工程科学学报
Chinese Journal of Engineering

基于深度强化学习的无人机集群数字孪生编队避障

张宇宸 段海滨 魏晨

Digital twin-based obstacle avoidance method for unmanned aerial vehicle formation control using deep reinforcement learning

ZHANG Yuchen, DUAN Haibin, WEI Chen

引用本文:
张宇宸, 段海滨, 魏晨. 基于深度强化学习的无人机集群数字孪生编队避障[J]. 北科大：工程科学学报, 2024, 46(7): 1187–1196. doi: 10.13374/j.issn2095–9389.2023.09.28.005
ZHANG Yuchen, DUAN Haibin, WEI Chen. Digital twin–based obstacle avoidance method for unmanned aerial vehicle formation control using deep reinforcement learning[J]. *Chinese Journal of Engineering*, 2024, 46(7): 1187–1196. doi: 10.13374/j.issn2095–9389.2023.09.28.005

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.09.28.005

工程科学学报, 第 46 卷, 第 7 期: 1187−1196, 2024 年 7 月
Chinese Journal of Engineering, Vol. 46, No. 7: 1187−1196, July 2024
https://doi.org/10.13374/j.issn2095-9389.2023.09.28.005; http://cje.ustb.edu.cn

基于深度强化学习的无人机集群数字孪生编队避障

张宇宸, 段海滨✉, 魏　晨

北京航空航天大学飞行器控制一体化技术重点实验室, 北京 100083
✉通信作者, E-mail: hbduan@buaa.edu.cn

摘　要　无人机集群在各个领域中扮演着重要角色, 具有丰富的应用场景. 然而, 将深度强化学习方法应用于自主无人机面临着诸多严峻挑战. 本文基于多智能体深度强化学习, 通过使用局部信息建立单个无人机的状态空间, 并使用多智能体近端策略优化(Multi-agent proximal policy optimization, MAPPO)的在线策略算法来训练策略网络, 从而克服了环境的不确定性和对全局信息的依赖. 同时, 引入了数字孪生的概念, 为资源紧张型算法提供了新思路. 为了解决采样困难和资源紧张的问题, 基于数字孪生技术, 构建了一个用于无人机编队避障策略模型训练的架构. 首先, 构建了多个数字孪生环境, 用于强化学习算法在任务开始之前进行交互采样的预训练, 以使集群具备基本的任务能力. 然后, 使用在真实环境中采集的数据进行补充训练, 使得集群能够更好地完成任务. 对采用这种两阶段训练架构的效果进行了对比, 同时与其他策略算法进行比较, 验证了 MAPPO 的样本效率性能. 最后, 设计了实际飞行验证测试, 验证了从孪生环境中获得的策略模型的实用性和可靠性.

关键词　数字孪生; 深度强化学习; 无人机; 编队控制; 避障
分类号　TG142.71

Digital twin-based obstacle avoidance method for unmanned aerial vehicle formation control using deep reinforcement learning

ZHANG Yuchen, *DUAN Haibin*✉, *WEI Chen*

China Science and Technology on Aircraft Control Laboratory, Beihang University, Beijing 100083, China
✉Corresponding author, E-mail: hbduan@buaa.edu.cn

ABSTRACT　Unmanned aerial vehicle (UAV) swarms have found extensive applications in various fields, playing a crucial role in cluster collaboration. These swarms involve multiple UAVs that work together to achieve common objectives. A key challenging task in swarm operations is collision-free formation control of UAVs. To solve this problem, applying deep reinforcement learning methods has received significant attention, but their application on autonomous UAVs poses challenges, including dependency on global information during training, difficulties in sampling, and excessive resource utilization. To overcome these challenges, in this work, a novel approach based on multi-agent deep reinforcement learning (MARL) is proposed for collision-free formation control of UAV swarms. MARL allows each UAV to interact with a dynamic environment that includes other UAVs, enabling collaborative decision-making and adaptive behavior. We focus on leveraging local information to establish a state space for individual UAVs. To train the policy network, we employ the multi-agent proximal policy optimization (MAPPO) algorithm, allowing robust learning and policy optimization in a multi-agent setting. Also, we address the issues of sampling difficulties and resource constraints by utilizing digital twin technology, serving as a bridge between physical entities and virtual models, which offers a novel approach to the intelligent collaborative control of drone swarms. By establishing models in virtual space, digital twin technology enables the simulation of real-world spaces for pre-training the reinforcement learning algorithm by generating synthetic experiences. We construct multiple digital twin environments to

收稿日期: 2023−09−28
基金项目: 科技创新 2030-“新一代人工智能”重大项目 (2018AAA0100803); 国家自然科学基金资助项目 (T2121003, 91948204, U20B2071)

facilitate interactive sampling and pre-train the swarm with basic task capabilities. Then, we supplement the training using real-world data collected in actual environments, enhancing the ability of the swarm to perform optimally in real-world scenarios. To evaluate the effectiveness of our approach, we compare the performance of the two-stage training architecture with other policy algorithms. To validate the sample efficiency of the on-policy algorithm MAPPO, we conducted a comparative analysis with other policy algorithms, particularly off-policy algorithms. The results reveal the superior sample efficiency and stability of MAPPO in addressing the challenges of collision-free formation control. Finally, we conduct a real-flight validation test to validate the practicality and reliability of the strategy model derived from the digital twin environments. Overall, this work demonstrates the effectiveness of our proposed approach in enabling UAV swarms to navigate complex environments and achieve collision-free formation control.

KEY WORDS digital twin; deep reinforcement learning; unmanned aerial vehicles (UAVs); formation control; collision avoidance

近年来, 无人机作为航空航天技术的一个关键应用领域, 已经在多个领域展现出了巨大的应用潜力. 目前, 无人机不论是在民用领域还是军事应用中, 都不断刷新人们对机器智能和协同作战的认知. 然而, 无人机集群的应用场景多种多样, 这就使得任务执行的复杂性显著增加, 如何进一步提升无人机集群的智能协同能力, 成了研究者们面临的重要课题[1]. 在无人机集群导航领域, 无须外部介入的条件下自主维持集群协调状态, 以实现特定行为或任务目标始终是一个基本且关键的挑战. 这一问题的解决早先是基于控制理论, 例如, Vicsek 等[2] 在其研究中探讨了粒子系统中的自组织运动, 随后出现了基于人工势场法的集群运动控制[3] 以及基于生物群体智能的无人机集群控制方法[4]. 随着计算能力的增长, 强化学习作为一种从交互中学习以达成特定目标的机器学习方法, 在无人机集群的自主决策和路径规划问题上提供了强大的支持[5]. Pu 等[6] 将传统基于模型控制和深度强化学习结合, 使用并行方案解决全向机器人的编队避撞问题. 尽管当前研究已探索在全局环境信息预先获取的条件下进行无人机集群的整体交互训练, 但这种方法往往忽略了真实环境下无人机依据个体局部观测进行决策的需求[7−8], 而真实环境下无人机往往仅依据个体的局部观测进行决策. 同时, 在真实环境中的采样效率过低, 以至训练效率低下[9]. 因此, 允许每个无人机作为个体与环境包括其他无人机动态交互的多智能体强化学习更契合无人集群的协作场景, 所有智能体通过联合动作的回报学习如何协同工作, 以实现群体智能和高效的任务执行, 这就意味着需要很高的并行度和庞大的算力支持. 在计算资源有限的情况下, off-policy 算法如多智能体深度确定性策略梯度[10](Multi-agent deep deterministic policy gradient, MADDPG) 和单调值函数分解方法[11](Monotonic value function factorisation, QMIX) 因其更高的采样效率更适合用来训练多智能体, 而 on-policy 算法如多智能体近端策略优化(Multi-agent proximal policy optimization, MAPPO) 具有明显高的算法运行效率和与之相当的数据样本效率[12].

数字孪生最初是在解决产品生命周期问题时提出的概念, 之后又被用于航空航天制造领域[13]. 如今, 基于数字孪生技术的部分成果得到了实际应用, 如数字孪生车间[14] 和人机协作系统[15]. 同时, 数字孪生技术作为连接物理实体与虚拟模型的桥梁, 为无人机集群的智能协同控制提供了新的思路. 通过在虚拟空间中建立高保真度的模型, 数字孪生技术基于物理模型、历史数据和实时传感器数据, 可以对实际空间进行准确模拟、监测, 并进一步对集群的行为进行预测和分析, 为无人机的路径规划、任务协调等决策行为提供支持. 基于这一理念, 部分研究者展开了无人机集群在数字孪生领域的研究. 例如, Pairet 等[16] 为海上油田建立了一个数字孪生框架, 集成了无人机、无人车和机器狗, 可应用于监视和紧急响应等远程人机协作场景. Lv 等[17] 基于 5G/B5G 通信构建了基于深度学习的无人机数字孪生通用信道模型, 旨在改善无人机通信信号的覆盖问题. Yang 等[18] 结合 Unity、ROS、Matlab 和 SimuIDE 构建了多旋翼无人机(unmanned aerial vehicle, UAV) 数字孪生平台, 实现了 UAV 整个生命周期的跟踪. Miao 等[19] 将数字孪生用于视觉导航的预学习中.

本文以无人机集群为研究对象, 基于数字孪生技术和多智能体深度强化学习结合的天然优势[20−21], 提出了一种无人机集群编队避撞决策模型的分阶段训练架构. 该方法通过任务前训练基础决策模型保障任务的完成, 并于任务执行时利用真实环境采样, 周期性获取补充样本以进一步训练优化模型.

1 问题建模和系统架构

1.1 问题建模

(1) 状态建模.

一般而言, 为追求高精度的路径规划, 完成编队和避障十分依赖全局信息, 现有的技术研究也往往很依赖包含任务场景内各友方单位、敌方单位的具体状态以及环境态势信息而进行决策, 这使得优化成本很高, 同时在环境恶劣的条件下难以确保优化效果[22], 而拓扑结构所体现的相对关系足以完成编队和避障任务, 且这些信息通常也便于获取.

考虑一架无人机可以通过传感器获取周围的单位状态. 如图 1 所示, 红色标记代表障碍物, 蓝色标记代表目标, 虚线表示无人机的有效探测范围. 无人机在任一时刻可以获取预设的目标位置, 以及探测范围内的邻机与障碍物所在的方向和距离.

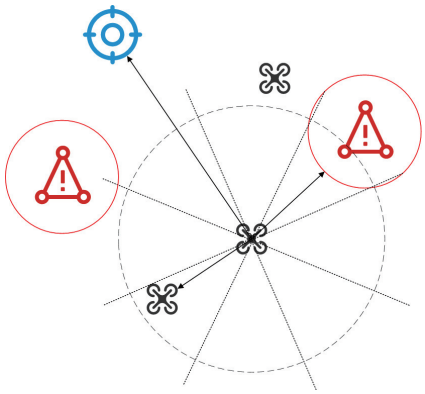

图 1 状态定义

Fig.1 State definition

无人机在给定其目标位置的前提下, 其状态向量$\boldsymbol{s}$定义为:

$$\begin{cases}\boldsymbol{s}=[\boldsymbol{g},\boldsymbol{n},\boldsymbol{o}]\\ \boldsymbol{g}=[g_0,g_1,\cdots,g_m]\\ \boldsymbol{n}=[n_0,n_1,\cdots,n_m]\\ \boldsymbol{o}=[o_0,o_1,\cdots,o_m]\end{cases}\tag{1}$$

其中, $\boldsymbol{g}$表示无人机在周围方向上的目标状态信息, $g_i(i=0,1,\cdots,m)$表示在 m 个方向上, 目标距离的倒数构成的向量, 若该方向上没有探测到目标, 则对应$g_i=0$. 同理$\boldsymbol{n}$和$\boldsymbol{o}$表示邻居无人机和障碍物的状态信息, 满足

$$\begin{cases}g_i=\dfrac{1}{d_{\boldsymbol{g}}+\xi}\\ n_i=\dfrac{1}{d_{\boldsymbol{n}}+\xi},\quad i=1,2,\cdots m\\ o_i=\dfrac{1}{d_{\boldsymbol{o}}+\xi}\end{cases}\tag{2}$$

其中, $d_{\boldsymbol{g}}$表示无人机当前位置到目标位置的距离, $d_{\boldsymbol{n}}$和$d_{\boldsymbol{o}}$则分别表示到邻居无人机以及障碍物物理边界的距离, 该距离通常可通过无人机搭载的传感器直接获取. 在本文的讨论范围中, 假定目标是一个事先已知的明确位置, 因而$\boldsymbol{g}$向量始终有一个元素不为零, 表示目标的状态, 而$\boldsymbol{n}$和$\boldsymbol{o}$则取决于每个时刻无人机的观测, 这样设计的目的是适用于障碍物和邻居无人机数量不确定或动态变化的场景. ξ是一个较小的正数常量, 防止分母为零的情况. 这种状态向量设计方法避免了任务过程中由于各时刻观测状态不同导致的状态向量长度不定问题.

考虑到实际无人机的控制律, 任一架无人机个体的动作空间定义如下

$$[h,v]\in\mathcal{A}_i\tag{3}$$

其中, h和v分别为无人机在水平方向的方向和速度, $\mathcal{A}_i$表示无人机 i 的动作策略集合, h的取值范围为水平面均匀划分的 m 个方向, v则为离散的速度值. 这是因为实际无人机的飞行控制律是基于速度环的串级比例微分控制器, 本文使用的实验平台经过期望的速度及其方向输入控制电机输出.

(2) 奖励函数.

首先定义编队奖励. 考虑一个无人集群编队, 包含 N 个无人机, 其拓扑关系由无向图$\mathcal{G}=(\mathcal{V},\mathcal{E})$表示, 其中$\mathcal{V}=(1,2,\cdots N)$表示点集, $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$为边集. 设无人机的水平位置向量$p_i=[x_i,y_i]\in\mathbf{R}^2$, 边$e_{ij}\in\mathcal{E}$表示无人机 i 和无人机 j 之间的拓扑关系, $e_{ij}=1$表示二者有拓扑连接, 为 0 则无连接. 每条边与一个非负权重值关联, 定义e_{ij}的权重w_{ij}如下

$$w_{ij}=\left\|p_i-p_j\right\|^2,\ (i,j)\in\mathcal{E}\tag{4}$$

其中, 权重w_{ij}构成了邻接矩阵$\boldsymbol{A}\in\mathbf{R}^{N\times N}$, 度矩阵$\boldsymbol{D}\in\mathbf{R}^{N\times N}$是一个对角阵, 对角线元素表示与对应无人机有拓扑连接的邻居权重之和. 则拉普拉斯矩阵($\boldsymbol{L}$)及其归一化拉普拉斯矩阵($\hat{\boldsymbol{L}}$)定义如下

$$\begin{cases}\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}\\ \hat{\boldsymbol{L}}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\end{cases}\tag{5}$$

其中, $\boldsymbol{I}\in\mathbf{R}^{N\times N}$为单位矩阵. 可见, 归一化拉普拉斯矩阵包含了一个队形拓扑的结构信息, 由此定义编队奖励函数r_{f}为

$$r_{\mathrm{f}}=\left\|\hat{\boldsymbol{L}}-\hat{\boldsymbol{L}}_{\mathrm{des}}\right\|_{\mathrm{F}}^2=\mathrm{tr}[(\hat{\boldsymbol{L}}-\hat{\boldsymbol{L}}_{\mathrm{des}})^{\mathrm{T}}(\hat{\boldsymbol{L}}-\hat{\boldsymbol{L}}_{\mathrm{des}})]\tag{6}$$

其中, $\mathrm{tr}[\cdot]$表示矩阵的迹, $\hat{\boldsymbol{L}}_{\mathrm{des}}$表示理想构型的归一化拉普拉斯矩阵. 由于$r_{\mathrm{f}}$是无人机之间绝对距离加权信息得到的, 故该指标不受编队的平移和旋

转影响. 值得注意的是, 由于r_f经过了归一化, 其仅仅反映编队在几何形状上是否满足要求, 但丢失了编队队形的缩放信息, 即编队内无人机间的紧凑和稀疏[23], 故引入规模奖励函数r_s, 定义如下

$$r_s = -\sum_{i=1}^{N}\sum_{j=1}^{\Phi}(D_{des} - \|p_i - p_j\|)^2 \tag{7}$$

其中, Φ为无人机 i 的邻居无人机, D_{des}为无人机间的期望间距, 它由两者间的理想间距和一个模拟误差的噪声构成.

任务的目标是到达一个既定的目标点$p_{tar} = [x_{tar}, y_{tar}]$, 设置目标奖励函数为

$$r_a = D_{t-1} - D_t \tag{8}$$

其中, D_t表示该时刻无人机距离目标的距离, 同样包括理想距离和噪声, D_{t-1}表示上一时刻无人机与目标的距离. r_a为正代表无人机正靠近目标, 反之则在远离目标.

对于避障, 考虑无人机可以通过传感器和机间信息传递获知周围障碍物碰撞表面以及邻机和自己的距离, 则碰撞奖励函数定义如下

$$r_o = \begin{cases} -1, & D_j < D_{avoid} \\ 0, & D_j \geqslant D_{avoid} \end{cases} \tag{9}$$

其中, D_j表示该时刻无人机 j 距探测到的障碍物碰撞表的实际距离, D_{avoid}为无人机到它们的理想保持距离. 由于本文采用多智能体强化学习, 个体的决策是相互独立的, 当某一个体发生碰撞后, 该个体会停止与环境交互, 但不会终止整个训练. 此时集群的各项奖励值会受到影响, 但其他个体的学习仍在继续, 这能保证状态正常的个体不会因为其他邻机出现意外情况而失去继续执行任务的能力. 最终无人机个体的奖励函数为各个子奖励函数的加权和, 表示如下:

$$R_i = w_1 r_f + w_2 r_s + w_3 r_a + w_4 r_o \tag{10}$$

1.2　系统架构

为了更有效地训练决策模型, 避免无人机完全依赖与现实交互以获得训练样本导致训练效率低下, 并希望模型每一次具体任务中针对性地获得更优的决策结果, 提出了一种基于深度强化学习的数字孪生分段训练架构. 如图 2 所示, 整个训练阶段分为 3 个部分:

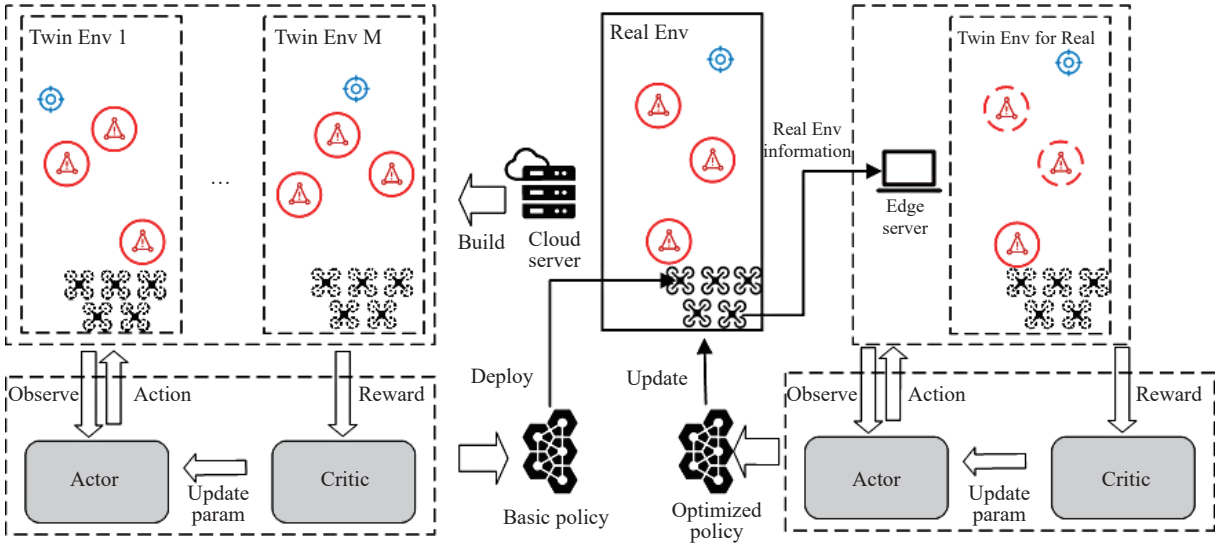

图 2　基于数字孪生的两阶段训练框架

Fig.2　Training frame with the two-phase digital twin

(1)现实环境中任务必要元素的物理实体. 物理实体整体分为两类, 即执行任务的无人机和与无人机产生交互的环境要素. 本文以小型四旋翼无人机作为研究对象, 故无人机机载计算资源有限, 很难胜任基于深度学习的网络训练任务, 仅用于运行决策模型, 处理传感器信息, 并通过通信模块与孪生系统交互以传递环境交互样本和更新模型.

(2)数字模型是现实环境中必要物理元素的代理模型, 是孪生环境的基本构成. 孪生平台服务器通过 4G/5G 手段与无人机进行远程通信, 获取实时的环境信息, 通过建模仿真构建任务场景的孪生模型.

(3)用于训练决策模型的深度强化学习网络. 训练网络获取孪生环境中的模型状态, 抽象为训练样本, 并进行集中式的训练, 这一过程在远端完成, 因此可以不受机载计算资源的限制, 其目的是

训练出每架无人机的分布式决策模型, 使无人机可以独立完成决策, 无须进行集中分配.

值得注意的是, 一般的深度强化学习架构受限于训练速度, 难以应用于实时任务. 然而本文的架构利用数字孪生技术, 事先构建了一个现实环境的孪生模型, 用于在任务开始前对典型任务场景训练基础决策模型. 一些研究已经实现了这一阶段所描述的效果, 事实上这些模型通常可以在类似于训练环境的条件下完成任务, 但当任务环境较复杂和特殊时, 往往无法以较优的策略完成任务甚至无法完成任务. 故本架构将被用于训练两阶段的模型, 第一阶段为任务开始前, 基于典型任务场景和历史任务数据库构成基础训练集, 使无人机具备基本完成任务的能力. 任务开始后, 通过现实世界获得新数据, 进一步训练并将更新后的模型远程部署在无人机上, 以针对当前任务环境进行更优的决策.

2 多无人机训练网络

传统的深度强化学习为单智能体交互, 在无人机群领域往往将整个集群视为一个智能体, 每架无人机的状态空间为整个集群状态空间的一个维度, 这使得问题建模在一个高维空间中, 使问题变得较为复杂, 提高了训练的难度和资源消耗. 多智能体近端策略优化(Multi-agent proximal policy optimization, MAPPO)算法是近端策略优化(Proximal policy optimization, PPO)算法在多智能体场景下的改进[24]. 针对本文研究的无人机集群, 在MAPPO中每个无人机被视为一个智能体, 通过上一节的建模, 可以将无人机集群编队过程视为一个马尔科夫决策过程, 可以定义为$\boldsymbol{M}=(\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R})$. 其中$\boldsymbol{S}$、$\boldsymbol{A}$、$\boldsymbol{P}$和$\boldsymbol{R}$分别为状态空间, 动作空间, 状态转移概率和奖励函数. 无人机作为分布式智能体基于局部观测和一个共享策略生成一个动作, 与环境交互并生成奖励. 它的策略网络只使用局部的状态信息, 而评价网络则使用全局状态信息. 具体而言, 每个智能体接收一个局部的观察, 并输出一个动作概率, 所有的智能体都采用一个策略网络. 每一个智能体的评价网络接收所有智能体的观测, 并输出一个价值, 这个价值用于策略网络的更新.

无人机i在时刻t的折扣奖励U_t定义如下:

$$U_t=R_t+\gamma R_{t+1}+\gamma^2R_{t+2}+\cdots+\gamma^kR_{t+k},k\to\infty \tag{11}$$

其中, γ为折扣系数, R_t为从状态s_{t-1}转移到状态s_t获得的奖励, 动作价值函数Q_π和状态价值函数V_π为

$$Q_\pi(s_t,a_t)=E[U_t|s_t,a_t] \tag{12}$$

$$V_\pi(s_t)=E[U_t|s_t]=\sum_a\pi(a|s_t)Q_\pi(s_t,a) \tag{13}$$

其中, a表示基于策略函数π产生的所有可能动作值, 用优势函数($A_\pi(s_t,a_t)$)表示在一个状态s下选择动作a的倾向, 若该动作的回报高于平均值, 则其优势函数为正, 否则为负.

$$A_\pi(s_t,a_t)=Q_\pi(s_t,a_t)-V_\pi(s_t) \tag{14}$$

对于多智能体任务, 由于折扣奖励无法直接得到, 因此求最大折扣奖励的问题转化为确定最优分散策略$\pi\in\Pi$下的折扣回报的期望, 即

$$\pi^*=\arg\max_\pi E_{s_0\sim\rho_0(s_0)}(V_\pi(s_0)) \tag{15}$$

其中, $V_\pi(s_0)$是初始状态的价值函数, $s_0\sim\rho_0(s_0)$表示初始状态的随机分布, E表示求期望. 由于MAPPO是在PPO基础上变化得到的, 因此, 类似于PPO, 它通过对代理目标函数执行梯度上升来重复更新策略参数θ_i, 策略网络的目标函数$L(\theta_i)$定义如下:

$$L(\theta_i)=E_{\tau\sim\pi_i}\left\{\min[r_t(\theta_i)\frac{\pi_i(a_t|s_t;\theta_i)}{\pi_i(a_t|s_t;\theta_i^{\text{old}})},\right.$$
$$\left.\text{clip}(r_t(\theta_i)),1-\sigma,1+\sigma]\right\} \tag{16}$$

其中, τ表示无人机的轨迹, $r_t(\theta_i)$表示新策略$\pi_i(a_t|s_t;\theta_i)$和旧策略$\pi_i(a_t|s_t;\theta_i^{old})$的似然比, σ是一个超参数, 它控制了更新的幅度.

MAPPO算法的流程如下:

Algorithm 1: MAPPO 算法

input: Each UAV's initial state $\boldsymbol{s}_t$, initial observation $\boldsymbol{o}_t$

output: Actor RNN π, Critic RNN V

initialize θ, the parameters for policy π, and ϕ, the parameters for critic V, using orthogonal initialization

set learning rate α

while step < $\text{step}_{\max}$ do

  set data buffer $D=\{\}$

  while batch ⩽ batch_size do

    τ = []empty list

    for $t=1,\cdots,T$ do

    for all agents i do

      $p_t^{(i)}=\pi\left(o_t^{(i)};\theta\right)$

      $a_t{}^{(a)}\sim p_t{}^{(a)}$

    end for

    execute actionsa_t, observe r_t, s_{t+1}, o_{t+1}

$\tau+=[s_t,o_t,a_t,r_t,s_{t+1},o_{t+1}]$;

end for

compute advantage estimate $\hat{A}$ *via* GAE on τ

compute reward-to-go $\hat{R}$ and normalize

end while

for mini-batch $k=1, \cdots, K$ do

$b \leftarrow$ random mini-batch from experience with all agent data

for each data chunk c in the mini-batch b do

update RNN hidden states for π and V from the first hidden state in the data chunk

end for

end for

Adam update θ on $L(\theta)$ with data b

Adam update ϕ on $L(\phi)$ with data b

end while

3　测试验证

3.1　预训练模型测试

考虑 3 架无人机进行三角编队飞行, 期望间隔为 5 m, 从一定区域内起飞, 以编队形式飞抵目标上方. 场地中的障碍物设定为一个无限高的圆柱形区域, 障碍物同时也可以理解为一个危险区域, 阻止无人机进入其上方一定范围的空域. 由式(4)、(5)可得期望编队的归一化拉普拉斯矩阵为:

$$\hat{\boldsymbol{L}}_{\mathrm{des}}=\begin{bmatrix}1 & -0.5 & -0.5\\ -0.5 & 1 & -0.5\\ -0.5 & -0.5 & 1\end{bmatrix} \tag{17}$$

首先根据典型场景, 随机生成若干组无人机和障碍物的初始状态信息, 本测试中目标位置设置为 (25, 120), 无人机初始状态设置如表 1. 共生成了二十组不同的初始状态用于深度强化学习训练基本的决策模型, 表 2 列举出其中的三组. 初始位置为训练网络的主要参数由表 3 列出.

在上述超参数设置下, 总回合数在 500 左右可以达到较好的收敛效果, 过大的训练回合会导致过拟合和不必要的资源浪费, 反之有可能深度强化学习训练不充分. 图 3 显示了在事先使用构造的训练用孪生模型进行模型训练过程中, 奖励函数值的变化过程, 对各无人机每个回合内的各项奖励取了平均值, 并与典型的多智能体深度强化学习方法[25]进行了对比, 可见, MAPPO 表现出更高的效率和更稳定的性能. 图 4 显示了单个无人机每个回合加权后的总奖励曲线.

3.2　实物验证

测试场地设置为一个户外 60 m×150 m 的缩比飞行试验场. 考虑到安全因素, 为了确保能在无人机不受损伤的状态下获取碰撞惩罚样本, 使用一个发信装置模拟障碍物实体, 发信装置设于地面, 以其为圆心的一定水平半径内视为碰撞区域. 发信装置和无人机集群通过无线数传进行组网, 当

表 1　无人机初始信息

Table 1　Initial status of UAVs

Case	UAV ID	*X*/m	*Y*/m	*Z*/m
1	0	52.6	50.4	0
	1	78.8	12.5	0
	2	22.7	38.0	0
2	0	27.4	43.5	0
	1	65.5	28.6	0
	2	78.3	33.1	0
3	0	25.4	16.5	0
	1	4.8	44.2	0
	2	19.2	30.5	0

表 2　障碍物初始信息

Table 2　Initial status of obstacles

Case	ID	*X*/m	*Y*/m	Radius/m
1	0	40	40	6
	1	26	60	4
	2	17	86.7	5
	3	17	100	3
	4	7	67	3
2	0	34	81	3
	1	6	70	8
	2	18	78	4
	3	37	52	5
	4	38	99	7
3	0	41	70	6
	1	46	52	7
	2	38	61	6
	3	4	37	5
	4	39	102	3

表 3　MAPPO 训练参数

Table 3　Parameters of the MAPPO training

Param	Value	Param	Value
Steps	50000	Max grad norm	10.0
Episode Length	250	Gamma	0.99
Mini batch	10	Gae lamda	0.95
Entropy coef	0.05	Clip param	0.2

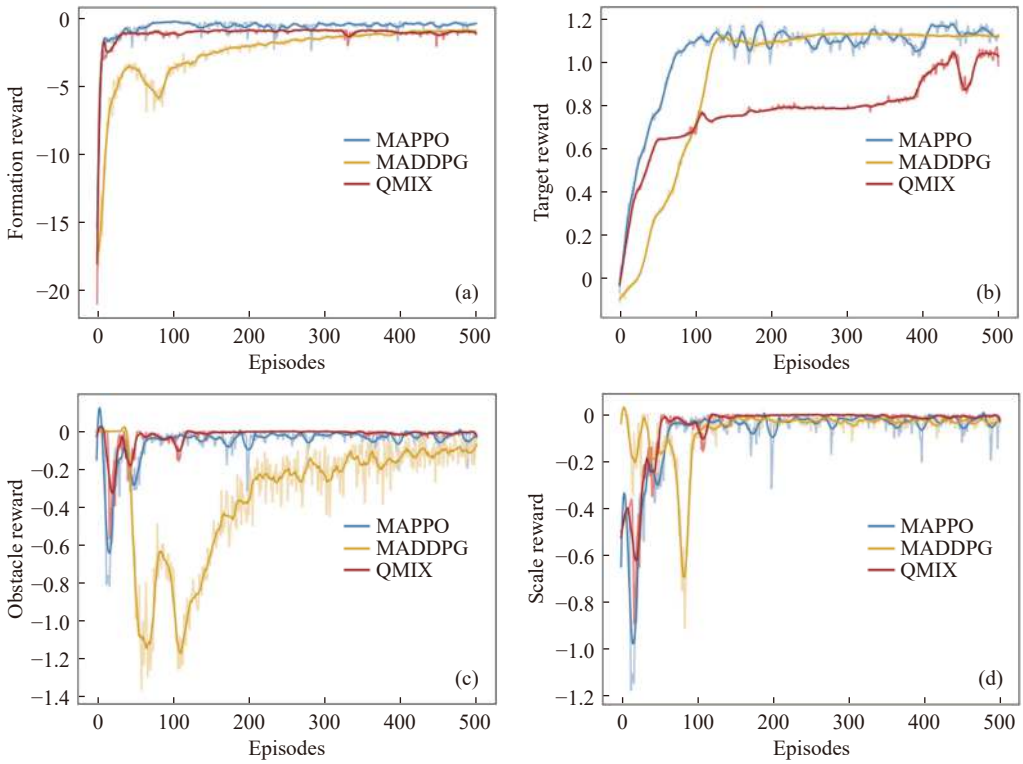

图 3 策略网络训练过程奖励曲线. (a) 队形奖励; (b) 目标奖励; (c) 避撞奖励; (d) 规模奖励

Fig.3 Rewards curves of the policy network training: (a) formation reward; (b) target reward; (c) obstacle reward; (d) scale reward

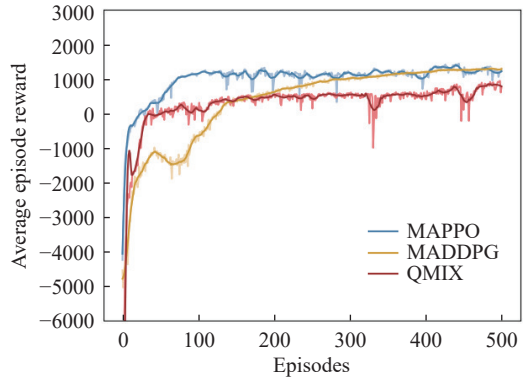

图 4 单无人机每回合平均奖励

Fig.4 Average reward of a single agent in each episode

发信装置出现在无人机的探测半径内时, 无人机机载的模拟探测器判定探测到障碍物, 并获得模拟障碍物的碰撞边缘范围. 此时无人机可以判断自己距离障碍物虚拟物理表面的距离, 进而判断是否“碰撞”到障碍物而受到惩罚. 设置的障碍物信息如表 4 所示.

基于孪生环境下得到的决策模型, 针对实物场景进行 20 个回合的补充训练. 相比于重新训练模型, 进行较短回合的补充训练即可得到对新场景更优的决策效果. 补充训练的飞行轨迹如图 5 所示, 补充训练后得到的轨迹明显更加平滑, 且过程中可以更快速安全地通过障碍物密集区域, 而

表 4 实飞测试障碍物初始信息

Table 4 Initial status of obstacles in the actual flight test

ID	*X*/m	*Y*/m	Radius/m
0	39	37	7
1	15	47	4
2	28	68	7
3	16	84	6
4	29	104	6

训练前的模型在同样的执行时间内并未完全穿越障碍区. 由图 6 中各奖励函数的对比曲线也可以看出, 其队形更加稳定, 同时无碰撞事件发生, 且能更早到达目标. 综合来看, 补充训练后的决策模型表现出更优的任务执行结果.

策略网络同样通过部署在无人机实机上进行试飞测试, 由于训练时考虑到了控制律逻辑, 因而通过孪生环境训练得到的策略网络可以方便地迁移到实物上, 图 7 给出了本试飞测试所使用的无人机平台. 该平台以搭载策略网络的树莓派作为任务机, 通过 Mavlink 协议与 CubeOrange 飞控模块交互, 进而控制无人机运动. 通过 XBee 无线数传组网模块获取邻机与模拟障碍物信息.

由图 8 可见, 试飞轨迹基本与孪生决策轨迹一致, 可知通过孪生环境训练出的深度强化学习决

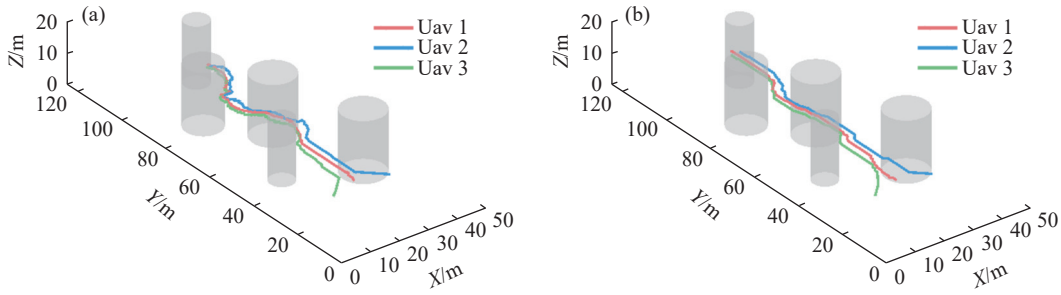

图 5　补充训练前后的决策轨迹比较. (a) 补充训练前决策轨迹; (b) 补充训练后决策轨迹

Fig.5　Comparison of the decision trajectories before and after supplementary training: (a) decision trajectory before supplementary training; (b) decision trajectory after supplementary training

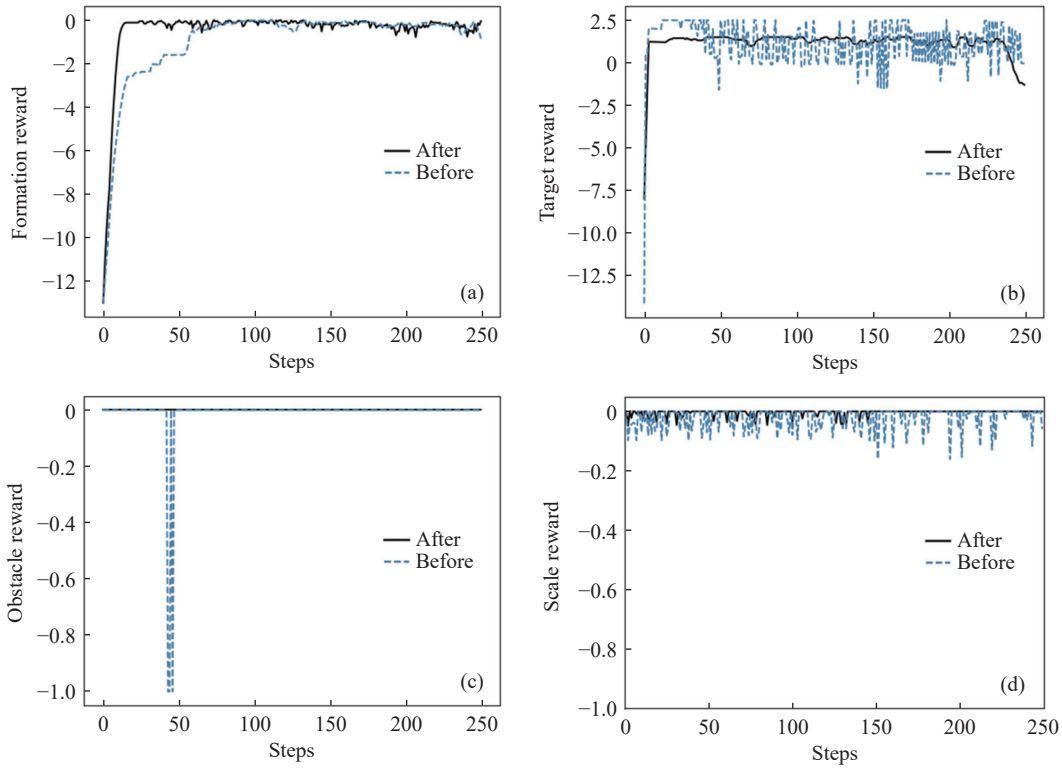

图 6　补充训练前后的过程奖励曲线比较. (a) 队形奖励; (b) 目标奖励; (c) 避撞奖励; (d) 规模奖励

Fig.6　Comparison of rewards curves before and after supplementary training in the actual scene: (a) formation reward; (b) target reward; (c) obstacle reward; (d) scale reward.

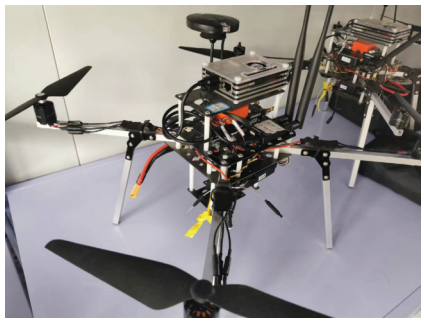

图 7　无人机验证平台

Fig.7　UAV platform for validation

策模型在现实世界同样具有可靠性. 借助于孪生模型, 在补充训练的同时, 可提前预测飞行的实际效果, 进而评估当前的任务执行能力, 可有助于确保编队避障任务的顺利完成.

4　结论

(1) 本文设计一种基于深度强化学习的无人机群协同数字孪生编队避障优化架构, 可有效克服深度强化学习算法难以应用于机载设备的难题, 通过建立孪生模型产生足量训练样本提供训练. 同时可以借助数字孪生的快速仿真验证能力预测任务执行结果, 通过实时对比预测结果和真实状态, 评估当前任务执行情况.

(2) 相比于完全借助事前训练得到的模型, 基

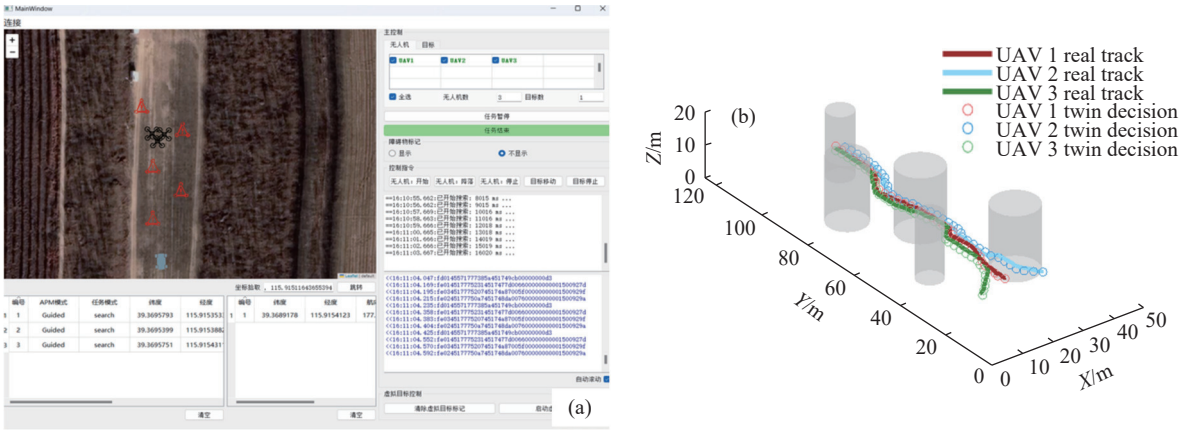

图 8 试飞验证轨迹. (a) 实验过程监控; (b) 孪生决策轨迹与实飞轨迹对比

Fig.8 Test flight validation trajectories: (a) monitoring of the experimental process; (b) comparison of the twin decision trajectory with the real flight trajectory;

于两阶段的数字孪生训练方法可以使策略模型在一般场景通用的同时, 在典型编队避障任务场景获得更优解. 通过对事先得到的训练模型进行少量的补充训练, 快速低成本地提升了实际场景中的任务执行效果, 避免了在真实世界中交互采样.

(3) 基于自身局部信息的状态向量构建策略, 使该深度强化学习优化策略能应用于友方单位和障碍物情况不定的多种场景, 避免了使用长度不定的全局信息, 场景中的成员数量和动态观测结果不影响策略网络的训练.

参 考 文 献

[1] Jia Y N, LI Q. Research development of multi-robot formation control. *Chin J Eng*, 2018, 40(8): 893
(贾永楠, 李擎. 多机器人编队控制研究进展. 工程科学学报, 2018, 40(8): 893)

[2] Vicsek T, Czirók A, Ben-Jacob E, et al. Novel type of phase transition in a system of self-driven particles. *Phys Rev Lett*, 1995, 75(6): 1226

[3] Barnes L E, Fields M A, Valavanis K P. Swarm formation control utilizing elliptical surfaces and limiting functions. *IEEE Trans Syst Man Cybern Part B Cybern*, 2009, 39(6): 1434

[4] Duan H B, Qiu H X. *Unmanned Aerial Vehicle Swarm Autonomous Control Based on Swarm Intelligence*. Beijing: Science Press, 2018
(段海滨, 邱华鑫. 基于群体智能的无人机集群自主控制. 北京: 科学出版社, 2018)

[5] Yuan X W, Wang H, Yu W W. A weighted mean field reinforcement learning algorithm for large-scale multi-agent collaboration. *Guid Navigat Control*, 2023, 3(2): 2350007

[6] Pu Z Q, Zhang T L, Ai X L, et al. A deep reinforcement learning approach combined with model-based paradigms for multiagent formation control with collision avoidance. *IEEE Trans Syst Man Cybern Syst*, 2023, 53(7): 4189

[7] Shi Y, Song J S, Hua Y Z, et al. Leader-follower formation control for fixed-wing UAVs using deep reinforcement learning // 2022 41*st Chinese Control Conference* (*CCC*). Hefei, 2022: 3456

[8] Yan Y Z, Li X X, Qiu X Y, et al. Relative distributed formation and obstacle avoidance with multi-agent reinforcement learning // 2022 *International Conference on Robotics and Automation* (*ICRA*). Philadelphia, 2022: 1661

[9] Zhang L F, Feng Y H, Liang X X, et al. Sample strategy based on TD-error for offline reinforcement learning. *Chin J Eng*, 2023, 45(12): 2118
(张龙飞, 冯旸赫, 梁星星, 等. 基于时间差分误差的离线强化学习采样策略. 工程科学学报, 2023, 45(12): 2118)

[10] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments // *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 6382

[11] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorization for deep multi-agent reinforcement learning. *J Mach Learn Res*, 2020, 21(1): 7234

[12] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of ppo in cooperative multi-agent games [J/OL]. *ArXiv* (2022–11–04) [2023–09–28]. https://arxiv.org/abs/2103.01955

[13] Grieves M. Digital Twin: Manufacturing Excellence through Virtual Factory Replication [J/OL]. *ResearchGate* (2015–04–20) [2023–09–28]. https://www.researchgate.net/publication/275211047_Digital_Twin_Manufacturing_Excellence_through_Virtual_Factory_Replication

[14] Zhuang C, Miao T, Liu J, et al. The connotation of digital twin, and the construction and application method of shopfloor digital twin. *Robot Comput Integr Manuf*, 2021, 68: 102075

[15] Malik AA, Brem A. Digital twins for collaborative robots: A case study in human-robot interaction. *Robot Comput Integr Manuf*, 2021, 68: 102092

[16] Pairet È, Ardón P, Liu X K, et al. A digital twin for human-robot

interaction // 2019 14*th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*). Daegu, 2019: 372

[17] Lv Z H, Chen D L, Feng H L, et al. Beyond 5G for digital twins of UAVs. *Comput Netw*, 2021, 197: 108366

[18] Yang Y L, Meng W, Zhu S Q. A digital twin simulation platform for multi-rotor UAV // 2020 7*th International Conference on Information, Cybernetics, and Computational Social Systems* (*ICCSS*). Guangzhou, 2020: 591

[19] Miao J S, Zhang P. UAV visual navigation system based on digital twin // 2022 18*th International Conference on Mobility, Sensing and Networking* (*MSN*). Guangzhou, 2022: 865

[20] Lei L, Shen G Q, Zhang L J, et al. Toward intelligent cooperation of UAV swarms: When machine learning meets digital twin. *IEEE Network*, 2021, 35(1): 386

[21] Xiong K, Wang Z H, Leng S P, et al. A digital-twin-empowered lightweight model-sharing scheme for multirobot systems. *IEEE Internet Things J*, 2023, 10(19): 17231

[22] Shen G Q, Lei L, Li Z L, et al. Deep reinforcement learning for flocking motion of multi-UAV systems: Learn from a digital twin. *IEEE Internet Things J*, 2022, 9(13): 11141

[23] Quan L, Yin L J, Xu C, et al. Distributed swarm trajectory optimization for formation flight in dense environments // 2022 *International Conference on Robotics and Automation* (*ICRA*). Philadelphia, 2022: 4979

[24] Liu Y X, Wang Q L. Game confrontation of 5v5 multi-agent based on MAPPO reinforcement learning algorithm // 2022 37*th Youth Academic Annual Conference of Chinese Association of Automation* (*YAC*). Beijing, 2022: 1395

[25] Papoudakis G, Christianos F, Schäfer L, et al. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks [J/OL]. *ArXiv* (2021–11–09) [2023–09–28]. https://arxiv.org/abs/2006.07869