用于巡航导弹突防航迹规划的改进深度强化学习算法

马子杰，高　杰，武沛羽，谢拥军
（北京航空航天大学 电子信息工程学院，北京 100191）

摘　要： 为了解决巡航导弹面临动态预警机雷达威胁下的突防航迹规划问题，提出一种改进深度强化学习智能航迹规划方法。针对巡航导弹面对预警威胁的突防任务，构建了典型的作战场景，给出了预警机雷达探测概率的预测公式，在此基础上设计了一种引入动态预警威胁的奖励函数，使用深度确定性策略梯度网络算法（Deep Deterministic Policy Gradient，DDPG）探究巡航导弹智能突防问题。针对传统DDPG算法中探索噪声时序不相关探索能力差的问题，引入了奥恩斯坦-乌伦贝克噪声，提高了算法的训练效率。计算结果表明，改进的DDPG算法训练收敛时间更短。

关键词： 巡航导弹；DDPG算法；突防策略；深度强化学习
中图分类号： TN959.1；TP181　　**文献标识码：** A　　**DOI：**10.16157/j.issn.0258-7998.211934

中文引用格式： 马子杰，高杰，武沛羽，等. 用于巡航导弹突防航迹规划的改进深度强化学习算法[J].电子技术应用，2021，47(8)：11-14，19.
英文引用格式： Ma Zijie，Gao Jie，Wu Peiyu，et al. An improved deep reinforcement learning algorithm for cruise missile penetration path planning[J]. Application of Electronic Technique，2021，47(8)：11-14，19.

An improved deep reinforcement learning algorithm for cruise missile penetration path planning

Ma Zijie，Gao Jie，Wu Peiyu，Xie Yongjun
（School of Electronics and Information Engineering，Beihang University，Beijing 100191，China）

Abstract： Aiming at the problem of cruise missile penetration trajectory planning under the threat of dynamic early of warning aircraft radar，an improved deep reinforcement learning intelligent trajectory planning method is proposed. Firstly，aiming at the penetration mission of cruise missiles facing early warning threats，a typical combat scenario is constructed，and a prediction formula of radar detection probability of early warning aircraft is given. On this basis，a reward function that introduces dynamic early warning threats is designed，and the deep deterministic policy gradient algorithm(DDPG) is used to explore the intelligent penetration of cruise missiles. And then，in response to the poor exploration ability of the traditional DDPG algorithm that explores the uncorrelated timing of noise，Ornstein-Uhlenbeck noise is introduced to improve the training efficiency of the algorithm. The simulation results show that the improved DDPG algorithm training convergence time is shorter.

Key words： cruise missile；deep deterministic policy gradient algorithm；penetration strategy；deep reinforcement learning

0 引言

巡航导弹是一种能机动发射、命中精度高、隐蔽性强、机动性能强的战术打击武器，但近年来由海陆空防御武器整合得到的体系化信息化反导防御系统态势感知能力和区域拒止能力都得到了极大的提升，巡航导弹的战场生存能力受到威胁，提升巡航导弹规避动态威胁的能力成为其能否成功打击目标的关键[1-3]。传统的巡航导弹航迹规划方法中将雷达威胁建模为一个静态的雷达检测区域，这难以适应对决策实时性要求较高的动态战场环境，而且其缺乏探索先验知识以外的突防策略的能力，需要研究能应对动态对抗的巡航导弹智能航迹规划算法。

深度强化学习是人工智能领域新的研究热点[4-6]。随着深度强化学习研究的深入，其开始被应用于武器装备智能突防，文献[7]利用深度强化学习提出了一种新的空空导弹制导律，提高了打击目标的能力。文献[8]针对目标、打击导弹、拦截导弹作战问题，探究了是否发射拦截导弹、拦截导弹的最佳发射时间和发射后的最佳导引律。文献[9]利用深度价值网络算法探究了静态预警威胁下的无人机航迹规划问题，提升了航迹规划的时间。文献[10]将雷达威胁建模为一个静态的雷达检测区域，在二维平面探究了巡飞弹动态突防控制决策问题，提高了巡飞弹的自主突防能力。

综上所述，目前巡航导弹智能突防研究中针对预警

雷达的威胁建模都属于静态建模，其设定预警机威胁区域固定，而实际战场环境中预警机是动态的，因而其威胁区域也是动态变化的。因此，本文提出了两点改进：(1)对预警机威胁进行动态建模，给出了预警机雷达探测概率的预测公式；(2)使用深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)算法训练时引入了时序相关的奥恩斯坦-乌伦贝克随机过程作为探索噪声，解决了收敛难度加大的问题，进而缩短了算法的训练时间。

1 DDPG 算法及其改进

DDPG[11-13]是深度强化学习应用于连续控制强化学习领域的一种重要算法，将确定性策略梯度算法与 Actor-Critic 框架相结合，提出了一个任务无关的模型，并可以使用相同的参数解决众多任务不同的连续控制问题。DDPG 采取经验回放机制，通过目标网络的参数不断与原网络的参数加权平均进行训练，以避免振荡。深度确定性策略梯度算法流程如下：

输入：环境；

输出：最优策略的估计；

参数：学习率 $\alpha^{(w)}$、$\alpha^{(\theta)}$、折扣因子 γ、控制回合数和回合内步数的参数、目标网络学习率 $\alpha_{目标}$。

(1) 初始化网络参数：$\theta \leftarrow$任意值，$\theta_{目标} \leftarrow \theta$，$w \leftarrow$任意值，$w_{目标} \leftarrow w$。

(2) For episode=1，M do(M 为仿真最大回合数)

(3) 用对 $\pi(S;\theta)$加扰动进而确定动作 A

(4) 执行动作 A，观测到收益 R 和下一状态 S'

(5) 将经验(S,A,R,S')储存在经储存空间 D

(6) For $t=1$，T do(T 为仿真终止时间)

(7) 从存储空间 D 采样出一批经验 B

(8) 为经验估计回报$U \leftarrow R+\gamma q(S',\pi(S';\theta_{目标});w_{目标})$

(9) 更新 w 以减小$-\frac{1}{|B|}\sum_{(S,A,R,S')\in B}[U-q(S,A;w)]^2$

(10) 更新 θ 以减小$-\frac{1}{|B|}\sum_{(S,A,R,S')\in B}[q(S,\pi(S;\theta);w)]$

(11) 更新目标网络和目标策略：

$w_{目标} \leftarrow (1-\alpha_{目标})w_{目标}+\alpha_{目标}w$，

$\theta_{目标} \leftarrow (1-\alpha_{目标})\theta_{目标}+\alpha_{目标}\theta$。

1.1 DDPG 算法求解流程

DDPG 算法的网络结构为 Actor-Critic 网络结构，其中 Actor 网络输入状态，输出动作，Critic 网络输入状态和动作，输出在这一状态下采取这个动作的评估 Q 值，其示意图如图 1 所示。由于巡航导弹、目标和预警机的状态动作信息是一个在时间上连续的序列，因此由状态构成的样本之间并不具备独立性，只使用单个神经网络结构学习过程很不稳定。为解决这个问题，DDPG 算法引入了经验回放机制，引入目标 Actor 网络和目标 Critic 网络，与现实网络独立训练。首先现实 Actor 网络与环境进行交互训练，得到状态 S、动作 a、奖励 r、下一时刻状态 S'，将这 4 个数据放入经验池中，得到一定的样本空间后，现实 Critic 网络从经验池中提取样本进行训练得到 Q 值；目标网络也进行同样的训练，每间隔一定时间就利用现实网络参数更新目标网络。训练完成后可以通过 Actor 网络得到高维的具体动作，可解决连续动作空间学习问题。其求解流程如图 2 所示。

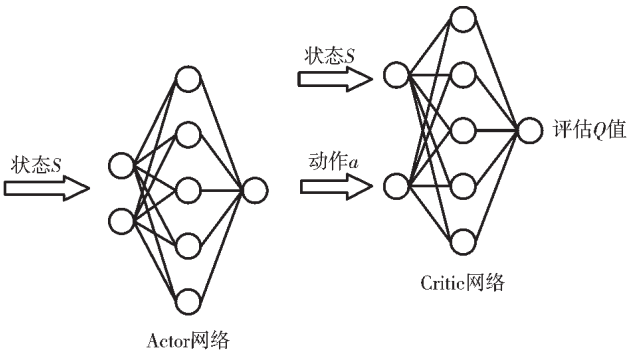

图 1 Actor-Critic 网络结构

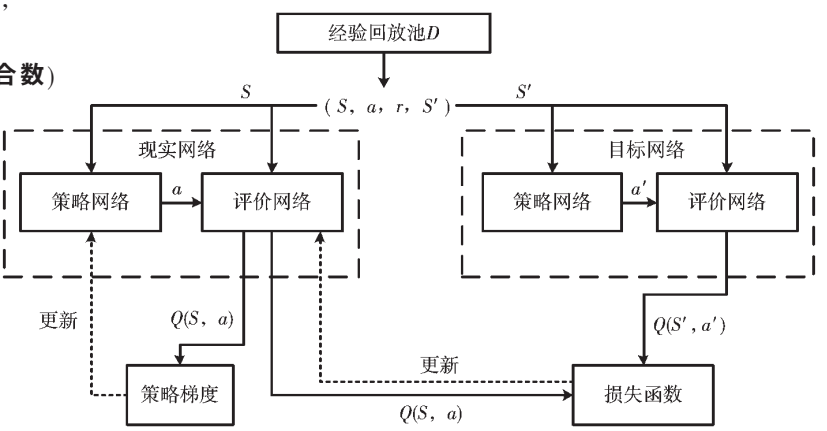

图 2 DDPG 算法流程图[10]

1.2 DDPG 算法的改进

1.2.1 时序相关的探索噪声

传统的 DDPG 算法中的探索噪声为高斯噪声，其在时序上不相关，对时序相关的问题探索能力差，探索时间长；导弹突防过程属于惯性过程，引入时序相关的奥恩斯坦-乌伦贝克随机过程可以提高在惯性系统中的控制任务的探索效率，使训练更快收敛。奥恩斯坦-乌伦贝克过程满足的微分方程为：

$$dx_t=\theta(\mu-x_t)dt+\sigma dW_t \tag{1}$$

其中，x_t 为过程刻画的量；θ 为比例系数；μ 是 x_t 的均值；W_t 为维纳过程，是一种随机噪声；σ 是随机噪声的权重。

1.2.2 动态预警威胁

预警机是一种装有远距离搜索雷达、数据处理、敌我识别以及通信导航、指挥控制、电子对抗等完善的电

《电子技术应用》http://www.chinaaet.com

子设备，用于搜索、监视与跟踪空中和海上目标并指挥、引导己方飞机执行作战任务的作战支援飞机，起到活动雷达站和空中指挥中心的作用，是现代战争中重要的武器装备。DDPG 算法应用于突防策略研究时，一般将预警机雷达威胁简化为一个静态的禁飞区，但这样无法反映真实作战场景下巡航导弹遇到的动态预警威胁，因此在解决巡航导弹突防航迹规划问题时，需要在 DDPG 算法中引入预警机雷达动态探测概率预测公式。

E2-D 预警机的雷达在一定的虚警概率下，一次扫描对目标的发现概率为[14]：

$$P_d=\int_0^{\infty} e^{-t}\left\{1-\varphi\left[\frac{y_0-n_o(1+\frac{S}{N}t)}{\sqrt{n_o(1+2\frac{S}{N}t)}}\right]\right\}dt \tag{2}$$

式中：

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt \tag{3}$$

其中，n_o 为一次扫描脉冲积累数，y_o 为虚警时的检测门限，S/N 为信噪比。

对其曲线进行拟合可得预警雷达探测瞬时概率与目标的雷达散射截面积值 σ 和目标与雷达的 R 的计算公式为：

$$P_d=\left[1+\left(\frac{c_2R^4}{\sigma}\right)^{c_1}\right]^{-1} \tag{4}$$

其中，R 的单位为 km，σ 的单位为 m^2；c_1 与 c_2 和雷达的工作模式和场景有关，本文分别取为 1.5 和 $1.5\times10^{-8.5}$。

2 巡航导弹突防决策模型

巡航导弹突防过程为一个马尔科夫决策过程(Markov Decision Process，MDP)，需要对导弹运动模型、状态空间、动作空间、奖励函数进行建模。

2.1 导弹运动模型

可以对突防过程的弹道进行简化：导弹和预警机均可视为质点，巡航导弹采用 3 自由度质点运动。

2.2 状态空间设计

由于对抗双方均设为质点，可以将巡航导弹、目标、预警机的质心位置 o_t、弹目质心位置的距离 l_t 以及航向角 φ_t 作为状态空间，即状态空间为 $s_t=[o_t,l_t,\varphi_t]$。

2.3 动作空间设计

巡航导弹处在一个连续的动作空间，其动作空间设为巡航导弹在 x、y、z 3 个方向的速度分量，即 v_x、v_y、v_z。

2.4 奖励函数设计

2.4.1 导弹成功击中目标奖励

巡航导弹采取突防策略的主要目的是在避开预警威胁的情况下，成功击中目标。其奖励函数为：

$$r_1=\begin{cases}1\ 000 & \text{击中目标}\\ 0 & \text{未击中目标}\end{cases} \tag{5}$$

2.4.2 导弹和目标相对距离奖励

导弹可目标的距离越近，导弹击中目标的可能性越大，其奖励函数为：

$$r_2=10e^{-\frac{l_t}{10}} \tag{6}$$

其中，l_t 为导弹与目标当回合的距离。

2.4.3 导弹速度和弹目连线夹角奖励

导弹速度和弹目连线夹角即为视线角，视线角越小，巡航导弹击中目标的可能性越大，其奖励函数为：

$$r_3=10e^{-\frac{\varphi_t}{\pi}} \tag{7}$$

其中，φ_t 为导弹与目标当回合的视线角。

2.4.4 视线角变化率奖励

视线角变化率奖励的具体形式为：

$$r_4=5\arctan(\varphi_t-\varphi_{t-1}) \tag{8}$$

其中，φ_{t-1} 为导弹与目标上一回合的视角。

2.4.5 探测概率降低奖励

$$r_5=\begin{cases}-kP_d & P_d\geqslant 50\%\\ 0 & P_d<50\%\end{cases} \tag{9}$$

其中，P_d 为雷达探测概率，k 为比例系数。

综合考虑上述 5 种奖励模型，每回合巡航导弹的动作奖励为：

$$r_i=r_1+r_2+r_3+r_4+r_5 \tag{10}$$

训练完成后的总奖励为：

$$R=\sum_{i=1}^{m} r_i \tag{11}$$

3 仿真结果与分析

3.1 仿真场景及武器性能参数

仿真场景主要对巡航导弹、攻击目标、预警机的位置、机动参数和机动范围进行设置。作战场景如图 3 所示，主要为巡航导弹、目标和预警机的空间位置关系。预警机在 7 500 m 高度以“跑道形”巡逻线探测巡航导弹，直边长度为 70 km，弧线半径为 15 km，航线中心点坐标为东经 119.5°、北纬 20°、海拔 7 500 m。巡航弹目标为位于东经 120°、北纬 20°、海拔 15 m 的宙斯盾舰船，巡航导弹的发射点位于东经 117.5°、北纬 20°、海拔 15 m。其中巡航导弹的最大巡航速度为300 m/s，目标的最大速度为 200 m/s；当巡航导弹和目标的相对距离小于 0.05 km 时假定巡航导弹击中目标。

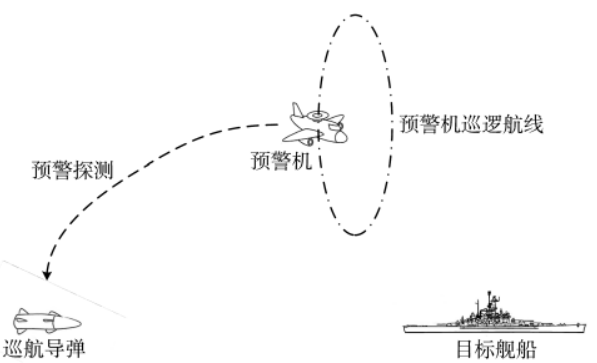

图 3 巡航导弹突防典型作战场景

3.2 软硬件环境及参数设置

仿真的软件环境为：Windows 10、Python3.7 以及 TensorFlow 架构，硬件环境为 GTX2060 和 64 GB DDR4 内存。Actor、Critic 神经网络结构均采用 2 层隐藏层的全连接神经网络，隐藏单元数为 256 和 32；超参数设置如下：学习率为 0.000 1，折扣因子为 0.95，目标网络更新系数为 0.005，经验回放池容量为 10 000。

3.3 仿真结果分析

分别使用传统 DDPG 算法和改进 DDPG 算法对巡航导弹应对动态预警威胁突防进行训练，其每回合奖励值曲线如图 4 所示，数据对比如表 1 所示。

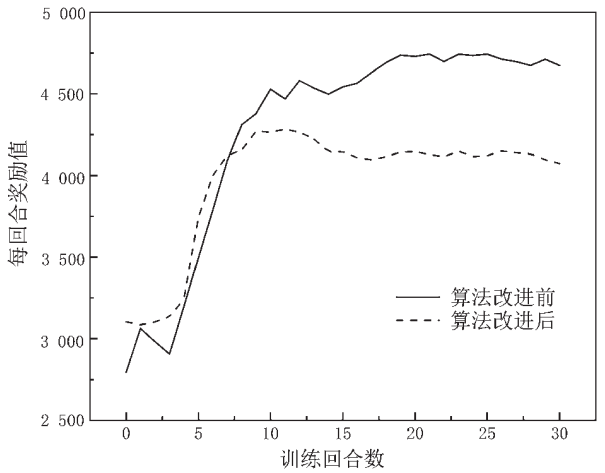

图 4 算法改进前后不同训练回合数下的奖励值

表 1 算法改进前后数据对比

算法	训练时间/min	收敛回合数	GPU 占用
传统 DDPG 算法	65	20	31%
改进 DDPG 算法	32	9	31%

改进后的 DDPG 算法由于其探索噪声时序相关，探索能力更高，收敛速度更快，相较于传统的算法模型训练达到稳定时间缩短了一半，训练收敛后改进算法每回合探索步数更少，因而其稳定每回合奖励值更低。训练完成后模型能在 1 s 内生成巡航导弹自主避开预警威胁打击目标的机动轨迹指令。

图 5 为模型训练完成后测试模型得到的一个攻防场景图，其中预警机巡航轨迹为跑道型轨迹，目标直线航行，巡航导弹避开预警威胁后成功击中目标。

4 结论

本文首先构建了巡航导弹突防时的典型作战场景，给出了预警机雷达探测概率的预测公式；然后采用一种基于时序相关探索噪声的改进 DDPG 算法求解得到了巡航导弹快速智能突防算法。仿真实验表明，在预警机雷达威胁下采用上述算法巡航导弹可以实现快速主动突防。该模型的训练时间大约为 30 min，训练完成后可在 1 s 内生成突防机动轨迹，远远超过传统航迹规划

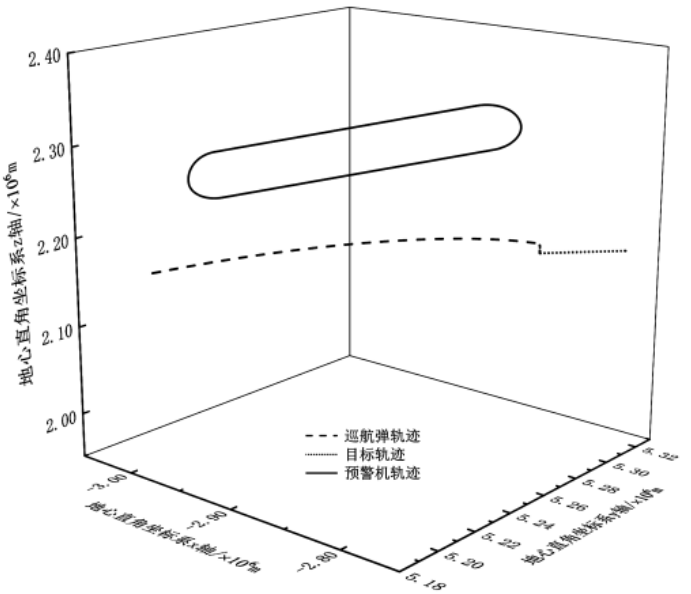

图 5 典型作战场景下训练后攻防轨迹图

算法的速度；而且该算法具备良好的适应性和延展性，可用于广泛的作战场景中。

参考文献

[1] 张凯杰，林浩申，夏冰.导弹集群智能突防技术的新发展[J].战术导弹技术，2018(5)：1-5，44.
[2] 戴全辉.巡航导弹武器系统伪装生存与隐身突防研究[J].战术导弹技术，2020(4)：41-46.
[3] 闫双卡，谭守林，滕和平，等.提高巡航导弹突防能力的技术途径[J].飞航导弹，2009(4)：26-29.
[4] 梁星星，冯旸赫，马扬，等.多 Agent 深度强化学习综述[J].自动化学报，2020，46(12)：2537-2557.
[5] SUTTON R S，BARTO A G.Reinforcement learning：an introduction[J].Trends in Cognitive Sciences，1999，3(9)：360-360.
[6] 曾家有，吴杰.智能反舰导弹发展需求及其关键技术[J].战术导弹技术，2018(2)：36-42.
[7] VAN HOORN，MARTIJN.Optimizing air-to-air missile guidance using reinforcement learning[D].Delft University of Technology，2019.
[8] SHALUMOV V.Cooperative online Guide-Launch-Guide policy in a target-missile-defender engagement using deep reinforcement learning[J].Aerospace Science and Technology，2020，104：105996.
[9] 邱月，郑柏通，蔡超.多约束复杂环境下 UAV 航迹规划策略自学习方法[J].计算机工程，2021，47(5)：44-51.
[10] 高昂，董志明，叶红兵，等.基于深度强化学习的巡飞弹突防控制决策[J]. 兵工学报，2021，42(5)：1101-1110.
[11] LILLICRAP T P，HUNT J J，PRITZEL A，et al.Continuous control with deep reinforcement learning[J].arXiv preprint arXiv：1509.02971.

(下转第 19 页)

wide stopband[J].IEEE Microw. Wireless Compon. Lett., 2014, 24(11): 742-744.
[7] CHEN X, YU X, Sun Sheng. Design of high-performance microstrip diplexers with stub-loaded parallel-coupled lines[J]. Electronics Letters, 2017, 53(15): 1052-1054.
[8] SHANG X, WANG Y, XIA W, et al. Novel multiplexer topologies based on all-resonator structures[J]. IEEE Trans. Microw. Theory Techn., 2013, 61(11): 3838-3845.
[9] GUAN X, YANG F, LIU H, et al. Compact and high-isolation diplexer using dual-mode stub-loaded resonators[J]. IEEE Microw. Wireless Compon. Lett., 2014, 24(6): 385-387.
[10] XIAO J K, ZHANG M, MA J G. A compact and high-isolated multiresonator-coupled diplexer[J]. IEEE Microw. Wireless Compon. Lett., 2018, 28(11): 999-1001.
[11] 黄巍. 面向无线通信的高性能平面双工器研究[D]. 南昌: 华东交通大学, 2018.
[12] VOSOOGH A, SORKHERIZI M S, ZAMAN A U, et al. An integrated Ka-band diplexer-antenna array module based on gap waveguide technology with simple mechanical assembly and no electrical contact requirements[J]. IEEE Transactions on Microwave Theory and Techniques, 2018, 66(2): 962-972.
[13] 李飞. 基于消失模的新型波导滤波器研究与设计[D]. 西安: 西安电子科技大学, 2017.
[14] DE PAOLIS F. Dimensional synthesis of evanescent-mode ridge waveguide bandpass filters[J]. IEEE Transactions on Microwave Theory and Techniques, 2018, 66(2): 954-961.
[15] 尤清春, 陆云龙, 尤阳, 等. 一种紧凑型的宽频带单脊波导功分器[J]. 电子元件与材料, 2018, 37(2): 50-54.

(收稿日期: 2021-04-02)

作者简介:
余亮(1996-), 男, 硕士研究生, 主要研究方向: 毫米波集成电路。
张健(1978-), 通信作者, 男, 博士, 研究员, 主要研究方向: 毫米波集成电路及无线通信系统设计, E-mail: zhangjian@hdu.edu.cn。
孙朋飞(1988-), 男, 博士, 讲师, 主要研究方向: 毫米波集成电路及应用。

扫码下载电子文档

(上接第 10 页)

magnetic wave reflected by reentry plasma sheath[J]. IEEE Transactions on Plasma Science, 2018, 46(5): 1755-1769.
[13] KUNDRAPU M, LOVERICH J, BECKWITH K, et al. Modeling radio communication blackout and blackout mitigation in hypersonic vehicles[J]. Journal of Spacecraft and Rockets, 2015, 52(3): 853-882.
[14] MICHAEL J R, STEPHENSON J J. A versatile three-dimensional ray tracing computer program for radio waves in the ionosphere[R]. Washington, DC: NTIA Technical Report OT-75-76, USA, 1975.
[15] YANG X, WEI B, YIN W. Relationship between the velocity of hypersonic vehicles and the transmission efficiency of radio waves[J]. Waves in Random and Complex Media, 2019, 29(2): 382-391.

(收稿日期: 2021-01-13)

作者简介:
杨鑫(1988-), 通信作者, 男, 博士, 副教授, 主要研究方向: 非均匀介质中电波传播的数值方法与应用, E-mail: e_yangx@126.com。

扫码下载电子文档

(上接第 14 页)

[12] TAN R, ZHOU J, DU H, et al. An modeling processing method for video games based on deep reinforcement learning[C]//2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, 2019: 939-942.
[13] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on Machine Learning, 2014: 387-395.
[14] 马东立, 罗勇, 裴旭. 低重频 PD 雷达的临界散射截面计算方法[J]. 北京航空航天大学学报, 2006, 32(9): 1011-1014, 1050.

(收稿日期: 2021-07-12)

作者简介:
马子杰(1997-), 男, 硕士研究生, 主要研究方向: 体系仿真、机器学习。
高杰(1997-), 男, 硕士研究生, 主要研究方向: 体系仿真、目标特性、计算电磁学。
谢拥军(1968-), 通信作者, 男, 教授, 博士生导师, 主要研究方向: 计算电磁学及应用、天线与微波工程、太赫兹技术等, E-mail: yjxie@buaa.edu.cn。

扫码下载电子文档

版权声明