

DOI:10.20189/j.cnki.CN/61-1527/E.202605013

基于共享参数近端策略优化的无人机编队变换策略

钱力滢¹, 鲜勇¹, 李少朋^{1,2*}, 刘智¹

(1. 火箭军工程大学, 西安 710025;

2. 清华大学自动化系, 北京 100084)

摘要: 针对受运动学约束的多无人机编队变换问题, 提出一种融合拍卖式目标分配与共享参数近端策略优化(proximal policy optimization, PPO)的端到端协同规划方法。首先, 针对编队变换过程中目标匹配不合理易导致的航迹交叉、路径负载不均等问题, 构建综合考虑航迹长度均衡性与交叉惩罚的分配代价矩阵, 并采用拍卖式分配机制完成目标指派。其次, 在中心化训练、去中心化执行框架下, 设计共享参数策略网络, 将自身运动状态、目标相对信息及邻居局部观测共同作为策略输入, 实现多无人机协同变换规划。仿真实验结果表明: 在不同编队规模条件下, 所提方法均能有效完成编队变换任务。与基线方法相比, 所提方法的最终平均训练奖励提高约67%; 相比未引用拍卖式分配的共享参数PPO, 收敛所需代数减少约19%。

关键词: 多无人机编队; 航迹规划; 强化学习; PPO; 拍卖分配

中图分类号: V279

文献标志码: A

开放科学标识码(OSID):



听
语
音
聊
科
研
与
作
者
互
动

文章编号: 2097-2741(2026)05-0140-10

无人机编队作为多智能体系统的典型应用, 因其在空中覆盖、机动灵活和任务并行能力上的优势, 正被广泛应用于民用与军事领域, 如协同监视巡逻、物流投送、快速勘察搜救等。这些应用不仅对编队的路径规划与控制提出了实时性与安全性要求, 也对任务分配、信息共享与协同策略提出了更高的鲁棒性与泛化能力需求^[1-5]。

同时, 编队变换规划问题面临若干本质挑战。首先, 动力学约束使得无人机不能任意转向或瞬间到达目标, 常需采用受约束的运动学轨迹模型; 其次, 变换过程中碰撞风险与安全距离限制要求在规划过程中显式考虑多体相互约束, 尤其在密集编队或复杂目标分布下容易诱发近距离冲突或碰撞事件^[6-10]; 再者, 编队变换不仅要求个体尽快到达目标, 还对队伍的同步性提出要求, 这在时间效率与安全之间存在显著的冲突与权衡。现有研究多从

传统路径规划算法^[11-13]、群智能优化算法^[14-17]、模型预测控制^[18]、一致性算法^[19-20]、强化学习算法^[21-23]等角度开展研究。传统规划方法可解释性强、约束表达明确, 但在多无人机、强耦合、动态场景下计算复杂、实时性不足; 一致性控制结构清晰, 适合闭环控制, 但易受规模扩展影响。顾伟等^[24]选择队形变换的能源消耗与完成时间作为效率衡量指标, 提出了一种面向时间协同的多无人机队形变换最优效率模型; 肖雁冰^[25]融合了长机-僚机法和虚拟结构法的部分特点, 提出了一种改进的长机-僚机法, 再引入分布式一致性算法的思想构建编队, 为多无人机系统的信息一致性提供保证。但是, 赵天隆等^[26]针对三维环境下采用传统人工势场法实现无人机编队路径规划存在的规划效率低及路径震荡问题, 提出了一种融合深度强化学习与改进人工势场法(artificial potential field, APF)的无人机

收稿日期: 2026-03-09

修回日期: 2026-04-03

录用日期: 2026-05-07

第一作者: 钱力滢(2003—), 男, 硕士研究生, 主要从事飞行力学与飞行器规划研究。

* 通信作者: 李少朋(1992—), 男, 副教授, 主要从事飞行力学与飞行器规划研究。

引用格式: 钱力滢, 鲜勇, 李少朋, 等. 基于共享参数近端策略优化的无人机编队变换策略[J]. 火箭军工程大学学报, 2026, 40(5): 140-149. Qian Liying, Xian Yong, Li Shaopeng, et al. Shared-parameter proximal policy optimization-based approach for unmanned aerial vehicle formation transformation[J]. Journal of Rocket Force University of Engineering, 2026, 40(5): 140-149.

编队路径规划策略,但是对同步性、安全性和泛化性的综合考虑不足。Pang 等^[27]设计了一种基于仿射变换的可变队形重构与避障控制方案,周从航等^[28]针对强化学习算法模型在编队控制问题中存在的收敛速度慢、奖励稀疏等不足,将人工势场法引入深度强化学习,建立了无人机编队路径规划网络训练框架,但是在智能体规模增大时训练不稳定、样本效率低。在多布局场景下如何具备安全高效且拥有良好泛化能力的变换策略,仍是一个待解决的研究难题。

针对上述问题,本文提出一种基于共享参数近端策略优化(proximal policy optimization, PPO)算法的无人机编队变换策略。在训练开始时,采用拍卖式目标分配模式以降低不良匹配带来的负面影响;同时,在观测与奖励设计的基础上,综合考虑同步性、碰撞惩罚与时间成本,设计共享参数模式的 PPO 策略,实现中心化训练与去中心化执行,以提升样本效率与可扩展性。

1 问题建模

1.1 任务描述

本工作研究无人机编队变换中的受限运动规划问题。设有 N 架同构无人机按照一定编队飞行,同时给定一组目标点集合 T 作为目标编队,任务是在每架无人机满足运动学约束与安全约束的前提下,引导所有无人机在时间协同窗口内到达目标编队对应位置。

设编队中无人机的位置为 $\mathbf{P}_i(t)$,每架无人机对应目标编队中的位置为 $\mathbf{P}_i'(t) = (x_i', y_i')$,定义位置误差为 $e_i(t) = \|\mathbf{P}_i - \mathbf{P}_i'\|$,给定到达误差阈值 ϵ_p ,定义到达时刻 $t_i^a = \inf\{t \geq 0 | e_i(t) \leq \epsilon_p\}$,若对所有 $i = 1, 2, \dots, N$,均满足时间协同约束 $T_0 \geq \max_i(t_i^a) - \min_i(t_i^a) = T_u$,且变换全过程满足安全距离约束 $\min_t(\min_{i \neq j} \|\mathbf{P}_i(t) - \mathbf{P}_j(t)\|) \geq D_s$,则判定编队变换成功。 $T_0 = k_0 \Delta t$ 为设定的时间协同窗口,其中, Δt 为仿真或控制周期, k_0 按允许的同步误差步数选取, D_s 为安全距离。为研究方便建立以下假设。

1) 仅考虑规划时无人机成员之间的避撞,各无人机均视为质点,当无人机之间欧氏距离小于安全距离时,则认定其发生碰撞;

2) 无人机编队过程中每架无人机能准确获得

自身状态,各无人机之间能够互相通信获取各自的相对位置信息,且不考虑通信过程中的延迟和丢包;

3) 忽略外界大气扰动对无人机运动的影响;

4) 无人机为同构群体,各项参数均相同,包括相同的质量、最大速度、最大加速度和最小转弯半径等。

5) 研究侧重于高层路径规划,底层追踪等细微操作依赖于无人机自身飞控系统。

其中,通信延迟丢包与气流扰动会影响执行性能,但不改变算法框架的有效性。中心化训练与去中心化执行的架构使得执行阶段策略只依赖局部观测,因此,对通信依赖较弱,延迟丢包主要影响“邻近观测”的准确性,可使用滤波技术处理;扰动可并入环境随机性,误差可作为状态转移噪声,强化学习可通过域随机化增强鲁棒性,本文先在不扰动条件下验证算法的核心机制,后续可扩展至鲁棒性训练。

1.2 无人机运动模型

将编队变换问题中无人机简化为二维平面 xOy 上的质点,无人机运动方程可表示为

$$\begin{cases} \dot{v}_i = b_i \\ \dot{x}_i = v_i \cos \beta_i \\ \dot{y}_i = v_i \sin \beta_i \end{cases} \quad (1)$$

式中: v_i 为第 i 架无人机的速率; b_i 为第 i 架无人机的切向加速度;航向角 β_i 为第 i 架无人机的速度和坐标系 Ox 轴的夹角, $i = 1, 2, \dots, N$, N 为无人机的数量。无人机的转向受到角速度变化率 ω 的约束。

2 基于 PPO 的无人机编队变换策略

2.1 PPO 算法原理

强化学习方法主要用于解决离散时间上的马尔可夫决策问题,智能体通过奖励驱动学习环境的状态 s 和动作 a 的映射关系,不断优化调整自身策略。策略梯度方法直接对参数化策略 $\pi_\theta(a | s)$ 关于期望回报求梯度,进而做梯度上升更新;但直接的策略梯度往往不稳定或需要非常小的学习率。

PPO 是一种深度强化学习方法^[29],通过构造“受限”的目标函数,在保证更新幅度不过大的同时,仍能高效利用样本。核心思想是近端更新,限制新旧策略之间的改变量。常见 PPO 形式有两种:一种是通过对新旧策略的 Kullback-Leibler 散度施加惩罚来限制更新幅度,另一种是通过裁剪的方式来约束更新。后者实现简洁且在众多连续控制任务中已被证明具有良好的稳定性与样本效率。因此,选用第二种

方式开展研究,算法结构示意图如图 1 所示。

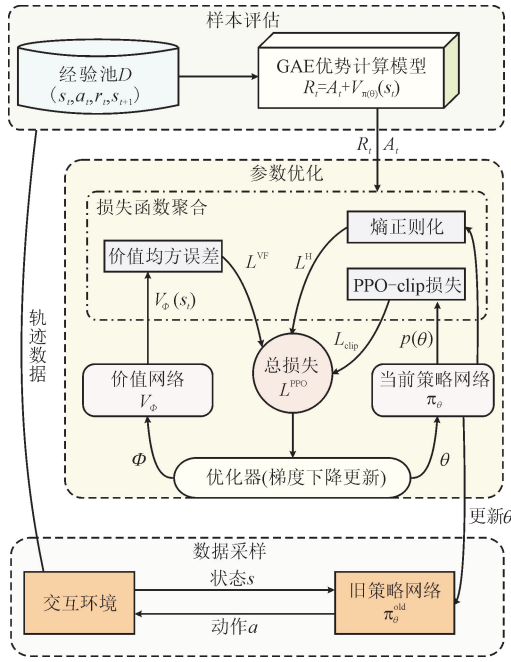


图 1 PPO 算法结构示意图^[29]

Fig. 1 Schematic of the PPO algorithm architecture

算法伪代码如下：

PPO(裁剪)算法

初始化策略网络参数 θ 和价值网络参数 Φ

对每一次迭代

利用策略网络 π_{θ} 在并行环境中收集轨迹数据 $D = \{s_t, a_t, r_t, s_{t+1}\}$

利用 GAE 计算优势函数 A_t , 对 A_t 做标准化, 计算回报目标

for $t = 1, 2, \dots, T_{\max}$

将 D 随机打乱分为若干 batch

对每个 batch 计算新旧策略概率比值 $p(\theta)$, 计算价值网络误差 L^V_{Φ} 、策略网络目标值 L^H_{clip} 、策略的熵 L^H , 计算总损失 $L^{\text{PPO}} = E[\tau_1 L^V_{\Phi} - \tau_2 L^H - L^H_{\text{clip}}]$

对 θ 和 Φ 做梯度下降更新

用 θ 更新旧策略网络参数 θ_{old}

结束循环

2.2 马尔可夫过程设计

为方便应用强化学习方法求解无人机编队变换问题,将其建模为马尔可夫过程,分别设计状态空间、动作空间、奖励函数。

2.2.1 状态空间设计

编队变换问题中,编队的初始构型和目标构型的位置信息均已知。计算单个无人机到其对应目标位置的欧式距离,可表示为

$$d_i = \|\mathbf{P}_i - \mathbf{P}_i'\| \quad (2)$$

计算无人机与目标点连线与 x 轴的夹角,其与无人机航向角 β_i 的差值 θ_i 可表示为

$$\theta_i = \arctan\left(\frac{y_i' - y_i}{x_i' - x_i}\right) - \beta_i \quad (3)$$

则环境的状态可表示为

$$s = (d_1, v_1, \beta_1, \theta_1, d_2, v_2, \beta_2, \theta_2, \dots, d_n, v_n, \beta_n, \theta_n) \quad (4)$$

式(4)中包含了每架无人机与其对应目标位置的距离和方位信息。

2.2.2 动作空间

无人机在每个仿真步长内执行一次连续值动作,通过切向加速度 b 和航向角变化率 ω 来调整自身状态,因此,多无人机系统的整体动作空间为 $A = (b_1, \omega_1, b_2, \omega_2, \dots, b_n, \omega_n)$ 。为便于强化学习算法收敛并满足物理约束,对每个动作分量施加上下界约束,可表示为

$$\begin{cases} b \in [-a_{\max}, a_{\max}] \\ \omega \in [-\omega_{\max}, \omega_{\max}] \end{cases} \quad (5)$$

所有无人机在每一个时间步长 Δt_0 内的运动学更新遵循以下规则:

$$v_i(t + \Delta t_0) = v_i(t) + b \Delta t_0$$

$$\beta_i(t + \Delta t_0) = \beta_i(t) + \omega_i \Delta t_0$$

$$\mathbf{P}_i(t + \Delta t_0) = \mathbf{P}_i(t) + v_i(t + \Delta t_0) \begin{pmatrix} \cos \beta_i(t + \Delta t_0) \\ \sin \beta_i(t + \Delta t_0) \end{pmatrix} \Delta t_0 \quad (6)$$

2.2.3 奖励函数设置

对于强化学习来说,奖励函数对于策略的学习起到决定性的影响,智能体的终极目标就是奖励最大化,良好的奖励机制可以促进智能体学习到较优的策略。无人机编队变换需要各无人机在安全约束下协同到达目标位置,因此,将奖励函数设置为如下两部分:

$$r = r_s + r_e \quad (7)$$

式中: r_s 为过程奖励,用于体现编队变换过程中的动态性能要求和约束; r_e 为终端奖励,用于衡量最终编队变换的效果。

过程奖励 r_s 由距离奖励 r_d 、时间惩罚 r_t 、碰撞惩罚 r_c 共同构成,具体计算公式为

$$\begin{cases} r_s = r_d - r_t - r_c \\ r_d = \alpha_d \sum_{i=1}^n (d_i' - d_i) \\ r_t = \alpha_t \\ r_c = \begin{cases} \alpha_{c_1} & \text{if } d_T \leq D_s \\ \alpha_{c_2} (3D_s - d_T) & \text{if } D_s < d_T \leq 3D_s \\ 0 & \text{otherwise} \end{cases} \end{cases} \quad (8)$$

式中: α_d 、 α_t 分别为距离奖励和时间惩罚的超参数; d_i' 为第 i 个无人机上个时间步的距离; D_s 为前文设置的碰撞安全距离; α_{c_1} 、 α_{c_2} 为碰撞惩罚的超参数; d_T 为无人机之间距离最小值, 当 d_T 小于安全距离时, 给智能体一次性大量惩罚, 当 d_T 接近安全距离, 存在碰撞风险时, 给予智能体线性递减惩罚。

终端奖励 r_e 由无人机到达奖励 r_a 、时间协同奖励 r_u 组成, 每架无人机成功到达目标点给予一定奖励, 所有无人机全部到达目标点给予一次性大量奖励。同时, 根据到达的时间节点计算时间协同奖励, 终端奖励计算公式可表示为

$$\begin{cases} r_e = r_a + r_u \\ r_a = \begin{cases} R_{a_0} k_a & k_a < n \\ R_{a_0} n + R_{a_1} & k_a = n \end{cases} \\ r_u = \begin{cases} \alpha_u & T_u \leq T_0 \\ 0 & T_u > T_0 \end{cases} \end{cases} \quad (9)$$

式中: R_{a_0} 、 R_{a_1} 是无人机到达奖励的超参数; k_a 为到达自己目标点的无人机数量; α_u 为时间协同奖励的超参数。

奖励函数各分项均采用无量纲化处理, 距离项以场景尺度归一化, 时间项以最长任务时间归一化, 增强权重可解释性。权重等超参数选择遵循“安全优先、效率其次、同步约束、最后微调”的原则。首先, 固定安全惩罚权重使训练中碰撞率快速下降到目标阈值以下; 随后, 在碰撞率稳定的前提下对时间、航迹项进行粗网格搜索, 再进行小幅度微调; 最后, 调节同步时间协同项权重使无人机群满足时间协同窗口约束。

3 共享参数 PPO 下的编队变换策略

3.1 拟解决的问题

3.1.1 无人机数量增多造成维度问题

随着无人机数量的增加, 依据所构建的马尔可夫过程, 环境状态空间的维数会迅速膨胀。若将每架飞机的观测全部拼接为全局状态量, 状态向量的维度随着 N 线性增长, 则会带来所谓的“维度灾难”, 决策网络需要更多的隐藏神经元来拟合高维输入, 样本复杂度和训练难度显著增加, 且更易发生过拟合或训练不稳定, 简单通过加深网络以缓解维度增长带来的负面影响, 并不是可靠的长期解决方案。

3.1.2 目标匹配方法影响规划效益

在本问题中, 原始编队的无人机与目标集合之间的对应关系通常由预先的分配算法确定, 从而对后续轨迹规划与策略学习产生深远影响, 若分配方案不合理, 例如多条航路交叉频繁, 会带来多种负面后果。对于基于强化学习的控制器而言, 这些因素会使得策略学习变得更加困难, 智能体需要补偿由不良分配引入的额外航程与复杂交互, 导致算法整体收敛速度降低且最终性能受限。

基于上述问题分析, 在标准 PPO 的框架上引入拍卖式目标分配机制和共享参数的 PPO 策略, 以解决无人机编队变换的航迹规划问题。

3.2 拍卖式目标分配

确定合理的无人机与目标点的匹配关系可以有效提升编队变换的稳定性, 减少航路冲突。采用拍卖算法对无人机目标点进行分配^[30]的算法步骤如下。

Step1 利用无人机与目标之间的直线距离快速估计航迹长度, 用全局长度统计量计算均衡性成本, 鼓励所分配的路径长度接近全局值, 从而避免少数无人机承担过长航程。

用 L_{ij} 表示第 i 架无人机与第 j 架无人机的直线距离, 则无人机全局均值可表示为

$$\bar{L} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N L_{ij} \quad (10)$$

均衡性成本可表示为

$$b_{ij} = |L_{ij} - \bar{L}| \quad (11)$$

经归一化后可得

$$\tilde{b}_{ij} = \frac{b_{ij}}{\max_{i,j} b_{ij} + \epsilon} \quad (12)$$

Step2 用线段相交计数近似计算交叉惩罚, 统计该线段与其他无人机的“代表线段”相交的次数, 候选线段 $s_{ij} = [P_i, P_j']$, 对每个无人机 k ($k \neq i$), 定义其代表线段为 $s_k^* = [P_k, P_{j_k}']$, 其中, $j_k^* = \operatorname{argmin}_j L_{kj}$, 交叉计数为

$$c_{ij} = \sum_{k \neq i} \mathbf{1}(\operatorname{cross}(s_{ij}, s_k^*) \cap j \neq j_k^*) \quad (13)$$

式中: $\mathbf{1}(\cdot)$ 为指示函数, $\operatorname{cross}(\cdot)$ 为二维线段相交判定算子, 采用向量叉乘的方法进行判定, 经过归一化后作为交叉代价, 可表示为

$$\tilde{c}_{ij} = \frac{c_{ij}}{\max_{i,j} c_{ij} + \epsilon} \quad (14)$$

Step3 将两项经归一化后的代价按权重组合得到总代价, 计算公式为

$$F_{ij} = \tilde{b}_{ij} + \tilde{c}_{ij} \quad (15)$$

式中: F_{ij} 为分配代价, 作为拍卖算法中分配代价矩阵的输入。

Step4 根据得到的分配代价矩阵, 使用拍卖算法进行无人机与目标位置的匹配。

通过引入拍卖算法预先确定无人机在目标队形中的位置, 使得无人机之间的航路冲突减少, 可以有效减少后续智能体策略学习的难度。

3.3 共享参数模式下的 PPO

在通过拍卖算法确定无人机与目标位置匹配关系的基础上, 采用共享参数的策略结构: 用参数共享的策略网络对每个无人机的局部观测分别输出动作分布的均值, 并用一个中心化价值网络对全局状态进行评价。该设计符合“中心化训练、去中心化执行 (CTDE)”范式, 训练阶段价值网络访问全局信息以提高估计精度与样本利用率, 执行阶段每个无人机仅依赖本地观测与共享参数的策略网络独立决策。

3.3.1 邻近观测

为使去中心化执行时每架无人机能够感知局部交互并做出与整体相协调的决策, 在每架无人机的局部观测中, 增加邻近观测信息, 按和第 i 个无人机的欧氏距离对其他无人机排序, 取距离最近的无人机的距离和相对方位信息作为邻近观测信息, 可表示为

$$\begin{cases} d_{ij} = \|\mathbf{P}_j - \mathbf{P}_i\| \\ \eta_{ij} = \arctan2(\det(\mathbf{P}_j, \mathbf{P}_i), \mathbf{P}_j - \mathbf{P}_i) \end{cases} \quad (16)$$

对第 i 架无人机的局部观测定义为

$$s_i = [d_i, v_i, \beta_i, \theta_i, \hat{d}_i^{(1)}, \hat{\eta}_i^{(1)}, \hat{d}_i^{(2)}, \hat{\eta}_i^{(2)}] \quad (17)$$

式中: $\hat{d}_i^{(1)}$ 、 $\hat{\eta}_i^{(1)}$ 、 $\hat{d}_i^{(2)}$ 、 $\hat{\eta}_i^{(2)}$ 分别为距离第 i 个无人机最近和次近的无人机相对距离和相对方位角。

3.3.2 网络结构

策略网络的整体结构如图 2 所示, 其中, s_{dim} 为局部观测维度数, a_{dim} 为动作维度数。可以看出, 网络将每架无人机的局部观测视为一个独立样本, 先用相同的前向网络对每个样本进行特征提取得到 μ_{head} 。隐藏层为 2 层全连接层, 每层 128 神经元, 激活函数为 ReLU, 再通过特定输出头模块得到每架无人机的动作均值 μ_i 。输出头包含一个全连接层, 64 个神经元, 激活函数为 Tanh。最终输出的动作均值与全局共享的对数标准差 $\log(std)$ 共同构成策略网络输出的连续动作分布的参数, 该连续高斯分布下的动作策略为 $\pi(a_i | s_i) = N(\mu_i, \text{diag}(\sigma^2))$, 根据动作上下边界 $[a_l, a_u]$ 映射到真实动作区间, 即

$$a_i^0 = a_l + \frac{\tanh(a_i) + 1}{2} (a_u - a_l) \quad (18)$$

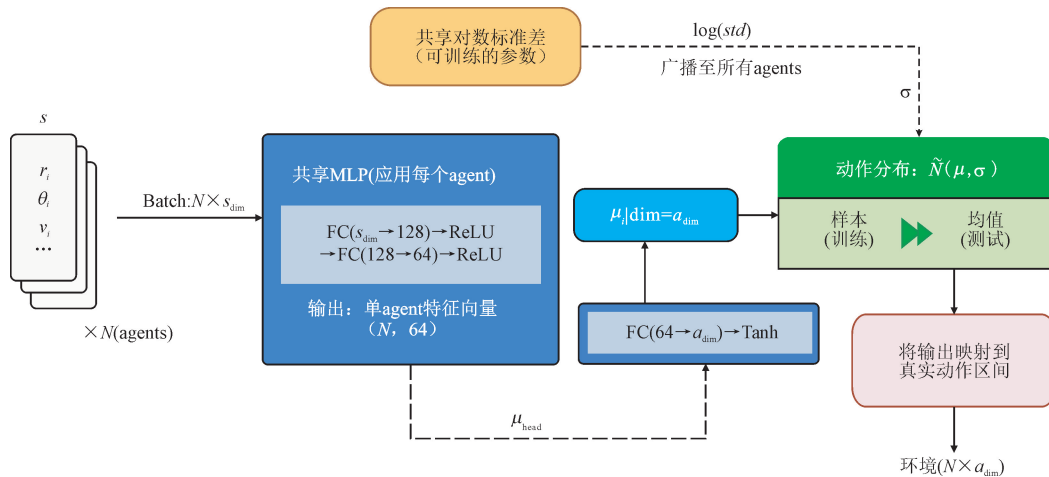


图 2 策略网络整体结构

Fig. 2 Overall architecture of the policy network

4 仿真实验验证

为验证所提算法的有效性, 采用 Gymnasium 架构建立交互环境, 仿真基于软件工具 Python3.11 和 torch2.9.0, 结合 Tianshou 第三方库, 进行无人

机编队变换的实验仿真验证。实验的硬件平台信息: CPU 为 intel ultra7 265KF, 64.0 GB RAM。

4.1 仿真参数设置

4.1.1 算法超参数设置

进行实验仿真时算法超参数设置如表 1 所示。

表 1 算法超参数设置	
Table 1 Hyperparameter settings of the algorithm	
算法超参数	数值
经验池容量	10 000
并行环境数量	8
学习率	0.000 3
折扣因子	0.99
价值损失系数	0.50
熵正则项的权重	0.005
GAE 超参数	0.95
裁剪阈值	0.20

4.1.2 环境参数设置

算法交互环境参数设置如表 2 所示。

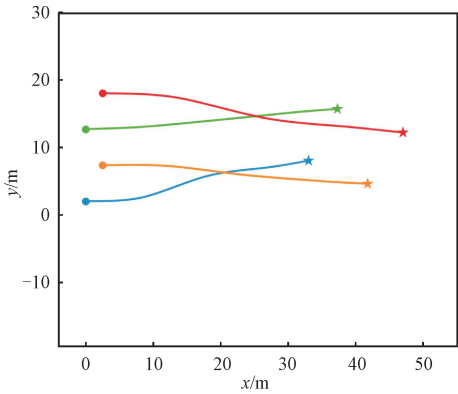
表 2 环境超参数设置	
Table 2 Hyperparameter settings of the environment	
环境超参数	数值
无人机安全距离/m	3
无人机速度区间/(m·s ⁻¹)	[3,8]
最大速率变化率/(m·s ⁻²)	5
最大航向变化率/(rad·s ⁻¹)	3
距离奖励超参数 α_d	3
时间惩罚超参数 α_t	0.1
到达奖励超参数 R_{a_0} 、 R_{a_1}	50、200
碰撞惩罚超参数 α_{c_1} 、 α_{c_2}	100、3
时间协同超参数 α_u	200
仿真步长 Δt /s	0.02

为提高策略的泛化能力与鲁棒性,在训练阶段,每次从预定义布局类型中随机抽取,将其作为目标队形,以增强对不同队形的适应能力。

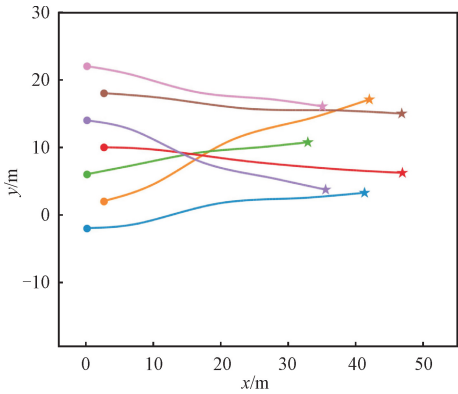
4.2 仿真分析

初始队形采用纵向交错编队,其中无人机航向均为 0°,速率为 3 m/s,随机抽取不同目标编队队形进行仿真,预定义线形编队、圆形编队、网格形状编队等,经过训练后在圆形编队上进行测试,结果如图 3 所示。

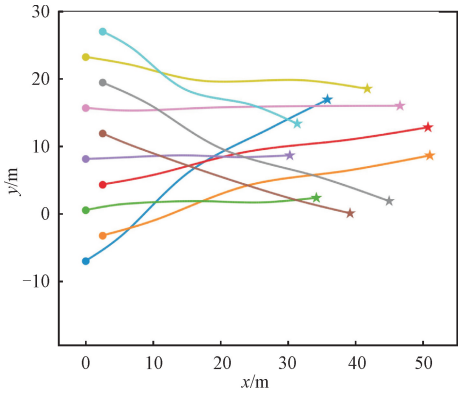
进行多次测试,统计不同无人机数量条件下编队变换的各指标,如表 3 所示。其中,平均航迹均值为所有无人机航迹长度的平均值。



(a)无人机数量为 4



(b)无人机数量为 7



(c)无人机数量为 10

● 起始点 ★ 目标点

图 3 编队变换测试结果

Fig. 3 Test results of round formation transformation

表 3 不同编队规模下的编队变换性能指标
Table 3 Formation transformation performance metrics across different formation scales

指标	不同规模下的指标值		
	N=4	N=7	N=10
碰撞率/%	2.6	6.7	11.4
平均航迹均值/m	44.5	51.7	59.8
平均最大终端误差/m	0.45	0.74	0.96
平均到达时间跨度/s	0.81	1.39	2.08

统计不同无人机数量下的训练奖励均值,绘制

训练奖励曲线如图 4 所示。可以看出,不同无人机数量的训练曲线均表现出较为一致的演化规律。在训练初期,策略处于强探索和不稳定调整阶段,平均奖励先经历短暂波动甚至明显下探;随着后续训练推进,奖励逐步回升,并在中后期进入持续增长和最终收敛阶段,表明智能体逐渐学习到有效的协同控制策略并形成稳定决策。

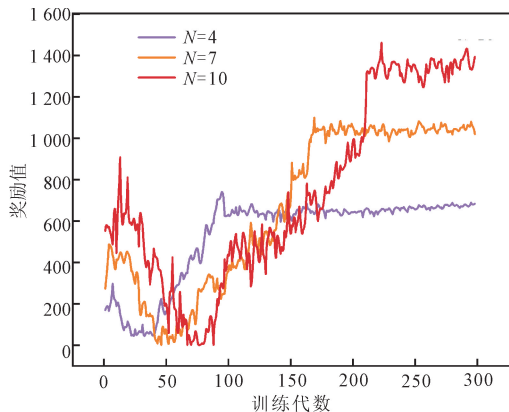


图 4 训练奖励值变化曲线

Fig. 4 Variation curves of training reward value

进一步比较不同规模场景可以发现,随着无人机数量增多,训练初期的奖励下沉更加明显,波动幅度也更大,且收敛速度有所减缓。这表明随着智能体数量增加,系统交互耦合关系更加复杂,训练过程中更容易受到碰撞规避、队形重构和同步约束等因素的影响,从而导致前期探索代价更高,策略学习难度更大。

4.3 消融实验

为量化各个模块对最终性能的影响,开展消融实验,验证所提方法各组成部分的作用与必要性。

4.3.1 消融项设置

各组实验设置如下:A1 为原始 PPO 算法作为基线;A2 为无拍卖式匹配模块的共享模式 PPO;A3 为采用拍卖式匹配、无共享参数模块的 PPO;A4 为所提方法。每个消融组均在相同的超参数、网络规模设置下重复多次取平均值。

4.3.2 实验结果

图 5 的训练曲线展示了各组在训练奖励随训练步数的变化。可以看出,A4 收敛更快,表明共享参数在提高样本效率与训练稳定性方面的贡献。A2 初期学习速度较快但最终收敛速度落后于 A4,表明拍卖分配提升了任务难度的可学习性并改善最终策略品质。A3 相比 A1 在训练稳定性上有改善,但收敛速度慢,最终效果较差,最终平均奖励分别仅为 524 和 504,相当于 A4 的 62%和 60%左右,表

明独立策略网络在样本利用上劣于参数共享。

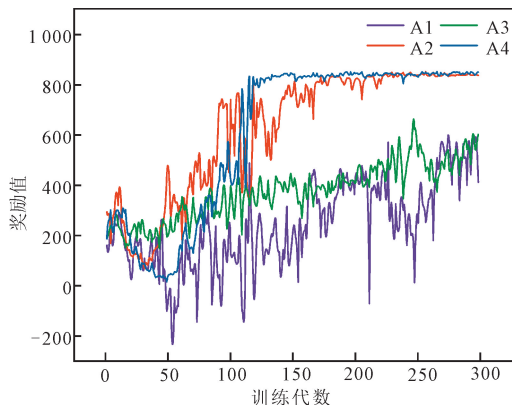


图 5 消融组训练奖励变化曲线

Fig. 5 Training reward curves for the ablation study groups

综上所述,A4 在成功率、碰撞率与完工时间等关键指标上均优于其余对照组,与基线方法 A1 相比,最终平均训练奖励提高约 67%,与 A2 相比,收敛所需代数减少约 19%,在复杂布局下优势更明显,验证了拍卖分配能改善任务可学习性并降低交叉风险,而参数共享可显著提高样本效率与训练稳定性。

4.3.3 奖励函数分项消融实验

为验证所设计奖励函数各组成部分的有效性,进一步开展奖励项消融实验,以完整奖励(B1)组为对照组,设置的消融组包括:去除安全项(B2)以及去除时间同步项(B3)。对照组和消融组均在无人机数量 $N=4$ 的条件下进行实验,超参数设置和 4.1 节保持一致。实验结果如表 4 所示。

表 4 奖励项消融结果

Table 4 Ablation study results for the reward components

组别	成功率/%	碰撞率/%	平均到达时间跨度/s
B1	90.8 ± 4.9	2.7 ± 0.8	0.62 ± 0.07
B2	51.6 ± 7.2	34.9 ± 8.1	0.88 ± 0.11
B3	71.9 ± 4.7	2.4 ± 1.0	6.76 ± 3.28

由表 4 可以看出,完整奖励在成功率、碰撞率、平均到达时间跨度指标上均取得了最优或次优的综合表现,说明各奖励项在多无人机编队重构任务中具有互补作用。比如,当去除安全项后,策略在测试阶段表现出更高的碰撞率和更低的任务成功率,说明安全惩罚对于约束智能体规避潜在碰撞、维持安全间隔具有重要作用。当去除时间同步项后,系统的平均到达时间跨度显著增大,表明同步奖励能够有效促进多无人机之间的协同行为,使编

队重构过程更加一致和平稳。

综上,完整奖励函数能够在安全性、协同性之间实现更优平衡,从而获得最佳综合性能。

5 结论

本文提出并实现了一种融合拍卖式目标分配与共享参数 PPO 的无人机编队变换方法,在规划开始采用拍卖式目标分配以降低不良匹配带来的额外航程与交叉风险,在控制层使用共享参数的 PPO 策略以及邻近观测实现中心化训练与去中心化执行,使得系统能更早地感知局部拥堵与潜在冲突,从而降低碰撞可能性。尽管所提方法在仿真中表现良好,但仍有若干可改进之处:当前分配为静态一次性分配,面对移动目标或突发故障时需重新分配;通信受限、环境噪声与异构无人机场景需进一步调整方法设计。未来拟进一步研究基于图神经网络的观测聚合和在受限通信等不完全信息条件下的规划框架。

参考文献

- [1] 鲜勇,郭靖,任乐亮,等. 无人机航迹规划智能算法研究综述[J]. 火箭军工程大学学报, 2025, 39(1): 106-121.
Xian Yong, Guo Jing, Ren Leliang, et al. Review of intelligent algorithms for unmanned aerial vehicle trajectory planning[J]. Journal of Rocket Force University of Engineering, 2025, 39(1):106-121.
- [2] 杨杰,席建祥,王成,等. 多无人机协同巡视任务规划方法综述[J]. 飞行力学, 2018, 36(5):1-6.
Yang Jie, Xi Jianxiang, Wang Cheng, et al. Summary of multi-UAV cooperative patrol task planning methods[J]. Flight Dynamics, 2018, 36(5):1-6.
- [3] 胡全鑫,高桂清,王成,等. 无人机集群分组多领航时变编队跟踪控制[J]. 火箭军工程大学学报, 2023, 37(2):27-33.
Hu Quanxin, Gao Guiqing, Wang Cheng, et al. Time-varying group formation tracking control for UAV swarm with multiple leaders[J]. Journal of Rocket Force University of Engineering, 2023, 37(2):27-33.
- [4] 高久安,李冰,席建祥,等. 间歇通信条件下双层编队四旋翼无人机集群跟踪控制[J]. 火箭军工程大学学报, 2025, 39(6):33-43.
Gao Jiuan, Li Bing, Xi Jianxiang, et al. Two-layer formation tracking control of quadrotor unmanned aerial vehicle swarms under intermittent communication[J]. Journal of Rocket Force University of Engineering, 2025, 39(6):33-43.
- [5] Yang Yu, Lee S. Multi-UAV coverage path assignment algorithm considering flight time and energy consumption[J]. IEEE Access, 2024, 12:26150-26162.
- [6] Hematulin W, Kamsing P, Torteeka P, et al. Trajectory planning for multiple UAVs and hierarchical collision avoidance based on nonlinear Kalman filters[J]. Drones, 2023, 7(2): 142.
- [7] 王文明. 固定翼无人机编队路径规划与队形保持方法研究[D]. 天津: 天津大学, 2022:15-18.
Wang Wenming. Research on formation path planning and formation keeping method of fixed-wing UAV [D]. Tianjin: Tianjin University, 2022:15-18.
- [8] 张青川,王乐,席建祥,等. 四旋翼无人机编队跟踪控制研究综述[J]. 火箭军工程大学学报, 2023, 37(2): 16-26.
Zhang Qingchuan, Wang Le, Xi Jianxiang, et al. Review on formation tracking control of quadrotor unmanned aerial vehicles[J]. Journal of Rocket Force University of Engineering, 2023, 37(2):16-26.
- [9] Chen Xi, Tang Jun, Ruan Yirun, et al. Path planning methods for UAVs: a survey[C]//Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. New York, USA: ACM, 2024:894-903.
- [10] Wang Li, Huang Weicheng, Li Haoxin, et al. A review of collaborative trajectory planning for multiple unmanned aerial vehicles[J]. Processes, 2024, 12(6): 1272.
- [11] Hart P, Nilsson N, Raphael B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE Transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.
- [12] Zhao Huimin, Gu Meixuan, Qiu Shaopeng, et al. Dynamic path planning for space-time optimization cooperative tasks of multiple unmanned aerial vehicles in uncertain environment[J]. IEEE Transactions on Consumer Electronics, 2025, 71(3): 7673-7682.
- [13] Goldenberg M, Felner A, Stern R, et al. Enhanced partial expansion A* [J]. Journal of Artificial Intelligence Research, 2014, 50(1):141-187.
- [14] Ullah S, Masood M. Genetic drift and its effects on the performance of genetic algorithm (GA)[C]//2023 International Conference on Robotics and Automation in Industry (ICRAI) . Piscataway, NJ, USA: IEEE, 2023:1-5.
- [15] Zhai Lixiang, Wu Husheng, Lai Linghong, et al. Intel-

- ligent optimization algorithms for multi-UAV path planning: a comprehensive review[J]. IEEE Access, 2025, 13: 101106-101130.
- [16] Liu Xiaoxiong, Su Yuzhan, Wu Yan, et al. Multi-conflict-based optimal algorithm for multi-UAV cooperative path planning[J]. Drones, 2023, 7(3): 217.
- [17] Yang Qingyun, Lu Jili, Zhang Yewei, et al. Multi-UAV cooperative path planning via graph neural network and proximal policy optimization[J]. IEEE Access, 2025, 13: 193575-193588.
- [18] Hidalgo-Paniagua A, Vega-Rodríguez M A, Ferruz J, et al. Solving the multi-objective path planning problem in mobile robotics with a firefly-based approach [J]. Soft Computing, 2017, 21(4): 949-964.
- [19] Chevet T, Vlad C, Maniu C S, et al. Decentralized MPC for UAVs formation deployment and reconfiguration with multiple outgoing agents[J]. Journal of Intelligent & Robotic Systems, 2020, 97(1): 155-170.
- [20] 刘云平, 徐泽, 赵中原, 等. 应对快速移动障碍物的多无人机编队避障控制[J]. 兵器装备工程学报, 2021, 42(3): 142-148, 156.
- Liu Yunping, Xu Ze, Zhao Zhongyuan, et al. Formation and collision avoidance control for multi-UAVs in response to fast moving obstacles[J]. Journal of Ordnance Equipment Engineering, 2021, 42(3): 142-148, 156.
- [21] Xiong Hang, Zhang Ying. Reinforcement learning - based formation-surrounding control for multiple quadrotor UAVs pursuit-evasion games [J]. ISA Transactions, 2024, 145: 205-224.
- [22] 司鹏搏, 吴兵, 杨睿哲, 等. 基于多智能体深度强化学习的无人机路径规划[J]. 北京工业大学学报, 2023, 49(4): 449-458.
- Si Pengbo, Wu Bing, Yang Ruizhe, et al. UAV path planning based on multi-agent deep reinforcement learning[J]. Journal of Beijing University of Technology, 2023, 49(4): 449-458.
- [23] Cattai T, Frattolillo F, Lacava A, et al. Multi-UAV reinforcement learning with realistic communication models: recent advances and challenges [J]. IEEE Open Journal of Vehicular Technology, 2025, 6: 2067-2081.
- [24] 顾伟, 汤俊, 白亮, 等. 面向时间协同的多无人机队形变换最优效率模型[J]. 航空学报, 2019, 40(6): 192-200.
- Gu Wei, Tang Jun, Bai Liang, et al. Time synergistic optimal efficiency model for formation transformation of multiple UAVs[J]. Acta Aeronautica et Astronautica Sinica, 2019, 40(6): 192-200.
- [25] 肖雁冰. 对抗环境下的多无人机编队方法和队形变换研究[D]. 南京: 南京邮电大学, 2020: 28-36.
- Xiao Yanbing. Research on formation method and formation transformation of multi-UAV in adversarial environment[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020: 28-36.
- [26] 赵天隆, 陈龙胜, 张存富, 等. 融合强化学习与改进人工势场的无人机编队路径规划[J]. 航空兵器, 2025, 32(5): 54-63.
- Zhao Tianlong, Chen Longsheng, Zhang Cunfu, et al. UAV formation path planning based on reinforcement learning and improved artificial potential field[J]. Aero Weaponry, 2025, 32(5): 54-63.
- [27] Pang Wen, Zhu Daqi, Sun Changyi. Multi-AUV formation reconfiguration obstacle avoidance algorithm based on affine transformation and improved artificial potential field under ocean currents disturbance [J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(2): 1469-1487.
- [28] 周从航, 李建兴, 石宇静, 等. 深度强化学习在无人机编队路径规划中的应用[J]. 电光与控制, 2024, 31(10): 27-33.
- Zhou Conghang, Li Jianxing, Shi Yujing, et al. Application of deep reinforcement learning in path planning of UAV formation [J]. Electronics Optics & Control, 2024, 31(10): 27-33.
- [29] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2026-02-15]. <https://doi.org/10.48550/arXiv.1707.06347>.
- [30] Barquet M, Bakolas E. Greedy decentralized auction-based task allocation for multi-agent systems[J]. IFAC-PapersOnLine, 2021, 54(20): 675-680.

Shared-Parameter Proximal Policy Optimization-Based Approach for Unmanned Aerial Vehicle Formation Transformation

QIAN Liying¹, XIAN Yong¹, LI Shaopeng^{1,2*}, LIU Zhi¹

(1. Rocket Force University of Engineering, Xi'an 710025, China;

2. Department of Automation, Tsinghua University, Beijing 100084, China)

ABSTRACT: To address the formation transformation problem in multi-unmanned aerial vehicle (UAV) systems subject to kinematic constraints, an end-to-end cooperative planning approach that integrates auction-based target assignment with shared-parameter proximal policy optimization was developed. First, to alleviate trajectory intersections and imbalanced path allocation arising from irrational target matching during formation transformation, an assignment cost matrix was constructed by comprehensively accounting for trajectory length balance and crossing penalties, and target assignment was performed using an auction-based mechanism. Second, under a centralized training and decentralized execution framework, a shared-parameter policy network was developed to facilitate coordinated formation transformation planning based on the motion state of each UAV, relative target information, and local neighbor observations. Simulation results demonstrate that the proposed method can effectively accomplish formation transformation tasks across various formation scales. Compared with baseline methods, the final training reward of the proposed approach is increased by about 67%. Compared to the shared parameter PPO without referencing auction allocation, the required number of iterations for convergence is reduced by about 19%.

KEYWORDS: multi-UAV formation; trajectory planning; reinforcement learning; PPO; auction allocation

(编辑:于福兴)