文章编号:1674-2974(2026)04-0001-09 DOI:10.16339/j.cnki.hdxbzkb.2026261

# 结合注意力机制与深度强化学习的无人机智能追踪方法

江未来[1,2],王文路[1,2†],王耀南[1,2]
(1. 湖南大学 电气与信息工程学院,湖南 长沙 410082;
2. 湖南大学粤港澳大湾区创新研究院,广东 广州 511300)

**摘 要**:针对深度强化学习算法在求解复杂环境下动态目标的无人机智能追踪问题时存在的收敛速度慢、成功率低、模型泛用性差等问题,将注意力机制与深度确定性策略梯度(deep deterministic policy gradient, DDPG)结合,提出了attention-DDPG模型,并在此基础上结合多经验池方法,搭建了一种新的深度强化学习算法——多经验池注意力深度确定性策略梯度(multi pool attention deep deterministic policy gradient, MPADDPG)算法.算法将注意力机制加入DDPG的Actor网络,赋予状态中各个分量不同的权重来突出重要且关键的信息,并引入多经验池机制分离失败经验,强化成功经验,提升了算法的收敛性;更进一步,通过赋予无人机环境感知能力,提升了算法的泛化能力.最后,在建立的无人机的连续状态空间与动作空间中验证了MPADDPG算法的有效性.仿真结果表明,MPADDPG算法的智能追踪成功率超过90%,较DDPG算法具有更高的追踪成功率与更强的泛化能力.

**关键词**:无人机;智能追踪;注意力机制;深度强化学习;多经验池
**中图分类号**:V279;V249 **文献标志码**:A

# Intelligent tracking method for UAV combining attention mechanism and deep reinforcement learning

JIANG Weilai[1,2], WANG Wenlu[1,2†], WANG Yaonan[1,2]
(1. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. Greater Bay Area Institute for Innovation, Hunan University, Guangzhou 511300, China)

**Abstract**: To address the challenges of slow convergence rates, low success rates, and poor model generalization in deep reinforcement learning algorithms for UAV intelligent tracking decision-making under a complex environment, this study combines the attention mechanism with the deep deterministic policy gradient (DDPG) approach and introduces an attention-DDPG model. Meanwhile, incorporating a multi-experience pool method, a proposed algorithm, named multi pool attention deep deterministic policy gradient (MPADDPG) is built up, and the attention mechanism is integrated into DDPG's Actor network. This allows for different weights to be assigned to various state components, so as to highlight the crucial information, while a multi-experience pool

* **收稿日期**:2024-11-28
**基金项目**:国家自然科学基金资助项目(62473138), National Natural Science Foundation of China(62473138)
**作者简介**:江未来(1989—),男,江西上饶人,湖南大学副教授,博士生导师
† 通信联系人,E-mail:3277109009@qq.com

strategy is employed to distinguish unsuccessful experiences from successful ones in order to improve convergence. Additionally, enhancing the UAV's environmental perception capabilities further boosts the algorithm's generalization ability. The effectiveness of MPADDPG is validated within a continuous state and action space framework established in this research. Simulation results demonstrate that MPADDPG achieves an intelligent tracking success rate exceeding 90%, outperforming DDPG in both tracking success and generalization capability.

**Key words**: unmanned aerial vehicle (UAV); intelligent tracking; attention mechanism; deep reinforcement learning; multi-experience pool

无人机智能追踪是一类典型的复杂环境下动态目标追踪的决策问题.这类问题要求无人机能够在动态、不确定或受限的复杂环境中自主做出决策,以实现对特定目标的持续、准确追踪.然而,传统的无人机追踪算法[1]面对复杂的环境存在适应性差、计算速度慢等问题,不利于无人机完成及时、准确的追踪任务.

为解决此类问题,学术界开展了卓有成效的研究.在文献[2-5]中,利用强化学习方法,通过将环境离散化并采用$Q$-learning算法成功实现了无人机对静态目标的追踪,但$Q$-learning算法的适用范围仅限于离散且低维的状态空间以及动作空间,在面对连续状态空间或动作空间的复杂情形,$Q$-learning算法效果不佳.文献[6-9]利用深度强化学习方法,实现了静态环境下基于DQN算法的无人机路径自动规划,文献[10]提出利用人工势场法与DQN结合的方法解决局部最优问题,提升了无人机在复杂环境下的自主避障与目标搜索能力.然而,上述算法主要适用于连续状态空间、离散动作空间类问题,对于连续状态空间和动作空间类问题仍存在局限.为了对连续状态空间和动作空间类问题进行有效求解,文献[11-12]采用Actor-Critic框架下的深度确定性策略梯度(DDPG)算法,通过输入周围环境的初始状态,求解每个状态下的确定性动作,实现了无人机在复杂室内环境下的搜索与救援.然而,上述文献主要讨论了无人机在静态环境下的追踪决策问题与动态环境下的避障问题,对于复杂环境下无人机针对动态目标的智能追踪决策问题,却未进一步探索.

本文针对复杂环境下无人机动态目标的智能追踪决策问题作出了三个方面的贡献.首先,不同于文献[13-14]将状态空间或动作空间离散化,本文考虑无人机的状态及动作均为连续变化.其次,将注意力机制[15]引入DDPG的Actor网络中,对无人机状态中各个分量赋予不同的权重,来突出重要且关键的信息,并结合多经验池方法将无人机与环境交互产生的经验数据划分为成功经验数据和失败经验数据,提升算法的收敛性.最后,赋予无人机环境感知能力,联合环境感知信息和目标的状态信息,提升算法的泛化能力.

## 1 问题描述

为了便于对复杂环境下无人机动态目标智能追踪决策问题进行分析与求解,本文提出两个设定:

1)设定无人机为质点,其飞行状态为低空且高度恒定,将无人机运动从三维空间简化到二维空间.

2)设定无人机在确定的长方形区域内运动,环境中存在不同半径的圆形障碍物.智能追踪决策示意图如图1所示.

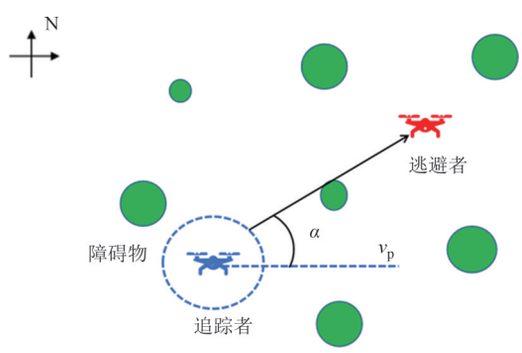



图1 智能追踪示意图

Fig.1 Intelligent tracking diagram

如图1所示,蓝色无人机为追踪者,蓝色圆环为追踪者感知半径,红色无人机为逃避者,绿色圆形表示障碍物.追踪者可通过自身搭载的传感器感知周围环境信息、逃避者信息,具有在环境中自主避障与追踪目标的能力.逃避者可在环境中随机运动并自由避障.障碍物的大小、位置随机分布.

本文的研究目标是实现追踪者在复杂环境下对逃避者的智能追踪.

## 2　问题建模

本文针对复杂环境下的无人机动态目标智能追踪问题,基于深度强化学习方法设计新的观测空间、动作空间以及奖励函数.

### 2.1　观测空间

观测空间涵盖了追踪者与逃避者的状态信息,包括坐标、速度大小和运动方向,以及环境障碍物信息,包括坐标、半径等.追踪者与逃避者的状态信息如图2所示.

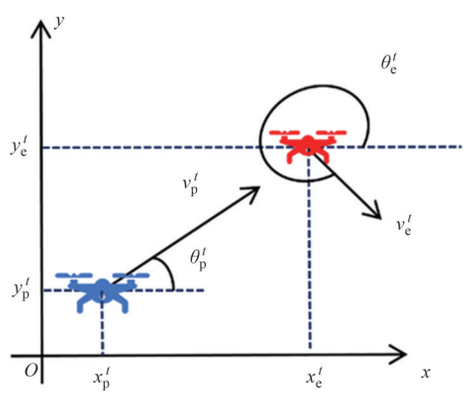



图2　无人机状态示意图

Fig.2　UAV status diagram

图2中,假设追踪者的状态信息为 $\boldsymbol{s}_p=[(x_p^t,y_p^t),v_p^t,\theta_p^t]$,其中 $(x_p^t,y_p^t)$、$v_p^t$、$\theta_p^t$ 分别为当前时刻 $t$ 的位置坐标、速度大小、运动方向角.逃避者的状态信息为 $\boldsymbol{s}_e=[(x_e^t,y_e^t),v_e^t,\theta_e^t]$,其中 $(x_e^t,y_e^t)$、$v_e^t$、$\theta_e^t$ 分别为位置坐标、速度大小、运动方向角.同时设定追踪者最大飞行速度为 $v_p^{\max}$,逃避者最大飞行速度为 $v_e^{\max}$.

追踪者可感知一定范围内的环境障碍物信息,如图3所示.

$\boldsymbol{d}_o=[d_1,d_2,\cdots,d_8]$ 表示追踪者与障碍物的距离.$\alpha$ 表示追踪者运动方向与逃避者之间的夹角.假设无人机运动区域的长和宽分别为 $h$ 和 $w$,$d_t$ 表示追踪者与逃避者之间的相对距离,如式(1)所示.

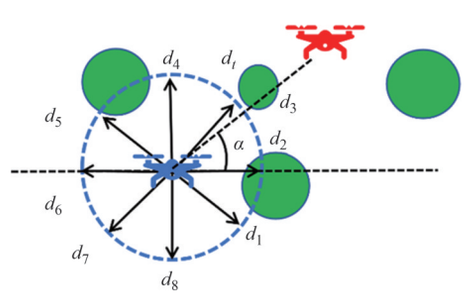



图3　无人机状态感知示意图

Fig.3　Schematic diagram of UAV state perception

$$d_t=\sqrt{\left(x_e^t-x_p^t\right)^2+\left(y_e^t-y_p^t\right)^2},\ d_t\in\left[0,\sqrt{w^2+h^2}\right] \tag{1}$$

因此,观测空间可以具体表示为:

$$\boldsymbol{O}=\left[\boldsymbol{s}_p,\boldsymbol{s}_e,\boldsymbol{d}_o,d_t,\alpha\right] \tag{2}$$

### 2.2　动作空间

追踪者的运动控制方式为速度大小控制与方向控制,本文设定在环境中追踪者各时刻的输出动作为,包括速度大小和角度的变化量.如式(3)所示,追踪者下一时刻的空间位置、速度大小和运动方向可更新为:

$$\begin{cases}\theta_p^{t+1}=\theta_p^t+a_t[0]\times\Delta\theta_{\max}\\ v_p^{t+1}=v_p^t+a_t[1]\times\Delta v_{\max}\\ x_p^{t+1}=x_p^t+v_p^{t+1}\times\cos\left(\theta_p^{t+1}\right)\\ y_p^{t+1}=y_p^t+v_p^{t+1}\times\sin\left(\theta_p^{t+1}\right)\end{cases} \tag{3}$$

式中:$a_t\in[-1,1]^2$ 为算法输出的归一化动作;$\Delta v_{\max}$ 为速度的最大变化量;$\Delta\theta_{\max}$ 为角度的最大变化量;$\theta_p^{t+1}$、$v_p^{t+1}$、$\left(x_p^{t+1},y_p^{t+1}\right)$ 分别表示追踪者在下一个时刻的运动方向角、速度大小和位置坐标.

### 2.3　奖励函数

不同于文献[14]采用的稀疏奖励机制,本文设计了一种连续奖励函数,细化追踪者每一步的奖励效果,有助于提高追踪者的学习效率,降低陷入局部最优的风险.

本文设计的奖励函数包括四个部分:终止奖惩、距离奖惩、角度奖惩、避障奖惩.

终止奖惩函数的数学表达式为:

$$r_T=-k_T \tag{4}$$

式中:$r_T$ 表示终止奖惩;$k_T$ 为正实数,即追踪者任务失败时施加惩罚.

距离奖惩函数的数学表达式为:

$$r_d=\begin{cases}\dfrac{k_1\times\left(d_{t+1}-d_t\right)}{d_{t+1}+9},\ d_{t+1}\geqslant 50\\ \dfrac{k_2\times\left(d_{t+1}-d_t\right)}{d_{t+1}+6},\ 30\leqslant d_{t+1}<50\\ \dfrac{k_3\times\left(d_{t+1}-d_t\right)}{d_{t+1}+3},\ 10\leqslant d_{t+1}<30\\ \dfrac{k_4\times\left(d_{t+1}-d_t\right)}{d_{t+1}+1},\ \text{其他}\end{cases} \tag{5}$$

式中:$r_{\mathrm{d}}$表示距离奖惩;$d_t$和$d_{t+1}$分别表示上一个时刻和当前时刻追踪者与逃避者之间的欧式距离;$k_1$、$k_2$、$k_3$、$k_4$均为距离奖惩的权重.

角度奖惩函数的数学表达式为:

$$r_\theta=\begin{cases}\dfrac{k_5}{\pi}\times\left(\dfrac{\pi}{2}-\alpha\right), & d_{t+1}\geqslant 50\\ \dfrac{k_6}{\pi}\times\left(\dfrac{\pi}{2}-\alpha\right), & 30\leqslant d_{t+1}<50\\ \dfrac{k_7}{\pi}\times\left(\dfrac{\pi}{2}-\alpha\right), & 10\leqslant d_{t+1}<30\\ \dfrac{k_8}{\pi}\times\left(\dfrac{\pi}{2}-\alpha\right), & \text{其他}\end{cases}\tag{6}$$

式中:$r_\theta$表示角度奖惩;$k_5$、$k_6$、$k_7$、$k_8$为角度奖惩权重.

避障奖惩函数的数学表达式为:

$$r=-k_{\mathrm{obs}}\times\sum_{i=1}^{9}\left(\frac{1}{d_i}-\frac{1}{r_{\mathrm{cap}}}\right),\ r_{\mathrm{obs}}=\begin{cases}r, & -r_{\mathrm{clip}}<r<0\\ -r_{\mathrm{clip}}, & r\leqslant -r_{\mathrm{clip}}\end{cases}\tag{7}$$

式中:$r_{\mathrm{obs}}$表示避障奖惩;$r_{\mathrm{clip}}$表示避障奖惩截断的最小值;$k_{\mathrm{obs}}$为正整数;$d_i$表示追踪者与障碍物的距离;$r_{\mathrm{cap}}$为距离上界阈值,当$d_i>r_{\mathrm{cap}}$时该障碍物对惩罚项贡献为零.

综合以上四种奖励函数,最终的奖励函数$r_{\mathrm{o}}$为:

$$r_{\mathrm{o}}=r_{\mathrm{T}}+r_{\mathrm{d}}+r_\theta+r_{\mathrm{obs}}\tag{8}$$

## 3 MPADDPG算法

针对DDPG算法单经验池模式下学习效率低,容易陷入局部最优解的问题,本文提出了MPADDPG算法,将单个经验池替换为多个经验池,并且利用注意力机制增强有效经验特征,提高算法的收敛速度.

### 3.1 DDPG算法

DDPG算法是Actor-Critic框架下的经典算法,算法通过这两个网络的协同工作,在连续状态空间、动作空间问题上实现高效的策略学习和优化.

在Critic网络训练阶段,首先计算出目标动作值和网络预测值,然后得到Critic损失值,最后利用梯度下降算法对Critic网络进行优化.目标动作值为

$$y_i=r(s_i,a_i)+\gamma Q_{\tilde{\omega}}\left[s_{i+1},\tilde{\pi}(s_{i+1},\tilde{\theta})+N(0,\sigma)\right]\tag{9}$$

式中:$y_i$表示目标动作值;$r(s_i,a_i)$表示即时奖励;$\gamma$表示折扣系数;$\tilde{\pi}\left(s_{i+1},\tilde{\theta}\right)$表示下一个状态下输出的目标动作;$\tilde{\theta}$表示目标Actor网络中的参数;$N(0,\sigma)$表示满足正态分布的平滑噪声;$\sigma$表示平滑噪声的方差;$Q_{\tilde{\omega}}\left[s_{i+1},\tilde{\pi}\left(s_{i+1},\tilde{\theta}\right)+N(0,\sigma)\right]$表示目标Critic网络中,下一个状态动作对的动作值.损失值的计算方式为:

$$L_{\mathrm{c}}=\frac{1}{N}\left[Q_{\mathrm{w}}(s_i,a_i)-y_i\right]\tag{10}$$

式中:$L_{\mathrm{c}}$表示Critic损失值;$N$表示每次更新时采样的数据总量;$Q_{\mathrm{w}}(s_i,a_i)$表示当前Critic网络中,当前状态动作对的动作值.

在Actor网络训练阶段,首先计算出当前状态下的预测动作,然后得到Actor网络损失值,最后利用梯度上升算法更新Actor网络参数.Actor网络损失值的计算方式为:

$$L_{\mathrm{a}}=-E\left[Q_{\mathrm{w}}\left[s,\pi(s_i,\theta)\right]\right]\tag{11}$$

式中:$L_{\mathrm{a}}$表示Actor网络损失值;$\pi(s_i,\theta)$表示当前状态下的预测动作;$\theta$表示Actor网络的参数;$Q_{\mathrm{w}}\left[s,\pi(s_i,\theta)\right]$表示新的当前状态动作对的动作值;$E[\cdot]$为期望运算符,表示对随机变量的加权平均.

完成Critic网络和Actor网络的更新后,利用软更新方式更新所有的目标网络,如式(12)所示.

$$\begin{cases}\theta=\tau\theta+(1-\tau)\tilde{\theta}\\ w=\tau w+(1-\tau)\tilde{w}\end{cases}\tag{12}$$

式中:$\tau$表示软更新的学习率;$w$表示Critic网络中的参数;$\tilde{\omega}$表示目标Critic网络中的参数.

### 3.2 注意力机制

注意力(Attention)机制具有增强神经网络信息处理的能力,其核心思想在于通过学习输入特征的权重分布,对原始特征进行加权处理,从而使模型能够聚焦于关键特征,提升学习的效率与准确性.

Attention机制的计算公式为:

$$\boldsymbol{E}'=\mathrm{Softmax}(\boldsymbol{WE})\cdot\boldsymbol{E}\tag{13}$$

式中:$\boldsymbol{E}'$表示新特征矩阵;$\boldsymbol{E}$表示为原特征矩阵;$\boldsymbol{W}$为表示为权重矩阵;Softmax(·)为激活函数.

本文设计了如图4所示的Attention网络结构,包含一个全连接层和一个Softmax激活函数.输入向量首先通过全连接层进行线性变换,以提取更深层次的特征信息.然后,经过Softmax激活函数处理,得到输入向量中各分量的权重$W_i$.这些权重反映了各分量在任务中的重要性.最后,将权重$W_i$与原始输入向量相乘,得到新的加权向量,从而强化了神经网络对有效特征的学习.

### 3.3 多经验池

MPADDPG算法中包括三个经验池:失败经验

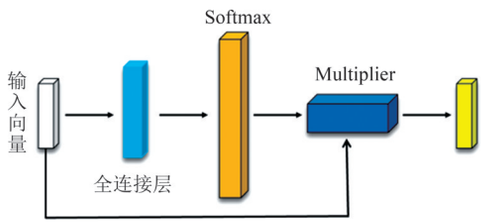



图 4　Attention 网络结构

Fig.4　Attention network structure

池、成功经验池和临时经验池. 失败经验池存储追踪者任务失败时产生的经验数据，设其容量大小为 $M_f$. 成功经验池存储追踪者任务成功时产生的经验数据,设其容量大小为 $M_s$. 而临时经验池则存储当前追踪者产生的临时经验数据，设其容量大小为 $M_t$. 当临时经验池溢出时，如果追踪者仍然没有到达终止状态，则认为过去的决策具有积极影响,将溢出数据存入成功经验池，然后根据最终任务的成败决定剩余所有数据的存储位置.

MPADDPG 算法更新时需要的数据分别从失败经验池与成功经验池中按照一定的比例采样. 这样即使追踪者陷入局部最优，仍然可以从成功经验池中获取任务成功的经验数据来学习，从而帮助追踪者快速脱离局部最优解. 假设追踪者每次从成功经验池中采样数据的比例为 $\beta$，随机采样的总数据量为 $B$，则从成功经验池与失败经验池中采样的数据量分别为:

$$\begin{cases} b_1 = \begin{cases} D_s, D_s \leqslant \beta B \\ \beta B, \text{其他} \end{cases} \\ b_2 = B - b_1 \end{cases} \tag{14}$$

式中:$b_1$ 表示从成功经验池中采样的数据量;$b_2$ 表示从失败经验池中采样的数据量;$D_s$ 表示当前成功经验池中的总数据量.

### 3.4　基于 Attention 机制的深度强化学习网络结构

MPADDPG 算法的网络架构沿袭了 DDPG 算法的设计范式，核心组件包括一个 Actor 网络及其对应的目标网络，以及一个 Critic 网络与其对应的目标网络. 在 Critic 网络的架构设计中，其输入整合了追踪器捕获的观测信息(维度为 20)与所执行的动作(维度为 2)，通过拼接操作形成一个 22 维的输入向量. 这一向量随后被送入两个连续的全连接层进行特征抽取.

第一个全连接层配置了 400 个神经元，旨在初步提炼输入数据的特征表示，其激活函数选用了 ReLU(rectified linear unit)，以增强网络的非线性映射能力并促进梯度传播.

第二个全连接层拥有 300 个神经元，进一步对前一层的输出进行深度特征提取，同样采用 ReLU 作为激活函数，以维持网络的非线性特性.

最终，经过这两层特征提取的向量被输入到一个输出层，该层为全连接结构，负责计算并输出对应的状态-动作值($Q$ 值)，其输出维度设定为 1，以适应值函数预测的需求. 值得注意的是，此输出层并未采用任何激活函数，以保持 $Q$ 值的线性输出特性，便于后续的决策与优化过程. 具体的网络结构如图 5 所示.

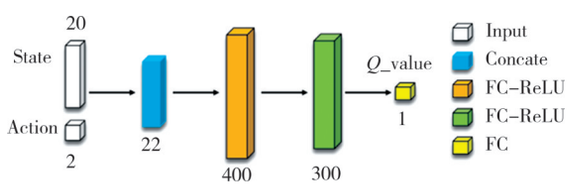



图 5　Critic 网络结构

Fig.5　Critic network architecture

在 MPADDPG 的 Actor 网络结构中，本文引入了一层 Attention 网络来强化输入向量的重要特征. 将 Attention 网络结构添加到图 6 中虚线框的位置，Actor 输入向量为追踪者观测到的状态信息 State，经过 Attention 网络对状态中的各个分量施以不同的权重以捕捉其中重要的信息. 然后依次经过两个全连接网络进行特征提取，第一个全连接网络的神经元数量为 400 个，第二个为 300 个，激活函数均采用 ReLU. 最后将提取到的特征信息送入最后一个全连接网络. 由于每个动作的输出在[-1,1]之间，因此最后一个全连接网络的激活函数采用 Tanh，如图 6 所示.

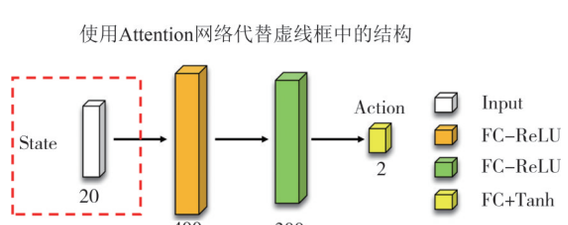



图 6　Actor 网络结构

Fig.6　Actor network architecture

本文采用 MPADDPG 算法实现追踪者对逃避者的实时追踪，并躲避环境中的障碍物. 追踪者在环境中运动时不仅可以获取自身与逃避者的状态信息以及相对位置信息，还可以利用自身搭载的传感器感知周围部分环境信息，然后将这些信息堆叠在一起，构成观测信息. 将观测信息送入追踪者的 Actor 网络中，输出动作，指导追踪者运动. 同时环境反馈追踪者回报以及下一个状态，并判断该状态是否为终止

状态.将一组交互的经验数据送入临时经验池中,根据临时经验池当前数据量以及追踪者当前状态信息判断是否要将经验数据转移到成功经验池或失败经验池中.在每一个时间步,追踪者都会从成功经验池和失败经验池中按照预先设定的比例随机采样一批数据进行训练.当追踪者进入终止状态时,结束本回合,并重新进入下一个回合的学习.具体的算法流程如图7所示.

## 4 实验结果

本文在Python环境下,依托PyTorch 1.10与CUDA 10.2计算框架,设计并实现了MPADDPG算法的网络架构.在参数优化方面,Critic网络与Actor网络均采用了Adam优化器,学习率设定为0.001,以确保网络训练的高效与稳定.

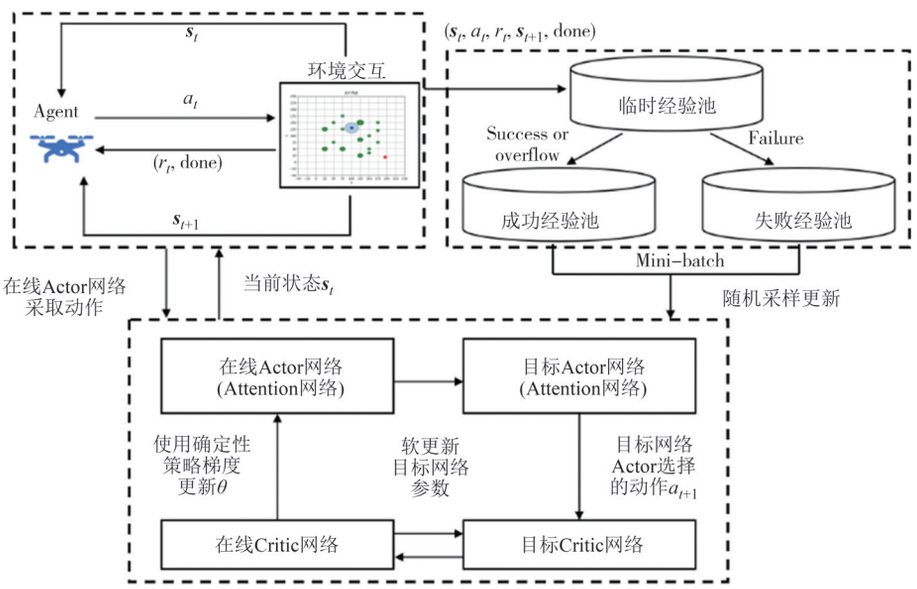



图7 MPADDPG算法流程

Fig.7 MPADDPG algorithm flow

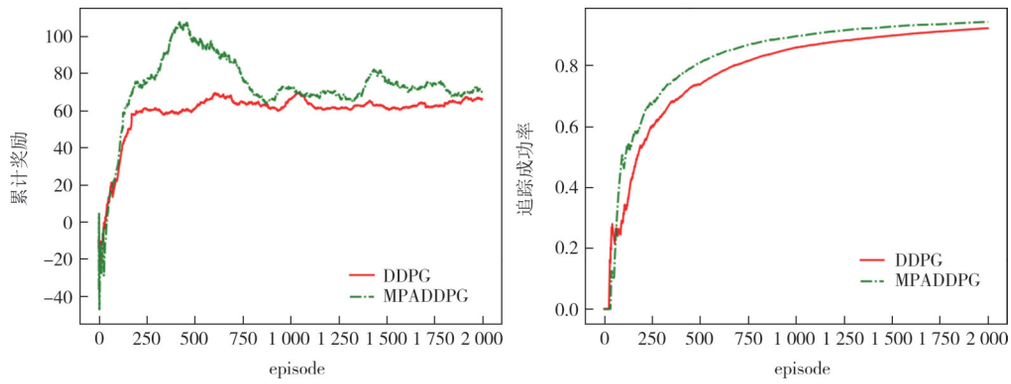
### 4.1 MPADDPG算法验证

仿真环境为一块方形区域,为了验证追踪者在复杂环境下的追踪决策能力,环境中存在若干障碍物.逃避者具备在环境中做随机运动并且可以自主规避环境中的障碍物的能力.追踪者的目标在于实时追踪逃避者,同时自主躲避环境中的障碍物.当追踪者出现以下情况时,当前回合结束:与障碍物发生碰撞;跨越边界;达到最大时间步.每次环境初始化时,追踪者与逃避者的位置也会随机初始化,以增强任务的随机性.本节采用DDPG、MPADDPG算法进行测试,测试结果如图8~图10和表1所示.

图8(a)、(b)分别为DDPG算法和MPADDPG算法在训练时的累积奖励曲线与追踪成功率曲线.实验结果表明,两种算法的奖励曲线均可以收敛,追踪成功率均随训练时间增加.然而,MPADDPG算法因为引入了多经验池机制与注意力机制,在训练过程中能够更加高效地提取训练经验中的有效特征,具有更快的收敛速度与更高的追踪成功率.为进一步验证算法的实际性能,本文将训练好的两种算法分别在相同环境中测试1 000轮.图9和表1是测试时的追踪成功率.测试结果表明,两种算法的追踪成功率均在90%以上,但MPADDPG算法的追踪成功率显著高于DDPG算法.图10是两个算法在简单环境下的追踪路径,其中红色圆圈表示逃避者,蓝色圆圈表示追踪者,绿色圆圈表示障碍物,蓝色圆圈周围的青色区域表示感知范围.从追踪路径可以看出,MPADDPG算法能够有效规划平滑且高效的路径,在避让感知范围内障碍物的同时,迅速向逃避者方向靠近.因此,简单环境下的仿真结果说明,MPADDPG算法具有更快的收敛速度以及更高的追踪成功率.

### 4.2 MPADDPG算法泛化能力验证

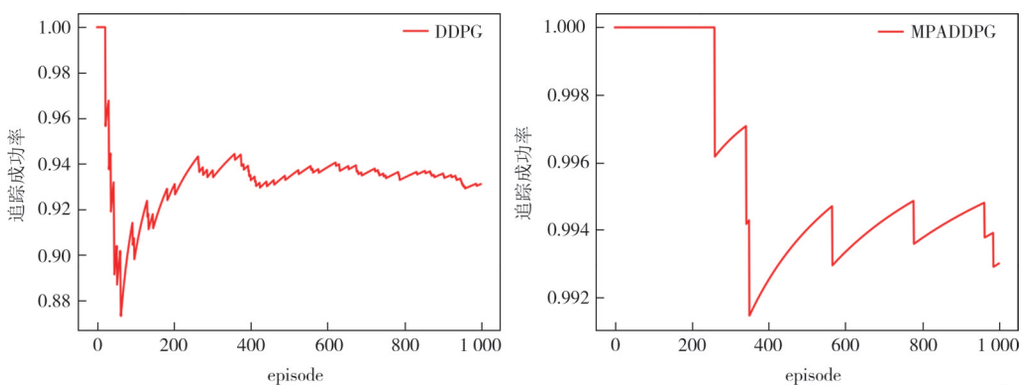
为了进一步验证训练出来的追踪者具备一定的环境泛化能力,本文将训练好的追踪者迁移至新的复杂环境中进行测试.该环境具有更高的障碍物密度,避障与追踪难度显著增加,且追踪者未在该环境下进行过训练.测试结果如图11、图12及表2所示.




(a)累积奖励曲线　　(b)追踪成功率曲线

图 8　简单环境下训练结果

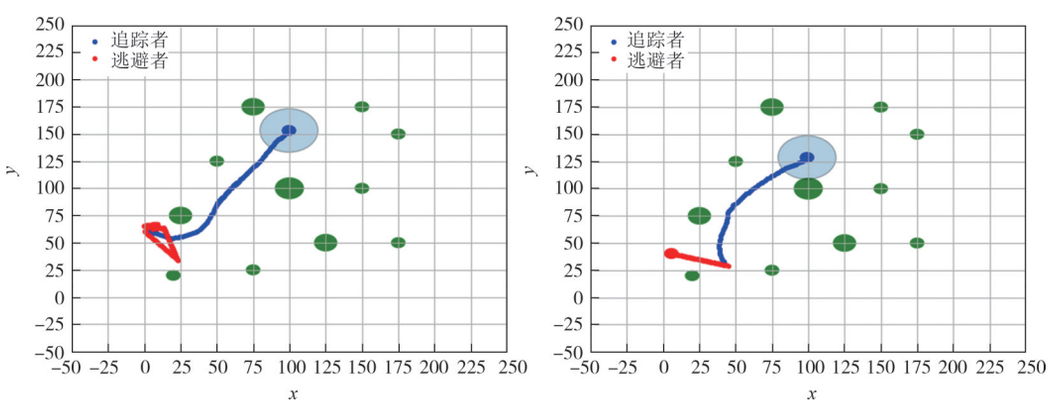
Fig.8　Training results in a simple environment




(a)DDPG算法　　(b)MPADDPG算法

图 9　简单环境追踪成功率曲线

Fig.9　Simple environment tracking success rate curves




(a)DDPG算法　　(b)MPADDPG算法

图 10　简单环境追踪路径

Fig.10　Simple environment tracking path

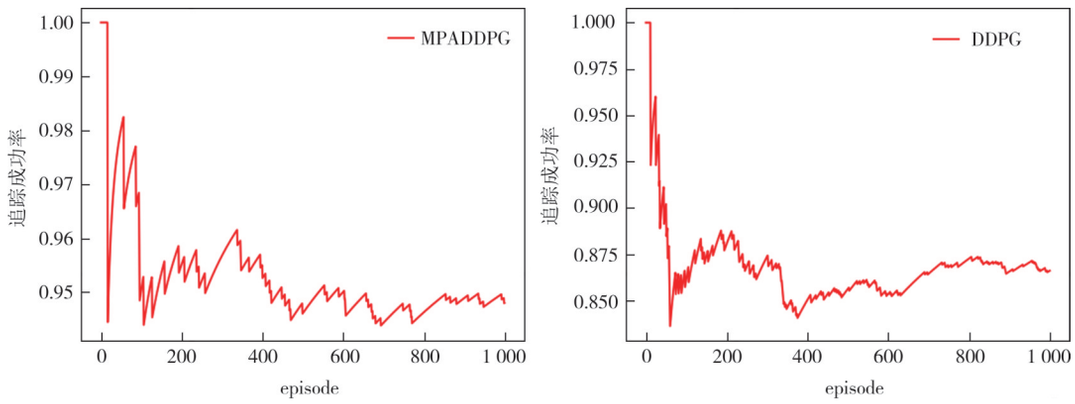
**表 1　简单环境下追踪成功率**

**Tab.1　Track success rates in simple environments**

| 算法 | 成功率 |
|---|---|
| MPADDPG | 0.993 |
| DDPG | 0.930 |

图 11 和表 2 展示了两个算法在复杂环境中分别测试 1 000 轮后的追踪成功率. 实验结果表明，MPADDPG 算法由于具备环境感知能力，在训练过程中侧重学习的是避障策略而非环境信息，因此在陌生的复杂环境下的追踪成功率仍然在 90% 以上，显著高于 DDPG 算法 86% 的追踪成功率. 图 12 进一步对比了两种算法在复杂环境中的追踪路径，结果显示，尽管两种算法均能完成任务，但 MPADDPG 算法的路径规划在避障效率和路径平滑度方面均优于 DDPG 算法，

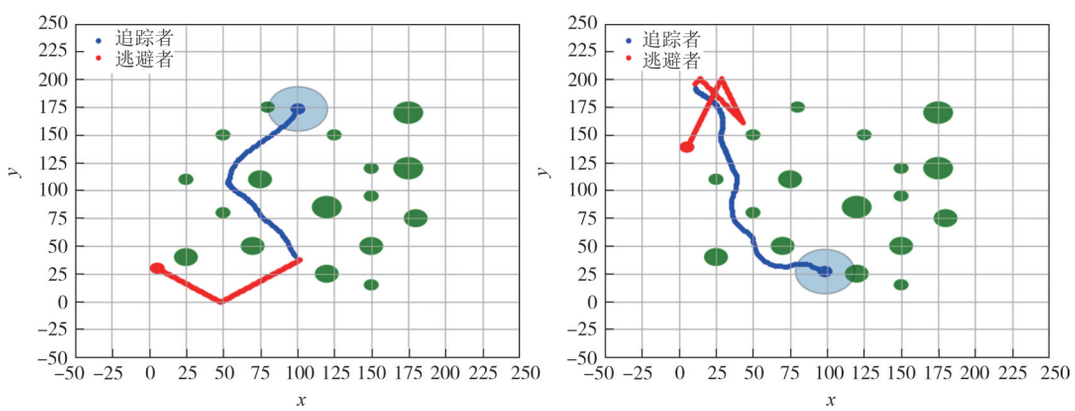
展现了更强的环境适应能力.在复杂环境下的仿真结果表明,MPADDPG算法较DDPG算法具备更好的环境泛化能力.因此,训练MPADDPG算法时只需要让追踪者在简单的环境中交互即可,缩短了训练时间.




(a)MPADDPG算法 (b)DDPG算法

图 11 复杂环境追踪成功率曲线

Fig.11 Track success curves for complex environments




(a)MPADDPG算法 (b)DDPG算法

图 12 复杂环境追踪路径

Fig.12 Complex environment tracking path

**表 2 复杂环境下追踪成功率**

**Tab.2 Tracking success rates in complex environments**

| 算法 | 成功率 |
|---|---|
| MPADDPG | 0.95 |
| DDPG | 0.86 |

## 5 结 论

针对复杂环境下无人机动态目标智能追踪决策问题,本文提出了一种改进的深度强化学习算法MPADDPG,实现无人机自主避障以及对动态目标的实时追踪.该算法在传统DDPG算法的基础上,结合了Attention机制,并加入了多经验池机制,以提升无人机在复杂环境下的自主避障与动态目标实时追踪能力.为增强仿真环境的真实性与实用性,本文设定无人机的状态和动作均为连续变化,并赋予无人机局部环境感知能力,使其能够实时获取周围环境信息.

仿真结果表明,MPADDPG与DDPG两种算法均具备稳定的动态目标智能追踪决策的能力,但是MPADDPG具有更快的收敛速度和更高的追踪成功率.此外,MPADDPG算法训练后的追踪者可以适应更加复杂的环境,表现出更加优秀的环境泛化能力.

## 参考文献

[1] 江未来,吴俊,王耀南. 基于元强化学习的无人机自主避障与目标追踪[J]. 湖南大学学报(自然科学版),2022,49(6):101-109.
JIANG W L, WU J, WANG Y N. Autonomous obstacle avoidance and target tracking of UAV based on meta-reinforcement learning[J]. Journal of Hunan University (Natural Sciences), 2022, 49(6):101-109.(in Chinese)

[2] 吴勇, 彭辉, 熊峰钥. 结合类脑导航的强化学习无人机自主导航[J]. 计算机测量与控制, 2024, 32(7):225-231.
WU Y, PENG H, XIONG F Y. Reinforcement learning algorithms combined with brain-inspired navigation [J]. Computer Measurement & Control, 2024, 32(7): 225-231. (in Chinese)

[3] 范芮滔,刘昊,程明,等.博弈环境下的多无人机系统协同路径规划[J/OL].北京航空航天大学学报,1-10[2024-11-28]. https://doi.org/10.13700/j.bh.1001-5965.2024.0481.
FAN R T, LIU H, CHENG M, et al. Game environment of multiple unmanned aerial vehicle system coordinated path planning [J/OL]. Journal of Beijing University of Aeronautics and Astronautics, 1-10 [2024-11-28]. https://doi.org/10.13700/j.bh.1001-5965.2024.0481. (in Chinese)

[4] 蒋坤, 操菁瑜, 柳文章, 等. 移动机器人导航与对抗控制的强化学习方法研究[J]. 控制理论与应用, 2025, 42(9):1757-1765.
JIANG K, CAO J Y, LIU W Z, et al. Research on reinforcement learning methods for navigation and adversarial control in mobile robots[J]. Control Theory & Applications, 2025, 42(9):1757-1765. (in Chinese)

[5] 娄智波, 彭越, 辛凯. 基于改进型Q-Learning算法的路径规划系统研究[J]. 计算机与数字工程, 2024, 52(8):2312-2316.
LOU Z B, PENG Y, XIN K. Research on path planning system based on improved Q-Learning algorithm [J]. Computer and Digital Engineering, 2024, 52(8): 2312-2316. (in Chinese)

[6] 王杰, 王高攀, 孙天杨. 基于DQN算法的无人机校园安全监控路径自动规划模型[J]. 自动化与仪器仪表, 2024(4): 193-196.
WANG J, WANG G P, SUN T Y. Automatic planning model of UAV campus security monitoring path based on DQN algorithm[J]. Automation & Instrumentation, 2024(4):193-196. (in Chinese)

[7] 沈骁, 赵彤洲. 基于DDQN的无人机区域覆盖路径规划策略[J]. 电子测量技术, 2023, 46(14):30-36.
SHEN X, ZHAO T Z. UAV regional coverage path planning strategy based on DDQN [J]. Electronic Measurement Technology, 2023, 46(14):30-36. (in Chinese)

[8] 赵启, 甄子洋, 龚华军, 等. 基于D3QN的无人机编队控制技术[J]. 北京航空航天大学学报, 2023, 49(8):2137-2146.
ZHAO Q, ZHEN Z Y, GONG H J, et al. UAV formation control based on dueling double DQN[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(8): 2137-2146. (in Chinese)

[9] 毕文豪,段晓波. 基于深度强化学习的无人机航路规划算法研究[J]. 航空科学技术,2023,34(12):118-124.
BI W H, DUAN X B. Research on UAV path planning method based on deep reinforcement learning[J]. Aeronautical Science & Technology, 2023, 34(12): 118-124. (in Chinese)

[10] 王子怡, 刘喆, 李玉婉, 等. 物联网背景下基于人工势场法和双目视觉的城市无人机配送补货研究[J]. 信息记录材料, 2024, 25(3):240-242.
WANG Z Y, LIU Z, LI Y W, et al. Research on replenishment of urban UAV distribution based on artificial potential field method and binocular vision under the background of Internet of Things[J]. Information Recording Materials, 2024, 25(3):240-242. (in Chinese)

[11] 展望晨, 郭乐江, 许世佳, 等. 基于贪婪DDPG的无人机智能避障算法[J]. 空天预警研究学报, 2024, 38(5):342-346.
ZHAN W C, GUO L J, XU S J, et al. UAV Intelligent obstacle avoidance algorithm based on greedy DDPG[J]. Journal of Air & Space Early Warning Research, 2024, 38(5): 342-346. (in Chinese)

[12] 荣垂霆, 朱恒伟, 张宾, 等. 基于深度强化学习的移动机器人路径规划研究[J]. 现代信息科技, 2024, 8(16):60-63.
RONG C T, ZHU H W, ZHANG B, et al. Research on path planning of mobile robots based on deep reinforcement learning [J]. Modern Informationn Technology, 2024, 8(16):60-63. (in Chinese)

[13] 赵学健, 叶昊, 李豪, 等. 基于改进DDPG的多AGV路径规划算法[J]. 计算机科学, 2025, 52(6):306-315.
ZHAO X J, YE H, LI H, et al. Multi-AGV path planning algorithm based on improved DDPG [J]. Computer Science, 2025, 52(6):306-315. (in Chinese)

[14] 牟文心, 时宏伟. 基于改进TD3算法的无人机轨迹规划[J]. 计算机系统应用, 2024, 33(12):197-209.
MOU W X, SHI H W. UAV trajectory planning based on improved TD3 algorithm [J]. Computer Systems & Applications, 2024, 33(12): 197-209. (in Chinese)

[15] 陈卓, 姜伟豪, 杜军威. 基于策略记忆的深度强化学习序列推荐算法研究[J]. 湖南大学学报(自然科学版), 2022, 49(8): 208-216.
CHEN Z, JIANG W H, DU J W. Research on deep reinforcement learning sequential recommendation algorithm based on policy memory [J]. Journal of Hunan University (Natural Sciences), 2022, 49(8):208-216. (in Chinese)