计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009938] [CSTR: 32024.14.csa.009938]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 EFRE-SAC 的无人机自主避障策略①

刘萌月, 时宏伟

(四川大学 计算机学院, 成都 610065)
通信作者: 时宏伟, E-mail: shihw001@126.com

摘　要: 在无人机自主避障任务中, 传统强化学习算法往往面临状态空间高维、信息稀疏以及探索效率低下等挑战. 现有的 SAC 算法虽然具备较强的稳定性和样本效率, 但在复杂环境下的表现仍显不足. 为此, 本文提出了一种基于注意力机制 SE 和随机网络蒸馏 RND 模块改进的 SAC 算法, 旨在提升无人机在三维地形环境中的自主避障能力. 注意力机制 SE 通过自适应调整特征图的通道权重, 增强了模型对重要信息的关注能力, 从而提升了特征表达的有效性; 而改进的 RND 网络则通过生成对抗目标, 鼓励探索新环境, 丰富了样本的多样性和改善了收集效率. 基于上述的 SE 和 RND, 我们构建了一个增强特征表达和探索的 SAC (EFRE-SAC) 框架, 使得无人机能够更有效地从深度图像中学习环境特征, 并在三维环境中快速适应. 在 AirSim+UE4 仿真平台的实验结果表明, 所提出的改进方法显著提高了无人机的避障成功率和训练效率, 验证了改进的 SE 和 RND 模块在强化学习任务中的有效性.

关键词: 无人机; 避障; SAC; 随机网络蒸馏; 注意力机制

引用格式: 刘萌月,时宏伟.基于 EFRE-SAC 的无人机自主避障策略.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9938.html

Autonomous Obstacle Avoidance Strategy for UAV Based on EFRE-SAC

LIU Meng-Yue, SHI Hong-Wei

(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: In the task of autonomous obstacle avoidance for UAVs, traditional reinforcement learning algorithms face challenges such as high-dimensional state spaces, sparse information, and low exploration efficiency. Although the existing soft Actor-Critic (SAC) algorithm demonstrates strong stability and sample efficiency, its performance in complex environments remains inadequate. To address these issues, this study proposes an improved SAC algorithm, incorporating a squeeze-and-excitation (SE) attention mechanism and random network distillation (RND) module, to enhance the obstacle avoidance capability of UAVs in three-dimensional terrain environments. The SE attention mechanism adaptively adjusts the channel weights of feature maps, enhancing the model's focus on Critical information and improving feature representation. Meanwhile, the improved RND network promotes the exploration of new environments by generating adversarial targets, thus increasing sample diversity and collection efficiency. Based on the integration of SE and RND, an enhanced feature representation and exploration SAC (EFRE-SAC) framework is constructed, enabling more effective learning of environmental features from depth images and rapid adaptation in three-dimensional environments. Experimental results on the AirSim+UE4 simulation platform demonstrate that the proposed method significantly improves the obstacle avoidance success rate and training efficiency of UAVs, validating the effectiveness of the improved SE and RND modules in reinforcement learning tasks.

Key words: unmanned aerial vehicle (UAV); obstacle avoidance; soft Actor-Critic (SAC); random network distillation (RND); attention mechanism

① 收稿时间: 2024-11-17; 修改时间: 2025-02-12; 采用时间: 2025-03-06; csa 在线出版时间: 2025-04-30

小型多旋翼无人机因其卓越的机动性、灵活的控制和低廉的成本等优势, 被广泛应用于农业[1]、物流[2]、监控和救援[3]等多个领域. 具体而言, 在农业领域, 无人机可实现精准的农药喷洒和作物检测, 显著提升生产效率; 在物流领域, 无人机提供了快速、灵活的货物运输方案; 在监控和救援领域, 无人机能够监测环境, 支持应急响应, 为救援行动提供有力支持. 随着无人机在各领域的应用不断拓展, 其飞行过程中面临的障碍物愈加复杂, 涵盖了多种环境挑战. 与此同时, 传感器技术、计算机视觉和人工智能的快速发展也为无人机的避障技术带来了新的机遇和挑战. 如何在复杂的环境中实现高效、稳定的避障, 依然是无人机技术发展中的重要课题.

目前, 针对无人机的自主避障技术, 研究人员提出了大量方法, 大致可以分为传统算法、机器学习算法和深度强化学习算法. 传统避障算法包括基于规则的避障策略和经典的路径规划方法. 基于规则的方法通过设定逻辑条件, 实现对障碍物的检测与路径调整. 例如, Mac 等人[4]利用机载视觉和惯性传感器提出了改进的潜在场方法, 通过调整目标吸引力和障碍物排斥力, 实现实时避障路径规划. Alharbi 等人[5]基于规则设计了无人机的共享空域冲突管理模型, 以确保空域的安全和高效使用. 这类算法逻辑清晰, 但缺乏灵活性. 经典的路径规划算法如 A*和 Dijkstra 算法通过图搜索进行规划路径, 确保无人机在约束条件下安全达到目标. 高九州等人[6]提出了改进的 A*算法, 李克玉等人[7]提出的改进 RRT 算法也用于无人机三维避障任务. 然而, 这类算法的路径更新效率较低, 难以满足实时需求.

随着机器学习技术的发展, 基于模型的避障方法和强化学习方法也被应用于无人机避障中. 基于模型的方法依赖环境的预建模, 如 Lindqvist 等人[8]提出的 NMPC 方法, 利用系统动态模型和障碍物轨迹的参数化描述, 通过模型预测控制优化无人机的避障策略. 代进进等人[9]同样提出了基于模型预测控制的避障路径规划方法, 采用一阶指数变化形式作为无人机飞行路径的参考轨迹, 解决无人机飞行路径的避障问题. 这类算法对未知环境适应性较差, 模型构建和维护成本高. 相比之下, 强化学习方法通过与环境的交互进行自我学习, 其中典型的 Q-Learning 算法通过学习状态-动作值函数来指导智能体在不同状态下选择最优动作, 从而实现自主学习和决策[10]. 但传统强化学习的训练过程样本需求量大、收敛速度慢且资源消耗高.

深度强化学习 (deep reinforcement learning, DRL) 的发展为无人机避障带来了新的机遇. DRL 结合了深度学习的特征提取能力和强化学习的决策能力, 使无人机能够在复杂环境中制定有效的避障策略, 尤其是在处理高维状态空间时, DRL 展现了显著优势. Duryea 等人[11]提出的 Double DQN 算法, 利用两个深度神经网络更精确地估计动作值. Xu 等人[12]通过引入更快的 R-CNN 模型来提取障碍物信息, 从而提高避障性能, 还提出了 MPTD3 算法, 将无人机的状态空间和动作空间建模为连续空间, 用于避障研究. 然而, 这些研究多在简化环境中进行, 如将无人机抽象为质点, 或障碍物不具代表性, 这在实际应用中存在局限性. Zhang 等人[13]利用激光测距传感器获取环境信息, 并通过 TD3 算法在 Unity 仿真平台中成功实现了自主导航和避障. Xue 等人[14]则利用 VAE 编码器将无人机第一视角图像编码为 SAC 算法的状态输入, 实现在不规则障碍物环境中的避障任务. 这些研究在 DRL 的探索阶段和利用阶段对环境的特征提取和探索能力的关注仍然有限. 探索与利用的平衡直接影响 DRL 学习效率, 是亟待改进的方向.

因此, 本文将在 AirSim 仿真平台中进一步探索基于 SAC 算法的无人机避障技术, 以增强无人机在陌生环境下的实时自主避障决策能力. 本文构建了基于 AirSim+UE4 仿真平台的三维地形环境, 提出了增强特征表达和探索的 SAC (enhanced feature representation and exploration SAC, EFRE-SAC) 框架, 实现无人机在限定飞行区域内的自主避障决策. 通过引入注意力机制挤压-激励 (squeeze-and-excitation, SE) 的卷积神经网络 (convolutional neural network, CNN) 处理无人机观测的图像信息, 减轻维度灾难问题, 提升环境特征表达质量, 从而增强决策效果; 通过在 SAC 中引入随机网络蒸馏 (random network distillation, RND) 改进 SAC 算法, 激励对新环境的探索, 丰富样本多样性和改善收集效率, 强化算法的探索能力; 基于连续动作空间控制实现无人机在三维环境下的自主避障能力, 并通过设计定制化奖励函数提升学习效率和稳定性.

1 算法原理

1.1 深度强化学习原理

在强化学习中, 问题通常被描述为一个马尔可夫

决策过程 (Markov decision process, MDP) , 其核心目标是让智能体找到一个策略π, 以最大化累积奖励, 奖励的累积目标定义为:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{1}$$

其中, γ是折扣因子, 控制远期奖励的重要性; $r_t = R(s_t, a_t)$为奖励函数, 表示智能体在状态s_t下, 执行动作a_t后获得的即时奖励. 为了评估策略π的优劣, 引入了状态-动作价值函数:

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a\right] \tag{2}$$

它表示智能体在特定状态下采取某一动作后期望获得的累积奖励. 若对$Q_\pi(s,a)$关于动作a求期望, 即在所有可能动作中加权求和, 可得到状态价值函数:

$$V_\pi(s) = \mathbb{E}_{a\sim\pi}[Q^\pi(s,a)] = \sum_{a\in\mathcal{A}} \pi(a|s) Q_\pi(s,a) \tag{3}$$

用来衡量在特定状态下策略π的优劣.

为应对高维连续状态空间中的复杂任务, DRL 结合了深度神经网络, 用以逼近策略和价值函数. DRL 中典型的行动者-评论家 (Actor-Critic, AC) 算法结合了 Actor 网络和 Critic 网络, 其中 Actor 网络负责在给定状态下选择动作, Critic 网络评估动作的预期奖励. 在软行动者-评论家 (soft Actor-Critic, SAC) 算法中, 通过引入熵正则化, 在提高探索能力的同时最大化期望回报. SAC 的优化目标如下:

$$J_\pi = \sum_{t=0}^{\infty} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))] \tag{4}$$

其中, $\mathcal{H}(\pi(\cdot|s_t)$表示策略$\pi$在状态$s_t$下的熵, 用于衡量策略的随机性, 参数$\alpha$则控制探索与利用的权衡. SAC 引入了两组独立的 Q 网络, 以减小估计偏差:

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})}\left[\frac{1}{2}(Q_\theta(s_t,a_t) - y_t)^2\right] \tag{5}$$

其中, y_t表示在当前状态下采取某一动作后, 考虑即时奖励以及下一步状态的价值评估后的目标值. SAC 通过最大化熵正则化目标来更新策略:

$$J_\pi(\phi) = \mathbb{E}_{s_t\sim D}\left[\mathbb{E}_{a_t\sim\pi_\phi}\left[\alpha\log\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right]\right] \tag{6}$$

其中, Actor 通过最大化熵正则化目标来学习最优策略. 此外, SAC 通过动态调整温度系数α, 以适应不同阶段的探索需求, 其更新公式为:

$$J(\alpha) = \mathbb{E}_{a_t\sim\pi}[-\alpha(\log\pi(a_t|s_t) + \mathcal{H}_0)] \tag{7}$$

SAC 的熵正则化和双 Q 网络机制使其适用于处理无人机避障任务中的复杂环境.

1.2 随机网络蒸馏

在无人机避障任务中, 由于奖励信号稀疏, 导致无人机在探索时容易陷入局部最优. RND 通过生成内在奖励, 鼓励智能体探索未访问过的状态[15]. RND 由目标网络和预测网络组成, 目标网络的参数在初始化后保证固定, 而预测网络则随着训练动态更新. 对于给定状态s, 目标网络生成固定特征向量$\hat{f}(s)$, 预测网络则试图对该特征向量进行估计. 在训练初期, 由于预测网络尚未学习目标网络的模式, 其误差较大. 随着训练进行, 预测网络对常访问的状态误差逐渐减小, 而对未访问过的状态误差较大. 内在奖励定义为预测误差的欧氏距离:

$$r_{\text{int}} = \left\|f(s) - \hat{f}(s)\right\|^2 \tag{8}$$

通过将内在奖励引入到强化学习的总奖励中, 智能体被鼓励去探索不常访问的状态, 从而覆盖更广的状态空间, 即便在稀疏奖励的环境中仍然可以获得有效的探索反馈.

1.3 注意力机制

在深度强化学习中, 特别是基于策略的算法由于其面对连续动作空间的挑战, 需要能够在训练中快速适应新数据的高精度模型. 与监督学习不同的是, 强化学习的训练数据较为不稳定, 无法像监督学习那样严格分离训练集和测试集以防止过拟合. 因此, 这类算法往往需要相对较浅的网络结构来保证训练效率和泛化能力. 在处理视觉输入时, CNN 就成为一种轻量又有效的网络选择. 无人机自主避障任务需要快速、高效地处理环境输入信息, 并提取有效的特征表示, 注意力机制 SE 的引入, 能够提升 CNN 特征提取的准确性[16].

SE 模块通过“挤压”和“激励”操作动态调整特征通道的重要性. 在“挤压”阶段, 对每个通道的进行全局平均池化, 压缩空间信息, 生成每个通道的全局响应. 随后, 在“激励”阶段, 利用卷积层生成每个通道的权重, 表示通道的重要性. 这一机制可使 CNN 更加关注对任务至关重要的通道特征. 在深度强化学习的策略网络中引入 SE 模块, 可提高环境状态的特征表达效果, 有助于无人机在复杂环境下做出高效的避障决策.

2 网络实现

在本文的无人机避障任务中, 任务目标是让无人机在飞行过程中与环境进行随机交互, 并成功通过障碍物区域而不发生碰撞. 环境的状态输入来自无人机搭载的摄像头捕获的深度图像. 由于无人机对其位置信息和环境状态知之甚少, 可以将此任务建模为一个部分可观测的马尔可夫决策过程.

为解决这一问题, 本文结合了 RND 和 SE 网络改进 SAC 算法, 提出了一个全新的无人机避障网络模型——EFRE-SAC. 该模型旨在帮助无人机在三维环境下实现自主避障. 特征提取部分通过 CNN 与 SE 模块的结合, 对环境状态进行处理, 提取有效特征表示, 从而帮助无人机做出决策. 在 EFRE-SAC 中, 如图 1 所示, 环境状态由无人机搭载的摄像头通过深度图像获取. 经过预处理, 图像尺寸为 (1, 128, 128), 表示一个单通道的 128×128 的二值化图像. 为了确保 CNN 网络能够高效训练并提取图像特征, 网络结构包括 3 层卷积层, 卷积核大小 3×3, 激活函数采用 ReLU. 为了增强 CNN 的特征表示能力, 引入了 SE 模块. SE 模块通过自适应地调整卷积特征图中每个通道的权重, 从而提升了对关键信息的关注度. SE 模块先通过全局平均池化 (AdaptiveAvgPool2d) 操作将每个通道的空间信息压缩为一个全局特征 (即“squeeze”过程), 然后通过两层全连接层 (Conv2d 层) 生成每个通道的权重系数 (即“excitation”过程), 最后这些权重系数与原卷积特征图进行逐通道的乘法操作, 从而实现对卷积特征的加权调整. 这样, SE 模块使得模型能够自动关注更为重要的特征通道, 进一步优化特征表示, 从而提高无人机自主避障任务的决策能力. 经过卷积层和 SE 模块处理后, 最终得到一个 512 维的特征向量, 作为环境状态的表示.

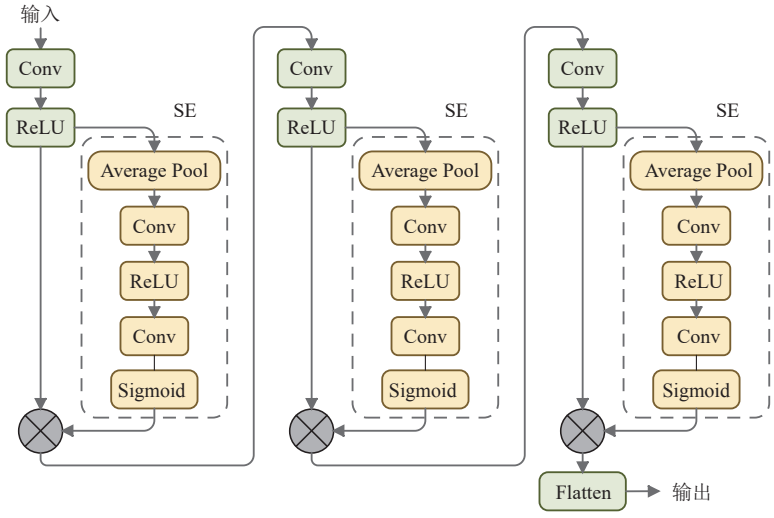

图 1 特征提取网络

EFRE-SAC 框架训练过程如图 2 所示, 包括 SAC 中的 Actor 网络、Critic 网络及 Critic Target 网络 3 部分, 其中, 每个网络均采用 256 个单元、两层的全连接结构. 为了保证探索与利用的平衡, SAC 算法中的熵系数 (entropy coefficient) 会在训练过程中自动调整. RND 中的预测网络和目标网络采用 3 层线性层结构, 目标网络在训练过程中保持不变 (即参数被冻结) , 而预测网络则通过对状态特征进行预测来产生输出. 预测网络的输出与目标网络的输出之间的均方误差 (MSE) 被用作内在奖励. 这个内在奖励能够引导模型去探索新的、未知的状态. 通过最小化预测网络和目标网络之间的差异, RND 强化了对新环境信息的探索, 使得模型能够发现更多潜在的有价值状态, 而不仅依赖外部奖励. RND 的引入不仅影响了探索行为, 还通过增强探索过程中的多样性, 间接地提高了策略的质量. 由于 RND 网络的内在奖励影响了经验回放池中的奖励分布, 策略在选择动作时, 会更加关注那些未被充分探索的状态区域, 从而在训练的早期阶段促进了策略的多样化, 并推动了全局最优解的发现. 模型中所有的网络 (Actor、Critic、Critic Target、RND 目标和预测网络) 共享特征提取网络, 从而确保状态信息的统一性和处理效率.

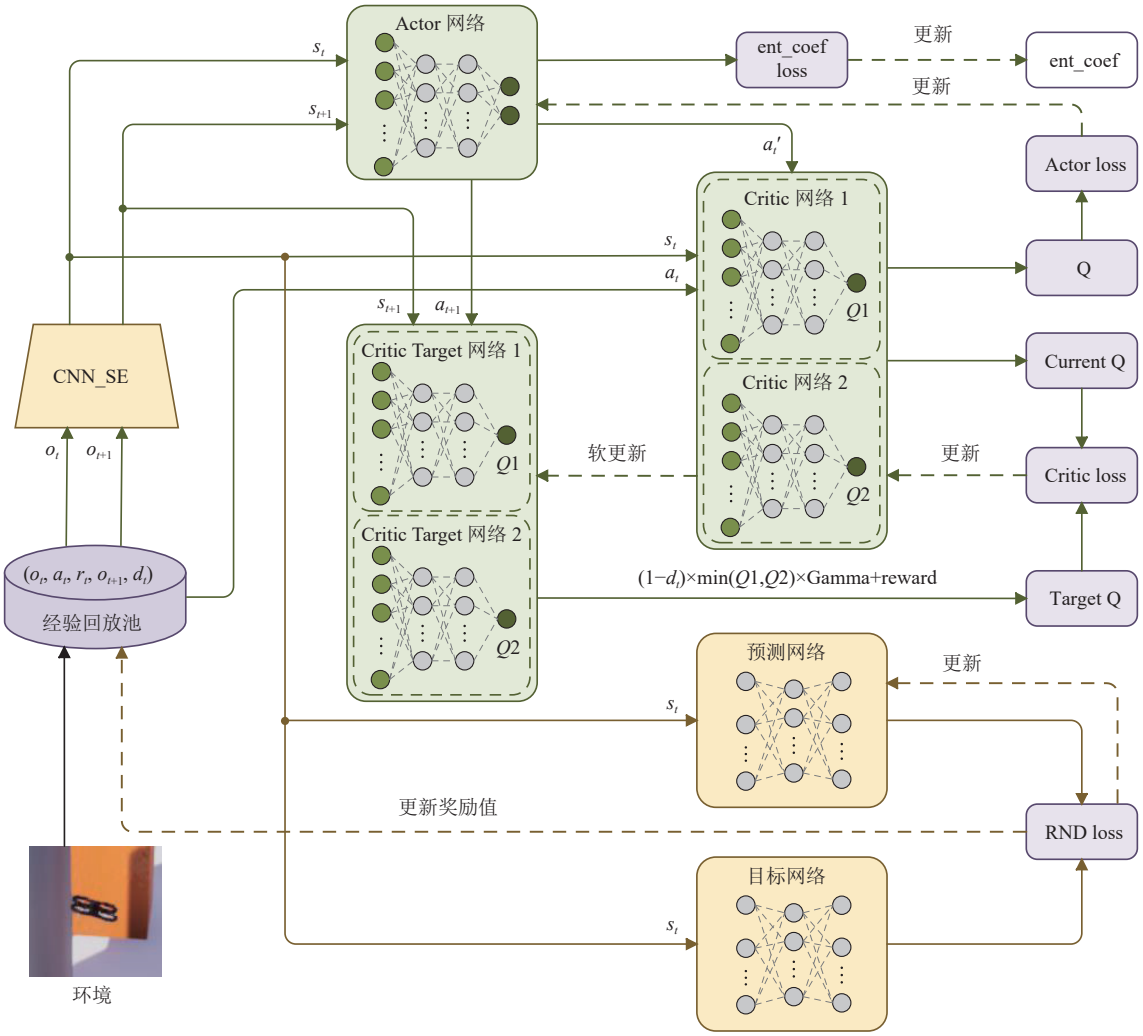

图 2 EFRE-SAC 框架训练过程

在训练过程中, 经验回放池用于存储数据$(o_t, a_t, r_t, o_{t+1}, d_t)$, 其中, o_t为时间步t时的环境观测值, a_t为时间步t时的执行动作, r_t为即时奖励, o_{t+1}为时间步$t+1$时的环境观测, d_t为到达目的地的标识. 一个训练回合中的网络参数更新过程如下.

1) 特征提取: 从经验回放池中批量采样数据后, 环境观测值o_t会经过特征提取网络处理, 得到状态表示s_t, 作为网络的输入.

2) 熵系数更新: 通过计算当前策略在状态s_t下的对数概率和目标熵之间的差异, 构造熵系数的损失函数, 通过梯度下降法调整熵系数α, 使策略的随机性适应任务需求.

3) Critic 网络更新: 计算目标 Q 值Q_{target}和当前 Q 值$Q_\pi(s_t, a_t)$之间的均方误差, 其中目标 Q 值通过 Critic Target 网络中的最小 Q 值计算得到, 并加入熵惩罚项. 通过反向传播更新 Critic 网络的参数, 提高策略评估的准确性.

4) Actor 网络更新: Actor 网络的损失函数包括熵系数和 Q 值. 损失函数鼓励策略选择能够最大化最小 Q 值的动作. 通过反向传播更新 Actor 网络的参数, 优化策略的期望回报.

5) 软更新 Critic Target 网络: 每隔一段时间, 采用软更新机制, 以加权平均的方式更新 Critic Target 网络的参数, 从而保证训练过程的稳定性.

6) RND 网络更新: 利用最小化预测网络和目标网络之间的均方误差来更新 RND 中的预测网络, 增强策略的探索行为, 这有助于在较为稀疏奖励环境中提供额外的内在奖励.

7) 内在奖励写入: 将 RND 计算出的内在奖励r_{int}与原始奖励r_t相加, 形成总奖励. 新的奖励被写入经验

回放池, 以便未来的训练中使用.

3 仿真及实验研究

3.1 仿真实验场景

AirSim 是由微软公司开发的基于虚幻引擎的无人机仿真平台, 因其开源、跨平台以及支持多种仿真定制化设置的特性, 逐渐成为深度强化学习、自动驾驶等领域的重要工具平台[17]. 在本文的研究中, 选择以 AirSim 为基础, 构建一个自定义的虚幻引擎环境, 进行三维环境下无人机自主避障任务的仿真研究.

无人机作为一种小型、灵活的飞行器, 常需要在复杂的三维环境中进行机动. 为了更好地模拟复杂的实际环境, 本文设计了一个多样化且具有挑战性的飞行场景如图 3 所示. 具体设置如下.

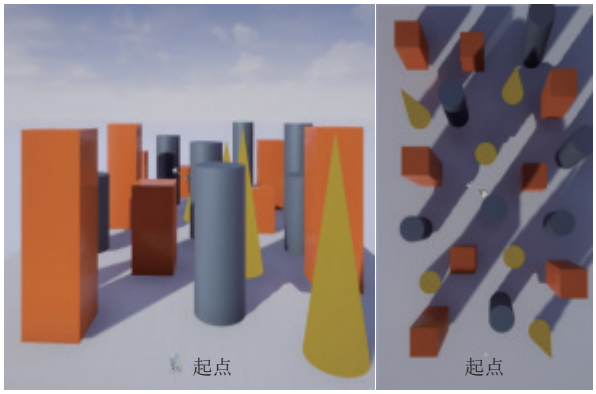

图 3 训练环境

1) 实验环境: 为一个平面区域, 尺寸为 30 m × 50 m, 模拟了一个包含障碍物的三维空间. 在该区域中, 飞行路径的复杂性增加, 通过引入不同类型的障碍物来增强环境的复杂性.

2) 障碍物设置: 为了更真实地反映现实中的障碍物类型, 实验中使用了 3 种不同形状的立体障碍物, 这些障碍物的大小和位置都进行了随机化. 具体包括: 直径为 3 m 的柱体、圆柱体以及底部直径为 3 m 的圆锥体. 障碍物的高度和类型随机放置在环境中, 从而考验无人机的自主避障能力.

3) 无人机飞行区域及任务: 无人机的飞行区域限定在水平方向 30 m, 垂直方向 10 m 的范围内. 在这个限定区域内, 无人机需要飞行并成功避开随机分布的障碍物通过障碍物区域, 保证飞行过程中不发生碰撞.

3.2 状态空间和动作空间设计

本文参照大疆精灵 4RTX 小型多旋翼高精度航测无人机的技术参数范围, 并考虑本文实验环境的特性, 设定 AirSim 中无人机的尺寸为 0.4 m×0.4 m×0.4 m, 飞行速度v_x设定为 3 m/s, 深度摄像头的水平视场角 (FOV) 为 60°.

状态空间: 在本文的仿真实验中, AirSim 无人机的状态空间由无人机搭载的深度相机捕获的深度图像经过预处理后构成. 为了在保证低延迟要求的同时提高环境感知的有效性, 摄像头的分辨率设定为 128×128, 深度感知的最大距离为 10 m. 如图 4 所示, 根据设定的响应阈值 (7 m, 像素值为 178) 进行二值化处理, 得到一个反映环境中障碍物信息的二值化图像. 该图像作为特征提取网络的输入, 经过卷积层和注意力机制处理后, 最终生成环境状态的特征表示. 通过这一设计, 保证了在较小的计算资源和低延迟要求下, 能够有效提升环境感知能力, 同时确保任务的执行效率.

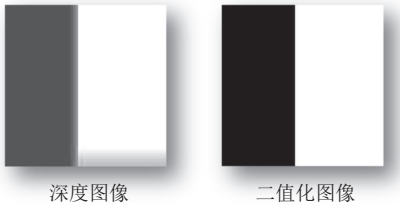

图 4 深度图预处理

动作空间: 无人机在复杂环境中的机动需求使得其动作空间需设计为连续空间. 为了简化并突出任务目标, 本文设定无人机的前进速度为 3 m/s, 动作空间由两个维度组成: 水平速度和垂直速度. 这些动作能够涵盖无人机在三维环境中进行多样化机动的需求, 并且 SAC 算法能够有效处理这种连续动作空间的决策问题. 表 1 列出了无人机动作空间的具体设计.

表 1 动作空间表

动作	动作范围	动作描述
v_y	[−2 m/s, 2 m/s]	水平速度
v_z	[−2 m/s, 2 m/s]	垂直速度

3.3 奖励函数设计

本文的研究目的是确保无人机在飞行过程中能够自主避开障碍物, 成功通过复杂的三维环境. 为了反映这一任务特性, 设计了一个非稀疏奖励函数, 以加速强化学习的训练并提高收敛性. 奖励函数包括碰撞/越界惩罚、边界奖惩、风险状态转移奖惩和到达目的地奖励. 具体设计如下.

1) 碰撞/越界惩罚: 若无人机发生碰撞或飞行越过

指定区域, 给予较大的惩戒: $r_c = -50$.

2) 边界奖惩: 在限定的飞行区域内, 将水平方向中心 10 m, 垂直方向中心 4 m 内区域设为最佳飞行区域, 在该区域内根据位置信息给予正向奖励, 其他区域给予负向惩罚:

$$r_b = (5 - |y|) + (7 - |z|) \tag{9}$$

3) 风险状态转移奖惩: 选择二值化处理后的图像中心区域 (32×32) 作为碰撞窗口. 若该区域内存在像素值 1, 表示当前状态存在碰撞风险, 否则, 表示当前状态无碰撞风险. 根据上一时刻和当前时刻的风险状态转移, 定义风险状态转移奖惩r_r如表 2 所示.

表 2 风险状态转移奖惩表

状态转移	奖惩值
有风险→有风险	−10
有风险→无风险	8
无风险→有风险	−8
无风险→无风险	5

4) 到达目的地奖励: 当无人机成功避开障碍物并完成任务时, 给予较大的奖励: $r_s = 100$.

综合以上 4 项, 最终的奖励函数定义为:

$$R = f_c \times r_c + 0.8 \times r_b + r_r + f_s \times r_s \tag{10}$$

其中, f_c与f_s分别代表是否发生碰撞/越界或成功标识, 该奖励函数充分反映任务特性, 能够有效鼓励无人机在遇到碰撞风险时自主采取躲避动作, 并最终成功完成避障任务.

3.4 仿真结果分析

在本实验中, 分别对 3 种算法进行了训练: 使用简单 CNN 网络的 SAC 算法、引入注意力机制的 SE-SAC 算法, 以及引入 RND 和 SE 机制的 EFRE-SAC 算法. SAC 算法凭借其高效的策略优化能力和稳定的训练过程, 广泛应用于连续控制任务, 采用了最大熵策略, 能够在考虑奖励的同时, 优化策略的随机性, 达到更好的探索性. 然而, SAC 的探索过程依赖于环境奖励的稀疏性和广度, 可能导致在复杂环境中探索不足. SE-SAC 引入了注意力机制, 旨在提高环境感知能力, 特别是对于复杂场景中的障碍物识别, 通过让模型关注重要的区域, SE-SAC 改善了 SAC 在感知和决策上的准确性, 尽管 SE-SAC 优化了感知模块, 依然缺乏有效的探索策略. EFRE-SAC 不仅继承了 SE-SAC 在感知上的优势, 而且通过 RND 机制进一步优化了探索策略, 使得模型能够更好地应对高维度、多障碍物的环境, 尤其是在障碍物形态和分布随机的情况下.

实验中, 经验回放池大小设置为 50 000, 批量采样大小为 64. 为了保证训练的稳定性, 在经验池中积累至少 1 000 步的数据后才开始训练. 我们对这 3 种算法的性能进行了对比, 重点分析了它们在复杂避障任务中的表现差异. 根据图 5、图 6 和图 7 的仿真结果, 可以从 3 个方面对 3 种算法 (SAC、SE-SAC、EFRE-SAC) 进行分析.

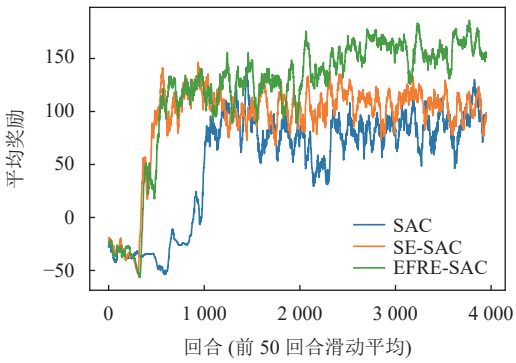

图 5 平均奖励值变化曲线

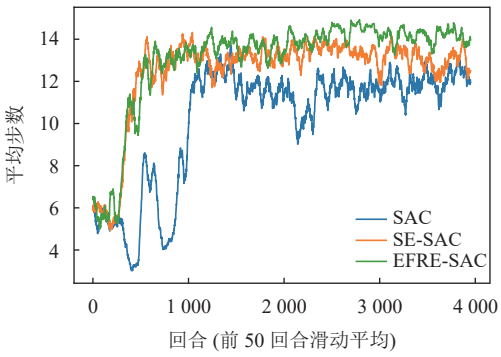

图 6 平均步数变化曲线

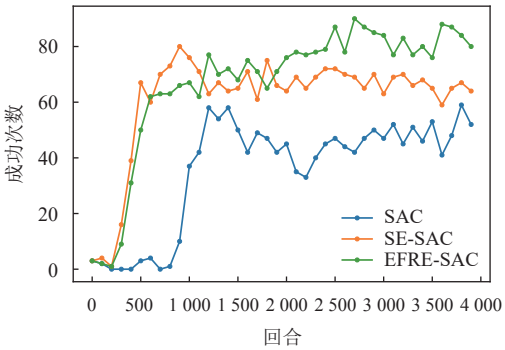

图 7 成功率变化曲线

图 5 展示了每前 50 回合的平均总得分. 从图 5 中趋势可以看出, SAC 算法在训练初期表现出较慢的上升速度和较大的波动性, 在约 2 500 回合后, 随着训练

的继续, 得分逐渐收敛在 50–100 分之间, 说明 SAC 模型在环境的特征提取上存在不足, 导致探索能力和对环境信息的捕捉不充分. 由于使用的简单 CNN 结构无法有效地提取环境中的关键特征, 进而影响了探索过程的效果, 最终未能实现显著的性能提升, 在较低得分水平收敛. 相比之下, SE-SAC 算法通过引入 SE 模块, 对环境的特征提取进行了优化. 经过约 300 回合的数据采集后, 上升速度明显加快, 且波动较小, 稳定性得到了显著提高. 在约 2 000 回合后, 得分维持在 100 左右, 尽管比 SAC 算法有所提升, 但仍未达到最佳状态. SE 模块对深度图像尤其是在二值化处理后的图像, 通过自适应调整通道的重要性, 强化了有用信息的表达, 由于对环境的探索仍然有限, 总奖励值还有待提高. 与 SE-SAC 相比, EFRE-SE 算法的得分上升速度较慢, 在约 1 000–2 000 回合维持在 100–150 分之间, 表明通过结合 RND 机制增强了对新环境的有效探索能力, 在前期对环境的有效探索和信息提取后, 开始在后期稳定上升, 在约 3 000 回合时收敛至约 150 分, 展现了相较于 SE-SAC 算法, 通过内在奖励机制的引入, 使得探索过程更加多样化, 最终获得了更高的性能表现.

图 6 展示了每前 50 回合的平均步数, 趋势与图 5 基本相同, 在任务完成时, 步数基本达到 15, 进一步验证了 EFRE-SAC 相较于 SAC 和 SE-SAC 的稳定性和优化效果.

图 7 展示了 3 种算法在每 100 回合的避障成功率. SAC 算法的成功率收敛后约在 50% 处波动, 表明在任务完成度上的低效率. SE-SAC 算法的成功率提升至约 65%, 证明 SE 模块在提升感知和决策能力方面的作用. EFRE-SAC 的成功率稳定在 85% 左右, 显著优于 SAC 和 SE-SAC, 证明引入 RND 网络对算法在复杂环境中的有效探索和任务完成度的提升.

在训练完成后, 本文对 SAC、SE-SAC 与 EFRE-SAC 算法进行了 500 回合的测试, 验证算法的有效性. 如图 8 是测试环境, 障碍物设置与训练环境具有较大差异, 飞行区域与训练环境保持一致. 表 3 展示了训练 500 回合后, 3 种算法在测试环境中的平均奖励值、平均步数、避障成功率以及成功避障时的平均奖励值 (成功回报) 这 4 个指标上的对比.

可以看出, EFRE-SAC 算法在 4 个指标上均优于 SAC 和 SE-SAC 算法, 避障成功率具有较大的提升, 表现出更强的自主避障能力和环境适应性. 测试结果进一步验证了 SE 模块增强了模型的特征提取能力, RND 网络提升了探索性并加速了训练过程, 最终提高了算法的稳定性和自主避障性能.

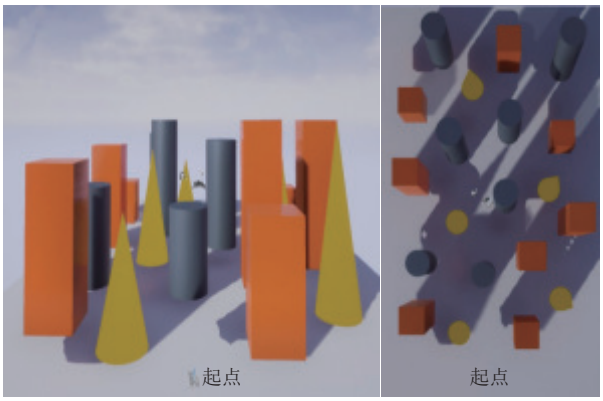

图 8 测试环境

表 3 测试环境指标对比

算法	平均奖励	平均步数	成功率 (%)	成功回报
SAC	59.77	10.44	32.40	184.41
SE-SAC	86.36	11.73	43.60	193.22
EFRE-SAC	154.37	13.64	67.00	217.37

4 结论

本文以无人机在未知三维环境中的自主避障为研究背景, 针对传统强化学习算法在高维状态空间、稀疏信息以及低效探索等方面的局限性, 提出了一种基于注意力机制 SE 和随机网络蒸馏 RND 模块改进的 SAC 算法——EFRE-SAC. 通过引入 SE 模块, 增强了对图像中重要信息的关注能力, 提升了特征表达的能力; 而 RND 模块的引入则鼓励探索新的环境状态, 提升了样本收集的多样性和效率. 本文研究基于 EFRE-SAC 算法, 设计了适应自主避障任务的状态空间、连续动作空间和非稀疏奖励函数. 通过仿真实验证, 所提出的算法能够有效应对高维状态空间下的自主避障任务, 并显著提高了收敛速度和训练稳定性. 实验结果表明, EFRE-SAC 在增强特征表达和探索效率方面具有显著优势, 能够显著提升无人机在复杂三维环境中的自主避障能力.

尽管 EFRE-SAC 算法在增强特征表达和探索效率方面取得了显著进展, 但算法在实际应用场景的表现尚未验证, 由于现实环境中存在更多的噪声和不可预测性, 算法的泛化能力需要进一步评估和改进. 未来可在真实环境中进一步探索, 提升其在复杂三维环境中

的自主避障能力和实际应用价值.

参考文献

1 冯海宽, 陶惠林, 赵钰, 等. 利用无人机高光谱估算冬小麦叶绿素含量. 光谱学与光谱分析, 2022, 42(11): 3575–3580.

2 孟姗姗, 郭秀萍. 卡车-无人机联合取送货模式下物流优化. 系统管理学报, 2022, 31(3): 555–566. [doi: 10.3969/j.issn 1005-2542.2022.03.013]

3 Liu CJ, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing, 2002, 11(4): 467–476.

4 Mac TT, Copot C, Hernandez A, *et al*. Improved potential field method for unknown obstacle avoidance using UAV in indoor environment. Proceedings of the 14th IEEE International Symposium on Applied Machine Intelligence and Informatics (SAMI). Herlany: IEEE, 2016. 345–350.

5 Alharbi A, Poujade A, Malandrakis K, *et al*. Rule-based conflict management for unmanned traffic management scenarios. Proceedings of the 39th AIAA/IEEE Digital Avionics Systems Conference (DASC). San Antonio: IEEE, 2020. 1–10.

6 高九州, 徐威峰, 张立辉, 等. 基于改进 A*算法的无人机避障航线规划. 现代电子技术, 2023, 46(8): 181–186. [doi: 10.16652/j.issn.1004-373x.2023.08.032]

7 李克玉, 陆永耕, 鲍世通, 等. 基于改进 RRT 算法的无人机三维避障规划. 计算机仿真, 2021, 38(8): 59–63, 96. [doi: 10.3969/j.issn.1006-9348.2021.08.011]

8 Lindqvist B, Mansouri SS, Agha-Mohammadi AA, *et al*. Nonlinear MPC for collision avoidance and control of UAVs with dynamic obstacles. IEEE Robotics and Automation Letters, 2020, 5(4): 6001–6008.

9 代进进, 李相民, 薄宁, 等. 基于模型预测控制的无人机避障路径规划方法. 火力与指挥控制, 2020, 45(1): 114–119. [doi: 10.3969/j.issn.1002-0640.2020.01.023]

10 Sonny A, Yeduri SR, Cenkeramaddi LR. Q-learning-based unmanned aerial vehicle path planning with dynamic obstacle avoidance. Applied Soft Computing, 2023, 147: 110773. [doi: 10.1016/j.asoc.2023.110773]

11 Duryea E, Ganger M, Hu W. Exploring deep reinforcement learning with multi Q-learning. Intelligent Control and Automation, 2016, 7(4): 129–144. [doi: 10.4236/ica.2016.74012]

12 Xu GQ, Jiang WL, Wang ZL, *et al*. Autonomous obstacle avoidance and target tracking of UAV based on deep reinforcement learning. Journal of Intelligent & Robotic Systems, 2022, 104(4): 60.

13 Zhang ST, Li YB, Dong QH. Autonomous navigation of UAV in multi-obstacle environments based on a deep reinforcement learning approach. Applied Soft Computing, 2022, 115: 108194.

14 Xue ZH, Gonsalves T. Vision based drone obstacle avoidance by deep reinforcement learning. AI, 2021, 2(3): 366–380.

15 Burda Y, Edwards H, Storkey AJ, *et al*. Exploration by random network distillation. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.

16 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

17 Shah S, Dey D, Lovett C, *et al*. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. Proceedings of the 11th International Conference on Field and Service Robotics. Cham: Springer, 2018. 621–635.

(校对责编: 张重毅)