航 空 学 报
Acta Aeronautica et Astronautica Sinica

Sep. 15 2024 Vol. 45 No. 17
ISSN 1000-6893 CN 11-1929/V

http://hkxb.buaa.edu.cn hkxb@buaa.edu.cn

引用格式：李佐龙，朱纪洪，匡敏驰，等. 基于混合动作的空战分层强化学习决策算法[J]. 航空学报，2024，45(17)：530053. LI Z L，ZHU J H，KUANG M C，et al. Hierarchical decision algorithm for air combat with hybrid action based on deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica，2024，45(17)：530053 (in Chinese). doi：10.7527/S1000-6893.2024.30053

高性能无人机关键技术专刊

基于混合动作的空战分层强化学习决策算法

李佐龙[1]，朱纪洪[1,*]，匡敏驰[1]，张杰[2]，任洁[2]

1. 清华大学 精密仪器系，北京 100084
2. 航空工业成都飞机设计研究所，成都 610091

摘 要：智能空战是世界主要军事强国的研究热点。为解决超视距空战博弈机动决策问题，提出了基于深度强化学习的超视距空战分层决策算法。在该决策算法中，使用适合于超视距空战的机动动作集，对飞机的航迹和姿态进行控制。为了扩大模型的动作空间，提升模型的决策能力，将空战的动作空间进行分层，建模为多维离散的动作空间。针对空战中稀疏奖励的问题，设计了一套综合考虑位置优势、武器发射和武器威胁等要素的奖励函数，用于引导智能体向最优策略收敛。搭建了完整的数字孪生空战仿真环境和空战专家系统，在仿真环境中训练决策算法，并通过与专家系统的对抗，对决策算法进行评估。实验结果表明：决策算法具备超视距空战自主决策的能力，能够根据战场态势，进行灵活的机动决策，在与专家系统对抗的过程中具有一定的优势。

关键词：超视距空战；智能决策；深度强化学习；近端策略优化；机动动作；分层决策

中图分类号：V249.4 **文献标识码**：A **文章编号**：1000-6893(2024)17-530053-18

空战是现代战争的重要环节，传统空战主要依靠飞行员的驾驶技术和空战经验。随着无人机技术的快速发展，无人机在军事领域得到了越来越多的应用，如执行中继通信、诱饵欺骗、侦查探测、目标打击等任务[1]。现阶段，军用无人机正朝着智能化方向发展。2016年，辛辛那提大学的Ernest等使用遗传模糊树的技术，设计了空战智能体ALPHA[2]。2020年，美国国防部举办AlphaDog-Fight比赛，在人机战斗中，苍鹭公司的算法5∶0击败了F-16教官Banger[3]。2023年8月，美国实现人工智能算法操纵XQ-58A隐身无人机实现自主飞行[4]。在未来，无人机空战将深刻改变空战形态，智能空战将会在未来成为现实。

根据交战距离，可以将空战划分为视距内空战和超视距空战。视距内空战中，交战双方距离较近，飞行员通过肉眼观察，即可确认对方位置，交战模式以近距格斗为主。超视距空战中，双方飞机距离较远，使用传感器等探测设备发现和锁定对方，由于距离很远，飞机的位置优势不明显，武器发射的条件较容易满足，经常出现多回合、周期性的进攻。目前，由于航空电子技术的快速发展，飞机的探测范围持续扩大，近距空战已不再是空战的主要形式，超视距空战已逐步成为现代空战的主要模式[5]。

智能空战基于军事理论中的OODA模型(观察、判断、决策、执行)[6]，可以建模为复杂的博弈过程。空战决策算法主要包括3类：基于博

收稿日期：2024-01-02；退修日期：2024-01-11；录用日期：2024-04-22；网络出版时间：2024-04-26 14:53
网络出版地址：https://hkxb.buaa.edu.cn/CN/Y2024/V45/I17/530053
* 通信作者. E-mail：jhzhu@tsinghua.edu.cn

弈理论的方法、基于优化理论的方法和基于人工智能的方法。基于博弈的方法主要包括矩阵博弈[7]、微分博弈[8]等，这些方法求解复杂度较高，难以应用到实际中。基于优化理论的方法将空战建模为序贯决策问题，使用贝叶斯网络[9]、遗传模糊树[2]等优化算法进行求解。最早基于人工智能的方法是空战专家系统，1975年Burgin和Owens[10]为NASA编写了名为自适应机动逻辑（Adaptive Maneuvering Logic，AML）的空战决策软件，主要使用if-else构建空战专家系统。Hubert针对基于专家系统的空战决策系统，设计了一种机动选择辅助系统。但是，专家系统需要大量的if-else语句刻画空战问题，依赖固定的空战规则，泛化能力较差，不具备自主学习的能力。

随着人工智能技术的快速发展，目前大多数研究者使用人工智能解决无人机空战自主决策问题。监督学习需要大量由专家标记过的空战数据，但获取这些空战数据集难度大且代价高，难以实现。此外，监督学习方法很可能出现过拟合和泛化能力不足的问题，导致训练得到的智能体难以应对复杂多变的战场环境[11]。因此，使用监督学习解决空战问题存在较大局限性。

近年来，强化学习在博弈类的游戏中取得了巨大的成功。Mnih等[12]使用Deep Q-Network算法，训练得到针对Atari游戏的智能体，达到了专业玩家的水平。Silver等[13]使用强化学习算法和蒙特卡洛树搜索方法，训练得到围棋智能体AlphaGo，击败了世界冠军。OpenAI团队针对Dota2开发了智能体OpenAI Five，以2∶0击败了人类冠军团队[14]。目前越来越多的研究者使用强化学习解决空战自主决策算法。Lockheed Martin公司使用分层强化学习算法，训练得到空战智能体PHANG-MAN，在AlphaDogFight比赛中击败F-16飞行教官Banger[3]。Sun等[11]基于PPO算法，提出了针对多智能体空战的决策算法MAHPG，解决了空战分层决策中的混合动作问题。章胜等[15]针对近距空战问题，提出了基于TD3算法和优先经验回放技术的空战决策算法，在虚拟仿真环境中进行人机对抗实验，并进行飞行试验，实现了智能空战算法从计算机仿真到真实飞行的迁移。张建东等[16]基于分层强化学习，并结合SAC算法和专家经验，建立空战的元策略决策模型，对Option-Critic算法进行改进，实现了空战中策略的无缝切换。邱妍等[17]针对无人机近距空战自主决策问题，基于近端策略优化（Proximal Policy Optimization，PPO）算法，针对距离、角度、速度等进行奖励塑造，分别对全连接网络和长短期记忆网络进行仿真训练。钱殿伟等[18]基于PPO算法，提出了基于双重观测和复合奖励的空战决策算法，在专家系统和矩阵博弈对手两类实验场景中对算法进行评估。

从公开发表的文献来看，目前大多数空战智能决策研究仅针对近距空战，对超视距空战的研究较少，大多数研究仅建立飞机的三自由度方程，没有考虑飞机的姿态信息，大多数研究使用NASA提出的基本动作集[7]，对实际空战中的常用机动考虑较少。

本文针对单机超视距空战决策问题，定义适用于超视距空战的机动动作集，控制飞机实现相应的机动动作。基于PPO算法，提出了自回归多维离散近端策略优化算法（Auto-Regressive Multi-discrete Proximal Policy Optimization，ARM-PPO），将动作空间进行分层，选择机动动作和相应的参数，并采用自回归的结构进行实现。将参数空间离散化，得到多维离散的动作空间。相较于固定参数，让算法选择参数可以扩大智能体决策范围，与连续参数相比，离散参数的设定又减小了智能体的探索空间，有利于智能体学习最优策略。同时，针对超视距空战问题，设计奖励函数，对智能体进行引导。搭建完整的空战仿真环境，算法训练的智能体在仿真环境中与专家系统进行对抗，对算法中的超参数取值进行讨论。并针对奖励函数和网络结构设计消融实验，验证了设计的奖励和网络结构的有效性。仿真实验表明，训练得到的智能体，具备超视距作战的能力，可以与专家系统相抗衡。

1 超视距空战问题建模

根据飞机所受的升力、阻力、重力和推力，在机体坐标系内建立飞机的线运动方程，计算得到飞机的加速度为

$$\begin{cases}\dot{u}=vr-wq-g\sin\theta+\dfrac{F_x}{m}\\ \dot{v}=-ur+wp+g\cos\theta\sin\phi+\dfrac{F_y}{m}\\ \dot{w}=uq-vp+g\cos\theta\cos\phi+\dfrac{F_z}{m}\end{cases}\tag{1}$$

式中：u、v、w为飞机速度$\boldsymbol{V}$在机体系x、y、z轴上的分量；p、q、r为飞机角速度$\boldsymbol{\Omega}$在机体系x、y、z轴上的分量；θ、ϕ分别为飞机的俯仰角和滚转角；F_x、F_y、F_z为飞机所受升力、阻力和发动机推力的合力在机体系坐标轴上的投影；m为飞机的质量。

在机体坐标系下建立飞机的角运动方程，计算得到飞机的角加速度为

$$\begin{cases}\bar{L}=\dot{p}I_x-\dot{r}I_{xz}+qr(I_z-I_y)-pqI_{xz}\\ M=\dot{q}I_y+pr(I_x-I_z)+(p^2-r^2)I_{xz}\\ N=\dot{r}I_z-\dot{p}I_{xz}+pq(I_y-I_x)+qrI_{xz}\end{cases}\tag{2}$$

式中：I_x、I_y、I_z为飞机绕机体系x、y、z轴的转动惯量；I_{xz}为飞机对x、z轴的惯性积；$\bar{L}$、M、N为飞机所受合力矩在机体系x、y、z轴的分量。根据飞机的运动学方程，计算飞机的位置和姿态。

2 深度强化学习

2.1 强化学习基本理论

强化学习(Reinforcement Learning, RL)是人工智能的重要组成部分。在强化学习算法中，智能体(Agent)不断与环境交互，在试错中学习最优策略。强化学习的数学基础是马尔科夫决策过程(Markov Decision Process, MDP)。马尔科夫决策过程包括五元组：$<\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma>$，其中$\mathcal{S}$为状态空间，$\mathcal{A}$为动作空间，$\mathcal{P}$为状态转移概率，$\mathcal{R}$为奖励函数，$\gamma$为折扣因子。

强化学习的目标是最大化期望累计回报：

$$J(\pi_\theta)=\mathbb{E}_{\pi_\theta}\left[\sum_{k=0}^{\infty}\gamma^k R_k\right]\tag{3}$$

式中：R_k为智能体在k时刻获得的奖励。在强化学习中，策略$\pi(a|s)$指在状态s下，选择动作a的概率，表达式为

$$\pi(a|s)=P(A_t=a|S_t=s)\tag{4}$$

强化学习算法使用价值函数来评价策略的优劣，包括状态价值函数$V_\pi(s)$和行动价值函数$Q_\pi(s,a)$，定义为

$$V_\pi(s)=\mathbb{E}_\pi[G_t|S_t=s]\tag{5}$$

$$Q_\pi(s,a)=\mathbb{E}_\pi[G_t|S_t=s,A_t=a]\tag{6}$$

式中：G_t为1幕的累积回报，计算式为

$$G_t=\sum_{k=t}^{\infty}\gamma^k R_k\tag{7}$$

可以将强化学习算法分为基于策略(Policy-Based)的方法和基于价值(Value-Based)的方法[19]，基于价值的方法通过值函数间接获取最优策略，包括策略迭代、价值迭代、Sarsa、Q-learning[19]、Deep Q-network[12]等，基于策略的方法包括Vanilla PG[20]、TRPO[21]、PPO[22]、SAC[23]等，介于二者之间的算法是DDPG[24]和TD3[25]。

2.2 PPO算法

PPO[22]是基于策略梯度的算法的一种，是目前最常使用的强化学习算法之一。PPO算法在TRPO算法的基础之上，优化如下的目标函数：

$$\mathcal{L}^{\mathrm{CLIP}}(\theta)=\mathbb{E}_t[\min\{r_t(\theta)\hat{A}_t,\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t\}]\tag{8}$$

式中：$r_t(\theta)$为由于重要性采样引入的新旧策略的概率比，计算方法为

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\mathrm{old}}(a_t|s_t)}\tag{9}$$

$A_\pi(s,a)$为优势函数，表示在当前状态s下，选择动作a的价值比平均值高出多少，计算式为

$$A_\pi(s,a)=Q_\pi(s,a)-V_\pi(s)\tag{10}$$

由于优势函数的真值未知，因此在实际中，通常使用广义优势估计(Generalized Advantage Estimation, GAE)[26]对动作优势函数进行估计：

$$\hat{A}_t=\sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V\tag{11}$$

式中：δ_{t+l}^V为一步时序差分误差，计算式为

$$\delta_{t+l}^V=r_t+\gamma V(s_{s+1})-V(s_t)\tag{12}$$

为了增强智能体对于未知策略的探索能力，在PPO算法中加入策略熵项[27]。策略熵的定义为

$$\mathcal{H}(\pi)=\mathbb{E}_{\pi}[-\log\pi(a|s)] \tag{13}$$

增加策略熵的优化函数为

$$J_{\theta}=\mathcal{L}_{\theta}^{\mathrm{CLIP}}+\beta_{\mathrm{E}}\mathcal{H}(\pi_{\theta}) \tag{14}$$

式中：β_{E}为温度系数。

3 超视距空战智能决策算法设计

在武器使用和机动动作方面，超视距空战和视距内空战有较大的差异。在武器使用方面，视距内空战使用航炮和近距格斗弹，而超视距空战使用中远程空空导弹。在机动动作方面，视距内空战强调占据位置优势，经常出现大过载机动，战斗节奏较快，飞行员需要准确把握时机。而超视距空战节奏较慢，对机动性的要求较低，同时由于双方距离较远，位置优势不明显。针对超视距空战的主要特征，在算法设计中进行体现，得到针对超视距空战的智能决策算法。

3.1 状态空间

根据空战知识，采用特征工程，引入空战领域知识，设计可以反映空战战场状态和特征的量，构建如图1所示空战的状态空间。根据经验和领域知识构造的特征，有利于智能体的学习，虽然会损失一些信息，但对空战决策的影响可以忽略。

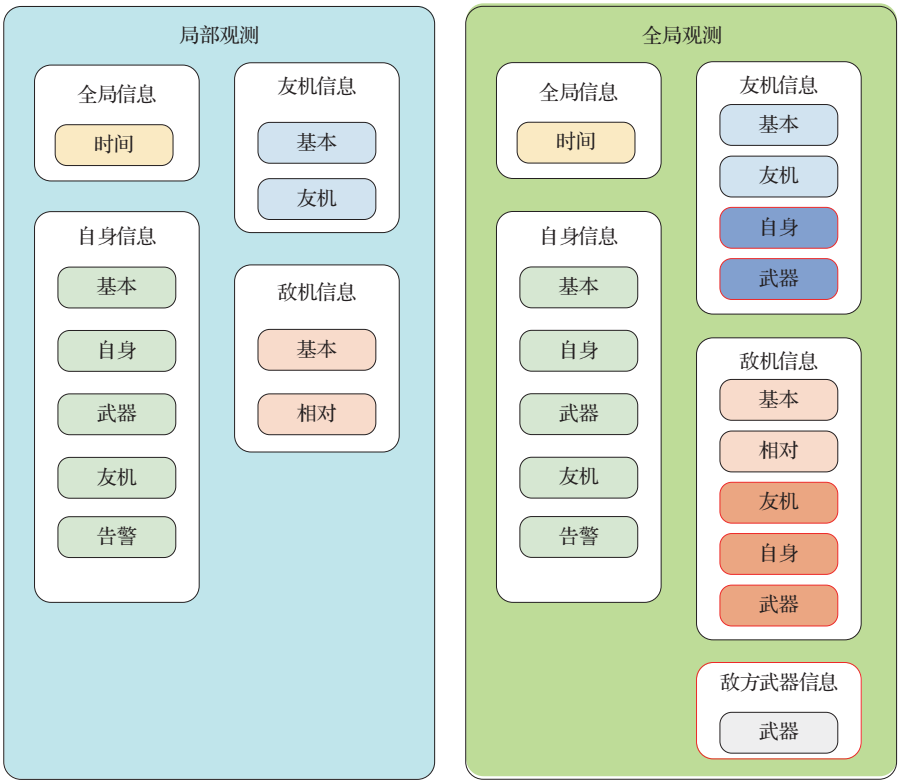


图 1 空战状态空间

Fig. 1 State space in air combat

空战是典型的部分观测马尔科夫决策过程，通常每架飞机只能获得局部观测，无法获得环境的全部状态。因此将空战的观测分为局部观测和全局观测两部分[28]，如图1所示。每个智能体获得的局部观测，包括飞机自身信息、友机信息和通过探测设备获得的敌方飞机信息。空战决策算法使用了Actor-Critic算法，算法中的Critic需要准确评估战场态势，即需要准确拟合值函数，因此使用全局观测。全局观测比局部观测多了敌方飞机自身和敌方武器的信息，

这些信息通过本机的传感器无法获得，只存在于全局信息之中。

3.2 动作空间

在使用强化学习解决空战问题的过程中，对空战动作空间的建模非常重要。经典强化学习算法的动作空间包括离散动作空间、连续动作空间、多维离散动作空间、混合动作空间等类型。一般来说，对空战的动作空间进行建模，有以下方法：

空战的动作空间具有连续的特性，常见的方法是使用飞机操纵杆和油门杆的偏移量作为控制量。Lockheed Martin公司在AlphaDogFight比赛中设计的智能体PHANG-MAN，使用飞机操纵杆作为动作参数，决策算法输出升降舵、方向舵、副翼和油门杆的偏移量[3]。文献[29]中，决策变量同样是操纵杆的偏移量。另一种方法是将飞机看作质点，建立飞机的三自由度方程，根据三自由度方程选择控制变量。文献[17-18, 30-31]等选择切向过载n_x、法向过载n_y和滚转角γ作为控制量。由于本文采用飞机的六自由度动力学方程，需要考虑飞机的轨迹和姿态，无法直接使用上述方法。在实验早期，也曾使用端到端的算法，网络直接输出飞机操纵杆和油门杆的偏移量。这样的设定增加了问题的难度，意味着强化学习在学习空战决策的同时，还要学习飞行控制，没有将已有的飞行控制相关的知识融入到算法中。实验结果表明：算法直接输出飞机操纵杆的偏移量，很难控制飞机平稳飞行，不利于智能体训练，智能体很难学到有效的空战策略。

还有一种方法是将空战动作空间离散化，建模为离散动作空间，使用机动动作集，每次做决策时，在动作集中选择最佳动作。这种方法起源于NASA提出的矩阵博弈算法，包括7种基本动作，分别是匀速、最大加速度加速、最大加速度减速、最大过载左转、最大过载右转、最大过载爬升和最大过载俯冲[7]。该机动动作集实现简单，通过排列组合就可以得到更复杂的机动。杨晟琦等[32]在NASA动作集的基础上，对转弯和爬升机动，按角度进一步细分为不同的动作，增强了动作集的表达能力。但是该动作集考虑的是飞机以最大加速度和最大过载飞行的极限情况，在实际空战过程中，飞机不可能一直按照极限情况飞行。因此，该动作集有一定的局限性。

基于已有研究，并结合超视距空战实际作战情况，设计了如表1所示的针对超视距空战的机动动作集。表1中，机动动作集包括拉起、追踪、攻击、盘旋、急盘旋、筋斗、平飞共7个动作。其中，引入追踪是为了显式定义飞机机头指向敌方的机动，从而避免智能体通过一系列机动拼凑出追踪机动的过程[33]。这样的处理方式降低了问题的复杂度和算法的搜索空间，在实验中取得了良好的效果。

表1 空战机动动作集

Table 1 Maneuver set in air combat

序号	动作	参数
0	拉起	过载，爬升率
1	追踪	速度
2	攻击	速度
3	盘旋	过载，盘旋角速度
4	急盘旋	过载，盘旋角速度
5	筋斗	过载，速度
6	平飞	速度

此外，每个动作都有动作参数，动作参数定量地给出了衡量机动的剧烈程度的指标，包括过载、爬升率、速度、盘旋角速度等。过载主要在筋斗、盘旋等机动中使用，衡量机动的剧烈程度。盘旋角速度是针对盘旋机动定义的，衡量转弯的快慢。爬升率针对拉起等爬升机动使用，衡量爬升的快慢。为了降低问题的复杂度，便于强化学习算法的训练，参考OpenAI针对Dota2开发的智能体OpenAI Five[14]，将所有的参数进行离散化。每次决策时，算法从离散化的参数列表中进行选择，离散化的参数如表2所示。参数离散化的方法是：根据空战经验，确定每个参数可能的取值范围，在取值范围内选择合适的步长，等间隔取值。值得注意的是，若参数的步长过小，智能体动作的准确度会得到提高，可以表示更复杂的动作，但搜索的空间更大，更不容易收敛；若步长过大，智能体更容易搜索到最优策略，但动作的精度较低，可能无法覆盖所有的动作，无法对状态空间进行充分探索。因此，经过反复的比较和尝试，最终确定了步长的取值。

表 2 空战机动动作参数

Table 2 Maneuver parameters in air combat

参数	数值
过载	2, 3, 4, 5, 6, 7
爬升率/(m·s^{-1})	0, 40, 80, 120, 160, 200
速度马赫数	0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4
盘旋角速度/((°)·s^{-1})	0, 5, 10, 15, 20, 25, 30

飞机执行部分机动的过程中,参数可以保持不变,如平飞时可以匀速前进,水平协调转弯可以保持恒定的过载和速度等。在这种情况下,给出目标参数后,飞机的实际参数可以和目标值保持一致。而在飞机执行某些机动的过程中,参数的取值会随时间变化,并且这种变化的规律有时难以给出。在这种情况下,表2给出的参数取值,只是表明参数的大致的范围和上限。

在定义了动作集之后,对飞机的轨迹和姿态进行控制,确保飞机实现相应的机动。飞行控制算法由高度控制模块、速度控制模块、滚转角控制模块等组成,根据每种机动动作的特点,调用相应的模块进行组合,从而控制飞机实现相应机动。

3.3 奖励函数

奖励函数在强化学习算法中起到至关重要的作用。空战问题固有的奖励函数为在战斗结束时,根据战争的结果给出的奖励。但是,这样的奖励过于稀疏,难以引导智能体习得最优策略。使用奖励塑造(Reward Shaping)[34]的技术,通过密集奖励,对智能体进行引导。根据文献[28,35],针对超视距空战,设计奖励函数,如表3所示。

奖励可分为事件奖励和状态奖励。事件奖励是稀疏的,在特定事件发生后才得到奖励。状态奖励是密集的,每个时间步都会计算,包括优势、威胁、速度和侧滑角过大惩罚。其中,优势奖励指的是飞机在位置上的优势,与双方飞机的距离和相对角度有关,其作用在于引导智能体将机头对准地方,追踪敌方飞机,占据位置优势。综合考虑距离和角度因素,结合超视距空战中位置优势不明显的特征,给出优势奖励

表 3 空战的奖励函数

Table 3 Reward function in air combat

类型	名称	数值
事件奖励	命中目标	+100
	平局	−10
	被命中	−100
	坠地	−100
	扫描到敌机	+10
	近距离躲避敌机	+50
	近距离经过敌机	+10
	发射导弹	−6～−2
状态奖励	优势	R_a
	威胁	R_t
	失速	R_s
	侧滑角过大	R_β

的计算方法为

$$A=\begin{cases}30\mathrm{e}^{-\frac{\theta_e^2}{1\,300}}\dfrac{d}{1\,000} & 0\leqslant d\leqslant 1\,000\\ 30\mathrm{e}^{-\frac{\theta_e^2}{1\,300}}\dfrac{39\,000-d}{38\,000} & 1\,000<d\leqslant 20\,000\\ 30\mathrm{e}^{-\frac{\theta_e^2}{1\,300}}\dfrac{10\,000}{d} & d>20\,000\end{cases}$$

$$R_a=A_{end}-A_{start} \tag{15}$$

式中:θ_e为敌机和本机的连线和本机机头方向的夹角;d为双方之间的距离。优势奖励随角度和距离的变化如图2(a)所示。

威胁程度是指对方武器对我方的威胁程度,与剩余时间t和相对方位角θ_a有关,威胁奖励的作用是引导智能体躲避来袭导弹,计算式为

$$T=\begin{cases}-50\mathrm{e}^{-\frac{\theta_a^2}{4\,900}}\dfrac{50-t}{50} & 0\leqslant t\leqslant 50\\ 0 & t>50\end{cases}$$

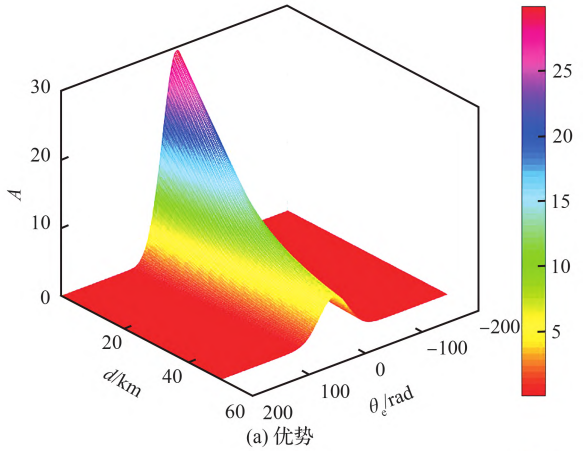
$$R_t=T_{end}-T_{start} \tag{16}$$

威胁奖励随时间和角度的变化如图2(b)所示。

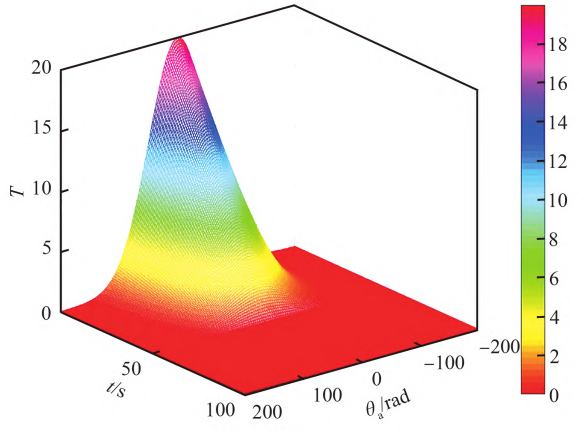
在实际训练过程中,增加了速度惩罚和侧滑角惩罚,计算方法如式(17)和式(18)所示。

$$S=\begin{cases}-20 & v\leqslant 80\\ \dfrac{v-160}{40} & 80<v\leqslant 160\\ 0 & v>160\end{cases}$$

$$R_s=S_{end}-S_{start} \tag{17}$$



(a) 优势



(b) 威胁

图 2 空间中的优势和威胁

Fig. 2 Advantage and threat in air combat

$$B=\begin{cases}-10 & |\beta|>20\\ -\dfrac{\beta^2}{40} & |\beta|\leqslant 20\end{cases}$$

$$R_\beta=B_{\text{end}}-B_{\text{start}} \tag{18}$$

式中：v为速度；β为侧滑角。

考虑到超视距空战中使用导弹作为进攻的武器，在奖励函数中，针对发射导弹设计奖励函数在实验中发现，智能体倾向于一次发射多枚导弹以提高胜率，但这样的做法不符合空战实际。为了让智能体学会使用武器，通过奖励函数的方式对智能体发射导弹进行引导。如表2所示，增加导弹发射的惩罚奖励，对导弹发射进行惩罚，且剩余的导弹数量越少，惩罚的数值越大。这样的设计能够使得智能体节约弹药，不轻易发射武器。此外，还增加了智能体发射导弹的距离和角度约束。这些设定可以使智能体较好地学会使用武器，提高作战能力。

3.4 智能决策算法设计

采用PPO算法结构，状态空间、动作空间、奖励函数如前所述，提出了自回归多维离散近端策略优化(Autoregressive Multi-discrete Proximal Policy Optimization, ARM-PPO)。算法的主要结构如图3所示。ARM-PPO算法保留了PPO算法大部分的设定，采用演员-评论家结构(Actor-Critic)，包括策略网络(Actor)和值网络(Critic)两部分，策略网络输出动作，值网络对当前状态进行评估，输出当前状态值函数的估计值。图4为传统的PPO算法和ARM-PPO算法的对比，与传统的PPO算法相比，ARM-PPO算法在以下方面做出改进：

1) 在策略网络和值网络中增加LSTM网络。由于空战机动决策问题具有序贯决策的特性，而且相较于全连接网络，LSTM更适合处理时间序列，因此使用LSTM处理空战观测量，提取空战时间序列的特征，对应图3中的空战特征提取网络。

2) 采用自回归的结构，将动作空间分为机动层和参数层，算法分别输出机动动作和参数。传统的PPO算法只适用于单一动作空间的决策，即只有连续动作或只有离散动作的情形。针对空战问题需要输出混合动作(包括机动动作和参数)的需求，对策略网络进行改进。策略网络包括机动动作网络和参数网络两部分，如图3所示，机动动作网络输出动作集中每个动作的概率，采样得到动作。参数网络采取类似自回归的结构，以LSTM提取的空战特征和采样得到的机动动作为输入，用子网络分别输出离散化的参数取值的概率，采样得到参数。在ARM-PPO算法中，完整的动作a包括机动动作m和相应参数x，不同动作对应的参数不同，因此参数依赖于机动动作的选取。机动和参数之间存在强相关性，为了强化这种相关性，采用自回归的形式，显式要求机动是参数网络的输入，即

$$x=f(z,m) \tag{19}$$

式中：z为LSTM输出的空战特征；m为采样得到的机动动作；x为机动对应的参数；$f(\cdot)$为参数网络表示的函数。由于机动动作m也是策略网络

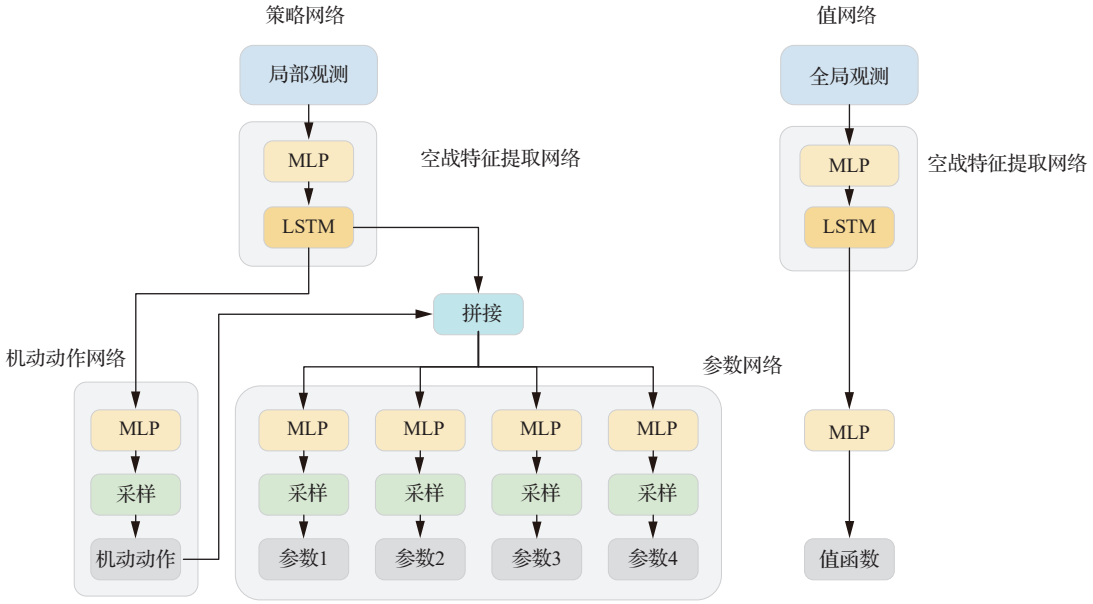


图 3 ARM-PPO算法结构

Fig. 3 ARM-PPO architecture

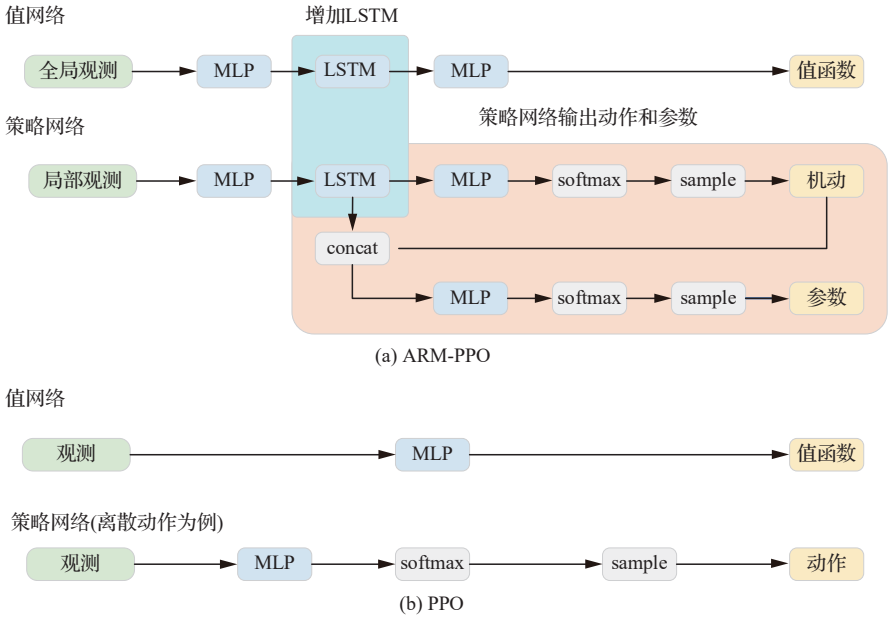


图 4 ARM-PPO算法和PPO算法的比较

Fig. 4 Comparison between PPO and ARM-PPO

的输出,即策略网络以自身的输出为输入,类似于自然语言处理中自回归生成文本的形式,因此借用自回归的概念,称算法具有自回归生成动作的特性。在4.3.4节中,通过对比实验,验证了自回归结构的有效性。

进一步地,认为动作和参数条件独立,满足:

$$P(a)=P(m)P(x|m) \tag{20}$$

根据式(20)计算动作对应的概率密度,进而计算PPO算法中的概率比$\pi_\theta(a|s)/\pi_{\text{old}}(a|s)$。

算法中的多维离散体现在:为了进一步缩小算法搜索的空间,降低动作的复杂度,将动作参数离散化,从离散化的参数列表中选择参数,

如表2所示。由于机动和参数都是离散的，因此称为多维离散的动作空间。将参数空间离散化，智能体的动作精度会有所下降，但是由于提供的参数取值覆盖了空战中大多数情形，因此智能体仍然能够以较高的精度和较大的可能性探索到环境中的不同状态。这样的设定，在基本不影响算法性能的前提下，减小了搜索空间，降低了问题的复杂度，有利于智能体更高效地学习空战策略。通过4.3.4节的对比实验，验证了参数空间离散化有利于智能体更好地搜索最优策略。

在策略网络中，采用了非法动作屏蔽机制[35]，针对进攻和躲避行为增加约束，相当于对算法进行剪枝，避免一些无效的决策，提高了算法的效率。

3.5 算法架构

系统的整体结构参考OpenAI Five[14]的训练框架，主要包括仿真环境（前端）和决策网络（后端），以及经验回放缓冲区，如图5所示。算法采用off-policy的结构，前端和后端进行网络通信，前端将观测发送给后端，后端根据从仿真环境中获得的观测，输出要执行的动作，返回给前端仿真环境进行模拟。在完成一定的决策步骤之后，将观测、动作、奖励等信息进行打包，存入经验回放缓冲区。当经验池中的新旧数据量之比超过阈值之后，混合新旧数据，开始训练过程。算法采用分布式的训练架构，训练过程中，多个优化器并行计算，从经验池中随机采样，更新网络参数。

ARM-PPO算法的流程如算法1所示。在

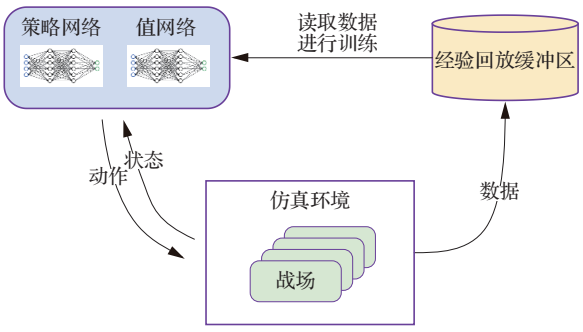


图5 系统结构

Fig. 5 Diagram of the system

每次迭代中，首先产生训练数据。决策算法根据当前局部观测，输出采取的机动动作，仿真环境中的智能体执行机动，进入下一个状态，将状态、动作、奖励等构成的轨迹数据存入Redis构建的经验回放缓冲区中。当缓冲区中新旧数据量之比超过20%时，开启训练进程，从缓冲区中读取数据，计算策略网络和值网络的梯度，更新网络参数，将更新后的网络参数存入Redis搭建的数据库中，用更新后的网络控制智能体决策，产生训练数据。如此往复，直到算法收敛。

算法1 单机空战强化学习算法：ARM-PPO

初始化策略网络参数θ_0，值网络参数ϕ_0；
初始化经验回放缓冲区$\mathcal{D}$；
for episode = 1, 2, …, M **do**
  for 不同空战战场的智能体 = 1, 2, …, n **do**
    if not 动作完成或触发中断机制 **then**
      智能体接收前端仿真环境发来的状态s_t；
      根据状态s_t，策略网络输出动作的概率分布，采样得到动作，$a_t=\text{sample}(\pi_\theta(s_t))$；
      智能体执行动作a_t，得到奖励r_t，进入状态s_{t+1}；
      将数据(s_t, a_t, r_t, s_{t+1})存入经验回放缓冲区$\mathcal{D}$；
    end if
    if 智能体所在战场结束或达到最大仿真时间 **then**
      将该飞机的决策序列分割为长度$l=16$步的若干序列$\tau=\{(s_t, a_t, r_t, s_{t+1})_i\}_{i=1}^{l}$；
      销毁战场内的飞机并随机初始化，生成新的作战场景；
    end if
  end for
  if $\mathcal{D}$中的新旧数据比>0.2 **then**
    从$\mathcal{D}$中批量采样轨迹数据$\mathcal{B}$；
    利用价值网络计算状态价值函数的估计值$V_\phi(s_t)$；
    根据$V_\phi(s_t)$，计算动作优势函数的估计值$\hat{A}(s_t, a_t)$；
    计算概率比$r_t=\dfrac{\pi_\theta(m_t)\pi_\theta(x_t)}{\pi_{\theta_k}(m_t)\pi_{\theta_k}(x_t)}$；
    计算策略网络的目标函数$J_{\text{actor}}(\theta)=\mathcal{L}^{\text{CLIP}}(\theta)+\beta_{\text{E}}\mathcal{H}(\pi_\theta)$；
    更新策略网络参数$\theta\leftarrow\theta+\alpha_a\nabla J_{\text{actor}}(\theta)$；
    计算价值网络的损失函数$J_{\text{critic}}(\phi)=\mathbb{E}_{\mathcal{B}}[r_t+\gamma V_\phi(s_{t+1})-V_\phi(s_t)]^2$；
    更新价值网络参数$\phi\leftarrow\phi-\alpha_c\nabla J_{\text{critic}}(\phi)$；
  end if
end for

4 仿真实验

算法在搭建的数字孪生空战仿真环境中进行训练，通过和专家系统的对抗，对算法性能进

行评估。针对奖励设计消融实验，验证了提出奖励的有效性。针对算法的网络结构，设计对比实验，验证算法设计的合理性。

4.1 仿真环境

使用Unity3D搭建了完整的数字孪生空战仿真环境，在仿真环境中，对提出的算法进行测试。通过对气动力学、电磁、光电等进行模拟计算，实现对空战的仿真。

在仿真环境中，构建了基于规则的空战专家系统[36]，用于与强化学习算法训练得到的智能体进行对抗。专家系统基于有限状态机的思想，根据空战的规则和经验，定义了巡逻、进攻、规避、逃逸、追踪上次目标和防撞地共6个状态，并给出了状态之间的转移关系，空战专家系统的状态转移如图6所示。

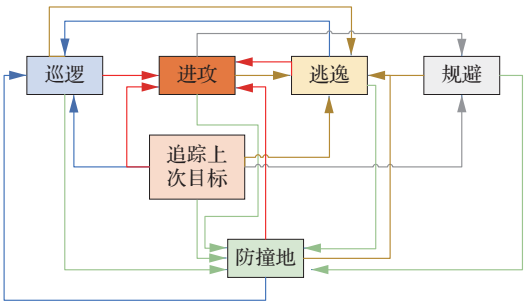


图6 空战专家系统状态转移

Fig. 6 Diagram of state transition of expert system for air combat

图6中，每个方框表示一个状态，状态之间的转移关系用箭头表示，与方框颜色相同的箭头表示从其他状态转移到该状态。状态转移的主要规则为：在没有发现敌机及敌方导弹的情况下，飞机进入巡逻状态；一旦获得敌方信息，飞机进入进攻状态，对敌机进行持续追踪，当满足导弹发射条件时，飞机发射导弹；当有敌方来袭导弹的告警时，飞机进入规避状态，采取合适的机动规避来袭导弹；当飞机的高度过低时，进入防撞地状态，飞机通过跃升提升高度，避免撞地。

目前，空战专家系统发展较为成熟，具有较高的可靠性和可解释性，是强化学习训练有效的基线算法。

4.2 实验设定

在搭建的空战仿真环境中训练算法，仿真环境中包括12个线程，每个线程中包含10个平行战场，每个战场包含红蓝双方各1架飞机，双方采用相同的机型和武器配置，以消除武器装备的影响。每架飞机的生成方式为：以某个固定点为中心，在半径为40 km的圆域内随机生成，双方的初始姿态随机生成。战场范围是半径为150 km的圆形区域，当有飞机被击落或出界或达到最大仿真时长后，重生得到新的战场，继续进行仿真。

表4为算法中采用的超参数。算法将W&B作为可视化工具，每迭代一轮，算法会将这一轮训练的累计奖励等上传到W&B网站上，在线生成训练曲线，并在线统计胜率、负率、平局率、命中率、末制导率等空战评价指标。

表4 算法使用的超参数

Table 4 Hyperparameters of the algorithm

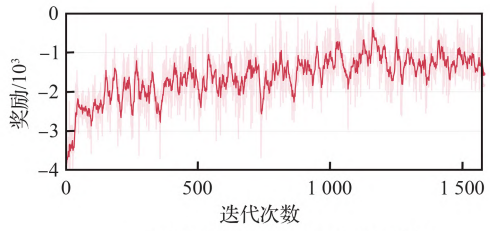
超参数	数值
折扣因子γ	0.99
GAE参数λ	0.95
裁剪系数ε	0.1
策略熵系数β_{E}	0.01
策略网络学习率α_{a}	5×10^{-4}
值网络学习率α_{c}	2×10^{-4}
Batch_Size	120
序列长度	16
经验池样本容量	480

4.3 实验结果与分析

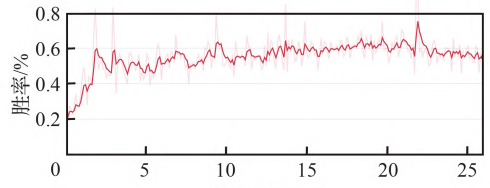
4.3.1 ARM-PPO算法

在仿真环境中，算法控制的智能体在和第4.1节中定义的专家系统对抗的过程中进行训练。经过大约1 500轮迭代后，算法趋于稳定，训练结果如图7所示。

图7(a)为累计奖励随迭代次数的变化，横坐标是迭代次数，图7(b)为胜率随战场数量的变化，横坐标是累计的战场数量，由于采用多线程和平行战场，每一次迭代产生的战场数量较多，因此两条曲线的横坐标不同。其中浅色的线条为原始



(a) 累积奖励随训练迭代次数的变化



(b) 胜率随迭代次数的变化

图 7 累积奖励和胜率随迭代次数的变化

Fig. 7 Variation of reward sum and win rate with training iteration

数据绘制的曲线,而深色线条是通过指数滑动平均后得到的平滑的结果。可以看到,在训练初期,累计奖励增长速度较快,智能体的决策能力得到较大幅度的提升,随着迭代次数进一步增加,累计奖励和胜率逐渐趋于稳定。由于强化学习算法具有一定的随机性,且仿真初始条件的随机性较大,而初始位置对空战结果的影响很大,因此曲线有较大幅度的波动。经过训练,智能体的平均胜率达到 60%,最高胜率超过 80%,算法控制的智能体和专家系统对抗的场景如图 8 所示。

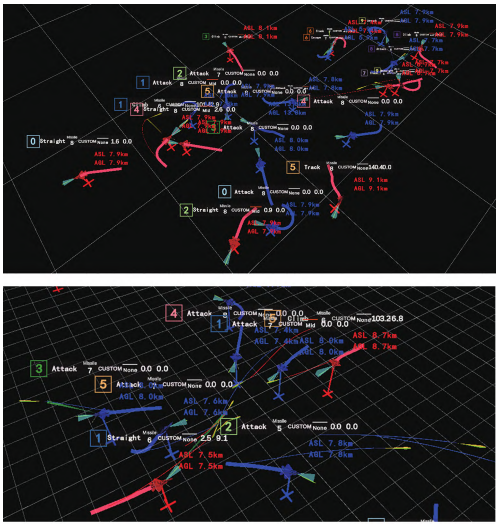
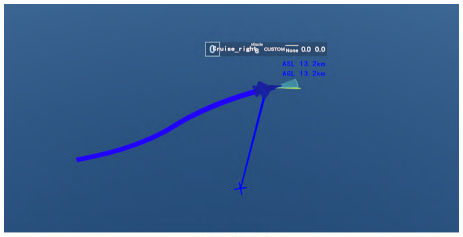


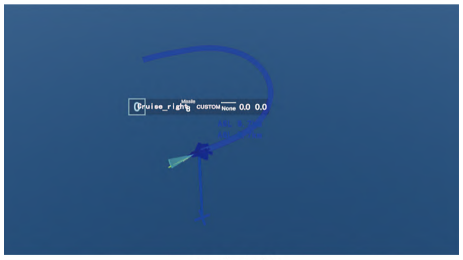
图 8 交战场景

Fig. 8 Battle scene

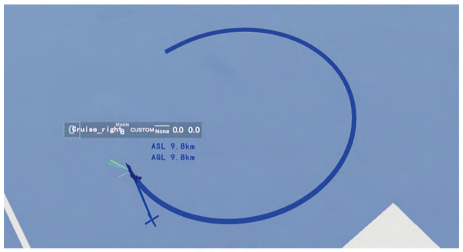
仿真环境中的实验结果表明:智能体学到了超视距空战的一些基本策略,具备了超视距空战的作战能力,可以很好地完成拉起、急盘旋等单个机动动作,如图 9 所示。



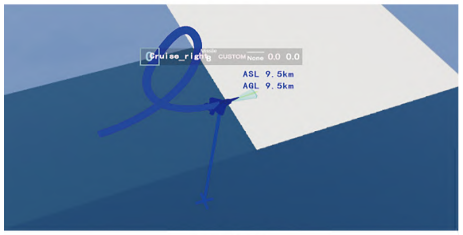
(a) 拉起



(b) 急盘旋



(c) 盘旋



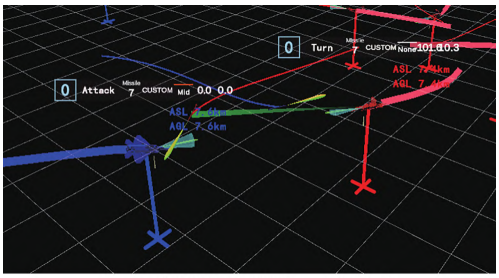
(d) 筋斗

图 9 智能体完成单个机动

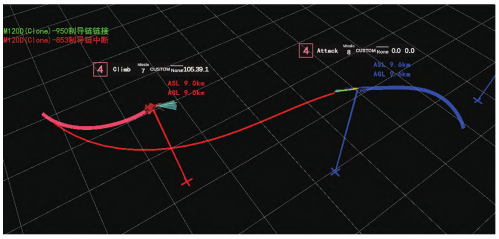
Fig. 9 The agent completes a single maneuver

除了完成单个机动之外,在奖励的正确引导下,智能体学会了选择合适的机动和武器的使用。在双方距离较远、相互接近的过程中,智能体会优先发起进攻,先发制人,如图 10 所示。

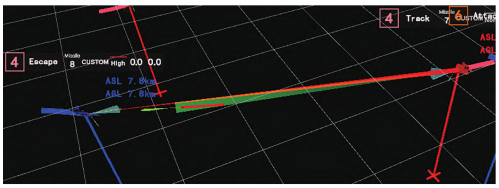
此外,智能体会根据场上我方对敌方的威胁程度以及收到的来自敌方的告警信息等要素,进行综合判断。如果告警等级较低,且我方武器威胁减弱,智能体会及时进行二次进攻。这样既可



(a) 智能体发射导弹，占据优势



(b) 智能体发射导弹，命中目标

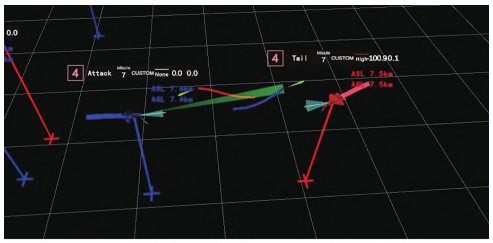


(c) 智能体为导弹提供制导

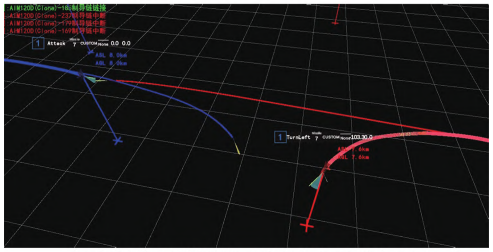
图 10 智能体优先进攻并精确打击

Fig. 10 The agent attacks accurately at the first time

以保证我方的命中率，又尽量节约武器。若告警等级较高，智能体通过告警信息，得知威胁的大致距离和方位，从而采取合适的机动进行躲避，如图 11 所示。若告警等级较低，智能体会优先进行回转再入，开启下一轮的进攻。通过训练，智能体学习到了在确保自身安全的前提下主动进攻、灵活进行攻防转换的策略。



(a) 智能体近距离躲避导弹

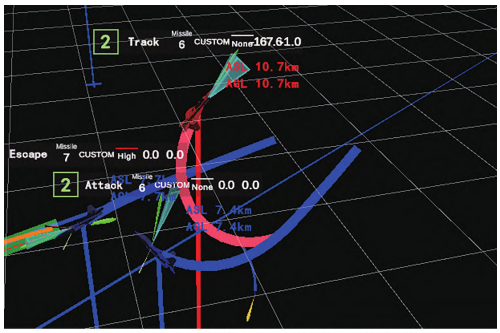


(b) 智能体提前躲避，脱离敌方导弹攻击区

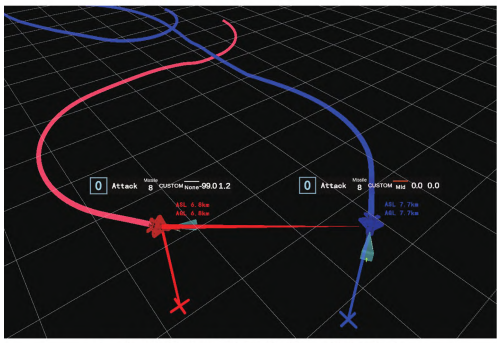
图 11 智能体进行躲避

Fig. 11 The agent escapes from the enemy

虽然仿真实验双方进行的是超视距空战的对抗，但随着战斗的进行，双方的距离越来越近，有可能满足视距内空战的条件。于是，超视距空战转化为视距内空战，出现了双方近距缠斗的现象，如图 12 所示。实验结果表明：在超视距场景下训练得到的智能体，同样具备进行视距内格斗的能力，具有一定的泛化性。



(a) 智能体近距缠斗，摆脱被尾追的劣势



(b) 智能体占据位置优势，进行攻击

图 12 智能体近距缠斗

Fig. 12 The agent dogfights with the enemy

总体来说，专家系统的规则是固定的，相对缺乏变化。而强化学习算法有一定的随机性，决策更灵活，通过机动动作的组合，可以得到更复杂的机动动作。实验表明：强化学习算法控制的智能体在空战中的表现可以与专家系统相抗衡，决策的灵活性、准确性都强于专家系统，具有更大的潜力。

4.3.2 奖励的消融实验

利用消融实验研究空战奖励函数中每项奖

励对算法性能的影响，每次分别去除一个奖励，探究该项奖励对于智能体决策的影响。实验结果如表5所示。

表 5 奖励消融实验

Table 5 Ablation experiments of rewards

奖励	胜率/%	负率/%	平局率/%
完整奖励	62	30	8
去除优势奖励	45	47	8
去除威胁奖励	37	62	1
去除躲避奖励	45	38	6
去除扫描奖励	57	39	4

表6为分别去除单个奖励的训练结果，其中胜率、负率和平局率取算法收敛之后的平均值。实验结果表明：提出的奖励符合空战的设定，分别去除优势奖励、威胁奖励、躲避奖励、扫描奖励之后，算法性能均出现下降。其中，去除优势奖励、威胁奖励、躲避奖励中的任何一个，算法性能都大幅度下降，胜率明显降低，说明优势、威胁、躲避奖励在智能体学习空战策略的过程中起到了关键作用。观察智能体的表现，去除优势奖励后，智能体难以学会主动对准敌方，占据优势。去除威胁奖励和躲避奖励后，智能体无法学会在存在敌方导弹威胁时进行躲避的策略。去除扫描奖励后，算法性能略有下降，但下降的幅度不明显。实验结果表明：与优势等奖励相比，扫描奖励起到的作用相对较小。

4.3.3 超参数对算法的影响

提出的ARM-PPO算法中，包含较多的超参数，需要根据环境，合理配置超参数的值。文献[22]中已经给出了裁剪因子ε、折扣因子γ、GAE算法中的参数λ等，因此在算法设计时，针对这些参数，没有进行进一步的探究。由于学习率的含义较为明确，在实际中已研究得较为充分，因此没有针对学习率的设置进行探究。在实验过程中发现，序列长度和batch_size对算法性能的影响并不明确，故通过实验，探究序列长度和batch_size对算法的影响。

序列长度是1幕的最大长度，也是将轨迹数据分割的长度。batch_size是指训练时从经验池中采样的数据条数。由于考虑重生环节，故在仿真环境中，空战变为无终态的过程。为了简化问题和便于算法训练，按照序列长度将轨迹数据进行分割，以便于智能体学习策略。

图13(a)为序列长度为8、16、32时，智能体的胜率变化情况。可以看到，改变序列长度，对于智能体性能的影响不大，但序列长度为16时的平均胜率略高于其他情形。如图13(b)所示，改变batch_size，对于胜率有一定程度的影响，batch_size为60和120时，训练后期胜率有一定程度的下降，batch_size为120时的胜率相对最高。经过以上的实验比较，最终确定了算法中超参数的取值。

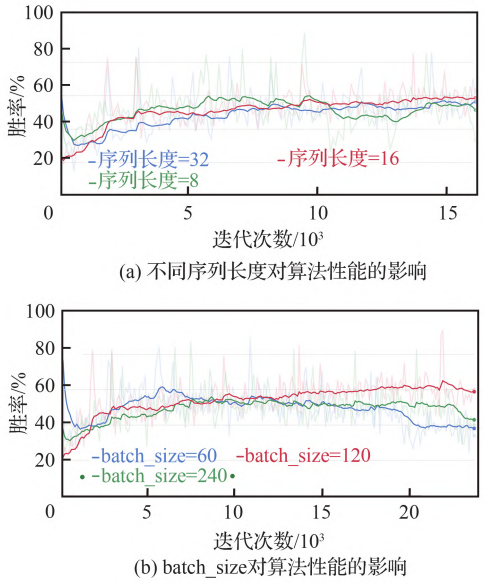


图 13 部分超参数对于智能体性能的影响

Fig. 13 Impact of some hyperparameters on the agent

4.3.4 不同算法的比较

针对ARM-PPO算法的网络结构设计中的关键要素，进行消融实验，改变相应要素，得到新的网络结构，与ARM-PPO算法进行比较，验证提出算法的有效性。图14为PPO算法中离散参数和连续参数的对比。图15为设计的不同网络结构，结构A去除ARM-PPO算法中的自回归结构，机动动作网络和参数网络分别独立输出机动动作和动作参数，彼此互不影响，结构B中的参数网络输出连续参数的均值，建立正态分布，采样得到参数。

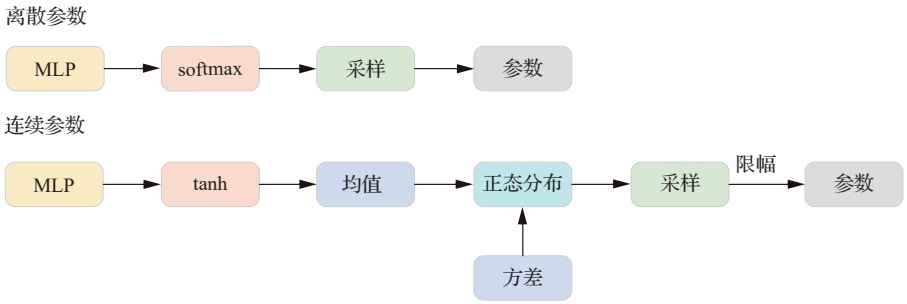
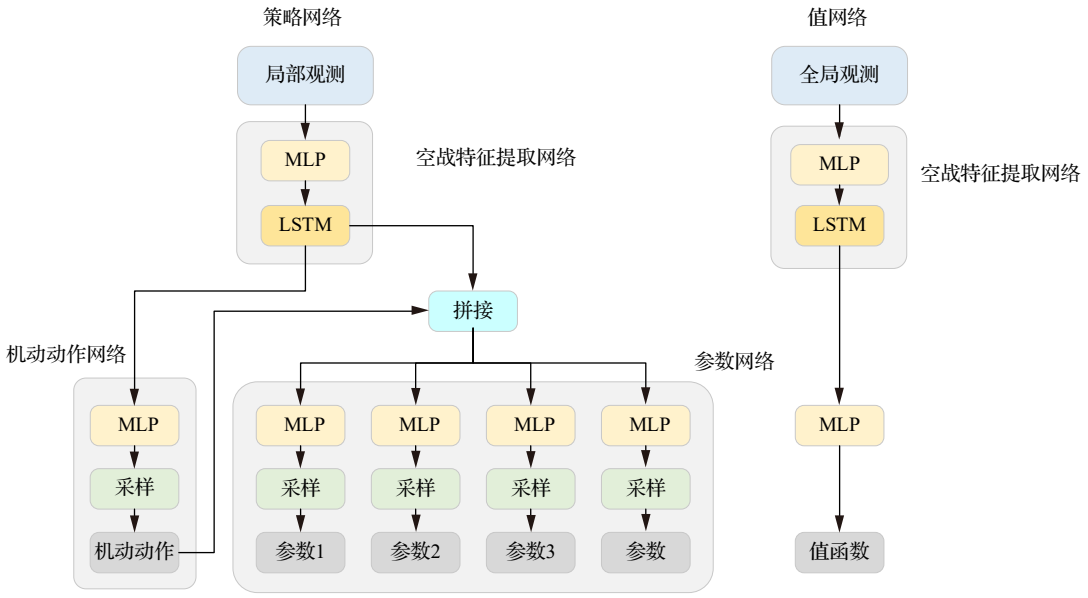
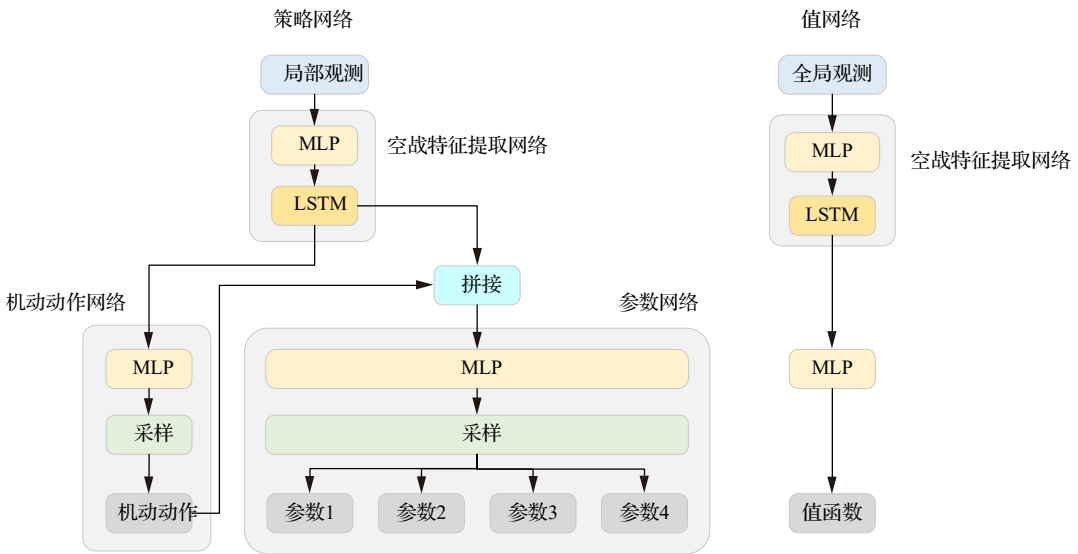


图 14 连续参数和离散参数对比

Fig. 14 Comparison between continuous and discrete parameters



(a) 结构A：策略网络分别独立输出机动动作和参数



(b) 结构B：参数网络输出连续参数的均值，采样得到参数

图 15 不同网络结构

Fig. 15 Different network structures

如图15（a）所示，结构A为H-PPO[37]的结构，机动网络和参数网络共享LSTM提取的特征，分别独立输出机动动作和参数，二者平行并立，去除了ARM-PPO算法中的自回归的结构，参数网络不以机动动作为输入。如图15（b）所示，结构B为连续参数的网络结构，和ARM-PPO算法的结构类似，但输出连续参数。在ARM-PPO算法中，将动作参数离散化，这里对连续参数的情形进行讨论。对于离散参数，使用softmax函数构建离散分布，采样得到参数；对于连续参数的情形，网络输出参数的均值，建立正态分布，采样得到参数值。设计结构A和结构B的目的在于验证ARM-PPO算法中自回归和离散参数的有效性。

图16为ARM-PPO、结构A和结构B的训练曲线比较。可以看到，结构A在训练中期的胜率和ARM-PPO算法基本接近，但在训练后期胜率有所下降，且结构A在训练初期胜率的上升速度比ARM-PPO慢。推测原因可能是因为采用自回归结构，对机动动作和参数之间的关系进行显式建模，便于智能体学习到动作和参数之间的关系，使得智能体更加快速地学习到更优的策略。

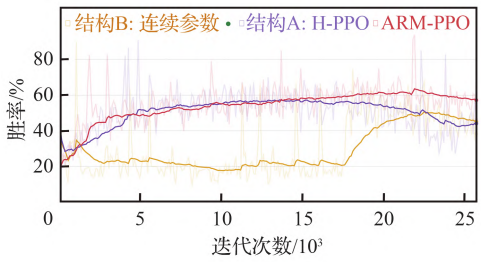


图16 不同网络结构训练曲线对比

Fig. 16 Comparison of training curves of different network structures

如图16所示，结构B的性能大幅下降，胜率降到了20%左右，虽然在后期探索到了一些较好的策略，胜率大幅提升，但最终的结果和结构A类似。原因可能在于：连续参数扩大了智能体的动作空间，智能体难以学习到最优策略，而且，离散参数的设定可以胜任空战动作空间的表达，连续参数反而增加了问题的复杂度。实验结果表明：采用自回归结构和动作参数离散化提升了算法的性能，证明了算法的有效性。

5 结 论

本文针对超视距空战决策问题，定义了超视距空战动作集，提出了自回归多维离散近端策略优化算法（ARM-PPO），并建立数字孪生空战仿真环境，在仿真环境中对算法进行训练和验证。结论如下：

1）针对超视距空战的特点，选择适合超视距空战的机动动作集，控制飞机实现机动动作，并将参数空间离散化，得到多维离散的动作空间。

2）针对超视距空战，设计奖励函数，定义位置优势和威胁奖励，引导智能体完成超视距空战进攻和防御的子目标。基于PPO算法，采用自回归结构，提出了自回归多维离散近端策略优化算法（ARM-PPO），显著提高了智能体的决策能力。

3）构建了空战的仿真环境，在仿真环境中对算法进行测试，并构建专家系统，和提出的算法进行比较。实验结果表明：提出的ARM-PPO算法，在双方势均力敌的情况下，可以胜过专家系统，能够较好地完成超视距空战的任务，为真实战场上战斗机的超视距作战提供可能的解决方案。

本文设计了针对超视距空战机动决策的算法，提出的算法在决策的灵活性、准确性等方面优于基于规则的专家系统。在后续的研究工作中，可以研究单机空战算法中近距和超视距空战融合的问题，还可以将本文的算法推广到多机的情形，解决多机协同作战的问题。此外，还将开展针对算法稳定性、可靠性等方面的研究，并进一步优化算法，为最终将算法部署到机载硬件设备上做准备。

参 考 文 献

[1] 喻煌超，牛轶峰，王祥科. 无人机系统发展阶段和智能化趋势[J]. 国防科技，2021，42(3)：18-24.
YU H C，NIU Y F，WANG X K. Stages of development of Unmanned Aerial Vehicles[J]. National Defense Technology，2021，42(3)：18-24 (in Chinese).

[2] ERNEST N，CARROLL D. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions[J]. Journal of Defense Management，2016，6(1)：1000144.

[3] POPE A P, IDE J S, MIĆOVIĆ D, et al. Hierarchical reinforcement learning for air-to-air combat[C]// 2021 International Conference on Unmanned Aircraft Systems (ICUAS). Piscataway: IEEE Press, 2021: 275-284.

[4] DINARDO G. Artificial intelligence flies XQ-58A Valkyrie drone [EB/OL] (2023-08-03) [2023-12-15]. https://www.defensenews.com/unmanned/2023/08/03/artificial-intelligence-flies-xq-58a-valkyrie-drone/.

[5] 赵志忠, 高正红, 刘行伟, 等. 用攻击点推移速率评估一对一超视距空战效能[J]. 系统仿真学报, 2005, 17(12): 2855-2857, 2862.
ZHAO Z Z, GAO Z H, LIU X W, et al. Using shooting point stepping pace for evaluating one-versus-one BVR combat effectiveness[J]. Acta Simulata Systematica Sinica, 2005, 17(12): 2855-2857, 2862 (in Chinese).

[6] 杜海文, 崔明朗, 韩统, 等. 基于多目标优化与强化学习的空战机动决策[J]. 北京航空航天大学学报, 2018, 44(11): 2247-2256.
DU H W, CUI M L, HAN T, et al. Maneuvering decision in air combat based on multi-objective optimization and reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(11): 2247-2256 (in Chinese).

[7] AUSTIN F, CARBONE G, FALCO M, et al. Automated maneuvering decisions for air-to-air combat[C]// Proceedings of the Guidance, Navigation and Control Conference. Reston: AIAA, 1987:2393.

[8] ISAACS R. Differential games: A mathematical theory with applications to warfare and pursuit, control and optimization[M]. Mineola: Dover Publications, 1999.

[9] 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): 86-97.

[10] 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: NASA-CR-2583 [R]. Washington, D. C.: NASA, 1975.

[11] SUN Z X, PIAO H Y, YANG Z, et al. Multi-agent hierarchical policy gradient for Air Combat Tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

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

[13] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529: 484-489.

[14] BERNER C, BROCKMAN G, CHAN B, et al. Dota2 with large scale deep reinforcement learning[DB/OL]. arXiv preprint: 1912.06680,2019.

[15] 章胜, 周攀, 何扬, 等. 基于深度强化学习的空战机动决策试验[J]. 航空学报, 2023, 44(10): 128094.
ZHANG S, ZHOU P, HE Y, et al. Air combat maneuver decision-making test based on deep reinforcement learning [J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(10): 128094 (in Chinese).

[16] 张建东, 王鼎涵, 杨啟明, 等. 基于分层强化学习的无人机空战多维决策[J]. 兵工学报, 2023, 44(6): 1547-1563.
ZHANG J D, WANG D H, YANG Q M, et al. Multi-dimensional decision-making for UAV air combat based on hierarchical reinforcement learning [J]. Acta Armamentarii, 2023, 44(6): 1547-1563 (in Chinese).

[17] 邱妍, 赵宝奇, 邹杰, 等. 基于PPO算法的无人机近距空战自主引导方法[J]. 电光与控制, 2023, 30(1): 8-14.
QIU Y, ZHAO B Q, ZOU J, et al. An autonomous guidance method of UAV in close air combat based on PPO algorithm [J]. Electronics Optics & Control, 2023, 30(1): 8-14 (in Chinese).

[18] 钱殿伟, 齐红敏, 刘振, 等. 基于改进近端策略优化的空战自主决策研究[J/OL]. 系统仿真学报,(2023-07-20)[2024-01-01]. https://doi.org/10.16182/j.issn1004731x.joss.23-0584.
QIAN D W, QI H M, LIU Z, et al. Research on autonomous decision-making in air-combat based on improved proximal policy optimization [J/OL]. Journal of System Simulation, (2023-07-20) [2024-01-01]. https://doi.org/10.16182/j.issn1004731x.joss.23-0584 (in Chinese).

[19] BARTO A G. Reinforcement learning[M]//OMIDVAR O, ELLIOTT D L. Neural Systems for Control. Amsterdam: Elsevier, 1997: 7-30.

[20] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation[C]// Proceedings of the 12th International Conference on Neural Information Processing Systems. New York: ACM, 1999: 1057 - 1063.

[21] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. New York:ACM,2015: 1889-1897.

[22] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[DB/OL]. arXiv preprint:1707.06347,2017.

[23] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al.

Soft actor-critic algorithms and applications[DB/OL]. arXiv preprint: 1812.05905,2018.

[24] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[DB/OL]. arXiv preprint :1509.02971, 2015.

[25] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// Proceedings of the 35th International Conference on Machine Learning,2018: 1587-1596.

[26] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[DB/OL]. arXiv preprint: 1506.02438, 2015.

[27] ENGSTROM L, ILYAS A, SANTURKAR S, et al. Implementation matters in deep policy gradients: A case study on PPO and TRPO[DB/OL]. arXiv preprint:2005.12729, 2020.

[28] ZHU J Y, KUANG M C, ZHOU W Q, et al. Mastering air combat game with deep reinforcement learning[J]. Defence Technology, 2024, 34: 295-312.

[29] 王宝来,高显忠,谢涛,等.基于强化学习与种群博弈的近距空战决策研究[J/OL].航空学报,(2023-11-02)[2024-01-01]. https://hkxb.buaa.edu.cn/CN/10.7527/S1000-6893.2023.29446.
WANG B L, GAO X Z, XIE T, et al. Research on decision-making in close-range air combat based on reinforcement learning and population game[J/OL]. Acta Aeronautica et Astronautica Sinica,(2023-11-02)[2024-01-01]. https://hkxb.buaa.edu.cn/CN/10.7527/S1000-6893.2023.29446 (in Chinese).

[30] 张婷玉, 孙明玮, 王永帅, 等. 基于深度Q网络的近距空战智能机动决策研究[J]. 航空兵器, 2023, 30(3): 41-48.
ZHANG T Y, SUN M W, WANG Y S, et al. Research on intelligent maneuvering decision-making in close air combat based on deep Q network[J]. Aero Weaponry, 2023, 30(3): 41-48 (in Chinese).

[31] ZHANG H P, WEI Y J, ZHOU H, et al. Maneuver decision-making for autonomous air combat based on FRE-PPO[J]. Applied Sciences, 2022, 12(20): 10230.

[32] 杨晟琦, 田明俊, 司迎利, 等. 基于分层强化学习的无人机机动决策[J]. 火力与指挥控制, 2023, 48(8): 48-52, 59.
YANG S Q, TIAN M J, SI Y L, et al. Research on UAV maneuver decision-making based on hierarchical reinforcement learning[J]. Fire Control & Command Control, 2023, 48(8): 48-52, 59 (in Chinese).

[33] 钟友武, 柳嘉润, 杨凌宇, 等. 自主近距空战中机动动作库及其综合控制系统[J]. 航空学报, 2008, 29(S1): 114-121.
ZHONG Y W, LIU J R, YANG L Y, et al. Maneuver library and integrated control system for autonomous close-in air combat [J]. Acta Aeronautica et Astronautica Sinica, 2008, 29(S1): 114-121 (in Chinese).

[34] NG A Y, HARADA D, RUSSELL S J. Policy invariance under reward transformations: theory and application to reward shaping[C]// Proceedings of the Sixteenth International Conference on Machine Learning. New York: ACM, 1999:278-287.

[35] 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.
ZHU J Y, ZHANG H L, KUANG M C, et al. Curriculum learning based simulation of UAV air combat under sparse rewards[J]. Journal of System Simulation, 2024, 36(6):1452-1467 (in Chinese).

[36] 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 W Q, ZHU J H, KUANG M C. An unmanned air combat system based on swarm intelligence[J]. Scientia Sinica (Informationis), 2020, 50(3): 363-374 (in Chinese).

[37] FAN Z, SU R, ZHANG W N, et al. Hybrid actor-critic reinforcement learning in parameterized action space[DB/OL]. arXiv preprint: 1903.01344,2019.

（责任编辑：李丹）

Hierarchical decision algorithm for air combat with hybrid action based on deep reinforcement learning

LI Zuolong[1], ZHU Jihong[1,*], KUANG Minchi[1], ZHANG Jie[2], REN Jie[2]

1. Department of Precision Instrument, Tsinghua University, Beijing 100084, China

2. AVIC Chengdu Flight Design and Research Institute, Chengdu 610091, China

Abstract: Intelligent air combat is a hot research topic among countries with strong military power in the world. To solve the maneuver decision problem of air combat Beyond Visual Range (BVR), we propose the hierarchical decision algorithm based on deep reinforcement learning. In the decision algorithm, we use the maneuver set appropriate to the BVR air combat to control the trajectory and the attitude of the aircraft. To expand the action space of the model and increase its decision-making ability, we hierarchize the action space and model it as the multi-discrete one. To solve the problem of sparse reward in air combat, we design a set of reward function taking into consideration the factors including the position advantage, weapon launching, and weapon threat, which can guide the agent to converge to the optimal policy. We also build a complete digital-twin simulation environment for air combat and an expert system. The decision algorithm is trained in the simulation environment, and is evaluated by fighting with the expert system. The experiment results indicate that the decision algorithm proposed has the ability to make autonomous and flexible decisions in BVR air combat based on current situations, and has some advantages against the expert system.

Keywords: air combat beyond visual range; intelligent decision; deep reinforcement learning; proximal policy optimization; maneuver; hierarchical decision

Received: 2024-01-02; **Revised**: 2024-01-11; **Accepted**: 2024-04-22; **Published online**: 2024-04-26 14:53
URL: https://hkxb.buaa.edu.cn/CN/Y2024/V45/I17/530053
* **Corresponding author.** E-mail: jhzhu@tsinghua.edu.cn