☞指挥控制与通信☜

基于深度强化学习的舰艇空中威胁行为建模*

房霄,曾贲,宋祥祥,贾正轩

(北京电子工程总体研究所,北京 100854)

摘要: 随着武器装备智能化发展的速度加快,传统武器装备的训练方法已经无法满足大规模现代战争的训练需求。在近十年中深度强化学习等人工智能方法在棋类以及电子竞技游戏中取得了极大突破,证明了人工智能方法在面对大搜索空间博弈问题的优势,能够有效解决军事对抗问题中的形势预判和临机调整问题。基于此背景,依托海军舰艇对空方面作战,开展了深度强化学习的方法研究。首先通过并行场景建模技术以及空中威胁决策行为建模技术实现深度学习模型的构建,之后通过单机突防场景的对抗迭代学习,得到收敛的突防策略。验证了深度强化学习方法在空中威胁行为构建场景的可行性,为后续深入开展编队联合防空训练场景构建提供支撑。

关键词: 深度强化学习; 人工智能; 舰艇防空; 空中威胁; 突防策略; 场景构建

doi: 10.3969/j.issn.1009-086x.2020.05.009

中图分类号: U674.7; E917; N945.12 **文献标志码:** A **文章编号:** 1009-086X(2020)-05-0059-08

Modeling of Air Target Threat to Warship Based on Deep Reinforcement Learning

FANG Xiao, ZENG Bi, SONG Xiang-xiang, JIA Zheng-xuan

(Beijing Institute of Electronic Engineering, Beijing 100854, China)

Abstract: With the development of intelligent weapons, the traditional training methods could not-meet the demands of large-scale modern warfare. In the past decade, artificial intelligence (AI) methods such as deep reinforcement learning have made great breakthroughs in chess and electronic competitive games. It proves that the AI methods have great advantages in solving large searching space problems. Furthermore, the problems of situation prediction and temporary adjustment could be solved more effectively by AI methods. A new method for modeling of air target threat is proposed based on the research of deep reinforcement learning. The parallel scene modeling technology and the air target behavior modeling technology are used to construct the model of deep reinforcement learning. The convergence penetration strategy is calculated with iterative learning under the scene of single airplane. The successful attempt verifies that the practicability of deep reinforcement learning in modeling of air target threat. It provides support for the further research on the modeling of fleet joint air defense.

Key words: deep reinforcement learning; artificial intelligence (AI); warship air defense; air threat; penetration strategy; modeling

* **收稿日期:** 2020-04-24; **修回日期:** 2020-05-18

第一作者简介: 房霄(1986-),男,北京人。高工,硕士,主要从事指挥控制及装备模拟训练技术研究。

通信地址: 100854 北京 142 信箱 30 分箱 **E-mail:** hmjs_0814@126.com

0 引言

在空中、水面、水下等方面作战中，空中威胁是当前海军舰艇作战所面临的最为严重的威胁，快速机动的各式作战飞机以及低空掠海反舰导弹能够为整个作战编队带来毁灭性的打击，因此海军舰艇对空方面作战是海军远海作战的基础和保障。

随着海战场作战样式和武器装备越来越复杂，海军舰艇对空方面作战无论从指挥训练还是对抗操作训练都面临极大挑战。这里面最为核心的就是对空中威胁场景的构建。当前主要的方式是对抗样式预先设计和规划好，在执行过程中不能根据战场条件变化进行实时调整，与实战状态下的空中威胁相差甚远，无法为我海军舰艇防空作战提供足够的训练支撑。

同时近些年智能化的快速发展，AI 智能越来越多的出现在美军的训练过程中，美军模拟训练正向实战化、智能化、体系化发展。其中智能化的训练是其实现实战化训练的重要手段和标志。特别是，智能化不仅为未来战场带来了彻底的颠覆，也为在智能化条件下的部队训练带来巨大挑战。NVIDA 公共事务部副总裁安托尼罗宾就指出“AI 将在创建合成环境，模拟对手军队，创建挑战性想定，等等模拟和训练的各个方面大显身手，能够有效推动战斗人员学习更多新技能”。

本文就是在这样的背景下，利用深度强化学习技术开展了对于海军舰艇空中威胁行为建模的方法研究，构建了攻防对抗场景并进行了单机突防的仿真验证，初步验证了深度强化学习方法在空中威胁行为构建场景的可行性，为后续深入开展编队联合防空训练场景构建提供支撑。

1 舰艇空中威胁场景分析

1.1 空袭体系的隐蔽与突袭战术

空袭体系攻击的隐蔽性包括广泛利用现代空袭兵器的低空飞行性能，利用地形和地物的遮挡绕过预警雷达场，从而使防空方的指控中心、防空导弹和高炮来不及作好战斗准备，防空歼击机来不及起飞迎敌。

为了对抗隐蔽飞行的空袭兵器和高速飞行的空袭武器，客观上要求防空体系指控中心和防空兵器进一步缩短战斗准备时间。

由于现代巡航导弹的隐身性能和应用地形匹配技术，被攻击方在没有准备的情况下发现它们的概率很低，即使个别被发现，甚至被击毁，也不会影响大批巡航导弹飞向拟攻击的目标。

1.2 对防空体系的火力压制战术

现代空袭体系突防的基本战术已不是逃避攻击，而是首先摧毁防空方的防空体系和防空兵器，用硬杀伤的方法夺出制电磁权和制空权。

携带空地反辐射导弹的防空压制飞机，投放空地反辐射导弹有 2 种方法：一是按预先测定的防空雷达所在位置的坐标和辐射电磁波的频率，或者按预警机或侦察机临时测定的参数，从视线外发射空舰反辐导弹进行打击；二是在没有预先给定的防空雷达坐标和辐射频率时，由载机自行探测目标，在防空雷达视线内捕获目标并发射空舰反辐射导弹，由空舰反辐射导弹的导引头自行跟踪目标，载机迅速转弯、降高退出防空雷达的视线。

1.3 大规模高密度攻击战术

对于海上大型舰船编队（如航母编队）和战区中心地带，现代空袭一次出动约 100 ~ 300 架飞机，并配以数百架无人机，再加上从地面、舰艇上发射的巡航导弹，可能达到 2 000 个以上的空袭兵器（含飞机上发射的空舰导弹），一次空袭作战过程只有 10 ~ 15 min。时间短，空中目标多，从而形成多层次的饱和攻击。防空体系面临提高火力强度（单位时间射击目标数）的强烈要求[1]。

2 典型的空中威胁建模方法

2.1 典型方法

对于海军舰艇空中威胁的模拟，一般从 4 个方面考虑，分别为平台运动特征模拟、探测能力模拟、决策能力模拟以及武器能力模拟。

（1）平台运动特征模拟

平台运动特征模拟典型的方法为点迹建航法和六自由度建模法。点迹建航法主要思路为将空袭平台抽象为一个质点，通过构建质点运动约束实现质点的运动模拟，考虑的约束一般包括平台升限、速度、转弯半径等条件。六自由度建模相比点迹建航法，能够更加精细地实现对飞行器运动特征的模拟[2]。

（2）探测能力模拟

探测能力模拟主要模拟机载平台探测雷达威力。分为雷达威力包络模拟以及信号注入模拟等。在航空兵仿真模拟中应用较多，在防空模拟训练中应用较少。

（3）决策能力模拟

决策能力模拟主要模拟作战中的指挥决策过程，往往体现了战役的战术意图以及飞行员或指挥员的战斗意志，在某些场景下往往采用博弈论或者优化算法对指挥决策行为进行建模。比较典型的方法为应用粒子群优化算法寻找最优突防路径以模拟飞行员突防寻优过程。但随着舰艇数量及飞机数量的递增，该优化问题的求解空间将逐渐增大至不可求解，而且极大消耗计算资源，很难适用于计算资源有限的武器装备模拟训练中。

（4）武器能力模拟

武器能力模拟则较为常用，一般也会采用点迹建航法和六自由度建模法进行建模。和飞行器不同的是，大部分打击武器的轨迹具备有较为明显的弹道特征，比如 TBM 类武器，ARM 武器，在进行建模时，会采用弹道拟合法，通过数据模型与实际靶弹数据进行拟合，抽象出弹道拟合公式进行弹道的模拟[2]。

2.2 存在问题

舰艇编队防空作战模拟训练的主要目的是作为实战训练的补充在优化训练成本的基础上实现常态化、实战化的作战训练。然而随着防空武器的信息化程度的快速发展，原有的目标威胁建模方法已经很难满足部队实战化训练需求，主要表现在：

（1）威胁建模方法不足，随着场景的复杂化，已经无法通过常规手段建立可靠、好用的模型；

（2）对抗实战程度不高，在防空训练过程中，假想敌往往由于对蓝军进攻战术战法和兵器的不了解而模拟的比较简单；

（3）对抗过程不够完善，没有考虑敌方进攻条件下的对抗过程，比如对于敌方反辐射导弹攻击下的要地防卫，或者在复杂干扰环境下的电磁对抗等。

3 面向舰艇空中威胁行为建模的智能设计技术

近年来，在大数据、云计算、机器视觉等技术突飞猛进的基础上，人工智能的应用前景得到了空前的发展，并逐步向着自主学习、数据驱动、虚实融合的方向演化，进而逐渐在应对多维度的复杂设计问题上实现了颠覆性的突破，甚至在一些领域上超越了人类，如面向围棋/中国象棋/国际象棋、DOTA2/星际争霸 II 等博弈对抗的系统设计上已经完美超越人类。

特别地，DeepMind 公司推出基于深度强化学习、联盟学习等新一代人工智能技术的 AlphaStar 智能体，在 DOTA2/星际争霸 II 这类博弈对抗游戏中，通过保持资源要素的合理调配、作战单元的临机决策为前提，短期、长期的目标规划，最终以精妙的战术规划、灵巧的进攻方式击败对手。类比到舰艇编队模拟训练场景中，诸如不完备信息条件下的对抗博弈，长远规划策略学习以及大规模交战及决策空间求解等问题已经在 AlphaStar 智能体上有所突破[3]。

因此，本文采用基于深度强化学习的方法，拟突破典型的空中威胁建模方法的约束，验证人工智能技术在军事模拟领域的可行性。

3.1 场景定义

为简化问题求解，本文考虑的作战场景为单机突防单舰防御的场景，在该场景下，单机按特定策略飞行靠近舰艇、飞抵可投弹区域、完成投弹并成功脱离战场。而舰艇则以发现来袭敌机，并对其进行防空打击为作战任务。

在此设定下，所需解决的问题可以抽象为在考虑飞机模型、舰艇模型以及交战条件模型约束下，对单机突防任务的策略进行寻优。

3.1.1 飞机模型

为进一步简化问题求解，本文将飞机模型考虑为质点模型。此外，考虑飞机飞行性能以及投弹能力的限制，本文对于飞机运动及投弹过程采取如下限制：

（1）运动特征模型

飞机采用点迹模拟法，飞机最小转弯半径限制，设为 $R_{\min}$，即任意时刻飞机的转弯半径 R 须满足 $R \geqslant R_{\min}$。飞机飞行高度约束为 $H \in [H_{\max}, H_{\min}]$。飞机加速度约束为单轴加速度 a_x, a_y, a_z 必须满足 $a_x, a_y, a_z \in [a_{\max}, a_{\min}]$，运动坐标系为北天东坐标系。飞行合速度限制在 $v \in [v_{\max}, v_{\min}]$ 范围内。

(2) 投弹能力模型

设定飞机在投弹过程中需沿当前速度方向继续飞行 t s 以保持发射过程稳定,且与舰艇间的夹角 θ 满足 $\theta \in [\theta_{\text{fire}}, \theta_{\min}]$ 方可完成投弹动作。

3.1.2 舰艇模型

对舰艇模型从探测模型、防御模型 2 个方面进行描述。

(1) 探测模型

探测模型主要用于模拟舰艇配备探测制导雷达发现跟踪空中威胁目标的能力。在舰艇北天东坐标系下,考虑雷达探测半径约束,雷达探测范围描述为

$$\begin{cases} x = r\cos\theta, z = r\sin\theta, r = \sqrt{k^2 y - y^2}, \\ y \in [0, R_0^2/k^2], \theta \in [0, 2\pi], \\ x^2 + y^2 + z^2 \leqslant R_0^2, y \in [R_0^2/k^2, R_0^2]. \end{cases} \tag{1}$$

(2) 防御模型

在本文中,舰艇防空打击采用简单策略实现:探测到来袭目标,则舰艇即发射防空导弹对目标进行拦截,拦截导弹预计飞行时间按导弹平均飞行速度以及目标首次被探测到时舰艇与目标距离进行折算,记作 $t_{\text{intercept}}$,该值即为预计防空拦截时间。通过该时间的计算以及交战条件模型中的突防成功条件的比较,实现对目标的防空拦截。

3.1.3 交战条件模型

交战条件模型主要涉及如下几个方面:

(1) 飞机投弹条件

飞机距舰的距离记作 $D_{\text{plane-ship}}$,满足 $D_{\text{plane-ship}} \leqslant D_{\text{fire}}$ 的条件,距海满足 $H \in [H_{\max}, H_{\text{fire}}]$,且需满足飞机速度方向的矢量与飞机和舰艇位置的夹角 $\theta \in [\theta_{\text{fire}}, \theta_{\min}]$ 的条件下方可执行投弹动作。

(2) 飞机突防成功条件

基于舰艇模型中的防御模型,考虑飞机的生存时间以其第一次被雷达检测到的位置与舰艇位置之间的距离除以导弹飞行速度进行近似。飞机被雷达探测到以后,记作 t_{detected}。飞机突防任务成功的条件为完成投弹动作以后,需在满足 $t_{\text{detected}} \leqslant t_{\text{intercept}}$ 的前提下,脱离雷达的探测范围。

3.2 智能体建模方法

采用深度强化学习、联盟学习等新一代智能技术,构建空中威胁智能体自学习的决策模型,并面向并行突防场景,充分生成不同初始状态下的作战场景,让空中威胁智能体并行地对抗不同作战场景下的舰艇,进而认知足够多的对抗样式,从而寻找不同对抗场景下的防御突破点,形成满足各对抗条件下的最优决策集合,建模架构如图 1 所示。

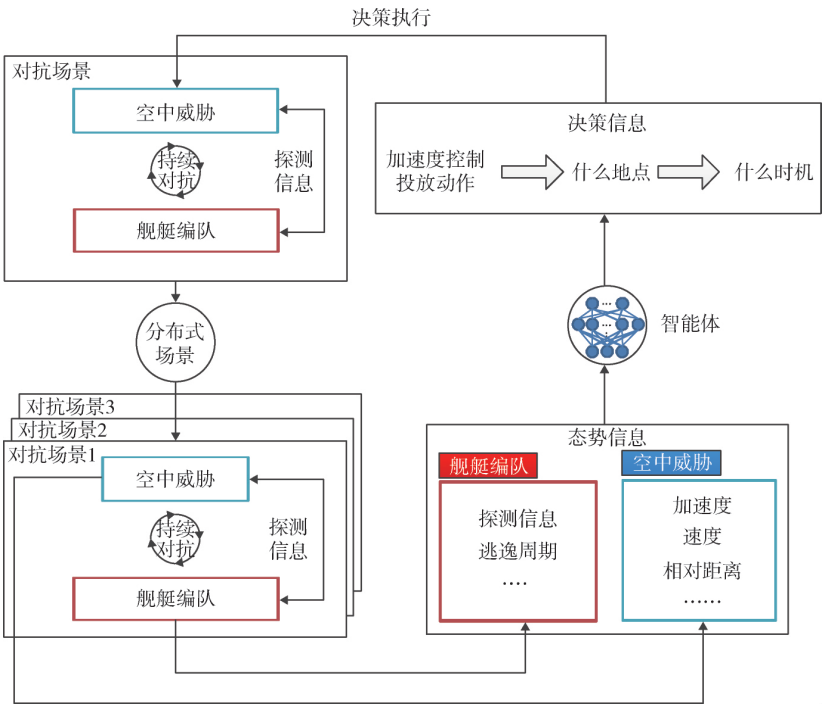

图 1 智能体建模架构

Fig. 1 Frame diagram of the agent modeling

3.2.1 基于空中威胁的临机决策建模技术

本文采用深度强化学习算法完成空中威胁智能体的建模过程，提升空中威胁智能体的决策能力。框架如图 2 所示[5]。智能体通过在环境中不断地探索生成动作、感知状态和获得回报，从大数据中获得复杂因素的关联性和问题处理的完备性，加强其对复杂关联关系的拟合能力。

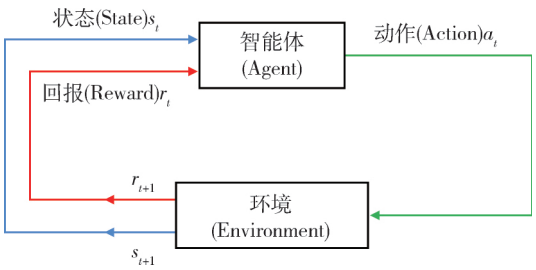

图 2 强化学习框架逻辑图

Fig. 2 Frame diagram of deep reinforcement learning

在本文中，考虑常规强化学习的配置，其中空中威胁智能体会与对抗场景产生互动。在每一个仿真间隔 t，空中威胁智能体都会观测到一组态势信息 $s_t \in S$，分析判断之后，让空中威胁做出一组动作 $a_t \in A$，然后会收到环境反馈的奖励值 $r(s_t, a_t) \in R$，经过一段时间的迭代训练，智能体会形成一个决策集合 $\pi: S \rightarrow A$[6]。

其中，每一个态势信息都对应空中威胁的一组动作。这样的一个态势信息与动作的映射函数反映出一种期望回馈，即依据每次获取到的态势信息 $s_t \in S$，从策略 π 中寻找最优的决策，直至对抗结束所产生的所有累计奖励的值函数。

$$Q_\pi(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right], \tag{2}$$

式中：$\gamma \in [0,1]$ 为衰减因子。

同样地，这个预期回馈也可以评估一个策略 π。因此，可以使用 Q_π 得到一种对 π 的更新方式。目标为使 $J(\theta)$ 最大化。

$$J(\theta) = E\left[Q_\pi(s, \pi_\theta(s))\right]. \tag{3}$$

根据确定型策略梯度算法[7-8]可得策略 π_θ 的参数更新算法为

$$\nabla_\theta J(\theta) \approx E\left[\nabla_\theta \pi_\theta(s) \quad \nabla_a Q_{\pi_\theta}(s,a) \mid_{a=\pi_\theta(s)}\right]. \tag{4}$$

进而规定 π_θ 的更新方向，从而就能确定策略集合 π 的最终形态，既扮演决策执行者的身份[9]，也称之为 actor 网络。同时，为了更好地评价其 π_θ 的演进方向与真实叠加产生的 $Q_\pi(s,a)$ 之间的关系，可以设置一位评价者(critic 网络)[10-12]，通过其观测、评估 actor 的决策质量，校正 actor 的演化方向。使用 Bellman 方程[13]。

$$(\tau_\pi Q)(s,a) = r(s,a) + \lambda E\left[Q_\pi(s', \pi(s'))\right], \tag{5}$$

式中：s' 为下一次的态势信息。

通过最小化 TD 误差[5]的方式，修正值函数与 Bellman 方程推导出来的期望值之间的误差，即二者标准差。

$$L(w) = E\left[\left(Q_w(s,a) - (\tau_{\pi_{\theta'}} Q_{w'})(s,a)\right)^2\right]. \tag{6}$$

依据 Bellman 方程的更新方式，确实能够找到最优解，但事实上这种建模方式不够合理，单纯利用期望值进行迭代，从某种程度上来说损失了 Q_π 作为分布的信息，因此，采用 N-Step 的分布 Bellman 方程[14]。

$$(\tau_\pi^N Z)(s_0,a_0) = r(s_0,a_0) + E\left[\sum_{n=1}^{N-1} \lambda^n r(s_n,a_n) + \lambda^N Z(s_N, \pi(s_N)) \mid s_0, a_0\right], \tag{7}$$

式中：$Z(s,a)$ 表示在状态 s 下执行动作 a 之后回报形成随机变量，具有概率分布的特性，则上述推导出来的更新的方程修改为

$$\begin{cases} L(w) = E\left[d\left((\tau_{\pi_{\theta'}} Z_{w'})(s,a), Z_w(s,a)\right)\right], \\ \nabla_\theta J(\theta) = E\left[\nabla_\theta \pi_\theta(s) E\left[\nabla_a Z_w(s,a)\right] \mid_{a=\pi_\theta(s)}\right] \end{cases}, \tag{8}$$

式中：d 表示分部之间的距离度量，采用交叉熵[15]求取。

3.2.2 基于并行作战场景的分布式建模技术

本文采用 Ring-AllReduce[9] 分布式架构，所有智能体组成单向环形架构，既第 $N-1$ 个智能体的梯度传输给第 N 个智能体，当所有智能体在其负责交互的仿真环境中收敛达到稳定，即可实现分布式训练，如图 3 所示。

3.3 算法流程

根据并行作战场景构建和智能体建模，选取了速度、距离、发射角度、是否被探测等数据作为每时刻获取的态势信息。

$$s_t = \{v_x, v_y, v_z, a_x, a_y, a_z, H, \theta_{\text{fire}}, detected, t_{\text{detected}}, approch, fire, back\}, \tag{9}$$

式中：(v_x, v_y, v_z) 为飞机的速度；(a_x, a_y, a_z) 为飞机的加速度；H 为飞机距海平面的高度；θ_{fire} 为飞机投弹的夹角；t_{detected} 为飞机被探测的时间总长且满足 $t_{\text{detected}} \leqslant t_{\text{intercept}}$ 的条件；*approch* 为飞机是否达到投弹的条件；*detected*，*fire*，*back* 均为标记变量，分别表示飞机是否被探测、飞机是否完成投弹，以及飞机是否脱离探测区域。具体算法流程如图 4 所示。

4 结果分析

在初始飞机位置、投弹条件等可随机设置情况下，开展训练任务。在训练过程中，智能体通过每一时刻收获的态势信息，在未输入任何先验知识的情况下，经过回馈函数的动态指导，动态调整其自身认知决策的能力。

经过一段时间的训练，得到空中威胁智能体的收敛模型，为更方便地检验算法的稳定性，随机选取投弹条件

$$\begin{cases} \theta_{\text{fire}} \leqslant \pi/6, \\ H_{\text{fire}} \geqslant 2\ \text{km}, \\ distance \leqslant 90\ \text{km}. \end{cases} \tag{10}$$

不同智能体随机抽取的 14 条飞行轨迹如图 5，6 所示。其中绿色轨迹表示在智能体能够完成任务时所生成的轨迹，其余各颜色的轨迹表示智能体训练不充分时决策出的飞行轨迹。从图 6 中可以看出，智能体存在逐步进化的现象。

对最终收敛结果进行详细分析，能够清楚看到飞机自行迭代出的投弹策略，在飞机满足对舰攻击条件后尽早投弹，在完成投弹后迅速降高逃逸，以避免防空导弹打击。通过表 1 逃出探测区时间与生存时间的对比，可以看出序号 14 逃逸的时间占比最少，也相对合理。

通过智能体飞行决策轨迹趋势能够直观看出智能体能够通过降低高度躲避雷达跟踪并尽量深入

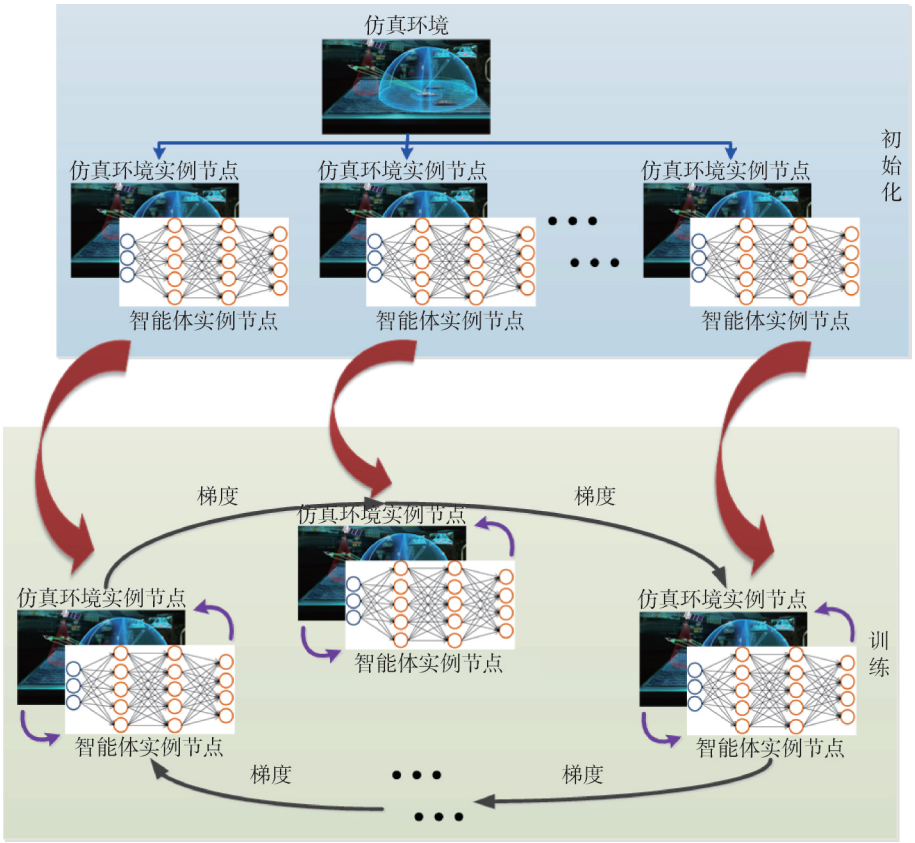

图 3 Ring-allreduce 分布式架构示意图

Fig. 3 Distributed architecture of ring-allreduce

到武器发射区内执行投弹过程。同时由于投弹限制，为了确保生存，智能体在投弹结束后会尽快降低高度以躲避舰艇防空导弹打击。学习的结果收敛且基本满足预期。

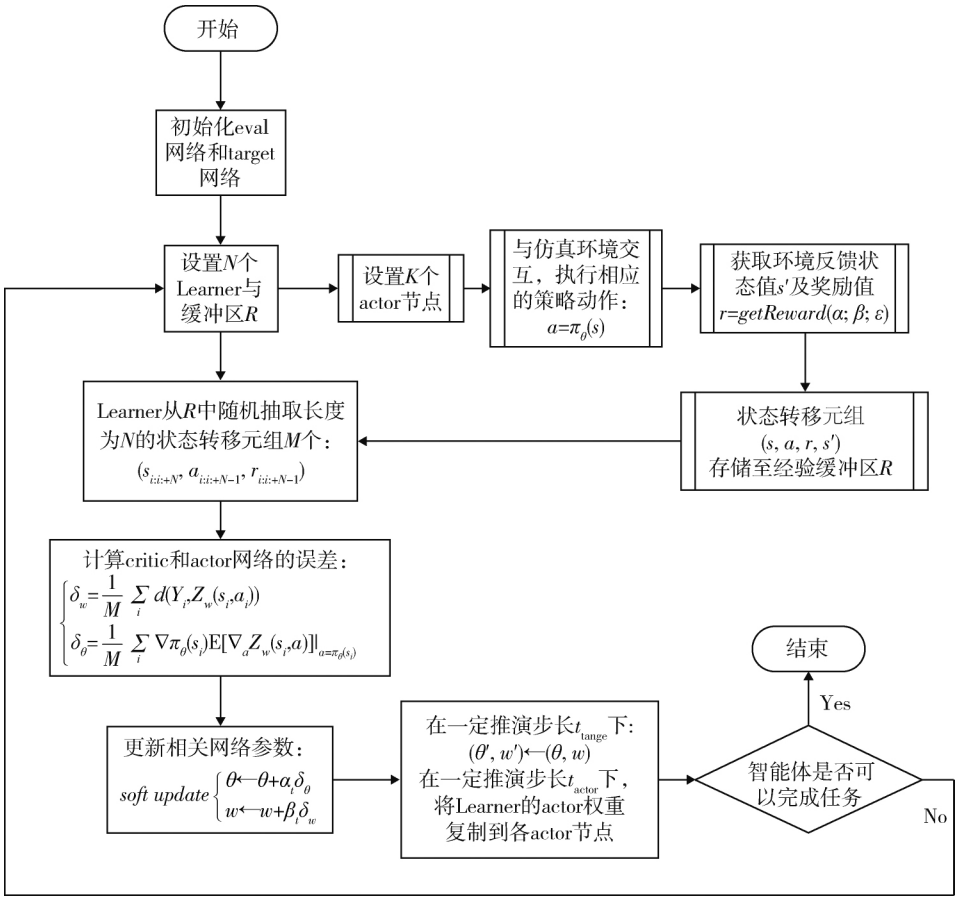

图 4　算法流程

Fig. 4　Algorithm flow chart

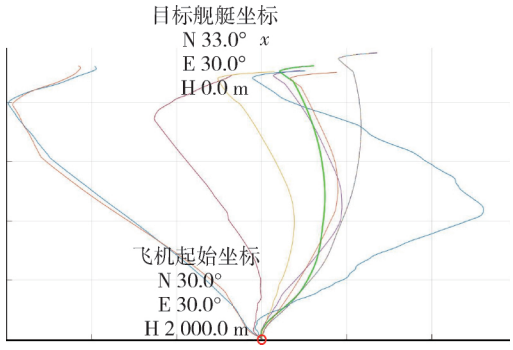

图 5　飞行轨迹对比图 1（地理坐标系俯视图）

Fig. 5　Flight path comparison chart 1 (Top view of geographical coordinate system)

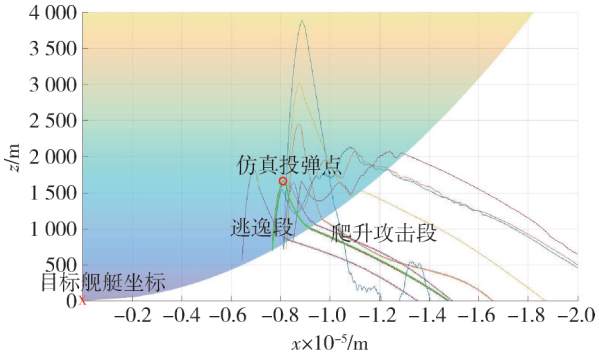

图 6　飞行轨迹对比图 2（雷达坐标系 RH 图）

Fig. 6　Flight path comparison chart 2 (RH view of Radar coordinate system)

表 1 逃出探测区时间与生存时间对比表

Table 1 Comparison of escape time and survival time

序号	逃逸时间	生存时间	逃逸时间/生存时间
1	80	102.087 937 00	0.783 638 130
2	59	96.768 490 39	0.609 702 598
3	111	112.526 573 00	0.986 433 666
4	83	93.679 597 52	0.885 998 683
5	72	97.691 475 36	0.737 014 153
6	64	84.711 531 85	0.755 505 167
7	130	130.733 426 30	0.994 389 910
8	112	125.975 651 90	0.889 060 690
9	118	126.645 197 60	0.931 736 870
10	64	84.711 531 85	0.755 505 167
11	64	90.886 893 79	0.704 171 936
12	120	130.234 100 20	0.921 417 661
13	78	109.098 098 30	0.714 952 884
14	52	91.600 691 48	0.567 681 304

5 结束语

随着深度强化学习、联盟学习等一大批新型智能技术发展，其在解决不完备信息条件下的对抗博弈，长远规划策略学习以及大规模交战及决策空间等问题的能力正在逐步得到认可。而在军事模拟训练领域，复杂程度虽然远远高于棋类游戏，但是人工智能技术依然成为了解决战争决策问题的一把金钥匙。

本文就在在这样的背景下，基于海军舰艇防空训练问题，构建了单机单舰的突防场景并开展了并行分布式场景仿真和空中威胁模型的迭代学习。通过大量学习并得到了相对收敛的结果。同时结果也基本符合单机单舰的突防过程，证明了基于深度学习的方法在海军舰艇防空模拟训练的可行性。

然而在开展验证过程中，也发现了一些问题，比如学习收敛速度在复杂场景条件下的急剧降低还有在多智能体的协同问题。后续的主要工作一方面将集中在优化并行仿真架构，提升学习的收敛速度。另一方面将主要开展多智能体的建模，利用人工智能架构实现多机编队攻击场景以及复杂对抗场景的学习和实现。

参考文献：

[1] 钟华. 贴近实战的外军军事训练 [J]. 国防科技，2014，35(4)：104.
ZHONG Hua. Foreign Military Training Close to Actual Combat [J]. National Defense Science and technology，2014，35(4)：104.

[2] 寇英信，李战武，李俊兵，等. 现代战斗机作战任务管理与决策 [M]. 北京：国防工业出版社，2017.
NI KOU Ying-xin，LI Zhan-wu，LI Jun-bing，et al. Mission Management and Decision of Modern Fighter [M]. Beijing：National Defense Industry Press，2017.

[3] 刘驰，王占健，戴子彭，等. 深度强化学习：学术前沿与实战应用 [M]. 北京：机械工业出版社.
NI LIU Chi，WANG Zhan-jian，DAI Zi-peng. Deep Reinforcement Learning：Research Frontiers and Practical Applications [M]. Mechanical Industry Press.

[4] POLI R，KENNEDY J，BLACKWELL T. Particle Swarm Optimization：An Overview [J]. Swarm Intelligence，2007，1(1).

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

[6] TESAURO G. Temporal Difference Learning and Td-Gammon [J]. Commmu. Acm 38，58-68(1995)

[7] Marc G Bellmare，Will Dabney，Remi Munos. A Distributional Perspective on Reinforcement Learning [C] // In International Conference on Machine Learning，2017：449-458.

[8] David Silver，Guy Lever，Nicolas Heess，et al. Deterministic policy gradient algorithms [C] // In International Conference on Machine Learning，2014.

[9] Diederik Kingma，Jimmy Ba Adam. A Method for Stochastic Optimization [C] // In International Conference on Learning Representations，2015.

[10] LILLICRAP T P，HUNT J J，PRITZEL A，et al. Continuous Control with Deep Reinforcement Learning [J]. arXiv Preprint arXiv：1509.02971 (2015).

[11] SILVER D，HUANG A，MADDISON C J，et al. Mastering the Game of go With Deep Neural Networks and Tree Search [J]. Nature 529 (2016) 484.

[12] VINYALS O，BABUSCHKIN I，CZARNECKI W M，et al. Grandmaster Level in StarCraft II Using Muti-agent Reinforcement Learning [J]. Nature 575，350-354 (2019) doi：10.1038/s41586-019-1724-z.

[13] GIBIANSKY A. BringHpc Techniques to Deep Learning. Technical report，Baidu Research，Tech. Rep.，2017，Zhang TESTS & CERTIFICATIONS IBM Certified Database Associate-DB2 Universal Database (2017)

[14] Gabriel，Barth-Maron，Matthew W，et al. Distributed Distributional Deterministic Policy Gradients [J]. arXiv preprint arXiv：1804.08617.

[15] John Schulman，Sergey Levine，Pieter Abbeel，et al. Trust Regon Policy Optimization [C] // In Proceedings of the 32nd International Conference on Machine Learning (ICML-15)，2015：1889-1897.