基于时空 Transformer 分布式强化学习的多无人机协同路径规划

张祥银[1,2†]，张曦梁[1,2]，史骏驰[1,2]

(1. 北京工业大学 信息科学技术学院，北京 100124；2. 数字社区教育部工程研究中心，北京 100124)

摘　要: 随着无人机技术的快速发展, 多无人机已广泛应用于应急救援、环境监测、军事侦察等复杂场景. 然而, 在复杂动态环境中, 多无人机往往面临路径规划效率低、个体间协同稳定性差、对环境及内部状态的时空特征处理能力不足等问题. 对此, 提出一种基于时空 Transformer 和动作块机制的分布式强化学习的路径规划框架. 所提出框架采用分布式 Actor-Critic 架构实现去中心化协同决策, 规避集中式决策存在的通信压力大、单点故障等弊端; 同时, 设计可扩展输入方案与多目标奖励函数, 全面平衡路径安全性、无人机间协同性及任务执行效率三大核心需求. 为更好地提取动态环境的时空特征, 减少特征传输中的冗余信息, 该网络融合时空 Transformer 与动作块机制, 有效提升了特征提取与协同决策的精度. 基于模拟复杂动态环境的消融实验验证了所提出方法的有效性与优越性, 研究成果为复杂场景下多无人机协同路径规划提供了高效的解决方案.

关键词: 多无人机；强化学习；注意力机制；分布式控制；协同路径规划；Actor-Critic 架构

中图分类号: TP24　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2026.0395

引用格式: 张祥银, 张曦梁, 史骏驰. 基于时空 Transformer 分布式强化学习的多无人机协同路径规划 [J]. 控制与决策.

Distributed reinforcement learning with spatio-temporal transformer based multi-UAV cooperative path planning

ZHANG Xiang-yin[1,2†]，*ZHANG Xi-liang*[1,2]，*SHI Jun-chi*[1,2]

(1. School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China；2. Engineering Research Centre of Digital Community of Ministry of Education，Beijing 100124，China)

Abstract: With the rapid development of UAV technology, multi-UAV are widely applied in emergency rescue, environmental monitoring, military reconnaissance and other complex scenarios. However, multi-UAV systems suffer from low path planning efficiency, unstable inter-agent coordination, and inadequate spatiotemporal feature perception of dynamic environments. To solve these problems, this paper proposes a distributed reinforcement learning path planning framework integrating spatiotemporal Transformer and action block mechanism. The framework adopts a distributed Actor-Critic structure to implement decentralized collaborative decision-making, overcoming high communication load and single-point failure of centralized strategies. A scalable input module and multi-objective reward function are constructed to balance path safety, multi-UAV coordination and task efficiency. Combined with spatiotemporal Transformer and action block mechanism, the model extracts key spatiotemporal features and reduces transmission redundancy, so as to enhance feature representation and collaborative decision-making performance. Ablation experiments in dynamic simulation environments demonstrate the superiority and effectiveness of the proposed method. This work provides an effective reference for cooperative path planning of multi-UAV in complex dynamic scenarios.

Keywords: multi-UAVs； reinforcement learning； attention mechanism； distributed control； cooperative path planning；Actor-Critic architecture

0　引　言

随着智能无人系统、多智能体理论及通信技术的快速迭代与深度融合, 多无人机系统作为一种新型分布式智能系统, 正日益广泛地应用于应急救援、

收稿日期: 2026-04-30；**录用日期:** 2026-09-10.
基金项目: 国家自然科学基金项目 (62373015).
责任编委: 柴利.
†通信作者. E-mail: xy_zhang@bjut.edu.cn.

环境监测、军事侦察、物流配送等复杂任务场景中. 与单个无人机相比, 多无人机系统凭借其分布式架构优势, 能够通过多机间的协同配合, 提高复杂任务的执行效率与环境适应能力[1]. 然而, 在实际应用过程中, 多无人机路径规划, 尤其是在障碍物动态变化、任务需求实时调整、通信链路不稳定的复杂动态环境中, 仍面临诸多挑战. 其中, 高效协同调度、智能冲突规避、任务执行效率优化以及系统鲁棒性提升等问题, 已成为制约无人机群体走向实际工程应用的关键瓶颈.

在多无人机协同控制与规划研究中, 鲜斌等将模型预测控制与改进人工势场法相结合, 在降低单步计算时间的同时, 实现了多无人机间的协同避碰与路径规划[2]. 然而, 此类传统启发式方法大多依赖于全局环境信息的已知性, 且对环境变化的适应性较差, 难以适配全局信息未知、障碍物动态移动且需实时动态规划的复杂场景, 其应用范围受到了显著限制.

近年来, 强化学习 (RL) 凭借其较强的环境感知、自主学习与决策优化能力, 被广泛应用于多无人机路径规划研究. 无人机能够通过与动态环境持续交互自主学习控制策略, 降低对先验环境模型的依赖. 相关研究已将深度强化学习用于多无人机轨迹规划与网络编队[3], 以及通信时延条件下的轨迹规划与传输控制[4]. 常见的多智能体强化学习 (MARL) 方法包括带启发式注意力的演员-评论家算法 (SACHA)[5]、多智能体深度确定性策略梯度算法 (MADDPG)[6] 及价值分解网络 (VDN)[7] 等. 其中, MADDPG 及其改进算法因适用于连续动作空间, 在多智能体路径规划中应用较为广泛. 然而, 此类方法在训练阶段通常仍依赖全局信息, 随着智能体数量增加, 计算与通信开销显著上升; 在信息缺失或通信时延条件下, 其实时性和鲁棒性也容易受到影响.

部分可观测条件下, 局部感知信息的可靠性同样影响多智能体系统的协同决策. Olfati-Saber 和 Xin 等分别研究了分布式 Kalman 滤波及故障传感器条件下的状态估计问题[8-9], 闫超等对多智能体深度强化学习的可扩展性与可迁移性进行了综述[10]. 此外, 相关研究分别围绕约束环境下的协同路径规划[11]、部分可观测条件下的多智能体协同[12]、GPS 拒止环境下的多无人机协同定位[13] 以及未知动态环境中的无人机无地图导航[14] 展开, 为复杂通信环境中的多智能体协同提供了研究基础.

在多无人机任务与通信协同方面, 李海峰等基于 MAPPO(Multi-Agent Proximal Policy Optimization) 实现了多无人机动态任务的分布式分配[15]; Javaid 等总结了协同无人机通信与控制领域的研究进展[16]; 王辰等研究了信道资源约束下的航迹与通信策略协同规划[17]. 此外, 深度强化学习还被用于无人机数据采集与路径规划[18] 以及自主多无人机无线网络规划[19]. 然而, 在部分可观测、通信时延和动态障碍物并存的环境中, 如何降低全局信息依赖并有效提取观测序列中的时空关联特征, 仍需进一步研究.

现有基于 DRL(Deep Reinforcement Learning) 的多无人机路径规划算法仍存在不足, 例如, 大多数算法并未构建起多无人机间动态、自适应的协同关系, 仅依赖人工设计的固定优化目标进行决策学习, 难以适应复杂动态环境中任务需求与环境状态的实时变化, 导致协同效率低下、决策精度不足. 本文提出了基于时空 Transformer 的强化学习路径规划控制方法, 主要研究贡献如下:

(1) 针对多无人机路径规划的需求, 设计了基于演员-评论家 (Actor-Critic) 架构的高效模型结构与输入编码方案, 考虑多无人机的位置信息、速度信息、环境障碍物信息, 为多无人机的协同决策、路径优化与动态避障提供了稳定、可靠且具有良好扩展性的框架.

(2) 构建了基于时间-空间 Transformer 的多无人机路径规划模型, 结合多无人机的运动特性与环境动态变化特征, 提出了针对性的注意力机制, 有效解决了现有算法中个体间交互信息利用率低、协同决策精度不足的问题.

(3) 提出了融入动作块的优化设计思路, 将多无人机的连续动作空间进行合理划分, 从而提升了路径规划的逻辑连续性、响应能力与协同执行效率.

1 问题建模

在复杂动态场景下, 多无人机系统由 N 架无人机组成, 记为 $\boldsymbol{X}(t)=[\boldsymbol{X}_1(t),\ldots,\boldsymbol{X}_i(t),\ldots,\boldsymbol{X}_N(t)]^{\mathrm{T}}$, 其中 $\boldsymbol{X}_i(t)$ 为单架无人机的状态, $i\in\{1,2,\ldots,N\}$. 将单架无人机抽象为二维平面内的质点, 定义第 i 架无人机的运动参数如下:

位置向量: $\boldsymbol{p}_i(t)=[x_i(t),y_i(t)]$, 其中 $x_i(t)$、$y_i(t)$ 分别为无人机 i 在 t 时刻在 x 轴、y 轴的位置坐标;

速度向量: $\boldsymbol{v}_i(t)=[v_{x_i}(t),v_{y_i}(t)]$, 其中 $v_{x_i}(t)$、$v_{y_i}(t)$ 分别为无人机 i 在 t 时刻在 x 轴、y 轴的速度分量;

加速度向量: $\boldsymbol{a}_i(t)=[a_{x_i}(t),a_{y_i}(t)]$, 其中 $a_{x_i}(t)$、$a_{y_i}(t)$ 分别为无人机 i 在 t 时刻 x 轴、y 轴的加速度分量.

单架无人机的状态 $\boldsymbol{X}_i(t)=[\boldsymbol{p}_i(t),\boldsymbol{v}_i(t)]=[x_i(t),y_i(t),v_{x_i}(t),v_{y_i}(t)]$, 结合运动学关系, 满足一阶微分方程组:

$$\dot{\boldsymbol{X}}_i(t)=[\dot{\boldsymbol{p}}_i(t),\dot{\boldsymbol{v}}_i(t)]=[\boldsymbol{v}_i(t),\boldsymbol{a}_i(t)]. \tag{1}$$

同时, 每架无人机需满足速度约束 $v_{min}\leqslant v\leqslant v_{max}$; 加速度约束 $a_{min}\leqslant a\leqslant a_{max}$; 无人机与障碍物和其余无人机的距离不能小于安全距离 d_S; 无人机间的通信距离为 d_c, 当无人机与其它所有无人机间的距离都大于 d_c, 则视该无人机脱离集群, 任务失败.

为客观评估多无人机路径规划算法的性能, 建立涵盖个体性能、协同性能与全局性能的多维度评价指标, 如下:

(1) 任务完成率 (TCR): 成功抵达目标点并完成任务的无人机数量占比 $\eta_{\text{task}}=\dfrac{N_{\text{task}}}{N}\times 100\%$ (N_{task} 为完成任务的无人机数量).

(2) 多无人机分散度 (MUD): 飞行时间内, 多无人机集群系统几何中心与个体的平均距离, $\overline{d}_{total}=\dfrac{1}{M}\sum_{t=1}^{M}(\dfrac{1}{N}\sum_{i=1}^{N}d_{i,t})$.

(3) 路径长度 (PL): 各架无人机的平均路径长度 $\overline{L}=\dfrac{1}{N}\sum_{i=1}^{N}L_i$.

(4) 飞行时间: 无人机平均飞行时间 $\overline{T}=\dfrac{1}{N}\sum_{i=1}^{N}T_i$, 以及任务完成时间 $T_{\text{total}}=max\{T_1,T_2,\ldots,T_N\}$.

(5) 避障成功率 (OASR): 成功规避所有障碍物的无人机数量占比 $\eta_{\text{obs}}=\dfrac{N_{\text{safe}}}{N}\times 100\%$, N_{safe} 为无碰撞无人机数量.

(6) 多无人机内部冲突率 (MUCR): 不同无人机之间发生路径冲突的帧数与总飞行时间的比值 $\eta_{\text{conf}}=\dfrac{C_{\text{conf}}}{T_{\text{total}}}$, C_{conf} 为冲突帧数.

(7) 算法实时性: 算法单次路径规划的平均耗时 t_{plan}.

2　基于强化学习的多无人机路径规划方法

2.1　Actor-Critic 框架

本文采用分布式演员-评论家 (Actor-Critic) 框架, 通过去中心化决策以及协同评估的设计, 兼顾了协同性、实时性、灵活性与工程实用性, 能够适配多无人机规模化、动态化的协同需求.

其中, 分布式 Actor 模块采用“去中心化”设计, 由多无人机中每架无人机作为独立的分布式 Actor 智能体, 各自独立生成动作并与自身感知范围内的局部环境进行交互, 并通过局部通信共享信息; 集中式 Critic 模块与仅接收局部观测、输出动作指令的分布式 Actor 不同, Critic 采用集中式价值评估架构, 以全局环境与多无人机状态信息作为输入, 对当前系统整体决策的优劣进行量化评价, 为各智能体的策略更新提供统一、可靠的梯度引导.

本文所用的分布式 Actor-Critic 架构的如图 1 所示.

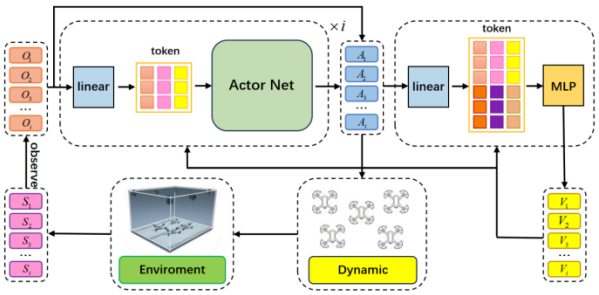


图 1　分布式 Actor-Critic 结构

本文对多无人机的观测信息进行了结构化建模与标记化处理, 为注意力机制的有效运行提供高质量输入.

每架无人机仅能观测到其感知范围内的目标对象 (包括目标点、障碍物及其他无人机), 为明确各观测对象的类型, 为每个对象分配专属标签, 具体分为目标点标签、障碍物标签和无人机标签 $\{flag_{goal},flag_{obs},flag_{agent}\}$, 以区分不同对象对避障决策的影响优先级, 记观测对象总数为 n. 采用 n 个特征窗口, 每个窗口对应 1 个目标对象的 $state$: 绝对位置 (x,y) 及速度 (v_x,v_y) (若是静态障碍物, 记为 $(0,\ 0)$) 和对应的 $flag$. 窗口填充选取规则为: 对感知范围内所有障碍物, 按与无人机的欧氏距离由近及远排序, 优先选取最近的目标对象填入窗口, 直至窗口填满; 若感知范围内目标对象不足 n 个, 则剩余窗口采用标记位填充, 确保该模块维度固定, 为注意力权重的合理分配提供依据, 具体过程如图 2 所示.

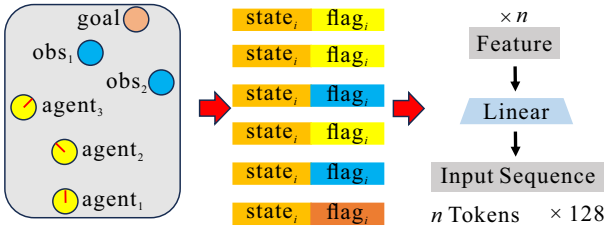


图 2　Actor 网络输入建模

为使观测信息能够被后续 Transformer 网络有效处理, 需对上述状态与标签信息进行嵌入 (Embedding) 转换, 即将每个观测对象的状态参数与

标签拼接形成原始特征向量, 再通过一层可学习的线性变换层将其映射为固定维度的高维嵌入向量. 这一嵌入过程实现了数值型状态与类别型标签的统一表征.

Actor 网络由时间-空间 Transformer 构成, 具体结构如图 3 所示.

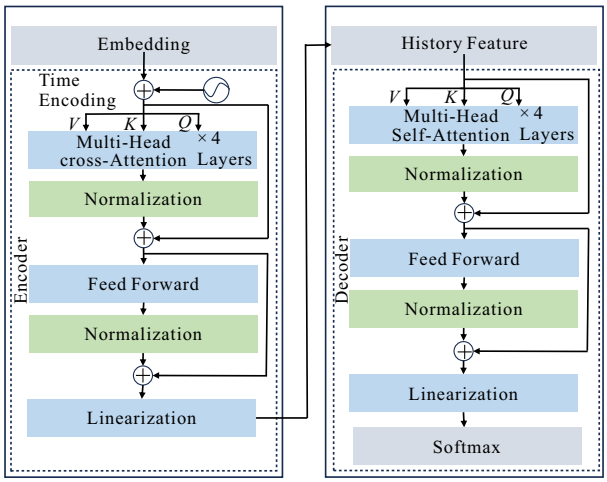


图 3 Actor 整体网络架构

Critic 网络由多层感知机构成, 实现特征降维与有效信息融合; 隐藏层间均采用 ReLU(Rectified Linear Unit) 激活函数, 整体结构如图 4 所示.

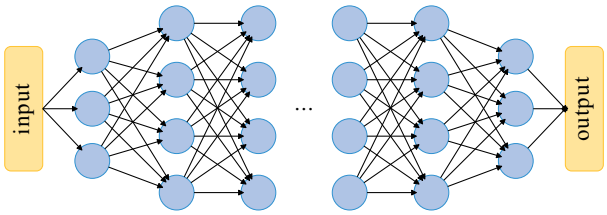


图 4 Critic 网络结构

2.2 时间 Transformer

时间 Transformer 用于描述单个无人机自身历史运动状态的长时程依赖关系, 捕捉个体运动的连续性与规律性, 为避障决策提供历史状态支撑. 时间 Transformer 的输入为单架无人机的历史状态序列嵌入向量, 包含以下部分:

(1) 网络输入

将每个时间步的个体状态$s_i(t-k)(k=0,1,\ldots,T_s-1)$通过全连接层映射为高维嵌入向量$e_t(k)\in R^d$, 其中$d$为嵌入维度, T_s为历史状态序列长度, 实现历史状态的高维表征与特征增强; 原始状态向量均经过归一化处理映射至[0,1]区间, 避免不同维度指标的数值差异导致模型训练失衡, 归一化公式为$\hat{s}=\dfrac{s-s_{\min}}{s_{\max}-s_{\min}}$, 其中$s_{\min}$、$s_{\max}$分别为各指标的最小、最大取值.

(2) 时间位置编码

由于 Transformer 架构本身不具备时序感知能力, 无法区分不同时间步状态的先后顺序, 因此引入时间位置编码$P_t(k)$, 表示每个历史状态在时间序列中的位置信息. 定义时间位置编码向量$P_t(k)\in\mathbb{R}^d$, 对应第k个历史时间步 $(k=0,1,\ldots,T_s-1)$, 编码向量的每个维度 i 的计算公式为:

$$P_t(k)=\begin{cases}\sin\left(\dfrac{k}{10000^{\frac{2i}{d}}}\right) & 当i为偶数时\\ \cos\left(\dfrac{k}{10000^{\frac{2i}{d}}}\right) & 当i为奇数时\end{cases}\tag{2}$$

其中, k为时序位置索引, $k=0$对应当前时刻, k越大, 对应历史时刻越久远; i为嵌入向量的维度索引; $10000^{2i/d}$为缩放因子, 用于控制不同维度编码的周期, 确保不同维度的编码具有不同的频率, 能够区分不同的时间步位置. 将位置编码与状态嵌入向量相加$e_t'(k)=e_t(k)+P_t(k)$, 融合后得到新的嵌入序列$E_t'=[e_t'(0),e_t'(1),\ldots,e_t'(T_s-1)]\in\mathbb{R}^{T_s\times d}$使模型能够精准捕捉历史状态的时序顺序;

(3) 自注意力计算

对处理后的嵌入序列施加因果掩码约束, 进行自注意力计算, 得到时间维度的注意力权重, 通用计算公式如下所示:

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V.\tag{3}$$

其中, Q、K、V分别对应查询向量、键向量和值向量, 三者均由对应的输入序列通过可学习的线性变换得到; d_k表示查询向量与键向量的维度, 引入$\sqrt{d_k}$进行缩放是为了避免QK^{T}计算后数值过大, 导致 Softmax 函数输出趋于极端, 从而影响梯度传播的稳定性.

(4) 前馈网络与残差连接

将自注意力输出通过两层全连接的前馈网络进行特征转换, 并引入残差连接与层归一化, 提升训练稳定性, 最终得到时间维度的特征向量$f_t(i)\in R^d$, 即第i架无人机融合历史状态信息后的特征.

时间 Transformer 的结构如图 5 所示, 通过上述四部分的操作, 能有效捕捉个体历史运动状态的长时程依赖, 为智能体把握自身运动规律、做出连贯的避障决策提供时间维度的特征支撑.

2.3 空间 Transformer

空间 Transformer 是 Actor 网络的空间特征融合核心, 采用多智能体交互建模与全局参数共享模式, 适配分布式网络一次训练、全机复用的核心需求, 同时解决无人机输入排序不一致的问题, 基于自身时序特征与多无人机内其他无人机的交互信息, 完成

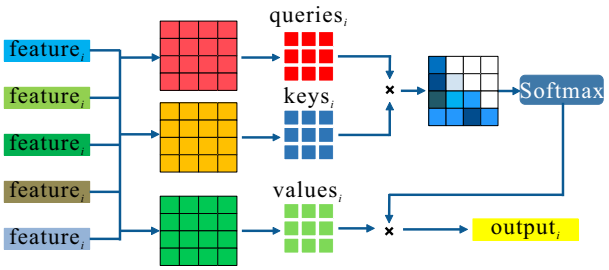


图 5　Actor 网络的时间 Transformer 层

空间特征融合, 包含以下部分:

(1) 网络输入

所有无人机经过时间 Transformer 处理后的时间维度特征向量集合 $F=[f_t(1),f_t(2),\ldots,f_t(N)]$ 通过全连接层进一步增强, 得到 $e_c(i)\in R^d$, 提升个体特征的表达能力, 挖掘个体状态的潜在关联信息.

(2) 空间位置编码

位置编码引入空间位置编码 $P_c(i)$, 表示第 i 架无人机在多无人机中的空间位置信息计算过程为: 首先, 每架无人机实时采集自身的绝对坐标 $\boldsymbol{p}_i=(x_i,y_i)$, 同时通过通信链路同步获取当前时刻内所有其他无人机的绝对坐标 $\boldsymbol{p}_j=(x_j,y_j)(j\neq i,j=1,2,\ldots,N)$, 然后, 基于当前时刻所有无人机的绝对坐标, 计算多无人机中心坐标 $\boldsymbol{p}_{center}=(x_{center},y_{center})$, 记:

$$\begin{cases} x_{center}=\dfrac{1}{N}\displaystyle\sum_{k=1}^{N}x_k \\ y_{center}=\dfrac{1}{N}\displaystyle\sum_{k=1}^{N}y_k \end{cases} \tag{4}$$

为全局空间参考点. 接着, 计算相对于所有其他无人机的平均相对坐标 $\Delta\boldsymbol{p}_{avg,i}$, 捕捉当前无人机与系统内所有邻机的整体交互距离, 强化空间交互关联性, 公式如下所示:

$$\Delta\boldsymbol{p}_{avg,i}=\frac{1}{N-1}\sum_{j\neq i}(x_i-x_j,y_i-y_j)\in\mathbb{R}^2. \tag{5}$$

最后, 融合双重相对坐标, 将上述两个 2 维相对坐标拼接, 得到 4 维的总相对坐标, 通过全连接层将 4 维的总相对坐标映射为与空间 Transformer 输出一致维度的高维编码向量, 得到空间位置编码 $P_c(i)$. 将空间位置编码与个体嵌入向量相加, 得到 $e_c'(i)=e_c(i)+P_c(i)$.

(3) 交叉注意力计算

采用交叉注意力机制建模个体间的交互关系, 计算第 i 架无人机与其他所有无人机 $j(j=1,2,\ldots,N,j\neq i)$ 的注意力权重. 交叉注意力权重反映了其他个体对当前个体决策的影响程度.

(4) 前馈网络与残差连接

与时间 Transformer 类似, 将交叉注意力输出通过前馈网络进行特征转换, 引入残差连接与层归一化 (Layer Normalization), 有效缓解梯度消失问题、提升训练稳定性, 最终得到集群维度的融合特征向量 $f_c(i)\in R^d$, 即第 i 架无人机融合自身历史状态与集群交互信息后的最终特征.

空间 Transformer 的结构如图 6 所示. 通过上述四部分的操作, 可实现个体间空间交互信息的高效聚合, 为避障决策提供协同支撑.

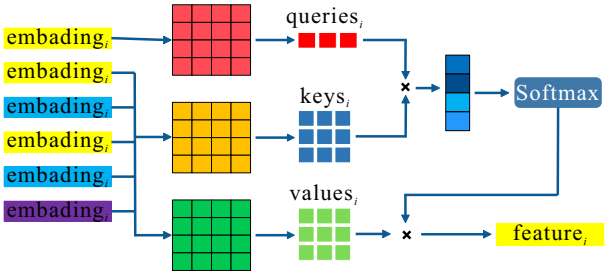


图 6　Actor 网络的空间 transformer 层

2.4　动作块机制

动作块 (Action Chunk) 是解决长视野任务的有效范式, 其核心是让智能体直接学习连续动作序列而非单步动作, 实现输入多帧观测信息、输出多帧连续动作的块级决策: 以双层 Transformer 输出的特征为基础, 将连续的避障动作序列按“动作连贯性、逻辑完整性”划分为若干个长度为 h 的动作块, 每个动作块对应一段连续、完整的避障动作片段, 通过一次模型推理, 直接输出 h 步连续动作, 而非单步动作.

动作块可记为 Chunk=$[a_t,a_{t+1},\ldots,a_{t+h-1}]$, 其中 t 为动作块的起始索引, h 为动作块长度, 即每个块包含的动作步数, 也是一次推理输出的连续动作帧数; a_t 表示第 t 步的动作集合, 包含多无人机系统内所有无人机的动作指令, 具体框架如图 7 所示.

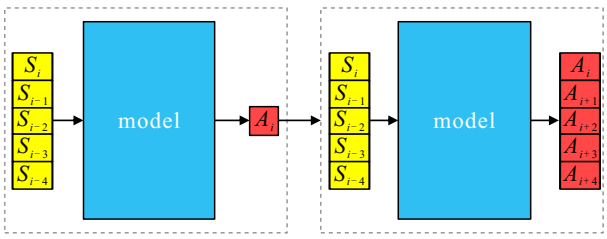


图 7　动作块机制

为进一步提升多无人机避障动作的平滑性与鲁棒性, 削弱单次推理噪声与环境突变对决策结果的影响, 本研究在 Action Chunk 机制的基础上, 提出逐帧推理–滑动窗口加权融合决策策略. 该策略每步均执行一次完整决策推理, 但实际输出动作并非直接

采用当前时刻推理结果, 而是将最近h步内所有时刻的推理输出进行加权融合, 利用历史推理信息对当前动作进行约束与平滑, 显著提升动作时序一致性.

设动作块长度为h, 在任意时刻k, 系统均会基于双层 Transformer 与当前环境观测, 完成一次前向推理, 得到该时刻推理输出的动作序列:

$$A_k = [a_k, a_{k+1}, \ldots, a_{k+h-1}]. \tag{6}$$

其中, A_k为时刻k推理得到的长度为h的动作块, a_{k+i}为该块中第i步的动作. 与传统“推理一次、执行h步”的模式不同, 本策略中每一步均执行推理, 并在时刻t输出最终执行动作a_t^{out}, 其由以t为终点、长度为h的滑动窗口内所有有效推理结果加权求和得到:

$$a_t^{\text{out}} = \sum_{i=t-h+1}^{t} \omega_{t-i} \cdot a_t^{(i)}, \sum \omega_{t-i} = 1. \tag{7}$$

其中, $a_t^{(i)}$表示在时刻i推理得到、对应执行时刻 t 的动作, ω_{t-i}为时刻i推理结果对应的权重, $t-i$为推理时刻与执行时刻的时间间隔; 满足权重归一化约束$\sum \omega_{t-i} = 1$. 其示意图如图 8 所示.

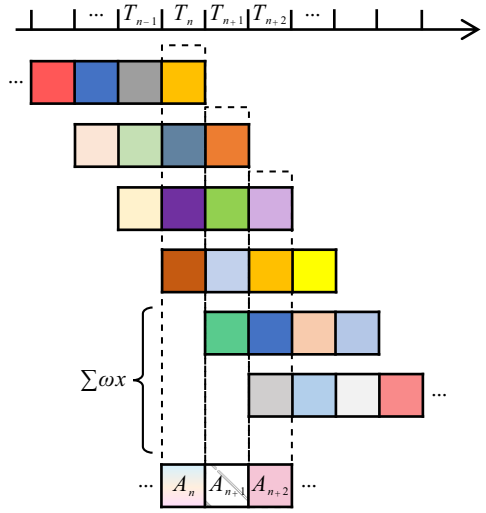


图 8 推理时的时序加权融合示意图

2.5 奖励函数

奖励函数是强化学习智能体学习最优策略的导向, 其设计需精准匹配多无人机集群避障、目标达成、队形保持及作业效率的任务需求. 本研究采用多目标融合式奖励函数, 通过分模块设计各子奖励项, 再经加权融合得到最终奖励, 具体设计如下:

(1) 避障奖励项

该奖励项的核心目标是引导多无人机规避障碍物, 保障飞行安全. 其设计基于无人机与障碍物的距离阈值划分风险等级, 实施差异化奖惩: 障碍物半径为R_O, 当无人机与障碍物的距离大于安全距离d_S时, 不施加奖惩; 当无人机与障碍物的距离处于安全距离至障碍物半径之间时, 判定为处于碰撞风险区域, 此时施加与距离负相关的惩罚项; 当无人机与障碍物的距离小于障碍物半径时, 判定为发生碰撞, 无人机损毁, 施加严重惩罚, 其公式如下:

$$r_{obs} = \begin{cases} 0, & d_O > R_O + d_S \\ -\dfrac{10}{d_O}, & R_O \leqslant d_O \leqslant R_O + d_S \\ -500, & d_O < R_O \end{cases} \tag{8}$$

(2) 目标达成奖励项

目标达成奖励项如公式 9 所示, 该奖励项用于引导多无人机向目标点协同移动, 确保任务的核心目标达成. 当系统内所有无人机均进入目标点预设范围时, 判定为任务成功, 施加高额正向奖励; 若未达成目标, 则基于多无人机系统平均中心与目标点的欧氏距离施加惩罚, 距离越远惩罚越重, 推动持续向目标点靠拢.

$$r_{goal} = \begin{cases} 5000, & d_G \leqslant D \\ -0.001 \times d_G, & d_G > D \end{cases} \tag{9}$$

(3) 队形保持奖励项

队形保持奖励项如公式 10 所示, 针对多无人机协同作业的要求, 设计该奖励项以维持内部的相对位置稳定性. 其判定标准为多无人机内任意两无人机间的距离是否处于预设合理范围: 若存在无人机间距离超出合理范围的情况, 即多无人机内个体间距小于最大通信距离d_{max}, 大于最小间距d_{min}, 则基于超出距离的程度施加惩罚.

$$r_{follow} = -\sum_{i \neq j} (\lambda_{close} \cdot \max(d_{min} - d_{ij}, 0)) + \lambda_{far} \cdot max(d_{ij} - d_{max}, 0)). \tag{10}$$

(4) 时间效率奖励项

为提升作业效率, 避免无效徘徊与冗余飞行, 引入每步固定时间惩罚, 引导其学习快速、高效、无冗余的协同策略. 其公式如下:

$$r_{time} = -1. \tag{11}$$

将上述四个子奖励项通过加权求和得到最终奖励, 公式为

$$r = \omega_{obs} r_{obs} + \omega_{goal} r_{goal} + \omega_{follow} r_{follow} + \omega_{time} r_{time}. \tag{12}$$

其中, α、β、γ、δ为各子奖励项的权重系数, 满足$\alpha + \beta + \gamma + \delta = 1$. 权重设置需根据任务优先级动态调整, 确保奖励函数的导向与实际任务需求匹配.

2.6 模型训练

训练方法选用 MADDPG 方法, 为适配动作块, 需对训练过程进行针对性改动, 确保训练过程与动

作块机制深度兼容, 将传统单步 TD 误差升级为块级 TD 误差, 实现无偏多步回溯, 缓解稀疏奖励下的价值估计偏差, 具体计算方式如下:

首先, 计算动作块Chunk_t的块级累积奖励$R_{\text{chunk},t}$, 融合块内h步单步奖励, 强化动作块与奖励的关联度, 公式为:

$$R_{\text{chunk},t} = \sum_{k=t}^{t+h-1} \gamma^{k-t} \cdot r_k. \tag{13}$$

其中, r_k为动作块内第k步的单步即时奖励, $\gamma \in [0,1]$为折扣因子, 用于衡量未来奖励的权重, 与传统 TD 方法的折扣因子设置逻辑一致, 确保时序差分的合理性.

在此基础上, 块级 TD 误差$\delta_{\text{chunk},t}$的计算参考论文无偏多步回溯思路, 以动作块为单元, 关联当前动作块的块级价值与下一动作块的块级价值, 公式为:

$$\delta_{chunk,t} = R_{chunk,t} + \gamma^h \cdot Q(O_{t+h}, Chunk_{t+1}) - Q(O_t, Chunk_t). \tag{14}$$

其中, $Q(O_{t+h}, \text{Chunk}_{t+1})$为下一动作块$\text{Chunk}_{t+1}$的块级价值估计, γ^h为折扣因子的h次方, 适配动作块长度h, 确保下一动作块价值的长期权重与动作块长度匹配, 实现“动作块-累积奖励-下一动作块价值”的强关联, 将传统单步 TD 的变为块级引导, 缓解稀疏奖励下的探索盲目性.

同时, 对于经验回放机制, 使用存储“块观测-动作块-块级累积奖励-下一块观测”的块级经验元组, 确保经验回放与动作块的输出逻辑一致. 训练流程如算法 1 所示.

算法 1　MADDPG训练流程.

多智能体深度确定性梯度算法(MADDPG)

使用随机权重μ和ω来初始化Actor网络和Critic网络

使用权重$\mu' = \mu$和$\omega' = \omega$来初始化目标Actor网络和Critic网络

初始化经验回放池D_{rl}

for 片段 = 1 : 最大片段个数 do

  for t = 1 : 最大步数 do

  执行器依据当前多帧观测O_t产生动作块 $Chunk_t = A^{\mu}(O_t)$

  在动作块中增加噪声来探索 $a_t' = a_t + v_t$

  执行探索动作块Chunk_t', 依次得到h步对应的新状态$s_t, s_{t+1}, \ldots, s_{t+h-1}$和单步奖励$r_t, r_{t+1}, \ldots, r_{t+h-1}$, 最终得到动作块对应的多帧新观测$O_{t+1}$

  计算当前动作块的累积奖励

$$R_{\text{chunk},t} = \sum_{k=t}^{t+h-1} \gamma^{k-t} \cdot r_k$$

  将块级转移经验$e(t) = (O_t, \text{Chunk}_t, R_{\text{chunk},t}, O_{t+1})$存入经验缓冲区$D_{rl}$

  从经验回放池D_{rl}中均匀抽取随机样本e_i

  计算每个块的时序差分误差 δ

$$\delta_{\text{chunk},i} = R_{\text{chunk},i} + \gamma^h \cdot Q_{\omega'}(O_{i+1}, A_{\mu'}(O_{i+1})) - Q_{\omega}(O_i, \text{Chunk}_i)$$

  计算损失函数 $L(w)$

$$L(w) = \frac{1}{N_{\text{rl}}} \sum_{i=1}^{N_{\text{rl}}} \delta_{\text{chunk},i}^2$$

  使用梯度下降法更新Critic网络

$$\nabla_w L(\omega) = \frac{1}{N_{\text{rl}}} \sum_{i=1}^{N_{\text{rl}}} \delta_{\text{chunk},i} \cdot \nabla_{\omega} Q_{\omega}(O_i, \text{Chunk}_i)$$

$$w_{t+1} = w_t - \alpha_w \nabla_w L(w)$$

  使用策略梯度下降法更新Actor网络

$$\nabla_{\mu} J(\mu) = \frac{1}{N_{\text{rl}}} \sum_{i=1}^{N_{\text{rl}}} w_i \cdot \nabla_{\mu} A_{\mu}(O_i) \cdot \nabla_{\text{Chunk}_i} Q_{\omega}(O_i, \text{Chunk}_i)$$

$$\mu_{t+1} = \mu_t - \alpha_{\mu} \cdot \nabla_{\mu} J(\mu)$$

  更新目标网络

$$\mu' = \tau\mu + (1-\tau)\mu'$$

$$w' = \tau w + (1-\tau)w'$$

  if 任务结束 then

    终止当前片段

  end if

 end for

end for

上述算法中, A和Q分别是由 μ, μ'和 w, w' 参数化定义的 Actor 和 Critic 神经网络, η_e为定义的折扣因子, τ为很小的目标网络更新速率, 它满足 $\tau \ll 1$, 而α_{μ}, α_w 分别为 Actor 和 Critic 网络的学习速率.

3　仿真实验及结果分析

3.1　实验环境

仿真实验的软硬件环境及仿真参数设置如下:

(1) 硬件环境: CPU 为 Intel Core i7-12700H, 显卡 NVIDIA RTX 4090.

(2) 软件环境: 操作系统为 Ubuntu 24.04, 仿真平台采用 Python 3.8, 结合 PyTorch 深度学习框架实现分布式 Actor-Critic 架构的搭建与训练; 采用 PyGame 搭建多无人机仿真场景; 利用 NumPy、Matplotlib 等完成实验数据的处理与可视化分析.

(3) 参数设置: 仿真场景设置为$100m \times 100m$的二维平面区域, 包含M个静态障碍物, 目标点随机设置; 无人机数量N范围取值为 3-10 架; 无人机初始位置随机分布于场景边缘, 初始速度随机, $v_{max} = 15$, $a_{min} = -1, a_{max} = 1, d_S = 2$, $n = 2$. 灰色为静

态障碍物, 红色为动态障碍物, 以随机方向和速度运动, 若撞到实验场景边缘会反弹, 主要参数设置如表 1 所示.

表 1 实验参数

参数	数值	参数说明
Actor网络学习率	0.000 1	控制网络参数更新速度
Critic网络学习率	0.000 5	控制网络参数更新速度
优化器	Adam	用于更新网络的优化器
价值折扣因子	0.95	权衡即时与长期价值
目标网络软更新系数	0.005	确保目标网络参数平稳更新
经验回放池容量	20 000	存储无人机与环境交互的时序经验
采样批次大小	32	平衡分布式训练的效率与稳定性
最大训练片段数	10 000	确保充分训练并收敛
每个片段最大步数	500	适配无人机协同任务的周期
动作探索噪声方差	0.1	提升网络环境探索能力
动作块长度	5	控制每帧输出未来多少帧动作

3.2 实验结果及分析

(1) 消融实验

消融实验目的是验证双层 Transformer 架构在时空信息分离与融合、建模时序依赖与空间交互中的必要性; 同时, 验证动作块机制在强化动作时序一致性、降低决策冗余、提升避障决策效率与稳定性中的作用, 并验证双层 Transformer 架构和动作块机制的兼容性.

本节设计了 3 组消融模型, 每组仅改变一个模块, 确保单一变量原则, 具体如下:

a) Baseline: Actor 和 Critic 网络结构为 MLP, 即传统 MADDPG.

b) 消融模型 1(Ablation-1): 在 Baseline 基础上, 增加双层 Transformer 架构, 建模多无人机内个体间的空间交互与个体历史状态的长时程依赖 h.

c) 消融模型 2(Ablation-2): 在 Baseline 基础上, 增加动作块机制, 进行动作块划分, 在每个决策帧直接输出未来多帧动作指令, 再按时间加权融合得到最终动作指令.

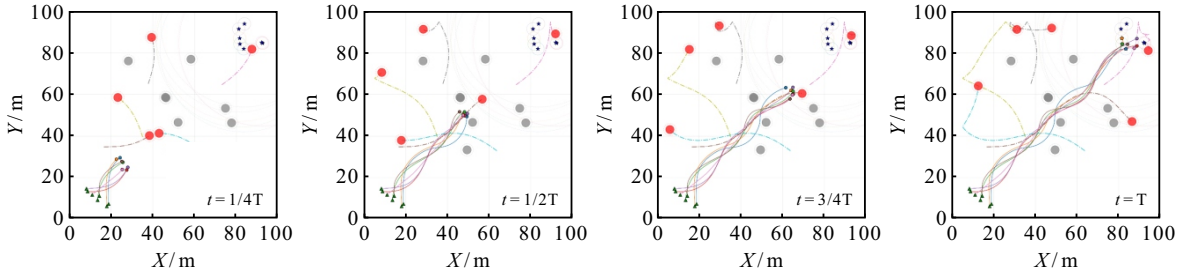
d) 消融模型 3(Ablation-3): 在 Baseline 基础上, 同时增加双层 Transformer 架构与动作块机制, 是本文所提出的方法 (简称为本文方法).

在不同无人机数量设置下, 均在随机场景中独立重复 100 组实验, 并以各指标的平均值作为最终结果. 图 9 给出了 3 种仿真场景下的路径规划结果, 其中图 9(a)～(d) 分别为 Baseline、Ablation-1、Ablation-2 和 Ablation-3 在场景 1 中的结果, 图 9(e) 和图 9(f) 分别为 Ablation-3 在场景 2 和场景 3 中的结果. 每个子图依次展示四个典型时刻; 虚线表示动态障碍物轨迹, 实线表示无人机轨迹. 图 10 进一步给出了不同无人机数量下各方法的指标对比.

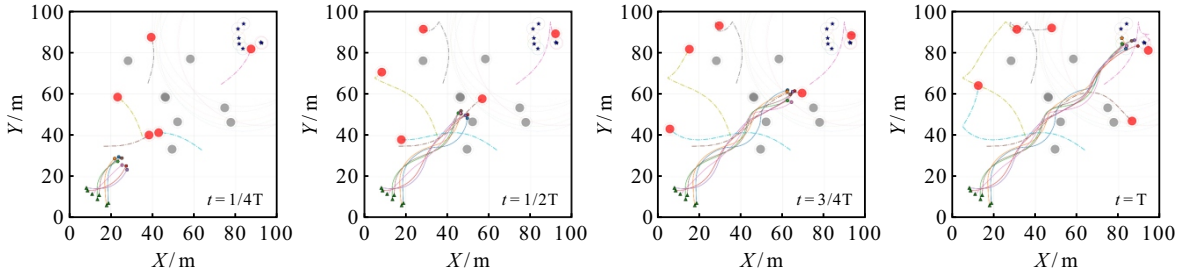
表 2 给出 Baseline 和另外三种模型的综合性能对比, 相关变化趋势在消融实验指标对比柱状图中进行了进一步展示. 从任务完成率来看, Baseline 的任务完成率为 90.2%, Ablation-1 引入双层 Transformer 架构后提升至 91.01%, 说明该架构通过分离并融合空间交互与时序依赖, 有效降低了因信息耦合导致的决策偏差, 提升了多无人机整体避障的准确性; Ablation-2 仅引入动作块机制后任务完成率降至 87.89%, 原因在于单一动作块机制虽强化了动作时序一致性, 但缺乏时空信息的有效建模, 无法应对内个体间的动态空间交互, 导致部分场景下避障决策状态不匹配; Ablation-3 融合两种机制后任务完成率达到 91.82%, 较 Baseline 提升 1.62 个百分点, 验证了双层 Transformer 与动作块机制的协同效应, 前者解决时空信息建模的完整性问题, 后者保障动作决策的时序一致性, 二者结合实现了精准感知到稳定决策的闭环优化. 图 10(a) 中, Ablation-3 在不同无人机数量下的任务完成率整体较高, 进一步验证了双层 Transformer 与动作块机制的协同作用.

由表 2 和图 10(d), 将避障成功率进一步拆解为静态与动态避障两个维度: 在静态避障场景下, 各模型成功率均保持在 87% 以上, 差异较小, 原因在于静态障碍物位置固定, Baseline 的基础框架已能满足感知需求, 新增模块的增益有限; 动态避障场景下, Ablation-1 较 Baseline 提升 0.9 个百分点, Ablation-3 进一步提升至最优, 结合图 10(e), 随着无人机数量增加, 各方法的动态避障成功率整体有所下降, 但 Ablation-3 在多数规模下仍保持较高水平, 这体现了双层 Transformer 对动态场景下时序依赖的建模能力, 通过捕捉个体历史运动状态, 可提前预判动态障碍物的运动趋势, 提升避障决策的前瞻性; 而 Ablation-2 略低于 Baseline, 说明缺乏时空建模的动作块机制无法适配动态场景的不确定性.

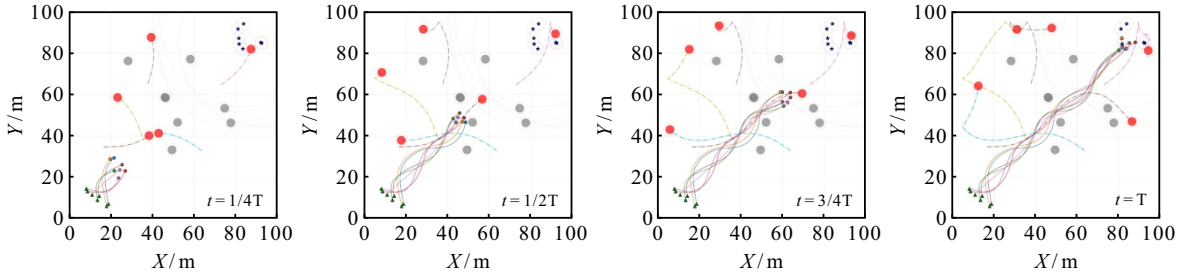
多无人机分散度反映避障过程中多无人机的队形保持能力, 路径长度则体现决策的路径优化水平. 表 2 中 Baseline 的分散度为 0.87m, Ablation-1 降至 0.84m, 说明双层 Transformer 通过建模个体间的空间交互, 可有效协调个体的运动轨迹, 避免因局部避障导致的队形分散; Ablation-2 的分散度骤升至 1.57m, 是所有模型中最差的, 原因在于动作块机制直接输出多帧动作指令, 缺乏对个体空间位置的实时调整, 导致部分个体过度偏离系统中心; Ablation-3 的分散度为 0.86m, 接近 Baseline 与 Ablation-1 的



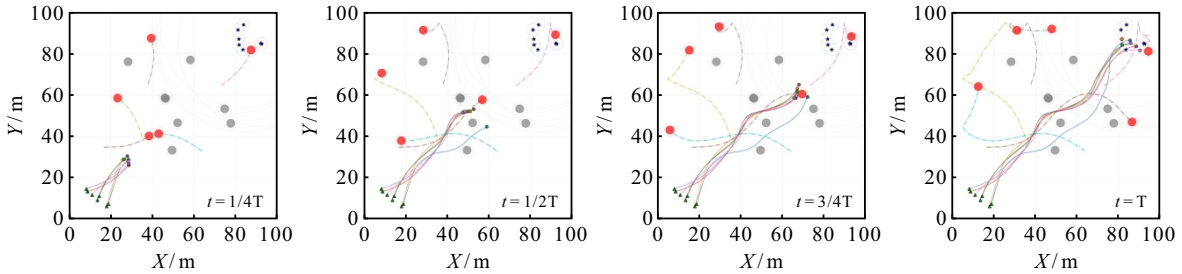
(a) Baseline 方法在场景1下的实验结果



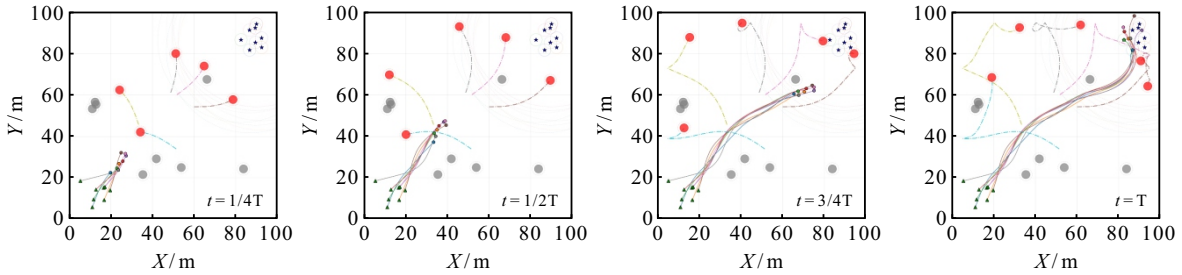
(b) Ablation1 方法在场景1下的实验结果



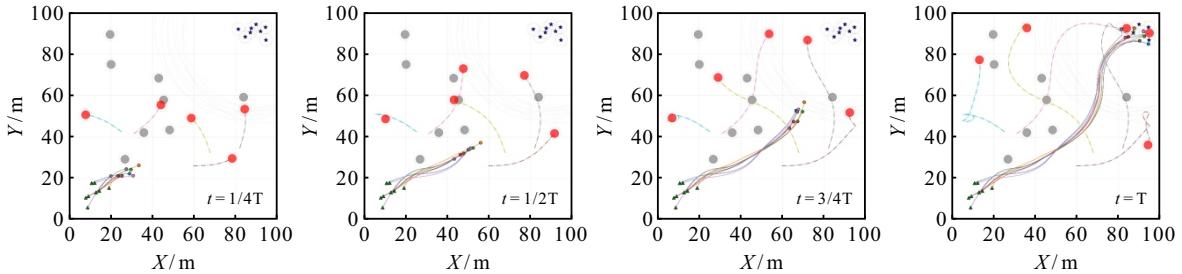
(c) Ablation2 方法在场景1下的实验结果



(d) Ablation3 方法在场景1下的实验结果



(e) Ablation3 方法在场景2下的实验结果



(f) Ablation3 方法在场景3下的实验结果

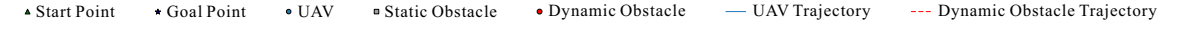


图 9　消融实验中不同方法及场景下的多无人机路径规划结果

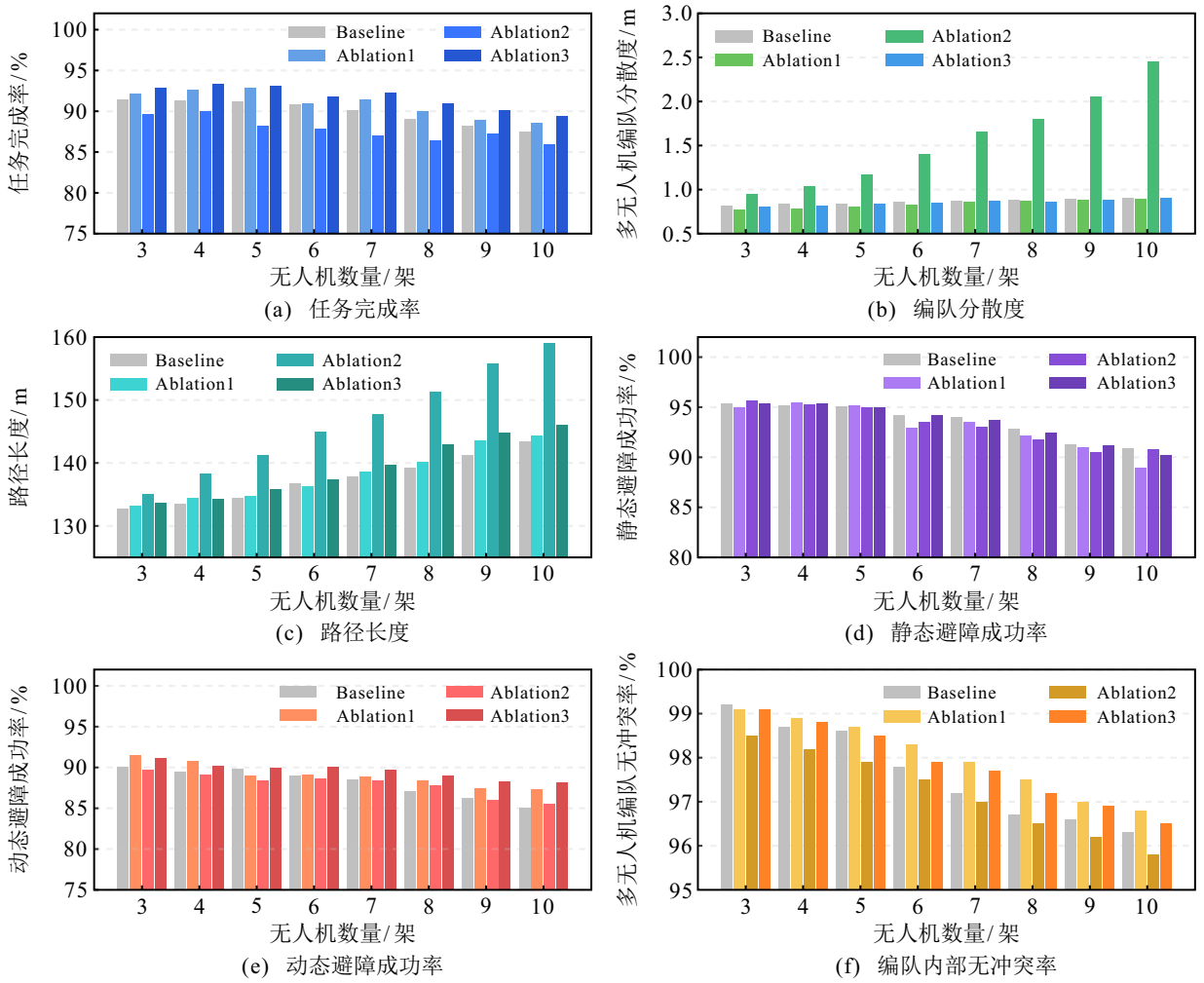


图 10 消融实验中多无人机路径规划指标的对比柱状图

表 2 消融实验结果对比

Methods	Baseline	Ablation1	Ablation2	Ablation3
TCR	90.20	91.01	87.89	91.82
MUD	0.87	0.84	1.57	0.86
PL	137.45	138.21	146.75	139.38
SOASR	93.62	93.06	93.21	93.48
DOASR	88.21	89.11	88.05	89.62
MUCR	97.64	98.01	97.20	97.82

最优水平, 结合图 10(b) 中 Ablation-3 变化相对平稳, 进一步表明融合架构可在保障动作时序一致性的同时, 通过 Transformer 的空间交互建模维持队形.

路径长度维度, Ablation-1 相比 Baseline 略有增加, 这是因为 Transformer 架构的时空建模需对个体轨迹进行精细化调整, 导致路径小幅延长; Ablation-2 路径长度显著增加, 源于动作块机制的多帧动作指令易导致“路径冗余”, 且缺乏时空优化的情况下, 个体易选择次优避障路径; Ablation-3 的路径长度为 139.38m, 虽略高于 Baseline, 但远低于 Ablation-2, 结合图 10(c), 随无人机数量增加, Ablation-2 的路径长度增长最为明显, 而 Ablation-3 仍能保持相对平稳, 说明双层 Transformer 可有效弥补动作块机制的路径冗余问题, 通过时空信息融合优化轨迹规划.

多无人机编队内部无冲突率反映个体间的运动协同性, 表 2 中 Baseline、Ablation-1、Ablation-2 和 Ablation-3 的内部无冲突率分别为 97.64%、98.01%、97.20% 和 97.82%. Ablation-1 与 Ablation-3 均高于 Baseline, 其中 Ablation-1 提升最为显著, 由图 10(f), 随着无人机数量增加, 各模型的内部无冲突率均有所下降, 但 Ablation-1 和 Ablation-3 仍保持了较好的协同性; 这验证了 Transformer 架构对个体空间交互的建模能力可有效降低编队内碰撞风险; Ablation-2 无冲突率略低, 体现单一动作块机制无法有效协调个体空间位置.

(2) 不同无人机数量下的性能分析

为分析无人机数量对算法性能的影响, 分别统计了本文算法在 3、5、7 和 9 架无人机工况下的各项评价指标, 结果如表 3 所示.

从表 3 可以看出, 随着无人机数量增大, 任务完成率、静动态障碍物避障成功率呈缓慢下降趋势, 多机内部冲突率小幅降低; 集群分散度、路径长度有所上升. 整体来看, 无人机数量提升后本文算法性能衰

表 3　不同无人机数量下的指标对比

	$N=3$	$N=5$	$N=7$	$N=9$
TCR	93.38	92.92	91.01	89.75
MUD	0.82	0.81	0.86	0.88
PL	134.58	137.47	137.92	142.94
SOASR	94.46	93.89	93.54	91.76
DOASR	91.21	91.31	89.13	88.84
MUCR	99.15	98.36	97.21	96.36

减幅度有限, 在多架无人机场景下仍可维持较高任务完成率与避障成功率, 验证了所提时空 Transformer 架构具备良好多无人机扩展能力.

(3) 不同动作块长度下的算法性能分析

为分析动作块长度对本文提出算法能力的影响, 分别设置动作块长度为 3, 5, 7, 针对每种无人机数量, 在随机场景下各重复 100 组实验, 取实验结果的平均值作为最终数据. 对比不同动作块长度下的算法性能, 具体结果见表 4.

表 4　动作块长度 h 取不同值时的指标对比

	$h=3$	$h=5$	$h=7$
TCR	91.03	91.82	87.24
MUD	0.86	0.86	0.88
PL	139.17	139.38	142.81
SOASR	93.72	93.48	90.31
DOASR	89.81	89.62	86.14
MUCR	96.93	97.82	95.32

从表 4 可以看出, 当动作块长度由 3 提升至 5 时, 任务完成率、内部无冲突率有所提升, 编队分散度基本持平, 路径长度、动静障碍物避障成功率仅有小幅波动; 块长度增至 7 后, TCR、SOASR、DOASR 均明显下降, 平均路径长度增加, 集群分散度变差. 较短块长预测视野不足, 难以充分发挥连续平滑决策优势; 块长过大会加剧多步动作预测漂移, 降低避障与任务完成性能.

(4) 不同路径规划算法性能对比

进一步地, 将本文方法与人工势场法 APF (artificial potential field)、基础 MADDPG 方法进行对比, 在每种无人机数量在随机场景下各重复 100 组实验, 取结果的平均值作为最终数据, 结果如表 5 所示.

表 5　不同方法的实验结果对比

Method	APF	MADDPG	MAPPO	本文方法
TCR	84.86	90.20	90.57	91.82
MUD	2.08	0.87	0.86	0.86
PL	153.62	137.45	138.92	139.38
SOASR	91.36	93.62	93.65	93.48
DOASR	82.65	88.21	88.31	89.62
MUCR	98.29	97.64	97.92	97.82

从表 5 可以看出, 本文方法的任务完成率和动态避障成功率分别达到 91.82% 和 89.62%, 均为所有方法中的最高值. 本文方法的编队分散度为 0.86 m, 与 MAPPO 相同, 略优于 MADDPG, 并明显优于 APF. 其平均路径长度为 139.38 m, 显著低于 APF, 但分别比 MADDPG 和 MAPPO 高 1.93 m 和 0.46 m. 本文方法的静态避障成功率为 93.48%, 与 MADDPG 和 MAPPO 接近; 内部无冲突率为 97.82%, 高于 MADDPG, 但略低于 APF 和 MAPPO. 综合来看, 本文方法在仅小幅增加路径长度的情况下, 取得了更高的任务完成率和动态避障成功率, 体现了其在复杂动态环境中的综合优势.

4　结　论

本文针对复杂动态环境下的多无人机协同路径规划问题, 提出了一种融合时空 Transformer 与动作块机制的分布式强化学习路径规划方法. 首先, 结合多无人机运动特性与协同任务需求完成问题建模, 并构建分布式 Actor-Critic 控制框架, 通过去中心化策略决策与集中式价值评估, 提高系统的协同性、实时性与可扩展性. 其次, 针对多无人机历史状态依赖、个体空间交互关系复杂以及观测信息冗余等问题, 设计时空双层 Transformer 网络及泛化性强的位置编码, 实现时间特征与空间交互特征的分离提取与融合. 在此基础上, 引入动作块机制并实现其与注意力网络的深度融合, 并对奖励函数、块级时序差分误差及训练流程进行相应设计. 多组消融实验验证了时空 Transformer 在历史状态建模、空间关系提取和队形保持方面的有效性, 同时说明动作块机制与时空特征提取网络具有较好的协同作用. 与传统路径规划方法及经典多智能体强化学习算法相比, 本文方法在任务完成率、动态障碍物避障成功率和集群协同稳定性等方面表现出较好的综合性能, 验证了所提方法的有效性与可行性.

下一步将考虑通信延迟/中断、传感器噪声及障碍物运动不确定性等实际因素, 通过半实物仿真与真实无人机平台实验, 验证所提方法在实际复杂环境中的应用能力.

参考文献 (References)

[1] 于彦鹏, 余墨多, 汤奇荣, 等. 面向城市应急物资配送的多无人机协同路径规划算法[J]. 控制与决策, 2025, 40(4): 1098-1106.
(Yu Y P, Yu M D, Tang Q R, et al. Multi-UAV cooperative path planning algorithm for urban emergency material distribution[J]. Control and Decision, 2025, 40(4): 1098-1106.)

[2] 鲜斌, 宋宁. 基于模型预测控制与改进人工势场法的多无人机路径规划[J]. 控制与决策, 2024, 39(7): 2133-2141.
(Xian B, Song N. A multiple UAVs path planning method based on model predictive control and improved artificial potential field[J]. Control and Decision, 2024, 39(7): 2133-2141.)

[3] Gong S M, Wang M, Gu B, et al. Bayesian optimization enhanced deep reinforcement learning for trajectory planning and network formation in multi-UAV networks[J]. IEEE Transactions on Vehicular Technology, 2023, 72(8): 10933-10948.

[4] Fan Z S, Gong S M, Long Y S, et al. Delay-tolerant multi-agent DRL for trajectory planning and transmission control in UAV-assisted wireless networks[C]. IEEE 99th Vehicular Technology Conference. Singapore, 2024: 1-5.

[5] Lin Q S, Ma H. SACHA: Soft actor-critic with heuristic-based attention for partially observable multi-agent path finding[J]. IEEE Robotics and Automation Letters, 2023, 8(8): 5100-5107.

[6] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative -competitive environments[C]. Proceedings of the 31st Int Conf on Neural Information Processing Systems. Long Beach: Curran Associates, 2017: 6379-6390.

[7] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C]. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm: IFAAMAS, 2018: 2085-2087.

[8] Olfati-Saber R. Distributed Kalman filtering for sensor networks[C]. Proceedings of the 46th IEEE Conference on Decision and Control. New Orleans, 2007: 5492-5498.

[9] Xin D J, Shi L F, Yu X K. Distributed Kalman filter with faulty/reliable sensors based on Wasserstein average consensus[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(4): 2371-2375.

[10] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 3083-3102.
(Yan C, Xiang X J, Xu X, et al. A survey on scalability and transferability of multi-agent deep reinforcement learning[J]. Control and Decision, 2022, 37(12): 3083-3102.)

[11] 董立静, 肖思哲, 牛思, 等. 约束环境下基于深度强化学习的协同路径规划研究[J]. 控制与决策, 2025, 40(6): 1838-1846.
(Dong L J, Xiao S Z, Niu S, et al. Collaborative pathfinding research based on deep reinforcement learning in constrained environments[J]. Control and Decision, 2025, 40(6): 1838-1846.)

[12] 王子豪, 张严心, 黄志清, 等. 部分可观测下基于RGMAAC算法的多智能体协同[J]. 控制与决策, 2023, 38(5): 1267-1277.
(Wang Z H, Zhang Y X, Huang Z Q, et al. Multi-agent collaboration based on RGMAAC algorithm under partial observability[J]. Control and Decision, 2023, 38(5): 1267-1277.)

[13] 万开方, 吴志林, 武韫晖, 等. 拒止环境下基于深度强化学习的多无人机协同定位[J]. 航空学报, 2025, 46(8): 331024.
(Wan K F, Wu Z L, Wu Y H, et al. Cooperative location of multiple UAVs with deep reinforcement learning in GPS-denied environment[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(8): 331024.)

[14] 姜凌峰, 李新凯, 张海, 等. 基于改进TD3算法的无人机动态环境无地图导航[J]. 航空学报, 2025, 46(8): 331035.
(Jiang L F, Li X K, Zhang H, et al. Mapless navigation of UAVs in dynamic environments based on an improved TD3 algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(8): 331035.)

[15] 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, 2025, 40(5): 1429-1437.
(Li H F, Yang H A, Sheng Z M, et al. Multi-UAV cooperative distributed dynamic task allocation based on MAPPO[J]. Control and Decision, 2025, 40(5): 1429-1437.)

[16] Javaid S, Saeed N, Qadir Z, et al. Communication and control in collaborative UAVs: Recent advances and future trends[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(6): 5719-5739.

[17] 王辰, 魏才盛, 殷泽阳, 等. 考虑信道资源约束的多无人机航迹与通信策略协同规划[J]. 航空学报, 2025, 46(18): 331837 .
(Wang C, Wei C S, Yin Z Y, et al. Collaborative planning of multi-UAV trajectories and communication strategies considering channel resource constraints[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(18): 331837.)

[18] 牟治宇, 张煜, 范典, 等. 基于深度强化学习的无人机数据采集和路径规划研究[J]. 物联网学报, 2020, 4(3): 42-51.
(Mou Z Y, Zhang Y, Fan D, et al. Research on the UAV-aided data collection and trajectory design based on deep reinforcement learning[J]. Chinese Journal on Internet of Things, 2020, 4(3): 42-51.)

[19] Bai Y, Zhao H, Zhang X, et al. Toward autonomous multi-UAV wireless network: A survey of reinforcement learning-based approaches[J]. IEEE Communications Surveys & Tutorials, 2023, 25(4): 3038-3067.

作者简介

张祥银 (1986−), 男, 副教授, 博士, 博士生导师, 主要研究方向为无人系统自主控制与决策优化、群智能优化、机器人控制, E-mail: xy_zhang@bjut.edu.cn;

张曦梁 (2000−), 男, 硕士生, 主要研究方向为多智能体集群, E-mail: 1595266195@qq.com;

史骏驰 (2002−), 男, 硕士生, 主要研究方向为多智能体集群, E-mail: 1446736290@qq.com.