文章编号:1671-4598(2024)07-0225-07 DOI:10.16526/j.cnki.11-4762/tp.2024.07.033 中图分类号:TP181;V249 文献标识码:A

结合类脑导航的强化学习无人机自主导航

吴 勇,彭 辉,熊峰钥
(成都信息工程大学 软件工程学院,成都 610228)

摘要:针对无人机自主导航常用的端到端强化学习方法存在训练效率低、泛化能力和通用性差等问题,引入了类脑导航模型,基于长短时记忆(LSTM)神经网络构建了类脑细胞导航模型,通过整合编码无人机智能体的自运动信息,实现了网格细胞和头朝向细胞的编码,进一步将这些信息作为深度强化学习算法 D3QN 的状态补充表示;通过在 AirSim 仿真环境的实验表明,类脑导航模型的引入能够有效提高算法的训练能力和无人机智能体的导航性能,相较于原 D3QN 算法,首次目标固定情况下,到达目标成功率提升了 2.54%,达到了 97.11%;而在目标改变后继续训练的情况下,到达目标成功率为 99.45%,而 D3QN 仅为 11.46%,未能找到新的目标点;表明算法的泛化能力得到有效提升。

关键词:无人机;深度强化学习;类脑导航;D3QN;自主导航

Reinforcement Learning Algorithms Combined with Brain-Inspired Navigation

WU Yong, PENG Hui, XIONG Fengyue
(School of Software Engineering, Chengdu University of Information Technology, Chengdu 610228, China)

Abstract: In response to the low training efficiency, poor generalization ability, and universality of widely used end-to-end reinforcement learning methods for autonomous navigation of UAV, a brain-inspired navigation model is introduced. Based on the long short-term memory (LSTM) neural network, a brain-inspired cell navigation model is constructed, the self-motion information of the UAV intelligent agent is integrated to encode grid cells and head direction cells, further supplement this information as the state of the deep reinforcement learning algorithm D3QN. The experiments in AirSim simulation environment show that the introduction of the brain-inspired navigation model can effectively improve the training ability of the algorithm and the navigation performance of the UAV intelligent agent. Compared with the original D3QN algorithm, the success rate of reaching the target is increased by 2.54% to 97.11% with the target first fixed. the success rate of reaching the target is 99.45% with the target continued to train after changed. The new target point misses with the success rate of the D3QN of only 11.46%. This indicates that the algorithm effectively improves generalization abilities.

Keywords: UAV; deep reinforcement learning; brain-inspired navigation; D3QN; autonomous navigation

0 引言

随着机器人技术的不断进步和成本的降低,搭载自主智能系统的无人机开始应用于各个领域,如农业、物流、交通监管和救灾等[1]。无人机为人类提供了高效、安全、便捷的服务。尤其在监测、侦查、通信等方面,无人机具有显著优势。然而,这些应用场景的多样性和复杂性要求无人机具备高度自主导航能力。因此,无人机自主导航技术的研究越来越受到学术界和工业界的关注。

无人机自主导航相关研究中,应用比较广泛的是同时定位与地图构建(SLAM, simultaneous localization and mapping)技术,其利用视觉里程计等机载传感器构建地图并定位,是路径规划算法的基础,根据 SLAM 建立的地图和定位信息,使用路径规划和避障相关算法,如传统启发式算法 A*、人工势场法等,以及智能优化算法如遗传算法、粒子群优化算法等[2]。

而随着深度强化学习(DRL, deep reinforcement learning)的研究和发展,DRL 算法在无人机自主导航中取得了显著的成果[3-4]。DRL 结合了深度学习和强化学习,可以直接从高维输入如图像中学习控制策略,无需精确定位和地图,有很高的自适应性,能够让无人机在复杂环境中实现端到端的学习,通过与环境的交互,无人机可以在训练过程中学习到有效的导航策略。然而,DRL 仍然面临着训练

收稿日期:2023-07-03; **修回日期**:2023-08-10。
基金项目:四川省科技计划资助项目(2019YJ0356)。
作者简介:吴 勇(1999-),男,硕士研究生。
通讯作者:彭 辉(1975-),男,博士,副教授。
引用格式: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.

样本不足、训练时间过长等问题。

类脑导航作为一种借鉴生物神经系统的导航方法，提供了一种全新的视角和思路。本文通过研究生物神经系统的导航原理，将其应用于无人机自主导航问题。利用类脑导航的生物启示特性，可以有效降低计算复杂度，提高实时性和鲁棒性。同时，结合深度强化学习算法，进一步优化无人机在复杂环境下的导航性能，提高其自适应能力和泛化能力。

深度强化学习 DRL 作为机器学习的子领域，其核心是训练智能体，在与环境交互的同时，通过最大化长期奖励依次做出决策。强化学习的基本数学框架是马尔可夫决策过程（MDP，markov decision process），Wang 等人[5]将无人机导航问题建模为部分可观测马尔可夫决策过程（POMDP，partially observable Markov decision process），改进了递归确定性策略梯度算法（RDPG，recurrent deterministic policy gradient），并提出在线 DRL 算法 Fast-RDPG 来完成无人机大规模复杂环境中的导航，但对奖励函数的设计要求较高。Shin 等人[6]设计了离散动作和连续动作的强化学习算法，并将 U-net 分割模型加入演员－评论家（Actor-Critic）结构，测试得到 D3QN 算法和 ACKTR 算法分别在离散和连续动作空间中表现较优。MAW[7]提出了一种混合路径规划算法，通过改进的随时动态 A＊（iADA＊，improved anytime dynamic A＊）算法求解全局路径规划，再使用强化学习算法进行路径点之间的局部规划，其比较了 DQN 和 DDPG 算法，发现 DQN 表现更好。Tajmihir[8]同样将 D3QN 算法用于无人机规避决策的导航控制问题，训练后能根据视觉信息做出可靠的规避控制决策。

在过去的二十多年里，生物和神经科学一直为人工智能提供灵感和指导，Tolman[9]提出生物大脑中存在反映环境结构的认知地图，后来，Behrens 等[10]深入探讨了认知地图，强调将知识组织成认知地图对灵活行为的重要性，以及与强化学习的共性。王继茹[11]则建立了认知地图构建系统并用于移动机器人的自主目标搜索。此外，杨闯等人[12-13]概述了类脑导航技术初步形成了环境感知、空间认知与自主路径规划决策一体化的端到端特征。同时有学者研究基于生物大脑导航机理启发的 SLAM 系统[14-15]。

而在与强化学习结合相关方面，谷歌 DeepMind 团队[16]利用网格细胞设计开发了深度强化学习智能体，证明了把循环神经网络（RNN，recurrent neural network）用于空间的定位和导航时，隐节点的物理意义类似于大脑的位置细胞、网格细胞、边界细胞等导航细胞。Stachenfeld K[17]从强化学习的角度来理解海马体表征方式，并表示预测性表征方式能最大化未来回报。Matthew Bot-vinick[18]对深度强化学习及其神经科学意义进行了概述，认为深度强化学习可能为神经科学研究提供重要作用。SUHAIMI A[19]同时进行了生物和深度强化学习实验，在相同任务中训练并比较了 DRL 智能体和小鼠各自神经网络中的表征学习过程后，发现相似的神经活动，即小鼠后顶叶皮层与人工神经网络具有相似的表征。

本文为了改善无人机在未知环境中的导航能力，基于卷积神经网络构建类脑导航模型，通过对无人机自运动信息进行编码整合后，将其作为 D3QN 的状态表示，完成两者的结合，最后通过仿真实验验证了方法的有效性。

1 模型设计

1.1 类脑导航模型

现有类脑导航模型的构建理论基础来源于海马体中的空间表征细胞，通过编码自运动信息并产生放电活动，从而表征环境。现阶段的表征模型主要为网格细胞、头朝向细胞和位置细胞。

本文通过训练一个循环神经网络来进行自运动信息的编码，并分别使用二维各向同性的高斯分布和 Von-Mises 分布来模拟网格细胞和头朝向细胞的放电激活状态：

$$c_i = \frac{e^{-\|\vec{x}-\mu_i\|^2/2\sigma^2}}{\sum_{j=1}^{N} e^{-\|\vec{x}-\mu_j\|^2/2\sigma^2}} \tag{1}$$

网格细胞放电激活公式如（1）所示，c_i 表示第 i 个网格细胞的活动，$\vec{x}$ 表示当前位置，u 表示按一定分布排列的 N 个二维向量，σ 为位置尺度单元。

$$h_i = \frac{e^{\kappa\cos(\varphi-\nu_i)}}{\sum_{j=1}^{M} e^{\kappa\cos(\varphi-\nu_j)}} \tag{2}$$

头朝向细胞放电激活公式如（2）所示，h_i 表示第 i 个头朝向细胞的活动，φ 表示当前方向角，ν 表示各方向中心角度，κ 代表方位角集中程度。

网格细胞在动物大脑中呈现规则的空间响应模式，被认为能够衡量动物在环境中的位移距离和方向，为认知地图提供度量标准[11]。可用 3 个参数来描述网格细胞的空间放电模式：空间尺度、方向和相位，分别表示相邻细胞间距离、细胞中心连线与固定方向的夹角和细胞与固定点的最短距离。为简化模型，方向设置为 60°以符合六边形阵列排列[20]，相位设置为 0，空间尺度则根据环境大小调整，如图 1 所示。

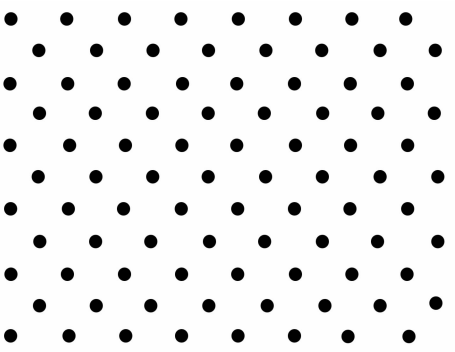

图 1 六边形网格分布

而头朝向细胞通过整合内源信息和外源信息中的角速度和方向角等信息，对当前方位进行估计，通过对特定方向角放电表示，且这种表示不受动物在环境中的位置影响。

最后通过循环神经网络进行监督学习，如图 2 所示。基于长短期记忆网络（LSTM，long short term memory）构建类脑导航模型，模型的输入为无人机的速度 v_t、方向角的正弦值 $\sin(\varphi_t)$ 和余弦值 $\cos(\varphi_t)$，输出为网格细胞和头朝向细胞的。其中，LSTM 的初始单元状态 $\overrightarrow{c_0}$ 和隐藏状态 $\overrightarrow{h_0}$，分别由初始地面坐标和方向计算得到，LSTM 输出 g_t，再通过 SoftMax 函数得到当前时刻网格细胞和头朝向细胞的活动。

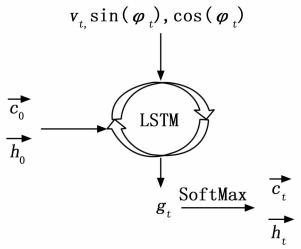

图 2　监督学习网络结构

对该网络模型进行训练，从而预测每个时间步长处的网格细胞和头朝向细胞激活情况，即 $\overrightarrow{c_t}$ 和 $\overrightarrow{h_t}$。在训练期间网络参数通过最小化预测值和目标值间的交叉熵来训练，损失函数如下：

$$L(\overrightarrow{c},\overrightarrow{h}) = -\sum_{i=1}^{N} c_i \log(\hat{c_i}) - \sum_{j=1}^{M} h_i \log(\hat{h_i}) \tag{3}$$

1.2　深度强化学习

强化学习主要包括智能体（Agent）、环境状态（State）和奖励函数（Reward），其对应马尔可夫决策过程（MDP），可用元组 (S,A,P,R,γ) 表示，S 为状态空间，A 为动作空间，P 为状态转移概率，R 为奖励函数，γ 为奖励折扣因子。

智能体通过与环境交互，在每个时刻 t，观察环境得到状态 $s(t)$ 和奖励 $r(t)$，然后根据策略 $\pi(a \mid s)$ 选择动作 $a(t)$，目的是学习到更好的策略 π，使得未来的累积奖励最大化。通常用状态 (s) 或者状态－动作 (s,a) 的价值函数来评估智能体的期望回报，对应的价值函数可以用贝尔曼方程表示：

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s') \tag{4}$$

$$V_\pi(s) = \sum_a \pi(a \mid s) Q_\pi(s,a) \tag{5}$$

$Q_\pi(s,a)$ 为状态－动作值函数，式中 R_s^a 为状态 s 下执行动作 a 的奖励；γ 为奖励折扣因子；$P_{ss'}^a$ 为状态转移概率；$V_\pi(s')$ 为后续状态的状态值函数。状态值函数 $V_\pi(s)$ 表示状态 s 时，所有状态－动作值函数 $Q_\pi(s,a)$ 在策略 π 下的期望。

由公式（4）和（5）可得

$$V_\pi(s) = \sum_a \pi(a \mid s) \left[R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s') \right] \tag{6}$$

为了最大化奖励期望，Q 学习通过时间差分（TD，temporal-difference）来得到最佳动作值的估计，而为了应对复杂的高维状态空间，使用深度神经网络来求解值函数的 DQN[21]（Deep Q Network）成为 DRL 的经典算法之一。

本文使用的 D3QN 算法，是 Double DQN[22] 和 Dueling DQN[23] 两个改进算法的结合，如图 3 所示。

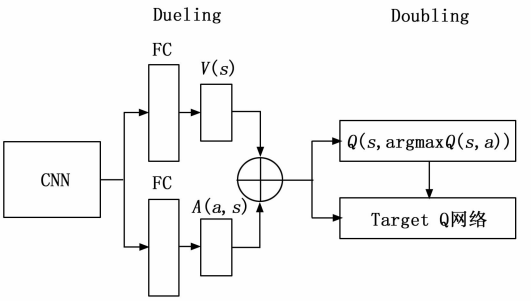

图 3　D3QN 结构

Dueling DQN 用两个独立的子网络，使用相同的卷积神经网络结构（CNN，convolutional neural network），经过全连接层（FC，fully connected）后，将 Q 函数拆解成状态值函数 $V(s)$ 和优势函数 $A(s,a)$，如式（7），能够提高性能和学习速度；Double DQN 通过使用两个独立神经网络来估计 Q 函数，一个用于选取最优动作，另一个 Target Q 网络保存策略网络的历史状态，并定期更新，用于评估动作的价值，能够减少过估计问题，提升算法的稳定性。

$$Q(s,a) = V(s) + A(s,a) - \frac{1}{A} \sum_{a'} A(s,a') \tag{7}$$

同时可以定义 $D3QN$ 的目标值为：

$$\Gamma D3QN_t = r_{t+1} + \gamma Q(s_{t+1}, \text{argmax}(s_{t+1}, a_t)) \tag{8}$$

其对应的损失函数：

$$Loss = \Gamma_t^{D3QN} - Q(s_t, a_t) \tag{9}$$

2　结合类脑导航的强化学习算法

无人机通过基础的端到端深度强化学习，使用传感器获取环境信息，训练得到控制决策，进行自主式探索导航；而类脑导航模型则在感知环境端进行信息加工，仿造生物大脑对自运动信息编码，得到有利于导航行为的环境信息，帮助强化学习算法更好的学习。

接下来将对深度强化学习算法所需的状态空间、动作空间和奖励函数进行设计。

2.1　状态空间

状态空间是智能体在强化学习训练过程中所使用的状态集合，代表了智能体所感知到的环境状态及其动态变化，为算法生成决策和评估奖励提供依据。本文设计的状态空间包括：

1）深度相机获取 RGB-D 深度图（Depth Map），图片大小为 108×192，具体通过将距离值归一化到 0～255 之间，对应黑色到白色，呈现类似灰度图效果，距离远近用灰度的深浅表示。

2）根据无人机运动时的速度和方向角计算得到类脑细胞模型编码；使用 1.1 节训练好的的 LSTM 模型计算。

最终得到的状态是多模态类型的，其包含一个二维的

图像和两个一维向量，还有两个标量：距离目标点的距离和与目标的方向角度。而为保证状态空间形式统一，需对二维图像进行特征提取，然后与其他向量以及标量进行拼接，最后才能用于强化学习算法训练。

2.2 动作空间

动作空间分为离散动作和连续动作，无人机智能体通过执行动作改变环境，得到不同的状态，从而获得对应奖励反馈。因连续动作空间带来的复杂解空间和维度诅咒等问题，此处选择离散动作，具体有 7 个离散动作，2 m/s 和 4 m/s 速度的前进动作，1 m/s 的后退动作，2 m/s 的上升以及下降，最后是 30 度的左转和右转动作。

2.3 奖励函数

奖励函数又称回报函数，作为智能体在与环境交互过程中得到的反馈信号，帮助改进策略。在导航任务中，主线奖励为无人机智能体到达终点时获得正向奖励，而在碰到障碍物或出界时给予负奖励，但在遇到智能体探索环境不足，难以到达目标点时，算法将收敛很慢甚至无法收敛。为缓解奖励稀疏问题，运用奖励塑形[24]（Reward Shaping）设计日常奖励，根据当前时刻与上一时刻目标相对距离计算得到，具体如下：

$$\begin{cases} r = +50 & \text{到达目标} \\ r = -50 & \text{碰撞 / 出界} \\ r = (S_{dist(t-1)} - S_{dist(t)}) * c \end{cases} \tag{10}$$

$S_{dist(t)}$ 表示 t 时刻与目标点的相对距离，c 为折扣超参数。解释为接近目标给予正奖励，远离目标给予负奖励，通过计算上一时刻的相对距离与当前时刻的相对距离差值来表达。

在训练过程中，为了防止无人机智能体进行无效的徘徊、转圈等重复操作，根据到达目标点所需最小步数，设置了单个训练回合最大执行步数，超过该步数将直接结束回合并给予负奖励。

2.4 整体架构与流程

无人机智能体的网络架构如图 4 所示，主要分两部分，第一部分如图上方所示，无人机获取深度图像，经过四层卷积神经网络进行特征提取；第二部分如图下方所示，根据速度和角度信息，使用训练好的类脑细胞模型进行编码，再将得到的网格细胞编码经过两层卷积神经网络。最后将深度图像的特征、网格编码的特征和头朝向编码特征进行拼接，再经过全连接层，最后输出的是动作空间中的某一个动作。

强化学习的算法训练流程如下所示：

1:初始化策略网络 policy 和目标网络参数,经验池 buffer
2:for epoch=1 to Maxepoch do
3: 初始化环境状态 s_0 ,设置无人机初始位置 p_0
4: for step=0 to Maxstep do
5: 获取环境状态 s_0
6: 根据状态得到动作 $a_t = policy(s_t)$
7: 执行动作并获得奖励 ,下一状态 s_{t+1}

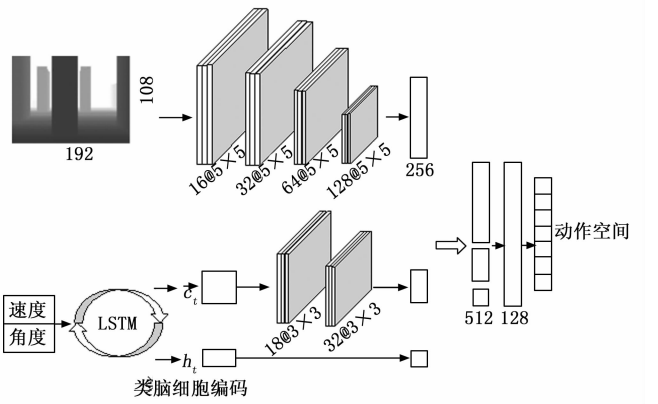

图 4 无人机智能体的网络架构

8: 将 (s_t, a_t, r_t, s_{t+1}) 存入经验池 buffer
9: if 到达更新步数 updatastep
10: 经验池采样并训练策略网络 policy
11: 计算策略网络损失值,更新参数
12: 定期复制策略网络参数到目标网络
13: end if
14: end for
15:end for

3 实验部分

本文使用微软针对无人机和自动驾驶汽车推出的 AirSim[25]仿真平台进行实验，其通过虚幻 4 引擎搭建物理环境，AirSim 提供相关接口，能够控制无人机进行移动和获取传感器数据。

算法程序使用 Python 语言编写，编程环境为 Python 3.7.2、Pytorch 1.8.2，计算机配置为 i7-8700@3.20 GHz，NVIDIA GeForce GTX 1070。

3.1 实验设置

为将类脑导航模型与深度强化学习相结合，首先需要对 1.1 节描述的类脑导航模型进行训练，在虚幻 4 中设立没有障碍物的空旷场地，大小为 150 m×150 m×30 m，无人机智能体在其中进行随机运动，模型中 LSTM 网络接收一个时间序列的数据，监督实验的标签数据通过真实坐标计算得。最后输出得到网格细胞编码和头朝向细胞编码，用于深度强化学习中的状态表示，实验相关参数如表 1：

表 1 监督学习算法参数

参数名称	数值
学习率	0.000 8
位置尺度单元 σ	3.5
方位角集中程度 κ	0.5
批次大小 batch size	1 024
LSTM 序列长度	100
LSTM 隐藏层单元	128

表中的位置尺度单元 σ 与环境大小成正比关系，方位角集中程度 κ 表示方位角的精准度，值越大，则对方向角度更

敏感。

然后在场景中加入长方体障碍物，并设置目标点，如图 5 所示，进行深度强化学习算法训练。

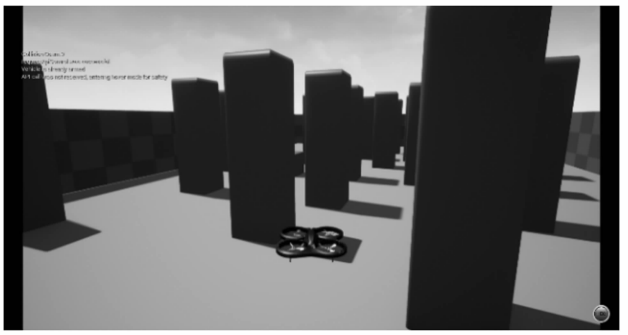

图 5　AirSim 中无人机训练场景

在环境中用三维坐标（x，y，z）表示位置，设置起点坐标为（46，10，0），第一个目标点为 G_1（100，104，15），第二个目标点为 G_2（7，88，15），强化学习算法训练中，其他的状态空间、动作空间和奖励函数与上文描述一致，训练相关参数如表 2 所示。

表 2　强化学习算法参数

参数名称	数值
学习率	0.000 4
奖励折扣因子 γ	0.95
训练轮数 epoch	3
每轮训练步数 Maxstep	20 000
经验池 buffer 大小	20 000
批次大小 batch size	64
每回合最大步数	250
仿真时间步长	1
距离奖励超参数 c	2.5
LSTM 隐藏层单元	128

表中奖励折扣因子 γ 表示智能体对未来一段时间累积回报奖励的看重比例，本实验无人机智能体最短在 40 步左右即可到达目标完成任务，相应的值设置较小，仅为 0.95；仿真时间步长则表示无人机的动作执行时间，反映无人机智能体的决策周期，设置较长的步长能够降低决策周期，从而简化任务复杂度，提升训练效率，若值过大，则会降低无人机智能体机动性，影响飞行避障。

然后通过 AirSim 的 Python 相关接口连接无人机模型，用于控制和获取传感器信息，最后进行强化学习算法训练。相关步骤如下：

使用 OpenAI 提供的 gym 库创建自定义环境，与 AirSim 环境对接；按照上文描述设置状态空间、动作空间和奖励函数；无人机智能体从起点开始，根据动作空间执行动作并得到对应奖励，状态为以下三者之一则视为回合结束：到达目标点、发生碰撞或出界、到达回合最大步数 250；回合结束后进行奖励统计和状态重置。总共训练 3 轮，每轮训练 20 000 步，回合数未设限制；具体算法流程见第 2.4 节。

3.2　结果与分析

本文算法是将类脑导航模型与 D3QN 算法相结合，为验证类脑导航模型的有效性，实验对比传统的 D3QN 算法，训练得到的奖励曲线如图 6 所示。强化学习中奖励值是无量纲的，只用于表示智能体行为的好坏，值越大，表现越好。由图可得，在第一轮（0～2 000 步）训练中，本算法得到的平均奖励值大于 D3QN 算法，后两轮训练得到的奖励相近，并一同收敛于 300 奖励值左右。

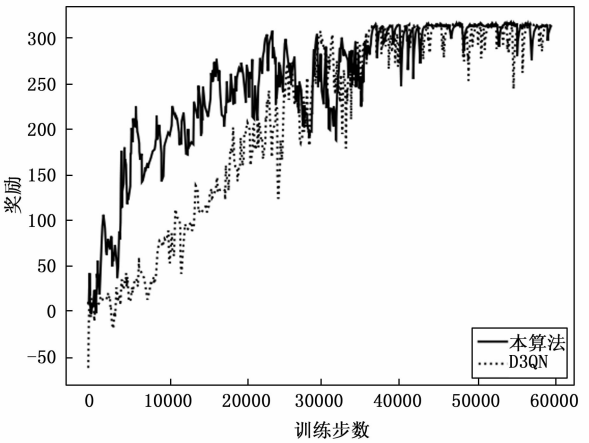

图 6　奖励曲线对比

其中目标设置为 G_1，通过保存奖励值最高的实验模型，进行实验验证得到的飞行路径如图 7 所示，可以看出，无人机智能体直接飞向目标，并顺利避开障碍物，最终得到的飞行路径也比较平滑，未出现急转弯或靠近障碍物等情况，两算法对比相差不大。

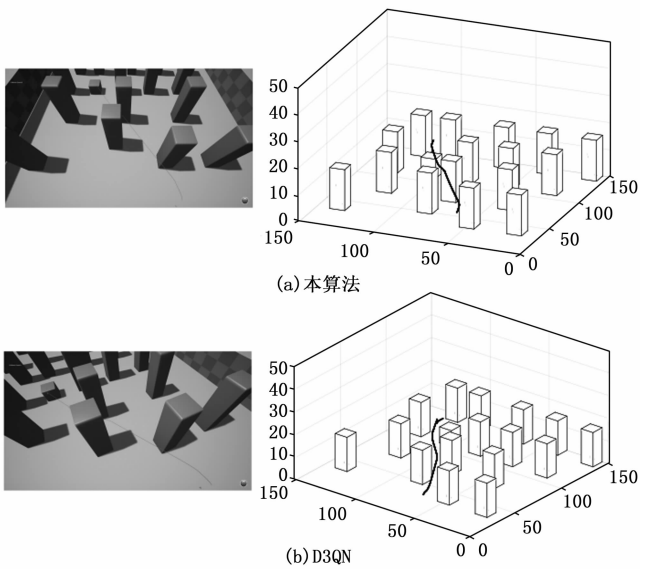

图 7　飞行路径示意图

为更好评价算法的导航的效果，使用了文献［26］提出的评价指标 SPL（SPL，success weighted by path length），即成功率和路径长度加权成功率，本文实验将步数 step 视作路径长度，计算公式如下：

$$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)} \tag{11}$$

式中，S_i 表示第 i 次实验是否成功到达目标，用 1 和 0 表示；l_i 表示第 i 次实验中无人机智能体到达目标需要的最少步数，本实验因起点和目标点固定，设置为 40 步；表示无人机实际到达目标所用步数。同时也记录了到达目标的成功率和与障碍物碰撞的碰撞率等数据，具体如表 3 所示。

表 3 实验评价指标对比

	成功率	碰撞率	SPL
Epoch 1			
本算法	6.16%	27.49%	2.42%
D3QN	2.68%	54.91%	0.90%
Epoch 2			
本算法	63.67%	4.30%	55.43%
D3QN	68.31%	2.46%	51.41%
Epoch 3			
本算法	97.11%	2.12%	79.19%
D3QN	94.57%	3.39%	77.67%

表中按照实验轮次（epoch）分开记录，整体来看，第一轮实验无人机并没有学习到较好的避障和导航策略，后两轮才学习到较优策略，并在最后一轮稳定下来，并且奖励值也收敛于同一位置。分析表中数据，能够发现本文算法收敛速度优于 D3QN，最后的指标数据也较好，但对比不明显。

为了验证类脑导航模型对目标点的导航能力提升效果，进行了泛化实验，保存上述训练完的实验模型，改变环境中的目标点为 G_2 后，重新加载模型进行训练，得到的奖励曲线如图 8 所示。因为是对上一实验模型的再训练，所以训练步数从 6 000 开始，由图可得，第一轮（6 000～8 000 步）训练，无人机智能体都因目标改变而无法得到较多奖励值，而到 9 000 步左右时，本文算法因找到新目标而迅速收敛至奖励值 250 左右，D3QN 算法直至结束仍然没有收敛。

同样保存奖励值最高的实验模型，进行实验验证，得到的飞行路径如图 9 所示，可以看出本算法最终能够找到目标，而 D3QN 算法仍然在旧目标和新目标之间徘徊，未能找到新的目标点。

与之对应的评价指标数据如表 4 所示，可以看出，经过学习后的无人机保留了避障策略的能力，碰撞率都比较低，并因为环境中目标点的改变，导致第一轮中，无人机到达新的目标成功率都为 0；而到了第二轮，本算法就能够到达目标并快速稳定到较高的奖励水平，D3QN 算法则因

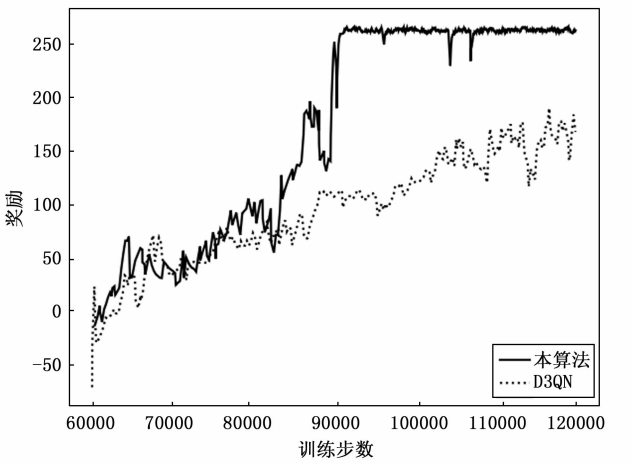

图 8 泛化实验奖励曲线对比

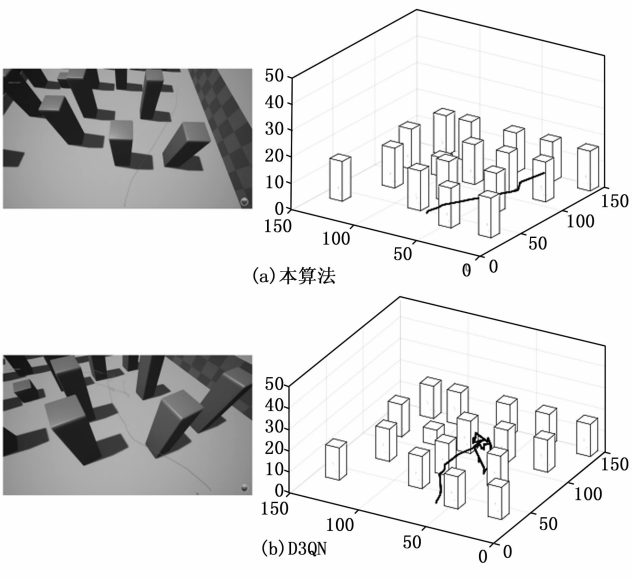

图 9 泛化实验的飞行路径

探索不足或是对上一目标过拟合，未能到达目标点，成功率依旧为 0，到了第三轮 D3QN 才开始偶尔到达目标，而本文算法已经收敛并趋于稳定。

表 4 泛化实验评价指标对比

	成功率	碰撞率	SPL
Epoch 1			
本算法	0%	6.00%	0%
D3QN	0%	2.19%	0%
Epoch 2			
本算法	86.55%	1.96%	86.06%
D3QN	0%	8.43%	0%
Epoch 3			
本算法	99.45%	0.14%	96.15%
D3QN	11.46%	18.75%	8.33%

4　结束语

针对无人机在未知的复杂环境中进行探索式导航飞行问题，本文将类脑导航模型与深度强化学习算法进行结合，通过类脑导航模型整合并编码无人机自运动信息，有效提升深度强化学习算法的训练能力和无人机智能体的导航能力。仿真实验表明，类脑导航细胞模型的加入，相较于原 D3QN 算法在目标固定情况下，到达目标成功率提升了 2.54%，达到了 97.11%；而在目标改变后继续训练的情况下，到达目标成功率为 99.45%，而 D3QN 仅为 11.46%，未能找到新的目标点。表明本文算法能够改善深度强化学习的泛化能力，为导航算法在不同环境之间迁移提供条件。

参考文献：

[1] AHMED F, MOHANTA J C, KESHARI A, et al. Re-cent advances in unmanned aerial vehicles: a re-view [J]. Arabian Journal for Science and Engineering, 2022, 47 (7): 7963 - 7984.

[2] ZHAO Y, ZHENG Z, LIU Y. Survey on computation-al-intelligence-based UAV path planning [J]. Knowledge-Based Systems, 2018, 158: 54 - 64.

[3] AZAR A T, KOUBAA A, ALI MOHAMED N, et al. Drone deep reinforcement learning: A review [J]. Electronics, 2021, 10 (9): 999.

[4] ALMAHAMID F, GROLINGER K. Autonomous unmanned aerial vehicle navigation using reinforcement learning: a systematic review [J]. Engineering Applications of Artificial Intelligence, 2022, 115: 105321.

[5] WANG C, WANG J, SHEN Y, et al. Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach [J]. IEEE Transactions on Vehicular Technology, 2019, 68 (3): 2124 - 2136.

[6] SHIN S Y, KANG Y W, KIM Y G. Obstacle avoidance drone by deep reinforcement learning and its racing with human pilot [J]. Applied Sciences, 2019, 9 (24): 5571.

[7] MAW A A, TYAN M, NGUYEN T A, et al. iADA*-RL: Anytime graph-based path planning with deep reinforcement learning for an autonomous UAV [J]. Ap-plied Sciences, 2021, 11 (9): 3948.

[8] TAJMIHIR ISLAM TEETHI, 卢　虎，闵　欢等. 基于改进强化学习的无人机规避决策控制算法 [J]. 探测与控制学报，2022, 44 (3): 68 - 73.

[9] TOLMAN E C. Cognitive maps in rats and men [J]. Psychological Review, 1948, 55 (4): 189

[10] BEHRENS T E J, MULLER T H, WHITTINGTON J C R, et al. What is a cognitive map? Organizing knowledge for flexible behavior [J]. Neuron, 2018, 100 (2): 490 - 509.

[11] 王继茹. 基于大脑空间认知机制的认知地图构建方法及应用研究 [D]. 成都：四川大学，2021.

[12] 陈雨荻，熊　智，刘建业等. 基于海马体的面向未知复杂环境类脑导航技术综述 [J]. 兵工学报，2022, 43 (11): 2965 - 2980

[13] 杨　闯，刘建业，熊　智等. 由感知到动作决策一体化的类脑导航技术研究现状与未来发展 [J]. 航空学报，2020, 41 (1): 35 - 49.

[14] YU F, SHANG J, HU Y, et al. NeuroSLAM: a brain-inspired SLAM system for 3D environments [J]. Biological Cybernetics, 2019, 113 (5 - 6): 515 - 545.

[15] ZENG T, TANG F, JI D, et al. NeuroBayesSLAM: Neurobiologically inspired Bayesian integration of multisensory information for robot navigation [J]. Neural Networks, 2020, 126: 21 - 35.

[16] BANINO A, BARRY C, URIA B, et al. Vector-based navigation using grid-like representations in artificial agents [J]. Nature, 2018, 557 (7705): 429 - 433.

[17] STACHENFELD K. Learning neural representations that support efficient reinforcement learning [D]. Princeton University, 2018, 14 - 41.

[18] BOTVINICK M, WANG J X, DABNEY W, et al. Deep reinforcement learning and its neuroscientific implications [J]. Neuron, 2020, 107 (4): 603 - 616.

[19] SUHAIMI A, LIM A W H, CHIA X W, et al. Representation learning in the artificial and biological neural net-works underlying sensorimotor integration [J]. Science Advances, 2022, 8 (22).

[20] KRUPIC J, BAUZA M, BURTON S, et al. Grid cell symmetry is shaped by environmental geometry [J]. Nature, 2015, 518 (7538): 232 - 235

[21] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529 - 533.

[22] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C] //International Conference on Machine Learning. PMLR, 2016: 1995 - 2003.

[23] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30 (1).

[24] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping [C] //Icml. 1999, 99: 278 - 287.

[25] SHAH S, DEY D, LOVETT C, et al. Airsim: High-fidelity visual and physical simulation for autonomous vehicles [C] // Field and Service Robotics: Results of the 11th International Conference. Springer Inter-national Publishing, 2018: 621 - 635.

[26] ANDERSON P, CHANG A, CHAPLOT D S, et al. On evaluation of embodied navigation agents [J]. arXiv preprint arXiv: 1807.06757, 2018.