引用格式：*吴婷，叶林奇，杨君，等.基于强化学习的多无人机避障编队控制[J].飞控与探测，2025，8(2)：9-17.*
Citation：*WU T，YE L Q，YANG J，et al. Reinforcement-learning-based multi-UAV formation control with collision avoidance [J]. Flight Control & Detection，2025，8(2)：9-17.*

基于强化学习的多无人机避障编队控制*

吴 婷[1]，叶林奇[1]，杨 君[2]，芦维宁[3]

（1. 上海大学 未来技术学院（人工智能研究院）·上海·200444；
2. 清华大学 自动化系·北京·100084；
3. 清华大学 北京信息科学与技术国家研究中心·北京·100084）

摘 要：针对多无人机编队飞行控制问题，提出了一种基于强化学习的解决方案。该方案融合了课程学习思想、Leader-Follower 模型和近端策略优化方法。首先，基于课程学习思想，将复杂的编队控制任务分解为两个学习阶段。在第一阶段，利用近端策略优化方法对 Leader 无人机进行训练，使其能够沿着预设轨迹飞行。随后进入第二阶段，对 Follower 无人机进行训练。此时 Leader 无人机的控制策略固定为第一阶段所得的神经网络，并将 Leader 的延时位置作为 Follower 的跟踪目标。根据部分可观测信息，设计了奖励函数，以引导无人机在编队飞行中保持稳定的线性队形。为了验证所提方法的有效性，在 Unity 环境中进行了四机编队空中立体 8 字形飞行仿真。结果表明，相较于传统控制方法，本方法在无须建立精确数学模型的前提下，通过相对简单的训练过程，就能使智能体在与环境的交互中学习到有效策略。这一成果简化了编队控制的复杂性，为无人机编队控制提供了一种新的解决方案。

关键词：多智能体强化学习；编队控制；近端策略优化；课程学习；Leader-Follower 模型；无人机飞行控制

中图分类号：TP273 **文献标志码**：A **文章编号**：2096-5974(2025)02-0009-9

DOI：10.20249/j.cnki.2096-5974.2025.02.002

Reinforcement-Learning-Based Multi-UAV Formation Control with Collision Avoidance

WU Ting[1], YE Linqi[1], YANG Jun[2], LU Weining[3]

(1. School of Future Technology (Institute of Artificial Intelligence), Shanghai University, Shanghai 200444;
2. Department of Automation, Tsinghua University, Beijing 100084;
3. Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084)

Abstract: This paper proposes a reinforcement learning-based method to address the flight control problem for multi-UAV formation. The proposed approach integrates curriculum learning, the leader-follower model, and the proximal policy optimization (PPO) method. Firstly, based on

* 基金项目：国家自然科学基金（62225308，62003188）
作者简介：吴婷，女，硕士生。
通信作者简介：叶林奇，男，博士，副研究员，硕士生导师。

curriculum learning, the complex formation control task is decomposed into two learning stages. In the first stage, the leader UAV is trained using the PPO method to fly along a preset trajectory. Subsequently, the follower UAVs are trained in the second stage. During this stage, the control policy of the leader UAV is fixed as the neural network obtained in the first stage, and the delayed position of the leader is used as the tracking target for the followers. Based on partially observable information, the reward function was elaborately designed to guide the UAVs to maintain a stable linear formation during flight. To validate the effectiveness of the proposed method, we conducted simulations of a four-UAV formation performing a complex "8" shaped flight pattern in a three-dimensional space using the Unity software. The results demonstrate that, compared to traditional control methods, our approach enables agents to learn effective strategies through interaction with the environment via a relatively simple training process, without the need to establish precise mathematical models. This method simplifies the complexity of formation control and provides a novel solution for UAV formation control.

Keywords: multi-agent reinforcement learning; formation control; proximal policy optimization; curriculum learning; leader-follower model; UAV control

0 引 言

相比于单智能体，多智能体系统具有更好的鲁棒性、更大的容错性，并且能协同完成更复杂的任务。近年来，多智能体的编队控制已成为研究的热点[1]。编队控制是研究多个智能体在有约束的条件下（如避障），向着特定的目标运动或者保持设定的几何形态的控制问题[2-3]。无人机具有更高的经济性、较低的成本，并且无人机集群在工业、军事等领域发挥着越来越重要的作用，因此，针对无人机的编队控制具有重要的研究意义。

在传统控制领域，有许多比较成熟的理论，如 PID 算法、状态观测器和滑模控制等，经过大量研究已验证这些方法具有鲁棒性。文献［4］根据李雅普诺夫方法推导反馈控制律，实现非完整系统的轨迹跟踪和避障，并扩展至多智能体。文献［5］采用传统的 PID 算法，先使无人机集群保持队形，再设计不同的控制策略使得无人机集群完成队形变化。文献［6］提出了基于扩张状态观测器的控制方法，通过重构人工势场法实现无人机和无人艇在平面上的目标跟踪，并将编队误差和跟踪误差限制在有限范围内。文献［7］中建立了准确的无人机动力学模型，设计了不受初始状态约束的固定时间的终端滑模和分布式固定时间的跟踪控制器，使得无人机集群在固定时间内组成预设的队形。文献［8］将编队问题转化为跟踪控制问题，将系统离散化再基于迭代学习设计编队控制器。以上方法都通过仿真验证了有效性，但是都需要建立准确的无人机模型，难以解决非线性和高阶模型的控制问题，在动态环境中表现不佳，且方法和原理相对复杂。

为应对以上问题和处理更复杂的非线性系统，近年来许多研究使用了强化学习来实现多智能体的编队控制。面对非线性多智能体系统，强化学习能有效求解哈密尔顿-雅各比-贝尔曼方程[9-10]。此外，强化学习方法不需要建立准确的模型，通过设计奖励函数训练智能体即可取得良好的控制效果。文献［11］中将深度强化学习引入到无人机集群编队控制中，利用改进的近端策略优化方法，将动态估计法作为评价机制，提升了训练效率，能有效解决高维度集群控制问题。文献［12］分别用深度确定性策略梯度、置信域策略优化方法和近端策略优化方法，在 GYMFC 环境中训练智能飞行器的内环控制，并迁移到四旋翼无人机的姿态控制中，结果表明，近端策略优化方法在精度和快速性方面要优于其他两种方法。除优化策略外，也有许多针对编队控制方法的研究，典型的编队控制方法有 Leader-Follower 法，基于行为法、虚拟结构法和图论法。Leader-Follower 法原理简单，当 Leader 的状态确定后，Follower 跟踪 Leader 的状态即可，经过很多研究证明，这种方法能够完成大部分集群任务[1-2]，如利用 Leader-Follower 法完成编队跟踪和路径规划任务。文献［13］构建了 Leader-Follower 的模型，将编队控

制问题视为最优输出调节问题，利用离策略强化学习求出折扣性能函数的解，使得编队的跟踪误差收敛为 0。文献［14］令未知动态的 Follower 跟踪虚拟的 Leader，基于强化学习的分布式最优控制实现异构无人机的编队轨迹跟踪。文献［15］使用平衡集中式和完全分布式的 Leader-Follower 架构实现多智能体的联合路径规划。

在以上工作基础上，本文考虑了更加复杂的地形环境以及更高难度的空中立体 8 字编队飞行控制问题。为实现多个无人机按预先设定的立体 8 字形轨迹飞行并保持一字形的编队队形，本文提出了一种基于 Leader-Follower 模型的分阶段强化学习方法，在无须建立准确无人机模型的条件下，将多智能体编队控制任务分成两个阶段，首先使用近端策略优化方法训练 Leader 沿预设轨迹飞行，接着将 Leader 的延时位置作为 Follower 虚拟的跟踪目标，根据局部观测设计奖励函数，再使用近端策略优化（Proximal Policy Optimization，PPO）算法训练多无人机的编队控制。最后，在 Unity 中构建仿真场景验证方法的有效性，即无人机能保持线性队形并有效避障。

1　设计思路

强化学习的一个重要理论基础是马尔可夫决策过程，智能体和环境交互并做出决策，且得到奖励，而动作决策只与当前的状态有关，与历史状态无关。马尔可夫决策过程[16-17]定义了一个序列 $\{N, S, (A_i)_{i\in N}, P, \{R_i\}_{i\in N}, \gamma\}$，其中 N 表示所有的智能体，S 代表所有智能体的观测集合，$\{A_i\}_{i\in N}$ 表示所有智能体的联合动作状态，P 表示智能体采取动作 $a\in A$ 的情况下，从一个状态转移到另一个状态的概率。$\{R_i\}_{i\in N}$ 表示状态转移得到的奖励。基于马尔可夫模型，每个无人机在优化自己的奖励函数的同时，整个无人机编队系统得到的奖励也受到联合策略的影响。

近端策略优化方法原理比较简单且训练效果稳定。典型的 PPO 算法是基于策略梯度以及信赖域策略梯度，通过构造 KL 惩罚系数和剪切替代目标来构造损失函数。剪切法是先比较新策略和旧策略在采样过程中生成的动作的概率比值，再比较这个比值与一个预先设定的阈值，来决定是否对梯度进行修剪。结合策略剪切目标和价值函数误差的损失函数，利用策略梯度和价值函数共享参数的网络架构，并增加熵奖励来保证迭代的效率[18]。本文在仿真中，采用在 Unity 中集成的 ML-Agents 模块调用 PPO 算法进行强化学习训练。

多无人机的空中立体 8 字编队飞行控制是一个具有挑战性的任务。立体 8 字编队飞行要求无人机在三维空间中完成复杂的轨迹动作，这增加了控制的难度。每架无人机都需要根据编队要求和自身位置进行实时调整，才能确保整个编队的稳定性和精确性。此外，在编队飞行过程中，无人机需要实时感知周围环境并进行避障，以确保在复杂环境中安全飞行。为实现该任务并保证良好的泛化性和收敛性，本文使用课程学习[19-20]，将无人机编队任务分成两个阶段。第一阶段，对无人机的速度、与路径点的相对位置、路径点的朝向进行观测，并将观测值作为输入量，将俯仰、偏航以及加速作为输出动作，利用 PPO 算法训练其中一个无人机沿着预设轨迹飞行，得到 Leader 无人机的策略网络。在第二阶段，将 Leader 的延时位置作为 Follower 的虚拟跟踪目标，对 Follower 的速度、与虚拟目标的相对位置、虚拟目标的朝向进行观测，并将观测值作为输入量，将俯仰、偏航以及加速作为输出动作，结合第一阶段训练好的网络，再次利用 PPO 算法训练得到 Follower 无人机的策略网络。

2　仿真模型与算法设计

2.1　仿真建模

无人机的动力学模型如下。

1）飞行时所受的推力大小

$$\boldsymbol{F}_t = \hat{\boldsymbol{f}} tb \tag{1}$$

其中，$\hat{\boldsymbol{f}}$ 表示飞行器的朝向，t 表示发动机的推力大小，本文中设置为 10^5，b 表示增压倍数，加速状态下为 2，否则为 1。

2）俯仰角

$$\theta = x_c + C_{\text{pitch}}\Delta TS_{\text{pitch}} \tag{2}$$

3）偏航角

$$\varphi = y_c + C_{\text{yaw}}\Delta TS_{\text{yaw}} \tag{3}$$

4）横滚角

$$\phi = z_c + C_{\text{roll}}\Delta TS_{\text{roll}} \tag{4}$$

其中，x_c，y_c，z_c 分别表示当前的俯仰角、偏航角和横滚角，式（2）、式（3）、式（4）后半部分表示各个姿态角的平滑增量，ΔT 表示固定的时间步长，本文设为 0.02 s。C_{pitch} 表示平滑的俯仰改变

量，其范围从当前的俯仰角度以 2 倍的 ΔT 时间步长平滑过渡到目标俯仰角度，通常范围为 [−1，1]。同理可知 C_{yaw}，C_{roll} 分别为平滑的偏航角和横滚角的增量。S_{pitch}，S_{yaw} 和 S_{roll} 表示俯仰速度、航向速度和横滚速度的大小，本文中均设置为 100 (°) /s，并且将俯仰角和横滚角的范围设置为 [−45°，45°]。这些角度的增量和速度是通过施加俯仰动作、偏航动作和增压控制的。

2.2 奖励设计

为实现多个无人机沿着绕过多种障碍物的预设轨迹，并保持一定的相对距离和线性队形飞行，采用 Leader-Follower 模型，将飞在最前面的无人机设定为 Leader，跟在其后面的飞行器定义为 Follower，整个控制框图如图 1 所示。

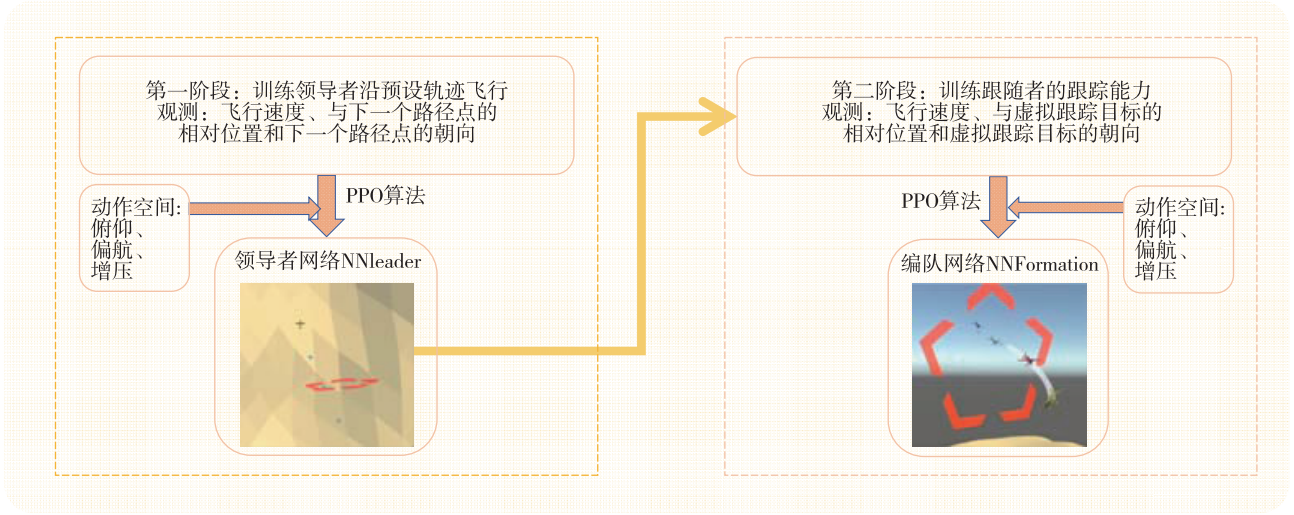


图 1 控制框图

Fig. 1 Control framework

基于构建的 Leader-Follower 模型，结合各自的观测值，设计无人机的奖励函数。在训练模式下，Follower 的奖励函数由以下部分组成：Follower 和虚拟跟踪目标之间距离 $d = \|\boldsymbol{p}_{follower} - \boldsymbol{p}_{target}\|$ 的奖惩、Follower 与跟踪目标前进方向的偏差 $f = \|\boldsymbol{f}_{follower} - \boldsymbol{f}_{target}\|$ 的奖惩

$$R_{follower} = \begin{cases} 1 - d \times 0.1 - f, & 0 < d \leqslant 10 \\ -0.5 - f, & d > 10 \end{cases} \tag{5}$$

对于 Leader，训练模式下，当距离下一个路径点的距离小于 20 m 时得到 0.5 的奖励，若超出了训练时间则得到 0.5 的惩罚。此外，若 Leader 或 Follower 撞到障碍物，将得到值为 1 的惩罚。

2.3 网络设计

在强化学习训练中，编写 Yaml 文件给出神经网络的超参数。采用 PPO 算法进行训练，将观测（维度为 9）作为神经网络的输入，将动作空间（维度为 3）作为神经网络的输出。神经网络包含策略网络和价值网络两部分，奖励信号为外部奖励，折扣因子为 0.995，强度为 1.0。两个神经网络的类型以及参数如表 1 所示。

表 1 神经网络参数

Tab. 1 Parameters of neural network

神经网络类型	参数名称	数值
策略网络	神经网络类型	全连接神经网络
	隐藏层数量	3
	隐藏层单元数量	512
价值网络	神经网络类型	全连接神经网络
	隐藏层数量	2
	隐藏层单元数量	128

表 2 列举了神经网络训练的超参数。其中 Batch size 为每次训练步数使用的样本数量；Buffer size 为在训练之前积累的经验数量；Learning rate 为学习率，该参数影响模型参数更新的步长；Learning rate schedule 为学习率调整，Linear 表明学习率为线性变化；Time horizon 表明了智能体影响未来行为的时间范围；Summary frequency 为进行一次统计总结的步数；Checkpoint interval 为保存一次检查点的步数；Keep checkpoints 为保留的检查点的数量；Max steps 为训练过程结束前在环境中采取观察和动作的总步数。

表 2　训练中默认的超参数

Tab. 2　Default hyperparameters for training

参数名称	数值
Batch size	2 048
Buffer size	20 480
Learning rate	0.000 3
Learning rate schedule	Linear
Time horizon	1 000
Summary frequency	50 000
Checkpoint interval	400 000
Keep checkpoints	100
Max steps	50 000 000

表 3 中列举了 PPO 训练策略所采用的特定的超参数。其中 Beta 为熵正则化的强度；Epsilon 影响策略在训练过程中的发展速度；Beta schedule 为 Linear，表明了 Beta 为线性变化；Epsilon schedule 为 Linear，表明了 Epsilon 为线性变化；Lambd 表示了计算广义优势估计使用的正则化参数。

表 3　PPO 特定的超参数

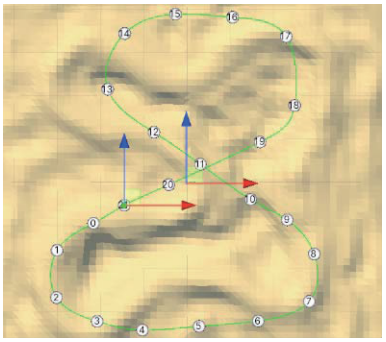
Tab. 3　Specific hyperparameters of PPO

参数名称	数值
Beta	0.005
Epsilon	0.2
Beta schedule	Linear
Epsilon schedule	Linear
Lambd	0.95

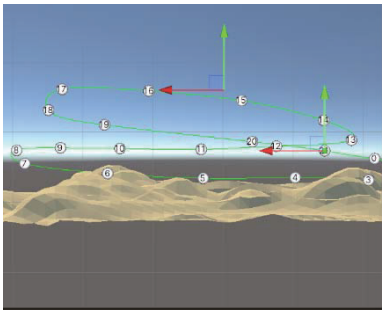
3　仿真实验及结果分析

3.1　仿真结果

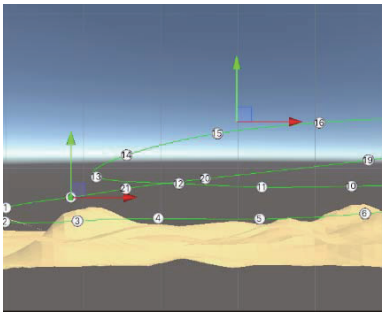
在 Unity 软件中，搭建一个含有多种障碍物的场景，如图 2 所示。在每隔约 100 m 的位置设置一个路径点，通过 22 个点位形成闭环的空中立体 8 字形轨迹，所有的路径点被存储在序列中。整个预设轨迹限制在宽度约 600 m、长度约 700 m、高度约 200 m 的范围内。在每一轮 episode 开始时，飞行器会置于随机的一个路径点附近并将其作为起始路径点。当 Leader 的训练时间超过预设的一个时间（本文为 500 s），或 Follower 的跟踪误差超过 10 m，或任意一个飞行器撞击障碍物时会结束当前的训练，重置环境，在新的 episode 中进行下一轮训练。本试验使用普通的笔记本电脑（Intel（R）Core（TM）i5-7300HQ CPU），由于训练任务的复杂性，第一阶段和第二阶段分别训练了约 15 min 和 2.6 h。



（a）飞行轨迹俯视图



（b）飞行轨迹侧视图



（c）飞行轨迹后视图

图 2　Unity 仿真场景

Fig. 2　Simulation scene in Unity

采用上文提出的方法进行强化学习训练，成功实现了四机空中立体 8 字形编队飞行控制，仿真效果见录制的视频：https://www.bilibili.com/video/BV1nz421y7dg/。

图 3、图 4、图 5 分别为训练 Leader 沿预设轨迹飞行的累积奖励、策略损失和价值损失曲线。从图中可以看出，累积奖励曲线在训练中期的上

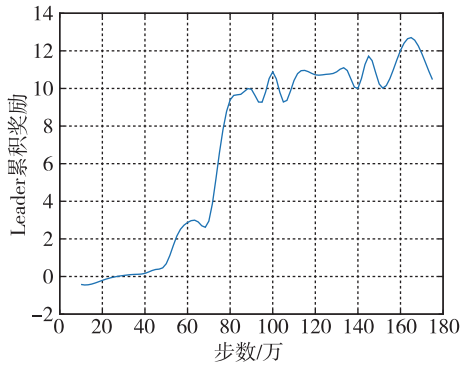


图 3 Leader 累积奖励曲线

Fig. 3 Curve of Leader's cumulative reward

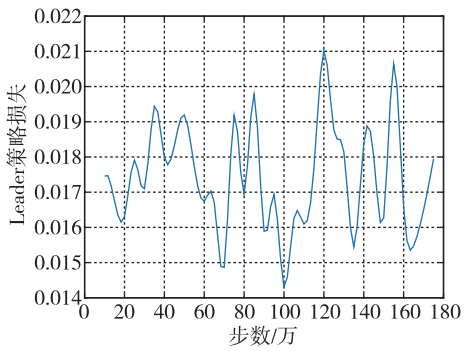


图 4 Leader 的策略损失

Fig. 4 Policy loss of Leader

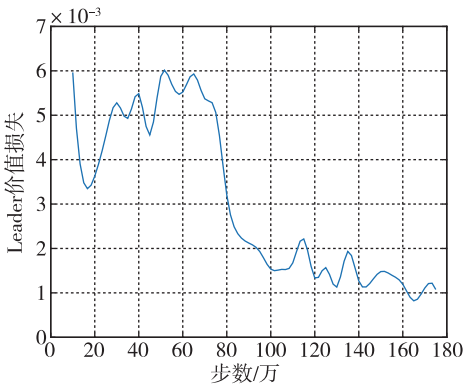
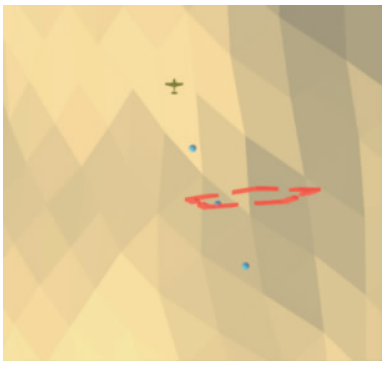


图 5 Leader 的价值损失

Fig. 5 Value loss of Leader



(a) Leader 直线飞行

(b) Leader 左转

(c) Leader 右转

图 6 Leader 绕 8 字形轨迹飞行仿真图

Fig. 6 Simulation results of Leader flying along the figure-eight trajectory

升趋势较快，并且策略损失在一个很小范围内浮动，价值损失也在训练后期收敛至很小的值，表明训练效果良好。图 6 为 Leader 沿直线飞行和转弯的仿真图。

图 7、图 8、图 9 分别为 Follower 的累积奖励、策略损失和价值损失曲线。从图中可以看出，累积奖励曲线在训练前期和中期的上升趋势较快，并且策略损失在一个很小范围内浮动，表明训练效果良好，验证了第二阶段的训练模型学习到了比较好的策略。

图 10（a）为整个编队轨迹的俯视图，是飞行器绕空中立体 8 字形轨迹飞行的整体路线，图 10（b），（c）为编队飞行的侧视图，展示飞行器之间保持固定距离分别向右和向左转飞行。图 11 为飞

行器 Leader 在任意初始位置情况下的编队跟踪轨迹。图 12 为 Followers 随时间变化的跟踪误差。可以看出，跟踪误差在大部分的时间段内均比较小，仅在某些转弯处发生突变，但之后仍能快速减小。

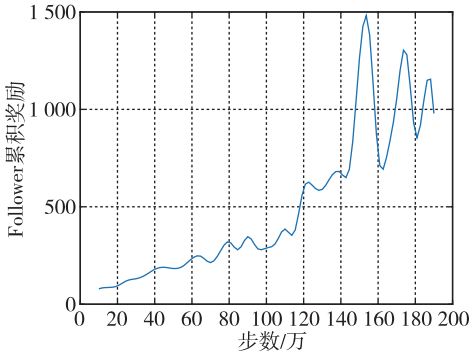


图 7　Follower 累积奖励曲线

Fig. 7　Curve of Follower's cumulative reward

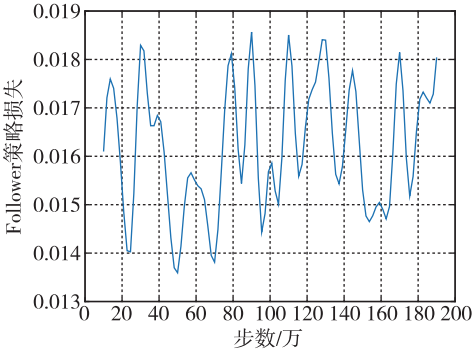


图 8　Follower 策略损失

Fig. 8　Follower's policy loss

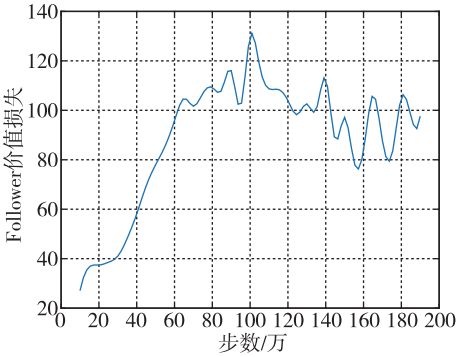
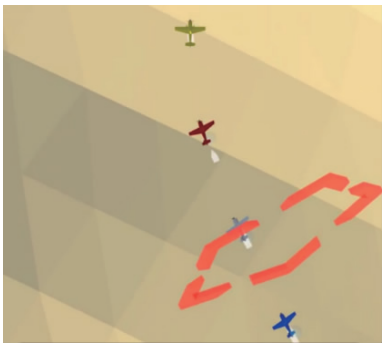
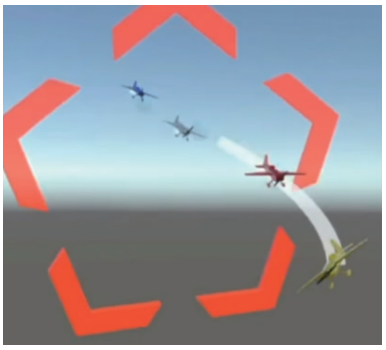


图 9　Follower 价值损失

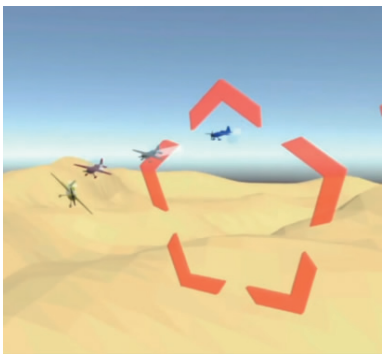
Fig. 9　Follower's value loss



(a) 编队直线飞行



(b) 编队右转飞行



(c) 编队左转飞行

图 10　保持预设距离编队飞行的仿真结果

Fig. 10　Simulation results of formation flying with the preset distance

3.2　结果分析

通过仿真可知，即使在比较复杂的环境以及较短的训练时间内（Leader，Follower 各训练约 200 万步），多架无人机在 Leader-Follower 框架基础上通过 PPO 算法进行两个阶段的训练，也能按照预设的轨迹飞行，同时还保持线性队形。仿真

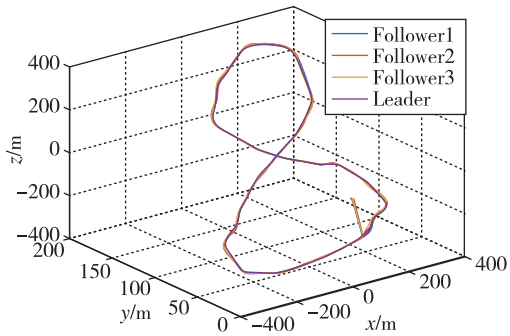


图 11 编队飞行轨迹

Fig. 11 Formation trajectory

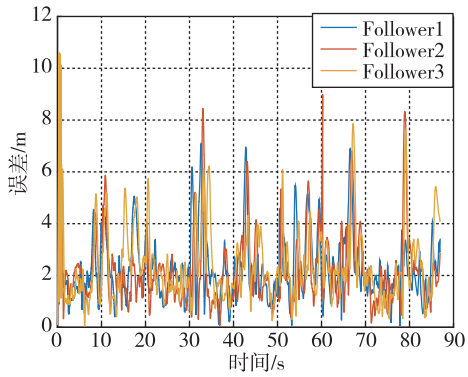


图 12 跟踪误差

Fig. 12 Tracking error

结果表明，该方法既实现了比较好的跟踪效果，又能够实现有效避障。

然而，该方法仍然存在一些不足，在转弯剧烈的情况下，飞行器的跟踪效果不甚理想。这可能是由于在奖励设置上没有考虑到 Follower 跟踪虚拟目标时两者的相对速度，仅仅考虑了两者的相对距离以及飞行方向的一致性。另一种可以避免在弯道处跟踪效果不理想的方法是更为准确地计算弯道处 Follower 和虚拟跟踪目标的圆周距离，并延长训练的时间，经过更长时间的训练可能达到更好的跟踪效果。

4 结 论

本文提出了一种基于 Leader-Follower 模型的分阶段强化学习方法，实现了无人机在复杂地形中按立体 8 字形轨迹飞行并保持一字形编队，通过 Unity 仿真验证了方法的有效性。本文主要贡献在于：1）基于多智能体强化学习，在不需要准确无人机模型的条件下精确控制无人机的运动，相比于传统控制方法避免了复杂的建模过程；2）基于 Leader-Follower 框架，并利用课程学习将较为复杂的编队任务简化为两个阶段的训练任务。最后在 Unity 中搭建 8 字形轨迹，并利用软件内部封装的 PPO 算法进行训练。从仿真结果来看，合理的奖励设置有助于提升多智能体保持线性队形的轨迹跟踪能力。在未来的研究中，可以考虑改善奖励设计，使得多智能体在转弯剧烈的情况下仍能保持比较好的跟踪效果，以及设计不同的队形以适应环境的变化。

参考文献 (References)

[1] 王祥科，李迅，郑志强．多智能体系统编队控制相关问题研究综述［J］．控制与决策，2013，28（11）：1601-1613.
WANG X K，LI X，ZHENG Z Q. Survey of developments on multi-agent formation control related problems［J］. Control and Decision，2013，28（11）：1601-1613（in Chinese）.

[2] 杨立炜，付丽霞，李萍．多智能体系统编队控制发展综述［J］．电子测量技术，2020，43（24）：18-27.
YANG L W，FU L X，LI P. Summary of development of multi-agent system formation control［J］. Electronic Measurement Technology，2020，43（24）：18-27（in Chinese）.

[3] 孙长银，穆朝絮．多智能体深度强化学习的若干关键科学问题［J］．自动化学报，2020，46（7）：1301-1312.
SUN C Y，MU C X. Important scientific problems of multi-agent deep reinforcement learning［J］. Acta Automatica Sinica，2020，46（7）：1301-1312（in Chinese）.

[4] MASTELLONE S，STIPANOVIĆ D M，GRAUNKE C R，et al. Formation control and collision avoidance for multi-agent non-holonomic systems：theory and experiments［J］. The International Journal of Robotics Research，2008，27（1）：107-126.

[5] LUO D，XU W，WU S，et al. UAV formation flight control and formation switch strategy［C］//2013 8th International Conference on Computer Science & Education. Colombo：IEEE，2013：264-269.

[6] HUANG Y，LI W，NING J，et al. Formation control for UAV-USVs heterogeneous system with collision avoidance performance［J］. Journal of Marine Science and Engineering，2023，11（12）：2332.

[7] 田霖，孙亮．固定翼无人机集群分布式固定时间编队跟踪控制［J］．测控技术，2023，42（11）：58-63，72.
TIAN L，SUN L. Distributed fixed-time formation tracking control for fixed-wing UAV swarm［J］. Measurement & Control Technology，2023，42（11）：58-63，72（in Chinese）.

[8] 顾立雯，王玉龙，马浪．基于迭代学习的多智能体系统协

同编队控制［J］. 控制工程，2021，28（11）：2178-2184.
GU L W，WANG Y L，MA L. Cooperative formation control of multi-agent systems based on iterative learning［J］. Control Engineering of China，2021，28（11）：2178-2184（in Chinese）.

［9］ WEN G，CHEN C L P，LI B. Optimized formation control using simplified reinforcement learning for a class of multiagent systems with unknown dynamics［J］. IEEE Transactions on Industrial Electronics，2019，67（9）：7879-7888.

［10］ ZHANG Y，CHADLI M，XIANG Z. Prescribed-time formation control for a class of multi-agent systems via fuzzy reinforcement learning［J］. IEEE Transactions on Fuzzy Systems，2023，31（12）：4195-4204.

［11］ 全家乐，马先龙，沈昱恒. 基于近端策略动态优化的多智能体编队方法［J］. 空天防御，2024，7（2）：52-62.
QUAN J L，MA X L，SHEN Y H. Multi-agent formation method based on dynamic optimization of proximal policies［J］. Air & Space Defense，2024，7（2）：52-62（in Chinese）.

［12］ KOCH W，MANCUSO R，WEST R，et al. Reinforcement learning for UAV attitude control［J］. ACM Transactions on Cyber-Physical Systems，2019，3（2）：1-21.

［13］ LIU H，PENG F，MODARES H，et al. Heterogeneous formation control of multiple rotorcrafts with unknown dynamics by reinforcement learning［J］. Information Sciences，2021，558：194-207.

［14］ LIU H，MENG Q，PENG F，et al. Heterogeneous formation control of multiple UAVs with limited-input leader via reinforcement learning［J］. Neurocomputing，2020，412：63-71.

［15］ DI CARO G A，YOUSAF A W Z. Multi-robot informative path planning using a leader-follower architecture［C］//2021 IEEE International Conference on Robotics and Automation（ICRA）. Xi'an，China：IEEE，2021：10045-10051.

［16］ ZHANG K，YANG Z，BAŞAR T. Multi-agent reinforcement learning：A selective overview of theories and algorithms［J］. Handbook of Reinforcement Learning and Control，2021，325：321-384.

［17］ GRONAUER S，DIEPOLD K. Multi-agent deep reinforcement learning：a survey［J］. Artificial Intelligence Review，2022，55（2）：895-943.

［18］ SCHULMAN J，WOLSKI F，DHARIWAL P，et al. Proximal policy optimization algorithms［J］. arXiv preprint arXiv：1707.06347，2017. https：//arxiv.org/abs/1707.06347.

［19］ BENGIO Y，LOURADOUR J，COLLOBERT R，et al. Curriculum learning［C］//Proceedings of the 26th Annual International Conference on Machine Learning. NY，United States：ICML，2009：41-48.

［20］ WANG X，CHEN Y，ZHU W. A survey on curriculum learning［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44（9）：4555-4576.