文章编号:1674-8190(2022)06-040-10

基于深度强化学习的多无人机协同进攻作战智能规划

李俊圣,岳龙飞,左家亮,俞利新,赵家乐
(空军工程大学 空管领航学院,西安 710051)

摘 要: 无人机依靠作战效费比高、灵活自主等优势逐步替代了有生力量作战,多无人机协同作战任务规划成为热点研究问题。针对传统任务规划采用的智能优化算法存在的依赖静态、低维的简单场景,机上计算较慢等不足,提出一种基于深度强化学习(DRL)的端到端的多无人机协同进攻智能规划方法;将压制敌防空(SEAD)作战任务规划过程建模为马尔科夫决策过程,建立基于近端策略优化(PPO)算法的SEAD智能规划模型,通过两组实验验证智能规划模型的有效性和鲁棒性。结果表明:基于DRL的智能规划方法可以实现快速、精细规划,适应未知、连续高维的环境态势,SEAD智能规划模型具有战术协同规划能力。

关键词: 多无人机;深度学习;深度强化学习;PPO算法;泛化性;协同作战
中图分类号: V219 **文献标识码:** A
DOI: 10.16615/j.cnki.1674-8190.2022.06.04 **开放科学(资源服务)标识码(OSID):**

Multi-UAV Cooperative Offensive Combat Intelligent Planning Based on Deep Reinforcement Learning

LI Junsheng, YUE Longfei, ZUO Jialiang, YU Lixin, ZHAO Jiale
(College of Air Traffic Control and Navigation, Air Force Engineering University, Xi'an 710051, China)

Abstract: Unmanned aerial vehicle (UAV) with the advantages of high effectiveness and flexible autonomy has gradually replaced manned aircraft to combat, and multi-UAV cooperative combat mission planning becomes the hot research issue. An end-to-end cooperative attack intelligent planning method for multi-UAV based on deep reinforcement learning (DRL) is presented to overcome the shortcomings of traditional mission planning algorithms, such as static dependence, low-dimensional simple scenarios and slow on-board computing speed. The suppression of enemy air defense (SEAD) mission planning is modeled as the Markov decision process. The SEAD intelligent planning model based on proximal policy optimization (PPO) algorithm is established, and two groups of experiments are used to verify the effectiveness and robustness of the intelligent planning model. The results show that the DRL-based intelligent planning method can realize fast and fine planning, adapt to unknown, continuous and high-dimensional environment situation. The SEAD intelligent planning model has the capacity of tactics cooperative planning.

Key words: multi-UAV; deep learning; deep reinforcement learning; PPO algorithm; generalization; cooperative combat

收稿日期: 2022-01-13; **修回日期:** 2022-04-25
基金项目: 国家自然科学基金(62106284);陕西省自然科学基金(2021JQ-370);军内科研项目(KJ20191A030153)
通信作者: 俞利新,975962507@qq.com
引用格式: 李俊圣,岳龙飞,左家亮,等. 基于深度强化学习的多无人机协同进攻作战智能规划[J]. 航空工程进展,2022,13(6):40-49,96.
LI Junsheng, YUE Longfei, ZUO Jialiang, et al. Multi-UAV cooperative offensive combat intelligent planning based on deep reinforcement learning[J]. Advances in Aeronautical Science and Engineering, 2022, 13(6): 40-49,96. (in Chinese)

0　引　言

近年来,随着传感器、功能载荷、控制和通信技术的进步,多无人机协同作战成为研究热点[1]。同时,由于计算机计算能力的提升、大数据的出现以及人工智能算法的发展,深度学习[2-3]、强化学习[4]等基于学习的算法掀起了第二波人工智能浪潮,人工智能在作战领域的发展也显得至关重要。2018年美国国防部颁布《国防部人工智能战略摘要》[5],强调人工智能技术在军事领域的应用,并于同年发布了无人集群系统并行作战场景。美国战略和预算评估中心连续发布针对中俄两国的马赛克式集群作战等颠覆性作战模式,打造全球范围内的武器系统协同作战[6]。

任务规划本质上是一个决策优化问题,目的是求解任务目标在约束条件下的最优解。目前对多无人机任务规划问题的求解大多采用静态的智能优化算法,如遗传算法、退火算法、粒子群算法、蚁群算法等,除此之外还有基于智能体(agent)建模的方法。张睿文等[7]采用基于agent的分层行为建模和组合模块化,实现了无人机集群动态自组织航路侦查,这种方式具有较强的可解释性和稳定性,但灵活性较差;谭威等[8]通过遗传算法、混合整数线性规划算法完成了多无人机基于任务的最优航迹规划;Zhang H等[9]研究了战术机动规划问题,以战机受威胁最小和地面目标毁伤性能最大为优化目标,考虑武器装备性能约束,采用基于分解的多目标优化算法[10]解出最优飞行航线和武器投放时机,但基于搜索的算法需要在线求解,实时性较差;潘楠等[11]考虑无人机的物理性能因素,搭载航迹最小适应度函数和威胁代价最小适应度函数,采用基于模拟退火的混合粒子群算法求解出多无人机最优任务分配策略,这种算法鲁棒性较差,计算效率随问题规模的增大呈指数降低;辛建霖等[12]通过改进标准蚁群算法,提高了无人机航迹规划的速度和精度。智能优化算法近年来通过改进初始化种群、更新策略优化个体迭代方式和混合多种算法已经使其能够并行求解出低维空间约束条件下的最优解或次优解,且具有较好的收敛性。但智能优化算法的本质仍是随机搜索,需要显式的目标函数,且每次只能在线求解,不能泛化到动态变化的未知环境,尤其是在计算大规模问题时计算量以指数形式增加且收敛性降低,因此对于未来大规模作战快速规划和动态变化场景,其应用具有一定局限性。

随着人工智能的飞速发展,从AlphaGo[13]、AlphaZero[14]到AlphaStar[15]等,深度强化学习(Deep Reinforcement learning,简称DRL)广泛应用。其中,深度学习解决高维映射问题,强化学习解决序贯决策问题,深度强化学习成功求解了一系列机器人控制[16]、自动驾驶[17]、游戏博弈[18]、优化与调度[19]、航空路径规划[20]等领域问题。基于学习的算法是一种数据驱动的算法,指通过“喂”数据训练,提高模型的预测或决策性能。基于学习的算法采用神经网络来学习或拟合输入与输出之间复杂的高维非线性关系,实现拟合误差最小,预测、决策结果最优等目标,并将映射关系以网络参数的形式保存,实现离线训练、在线规划,对新输入数据也具有一定的鲁棒性和内插泛化性,非常适合于求解快速动态任务规划问题。

因此,本文选择压制敌防空(Suppression of Enemy Air Defense,简称SEAD)作战任务规划[21-22]作为研究背景,提出一种基于DRL的端到端的多无人机智能任务规划方法。首先对SEAD问题进行描述并提出打击任务想定;然后介绍DRL算法原理以及近端策略优化(Proximal Policy Optimization,简称PPO)算法,并建立基于PPO的智能任务规划模型;最后通过设置不同复杂度的模拟仿真和对比试验,分析验证智能规划模型的优越性与潜在价值。

1　SEAD问题描述与打击任务想定

1.1　SEAD问题描述

压制敌防空作战是以软杀伤或硬摧毁的方式打击敌方防空系统的进攻性制空战斗,主要目的是致使敌方防空系统瘫痪,从而使我方夺得制空权,保护我方攻击机顺利完成打击任务。传统战术中,干扰机进行压制,通过电子干扰压缩敌雷达探测距离,为攻击机创造出射程进入条件,随后攻击机在敌雷达探测盲区前出进行攻击。本文提出新的战术:作战中无人机诱饵价格低廉、作战效费比高、机动性能好,能欺骗敌方火力,为攻击机争取时间。因此选择释放攻击诱饵,通过挂载龙波

透镜增大雷达散射截面积(Radar Cross Section,简称RCS)引诱敌方雷达开机,并进入敌攻击范围引诱敌方对其开火,此时攻击机与攻击诱饵协同配合,在敌方地导攻击的间歇发起攻击,达到摧毁的目的,如图1所示。由于敌方雷达探测距离比我方攻击机攻击距离远,攻击机无法独自进攻,因此攻击机与诱饵需要密切协同,在合适的位置进入,同时攻击机在合适的位置进入、发射导弹,两者协同完成任务[23]。

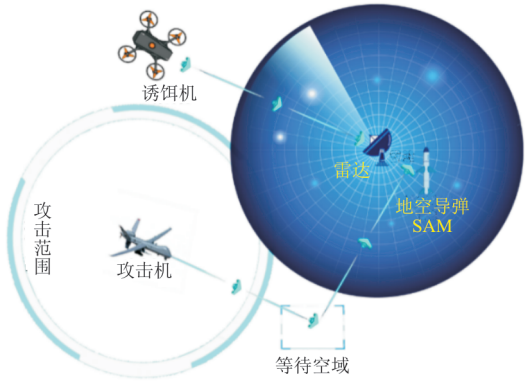

图1 压制敌防空作战

Fig. 1 Suppression of enemy air defense mission

协同作战规划本质上是一个序贯决策问题,在不同时空序列状态下,作战单元采取最优决策序列,从初始态势转移到终止态势,完成任务目标。因此,SEAD作战任务规划可以归结为一个端到端的从状态(位置、态势)到决策(机动、攻击、诱饵牵制)的序列优化问题,优化目标为求解一个最优状态—决策序列,满足我方诱饵机和攻击机通过战术协同,打掉敌方雷达且保证自身攻击机安全,如图2所示。

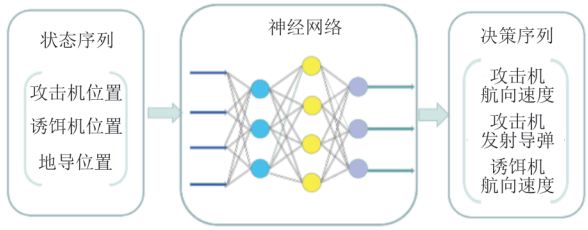

图2 SEAD智能规划模型

Fig. 2 Intelligent mission planning model of SEAD

1.2 SEAD打击任务想定

本文研究如何使攻击机与诱饵机有效协同合作,并检验复杂作战情景下任务的完成情况,因此具体想定(红方为己方,蓝方为敌方)有以下三个方面。

(1)在蓝方防御区域设置多个地导阵地,射程为20 km,能够击毁进入其攻击范围的无人机,但具有攻击间隔,需要一定时间准备才能继续瞄准制导;(2)在地导阵地之间设置1个高价值目标,规定高价值目标处于红方攻击机射程之内一段时间则认为被摧毁,为红方的主要突袭目标;(3)红方兵力为多架攻击机和诱饵机,攻击机射程20 km,初始部署在红色空域,主要任务是躲避地导攻击的同时突袭敌方高价值目标。诱饵机初始部署在绿色空域,主要任务为适时进入敌方地导攻击范围吸引火力,掩护红方攻击机,保证红方攻击机的安全。任务完成的条件为击毁蓝方高价值目标且攻击机存活,示意图如图3所示。

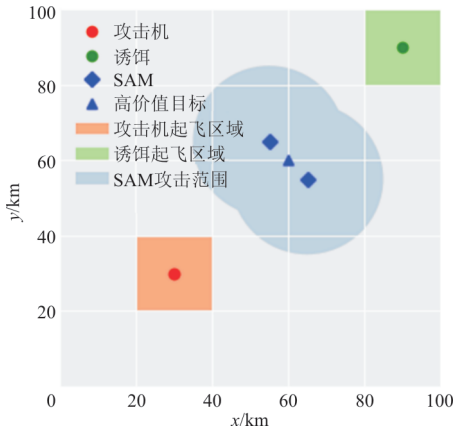

图3 任务场景

Fig. 3 Mission scenario

2 强化学习

2.1 基本原理

强化学习又称试错学习,旨在智能体通过试错机制不断与环境交互得到反馈,从而得到累积奖励最大化的最优策略。强化学习包括智能体和环境两部分。智能体感知初始状态从动作列表集选择动作,环境接受该动作后给予智能体一个即时奖励,同时智能体转移到下一个状态,继续选择新的动作直至到达终止状态。智能体的目标是找到最优状态—决策序列,因此强化学习是一个决策优化算法,强化学习框架如图4所示。

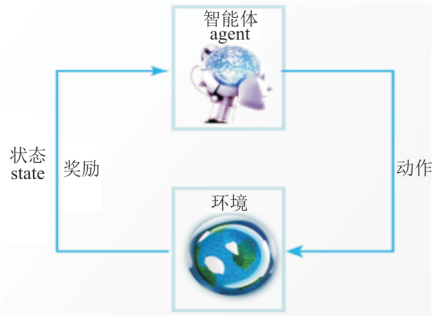

图 4　强化学习框架

Fig. 4　Reinforcement learning framework

通常将强化学习建模为一个有限马尔科夫决策过程,用(S,A,R,P,γ)五元组表示,其中:S为可选的状态集,A为可选的动作集,R为奖励函数,P为状态转移函数,γ为折扣因子,用来计算长期折扣奖励和。假设t时刻智能体的状态为$s\in S$,根据策略$\pi: S\to A$采取动作$a\in A$,则环境反馈给智能体一个即时奖励$r\in R$,同时智能体转移到新的状态$s'\in S$。传统强化学习采用离散的状态—动作值表(Q table)评估状态—动作的好坏,但对于连续高维问题,遇到了“维度灾难”,因此研究者提出了深度强化学习[24-26]。

2.2　深度强化学习

深度强化学习引入神经网络(Neural Network,简称NN)替换离散状态的状态—动作值表,将高维连续的状态与动作采用神经网络近似,以达到降低任务复杂度、提高学习速度的目的[27]。深度强化学习采用神经网络近似策略和值函数,解决了高维映射问题。智能体的目标是得到期望回报$J(\pi_\theta)=E_{\tau\sim\pi_\theta}[R(\tau)]$最大的策略$\pi_\theta$,$\theta$为策略参数,回报轨迹$\tau=(s_0,a_0,s_1,a_1,\cdots)$上的折扣奖励和$R(\tau)=\sum_{t=0}^{\infty}\gamma^t r_t$,最优策略为

$$\pi_\theta^*=\arg\max_{\pi_\theta} E_{\tau\sim\pi_\theta}\left(\sum_{t=0}^{\infty}\gamma^t r_t\right) \tag{1}$$

式中:r_t为t时刻奖励;γ为折扣因子。

深度强化学习算法分为三种学习框架:值函数(Value Based)、策略梯度(Policy Gradient)和行动者—评论家(Actor-Critic)。其中Actor-Critic类强化学习算法综合了值函数和策略梯度,用值函数误差指导策略更新,加快学习速度。策略π_θ通过期望回报的梯度$\nabla_\theta J(\pi_\theta)$更新,期望回报梯度为

$$\nabla_\theta J(\pi_\theta)=E_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\nabla_\theta\log\pi_\theta(a|s)R(\tau)\right] \tag{2}$$

式中:$\pi_\theta(a|s)$为Actor;$R(\tau)$为Critic,$R(\tau)$也可采取其他形式如状态—行为值函数$Q^\pi(s_t,a_t)$、优势函数$A^\pi(s_t,a_t)$、TD残差$G_t=r_t+V^\pi(s_{t+1})-V^\pi(s_t)$等。

当$R(\tau)$取TD残差,并且值函数$V^\pi(s_t)$由参数为ω的神经网络进行逼近时,对式(2)求导,$R(\tau)$按照式(3)、式(4)更新,$\pi_\theta(a|s)$按照式(5)更新。

$$G_t-\hat{v}(s_t,\omega)\to\delta \tag{3}$$

$$\omega+\beta\delta\nabla_w\hat{v}(s_t,\omega)\to\omega \tag{4}$$

$$\theta+\alpha\delta\nabla_\theta\log\pi_\theta(a_t|s_t)\to\theta \tag{5}$$

2.3　近端策略优化

PPO算法是一种简单稳定、性能强大且易于实现的Actor-Critic框架算法。OpenAI DOTA2智能体OpenAI Five[28]和腾讯的王者荣耀智能体JueWu[29]都采用PPO实现。PPO算法针对信赖域策略优化(Trust Region Policy Optimization,简称TRPO)[30]算法计算量巨大(为保证策略性能单调非减,将目标函数进行一阶近似,约束条件进行二阶泰勒展开,利用共轭梯度法求解策略参数)的问题,通过一阶近似,优化替代损失函数,在每一次迭代中计算新策略,并且保证新策略和旧策略相近,朝着损失函数最小化(期望回报最大化)的方向优化策略,算法在采样效率、最终性能、工程实现和调试复杂度之间取得了平衡。

PPO算法中Critic采用优势函数来评估动作的好坏,式(2)转化为

$$\nabla_\theta J(\pi_\theta)=E_{a\sim\pi_\theta}[\nabla_\theta\log\pi_\theta(a|s)A^\pi(s,a)] \tag{6}$$

由于PPO是一种同策略(on-policy)算法,为了提高样本利用率,引入重要性采样,用旧策略$\pi_{\theta'}$进行采样,得到:

$$\nabla_\theta J(\pi_\theta)=E_{a\sim\pi_{\theta'}}\left[\frac{\pi_\theta}{\pi_{\theta'}}A^{\pi_{\theta'}}(s,a)\nabla_\theta\log\pi_\theta\right] \tag{7}$$

又由于$\pi_\theta\nabla_\theta\log\pi_\theta=\nabla\pi_\theta$,式(7)转换为

$$\nabla_\theta J(\pi_\theta)=E_{a\sim\pi_{\theta'}}\left[\frac{\nabla\pi_\theta}{\pi_{\theta'}}A^{\pi_{\theta'}}(s,a)\right] \tag{8}$$

该梯度对应的优化目标函数:

$$J(\pi_\theta)=E_{a\sim\pi_{\theta'}}\left[\frac{\pi_\theta}{\pi_{\theta'}}A^{\pi_{\theta'}}(s,a)\right] \tag{9}$$

在实际应用中,基于采样估计期望,简化得到PPO的优化目标替代损失函数,如式(10)所示,通过裁剪操作(clip)来限制策略更新幅度,保证训练

稳定性。

$$J^{\text{clip}}(\theta)=\sum_{(s,a)}\min\left\{r(\theta)\hat{A},\text{clip}\left[r(\theta),1-\varepsilon,1+\varepsilon\right]\hat{A}\right\} \tag{10}$$

式中：$r(\theta)=\dfrac{\pi_\theta(a|s)}{\pi_{\theta'}(a|s)}$为新旧策略之比；$\varepsilon$为裁剪幅度超参数。

优势函数采用泛化优势估计(Generalized Advantage Estimation，简称GAE)[31]以平衡值函数估计的方差与偏差，如式(11)所示。

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)}=\sum_{l=1}^{\infty}(\gamma\lambda)^l\left[r_t+\gamma V(s_{s+t+l})-V(s_{t+l})\right] \tag{11}$$

3 基于PPO的SEAD智能规划建模

3.1 飞机运动学方程

本文主要研究深度强化学习算法在多无人机智能规划上的可行性和潜在价值，同时考虑到无人机在执行任务时为了精准完成任务不能采取大机动，因此建立简单二维环境，攻击机和诱饵机用简化的三自由度模型，运动学方程为

$$\begin{cases}\dot{x}_{\text{f}}=v_{\text{f}}\cos\varphi_{\text{f}}\\ \dot{y}_{\text{f}}=v_{\text{f}}\sin\varphi_{\text{f}}\\ \dot{x}_{\text{j}}=v_{\text{j}}\cos\varphi_{\text{j}}\\ \dot{y}_{\text{j}}=v_{\text{j}}\sin\varphi_{\text{j}}\end{cases} \tag{12}$$

式中：$\dot{x}_{\text{f}},\dot{y}_{\text{f}},v_{\text{f}},\varphi_{\text{f}}$分别为攻击机的位置变化量，速度和航向；$\dot{x}_{\text{j}},\dot{y}_{\text{j}},v_{\text{j}},\varphi_{\text{j}}$分别为诱饵机的位置变化量，速度和航向，航向取值范围为$(-\pi,\pi)$中的任意连续值，速度取值范围为$(0,1]$中的连续值。

3.2 马尔科夫决策过程建模

3.2.1 状态空间

攻击机和诱饵机状态定义为二维空间中的坐标位置，取连续值，分别用$(x_{\text{f}},y_{\text{f}})$、$(x_{\text{j}},y_{\text{j}})$表示。

3.2.2 动作空间

攻击机和诱饵机的动作由各自的航向角、速度约束，通过航向的控制连续改变飞机在二维空间的位置，通过改变速度控制飞机协同到达任务位置。攻击机发射导弹，诱饵机吸引敌方火力，通过设置距离条件默认自动完成，实际任务中需要进行发射时机、位置、参数的详细计算。

3.2.3 奖励函数

奖励函数设计遵循奥卡姆剃刀原理——简单有效原则，即如果诱饵机能进入地空导弹射程内，且攻击机能打掉敌雷达而不进入其导弹射程内，则得到+1奖励；如果攻击机进入地空导航射程或飞出环境边界，则得到−1的奖励；对于其他中间状态，借助专家经验知识采取奖励塑形，给予一个大小为相对距离的连续奖励，引导智能体学习，奖励函数公式如式(13)所示。

$$\begin{cases}r=1 & (d_{\text{fs}}\leqslant d_{\text{f}},\ d_{\text{ds}}\leqslant d_{\text{s}})\\ r=-1 & \begin{array}{l}(d_{\text{ds}}>d_{\text{s}}\text{或}\\ x_{\text{f,d}},y_{\text{f,d}}\leqslant 0\text{或}\\ x_{\text{f,d}},y_{\text{f,d}}\geqslant x_{\max})\end{array}\\ r=x_{\text{f}}+y_{\text{f}}-x_{\text{d}}-y_{\text{d}}+0.8 & (\text{其他})\end{cases} \tag{13}$$

式中：d_{fs}和d_{ds}分别为攻击机与地导的距离、诱饵与地导的距离；d_{s}为地导的攻击范围；d_{f}为攻击机的攻击范围；x_{d}和y_{d}分别为诱饵的横坐标和纵坐标。

3.3 智能规划模型构建

综上，建立端到端的SEAD作战智能规划模型，输入为攻击机、诱饵机的位置状态，对其根据式(14)进行零均值(Z-Score)标准化，然后输入到两层全连接神经网络进行训练，最后输出攻击机和诱饵机的航向角决策量和速度决策量，优化目标为最大化累积奖励，解为最优或次优的参数化策略，智能规划模型如图5所示。

$$x^*=\frac{x-\mu}{\sigma} \tag{14}$$

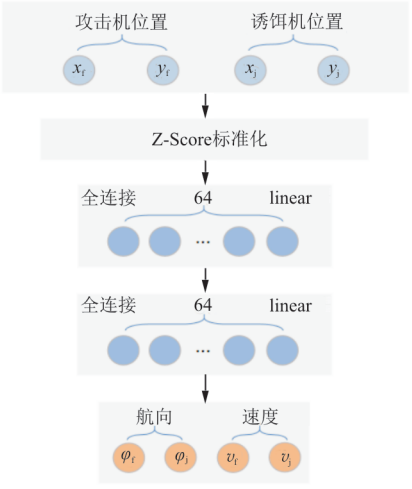

图5 本文策略网络模型

Fig. 5 Policy network model in this work

因此,本文采用先训练、后规划的思路。首先在环境中离线训练智能体,训练网络与环境交互采样,将经验存储下来,根据PPO损失函数进行梯度下降训练和优化。待训练稳定后,得到训练好的网络即推理网络,进行在线规划测试,检验智能体的规划能力,研究框架如图6所示。

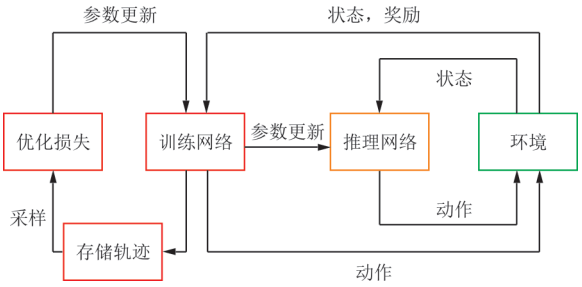

图6　本文研究框架

Fig. 6　The research framework of this paper

基于PPO的智能规划算法流程如下:

初始化攻击机和诱饵的位置
初始化actor和critic网络参数、经验库
for iteration=1,2,…,N do
　运行策略$\pi_{\theta'}$,采样得到动作a
　环境收到a,根据式(12)返回状态s',根据式(13)返回奖励r
　存储轨迹$(s,a,r,s',d,v,\log p(a))$到经验库
　根据存储轨迹计算回报$R(\tau)$和优势$\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)}$
　for k=1,2,…,K do
　　从轨迹中采样,根据式(10)计算策略损失函数$J(\theta)$
　　采用优化器优化策略损失函数$J(\theta)$
　　更新策略参数$\theta'\leftarrow\theta$
　end for
end for

首先环境初始化攻击机和诱饵的位置,初始化神经网络参数和经验库;然后智能体开始与环境交互,存储经验,从经验中采样小批量轨迹,优化网络损失,更新网络参数,循环至最大回合,得到优化后的策略网络,即智能规划模型。

本文提出一种通用智能规划框架,包括环境、规划器(智能体)和控制器三部分,规划器首先输入初始态势,与环境进行大量交互即离线训练,得到训练好的规划器。训练好的规划器可以直接部署,输入新的初始态势进行在线推理,将推理出来的决策序列即规划结果输入控制器执行,从而实现无人系统智能自主规划。智能规划的完整流程如图7所示。

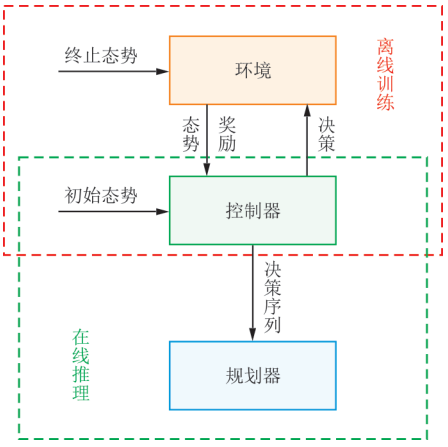

图7　智能规划框架

Fig. 7　The architecture of intelligent planning

4　仿真实验与分析

4.1　实验设置

SEAD环境设置为100 km×100 km的正方形区域,攻击机坐标为(30,30),攻击范围是20 km;诱饵机坐标为(90,90);地导坐标为(65,55)、(55,65),攻击范围为20 km;高价值目标坐标为(60,60),如图8所示。实验中对上述距离缩小100倍进行归一化,易于神经网络训练,防止梯度消失。仿真环境用python3.6和PyCharm,深度学习库用pytorch实现,与其他经典深度强化学习算法进行对比,并进行鲁棒性测试和消融实验,验证算法的有效性。

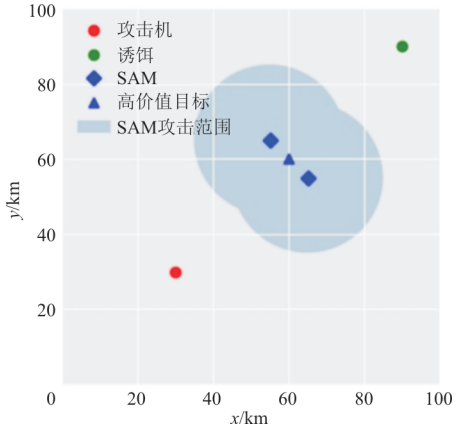

图8　初始态势

Fig. 8　Initial situation

本文超参数设置如表1所示,神经网络采用正交初始化,优化器选择Adam[31]。

表 1 超参数设置
Table 1 Hyperparameter setting

参数	数值
最大回合数	150
学习率	0.000 3
新旧策略裁剪幅度	0.2
奖励折扣系数	0.995
泛化优势估计参数	0.95
每回合最大步长	1 000
策略损失系数	1
值损失系数	0.5
熵系数	0.01
小批量采样数量	64
随机种子个数	3

4.1.1 数据标准化

(1) 优势函数标准化

将优势函数值进行标准化处理，提高训练稳定性，增强策略训练技巧，公式为

$$A'=\frac{A-\mu_A}{\sigma_A} \tag{15}$$

式中：A 为优势函数值；μ_A 为优势函数值的均值；σ_A 为优势函数值的标准差。

(2) 值函数损失标准化

同理，值函数损失也进行标准化处理，公式为

$$V'_{\text{loss}}=\sum_{i}^{N}\frac{(V_i-\mu_{Ri})^2}{6\sigma_{R_i}} \tag{16}$$

式中：V 为状态值；μ_{Ri} 为回报值的均值；σ_{R_i} 为回报值的标准差。

4.1.2 自适应调整参数大小

(1) 自适应学习率

训练前期采取较大的学习率加速收敛，训练后期采取较小的学习率便于找到最优值，公式如式(17)所示。

$$l_r=R_L\times\left(1-\frac{i_{\text{EP}}}{i_{\text{EP}}-i_{\text{EP-max}}}\right) \tag{17}$$

式中：R_L 为初始学习率；i_{EP} 为当前训练回合。

(2) 自适应 clip 值

clip 自适应变化原理和学习率自适应变化一致，训练前期采取较大的 clip 值，允许接受差异较大的新策略，加快策略学习；训练后期采用较小的 clip 值，只接受差异小的新策略，保证策略稳定更新，如式(18)所示。

$$A_{\text{clip}}=A_{\text{CLIP}}\times\left(1-\frac{i}{C_{\text{max\_Episode}}}\right) \tag{18}$$

式中：A_{CLIP} 为初始裁剪幅度；i 为当前训练回合；$C_{\text{max\_Episode}}$ 为最大回合数。

4.2 策略训练技巧

4.2.1 域随机化

为提高智能体策略的鲁棒性，适应多样化输入，对训练阶段的状态输入增加扰动[32]，如式(19)所示，即在每个随机种子上运行参数扰动的不同环境训练智能体，使得智能体能够抽象出更高层的策略特征，避免过拟合到一种环境和策略，最终学到的策略更加鲁棒，更好泛化到未知环境。

$$\begin{cases}x'=x+\delta\\ y'=y+\delta\end{cases} \tag{19}$$

式中：x'，y' 为扰动后的坐标；x，y 为扰动前的坐标；δ 为扰动量。

4.2.2 最大化策略熵

熵用于衡量随机变量的随机性，实际计算时考虑其服从的随机分布，熵越大越随机。因此在最大化累计收益的同时，最大化策略的熵值，让策略尽可能随机，智能体可以充分探索状态空间，避免策略陷入局部最优，并且可以探索到多个可行方案来完成任务，增强了策略的探索能力、鲁棒性和抗干扰能力，策略熵计算公式如式(20)所示。

$$H\left[\pi(\cdot|s_t)\right]=-\sum\pi(\cdot|s_t)\log\pi(\cdot|s_t) \tag{20}$$

4.2.3 网络参数共享

Actor 和 Critic 采用网络参数共享，损失函数为策略损失、值函数损失和策略熵之和，对损失函数梯度进行回传。通过参数共享共用底层网络特征，共享了一部分底层特征，降低训练难度，损失函数如式(21)所示。

$$L(\theta,\phi)=L(\theta)+\lambda_{\text{critic}}L(\phi)+\lambda_{\text{entropy}}H(\pi) \tag{21}$$

式中：$L(\theta)$ 为策略损失；$L(\phi)$ 为值损失；λ_{critic} 为值损失系数；λ_{entropy} 为熵系数。

4.3 实验一：静态场景有效性测试

实验场景设置为 SEAD，其中攻击机 2 架，攻击距离 20 km，诱饵机 2 架，SAM 两套，攻击距离为 20 km。诱饵机作为成本低廉、机动性强的无人

机,吸引敌方火力;攻击机利用敌攻击诱饵的间隙打掉地导和目标,完成任务。进行 150 回合仿真来训练策略网络模型,并利用训练好的模型进行在线规划,检验在线智能规划能力。

4.3.1　离线训练

在三个不同的随机种子上训练 PPO 的策略网络和值函数网络,记录训练过程中的每回合时间步长和累计奖励,并与基于 A2C 和 TRPO 的智能规划模型[30]进行对比,得到平均奖励学习曲线,如图 9 所示。

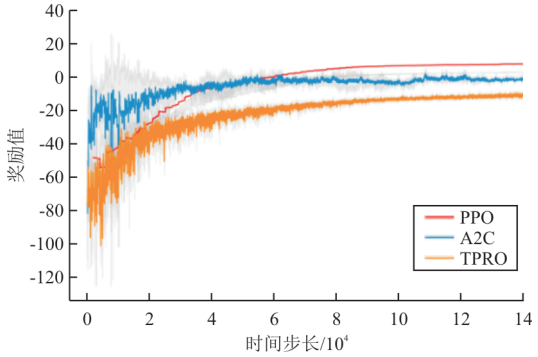

图 9　奖励曲线性能对比

Fig. 9　Comparison of episode rewards

从图 9 可以看出:A2C 和 TRPO 模型的方差大、训练稳定性低、收敛效果差,而 PPO 模型获得平均奖励更高,且曲线上升平滑稳定,收敛性能优异。

4.3.2　在线规划

在环境中对训练好的 PPO 模型进行测试,输入初始状态,检验模型的规划能力,结果如图 10~图 11 所示。

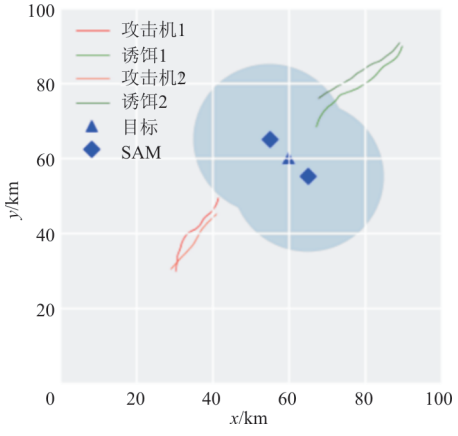

图 10　攻击过程Ⅰ(诱饵先进入)

Fig. 10　Attack process Ⅰ (decoy entry early)

从图 10 可以看出:诱饵径直飞向地导,吸引敌方雷达跟踪和导弹发射,攻击机伺机绕飞等待。

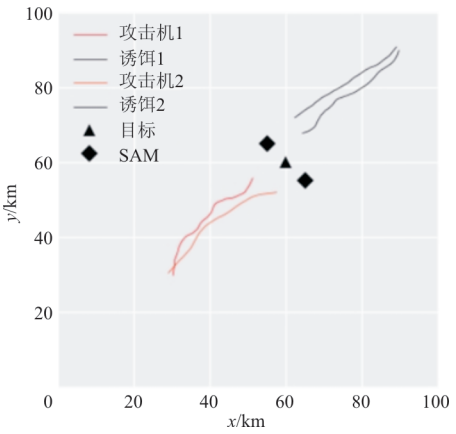

图 11　攻击过程Ⅰ(攻击机趁机攻击)

Fig. 11　Attack progress Ⅰ (attacker takes opportunity to attack)

从图 11 可以看出:攻击机利用地导跟踪锁定诱饵并发射导弹攻击的时间差,快速发起攻击,成功打掉地导和敌方保护目标,诱饵牺牲,任务完成,表明本文所提 PPO 智能规划模型具有一定战术协同规划能力。

4.4　实验二:动态场景鲁棒性测试

为测试本文智能规划模型的鲁棒性,对训练环境增加一定的随机性,如图 12 所示。

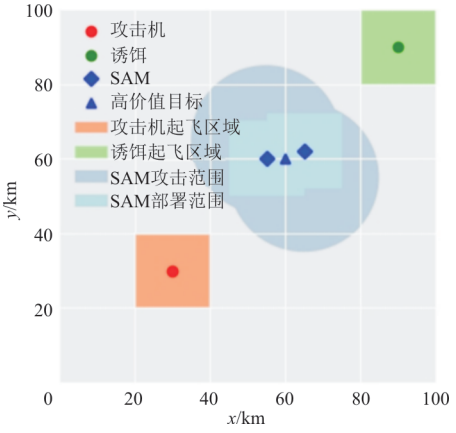

图 12　不同初始位置的 SEAD 环境

Fig. 12　Environment of SEAD in different initial area in this work

攻击机、诱饵机和地导起始位置可以在周围一定区域中随机变化,检验训练好的模型在未知

环境中的泛化能力和鲁棒性，结果如图 13～图 14 所示。

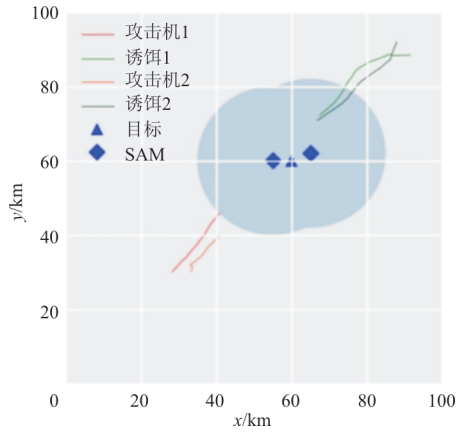

图 13 攻击过程Ⅱ（诱饵先进入）

Fig. 13 Attack process Ⅱ (decoy entry early)

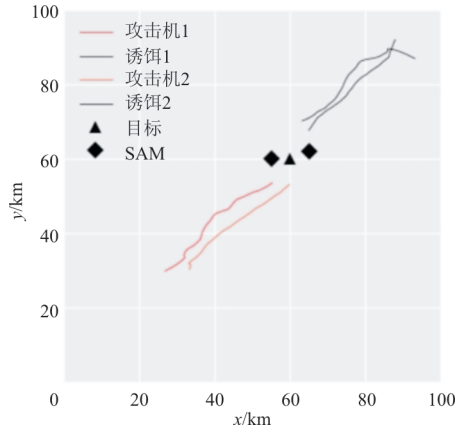

图 14 攻击过程Ⅱ（攻击机趁机攻击）

Fig. 14 Attack process Ⅱ (attacker takes opportunity to attack)

图 13 中，将攻击机初始位置(30,30)，诱饵初始位置(90,90)，地导位置(65,62)、(55,60)，高价值目标位置(60,60)输入到训练好的模型中，发现攻击机依然会智能地等待诱饵先进入地导范围，然后趁机迅速攻击，成功打掉地导和目标(如图 14 所示)。智能规划模型对未知的输入具有一定鲁棒性和泛化能力，可以适应未知的、不确定的环境，具有较强的实际应用价值。

4.5 消融实验

为了对比本文智能规划模型训练中的不同训练技巧对算法性能的影响，即对比使用全部技巧的模型和优势函数标准化、值函数标准化、自适应 clip 的技巧对模型性能的影响，进行消融实验，结果如图 15 所示。

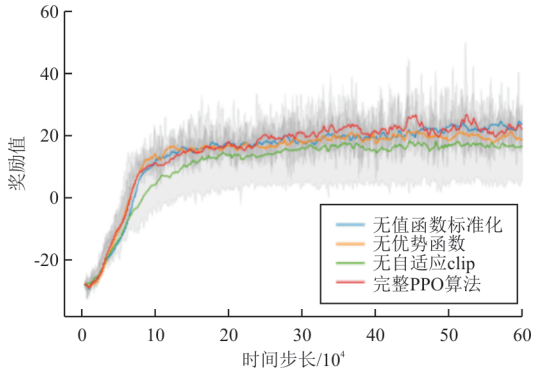

图 15 不同训练技巧的平均奖励值对比

Fig. 15 Averaged reward comparison of different tricks

从图 15 可以看出：使用全部训练技巧的模型平均奖励值更高，最终性能更好，自适应 clip 会大幅提高模型前期收敛速度，优势函数标准化可以提高模型的最终性能，值函数标准化对模型性能影响相对较小。

5 结 论

(1) 针对多无人机协同 SEAD 作战任务，本文提出了一种基于深度强化学习的端到端的协同作战智能规划方法，建立了基于 PPO 算法的 SEAD 作战智能规划模型，通过离线训练—在线规划框架，实现快速任务规划。

(2) 仿真验证了所提模型的有效性和鲁棒性，得出本文提出的基于 DRL 的多无人机协同智能规划方法具有快速、精细、协同的优点。

(3) 引入了域随机化、最大化策略熵和底层网络参数共享三种策略训练技巧，消融实验发现自适应 clip 可以提高模型的收敛速度，优势函数标准化能够大幅提高模型的最终性能。

接下来的工作是将这种端到端的方法推广至大规模复杂作战场景，建立高保真的飞机模型和导弹模型，针对更复杂的多智能体规划问题展开深入研究。

参考文献

[1] STEVENS R, SADJADI F. Small unmanned aerial vehicle real-time intelligence, surveillance and reconnaissance (ISR) using onboard pre-processing [C] // SPIE Defense and Security Symposium. Orlando: SPIE, 2008: 1-8.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] SUTTON R, BARTO A. Reinforcement learning: an introduction[M]. US: MIT Press, 1998.

[5] US Department of Defense. Summary of the 2018 department of defense artificial intelligence strategy [EB/OL]. (2019-02-12)[2022-01-13]. https:// media. defense. gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY. PDF.

[6] CLARK B, PATT D, SCHRAMM H. Mosaic warfare: exploiting artificial intelligence and autonomous systems to implement decision-centric operations [M]. Washington DC: Center for Strategic and Budgetary Assessments, 2020.

[7] 张睿文, 宋笔锋, 裴扬, 等. 复杂任务场景无人机集群自组织侦察建模与仿真[J]. 航空工程进展, 2020,11(3): 316-325,343.
ZHANG Ruiwen, SONG Bifeng, PEI Yang, et al. Modeling and simulation of UAV swarm self-organized surveillance in complex mission scenarios[J]. Advances in Aeronautical Science and Engineering, 2020, 11(3): 316-325, 343. (in Chinese)

[8] 谭威, 胡永江, 李文广, 等. 多无人机协同任务规划研究综述[J]. 微型电脑应用, 2021, 37(9): 189-192.
TAN Wei, HU Yongjiang, LI Wenguang, et al. A survey of multi UAV cooperative mission planning[J]. Microcomputer Application, 2021, 37(9): 189-192. (in Chinese)

[9] ZHANG H, YANG R N, WU J, et al. Research on multi-aircraft cooperative suppressing jamming embattling in electronic warfare planning[J]. System Engineering and Electronics, 2017, 39(3): 542-548.

[10] SMIRNOV E A, TIMOSHENKO D M, ANDRIANOV S N. Comparison of regularization methods for ImageNet classification with deep convolutional neural networks[J]. AASRI Procedia, 2014(6): 84-96.

[11] 潘楠, 刘海石, 陈启用, 等. 多基地多目标无人机协同任务规划算法研究[J]. 现代防御技术, 2021, 49(2): 49-56.
PAN Nan, LIU Haishi, CHEN Qiyong, et al. Research on cooperative mission planning algorithm of multistatic and multi-target UAV[J]. Modern Defense Technology, 2021, 49(2): 49-56. (in Chinese)

[12] 辛建霖, 左家亮, 岳龙飞, 等. 基于改进启发式蚁群算法的无人机自主航迹规划[J]. 航空工程进展, 2022, 13(1): 60-67.
XIN Jianlin, ZUO Jialiang, YUE Longfei, et al. Autonomous path planning for unmanned aerial vehicle based on improved heuristic ant colony algorithm[J]. Advances in Aeronautical Science and Engineering, 2022, 13(1): 60-67. (in Chinese)

[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] SILVER D, HUBERT T, SCHRITTWIESER J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm [EB/OL]. [2022-01-13]. https: //ai. dmi. unibas. ch/research/reading_group/silver-et-al-arxiv2017. pdf.

[15] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft Ⅱ using multi-agent reinforcement learning[J]. Nature, 2019, 575: 350-354.

[16] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science Robot, 2019, 4(26): 1-8.

[17] KENDALL A, HAWKE J, JANZ D, et al. Learning to drive in a day [EB/OL]. [2022-01-13]. https: // arxiv. org/pdf/1807. 00412. pdf.

[18] YE D H, LIU Z, SUN M F, et al. Mastering complex control in MOBA games with deep reinforcement learning[C]// 2020 AAAI Conference on Artificial Intelligence. [S. l.]: AIAA, 2020: 1-10.

[19] HU H, ZHANG X, YAN X, et al. Solving a new 3D bin packing problem with deep reinforcement learning method [EB/OL]. [2022-01-13]. https: // arxiv. org/pdf/1708. 05930. pdf.

[20] 疏利生, 李桂芳, 嵇胜. 基于强化学习的航空器机场智能静态路径规划[J]. 航空工程进展, 2021, 12(3): 65-70.
SHU Lisheng, LI Guifang, JI Sheng. Aircraft AI static path planning on airport ground based on reinforcement learning [J]. Advances in Aeronautical Science and Engineering, 2021, 12(3): 65-70. (in Chinese)

[21] BARKDOLL T C, GAVER D P, GLAZEBROOK K D, et al. Suppression of enemy air defenses (SEAD) as an information duel[J]. Naval Research Logistics, 2002, 49: 723-742.

[22] HAQUE M, EGERSTEDT M, RAHMANI A. Multilevel coalition formation strategy for suppression of enemy air defenses missions[J]. Journal of Aerospace Information Systems, 2013, 10(6): 287-296.

[23] ZHANG L A, JIA X, DARA G, et al. Air dominance through machine learning: a preliminary exploration of artificial intelligence-assisted mission planning[EB/OL]. [2022-01-13]. https: //www. rand. org/pubs/research_reports/RR4311. html.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning [EB/OL]. [2022-01-13]. http: // citeseerx. ist. psu. edu/viewdoc/download; jsessionid=06FDDD274D336653D404F1C3B3C28787? doi=10. 1. 1. 467. 6642&rep=rep1&type=pdf.

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

(下转第 96 页)