文章编号:1671-1637(2026)03-0303-14

面向高层建筑应急救援的多无人机搜索轨迹协同控制方法

陈德启,张自设,张文会*,王宪彬
(东北林业大学 土木与交通学院,黑龙江 哈尔滨 150040)

摘 要:为解决多无人机协同执行高层建筑应急搜寻任务时,因智能体间近距离碰撞、队形重构等关键协同经验匮乏,致使学习效率低下、策略鲁棒性不足,提出了一种融合优先经验回放的多智能体深度确定性策略梯度模型(PER-MADDPG);构建了集成六自由度动力学模型的无人机集群仿真环境,将多机协同搜寻任务抽象为多智能体马尔可夫决策过程,并设计了融合个体轨迹跟踪、能耗约束以及团队队形保持、避碰需求的多层次奖励函数;通过中心化评论家网络计算团队联合动作的时序差分误差,对联合经验进行量化并实施优先采样,引导算法聚焦于高价值的稀疏协同样本,从而加速了鲁棒协同策略的收敛。研究结果表明:PER-MADDPG 算法任务成功率达 98%,较基准 MADDPG 算法提升 15.3%;智能体间碰撞率由 8%降低至 1%;在协同与控制精度方面,平均队形误差由 0.07 m 降至 0.03 m,平均轨迹跟踪误差由 0.12 m 降至 0.05 m;在四机及六机编队的扩展性测试中能有效克服物理空间拥挤导致的性能衰减,展现出优于基准算法的鲁棒性。建立的 PER-MADDPG 能够有效平衡个体控制精度与团队协同稳定性,提升高层建筑应急救援的搜寻效率。

关键词:航空运输;低空交通;多智能体;高层建筑应急救援;无人机集群;轨迹协同控制
中图分类号:U8 **文献标志码**:A **DOI**:10.19818/j.cnki.1671-1637.2026.159

Collaborative control method for multi-UAV search trajectory in high-rise building emergency rescue

CHEN De-qi, ZHANG Zi-she, ZHANG Wen-hui*, WANG Xian-bin
(School of Civil Engineering and Transportation, Northeast Forestry University, Harbin 150040, Heilongjiang, China)

Abstract: To address the issues of low learning efficiency and poor strategy robustness in multi-UAV systems during their collaborative emergency search task for high-rise buildings, caused by insufficient experience in close-range inter-agent collision and formation reconfiguration, a multi-agent deep deterministic policy gradient model integrated with prioritized experience replay (PER-MADDPG) was proposed. A UAV swarm simulation environment with a six-DOF dynamic model was established. The multi-UAV collaborative search task was formulated as a multi-agent Markov decision process, and a hierarchical reward function integrating individual trajectory tracking, energy constraints, team formation keeping, and collision avoidance requirements was

出版历程:2025-10-10 收稿,2026-01-03 修回,2026-01-23 录用
基金项目:黑龙江省哲学社会科学研究规划项目(23GLC022);国家自然科学基金项目(52572369)
作者简介:陈德启(1990-),男,黑龙江哈尔滨人,讲师,工学博士,E-mail:chendeqi@nefu.edu.cn。
* **通信作者**:张文会(1978-),男,黑龙江哈尔滨人,教授,工学博士,博士后,E-mail:zhangwenhui@nefu.edu.cn。
引用格式:陈德启,张自设,张文会,等.面向高层建筑应急救援的多无人机搜索轨迹协同控制方法[J].交通运输工程学报,2026,26(3):303-316.
Citation: CHEN De-qi, ZHANG Zi-she, ZHANG Wen-hui, et al. Collaborative control method for multi-UAV search trajectory in high-rise building emergency rescue[J]. Journal of Traffic and Transportation Engineering, 2026, 26(3): 303-316.

designed. A centralized critic network was used to calculate the temporal difference errors of the team's joint actions. The joint experiences were quantified, and the prioritized sampling was implemented. The algorithm was guided to focus on high-value sparse collaborative samples. As a result, the convergence to robust collaborative policies was accelerated. Experimental results show that a 98% task success rate is achieved by the PER-MADDPG algorithm, 15.3% higher than the baseline MADDPG algorithm, and the inter-agent collision rate is reduced from 8% to 1%. In terms of collaboration and control accuracy, the average formation error is decreased from 0.07 m to 0.03 m, and the average trajectory tracking error is lowered from 0.12 m to 0.05 m. In the scalability tests on four-UAV and six-UAV formations, the performance degradation caused by physical space congestion is effectively overcome, demonstrating superior robustness to baseline algorithms. The established PER-MADDPG can effectively balance individual control accuracy and team collaboration stability, enhancing search efficiency in high-rise building emergency rescue.

Keywords: aviation transportation; low-altitude traffic; multi-agent; high-rise building emergency rescue; UAV swarm; trajectory collaborative control

Publication history: Received 2025-10-10; Received in revised form 2026-01-03; Accepted 2026-01-23

Funding: Heilongjiang Province Philosophy and Social Science Research Planning Project (23GLC022); National Natural Science Foundation of China (52572369)

* **Corresponding author**: ZHANG Wen-hui, professor, PhD, E-mail: zhangwenhui@nefu.edu.cn.

0 引 言

随着城市化进程的加速,高层建筑的应急救援需求日益凸显。据住建部 2024 年统计数据显示,中国超过 18 层的楼宇占比已达 40%,而常规消防云梯的作业高度普遍难以逾越 15 层,这使得更高楼层的垂直救援面临巨大挑战[1]。在此背景下,多无人机(Unmanned Aerial Vehicles, UAVs)编队以其更高的搜寻效率和更广的覆盖范围,为高层建筑应急救援快速搜索提供新型手段。

多智能体强化学习(Multi-agent Reinforcement Learning, MARL)通过交互试错学习,能够自主演化出多无人机的协同控制策略[2],将多无人机协同搜索任务转化为基于 MARL 的协同控制问题[3-4]。早期的 MARL 中每个智能体将其他智能体视为环境的固有组成部分,独立运用单智能体强化学习算法开展学习[5]。但后续研究表明,由于各智能体均在独立更新策略,环境会呈现非平稳性特征,使得智能体的学习过程难以收敛[6]。为缓解此问题,研究者将独立近端策略优化(Independent Proximal Policy Optimization, IPPO)算法应用于无人机集群控制[7-8],尽管在部分简单任务中取得了一定成效,但其无法实现智能体间的明确协调,因此在需要精细协作的任务中性能仍存在局限[9]。

针对非平稳性问题,研究者提出了"中心化训练-去中心化执行"(Centralized Training with Decentralized Execution, CTDE)框架,现已成为 MARL 领域的主流范式[10-11]。CTDE 框架在训练阶段允许算法获取所有智能体的全局信息,从而得到稳定的学习信号;执行阶段则让每个智能体基于自身局部观测独立决策,兼顾了策略的可部署性[12]。基于该框架,涌现出一系列代表性算法,其中 Lowe 等[13]提出的多智能体深度确定性策略梯度(Multi-agent Deep Deterministic Policy Gradient, MADDPG)堪称开创性成果,现已被广泛应用于无人机协同导航、目标跟踪等任务场景[14]。近期研究将注意力机制与 MADDPG 结合,以处理更复杂的智能体间交互[15]。不过,当智能体数量增加时,MADDPG 的中心化 Critic 输入维度会急剧膨胀,进而引发"维度灾难"问题[16]。

为提升算法扩展性,基于单调值函数分解(Q-value Mixing Network, QMIX)方法受到了广泛关注[17-18]。QMIX 及其变体在无人机集群的协同探索、资源分配等任务中表现出色[19-20]。然而,这类方法通常更适用于纯合作型任务,在处理智能体间交互复杂的任务时,其值函数分解的假设可能过于简化[21]。此外,基于 Actor-critic 框架的 On-policy 算法也在 CTDE 框架下逐步发展。研究人员提出的

多智能体近端策略优化（Multi-agent Proximal Policy Optimization，MAPPO）算法将 PPO 扩展至多智能体领域[22]。MAPPO 因其出色的稳定性和性能，在众多 MARL 测试中取得了较好效果[23]，并已成功应用于无人机编队控制、自动驾驶车队协同等领域[24-25]。例如，Wu 等[26]将 MAPPO 算法应用于大规模无人机集群的对抗任务。不过，作为 On-policy 算法，MAPPO 与 IPPO 类似，存在样本效率较低的固有局限；在需要与环境进行大量交互以学习有效策略的复杂任务中，其训练成本显著增加[27]。

在无人机编队飞行的大部分时间里，编队执行的是相对常规的轨迹跟踪与队形保持任务，由此产生的经验数据量大但信息价值密度较低[28]。而那些对学习安全、高效的协同策略至关重要的关键事件（如成功规避潜在的智能体间近距离碰撞，从严重的队形破坏中快速恢复，或在狭窄空间内完成精妙的联合机动），在整个经验流中却很稀疏[29]。标准经验回放机制会等概率复用所有历史数据，导致这些宝贵的稀疏经验极易被海量常规经验“淹没”，使得算法难以从这些决定任务成败的关键瞬间中高效学习[30]。因此，如何引导 MARL 算法聚焦于这类高价值的稀疏协同经验，已成为提升多无人机编队在复杂环境中智能决策水平的关键突破口[31]。

为应对现有 MARL 算法处理稀疏关键协同事件时的学习效率瓶颈，文中提出了融合优先经验回放（Priority Experience Replay，PER）的 PER-MADDPG 模型，并设计高保真多无人机协同螺旋扫描任务，构建局部观测空间及兼顾个体行为与团队目标的多层次奖励机制；通过搭建仿真平台，开展不同编队规模的扩展性试验以探究算法在复杂拥挤环境下的性能极限，并与基线方法对比验证其在任务效率与协同稳定性上的优势。

1　问题描述与模型建立

多无人机协同编队能显著提升高层建筑搜救效率，但其高维、紧耦合特性对维持队形与个体控制提出了严峻挑战。为此，首先构建高保真六自由度动力学模型作为物理基础，进而将多智能体的协同序贯决策过程转化为马尔可夫决策过程。

1.1　无人机动力学模型

1.1.1　运动学方程

无人机的运动学描述了其几何运动特性，定义了 2 个坐标系：一是固定的地面坐标系（$O_g x_g y_g z_g$），用于描述无人机的全局位置；二是与无人机固连的机体坐标系（$O_b xyz$），用于描述无人机自身的姿态和速度。无人机在地面坐标系中的位置由（x_g，y_g，z_g）表示，其姿态由欧拉角（滚转角 φ，俯仰角 θ，偏航角 ψ）表示。无人机的位置和姿态变化率由其在机体坐标系下的速度分量（u，v，ω）和角速度分量（p，q，r）通过式(1)、(2)转换得到

$$\begin{cases}\dot{x}_g = u\cos(\theta)\cos(\psi)+v[\sin(\varphi)\sin(\theta)\cos(\psi)-\cos(\varphi)\sin(\psi)]+\omega[\cos(\varphi)\sin(\theta)\cos(\psi)+\sin(\varphi)\sin(\psi)]\\ \dot{y}_g = u\cos(\theta)\sin(\psi)+v[\sin(\varphi)\sin(\theta)\sin(\psi)+\cos(\varphi)\cos(\psi)]+\omega[\cos(\varphi)\sin(\theta)\sin(\psi)-\sin(\varphi)\cos(\psi)]\\ \dot{z}_g = -u\sin(\theta)+v\sin(\varphi)\cos(\theta)+\omega\cos(\varphi)\cos(\theta)\end{cases} \tag{1}$$

$$\begin{cases}\dot{\varphi} = p+q\sin(\varphi)\tan(\theta)+r\cos(\varphi)\tan(\theta)\\ \dot{\theta} = q\cos(\varphi)-r\sin(\varphi)\\ \dot{\psi} = [q\sin(\varphi)+r\cos(\varphi)]/\cos(\theta)\end{cases} \tag{2}$$

式中：u、v、ω 分别为无人机在机体坐标系下沿其前向轴、右向轴、下向轴的线速度分量；p、q、r 分别为无人机在机体坐标系下沿其前向轴、右向轴、下向轴的角速度分量。

1.1.2　动力学方程

无人机的动力学描述了力、力矩与无人机运动状态改变之间的关系。作用在无人机上的物理力的矢量和构成了作用在机体上的合外力（F_x，F_y，F_z），并产生了滚转、俯仰、偏航等合外力矩（L，M，C）。根据牛顿-欧拉方程，机体坐标系下的线加速度（$\dot{u}$，$\dot{v}$，$\dot{\omega}$）和角加速度（$\dot{p}$，$\dot{q}$，$\dot{r}$）由式(3)、(4)的合外力与合外力矩决定

$$\begin{cases}\dot{u} = rv - q\omega + (F_x/m)\\ \dot{v} = p\omega - ru + (F_y/m)\\ \dot{\omega} = qu - pv + (F_z/m)\end{cases} \tag{3}$$

$$\begin{cases}\dot{p} = [(I_y - I_z)qr + L]/I_x\\ \dot{q} = [(I_z - I_x)rp + M]/I_y\\ \dot{r} = [(I_x - I_y)pq + C]/I_z\end{cases} \tag{4}$$

式中：m 为无人机质量；（I_x，I_y，I_z）为无人机绕机体坐标系各轴的转动惯量。

1.2　马尔可夫决策过程建模

鉴于实际通信与感知受限，本文将多无人机协同扫描任务建模为分散式部分可观测马尔可夫决策过程。构建了以局部观测空间，使智能体能利用有限信息推断全局状态并习得协同策略，其核心要素定义如下。

1.2.1　状态空间

在多智能体协同任务中，每个智能体（无人机）i 的决策依据是其局部观测向量 $\boldsymbol{o}_i$。为实现高效协

同,该观测向量被设计为包含自身任务状态、自身物理状态以及对邻居智能体的感知三部分,计算如下

$$\begin{cases}\boldsymbol{o}_i=(\Delta\boldsymbol{p}_i,\boldsymbol{v}_i,\Delta\boldsymbol{p}_{i,1},\cdots,\Delta\boldsymbol{p}_{i,m},\cdots,\Delta\boldsymbol{p}_{i,N-1})^{\mathrm{T}}\\ \Delta\boldsymbol{p}_i=\boldsymbol{p}_i-\boldsymbol{p}_{\mathrm{d},i}\\ \Delta\boldsymbol{p}_{i,m}=\boldsymbol{p}_i-\boldsymbol{p}_m\end{cases}\tag{5}$$

式中:N 为无人机总数;$\Delta\boldsymbol{p}_i$ 为无人机 i 的当前位置 $\boldsymbol{p}_i$ 与其动态期望目标点 $\boldsymbol{p}_{\mathrm{d},i}$ 之间的三维误差向量;$\boldsymbol{v}_i\in\mathbb{R}^3$ 为无人机 i 的三维线速度向量;$\Delta\boldsymbol{p}_{i,\mathrm{m}}$ 为无人机 i 与无人机 m 之间的相对位置向量。

1.2.2 动作空间

本研究采用同质智能体设置,所有无人机共享相同的动作空间定义。每个智能体 i 的动作 $\boldsymbol{a}_i$ 被定义为一个归一化的三维力向量,即

$$\begin{cases}\boldsymbol{a}_i=(f_x,f_y,f_z)^{\mathrm{T}}\\ f_x,f_y,f_z\in[-1,1]\end{cases}\tag{6}$$

该归一化动作向量 $\boldsymbol{a}_i$ 在物理引擎中会经过最大推力系数 $F_{\max}$ 的缩放,并叠加一个恒定的重力补偿力,最终形成施加于无人机质心的合外力,从而驱动其在三维空间中运动。在任意时刻 t,所有 N 个无人机的智能体瞬时动作共同构成了系统的联合动作 $\boldsymbol{a}_{\mathrm{z},t}=\{\boldsymbol{a}_{1,t},\boldsymbol{a}_{2,t},\cdots,\boldsymbol{a}_{N,t}\}$。

1.2.3 奖励函数

奖励函数的设计是引导智能体从独立的个体行为演化为高效团队协作的关键。为实现这一目标,本文设计了一个融合了个体奖励与团队奖励的多层次结构化奖励函数。

(1)个体奖励 R_{i} 主要激励智能体完成其基本职责,每个智能体独立计算。

①轨迹跟踪奖励 r_{t} 旨在引导无人机紧密跟随其预设螺旋轨迹的核心正向激励,被设计为轨迹误差的指数衰减函数,误差越小,奖励值越高,从而鼓励高精度控制,计算如下

$$r_{\mathrm{t}}=\exp(-\tau_1\|\Delta\boldsymbol{p}_i\|_2)\tag{7}$$

式中:τ_1 为调节奖励函数陡峭程度的参数;$\|\cdot\|$ 表示求 L2 范数(欧几里得范数)。

②能耗惩罚 r_{e} 旨在为激励无人机以更平稳、节能的方式飞行,通过对控制动作的幅值(L2 范数)施加负向惩罚,引导智能体最小化控制输出,其定义如下

$$r_{\mathrm{e}}=-\tau_2\|\boldsymbol{a}_i\|_2^2\tag{8}$$

式中:τ_2 为能耗惩罚系数。

(2)团队奖励 R_{t} 该部分奖励由系统根据所有智能体的全局状态计算得出,并广播给所有当前存活的智能体,用于促进协同行为的涌现。

①队形保持惩罚 r_{f} 为维持编队稳定性,系统会计算无人机编队当前构型与期望构型之间的几何偏差,并施加惩罚。具体而言,该惩罚基于无人机对 (i,m) 的当前距离 $d_{i,m}$ 和其各自目标点之间的期望距离 $d_{\mathrm{d},i}$ 误差的绝对值总和,其定义如下

$$r_{\mathrm{f}}=-\tau_3\sum_{i<m}|d_{i,m}-d_{\mathrm{d},i}|\tag{9}$$

式中:τ_3 为队形惩罚系数。

需要特别说明的是,该机制是驱动“队形恢复”行为的核心动力,算法并未预设特定的恢复逻辑或状态判断,而是通过最小化该惩罚项,驱动智能体在探索过程中自主涌现出从队形破坏状态快速回归期望构型的协同策略。

②事件驱动的团队奖惩 r_{event} 仅在回合结束时触发,用于定义团队层面的最终目标。其包含共同成功奖励与共同失败惩罚:共同成功奖励 r_{s},仅当所有无人机都无碰撞地完成了各自的扫描任务时,系统会给予一个巨大的正向团队奖励;共同失败惩罚 r_{fail},仅当任何一个无人机发生碰撞或飞出任务边界时,整个团队会立即受到一个巨大的负向惩罚,并终止当前回合。

最终,智能体 i 在任意时刻 t 所接收到的总奖励 $R_{\mathrm{total},i,t}$ 由其个体奖励和共享的团队奖励共同构成,计算如下

$$\begin{aligned}R_{\mathrm{total},i,t}&=R_{\mathrm{i}}+R_{\mathrm{t}}=\\&r_{\mathrm{t},i,t}+r_{\mathrm{e},i,t}+r_{\mathrm{f},t}+r_{\mathrm{event},t}\end{aligned}\tag{10}$$

这种精细化的多层次奖励设计,是引导多智能体在个体利益和集体目标之间做出有效权衡、学习复杂协同策略的关键所在。

2 基于优先经验回放的多智能体协同控制方法

2.1 PER-MADDPG 框架

PER-MADDPG 框架主要包含 2 个相互关联的核心循环:左半区的多智能体环境交互与数据采集循环,以及位于右半区的中心化网络训练与参数更新循环,如图 1 所示。

在交互循环中,多个独立的智能体(Actor 网络 μ_i)根据各自的局部观测生成动作 $\boldsymbol{a}_i=\mu_i(\boldsymbol{o}_i)$ 并与环境交互,产生联合经验 $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{r},\boldsymbol{s}')$,这些经验被存入一个共享的“优先经验回放池”中。在训练循环中,PER-MADDPG 的核心机制得以体现。首先,“优先采样”模块根据经验的重要性(Temporal-difference Error, TD-error)从回放池中抽取小批量联合经

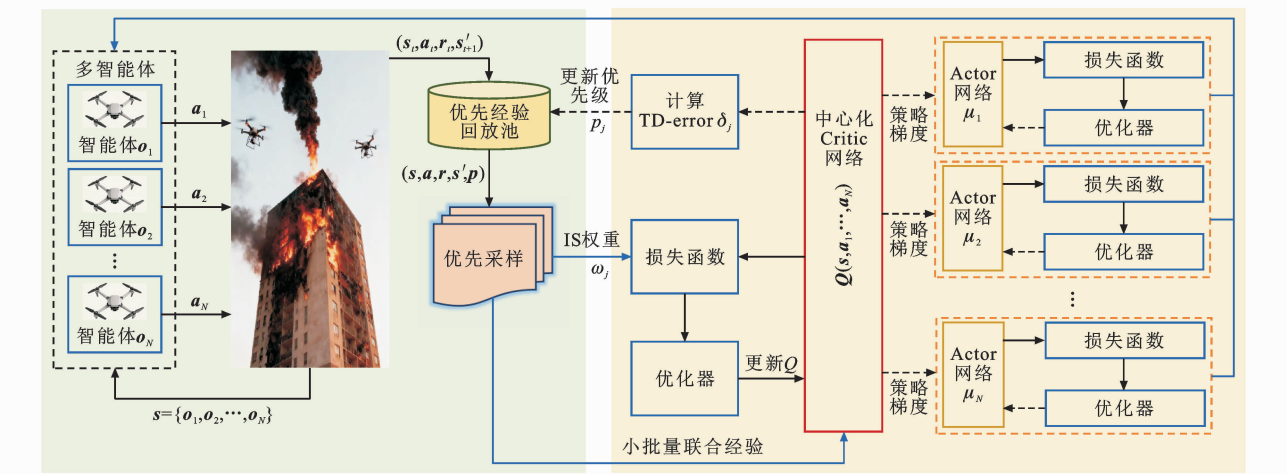

图 1　PER-MADDPG 算法框架

Fig. 1　PER-MADDPG algorithm framework

验，并将其送入中心化的 Critic 网络 $Q(\boldsymbol{s},\boldsymbol{a}_1,\cdots,\boldsymbol{a}_N)$。该 Critic 网络在训练时可以访问所有智能体的联合观测 $\boldsymbol{s}=\{\boldsymbol{o}_1,\cdots,\boldsymbol{o}_N\}$ 和联合动作 $\{\boldsymbol{a}_1,\cdots,\boldsymbol{a}_N\}$。Critic 网络通过左侧的更新回路（Loss 函数与优化器）进行学习，并产生用于评估联合动作价值的 TD-error。这个 TD-error 一方面通过关键的反馈回路（图 1 中虚线），反向更新经验池中样本的优先级；另一方面，中心化 Critic 产生的策略梯度，则被用于指导左侧每一个独立的 Actor 网络进行更新。

2.2　融合优先经验回放的核心机制

在标准 MADDPG 算法中引入优先经验回放 PER 机制，利用中心化 Critic 产生的团队 TD-error 量化联合经验的重要性，对预测偏差大、信息增益高的关键样本赋予高采样优先级，引导智能体聚焦于高价值稀疏事件的学习，从而克服样本效率瓶颈并加速鲁棒协同策略的收敛。PER-MADDPG 的实现主要围绕经验的优先级定义、优先采样和带权重更新 3 个环节展开。

2.2.1　经验优先级

PER 的核心在于利用时序差分误差 TD-error 来量化经验样本的重要性。在 PER-MADDPG 中，文中利用中心化 Critic 的 TD-error 来评估一个联合经验的价值。对于经验池中的第 j 个样本 $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{r},\boldsymbol{s}')$，其 TD-error δ_j 公式如下

$$\delta_j = r_j + \gamma Q'_j(\boldsymbol{s}'_j,\boldsymbol{a}'_j) - Q_j(\boldsymbol{s}_j,\boldsymbol{a}_j) \tag{11}$$

式中：r_j、$\boldsymbol{s}_j$、$\boldsymbol{a}_j$、$\boldsymbol{s}'_j$ 分别为经验样本 j 中记录的奖励、联合观测、联合动作和下一联合观测；$\boldsymbol{a}'_j$ 为根据目标 Actor 网络生成的下一联合动作向量；$Q_j(\cdot)$ 和 $Q'_j(\cdot)$ 分别为 Critic 网络和目标 Critic 网络；γ 为折扣因子。

这个 TD-error 反映了中心化 Critic 对整个团队在该时刻联合动作价值预测程度，其绝对值 $|\delta_j|$ 越大，意味着该经验包含更重要的协同信息。

随后，样本 j 的优先级 p_j 根据其 TD-error 的绝对值计算得出

$$p_j = (|\delta_j|+\varepsilon)^{\kappa} \tag{12}$$

式中：ε 为一个极小的正常数，用于保证 TD-error 为 0 的样本也具有一个基础的被采样概率；κ 为超参数，决定了优先级的“强度”。

2.2.2　优先采样与重要性采样权重

基于上述优先级，样本 j 被采样的概率 $P(j)$ 计算如下

$$P(j) = \frac{p_j}{\sum_k p_k} \tag{13}$$

这种有偏采样会改变样本的原始分布，直接用于训练将导致价值函数的估计产生偏差。为修正此偏差，PER 引入了重要性采样 IS 权重 ω_j 来对损失函数的梯度更新进行缩放，计算如下

$$\omega_j = \left[\frac{1}{\mathcal{D}P(j)}\right]^{\beta} \tag{14}$$

式中：$\mathcal{D}$ 为经验回放池的大小；超参数 β 用于调节补偿的程度。

最终，中心化 Critic 网络的损失函数 $L(Q_j)$ 被该权重所加权，重新定义为

$$L(Q_j) = E_{j\sim P(j)}\left\{\omega_j[y_j - Q_j(\boldsymbol{s}_j,\boldsymbol{a}_j)]^2\right\} \tag{15}$$

$$y_j = r_j + \gamma Q'_j(\boldsymbol{s}'_j,\boldsymbol{a}'_j)$$

$$\boldsymbol{a}'_j = (\mu'_1(\boldsymbol{o}'_{1,j}),\cdots,\mu'_N(\boldsymbol{o}'_{N,j}))$$

式中：$E_{j\sim P(j)}$ 为优先经验回放机制下，经验样本 j 服从优先级概率分布 $P(j)$ 时的数学期望；μ'_k 为智能

体 k 的目标 Actor 网络。

通过上述机制，PER-MADDPG 能够聚焦于高 TD-error 的关键协同经验，显著提升学习效率和最终策略的鲁棒性。

3 仿真试验与结果

3.1 试验设计

本节将详细阐述仿真试验的各项配置，包括软硬件平台、核心仿真参数、关键超参数及基线算法等。

3.1.1 仿真平台与参数设置

为保证研究的可重复性和结果的可靠性，所有模型的训练与测试均在统一的仿真平台下完成。基于 PyBullet 物理引擎搭建高保真仿真平台，算法训练框架采用 Ray RLlib(PyTorch 后端)，试验运行于配置 Intel Xeon E5 CPU 与 NVIDIA GeForce 3090 GPU 的高性能服务器。试验任务设定为双机/多机协同螺旋扫描。

为在保证动力学仿真精度的同时兼顾大规模蒙特卡洛测试的计算效率，本文选取典型高层建筑切片作为作业场景。各无人机基于时间参数化的圆柱螺旋线方程生成参考轨迹，对于第 i 个无人机，其在时刻 t 的期望位置向量 $\boldsymbol{p}_{\mathrm{d},i}(t)$ 计算如下

$$\boldsymbol{p}_{\mathrm{d},i}(t)=\left(R_{\mathrm{h}}\cos\left(\frac{2\pi Kt}{T_{\mathrm{task}}}+\varphi_i\right),R_{\mathrm{h}}\sin\left(\frac{2\pi Kt}{T_{\mathrm{task}}}+\varphi_i\right),\frac{H_{\mathrm{h}}t}{T_{\mathrm{task}}}+z_0\right)^{\mathrm{T}} \tag{16}$$

式中：R_{h}、H_{h} 分别为螺旋半径、螺旋总高度；T_{task} 为任务总时长；K 为螺旋总圈数；φ_i 为第 i 架无人机为保持编队构型而设定的初始相位偏移；z_0 为初始起飞高度。

为保证对比试验的公平性，所有对比算法的关键超参数均经过初步调优并保持一致。超参数设置如表 1 所示。

表 1 算法关键超参数设置

Table 1 Key hyperparameter settings for algorithm

超参数	描述	PER-MADDPG	MADDPG	MAPPO	IPPO
学习率	Actor/Critic 网络学习率	3.0×10^{-5}			
折扣因子	未来奖励的折扣系数	0.99			
经验池大小	存储历史经验的最大数量	1.0×10^{6}			
κ	采样优先级的指数	0.6			
β	重要性采样的指数	0.4			
软更新率	目标网络更新速率	0.005	0.005		

3.1.2 评价指标

为全面评估算法在多智能体协同任务中的综合性能，本文设计了涵盖任务完成度、协同与个体性能、安全性及效率 5 个维度的量化指标。所有指标均通过 100 次独立的蒙特卡洛测试计算平均值与标准差，以确保结果的统计显著性。

任务成功率：统计在规定时间内，编队以高质量协同状态完成螺旋扫描任务的回合百分比。定义“成功”必须同时满足无碰撞、无越界且全程保持有效轨迹跟踪。其中，有效跟踪是指无人机的实时位置跟踪误差(欧几里得距离)在飞行全过程中始终低于设定的安全阈值(本文设定有效跟踪误差阈值 $\delta_{\mathrm{th}}=1.0$ m)。

平均任务完成时间：仅统计成功回合的平均耗时，衡量任务效率。

平均队形误差：衡量无人机编队在飞行过程中与理想几何构型的平均偏差，是评估算法协同能力的核心指标。

平均轨迹跟踪误差：单个无人机与其自身预设轨迹的平均偏差，反映控制精度。

智能体间碰撞率：在所有测试回合中，发生无人机之间碰撞的回合所占的百分比，是评估算法安全性的核心指标。

平均控制能耗：所有无人机电机推力指令的平方积分均值，衡量经济性与平稳性。

3.1.3 对比算法

为充分验证本文所提 PER-MADDPG 算法的先进性，选取了涵盖非协作学习、主流协同算法等多种基线进行系统性比较。

IPPO：每个智能体独立使用 PPO 算法进行学习，互不通信和观察。该方法作为非协作学习的下限基准，用以衡量协同的必要性。

MAPPO：基于 On-policy 的先进多智能体协同算法，代表了同类算法的先进水平。

MADDPG：本文所提算法的基础，一个基于 Off-policy 的先进多智能体协同算法，用于直接对比 PER 机制带来的性能增益。

3.2 PER 关键超参数敏感性分析

在不同算法间的横向对比之前，首先需要确定本研究所提 PER-MADDPG 算法中关键超参数(优先级指数 κ 与重要性采样指数 β)的最优取值。为此，本节采用控制变量法进行预试验，以任务成功率和队形误差为指标，探究这 2 个参数对算法性能的影响。

3.2.1　优先级指数的影响分析

优先级指数 $\kappa\in[0,1]$ 决定了 TD-error 转化为优先级的强度，$\kappa=0$ 时，PER 退化为均匀采样。固定 $\beta=0.4$，分别测试了 κ 为 0、0.3、0.6、0.9 时模型的性能，如图 2 所示。

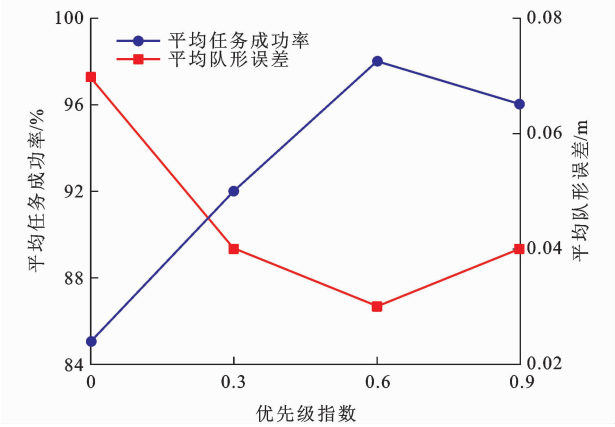

图 2　优先级指数的敏感性分析

Fig. 2　Sensitivity analysis of priority indices

由图 2 可见：当 $\kappa=0$（即标准 MADDPG）时，模型性能最低。随着 κ 的增加，任务成功率显著提升，平均队形误差显著下降，这证明了优先经验回放机制对于提升学习效率和协同精度的有效性；当 $\kappa=0.6$ 时，2 项指标均达到最优；然而，当 κ 进一步增加到 0.9 时，性能出现轻微下降，这可能是因为过高的优先级强度导致算法过于集中学习少数 TD-error 极大的样本，容易陷入局部最优，破坏了策略探索的多样性，反而导致训练不稳定和最终性能的下降。因此，$\kappa=0.6$ 是平衡学习效率与策略多样性的最佳选择。

3.2.2　重要性采样指数的影响分析

重要性采样指数 $\beta\in[0,1]$ 用于修正由优先采样带来的数据分布偏差。本研究固定 $\kappa=0.6$，分别测试了 β 为 0、0.4、0.7、1.0 时模型的性能，结果如图 3 所示。

由图 3 可见：当 $\beta=0$ 时，尽管采用了优先采样，但由于价值函数估计存在严重偏差，导致学习效果很差，任务成功率仅为 65%，队形误差也处于最高点；随着 β 的增加，性能显著改善；当 $\beta=0.4$ 时，模型取得了最优的综合性能；当 β 继续增大至 0.7 和 1.0 时，虽然性能依然维持在较高水平，但相比 $\beta=0.4$ 有微弱的下降，这可能是因为在学习初期，TD-error 的估计本身波动较大，过高的 β 值会放大这种不稳定性，而一个适中的值能够在修正偏差和维持训练稳定性之间取得更好的平衡。综上所述，通过

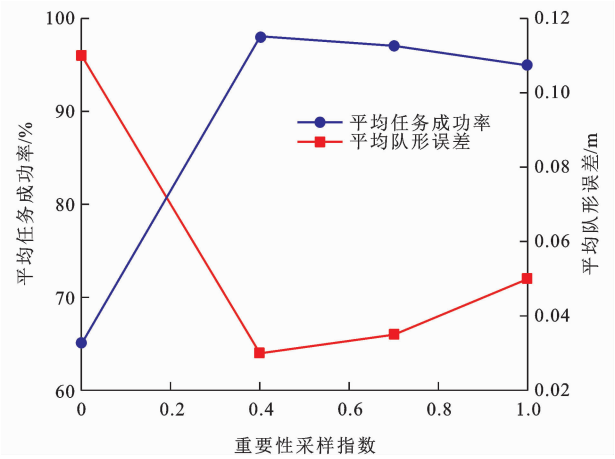

图 3　重要性采样指数的敏感性分析

Fig. 3　Sensitivity analysis of importance sampling indices

参数敏感性分析试验验证，$\kappa=0.6$ 与 $\beta=0.4$ 的超参数组合在本研究的多无人机协同扫描任务中，能够实现最优的综合性能。

3.3　训练过程分析

为深入探究不同算法在学习效率、收敛性和稳定性方面的内在差异，本节将从 2 个维度对训练过程进行动态分析：首先，通过平均回合奖励曲线评估各算法的整体性能演进；其次，借助策略熵（Policy Entropy）曲线，对基线算法在学习过程中的探索行为与策略稳定性进行剖析。

3.3.1　收敛性与性能演进分析

平均回合奖励是衡量多智能体协同策略学习效率与收敛性能的核心指标。图 4 展示了 4 种算法在 600 个训练回合内的平均奖励变化趋势。

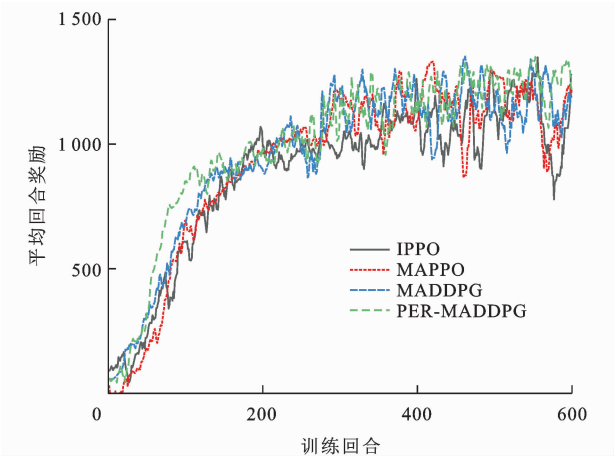

图 4　各算法的平均回合奖励学习曲线

Fig. 4　Average episode reward learning curves for each algorithm

由图 4 可见：各算法性能差异显著，PER-MADDPG 凭借优先经验回放机制聚焦高价值样本，展现出最快的收敛速度和最优的策略鲁棒性，平

均奖励稳定维持在最高水平(约 1 250);同属 Off-policy 的 MADDPG 虽能实现较高回报,但因均匀采样限制了对稀疏经验的学习效率,导致收敛稍缓且后期波动较大;相比之下,MAPPO 受限于 On-policy 样本效率,初期收敛慢且伴随剧烈振荡,而 IPPO 因独立学习范式难以应对环境非平稳性,表现最差并在后期出现显著的性能退化。

3.3.2 策略探索与不稳定性分析

为揭示不同算法在策略学习过程中的收敛特性,本文引入 Policy Entropy 作为分析指标。策略熵反映了策略分布的随机性,对于连续动作空间,其计算方式为策略概率密度函数的微分熵,具体而言,对于第 i 个智能体在时刻 t 的局部观测 $\boldsymbol{o}_{i,t}$,其策略熵 $H_{i,t}$ 计算如下

$$H_{i,t}=-\int\pi_i(\boldsymbol{a}_i\mid\boldsymbol{o}_{i,t})\lg[\pi_i(\boldsymbol{a}_i\mid\boldsymbol{o}_{i,t})\mathrm{d}\boldsymbol{a}_i] \tag{17}$$

式中:$\pi_i(\boldsymbol{a}_i\mid\boldsymbol{o}_{i,t})$为智能体 i 在时刻 t 的策略概率密度函数(对于随机策略算法为动作分布,对于确定性策略算法则包含探索噪声分布)。

如图 5 所示,策略熵分析进一步揭示了算法的收敛特性:IPPO 快速陷入低熵区间(3.75～4.00),表明其过早收敛至确定性策略导致探索不足与局部最优;反之,MAPPO 与 MADDPG 维持高熵值(大于 4.5)且伴随剧烈振荡,反映出动作选择的高随机性与策略的不稳定性。相比之下,PER-MADDPG 稳定于适中区间(4.2～4.5),既避免了盲目振荡又保留了必要探索,证明优先经验回放机制有效平衡了探索与利用矛盾,促使算法向鲁棒协同策略平滑收敛。

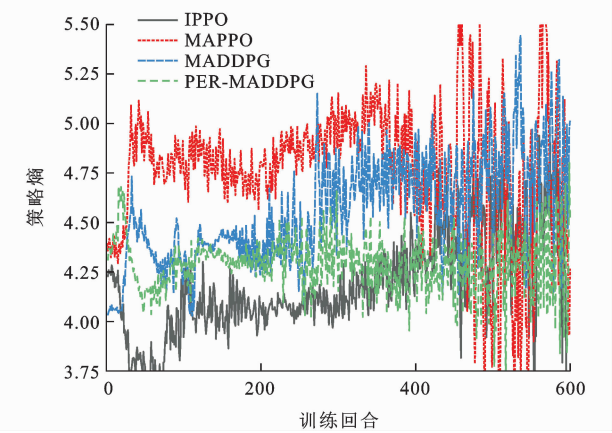

图 5 各算法的策略熵变化曲线

Fig. 5 Policy entropy variation curves for each algorithm

3.4 核心性能对比分析

为全面评估各算法在协同扫描任务中的最终性能,本节将对 PER-MADDPG 算法与 IPPO、MAPPO、MADDPG 三个基线算法在第 3.1.2 节所定义的各项评价指标上的表现进行详细的定性与定量分析。所有统计性结论均基于 100 次独立的蒙特卡洛测试结果。

3.4.1 综合性能量化对比

表 2 显示,PER-MADDPG 在各项关键指标上均显著优于基准算法。其任务成功率高达 98.0% 且碰撞率低至 1.0%,远超非协作基准 IPPO(成功率仅 35.0%,碰撞率高达 25.0%),充分验证了协同机制的必要性与策略鲁棒性。此外,该算法还实现了接近理论极限的最短完成时间(30.5 s)与最低能耗,表明其能有效兼顾高精度控制与作业效率。

表 2 各算法核心性能指标对比

Table 2 Comparison of core performance metrics for each algorithm

性能指标	IPPO	MAPPO	MADDPG	PER-MADDPG
任务成功率/%	35.0 ± 4.8	70.0 ± 4.6	85.0 ± 3.6	98.0 ± 1.4
智能体间碰撞率/%	25.0 ± 4.3	12.0 ± 3.2	8.0 ± 2.7	1.0 ± 1.0
任务完成时间/s	38.0 ± 4.0	34.0 ± 2.5	32.0 ± 1.5	30.5 ± 0.8
轨迹跟踪误差/m	0.28 ± 0.10	0.18 ± 0.06	0.12 ± 0.04	0.05 ± 0.02
队形误差/m	0.18 ± 0.08	0.10 ± 0.05	0.07 ± 0.04	0.03 ± 0.015
控制指令代价	31 500 ± 2 500	26 800 ± 1 500	24 500 ± 900	22 800 ± 650

3.4.2 飞行轨迹定性分析

为直观地展示不同算法在协同行为上的差异,图 6 呈现了 4 种算法在一个典型测试回合中的三维协同轨迹。由图 6 可见,各算法轨迹呈现显著的性能分层:独立学习基准 IPPO 的轨迹严重偏离理想路径且相对距离剧烈波动,表明其缺乏有效的协同机制;引入协同训练的 MAPPO 与 MADDPG 显著改善了轨迹平滑度与队形稳定性,验证了协同训练的有效性;而 PER-MADDPG 表现最优,其实际轨迹与理想路径高度重合且全程保持恒定的编队构型,直观证明了其在兼顾个体控制精度与团队协同稳定性方面的卓越性能。

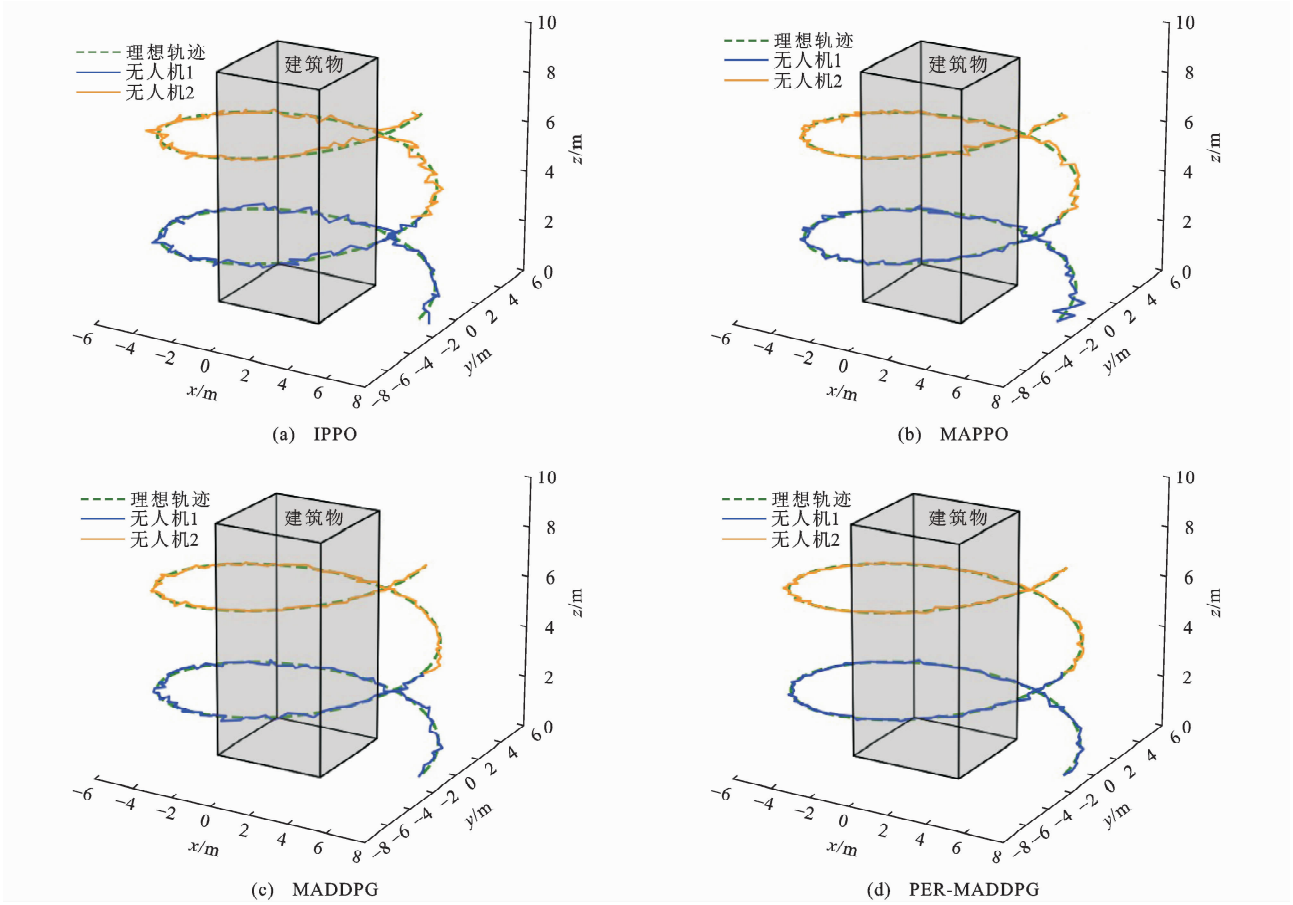

图 6　不同算法的三维协同轨迹对比

Fig. 6　Comparison of three-dimensional collaborative trajectories among different algorithms

3.4.3　误差时序量化分析

图 7 和图 8 分别展示了 4 种算法的轨迹跟踪误差和队形误差随时间变化的曲线。

由图 7 和图 8 可知，IPPO 在 2 项指标上均表现出高均值与剧烈振荡，反映其个体控制与协同能力的双重缺失；相比之下，PER-MADDPG 的轨迹跟踪误差始终维持在最低水平，队形误差更是收敛至近零状态，定量证明了该方法在保障个体高精度飞行的同时，实现了极高稳定性的编队协同。

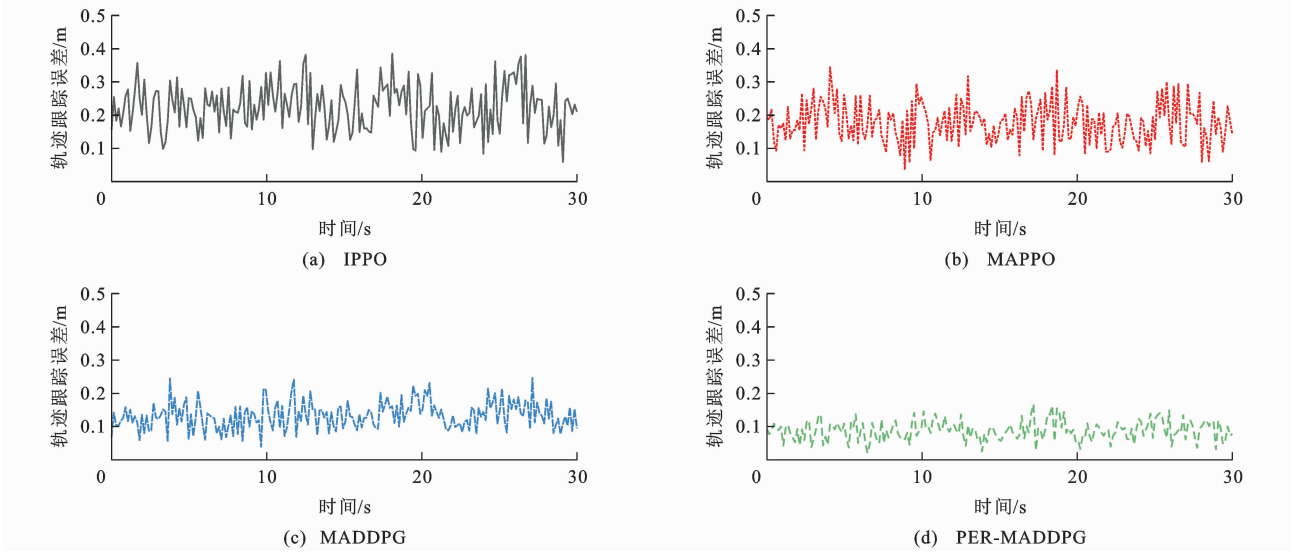

图 7　各算法的轨迹跟踪误差时序对比

Fig. 7　Comparison of times series of trajectory tracking errors for each algorithm

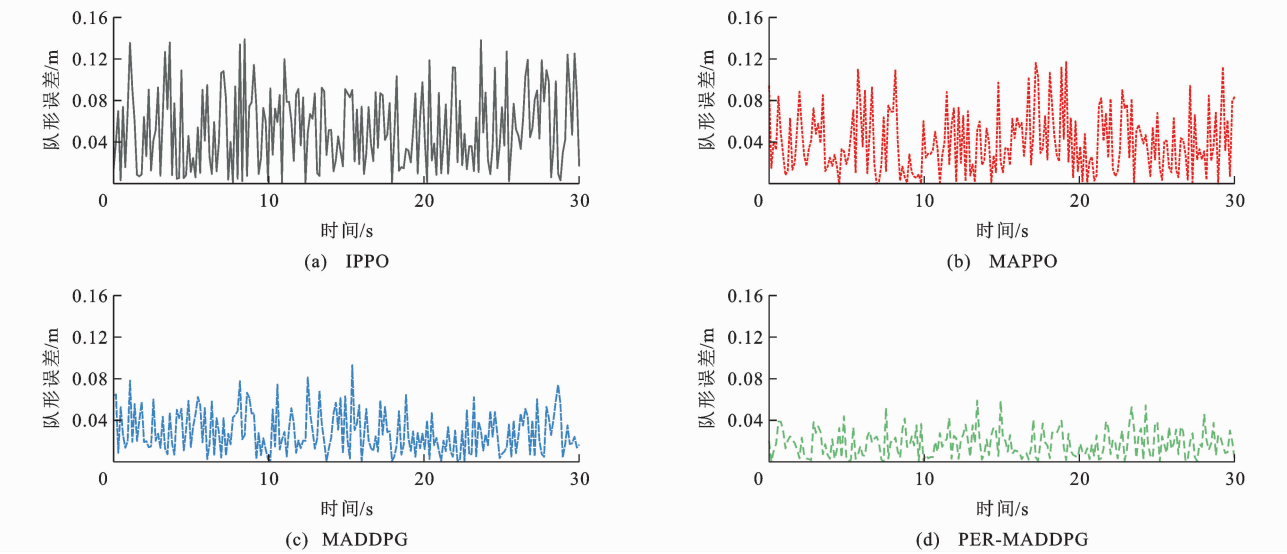

图 8 各算法的队形误差时序对比

Fig. 8 Comparison of time series of formation errors for each algorithm

3.4.4 性能统计分布分析

在对各算法的性能进行统计分布分析时，需要特别说明的是，任务完成时间指标与前序误差时序图(图 7、8)中的 30 s 时间轴的关系。误差时序图的横坐标是一个标准化的“参考时间轴”，其目的是在同一理想时间基准下，公平地对比各算法在任意时刻的瞬时控制性能，然而，任务完成时间则是一个衡量最终结果的实际物理时间。

为此，从统计学层面验证算法的性能与鲁棒性，图 9 展示了基于 100 次独立测试结果绘制的箱形图，对比了各算法在“任务完成时间”和“轨迹跟踪误差”2 个核心指标上的数据分布。无论是任务完成时间还是轨迹跟踪误差，4 个算法的性能都呈现出清晰的阶梯式提升。对于本文提出的 PER-MADDPG 算法，其中位数在 2 个指标上均处于最优位置(完成时间最短，跟踪误差最低)，证明了其平均性能的优越性，且箱体四分位距(IQR)在 2 个指标上均为最窄，这表明其在 100 次测试中的性能波动最小，展现出极高的稳定性和鲁棒性。

3.5 可扩展性与性能边界分析

为验证 PER-MADDPG 算法在更大规模无人机编队中的适用性，并探究其在复杂交互环境下的性能边界，本节在原有双机试验的基础上，将编队规模扩展至 $N=4$ 和 $N=6$，开展了进一步的对比测试。

3.5.1 不同规模性能与效率分析

图 10 展示了不同无人机数量下，各算法在任务成功率与系统搜寻效率的变化趋势。

如图 10(a)所示，尽管物理空间拥挤效应导致所有算法的任务成功率随编队规模扩大而下降，

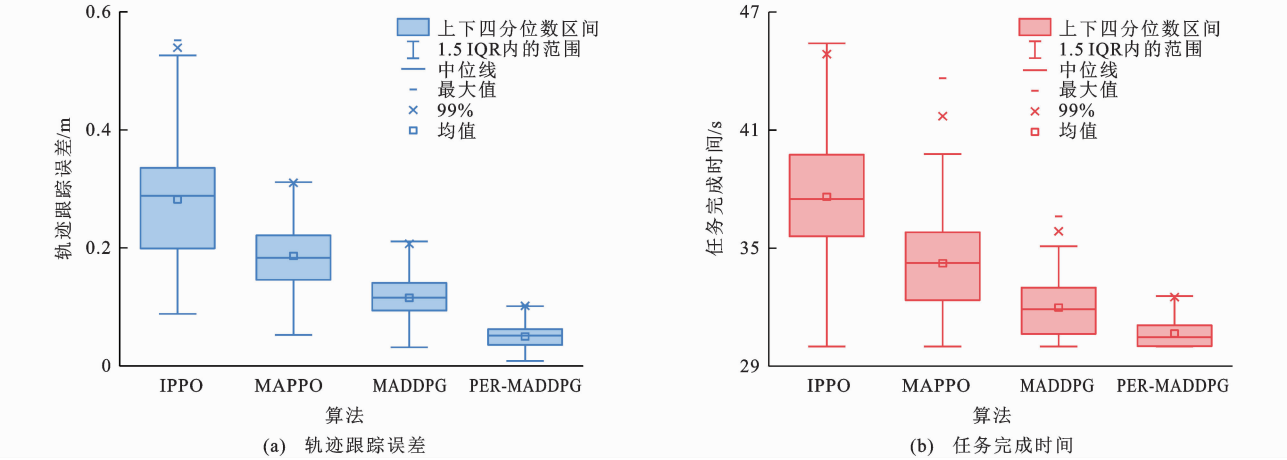

图 9 各算法核心性能指标的统计分布箱形图

Fig. 9 Box plot of statistical distributions for core performance metrics of each algorithm

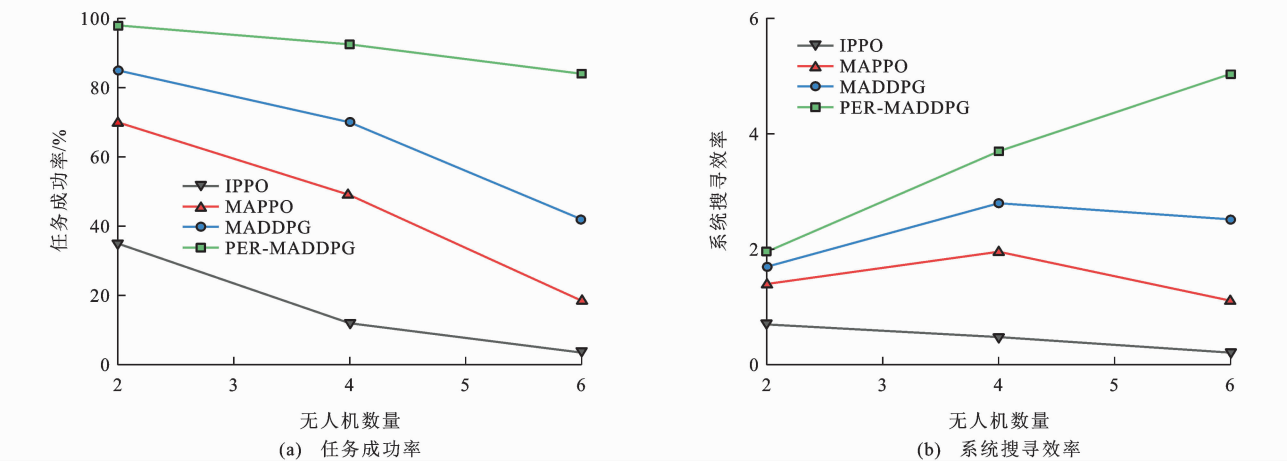

图 10　不同编队规模下的任务成功率与系统搜寻效率对比

Fig. 10　Comparison of task success rate and system search efficiency under different formation sizes

PER-MADDPG 在 $N=6$ 的极限场景下仍维持 84.0% 的高成功率（远优于基准 MADDPG 的 42.0%），展现出最优的抗干扰鲁棒性。为更全面地衡量编队规模扩大的实际收益，本文引入系统搜寻效率作为评价指标，定义为单位时间内有效完成任务的无人机期望数量。如图 10(b)所示，该方法有效克服了拥塞崩溃风险，实现了近乎线性的协同增益；反观基准 MADDPG 在 $N=6$ 时因碰撞激增已出现效率显著负增长，MAPPO 与 IPPO 则更早进入衰退期。

然而，值得注意的是，编队规模的选取并非“多多益善”，而需在系统效率、控制精度与部署成本之间进行权衡。试验数据显示，当编队从双机扩展至四机时，系统搜寻效率提升了近一倍（1.96→3.70），而单机轨迹跟踪误差仅有微小增加（0.03 m→0.05 m），任务成功率仍保持在 92.5% 的高可靠区间。相比之下，当进一步扩展至六机时，虽然总效率提升至 5.04，但受限于物理空间的极度收窄，单机平均误差上升至 0.08 m，且任务成功率下降至 84.0%。根据边际效用递减法则，在本文设定的高层建筑搜救场景下，四机编队表现出了最优的效费比，是兼顾作业效率与单机控制品质的最佳平衡点。在实际应用中，决策者应根据任务对时效性与精度的具体优先级，参照此规律灵活选择编队规模。

3.5.2　四机协同轨迹的定性分析

为直观评估多机编队的协同效果，图 11 和图 12 分别展示了 $N=4$ 时 MADDPG 与 PER-MADDPG 的三维轨迹与侧视轨迹对比。

如图 11 可知，PER-MADDPG[图 11(b)]控制下的 4 架无人机轨迹平滑且紧密贴合理想螺旋线（虚线），且 4 条轨迹之间保持了清晰的间隔。反观 MADDPG[图 11(a)]，其轨迹在螺旋上升过程中出现了明显的抖动和局部偏离，反映了智能体在处理复杂的邻居避碰时决策的不稳定性。

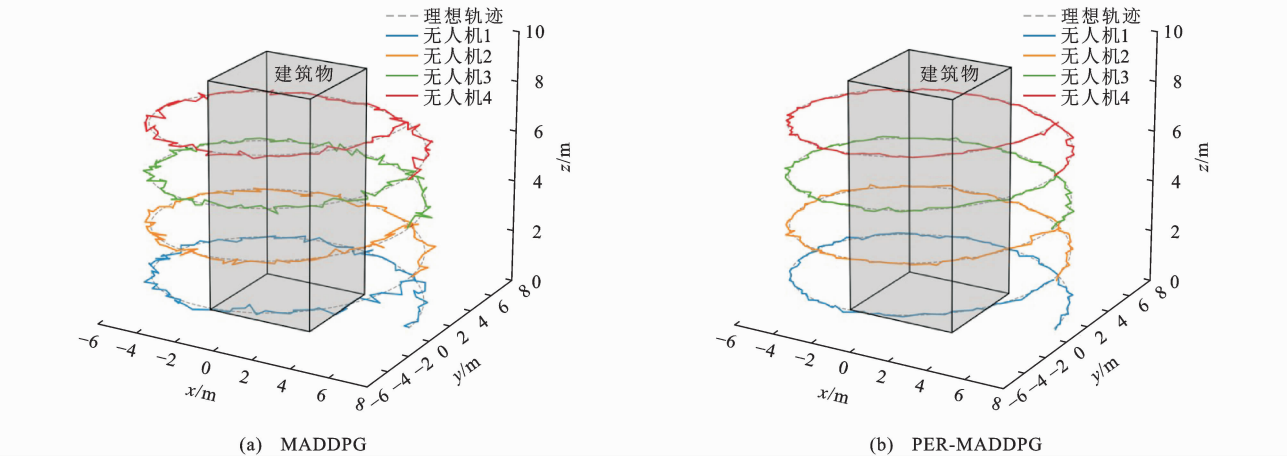

图 11　四机编队协同螺旋扫描三维轨迹对比

Fig. 11　Comparison of three-dimensional trajectories for collaborative spiral scanning of four-UAV formation

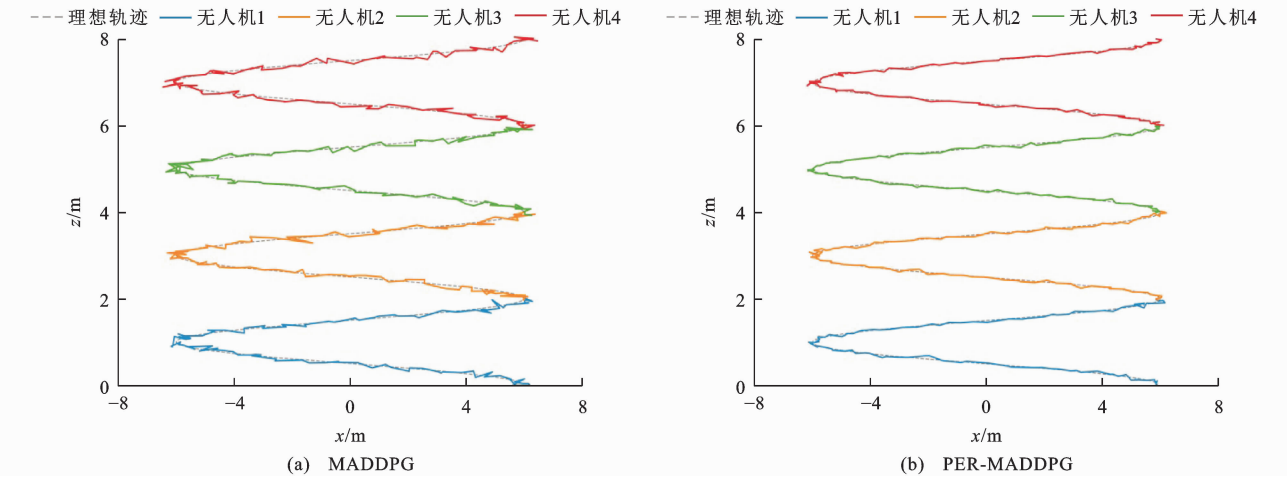

图 12 四机编队轨迹跟踪侧视图对比

Fig. 12 Side-view comparison of trajectory tracking for four-UAV formation

而图 12 的侧视图进一步揭示了垂直方向的协同效果。PER-MADDPG 的轨迹在垂直高度上分层清晰，4 架无人机始终保持同步上升的姿态，未发生高度层面的混淆或干扰。

3.5.3 *误差分布与精度分析*

图 13 通过箱形图进一步量化了不同规模下的轨迹跟踪误差分布。

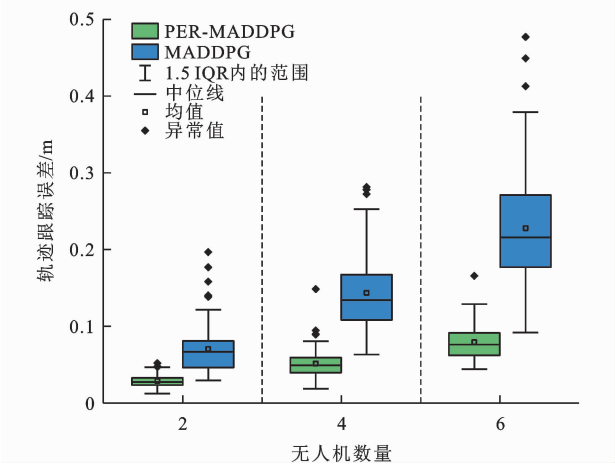

图 13 不同编队规模下的轨迹跟踪误差统计分布

Fig. 13 Statistical distribution of trajectory tracking errors under different formation sizes

从统计结果可以看出，随着 N 从 2 增加到 6，所有算法的误差中位数和离散度均呈上升趋势。然而，PER-MADDPG(绿色箱体)在所有测试组中均表现最优：即使在 $N=6$ 的高难度组，PER-MADDPG 的平均误差(约 0.08 m)仍低于 MADDPG 在 $N=4$ 时的水平(约 0.14 m)；PER-MADDPG 的异常值显著少于 MADDPG，说明其在应对极端拥挤状况时，能够更稳定地输出安全控制策略，极少出现严重的轨迹失控。综上所述，扩展性试验有力证明了 PER-MADDPG 算法不仅适用于双机协同，在更复杂的四机及六机编队任务中同样具备协同控制能力与鲁棒性。

4 结 语

(1)本文针对多智能体强化学习在处理无人机协同任务中稀疏关键事件效率低下的问题，提出了 PER-MADDPG 方法。该方法通过一个中心化评论家网络评估团队的联合表现，并以此为依据优先学习近距离避碰、队形恢复等关键经验，从而提升了无人机编队学习复杂协同策略的效率与安全性。提出的协同控制策略，能提升无人机编队在应急搜救等高风险任务中的自主性与可靠性。稳定的飞行控制不仅保障了任务安全，也能确保机载设备获取高质量的灾情数据，为实现高效、智能的无人化应急响应提供了关键技术支撑。

(2)扩展性分析表明，随着编队数量增加，物理空间的拥挤效应虽然会增加控制难度，但 PER-MADDPG 算法在四机及六机编队中仍能有效维持协同构型，其性能衰减速率显著低于基准算法，证明了该策略在处理状态空间维度膨胀时的有效性。然而，当前研究仍存在“仿真到现实”的鸿沟，且仿真环境相对理想化，并未充分考虑真实世界的通信延迟与复杂物理干扰。此外，算法在更大规模无人机集群下的可扩展性也有待进一步验证。

(3)未来的研究将重点围绕 3 个方向展开：一是开展虚实迁移研究，将算法部署至物理平台；二是在更复杂、不确定的动态环境中测试算法的鲁棒性；三是探索更具可扩展性的 MARL 框架，以支持更大规模的无人机集群协同作业。

参考文献：

References：

[1] 陈德启，张自设，张文会，等. 面向高层建筑应急救援的无人机螺旋搜索轨迹控制方法[J]. 交通运输系统工程与信息，2025，25(6)：87-100.
CHEN De-qi，ZHANG Zi-she，ZHANG Wen-hui，et al. Trajectory control method for UAV spiral search oriented to high-rise building emergency rescue[J]. Journal of Transportation Systems Engineering and Information Technology，2025，25(6)：87-100.

[2] LI C，CHANG Q，FAN H T. Multi-agent reinforcement learning for integrated manufacturing system-process control[J]. Journal of Manufacturing Systems，2024，76：585-598.

[3] ZHANG K Q，YANG Z R，BAŞAR T. Multi-agent reinforcement learning：A selective overview of theories and algorithms[M]. Handbook of Reinforcement Learning and Control. Cham：Springer International Publishing，2021：321-384.

[4] LI T X，ZHU K，LUONG N C，et al. Applications of multi-agent reinforcement learning in future Internet：A comprehensive survey[J]. IEEE Communications Surveys & Tutorials，2022，24(2)：1240-1279.

[5] 伍景琼，陈子伟，岑明睿，等. 无人机配送模式及关键技术研究综述[J]. 交通信息与安全，2025，43(3)：112-127.
WU Jing-qiong，CHEN Zi-wei，CEN Ming-rui，et al. A review of drone delivery models and key technologies[J]. Journal of Transport Information and Safety，2025，43(3)：112-127.

[6] 刘京奥，姜晓爱，王永超，等. 复杂环境下旋翼无人机集群协同多目标跟踪[J/OL]. 北京航空航天大学学报，2025，https：//doi.org/10.13700/j.bh.1001-5965.2025.0621.
LIU Jing-ao，JIANG Xiao-ai，WANG Yong-chao，et al. Cooperative multi-target aerial tracking for multicopter swarm in cluttered environment[J/OL]. Journal of Beijing University of Aeronautics and Astronautics，2025，https：//doi.org/10.13700/j.bh.1001-5965.2025.0621.

[7] KANG Y，DI J，LI M，et al. Autonomous multi-drone racing method based on deep reinforcement learning[J]. Science China Information Sciences，2024，67(8)：180203.

[8] GUAN Y，ZOU S，PENG H X，et al. Cooperative UAV trajectory design for disaster area emergency communications：A multiagent PPO method[J]. IEEE Internet of Things Journal，2024，11(5)：8848-8859.

[9] YU C，VELU A，VINITSKY E，et al. The surprising effectiveness of PPO in cooperative，multi-agent games[EB/OL]. 2021，arXiv：2103.01955.

[10] 盖思博，马　蓓，郑　莉，等. 多智能体强化学习框架下机器人集群通信技术综述[J]. 兵工学报，2026，47(1)：3-19.
GAI Si-bo，MA Bei，ZHENG Li，et al. A survey of communication technology for MARL-based robot swarm[J]. Acta Armamentarii，2026，47(1)：3-19.

[11] 陈冠良，刘　义，余　意. 异构多智能体强化学习驱动的无人机三维避障与边缘计算协同优化[J/OL]. 计算机应用，2025，https：//link.cnki.net/urlid/51.1307.TP.20251216.1320.002.
CHEN Guan-liang，LIU Yi，YU Yi. Heterogeneous multi-agent reinforcement learning enabled co-optimization of UAV 3D obstacle avoidance and edge computing[J/OL]. Journal of Computer Applications，2025，https：//link.cnki.net/urlid/51.1307.TP.20251216.1320.002.

[12] AN T X，LEE J，BJELONIC M，et al. Scalable multi-robot cooperation for multi-goal tasks using reinforcement learning[J]. IEEE Robotics and Automation Letters，2025，10(2)：1585-1592.

[13] LOWE R，WU Y，TAMAR A，et al. Multi-agent actor-critic for mixed cooperative-competitive environments[EB/OL]. 2017，arXiv：1706.02275.

[14] ZHAO G D，WANG Y，MU T，et al. Reinforcement-learning-assisted multi-UAV task allocation and path planning for IIoT[J]. IEEE Internet of Things Journal，2024，11(16)：26766-26777.

[15] WU J H，LI D Y，YU Y Z，et al. An attention mechanism and adaptive accuracy triple-dependent MADDPG formation control method for hybrid UAVs[J]. IEEE Transactions on Intelligent Transportation Systems，2024，25(9)：11648-11663.

[16] HOOK J，DE SILVA V，KONDOZ A. Deep multi-critic network for accelerating policy learning in multi-agent environments[J]. Neural Networks，2020，128：97-106.

[17] 吴　旭，赵中原. 基于 NRO-QMIX 的战场环境多无人机协同目标搜索[J]. 山东理工大学学报(自然科学版)，2025，39(6)：32-40，49.
WU Xu，ZHAO Zhong-yuan. Multi-UAV cooperative target search in battlefield environments based on NRO-QMIX[J]. Journal of Shandong University of Technology (Natural Science Edition)，2025，39(6)：32-40，49.

[18] RASHID T，SAMVELYAN M，DE WITT C S，et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[EB/OL]. 2020，arXiv：2003.08839.

[19] DING R J，CHEN J W，WU W，et al. Packet routing in dynamic multi-hop UAV relay network：A multi-agent learning approach[J]. IEEE Transactions on Vehicular Technology，2022，71(9)：10059-10072.

[20] 魏　麟，杨济睿，李秀易，等. 面向融合运行的飞行员在环建模技术综述[J]. 交通运输工程学报，2024，24(4)：208-227.
WEI Lin，YANG Ji-rui，LI Xiu-yi，et al. Review on pilot-in-the-loop modeling techniques facing integrated operation[J]. Journal of Traffic and Transportation Engineering，2024，24(4)：208-227.

[21] RASHID T，FARQUHAR G，PENG B，et al. Weighted QMIX：expanding monotonic value function factorisation for deep multi-agent reinforcement learning[EB/OL]. 2020，arXiv：2006.10800.

[22] KANG H Y，CHANG X L，MIŠIĆ J，et al. Cooperative UAV resource allocation and task offloading in hierarchical aerial computing systems：A MAPPO-based approach[J].

IEEE Internet of Things Journal, 2023, 10(12): 10497-10509.

[23] 关 巍,胡彤博,张显库,等.基于改进 PPO 算法的无人机/无人船协同导航方法[J/OL].系统工程与电子技术,2025, https://link.cnki.net/urlid/11.2422.TN.20251106.1955.046.
GUAN Wei, HU Tong-bo, ZHANG Xian-ku, et al. Cooperative navigation method for UAV/ USV based on the improved PPO algorithm[J/OL]. Systems Engineering and Electronics, 2025, https://link.cnki.net/urlid/11.2422.TN.20251106.1955.046.

[24] TIAN J J, JIA H F, WANG G F, et al. Optimal scheduling of shared autonomous electric vehicles with multi-agent reinforcement learning: A MAPPO-based approach [J]. Neurocomputing, 2025, 622: 129343.

[25] DAI S H, LI S K, TANG H C, et al. MARP: A cooperative multiagent DRL system for connected autonomous vehicle platooning[J]. IEEE Internet of Things Journal, 2024, 11(20): 32454-32463.

[26] WU X, YAN Q Z, WANG J C, et al. Dynamic task allocation for UAV swarms in maritime rescue scenarios based on PG-MAPPO[J]. IEEE Internet of Things Journal, 2025, 12(18): 38073-38087.

[27] ZHONG R J, ZHANG D H, SCHÄFER L, et al. Robust on-policy sampling for data-efficient policy evaluation in reinforcement learning[EB/OL]. 2021, arXiv: 2111.14552.

[28] 骞晨旭,张雪波,李 论,等.面向智能空中博弈的大语言模型-强化学习分层决策算法[J].控制与决策,2026,41(3): 855-864.
QIAN Chen-xu, ZHANG Xue-bo, LI Lun, et al. Research on LLM-RL hierarchical decision-making algorithm for intelligent aerial combat[J]. Control and Decision, 2026, 41(3): 855-864.

[29] HU G Z, ZHU Y H, ZHAO D B, et al. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 3966-3978.

[30] 谢海文.多智能体分布式协同任务分配策略设计[D].北京:北方工业大学,2025.
XIE Hai-wen. Multi-agent distributed cooperative task allocation strategy design [D]. Beijing: North China University of Technology, 2025.

[31] ORR J, DUTTA A. Multi-agent deep reinforcement learning for multi-robot applications: A survey[J]. Sensors, 2023, 23(7): 3625.