计算机科学
COMPUTER SCIENCE

基于深度强化学习的无人机自主探索方法

唐嘉宁, 李成阳, 周思达, 马孟星, 施炀

引用本文

唐嘉宁, 李成阳, 周思达, 马孟星, 施炀. 基于深度强化学习的无人机自主探索方法[J]. 计算机科学, 2024, 51(11A): 231100139-6.

TANG Jianing, LI Chengyang, ZHOU Sida, MA Mengxing, SHI Yang. Autonomous Exploration Methods for Unmanned Aerial Vehicles Based on Deep ReinforcementLearning [J]. Computer Science, 2024, 51(11A): 231100139-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100139

基于深度强化学习的无人机自主探索方法

唐嘉宁[1,2,3]　李成阳[1,2]　周思达[2,3]　马孟星[1,2,3]　施　炀[1,2]
1 云南民族大学电气信息工程学院　昆明 650031
2 云南省无人自主系统重点实验室　昆明 650031
3 云南民族大学无人自主系统研究院　昆明 650031
(041749@yum.edu.cn)

摘　要　无人机面对非结构化未知环境,如山地和丛林等场景进行探索时,必须在缺乏先验条件的情况下同时进行环境感知和航迹规划。传统方法受制于算法和传感器等多重因素的制约,探索范围有限,效率低下,并易受到环境变化的干扰。为解决这一问题,提出了一种基于深度强化学习的无人机自主探索方法。该方法以归一化优势函数(Normalized Advantage Functions,NAF)算法为基础,引入了3种算法增强机制,以提升无人机在非结构化未知环境中的探索范围和效率。在自行设计的仿真环境中进行实验,结果表明,改进后的NAF算法相较于原始版本,具有更大的探索范围和更高的效率,同时表现出优越的收敛性和鲁棒性。

关键词:无人机自主探索;智能决策;深度强化学习;NAF算法;增强机制
中图分类号　V249

Autonomous Exploration Methods for Unmanned Aerial Vehicles Based on Deep Reinforcement Learning

TANG Jianing[1,2,3],LI Chengyang[1,2],ZHOU Sida[2,3],MA Mengxing[1,2,3] and SHI Yang[1,2]
1 School of Electrical and Information Technology,Yunnan Minzu University,Kunming 650031,China
2 Yunnan Key Laboratory of Unmanned Autonomous System,Kunming 650031,China
3 Institute of Unmanned Autonomous Systems,Yunnan Minzu University,Kunming 650031,China

Abstract　Faced with unstructured and unknown environments,such as exploring in mountains and jungles,UAVs must simultaneously perform environment sensing and trajectory planning in the absence of a priori conditions. Traditional methods are constrained by multiple factors such as algorithms and sensors,resulting in limited exploration range,low efficiency,and susceptibility to interference from environmental changes. To solve this problem,this study proposes an autonomous exploration method for UAVs based on deep reinforcement learning. The method is based on the normalized advantage functions(NAF) algorithm and introduces three algorithmic enhancement mechanisms to improve the exploration range and efficiency of UAVs in unstructured and unknown environments. By conducting experiments in a self-designed simulation environment,the results of simulation experiments and analysis show that the improved NAF algorithm has a larger exploration range and higher efficiency compared to the original version,while exhibiting superior convergence and robustness.

Keywords　Autonomous UAV exploration,Intelligent decision making,Deep reinforcement learning,NAF algorithm,Augmentation mechanism

1　引言

近年来,无人机因其出色的悬停能力及较低的起降条件,逐渐成为探索未知复杂环境的最佳对象,被广泛应用于各种危险场景,例如搜救[1]、环境监测[2]等领域。无人机自主探索是指无人机在没有环境先验条件的情况下,对指定的未知区域展开尽可能快速、完整的探索,完成对环境的感知和路径规划。根据算法本身的不同,主要分为两类传统方法(基于边界检测和基于采样的方法)和基于强化学习的方法。

基于边界检测的方法,是目前大多数未知环境探索算法的基础。Yamuchi于1997年将探索定义为在未知环境中移动,同时构建后续用于导航的地图[3]。Keidar为了提高边界检测的效率,提出了两种新的边界检测算法WFD(波前边界检测器)和FFD(快速边界检测器)[4],在检测边界时避免搜索地图中的已知和未知区域,将搜索范围缩小至只有可能包含边界的区域。Zhou提出了一种无人机快速探索的层次结构FUEL[5],通过探索时增量式地更新并存储关键的前沿信息,提高了复杂环境下探索的速度。

基于采样的方法,通常以RRT[6]及其变种算法进行环境地图的随机采样。该方法与后续的路径规划相互耦合,因此

基金项目:国家自然科学基金(61963038,62063035)
This work was supported by the National Natural Science Foundation of China(61963038,62063035).
通信作者:李成阳(1521148692@qq.com)

受到了广大研究者的青睐。Bircher 将 NBV(Next-Best View)思想引入未知环境自主探索[7],在在线计算的随机树中找到效用最高的分支为最佳分支并作为下一步的探索路径,其中每条分支的效用评估取决于该分支上所有采样点的累积信息量。Wang 将二维探索扩展到三维环境,通过在传感器范围内随机采样,并在探索过程逐步增量式扩展和维护概率路线图(PRM)来构建三维环境的拓扑结构[8]。

基于强化学习的探索方法方面,Kulkarni 等对比了随机路标图和强化学习方法两种方法[9],证明了强化学习在路径规划方向的优势;Ma 等提出了一种添加状态链顺序反馈的强化学习算法[10],优化了机器人在静态复杂未知环境中的搜索方法。随着神经网络的提出并被广泛应用于各个领域,Panov 等在栅格地图上利用深度强化学习算法极大程度地优化了机器人的探索方法[11];Wang 等在无人机自主避障及路径规划问题中采用 DQN 算法[12],相较于未训练的无人机缩短了一半的巡航时间。Guo 等提出一种结合长短期记忆神经网络的移动机器人局部规划算法[13],增强了其算法的泛化性及探索效率。

然而上述 3 类未知环境的探索方法也存在如下问题:

(1)基于边界检测的探索方法虽然相对容易实现且适应性强,但在未知的非结构化环境中容易发生重复探索,从而降低探索效率。

(2)基于采样的探索方法由于前期采样具有随机性,很难在复杂环境中的狭窄入口等情况下实现快速且全面的探索,因此不完全适用于复杂的非结构化环境。

(3)基于强化学习的探索方法虽然在一定程度上提高了传统算法的探索效率和完整度,但由于非结构化环境过于复杂,部分算法选择不考虑飞行动力学约束,离散化动作空间进行建模,导致算法迁移到真实场景中的使用能力和泛化性较弱。

鉴于上述方法的不同缺点,本文提出一种基于深度强化学习的无人机自主探索方法。该方法在考虑飞行动力学约束的前提下,将动作空间建模为连续动作空间,减小了算法的仿真误差,增大了仿真到现实的迁移能力。本文提出的 3 种算法改进机制能有效提高模型的训练效率和鲁棒性,实现了使用强化学习算法解决未知环境下无人机自主探索问题的技术途径。

本文首先从构建自主探索流程框架入手,对无人机及整体环境进行建模;然后针对非结构化未知环境的特点,设计了 3 种改进归一化优势函数(Normalized Advantage Functions, NAF)算法的机制,针对性地提高了未知环境下深度强化学习算法的效果;最后在自行设计的非结构化环境进行仿真实验,实验结果验证了本文方法的有效性和实用性。

2 深度强化学习的背景知识

强化学习作为机器学习的一个重要分支,其本质是智能体以“试错”的方式在与环境交互中学习策略。与常见的监督学习和非监督学习不同,强化学习强调智能体与环境之间的交互,在交互过程中通过不断学习来改变策略获取最大回报,以得到最优策略[14]。其交互方式如图 1 所示。强化学习算法通常利用马尔可夫决策过程(Markov Decision Process, MDP)进行建模。MDP 由五元组$\{S,A,P,R,\gamma\}$组成,其中,S 是智能体当前环境下有限状态的集合;A 是有限动作集;P 是转移概率:$P_{ss'}^{a}=P[S_{t+1}=s'\mid S_t=s,A_t=a]$;$R$ 是奖励函数:$R_s^a=E[R_t\mid S_t=s,A_t=a]$,$\gamma$ 是折扣因子:$\gamma\in[0,1]$,其表示智能体对于当前奖励及未来奖励的权衡关系。

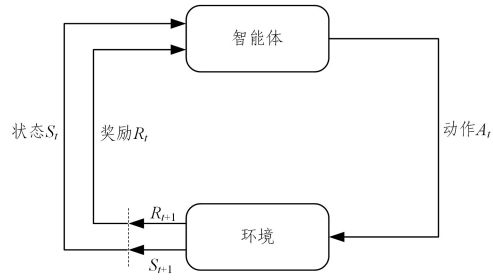


图 1 强化学习示意图

Fig. 1 Schematic of reinforcement learning

近来,深度学习(Deep Learning, DL)的迅速发展引领了人工智能领域的创新浪潮。随着深度学习技术在各个领域的卓越表现,融合深度神经网络和强化学习成为学术界和工业界关注的焦点。借助深度神经网络强大的表征能力去拟合强化学习的主要组成部分,包括状态价值函数、动作价值函数、策略、模型等,大幅提升了强化学习算法的性能。2013 年,DeepMind 公司的 Mnih 等提出了开创性的深度 Q 网络(Deep Q-Network, DQN)[15],该算法用神经网络取代了存放值函数的表格,智能体仅通过从图像中获取信息就能学会玩视频游戏。

尽管 DQN 在处理离散动作空间的问题上取得了巨大成功,但在面对具有非凸优化特性和连续动作空间的复杂任务时,其在梯度计算方面的复杂性成为了限制因素。归一化优势函数(Normalized Advantage Functions, NAF)算法[16]的提出对这一问题作出了创新性回应,其采用归一化优势函数的方法,有效地应对了在连续动作空间中进行策略优化的挑战。与 DQN 相比,NAF 算法不仅提供了更精确的策略价值估计,而且消除了繁琐的学习率调整过程,从而显著提高了稳定性和性能。

图 2 描述了 NAF 算法中网络的更新流程。NAF 算法的基本思想是在 DQN 算法的基础上,神经网络可以直接输出动作,并且保证输出的动作具有最大的 Q 值。算法借鉴了 Dueling DQN 的思想,将 Q 拆分为优势函数 A 和价值函数 V,即:

$$Q(s_t,a_t\mid\theta^Q)=V(s_t\mid\theta^V)+A(s_t,a_t\mid\theta^A) \tag{1}$$

其中,θ^Q,θ^V 和 θ^A 分别表示 Q 值函数、状态值函数 $V(s_t\mid\theta^V)$ 和动作优势函数 $A(s_t,a_t\mid\theta^A)$ 的参数,s_t 和 a_t 分别为 t 时刻无人机的当前状态和控制动作。

神经网络部分有 3 个输出,分别是状态价值 $V(s_t\mid\theta^V)$、完全贪婪策略下的动作 $\mu(s\mid\theta^\mu)$ 以及下三角矩阵 $L(s\mid\theta^P)$。输出下三角矩阵 $L(s\mid\theta^P)$ 的目的是根据乔列斯基(Cholesky)分解,构造正定矩阵 $P(s\mid\theta^P)$,即:

$$P(s\mid\theta^P)=L(s\mid\theta^P)L(s\mid\theta^P)^{\mathrm{T}} \tag{2}$$

正定矩阵 $P(s\mid\theta^P)$ 用于构造优势函数 A,即:

$$A(s_t,a_t\mid\theta^A)=-\frac{1}{2}(a-\mu(s\mid\theta^\mu))^{\mathrm{T}}P(s\mid\theta^P)(a-\mu(s\mid\theta^\mu)) \tag{3}$$

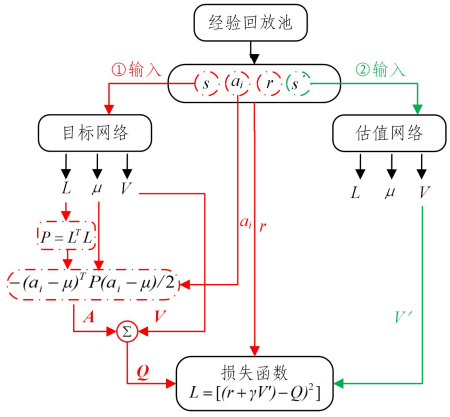


图 2　NAF 算法流程图

Fig. 2　Flowchart of NAF algorithm

由于 $P(s|\theta^P)$ 为正定矩阵，故式(3)的最大值为 0，即当动作 $a=\mu(s|\theta^\mu)$ 时优势函数有最大值 0。根据式(1)可得出，当动作 $a=\mu(s|\theta^\mu)$ 时，有最大的 $Q(s_t,a_t|\theta^Q)$。所以 NAF 算法能在输出连续动作的同时，输出当前动作对应的 Q 值，以此完成网络参数 θ^A 和 θ^V 的更新。

3　无人机自主探索流程设计

本节首先介绍无人机自主探索流程的总体框架，之后在 NAF 算法基础上，设计 3 种算法增强机制，用于提升算法和整体框架的探索性能。

3.1　总体框架设计

图 3 给出了基于深度强化学习算法的无人机自主探索框架，以下统称算法框架为 GNAF。整个框架共包括 4 个模块，分别是深度强化学习模块、经验回放模块、收敛探索模块以及神经网络训练模块。

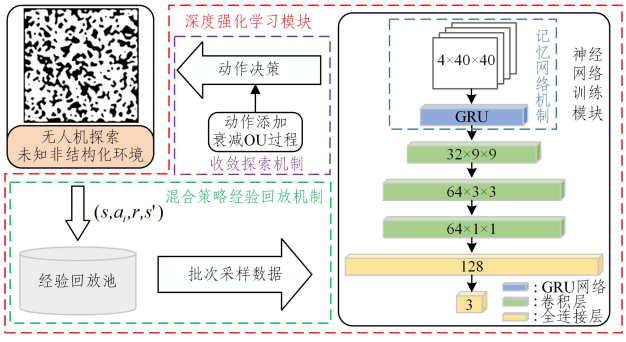


图 3　无人机自主探索框架

Fig. 3　Autonomous exploration framework of UAV

在非结构化未知环境中，无人机通过自主探索的方式获取信息，并将所得信息存储于经验回放池。当经验池的存储量达到一定阈值时，采用混合策略从经验池中进行数据采样，并输入神经网络供训练模块使用。该模块的操作流程包括将无人机当前状态输入门控循环单元(GRU)网络，使网络对无人机所处环境的历史信息进行记忆。随后，通过三层卷积网络对环境状态特征进行提取，最终通过全连接网络输出状态价值、动作以及下三角矩阵。输出的动作与收敛探索机制相结合，控制无人机在环境的下一个动作，持续推动其与环境的交互与自主探索。

3.2　记忆网络机制

未知环境下，基于深度强化学习的无人机实现自主探索，在不同环境状态下采取的策略极为复杂，如何使智能体再次遇到相同或相似的状态时可以快速学习之前的参数？添加记忆网络能够有效地解决这个问题。传统的长短期记忆(Long Short-Term Memory，LSTM)[17]网络，基于 RNN 做出改进，能够有效捕捉长序列之间的关联，但其参数量较多，容易导致过拟合。本文采用基于 LSTM 改进的门控循环单元结构(Gated Recurrent Unit，GRU)[18]，全局思想是将当前时刻输入的 x_t 与上一时刻的隐藏信息 h_{t-1} 进行一系列耦合，得到下一时刻的隐藏状态 $\tilde{h}_t$。其结构如图 4 所示，核心可以分为重置门与更新门两个部分。

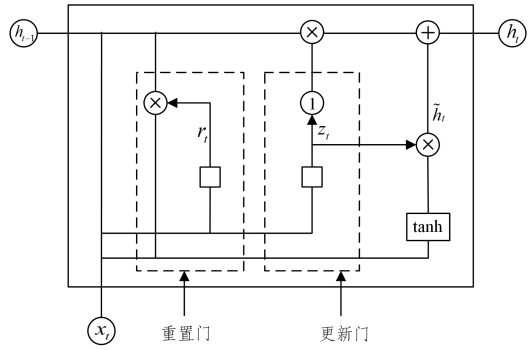


图 4　门控循环单元结构

Fig. 4　Gated recurrent unit structure

重置门部分将 LSTM 的遗忘门和输出门结合，降低了参数数量，其决定了如何将新的输入信息与前面的记忆信息结合：

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]) \tag{4}$$

其中，r_t 是重置门的输出；W_r 是一个权重矩阵，用于将 $[h_{t-1},x_t]$ 转换成与拼接前相同的维度；输出的 r_t 用于计算候选隐藏状态 $\tilde{h}_t$：

$$\tilde{h}_t=\tanh(W\cdot[r_t*h_{t-1},x_t]) \tag{5}$$

更新门用于更新记忆，即用于决定传递多少信息到下一时刻：

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]) \tag{6}$$

计算得到 z_t 后，将其代入下式，得到最终传递到下一时刻的隐藏态 h_t：

$$h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t \tag{7}$$

3.3　收敛探索机制

传统强化学习算法通常添加独立的高斯噪声，然而 Pawel 提出相比于独立噪声，在惯性系统中使用高斯噪声会产生一系列独立的在 0 均值附近的高斯噪声，使得随机噪声的作用被抵消[19]。如果使用 Ornstein-Uhlenbeck 过程(即 OU 噪声)，则会顺着惯性向一个方向多探索几步，使得探索效果更优，因此 OU 过程更适合于时间离散化粒度较小的惯性系统。

针对上述情况，本文设计了一种基于 OU 过程可变噪声的收敛探索机制，其在增强无人机探索能力的同时提高了其收敛速度。

OU 过程是一种随机过程，其微分方程定义为：

$$dx_t=-\theta(x_t-\mu)dt-\sigma dW_t \tag{8}$$

其中，μ 为均值，θ 和 σ(均值)均为大于 0 的参数，σdW_t 是扰动项，W_t 是维纳过程。对微分方程求解后可得：

$$x_t=\mu+(x_0-\mu)e^{-\theta t}+\sigma\int_0^t e^{-\theta(t-s)}W_t \tag{9}$$

其中，θ 表示系统对干扰的反应程度，$\sigma\int_0^t e^{-\theta(t-s)}W_t$ 为扰动项，其增量$(W_{(t)}-W_{(s)})$服从高斯分布：$(W(t)-W(s))\sim N(0,\sigma^2(t-s))$，$\sigma$ 表示扰动的放大系数。本文提出的收敛探索机制，对扰动系数 σ 添加一个衰减系数 β。训练时 σ 定义为：

$$\sigma=\max(\sigma_{\min},-\beta(\sigma_{\text{init}}-\sigma_{\min})+\sigma) \tag{10}$$

其中，σ 为 OU 过程中扰动项的系数，σ_{init} 为初始值。随着训练步数的增加，σ 逐渐减小直到 $\sigma_{\min}$，从而达到收敛探索的目的。

收敛探索机制很好地平衡了强化学习问题中探索与利用的关系。智能体在前期具备更强大的探索能力，随着训练次数的增加，在后期更多地利用无人机与环境交互得到的信息来控制智能体的行为。该机制提高了前期的探索效率，加快了无人机训练的收敛速度，避免了无人机陷入局部最优解，导致探索不完全。

3.4 混合策略经验回放机制

在无人机探索未知环境的过程中，无人机与环境交互会产生大量的经验样本，然而这些样本的重要性存在差异。若仅采用传统的经验回放机制，无法有效地促使无人机更快地学习到重要的经验；若仅采用传统的优先级别经验回放机制，过度频繁地抽取时间差分误差（Temporal Difference Error，TD-Error）较大的经验样本，可能会导致训练结果过拟合，尤其是在训练中后期，智能体的探索能力明显下降。因此，如何在智能体的探索速度与探索能力之间权衡是值得思考的问题。

为解决上述问题，基于优先经验回放[20]，设计了一种混合策略经验回放机制。该机制在前期采用优先经验回放，在中后期采用均匀随机采样。在优先经验回放中，TD-Error 用于衡量估计值与实际值之间的差异。NAF 算法的 TD Error 计算公式如下：

$$\delta_i=|Q(s,a)-(r+\gamma Q(s',\arg\max_{a'}Q(s',a')))| \tag{11}$$

其中，$Q(s,a)$表示实际值，$Q(s',a')$表示估计值。在优先级别经验回放中，定义经验样本的优先级别为：

$$p_i=|\delta_i|+\varepsilon \tag{12}$$

其中，p_i是经验样本的优先级；δ_i表示 TD Error；ε 是一个大于零的常数，用于保证无论 TD-error 取值如何，采样概率 p_i 仍大于 0，仍有概率会被采样到。我们将经验样本的优先级别转化为采样概率，如下：

$$P(i)=\frac{p_i^a}{\sum_k p_k^a} \tag{13}$$

4 仿真实验结果及分析

本文环境为自行设计的非结构化仿真环境，可以根据需要切换地图大小以及障碍物密度等，可以反馈得到无人机在环境中的各类信息，适用于无人机自主探索算法的研究和验证。实验台式机操作系统为 Ubuntu 20.04，搭载的 CPU 为 i9-12900KF，显卡为 ZOTAC RTX 3090。

4.1 无人机飞行动力学约束

本文将非结构化三维地图切片，不考虑无人机高度的爬升，仅考虑无人机在二维平面的飞行动力学约束，将最小允许转弯半径$R_{\min}$和最大航向半角 ψ 建模为：

$$R_{\min}=\frac{V_{\text{UAV}}^2}{g\sqrt{\mu^2-1}} \tag{14}$$

$$\psi=\arcsin\left(\frac{1}{2}\cdot\frac{\lambda}{R_{\min}}\right) \tag{15}$$

其中，V_{UAV}为无人机速度；μ 为允许的最大正常过载系数；g 为重力加速度，λ 为步长。

无人机规划下一步的位置和方向为：

$$\begin{cases}x_u'=x_u+\lambda\cdot\cos(\psi+\varphi_u)\\ y_u'=y_u+\lambda\cdot\sin(\psi+\varphi_u)\\ \varphi_u'=\psi+\varphi_u\end{cases} \tag{16}$$

其中，(x_u,y_u)和(x_u',y_u')分别为无人机在当前时刻与下一时刻的位置，φ_u和φ_u'分别为无人机在当前时刻与下一时刻的方向。

4.2 强化学习建模

4.2.1 状态空间设计

在非结构化未知环境中，无人机受到传感器的影响，只能观测到局部且有限的环境信息。本文在二维切片地图环境下进行实验，即真实高原山地环境按照某一固定海拔高度水平切片后的模拟仿真地图，如图 5 所示。图 5(a)中白色为可行区域，黑色为障碍物区域，灰色为未知区域为。全局地图由随机种子控制生成，可通过修改随机种子大小修改地图。图 5(b)灰色区域为开始探索前人为设置的蒙版，对无人机来说是不可见的未知区域。

(a)全局地图(假如未知)　(b)探索地图(已探索)

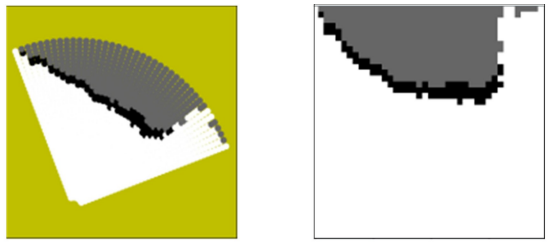
图 5 仿真训练地图

Fig. 5 Simulation training map

将无人机的扇形视野范围局部观测地图(见图 6(a))进行预处理，得到 40×40 大小的矩形观测地图(见图 6(b))，其中白色为无人机已探索区域，黑色为障碍物区域，灰色为未知区域。无人机在探索过程中的状态空间描述为：

$$S=\{o_t,h_t\} \tag{17}$$

其中，o_t为观测地图；h_t为 GRU 网络的历史信息。



(a)局部地图(当前观测)　(b)输入网络的当前状态

图 6 无人机视野观测地图

Fig. 6 Map of UAV field of view observation

4.2.2 动作空间设计

假定无人机在定高飞行条件下进行探索，飞行高度不变的前提下本文的动作空间设计只考虑无人机的相对航向角

度。在速度已知且匀速的前提下，根据飞行动力学约束，设置其最大航向角范围为[－29.3°，29.3°]。区别于传统 DDQN 算法，本文通过 NAF 算法将航向角度函数映射到[－1，1]范围内进行训练，将离散动作空间扩展到连续动作空间，提升了模型的泛化性以及仿真到现实的迁移能力。

4.2.3　奖励函数设计

区别于普通航迹规划，无人机在未知环境下的探索不仅需要对环境进行探索，同时需要在避障的情况下进行规划，所以针对奖励函数不能设计到达目标点的奖励。在无人机与环境交互过程中，通过比较前后两幅全局地图数据，即全局地图内视障碍物等信息的前后状态，计算探索值，最终得到奖励 r；为了鼓励无人机学会避开障碍，设立一个较大的碰撞惩罚，奖励函数设计如下：

$$\begin{cases} -1000, & \text{发生碰撞} \\ r, & \text{未发生碰撞} \end{cases} \tag{18}$$

4.2.4　网络设计

本文采用的目标网络与估值网络，除各网络层参数外，结构上保持一致。具体而言，图 3 所示神经网络模块中，连续 4 帧图片被整合并输入神经网络进行训练。这一过程经过门控循环单元(GRU)模块进行特征记忆，卷积层用于对特征进行压缩与提取，最终通过全连接层输出结果。为了保持参数的更新，估值网络会定期从目标网络复制参数。神经网络的示意图如图 3 所示。

4.2.5　超参数设计

本文各项算法的超参数设计情况如表 1 所列。

表 1　训练参数设置

Table 1　Training parameter settings

参数名称	参数大小
学习率	1×10^{-3}
折扣系数	0.98
批次大小	64
经验池容量	50000
σ(扰动系数)	0.9→0.1

4.3　算法有效性实验

为验证本文所提算法框架的有效性，根据上述设计方法进行对比实验，分别记录本文 GNAF 算法、传统 NAF 算法、DDQN 算法在相同地图下每 10 个回合中一次单步探索奖励的均值，最后画出单步平均奖励曲线，如图 7 所示。在训练过程中，NAF 算法同样使用连续动作空间建模，DDQN 算法则将动作空间均匀划分，建模为离散动作空间。

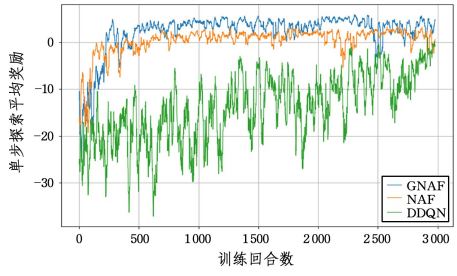


图 7　单步探索的平均奖励

Fig. 7　Average reward for single-step exploration

根据图 7 结果可知，本文所提出的自主探索算法框架在前 500 个回合的训练中就能达到较高的探索奖励值。单步探索平均奖励曲线逐渐趋于平稳，表示无人机在环境中能够较稳定地探索，受到惩罚(撞到障碍)的次数减少。可以看出在整个过程中，GNAF 算法单步探索平均奖励值能够更快收敛，并且收敛后奖励波动小，具有较好的稳定性。

4.4　测试对比实验

在相同的随机地图下，对 3 种算法进行测试。设置探索步数范围为 20000 步至 30000 步，每次实验间隔为 2000 步。每个特定探索步数进行 5 次实验，并将各实验的探索结果指标取均值，以降低由随机因素引起的误差。最终，绘制探索面积比指标图进行对比分析。

从图 8 可以看出，本文提出的算法框架 GNAF 在相同探索步数下的探索面积比更大，与传统 NAF 算法相比平均提升了 8.83%，与传统 DDQN 算法相比平均提高了25.17%。

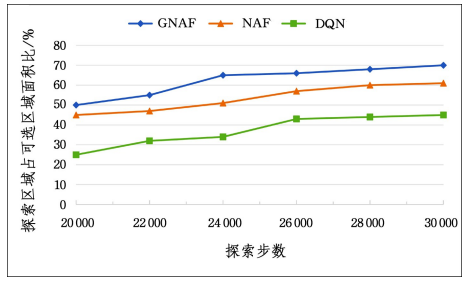
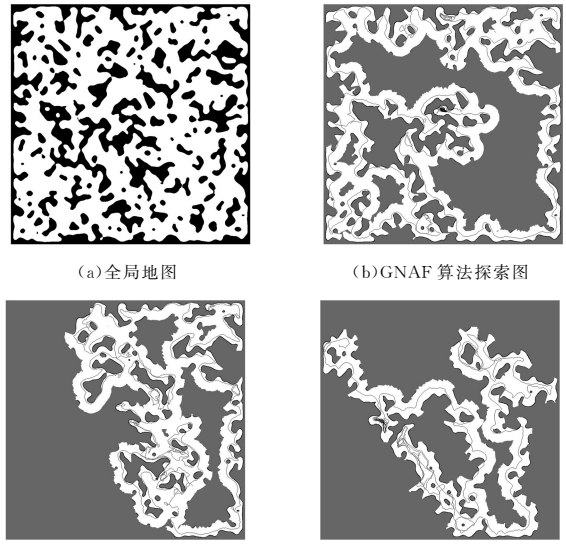


图 8　探索面积对比图

Fig. 8　Explore area comparison charts

图 9 展示了在同一个随机地图上，3 种算法在探索 30000 步时获得的探索结果。在同一地图下，GNAF 算法在 30000 步下实现的探索面积比为 69.39%，而 NAF 算法和 DDQN 算法的探索面积比为 58.37%和 42.12%，分别提高了 11.02%和 27.27%。



(a)全局地图　(b)GNAF 算法探索图

(c)NAF 算法探索图　(d)DDQN 算法探索图

图 9　30000 步时各算法探索图

Fig. 9　Exploration map of each algorithm at 30000 steps

4.5　算法泛化性实验分析

算法的泛化性是评估算法性能的重要指标。我们随机生成了 3 个包含非结构化障碍物的地图，并分别采用 3 种训练好的算法进行 5 轮测试。每个地图的探索步数均为 30000 步，将取得的探索面积比进行平均，最终的结果如表 2 所列。

表 2 随机地图下 3 种算法探索面积的对比

Table 2 Comparison of three algorithms for exploring area under randomized maps

(%)

测试地图	GNAF 算法探索面积比	NAF 算法探索面积比	DDQN 算法探索面积比
地图 1	68.95	55.12	38.12
地图 2	66.72	58.29	39.27
地图 3	63.29	54.19	41.32

可以看出，GNAF 算法相较于 NAF 算法，探索面积平均提高了 10.45%，相较于 DDQN 算法平均提高了 16.29%。

结束语 针对无人机在非结构化未知环境中探索的挑战，本文提出了一种基于深度强化学习的自主探索方法，以归一化优势函数(NAF)算法为基础，引入了 3 种算法增强机制，旨在提升无人机在缺乏先验条件情况下的环境感知能力和探索效率。通过在自行设计的仿真环境中进行实验，验证了改进后的 NAF 算法相较于原始版本的显著优势。这一方法在非结构化未知环境中显著扩大了无人机的探索范围，提高了探索的效率。同时，我们观察到改进后的算法表现出更好的收敛性和鲁棒性，这对于在实际应用中面对环境变化的情况具有重要意义。

然而，我们也认识到在实现这一改进的过程中引入了额外的计算负担，从而对实时性能产生一定的影响。未来将着重于优化算法，以在不引入过多计算负担的前提下继续提升无人机在复杂环境中的探索性能。

参 考 文 献

[1] VALENTE J, DEL CERRO J, BARRIENTOS A, et al. Aerial coverage optimization in precision agriculture management: A musical harmony inspired approach[J]. Computers and Electronics in Agriculture, 2013, 99: 153-159.

[2] TOMIC T, SCHMID K, LUTZ P, et al. Toward a fully autonomous UAV: Research platform for indoor and outdoor urban search and rescue[J]. IEEE Robotics & Automation Magazine, 2012, 19(3): 46-56.

[3] YAMAUCHI B. A frontier-based approach for autonomous exploration[C]// Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. IEEE, 1997: 146-151.

[4] KEIDAR M, KAMINKA G A. Efficient frontier detection for robot exploration[J]. The International Journal of Robotics Research, 2014, 33(2): 215-236.

[5] ZHOU B, ZHANG Y, CHEN X, et al. FUEL: Fast UAV exploration using incremental frontier structure and hierarchical planning[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 779-786.

[6] LAVALLE S M. Rapidly-exploring random trees: A new tool for path planning[R]. Research Report, 1998.

[7] BIRCHER A, KAMEL M, ALEXIS K, et al. Receding horizon "next-bestview" planner for 3d exploration[C]// 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016: 1462-1468.

[8] WANG C, MA H, CHEN W, et al. Efficient autonomous exploration with incrementally built topological map in 3Denvironments[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9853-9865.

[9] KULKARNI P, GOSWAMI D, GUHA P, et al. Path planning for a statically stable biped robot using PRM and reinforcement learning[J]. Journal of Intelligent and Robotic Systems, 2006, 47(3): 197-214.

[10] MA X, XU Y, SUN G, et al. State-chain sequential feedback reinforcement learning for path planning of autonomous mobile robots[J]. Journal of Zhejiang University Science C, 2013, 14(3): 167-178.

[11] PANOV A I, YAKOVLEV K S, SUVOROV R. Grid path planning with deep reinforcement learning: Preliminary results[J]. Procedia Computer Science, 2018, 123: 347-353.

[12] WANG G, ZHENG X, ZHAO H, et al. Unmanned aerial vehicles path planning based on deep reinforcement learning[C]// The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Cham: Springer, 2019: 81-88.

[13] GUO N, LI C, WANG D, et al. Local path planning of mobile robot based on long short-term memory neural network[J]. Automatic Control and Computer Sciences, 2021, 55(1): 53-65.

[14] SUTTON R S, BARTO A G. Reinforcement Learning: An I-ntroduction(2nd ed)[M]. Massachusetts: MIT Press, 2018.

[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[J]. arXiv: 1312.5602, 2013.

[16] GU S, LILLICRAP T, SUTSKEVER I, et al. Continuous deep q-learning with model-based acceleration[C]// International Conference on Machine Learning. PMLR, 2016: 2829-2838.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[18] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learn-ing phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv: 1406.1078, 2014.

[19] WAWRZYNSKI P. Control policy with autocorrelated noise in reinforcement learning for robotics[J]. International Journal of Machine Learning and Computing, 2015, 5(2): 91.

[20] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv: 1511.05952, 2015.

TANG Jianing, born in 1984, Ph.D, professor, Ph.D supervisor. Her main research interest is cooperative guidance and control.

LI Chengyang, born in 1999, postgraduate. His main research interests include deep reinforcement learning and auto-nomous exploration of unmanned aerial vehicles.