,奖励机制也对训练起到了重要作用,其中击毁敌机的奖励为 100,越界或超速的惩罚为 10。训练的目标是通过这些配置最大化红方无人机的协作与对抗能力。基于改进注意力机制的 MADDPG 算法伪代码见表 1。

3.2 实验结果

本次实验采用两组对比实验,实验 a:红方基于 MADDPG 算法;实验 b:红方基于改进注意力机制的 MADDPG 算法,两组实验参数及环境配置均一致,结果如图 5、图 6 所示。

表 1　伪代码

Table 1　Pseudocode

	伪代码：MADDPG 添加 attention
1	Initialize N agent with replay buffer D、Actor and Critic#初始化 N 个智能体
2	For episode = 1, M do #开始进行训练，训练 M 轮
3	Reset environment#每轮开始的时候重置环境
4	For step = 1, J# 每轮训练 J 步
5	$s_{t\_attention} \leftarrow s_t$ #计算当前状态对应的 attention
6	For agent i = 1 to N do #对每一个智能体进行操作
7	a_t = Actor($s_{t\_attention}$) + N_t# 为每一个智能体根据其当前的状态 $s_{t\_attention}$ 和其 actor 并添加随机噪声 N_t，来选择一个动作进行探索
8	Execute actions a_t and observe reward r_t and new state s_{t+1}#依次执行动作，并获取各自对应的奖励和新的状态
9	$s_{t+1\_attention} \leftarrow s_{t+1}$#计算下一个状态对应的 attention
10	Store ($s_{t\_attention}$, a_t, r_t, $s_{t+1\_attention}$) in replay buffer D# 将每个智能体的当前状态、动作、奖励和新的状态放入各自的经验池
11	$s_{t\_attention} \leftarrow s_{t+1\_attention}$ #更新每个智能体的状态
12	End For
13	For agent i = 1 to N do #对每一个智能体进行操作
14	Sample transitions ($s_{t\_attention}$, a_t, r_t, $s_{t+1\_attention}$) from replay buffer D #从 buffer 里取出保存的数据
15	Set y_t = r_t + λCritic($s_{t+1\_attention+1}$) #根据奖励计算状态估值
16	Set y'_t = Critic($s_{t\_attention}$) #直接计算 critic 的估值
17	Perform a gradient descent step on $(y'_t - y_t)^2$with respect to the network parameters #最小化 $y'_t - y_t$，利用梯度更新 Critic 网络参数
18	Perform a gradient descent step on $-y_t$ with respect to the network parameters #最大化 Critic ($s_{t\_attention}$) 估值来更新 Actor 网络
19	End For
20	End For
21	End For

	伪代码：添加注意力机制
1	For agent i = 1 to N do #对每一个智能体进行操作
2	Caculate the distance from every other agent d_{ij} #计算智能体和其他智能体的距离
3	Merge $s_{it} \leftarrow s_{it}$ with $s_{it} * \frac{d_{ij}}{\sum_{j=1}^{N} d_{ij}, j \neq i}$ #根据距离权重来将其他智能体的状态和当前智能体的状态进行合并，得到添加了注意力机制的状态
4	End For

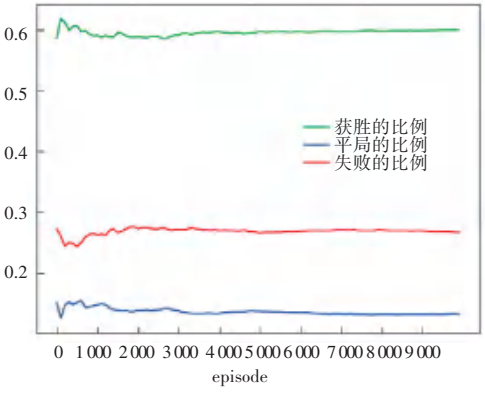

图 5　红方基于 MADDPG 算法

Fig. 5　Red side based on MADDPG algorithm

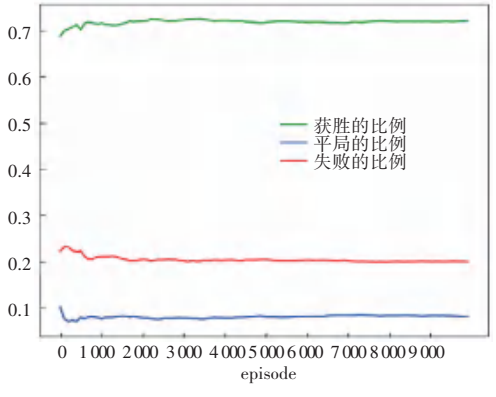

图 6　红方基于 ATT-MADDPG 算法

Fig. 6　Red side based on ATT-MADDPG algorithm

如图 5 所示为红方基于 MADDPG 算法与蓝方无人机群进行红蓝对抗一万轮次的结果，在 3 000 轮时胜率收敛于 60%左右；如图 6 所示为红方无人机群采用添加注意力机制的 MADDPG 算法与蓝方无人机群进行红蓝对抗一万轮次的结果，在 2 000 轮时胜率收敛于 72%左右。

4　结束语

通过第 3 节的实验结果可以看出，本文所提添加注意力机制的 MADDPG 算法，在适用于红蓝双方多

无人机对抗的场景中能显著提高红方胜率,胜率提升达到了 12 个百分点,并且收敛速度也得到提高。值得注意的是,尽管基于注意力机制的 MADDPG 算法在实验中表现出了明显的优势,但仍然存在进一步优化的空间。例如,如何进一步提升算法的鲁棒性,使其在更为复杂和动态的环境中保持较高的性能,仍然是未来的研究方向。此外,考虑到实际应用中无人机群的规模和任务的复杂性,如何扩展该算法以适应大规模无人机集群的实时对抗,也将是值得深入探讨的课题。

参考文献

[1] 牛轶峰, 肖湘江, 柯冠岩. 无人机集群作战概念及关键技术分析[J]. 国防科技, 2013, 34 (5): 37-43.

[2] STOLFI D H, BRUST M R, DANOY G, et al. Optimizing the performance of an unpredictable UAV swarm for intruder detection [C]//Proceedings of the 3th International Conference on Optimization and Learning. Cham:Springer,2020: 37-48.

[3] CEVIK P, KOCAMAN I, AKGUL A S, et al. The small and silent force multiplier: A swarm UAV-electronic attack [J]. Journal of Intelligent & Robotic Systems, 2013, 70: 595-608.

[4] 周欢, 赵辉, 韩统, 等. 基于规则的无人机集群飞行与规避协同控制[J]. 系统工程与电子技术,2016,38(6):1374-1382.

[5] 杨书恒, 张栋, 任智, 等. 基于多智能体强化学习的无人机集群对抗方法研究[J]. 无人系统技术,2022,5(5):51-62. DOI: 10. 19942/j. issn. 2096-5915. 2022. 5. 049.

[6] 徐华东, 王世勇, 杨轻, 等. 基于柱状空间和改进 A* 算法的无人机规避方法[J]. 测控技术, 2014, 33(7): 132-135.

[7] PHAM L B, DICKERSON B, SANDERS J, et al. UAV swarm attack: Protection system alternatives for Destroyers [D]. Monterey ,California:Naval Postgraduate School, 2012.

[8] YU M Y, YANG Z, KOLAR M, et al. Convergent policy optimization for safe reinforcement learnin[C]//Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2019: 3121-3133.

[9] 王祥科, 李迅, 郑志强. 多智能体系统编队控制相关问题研究综述[J]. 控制与决策, 2013, 28 (11): 1601-1613. DOI: 10. 13195/j. kzyjc. 2013. 11. 026.

[10] ABDOOS M, MOZAYANI N, BAZZAN A L. Holonic multi-agent system for traffic signals control [J]. Engineering Applications of Artificial Intelligence, 2013, 26 (5/6): 1575-1587.

[11] CHU T, WANG J, CODECÀ L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21: 1086-1095.

[12] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41 (S1):1-6.

[13] 周绍磊, 康宇航, 秦亮,等. 多无人机协同控制的研究现状与主要挑战[J]. 飞航导弹, 2015, 43(7): 31-35.

[14] 陈军, 高晓光. 机群协同空战中的指控系统建模与分析[J]. 计算机工程与应用, 2009,45(10): 195-198.

[15] XIE R, YING L J, LIN L D. Research on maneuvering decisions for multiUAVs Air combat [C]//Proceedings of 2014 IEEE International Conference on Control & Automation (ICCA). Piscataway,NJ:IEEE,2014:1598-1603.

[16] 轩书哲, 柯良军. 基于多智能体强化学习的无人机集群攻防对抗策略研究[J]. 无线电工程, 2021, 51 (5): 360-366.

[17] 薛俊杰, 王瑛, 祝捷, 等. 美国无人机分布式处理系统研究现状综述[J]. 飞航导弹, 2015, 43(10): 30-34.

[18] 刘峰, 魏瑞轩, 丁超, 等. 面向多机协同的 Att-MADDPG 围捕控制方法设计[J]. 空军工程大学学报(自然科学版),2021,22 (3):9-14.

[19] 羊淼海, 陈丹. 基于改进 YOLOv7 的抓取图像小目标检测算法[J]. 智能计算机与应用,2025,15(7):99-103. DOI:10. 20169/j. issn. 2095-2163. 250714.

[20] SUTTON R S, MCALLESTER D A, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation [C]//Proceedings of Advances in Neural Information Processing Systems. NeurIPS ,1999: 1057-1063.

[21] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [J]. arXiv preprint arXiv,1509. 02971,2015.

[22] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[J]. IEEE Transactions on Neural Networks, 2005, 16: 285-286.