第 43 卷 第 2 期
2021 年 2 月

系统工程与电子技术
Systems Engineering and Electronics

Vol. 43 No. 2
February 2021

文章编号：1001-506X(2021)02-0443-09

网址：www. sys-ele. com

基于深度随机博弈的近距空战机动决策

马 文[1]，李 辉[1,2]，王 壮[1]，黄志勇[1]，吴昭欣[2]，陈希亮[3]

（1. 四川大学计算机学院，四川 成都 610065；2. 四川大学视觉合成图形图像技术国家级重点实验室，四川 成都 610065；3. 陆军工程大学指挥控制工程学院，江苏 南京 210007）

摘 要：针对空战中作战信息复杂、难以快速准确地感知态势做出决策的问题，提出一种博弈论与深度强化学习相结合的算法。首先，依据一对一典型空战流程，以随机博弈为标准，构建近距空战中红蓝双方对抗条件下的双机多状态博弈模型。其次，利用深度 Q 网络（deep Q network，DQN）处理战机的连续无限状态空间。然后，使用 Minimax 算法构建线性规划来求解每个特定状态下阶段博弈的最优值函数，并训练网络逼近值函数。最后，训练完成后根据网络输出求得最优机动策略。空战仿真实验表明，该算法具有较好的适应性和智能性，能够有效地针对空战对手的行动策略实时选择有利的机动动作并占据优势地位。

关键词：博弈论；深度强化学习；随机博弈；空战决策

中图分类号：TP 181 **文献标志码**：A **DOI**：10. 12305/j. issn. 1001-506X. 2021. 02. 19

Close air combat maneuver decision based on deep stochastic game

MA Wen[1], LI Hui[1,2], WANG Zhuang[1], HUANG Zhiyong[1], WU Zhaoxin[2], CHEN Xiliang[3]

(*1. College of Computer Science, Sichuan University, Chengdu 610065, China; 2. National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China; 3. College of Command and Control Engineering, Army Engineering University, Nanjing 210007, China*)

Abstract: In order to solve the problem of complex combat information and difficult to quickly and accurately perceive situation and make decision in air combat, an algorithm combining game theory and deep reinforcement learning is proposed. Firstly, according to the typical one-to-one air combat process and the standard of random game, a two machine multi-state game model under the condition of red and blue confrontation in close air combat is constructed. Secondly, deep Q network (DQN) is used to deal with the continuous infinite state space of fighter. Then, the Minimax algorithm is used to construct a linear programming to solve the optimal value function of the stage game in each specific state, and the network approximation value function is trained. Finally, the optimal maneuver strategy is obtained according to the output of the network after training. The simulation results show that the algorithm has good adaptability and intelligence for the air combat. It can effectively select the favorable maneuver action and occupy the dominant position according to the air combat opponent's action strategy.

Keywords: game theory; deep reinforcement learning; stochastic game; air combat strategy

0 引 言

自 1991 年海湾战争开创了空中力量为主赢得战争胜利的历史，空中力量在现代战争中起着越来越重要的作用，制空权的争夺很大程度上决定了战争的胜负[1]。然而空中作战形势瞬息万变，需要采集的信息极为复杂，使得作战方在感知空战态势后做出决策变得困难，传统方法无法实现快速准确的空战策略[2]。因此，如何根据双方作战态势选取有利且精确有效的空战机动策略是空中作战的重要研究方向。

国内外学者进行了相关研究，并提出了很多空战策略

收稿日期：2020－03－06；修回日期：2020－07－06；网络优先出版日期：2020－09－11。

网络优先出版地址：https://kns. cnki. net/kcms/detail/11. 2422. TN. 20200911. 1450. 006. html

基金项目：全军装备预研项目(31505550302)资助课题

引用格式：马文，李辉，王壮，等. 基于深度随机博弈的近距空战机动决策[J]. 系统工程与电子技术，2021，43(2)：443-451.

Reference format: MA W, LI H, WANG Z, et al. Close air combat maneuver decision based on deep stochastic game[J]. Systems Engineering and Electronics, 2021, 43(2): 443-451.

生成方法，传统方法包括：影响图[3]、专家系统法[4]、微分对策法[5]、矩阵对策法[6]等。这些方法在一定程度上为空战决策提供了有效的解决方案，但也有较大的局限。如遗传算法的机动决策具有一定主观性；专家系统法的可适应性较差；微分对策法计算量过大，求解困难；矩阵对策法存在难以确保实时性等。

近年来，空战决策的研究也日益增多，文献[7]引入博弈论的思想，提出了构建自由空战指挥引导对策模型的思路，但并未给出仿真结果。文献[8]提出了一种结合近似动态规划与零和博弈的在线积分策略迭代算法，解决了机动决策建模存在的“维数灾难”问题，但有学习周期偏长、难以应付复杂机动等缺点。文献[9]利用多状态转移马尔可夫网络构建机动决策网络，满足了空战决策的实时性要求，但未利用网络参数进行学习。文献[10]提出一种进化式的专家系统树方法来研究空战决策，解决了传统专家系统方法无法应付非预期情况的问题，但其设定的仿真环境较为简单。文献[11]提出了一种矩阵对策法与遗传算法相结合的空战决策算法，建立了无人机空战决策模型，满足空战合理性与实时性需求，但只对直线和 S 型飞行两种蓝色战机模型进行了仿真研究。

随着人工智能的发展，出现了越来越多人工智能在博弈中战胜人类的事例：2016 年人工智能 Alpha 飞行员打败了一位美国空军上校[12]。2018 年，AlphaGo Zero 在 3 天内自学了 3 种棋类游戏，轻而易举战胜了最优秀的人类棋手[13]。这些事例表现出了人工智能在智能决策方面的极大潜能。以最著名的 AlphaGo 为例，其主要运用了深度强化学习方法。深度强化学习将深度学习的感知能力和强化学习的决策能力相结合，是一种更接近人类思维方式的人工智能方法[14]。深度强化学习适合解决连续决策问题，而空战博弈正属于此类问题[15]，因此从深度强化学习入手研究空战策略方法是一种可行思路。

本文将深度强化学习与博弈相结合，提出了一种基于深度强化学习的算法——Minimax-深度 Q 网络（deep Q network，DQN）。该方法使用 Minimax 算法[16]构建线性规划来求解每个特定状态阶段博弈的纳什均衡策略，并引入 DQN 来更新动作状态值函数，以得到一种针对高决策水平对手的最优策略。

1 随机博弈与深度强化学习

本文为研究空战中红蓝双方战机的对抗情况，获得一种对红方有利的最优空战策略，需要用到博弈论以及深度强化学习的理论知识。

1.1 博弈论

本文研究的空战博弈实际上就是红蓝双方的对抗过程，双方的竞争性质可以利用博弈论的知识概括。博弈论是研究决策者在决策主体各方相互作用的情况下，如何进行决策以及有关该决策的均衡问题的理论[17]，被广泛应用于军事问题研究。

博弈论的一个重要策略组合叫作纳什均衡[18]，即在一组策略中，所有的局中人在其他人不改变策略的情况下，此时的策略是最优的。即联立策略$(\pi_1^*,\pi_2^*,\cdots,\pi_n^*)$满足

$$V_i(\pi_1^*,\pi_2^*,\cdots,\pi_i^*,\cdots,\pi_n^*)\geqslant V_i(\pi_1^*,\pi_2^*,\cdots,\pi_i,\cdots,\pi_n^*)$$
$$\forall\pi_i\in\Pi_i,\ i=1,2,\cdots,n \tag{1}$$

则为一个纳什均衡。

本文研究的红蓝双方战机对抗情况与追逃博弈有密切联系，追逃博弈将参与双方定义为追踪者和逃脱者，在博弈过程中博弈各方均以红方最大利益为目标，一方的得益必然导致另一方的损失，二者的得失总和为零[19]，本文涉及的博弈类型也称二人零和博弈。并且追逃博弈考虑的是红方在最差情况下的对抗策略优化设计，即纵使蓝方采用非常智能的机动方式，红方仍可获得可以接受的对抗效果，且得到的结果是全局最优[20]。本文以此为研究思路，处理空战中红蓝双方战机的对抗问题，并引入深度强化学习知识。

1.2 深度强化学习

马尔可夫决策过程（Markov decision process，MDP）是指决策者周期性地或连续性地观察具有马尔可夫性的随机动态系统，序贯地作出决策[21]，即根据当前的观测状态选择一个动作执行到达下一步的状态，下一步的状态只与当前的状态和动作有关。

MDP 是强化学习的基础。强化学习是智能体以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，使智能体获得最大的奖赏[22]。强化学习适用于解决连续决策问题，因此可以应用于解决空战中红蓝双方战机对抗的决策问题[23]。

深度学习通过构建基于表示的多层机器学习模型，训练海量数据，学习有用特征，以达到提升识别、分类或预测的准确性[24]。深度学习具有较强的感知能力，但缺乏一定的决策力，因此将深度学习与强化学习相结合，为系统的感知决策问题提供了解决思路。

深度强化学习将深度学习与强化学习结合，用神经网络来拟合强化学习中的价值函数和策略函数，解决了强化学习状态空间较小的局限性[25]。由于本文研究的红蓝双方战机对抗的空战策略问题中，状态特征是连续多维的，因此可采用深度强化学习中基于价值函数的经典 DQN 算法[26]解决该问题。

1.3 随机博弈

由于空战博弈是一个动态过程，而传统博弈一般是单步的，因此需要从传统博弈拓展到随机博弈。

MDP 包含一个玩家和多个状态，而矩阵博弈包含多个玩家和一个状态。对于具有多个玩家和多个状态的博弈，定义了一种 MDP 与矩阵博弈相结合的博弈方法，称为马尔可夫博弈，即随机博弈[27]。

随机博弈可表示为一个元组[28]，$(n,S,A_1,A_2,\cdots,A_n,T,\gamma,R_1,R_2,\cdots,R_n)$，其中包含的要素如下。

（1）个数 n：表示玩家数量。

（2）状态 S：状态是对环境的描述，在智能体做出动作后，状态会发生改变，其演变具有马尔可夫性。

（3）动作 A：动作是对智能体行为的描述，是决策的结果。动作空间可以是离散或连续的。

(4) 转移函数 T：由给定玩家当前状态 s 和每个智能体的一个动作 A_i 控制，转移概率在[0,1]之间。

(5) 折扣因子 γ：折扣因子是对未来奖励的衰减，$\gamma\in[0,1]$。

(6) 回报函数 R：表示指定玩家在状态 s 采取联合行为 $(A_1,A_2,\cdots,A_n)$ 后在状态 s' 处取得的回报。

随机博弈环境中的每个智能体都由一组状态 S 和一组动作集 $A_1,A_2,\cdots,A_k$ 定义，状态转换由当前状态 s 和每个智能体的一个动作 A_i 控制，每个智能体都有一个相关的奖励函数，试图最大化其预期的折扣奖励之和。与 MDP 类似，随机博弈中玩家下一状态和回报只取决于当前状态和所有玩家的当前行为。求解随机博弈需要找到一个策略 π，使得具有折扣因子 γ 的玩家的未来折扣回报最大化。

2　空战博弈建模

2.1　空战问题描述

本文红蓝双方战机采用文献[29]中开发的模拟无人机的动力学方程，在笛卡尔坐标系下构建战机的运动模型。空战态势主要由红蓝战机的位置 $(x^{\text{pos}},y^{\text{pos}})$、航迹偏角 ψ、滚转角 ϕ、滚转角变化率 $\dot{\phi}$、速度 v 和时间变化量 δt 定义：

$$\begin{cases}\phi=\phi+\dot{\phi}\delta t\\ \phi=\max(\phi,-\phi_{\max})\\ \phi=\min(\phi,\phi_{\max})\\ \dot{\psi}=\dfrac{9.81}{v}\tan\phi\\ \psi=\psi+\dot{\psi}\delta t\\ x^{\text{pos}}=x^{\text{pos}}+v\delta t\sin\psi\\ y^{\text{pos}}=y^{\text{pos}}+v\delta t\cos\psi\end{cases}\tag{2}$$

其中航迹偏角的限制范围为[−180°,180°]，滚转角的范围受实际飞机最大转弯能力限制，具体如图 1 所示。红方战机的目标是在蓝方战机背后取得并保持优势地位，可使用视界角(angle of aspect，AA)和天线偏转角(antenna train angle，ATA)来量化此优势位置。此外，航向交叉角(heading crossing angle，HCA)也用于描述红蓝战机之间的朝向差异。

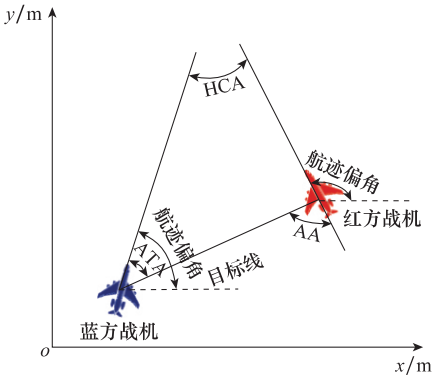


图 1　红蓝双方战机相对几何关系

Fig. 1　Relative geometric relationship between the red and the blue fighters

2.2　随机博弈建模

本文将红蓝双方战机作为智能体，以二人零和博弈为条件对空战博弈进行建模。根据第 1.2 节可知，随机需要确定一个元组 $(n,S,A_1,A_2,\cdots,A_n,T,\gamma,R_1,R_2,\cdots,R_n)$，根据此一元组来构建空战中的随机博弈模型。

2.2.1　随机博弈模型

首先需要确定随机博弈环境中每个智能体需要的状态空间 S、动作空间 A 和奖励函数 R，智能体为当前状态 s 决策选择一个动作 A_i 到达下一个状态 s'，并得到与环境交互后反馈奖励 r，然后进行下一轮交互，由此实现循环。

(1) 个数 n：红蓝双方战机对抗中玩家数量 n 为 2。

(2) 状态空间 S：根据影响战机空战态势的因素，可以确定战机的状态特征，主要由红方战机的坐标 $(x_r^{\text{pos}},y_r^{\text{pos}})$、蓝方战机坐标 $(x_b^{\text{pos}},y_b^{\text{pos}})$、航迹偏角 φ 和滚转角 ϕ 组成。由此可得博弈的状态空间可表示为

$$S=(x_r^{\text{pos}},y_r^{\text{pos}},\varphi_r,\phi_r,x_b^{\text{pos}},y_b^{\text{pos}},\varphi_b,\phi_b)$$

由于战机的状态空间是连续无限空间，所以需要用到深度学习神经网络来处理这些特征。

(3) 动作 A：战机的可选机动动作设置为向左滚转、维持滚转和向右滚转，分别用 L，S，R 代表这 3 种可选动作，构建离散的动作空间，则红方的动作空间为 $A_r=\{\text{L},\text{S},\text{R}\}$，同理蓝方动作空间为 $A_b=\{\text{L},\text{S},\text{R}\}$。

(4) 转移函数 T：以红方为例，红方当前状态 s 在红方根据策略选择的动作 a 与对手蓝方选择的动作 o 的联合行为 (a,o) 影响下，转移到下一状态 s' 的概率。

(5) 折扣因子 γ：折扣因子在[0,1]中选取，一般为 0.9 左右。

(6) 回报函数 R：在随机博弈中，使用 MDP 的 Q 值来表示即时收益。以 $Q(s,a,o)$ 表示每个状态 s 下，己方采取动作 a 及蓝方采取动作 o 的预期奖励。根据导弹的攻击区域，设定到达导弹可攻击范围为有利态势。对于红方的奖励值 r，若红方到达有利态势返回 $r=1$，若对手蓝方到达有利态势则 $r=-1$，其余情况 $r=0$。

2.2.2　战机优势奖励函数

本文参考文献[29]中定义的有利态势区域来选择占位，且暂不涉及 4 代、5 代红外近距格斗导弹的前向攻击能力[30]。以二维平面内的空战对抗为例，红方战机的优势区域，如图 2 所示。

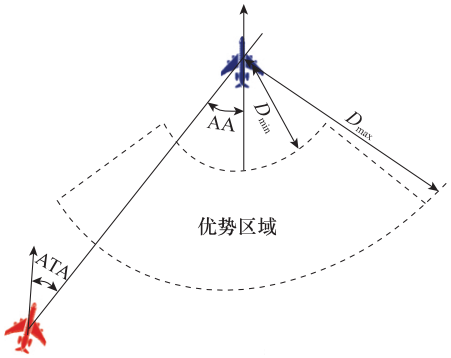


图 2　红方战机优势区域

Fig. 2　Dominant area of the red fighter

红方战机取得优势需要满足 4 个条件：① 红方战机与蓝方战机的欧氏距离 D 在 $D_{\min}$ 到 $D_{\max}$ 范围内，该区域根据战机的速度和武器攻击范围决定；② 红方战机与蓝方战机的高度差 H 在 $H_{\min}$ 到 $H_{\max}$ 范围内，该范围由战机的速度和武器攻击范围决定；③ 红方战机的 AA 在指定视界范围内；④ 红方战机的 ATA 在指定 ATA 范围内。同时满足以上 4 个条件则判定红方取得优势，并获得奖励值 $r=1$，即占据有利态势的要求如下所示：

$$\begin{cases} D_{\min}<D<D_{\max} \\ H_{\min}\leqslant H\leqslant H_{\max} \\ |\mathrm{AA}|<\mathrm{AA}_{\max} \\ |\mathrm{ATA}|<\mathrm{ATA}_{\max} \end{cases} \tag{3}$$

2.2.3 随机博弈价值函数

对于多人随机博弈，已知回报函数和转移函数，期望求得其纳什均衡解，即每个智能体的联合策略，智能体的策略就是动作空间的概率分布。由于在博弈环境下，预期回报会受到对手策略的影响，而在红蓝战机空战博弈中，一般无法预测到对手的动作。在此基础上，本文采用 Minimax 算法选取随机博弈的最优策略。假设对手拥有高水平决策能力，在蓝方选取使红方收益最小的动作的前提下，红方选取使自己收益最大的动作，该思想与追逃博弈类似。Minimax 算法的意义在于，在最坏的情况下获取最大的回报。

MDP 的价值函数表示最优策略所获得的预期折扣回报和，状态值函数 $V(s)$ 和状态动作值函数 $Q(s,a)$ 的公式如下：

$$\begin{cases} V(s)=\max\limits_{a'\in A}Q(s,a') \\ Q(s,a)=R(s,a)+\gamma\sum\limits_{s'\in S}T(s,a,o,s')V(s') \end{cases} \tag{4}$$

式中，$T(s,a,o,s')$ 表示状态 s 经过动作 a，o 到达状态 s' 的转移概率。

由此可得，随机博弈状态 s 下的最优值函数 $V(s)$ 可表示为

$$V(s)=\max_{\pi\in \mathrm{PD}(A)}\min_{o\in O}\sum_{a\in A}Q(s,a,o)\pi_a \tag{5}$$

式中，$\mathrm{PD}(A)$ 表示动作的离散概率分布。根据式(5)可以使用线性规划约束方法求得状态 s 下的最优策略 π 和最优值函数 V。

对于状态 s 下红方动作 a 及蓝方动作 o 的动作状态值函数 $Q(s,a,o)$ 为

$$Q(s,a,o)=R(s,a,o)+\gamma\sum_{s'}T(s,a,o,s')V(s') \tag{6}$$

通过上述的递归方程可以经过迭代求得收敛的最优值函数，进而得到最优策略 π。

由于红蓝双方战机对抗属于混合策略博弈，即博弈双方选择某一动作并不是确定的，而是对所有动作都有一个选择概率，此概率就是通过线性规划求得的最优策略 π。因此本文采用轮盘赌选择法进行动作选择，个体的适应度越高，被选择的概率越大。

3 深度强化学习空战博弈实现

3.1 空战深度强化学习模型

由于在博弈情境下，转移函数难以确定，对于式(6)中使用值迭代求解 MDP 的传统方法涉及的状态转移函数 T，可以利用强化学习中的异步更新方式 Q-learning[31] 替代。

Q-learning 利用时间差分目标来更新当前行为值函数，每当状态 s 采取动作 a 转换到状态 s' 时得到奖励 r 进行更新：

$$Q(s,a)=r+\gamma V(s') \tag{7}$$

由于执行更新的概率正是 $T(s,a,s')$，所以可以取代转移函数。将 Q-learning 的方法应用到随机博弈中，式(6)可转化为

$$Q_t(s,a,o)=(1-\alpha)Q_{t-1}(s,a,o)+\alpha(r+\gamma V(s')) \tag{8}$$

式中，α 代表学习效率。

与传统的 Q-learning 相比，Minimax-Q 方法结合了博弈论的思想，用 Minimax 值替换了 Q-learning 中的最大值，以得到博弈条件下需要的最优策略。

此外，由第 2.2 节可知，红蓝双方战机对抗的空战博弈所涉及的状态为连续无限空间，所以需要用到深度学习神经网络处理特征。因此，将 Minimax-Q 方法进一步拓展，加入深度神经网络来逼近值函数，利用经验回放训练强化学习的学习过程，并设置独立的目标网络来处理时间差分目标。

DQN 将 Q-learning 中的线性函数逼近以神经网络参数形式非线性逼近，可以处理空战博弈下高维度的非线性输入数据。DQN 的行为值函数对应一组参数，在神经网络里对应每层网络的权重，用 θ 表示，更新值函数实际上就是更新 θ 参数[32]。

Q 网络由输入层、隐藏层和输出层构成，本文构建的神经网络输入为战机的状态特征，即 $S=(x_r^{\mathrm{pos}},y_r^{\mathrm{pos}},\varphi_r,\phi_r,x_b^{\mathrm{pos}},y_b^{\mathrm{pos}},\varphi_b,\phi_b)$ 共 8 个节点，输出为状态 s 下所有红方可选动作 a 及蓝方动作 o 对应的 $Q(s,a,o)$。由于红蓝双方战机可选动作共有 3 种，所以输出为 9 个节点。

将智能体与环境交互得到的当前状态 s、红方采取的动作 a、蓝方采取的动作 o、对应的奖励值 r 以及执行动作到达的下一状态 s' 作为一个五元组 $\{s,a,o,r,s'\}$ 存储到记忆库。记忆库的大小是有限的，当记录库存储满后，新一组数据会覆盖记忆库中的第一组数据。从记忆库中随机抽取一定大小的数据作为训练样本，并计算出目标 Q 值来训练神经网络，计算目标 Q 值的方式即式(8)。

3.2 DQN 网络训练过程

由于强化学习是试错学习，要通过环境反馈的奖励 Reward 来优化损失函数，损失函数为 $\mathrm{loss}=(\mathrm{target\_}q-q)^2$，采用的优化方法为梯度下降。用神经网络逼近值函数

时，若计算目标值函数的网络与梯度计算逼近值函数的网络参数相同，会因为数据的关联性导致训练结果不稳定。因此需要定义两个神经网络，目标网络与 Q 网络的结构完全相同，但内部的参数不同。目标网络拥有 Q 网络一段时间以前的参数，这组参数被固定一段时间后，再将 Q 网络的最新参数传递给目标网络[31]。

Minimax-DQN 具体训练过程如图 3 所示。

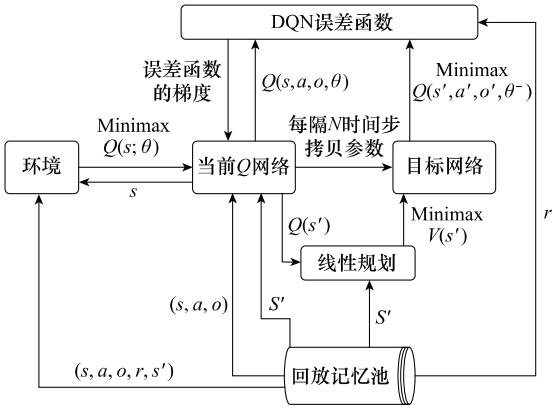


图 3　Minimax-DQN 训练过程示意图

Fig. 3　Schematic diagram of Minimax-DQN training process

3.3　Minimax-DQN 算法流程

总结上述内容给出 Minimax-DQN 的算法步骤如下。

步骤 1　初始化：给定红蓝双方一个初始状态，初始化记忆库，设置观察值。

步骤 2　创建两个神经网络分别为 Q 网络和目标网络，Q 网络参数为 θ，目标网络参数 $\theta^- = \theta$。神经网络输入为状态 s，输出为动作状态值函数 Q，学习一定次数后，将 Q 网络的参数拷贝给目标网络。

循环遍历：

步骤 3　红方智能体根据当前状态 s 按照策略 π 选择动作 a 并执行，得到下一状态 s' 以及获得的奖励 r。观测蓝方智能体在状态 s 下选取的动作 o，将 $\{s,a,o,r,s'\}$ 五元组存储到记忆库中。依据空战态势的复杂多样性，设置记忆库存储上限为 100 000 组数据。

步骤 4　从中记忆库中随机抽取 256 组数据作为一个训练样本。将训练样本的 s' 值作为神经网络的输入，根据神经网络输出得到状态 s' 下的 $Q(s')$。

步骤 5　根据式(5)使用线性规划得到 Minimax 状态值 $V(s')$，再根据式(8)计算出目标 Q 值 target_q。

步骤 6　计算损失函数

$\text{loss}=(\text{target\_}q-Q(s,a,o,\theta))^2$，采用梯度下降法进行优化，更新 Q 网络参数。

循环结束

步骤 7　根据式(5)使用训练好的神经网络输出的 Q 值进行线性规划求解得到最优策略 π。

根据上述算法，可得到算法流程图如图 4 所示。

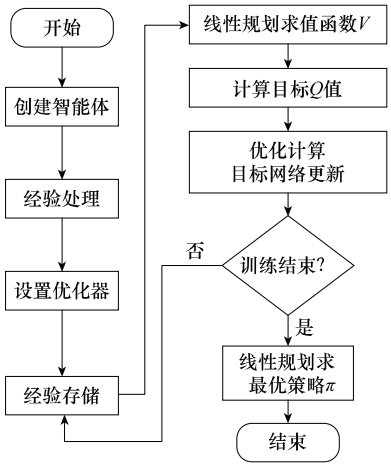


图 4　Minimax-DQN 算法流程图

Fig. 4　Flow chart of Minimax-DQN algorithm

4　仿真实验

假设红蓝双方处于同一高度水平，将仿真环境的空域范围限制在水平面内，横坐标 $x\in[-10\ \text{km},10\ \text{km}]$，纵坐标 $y\in[-10\ \text{km},10\ \text{km}]$，战机滚转角变化率 $\dot{\phi}=40$。给定红蓝双方的初始态势：蓝方战机的初始坐标为地图中央 $(x_b^{\text{pos}},y_b^{\text{pos}})=(0,0)$，速度 $v_b=300\ \text{m/s}$，航迹偏角和滚转角分别为 $\varphi_b=0°$，$\phi_b=0°$；红方战机速度 $v_r=300\ \text{m/s}$，滚转角 $\phi_r=0°$，为了模拟不同初始态势下的对抗情况，红方战机的初始坐标 $(x_r^{\text{pos}},y_r^{\text{pos}})$ 在 $x_r^{\text{pos}}\in[-1\ 000\ \text{km},1\ 000\ \text{km}]$，$y_r^{\text{pos}}\in[-500\ \text{km},500\ \text{km}]$ 范围内随机生成，航迹偏角 φ_b 在 $[-180°,180°]$ 随机取值。深度强化学习中折扣因子 $\gamma=0.95$，神经网络学习率 $\text{lr}=0.000\ 5$。使用上述参数进行红蓝双方战机博弈对抗仿真实验。

首先，红方根据 Minimax-DQN 算法选取策略，蓝方采用随机策略，若红方胜利则奖励值 $r=1$，若蓝方胜利则 $r=-1$，若飞出限定地图范围则 $r=0$。训练 10 000 个回合后，停止神经网络的学习。图 5 为算法训练过程中的损失变化图，横坐标为训练步数，纵坐标为神经网络每次训练的损失，可以看出随着训练步数的增加，训练损失逐渐下降最终收敛趋近于 0，满足了训练要求。

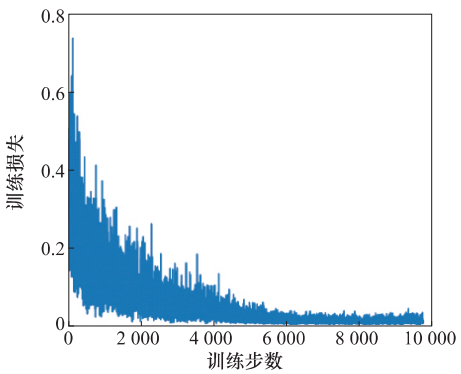


图 5　网络训练损失变化图

Fig. 5　Change diagram of network training loss

训练完成后，红蓝双方根据价值网络得出各自策略进行 1 000 次博弈对抗，最终得到的博弈获胜结果如图 6 所示。

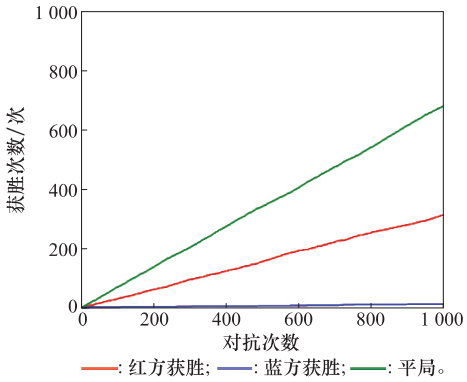


图 6 Minimax-DQN 对抗随机策略

Fig. 6 Confrontation between Minimax-DQN and random strategy

实验结果中，采用 Minimax-DQN 算法的红方获胜次数为 307 次，采用随机策略的蓝方获胜 12 次，另外有 681 局平局。红方在对抗中取得了较为优秀的成绩，证明了算法在博弈条件下的可行性。需要注意的是，该算法的重点是利用了线性规划求出的 Minimax 值函数 V 去更新神经网络中的 Q 值。本文在同样的情景条件及相同网络参数情况下，红方采用传统 DQN 算法选取最优策略，蓝方采用随机策略，得到的博弈结果如图 7 所示。其中，采用 DQN 算法的红方获胜次数为 212 次，采用随机策略蓝方获胜 15 次，另外有 773 局平局。对比图 6 与图 7 可以看出，蓝方采用的随机策略对抗性较差，而 Minimax-DQN 和 DQN 两种算法都可以在博弈情景下生成对抗策略，但 Minimax-DQN 算法的胜率比 DQN 更高，说明该算法相较 DQN 算法能够更准确有效地作出决策，引导战机占领有利的态势位置。

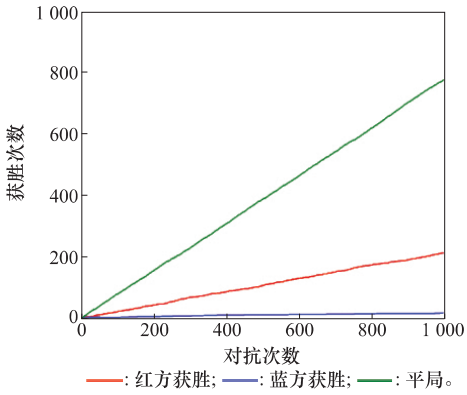


图 7 DQN 对抗随机策略

Fig. 7 Confrontation between DQN and random strategy

由于对局中平局次数较多，本文对两种算法的结果进行进一步分析，发现 Minimax-DQN 算法博弈结果中奖励值为 0 的对局，消耗的平均对抗步数约为 45 步，而 DQN 算法相同情况消耗的平均对抗步数约为 82 步。下面对双方奖励值为 0 的具有代表性的对局进行可视化，如图 8 所示。对比图 8 两种算法的对局可以发现，Minimax-DQN 算法达成平局主要是由于在红方达到优势区域前飞出了限定地图范围，因此平均步数较小。而 DQN 算法存在无法针对对手动作采取有效对抗策略的问题，因此在地图中消耗的步数更大。

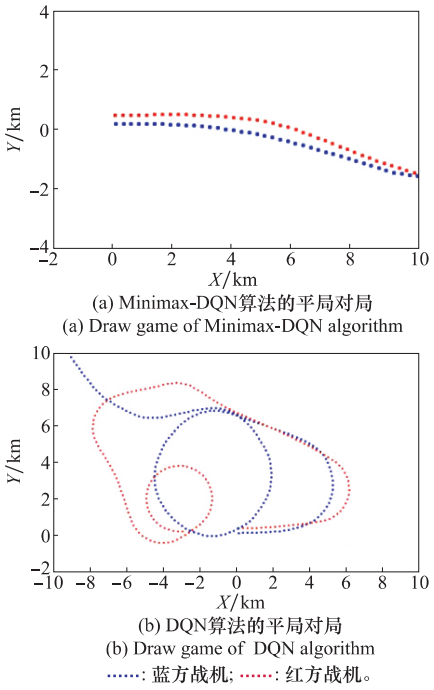


图 8 两种算法的平局对局

Fig. 8 Draw game of two algorithms

接下来，将上述两种算法进行对抗比较，红方智能体采用 Minimax-DQN 算法训练后的网络生成策略，而蓝方智能体采取 DQN 算法生成策略进行博弈对抗 1 000 次，得到的博弈结果如下图 9 所示。

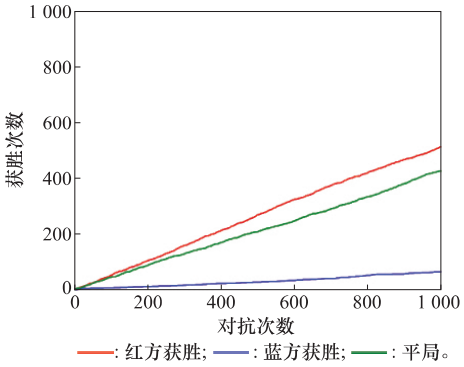


图 9 Minimax-DQN 对抗 DQN

Fig. 9 Confrontation between Minimax-DQN and DQN

图 9 的实验结果中，采用 Minimax-DQN 算法的红方获胜次数为 501 次，采用 DQN 算法的蓝方获胜 49 次，另外有 450 局平局。对比图 6 可以发现，Minimax-DQN 算法的获胜次数增加，平局数减少，在对手实力越强的情况下算法表现越好，验证了 Minimax 算法在最坏情况下表现出最大回报的性质。

图 10 为红方随机生成一个初始态势后与蓝方进行博弈对抗的仿真可视化轨迹图。图 11 为对抗过程中红方战机的状态变量图。可以看到，红方战机不断调整自己的状态最终占据了优势地位。

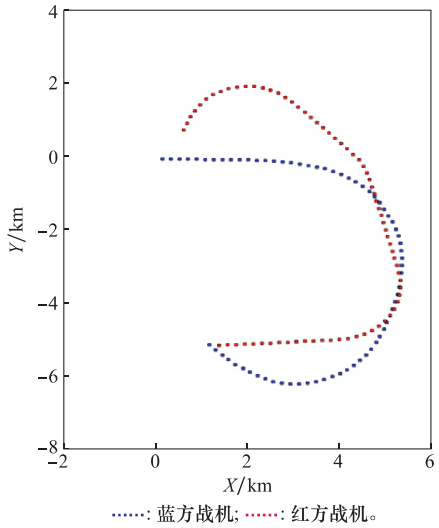
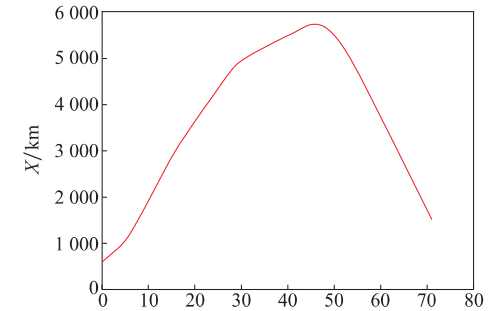


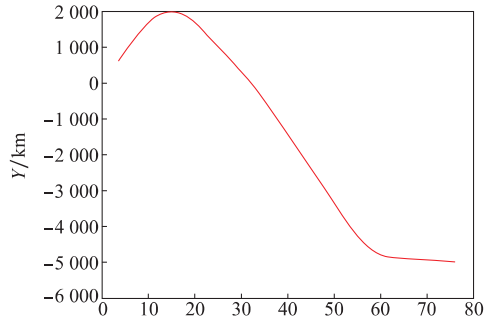
图 10　红方随机态势下的博弈对抗轨迹

Fig. 10　Game confrontation trajectory under the red random state



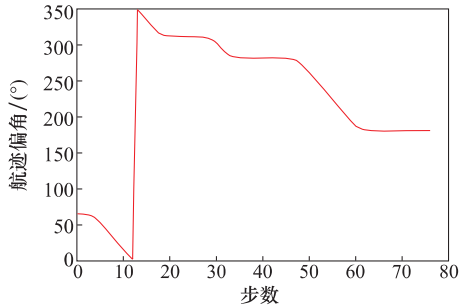
(a) 红方X坐标

(a) X-coordinate of the red



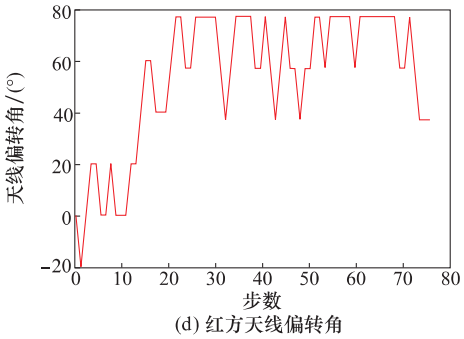
(b) 红方Y坐标

(b) Y-coordinate of the red



(c) 红方航迹偏角

(c) Track deflection angle of the red



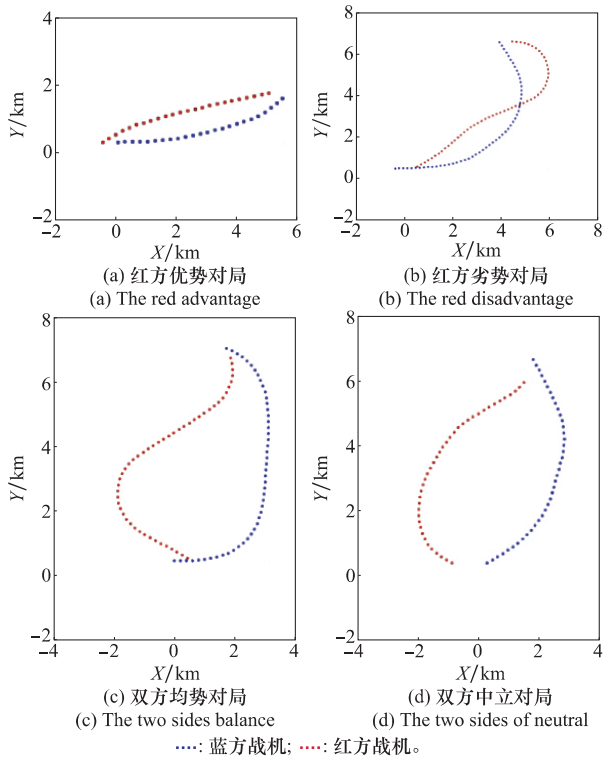
(d) 红方天线偏转角

(d) Antenna deflection angle of the red

图 11　红方战机状态变量

Fig. 11　State variables of the red fighter

为了测试 4 种典型初始态势下，红蓝双方的博弈对抗情况，进行了仿真测试如图 12 所示，4 组对局分别为红方优势对局、红方劣势对局、双方均势对局和双方中立对局。



(a) 红方优势对局

(a) The red advantage

(b) 红方劣势对局

(b) The red disadvantage

(c) 双方均势对局

(c) The two sides balance

(d) 双方中立对局

(d) The two sides of neutral

图 12　博弈对抗轨迹

Fig. 12　Trajectory of game confrontation

图 12(a)中，红方战机初始坐标位于蓝方战机西方 500 m 处，航迹偏角 $\varphi_b=30°$，两机距离小，红方战机达到攻击蓝机条件，红方处于优势。可以看出蓝方战机想掉头改变局势，但被红方战机利用初始位置的优势拦截，红方率先达到优势态势。图 12(b)中，红方战机初始坐标位于蓝方战机东方 500 m 处，航迹偏角 $\varphi_b=30°$，两机距离小，蓝方战机达到攻击红机条件，红方处于劣势。可以看出红方战机率先掉头改变局势，利用角度偏差最终扭转局面获得胜利。图 12(c)中，红方战机初始坐标位于蓝方战机东方 500 m

处，航迹偏角 $\varphi_b=150°$，两机距离小，双方均达到攻击对方条件，两机处于均势。红蓝双方同时转向，红方率先拉近距离并取得角度优势，最终获胜。图 12(d)中红方战机初始坐标位于蓝方战机西方 1 000 m 处，航迹偏角 $\varphi_b=150°$，两机距离大，双方均不满足攻击对方条件，两机处于均势。可以看出依旧是红方率先拉近距离，从左后方占据优势态势。

由 4 组仿真实验可以看出，红方能在任意初始态势下，利用决策最终占据有利态势，验证了算法的有效性。

结合上述仿真实验结果及对比分析，Minimax-DQN 算法结合了博弈思想来更新神经网络值函数，具有较好的适应性和智能性，能够在博弈场景下准确地作出有效决策来引导战机占据有利的态势位置。

5 结 论

本文采用深度强化学习与博弈相结合的方法，提出了一种可解决空战中红蓝双方战机对抗机动决策问题的算法，即 Minimax-DQN 算法。该算法利用深度学习神经网络来处理空战中战机高维连续的态势特征，并引入强化学习决策模型，通过与环境交互训练智能体，最后采用 Minimax 算法构建线性规划求解出纳什均衡策略，得到战机的最优机动决策。

目前本文通过空战仿真实验验证了算法的可行性，下一步的工作将加入高度影响，将仿真环境从二维拓展到三维，并考虑雷达和武器情况进行现代空战的研究，使算法适应更加复杂的战场环境，还将由一对一空战决策问题拓展到多对多的集群协同作战中的博弈智能问题研究。

参考文献：

[1] 孔江涛. 面向双机空战机动决策的置信规则推理技术研究[D]. 长沙：国防科学技术大学，2015.
KONG J T. Research of belief-rule-based reasoning technology for learning air combat maneuvers[D]. Changsha: National University of Defense Technology, 2015.

[2] HUANG C Q, DONG K S, HUANG H Q, et al. Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization[J]. Journal of Systems Engineering and Electronics, 2018, 29(1): 90 - 101.

[3] VIRTANEN K, KARELAHTI J, RAIVIO T. Modeling air combat by a moving horizon influence diagram game[J]. Journal of Guidance Control & Dynamics, 2006, 29(5): 1080 - 1091.

[4] CHAPPELL A R. Knowledge-based reasoning in the Paladin tactical decision generation system[C]// Proc. of the 11th Digital Avionics Systems Conference, 1992: 155 - 160.

[5] HORIE K, CONWAY B A. Optimal fighter pursuit-evasion maneuvers found via two-sided optimization[J]. Journal of Guidance, Control, and Dynamics, 2006, 29(1): 105 - 112.

[6] SU M C, LAI S C, LIN S C, et al. A new approach to multi-aircraft air combat assignments[J]. Swarm and Evolutionary Computation, 2012, 6: 39 - 46.

[7] 董肖杰，余敏建. 基于博弈论的自由空战指挥引导对策问题研究[J]. 航空计算技术，2017，47(2)：80 - 84，88.
DONG X J, YU M J. Study on countermeasure of free air combat command and guide based on game theory[J]. Aeronautical Computing Technique, 2017, 47(2): 80 - 84, 88.

[8] 梅丹，刘锦涛，高丽. 基于近似动态规划与零和博弈的空战机动决策[J]. 兵工自动化，2017，36(3)：35 - 39.
MEI D, LIU J T, GAO L. Maneuver decision of air combat based on approximate dynamic programming and zero-sum game[J]. Ordnance Industry Automation, 2017, 36(3): 35 - 39.

[9] 罗元强，孟光磊. 基于马尔可夫网络的无人机机动决策方法研究[J]. 系统仿真学报，2017，29(S1)：110 - 116.
LUO Y Q, MENG G L. Research on UAV maneuver decision-making method based on markov network[J]. Journal of System Simulation, 2017, 29(S1): 110 - 116.

[10] 王炫，王维嘉，宋科璞，等. 基于进化式专家系统树的无人机空战决策技术[J]. 兵工自动化，2019，38(1)：48 - 53.
WANG X, WANG W J, SONG K P, et al. UAV air combat decision based on evolutionary expert system tree[J]. Ordnance Industry Automation, 2019, 38(1): 48 - 53.

[11] 邓可，彭宣淇，周德云. 基于矩阵对策与遗传算法的无人机空战决策[J]. 火力与指挥控制，2019，44(12)：61 - 66，71.
DENG K, PENG X Q, ZHOU D Y. Study on air combat decision method of UAV based on matrix game and genetic algorithm[J]. Fire Control & Command Control, 2019, 44(12): 61 - 66, 71.

[12] 周光霞，周方. 美军人工智能空战系统阿尔法初探[C]// 第六届中国指挥控制大会，2018：66 - 70.
ZHOU G X, ZHOU F. A preliminary study of the alpha of the US army's Artificial intelligence air combat system[C]// Proc. of the 6th China Command and Control Conference, 2018: 66 - 70.

[13] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354 - 359.

[14] VINCENT F L, PETER H, RIASHAT I, et al. An introduction to deep reinforcement learning [J]. Foundations and Trends in Machine Learning, 2018, 11(3/4): 219 - 354.

[15] MA Y F, MA X L, SONG X. A case study on air combat decision using approximated dynamic programming[J]. Mathematical Problems in Engineering, 2014(4): 183 - 193.

[16] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[C] // Proc. of the 11th International Conference on Machine Learning, 1994: 157 - 163.

[17] CORCHON L C, MARINI M A. Handbook of game theory and industrial organization, volume I, theory[M]. Miami: Edward Elgar, 2018.

[18] PAVLIDIS N G, PARSOPOULOS K E, VRAHATIS M N. Computing Nash equilibria through computational intelligence methods[J]. Journal of Computational & Applied Mathematics, 2005, 175(1): 113 - 136.

[19] BARDHAN R. An SDRE based differential game approach for maneuvering target interception[C]// Proc. of the AIAA Guidance, Navigation and Control Conference, 2015: 704 - 711.

[20] OYLER D W, KABAMBA P T, GIRARDA R. Pursuit-evasion games in the presence of obstacles[J]. Automatica, 2016, 65(c): 1 - 11.

[21] MCGREGOR S, BUCKINGHAM H, DIETTERICH T G, et al. Interactive visualization for testing Markov decision processes: MDPVIS[J]. Journal of Visual Languages & Computing, 2017, 39(4): 93 - 106.

[22] 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]. Systems Engineering and Electronics, 2020, 42(7): 1567 - 1574.

[23] MAO M Y, ZHANG A, ZHOU D, et al. Reinforcement learning of UCAV air combat based on maneuver prediction[J]. Electronics Optics & Control, 2019, 26(2): 5 - 10.

[24] LECUN Y, BENGIO Y, HINTON G. Deep learning.[J]. Nature, 2015, 521(7553): 436 - 444.

[25] NGUYEN T, NGUYEN N D, NAHAVANDI S. Multi-agent deep reinforcement learning with human strategies[C]// Proc. of the IEEE International Conference on Industrial Technology, 2019: 1357 - 1362.

[26] SEWAK M. Deep reinforcement learning: frontiers of artificial intelligence[M]. Singapore: Springer, 2019: 95 - 108.

[27] SHAPLEY L S. Stochastic games[J]. Proceedings of the National Academy of Sciences, 1953, 39(10): 1095 - 1100.

[28] SCHWARTZ H M. Multi-agent machine learning: a reinforcement approach[M]. New Jersey: Wiley Publishing, 2014.

[29] JAMES S M, JONATHAN P H, BRIAN W, et al. Air-combat strategy using approximate dynamic programming[J]. Journal of Guidance, Control, and Dynamics, 2010, 33(5): 1641 - 1654.

[30] 樊会涛. 第五代空空导弹的特点及关键技术[J]. 航空科学技术, 2011(3): 1 - 5.
FAN H T. Characteristics and key technologies of the fifth generation of air to air missiles[J]. Aeronautical Science & Technology, 2011(3): 1 - 5.

[31] WATKINS C J C H. Learning from delayed rewards[D]. London: University of Cambridge, 1989.

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

作者简介：

马　文(1997 -)，女，硕士研究生，主要研究方向为深度强化学习技术。
E-mail:1262578027@qq.com

李　辉(1970 -)，男，教授，博士，主要研究方向为智能计算、战场仿真、虚拟现实。
E-mail:lihuib@scu.edu.cn

王　壮(1987 -)，男，博士研究生，主要研究方向为军事人工智能、深度强化学习技术。
E-mail:zhuang_wang@qq.com

黄志勇(1995 -)，男，硕士研究生，主要研究方向为深度强化学习技术。
E-mail:771048263@qq.com

吴昭欣(1996 -)，男，硕士研究生，主要研究方向为战场仿真、深度强化学习技术。
E-mail:597779499@qq.com

陈希亮(1985 -)，男，副教授，博士，主要研究方向为深度强化学习、指挥信息系统工程。
E-mail:383618393@qq.com