控制与决策
Control and Decision
ISSN 1001-0920,CN 21-1124/TP

《控制与决策》网络首发论文

题目：基于 MAPPO 的多无人机协同分布式动态任务分配
作者：李海峰，杨宏安，盛梓茂，刘超，陈逸新
DOI：10.13195/j.kzyjc.2024.0784
收稿日期：2024-07-02
网络首发日期：2024-10-23
引用格式：李海峰，杨宏安，盛梓茂，刘超，陈逸新．基于 MAPPO 的多无人机协同分布式动态任务分配[J/OL]．控制与决策．
https://doi.org/10.13195/j.kzyjc.2024.0784

网络首发时间：2024-10-23 10:23:54
网络首发地址：https://link.cnki.net/urlid/21.1124.TP.20241022.0657.002

基于 MAPPO 的多无人机协同分布式动态任务分配

李海峰，杨宏安†，盛梓茂，刘 超，陈逸新

(西北工业大学 机电学院，陕西 西安 710000)

摘 要: 针对多无人机在高动态近距空战中自主决策困难且协同性差等问题, 提出了一种基于 MAPPO 的多无人机分布式动态任务分配方法. 首先, 考虑任务可执行约束和无人机载荷约束, 建立以敌方无人机为目标、攻击战术为任务的多无人机动态任务分配模型; 其次, 设计了包含分离式状态滑动标准化机制、动作屏蔽机制以及注意力机制的任务重分配网络, 该网络可有效处理 MAPPO 算法在状态滑动标准化过程中的信息失真问题, 并确保任务分配过程严格满足任务约束, 同时可基于攻击目标专注于攻击战术的选择, 实现多无人机的协同分布式动态任务分配; 最后, 在 3v3 近距空战场景中, 搭载所提算法的我方无人机与搭载空战决策专家系统的敌方无人机进行空战对抗, 其作战胜率高达 98.5%, 所得结果验证了该方法的有效性.

关键词: 多无人机；动态任务分配；近距空战；MAPPO；分布式

中图分类号: V279 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0784

引用格式: 李海峰, 杨宏安, 盛梓茂, 等. 基于 MAPPO 的多无人机协同分布式动态任务分配 [J]. 控制与决策, 2024, 39(0): xxxx-xxxx.

Multi-UAV Collaborative Distributed Dynamic Task Allocation Based on MAPPO

LI Hai-feng，*YANG Hong-an*†，*SHENG Zi-mao*，*LIU Chao*，*CHEN Yi-xin*

(School of Mechanical Engineering，Northwestern Polytechnical University，Xi'an 710000 ,China)

Abstract: To address the issues of autonomous decision-making difficulties and poor coordination in highly dynamic close-range air combat involving multiple UAVs, a distributed dynamic task allocation method based on MAPPO is proposed. Firstly, considering the executable constraints of tasks and the payload constraints of UAVs, a dynamic task allocation model for multi-UAV is established, targeting enemy UAVs and attack tactics as tasks; Secondly, a task reallocation network is designed, incorporating a seperated state sliding normalization mechanism, action mask mechanism, and attention mechanism. This network effectively addresses the information distortion issues that occur during the state sliding normalization process in the MAPPO algorithm, ensuring that the task allocation process strictly adheres to task constraints. Additionally, it focuses on the selection of attack tactics based on attack target, enabling distributed dynamic task allocation for multi-UAV; Finally, in a 3v3 air combat scenario, our UAVs equipped with the proposed algorithm engage in air combat against enemy UAVs equipped with an expert system, The combat victory rate of our UAVs reaches 98.5%, validating the effectiveness of the proposed method.

Keywords: multi-UAV；dynamic task allocation；close range air combat；MAPPO；distributed

0 引 言

在复杂多变的战场中, 无人机技术的迅猛发展为现代军事战略带来了巨大变革[1-2]. 然而, 单一无人机存在信息感知能力有限、任务执行效率低等不足, 因此, 多无人机协同作战在未来战场中拥有广阔的应用前景[3]. 多无人机的任务分配技术是无人机高效、可靠完成任务的重要技术支撑[4-6]. 在高动态复杂近距空战场景中, 静态任务分配难以应对快速变化的战场态势, 集中式任务分配受限于中心节点的决策速度, 两者均可能导致任务分配的滞后和失败. 因此, 亟需多无人机系统能够根据当前战场态势自主调整任务, 实现分布式动态任务分配, 以协同打击

收稿日期: 2024-07-02；**录用日期:** 2024-10-15.
基金项目: 国家自然科学基金面上项目: 可编程的集群机器人自主鲁棒成型 (51775435).
责任编委: 谢晖.
†通讯作者. E-mail: yhongan@nwpu.edu.cn
本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

敌方军事目标.

目前, 多无人机分布式任务分配算法的研究主要集中于基于市场机制的拍卖算法、合同网算法等. 文献 [7-13] 针对不同类型的无人机系统和任务需求, 提出了多种分布式任务分配方法, 包括改进的合同网协议、带有共享存储中心的分布式拍卖算法、考虑任务时间窗和无人机能力约束的分布式协同任务分配方法等. 然而, 上述算法普遍存在灵活性不足、计算和通信开销大等局限性.

强化学习作为一种前沿的人工智能算法, 具有实时性高、适应性强等特点, 在无人机空战智能决策中已有较多研究成果. 文献 [14-18] 针对不同类型的空战场景和战机动作空间, 提出了基于 DQN、PPO 等强化学习算法的智能决策方法. 然而, 这些研究大多集中在 1v1 的空战场景, 对多机协同空战的研究相对较少, 且存在稀疏奖励和动作空间巨大等问题.

本文将无人机空战问题看作是无人机在高度动态、激烈对抗的环境中的在线任务重规划问题, 而分布式任务重分配是在线任务重规划的首要任务. 将多机空战问题分解为任务分配、航迹规划和轨迹跟踪三个阶段, 并基于 MAPPO[19] 算法, 在多无人机近距空战的背景下, 考虑任务可执行约束和无人机载荷约束, 提出一种在高动态空战场景下的多无人机分布式动态任务分配方法. 主要创新点包括:

(1) 设计了一种分离式状态滑动标准化机制, 有效提高了 MAPPO 算法的奖励收敛值;

(2) 引入动作掩码机制[20], 解决了特定空战场景下的任务约束挑战, 确保任务分配决策的有效性;

(3) 设计基于注意力机制[21] 的攻击战术决策网络, 提高了无人机战术分配的合理性.

1 问题描述与建模

1.1 任务场景

在多无人机近距空战中, 无人机采用分布式架构, 无统一指挥中心, 通过交流协商的方式完成任务分配. 在不考虑障碍物时, 设有N架我方无人机$\mathbf{U}=\{u_1,u_2,\ldots,u_N\}$对阵$M$架敌方无人机$\mathbf{E}=\{e_1,e_2,\ldots,e_M\}$. 为突出核心算法, 本文简化文献 [16] 中的 14 种机动动作, 并修改其为 3 种对敌战术动作$\mathbf{T}=\{t_1,t_2,t_3\}$, 分别为左突袭、右突袭和尾后攻击. 这些战术分别指我方无人机从目标的正左、正右或尾部发起攻击.

多无人机近距空战流程如下: 任意时刻, $u_i\in\mathbf{U}$首先根据自身对环境的观测信息$\mathbf{o}_i$进行任务重分配, 得到攻击目标$T_{\mathrm{tgt},i}$和攻击战术$T_{\mathrm{tac},i}$; 然后根据任务分配结果, 利用 Dubins 曲线[22] 进行航迹规划, 得到航迹$\mathbf{p}_i$; 随后采用 DWA[23] 进行局部航迹规划, 跟踪该 Dubins 路径并避免与其余无人机发生碰撞. 鉴于任务重分配是本文探讨的重点, 因此本文对于航迹规划和轨迹跟踪部分的细节不作过多的描述.

1.2 任务约束

由多智能体强化学习 (Multi-Agent Reinforcement Learning, MARL) 策略驱动的多无人机空战中, 各无人机需每步决策, 并考虑战斗状态、弹药量等约束. 为了方便描述各项约束条件, 引入决策变量X_{ij}和Y_{ik}, 分别表示u_i是否攻击e_j、u_i是否采取战术t_k. 当$X_{ij}=0$, $Y_{jk}=0$时, 表示u_i不执行任何任务; 当$X_{ij}=1$, $Y_{jk}=1$时, 表示u_i以战术t_k攻击e_j.

1.2.1 任务可执行约束

在以下两种情况下, 无人机u_i无法执行任务:

(1) 当$u_i\in\mathbf{U}$被摧毁时 ($D_i=1$), 它在任何时间步都无法继续参与决策. 这需在多智能体强化学习策略更新中明确处理, 避免干扰任务分配. 约束表示为

$$C_1:\begin{cases}X_{ij}=0\\ Y_{ik}=0\end{cases},\quad \forall e_j\in\mathbf{E},\forall t_k\in\mathbf{T},D_i=1. \tag{1}$$

(2) 当$e_j\in\mathbf{E}$被摧毁时 ($D_j=1$), 我方任何无人机u_i都不应选择e_j为攻击目标. 这对于确保决策的有效性至关重要, 避免无人机浪费资源或做出无效决策. 约束表示为

$$C_2:X_{ij}=0,\forall u_i\in\mathbf{U},D_j=1. \tag{2}$$

1.2.2 无人机载荷约束

在多无人机空战中, 弹药量是执行任务的关键约束. 设无人机u_i携带w_i枚导弹, 当弹药耗尽时, 无人机无法继续执行攻击任务. 约束表示为

$$C_3:\begin{cases}X_{ij}=0\\ Y_{ik}=0\end{cases},\quad w_i=0. \tag{3}$$

引入这些约束后, 无人机能基于实际状态和资源做出合理任务分配, 提升任务分配的合理性和有效性. 经分析, 在u_i无法攻击e_j时, 即$X_{ij}=0$, 其对应的战术决策变量必为$Y_{ik}=0$, 其中$t_k\in\mathbf{T}$. 所以, 约束$C_1\sim C_3$均可视为对无人机攻击目标的约束.

1.3 无人机简化运动学模型

无人机运动是一个复杂的非线性动态过程, 若在多无人机动态任务分配中引入过于复杂的运动模型, 将显著提升求解难度, 且往往并非必要. 因此, 选取一个恰当的无人机运动模型显得尤为重要. 在任务分配问题中, 可将无人机运动模型认为是 Dubins 模型. Dubins 模型是无人机在定高等速飞行且不考虑障碍物假设下的简化模型, 能够刻画无人机的转

弯约束, 在无人机协同任务分配问题的研究中得到了广泛的应用. 因此, 在无障碍物的空域环境下, 可以将无人机运动学模型简化到二维平面内. 所以, 无人机的运动学模型为

$$\begin{cases} \dot{x}_i^{(t)} = v_{ix}^{(t)} = v_i \cos\theta_i^{(t)} \\ \dot{y}_i^{(t)} = v_{iy}^{(t)} = v_i \sin\theta_i^{(t)} \\ \dot{\theta}_i^{(t)} = \omega_i^{(t)} \\ |\omega_i^{(t)}| \leqslant \omega_{\max} \end{cases}. \tag{4}$$

对于t时刻的无人机u_i, $x_i^{(t)}$, $y_i^{(t)}$分别表示其沿x, y轴方向的位置; $v_{ix}^{(t)}$, $v_{iy}^{(t)}$分别表示其沿x, y轴方向的速度, 速度大小恒定为v_i; $\theta_i^{(t)}$, $\omega_i^{(t)}$分别表示其航向角以及航向角角速度. 此外, 无人机仅能攻击位于其前方雷达探测范围内的敌方目标, 该雷达的探测范围呈圆锥形, 其半径为l, 角度为ζ.

2　基于 MAPPO 的多无人机分布式动态任务分配

2.1　MAPPO 算法

近端策略优化算法 (Proximal Policy Optimization, PPO) 是一种基于策略梯度算法 (Policy Gradient, PG)[24] 和信赖域策略优化算法 (Trust Region Policy Optimazation, TRPO)[25] 的改进算法. PPO 算法通过引入一个裁剪机制来限制策略更新的幅度, 解决了 PG 中的稳定性和收敛性问题、简化了 TRPO 中的计算过程. MAPPO 算法是 PPO 算法的多智能体版本, 其在训练阶段利用全局信息进行集中式训练, 而在执行阶段仅利用局部观测信息分布式执行, 其能够在复杂的多智能体场景中提高学习效率和稳定性. 因此, 本文采用 MAPPO 进行多无人机分布式动态任务分配.

2.2　MAPPO 基本要素

MAPPO 主要基于部分可观测马尔可夫决策过程 (Partially Observable Markov Decision Process, POMDP)[26] 对多个智能体在动态环境中的交互和决策过程进行建模. 因此, 将多无人机近距空战分布式动态任务分配问题转换为基于 POMDP 的序贯决策问题, 定义 MAPPO 的基本要素如下.

2.2.1　联合观测O

联合观测描述了当前所有参加任务的我方无人机对环境中其他所有无人机观测的笛卡尔积

$$\mathbf{O} = \mathbf{O}_1 \times \mathbf{O}_2 \times \ldots \times \mathbf{O}_N \tag{5}$$

其中$\mathbf{O}_i$为u_i的观测空间. $\mathbf{O}_i$可以表述为

$$\mathbf{O}_i = (\mathbf{O}_{i1}, \mathbf{O}_{i2}, \ldots, \mathbf{O}_{i(N+M)}) \tag{6}$$

其中$\mathbf{O}_{ij}$表示u_i对无人机j的观测空间, $N+M$为环境中所有无人机数目. 对于u_i, 其在t时刻对第j个无人机的观测量$\mathbf{o}_{ij}^{(t)} \in \mathbf{O}_{ij}$可以表述为

$$\mathbf{o}_{ij}^{(t)} = \begin{cases} (x_{ij}^{(t)}, y_{ij}^{(t)}, \theta_{ij}^{(t)}, T_{\text{tgt},j}^{(t)}, T_{\text{tac},j}^{(t)}, D_j^{(t)}), & j \in \mathbf{U} \\ (x_{ij}^{(t)}, y_{ij}^{(t)}, \theta_{ij}^{(t)}, D_j^{(t)}), & j \in \mathbf{E} \end{cases} \tag{7}$$

每架无人机均基于观测量分析当前战场态势, 从而自主做出决策, 而不依赖于中心节点, 以达到协同摧毁敌机的目的, “分布式” 任务分配由此而来.

2.2.2　全局状态S

环境的全局状态$\mathbf{S}$为所有我方无人机局部观测信息的拼接, t时刻, $\mathbf{s}^{(t)} \in \mathbf{S}$为

$$\mathbf{s}^{(t)} = (\mathbf{o}_1^{(t)}, \mathbf{o}_2^{(t)}, \ldots, \mathbf{o}_N^{(t)}). \tag{8}$$

2.2.3　联合动作A

联合动作$\mathbf{A}$描述了当前所有参加作战任务的我方所有无人机的动作的笛卡尔积

$$\mathbf{A} = \mathbf{A}_1 \times \mathbf{A}_2 \times \ldots \times \mathbf{A}_N. \tag{9}$$

其中$\mathbf{A}_i$表示u_i的动作空间. 对于u_i, 其在t时刻的动作$\mathbf{a}_i^{(t)} \in \mathbf{A}_i$为

$$\mathbf{a}_i^{(t)} = (T_{\text{tgt},i}, T_{\text{tac},i}), T_{\text{tgt},i} = e_j \in \mathbf{E}, T_{\text{tac},i} = t_k \in \mathbf{T}. \tag{10}$$

所以t时刻无人机的联合动作$\mathbf{a}^{(t)} \in \mathbf{A}$为

$$\mathbf{a}^{(t)} = (\mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)}, \ldots, \mathbf{a}_N^{(t)}). \tag{11}$$

2.2.4　全局奖励函数R

全局奖励函数R定义为: 在全局状态$\mathbf{s}^{(t)}$下, 所有我方无人机执行联合动作$\mathbf{a}^{(t)}$, 无人机团队所获得的即时收益. R对于无人机系统的行为和决策起着关键的引导作用.

(1) 目标、战术切换惩罚. 对无人机频繁切换目标或针对同一目标频繁切换战术进行惩罚 (分别为$r_1^{(t)}$和$r_2^{(t)}$), 确保任务执行的稳定性和连贯性, 提高作战协同性. 通常, 该惩罚项的重要性因子设置较低, 以确保其在必要时具有应对动态变化战场的能力.

$$\begin{cases} r_1^{(t)} = -\sum_{i=1}^{N} I(T_{\text{tgt},i}^{(t)} \neq T_{\text{tgt},i}^{(t-\Delta t)}) \\ r_2^{(t)} = -\sum_{i=1}^{N} I(T_{\text{tac},i}^{(t)} \neq T_{\text{tac},i}^{(t-\Delta t)} \wedge T_{\text{tgt},i}^{(t)} = T_{\text{tgt},i}^{(t-\Delta t)}) \end{cases}. \tag{12}$$

其中, $I(x)$为指示函数, 当条件x成立时, $I(x) = 1$, 否则, $I(x) = 0$.

(2) 航迹质量惩罚. 鼓励无人机就近选择平滑的攻击路线

$$r_3^{(t)} = \sum_{i=1}^{N} ((l_i^{(t)} + l_i'^{(t)}) - (l_i^{(t-\Delta t)} - l_i'^{(t-\Delta t)})). \tag{13}$$

其中, $l_i^{(t)}$和$l_i'^{(t)}$分别为针对u_i在t时刻分配方案的航

迹总长度和航迹中的非直线部分的长度.

(3) 战斗时间消耗惩罚. 每个时间步都进行惩罚, 以督促无人机快速解决战斗 ($r_4^{(t)}=-1$).

(4) 战术动作重复惩罚. 惩罚多架无人机对同一敌机从同一方向发起进攻的行为, 避免浪费资源.

$$r_5^{(t)}=-\sum_{j=1}^{M}\sum_{k=1}^{3}I(T_{\mathrm{tac},jk}^{(t)}>1). \tag{14}$$

其中, $T_{\mathrm{tac},jk}^{(t)}$为$t$时刻针对敌机$j$的$k$战术的我方无人机数目之和.

(5) 目标遗漏惩罚. 若有敌机没有被我方任何无人机分配, 将进行目标遗漏惩罚

$$r_6^{(t)}=-\sum_{j=1}^{M}I(\sum_{i=1}^{N}I(T_{\mathrm{tgt},i}^{(t)}=e_j)=0). \tag{15}$$

(6) 被摧毁惩罚. 若我方无人机被摧毁, 将得到被摧毁惩罚

$$r_7^{(t)}=-\sum_{i=1}^{N}I(D_i^{(t)}=1\wedge D_i^{(t-\Delta t)}=0). \tag{16}$$

(7) 摧毁敌方单位奖励. 若我方无人机摧毁敌方无人机, 将得到奖励

$$r_8^{(t)}=\sum_{j=1}^{M}1-I(D_j^{(t)}=1\wedge D_j^{(t-\Delta t)}=0). \tag{17}$$

(8) 协同作战奖励. 鼓励多个无人机针对同一敌方目标从不同战术方向发起进攻, 形成夹击之势.

$$r_9^{(t)}=\sum_{j=1}^{M}\sum_{i}^{N}\sum_{k=1}^{3}X_{ij}^{(t)}X_{ik}^{(t)}. \tag{18}$$

综上, t时刻无人机系统获得的总奖励$r^{(t)}$为以上 9 个奖励的和

$$r^{(t)}=\sum_{k=1}^{9}\omega_k r_k^{(t)}. \tag{19}$$

其中, ω_k为各项奖励的缩放因子, 确保各子奖励取值范围相当. 由于某些子奖励值 (如目标、战术切换惩罚) 的具体范围在事先无法获知, 因此各项缩放因子的确定由实验调参给定.

2.2.5 状态转移概率

武器系统发射空对空导弹之后, 并不是 100% 摧毁目标. 以敌机攻击我机为例, 对于任意$u_i\in\mathbf{U}$而言, 其被击毁的概率取决于该时刻朝u_i开火的敌方无人机的数量, 如下

$$P(D_i^{(t+1)}=1|D_i^{(t)}=0)=1-\prod_{e_j\in\mathbf{F}_i}(1-p_{\mathrm{attack}}^{e_j}). \tag{20}$$

其中, $\mathbf{F}_i$为朝u_i开火的敌方无人机的集合, $p_{\mathrm{attack}}^{e_j}$为敌机发射导弹后摧毁我机的概率, 取$p_{\mathrm{attack}}^{e_j}=0.8$.

2.3 任务重分配网络

无人机u_i的任务重分配网络由数据处理层、攻击目标决策层以及攻击战术决策层组成, 如图 1.

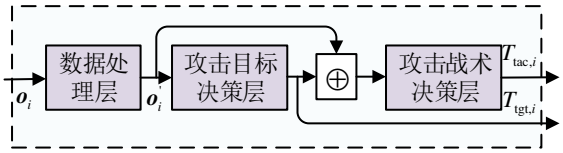


图1 任务重分配网络

该网络接收u_i的观测信息$\mathbf{o}_i$, 输出u_i的任务分配结果$(T_{\mathrm{tgt},i},T_{\mathrm{tac},i})$. 具体的, 首先数据处理层对$\mathbf{o}_i$进行初步的处理, 得到$\mathbf{o}_i'$; 然后, 攻击目标决策层根据$\mathbf{o}_i'$, 生成攻击目标$T_{\mathrm{tgt},i}$, 在该层会屏蔽掉不满足任务约束的攻击目标; 随后, 攻击战术决策层根据$T_{\mathrm{tgt},i}$和$\mathbf{o}_i'$, 生成对应的攻击战术$T_{\mathrm{tac},i}$. 无人机u_i可利用该网络在每一时刻t进行一次任务重分配, 当敌机被摧毁退出战场时, 攻击目标决策层会自动屏蔽该敌机, 且由于每个无人机都具有自身的任务重分配网络, 所以该网络实现了满足约束的多无人机分布式动态任务分配. 本文将基于该网络的 MAPPO 算法命名为 MAPPO-AAS(Multi-Agent Proximal Policy Optimization with Attention and Action Mask Separation), 具体如下.

2.3.1 基于分离式状态滑动标准化的数据处理层

智能体在与环境交互时, 为维护训练稳定性, 会对全局状态和局部观测进行滑动标准化, 使其符合均值为 0、方差为 1 的正态分布. 经过标准化后的全局状态和局部观测, 更有利于网络的训练[27]. 由于全局状态和局部观测的滑动标准化步骤完全一致, 为了简化描述, 将它们都简称为样本$\mathbf{x}$. 因此, $\mathbf{x}$的滑动标准化公式可以表示为

$$\begin{cases}\bar{\mathbf{x}}_{n+1}=\bar{\mathbf{x}}_n+\dfrac{1}{n+1}(\mathbf{x}_{n+1}-\bar{\mathbf{x}}_n)\\ \sigma_{n+1}^2=\sigma_n^2+(\mathbf{x}_{n+1}-\bar{\mathbf{x}}_n)(\mathbf{x}_{n+1}-\bar{\mathbf{x}}_{n+1})\\ \mathbf{x}'_{n+1}=\dfrac{\bar{\mathbf{x}}_{n+1}}{\sigma_{n+1}}\end{cases}. \tag{21}$$

其中, $\mathbf{x}_{n+1}$为新样本数据, $\bar{\mathbf{x}}_n$为上一时刻的样本均值, $\bar{\mathbf{x}}_{n+1}$为考虑新样本后的样本均值, σ_n^2为上一时刻的样本方差, σ_{n+1}^2为考虑新样本的样本方差, $\mathbf{x}'_{n+1}$为标准化之后的 $\mathbf{x}_{n+1}$. 然后, 将标准化之后的样本数据$\mathbf{x}'_{n+1}=(\mathbf{s}'_{n+1}\vee\mathbf{o}'_{n+1})$输入网络进行训练.

然而, 标准化离散信息 (如无人机存活状态 0 和 1) 可能导致信息失真, 影响模型性能. 因此, 提出分离式状态滑动标准化机制: 将原始数据分为离散数据$\mathbf{x}_{\mathrm{d}}$和连续数据$\mathbf{x}_{\mathrm{c}}$, 仅对连续数据进行标准化处理, 得到$\mathbf{x}'_{\mathrm{c}}$, 确保其符合正态分布, 而离散信息

$\mathbf{x}_{\mathrm{d}}$保持不变, 最后将$\mathbf{x}_{\mathrm{d}}$和$\mathbf{x}_{\mathrm{c}}'$合并为$\mathbf{x}'$.

2.3.2　基于动作屏蔽机制的攻击目标决策层

在无人机近距空战过程中, 将不满足任务约束的分配方案规定为非法动作. 在强化学习中, 处理非法动作的方法通常有两种: 一是给予非法动作负奖励来引导智能体避免执行任何非法动作; 二是通过动作屏蔽机制, 调整策略网络的输出, 使得非法动作被选中的概率为零. 本文为了解决无人机空战中的非法动作问题, 设计了动作屏蔽机制. 假设环境中有三个敌方无人机$\mathbf{E}=\{e_1,e_2,e_3\}$, 所有我方无人机弹药充足且处于挂载状态, 其中$e_2$已被摧毁.

(1) 攻击目标掩码

此时, $u_i\in\mathbf{U}$的攻击目标掩码$\mathbf{m}_i$为一个多维向量$\mathbf{m}_i=(m_{i1},m_{i2},m_{i3})^{\mathrm{T}}$, 其中$m_{ij}$代表$u_i$对敌方无人机$e_j$的目标选择掩码值. 根据合法性, 在对应位置设置相应的目标掩码值. 若非法, $m_{ij}=-\infty$; 若合法, $m_{ij}=0$. 所以, u_i的攻击目标掩码为

$$\mathbf{m}_i=(0,-\infty,0)^{\mathrm{T}}. \tag{22}$$

式 (22) 表示u_i可以选择e_1和e_3作为攻击目标, 而不可选择已被摧毁的e_2作为攻击目标. 上述掩码值的设置仅为理论值, 在算法实现时取值如下: 若非法, $m_{ij}=-10^{12}$; 若合法, $m_{ij}=0$.

(2) 攻击目标决策网络

设计攻击目标决策网络如图 2 所示. 首先, 基于$\mathbf{o}_i'$, 通过全连接层FC对各e_j进行打分, 得到S_{ij}; 然后, 利用攻击目标掩码$\mathbf{m}_i=(m_{i1},m_{i2},m_{i3})^{\mathrm{T}}$对非法攻击目标进行屏蔽; 随后, 经过一次 softmax 操作, 得到针对各e_j的选择概率, 概率公式如下

$$\pi(T_{\mathrm{tgt},i}=e_j|\mathbf{o}_i',\mathbf{m}_i)=\frac{\exp(S_{ij}+m_{ij})}{\sum_{k=1}^{3}\exp(S_{ik}+m_{ik})}. \tag{23}$$

由式 (22) 知$m_{i2}=-\infty$, 所以$\pi(T_{\mathrm{tgt},i}=e_2|o_i',m_i)=0$, 即$u_i$选择$e_2$作为攻击目标的概率为 0, 从而实现对非法攻击目标—e_2的屏蔽, 满足任务可执行约束.

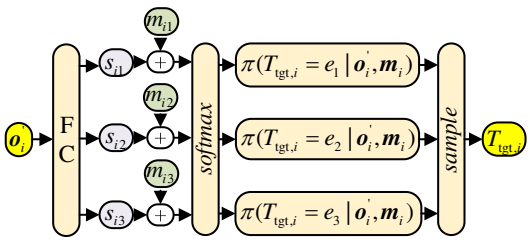


图2　基于动作屏蔽机制的攻击目标决策网络

2.3.3　基于注意力机制的攻击战术决策层

攻击战术决策层运用了注意力机制, 在攻击目标的基础上, 更加聚焦于攻击战术的决策, 提高攻击战术的合理性, 如图 3.

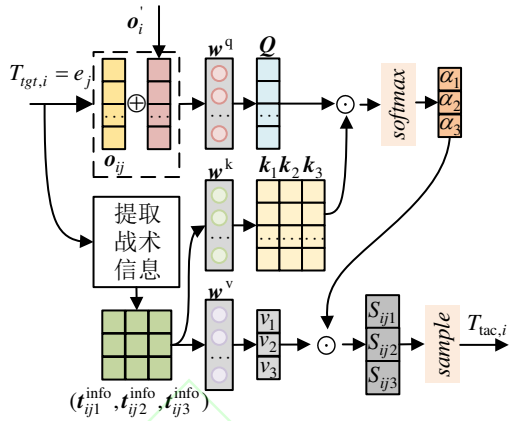


图3　基于注意力机制的战术动作决策网络

首先, 将$\mathbf{o}_i'$与攻击目标e_j的相关信息$\mathbf{o}_{ij}$进行拼接, 作为注意力网络的$\mathbf{q}$

$$\mathbf{q}=\mathbf{o}_i'\oplus\mathbf{o}_{ij}. \tag{24}$$

其中$\oplus$表示两个向量的拼接. 其次, 提取无人机u_i关于攻击目标e_j各个战术k的战术信息$\mathbf{t}_{ijk}^{\mathrm{info}}$作为注意力网络的$\mathbf{k}$和$\mathbf{v}$

$$\mathbf{k}=\mathbf{v}=(\mathbf{t}_{ij1}^{\mathrm{info}},\mathbf{t}_{ij2}^{\mathrm{info}},\mathbf{t}_{ij3}^{\mathrm{info}}). \tag{25}$$

所以, 注意力网络的$\mathbf{Q},\mathbf{K},\mathbf{V}$为

$$\begin{cases}\mathbf{Q}=\mathbf{q}\otimes\mathbf{w}^{\mathrm{q}}\\ \mathbf{K}=(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3)=(\mathbf{t}_{ij1}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{k}},\mathbf{t}_{ij2}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{k}},\mathbf{t}_{ij3}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{k}})\\ \mathbf{V}=(v_1,v_2,v_3)=(\mathbf{t}_{ij1}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{v}},\mathbf{t}_{ij2}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{v}},\mathbf{t}_{ij3}^{\mathrm{info}}\otimes\mathbf{w}^{\mathrm{v}})\end{cases}. \tag{26}$$

其中, $\mathbf{w}^{\mathrm{q}},\mathbf{w}^{\mathrm{k}},\mathbf{w}^{\mathrm{v}}$为注意力网络的权重, $\mathbf{q},\mathbf{k}_i$为向量, v_i为标量, $i=1,2,3$, $\otimes$表示矩阵乘法. 然后, 生成注意力权重

$$\begin{cases}\alpha_1=\mathbf{Q}\odot\mathbf{k}_1\\ \alpha_2=\mathbf{Q}\odot\mathbf{k}_2\\ \alpha_3=\mathbf{Q}\odot\mathbf{k}_3\end{cases}. \tag{27}$$

其中, $\odot$表示向量点乘. 最后, 根据注意力权重, 无人机u_i对攻击目标e_j各攻击战术进行打分

$$\begin{cases}s_{ij1}=\alpha_1v_1\\ s_{ij2}=\alpha_2v_2\\ s_{ij3}=\alpha_3v_3\end{cases}. \tag{28}$$

在打分的基础上, 对其进行 softmax 操作, 采样得到无人机u_i针对攻击目标e_j的攻击战术$T_{\mathrm{tac},i}$.

2.4　MAPPO-AAS 的并行训练框架

在同构无人机系统中, 无人机之间可以共享网络参数, 故设全局 actor 和 critic, 参数分别为θ, ϕ. 为提升数据收集效率, 采用分布式数据收集, 集中式算法训练的并行训练架构. 该架构分为两个核心部分:

(1) 分布式数据收集模块. 该模块包含N个worker, 分别与N个环境独立交互, 各有局部 actor 和 critic, 其参数为θ_n, ϕ_n. 各 worker 加载最新的全局参数并与环境进行交互, 获得经验数据.

(2) 集中式算法训练模块. 该模块接收分布式数据收集模块的经验数据. 用于更新全局的 actor 和 critic 网络参数θ, ϕ.

以上两个关键模块交替执行, 直至完成训练. 算法的训练流程可以归纳为以下几个关键步骤:

step 1: 初始化全局网络参数θ, ϕ、局部网络参数θ_n, ϕ_n; 初始化N个独立环境.

step 2: 分布式数据收集模块同步全局网络参数, 并在各自独立环境进行交互以收集数据.

step 3: 集中式算法训练模块将各环境收集的数据进行整合, 作为全局网络的训练数据, 更新全局网络参数. 若训练结束, 转至 step 4, 否则转至 step 2.

step 4: 结束训练, 保存相关数据.

3 仿真实验

本实验平台采用自主研发的“无人机任务规划系统”, 该系统为无人机空战过程中的任务规划提供全方位的算法支持, 包括任务分配、航迹规划和轨迹跟踪等算法的训练、测试及可视化. 实验室服务器配置为 CPU Intel(R) Xeon(R)Sliver 4215R 和 GPU NVIDIA GeForce GTX 3 090 24GB.

3.1 实验设定

在公海区域, 红蓝双方 (蓝方为我方无人机) 各有三架无人机进行100 km × 100 km范围内的空中对抗. 作战结果的评估基于以下两个标准:

(1) 胜负评估: 回合结束时, 若我方存活的无人机数量多于敌方, 则判定为胜利; 否则, 判定为失败. 通过整体胜率来评估作战结果.

(2) 战损比评估: 回合结束时, 计算敌我双方的战损比比值K/L(Kill-to-Loss Radio). 通过具体的作战数据来评估作战结果. K/L的计算公式如下:

$$K/L = \frac{M_{\text{kill}}/M}{N_{\text{loss}}/N + 1}. \tag{29}$$

式中, M_{kill}表示被我方无人机击毁的敌机数量, N_{loss}表示作战过程中我方无人机被击毁的数量. 上式表示, K/L越大, 我方无人机击杀 1 个敌方无人机单位时, 所损失的无人机越少. 为确保实验的公平性, 双方无人机的运动学模型和各项性能参数均保持一致.

3.2 算法有效性验证

为验证 MAPPO-AAS 算法的有效性, 将其与传统 MAPPO 算法、合同网协议算法[7](Contract Net Protocal, CNP) 进行了对比试验, 并以空战决策专家系统 (Expert System)[28] 为基准 (敌机同样搭载此系统). 实验过程中, MAPPO-AAS 和 MAPPO 算法超参数保持一致; 合同网协议算法采用文献 [7] 的改进方式进行任务重分配决策, 但回合结束后, 仍以上述 MAPPO-AAS 的奖励计算方式来计算合同网协议算法的奖励值, 以便进行算法对比. 其中, MAPPO 未采用动作屏蔽机制, 而是引入了负奖励来惩罚非法动作. 对非法动作的惩罚函数设定如下

$$r_{10}^{(t)} = -\sum_{i=1}^{N} I(\mathbf{a}_i^{(t)} \in \overline{\mathbf{A}}_i^{(t)}). \tag{30}$$

其中, $\mathbf{a}_i^{(t)}$为无人机u_i在t时刻的动作, $\overline{\mathbf{A}}_i^{(t)}$为无人机$i$在$t$时刻的非法动作空间, 所以 MAPPO 的总奖励为

$$r^{(t)} = \sum_{k=1}^{10} \omega_k r_k^{(t)}. \tag{31}$$

上述各算法性能评估包含: 总奖励、胜率和战损比. 总体训练轮数为 10 000 轮, 每训练 1 轮统计平均奖励值, 每训练 50 轮, 模拟 200 回合对战, 统计其胜率和战损比. 算法有效性验证实验图如图 4 所示, 相关数据分析如下:

奖励方面, MAPPO-AAS 策略奖励随训练轮数增加逐步上升, 50 轮后超 Expert System(−200), 4 000 轮后收敛至约 500. MAPPO 策略初始奖励低, 然后迅速上升, 但收敛值 (−1 500) 远低于 Expert System. CNP 策略平均奖励 (−70) 略高于 Expert System. 胜率方面, MAPPO-AAS 策略初始胜率几乎为 0, 50 轮后迅速攀升并超过 Expert System(42%), 2000 轮后稳定在 98.5% 左右. MAPPO 胜率增长缓慢, 最终约 5%. CNP 胜率略高于 Expert System, 约 55%. 战损比方面, MAPPO-AAS 策略战损比比值随训练迅速攀升, 50 轮后超过 Expert System(约 0.35), 最后收敛均值约 2.5. MAPPO 策略战损比比值基本为 0, CNP 策略约 0.5, 均远低于 MAPPO-AAS.

奖励、胜率和战损比对比结果验证了 MAPPO-AAS 算法在无人机近距空战动态任务分配中的有效性和优越性.

3.3 消融实验

为研究任务重分配网络中, 不同机制对算法性能的影响, 设计消融实验, 分别比较其总奖励曲线、胜率和战损比. 4 种对比算法如表 1, 为了方便描述, 将省略 MAPPO 前缀, 如将 MAPPO-AAS 称为 AAS. 消融实验对比图如图 5 所示, 相关数据分析如下:

奖励方面, 四种算法的奖励值在训练初期迅速

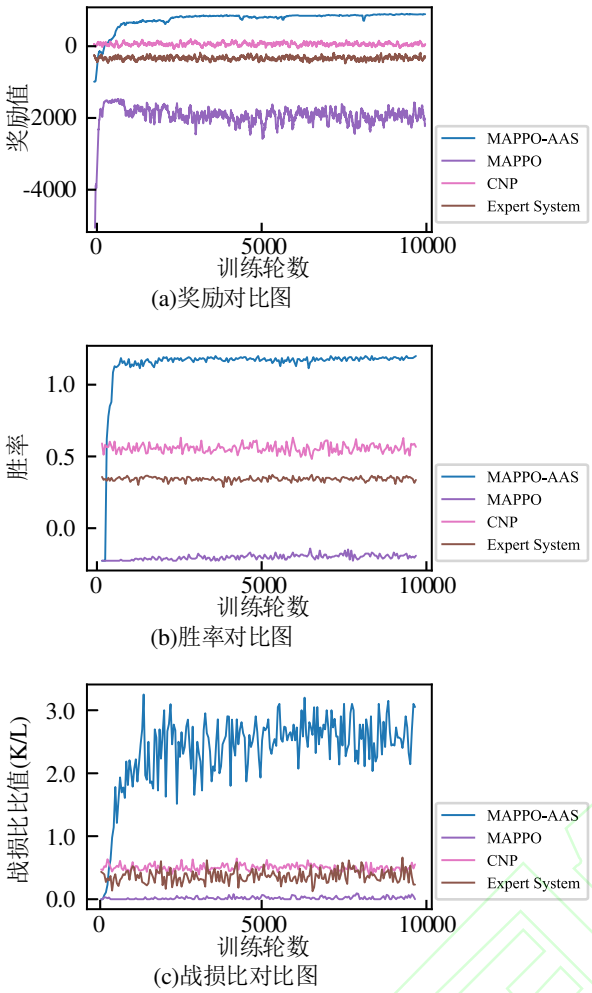


图4　算法有效性验证实验图

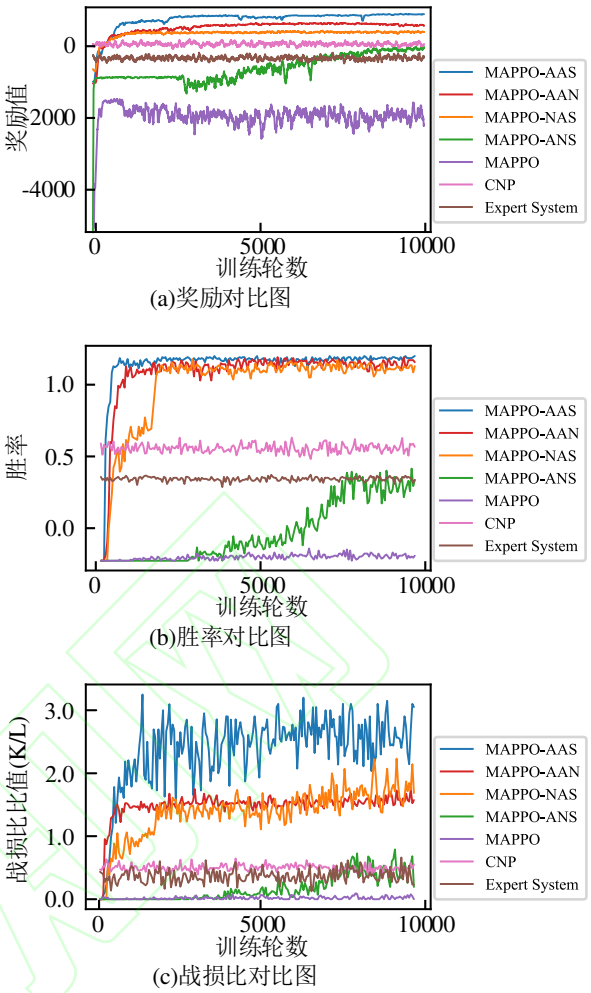


图5　算法的消融实验图

表1　消融实验算法设计

模型	注意力	动作屏蔽	分离式标准化
MAPPO-AAS	YES	YES	YES
MAPPO-NAS	NO	YES	YES
MAPPO-ANS	YES	NO	YES
MAPPO-AAN	YES	YES	NO

注: YES代表使用了该机制, NO代表没有使用

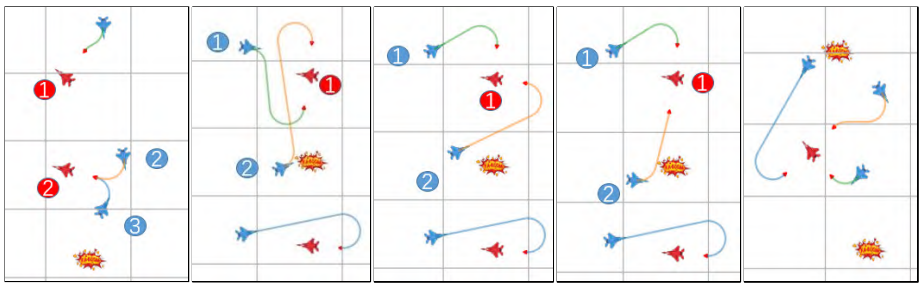
上升并收敛. AAS 收敛值最高, AAN 次之但收敛速度略快, 表明未使用分离式状态滑动标准化机制会导致提前收敛, NAS 收敛值低于 AAS 和 AAN, 说明注意力机制对性能有显著提升. ANS 训练后期收敛慢且不稳定, 最终收敛值较低, 表明动作屏蔽机制至关重要. 胜率方面, AAS、AAN、NAS 分别在 50、100、150 轮左右胜率超过 Expert System 和 CNP. 训练 10000 轮时, AAS 胜率保持在 98.5% 以上, AAN 和 NAS 胜率分别为 93% 和 90% 左右, 均远超 Expert System 和 CNP. 从胜率角度看, 三种机制均对算法有影响, 其中动作屏蔽机制和注意力机制最为关键. 战损比方面, AAS 在训练 200 轮左右与 AAN、NAS 拉开差距. 从战损比角度看, AAS 较其他算法有较大提升, 在击毁同等数量敌机的情况下自身损失较小.

3.4　行为分析

如图 6, 分别选择 MAPPO-AAS 模型在训练 100 轮、500 轮、1000 轮、2000 轮和 4000 轮时的参数, 对各阶段的任务分配策略进行了复盘. 在训练 100 轮时, 我方无人机战术重复, 导致资源浪费. 训练 500 轮时, 无人机选择非最优路径进攻, 未就近攻击敌机. 训练 1000 轮时, 虽然有所改进, 但仍有路径过长的问题. 训练 2000 轮时, 无人机展现出更明智的战术选择, 形成夹击之势, 大大缩短路径距离. 训练 4000 轮时, 无人机已学会从多个战术方向围攻敌机, 形成战术协作, 表明 MAPPO-AAS 模型已具备显著的战术意识和协作能力.

4　结　论

本文研究了近距空战场景下的多无人机协同分布式动态任务分配问题, 提出了一种基于 MAPPO 的分布式动态任务分配方法. 该方法通过引入任务可执行和载荷约束, 设计任务重分配网络, 解决了状态标准化信息失真、任务分配约束和攻击战术决策问题. 算法采用分布式数据收集与集中式训练的架构, 提高训练效率. 在 3v3 空战仿真中, 取得 98.5%



(a)100轮 (b)500轮 (c)1000轮 (d)2000轮 (e)4000轮

图6 MAPPO-AAS 各阶段任务分配策略

胜率, 远超传统 MAPPO、Expert System、CNP. 然而, 本研究仍存在一定局限性:

(1) 本文将空战环境简化为二维, 且仅考虑了 3 种固定战术动作, 并设置无人机速度大小为恒定, 这在一定程度上限制了模型的实用性;

(2) 在无人机空战过程中, 未考虑障碍物, 导致在实际应用中无人机遭遇障碍物时可能无法做出有效应对;

(3) 由于任务重分配网络无法处理变维度的数据, 因此所提方法无法妥善处理无人机动态加入的情况, 限制了该方法的可扩展性.

(4) 各子奖励的缩放因子由调参给定, 在不同作战场景下可能无法确保缩放因子的最优性. 此外, 在多达 9 个参数的情况下, 调参过程复杂.

在后续研究中, 将逐步引入三维模型, 考虑更多实际因素, 如高度变化和速度变化等, 以提高模型的真实性和实用性; 同时, 无人机的空战场景也将由无障碍物的公海区域拓展至复杂的山地地形, 以有效应对实际作战中的障碍物; 然后, 将进一步改进和完善多智能体强化学习模型, 以妥善处理无人机动态加入的问题; 最后, 探索自动化调参方法, 以便在不同作战场景中找到更合适的缩放因子组合.

参考文献 (References)

[1] 金钰, 谷全祥. 2023 年国外军用无人机装备技术发展综述[J]. 战术导弹技术, 2024(1): 33-47.
(Jin Y, Gu Q X. Overview of the development of foreign military UAV systems and technology in 2023[J]. Tactical Missile Technology, 2024(1): 33-47.)

[2] 李军, 陈士超. 无人机蜂群关键技术发展综述[J]. 兵工学报, 2023, 44(9): 2533-2545.
(Li J, Chen S C. Overview of key technology and its development of drone swarm[J]. Acta Armamentarii, 2023, 44(9): 2533-2545.)

[3] 张瑞鹏, 冯彦翔, 杨宜康. 多无人机协同任务分配混合粒子群算法[J]. 航空学报, 2022, 43(12): 412-427.
(Zhang R P, Feng Y X, Yang Y K. Hybrid particle swarm algorithm for multi-UAV cooperative task allocation[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(12): 412-427.)

[4] 孙彧, 潘宣宏, 戴定成, 等. 无人机蜂群作战任务规划研究现状与展望[J]. 火力与指挥控制, 2024, 49(1): 1-15.
(Sun Y, Pan X H, Dai D C, et al. Research status and prospect of UAV swarm combat mission planning[J]. Fire Control & Command Control, 2024, 49(1): 1-15.)

[5] 樊洁茹, 李东光. 有人机/无人机协同作战研究现状及关键技术浅析[J]. 无人系统技术, 2019, 2(1): 39-47.
(Fan J R, Li D G. Overview of MAV/UAV collaborative combat and its key technologies[J]. Unmanned Systems Technology, 2019, 2(1): 39-47.)

[6] 段海滨, 张岱峰, 范彦铭, 等. 从狼群智能到无人机集群协同决策[J]. 中国科学: 信息科学, 2019, 49(1): 112-118.
(Duan H B, Zhang D F, Fan Y M, et al. From wolf pack intelligence to UAV swarm cooperative decision-making[J]. Scientia Sinica: Informationis, 2019, 49(1): 112-118.)

[7] 王强, 贾强. 基于改进合同网的多无人机动态任务分配[J]. 火炮发射与控制学报, DOI: 10.19323/j.issn.1673-6524.202304010.
(Wang Q, Jia Q. A multi-UAV dynamic task allocation method based on improved contract net protocol[J]. Journal of Gun Launch and Control, DOI: 10.19323/j.issn.1673-6524.202304010.)

[8] 许可, 宫华, 秦新立, 等. 基于分布式拍卖算法的多无人机分组任务分配[J]. 信息与控制, 2018, 47(3): 341-346.
(Xu K, Gong H, Qin X L, et al. Multi-UAV task assignment for grouped tasks based on distribution auction algorithm[J]. Information and Control, 2018, 47(3): 341-346.)

[9] 李瑞琳, 崔巍, 冯彦翔, 等. 时间窗约束下的无人集群分布式任务分配算法[J]. 弹箭与制导学报, 2023, 43(03): 16-25.
(Li R L, Cui W, Feng Y X, et al. Distributed task allocation algorithm for unmanned cluster with time window constraints[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2023, 43(03): 16-25.)

[10] 邸斌, 周锐, 丁全心. 多无人机分布式协同异构任务分配[J]. 控制与决策, 2013, 28(2): 274-278.
(Di B, Zhou R, Ding Q X. Distributed coordinated heterogeneous task allocation for unmanned aerial

vehicles[J]. Control and Decision, 2013, 28(2): 274-278.)

[11] 高程, 都延丽, 步雨浓, 等. 基于顺序扩展一致性包算法的多无人机分布式任务分配[J]. 控制与决策, 2023, 38(11): 3242-3250.
(Gao C, Du Y L, Bu Y N, et al. Distributed task allocation of multiple UAVs based on sequential extended consensus based bundle algorithm[J]. Control and Decision, 2023, 38(11): 3242-3250.)

[12] 颜骥, 李相民, 刘波. 考虑时序约束的多智能体协同任务分配[J]. 控制与决策, 2015, 30(11): 1999-2003.
(Yan J, Li X M, Liu B. Multi-agents cooperative task allocation with precedence constrains[J]. Control and Decision, 2015, 30(11): 1999-2003.)

[13] 张祥银, 夏爽, 张天. 基于自适应遗传学习粒子群算法的多无人机协同任务分配[J]. 控制与决策, 2023, 38(11): 3103-3111.
(Zhang X Y, Xia S, Zhang T. Adaptive genetic learning particle swarm optimization based cooperative task allocation for multi-UAVs[J]. Control and Decision, 2023, 38(11): 3103-3111.)

[14] 潘耀宗, 张健, 杨海涛, 等. 战机自主作战机动双网络智能决策方法[J]. 哈尔滨工业大学学报, 2019, 51(11): 144-151.
(Pan Y Z, Zhang J, Yang H T, et al. Dual network intelligent decision method for fighter autonomous combat maneuver[J]. Journal of Harbin Institute of Technology, 2019, 51(11): 144-151.)

[15] 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.
(Shi W, Feng Y H, Cheng G Q, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623.)

[16] Piao H Y, Sun Z X, Meng G L, et al. Beyond-visual-range air combat tactics auto-generation by reinforcement learning[C]. 2020 International Joint Conference on Neural Networks. Glasgow, 2020: 1-8.

[17] Ding N, Soricut R. Cold-start reinforcement learning with softmax policy gradient[C]. Advances in Neural Information Processing Systems. Long Beach, 2017.

[18] 唐文泉, 孙莹, 杨奇, 等. 一种面向 2V2 近距空战的强化学习算法[J]. 战术导弹技术, 2022(1): 120-130.
(Tang W Q, Sun Y, Yang Q, et al. A reinforcement learning algorithm for 2V2 close-range air combat[J]. Tactical Missile Technology, 2022(1): 120-130.)

[19] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[C]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

[20] Huang S, Ontañón S. A closer look at invalid action masking in policy gradient algorithms[J/OL]. 2020, arXiv: 2006.14171.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems. Long Beach, 2017: 5998-6008.

[22] Stephan J, Pfeifle O, Notter S, et al. Precise tracking of extended three-dimensional dubins paths for fixed-wing aircraft[J]. Journal of Guidance, Control, and Dynamics, 2020, 43(12): 2399-2405.

[23] Li X X, Hu X G, Wang Z Q, et al. Path planning based on combinaion of improved A-STAR algorithm and DWA algorithm[C]. 2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture. Manchester, 2020: 99-103.

[24] Lin Z W, Xue C Y, Deng Q, et al. A single-loop robust policy gradient method for robust Markov decision processes[J/OL]. 2024, arxiv: 2406.00274.

[25] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[C]. International Conference on Machine Learning. Lille: PMLR, 2015: 1889-1897.

[26] Kurniawati H. Partially observable Markov decision processes and robotics[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 253-277.

[27] Engstrom L, Ilyas A, Santurkar S, et al. Implementation matters in deep policy gradients: A case study on ppo and trpo[J/OL]. 2020, arXiv: 2005.12729.

[28] 锐平, 高正红. 无人机空战仿真中基于机动动作库的决策模型[J]. 飞行力学, 2009, 27(6): 72-75.
(Rui P, Gao Z H. Research on decision system in air combat simulation using maneuver library[J]. Flight Dynamics, 2009, 27(6): 72-75.)

作者简介

李海峰 (1998−), 男, 硕士生, 从事多无人机协同任务规划的研究, E-mail: lihaifeng750@163.com;

杨宏安 (1972−), 男, 教授, 博士, 从事机器人技术及其工程应用、特种机电装备研发等研究, E-mail: yhongan@nwpu.edu.cn;

盛梓茂 (1999−), 男, 博士生, 从事多无人机的分布式任务规划的研究, E-mail: hpShengZimao@163.com;

刘超 (2000−), 男, 硕士生, 从事机器视觉及其工业应用的研究, E-mail: 13572384099@163.com;

陈逸新 (1999−), 男, 硕士生, 从事 SlAM 稠密建图的研究, E-mail: 1832708582@qq.com.