基于深度强化学习的巡飞弹突防控制决策

高昂[1]，董志明[1]，叶红兵[2]，宋敬华[1]，郭齐胜[1]
(1. 陆军装甲兵学院 演训中心，北京 100072；2. 湘南学院，湖南 郴州 423099)

摘要： 巡飞弹突防控制决策(LMPCD)问题是“多域战”作战概念背景下的重要研究方向。针对该问题，建立基于马尔可夫决策过程的LMPCD模型。拟合LMPCD函数与飞行状态-动作值函数，构建基于演员-评论家方法的LMPCD框架，给出基于深度确定性策略梯度算法的深度强化学习模型求解方法，生成巡飞弹突防控制最优决策网络。通过1 000次巡飞弹突防仿真测试，结果表明，巡飞弹执行任务成功率为82.1%，平均决策时间为1.48 ms，验证了LMPCD模型及其求解过程的有效性。

关键词： 巡飞弹；深度强化学习；马尔可夫决策过程；突防；控制决策
中图分类号： E911 **文献标志码：** A **文章编号：** 1000-1093(2021)05-1101-10
DOI： 10.3969/j.issn.1000-1093.2021.05.023

Loitering Munition Penetration Control Decision Based on Deep Reinforcement Learning

GAO Ang[1], DONG Zhiming[1], YE Hongbing[2], SONG Jinghua[1], GUO Qisheng[1]
(1. Military Exercise and Training Center, Army Academy of Armored Forces, Beijing 100072, China;
2. Xiangnan University, Chenzhou 423099, Hunan, China)

Abstract: Loitering munition penetration control decision (LMPCD) is an important research direction under the concept of “multi-domain war”. The research on real-time route planning of loitering munition penetration has important military significance. Traditional knowledge, reasoning, and planning methods do not have the ability to explore and discover new knowledge outside the framework. The bionic optimization method is suitable for solving the path planning problem in static environment, such as traveling salesman problem, and is difficult to be applied to the penetration problem of loitering munition with high requirement of environmental dynamics and real-time decision-making. For the limitations of the first two methods, the applicability of the deep reinforcement learning method is analyzed, and the domain knowledge of loitering munition is introduced into each element of the deep reinforcement learning algorithm. The flight motion model of loitering munition is analyzed, the state space, action space and reward function of loitering munition are designed, the algorithm framework of loitering munition penetration control decision is analyzed, and the training process of loitering munition penetration control decision algorithm is designed. Through the penetration simulation test of 1 000 rounds of loitering munition, the result shows that the penetration success rate of loitering munition is 82.1% and the average decision time is 1.48 ms, which verifies the effectiveness of the algorithm training process and the

收稿日期： 2020-07-21
基金项目： 军队科研计划项目(41405030302、41401020301)
作者简介： 高昂(1988—)，男，博士研究生。E-mail: 236211566@qq.com
通信作者： 董志明(1977—)，男，教授，硕士生导师。E-mail: 15689783388@163.com

control decision model.

Keywords: loitering munition; deep reinforcement learning; Markov decision process; penetration; control decision

0 引言

按照全域机动,全域力量投送,创造领域优势,确保行动自由的“多域战”作战理念,巡飞弹这种飞航式智能弹药成为军事领域的重要发展方向[1-3]。巡飞弹如何在动态对抗环境中有效规避威胁、提高生存力是其执行作战任务成功与否的关键[4-5]。目前,巡飞弹航迹规划方法主要分为基于知识、推理、规划,仿生优化,学习3类方法[6]。第1类方法缺乏探索及发现框架之外新知识能力;第2类方法适用于求解旅行商这类静态环境下的路径规划问题,难以应用于动态对抗、决策实时性要求较高的环境;深度强化学习(DRL)属于第3类方法,DRL可以突破专家先验知识的限制,直接从高维战场空间中感知信息,并通过与环境不断交互优化模型。目前,采用DRL方法进行飞行器航迹规划的工作并不多。文献[7]在航迹终端约束条件下,基于DRL实现无人机从终端附近任意位置向目标点自主机动;文献[8]在城市环境中,基于DRL实现无人机从静态障碍物中通过,并到达指定目标区域。尽管飞行器控制在自主化方面已经取得了一定进展,但上述方法仍需要在更复杂的环境下进行进一步测试,例如动态环境中的航迹规划对飞行器来说仍然具有挑战性。本文考虑了存在潜在敌人威胁条件下,飞行器自主航迹规划问题,其难点在于飞行器在完成任务之前,并不知道威胁的数量、位置、策略,因此,必须学习一个合适的策略来对动态环境做出反应。具体来说,假设敌人的地空导弹雷达能够探测到一定范围内的巡飞弹,并能够影响巡飞弹在一定空间内的生存概率,因此巡飞弹必须学会在保证其自身不被摧毁的前提下完成突防任务。

1 基于马尔可夫决策过程的巡飞弹突防控制决策模型

巡飞弹的作战运用方式为,当其收到控制平台发出的敌目标信息后,会绕过威胁区域,选择高效飞行搜寻路线,对固定目标实施打击。本节将巡飞弹机动突防建模为马尔可夫决策过程(MDP),建立巡飞弹飞行运动模型,设计巡飞弹状态空间、动作空间、奖励函数。MDP可由元组(S,A,P,R,γ)描述,S表示有限状态集;A表示有限动作集;$P=P(s_{t+1}\mid s_t,a)$表示状态s_t下,采取动作a后,转移到下一状态s_{t+1}的概率,t为仿真时间;巡飞弹在与环境交互过程中,在每个时间步长内,根据状态s_t执行动作a,通过与环境交互,生成下一时间步长的状态s_{t+1};$R(s,a)$表示状态s下采取动作a获得的累积奖励,$r(s,a)$表示状态s下采取动作a获得的即时奖励;γ为折扣因子,用来计算累积奖励E. 定义状态值函数$v_\pi(s)$和状态-行为值函数$q_\pi(s,a)$分别如(1)式和(2)式。

$$v_\pi(s)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\mid S_t=s\right], \tag{1}$$

式中:k为仿真时间间隔;$v_\pi(s)$能够衡量策略π下状态s有多好。相应地,状态-行为值函数定义为

$$q_\pi(s,a)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\mid S_t=s,a_t=a\right]. \tag{2}$$

由上述可以看出,$q_\pi(s,a)$衡量的是采用策略π时,在状态s下采取动作a有多好。

1.1 巡飞弹飞行运动模型

巡飞弹的空间质心运动采用3自由度质点运动模型[9-10],假设巡飞弹发动机推力和速度方向一致,采用北东地大地坐标系,建立巡飞弹质点动力学运动模型$f_m(t)$如(3)式所示,系统转移概率$P(\cdot\mid s,a)=1$.

$$f_m(t)=\begin{cases}v_x(t)=v(t)\cos\beta(t)\cos\varphi(t),\\ v_y(t)=v(t)\cos\beta(t)\sin\varphi(t),\\ v_z(t)=v(t)\sin\beta(t),\\ \dfrac{\mathrm{d}v(t)}{\mathrm{d}t}=g(n_x(t)-\sin\beta(t)),\\ \dfrac{\mathrm{d}\gamma(t)}{\mathrm{d}t}=\dfrac{g}{v(t)}(n_z(t)\cos\phi(t)-\cos\beta(t)),\\ \dfrac{\mathrm{d}\varphi(t)}{\mathrm{d}t}=\dfrac{gn_z(t)\sin\phi(t)}{v(t)\cos\beta(t)},\end{cases} \tag{3}$$

式中:x、y、z表示大地坐标系下坐标分量;v表示速度矢量;v_x、v_y、v_z分别表示巡飞弹在x轴、y轴、z轴3个方向的分量速度;g表示重力加速度;β、φ、ϕ分别表示航迹倾角、航向角、滚转角;n_x、n_z分别表示巡飞弹切向过载和法向过载。

假设巡飞弹在 Oxy 平面以固定速度 v 高速突防，则控制巡飞弹航迹倾角 $\beta=0°$，滚转角 $\phi=0°$，运动模型简化为

$$f_{\mathrm{m}}(t)=\begin{cases}v_x(t)=v(t)\cos\varphi(t),\\ v_y(t)=v(t)\sin\varphi(t).\end{cases}\tag{4}$$

巡飞弹飞行轨迹受最小航迹段长度 l_{AB}、最小转弯半径 $R_{\min}$ 的限制如图 1 所示，最小航迹段长度为飞行器在开始改变飞行姿态前必须保持直飞的最短距离，图 1 中：$\varphi_{\max}$ 为最小转弯半径 $R_{\min}$ 对应最大转弯角[4]；弧 $\overset{\frown}{ABCD}$ 为巡飞弹的飞行轨迹，A、B、C、D 分别为巡飞弹的起点位置、转弯开始位置、转弯结束位置、终点位置。

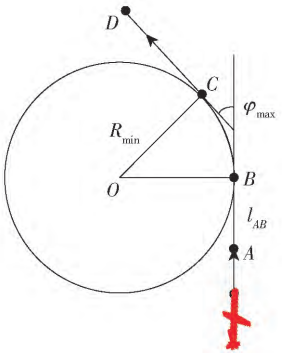


图 1　巡飞弹飞行航迹示意图

Fig. 1　Schematic diagram of flight path of loitering munition

1.2　状态空间设计

针对巡飞弹突防场景，将各作战平台视为质点，巡飞弹的质心位置 $o(t)$、与目标区域中心位置的距离 $l(t)$、航向角 $\varphi(t)$ 作为状态空间，定义 $s(t)=[o(t),l(t),\varphi(t)]$，其中，$o(t)=[x(t),y(t)]$ 分别为在地理坐标系下，巡飞弹在 t 时刻的纬度、经度，$x(t)\in\left[-\frac{\pi}{2}\ \mathrm{rad},\frac{\pi}{2}\ \mathrm{rad}\right]$，$y(t)\in[-\pi\ \mathrm{rad},\pi\ \mathrm{rad}]$，$l(t)$ 计算如（5）式[11]：

$$l(t)=2\arcsin\sqrt{\sin^2\left[\frac{\alpha(t)}{2}\right]+\cos[x(t)]\times\cos(y(t))\times\sin^2\left[\frac{\beta(t)}{2}\right]}\times 6\,371\times1\,000,\tag{5}$$

式中：$\alpha=x(t)-x_{\mathrm{g}}$，$\beta=y(t)-y_{\mathrm{g}}$；$x_{\mathrm{g}}$、$y_{\mathrm{g}}$ 分别为目标区域中心点的经度、纬度坐标。

1.3　动作空间设计

根据巡飞弹飞行运动模型控制量的定义，飞行动作空间定义如（6）式所示。

$$A_{\mathrm{f}}=\{\Delta\varphi\},\Delta\varphi=\varphi(t)-\varphi(t-1),\quad-\varphi_{\max}<\Delta\varphi<\varphi_{\max},\tag{6}$$

式中：$\Delta\varphi$ 表示两个相邻仿真时间步长间航向角的改变量。设置巡飞弹作战条令与交战规则如图 2 所示，主要为巡飞弹可接战临机出现目标，武器控制状态为对地自由开火，即发现目标即摧毁，开火动作不受算法控制。

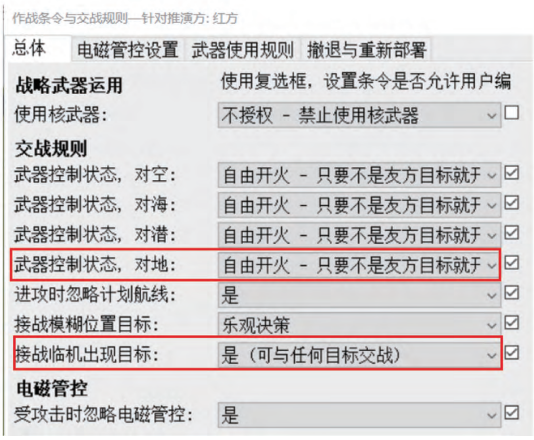


图 2　巡飞弹作战条令与交战规则设置

Fig. 2　Doctrine and engagement rules of loitering munition

1.4　奖励函数设计

巡飞弹的突防目的是机动到目标地域执行任务，设巡飞弹完成突防控制任务的条件，如（7）式所示。

$$\lim_{t\to\max t}d(t)\leqslant l,\tag{7}$$

式中：在巡飞弹初始发射时刻，$t=0$ s，t 为离散值，以 1 s 为 1 个仿真时间步长；max$t$ 为每轮训练最大仿真时间；$d(t)$ 表示 t 时刻，巡飞弹与目标区域中心位置 A_{T} 的距离；l 表示巡飞弹的探测半径。目标区域的范围是以目标点 A_{T} 为圆心，以 l 为半径的圆形区域，如图 3 所示。根据巡飞弹突防控制任务完成的条件，设计巡飞弹突防控制评价函数，如（8）式所示。

$$\mathrm{reward}(t)=\begin{cases}\dfrac{\cos\varphi_{\mathrm{r}}(t)}{d(t)}=\dfrac{\cos[\varphi_{\mathrm{t}}(t)-\varphi(t)]}{d(t)},\ d(t)\neq0\ \mathrm{km};\\ -\dfrac{1}{t},\text{如果巡飞弹进入威胁区域超出探索边界，}\\ \quad t\neq0\ \mathrm{s}.\end{cases}\tag{8}$$

式中：$\varphi_{\mathrm{t}}(t)$ 表示 t 时刻巡飞弹质心位置 $o(t)$ 到 A_{T} 的任务方向 $\overrightarrow{o(t)A_{\mathrm{T}}}$ 与正北方向的夹角；$\varphi_{\mathrm{r}}(t)$ 表示 t 时刻巡飞弹航向与任务方向 $\varphi_{\mathrm{t}}(t)$ 的偏差。当 $\varphi_{\mathrm{r}}(t)$ 控制在 $\left(-\frac{\pi}{2}\ \mathrm{rad},\frac{\pi}{2}\ \mathrm{rad}\right)$ 时，巡飞弹将朝着目标区域方向机动，且 $\varphi_{\mathrm{r}}(t)$、$d(t)$ 越趋近于 0 rad、

0 km，越容易到达目标区，奖励值越高；为减小巡飞弹没有价值的探索空间，提高学习效率，划定巡飞弹突防区域边界，对超过边界的巡飞弹施加持续惩罚 $-\frac{1}{t}$；巡飞弹在机动至任务区域过程中，需要探索地空导弹防御区漏洞，因此，对进入威胁区域的巡飞弹施加持续惩罚 $-\frac{1}{t}$. 除上述巡飞弹突防控制评价函数，需要定义巡飞弹突防任务成功或失败时的回报函数，如（9）式所示。

$$\text{reward}(t)=\begin{cases}100, & \text{到达目标区域，执行任务成功；}\\ -100, & \text{被地空导弹打掉，执行任务失败。}\end{cases} \tag{9}$$

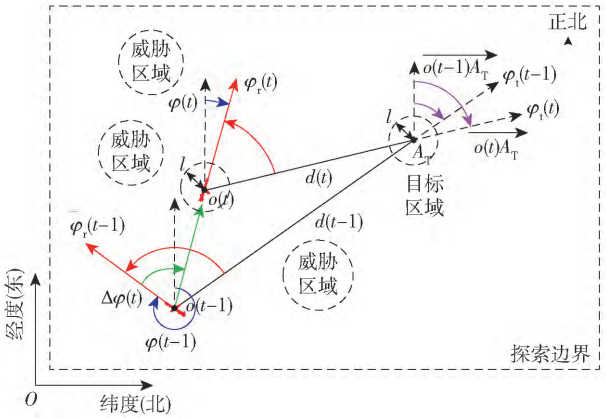


图 3 巡飞弹突防场景几何关系示意图

Fig. 3 Schematic diagram of geometric relationship of loitering munition penetration scene

2 基于深度强化学习的巡飞弹突防控制决策模型求解

强化学习是在给定的 MDP 中寻找最优策略 $\pi^*(a|s)=P(a_t=a|s_t=s)$ 的过程。DRL 主要是在给出状态 s 和 $q_\pi(s,a)$，或 s 和 $v_\pi(s)$ 的值后，可以借助深度神经网络（DNN）较强的拟合能力，通过模型实现 $s\to q_\pi(s,a)$ 或 $s\to v_\pi(s,a)$ 的映射关系。

2.1 基于演员-评论家的巡飞弹突防决策框架

DRL 基本可分为基于策略梯度（PG）与基于值函数两类，基于 PG 的 DRL 够直接优化策略的期望总奖励值并在策略空间搜索最优策略，适用范围更广[12-13]，因此，本节基于 PG 设计算法框架。

为实现巡飞弹突防端到端的感知与决策控制，利用 DNN 技术设计巡飞弹策略网络，如图 4 所示，神经网络的参数集合为 θ，输入层为巡飞弹状态空间，输出层为巡飞弹的策略 π，即取值范围在 $[-1,1]$ 区间的连续值 x，并将 x 实时映射为巡飞弹动作 a，即 $x\to\Delta\varphi$，$\Delta\varphi\in\left(-\frac{\pi}{2}\text{ rad},\frac{\pi}{2}\text{ rad}\right)$。

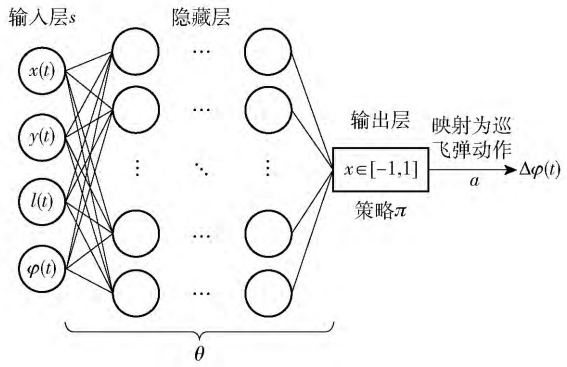


图 4 巡飞弹决策网络结构

Fig. 4 Network structure of loitering munition penetration decision

巡飞弹在战场环境中的状态、动作、奖励值探索轨迹 τ 可描述为

$$\tau=\{s_1,a_1,r_1,s_2,a_2,r_2,\cdots,s_t,a_t,r_t,s_{t+1},a_{t+1},r_{t+1},\cdots,s_T,a_T,r_T\},$$

式中：s_t、a_t、t_t 分别为仿真时间，巡飞弹的状态、动作、奖励值；$t=1,2,3,\cdots,T$，T 为仿真终止时间。

如图 5 所示，τ 发生的概率为

$$p_\theta(\tau)=p(s_1)\,p_\theta(a_1|s_1)\,p(s_2|s_1,a_1)\cdot p_\theta(a_2|s_2)\,p(s_3|s_2,a_2)\cdots=p(s_1)\prod_{t=1}^{T}p_\theta(a_t|s_t)\,p(s_{t+1}|s_t,a_t), \tag{10}$$

因此，在巡飞弹的突防策略为 π 情况下，所能获得的期望奖励为

$$\bar{R}_\theta=\sum_\tau R(\tau)\,p_\theta(\tau)=E_{\tau\sim p_\theta(\tau)}[R(\tau)], \tag{11}$$

式中：$R(\tau)$ 为序列 τ 在 t 时刻所得到的奖励和，$R(\tau)=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'}$，$t'$ 表示仿真时间，$t'>t$.

本节期望通过调整巡飞弹的突防策略 π，使得期望奖励最大，于是对期望函数使用梯度提升方法更新巡飞弹策略网络参数 θ，求解过程如下：

$$\nabla\bar{R}_\theta=\sum_\tau R(\tau)\,\nabla p_\theta(\tau)=\sum_\tau R(\tau)\,p_\theta(\tau)\frac{\nabla p_\theta(\tau)}{p_\theta(\tau)}=E_{\tau\sim p_\theta(\tau)}[R(\tau)\,\nabla\lg p_\theta(\tau)]\approx\frac{1}{N}\sum_{n=1}^{N}R(\tau^n)\,\nabla\lg p_\theta(\tau^n)=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}R(\tau^n)\,\nabla\lg p_\theta(a_t^n|s_t^n)=$$

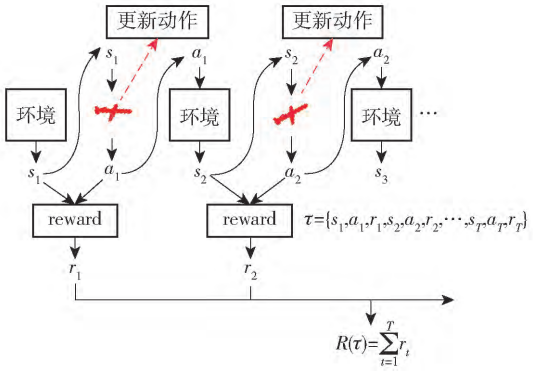


图 5　巡飞弹探索轨迹示意图

Fig. 5　Schematic diagram of loitering munition exploration trajectory

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}\right)\nabla\lg p_\theta(a_t^n\mid s_t^n)\ ,\quad(12)$$

式中：N 表示仿真的最大经验序列数；T_n 表示第 n 经验序列的仿真终止时间。

利用该梯度调整策略参数 θ，如（13）式：

$$\theta\leftarrow\theta+\eta\ \nabla\overline{R}_\theta\ ,\quad(13)$$

式中：η 为学习率。

由于（12）式中 $\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}$ 采样的不稳定，所以考虑用 $\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}$ 的期望值 $E\left(\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}\right)$ 来代替，而 Q-learning 算法中，$Q^{\pi_\theta}(s_t^n,a_t^n)$ 的定义即为在某状态 s，采取某动作 a，假设策略为 π_θ 情况下，奖励值的期望，因而有

$$E\left(\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}\right)=Q^{\pi_\theta}(s_t^n,a_t^n)\ .\quad(14)$$

因此，采用 Q 函数来估算 R 的期望值，同时，创建一个评价网络来计算 Q 函数值。为提升巡飞弹突防学习效率，设计巡飞弹评价网络结构如图 6 所示，输入层为 t 时刻巡飞弹状态空间、动作值，输出为 Q 函数值。

此时，巡飞弹策略网络的参数梯度变为

$$\nabla\overline{R}_\theta=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}Q^{\pi_\theta}(s_t^n,a_t^n)\ \nabla\lg p_\theta(a_t^n\mid s_t^n)\ .\quad(15)$$

巡飞弹评价网络根据估计的 Q 值和实际 Q 值的平方误差进行更新，对评价网络来说，其损失值为

$$loss=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(r_t^n+\max_{a_{t+1}^n}Q^{\pi_\theta}(s_{t+1}^n,a_{t+1}^n)-Q^{\pi_\theta}(s_t^n,a_t^n)\right)^2.\quad(16)$$

设计巡飞弹突防控制决策算法框架设计如图 7 所示。

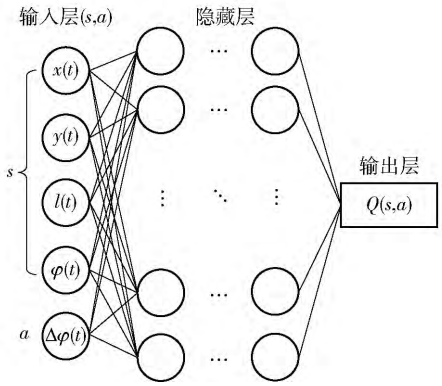


图 6　巡飞弹评价网络结构

Fig. 6　Network structure of loitering munition evaluation

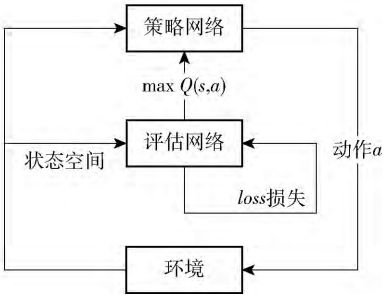


图 7　巡飞弹突防控制决策算法框架

Fig. 7　Algorithm framework of loitering munition penetration control

以上为基于演员-评论家（AC）的 DRL 框架建模，属于 PG 方法类，但可以进行单步更新，比传统 PG 效率更高。

2.2　基于深度确定性策略梯度的巡飞弹突防控制决策求解

深度确定性策略梯度（DDPG）是 AC 框架下的算法[14]，但融合了 DQN 的优势，提高了 AC 的稳定性、收敛性，其流程示意图 8[15] 所示。图 8 中：s'、a' 分别表示更新后的状态值、动作值。

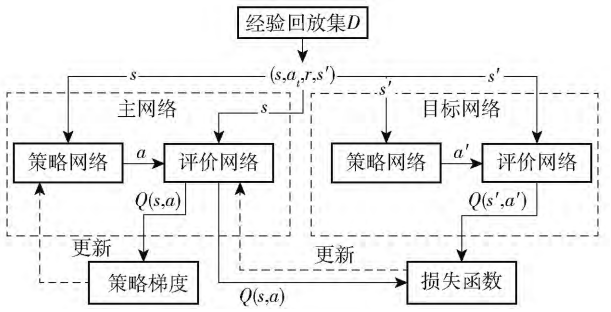


图 8　DDPG 算法流程图

Fig. 8　Flow chart of DDPG algorithm

根据上述流程，基于 DDPG 的巡飞弹突防控制　　决策算法训练流程如表 1 所示。

表 1　巡飞弹突防控制决策算法训练流程

Tab. 1　Training process of loitering munition penetration control algorithm

巡飞弹突防控制决策算法
初始化经验回放集 D
初始化评价网络 $Q(s,a\mid\theta^Q)$，策略网络 $\mu(s\mid\theta^\mu)$ 的权重 θ^Q、θ^μ，μ 为当前策略
用 θ^Q、θ^μ 初始化目标网络 Q'，μ' 的权重 $\theta^{Q'}\leftarrow\theta^Q$，$\theta^{\mu'}\leftarrow\theta^\mu$
For　$N_e=1$，…，M do：M 为仿真最大回合数
为巡飞弹动作探索初始化一个随机噪声分布 ε
接收巡飞弹初始飞行状态 s_1
For $t=1$，T do：T 为仿真终止时间
根据当前策略和探索噪声选择动作，$a_t=\mu(s_t\mid\theta^\mu)+\varepsilon_t$　　(17)
巡飞弹执行动作 a_t，与仿真环境交互获得即时奖励值 r_t，并得到下一时刻巡飞弹飞行状态 s_{t+1}
在 D 中存储经验序列 (s_t,a_t,r_t,s_{t+1})
//训练部分
从 D 中随机采样得到 N 个经验序列 (s_t,a_t,r_t,s_{t+1})
设置 $y_i=r_i+\gamma Q'(s_{i+1},\mu'(s_{i+1}\mid\theta^{\mu'})\mid\theta^{Q'})$　　(18)
式中：i 表示第个 i 经验序列
根据评价网络的损失值更新评价网络：$L=\frac{1}{N}\sum_i(y_i-Q(S_i,a_i\mid\theta^Q)^2)$　　(19)
根据巡飞弹采样的梯度更新策略网络：$\nabla_{\theta^\mu}\mu\mid_{s_i}\approx\frac{1}{N}\sum_i\nabla_aQ(s,a\mid\theta^Q)\mid_{s=s_i,a=\mu(s_i)}\nabla_{\theta^\mu}\mu(s\mid\theta^\mu)\mid_{s_i}$　　(20)
更新目标网络： $\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'}$ $\theta^{\mu'}\leftarrow\tau\theta^\mu+(1-\tau)\theta^{\mu'}$　　(21)

关于巡飞弹策略网络部分，参数更新会涉及到评价网络，(19) 式是关巡飞弹策略网络参数的更新：$\nabla_aQ(s,a\mid\theta^Q)\mid_{s=s_i,a=\mu(s_i)}$ 来自评价网络，表示巡飞弹采取什么动作能获得更大的 Q；$\nabla_{\theta^\mu}\mu(s\mid\theta^\mu)\mid_{s_i}$ 来自策略网络，表示巡飞弹要怎么样修改自身参数更有可能做这个动作。所以，$\nabla_aQ(s,a\mid\theta^Q)\mid_{s=s_i,a=\mu(s_i)}\nabla_{\theta^\mu}\mu(s\mid\theta^\mu)\mid_{s_i}$ 表示巡飞弹要朝着更有可能增大 Q 值的方向修改动作参数。

巡飞弹突防控制决策算法流程训练完毕后，得到最优决策网络 $\mu(s\mid\theta^\mu)$，直接使用 $\mu(s\mid\theta^\mu)$ 输出作为决策结果，即 $a=\mu(s\mid\theta^\mu)$，$s\in S$.

3　实验设计及结果分析

图 9 所示为巡飞弹突防敌地空导弹防御阵地，到某地域实施“斩首”行动仿真实验。

3.1　实验场景及武器性能参数设置

实验场景主要对巡飞弹及 3 个地空导弹阵地的初始位置，以及与巡飞弹突防相关的红方和蓝方主要武器性能参数进行了设置。由表 2 可知：地空导弹的火力射程为 6.0 ~ 7.6 km，巡飞弹的飞行高度为3.658 km，当巡飞弹进入地空导弹火力范围时，即进入威胁区域；巡飞弹的侦察距离为 10 km，地空导弹的火力范围为 10 km，当巡飞弹距地空导弹阵地发射点 10 km 时，会相互探测到对方的位置坐标。导弹的爬升速度为 323 m/s，爬升至巡飞弹的飞行高度需要约 11.3 s 时间，此时，巡飞弹以 250 km/h 速

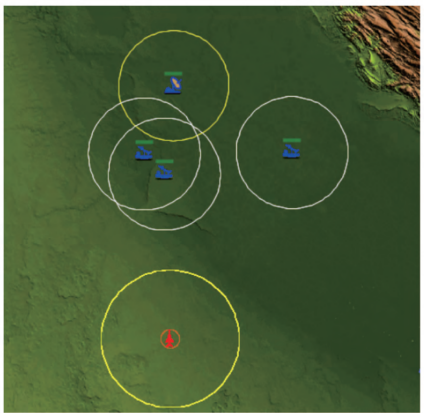


图 9　巡飞弹突防想定示意图

Fig. 9　Schematic diagram of loitering munition penetration scenario

度可机动约 785 m. 由于导弹的巡航速度为 2 185 km/h，远大于巡飞弹的机动速度，因此，在导弹爬升至巡飞弹飞行高度前，巡飞弹如果没有规避到地空导弹阵地火力范围以外，就会面临被摧毁的危险；目标区域设置为：以目标点坐标为圆心，巡飞弹侦察距离为半径圆形区域，是因为这里假定巡飞弹进入该区域，即可在一定探测时间发现目标，并自动锁定将其摧毁。

表 2　实验场景及主要武器性能参数设置表

Tab. 2　Experimental scene and weapon performance parameter setting

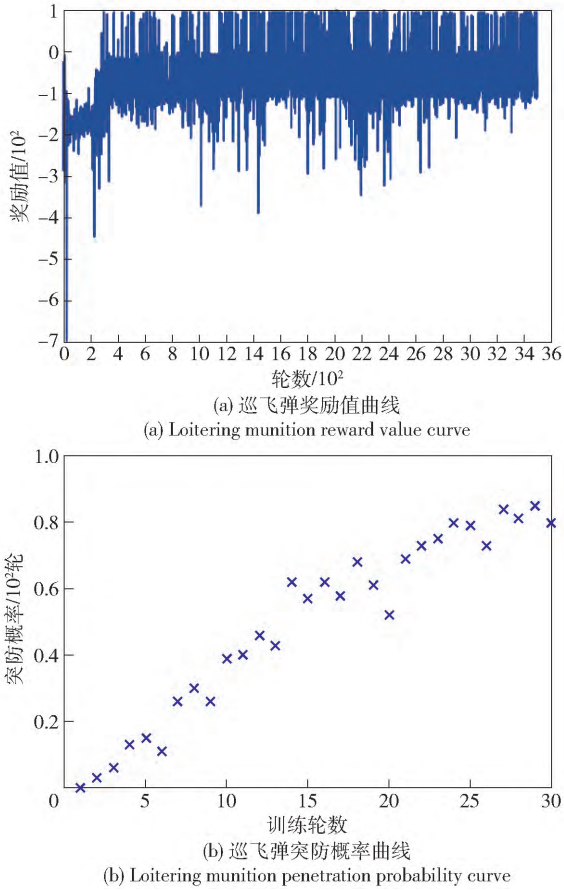
对抗双方	初始位置设定	性能参数	属性值
红方巡飞弹	初始位：以 N43.08° E31.05° 为中心，10 km 内区域随机初始化	飞行高度/km	3.658
		飞行速度/(km·h^{-1})	250
		侦察距离/km	10
蓝方地空导弹阵地	1 号位：以 N43.43° E32.39° 为中心，2 000 m 范围内随机初始化	探测范围/km	10
		火力射程/km	6.0～7.6
	2 号位：以 N44.04° E32.29° 为中心，2 000 m 范围内随机初始化	火力范围/km	10
		巡航速度/(km·h^{-1})	2 185
	3 号位：以 N44.27° E32.39° 为中心，2 000 m 范围内随机初始化	爬升速度/(m·s^{-1})	323
蓝方目标区域	以目标(N44.05°，E32.35°) 为圆心，10 km 为半径圆形区域		

3.2　仿真流程及参数设置

实验软件环境：ubuntu18.04 + pytorch. 硬件环境：Intel core i7 + GeForce GTX 1060Ti + 64G. actor、critic 神经网络结构分别采用 2 层、3 层隐藏层的全连接神经网络，隐藏单元数分别为(256，128)、(256，128，64)，并使用 relu 激活函数。网络主要超参数设置：actor、critic 网络学习率 $\eta=0.001$，折扣因子 $\Gamma=0.99$，目标网络更新系数 $\tau=0.001$，经验回放池容量 $D=100\,000$，当经验回放池数据达到 *scale* = 10 000 规模时，开始采用更新策略网络，采样数据规模 *batchsize* = 1 000，探索噪声 $\varepsilon=0.2$.

3.3　实验结果分析

图 10(a) 和图 10(b) 所示分别为巡飞弹奖励值曲线图、巡飞弹突防概率曲线图。统计每训练 100 轮的巡飞弹突防次数，可以看出，随着训练轮数的增加，奖励值和突防概率呈逐渐上升趋势。这是巡飞弹根据(13) 式不断调整策略参数的结果，即巡飞弹与环境交互采集数据，然后基于 PG 更新参数，随后根据更新后的策略再次采集数据、更新参数，如此循环进行。巡飞弹策略网络参数梯度变化(15) 式中，$\sum_{t=1}^{T_n}\nabla\lg p_\theta(a_t^n\mid s_t^n)$ 梯度项表示能够提高轨迹 τ 出现概率的方向，乘上 R 之后，在模型求解过程中会使概率密度向总奖励值更高的轨迹方向移动，最大化高奖励轨迹 τ 出现的概率。由于巡飞弹总奖励值与突防概率呈正相关性，因此，突防概率也会随奖励值升高而提升。为了探索更多的突防方案，将噪声值设置为 $\varepsilon=0.2$，即巡飞弹每训练 100 轮，会有 80 轮按照当前的突防策略执行，20 轮随机探索。由于噪声值设置相对较大，因此图 10(a) 中



(a) 巡飞弹奖励值曲线

(a) Loitering munition reward value curve

(b) 巡飞弹突防概率曲线

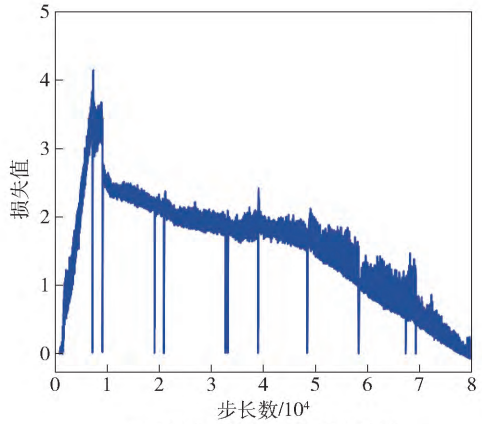
(b) Loitering munition penetration probability curve

图 10　训练数据统计图

Fig. 10　Statistical graph of training data

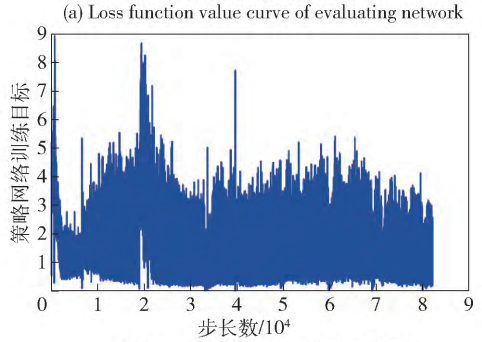
奖励值曲线波动较大。由于噪声值设置为 $\varepsilon = 0.2$，因此突防概率会收敛于 80% 上下。

图 11(a) 为巡飞弹评价网络损失函数值曲线，由评价网络损失值函数(16)式可知：横坐标为训练周期；纵坐标为目标评价网络与主评价网络对巡飞弹状态-动作值的估计在每个训练周期内的累积偏差，即损失值。本文以 1 s 为仿真时间步长，巡飞弹在每个时间步长内与环境交互采集一次数据，当经验回放池数据量达到规模 $scale = 10\ 000$ 之后，每 $batchsize = 1\ 000$ 条经验数据根据(16)式计算一次损失函数值，从图 11(a) 中可以看出，评价网络的损失值随训练进行不断减小，并趋近于 0，这说明评价网络对巡飞弹状态-动作的估计值趋于准确。图 11(b) 为巡飞弹策略网络训练目标变化图，横坐标为训练周期，纵坐标为策略网络在每次训练时目标，巡飞弹根据(21)式更新训练目标网络。从图 11(b) 中可以看出，策略网络训练目标随训练进行，逐渐维持在一个较小的值，说明巡飞弹突防控制策略在逐步优化并趋于稳定。



(a) 评价网络损失函数值曲线图

(a) Loss function value curve of evaluating network



(b) 策略网络训练目标变化曲线图

(b) Changing curve of training target of policy network

图 11 巡飞弹突防控制决策模型最优策略求解过程

Fig. 11 Process of solving the optimal policy of loitering munition penetration control decision model

统计巡飞弹每训练 M 轮的平均奖励值，即

$$\overline{R}_{N_e,N_e+M-1} = \frac{1}{M}\sum_{N_e}^{N_e+M-1} R_{N_e}, \tag{21}$$

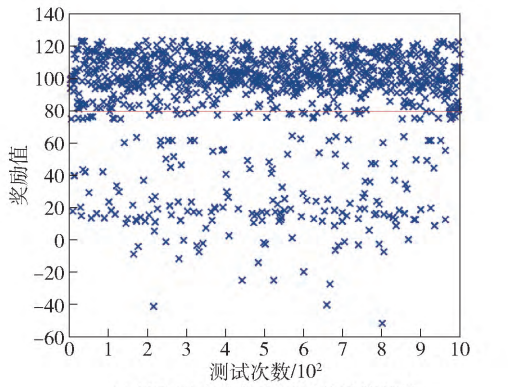
式中：$N_e = 100k + 1$，$k = 0,1,2,\cdots$。由于巡飞弹到达以 A_T 圆心的目标区域，会获得一个远大于其他奖励的奖励值加 100，因此，把 $\overline{R}_{N_e,N_e+M-1} > R_\varepsilon$ 作为收敛条件，R_ε 为正实数阈值，取 $R_\varepsilon = 80$，$M = 100$。如表 3 所示，由于 $\overline{R}_{3\,401,3\,500} = 85.2 > R_\varepsilon$，因此，在 $N_e = 3\ 500$ 轮将算法终止。

表 3 巡飞弹突防平均奖励值统计

Tab. 3 Average reward values of loitering munition penetration

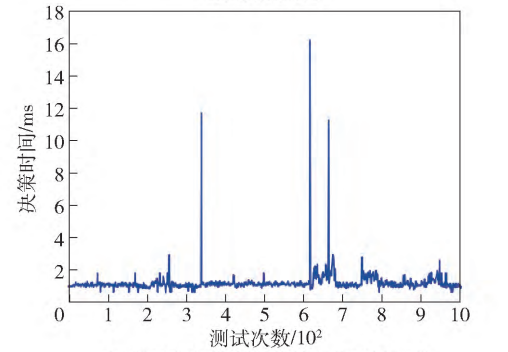
训练轮数				
3 001 ~ 3 100	3 101 ~ 3 200	3 201 ~ 3 300	3 301 ~ 3 400	3 401 ~ 3 500
50.3	70.2	68.4	84.6	85.2

训练完成后，取 $N_e = 3\ 500$ 的巡飞弹策略模型 $\pi_{3\,500}$ 进行 1 000 次突防仿真测试，数据统计结果如图 12 所示。



(a) 巡飞弹突防仿真测试奖励值统计

(a) Reward value statistics of loitering munition penetration simulation test



(b) 巡飞弹突防仿真测试决策时间统计

(b) Decision-making time of loitering munition penetration simulation test

图 12 巡飞弹突防仿真测试数据统计

Fig. 12 Data statistics of penetration simulation test for loitering munition

巡飞弹决策控制模型测试统计结果如表 4 所

示，1 000 次突防仿真测试实验，共成功突防 821 次，成功率为 82.1%，平均决策时间 1.48 ms，满足巡飞弹控制决策指标要求。

表 4　决策控制模型测试统计结果

Tab. 4　Statistical results of decision control model test

测试次数	成功次数	成功率/%	平均决策时间/ms
1 000	821	82.1	1.48

从 1 000 次突防仿真测试实验中，选择 3 组具有代表性的巡飞弹突防轨迹样例，如图 13 所示。巡飞弹的初始位置在图 13 中绿色圆形区域内随机初始化，进而反应训练结果在该发射区域的泛化性能。目标区域为图 13 中橙色圆形区域，巡飞弹进入该区域成功摧毁目标，即为成功完成突防任务。图 13 中蓝色区域为地空导弹威胁区域，巡飞弹实施突防任务时需要即时调整突防路线，避开威胁区域。从图 13 中可以看出有红、绿、蓝 3 条不同颜色的巡飞弹突防轨迹，分别记为 1 号、2 号、3 号突防路线。

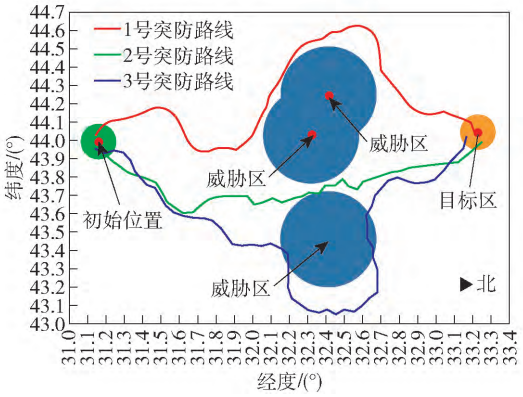


图 13　巡飞弹突防仿真测试轨迹样例

Fig. 13　Sample trajectories of loitering munition in penetration simulation test

图 14 为巡飞弹突防仿真测试奖励值曲线，从图 14 中3 条突防路线的即时奖励值始终大于 0 可知，巡飞弹航向 $\varphi(t)$ 与任务方向 $\varphi_t(t)$ 的偏差始终在 $\left(-\frac{\pi}{2}\text{ rad},\frac{\pi}{2}\text{ rad}\right)$ 内，即巡飞弹始终朝着目标区域方向机动。另外，3 条突防路线在最后仿真时间内，即时奖励值骤增，结合图 13 可知，这是巡飞弹与目标点距离 $d(t)\to 0$ km，飞弹航向 $\varphi(t)$ 与任务方向 $\varphi_t(t)$ 的偏差 $|\varphi_t(t)-\varphi(t)|\to 0$ rad，$\cos[|\varphi_t(t)-\varphi(t)|]\to 1$ 导致，即巡飞弹在突破防空威胁区之后，会即时调整航向，始终与任务方向保持大致一致，向目标区域机动。

图 15 为巡飞弹动作控制参数变化曲线，结合

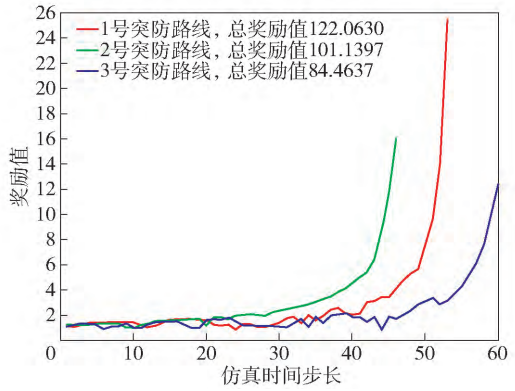


图 14　巡飞弹突防仿真测试奖励值曲线

Fig. 14　Reward curves of loitering munition in penetration simulation test

图 13 可知：在 1 号突防路线中，巡飞弹在突破威胁区之前，$\Delta\varphi>0$ rad，并且 $\Delta\varphi$ 逐渐增大，后逐渐减小，实现向东平稳转向；巡飞弹临近威胁区域，$\Delta\varphi$ 减小至 0 rad，并且随着距离的进一步临近，$\Delta\varphi$ 继续减小，实现向西平稳转向，从而在威胁区西侧边缘绕过；巡飞弹突破威胁区域，$\Delta\varphi$ 逐渐增大至大于 0 rad，实现向东平稳转向之后，始终控制航向与任务方向保持一致，机动至目标区，实现突防。

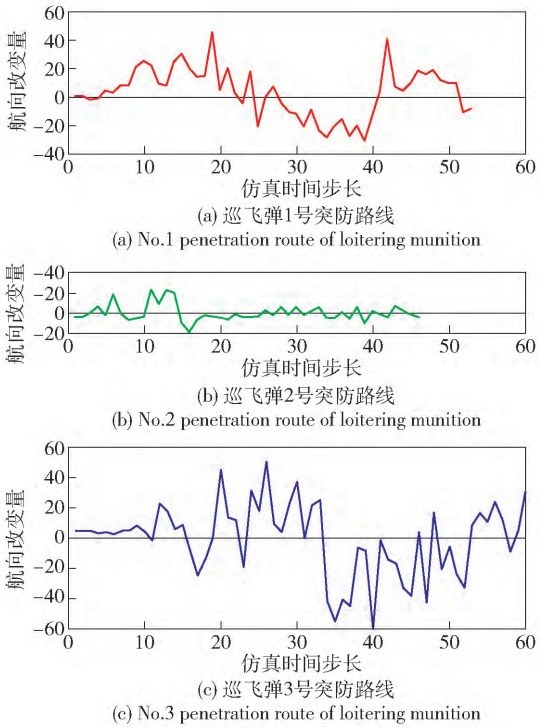


图 15　巡飞弹动作控制参数变化曲线

Fig. 15　Sample diagram of penetration trajectories

在 2 号突防路线中，巡飞弹在突破威胁区前，$\Delta\varphi>0$ rad，进而向东机动至临近威胁区域，随后控

制航向与任务方向保持一致; $\Delta\varphi$ 在没有大的变动情况下,始终朝目标区域方向机动,从防御体系漏洞突破威胁区,实现突防。

在 3 号突防路线中,$\Delta\varphi$ 的变动范围较大,特别是在即将进入威胁区时,$\Delta\varphi > 0$ rad 持续增大,后持续减小至 $\Delta\varphi < 0$ rad,从而在威胁区东侧边缘绕过;在突破威胁区后,又调整 $\Delta\varphi$,向目标区域机动,实现突防。

综上所述,3 组具有代表性的突防仿真样例中,巡飞弹均能从发射区域的任意位置机动至目标区域,并将目标摧毁,决策网络具有较好的泛化能力,奖励值均层指数级增长。由此可以看出,本文所提模型可有效实现巡飞弹突防控制决策,在一定程度上提高了巡飞弹的自主性。

4 结论

本文针对巡飞弹动态突防控制决策问题,采用 MDP 描述了巡飞弹飞行运动模型,设计了飞行状态空间、动作空间、奖励函数等,提出基于 DRL 的 LMPCD 模型及其求解方法。仿真实验结果表明,巡飞弹在动态对抗环境中,能够实现自主突防,证明了模型及求解方法的有效性。该方法可为预测“蓝军”巡飞弹突防路线提供了技术借鉴,以及该方法以实际武器装备可获取的数据为输入,对下一步在真实环境中应用具有重要军事意义。

参考文献(References)

[1] 庞艳珂,韩磊,张民权,等. 攻击型巡飞弹技术现状及发展趋势[J]. 兵工学报,2010,31(增刊 2):149 - 152.
PANG Y K, HAN L, ZHANG M Q, et al. Status and development trends of loitering attack missiles [J]. Acta Armamentarii, 2010, 31(S2): 149 - 152. (in Chinese)

[2] 郭美芳,范宁军,袁志华. 巡飞弹战场运用策略[J]. 兵工学报,2006,27(5):944 - 947.
GUO M F, FAN N J, YUAN Z H. Battlefield operational strategy of loitering munition [J]. Acta Armamentarii, 2006, 27(5): 944 - 947. (in Chinese)

[3] 刘杨,王华,王昊宇. 巡飞弹发展背后的作战理论与概念支撑[J]. 飞航导弹,2018(10):51 - 55.
LIU Y, WANG H, WANG H Y. Operational theory and conceptual support behind the development of loitering munition [J]. Aerodynamic Missile Journal, 2018(10): 51 - 55. (in Chinese)

[4] 郝峰,张栋,唐硕,等. 基于改进 RRT 算法的巡飞弹快速航迹规划方法[J]. 飞行力学,2019,37(3):58 - 63.
HAO F, ZHANG D, TANG S, et al. A rapid route planning method of loitering munitions based on improved RRT algorithm [J]. Flight Mechanics, 2019, 37(3): 58 - 63. (in Chinese)

[5] 欧继洲,黄波. 巡飞弹在陆上无人作战体系中的应用初探[J]. 飞航导弹,2019(5):20 - 24.
OU J Z, HUANG B. Application of loitering munition in land unmanned combat system [J]. Aerodynamic Missile Journal, 2019(5): 20 - 24. (in Chinese)

[6] 王琼,刘美万,任伟建,等. 无人机航迹规划常用算法综述[J]. 吉林大学学报(信息科学版),2019,37(1):58 - 67.
WANG Q, LIU M W, REN W J, et al. Overview of common algorithms for UAV path planning [J]. Journal of Jilin University (Information Science Edition), 2019, 37(1): 58 - 67. (in Chinese)

[7] 张堃,李珂,时昊天,等. 基于深度强化学习的 UAV 航路自主引导机动控制决策算法[J]. 系统工程与电子技术,2020,42(7):1567 - 1574.
ZHANG K, LI K, SHI H T, et al. Autonomous guidance maneuver control and decision-making algorithm based on deep reinforcement learning UAV route [J]. Journal of Systems Engineering and Electronics, 2020, 42(7): 1567 - 1574. (in Chinese)

[8] Bouhamed O, Ghazzai H, Besbes H, et al. Autonomous UAV navigation: a DDPG-based deep reinforcement learning approach [EB/OL]. [2020-07-11]. http://arxiv.org/pdf/1509.02971.pdf.

[9] 张建生. 国外巡飞弹发展概述[J]. 飞航导弹,2015(6):19 - 26.
ZHANG J S. Overview of foreign cruise missile development [J]. Aerodynamic Missile Journal, 2015(6): 19 - 26. (in Chinese)

[10] 李增彦,李小民,刘秋生. 风场环境下的巡飞弹航迹跟踪运动补偿算法[J]. 兵工学报,2016,37(12):2377 - 2384.
LI Z Y, LI X M, LIU Q S. Trajectory tracking algorithm for motion compensation of loitering munition under wind environment [J]. Acta Armamentarii, 2016, 37(12): 2377 - 2384. (in Chinese)

[11] 黎珍惜,黎家勋. 基于经纬度快速计算两点间距离及测量误差[J]. 测绘与空间地理信息,2013,36(11):235 - 237.
LI Z X, LI J X. Quickly calculate the distance between two points and measurement error based on latitude and longitude [J]. Geomatics & Spatial Information Technology, 2013, 36(11): 235 - 237.

[12] 刘建伟,高峰,罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报,2019,42(6):1406 - 1438.
LIU J W, GAO F, LUO X L. A review of deep reinforcement learning based on value function and strategy gradient [J]. Chinese Journal of Computers, 2019, 42(6): 1406 - 1438. (in Chinese)

[13] 刘全,翟建伟,章宗长. 深度强化学习综述[J]. 计算机学报,2018,41(1):1 - 27.
LIU Q, ZHAI J W, ZHANG Z C. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41(1): 1 - 27. (in Chinese)

[14] KONDA V R, TSITSIKLIS J N. Actor-Critic algorithms [C]// Proceedings of Advances in Neural Information Processing Systems. Denver, CO, US: NIPS Foundation, 2000: 1008 - 1014.

[15] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [EB/OL]. [2020-07-11]. http://arxiv.org/pdf/1509.02971.pdf.