Journal of System Simulation

Volume 36 | Issue 6 Article 18

6-28-2024

Curriculum Learning-based Simulation of UAV Air Combat Under Sparse Rewards

Jingyu Zhu
School of Electrical Engineering, Xinjiang University, Urumqi 830000, China, zhujingyu@stu.xju.edu.cn

Hongli Zhang
School of Electrical Engineering, Xinjiang University, Urumqi 830000, China, zhl@xju.edu.cn

Minchi Kuang
Department of Precision Instrument, Tsinghua University, Beijing 100084, China

Heng Shi
Department of Precision Instrument, Tsinghua University, Beijing 100084, China

See next page for additional authors

Follow this and additional works at: https://dc-china-simulation.researchcommons.org/journal

 Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation. For more information, please contact xtfzxb@126.com.

Curriculum Learning-based Simulation of UAV Air Combat Under Sparse Rewards

Abstract

Abstract: To address the limited exploration capabilities and sparse rewards of conventional reinforcement learning methods in air combat environment, a curriculum learning distributed proximal policy optimization (CLDPPO) reinforcement learning algorithm is proposed. A reward function informed by professional empirical knowledge is integrated, a discrete action space is developed, and a global observation and local value and decision network featuring separated global and local observations is established. A methodology for unmanned aerial vehicles UAVs is presented to acquire combat expertise through a sequence of fundamental courses that progressively intensify in their offensive, defensive, and comprehensive content. The experimental results show that the methodology surpasses the specialist system and the other mainstream reinforcement learning algorithms, which has the ability of the autonomous acquisition of air warfare tactics and can enhance the sparse rewards.

Keywords

UAVs, air combat, sparse reward, curriculum learning, distributed proximal policy optimization (DPPO)

Authors

Jingyu Zhu, Hongli Zhang, Minchi Kuang, Heng Shi, Jihong Zhu, zhi Qiao, and Wenqing Zhou

Recommended Citation

Zhu Jingyu, Zhang Hongli, Kuang Minchi, et al. Curriculum Learning-based Simulation of UAV Air Combat Under Sparse Rewards[J]. Journal of System Simulation, 2024, 36(6): 1452-1467.

This paper is available in Journal of System Simulation: https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18

第 36 卷第 6 期 系统仿真学报© Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

稀疏奖励下基于课程学习的无人机空战仿真

祝靖宇[1]，张宏立[1*]，匡敏驰[2]，史恒[2]，朱纪洪[2]，乔直[2]，周文卿[3]
(1. 新疆大学 电气工程学院，新疆 乌鲁木齐 830000；2. 清华大学 精密仪器系，北京 100084；3. 清华大学 计算机科学技术系，北京 100084)

摘要：针对传统强化学习在空战环境下探索能力差和奖励稀疏的问题，提出了一种基于课程学习的分布式近端策略优化(curriculum learning distributed proximal policy optimization，CLDPPO)强化学习算法。嵌入包含专家经验知识的奖励函数，设计了离散化的动作空间，构建了局部观测与全局观测分离的演员评论家网络。通过为无人机制定进攻、防御以及综合课程，让无人机从基本课程由浅入深开始学习作战技能，阶段性提升无人机作战能力。实验结果表明：以课程学习方式训练的无人机能以一定的优势击败专家系统和主流强化学习算法，同时具有空战战术的自我学习能力，有效改善稀疏奖励的问题。

关键词：UAVs；空战；稀疏奖励；课程学习；分布式近端策略优化

中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2024)06-1452-16

DOI: 10.16182/j.issn1004731x.joss.23-0349

引用格式: 祝靖宇, 张宏立, 匡敏驰, 等. 稀疏奖励下基于课程学习的无人机空战仿真[J]. 系统仿真学报, 2024, 36(6): 1452-1467.

Reference format: Zhu Jingyu, Zhang Hongli, Kuang Minchi, et al. Curriculum Learning-based Simulation of UAV Air Combat Under Sparse Rewards[J]. Journal of System Simulation, 2024, 36(6): 1452-1467.

Curriculum Learning-based Simulation of UAV Air Combat Under Sparse Rewards

Zhu Jingyu[1], *Zhang Hongli*[1*], *Kuang Minchi*[2], *Shi Heng*[2], *Zhu Jihong*[2], *Qiao zhi*[2], *Zhou Wenqing*[3]

(1. School of Electrical Engineering, Xinjiang University, Urumqi 830000, China; 2. Department of Precision Instrument, Tsinghua University, Beijing 100084, China; 3. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: To address the limited exploration capabilities and sparse rewards of conventional reinforcement learning methods in air combat environment, *a curriculum learning distributed proximal policy optimization (CLDPPO) reinforcement learning algorithm is proposed. A reward function informed by professional empirical knowledge is integrated, a discrete action space is developed, and a global observation and local value and decision network featuring separated global and local observations is established. A methodology for unmanned aerial vehicles UAVs is presented to acquire combat expertise through a sequence of fundamental courses that progressively intensify in their offensive, defensive, and comprehensive content.* The experimental results show that the methodology surpasses the specialist system and the other mainstream reinforcement learning algorithms, which has the ability of the autonomous acquisition of air warfare tactics and can enhance the sparse rewards.

Keywords: UAVs; air combat; sparse reward; curriculum learning; distributed proximal policy optimization (DPPO)

收稿日期：2023-03-29 修回日期：2023-05-18

第一作者：祝靖宇(1998-)，男，硕士生，研究方向为无人系统、深度强化学习。E-mail：zhujingyu@stu.xju.edu.cn
通讯作者：张宏立(1972-)，男，教授，博士，研究方向为智能信息处理。E-mail：zhl@xju.edu.cn

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

0 引言

无人机空战是最具复杂性的任务之一，世界各国投入了大量的精力对无人机空战进行研究，旨在替代人类飞行员完成复杂任务。随着传感器技术和计算能力的飞速提升，空战系统自主决策能力也逐渐成为可能。尽管当前完全自主的空战系统并未完全实现，但已经成为研究热点。为了应对人工智能对无人机空战对抗技术的挑战，提升未来空天战场的核心作战能力[1]，研发新一代无人机空战系统就显得尤为重要。现代空战决策方法主要分为两类：

第一类是基于规则化的机器搜索方法。最初文献[2]开发了自适应机动决策的专家系统，其主要的逻辑为IF-ELSE规则式决策方式；文献[3]针对空战决策中不确定性和实时性，提出了一种基于规则集和模糊贝叶斯网络的混合战术决策方法；文献[4]提出了一个时间敏感信息的空战环境的约束策略博弈模型，并使用线性编程以及线性不等式来解决该问题；文献[5]将一对一空战的优势函数和环绕优势函数转化为混合整数非线性规划问题，通过粒子群算法来确定空战决策方案。

第二类是基于深度强化学习的自演进方法。文献[6]将专家知识作为启发信号，将启发式算法和强化学习相结合；文献[7]提出了基于DQN的智能战术决策方法；文献[8]提出了基于确定性策略梯度的无人机自主空战机动决策学习模型；文献[9]提出一种结合动态关系权重算法和移动时间策略的决策方法，并且在机动决策中加入轨迹预测；洛克希德–马丁公司[10]将分层架构与最大熵强化学习相结合，在阿尔法狗斗比赛中，击败了美国空军F-16武器教练课程的毕业生，但是在比赛中并非使用真实的导弹，而是使用了攻击区域的设定来简化了问题；文献[11]提出了近似动态规划得到了一个近似策略迭代算法，用于解决空战机动任务建模的马尔可夫决策过程模型。近年来，一些学者对空战复杂环境进行了研究[12-13]，文献[14]在复杂空战环境下使用PPO算法来做出空战决策，但是没有使用专家系统作为基准来判断其能力；文献[15]设计了四种近端策略优化的算法增强机制，用于解决多机空战问题多机协同对抗问题；文献[16]提出奖励整形的多智能体深度确定性策略梯度算法(multi-agent deep deterministic policy gradient, MADDPG)的空战决策算法，能够得到有效空战策略优于原MADDPG和DPPG算法。

上述无人机空战算法的研究存在的问题：

(1) 基于规则化的机器搜索方法核心在于设计专家知识的知识库或优良的态势函数评估战场环境态势，继而使用规则判断或优化算法搜索得到空战决策策略。因此，此类方法常有对于专家知识要求极为严格、态势函数难以设计、优化算法搜索实时性难以保证、建模场景相对简单且缺乏自我探索创新能力的问题。

(2) 基于深度强化学习的自演进方法核心在于通过神经网络极强的拟合能力，结合强化学习的优化能力，能够有效解决实时性的问题。然而，由于正常空战对抗时间过长，对抗过程奖励稀疏，智能体难以显著地提升性能，各个研究单位构建的仿真环境并不一致，所以，在空战领域鲜有对比的基准实验，未能与专家系统、专业飞行员对抗，难以评估其真实作战水平。

鉴于上述各类方法的不足之处，本文提出了一种基于课程学习的分布式近端策略优化(curriculum learning distributed proximal policy optimization, CLDPPO)强化学习算法。通过给智能体设置进攻、防御以及综合课程，让智能体分阶段地提升作战能力。嵌入了包含专家知识的奖励函数，能够有效地缓解奖励的稀疏性，提升智能体的作战性能。为了能够合理地评估智能体的作战性能，将专家系统作为基准算法来评估智能体的水平，使用主流的强化学习算法来对比算法性能，验证了所提课程学习能够有效地提升智能体的性能表现。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

第 36 卷第 6 期 系统仿真学报 Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

1 空战强化学习模型

1.1 空战基本模型

对于单无人机对抗系统，将单无人机博弈对抗过程建模为POMDP(partially observable markov decision process)问题，主要由五元组$<S,A,R,P,\gamma>$组成，S为一系列的离散状态$s_0,s_1,\cdots,s_t,s_{t+1},\cdots$；$A$为一系列的离散动作$a_0,a_1,\cdots,a_t,a_{t+1},\cdots$；$R$为一系列的离散奖励$r_0,r_1,\cdots,r_t,r_{t+1},\cdots$；$P$为状态转移函数，$P(s_{t+1}|s_t,a_t)$表示为当前状态动作对$(s_t,a_t)$映射到能够到达的后继状态$s_{t+1}$的概率分布；$\gamma$为折扣因子，用于定义未来奖励的重要性。

如图1所示，空战基本模型由空战智能体、空战环境、状态空间、动作空间和态势奖励组成。其中，空战智能体获取当前空战环境的状态，并通过感知状态特征选择当前决策动作，空战环境根据智能体所选择的动作，反馈当前状态选择该动作的奖励值给智能体。当智能体完成当前动作之后，进入到下一个状态，重新选择动作执行并获得奖励，不断循环该过程直至对抗结束。智能体通过获取环境反馈的奖励不断更新迭代自我的策略，从而学习到态势奖励所引导的空战对抗策略。

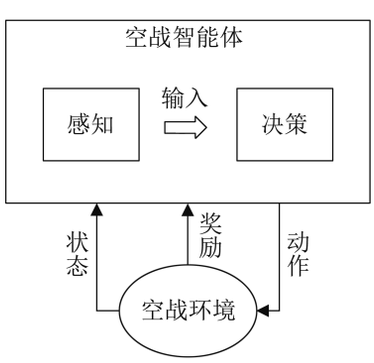


图1 空战基本模型
Fig. 1 Basic model of air combat

1.2 状态空间建模

在实际空战中，由于战机迷雾和雷达探测能力的限制，当前战机所能得到的信息是不完美信息，完备的状态空间设计则显得尤为重要。观测空间中除了对自身和敌人的状态观测，还有对空战态势的表示。为了便于智能体分析敌我态势信息，增加了经过处理的态势信息，并对原始数据和敌我态势建立几何关系映射，如图2所示。其中，使用McGrew[17]对空战态势的几何表示方法作为参考。本机与蓝机的位置与速度为X_{self}、X_{enemy}、v_{self}、v_{enemy}，无人机的姿态由俯仰角φ、滚转角θ和偏航角ϕ组成，天线列角为θ_{ATA}、纵横角为θ_{AA}、水平交叉角为θ_{HCA}、仰角为θ_{EL}，距离为d、雷达锁定信号为L、雷达等级为R_{level}、导弹来临的雷达告警信息为R_{alarm}、剩余导弹数量为M_{left}、当前是否可发射导弹为M_{castable}。总空战状态空间定义如下：

$$S=[X_{\text{self}},X_{\text{enemy}},v_{\text{self}},v_{\text{enemy}},\varphi,\theta_{\text{roll}},\phi,d,\theta_{\text{ATA}},\theta_{\text{AA}},\theta_{\text{HCA}},\theta_{\text{EL}},L,R_{\text{level}},R_{\text{alarm}},M_{\text{left}},M_{\text{castable}}] \quad (1)$$

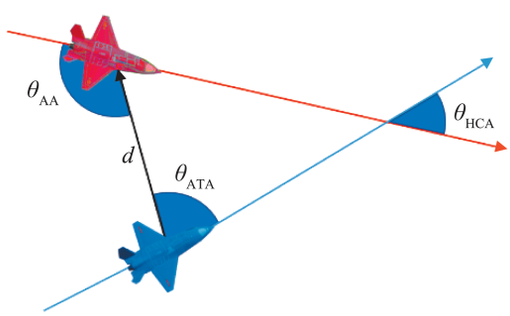


图2 空战角度
Fig. 2 Air combat angles

1.3 动作空间建模

为了使代理充分自我探索了空战中可能出现的战术，设定给智能体为空战中具备专家知识的基础动作[18]。通过合理灵活地运用这些基础动作，能够做出一些空战中常见的机动，如殷麦曼机动和眼镜蛇机动等。与此同时，通过限定动作的种类，从而缩减了动作的搜索空间。一共具有6类动作，如表1所示。

动作空间定义如下：

$$A=[\text{Straight, Track, Circle, Loop, Attack, Escape}] \quad (2)$$

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

表1 动作空间设计
Table 1 Action design

类型	参数	描述
直飞	目标俯仰角	直飞、拉升和俯冲
追踪	目标位置和方位	追踪敌人或预测点
盘旋	滚转角和俯仰轴	制定了一系列的固定参数
筋斗	俯仰角	用于快速改变方向和位置
攻击	目标预测位置	攻击目标预测位置
躲避	报警信息	调整机体垂直于报警方向飞行或下高回转

1.4 态势奖励建模

空战对抗过程中具有一些具体的节点事件，对于代理来说具有充分的鼓励性。通过加入包含飞行员经验的节点事件奖励，并结合空战中连续状态变化[19]的奖励，从而组成了具备专家经验的奖励函数。

节点事件奖励：在所有奖励中，最为重要的奖励是命中蓝方以及被击中奖励，标志着空战对局的胜负；导弹发射奖励用于鼓励智能体在合适的位置发射导弹；近距离躲避蓝方导弹的奖励，用于鼓励智能体躲避蓝方导弹；获取蓝方视野奖励，用于鼓励智能体寻找蓝方位置，对于空战对抗是至关重要的；导弹近距离略过蓝机迫使其做大规模机动，为后续进攻铺垫，此类进攻具备一定战略意义的。通过加入节点事件奖励给代理一定的鼓励，使智能体便于探索环境。

连续状态变化奖励：当无人机做机动过程中，有很多连续状态的观测量是不断变化的，这些观测量是空战对抗中常用的专家知识。连续观测量主要由速度和侧滑角限幅惩罚、无人机之间相对姿态优势和导弹的威胁组成。

无人机飞行的速度于空战博弈过程尤为重要，R_{velocity}用于鼓励代理保持高速飞行，过低的速度容易被蓝方所击中。为了能够让无人机保持相对良好的姿态，给以无人机的侧滑角一定的惩罚，R_β用于限制无人机的侧滑角。

$$R_{\text{velocity}}=\begin{cases}-80, & v<80\\ v-160, & 80\leqslant v\leqslant 160\\ 0, & v>160\end{cases} \tag{3}$$

$$R_\beta=\begin{cases}-40, & \beta>20^\circ\\ -0.1\beta^2, & \beta\leqslant 20^\circ\end{cases} \tag{4}$$

为了能够合理的评估无人机之间姿态优势，通过无人机之间的相对角度和距离综合来评估其姿态优势奖励$R_{\text{advantage}}$。正如图3所示，当两机相对角度θ为0时，此时己方无人机尾追蓝机。随着两机之间的距离d不断的减少，己方无人机的姿态优势也相应的上升。

$$R_{\text{advantage}}=\begin{cases}A(d/1\,000), & d\leqslant 1\,000\\ A\left((39\,000-d)/38\,000\right), & 1\,000<d\leqslant 20\,000\\ A(10\,000/d), & 20\,000<d\\ A=80\mathrm{e}^{-\frac{\theta^2}{1\,300}}\end{cases} \tag{5}$$

式中：θ为两无人机之间的夹角；d为两无人机之间的距离。

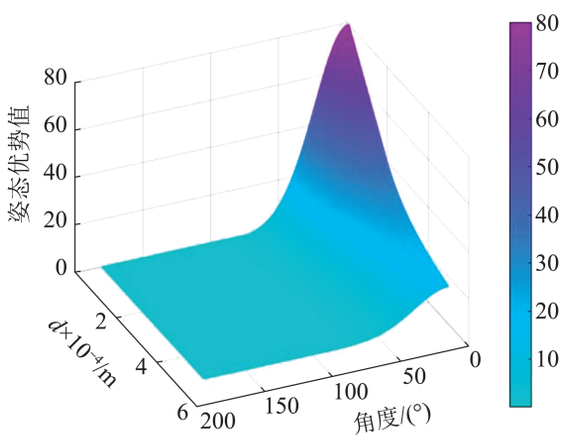


图3 无人机姿态优势
Fig. 3 Aircraft position advantage

R_{threat}用于评估蓝方导弹对于无人机的威胁。导弹与无人机之间的角度越大，蓝机通过使用机动易于躲避来袭导弹，尾追蓝机的导弹难以躲避。由图4可知，当导弹与蓝机角度为0时，且具备接近蓝机能力的导弹能够得到最高的威胁。

$$R_{\text{threat}}=\begin{cases}160\mathrm{e}^{-\frac{\theta^2}{144}}(1-t/20), & 0\leqslant t<20\\ 0, & 20\leqslant t\end{cases} \tag{6}$$

式中：θ为导弹与无人机之间的夹角；t为导弹预计接近无人机的时间。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

Zhu et al.: Curriculum Learning-based Simulation of UAV Air Combat Under Spar

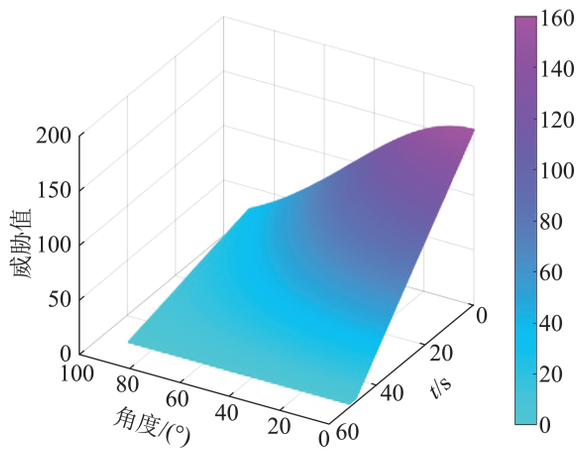


图4 导弹威胁奖励
Fig. 4 Missile threat reward

空战智能体中具备专家知识的奖励设计，各个奖励量的设计根据奖励类型的重要程度，动态地调整奖励量的数值，具体奖励设计详见表2。

2 基于课程学习的空战模型

2.1 课程设计

文献[19]提出了课程学习的理念，该理念借鉴了人类学习的方式，从简单任务开始逐步深入学习，以提高学习效率。根据该理念为智能体设置了3个课程：攻击课程、防御课程和综合课程[19]。其中，攻击课程通过让智能体与无攻击性的靶机对抗来训练其导弹攻击技能；防御课程则通过关闭智能体的攻击功能，让其学习如何躲避蓝方导弹的攻击；综合课程则将前两个课程的技能结合起来，并允许智能体与专家系统进行对抗，以平衡攻击和防御技能的运用。通过三个课程的有序训练，能够有效地提升智能体的学习效率，并让其掌握专业的空战技能。

课程学习的空战模型如图5所示。空战环境经过课程学习模块设定训练课程，智能体在设定课程的空战环境下获取状态量，通过对状态变量编码输入到决策网络中输出动作给智能体执行，进而得到该动作的奖励。收集状态、动作和奖励数据，通过分布式近端策略优化算法(distributed proximal policy optimization, DPPO)计算梯度来更新决策网络参数，实现策略的更新。当观测智能体课程完成之后，设定下一阶段的课程，不断定制阶段化的课程帮助智能体提升作战性能。

2.2 决策网络设计

智能体的目标是得到一个空战的最优策略，将其定义为从观测空间到动作的概率分布函数。其核心神经网络主要由深层长短期记忆网络(LSTM)组成，为了能够解决空战问题中部分观测的问题，使用了深层LSTM核心网络来处理时序数据，设计了具有全局观测量与局部观测量相分离的Actor-Critic网络架构。

表2 奖励设计
Table 2 Reward design

名称	奖励	奖励类型	描述	行为意图
击中	640	节点事件奖励	导弹击中对方敌机	鼓励击败敌方
被击中	−640	节点事件奖励	被敌方导弹击中	惩罚自身被击中
发射导弹	−30~−10	节点事件奖励	根据剩余导弹变化	惩罚发射导弹
视野	20	节点事件奖励	通过雷达获取到敌方位置	鼓励获取敌方视野
丢失视野	−20	节点事件奖励	本机丢失敌方视野	惩罚失去敌方视野
导弹近距离略过	50	节点事件奖励	己方导弹近距离略过敌机	鼓励有效发射导弹
躲避敌方导弹	50	节点事件奖励	己方近距离躲避敌方导弹	鼓励躲避来袭导弹
机体失速	−80~0	连续变化奖励	机体失速	惩罚机体失速
侧滑角限幅	−40~0	连续变化奖励	限制侧滑角度	惩罚过大侧滑角
导弹威胁	0~160	连续变化奖励	由相对角和接近时间计算	鼓励导弹给敌方带来威胁
姿态优势	0~200	连续变化奖励	根据弹目距离和角度计算	鼓励本机出于优势态势

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

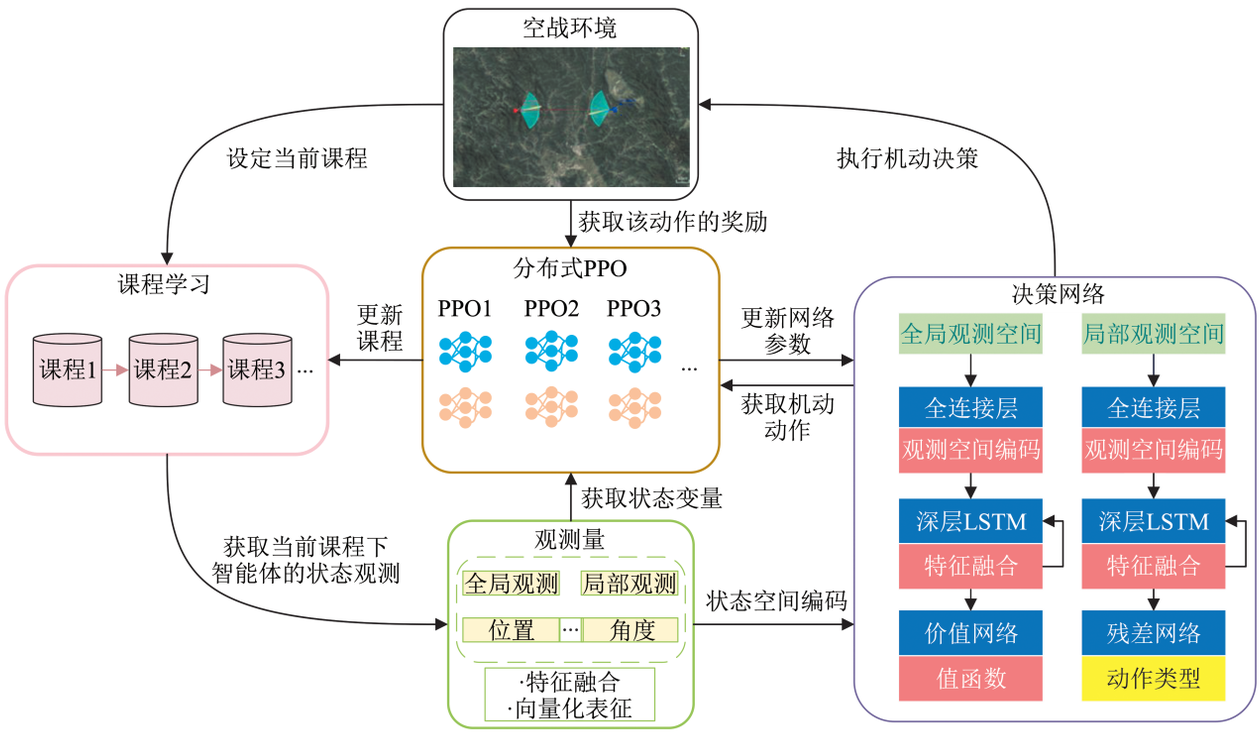


图5　课程学习的空战模型

Fig. 5　Air combat model for curriculum learning

智能体的网络架构如图6所示，价值网络和策略网络分别使用全局观测空间和局部观测空间，通过使用全连接层获取观测量的特征，进而使用深层的LSTM来提取时序信息的特征。价值网络将全局观测量所提取的特征通过全连接层输出当前状态量的价值，策略网络通过将提出的局部观测量的特征作为残差网络的输入，最终输出当前状态所选择的动作类型。在给定策略的情况下，通过不断将当前观测作为输入，并在每个时间步从输出分布采样出动作来控制代理与环境交互博弈。将复杂的多维观测值通过编码输入到LSTM网络，使用LSTM网络状态提取时序特征，继而通过使用全连接层和残差网络来预测当前执行的策略(动作和值函数)。

2.3 算法设计

在空战中，状态s_t表示 t 时刻无人机的状态。例如，无人机当前位置、发射导弹信息，以及蓝方告警等空战状态。无人机使用当前的观测量s_t来选择动作a_t。通过执行动作a_t结合所设定的奖励函数，得到对应当前动作的奖励r_t。无人机执行完动作a_t之后，到达新的状态s_{t+1}，根据新的状态做出新的决策a_{t+1}，得到下个状态s_{t+1}的奖励r_{t+1}。智能体的目标是最大化无人机所得到的总回报：

$$G_t=\sum_{t=0}^{\infty}\gamma^t r_t \tag{7}$$

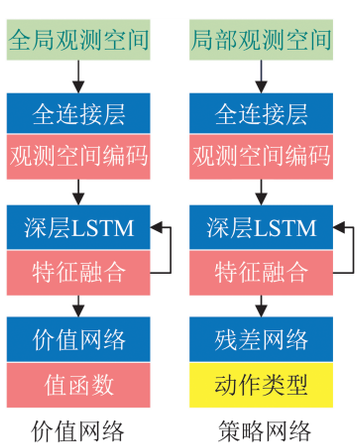


图6　决策网络架构

Fig. 6　Decision network architecture

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

Zhu et al.: Curriculum Learning-based Simulation of UAV Air Combat Under Spar

为了能够最大化所得总回报，采用策略学习的方法来学习一个最优策略π^*，对无人机进行动作决策。为了能够充分表示当前状态s_t和动作a_t的价值，使用状态价值函数$V^{\pi}(s_t)$和动作价值函数$Q_{\pi}(s_t,a_t)$来表示。基于策略π的状态价值函数，表示为$V^{\pi}(s_t)=E_{\pi}[G_t|S=s_t]$。动作价值函数定义为$Q_{\pi}(s_t,a_t)=E_{\pi}[G_t|S=s_t,A=a_t]$。优势函数定义为$A_{\pi}(s_t,a_t)=Q_{\pi}(s_t,a_t)-V_{\pi}(s_t)$。通过使用参数化$\theta$代理的策略$\pi_{\theta}$，并且设计衡量策略优劣的目标函数：

$$J(\theta)=E_{S_0}[V_{\pi}(s_0)] \tag{8}$$

将目标函数对策略θ导数之后，使用梯度上升的方法来最大化这个目标函数，从而使策略最优。策略梯度的目标是寻找到一个最优策略π^*并且最大化这个策略在环境中的期望回报。

$$\pi^*=\arg\max_{\pi} E_{\pi}\left\{\sum_{t}^{H}\gamma^t r_{t+1}|S=s_0\right\} \tag{9}$$

策略梯度的方法主要沿着梯度方向$\nabla J(\theta)$迭代更新参数θ，但是这个算法无法保证每次更新步幅大小，可能由于步幅过大导致策略突然变差从而影响训练效果。为了解决这个问题文献[20]提出一阶导数的近端策略优化(proximal policy optimization, PPO)算法，使用的框架为Actor-Critic(AC)框架。PPO的Actor网络更新部分是最大化“surrogate”目标：

$$\begin{aligned}&\max L^{\mathrm{CPI}}(\theta)=\hat{E}_t[r_t(\theta)\hat{A}_t]\\&r_t(\theta)=\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)}\end{aligned} \tag{10}$$

式中：$r_t(\theta)$为当前策略π_{θ}和旧策略$\pi_{\theta_{\mathrm{old}}}$的概率比例；$\hat{A}_t$为使用广义优势估计(generalized advantage estimation，GAE[21])来估计在状态s_t下动作a_t的优势值。广义优势估计借鉴了时序差分(temporal-difference, TD)TD(λ)的思想，将TD(λ)中的值函数换成了优势函数通过调整λ的大小可以得到不同的近似估计。在PPO中，对surrogate目标进行裁剪：

$$L^{\mathrm{clip}}(\theta)=\hat{E}_t[\min(r_t(\theta)\hat{A}_t,\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{11}$$

其中，clip的实现如式12所示，clip的作用在于将x限制在$[x_{\min},x_{\max}]$，从而使策略在有限范围内变化，通过使用KL损失来大大简化算法。

$$\mathrm{clip}(x,x_{\min},x_{\max})=\max(\min(x,x_{\max}),x_{\min}) \tag{12}$$

熵是随机变量不确定性的度量。在强化学习中，为了帮助智能体提升探索能力，通常在Actor的loss上添加一项策略熵，并且乘以一个系数，设置为0.01。策略的熵可以表示为

$$\begin{aligned}H(\pi(\cdot|s_t))=&-\sum_{a_t}\pi(a_t|s_t)\log(\pi(a_t|s_t))=\\&E_{a_t\sim\pi}[-\log(\pi(a_t|s_t))]\end{aligned} \tag{13}$$

Critic网络使用时序差分误差来更新网络参数ϕ，其中$V_{\phi}(s_t)$估计了状态s_t的状态价值函数：

$$\delta_t=\gamma V_{\phi}(s_{t+1})+r_{t+1}-V_{\phi}(s_t) \tag{14}$$

设定智能体当前的课程，智能体通过获取当前设定空战状态，使用决策网络选择动作，并得到该动作的奖励，直至智能体获胜或者被击败，此过程为工作，智能体称之为工人。一个GPU上具有M个工人，一共具有N个GPU，每个工人在不断空战环境交互生成经验数据，并保存于经验回放区中，每当满足经验回放区的阈值时，每个GPU采样数据，根据分布式近端策略优化算法[22]得到价值网络和策略网络的局部梯度。最终，平均每个GPU上的梯度得到全局梯度，用于更新价值网络和策略网络。当智能体完成设定课程时，切换到下个课程，否则，保持当前课程直至完成。课程学习分布式近端策略优化算法的伪代码如算法1~2所示。

算法1　课程学习分布式近端策略优化(工人)

C个课程，N个GPU，一个GPU上M个工人，B次迭代次数

1 for k = 1 to C do
2　for i = 1 to N do
3　　for j = 1 to M do
4　　智能体执行策略π_{θ}，收集数据$\{s_t,a_t,r_t,s_{t+1},\cdots\}$
5　　存储智能体交互信息

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

6 end

7 更新策略 $\pi_{\text{old}} \leftarrow \pi_\theta$

8 for b = 1 to B do

9 从缓存区 buffer 中获取数据 $\{s_t, a_t, r_t, s_{t+1}, \cdots\}$

10 使用广义优势估计来估计优势函数 A^{π_θ}

11 计算其策略函数和价值函数

12 $$J_{\text{ppo}}(\theta)=\sum_{t=1}^{T}[\min(r(\theta),\text{clip}(r(\theta),1-\epsilon,1+\epsilon))A^{\pi_{\theta_k}}(s,a)]$$

13 策略更新率 $r(\theta)=\dfrac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}$

14 $$V_{\text{loss}}=\sum_{t=1}^{T}(-V(s_t)+r_t+\gamma V(s_{t+1}))^2$$

15 计算 $\nabla J_{\text{ppo}}, \nabla V_{\text{loss}}$，并且发送梯度 θ, ϕ 给主人

16 end

17 end

18 未完成设定课程保持 k 不变

19 end

算法2 课程学习分布式近端策略优化(主人)

1 for k = 1 to C do

2 for i = 1 to N do

3 for j = 1 to M do

4 等待获取设定的梯度 θ, ϕ

5 平均获取的梯度并且更新全局梯度 θ

6 平均获取的梯度并且更新全局梯度 ϕ

7 end

8 end

9 未完成设定课程保持 k 不变

10 end

3 无人机空战训练系统

3.1 空战决策流程

基于课程学习的空战模型构建了无人机的整体决策流程框架系统。首先，给智能体设定阶段化的课程，其次，将空战环境的原始数据进行特征融合，转化为有效的状态空间，并通过网络输出状态价值和动作决策。同时将智能体与环境交互的数据保存到经验回放区中，每当获取到一定量当前策略生成的数据时，使用分布式近端策略优化更新价值网络和策略网络，提供新的策略网络和价值网络给智能体使用。当某一课程完成，将开展后续的课程，从而不断提升智能体的作战性能。单机空战决策流程框架如图7所示。

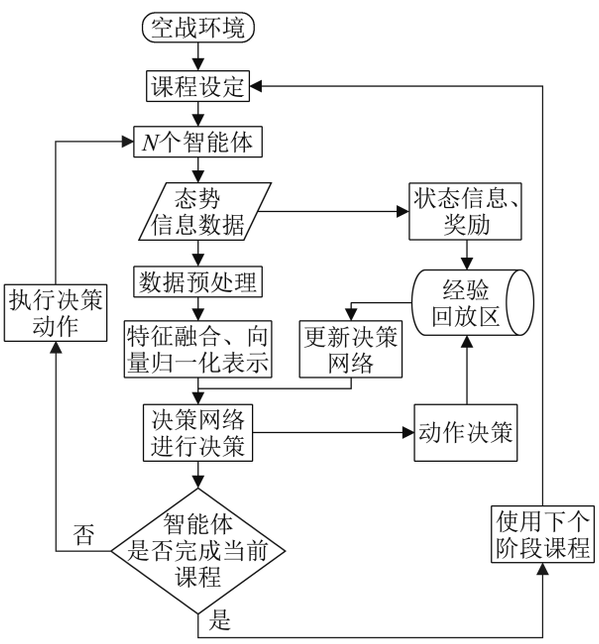


图7 单机空战决策流程

Fig. 7 Single aircraft air combat decision

3.2 训练系统框架

训练系统框架主要由四部分组成，如图8所示。战场仿真环境通过环境控制器并行运行在CPU上，并且与前向传播器建立了高速的通信。前向传播器主要由特征融合和决策网络构成，获取战场仿真环境的状态观测，状态观测通过特征工程预处理后通过决策网络前向传播产生机动决策。智能体依据获取的机动决策与空战环境进行交互，从而改变智能体状态得到奖励。将这一过程产生的数据整理打包，包括当前智能体状态观测、机动决策、奖励、LSTM隐含层每16步条打

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

包为一条序列，异步发送到由Redis构建的经验回放池中。优化器中GPU采样数据来更新系统模型参数，由控制器进行版本分发。

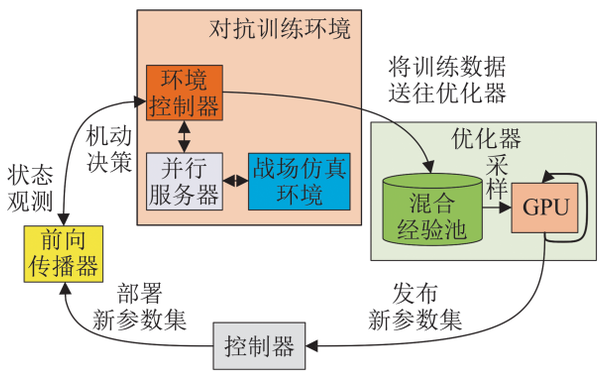


图8　系统训练架构
Fig. 8　Systematic training architecture

每个优化器的GPU对经验回放池中的数据采样得到mini-batch，并对其梯度进行计算。使用NVIDIA NCCL2协议的MPI标准函数对多个优化器间所计算的梯度进行平均，得到平均梯度之后返回给每个GPU模型并自行进行梯度下降，以此保证分布式系统上的GPU模型同步更新。在每个小批量的数据中具有120个样本，每一个样本具有16个动作、状态组合的序列对。使用Adam优化器通过在小批量的16个动作的样本上反向传播计算梯度，来更新模型参数，并设计了全局统计器以更新次数来判断是否分发新版本给控制器。通过分布式并行处理使模型动态快速进化，通过采用全局平均梯度，让智能体能够稳步提升性能。

系统架构如图9所示，其使用了Redis数据库作为前端和后端之间的桥梁，实现高速态势数据传输。Redis数据库用于保存模型和仿真环境产生的数据。仿真环境产生的数据通过空战模拟战场产生，战场上具有平行战场。在训练过程中，利用多核处理器增加效率，同时利用NCCL通信同步多训练进程的结果，实现超大规模训练。

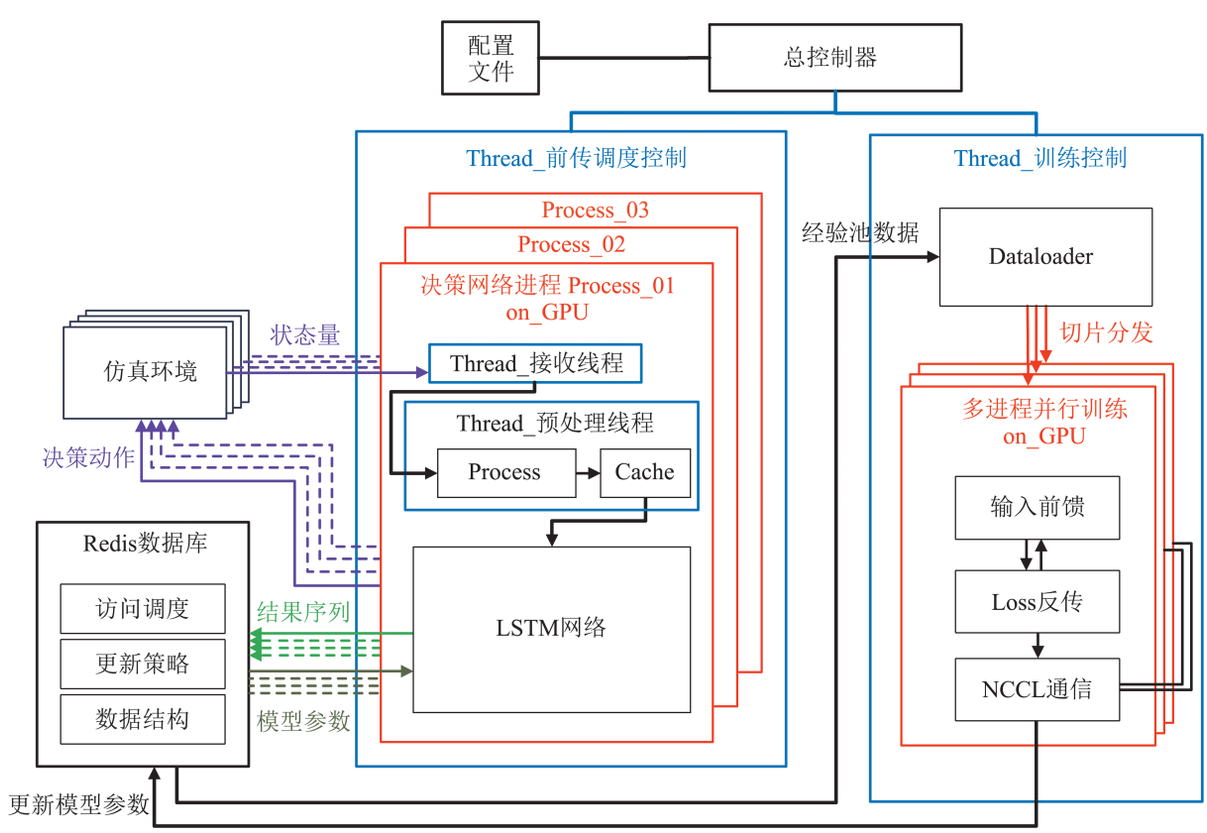


图9　系统架构
Fig. 9　System architecture

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

首先设置好含有系统架构数据的配置文件，打开启动器，同时开启两个部分。第一个部分用于启动进程管理器，进程管理器启动两种类型的线程，第一种线程为网络模型前向传播的线程，其主要功能是用于获取仿真环境的观测来进行前向传播产生决策动作返回仿真环境给智能体执行。第二种线程为获取经验回放区的数据进行反向传播的线程，其最主要的功能是用于使用数据梯度上升更新网络模型。这些进程通过配置器的分布式配置在各个GPU上。第二部分用于网络通信连接，开放端口给各个仿真环境前端连接。通过监听线程来启动传输数据的功能，主要功能为产生通信队列、端口分发和通信上下文资源。能够同时开启多个端口监听多个仿真环境前端，从而建立分布式架构，极大地增加数据的吞吐量，加速智能体进化过程。

4 红蓝对抗实验仿真

4.1 实验想定

实验想定如图10所示，红蓝双方配置相等，各自包含6枚导弹，其初始生成位置由系统随机生成。课程学习实验一共分为攻击、防御和综合课程，红机由强化学习训练的智能体控制，蓝机根据课程设定为靶机或专家系统–有限状态机(finite state machine, FSM)。

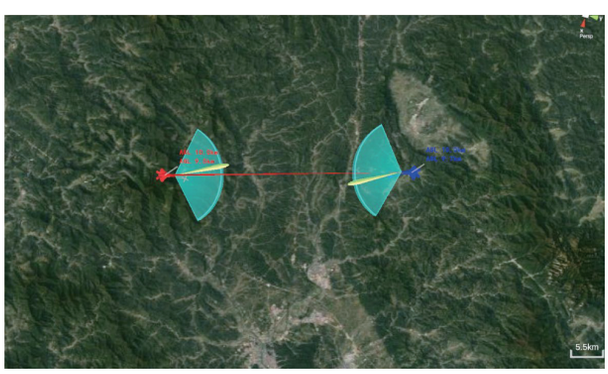


图10 红蓝机实战对抗
Fig. 10 Red and blue drone combat

想定过程由三种课程组成，首先，设定攻击课程，使智能体对抗没有攻击能力的靶机，靶机是指按照设定轨迹飞行的无人机。当攻击课程完成后，关闭智能体的攻击动作，进而与状态机对抗来学习防御课程。最终，开放智能体动作，使其与状态机学习综合课程。

4.2 实验条件

实验设备一共为两类设备：前端服务器和后端服务器，前端服务器用于并行空战仿真产生对抗数据，后端服务器将数据采样用于更新优化网络。实验的前端台式机搭载的CPU为AMD-5950X、显卡为NVIDIA GeForce RTX3080ti、内存为64 GB，后端服务器搭载的CPU为Intel Xeon Gold 6230R×2、GPU为NVIDIA GeForce RTX3090×8、内存为512 GB。

4.3 实验环境

实验环境是面向强化学习的空战仿真平台[23]，该仿真平台建立了数字孪生空战环境，具备作战单元配置、批量对战等功能。考虑真实空战场景战机携带的导弹数量有限，每架战机携6枚AIM-120中距弹，该导弹携带雷达导引头，可通过无人机提供制导链。

对抗环境具有平行战场机制，每个平行战场独立对战，如图11所示。前端设备可以同时开启多个对战环境，通过TCP/IP协议与后端模型建立Socket通信，有效地增加了数据的吞吐量，能够加速代理训练。后端服务器实时将经验池中的数据采样出来，使用分布式近端策略优化来计算梯度和均方差，用于更新Actor和Critic网络。在空战环境中具有多个平行战场，通过多进程和多线程管理调度各空战环境，可以成百上千地提升空战环境与智能体交互所产生的过程数据，加快智能体的性能提升。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

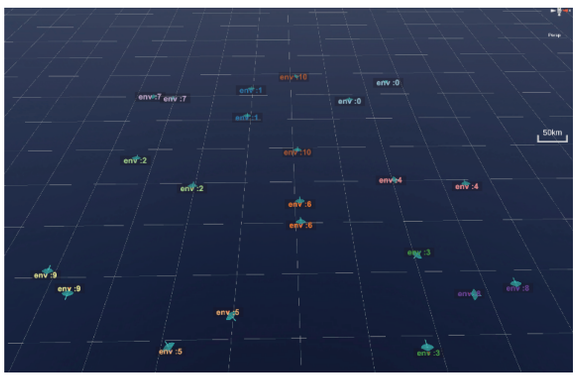


图 11 分布式平行战场
Fig. 11 Distributed parallel battlefield

4.4 实验仿真与分析

4.4.1 课程学习实验

课程学习实验中一共给智能体设置了三个阶段的课程—攻击、防御和综合三类课程。在学习攻击的课程中，使用靶机和智能体对抗。通过控制智能体和靶机对抗博弈，充分鼓励智能体与蓝方无人机对抗，给以积极的鼓励。让智能体学会在合适的时间和位置发射导弹。通过一系列的性能指标来观测智能体的状态。

图 12 为攻击课程的总奖励曲线。总奖励的计算是由经验池采样的 120 条序列中 16 个动作所组成的。随着智能体的不断训练，所获取的奖励一直在提升。由于智能体对抗的是靶机，靶机并没有攻击能力，所以初始奖励的基准就较为可观。

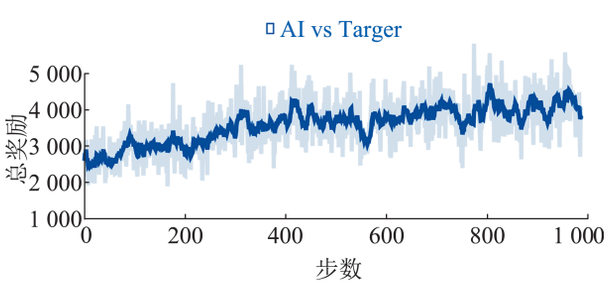


图 12 攻击课程总奖励
Fig. 12 Total reward for attack curriculum

通过观察图 13 的价值网络的损失函数，可以看出当前智能体已经收敛。为了能够合理评估智能体的实时性能，通过统计每 100 场实时对抗蓝机的实时胜率来评估，由图 14 可知，智能体的实时胜率已经达到了 93% 左右，因为随机初始化红蓝双机位置，对于一些初始化比较远的蓝机，雷达无法扫描到，所以很难达到 100% 的胜率。

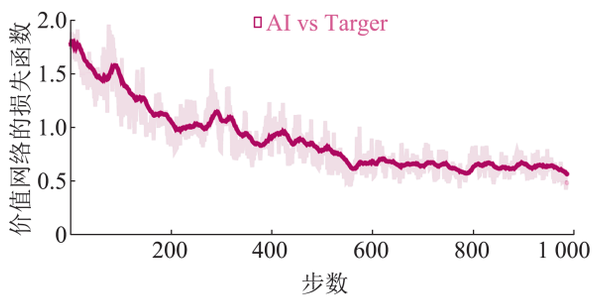


图 13 价值网络的损失函数
Fig. 13 Loss function of value network

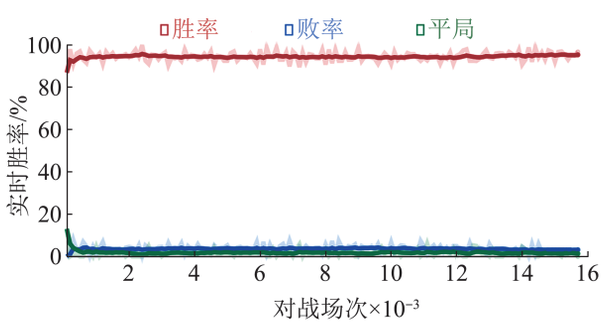


图 14 对抗靶机的实时胜率
Fig. 14 Real-time win rate against target machines

在实际对抗场景中，智能体所控制的红机充分展现了攻击性，并且倾向于大量发射导弹，如图 15 所示。因为在实际对抗过程中，靶机不会躲避来袭导弹，所以智能体的攻击容易命中蓝方。靶机不具备攻击性，所以智能体通常不会做出躲避的动作。为了能够培养智能体的防御意识，设计了相应的防御课程。

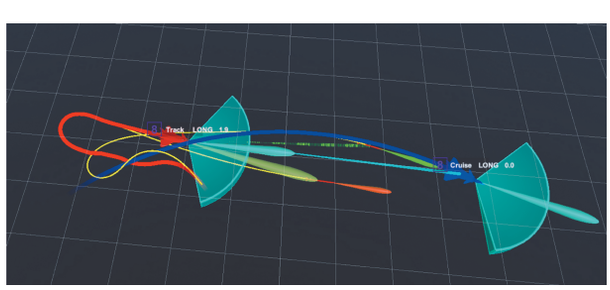


图 15 对抗靶机的实战场景
Fig. 15 Real-world scenarios against target aircraft

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

防御课程是将智能体的攻击关闭，让智能体对抗具有攻击能力的状态机。之所以将智能体的攻击关闭，是因为经过进攻课程之后，其他动作选择的概率都已经较小。为了能够选择更多样化的动作，网络输出中攻击动作关闭，让智能体专注于躲避蓝方导弹。通过一定时间的训练，智能体成功地学会了躲避蓝方导弹。在图16中，智能体通过复杂的机动躲避了状态机大量的导弹。

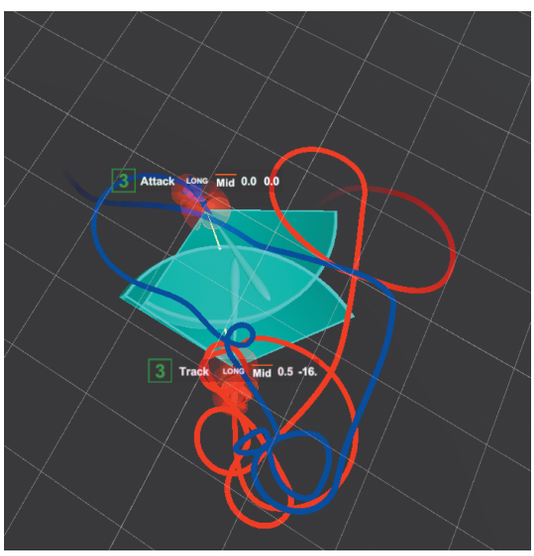


图 16　防御课程实战
Fig. 16　Defense curriculum practice

当智能体学会躲避蓝方导弹时，就需要给智能体设置综合的课程。将智能体的进攻动作开放，让智能体和状态机进行全面的对抗。如图17所示，最终智能体能够以58%的胜率、36%的败率和6%的平局率，以较大的优势击败专家系统。

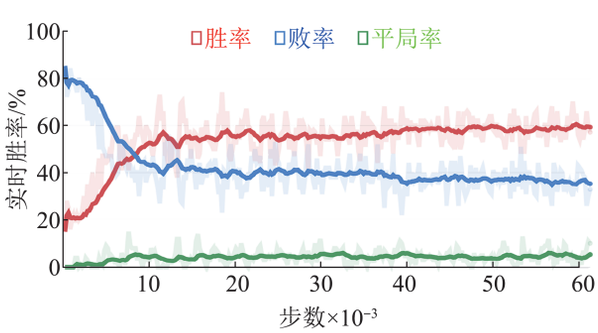


图 17　综合课程的实时胜率
Fig. 17　Real-time win rate for integrated curriculum

在进行综合课程的训练中，智能体需要学习合理的平衡所学习的进攻和防御之间的关系。从图18可知，随着综合课程的不断训练，智能体在与状态机的对抗中逐渐地取得了优势，奖励也从负转正。

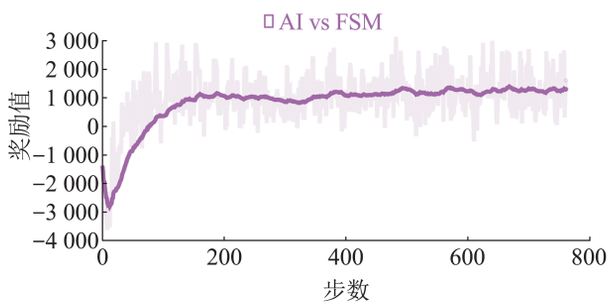


图 18　综合课程的总奖励
Fig. 18　Total award for integrated curriculum

4.4.2 算法有效性检验

为了能够充分对比各类算法之间的性能差距，使用了专家系统–有限状态机作为基准算法，使各类主流强化学习算法与专家系统进行1v1空战仿真对抗，以各个算法收敛之后与专家系统对抗的最终胜率作为性能指标，具体各类算法[24-25]的实际性能如表3所示。

表 3　算法性能基准对比
Table 3　Results of each algorithm against FSM　%

算法	胜率	失败率	平局率
DQN	35	64	1
SAC	43	53	4
DDPG	40	55	5
DPPO	47	50	3
CLDPPO	**58**	**36**	**6**

由表3可知，本文算法CLDPPO的性能优于当前主流的算法和专家系统。从实际胜率分析，仍然具备一定提升的空间。CLDPPO的超参数具体设置如表4所示。

智能体经过一系列的课程之后，演化出了一定的战术能力，并且能够自主学习到一些常用的空战战术。经过综合课程之后智能体的对抗表现，如图19所示。在图19(a)中，智能体采用了车轮战

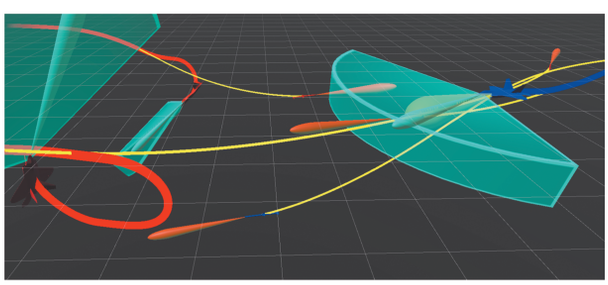
https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

Zhu et al.: Curriculum Learning-based Simulation of UAV Air Combat Under Spar

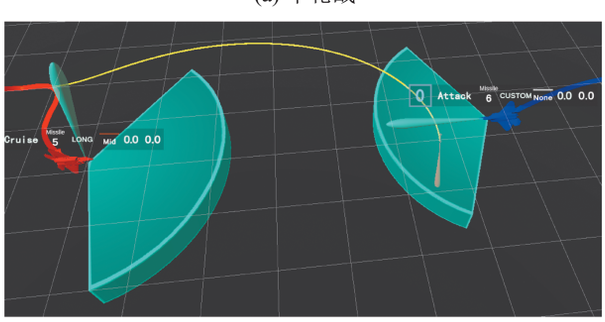
的战术，向蓝机发射导弹之后，使用躲避机动来面对蓝方来袭导弹。在图19(b)中，智能体通过使用风筝战的战术，边打边进行高速的位置切换，使蓝机难以命中本机。

表 4　CLDPPO的超参数设计
Table 4　Hyperparameter design of CLDPPO

参数名	参数值
用于优化的GPU数量	5
批次数量	120
经验库容量	480
衰减系数	0.99
GAE λ	0.95
PPO 裁剪系数	0.2
数据复用率	1
模型版本替换频率	2
Actor 学习率	0.000 5
Critic 学习率	0.000 2
动作的熵系数	0.01



(a) 车轮战



(b) 风筝战

图 19　战术表现
Fig. 19　Tactical performance

4.4.3 消融实验

对1.4节中所设计的态势奖励，每次移除其中一个奖励，并进行课程学习对抗直至算法完全收敛。通过对比收敛的智能体与状态机对抗的最终胜率，得到了如表5所示的结果。

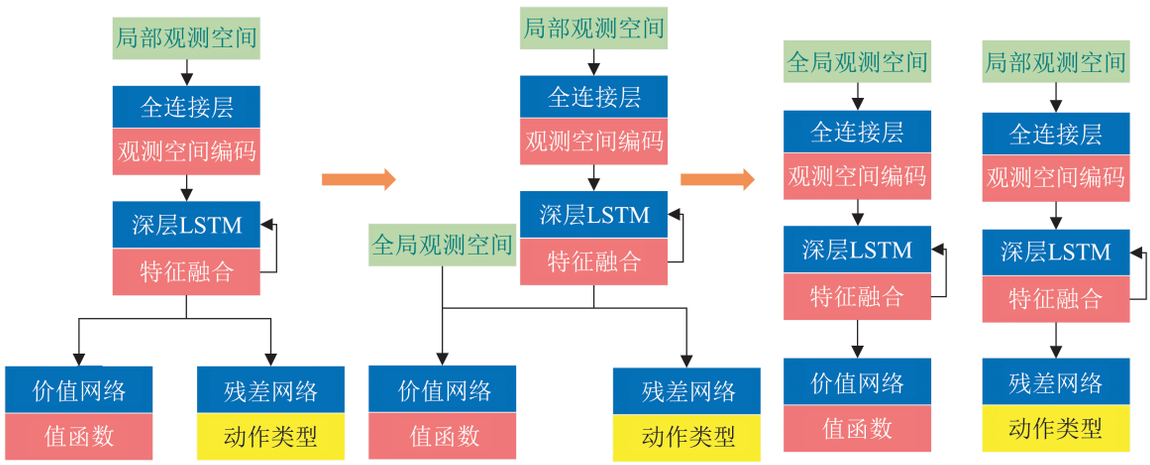
表 5　关键奖励消融实验
Table 5　Ablation experiments for key rewards　%

算法	胜率	失败率	平局率
本文算法	**58**	**36**	**6**
−视野奖励	56	37	7
−失速	55	38	7
−侧滑角	55	39	6
−躲避导弹	54	40	6
−近距离导弹	53	40	7
−姿态优势	51	44	5
−导弹威胁奖励	48	50	2

由表5可知，所添加具备专家知识的奖励都能够带来正向的增益。由于奖励的稀疏性，部分奖励对于胜率的影响相对较小，但是在整体的实战表现可以看出，所添加对应专家知识的奖励一定程度上改善了机动的表现。在空战对抗过程中，添加了导弹近距离躲避蓝方导弹奖励的智能体通常能够更好地躲避蓝方的导弹。节点事件奖励的加入，能够使智能体感知到节点事件所带来的影响，并针对这些事件结合当前的观测量做出合理的机动决策。添加包含专家知识奖励的智能体，能够一定程度地提升智能体整体的对抗性能，有效缓解奖励稀疏的问题。

对于网络结构的设计进行了一定的对比实验，网络结构如图20所示。网络结构A是传统的演员−评论家的结构。在网络结构B中给价值网络添加了全局观测量帮助进行特征提取，但是未通过深度LSTM网络提取时序特征。最终，提出全局观测空间和局部观测空间完全分离的演员−评论家结构C。由表6可知，对演员−评论家进行网络结构上的分离，并将全局观测添加LSTM网络的网络结构C能够产生更优的无人机自主决策性能。

Published by Journal of System Simulation, 2024



(a) 网络结构A　　(b) 网络结构B　　(C) 网络结构C

图 20　网络结构对比实验

Fig. 20　Comparative experiment of network structure

表 6　不同网络结构的性能

Table 6　Performance of different network architectures

%

网络结构	胜率	失败率	平局率
A	50	45	5
B	54	41	5
C	**58**	**36**	**6**

5　结论

本文旨在研究基于单一无人机对抗的自主决策能力，并提出一种基于课程学习的分布式近似策略优化算法，以解决传统强化学习中探索能力不足和奖励稀疏的问题。通过采用定制化阶段性课程的方法，让智能体从简单课程开始学习，不断使用定制化的课程以提高智能体的作战能力并达到阶段性突破。为了更好地模拟真实空战对抗情境，利用数字孪生技术建立了一个接近真实空战场景的模拟环境。同时，提出了一个基于分布式近端策略优化的空战代理框架，该框架充分利用了分布式架构的特点，提出了平行战场的概念，使多个对抗环境同时存在，能够成倍地加速智能体的收敛。

为了评估算法的有效性，采用专家系统作为基准，并利用智能体与专家系统进行对抗。同时，使用主流的强化学习算法作为对比，实验结果表明：本文算法优于当前主流强化学习方法。最终智能体能够以58%的胜率、36%的败率和6%的平局率，以较大的优势击败专家系统，并能够自主学习有效的空战战术。此外，消融实验证明，本文使用的奖励设计和网络结构改进能够显著提升算法的效果。

基于课程学习的空战对抗博弈是空战强化学习研究的改进与拓展，虽然取得了一定的效果，但与算法落地仍然具有一定的距离。后续开放动作集、改良奖励函数，以及结合先进的分层强化学习，让智能体具有更强的作战能力，是今后进一步研究的方向。

参考文献：

[1] 孙智孝, 杨晟琦, 朴海音, 等. 未来智能空战发展综述[J]. 航空学报, 2021, 42(8): 28-42.
Sun Zhixiao, Yang Shengqi, Piao Haiyin, et al. A Survey of Air Combat Artificial Intelligence[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 28-42.

[2] Burgin G H, Owens A J. An Adaptive Maneuvering Logic Computer Program for the Simulation of One-to-one Air-to-air Combat. Volume 2: Program Description [EB/OL]. (1975-09-01) [2020-10-02]. https://ntrs. nasa. gov/citations/197500247 17.

[3] Geng Wenxue, Kong Fan'e, Ma Dongqian. Study on Tactical Decision of UAV Medium-range Air Combat[C]// The 26th Chinese Control and Decision Conference

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349

Zhu et al.: Curriculum Learning-based Simulation of UAV Air Combat Under Spar

(2014 CCDC). Piscataway, NJ, USA: IEEE, 2014: 135-139.

[4] Li Shouyi, Chen Mou, Wang Yuhui, et al. Air Combat Decision-making of Multiple UCAVs Based on Constraint Strategy Games[J]. Defence Technology, 2022, 18(3): 368-383.

[5] Li Weihua, Shi Jingping, Wu Yunyan, et al. A Multi-UCAV Cooperative Occupation Method Based on Weapon Engagement Zones for Beyond-visual-range Air Combat[J]. Defence Technology, 2022, 18(6): 1006-1022.

[6] 左家亮, 杨任农, 张滢, 等. 基于启发式强化学习的空战机动智能决策[J]. 航空学报, 2017, 38(10): 212-225.
Zuo Jialiang, Yang Rennong, Zhang Ying, et al. Intelligent Decision-making in Air Combat Maneuvering Based on Heuristic Reinforcement Learning[J]. Acta Aeronautica et Astronautica Sinica, 2017, 38(10): 212-225.

[7] Liu Pin, Ma Yaofei. A Deep Reinforcement Learning Based Intelligent Decision Method for UCAV Air Combat[C]//Modeling, Design and Simulation of Systems. Singapore: Springer Singapore, 2017: 274-286.

[8] Yang Qiming, Zhu Yan, Zhang Jiandong, et al. UAV Air Combat Autonomous Maneuver Decision Based on DDPG Algorithm[C]//2019 IEEE 15th International Conference on Control and Automation (ICCA). Piscataway, NJ, USA: IEEE, 2019: 37-42.

[9] Lei Xie, Dali Ding, Zhenglei Wei, et al. Moving Time UCAV Maneuver Decision Based on the Dynamic Relational Weight Algorithm and Trajectory Prediction[J]. Mathematical Problems in Engineering, 2021, 2021: 6641567.

[10] Pope A P, Ide J S, Daria Mićović, et al. Hierarchical Reinforcement Learning for Air-to-air Combat[C]//2021 International Conference on Unmanned Aircraft Systems (ICUAS). Piscataway, NJ, USA: IEEE, 2021: 275-284.

[11] Crumpacker J B, Robbins M J, Jenkins P R. An Approximate Dynamic Programming Approach for Solving an Air Combat Maneuvering Problem[J]. Expert Systems with Applications, 2022, 203: 117448.

[12] 曾贲, 房霄, 孔德帅, 等. 一种数据驱动的对抗博弈智能体建模方法[J]. 系统仿真学报, 2021, 33(12): 2838-2845.
Zeng Ben, Fang Xiao, Kong Deshuai, et al. A Data-driven Modeling Method for Game Adversity Agent[J]. Journal of System Simulation, 2021, 33(12): 2838-2845.

[13] 赵毓, 郭继峰, 颜鹏, 等. 稀疏奖励下多航天器规避决策自学习仿真[J]. 系统仿真学报, 2021, 33(8): 1766-1774.
Zhao Yu, Guo Jifeng, Yan Peng, et al. Self-learning-based Multiple Spacecraft Evasion Decision Making Simulation Under Sparse Reward Condition[J]. Journal of System Simulation, 2021, 33(8): 1766-1774.

[14] Piao Haiyin, Sun Zhixiao, Meng Guanglei, et al. Beyond-visual-range Air Combat Tactics Auto-generation by Reinforcement Learning[C]//2020 International Joint Conference on Neural Networks (IJCNN). Piscataway, NJ, USA: IEEE, 2020: 1-8.

[15] 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.

[16] Kong Weiren, Zhou Deyun, Yang Zhen. Air Combat Strategies Generation of CGF Based on MADDPG and Reward Shaping[C]//2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL). Piscataway, NJ, USA: IEEE, 2020: 651-655.

[17] McGrew J S, How J P, Williams B, et al. Air-combat Strategy Using Approximate Dynamic Programming[J]. Journal of Guidance, Control, and Dynamics, 2010, 33 (5): 1641-1654.

[18] 周文卿, 朱纪洪, 匡敏驰, 等. 基于预知博弈树的多无人机群智协同空战算法[J]. 中国科学(技术科学), 2023, 53 (2): 187-199.
Zhou Wenqing, Zhu Jihong, Kuang Minchi, et al. Multi-UAV Cooperative Swarm Algorithm in Air Combat Based on Predictive Game Tree[J]. Scientia Sinica (Technologica), 2023, 53(2): 187-199.

[19] Bengio Y, Jérôme Louradour, Collobert R, et al. Curriculum Learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2009: 41-48.

[20] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2021-01-21]. https://arxiv.org/abs/1707.06347.

[21] Schulman J, Moritz P, Levine S, et al. High-dimensional Continuous Control Using Generalized Advantage Estimation[EB/OL]. (2018-10-20) [2021-02-22]. https://arxiv.org/abs/1506.02438.

[22] Berner C, Brockman G, Chan B, et al. Dota 2 with Large Scale Deep Reinforcement Learning[J]. (2019-12-13) [2021-12-03]. https://arxiv.org/abs/1912.06680.

[23] 周文卿, 朱纪洪, 匡敏驰. 一种基于群体智能的无人空战系统[J]. 中国科学(信息科学), 2020, 50(3): 363-374.
Zhou Wenqing, Zhu Jihong, Kuang Minchi. An Unmanned Air Combat System Based on Swarm Intelligence[J]. Scientia Sinica(Informationis), 2020, 50

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 6, Art. 18

(3): 363-374.

[24] Silver D, Lever G, Heess N, et al. Deterministic Policy Gradient Algorithms[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2014: 387-395.

[25] Haarnoja T, Zhou A, Abbeel P, et al. Soft Actor-critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C]//Proceedings of the 35th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2018: 1861-1870.

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/18
DOI: 10.16182/j.issn1004731x.joss.23-0349