## 无人自主系统及智能决策专栏

doi: 10.11809/bqzbgcxb2024.07.001

专栏主编: 李波 博士( 西北工业大学 教授、博导)

导语: 伴随着人工智能技术加速落地应用,无人自主系统成为现代先进科技发展的重要组成部分,因其自主性、智能性与无人化的特点,无人自主系统深刻影响着军事、经济、社会和人类生活方式等,颠覆传统规则。作为多学科领域的集大成者,无人自主系统涉及智能感知与理解、多任务集群协同、人机交互与人机融合协同控制等诸多关键科学问题需要深入研究,已成为目前学术界和工业界重要的研究和应用领域之一,具有重要的研究意义。

为了探讨和交流无人自主系统及智能决策的新理论、新方法和新应用,本专栏收录了领域内相关研究机构的 7 篇论文,内容涵盖智能感知与定位、自主轨迹跟踪控制、无人机集群自主协同对抗、目标威胁智能评估等方面,希望能够促进业内相关科研人员的深入合作与交流,推动智能无人系统的前沿技术创新与高质量发展。

# 基于强化学习的多智能体协同电子对抗方法

杨　洋[1,2],王　烨[1,2],康大勇[3],陈嘉玉[1],李　姜[1,2],赵华栋[1,2]

( 1. 中国科学院长春光学精密机械与物理研究所, 长春　130033; 2. 中国科学院大学, 北京　100049;
3. 光电对抗测试评估技术重点实验室, 河南 洛阳　471000)

摘要: 传统电子战正逐步向融合人工智能技术的智能电子战演变,基于强化学习的多无人机电子协同对抗为主要场景,针对复杂高维的状态动作空间下多智能体强化学习算法不容易收敛问题,提出了一种基于优先经验回放的多智能体双对抗策略梯度算法。该算法通过引入优先经验回放机制,并提出对抗 Critic 网络和双 Critic 网络来平衡动作及价值间的关系和减小单一 Critic 网络估计不确定性的问题。仿真实验结果表明: 在同一仿真场景下相较于其他强化学习算法,PerMaD4 算法具有更好的收敛效果且任务完成度提高了 8.9%。

关键词: 协同决策; 强化学习; 策略梯度; 电子对抗仿真

本文引用格式: 杨洋,王烨,康大勇,等. 基于强化学习的多智能体协同电子对抗方法[J]. 兵器装备工程学报,2024,45(7):1-10.

Citation format: YANG Yang, WANG Ye, KANG Dayong, et al. Multi-agent cooperative electronic countermeasure method based on reinforcement learning[J]. Journal of Ordnance Equipment Engineering,2024,45(7):1-10.

中图分类号: TP18　　文献标识码: A　　文章编号: 2096-2304(2024)07-0001-10

收稿日期: 2023-10-18; 修回日期: 2023-11-09; 录用日期: 2023-12-12

基金项目: 国家自然科学基金项目(61977059)

作者简介: 杨洋(1998—),男,硕士研究生,E-mail: student_yangyang@163.com。

通信作者: 李姜(1982—),男,博士,研究员,博士生导师,E-mail: cclijiang@163.com。

# Multi-agent cooperative electronic countermeasure method based on reinforcement learning

YANG Yang[1,2], WANG Ye[1,2], KANG Dayong[3], CHEN Jiayu[1], LI Jiang[1,2], ZHAO Huadong[1,2]

(1. Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Key Laboratory of Electro-Optical Countermeasures Test & Evaluation Technology, Luoyang 471000, China)

**Abstract:** Traditional electronic warfare is gradually evolving into intelligent electronic warfare that integrates artificial intelligence technology. In view of the problem that multi-agent reinforcement learning algorithm is not easy to converge in complex and high-dimensional state action space, a multi-agent dual adversarial strategy gradient algorithm based on preferential experience playback is proposed. The algorithm introduces a preferential experience playback mechanism, and presents a counter Critic network and a dual Critic network to balance the relationship between action and value and to reduce the uncertainty of a single Critic network. The simulation results show that compared with other reinforcement learning algorithms, the PerMaD4 algorithm has better convergence effect and the task completion degree is increased by 8.9% in the same simulation scene.

**Key words:** collaborative decision-making; reinforcement learning; policy gradient; electronic countermeasure simulation

## 0 引言

将人工智能与传统电子战相结合，可以有效提升在复杂的网络电子对抗环境中的适应能力。智能化已经成为未来电子战系统发展的趋势和重点，而无人化协同作战则开创了新的电子战作战模式[1]。通过将人工智能技术应用于电子战，可以使系统具备更高的自主性和智能化程度。电子对抗技术利用技术手段来削弱、破坏敌方的电子信息设备、系统、网络以及相关武器系统或人员的作战效能，并同时保护己方的电子信息设备、系统、网络以及相关武器系统或人员的作战效能[2]。美国的“第三次抵消战略”中提出了智能武器、自动化无人武器系统等新概念武器，旨在通过发展颠覆性技术来改变未来战局[3]。这些武器的发展与无人作战平台、电子战、辅助决策[4]等技术领域的应用密切相关。

目前为止主流的辅助决策方法有3类：第1类是基于专家系统的决策方法，专家系统模拟人类专家在特定领域中决策的思考过程，并从中学习经验和规则，从而为指挥官提供更有针对性的建议[5]，该系统有助于优化作战计划、提升对局意识、提高战场有效性。孙乐等[6]介绍了一种基于专家系统的联合作战决策支持系统，该系统包含了多个领域的知识和规则，并利用专家系统的知识表示和推理能力为决策提供支持。专家系统具有不需要人工过多干预的自动化特点[7]，但由于其几乎没有自主学习特性，使得专家系统在运行时依然受人类已知规则的束缚，其实质是利用计算机替代人工实现重复性，基础性工作。第2类是基于模型的群体智能决策算法，其中不乏具有代表性的研究，马也等[8]将改进的蜂群算法用于无人机群协同防御作战场景可有效的提升防御的成功率；胡振震等[9]基于改进的粒子群算法能非常有效获得对手的针对性策略以实现其最大化利用；李杰等[10]基于改进的蚁群算法实现了高效高质的作战环推荐优化求解。随着科技的发展，未来的战争形势必定复杂多变，作战行动中需要考虑的因素呈指数增长，需要对海量数据进行处理和分析[11]。因此，近些年来研究者们一直努力探讨第3种途径，即基于“人工智能”的决策技术[12]。

强化学习技术具有自动化程度高，很好适应动态变化的环境，能应用于各种复杂领域的特点，是基于人工智能的决策任务的技术首选。目前有一些主流的强化学习算法，Mnih等[13]提出的DQN算法，使用深度学习技术通过多方面的改进提高深度强化学习的效果，并解决了一些传统强化学习方法存在的问题。Lillicrap等[14]基于深度学习方法和策略梯度算法提出DDPG，通过采用经验回放和目标网络技巧解决连续动作空间中的强化学习问题。Lowe等[15]提出的MADDPG，基于深度学习多智能体强化学习算法，通过分布式经验回放，中心化训练、分布式执行方式解决多智能体的协同决策问题。在一些工程问题中，唐峯竹等[16]根据强化学习提出一种多无人机动态任务分配算法，对任务执行的优先级顺序和执行时间加以约束，提高了有限时间内总体的任务完成度。施伟等[17]基于强化学习提出一种多无人机协同空战方法，该方法提高了在协同对抗场景下多无人机、多智能体间的协同程度。薛喜地[18]提出一种基于强化学习的无人机室内避开障碍物算法，该方法提高了训练速度以及导航过程中对环境的适应能力。

通过针对目前成熟算法研究和工程实践遇到的问题，发

现现有研究存在以下不足:

1) 由于仿真环境复杂,智能体的观测空间和动作空间纬度增加,容易出现经验回放池收集的经验质量低,采样效率不高,算法难以收敛等问题。

2) 由于智能体个数和环境复杂度增加,经典算法评估模块需消耗大量计算资源,容易出现评估不准确,经典算法适应能力下降,任务完成度不高,难以收敛等问题。

3) 在算法结果可视化方面,现有研究目前很少有操作便捷,建模简单的三维仿真场景可视化平台。

针对经典算法的上述不足,本文中基于多智能体深度策略梯度进行改进,提出对抗判别网络对智能体输入状态和动作的相对贡献进行准确有效的分析; 提出了双 Critic 网络共同估计动作值函数,优化值函数估计的准确性,并引入优先经验回放机制使智能体充分地利用经验信息,从而提高算法的学习效率和稳定性。试验结果表明,改进的算法相较于其他强化学习算法具有更好的收敛效果和任务完成度。

## 1　任务描述与模型建立

### 1.1　任务描述

开展红蓝双方电子战仿真,红方单位装备为无人机集群和轰炸机,蓝方单位装备为地面雷达和导弹系统。无人机集群主要任务是利用搭载的光电探测设备和雷达干扰装置对蓝方雷达进行协同侦察,干扰并找到目标单位,为后面轰炸机开辟出一条合理的投弹路线。蓝军作战单位主要是雷达以及导弹系统,蓝军防空雷达主要用于远距离探测空中目标,为导弹系统提供远距离的目标跟踪数据,另外该雷达单位附近存在火炮系统,当目标距离近时,火炮系统被动侦察并打击空中飞行目标。电子对抗场景如图 1 所示。


图 1　电子对抗场景示意图

Fig. 1 Schematic diagram of electronic countermeasures scenarios

### 1.2　模型建立

#### 1.2.1　蓝方模型的建立

目标进入雷达的探测范围内,雷达基于一定概率发现目标,该概率取决于雷达和目标之间的能量接触。为便于讨论,仅分析雷达的慢速扫描与快速扫描。

1) 雷达慢速扫描。雷达实施慢速扫描时,可将其针对目标的侦察行为视作分散性观测,雷达此刻的探测概率记作 $P_D$:

$$P_D = 1 - \prod_{i=1}^{m}(1 - P_{dt}) \tag{1}$$

2) 雷达快速扫描。雷达快速扫描时,可视为连续观察,在无噪音干扰的情况下,雷达对点目标的发现概率为

$$P = 1 - \exp\{-C\int_0^t R(t)^{-4}\mathrm{d}t\} \tag{2}$$

记 $U = C\int_0^t R(t)^{-4}\mathrm{d}t$ 为发现势,则到 $t$ 时刻,雷达到目标探测区的发现势为

$$U = C\int_0^t R(t)^{-4}\mathrm{d}t = C\int_0^t [X_0^2 + (y_0 + V_\delta t)^2]^{-2}\mathrm{d}t \tag{3}$$

令,$y_0 + V_\delta t = X_0\tan\varphi$,则:

$$U(X_0) = \frac{C}{V_\delta X_0^3}\int_{\phi_0}^{\phi}\cos^2\phi\mathrm{d}\phi =$$

$$\frac{C}{2V_\delta X_0^3}[(\phi - \phi_0) + 0.5(\sin 2\phi - \sin 2\phi_0)] \tag{4}$$

在该段上发现的目标概率是:

$$P_{X_0} = 1 - \mathrm{e}^{-U(x_0)} \tag{5}$$

单发防空导弹打击的概率为

$$P = 1 - \exp\left\{\frac{\alpha W_d^\beta}{\sigma_d^\beta}\right\} \tag{6}$$

式(6)中: $W_d$ 为导弹的战斗部质量; $\sigma_d$ 为没有干扰情况下的到单精度误差的均方差; $\alpha$、$\beta$、$\gamma$ 为比例系数。

#### 1.2.2　红方模型的建立

1) 检测雷达信号概率模型。为模拟红方作战单位,需要综合考虑各种因素,不只涵盖了气候状况、地势地貌等自然要素,同时涵盖蓝方战斗部队中的防空雷达、导弹系统所带来的危险。红方无人机侦察到蓝方雷达信号的概率为

$$P = 1 - \exp\left\{-C\int_{t_0}^{t_1}[(x_t - \tau)^2 + (y_t - \xi)^2]^{-2}\mathrm{d}t\right\} \tag{7}$$

式(7)中: $\tau$ 为目标横坐标,$\xi$ 为目标纵坐标。

2) 探测设备定位模型。图 2 所示对地定位数学模型中,地球模型采用 WGS-84 旋转椭球体模型。$O'_g$ 表示地理坐标系 $O_gX_gY_gZ_g$ 的原点 $O_g$ 和地球坐标系 $O_eX_eY_eZ_e$ 的原点 $O_e$ 之间连线在地球表面的交点,$\lambda$、$L$ 分别表示地面目标 $T$ 的经度和纬度。

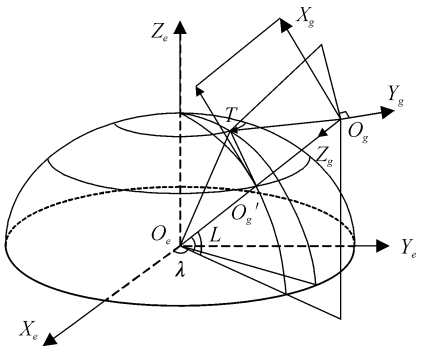



图 2　对地定位数学模型

Fig. 2 Mathematical model of ground to ground positioning

使光电探测设备的视轴方向以及吊舱视频中心的瞄准线方向均保持为 $X_a$ 轴正向，假定载机与地面目标相距 $l_{\mathrm{der}}$，通过激光测距机可以测出该距离。设光电平台坐标系原点 $O_a$ 到地面目标之间的矢量为 $\boldsymbol{V}_a$。在目标恰好位于瞄准线十字的中心点时，矢量 $\boldsymbol{V}_a$ 表示目标在光电平台坐标系 $O_aX_aY_aZ_a$ 中的空间位置向量，由此得到：

$$\boldsymbol{V}_a = [l_{\mathrm{der}},0,0]^{\mathrm{T}} \tag{8}$$

基于光电平台的构造特性，能够通过 2 次角度旋转将 $\boldsymbol{V}_a$ 映射到机体坐标系 $O_bX_bY_bZ_b$。由此得到：

$$\boldsymbol{V}_b = \boldsymbol{C}_a^b \cdot \boldsymbol{V}_a = \begin{bmatrix} \cos\theta_a & -\sin\theta_a & 0 \\ \sin\theta_a & \cos\theta_a & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos\theta_p & 0 & \sin\theta_p \\ 0 & 1 & 0 \\ -\sin\theta_p & 0 & \cos\theta_p \end{bmatrix} \cdot \boldsymbol{V}_a \tag{9}$$

式(9)中：$\theta_p$、$\theta_a$ 分别代表从机体坐标系 $O_bX_bY_bZ_b$ 到光电平台坐标系 $O_aX_aY_aZ_a$ 的仰角和方位角，这些角度值可以通过吊舱内的测量设备直接测量得出。随后将矢量 $\boldsymbol{V}_b$ 通过 3 次角度旋转映射到地理坐标系 $O_gX_gY_gZ_g$ 中，此过程中，$\alpha$、$\beta$、$\gamma$ 分别表示地理坐标系 $O_gX_gY_gZ_g$ 到机体坐标系 $O_bX_bY_bZ_b$ 的飞行器航向角、仰角和横滚角，这些角度值由飞行器导航组件提供。

$$\boldsymbol{V}_g = \boldsymbol{C}_b^g \cdot \boldsymbol{V}_b = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -sin\beta & 0 & \cos\beta \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{bmatrix} \cdot \boldsymbol{V}_b \tag{10}$$

将矢量 $\boldsymbol{V}_g$ 经过 3 次角度旋转变换到地球坐标系 $O_eX_eY_eZ_e$ 中，其中 $\lambda_p$、$L_p$ 分别为飞机的经度和纬度；$\boldsymbol{C}_e^b$ 为从地理坐标系 $O_gX_gY_gZ_g$ 到地球坐标系 $O_eX_eY_eZ_e$ 的坐标转换矩阵。

$$\boldsymbol{V}_e = \boldsymbol{C}_g^e \cdot \boldsymbol{V}_g = \begin{bmatrix} \cos\lambda_p & -\sin\lambda_p & 0 \\ \sin\lambda_p & \cos\lambda_p & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos L_p & 0 & -\sin L_p \\ 0 & 1 & 0 \\ \sin L_p & 0 & \cos L_p \end{bmatrix} \cdot \begin{bmatrix} \cos\frac{\pi}{2} & 0 & -\sin\frac{\pi}{2} \\ 0 & 1 & 0 \\ \sin\frac{\pi}{2} & 0 & \cos\frac{\pi}{2} \end{bmatrix} \boldsymbol{V}_g \tag{11}$$

联立式(9)—式(11)，可得矢量 $\boldsymbol{V}_a$ 转换到地球坐标系 $O_eX_eY_eZ_e$ 的表达式为

$$\boldsymbol{V}_e = \boldsymbol{C}_g^e \boldsymbol{C}_b^g \boldsymbol{C}_a^b \cdot \boldsymbol{V}_a = \boldsymbol{C}_a^e \cdot \boldsymbol{V}_a \tag{12}$$

由机载 INS 或 GPS 设备计算得出当前飞机在地球坐标系 $O_eX_eY_eZ_e$ 中的坐标矢量 $\boldsymbol{V}_{ep}$ 后可计算目标在地球坐标系中的坐标矢量为

$$\boldsymbol{V}_{et} = \boldsymbol{V}_{ep} + \boldsymbol{V}_e \tag{13}$$

将式(13)重写为

$$\boldsymbol{V}_{et} = [X_{et}Y_{et}Z_{et}]^{\mathrm{T}} \tag{14}$$

根据式(14)，可求得目标的经度 $\lambda_t$ 为

$$\lambda_l = \tan^{-1}\left(\frac{Y_{et}}{X_{et}}\right) \tag{15}$$

纬度 $L_t$ 使用迭代收敛的方法求得：

$$(R_N + H_i)_{i+1} = \frac{X_{el}}{\cos L_{ii}\cos\lambda_i} \tag{16}$$

$$R_{N(i+1)} = \frac{R_e}{\sqrt{\cos^2 L_{ii} + (1-e^2)\sin^2 L_{ii}}} \tag{17}$$

$$L_{i(i+1)} = \mathrm{tg}^{-1}\left(\frac{(R_N + H)_{i+1}}{(R_N + H)_{i+1} - R_{N(i+1)} \cdot e^2} \cdot \frac{Z_t^e}{\sqrt{(X_{el})^2 + (Y_{el})^2}}\right) \tag{18}$$

式(17)中：$R^e$ 为地球椭球长半轴；$e$ 为地球椭球的第 1 偏心率；$R^N$ 为 $O_g$ 点处的卯酉圈曲率半径；$H$ 为 $O_g$ 点的飞行高度；式(18)中的下标 $i$ 表示第 $i$ 次迭代，迭代开始时的参数 $L_0$ 从下式求得：

$$L_{t0} = \mathrm{tg}^{-1}\left(\frac{1}{1-e^2} \cdot \frac{Z_t^e}{\sqrt{(X_t^e)^2 + (Y_t^e)^2}}\right) \tag{19}$$

迭代 $k$ 次基本达到稳定：

$$H_t = (R_N + H_t)_k - R_N k \tag{20}$$

## 2 基于优先经验回放的多智能体双对抗策略梯度

### 2.1 MADDPG 算法

多智能体深度确定性策略梯度中使用了 2 个主要的网络：动作网络和判别网络。MADDPG 算法的结构如图 3 所示，每个智能体都拥有自己的 Actor 网络和 Critic 网络。

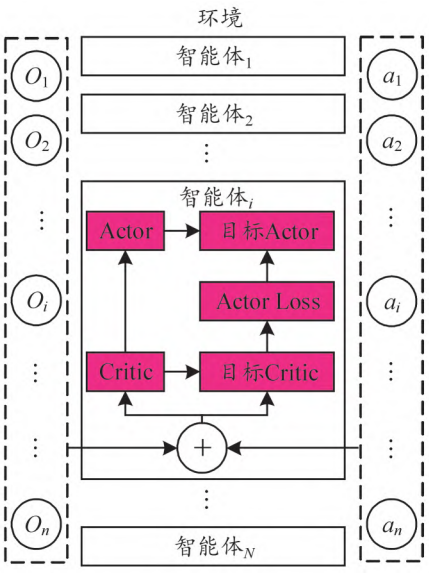



图 3 MADDPG 算法的结构

Fig. 3 Structure of MADDPG algorithm

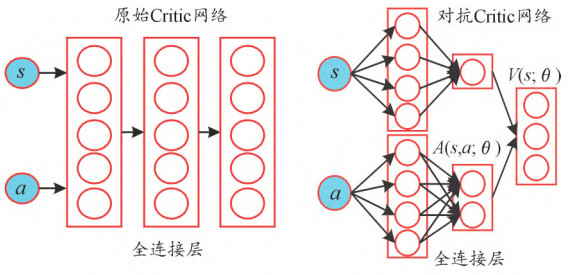
以第 $i$ 个智能体为例,该智能体的输入除了自身状态动作对信息外还有其他智能体的动作$(a_1,a_2,a_i,\cdots,a_N)$与状态$(o_1,o_2,o_i,\cdots,o_N)$。此外,MADDPG 算法使用了经验池机制用于存储每个智能体与环境交互生成的数据$(s,a,r,s')$,每当新的数据被生成,这些数据会存储在经验池内,通常用 $D$ 表示。训练过程中,实行集中式训练与分布式执行的策略,即依据各自的策略每一个智能体都会根据当前状态执行相应的动作,并在与环境互动过程中获取经验,之后存储至各自的经验缓冲区 $D$。所有智能体与环境完成互动后,它们将从经验池中随机选取经验来分别训练各自的神经网络。

### 2.2　PerMaD4 算法

由于多智能体强化学习环境存在非稳定性、高维度和连续性空间的特点。多智能体强化学习算法需要使用大量的计算资源,同时需要更新大量的参数,使得网络估值累计的偏差增大; 而且合适的采样点往往很难确定,如果采样点不够多或不够均匀,就会出现样本不充分的问题,从而影响算法的效果; 也很有可能出现状态和动作关联性低、稀疏等问题,使得算法难以找到合适的解。相较于 MADDPG,PerMaD4 算法有以下 3 处改进。

#### 2.2.1　对抗判别网络

如图 4(a)所示,MADDPG 算法的 Critic 网络的输入被压缩到几层并行的全联接网络中,这种结构无法对各个状态和动作的相对贡献进行准确有效的分析; 同时由于状态值和动作值被混合在一起,当算法处理高维状态空间时,Critic 网络需要学习大量的参数来评价动作网络的策略,必然会导致计算成本过高,从而会出现过度估计情况。因此,为了 Critic 网络可以对各个动作的相对贡献进行准确有效的分析,引入优势函数:




(a) 原始判别网络结构　(b) 对抗判别网络结构

图 4　改进前后的 MADDPG 判别网络结构

Fig. 4 The Critic network structure in MADDDPG has been improved

1) 定义最优优势函数。设 $Q^*(s,a)$ 为最优动作价值函数,$V^*(s)$ 为最优状态价值函数,它们的计算公式如下:

$$\begin{cases} Q^*(s,a) = \max\limits_{\pi} Q_\pi(s,a) \\ V^*(s) = \max\limits_{\pi} V_\pi(s) \end{cases} \tag{21}$$

最优优势函数 $A^*(s,a)$ 计算公式为

$$A^*(s,a) = Q^*(s,a) - V^*(s) \tag{22}$$

式(22)中: $V^*(s)$ 评价状态 $s$ 的好坏程度,$Q^*(s,a)$ 评价在状态 $s$ 下智能体执行动作 $a$ 的好坏程度,这里相当于将 $V^*(s)$ 做为基线,$A^*(s,a)$ 为动作 $a$ 相对于基线 $V^*(s)$ 的优势程度。

2) 优势函数的性质。在强化学习理论中,$V^*(s)$ 函数是 $Q^*(s,a)$ 函数关于 $a$ 的最大化,即公式(23):

$$V^*(s) = \max_a Q^*(s,a) \tag{23}$$

对式(20)两边同时最大化动作 $a$,得到式(24):

$$\max_a A^*(s,a) = \max_a Q^*(s,a) - V^*(s) \tag{24}$$

再将式(23)代入式(24)可得式(25):

$$\max_a A^*(s,a) = 0 \tag{25}$$

3) 对抗判别函数。根据式(22),还可以将其做变换得到如下等式:

$$Q^*(s,a) = V^*(s) + A^*(s,a) \tag{26}$$

将式(25)代入式(26)可得到最优价值函数的计算式为

$$Q^*(s,a) = V^*(s) + A^*(s,a) - \max_a A^*(s,a) \tag{27}$$

根据式(27)可将图 4(a)所示结构改进为图 4(b)所示的对抗判别网络结构。在此结构中,Critic 网络相当于维护了 2 个值函数。状态值函数表示在给定状态下做出任何动作的期望回报,而动作值函数表示在给定状态下采取某个动作的期望回报。该结构可以让智能体的判别网络在学习过程中能够更好地区分状态值和动作值,强化两者之间的关联从而更准确地估计 $Q$ 值,同时也可缓解状态值和动作值不平衡的问题。

#### 2.2.2　引入优先经验回放机制

强化学习算法随着实验环境复杂度的增加,导致系统内智能体的个数增加,状态,观测维度扩张,这严重影响着算法训练效率和质量。传统的经验回放方法将智能体与环境交互得到的经验直接全部存储在经验池[19],在训练时从经验池中随机采样选取经验进行学习。原始的经验回放机制不能分辨哪些经验更重要,毫无保留的存储所有经验,其中高质量的经研虽然有利于算法的进一步训练,但由于采样会抽取大量低质量经验,这就导致算法的训练效率很低,消耗大量时长。进步的,优先经验回放[20]缓解了前面所述问题。优先经验回放的基本思想是通过一个重放缓存器来存储经验,并为每个经验设置一个优先级,以非均匀抽样代替均匀抽样。该机制中优先级由 TD 误差的绝对值 $|\delta_j|$ 表示,当 $|\delta_j|$ 越大说明算法对此时状态动作价值评估不准确,应该给该经验较高的权重。在抽样时,有 2 种方法可计算抽样概率,方法一用式(28)计算概率:

$$p_j \propto |\delta_j| + \varepsilon \tag{28}$$

式(28)中: $\varepsilon$ 定义为很小的数,防止抽样概率接近 0,用于保证所有样本都以非 0 的概率抽到。第 2 种抽样方式对 $|\delta_j|$ 降序排序,然后以式(29)计算抽样概率:

$$p_j \propto \frac{1}{rank(j)} \tag{29}$$

式(29)中:$rank(j)$是$|\delta_j|$的序号,$|\delta_j|$越大$rank(j)$越小。上述2种抽样方式原理一致,即$|\delta_j|$越大样本被抽到的概率越大。由于是非均匀抽样,不同的样本有不同的抽样概率,这导致算法的预测存在偏差,此时应该调整相应的学习率来抵消不同抽样概率造成的偏差。优先经验回放数组如表1所示。

表1中$b$为数组大小,需手动设置,如果样本数量超过$b$,则需删除回放池中最旧的样本。通过该方式,算法在学习阶段可以按照优先级从高到低依次在重放缓存器中选取经验进行训练,从而更加充分地利用经验信息,提高学习效率和稳定性。这些经验在当前策略下具有较高的优先级,能够优先进行学习,避免错误信息的传递。传统的经验回放过程中,由于采样的经验是完全随机的,有时会导致某些重要的经验没有被选到,使得学习效果不如预期,而优先经验回放通过设置优先级来保持经验库的多样性,使得散度更大、更具代表性的经验更容易被重复使用,从而提高了学习的稳定性。

表1 优先经验回放数组

Table 1 Priority experience playback array

| 次数 | 经验序列 | TD目标 | 抽样概率 | 学习率 |
|---|---|---|---|---|
| … | … | … | … | … |
| $j-1$ | $(s_{j-1},a_{j-1},r_{j-1},s_j)$ | $\delta_{j-1}$ | $p_j\propto\lvert\delta_{j-1}\rvert+\varepsilon$ | $\alpha\cdot(b\cdot p_{j-1})^{-\beta}$ |
| $j$ | $(s_j,a_j,r_j,s_{j+1})$ | $\delta_j$ | $p_j\propto\lvert\delta_j\rvert+\varepsilon$ | $\alpha\cdot(b\cdot p)^{-\beta}$ |
| $j+1$ | $(s_{j+1},a_{j+1},r_{j+1},s_{j+2})$ | $\delta_{j+1}$ | $p_j\propto\lvert\delta_{j+1}\rvert+\varepsilon$ | $\alpha\cdot(b\cdot p_{j+1})^{-\beta}$ |
| … | … | … | … | … |

### 2.2.3 基于双Critic网络的价值评估

由于单一Critic网络可能会估计不准确,导致算法训练不稳定性和难以收敛等问题,本文中提出同时训练2个Critic网络来估计动作值函数,从而提高了对动作值函数的估计准确性。图5所示PerMaD4算法中的双价值网络结构,该算法同时维护了2个价值网络以及一个策略网络,分别如式(30)所示:

$$\begin{cases} q(s,a;\omega_1) \\ q(s,a;\omega_2) \\ \mu(s;\theta) \end{cases} \tag{30}$$

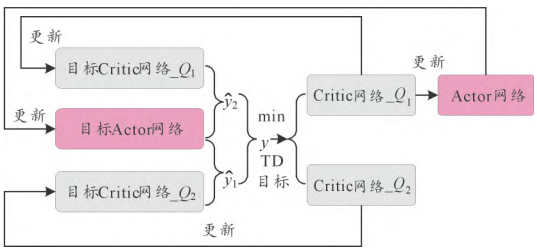



图5 PerMaD4算法的双价值网络结构

Fig. 5 The double-valued network structure of the PerMaD4

上述3个网络各自对应一个目标网络:

$$\begin{cases} q(s,a;\omega_1^-) \\ q(s,a;\omega_2^-) \\ \mu(s;\theta^-) \end{cases} \tag{31}$$

算法使用目标策略网络来预测动作:

$$\hat{a}_{j+1}^- = \mu(s_{j+1};\theta^-) \tag{32}$$

由于该结构中存在2个目标价值网络,因此计算TD误差时,先用式(33)计算:

$$\begin{cases} \hat{y}_{j,1} = r_i + \gamma\cdot q(s_{j+1},\hat{a}_{j+1}^-;\omega_1^-) \\ \hat{y}_{j,2} = r_i + \gamma\cdot q(s_{j+1},\hat{a}_{j+1}^-;\omega_2^-) \end{cases} \tag{33}$$

然后,将计算出来的最小者定义为TD目标:

$$\hat{y}_j = \min\{\hat{y}_{j,1},\hat{y}_{j,2}\} \tag{34}$$

## 3 基于PerMaD4的协同电子对抗

### 3.1 动作空间设计

无人机搭载侦察探测设备,并且其可发射干扰信号干扰地面雷达的正常工作。为缩减无人机集群的操作空间规模,本研究对某些行为进行了分离简化的处理方法,智能体动作空间见表2。

表2 动作空间

Table 2 Action state

| 字段 | 说明 |
|---|---|
| 飞行动作 | 前,后,左,右,悬停 |
| 飞行速度 | 低速,中速,高速 |
| 侦察方向 | 左前方,正前方,右前方 |
| 干扰强度 | 0,低,中,高 |
| 干扰频段 | 低,中,高 |
| 干扰目标 | 与雷达个数一致 |

### 3.2 状态空间设计

无人机的状态空间分为2个区域,一是代表全局环境状况的环境状态空间,如表3所示。二是代表无人机自身状态以及对环境监测结果的智能体观测空间,如表4所示。

表 3　环境空间

Table 3 Environment state table

| 字段 | 说明 |
|---|---|
| 雷达位置 | 地面雷达的坐标 |
| 雷达频率 | 地面雷达信号发射的频率 |
| 雷达探测范围 | 雷达实时的探测距离 |
| 导弹系统位置 | 防空导弹系统的坐标 |
| 目标位置 | 敌方基地的坐标 |

表 4　智能体观测空间

Table 4 Agent observation table

| 字段 | 说明 |
|---|---|
| 位置 | 无人机的坐标 |
| 朝向 | 无人机的飞行方向 |
| 速度 | 无人机的飞行速度 |
| 方向 | 无人机的定向侦察方向 |
| 强度 | 无人机的干扰强度 |
| 频段 | 无人机的干扰频段 |
| 续航 | 无人机剩余生命值 |
| 标志位 | 是否开启干扰 |
| 目标 | 定位到的雷达与敌方基地位置 |

### 3.3　奖励函数设计

无人机之间需要协同完成任务,如果距离太远将无法完成通信,因此需要设置无人机之间的距离奖励:

$$R_1 = \begin{cases} 100 \times \left(\frac{C}{D} - 1\right), & D > C \\ 0, & D \leqslant C \end{cases} \tag{35}$$

式(35)中: $C$ 表示无人机之间的通信距离; $D$ 表示无人机之间的距离。

接近目标时奖励为

$$R_2 = \begin{cases} 500 \times \dfrac{D_{\text{blue}} - d_{\text{now}}}{d_{\text{last}} - D_{\text{red}}} & d_{\text{now}} > D_{\text{blue}} - 2 \\ 100 \times \left(1 - \dfrac{d_{\text{now}}}{D_{\text{blue}}}\right), & \text{other} \end{cases} \tag{36}$$

被雷达探测的奖励:

$$R_3 = -10 \tag{37}$$

发现雷达的奖励:

$$R_4 = 20 \tag{38}$$

干扰到雷达的奖励:

$$R_5 = 1\,000 \times \left(1 - \frac{D_{\text{blue\_now}}}{D_{\text{blue}}}\right) \tag{39}$$

式(39)中: $D_{\text{blue\_now}}$ 表示被干扰后雷达的探测距离, $D_{\text{blue}}$ 表示雷达最大探测距离。

无人机被击落的奖励:

$$R_6 = -100 \tag{40}$$

开辟投弹区域的奖励:

$$R_7 = 100 \tag{41}$$

总奖励 $R$:

$$R = R_1 + R_2 + R_3 + R_4 + R_5 + R_6 + R_7 \tag{42}$$

### 3.4　任务完成度设计

根据仿真任务可设计在每个仿真回合内,观察蓝方阵地是否被打击成功,如果打击成功设 $D_i = 1$,否则 $D_i = 0$。任务完成度计算如下:

$$\sigma_D = \frac{n}{N} \tag{43}$$

式(43)中: $n$ 表示在所有回合中, $D_i = 1$ 的回合次数; $N$ 表示所有回合次数。

### 3.5　仿真实验参数配置

在保证实验场景复杂度,基本算法参数一致的情况下,在自主设计的多智能体电子协同对抗作战环境中分别采用 PerMaD4 算法, MaTD3 算法, MADDPG 算法, PerMADDPG 4 组算法进行对比实验。4 种算法基本的公共超参数如表 5 所示。

表 5　超参数

Table 5 Hyperparameters

| 参数 | 数值 | 参数 | 数值 |
|---|---|---|---|
| Actor 学习率 | 5e-4 | Buffer Size | 5e5 |
| Critic 学习率 | 5e-4 | Batch Size | 256 |
| 折扣率 | 0.9 | | |

该实验环境部分运行在 Core i7-11700K 处理器, Win10 系统的 Unity3D 软件中; 算法部分基于 Python3.8 版本,使用 Pytorch 深度学习框架以及包括 Numpy, Pygame 在内的第三方库编写。

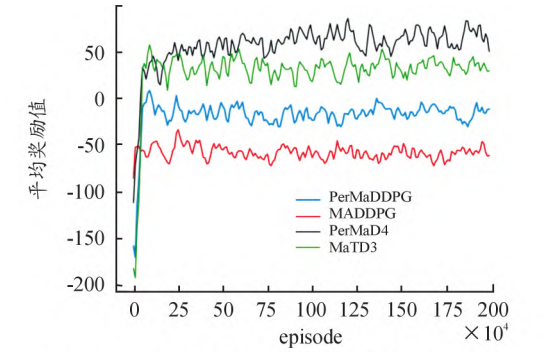
## 4　实验结果及分析

本文中将自主设计的电子对抗仿真作战平台用以算法的训练实验环境,可以验证算法的可靠性与进步性。在该平台中,将 PerMaD4 算法, PerMADDPG 算法, MADDPG 算法, MATD3 算法 4 组算法分别进行 2 000 000 个 episode 实验周期的训练,记录并保存实验结果。在这个实验中,用于衡量算法优越性的评估标准是在每个实验周期内获得的总体平均奖励以及任务的完成程度。

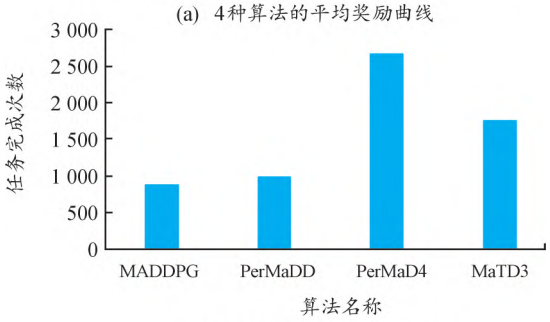
### 4.1　实验结果与分析

由于训练次数庞大,获得的奖励值数值较多,直接绘图会导致曲线很粗无法区分。因此,做平滑处理得到图 6(a)。图 6 表明: 在同一复杂度任务中, 4 种算法均大致在第 250 000回合之后平均回报趋于平缓,仿真结束均可达到收敛,说明该电子对抗作战仿真平台设计合理。统计第

250 000 回合后的奖励值平均结果见表 6。




(a) 4种算法的平均奖励曲线




(b) 4种算法的任务完成度统计

**图 6 4 种算法的实验结果**

Fig. 6 Experimental results of four algorithms

数值表明: PerMaD4 算法在电子对抗环境中的表现优于其他算法,验证了该算法的进步性以及稳定性; 另一方面,图 6( b) 统计了在统一的回合次数下,各算法的任务完成次数。该结果表明: 智能体使用 PerMaD4 算法训练的任务完成次数是 4 种算法中最高的,因此验证该算法的进步性以及稳定性。

**表 6 4 种算法的平均奖励值统计**

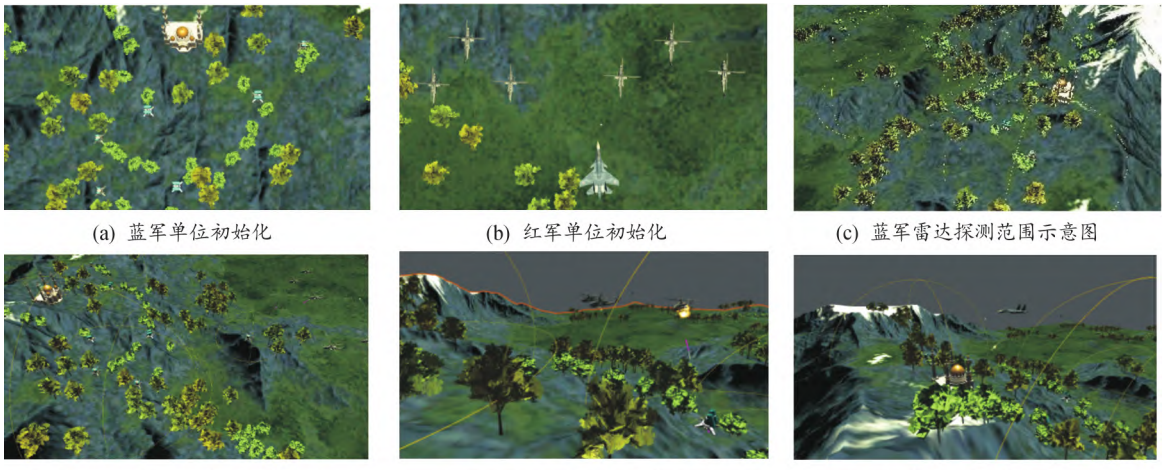
Table 6 The average return

| 算法名称 | 平均奖励 | 算法名称 | 平均奖励 |
|---|---|---|---|
| PerMaD4 | 52 | MaTD3 | 39.5 |
| PerMaDDPG | -12.5 | MADDPG | -60.5 |

### 4.2 实验结果可视化

实验开始前,初始化仿真环境,红军 6 架无人机,1 架轰炸机,蓝军有 4 个防空雷达( 自带火炮系统) ,3 个导弹系统。红军面临的特定职责包括: 无人机集群根据训练好的策略协同工作,找出蓝方目标单位,此期间需探测蓝方雷达的位置并实施干扰,使蓝军雷达暂时失去探测能力,为后方的轰炸机开辟投弹通道,以完成打击蓝军目标的任务。图 7( a) 为初始化蓝军单位; 图 7( b) 为初始化红军单位。图 7( c) 模拟蓝军防空雷达的探测范围,图 7( d) 模拟蓝军部分雷达被干扰或打击,其探测功能减弱,仿真效果表现为其探测范围减小。图 7( e) 模拟红军无人机集群遭受蓝军火炮拦截的场景,图中部分红军无人机被发现并被击落。图 7( f) 模拟蓝军重点单位被摧毁前的状态,此刻红军导弹正在逼近。

图 8 统计了使用 2 种算法分别迭代 2 000 000 次且趋于稳定时候的无人机集群飞行路径。图 8( a) 中,第 1 号无人机并没有探索到雷达,最终碰壁坠毁; 第 5 和第 6 号无人机在探索途中被击落; 仅第 2、3、4 号无人机发现并成功干扰到雷达,由于无人机集群干扰雷达的成功率不高,导致该算法的任务完成度相对较低。图 8( b) 中,仅有第 2、6 号无人机被摧毁; 第 1、3、4、5 号无人机成功探测并持续干扰雷达,大大提升该算法的任务完成度。

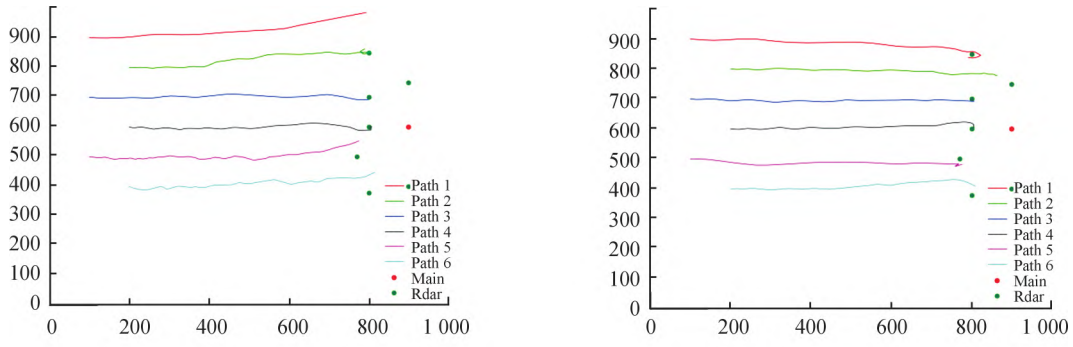


(a) 蓝军单位初始化 (b) 红军单位初始化 (c) 蓝军雷达探测范围示意图

(d) 蓝军部分雷达被干扰或打击 (e) 红军无人机被打击画面 (f) 打击蓝军重点保护单位

**图 7 电子对抗作战平台**

Fig. 7 Electronic warfare platform




(a) 使用MADDPG算法的无人机路径　(b) 使用PerMaD4算法的无人机路径

**图8　无人机集群飞行路径**

Fig. 8 Drone cluster flight path

## 5　结论

针对仿真环境复杂,智能体观测空间、动作空间维度增加,经验回放池收集的经验质量低,采样效率不高,评估模块消耗大量计算资源,出现评估不准确,算法难以收敛等问题,提出基于优先经验回放的多智能体双对抗策略梯度算法:

1) 通过引入优先经验回放机制,充分地利用智能体与环境交互得到的经验信息,从而提高算法的学习效率和稳定性。

2) 提出双对抗判别网络,可准确有效的分析智能体输入状态和动作的相对贡献和估计优化价值网络。

3) 该算法较于其他经典强化学习算法任务完成度提高8.9%。

## 参考文献:

[1] 李洪,王超,王睿. 关于电子战发展趋势的一些思考[J]. 中国军转民,2023(1):57-59.
LI Hong,WANG Chao,WANG Rui. Some reflections on the development trends of electronic warfare[J]. China Journal of Military to Civilian Transition,2023(1):57-59.

[2] 王健,杨渡佳,黄科举,等. 认知电子战发展趋势:从单体智能到群体智能[J]. 信息对抗技术,2023,2(Z1):151-170.
WANG Jian,YANG Dujia,HUANG Keju,et al. The development trend of cognitive electronic warfare: From single agent intelligence to group intelligence[J]. Information Countermeasures Technology,2023,2(Z1):151-170.

[3] 苏周,刘飞,许晓剑,等. 智能化电子战装备发展研究[J]. 舰船电子对抗,2023,46(4):9-13,18.
SU Zhou,LIU Fei,XU Xiaojian,et al. Research on the development of intelligent electronic warfare equipment[J]. Ship Electronic Countermeasures,2023,46(4):9-13,18.

[4] 李博骁,张峰,李奇峰,等. 人工智能技术在军事领域的应用思考[J]. 中国电子科学研究院学报,2022,17(3):238-246.
LI Boxiao,ZHANG Feng,LI Qifeng,et al. Reflections on the application of artificial intelligence technology in the military field[J]. Journal of the Chinese Academy of Electronic Sciences,2022,17(3):238-246.

[5] 崔世亮,刘广斌. 基于专家系统的船舶结构优化设计[J]. 船舶物资与市场,2022,30(5):24-26.
CUI Shiliang,LIU Guangbing. Ship structure optimization design based on expert system[J]. Ship Materials & Market,2022,30(5):24-26.

[6] 孙乐. 军事领域中科技专家系统的应用与效能评估[J]. 舰船电子工程,2022,42(1):16-18.
SUN Yue. Application and effectiveness evaluation of scientific and technological expert system under the background of joint operations[J]. Ship Electronic Engineering,2022,42(1):16-18.

[7] CASAL-GUISANDE,COMESAÑA-CAMPOS,PEREIRA,et al. A decision-making methodology based on expert systems applied to machining tools condition monitoring[J]. Mathematics,2022,10(3):520.

[8] 马也,范文慧,常天庆. 基于智能算法的无人集群防御作战方案优化方法[J]. 兵工学报,2022,43(6):1415-1425.
MA Ye,FAN Wenhui,CHANG Tianqing. Optimization method of unmanned swarm defensive combat scheme based on intelligent algorithm[J]. Acta Armamentarii,2022,43(6):1415-1425.

[9] 胡振震,陈少飞,袁维淋,等. 基于粒子群优化的德州扑克在线对手利用[J]. 控制与决策,2023,39(5):1687-1696.
HU Zhenzhen,CHEN Shaofei,YUAN Weilin,et al. Based on particle swarm optimization texas poker online opponents

use [J]. Control and Decision,2023,39(5):1687 - 1696.

[10] 李杰,谭跃进. 基于集成改进蚁群算法的作战环推荐方法 [J]. 系统工程与电子技术,2023,31(5):1 - 13.
LI Jie,TAN Yuejin. Combat ring recommendation method based on integrated improved ant colony algorithm [J]. Systems Engineering and Electronics,2023,31(5):1 - 13.

[11] 杨益,任辉启. 智能化战争条件下军事设施拓扑防护构想 [J]. 防护工程,2021,43(5):68 - 73.
YANG Yi,REN Huiqi. Concept of topological protection for military facilities under intelligent warfare conditions [J]. Protective Engineering,2021,43(5):68 - 73.

[12] 韦强,赵书文. 人工智能推动战争形态演变 [J]. 军事文摘,2017(13):54 - 57.
WEI Qiang,ZHAO Shuwen. Artificial intelligence promotes the evolution of war forms [J]. Military Abstracts,2017(13):54 - 57.

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

[14] LILLICRAP T P,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning [C]//The International Conference on Learning Representations,San Juan,Puerto Rico,2016:174 - 193.

[15] LOWE R,WU Y,TAMAR A,et al. Multi-agent actor-critic for mixed cooperative competitive environments [C]//USA:MIT Press,2017:6379 - 6390.

[16] 唐峯竹,唐欣,李春海,等. 基于深度强化学习的多无人机任务动态分配 [J]. 广西师范大学学报(自然科学版),2021,39(6):63 - 71.
TANG Fengzhu,TANG Xin,LI Chunhai,et al. Dynamic assignment of multiple unmanned aerial vehicle tasks based on deep reinforcement learning [J]. Journal of Guangxi Normal University (Natural Science Edition),2021,39(6):63 - 71.

[17] 施伟,冯旸赫,程光权,等. 基于深度强化学习的多机协同空战方法研究 [J]. 自动化学报,2021,47(7):1610 - 1623.
SHI Wei,FENG Yanghe,CHENG Guangquan,et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning [J]. Acta Automatica Sinica,2021,47(7):1610 - 1623.

[18] 薛喜地. 基于深度强化学习的室内无人机避障 [D]. 哈尔滨:哈尔滨工业大学 2021.
XUE Xidi. Collision avoidance for indoor uav based on deep reinforcement learning [D]. Harbin:Harbin Institute of Technology,2021.

[19] LATHUILIERE S,MASSE B,MESEJO P,et al. Neural network based reinforcement learning for audio visual gaze control in human robot interaction [J]. Pattern Recognition Letters,2017(4):1 - 10.

[20] 代学武,吴越,石琦,等. 基于优先经验回放可迁移深度强化学习的高铁调度 [J]. 控制与决策,2023,38(8):2375 - 2388.
DAI Xuewu,WU Yue,SHI Qi,et al. Based on reinforcement learning experience playback can be migrated depth priority high-speed rail dispatching [J]. Control and decision,2023,38(8):2375 - 2388.

科学编辑 李波 博士(西北工业大学 教授)
责任编辑 唐定国