基于非对称冲突区域的多无人机协同路径规划

杨群亭，文章，高伟，刘冰清，杨倩
（中国民航大学空中交通管理学院，天津 300300）

摘要：为了解决低空融合空域多无人机协同路径规划中飞行安全与运行规则难以兼顾的问题，提出了一种基于非对称冲突区域（ACR）与柔性演员-评论家（SAC）算法的多无人机协同路径规划方法。该方法的主要思想是在无人机机体坐标系下构建方向偏置的六边形冲突区域，利用其几何不对称性使不同通行方向产生差异化重叠惩罚，将靠右避让、靠右超越和顺时针环流等运行规则隐式嵌入奖励信号，并采用集中训练-分散执行框架完成策略学习与执行。仿真结果表明，所提方法在三类典型冲突场景下任务成功率均可达 99.83% 以上，平均冲突次数接近于零，并使多无人机形成靠右避让、靠右超越和顺时针环流的有序通行秩序，降低了通行方向的随机性。

关键词：低空融合空域；多无人机；协同路径规划；深度强化学习；非对称冲突区域
中图分类号：TP309
文献标志码：A
doi：10.11959/j.issn.1000

Multi-UAV cooperative path planning based on asymmetric conflict region

Yang Qunting, Wen Zhang, Gao Wei, Liu Bingqing, Yang Qian
College of Air Traffic Control, Civil Aviation University of China, Tianjin 300300, China

Abstract: To solve the problem that flight safety and operating rules were difficult to balance simultaneously in multi-UAV cooperative path planning in low-altitude integrated airspace, a multi-UAV cooperative path planning method integrating the asymmetric conflict region (ACR) and the soft actor-critic (SAC) algorithm was proposed. The main idea of the method was that a direction-biased hexagonal conflict region was constructed in the body frame of each UAV, and its geometric asymmetry was exploited to produce differentiated overlap penalties for different passing directions, by which the right-side avoidance, right-side overtaking and clockwise circulation rules were implicitly embedded into the reward signal. A centralized training with decentralized execution framework was adopted for policy learning and deployment. Simulation results show that, in the three typical conflict scenarios, the proposed method achieves a task success rate above 99.83% with the average number of conflicts approaching zero, and induces an orderly passing pattern of right-side avoidance, right-side overtaking, and clockwise circulation, thereby reducing the randomness of passing directions.

Key words: low-altitude integrated airspace, multi-UAV, cooperative path planning, deep reinforcement learning, asymmetric conflict region

---

收稿日期：XXXX–XX–XX；**修回日期：**XXXX–XX–XX
通信作者：高伟，邮箱 atmgao@126.com
基金项目：国家自然科学基金资助项目（No. 52202404）；西藏自治区重大科技专项基金资助项目（No. XZ202402ZD0004）；中央高校科研基金（No. 3122017064）
Foundation Items: The National Natural Science Foundation of China (No. 52202404), The Major Science and Technology Project of Xizang Autonomous Region(No. XZ202402ZD0004), The Fundamental Research Funds for the Central Universities(No. 3122017064)

0 引言

低空空域中无人机（unmanned aerial vehicle, UAV）的运行密度持续提升，未来低空运行将逐步从隔离运行向融合运行演进[1-3]。融合空域内多架无人机同时执行任务时，无人机种类多、速度差异大，机间交互频次急剧上升，低空环境网络覆盖不连续、链路质量波动，各机往往只能依赖局部观测与有限信息交互进行分散决策，路径规划必须在保障飞行安全的同时，满足必要的规则约束，否则高密度场景下会陷入无序甚至冲突状态。可见，安全间隔与通行秩序已成为低空多机协同规划中并行的刚性需求[4]。

围绕这一需求，现有研究主要沿两条路径展开。一条以保持安全间隔为核心，将冲突简化为基于机间距离的惩罚或势场约束[5-6]，此类信号关于无人机质心对称，不携带方向语义，对通行方向、绕行规则缺乏统一表征，群体行为虽能避免碰撞却难以形成有序的交通流[7]。另一条尝试显式引入运行规则，但多采用规则编码或硬约束的方式[8]，规则以离散判断形式作用于决策，无法提供反映违规程度的连续反馈，规则模块与策略学习过程相互独立，难以根据动态交互自适应调整。从决策架构来看，集中式方法虽可获得全局协调，但其计算与通信开销大，且高度依赖可靠骨干网，难以适应链路质量波动的低空环境；分布式方法扩展性好，但各机仅凭局部观测独立决策，在通信受限时容易产生规则不一致的群体行为。更本质的原因在于，空域运行规则以离散条款形式给出，而学习型路径规划依赖连续的交互反馈优化策略，两者缺少统一的表达方式。因此，如何将离散的通行规则自然地融入连续决策过程，是使多机涌现有序协同行为的关键难点[9]。

针对上述问题，本文构建一种面向多无人机协同路径规划的非对称冲突区域（asymmetric conflict region, ACR），将运行规则由离散条款转化为强化学习过程中的连续几何反馈。与仅依据机间距离施加避碰惩罚的方法不同，ACR 在无人机机体坐标系下引入方向偏置的六边形结构，使不同接近方向和通行方式对应不同的区域重叠状态，从而将靠右避让、靠右超越和顺时针环流等规则表达为可累积的重叠惩罚信号。在此基础上，本文采用柔性演员-评论家（soft actor-critic, SAC）算法完成连续动作空间下的策略优化，并通过集中训练-分散执行框架实现多机策略学习与独立决策，规则一致性在训练阶段内化于共享策略，执行阶段各机仅凭局部观测独立决策，无需集中调度节点、全局信息广播或机间协商交互，即可涌现方向一致的通行秩序。该特性使所提方法能够以较低的信息交互开销降低对集中式调度与可靠骨干链路的依赖，适配低空融合空域通信受限条件下的分散协同运行需求。对头、超越和多机交汇三类典型场景的仿真结果表明，所提方法能够在抑制机间冲突的同时，引导多无人机形成符合通行规则的协同避让行为。

1 相关工作

围绕融合空域下多无人机协同路径规划，现有研究主要分为无人机路径规划方法、多无人机协同决策框架和空域运行规则融入三个方向。

传统路径规划方法主要包括 A* 搜索、人工势场（artificial potential field, APF）以及各类启发式优化算法[10-12]。文献[10]通过改进 A* 启发函数与 APF 引斥力机制，实现了复杂城市低空环境下的混合路径规划；文献[13]基于分段势场方法缓解了固定翼无人机编队规划中的局部极小问题；Phung 等[14]提出球面向量粒子群优化方法，提高了三维路径规划中的安全性与轨迹平滑性。在多目标任务场景下，付澍等[15]将无人机数据收集路径规划建模为定向问题，并通过指针网络进行求解；许云鹏等[16]结合时空依赖图卷积网络与禁忌搜索，实现了应急救援场景下的动态路径优化。深度强化学习（deep reinforcement learning, DRL）通过环境交互优化策略，为无人机自主导航与路径规划提供了新的求解框架[17-21]。Aggarwal 等[3]采用 DRL 求解多无人机数据收集任务中的路径规划问题；文献[22]针对车联网服务迁移问题，提出了一种摒弃显式 Critic 网络的多智能体组相对策略优化方法。Bouhamed 等[23]基于 DDPG 实现了城市环境中的自主无人机导航；Lee 等[24]结合 HER 与 SAC 缓解了奖励稀疏问题；Gong 等[25]从合规约束角度研究了低空空域中的无人机轨迹规划。

为提升多机间的协同决策能力，多智能体强化学习（multi-agent reinforcement learning, MARL）逐渐被应用于多无人机协同路径规划。Brittain 等[6]提出结合注意力机制的深度多智能体强化学习

框架，将集中训练—分散执行机制引入多智能体协同决策；Yu 等[26]提出的 MAPPO 在合作型多智能体任务中表现出良好的稳定性，并被广泛应用于多无人机协同搜索[27]与目标跟踪[28]任务。Chen 等[5]提出的 CADRL 利用值网络对邻机交互进行建模，实现了多智能体场景下的协同避碰；Everett 等[29]进一步提出 GA3C-CADRL，提高了复杂动态环境中的避碰能力；Lin 等[30]将 SAC 推广至部分可观测多智能体路径规划任务；Yue 等[31]采用约束马尔可夫决策过程与 SAC-Lagrangian 算法求解多机安全决策问题。

围绕空域运行约束的融入，部分研究尝试在冲突解脱过程中引入规则约束。Kuchar 等[32]系统综述了冲突探测与解脱领域的早期研究；文献[8]基于模型预测控制构建了遵循 Right-of-Way 规则的实时冲突解脱算法；在低空交通管理方面，Kopardekar 等[33]提出 NASA 的 UTM 概念框架，Rastgoftar 等[34]建立了基于固定走廊的有限状态 UTM 模型，Luo 等[1]提出统一蜂窝原生网络架构 UTICN，为低空多机协同运行提供了基础设施支持。综上，现有多无人机路径规划与协同避碰研究已在安全间隔保持、动态避障和多智能体协同决策方面取得较多进展，但多数方法仍以距离阈值、碰撞惩罚或安全约束刻画冲突风险，难以直接表达运行规则中的方向性通行关系。已有显式规则方法通常依赖规则判定、优先权判断或硬约束修正，规则更多作为独立模块参与决策过程。与上述方法不同，ACR 将方向性运行规则编码为非对称几何区域的重叠差异，使规则约束以连续惩罚信号的形式进入强化学习奖励函数，从而在策略学习过程中引导多无人机形成规则一致的协同行为。

2　非对称冲突区域

2.1　冲突区域建模

运行规则具有明确的方向性约束，而常规距离阈值或对称安全区域主要刻画机间间隔大小，难以区分不同相对方位下的通行关系。因此，需要构造具有方向偏置的冲突区域，使不同接近方向和通行方式对应差异化的区域重叠结果。

区域形状的设计直接决定方向偏置的表达能力。圆形与椭圆关于质心对称，无法表达方向相关的通行差异；四边形虽可实现单一方向上的非对称，但难以同时承载三类典型场景所需的多向偏置。六边形在保持顶点数较少的前提下，可提供三组承载不同方向偏置的边界特征：侧向两边用于区分对 头接近时的左右通过方向，后向两边用于区分同向 超越时的接近侧，中弦偏置用于区分多机交汇时的 绕行方向。针对每一架无人机 i，假设其在平面内为圆形，半径为 R，在其机体局部坐标系下构建一个六边形非对称冲突区域 $A_i^{Conflict}$。该区域以无人机质心为原点，其形状由以下六个顶点定义，如图 1 所示：

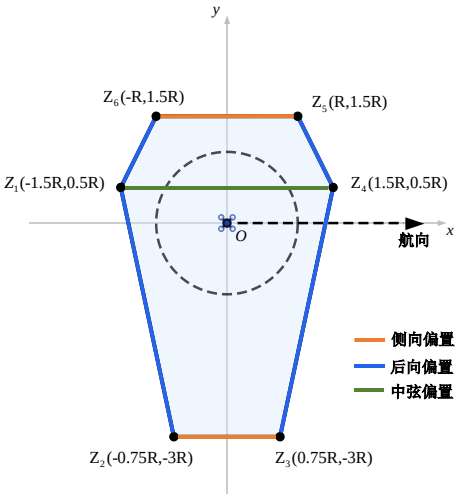


图 1　非对称冲突区域几何结构示意图

设无人机位置为 $p_i=(x_i,y_i)$，航向角为 ψ_i，在世界坐标系中，冲突区域随无人机的位置 p_i 进行平移、随航向角 ψ_i 进行旋转变换。第 k 个顶点在世界坐标系中的位置为:

$$Z_k^{\prime}=R(\psi_i)Z_k+p_i \tag{1}$$

$$R(\psi_i)=\begin{bmatrix}\cos\psi_i & -\sin\psi_i\\ \sin\psi_i & \cos\psi_i\end{bmatrix} \tag{2}$$

其中 $R(\psi_i)$ 为二维旋转矩阵，确保冲突区域始终与无人机的飞行方向保持一致的相对位置关系。

2.2　冲突区域特征分析

ACR 几何特征需同时体现安全分离与方向判别两类要求，核心在于通过不同方向之间的偏置关系改变区域重叠状态。方向偏置较弱时，不同通行方向对应的重叠状态差异有限，难以稳定区分期望通行方向与非期望通行方向；方向偏置较强时，冲突区域占用范围过大，将对已满足正常间隔的交通产生额外惩罚。针对三类典型冲突场景分析其对策

略行为的诱导效果。

（1）对头飞行

对头飞行场景中的方向区分主要由 ACR 在航向中轴线两侧的侧向偏置体现。记合规通行侧的最大延伸为δ_1，违规通行侧的最大延伸为δ_2。两机以横向间距d_{avoid}完成对头穿越时，相邻一侧的延伸之和决定 ACR 重叠结果，δ_1与δ_2需分别满足：

$$R \leqslant \delta_1 \leqslant \frac{d_{avoid}}{2}, \frac{d_{avoid}}{2} \leqslant \delta_2 \leqslant 2d_{avoid} \tag{3}$$

其中的δ_1下界保证 ACR 覆盖无人机机体，上界保证靠右通行不触发重叠，δ_2的下界保证靠左通行触发重叠，上界为正常巡航间距（通常取最小避让距离的 2 倍）。本文取$\delta_1 = 1.5R$、$\delta_2 = 3R$，即Z_5、Z_6的纵坐标为$1.5R$，Z_2、Z_3的纵坐标为$-3R$，如图 2 所示，靠右通行时相邻侧延伸之和$2\delta_1 = 3R$恰好等于$d_{avoid} = 3R$，ACR 处于无重叠的临界状态（safe）；靠左通行时延伸之和$2\delta_2 = 6R$，超出d_{avoid}一倍，沿横向产生约 3R 的重叠深度（invade）。两种通行方向在惩罚信号上呈现明显的左右不对称。

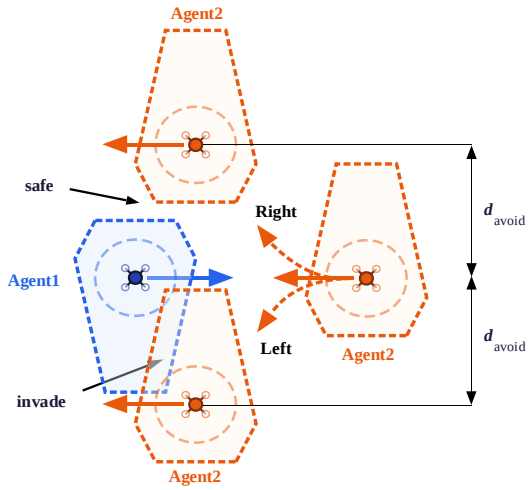


图2　对头飞行场景下的靠右避让机制

（2）同向超越

在同向超越场景中，ACR 的突出部分相对航向中轴线呈现上下非对称特征，顶点□₁与□₄的纵坐标为$0.5R$，位于航向中轴线上方，使冲突区域后部偏左、偏右两侧的突出面积存在差异。这一几何特征使后机从不同侧超越时产生差异化的重叠结果，如图 3 所示。当 Agent2 以 15°偏角从 Agent1 的左侧超越时，其前向区域侵入 Agent1 后方左侧面积较大的突出区域，发生重叠（invade）；以相同偏角从右侧超越时，其前向区域与 Agent1 后方右侧较小的区域相邻，不发生重叠（safe）。在$d_{avoid} = 3R$间距下，左侧超越路径穿越的 ACR 区域宽度约为 1.5R，而右侧超越路径与 ACR 边界相切，重叠宽度近似为 0。该差异在累积惩罚信号引导下，使策略在训练过程中逐渐形成靠右超越的行为偏好。

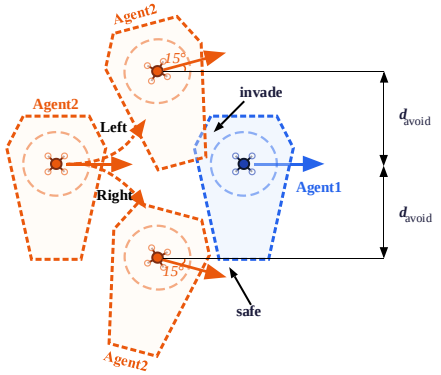


图3　同向飞行场景下的靠右超越机制

（3）多机交汇

在多机交汇场景中，ACR 的最宽边界由顶点$Z_1(-1.5R, 0.5R)$至$Z_4(1.5R, 0.5R)$的连线构成，位于质心上方中弦偏置$h_{中弦偏置} = 0.5R$处，是相邻无人机在密集交互中最先发生接触的冲突边。这一几何特征使顺时针与逆时针两种绕行方向具有不同的安全裕度，如图 4 所示。以该冲突边上各点为起点、以相同半径绘制圆弧，可得到一条恰好不发生区域侵入的极限弧线：顺时针方向绕行的无人机位于该极限弧线下方一个$h_{中弦偏置}$处，可用的调整空间更大；逆时针方向绕行的无人机则位于极限弧线上方$h_{中弦偏置}$处，调整余量更小。在相同绕行半径下，两条绕行轨迹的径向间距差为$2h = R$，顺时针方向具有更大的安全裕度，策略在训练中更易维持无冲突状态，系统在多机交汇场景中自发收敛为顺时针环流模式。

3　基于 SAC 的多无人机自主飞行决策算法

3.1　问题描述与建模

在低空空域中，多架无人机需要在共享空间内同时执行飞行任务，各机从指定起点飞往目标点。飞行过程中，无人机需要自主完成路径规划、动态避障、以及符合空域运行规则的协同决策。本文将

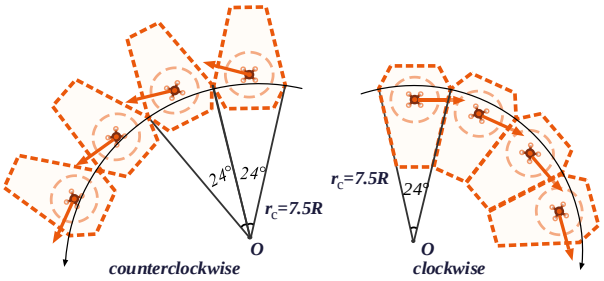


图4　多机交汇场景下的顺时针环流机制

多无人机自主飞行决策问题建模为马尔可夫决策过程（Markov Decision Process, MDP），并采用深度强化学习方法求解。

假设空域中存在 N 架无人机，各机在固定高度层巡航，将问题简化为二维平面运行。在MDP框架下，定义五元组(S,A,P,R,γ)，其中S为状态空间，包含无人机自身的位置、航向、速度以及对周围环境和其他飞行器的局部观测；A为动作空间，由无人机的线速度v和角速度ω构成；$P(s_{t+1}|s_t,a_t)$为状态转移概率，描述无人机在当前状态下执行动作后到达下一状态的动态过程；R为奖励函数，综合评价导航效率、飞行安全与规则遵从程度；$\gamma\in(0,1)$为折扣因子，用于平衡即时奖励与长期回报。由于多无人机交互环境的复杂性，状态转移概率P未知的，需要通过与环境的交互来学习策略。

在多智能体架构层面，本文采用参数共享式集中训练-分散执行（Centralized Training with Decentralized Execution, CTDE）框架。训练阶段N架无人机共享同一套策略网络参数与经验回放缓冲区，各机在每个时间步产生的经验四元组$(s_t^i,a_t^i,r_t^i,s_{t+1}^i)$统一写入公共缓冲区，由同一训练流程采样并更新网络参数；执行阶段，各机以自身状态及对邻机位置、速度与航向的局部观测为输入独立决策。该框架在保证多机策略一致性的同时降低了执行端的通信与调度开销，适用于低空空域中各机自主决策的协同运行场景。

3.2　SAC算法

基于上述MDP建模，本文采用SAC（Soft Actor-Critic）算法求解多无人机自主飞行决策问题，算法整体框架如图5所示。

SAC在Actor-Critic架构基础上引入最大熵正则项，其优化目标为最大化预期"软回报"，目标函数可以表示为：

$$J(\pi)=\sum_{t=0}^{T}\mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t)+\alpha\mathcal{H}(\pi(\cdot|s_t))\right] \quad (4)$$

其中$\mathbb{E}$表示在策略π所诱导的轨迹分布下对奖励与熵之和的期望，$r(s_t,a_t)$表示智能体在状态s_t下执行动作a_t所获得的即时环境奖励，$\mathcal{H}(\pi(\cdot|s_t))=-\log\pi(a_t|s_t)$为策略在状态下的信息熵，温度系数$\alpha$用于平衡奖励项与熵项的相对权重，$\alpha$较大时策略偏向探索，$\alpha$趋近于零时退化为标准强化学习的累积回报最大化形式。

为实现上述最大熵优化目标，SAC将其分解为环境、Critic、Target、Actor与温度系数五个模块协同完成，各模块的实现细节如下。

（1）环境交互与经验回放

在每个决策时步t，Agent依据当前策略π_ϕ对状态s_t采样动作a_t，并输出至环境，环境返回即时奖励r_t和下一状态s_{t+1}，形成四元组经验(s_t,a_t,r_t,s_{t+1})，产生的经验连同策略的对数概率$\log\pi_\phi(s)$一并写入经验缓冲池，训练时从中均匀采样小批量数据，用于同步更新Critic网络和Actor网络。

（2）Critic网络

SAC使用两个独立参数化的Q网络Q_{θ_1}和Q_{θ_2}来估计状态-动作值函数，训练目标为最小化软贝尔曼误差：

$$J_Q(\theta_i)=\mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_{\theta_i}(s_t,a_t)-y_t\right)^2\right] \quad (5)$$

其中$i=1,2$，在计算目标值时取两网络输出的较小值，抑制值函数过高估计，本文奖励函数以导航效率为主，冲突规则惩罚为辅，引导策略在掌握基本导航能力的基础上逐步收敛至符合规范的飞行行为；若Q值被高估，策略将倾向于忽略冲突规则惩罚信号，规则遵从行为难以习得。

（3）Target网络

Target网络与Critic网络结构相同，但其参数通过软更新缓慢跟踪主网络，专门用于生成稳定的目标Q值。

$$y_t=r_t+\gamma\left(\min_{j=1,2}Q_{\bar{\theta}_j}(a_{t+1},s_{t+1})-\alpha\log\pi_\phi(a_{t+1}|s_{t+1})\right) \quad (6)$$

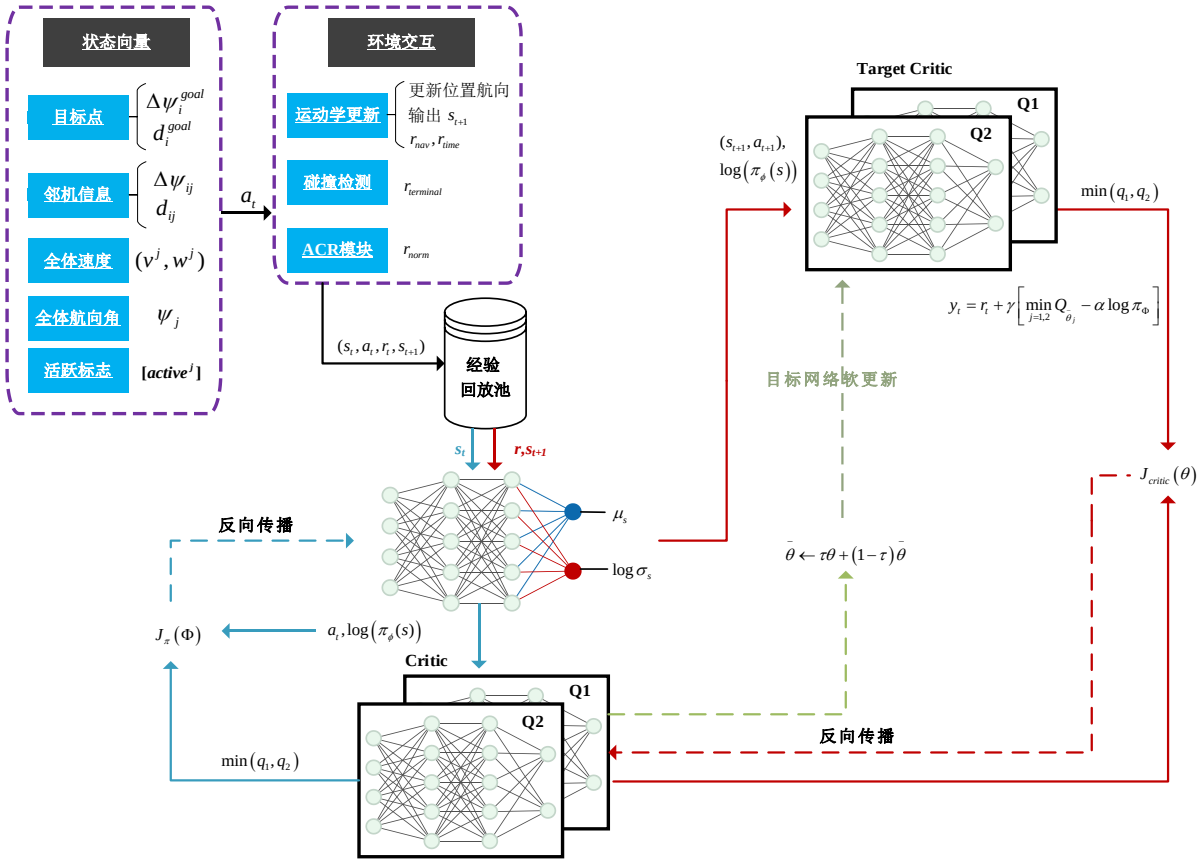


图 5　SAC 算法整体框架

其中 $a_{t+1}\sim\pi_\phi(\cdot|s_{t+1})$，熵正则项 $-\alpha\log\pi_\phi$ 对确定性动作施加隐式惩罚：若策略在某状态下的输出高度集中，$\log\pi_\phi$ 接近 0，熵 $\mathcal{H}$ 较小，目标 Q 值随之降低，从而引导策略保持一定随机性。Target 网络参数按如下软更新规则迭代：

$$\bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta} \tag{7}$$

（4）Actor 网络

Actor 网络将策略 π_ϕ 参数化为条件高斯分布，输出线速度与角速度的均值和方差。通过重参数化技巧，策略梯度可沿采样动作反向传播，优化目标为：

$$J_\pi(\phi) = \mathbb{E}_{s_t\sim\mathcal{D},\epsilon_t\sim\mathcal{N}}\left[\alpha\log\pi_\phi(a_t|s_t) - \min_{j=1,2}Q_{\theta_j}(s_t,a_t)\right] \tag{8}$$

该目标同时驱动策略向 Q 值高和熵值大两个方向优化，二者通过 α 加权平衡。

（5）温度系数 α

α 通过最小化以下损失自动调节：

$$J(\alpha) = \mathbb{E}_{a_t\sim\pi_\phi}\left[-\alpha\log\pi_\phi(a_t|s_t) - \alpha\overline{\mathcal{H}}\right] \tag{9}$$

当策略熵低于目标熵 $\overline{\mathcal{H}}$ 时，损失驱动 α 增大加强探索；当策略熵高于目标熵时，α 减小侧重利用。目标熵设为 $\overline{\mathcal{H}}=-2$，与二维动作空间的维度对应，确保策略在速度与角速度两个维度上均维持适度的随机性。

3.3　状态空间

针对空域中 N 架无人机，本文为每架无人机 i 设计融合结构化向量与局部栅格地图的混合状态空间。结构化向量由目标相对状态、邻机相对状态与群体运动状态三部分组成，分别表示无人机的目标导航、冲突感知与协同避让。其中目标方向偏差角 $\Delta\psi_i^{goal}$ 与目标距离 d_i^{goal}，邻机相对方位角 $\Delta\psi_{ij}$ 与相对距离 d_{ij} 均在机体坐标系下表达，计算方式如下：

$$\Delta\psi_i^{goal} = \arctan 2(g_y - y_i, g_x - x_i) - \psi_i \tag{10}$$

$$d_i^{goal} = \| p_i^{goal} - p_i \|_2 \tag{11}$$

$$\Delta\psi_{ij} = \arctan 2(y_j - y_i, x_j - x_i) - \psi_i \tag{12}$$

$$d_{ij} = \| p_i - p_j \|_2 \tag{13}$$

群体运动状态由空域内全部 N 架无人机的线速度 v_j、角速度 ω_j、绝对航向角 ψ_j 以及活跃状态标志

$[active^j]$拼接而成，刻画群体的整体运动态势，预判潜在冲突。其中，活跃状态标志区分仍在执行任务的无人机与已到达目标或已退出运行的无人机，避免策略对无效目标产生不必要的避让动作；对于非活跃无人机，其对应的速度与航向分量统一填充为零向量。

表1　　状态向量构成

分量	含义	维数
$\Delta\psi_i^{goal},d_i^{goal}$	目标相对状态	2
v_j,ω_j	全体线速度角速度	$2N$
$\Delta\psi_{ij},d_{ij}$	邻机相对状态	$2(N-1)$
$[active^j]$	活跃标志	N
$Gridmap_i$	局部栅格地图	121

局部栅格地图$Gridmap_i$以无人机自身为中心、当前航向为正方向对齐，栅格地图大小为 11 × 11，分辨率为每格 2 米，覆盖边长为 22 米的正方形区域。栅格根据按照内容赋值：自由空间为 0，静态障碍物为 1，动态障碍为 0.5。展平后的栅格地图为 121 维向量，与结构化向量拼接构成完整状态向量s_t^i，各分量构成如表 1 所示。

3.4　动作空间

无人机在二维平面内的运动由线速度v与角速度ω共同决定。在每个控制周期$\Delta t = 0.1s$内，无人机按照如下运动学模型更新位置和航向。

$$v = v_{\min} + \frac{a_1 + 1}{2}\left(v_{\max} - v_{\min}\right) \tag{14}$$

$$\omega = \omega_{\min} + \frac{a_2 + 1}{2}\left(\omega_{\max} - \omega_{\min}\right) \tag{15}$$

取$v \in (0,15)m/s$、$\omega \in (-1,1)rad/s$。线速度下界设为零，允许无人机在严重冲突情况下悬停等待，角速度关于零对称，确保冲突规则的学习过程不受动作空间本身的方向偏置干扰。

3.5　奖励函数

奖励函数决定智能体学到的行为。本文兼顾导航效率、飞行安全与空域冲突规则三类目标，设计了一套多分量加权奖励函数，由导航奖励r_{nav}、时间惩罚r_{time}、规则惩罚r_{norm}与终止奖励r_{terminal}四部分组成：

$$r_t^i = r_{\text{nav}} + r_{\text{time}} + r_{\text{norm}} + r_{\text{terminal}} \tag{16}$$

规则惩罚由无人机间 ACR 的几何重叠状态确定。设C_i^t和C_j^t分别为无人机i与j在时刻t的非对称冲突区域，其顶点根据当前位置与航向角从机体坐标系变换至世界坐标系。两机 ACR 的重叠面积定义为：

$$S_{ij}^t = Area\left(C_i^t \cap C_j^t\right) \tag{17}$$

其中，$Area$表示两个凸六边形相交区域的面积。当重叠面积大于给定容差ε时，认为两机发生 ACR 重叠，指示函数为：

$$I_{\text{overlap}}^t(i,j) = \begin{cases} 1, & S_{ij}^t > \varepsilon \\ 0, & S_{ij}^t \leqslant \varepsilon \end{cases} \tag{18}$$

其中，ε为极小正数，用于避免数值误差导致的边界误判，仅边界相切或重叠面积不超过ε时，不计入 ACR 重叠惩罚。各连续分量的表达式与系数如表 2 所示。

表2　　连续奖励

分量	符号	表达式	系数
导航奖励	r_{nav}	$c_{\text{fwd}}(d_{t-1}^{\text{goal}} - d_t^{\text{goal}})$	$c_{\text{fwd}} = 5$
时间惩罚	r_{time}	c_{time}	$c_{\text{time}} = -2$
规则惩罚	r_{norm}	$c_{\text{norm}}\sum_{j\neq i} I_{overlap}^t(i,j)$	$c_{\text{norm}} = -10$

奖励项系数根据单步运动尺度、规则约束强度与回合终止结果共同确定。导航奖励采用距离差分形式$c_{\text{fwd}}(d_{t-1}^{\text{goal}} - d_t^{\text{goal}})$，用于衡量无人机朝目标方向运动所获得的即时收益。在控制周期$\Delta t = 0.1s$、最大线速度$v_{\max} = 15m/s$的条件下，单步最大距离缩减约为$1.5m$，对应最大导航奖励约为 7.5。时间惩罚取$c_{time} = -2$，其绝对值小于有效前进时的导航收益，用于抑制徘徊和无效绕行，但不改变策略趋向目标的基本优化方向。规则惩罚取$c_{norm} = -10$，略大于单步最大导航奖励，确保无人机在违规通行并触发 ACR 重叠时难以通过缩短航迹获得正向净收益，抑制违规近路策略；同时，该惩罚未被设置得过大，避免规则约束持续主导奖励信号而削弱目标引导作用。

回合一旦触发任一终止条件立即停止上述连续累计，转而给予一次性终止奖励r_{terminal}。终止奖励按事件性质分为成功、碰撞、超时与异常四类，以较大的数值对回合结果作出明确反馈，取值如表 3 所示。

表3　终止奖励

终止事件	类型	奖励值
到达目标	成功	120
碰撞静态障碍	静态碰撞	-40
达到最大步数	超时	-40
碰撞其他无人机	动态碰撞	-15
长时间停滞	异常	-10
偏航累计($>4\pi$)	异常	-10

表4　训练超参数

参数	符号	数值
学习率	lr	1×10^{-4}
批量大小	$batch$	4096
折扣因子	γ	0.99
软更新系数	τ	0.01
目标熵	$\mathcal{H}$	-2

4　实验与结果分析

为验证所提 SAC-ACR 方法的有效性，在对头飞行、同向超越与多机交汇三类典型冲突场景下开展仿真实验，从训练收敛性、冲突解脱能力与规则一致性三个维度评估性能，并与 A*+APF、DRL、CADRL 及无 ACR 约束的 SAC 方法进行对比。所有实验在相同的状态空间、动作空间与网络结构下进行，以 3 个独立随机种子（12、34、42）重复训练，保证结论的可靠性与可复现性。

4.1　实验设置

实验基于 PyTorch 实现，训练在单卡 NVIDIA RTX 4060 GPU 上完成。仿真空域为 300m × 150m 矩形区域，环境以固定步长 $\Delta t = 0.1$s 更新，越界视为静态碰撞。无人机物理参数参照 DJI FlyCart 30 大型载重机型设定，最大线速度为 15 m/s，最大角速度为 1.0 rad/s，碰撞半径为 1.5 m。SAC 网络采用两层各 500 个神经元的全连接结构，相关训练超参数如表 4 所示。

针对低空运行中常见的对头飞行、同向超越与多机交汇三类冲突情形，分别构建对应的仿真场景，如图 6 所示。对头飞行场景对应双向航路、廊道交汇或相对航向任务中的迎面冲突；同向超越场景对应同一航路或相近航迹上不同速度无人机前后接近产生的超越冲突；多机交汇场景对应低空航路节点、任务集结区或进出场交汇区域中的多方向汇聚冲突。对头飞行与同向超越场景各训练 2500 个回合，多机交汇场景训练 1000 个回合。

4.2　收敛性分析

三种场景 SAC-ACR 的训练曲线如图 7 所示，每个场景曲线依次为回合累计奖励 R、ACR 重叠次数 N_c 与任务成功率 P_s，分别反映策略整体性能、规则遵从水平与任务完成情况，每条曲线以实线表示种子间均值、阴影带表示标准差。三类场景曲线呈现一致的三阶段演化规律，训练初期，策略尚未建立有效的状态—动作映射，无人机在起点附近徘徊，R、N_c 与 P_s 均处于低位；中期，策略在距离缩减奖励引导下习得趋向目标的动作，R 稳步上升，但因尚未掌握避让规则，各机航路频繁交汇，N_c 攀升至训练峰值；后期，导航策略逐渐稳定，ACR 区域重叠惩罚成为影响 R 的主要因素，策略沿几何不对称性所指示的方向收敛至符合冲突规则的交互模式，N_c 由峰值衰减并趋近于零，P_s 稳定在 100% 附近，R 同步收敛。三种子曲线阴影带在末期明显收窄，表明结果具有良好的可复现性。

三种场景的训练收敛情况如表 5 所示，取每个种子最后 200 个训练回合的均值作为统计样本，训练指标以均值±标准差的形式给出，训练末期的冲突次数 N_c^f 均接近于零，成功率 P_s^f 稳定在 99.83 以

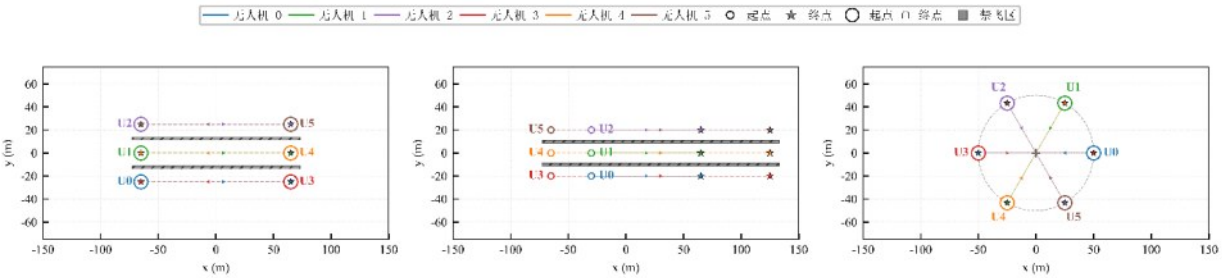


图6　三类典型冲突场景设置

(a) 对头飞行 (b) 同向超越 (c) 多机交汇

上。其中，同向超越场景无人机相对速度较小，峰值冲突数N_c^{peak}显著高于其他场景。多交汇场景中，ACR 的几何不对称性为各机指示了一致的顺时针绕行方向，群体迅速形成统一避让模式，收敛回合最早，策略学习效率最高。

4.3　飞行轨迹与规则一致性分析

三类场景下 SAC-ACR 的收敛飞行轨迹如图 8 所示。对头飞行场景中，两组编队相向接近时均向各自右侧偏转，分占空域右半侧完成穿越；同向超越场景中，快速无人机编队接近慢速编队时向右侧形成横向偏移，从右侧完成超越后回归航路；多机交汇场景中，六架无人机在接近中心区域时提前向同侧偏转，形成顺时针环流后飞向各自目标，整体与 ACR 几何不对称性对应的靠右避让、靠右超越和顺时针环流规则一致。

进一步分析 ACR 对运行规则的诱导效果，引入规则合规率指标，即成功测试回合中满足对应运行规则的回合占比，设置无 ACR、圆形对称冲突区域、本文 ACR 和反向偏置 ACR 四组模型进行对比。其中，反向偏置 ACR 是将本文 ACR 的非对称结构关于航向轴翻转后得到的对照模型。

不同冲突区域形式下的典型轨迹如图 9 所示。无 ACR 和圆形 ACR 缺少明确的方向偏置，收敛后的通行方向具有随机性；本文 ACR 与反向偏置 ACR 分别诱导出相反方向的通行趋势，符合规则的行为在测试回合中出现的概率如表 6 所示。

4.4　多机交汇场景下的算法对比

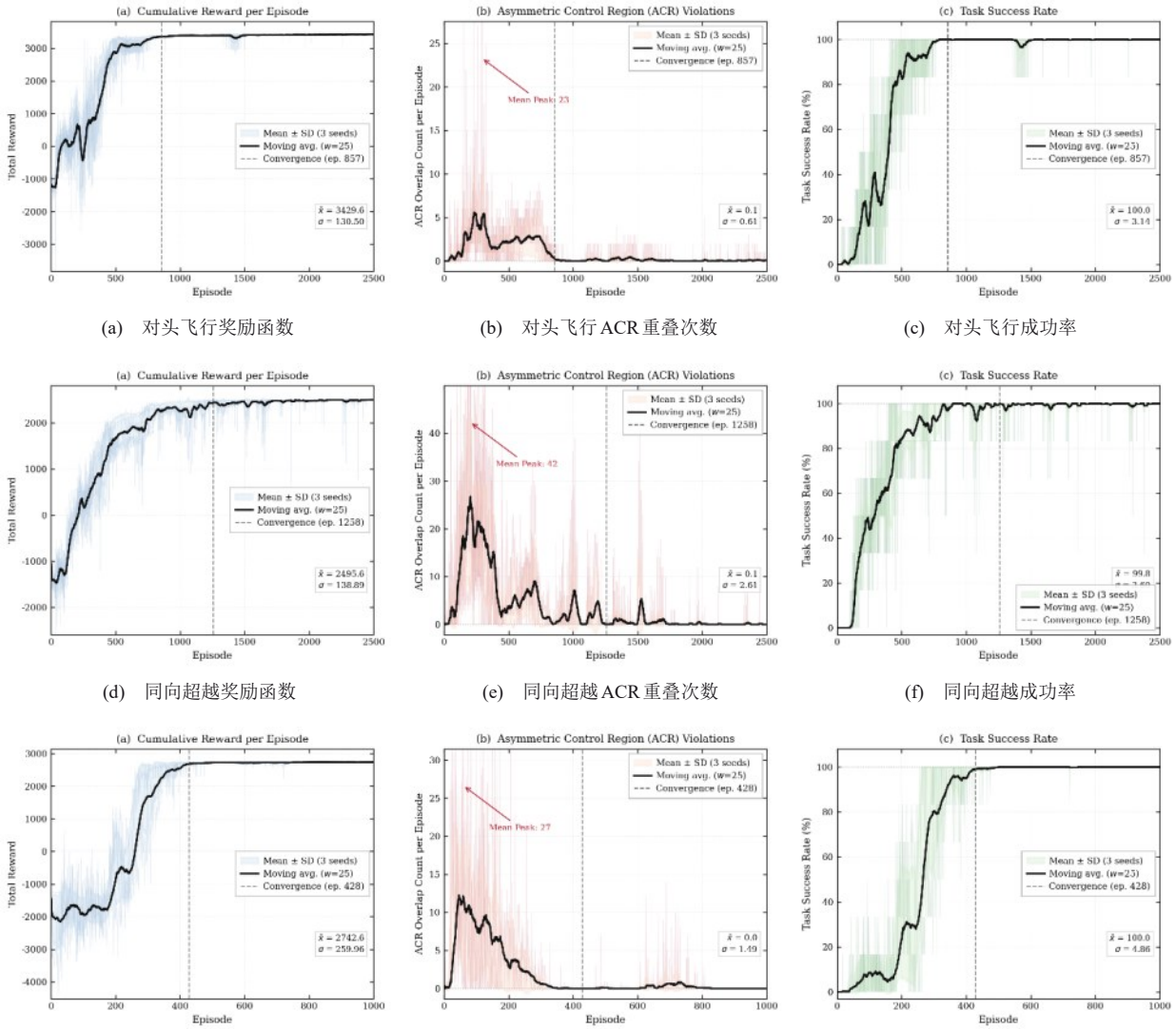
为验证所提 SAC-ACR 方法的有效性，在多机交汇场景下与 A*+APF、DRL、CADRL 以及无 ACR 约束的 SAC 算法进行对比实验。多机交汇场景中多架无人机同时从不同方向接近，局部交互关系随运动过程不断变化，若缺少统一的规则引导，容易出现绕行方向随机、局部冲突传递和群体行为无序等问题，该场景能够集中体现 ACR 约束与 SAC 策略学习结合后的群体协同效果。其中，A*+APF 为传统规划方法，DRL 与 CADRL 为基于策略梯度的多智能体强化学习方法，SAC-noACR 在网络结构与超参数上与本文方法完全一致，仅屏蔽 ACR 惩罚项。A*+APF 不依赖学习机制，其余深度强化学习方法在相同环境与相同随机种子条件下训练 1000 回合，训练曲线对比如图 10 所示。

训练过程表明，不同方法在收敛特性上存在明显差异。SAC-ACR 在较少训练回合内即可实现稳定收敛，累计奖励提升较为平稳，整体波动较小；SAC-noACR 在累计奖励上虽能逐步逼近 SAC-ACR 的水平，但训练中期冲突次数明显高于 SAC-ACR，安全性提升滞后。DRL 收敛速度较慢，在训练结束时仍未达到稳定状态；CADRL 在前期表现出一定提升，但中后期存在较明显波动，难以维持稳定的协同避障策略。在冲突控制方面，SAC-ACR 能够在训练过程中逐步抑制冲突并最终趋近于零，而其余方法冲突水平整体较高或在训练后期仍存在反复。在任务成功率上，SAC-ACR 同样表现出更快的提升速度与更稳定的收敛趋势。

为评估各算法训练完成后的实际飞行决策能力，在多机交汇场景下对训练得到的模型进行 30 次独立评估回合测试，每回合采用不同随机初始化条件，统计任务成功率P_s、平均累计奖励R平均冲突次数N_c与顺时针合规率P_n，结果如表 7 所示。A*+APF 按既定路径与势场斥力修正航迹，在多机汇聚时缺乏统一的让行规则而出现互不相让的僵持，30 次评估中无全体成功回合，合规率无法统计。DRL 为面向单机避让的算法，各机仅依据自身观测独立决策，机间不存在群体层面的协同机制，通行方向互不关联，未能形成稳定的环流模式，合规率同样无法统计。CADRL 合规率为 86.7%，但方向选择仍有波动。SAC-noACR 与 SAC-ACR 均实现零冲突与 100% 成功率，避碰层面已无差异。SAC-noACR 缺少方向性约束，策略收敛后的绕行方向由训练随机性决定，不同回合间通行方向不一致，合规率仅为 43.3%，接近随机水平。SAC-ACR 借助 ACR 合规率达到 93.3%。

表 5　　**三类场景训练收敛结果**

场景	N_c^{peak} 峰值回合	N_c^{peak} 冲突次数	收敛回合	R^f	N_c^f	P_s^f
对头飞行	291	23	857	3429.60 ± 1.76	0.077 ± 0.080	100 ± 0.00
同向超越	201	42	1258	2495.63 ± 10.04	0.075 ± 0.117	99.83 ± 0.22
多机交汇	59	37	415	2764.21 ± 6.97	0.008 ± 0.014	100 ± 0.00



(a)　对头飞行奖励函数　　(b)　对头飞行 ACR 重叠次数　　(c)　对头飞行成功率

(d)　同向超越奖励函数　　(e)　同向超越 ACR 重叠次数　　(f)　同向超越成功率

(g)　多机交汇奖励函数　　(h)　多机交汇 ACR 重叠次数　　(i)　多机交汇成功率

图 7　三类典型场景收敛性曲线

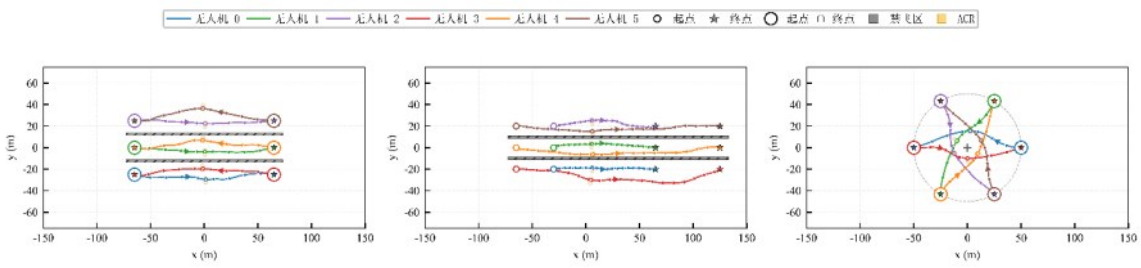


图 8　三类典型场景下的飞行轨迹

(a) 对头飞行 (b) 同向超越 (c) 多机交汇

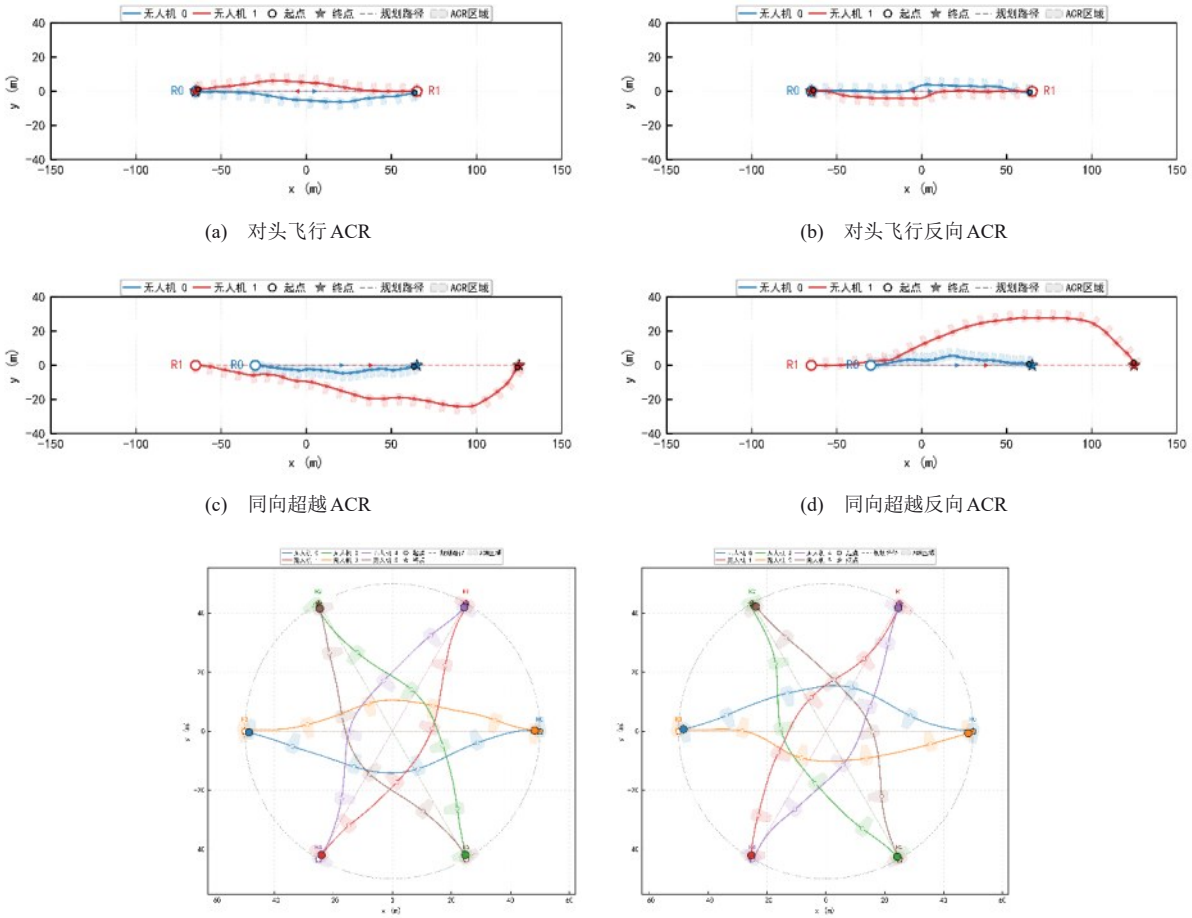
5　结束语

本文针对低空融合空域中多无人机协同运行过程中规则约束与自主决策难以统一的问题，构建了基于非对称冲突区域的协同路径规划方法，通过方向相关的几何区域重叠关系，将传统以离散条款描述的运行规则转化为强化学习可感知的连续惩罚信号，使智能体在交互过程中逐步形成具有一致性的协同避让行为。仿真实验结果表明，在所设置的典

表6　三种典型场景规则合规率

场景	运行规则	合规率P_n
对头飞行	靠右避让	94.8
同向超越	靠右超越	90.4
多机交汇	顺时针环流	91.7

型冲突场景下，所提方法在保持接近零冲突与较高任务成功率的同时，使群体行为呈现靠右避让、靠右超越和顺时针环流的有序通行秩序，降低了通行方向的随机性。



(a)　对头飞行 ACR　　(b)　对头飞行反向 ACR

(c)　同向超越 ACR　　(d)　同向超越反向 ACR

(e)　多机交汇 ACR　　(f)　多机交汇反向 ACR

图9　不同冲突区域下的规则诱导轨迹

表7　多算法性能对比

算法	冲突N_c	累计奖励R	成功率P_s	合规率P_n
A*+APF	54.90 ± 14.86	-479.77 ± 459.75	25.56 ± 9.52	—
DRL	0.03 ± 0.18	1603.33 ± 519.69	93.89 ± 10.25	—
CADRL	0.77 ± 1.38	2586.96 ± 99.35	98.33 ± 5.09	86.7
SAC-noACR	0.0 ± 0.00	2760.33 ± 6.81	100 ± 0.00	36.7
SAC-ACR	0.0 ± 0.00	2760.13 ± 6.57	100 ± 0.00	93.3

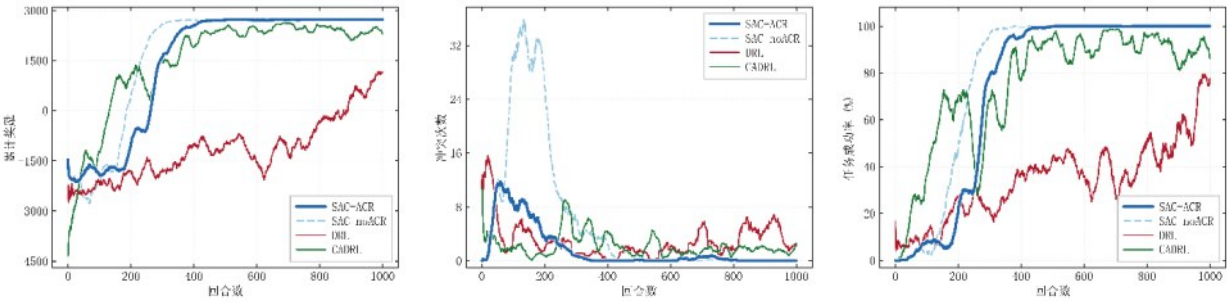


图 10　多算法训练曲线对比

(a) 回合累计奖励 (b) 冲突区域冲突次数 (c) 任务成功率

参考文献：

[1] Luo G, Li J, Zhang Q, et al. Toward low-altitude airspace management and UAV operations: Requirements, architecture and enabling technologies[J]. IEEE Wireless Communications, 2025.

[2] Guan X, Shi H, Xu D, et al. The exploration and practice of low-altitude airspace flight service and traffic management in China[J]. Green Energy and Intelligent Transportation, 2024, 3(2): 100149.

[3] Aggarwal S, Kumar N. Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges[J]. Computer communications, 2020, 149: 270-299.

[4] FEDERAL AVIATION ADMINISTRATION, DEPARTMENT OF TRANSPORTATION. Title 14 of the Code of Federal Regulations, Part 91: General operating and flight rules, §91.113 Right-of-way rules: Except water operations[S]. Washington, DC: U.S. Government Publishing Office, 2025.

[5] Chen Y F, Liu M, Everett M, et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]// 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017: 285-292.

[6] Brittain M, Yang X, Wei P. A deep multi-agent reinforcement learning approach to autonomous separation assurance[J]. arXiv preprint arXiv: 2003.08353, 2020.

[7] Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning[C]//International conference on autonomous agents and multiagent systems. Cham: Springer International Publishing, 2017: 66-83.

[8] Mukherjee A, Mondal S, Tsourdos A. Autonomous Detect and Avoid algorithm respecting airborne Right of Way rules[C]//AIAA SCITECH 2024 Forum. 2024: 1080.

[9] 尹公主,张宏莉,田逸艺,等. 网络群体认知计算建模研究综述[J]. 通信学报,2026,47(3):233-255.
YIN G Z, ZHANG H L, TIAN Y Y, et al. Survey on computational modeling of network group cognition[J]. Journal on Communications, 2026, 47(3): 233-255.

[10] Gao W, Li L, Pang D. Urban low-altitude UAV path planning by fusing an enhanced A* algorithm with an adaptive artificial potential field method[J]. Scientific Reports, 2026.

[11] Zhao J, Zhang Y, Ning L, et al. Adaptive Path Planning of UAV Based on A* Algorithm and Artificial Potential Field Method[J]. Drones (2504-446X), 2026, 10(2): 93.

[12] Duong T T N, Bui D N, Phung M D. Navigation variable-based multi-objective particle swarm optimization for UAV path planning with kinematic constraints[J]. Neural Computing and Applications, 2025, 37 (7): 5683-5697.

[13] Fang Y, Yao Y, Zhu F, et al. Piecewise-potential-field-based path planning method for fixed-wing UAV formation[J]. Scientific Reports, 2023, 13(1): 2234.

[14] Phung M D, Ha Q P. Safety-enhanced UAV path planning with spherical vector-based particle swarm optimization[J]. Applied Soft Computing, 2021, 107: 107376.

[15] 付澍, 杨祥月, 张海君, 等. 物联网数据收集中无人机路径智能规划[J]. 通信学报, 2024, 42(2): 124-133.
FU S, YANG X Y, ZHANG H J, et al. UAV path intelligent planning in IoT data collection[J]. Journal on Communications, 2021, 42(2): 124-133.

[16] 许云鹏,谢雅琪,于然,等. 感-通-物多目标融合应急无人机路径规划方法[J]. 通信学报,2024,45(4) :1-12.
XU Y P, XIE Y Q, YU R, et al. Integrated perception-communication-logistics multi-objective oriented path planning for emergency UAVs [J]. Journal on Communications, 2024, 45(4): 1-12.

[17] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[18] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv 2015[J]. arXiv preprint arXiv:1509.02971, 2015.

[19] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[20] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]// International Conference on Machine Learning. Pmlr, 2018: 1861-1870.

[21] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[J]. arXiv preprint arXiv:1812.05905, 2018.

[22] 芮兰兰,邓淑予,陈子轩,等. 基于多智能体深度强化学习的智能网联汽车服务迁移优化方法[J]. 通信学报,2026,47(1) :141-155.
RUI L L, DENG S Y, CHEN Z X, et al. Service migration optimization method for intelligent connected vehicles based on multi-agent deep reinforcement learning[J]. Journal on Communications, 2026, 47(1): 141-155.

[23] Bouhamed O, Ghazzai H, Besbes H, et al. Autonomous UAV navigation: A DDPG-based deep reinforcement learning approach[C]//2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020: 1-5.

[24] Lee M H, Moon J. Deep reinforcement learning-based model-free path planning and collision avoidance for UAVs: A soft actor–critic with hindsight experience replay approach[J]. ICT Express, 2023, 9(3): 403-408.

[25] Gong Y, Fan J, Zhang R, et al. Safe and economical UAV trajectory planning in low-altitude airspace: a hybrid DRL-LLM approach with compliance awareness[J]. arXiv preprint arXiv:2506.08532, 2025.

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

[27] Wei D, Zhang L, Liu Q, et al. UAV swarm cooperative dynamic target search: a MAPPO-based discrete optimal control method[J]. Drones, 2024, 8(6): 214.

[28] Yue L, Yang R, Zuo J, et al. Factored multi-agent soft actor-critic for cooperative multi-target tracking of UAV swarms[J]. Drones, 2023, 7 (3): 150.

[29] Everett M, Chen Y F, How J P. Motion planning among dynamic, decision-making agents with deep reinforcement learning[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 3052-3059.

[30] Lin Q, 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.

[31] Yue L, Yang R, Zhang Y, et al. Research on reinforcement learning-based safe decision-making methodology for multiple unmanned aerial vehicles[J]. Frontiers in Neurorobotics, 2023, 16: 1105480.

[32] Kuchar J K, Yang L C. A review of conflict detection and resolution modeling methods[J]. IEEE Transactions on intelligent transportation systems, 2000, 1(4): 179-189.

[33] Kopardekar P H. Unmanned aerial system (UAS) traffic management (UTM): Enabling low-altitude airspace and UAS operations[R]. National Aeronautics and Space Administration, 2014.

[34] Rastgoftar H, Emadi H, Atkins E M. A finite-state fixed-corridor model for UAS traffic management[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 25(3): 2322-2330.

杨群亭（1983− ），男，博士，中国民航大学高级工程师，硕士生导师，主要研究方向为低空运行、空域规划、安全评估。

文章（2002− ），男，黑龙江齐齐哈尔人，中国民航大学硕士生，主要研究方向为无人机路径规划、多智能体强化学习。

高伟（1971− ），男，天津人，硕士，中国民航大学副教授，主要研究方向为智慧交通、容量评估、低空交通管理。

刘冰清（1997− ），女，山东泰安人，中国民航大学硕士生，中国市政工程华北设计研究总院工程师，主要研究方向为冲突探测与解脱、协同控制。

杨倩（2004− ），女，河南南阳人，中国民航大学硕士生，主要研究方向为机场运行优化、运行规则设计。