航空科学技术
Aeronautical Science & Technology
Dec. 25 2023 Vol. 34 No.12 118-124

# 基于深度强化学习的无人机航路规划算法研究

毕文豪，段晓波
西北工业大学，陕西 西安 710072

**摘 要：**航路规划是无人机在复杂战场环境中完成作战任务的关键技术之一。本文提出了一种基于PER-D3QN的无人机航路规划算法，通过网络模型设计、状态空间设计、动作空间设计和收益函数设计实现无人机在战场环境下的航路规划。PER-D3QN算法将目标网络模型、竞争网络模型和优先级经验重现策略进行结合，有效地解决了深度强化学习方法存在的过拟合问题和网络优化不稳定问题。最后，通过仿真试验验证了所提算法相较于Double DQN和DQN算法具有更好的收敛性、稳定性和适用性，相较于A*算法具有较好的实时性，可高效实现无人机在复杂战场环境下的航路规划，有效帮助无人机遂行作战任务。

**关键词：**无人机；航路规划；深度强化学习；战场环境建模；PER-D3QN

**中图分类号：**V249.3 **文献标识码：**A DOI：10.19452/j.issn1007-5453.2023.12.014

近年来，随着信息工程、控制理论、人工智能等技术的不断发展，无人机（UAV）在军事领域的应用越来越广泛。目前，军用无人机可以完成对地侦察、对地打击、无人货运等战术功能。同时，无人僚机、无人机编队协同作战和无人机集群作战等更高层次作战功能也在不断发展中。为了更好地完成上述任务，无人机需要具有根据战场环境和任务需求进行自主航路规划的能力。

航路规划是指在环境约束、飞行性能约束等约束条件下，寻找无人机从起始点到目标点、满足特定任务指标的飞行航路。航路规划可以帮助无人机在面对不同任务功能时，以最小的损失（如时间、受损程度等）完成相应的作战任务，其中主要需要考虑三个方面：（1）面向地形约束，考虑地形变化产生的碰撞威胁；（2）在飞机性能、燃油量和武器性能等约束下规划合理的飞行航路；（3）降低作战环境中无人机遇到不确定风险的可能性，如敌方雷达、防空导弹等战场威胁，提高飞行器安全冗余和作战效能[1]。

目前，航路规划主要包含三种方法：一是基于几何模型搜索的方法，如A*算法[2-3]、Dijkstra算法[4]等；二是基于虚拟势场的方法，如人工势场法[5]等；三是基于优化算法的方法，如蚁群算法[6]、遗传算法[7]、粒子群算法[8]、灰狼算法[9]等。

近年来，人工智能技术的快速发展为无人机航路规划技术的研究提供了新的思路。参考文献[10]提出了一种基于Layered PER-DDQN的无人机航路规划算法，通过战场环境分层将航路规划问题分解为地形规避问题和威胁躲避问题，有效解决了无人机在战场环境中的航路规划问题。参考文献[11]提出了一种基于REL-DDPG的无人机航路规划算法，有效解决了复杂三维地形环境下的航路规划问题。

本文利用不同深度强化学习算法的优势，结合目标网络模型、竞争网络模型和优先级经验重现策略构建PER-D3QN模型，并将其应用于解决战场环境下无人机航路规划问题。搭建了包含战场三维地形和战场威胁的战场环境模型，通过与Double DQN、DQN和A*算法的仿真结果对比，验证本文算法具有较好的收敛性、稳定性、适用性和实时性。

## 1 战场环境建模

战场环境建模是利用数学建模的方法对无人机作战场景进行抽象化表述。对于战场环境下无人机航路规划问题，需要考虑战场三维地形和战场威胁。

### 1.1 战场三维地形建模

本文利用数字高程模型（DEM）描述战场三维地形，该方

收稿日期：2023-06-16；退修日期：2023-10-11；录用日期：2023-11-08
基金项目：航空科学基金（201905053001）；国家自然科学基金（62073267，61903305）
*引用格式：Bi Wenhao，Duan Xiaobo. Research on UAV path planning method based on deep reinforcement learning [J]. Aeronautical Science & Technology，2023，34(12)：118-124. 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.*

法利用有限数量的地形高程数据对地面的地形构造进行数字化模拟，可表示为$\{V_i=(x_i,y_i,z_i)|i=1,2,\cdots,n,(x_i,y_i)\in D\}$，其中，$(x_i,y_i)\in D$表示平面坐标，$z_i$表示$(x_i,y_i)$位置的高程信息。

数字高程模型可以通过简单的数据结构实现地形信息的表示，但由于数据总量和数据密度受数据来源的限制，无法获取地形上每一点的具体数据。本文采用双线性内插法[12]来解决这一问题，即根据待采样点与高程数据库中已有相邻点之间的距离确定相应权值，计算出待采样点的高度值，如图1所示。待采样点$p$的高程信息计算公式为

$$z_p=(l-u)(m-v)z_a+(l-u)vz_b+uvz_c+u(m-v)z_d \quad (1)$$

其中，$z_a,z_b,z_c,z_d$表示相邻点的高程数据。

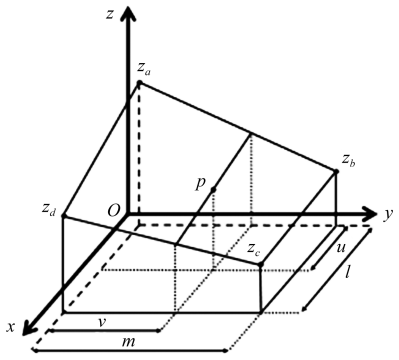


图1 数字高程模型双线性内插法示意图

Fig.1 Bilinear interpolation for DEM

### 1.2 战场威胁建模

在战场环境中，战场威胁是无人机航路规划中的重要约束之一，主要包括雷达探测、防空导弹、防空高炮等。无人机在执行航路规划时，需要对战场威胁区域进行规避，在自身不受损伤的前提下完成指定的任务。本文以防空武器所处位置为圆心，依据防空武器作战半径设定圆形威胁区。

## 2 基于PER-D3QN的无人机航路规划算法

### 2.1 深度强化学习

强化学习是一种基于马尔可夫决策过程框架，使用状态、动作和收益定义智能体与环境之间的交互过程，通过智能体在环境中不断探索、试错的方式进行学习，根据环境对当前状态下所选取动作的反馈收益调整探索策略的机器学习方法。

深度强化学习（DQN）在强化学习的基础上，通过构建深度神经网络（DNN）表征智能体的动作价值函数。动作价值函数表示为$Q(\boldsymbol{s}_t,a_t|\theta)$，其中，$\boldsymbol{s}_t$和$a_t$表示$t$时刻智能体的状态和动作，$\theta$表示网络参数。DQN算法通过最小化损失函数$L(\theta)$的方式更新网络参数$\theta$，使得网络$Q(\boldsymbol{s}_t,a_t|\theta)$的输出不断逼近最优动作价值。损失函数$L(\theta)$的定义为

$$L(\theta)=\frac{1}{2}[y-Q(\boldsymbol{s}_t,a_t|\theta)]^2 \quad (2)$$

$$y=R_{t+1}+\gamma\max_{a_{t+1}\in A}Q(\boldsymbol{s}_{t+1},a_{t+1}|\theta) \quad (3)$$

式中，$y$为网络更新的目标值；$\gamma\in[0,1]$为折扣系数；$R_{t+1}$为收益函数。

### 2.2 PER-D3QN算法

在DQN算法中，智能体的动作价值函数网络在更新过程中存在由于训练样本相关性强导致的过拟合问题和网络更新参数相关联导致的网络优化不稳定问题。针对上述问题，相关研究人员提出了Double DQN、Dueling DQN等方法。

Double DQN[13]在DQN算法的基础上引入目标网络$Q(\boldsymbol{s}_t,a_t|\theta^-)$解决过拟合问题。Double DQN首先利用预测网络$Q(\boldsymbol{s}_t,a_t|\theta)$选取最大价值的动作，然后利用目标函数计算该动作对应的价值，并与收益函数共同构成目标更新值$y'$，其表达式为

$$y'=R_{t+1}+\gamma Q(\boldsymbol{s}_{t+1},\operatorname*{argmax}_{a\in A}Q(\boldsymbol{s}_{t+1},a|\theta)|\theta^-) \quad (4)$$

Dueling DQN[14]将竞争网络模型引入DQN算法中，通过将动作价值函数分解为状态价值函数和动作优势函数，提高对最优策略的学习效率。竞争网络的表达式为

$$Q(\boldsymbol{s}_t,a_t|\theta)=V(\boldsymbol{s}_t|\theta,\alpha)+A(\boldsymbol{s}_t,a_t|\theta,\beta) \quad (5)$$

式中，$V(\boldsymbol{s}_t|\theta,\alpha)$为状态价值函数，$A(\boldsymbol{s}_t,a_t|\theta,\beta)$为动作优势函数，$\alpha$和$\beta$为对应网络的参数。

此外，相关研究人员提出在DQN算法中加入经验重现策略，将智能体与环境交互得到的样本数据存储至经验重现池，当进行网络更新时，从经验重现池随机提取小批量的样本数据进行训练，消除样本之间的关联性，从而提升算法的收敛效果和稳定性。

优先级经验重现（PER）[15]是指在经验重现策略的基础上，根据经验重现池中数据样本的优先级进行采样学习，从而提升网络的训练效率。在网络训练过程中，目标值与预测值之间的误差越大，预测网络对该状态的预测精度越低，则其学习的优先级越高。定义第$i$组数据样本的优先度为

$$p_i=|y'-Q(\boldsymbol{s}_t,a_t|\theta)| \quad (6)$$

根据数据样本的优先度，定义其优先级为

$$P_i=\frac{p_i}{\sum_{j=1}^{N}p_j} \quad (7)$$

PER-D3QN算法是一种结合Double DQN、Dueling DQN和优先级重现策略的深度强化学习算法。该方法联合了上述方法的优势，有效地提高了复杂环境下智能体的学习效率和准确性。

本文利用PER–D3QN算法解决无人机航路规划问题，构建如图2所示的算法模型。本文所提算法旨在让无人机能够根据当前在战场环境中的所处状态$\boldsymbol{s}_t$，利用PER–D3QN算法选择飞行动作$a_t$并获得相应收益$R_{t+1}$，实现无人机在战场环境中的全局航路规划。

### 2.3 航路规划算法实现

#### 2.3.1 状态空间

在战场环境下，无人机的状态空间包含地形环境状态、威胁区域状态和与目标点的相对状态。

地形环境状态用于描述无人机当前时刻与其周围地形环境的相对位置关系，其表达式为

$$\boldsymbol{s}_t^1 = \text{Flatten}\left(\text{Relu}\left(z_t - \boldsymbol{D}_t\right)\right) \tag{8}$$

式中，Flatten (·) 函数用于将矩阵拉伸为一维数组；Relu (·) 表示线性整流函数；$z_t$表示无人机在$t$时刻的飞行高度；$\boldsymbol{D}_t$表示无人机周围区域的数字高程信息矩阵。

威胁区域状态用于表示无人机当前时刻与战场威胁区域的相对位置关系，如图3所示，其表达式为

$$\boldsymbol{s}_t^2 = \left[\frac{d_1}{R} \quad \frac{d_2}{R} \quad \frac{d_3}{R} \quad \frac{d_4}{R} \quad \frac{d_5}{R}\right] \tag{9}$$

式中，$R$表示威胁区感知半径。

无人机与目标点的相对状态用于引导无人机飞向目标位置，其表达式为

$$\boldsymbol{s}_t^3 = [\varphi_t, \gamma_t] \tag{10}$$

式中，$\varphi_t$和$\gamma_t$分别为无人机所处位置与目标点的航向角度差和俯仰角度差，如图4所示。

根据以上三个状态即可得到无人机在$t$时刻的状态矢量，即

$$\boldsymbol{s}_t = [\boldsymbol{s}_t^1, \boldsymbol{s}_t^2, \boldsymbol{s}_t^3] \tag{11}$$

#### 2.3.2 动作空间

动作空间是无人机可采取的行为集合，本文采用的动作空间共包含9个离散动作，分别为向右下方飞行$a_1$、向下方飞行$a_2$、向左下方飞行$a_3$、向右方飞行$a_4$、向前方飞行$a_5$、向左方飞行$a_6$、向右上方飞行$a_7$、向上方飞行$a_8$和向左上方飞行$a_9$，如图5所示。无人机可根据当前状态从上述离散动作中进行选择。动作空间$A$可表示为

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9\} \tag{12}$$

#### 2.3.3 收益函数

针对无人机规划问题的特点，从任务完成、地形避障、威胁区域躲避、飞行高度控制、飞行引导等角度设计深度强化学习的收益函数。

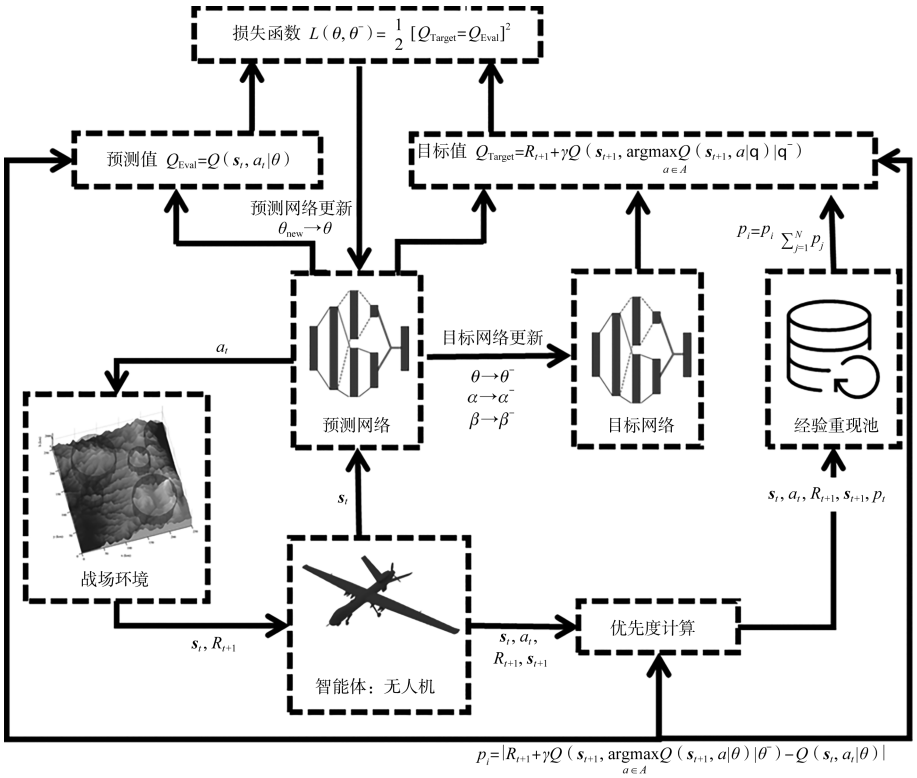


图2 基于PER-D3QN的无人机航路规划算法网络模型

Fig.2 Network model of UAV path planning method based on PER-D3QN

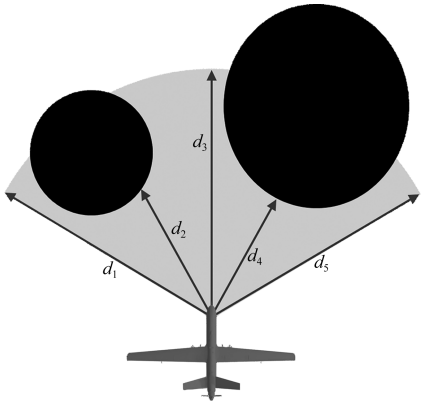


图3 威胁区域状态示意图

Fig.3 Diagram of threat area state

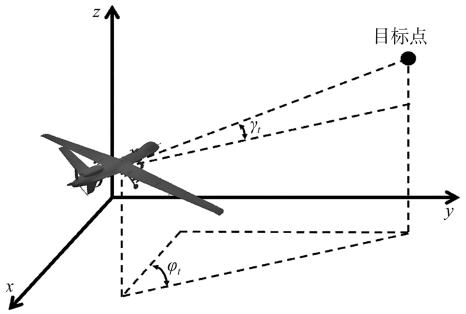


图4 无人机与目标点相对状态示意图

Fig.4 Diagram of the relative state of the UAV and the target point

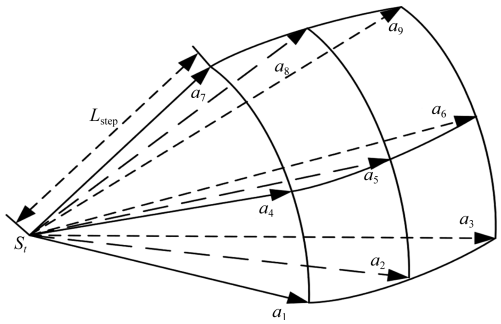


图5 动作空间示意图

Fig.5 Diagram of action space

（1）任务完成收益函数

无人机航路规划的首要目标是规划出从初始点到目标点的飞行航路，当无人机完成航路规划时给予其正向奖励，表达式为

$$R_M=\begin{cases}k_M & \boldsymbol{s}_t=\boldsymbol{s}_{\text{Target}},k_M\gg 0\\ 0 & \boldsymbol{s}_t\neq\boldsymbol{s}_{\text{Target}}\end{cases}\tag{13}$$

式中，$k_M$为无人机抵达目标点给予后的激励值，$\boldsymbol{s}_{\text{Target}}$为无人机抵达目标点时的期望状态。

（2）地形避障收益函数

无人机在飞行过程中需要及时躲避地形障碍，当无人机与地面发生碰撞时给予惩罚反馈，表达式为

$$R_G=\begin{cases}0 & (z_t-h_{\text{safe}})>z_{t,G}\\ k_G & (z_t-h_{\text{safe}})\leqslant z_{t,G},k_G\ll 0\end{cases}\tag{14}$$

式中，$k_G$表示惩罚反馈；$z_t$表示无人机当前时刻飞行高度；$h_{\text{safe}}$表示无人机与地面的安全距离；$z_{t,G}$表示无人机所处位置的地形高度。

（3）威胁区域躲避收益函数

在战场环境下，无人机需要尽量避免进入敌方防空武器的威胁区，根据进入威胁区的距离设计反馈函数，表达式为

$$R_T^i=\begin{cases}0 & D_t^i>r_T^i\\ -\left|\dfrac{r_T^i-D_t^i}{r_T^i-r_T^{i\prime}}\right| & r_T^i\geqslant D_t^i>r_T^{i\prime}\\ k_T & D_t^i\leqslant r_T^{i\prime},k_T\ll -1\end{cases}\tag{15}$$

$$R_T=\sum_{i=1}^{N_T}R_T^i\tag{16}$$

式中，$R_T^i$为第$i$个威胁区的反馈函数；$r_T^i$和$r_T^{i\prime}$分别为敌方威胁作战半径和不可逃逸半径；$D_t^i$为无人机距第$i$个威胁区中心的距离；$k_T$为无人机进入不可逃逸区域后的惩罚反馈。

（4）飞行高度控制收益函数

为避免在作战过程中被敌方雷达过早发现，无人机需要在飞行过程中尽可能保持较低的高度，当飞行高度高于预期高度后给予惩罚反馈，表达式为

$$R_H=\begin{cases}k_H & z_t\geqslant H_{\text{up}},k_H\ll 0\\ 0 & z_t<H_{\text{up}}\end{cases}\tag{17}$$

式中，$k_H$为惩罚反馈；$H_{\text{up}}$为预期高度。

（5）飞行引导收益函数

为帮助无人机快速抵达目标位置，设计飞行引导函数，表达式为

$$R_L=\frac{k_L}{\text{sigmoid}\left(\text{Distance}\left(\boldsymbol{s}_t,\boldsymbol{s}_{\text{Target}}\right)\right)}\tag{18}$$

式中，$k_L>0$为飞行引导奖励系数，$\text{Distance}\left(\boldsymbol{s}_t,\boldsymbol{s}_{\text{Target}}\right)$为无人机当前位置与目标位置的距离，$\text{sigmoid}\left(\cdot\right)$用于调整数值区间，其表达式为$\text{sigmoid}\left(x\right)=\left(1+\text{e}^{-x}\right)^{-1}$。

综合以上收益函数，得到总收益函数，即

$$R=\gamma_M R_M+\gamma_G R_G+\gamma_T R_T+\gamma_H R_H+\gamma_L R_L\tag{19}$$

式中，$\gamma_M$，$\gamma_G$，$\gamma_T$，$\gamma_H$和$\gamma_L$分别为任务完成收益函数、地形避障收益函数、威胁区域躲避收益函数、飞行高度控制收益函数和飞行引导收益函数的权重系数。

## 3 试验与结果分析

本文在Python编程环境中利用PyTorch框架构建PER-D3QN网络，用于算法训练和测试的计算机配置为Intel i7-

12700H CPU，NVIDIA RTX 3060 GPU，32GB运行内存。

为验证本文算法的可行性和有效性，使用Double DQN、DQN和传统A*算法进行对比试验。试验在随机生成的100幅战场环境地图中进行，其中，训练过程使用80幅战场环境地图，测试过程使用20幅战场环境地图。A*算法只在测试过程的战场环境地图中进行仿真。PER-D3QN、Double DQN和DQN在训练过程中的学习曲线如图6所示。

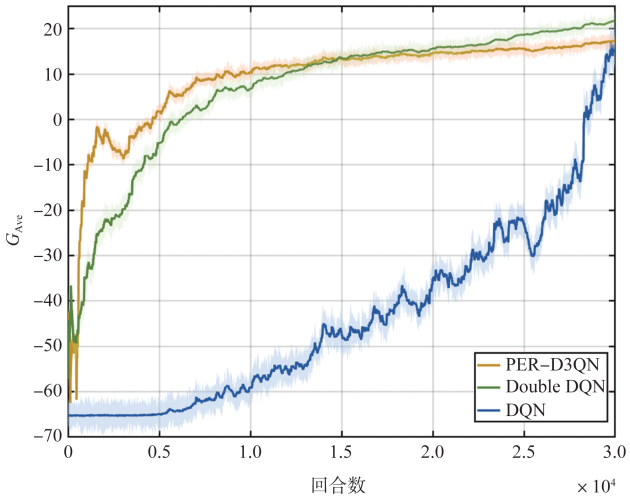


图6 PER-D3QN，Double DQN和DQN的学习曲线

Fig.6 Reward curve for PER-D3QN, Double DQN and DQN

在训练过程中，智能体采用PER-D3QN、Double DQN和DQN算法各进行了30000回合的探索与学习。图6展示了智能体在训练过程中获得的平均收益，其中，PER-D3QN在训练至7500回合后达到收敛状态，收敛效果优于Double DQN和DQN。

为验证智能体训练过程的有效性，对训练好的算法模型进行测试。在测试过程中，对20幅战场环境地图各随机生成50组起始点和目标点，共构建1000个仿真场景。将PER-D3QN、Double DQN、DQN和A*算法在上述仿真场景中进行测试，对航路规划成功率、平均航程和算法平均计算时间进行计算和统计。

航路规划成功率是指测试过程中成功完成航路规划的次数与总测试次数的比值，其计算公式为

$$R_s = \frac{N_s}{N_a} \tag{20}$$

式中，$N_s$表示成功完成航路规划的次数；$N_a$表示总测试次数。

平均航程和平均计算时间在4种算法均成功完成航路规划的仿真场景中计算并统计。平均航程是指各算法规划得到的飞行航路航程的平均值，其计算公式为

$$L_{ave} = \frac{1}{N_v}\sum_{i=1}^{N_v}\sum_{j=1}^{M_i-1}\text{Distance}\left(s_{i,j}, s_{i,j+1}\right) \tag{21}$$

式中，$L_{ave}$为平均航程；$N_v$指飞行航路的个数；$M_i$为第$i$条飞行航路的节点数量；$s_{i,j}$为第$i$条飞行航路的第$j$个航路节点。

平均计算时间是指算法运行所消耗时间的平均值，其计算公式为

$$T_{ave} = \frac{1}{N_v}\sum_{i=1}^{N_v} t_{c,i} \tag{22}$$

式中，$T_{ave}$为平均计算时间；$t_{c,i}$为规划第$i$条飞行航路的计算耗时。

测试过程的统计结果见表1，得到的部分航路规划结果如图7所示。

由表1可以看出：在成功率和平均航程上，传统A*算法的成功率为100%且平均航程最短，可在任一仿真场景中实现航路规划；PER-D3QN算法的成功率和平均航程稍逊于A*算法，但优于Double DQN和DQN算法，具有较好的稳定性，可胜任绝大部分仿真场景。在平均计算时间上，PER-D3QN算法相较其他三种算法具有较大优势，可在较短时间内完成航路规划任务，更符合无人机作战的实时性需求，具有更好的适用性。

表1 算法测试结果统计

Table 1 Method test result statistics

| 算法 | 航路规划成功率/% | 平均航程/km | 平均计算时间/s |
|---|---|---|---|
| PER-D3QN | 93.6 | 377.27 | 1.32 |
| Double DQN | 87.5 | 399.52 | 2.11 |
| DQN | 34.6 | 420.56 | 3.67 |
| A* | 100 | 361.81 | 7.86 |

仿真结果表明，本文所提出的基于PER-D3QN的无人机航路规划算法，相较于Double DQN和DQN，在训练过程和测试过程中均具有较好的表现；相较于传统A*算法，虽在航路规划成功率和平均航程上稍有劣势，但在平均计算时间上具有较大优势。

## 4 结束语

为解决无人机在战场环境下的航路规划问题，本文提出了基于PER-D3QN的无人机航路规划算法。构建了包含战场三维地形和战场威胁的战场环境模型。通过设计网络模型、状态空间、动作空间和收益函数实现无人机航路规划算法。在仿真验证中，验证了所提算法相较于Double DQN和DQN具有更好的收敛性、稳定性和适用性，相较于传统A*算法具有较好的实时性。在下一步研究中，应考虑多无人机协同作战的任务场景，基于多智能体深度强化学

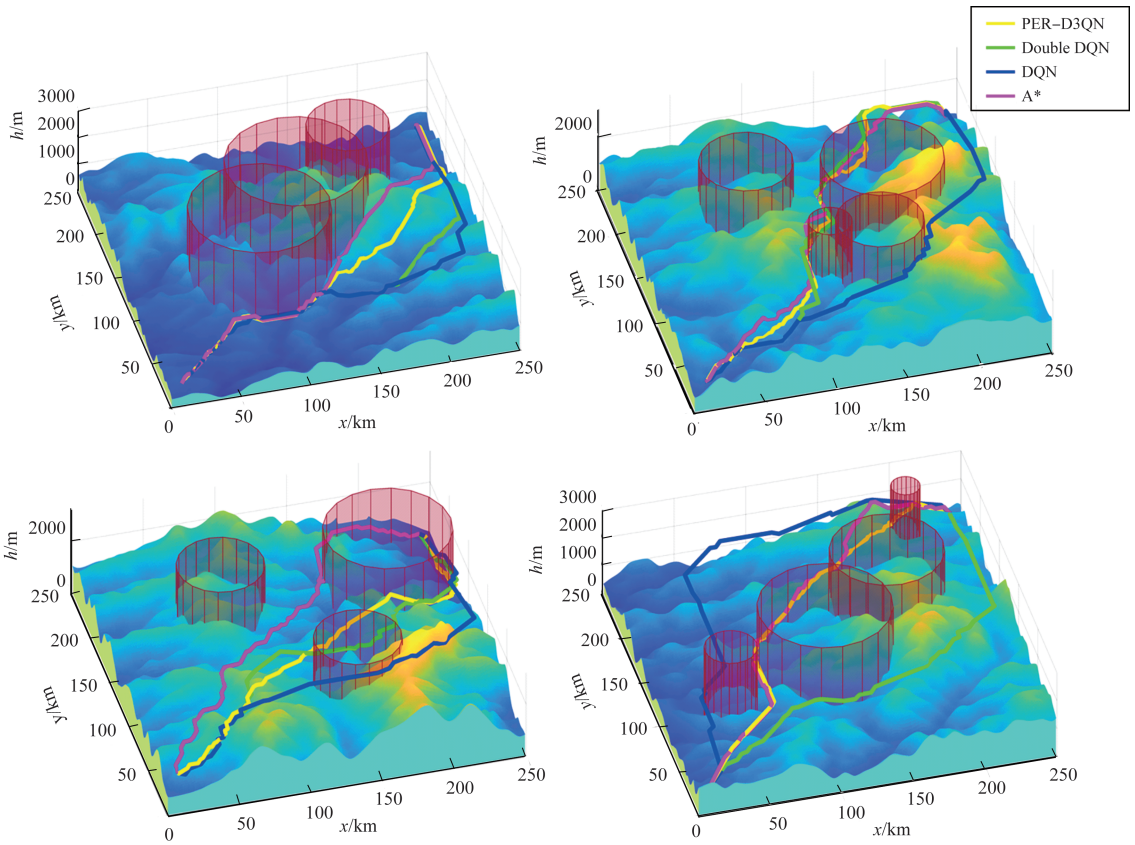


图7 战场环境下PER-D3QN，Double DQN，DQN和A*算法航路规划结果

Fig.7 Results of path planning based on PER-D3QN, Double DQN, DQN and A* in battlefield environment

习方法研究复杂战场环境下多无人机协同航路规划问题，以进一步提升无人机作战效能。

## 参考文献


[1] 李冲. 多UCAV协同作战航路规划研究[D]. 西安：西北工业大学，2018.
Li Chong. Research on the route planning of multiple UCAV cooperation[D]. Xi'an: Northwestern Polytechnical University, 2018. (in Chinese)

[2] Zhang An, Li Chong, Bi Wenhao. Rectangle expansion A* pathfinding for grid maps[J]. Chinese Journal of Aeronautics, 2016, 29(5): 1385-1396.

[3] 李海，郭水林，周晔. 融合动态风险图和改进A*算法的动态改航规划[J]. 航空科学技术，2021，32(5): 61-71.
Li Hai, Guo Shuilin, Zhou Ye. Dynamic diversion combining dynamic risk map and improved A* algorithm[J]. Aeronautical Science & Technology, 2021, 32(5): 61-71. (in Chinese)

[4] 巩慧，倪翠，王朋，等. 基于Dijkstra算法的平滑路径规划方法[J/OL]. (2023-06-15). 北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2022.0377.
Gong Hui, Ni Cui, Wang Peng, et al. A smooth path planning method based on Dijkstra algorithm[J/OL]. (2023-06-15). Journal of Beijing University of Aeronautics and Astronautics. https://doi.org/10.13700/j.bh.1001-5965.2022.0377.(in Chinese)

[5] 王庆禄，吴冯国，郑成辰，等. 基于优化人工势场法的无人机航迹规划[J]. 系统工程与电子技术，2023，45(5): 1461-1468.
Wang Qinglu, Wu Fengguo, Zheng Chengchen, et al. UAV path planning based on optimized artificial potential field method[J]. Systems Engineering and Electronics, 2023, 45(5): 1461-1468. (in Chinese)

[6] 梁宵，王宏伦，孟光磊，等. 三维真实地形环境下无人机救援航路规划方法[J]. 北京航空航天大学学报，2015，41(7): 1183-1187.
Liang Xiao, Wang Honglun, Meng Guanglei, et a1. Path planning for UAV under three-dimensional real terrain in rescue mission[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(7): 1183-1187. (in Chinese)

[7] 杨咪,张安,毕文豪,等. 基于改进遗传算法的多弹协同攻击航路规划[J]. 兵工自动化, 2020,39(2): 28-32+40.
Yang Mi, Zhang An, Bi Wenhao, et al. Multi-missile cooperative attacking routing planning based on improved genetic algorithm[J]. Ordnance Industry Automation, 2020, 39(2): 28-32+40. (in Chinese)

[8] 胡美富,宁芊,陈炳才,等. RWPSO与马尔科夫链的无人机航路规划[J]. 哈尔滨工业大学学报,2019,51(11): 75-81.
Hu Meifu, Ning Qian, Chen Bingcai, et al. UAV route planning based on RWPSO and Markov chain[J]. Journal of Harbin Institute of Technology, 2019, 51(11): 75-81. (in Chinese)

[9] 吴坤,池沛,王英勋,等. 基于混沌灰狼优化的多无人机协同航路规划[J]. 航空科学技术, 2022,33(10): 82-95.
Wu Kun, Chi Pei, Wang Yingxun, et al. Cooperative path planning for multi-UAVs based on chaos gray wolf optimization[J]. Aeronautical Science & Technology, 2022, 33(10): 82-95. (in Chinese)

[10] Wang Weixiang, Zhang An, Bi Wenhao, et al. A novel UAV path planning method based on layered PER-DDQN[C]// Proceedings of the 2021 Asia-Pacific International Symposium on Aerospace Technology (APISAT 2021), 2023: 693-702.

[11] Hu Zijian, Gao Xiaoguang, Wan Kaifang, et al. Relevant experience learning: A deep reinforcement learning method for UAV autonomous motion planning in complex unknown environments[J]. Chinese Journal of Aeronautics, 2021, 34 (12): 187-204.

[12] 张腾腾. 1∶10000 DEM精细化处理方法及其应用效果分析[J]. 北京测绘,2021,35(2): 212-216.
Zhang Tengteng. 1∶10000 DEM fine processing method and its application effect analysis[J]. Beijing Surveying and Mapping, 2021, 35(2): 212-216. (in Chinese)

[13] Van Hasselt H, Hado Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]. Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[14] Wang Ziyu, Tom S, Matteo H, et al. Dueling network architectures for deep reinforcement learning[C]. The 33rd International Conference on Machine Learning, 2016.

[15] Tom S, John Q, Ioannis A, et al. Prioritized experience replay [C]. The 4th International Conference on Learning Representations,2016.

# Research on UAV Path Planning Method Based on Deep Reinforcement Learning

Bi Wenhao, Duan Xiaobo

*Northwestern Polytechnical University, Xi'an 710072, China*

**Abstract:** Path planning is one of the key technologies for UAVs to accomplish operation missions in complex battlefield environments. In this paper, we propose a UAV path planning method based on PER-D3QN, which realizes the path planning for UAVs in the battlefield environment through network model design, state space design, action space design and reward function design. The PER-D3QN algorithm combines the target network, dueling network and prioritized experience replay, which effectively solves the overfitting problem and unstable problem in deep reinforcement learning. In the end, it is verified through simulation experiments that the proposed method achieved better convergence, stability and applicability compared with double DQN and DQN algorithms, and better real-time performance compared with A* algorithm, which can efficiently realize the path planning of UAVs in the complex battlefield environment, and effectively help the UAVs to attempt the operational mission.

**Key Words:** UAV; path planning; deep reinforcement learning; battlefield environment modeling; PER-D3QN

**Received:** 2023-06-16; **Revised:** 2023-10-11; **Accepted:** 2023-11-08

**Foundation item:** Aeronautical Science Foundation of China(201905053001); National Natural Science Foundation of China (62073267, 61903305)