Journal of System Simulation

Volume 34 | Issue 6 Article 11

6-16-2022

Multi-UAVs 3D Path Planning Method Based on Random Strategy Search

Sen Zhang
College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China;,
zhangsen_hust@163.com

Mengyan Zhang
College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China;

Jingping Shao
College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China;

Jiexin Pu
College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China;

Multi-UAVs 3D Path Planning Method Based on Random Strategy Search

Abstract

Abstract: In view of the difficulty of the traditional path planning method without energy consumption constraints to meet the emergency rescue requirements in the complex mountain operation environment, *a three-dimensional path planning algorithm for multi-UAVs is proposed based on LSTM-DPPO(long short-term memory-distributed proximal policy optimization) framework. The LSTM long and short-term memory neural network is used to extract the important characteristic state information sequence of the multiple unmanned aerial vehicles in their respective flight process. After repeated iteration and updating, an optimal network parameter model is obtained. Combined with the energy consumption, the optimal 3D detection path is generated.* Simulation experiments verify that the proposed method is more effective than the traditional path planning method and can plan the optimal detection path with the minimum energy consumption.

Keywords

multi-UAVs, deep reinforcement learning algorithms, neural network, 3D path planning, energy consumption

Recommended Citation

Sen Zhang, Mengyan Zhang, Jingping Shao, Jiexin Pu. Multi-UAVs 3D Path Planning Method Based on Random Strategy Search[J]. Journal of System Simulation, 2022, 34(6): 1286-1295.

第 34 卷第 6 期 系统仿真学报© Vol. 34 No. 6
2022 年 6 月 Journal of System Simulation Jun. 2022

基于随机策略搜索的多机三维路径规划方法

张森，张孟炎，邵敬平，普杰信
(河南科技大学 信息工程学院，河南 洛阳 471023)

摘要：针对传统无能耗约束的多无人机路径规划方法难以适应复杂山地作业环境的应急救援要求，*提出了一种基于LSTM-DPPO(long short-term memory-distributed proximal policy optimization)框架的多无人机三维路径规划算法。利用LSTM长短期记忆神经网络提取出多无人机在各自飞行过程中的重要特征状态信息序列，经过多次迭代更新后得到一个最优网络参数模型，结合能耗生成最优的三维探测路径。*实验结果表明：该方法相对于传统路径规划方法效果显著，能在能耗最小的前提下规划出最优探测路径。

关键词：多无人机；深度强化学习算法；神经网络；三维路径规划；能耗
中图分类号：TP183 文献标志码：A 文章编号：1004-731X(2022)06-1286-10
DOI: 10.16182/j.issn1004731x.joss.21-0112

Multi-UAVs 3D Path Planning Method Based on Random Strategy Search

Zhang Sen, *Zhang Mengyan*, *Shao Jingping*, *Pu Jiexin*
(College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China)

Abstract: In view of the difficulty of the traditional path planning method without energy consumption constraints to meet the emergency rescue requirements in the complex mountain operation environment, *a three-dimensional path planning algorithm for multi-UAVs is proposed based on LSTM-DPPO(long short-term memory-distributed proximal policy optimization) framework. The LSTM long and short-term memory neural network is used to extract the important characteristic state information sequence of the multiple unmanned aerial vehicles in their respective flight process. After repeated iteration and updating, an optimal network parameter model is obtained. Combined with the energy consumption, the optimal 3D detection path is generated*. Simulation experiments verify that the proposed method is more effective than the traditional path planning method and can plan the optimal detection path with the minimum energy consumption.

Keywords: multi-UAVs; deep reinforcement learning algorithms; neural network; 3D path planning; energy consumption

引言

地震通常会造成受灾区域交通道路及建筑物的严重损坏。此时，救灾人员能否第一时间到达现场起着至关重要的作用。随着无人机探测技术的快速发展，利用其快速获取、评估受灾情况成为可能。当灾情发生时，无人机可以在第一时间对受灾区域进行合理探测。此时，合理的路径规划显得尤为重要。从目前看来，虽然学者在无人机路径规划方面开展了大量研究，但仍有很多难点需要突破。

Radmanesh，Mohammadreza 等[1]提出了一种

收稿日期：2021-02-05 修回日期：2021-06-16
第一作者：张森(1984-)，男，博士，副教授，研究方向为智能机器人控制，水下光场建模与仿真，智能图像处理。
E-mail：zhangsen_hust@163.com

GWO(grey wolf optimization)路径优化算法，该算法通过贝叶斯变化和基于距离的价值函数的单元加权，实现了最优路径规划的有效性；阚平等[2]提出了一种改进的粒子群优化算法，在保证各架植保无人机的补给时间满足间隔分布的约束下，构建目标函数，成功验证了路径规划算法的可行性；戴健等[3]采用均衡划分和凹点凸分解法相结合，通过“Z”字形覆盖法以及 Dubins 转弯路径来实现多无人机最优路径的搜索，并验证了该方法的有效性和实用性；Yoon 等[4]提出将交会路径规划与导引律路径算法相结合，有效解决了无人机空中加油的路径规划问题，并通过数值仿真实验验证了该方法的有效性；Yang 等[5]提出了一种在时间和角度约束下的毕达哥拉斯(PH)速度曲线图的路径规划方法。该方法通过考虑无人机路径曲率约束，将多无人机到达目标点的角度协同作为路径规划的初始条件，并通过仿真验证了在时间和角度约束下的 PH 曲线多无人机路径协同规划的可行性。

本文针对多无人机在三维复杂环境下的最优路径规划问题，提出了一种将 LSTM(long short term memory)神经网络与随机策略搜索算法 DPPO (distributed proximal policy optimization)相结合的架构。DPPO 算法是在 PPO 算法的基础上进行多线程分支计算[6]，很大程度上提高了 PPO 算法的运算性能。三维仿真地形通过 Matlab 仿真软件进行搭建。经过多次训练迭代得出最优路径，并与其他三维路径算法对比，进一步验证本文算法的有效性和合理性。

1 三维路径规划算法

目前常用的路径规划算法主要为：Dijkstra 算法[7]、A*算法[8]、粒子群算法[9]、蚁群算法[10]和遗传算法[11]等。二维路径算法是指基于二维平面，即 x–y 平面上的路径规划[12]。图 1 为常见的二维路径规划图，三维仿真环境如图 2 所示。无人机碰到山丘后，有①和② 2 种越过山丘的路线，此时可通过计算无人机的能耗值来判断、控制无人机按照哪种路线飞过山丘。

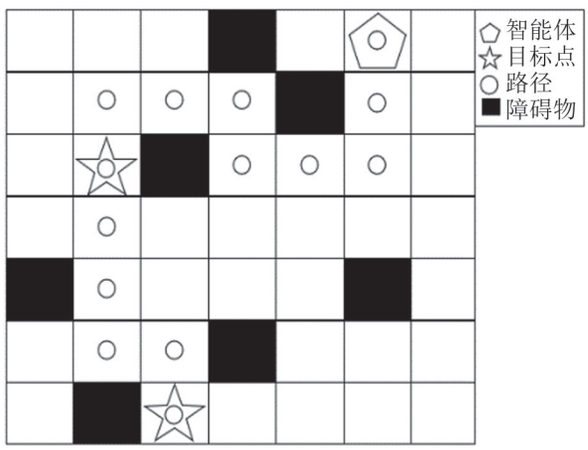


图 1 二维路径规划图
Fig. 1 Two-dimensional path planning

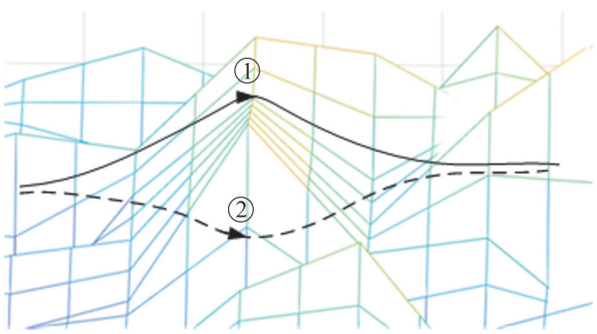


图 2 三维路径规划图
Fig. 2 Three-dimensional path planning

2 多无人机协同探测建模

针对多无人机在三维环境下探测多个目标点的问题，首先设置无人机起飞点、降落点及探测点的相应位置坐标[13]。其中，无人机设为 $v_n(n=1,2,3,4,5)$，节点 $u_n=(u_0,u_1,\cdots,u_{11})$，$u_0$ 为起飞点，坐标为 (x_o,y_o,z_o)，u_{11} 为降落点，坐标为 (x_d,y_d,z_d)，目标点 $u_i\in u(i=1,2,\cdots,10)$ 的坐标为 (x_i,y_i,z_i)。路线集合为 $\eta^{v_n}=\{(i,j)|i,j\in u\}$，表示每个无人机 v_n 的探测路径。定义多个无人机协同探查多个目标点的决策变量为

DOI: 10.16182/j.issn1004731x.joss.21-0112

$$x_{i,j}^{v_n}=\begin{cases}1, & \text{无人机}v_n\text{从}i\text{飞到}j\\ 0, & \text{否则}\end{cases} \tag{1}$$

探测模型可设为

$$\min \mathrm{L}(\eta)=af_1+bf_2,\ 0<a,b<a+b=1 \tag{2}$$

式中：f_1，f_2为2个优化函数；a，b为优化函数在损失函数$\mathrm{L}(\eta)$中的比例系数。

每架无人机成功探测目标点一次：

$$\sum_{i=1}^{u_i}\sum_{v_n=1}^{v_n}x_{i,j}^{v_n}=1,\ \forall j=1,2,\cdots,u_i \tag{3}$$

每架无人机离开节点一次：

$$\sum_{j=1}^{u_i}\sum_{v_n=1}^{v_n}x_{i,j}^{v_n}=1,\ \forall i=1,2,\cdots,u_i \tag{4}$$

多个无人机在完成各自探测任务的前提下，航程越短越好。无人机的航程不能超过其最大里程：

$$s_{i,j}x_{i,j}^{v_n}\leqslant G,\ \forall i,j\in u_i,i\neq j \tag{5}$$

式中：$s_{i,j}$为节点i~j之间的路程。

所有无人机的总飞行路程，其最终值越短越好，是多无人机探测目标点的优化函数之一：

$$\min f_1=\sum_{i=1}^{u}\sum_{j=1}^{u}\sum_{v_n=1}^{v_n}x_{i,j}^{v_n}s_{i,j} \tag{6}$$

在多无人机协同探测过程中，由于单个无人机的执行能力有限，需要通过式(7)来引入节点方差，调节各个无人机的任务负载：

$$\min f_2=\mathrm{var}\,(\eta^{v_n}) \tag{7}$$

3 深度强化学习算法

深度强化学习算法主要由神经网络和强化学习算法2部分组成[14]。现有的强化学习算法主要可分为2类：基于策略(policy-based，RL)和基于价值(value-based，RL)[15]。本文研究的是多无人机路径规划，PPO和DPPO算法[16]均有很好的探索性，因此较为符合本文研究。

3.1 PPO算法

策略梯度算法的难点在于步长的选择[17]，在训练过程中新旧策略的变化差异过大不利于智能体的学习[18]。而PPO算法很好地解决了步长这一问题。在Policy Grandient算法中，可以将每个无人机设为actor，该算法将策略参数化为c_θ^n，如式(8)所示，无人机数量设为$v_n(n=1,2,3,4,5)$架，且每个回合eposide设有T个时间步，即：

$$\begin{aligned}&c_\theta^n=[s_1^n,a_1^n,s_2^n,a_2^n,\cdots,s_T^n,a_T^n]\\&V=[v_1,v_2,\cdots,v_n]\end{aligned} \tag{8}$$

由于无人机在不同状态下所采取的动作可能是不同的，因此一个序列τ的发生概率为

$$\begin{aligned}p_\theta^{v_n}(\tau)=&p^{v_n}(s_1^{v_n})p_\theta^{v_n}(a_1^{v_n}s_1^{v_n})p^{v_n}(s_2^{v_n}s_1^{v_n},a_1^{v_n})\\&p_\theta^{v_n}(a_2^{v_n}s_2^{v_n})p^{v_n}(s_1^{v_n}s_2^{v_n},a_2^{v_n})\cdots=\\&p^{v_n}(s_1^{v_n})\prod_{t=1}^{T}p_\theta^{v_n}(a_t^{v_n}s_t)p^{v_n}(s_{t+1}^{v_n}s_t^{v_n},a_t^{v_n})\end{aligned} \tag{9}$$

Policy Grandient策略梯度算法不通过误差反向传播，它主要通过奖励值reward来增加或减少下一次选中该动作的概率值。设优势函数为

$$A_\theta^{v_n}=Q_{\mathrm{c}}^{v_n}(s_t^{v_n},a_t^{v_n})-V_{\mathrm{c}}^{v_n}(s_t^{v_n}) \tag{10}$$

式中：$V_{\mathrm{c}}^{v_n}(s_t^{v_n})$为无人机的状态价值函数；$Q_{\mathrm{c}}^{v_n}(s,a)$为每个无人机当前动作所对应的值函数。

Policy Grandient算法的模型表示为

$$\begin{aligned}\nabla\hat{R}_\theta^{v_n}=&\frac{1}{Z}\sum_{z=1}^{Z}\sum_{t=1}^{T}A_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})\nabla\lg(p_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n}))=\\&E_{x\sim p^{v_n}(\theta)}[A_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})\nabla\lg(p_\theta^{v}(a_t^{zv_n}s_t^{zv_n})]\end{aligned} \tag{11}$$

PPO算法架构如图3所示。

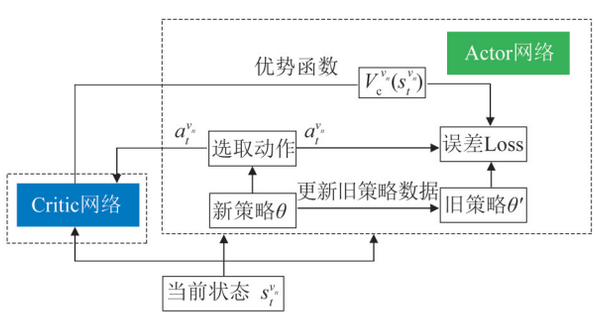


图3 PPO算法结构框架

Fig. 3 PPO algorithm structure framework

首先，多无人机各自状态$s_t^{v_n}$分别输入到critic网络、actor网络中的新策略网络θ和旧策略网络θ'中。然后新策略网络θ将自身参数复制给旧策略θ'，无人机与环境交互并在新策略θ中产生动作$a_t^{v_n}$，将动作传送给critic网络，并作为误差网络分

析的数据。critic网络根据输入的$s_t^{v_n}$和$a_t^{v_n}$信息计算出优势函数$A_\theta^{v_n}$。新策略网络θ通过设置每走t步来更新一次旧策略θ'。通过梯度下降法实现策略的最优化。

PPO算法的核心理念是重采样，如果采样序列p的分布无法直接计算时，可另假设一个采样序列来近似p分布：

$$p_\theta^{v_n}(\tau)=p^{v_n}(s_1^{v_n})p_\theta^{v_n}(a_1^{v_n}s_1^{v_n})p^{v_n}(s_2^{v_n}s_1^{v_n},a_1^{v_n})\\ p_\theta^{v_n}(a_2^{v_n}s_2^{v_n})p^{v_n}(s_1^{v_n}s_2^{v_n},a_2^{v_n})\cdots=\\ p^{v_n}(s_1^{v_n})\prod_{t=1}^{T}p_\theta^{v_n}(a_t^{v_n}s_t)p^{v_n}(s_{t+1}^{v_n}s_t^{v_n},a_t^{v_n}) \tag{12}$$

式中：$p_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})p_{\theta'}^{v_n}(a_t^{zv_n}s_t^{zv_n})$为权重因子。

回报函数为

$$\nabla\hat{R}_\theta^{v_n}=E_{(s_t^{v_n},a_t^{v_n})\sim c_\theta^{v_n}}[A_\theta(a_t^{zv_n}s_t^{zv_n})\nabla\lg(p_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})]=\\ E_{(s_t^{v_n},a_t^{v_n})\sim c_\theta^{v_n}}[\frac{p_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})}{p_{\theta'}^{v_n}(a_t^{zv_n}s_t^{zv_n})}A_\theta^{v_n}(a_t^{zv_n}s_t^{zv_n})] \tag{13}$$

定义好该模型后，为了训练模型，可定义损失函数为

$$J_{\theta'}^{n}(\theta)=E_{(s_t^{v_n},a_t^{v_n})\sim c_\theta^{v_n}}[A_\theta(a_t^{zv_n}\mathrm{s}_t^{zv_n})\nabla\lg(p_\theta^{v_n}(a_t^{zv_n}\mathrm{s}_t^{zv_n})]=\\ E_{(s_t^{v_n},a_t^{v_n})\sim c_\theta^{v_n}}\left[\frac{p_\theta^{v_n}(a_t^{zv_n}\mathrm{s}_t^{zv_n})}{p_{\theta'}^{v_n}(a_t^{zv_n}\mathrm{s}_t^{zv_n})}(\mathrm{L}(\eta)-b(\mathrm{s}_t^{zv_n}))\right] \tag{14}$$

在推导回报奖励的过程中，已假设新旧策略的分布度接近，因此增加一个约束值KL：

$$J_{\mathrm{PPO}}^{\theta'}(\theta)=J^{\theta'}(\theta)-\alpha KL(\theta,\theta') \tag{15}$$

KL为散度也叫相对熵，新旧策略的偏差比较大时，为了防止步长大幅度增长而导致波动性，将KL散度作为惩罚项，有效抑制步长，保证步长的稳定性。

3.2 DPPO算法

DPPO是将PPO算法进行了多线程处理[19]，如图4所示。在PPO算法的基础上分支出n个线程区域，每个线程独立与环境进行交互，采集相关数据，并运算梯度。当n个线程完成各自的梯度计算后，便可将各自的数据一并传输至全域网，供全域网参数的快速更新。DPPO可以避免experience间的相关性，所以明显优于PPO算法。

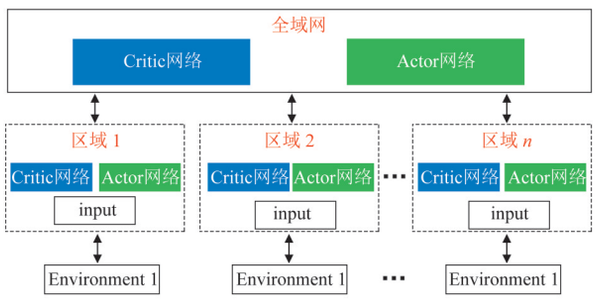


图4 DPPO算法结构框架
Fig. 4 DPPO algorithm structure framework

4 三维路径算法设计

4.1 LSTM神经网络

在传统神经网络中，NN(neural network)模型会受到短期记忆的影响，如果一条序列足够长，则早期的信息将很难传送至后面的时间步[20]。因此，很可能会丢失一些重要信息。LSTM的核心理念在于“门”，分别为遗忘门、输入门和输出门[21]。如图5所示，激活函数主要有Sigmoid函数和tanh函数[22]。

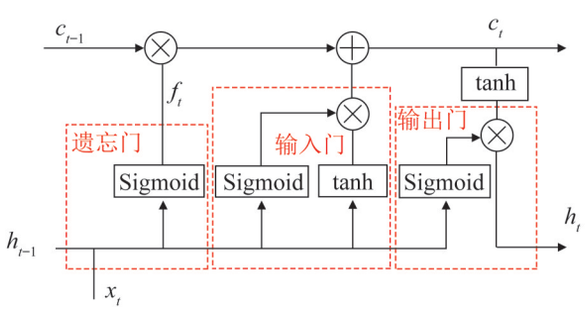


图5 LSTM神经网络框架
Fig. 5 LSTM neural network framework

4.2 LSTM-DPPO算法

图6为LSTM-DPPO算法结构框架，LSTM为循环神经网络，可以处理连续动作信息。每个无人机将自己此刻的状态信息s_i依次传送至LSTM神经网络中，首先，通过遗忘门去除状态信息s_i中存在的一些无用信息，然后通过输入门更新有

效信息，最后由输出门输出有效信息并保存至隐藏状态中，并将处理过后的状态信息编码成一个固定序列，发送至DPPO网络中进行训练。

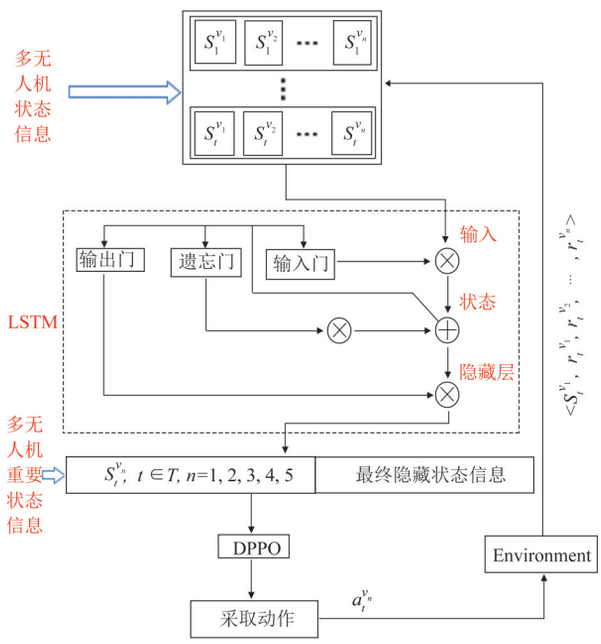


图6 LSTM-DPPO算法结构框架
Fig. 6 LSTM-DPPO algorithm structure framework

4.3 状态空间

状态信息为无人机在每个时刻的瞬时速度$[v_{xa}, v_{ya}, v_{za}]$、所处位置$[p_{xd}, p_{yd}, p_{zd}]$、目标点所在位置$[p_{xg}, p_{yg}, p_{zg}]$、障碍物的位置$[p_{xo}, p_{yo}, p_{zo}]$，每个无人机设置一个等距的测距射线，设最长测距为$l_{\max}$，通过射线测得的障碍物信息可定义为$[l_i, \alpha, l_t, l_h]$，其中，l_i和α为无人机发出的射线被障碍遮挡后的长度和角度；l_t为障碍物类型；l_h为障碍物的垂直高度。因此可定义输入状态为

$$s_i = \{ v_{xa}, v_{ya}, v_{za}, p_{xd}, p_{yd}, p_{zd}, p_{xo}, p_{yo}, p_{zo}, l_i, \alpha, l_t, l_h \} \tag{16}$$

4.4 动作空间

多无人机的动作为联合动作[23]。每架无人机需要通过分享各自的动作信息来作为更新策略的依据。

无人机的动作空间可分为前向运动、纵向运动、侧向运动和悬停。前向运动分为前进1和后退2，纵向运动分为上升3和下降4，侧向运动分为左转5和右转6，悬停为7。

4.5 奖罚函数

本文研究的是多无人机三维空间的路径规划，与二维相比，障碍物均为不规则的三维实体，地况信息会更加复杂。根据无人机是否到达目标点、是否与障碍物发生碰撞来合理设置奖励值。基于碰撞的奖励函数值为

$$r_t = \begin{cases} 0, & \text{未发生碰撞也未到达目标点} \\ -10, & \text{与障碍物发生碰撞} \\ 5, & \text{到达目标点} \end{cases} \tag{17}$$

Algorithm 1DPPO

for $i \in \{1, 2, \cdots, N\}$ do

for $w \in \{1, 2, \cdots, T/K\}$ do

run policy $c_\theta$for K time steps, collecting $\{s_t, a_t, r_t\}$for $t \in \{(i-1)K, \cdots, iK-1\}$

estimate return $\hat{R}_t = \sum_{t=(i-1)k}^{iK-1} \gamma^{t-(i-1)K} r_t + \gamma^K V_\phi(s_i K)$

estimate advantage $\bar{A}_t = \bar{R}_t - V_\phi(s_t)$

store partial trajectory information

end for

$c_{\text{old}} \leftarrow c_\theta$

for $m \in \{1, 2, \cdots, M\}$ do

$J_{\text{PPO}}(\theta) = \sum_{t=1}^{T} \frac{c_\theta(a_t|s_t)}{c_{\text{old}}(a_t|s_t)} \bar{A}_t - \lambda KL[c_{\text{old}}|c_\theta] - \xi \max(0, KL[c_{\text{old}}|c_\theta] - 2KL_{\text{target}})^2$

if $KL[c_{\text{old}}|c_\theta] > 4KL_{\text{target}}$ then

break and continue with next outer iteration $i+1$

end if

Compute $\nabla_\theta J_{\text{PPO}}$

send gradient wrt. to θ to chief

wait until gradient accepted or dropped; update

parameters

end for

for $b \in \{1, 2, \cdots, B\}$, do

$L_{\mathrm{BL}}(\phi) = -\sum_{t-1}^{T}(\bar{R}_t - V_\phi(s_t))^2$

Compute $\nabla_\theta L_{\mathrm{BL}}$

send gradient wrt. to θ to chief

wait until gradient accepted or dropped; update parameters

end for

if $KL[c_{\mathrm{old}}|c_\theta] > \beta_{\mathrm{high}} KL_{\mathrm{target}}$ then

$\lambda \leftarrow \bar{\alpha}\lambda$

else if $KL[c_{\mathrm{old}}|c_\theta] < \beta_{\mathrm{high}} KL_{\mathrm{target}}$ then

end if

end for

5 能耗模型

无人机的通用能耗模型[24]为

$$Q_U = \int_{t_d}^{t_f} \sum_{m=1}^{4} T_m(t)\, w_m(t)\, \mathrm{d}t \tag{18}$$

式中：T_m和w_m为无人机4个电机各自的转矩和转速大小，然后将4个电机的量值进行累加；m为无人机的旋翼数量；t_{d}和t_{f}分别为无人机飞行的开始和结束时间点，最后再通过对时间积分来求得无人机的总能耗。

无人机能耗模型的构建如式(19)所示。本文通过将无人机在高空飞行时的空气阻力以及自身材料等能耗因素考虑在内。得EC能耗公式为

$$\begin{aligned} S_U = &\int_{t_d}^{t_f} (q_{\mathrm{a}} + q_{\mathrm{l}} + q_{\mathrm{ud}} + \sum_{m=1}^{4} q_{\mathrm{pro}})\mathrm{d}t = \\ &\int_{t_d}^{t_f} (\frac{1}{2}\rho v_{\mathrm{o}}^3 A c_{\mathrm{air}} + k\xi T_{\mathrm{o}} + mgv_{\mathrm{o}}\sin\gamma + \\ &\sum_{m=1}^{4} \rho R w_m^3 (1 + 2(\frac{v_{\mathrm{o}}}{w_m})^2)\frac{\sigma c_{\mathrm{bd}}}{8})\mathrm{d}t \end{aligned} \tag{19}$$

式中：q_{a}为结合空气阻力等环境因素所产生的功率；ρ为当地探测时的空气密度；v_{o}为无人机飞行时三轴方向上的合速度，是一个实时变量；A为无人机与空气的接触面积；c_{air}为空气阻力；q_{l}为无人机起飞时的功率；k为功率因数；ξ为旋翼转动时的下流值；q_{pro}为无人机飞行时桨叶自身产生的功率；R为无人机旋翼的桨叶面积；σ为旋翼的刚度系数；c_{bd}为桨叶的阻力系数。4个旋翼所产生拉力的合力：

$$T_{\mathrm{o}} = \sqrt{m^2 g^2 + D_{\mathrm{B}}^2 + 2D_{\mathrm{B}} q_{\mathrm{ud}}} \tag{20}$$

式中：q_{ud}为无人机改变当时状态，以飞行角度γ爬升或下降时产生的功率；D_{B}为机体阻力，表示为

$$D_{\mathrm{B}} = \frac{1}{2}\rho v_{\mathrm{o}}^2 A c_{\mathrm{air}} \tag{21}$$

$$v_{\mathrm{o}} = \sqrt{\dot{p}_x^2 + \dot{p}_y^2 + \dot{p}_z^2} \tag{22}$$

为了更结合实际情况，Goeke等[25]得出锂电池的充电效率约为μ_{o}=90%，发动机的转换效率为μ_{self}=92%，传输效率为μ_{tran}=90%。则最终能耗为

$$S_{U_{\mathrm{d}}} = S_U/(\mu_{\mathrm{o}}\mu_{\mathrm{self}}\mu_{\mathrm{tran}}) \tag{23}$$

6 实验验证

6.1 有无能耗约束下的无人机路径规划

为了体现仿真实验对于无人机实飞的有效性和实用性，将实验分为有能耗约束和无能耗约束2种情况，并分别进行Matlab仿真测试。

在该区域中设定10个需要探测的目标，设定无人机抵达相应目标区域后便可完成探测任务，而对于在目标区域内如何完成探测不予考虑。

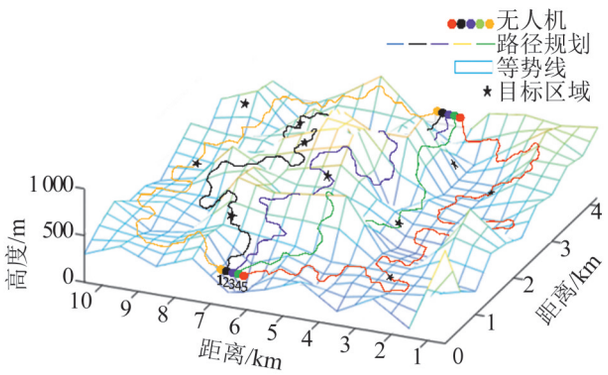
本文选择的对象为多无人机，属于多智能体协同控制系统。在强化学习过程中，由于每架无人机所处的环境都具有复杂度高等特点，导致无人机的状态空间和动作空间会随着无人机数量呈指数型增长，计算难度大幅上升[26]。因此，本文算法主要从无人机的可控性方面考虑，选取5架无人机作为实验对象。

图7为无能耗约束下的路径效果图。5架无人机在没有自身能耗约束的情况下，从起飞点出发，在协同探测完毕后成功抵达降落点。由于没有能耗的约束，使得无人机可以对路径进行相对更加

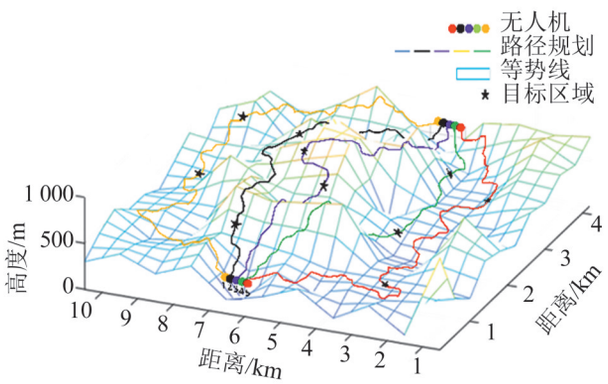
DOI: 10.16182/j.issn1004731x.joss.21-0112

第 34 卷第 6 期 系统仿真学报 Vol. 34 No. 6
2022 年 6 月 Journal of System Simulation Jun. 2022

充分的探索，因此在第300次训练时，5架无人机均成功抵达终点，但与第1 000次训练相比，路径上存在过多的能量消耗。



(a) 训练第300次



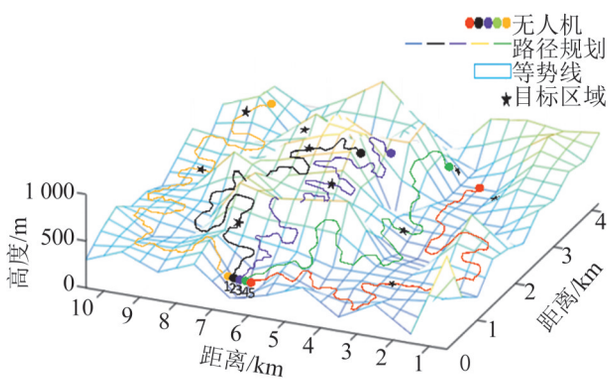
(b) 训练第1 000次

图7 无能耗约束下的多机路径训练

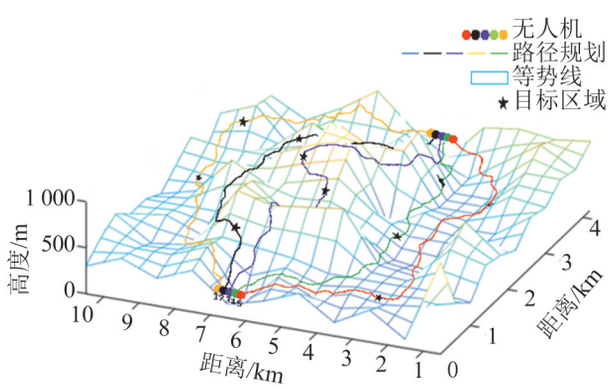
Fig. 7 Multi-machine path training without energy consumption constraints

图8结合了无人机在高空飞行时的实际问题，将能耗量设定为限制条件，分别再次训练至第300次，第1 000次。从图8(a)可以明显看出，在刚开始的训练过程中，由于训练经验有限，导致无人机的飞行动作不够准确，产生了过多的路径损耗，最终均未抵达终点。图8(b)为训练第1 000次后的飞行图示，可以看出，无人机均能到达终点，且路径上并未出现过多的能量消耗。

从表1分析可得，有能耗约束下的无人机总路径明显大于后者，且二者在路径上和目标区域的检测程度上也未进行优化。在训练第1 000次时，有能耗约束下的无人机已覆盖所有目标区域，且从图7~8的路径效果以及表中数据进行计算对比，飞行总路径长度缩短了4.98 km，且有能耗约束的无人机相对于前者节省了8.712×10^3 kJ的能耗，增强了无人机在实飞过程中的续航能力。



(a) 训练第300次



(b) 训练第1 000次

图8 有能耗约束下的多机路径训练

Fig. 8 Multi-machine path training with energy consumption constraint

表1 有能耗约束与无能耗约束下的指标对比

Table 1 Comparison of indicators with and without energy consumption constraints

指标		第300次	第1 000次
所有无人机总路径/km	无能耗约束	76.59	64.12
	有能耗约束	65.23	59.14
已检测目标区域量	无能耗约束	8	10
	有能耗约束	7	10
所有无人机总能耗×10^{-3}/kJ	无能耗约束	42.486	36.128
	有能耗约束	35.000	27.416

6.2 与传统路径算法比较

本文算法与传统算法在三维路径规划上的效

果如图9~10所示。从路径效果上可初步看出二者在能量消耗上的区别。

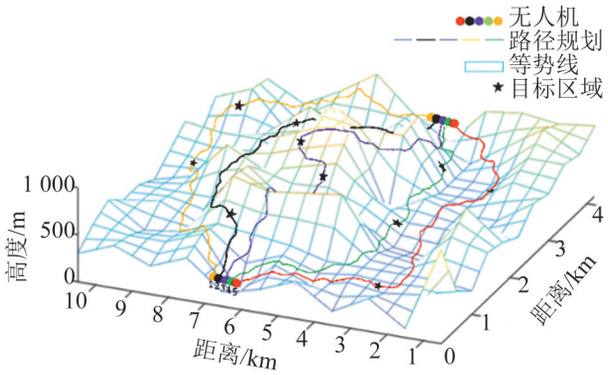


图9 基于LSTM-DDPO算法的三维路径规划
Fig. 9 3D Path Planning Based on LSTM-DDPO Algorithm

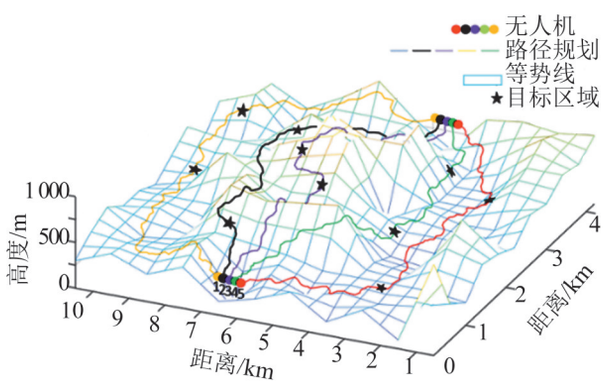
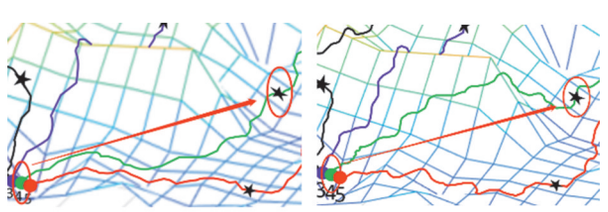


图10 基于A*算法的三维路径规划
Fig. 10 3D Path Planning Based on A* Algorithm

本研究可通过局部相关数据来突出路径优化的效果，如图11所示。其中以第4架无人机在两种算法下分别到达第1个目标区的路径飞行角度以及能耗作为代表，进行进一步的数据分析论证。



(a) 基于LSTM-DDPO算法 (b) 基于A*算法

图11 LSTM-DPPO算法与A*算法局部图对比
Fig. 11 LSTM-DPPO algorithm and A* algorithm local graph comparison

通过表2和图11可看出，无人机是以2种路径飞过了山峰，通过数据比较，本文算法虽然在路程上大于A*算法，但是飞行时间更短，且由于无人机在升高过程中需要更多的能量消耗，因此A*算法相对于本文算法产生了相对更多的能耗。

表2 LSTM-DPPO和A*算法的局部数据对比
Table 2 Local data comparison between LSTM-DPO and A* algorithm

算法	路径/km	飞行时长/min	能耗×10^{-3}/kJ
LSTM-DPPO	5.613	11.25	5.283
A*	5.552	11.38	6.192

从图12可以看出，在训练大约第400次之前，奖励值一直持续增加，说明无人机正在不断的探索环境，在第400次之后，无人机开始对自身的路径效果进行不断优化，并在最终趋于奖励的稳定值。

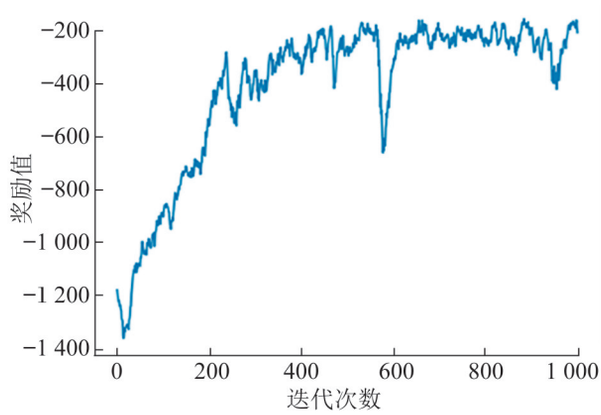
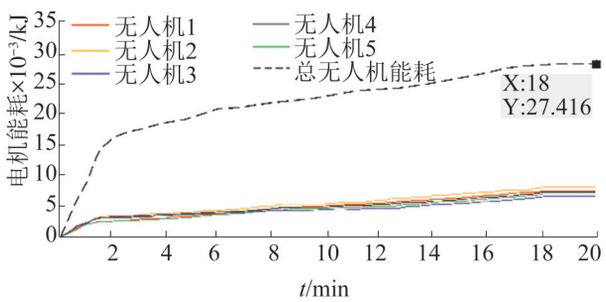


图12 LSTM-DPPO algorithm算法训练过程中的奖励值
Fig. 12 Reward value in the training process of LSTM-DPPO algorithm

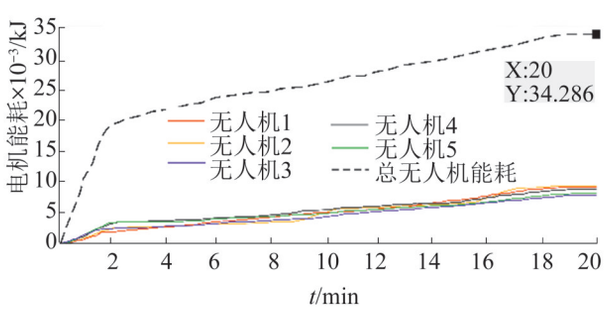
从图13可以看出，基于2种算法下的无人能耗变化值相差不大。且从最终能耗值可看出，本文算法所规划的路径相对于A*算法节省了能耗，极大地优化了无人机的能量消耗，增强了无人机的续航能力。

表3为多种算法下的无人机数据对比。4种算法均对所有目标完成了探索任务。路程上，本文算法仅次于A*算法，且明显小于蚁群算法和

Dijkstra 算法；飞行时长上，本文算法明显小于其他算法；总能耗上，本文算法所规划的路径使无人机产生了最少的能耗值。结合实际出发，本文算法在无人机的飞行时长和飞行能耗上，相对于其他算法均有了显著的优化效果。



(a) LSTM-DPPO算法



(b) A*算法

图13 多无人机能耗曲线图

Fig. 13 Multi-UAVs energy consumption curve

表3 多种算法下的指标对比

Table 3 Comparison of indicators under a variety of algorithms

算法	所有无人机总路径/km	已检测目标区域量	所有飞行时间总长/min	所有无人机总能耗×10⁻³/kJ
LSTM-DPPO	59.14	10	18.75	27.416
A*	58.12	10	19.22	33.286
蚁群	63.85	10	21.27	35.822
Dijkstra	64.67	10	21.87	32.468

7 结论

本文研究了基于三维空间环境下的随机策略梯度算法路径规划问题，并通过将仿真实验与实际环境相结合，对无人机能耗限制与否进行了数据比对，表明了该研究的实用性和可行性。通过将本文算法与传统算法进行比较，分析论证了无人机的路径数据以及能耗变化数据曲线图。实验结果表明：相对于无能耗约束情况，基于有能耗约束下的无人机路径以及能耗值均得到了进一步优化。

本文受限于实验硬件资源，无法对更多架无人机进行测试。虽然运用深度强化学习算法实现了 5 架无人机的最优路径规划，但是尚未讨论该算法支持多智能体的承载能力，因此本文的研究成果尚存在一定的局限性。同时，根据具体实验条件，对本文算法支持多智能体的容量和效率进行更深层次的量化分析，将是下一步的研究重点。

参考文献：

[1] Radmanesh Mohammadreza, Kumar Manish, Sarim Moha Mmad. Grey Wolf Optimization Based Sense and Avoid Algorithm in a Bayesian Framework for Multi-UAVs Path Planning in an Uncertain Environment[J]. Aerospace Science and Technology(S1270-9638), 2018, 77: 168-179.

[2] 阚平, 姜兆亮, 刘玉浩, 等. 多植保无人机协同路径规划[J]. 航空学报, 2020, 41(4): 255-265.
Kan Ping, Jiang Zhaoliang, Liu Yuhao, et al. Collaborative Path Planning for Multi-Plant Protection UAV[J]. Journal of Aviation, 2020, 41(4): 255-265.

[3] 戴健, 许菲, 陈琪锋. 多无人机协同搜索区域划分与路径规划[J]. 航空学报, 2020, 41(增1): 146-153.
Dai Jian, Xu Fei, Chen Qifeng. Multi-UAVs Cooperative Search Area Division and Path Planning[J]. Journal of Aviation, 2020, 41(S1): 146-153.

[4] Yoon Y, Kim M, Kim Y. Three-Dimensional Path Planning for Aerial Refueling Between One Tanker and Multi-UAVs[J]. International Journal of Aeronautical and Space(S2093-274X), 2018, 19(4): 1027-1040.

[5] Yang X, Zhou W, Zhang Y. On Collaborative Path Planning for Multi-UAVs Based on Pythagorean Hodograph curve[C]// Guidance, Navigation & Control Conference. Nanjing: IEEE, 2016: 12-14.

[6] 黄东晋, 蒋晨凤, 韩凯丽. 基于深度强化学习的三维路径规划算法[J]. 计算机工程与应用, 2020, 56(15): 30-36.
Huang Dongjin, Jiang Chenfeng, Han Kaili. 3D Path Planning Algorithm Based on Deep Reinforcement Learning[J]. Computer Engineering and Applications, 2020, 56(15): 30-36.

[7] Sun P, Shan R. Predictive Control with Velocity Observer for Cushion Robot Based on PSO for Path

Planning[J]. Journal of Systems Science & Complexity (S1009-6124), 2020, 33(4): 988-1011.

[8] Votion Johnathan, Cao Yongcan. Diversity-Based Cooperative Multivehicle Path Planning for Risk Management in Costmap Environments[J]. IEEE Transactions on Industrial Electronics(S0278-0046), 2019, 66(8): 6117-6127.

[9] Yu W, Low Kin Huat, Lü Chen. Cooperative Path Planning for Heterogeneous Unmanned Vehicles in a Search-and-Track Mission Aiming at an Underwater Target[J]. IEEE Transactions on Vehicular Technology (S0018-9545), 2020, 69(6): 6782-6787.

[10] Chnjiang W, Shijie Z, Licai X. Dynamic Path Planning Based on Improved Ant Colony Algorithm in Traffic Congestion[J]. IEEE Access(S2169-3536), 2020, 8: 180773-180783.

[11] Yi Jun, Bai Junren, He Haibo. A Multifactorial Evolutionary Algorithm for Multitasking Under Interval Uncertainties[J]. IEEE Transactions on Evolutionary Computation(S1089-778X), 2020, 24(5): 908-922.

[12] 陈海, 何开锋, 钱炜祺. 多无人机协同覆盖路径规划[J]. 航空学报, 2016, 37(3): 928-935.
Chen Hai, He Kaifeng, Qian Weiqi. Multi-UAV Collaborative Coverage Path Planning [J]. Journal of Aviation, 2016, 37(3): 928-935.

[13] Yao X, Wang X, Zhang L. Model Predictive and Adaptive Neural Sliding Mode Control for 3D Path Following of Autonomous Underwater Vehicle with Input Saturation[J]. Neural Computing and Applications (S0941-0643), 2020, 32(22): 16875-16889.

[14] Delin G, T Lan, Z Xinggan. Joint Optimization of Handover Control and Power Allocation Based on Multi-Agent Deep Reinforcement Learning[J]. IEEE Transactions on Vehicular Technology(S0018-9545), 2020, 69(11): 13124-13138.

[15] Jonggyu J, Hyun Jong Y. Deep Reinforcement Learning-Based Resource Allocation and Power Control in Small Cells with Limited Information Exchange[J]. IEEE Transactions on Vehicular Technology(S0018-9545), 2020, 69(11): 13768-13783.

[16] Fengxiao T, Z Yibo, Kato Neia. Deep Reinforcement Learning for Dynamic Uplink/DowNlink Resource Allocation in High Mobility 5G HetNet [J]. IEEE Journal on Selected Areas in Communications(S0733-8716), 2020, 38(12): 2773-2782.

[17] Y Guan, R Yangang, L Shengbo Eben. Centralized Cooperation for Connected and Automated Vehicles at Intersections by Proximal Policy Optimization [J]. IEEE Transactions on Vehicular Technology(S0018-9545), 2020, 69(11): 12597-12608.

[18] Z Wenhan, L Chunbo, W Jin. Deep-Reinfo Rcement-Learning-Based Offloading Scheduling for Vehicular Edge Computing[J]. IEEE Internet of Things Journal (S2327-4662), 2020, 7(6): 5449-5465.

[19] C Guangda, Y Shunyi, M Jun. Distributed Non-Communicating Multi-Robot Collision Avoidance Via Map-Based Deep Reinforcement Learning[J]. International Journal of Electrical Power & Energy Systems(S0142-0615), 2020, 20(17): 4836.

[20] L Da, Z Zhaosheng, L Peng. Battery Fault Diagnosis for Electric Vehicles Based on Voltage Abnormality by Combining the Long Short-Term Memory Neural Network and the Equivalent Circuit Model[J]. IEEE Transactions on Power Electronics(S0885-8993), 2020, 36(12): 1303-131.

[21] L Tao, H Yongjin, J Ankang. Adversarial Active Learning for Named Entity Recognition in Cybersecurity[J]. CMC-Computers Materials & Continua(S1546-2218), 2020, 66 (1): 407-420.

[22] L Yang Yuan, Do Tien Van, Nguyen Hai T. A Comparison of Forecasting Models for the Resource Usage of MapReduce Applications[J]. Neurocomputing(S0925-2312), 2020, 418: 36-55.

[23] Y Ziming, X Yan. A Multi-Agent Deep Reinforcement Learning Method for Cooperative Load Frequency Control of a Multi-Area Power System[J]. IEEE Transactions on Power Systems(S0885-8950), 2020, 35 (6): 4180-4192.

[24] Fouad Y, Nassim R, Laid D, et al.Trajectory Optimisation for a Quadrotor Helicopter Considering Energy Consumption[C]// 2017 4th International Conference on Control, Decision and Information Technologies. Barcelona, Spain: IEEE, 2017: 5-7.

[25] Goeke D, Schneider M. Routing a Mixed Fleet of Electric and Conventional Vehicles[J]. European Journal of Operational Research(S0377-2217), 2015, 245(1): 81-99.

[26] 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 Changyin, Mu Zhaoxu. Some Key Scientific Problems of Deep Reinforcement Learning for Multi-Agent[J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312.

DOI: 10.16182/j.issn1004731x.joss.21-0112