Journal of System Simulation

Volume 35 | Issue 8 Article 7

8-25-2023

Intelligent Air Defense Task Assignment Based on Assignment Strategy Optimization Algorithm

Jiayi Liu
Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China; Graduate College, Air Force Engineering University, Xi'an 710051, China, sixandone1@163.com

Gang Wang
Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China

Qiang Fu
Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China

Xiangke Guo
Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China

See next page for additional authors

Intelligent Air Defense Task Assignment Based on Assignment Strategy Optimization Algorithm

Abstract

Abstract: Aiming at the insufficient solving speed of assignment strategy optimization algorithm in largescale scenarios, *deep reinforcement learning is combined with Markov decision process to carry out the intelligent large-scale air defense task assignment*. According to the characteristics of large-scale air defense operations, *Markov decision process is used to model the agent and a digital battlefield simulation environment is built. Air defense task assignment agent is designed and trained in digital battlefield simulation environment through proximal policy optimization algorithm*. The feasibility and advantage of the method are verified by taking a large-scale ground-to-air countermeasure mission as an example.

Keywords

assignment strategy optimization algorithm, task assignment, Markov decision process, deep reinforcement learning, agent

Authors

Jiayi Liu, Gang Wang, Qiang Fu, Xiangke Guo, and Siyuan Wang

Recommended Citation

Liu Jiayi, Wang Gang, Fu Qiang, et al. Intelligent Air Defense Task Assignment Based on Assignment Strategy Optimization Algorithm[J]. Journal of System Simulation, 2023, 35(8): 1705-1716.

第 35 卷第 8 期 系统仿真学报© Vol. 35 No. 8
2023 年 8 月 Journal of System Simulation Aug. 2023

基于分配策略优化算法的智能防空任务分配

刘家义[1,2]，王刚[1]，付强[1*]，郭相科[1]，王思远[1,2]
(1. 空军工程大学 防空反导学院，陕西 西安 710051；2. 空军工程大学 研究生院，陕西 西安 710051)

摘要：针对分配策略最优算法在大规模场景中求解速度不足的问题，*基于马尔可夫决策过程，将深度强化学习与其相结合，将大规模防空任务分配问题进行智能化求解*。根据大规模防空作战特点，*利用马尔可夫决策过程对智能体进行建模，构建数字战场仿真环境；设计防空任务分配智能体，通过近端策略优化算法，在数字战场仿真环境中进行训练*。以大规模防空对抗任务为例，验证了该方法的可行性和优越性。

关键词：分配策略优化算法；任务分配；马尔可夫决策过程；深度强化学习；智能体
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2023)08-1705-12
DOI: 10.16182/j.issn1004731x.joss.22-0432
引用格式: 刘家义, 王刚, 付强, 等. 基于分配策略优化算法的智能防空任务分配[J]. 系统仿真学报, 2023, 35(8): 1705-1716.
Reference format: Liu Jiayi, Wang Gang, Fu Qiang, et al. Intelligent Air Defense Task Assignment Based on Assignment Strategy Optimization Algorithm[J]. Journal of System Simulation, 2023, 35(8): 1705-1716.

Intelligent Air Defense Task Assignment Based on Assignment Strategy Optimization Algorithm

Liu Jiayi[1,2], *Wang Gang*[1], *Fu Qiang*[1*], *Guo Xiangke*[1], *Wang Siyuan*[1,2]
(1. Air and Missile Defense College, Air Force Engineering University, Xi'an 710051, China;
2. Graduate College, Air Force Engineering University, Xi'an 710051, China)

Abstract: Aiming at the insufficient solving speed of assignment strategy optimization algorithm in large-scale scenarios, *deep reinforcement learning is combined with Markov decision process to carry out the intelligent large-scale air defense task assignment*. According to the characteristics of large-scale air defense operations, *Markov decision process is used to model the agent and a digital battlefield simulation environment is built. Air defense task assignment agent is designed and trained in digital battlefield simulation environment through proximal policy optimization algorithm*. The feasibility and advantage of the method are verified by taking a large-scale ground-to-air countermeasure mission as an example.

Keywords: assignment strategy optimization algorithm; task assignment; Markov decision process; deep reinforcement learning; agent

0 引言

防空反导作战实际上是一个持续决策的过程，需要针对战场局势的变化，作出适应性较好的决策，任务分配是其中的重要一环，其目的是合理分配资源、最大化防空作战效能。现有的研究中常提及目标分配和任务分配[1-4]两个概念，二者存

收稿日期：2022-04-29 修回日期：2022-06-30
基金项目：国家自然科学基金(62106283)
第一作者：刘家义(1996-)，男，博士生，研究方向为深度强化学习、智能辅助决策。E-mail：sixandone1@163.com
通讯作者：付强(1988-)，男，副教授，博士，研究方向为智能辅助决策、指控模型。E-mail：fuqiang_66688@163.com

在很多共性，但又不完全相同。任务分配可看作是在目标分配基础上提出的概念，当作战任务被分解为不同类型的任务后，目标分配将转化为任务分配[5]。本文结合目标分配和任务分配的研究成果，针对大规模防空作战的任务分配问题进行研究。

目前，大多数研究都是单次静/动态打击式目标分配，但防空作战是一个动态过程，在此过程中，其面临的威胁可能是大规模的体系空袭，也可能是小规模的战术偷袭，同时，火力单元和来袭目标的数量也在不断变化。因此，动态武器目标分配(dynamic weapon target assignment, DWTA)是防空反导指控系统亟待解决的重要理论问题[6]。DWTA的研究主要有多级武器–目标分配[7-8]、基于马尔可夫决策过程最优化的分配策略优化算法[9]。尽管这些方法在不断改进，但是对大规模的武器目标分配问题的求解速度仍然略显不足[10]。

深度强化学习(DRL)是深度学习(DL)与强化学习(RL)的结合，20世纪90年代以来，其发展为指控系统的智能化提供了动力，其和指控系统的结合在协同作战、精准制导等方面产生了巨大的效应[11]。其利用马尔可夫决策过程(Markov decision process, MDP)对智能体及其交互环境完成建模后，即可利用相应的方法对问题进行求解，具有较快的反应性和较高的动态性[12]。因此，本研究基于MDP将分配策略最优算法结合DRL方法，利用深度神经网络的高速运算能力求解MDP，弥补了分配策略最优算法在求解速度上的不足，解决了大规模防空任务分配问题。

1 相关工作

1.1 分配策略优化算法

目标分配可以分为静态和动态。其中，DWTA考虑了战场态势随时间而变化，比静态的目标分配更切合实际问题的需要，逐渐成为研究的热点。但DWTA的求解也因为约束条件多而面临着计算复杂度的挑战。在DWTA的求解方法中，有一类方法称为分配策略优化算法，此类方法利用了MDP的动态性来求解该问题[13-14]。其中，影响较大的是韩松臣的基于马尔可夫决策的动态WTA过程[15]，提出可基于马尔可夫动态系统，通过随机服务系统输入过程的最优控制，建立目标分配决策模型[15]，在一定假设条件下，将DWTA分为策略优化和匹配优化2个阶段。陈英武等在此基础上用五元组$\{\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}, r, V\}$定义DWTA的MDP，提出了一种混合的最优策略改进算法，其中，$\boldsymbol{S}$为状态空间，$\boldsymbol{A}$为方案集合，$\boldsymbol{P}$为转移概率矩阵，r为收益函数，V为目标函数。用MDP的无限阶段平均模型(式1)来描述目标函数V，用来求解大规模的DWTA问题[16]。

$$V(\pi,i)=\lim_{N\to\infty}\frac{1}{N}\sum_{s=1}^{N}E(R_s(i,\pi)|n_1=i,\pi)=\lim_{N\to\infty}\frac{1}{N}\sum_{s=1}^{N}\sum_{\pi_j\in\pi,j\in S}r(\pi_j)p\{n_s=j|n_1=i,\pi_j\} \tag{1}$$

式中：$V(\pi,i)$为武器系统从状态i出发所获得的长期平均收益；$p\{n_s=j|n_1=i,\pi_j\}$为武器系统采用策略π在第一个目标到达时所处状态为i，于第s个目标到达时转移到状态j的$s-1$步转移概率。

何鹏等[17]将策略分配优化算法应用于任务分配问题中，将其描述为一个分阶段的序列决策过程，在小规模任务分配寻优中效果较为理想。尽管许多研究在不断改进分配策略最优算法[18-20]，但依然无法完全解决计算复杂度的难题，在求解大规模DWTA时速度仍略显不足，实时性不太理想[21]。

1.2 深度强化学习

RL的思路是利用试错法和奖励来训练智能体学习行为。RL的基本环境是一个马尔可夫决策过程。一个马尔可夫决策过程有五元素，即$\langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}, \boldsymbol{P}, \gamma \rangle$，其中，$\boldsymbol{S}$代表状态集合，$\boldsymbol{A}$代表动作集合，$\boldsymbol{R}$代表奖励函数，$\boldsymbol{P}$代表状态转移概率，$\gamma$代表折扣因子。其基本框架如图1所示。

DOI: 10.16182/j.issn1004731x.joss.22-0432

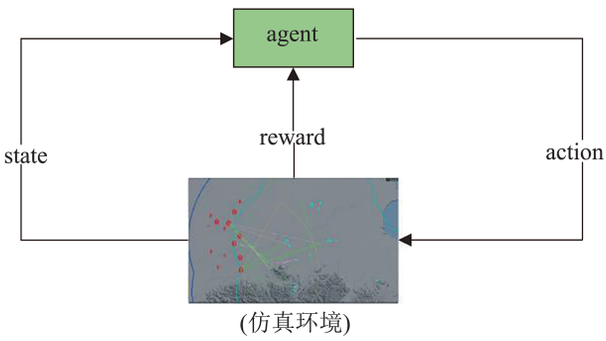


图1 强化学习基本框架
Fig. 1 Basic framework of reinforcement learning

智能体(agent)从环境中感知当前状态(state)，然后做出相应的行为(action)，得到对应的奖励(reward)。然而在实际问题中，状态往往十分复杂，导致传统RL存在维数灾难的问题[22]。DL利用深度神经网络作为函数拟合器，与RL结合形成了DRL[23]，有效解决了维数灾难的问题[24]。其中，DQN算法[25]将卷积神经网络和Q学习结合用于决策，在自动驾驶、机器人控制、无人机导航等多个领域取得应用成果[26-28]。本研究旨在将解决动态目标分配问题的分配策略优化算法，用于求解任务分配问题，同时结合DRL方法，克服分配策略优化算法在大规模场景中求解速度上的不足。

2 问题描述

2.1 目标分配与任务分配

工作任务分配与工作目标分配两者求解的问题模型以及解决问题的方法大同小异，分配本质是一样的[29]。随着武器系统和作战方式的不断发展，目标分配问题显示出一些局限性，而任务分配改变了目标分配火力单元–目标的模式，形成任务–目标的分配模式，在火力单元和目标数都相同的情况下，任务分配较目标分配有以下优势：

(1) 任务分配更加灵活，有更多分配结果供选择。将任务分解为跟踪任务和拦截任务，此时将传感器和拦截器灵活组合，可以虚拟出更多的火力单元。

(2) 任务分配抗毁性更强。在目标分配中，若火力单元的传感器或发射装置遭摧毁，这个火力单元将不能继续作战。而在任务分配模式下，只要该火力单元还可以完成部分协同作战任务，就可以继续参加任务分配。

(3) 任务分配可实施性更强。具体的拦截过程涉及到多个子任务，这些任务之间有较强的时间与空间的约束，任务分配可以对这些子任务进行合理配置，最大化作战效能。

虽然任务分配有许多优势，但面对大规模复杂场景，还需要具有以下几种能力：

(1) 实时的态势处理能力

随着空袭网络化作战的发展，高实时、高动态的战场态势成为防空反导作战的主要挑战之一。因此，必须具有实时的战场态势分析和处理能力。

(2) 动态的要素调配能力

基于要素的集成分布式协同作战是应对空域网络化的发展趋势。分散部署的要素资源需要进行协同作战，形成虚拟作战联盟，以作战要素集成的方式动态调配、灵活组合。需要动态的要素调配能力。

(3) 高速的信息计算能力

基于要素集成的作战模式带来了武器组合的爆炸式增长，大量的实时信息数据处理成为主要挑战之一。高速的信息计算能力是实时地在众多组合之中快速寻找最优结果、最大化作战效能的根本保证。

2.2 智能防空任务分配

为充分发挥任务分配的优势并最大程度上达到上述3种能力，本文基于分配策略最优算法的思想，将该问题建模为MDP并用DRL来求解，用智能化的方法增强实时性和计算能力。本文研究的是大规模防空任务分配问题，目的是在保护对象受损最小时使用最少的资源。因此本研究的优化目标为求解最优的策略函数π^*，最大化期望累积奖励值为

$$\max_{\pi} E\left[\sum_{t=0}^{T}\gamma_t r_t\right]$$
$$\text{s.t.}\quad s_{t+1}\sim p(\cdot|s_t,a_t), a_t\sim\pi(\cdot|s_t), t=0,1,\cdots,T-1 \tag{2}$$

式中：$p(\cdot|s_t,a_t)$为t时刻的状态转移概率。此时任务分配问题转化成了利用RL算法在状态转移概率未知情况下求解MDP，RL算法求解的核心思路是采用时间差分方法估计动作–值函数：

$$Q^{\pi}(s,a)=E\left[\sum_{t}^{T}\gamma_t r_t|s_0=s,a_0=a\right] \tag{3}$$

$$Q'(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r(s_t,a_t)+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{4}$$

$$\pi'(s)=\arg\max_a Q^{\pi}(s,a) \tag{5}$$

式中：$Q^{\pi}(s,a)$为状态动作值–函数，表示在状态s下执行动作a，后续动作选取遵从策略π所获得的期望总奖励；α为学习率，表示新信息对旧信息的影响程度；Q'为更新后的估计值。

2.3 MDP建模

对于DRL而言，状态空间、动作空间和奖励函数的定义都十分重要，必须满足合理性和完整性，本文的状态空间、动作空间和奖励函数设计如下。

状态空间：红方受保护的单位状态、传感器状态、拦截器状态；蓝方单位基本信息以及可跟踪和可拦截的蓝方单位的状态。

动作空间：动作分为选择跟踪的单位、选择拦截的单位、选择拦截的时机和用于拦截的资源数量。

奖励函数：如果只在每局最后一步给出胜利或者失败的奖励值，可以给智能体最大限度的学习空间，但会导致奖励值过于稀疏，智能体探索到获胜状态的概率很低。为了较好地平衡智能体的探索和学习，本文的奖励函数为

$$R=\begin{cases}5m+2n-5i+j-100, & \text{失败}\\ 5m+2n-5i+j, & \text{胜利}\end{cases} \tag{6}$$

式中：m为拦截高价值数量；n为拦截高威胁单位数量；j为拦截空对地导弹数量；i为要地被攻击次数。拦截高价值单位加5分，拦截高威胁目标加2分，拦截空对地导弹加1分，要地被攻击1次扣5分，超过3次判定为失败，扣100分。

3 基于保卫要地任务的环境设计

在DRL的训练中，智能体与环境交互进行试错是十分关键的环节。为了解决军事博弈对抗场景交互试错成本高的难题，在前期工作中[30]已构建了一个高仿真度的数字战场，将物理环境较好地映射到虚拟环境中。本研究在智能化目标分配的基础上，依据任务分配问题的需求将仿真环境设计进一步完善。

3.1 交互场景

数字战场主要负责战场环境的呈现和交互过程的模拟，包括模拟每个单位的行为逻辑和互相攻击的毁伤计算。根据任务分配特点将各个单位分为传感器与拦截器，具体交互环境如图2所示。

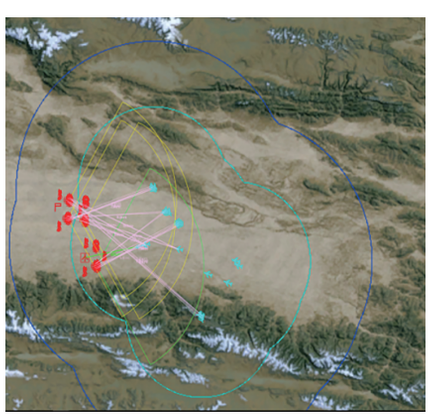


图2 交互环境
Fig. 2 Interaction environment

3.2 数据交互

本研究在数据交互流程中增加了协议模块，包含了数字战场与智能体交互的接口，主要作用是将数字战场与智能体之间交互信息的序列化、传输和反序列化。一次完整的数据交互流程如图3所示。

DOI: 10.16182/j.issn1004731x.joss.22-0432

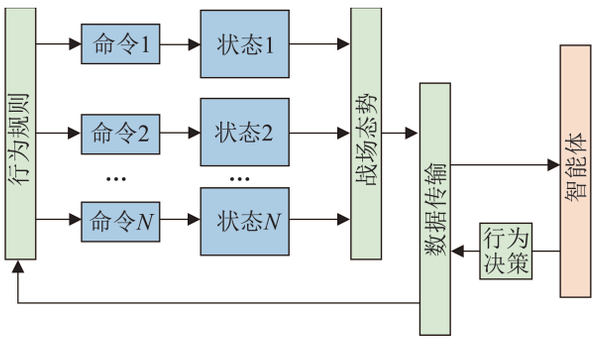


图3　数据流程
Fig. 3　Data flow chart

4　面向防空任务分配的深度强化学习方法

4.1　训练框架设计

在使用DRL方法求解问题之前，需要先对智能体进行训练，通过不断与环境交互，让智能体学习到有效策略，优化神经网络参数。本文的智能体训练框架如图4所示。

在交互方面，智能体需要输入的是状态信息和奖励，输出则是动作信息，而仿真环境需要输入的是作战指令，输出的是战场态势信息。因此，智能体要和环境进行交互，需要根据定义的MDP模型，将环境输出的数据转换为状态信息，将智能体输出的动作转换为作战指令。在训练方面，智能体将与环境交互得到的数据输入RL算法，通过计算出的*loss*来更新网络参数。如此迭代，不断优化智能体的策略。

4.2　训练网络结构设计

深度神经网络是DRL方法解决大规模复杂问题的关键，网络结构的设计必须符合场景需求。结合3.2节中的MDP模型和大规模防空任务分配问题需要，设计网络结构如图5所示。

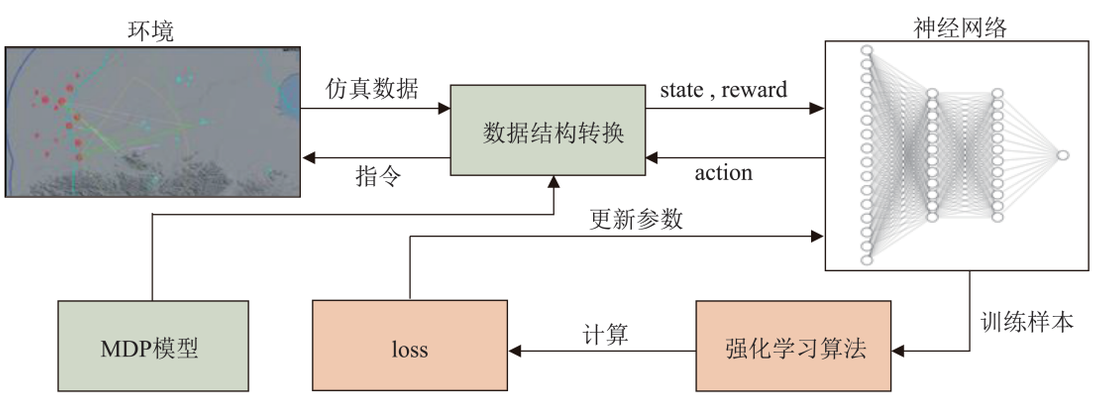


图4　智能体训练框架
Fig. 4　Agent training framework

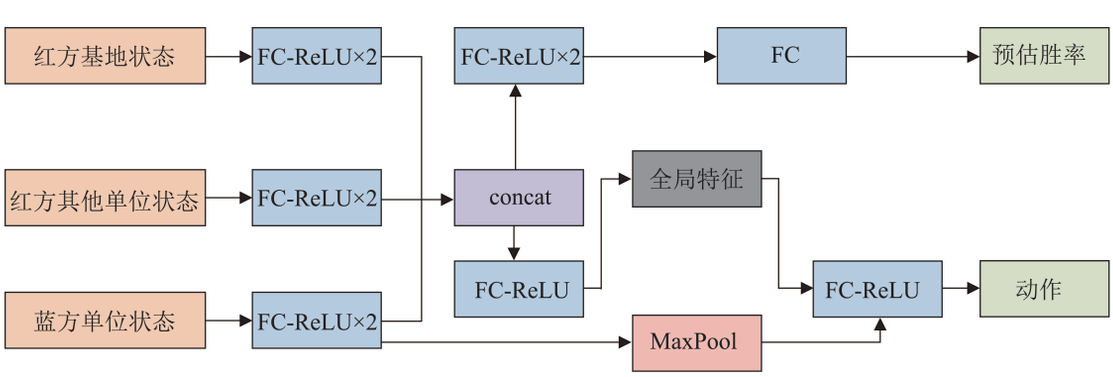


图5　神经网络结构
Fig. 5　Neural network structure

分别输入状态空间定义的几种状态，经过2层FC-ReLU层进行特征提取后再合并作为基础数据，分别输入到价值网络和策略网络。在价值网络中，基础数据再经过2层FC-ReLU层和1层FC层，输出当前态势下的预估胜率，作为评价此阶段决策好坏的一个指标。在策略网络中，基础数据经过1层FC-ReLU层形成全局特征，与经过特征提取后的蓝方单位状态信息一起输入到FC-ReLU层，继而输出动作。

4.3 近端策略优化算法

如何快速训练智能体，优化神经网络参数，让智能体输出高水平的策略，也是本研究的核心问题之一。本研究选用近端策略优化(proximal policy optimization, PPO)算法作为图5中的RL算法，用于优化神经网络参数。PPO算法直接优化策略函数$\pi_\theta(a|s)$，其中，s为状态，a为动作，计算累积期望回报的策略梯度，保证每步迭代获得一个“更好”的策略，进而得到使整体回报最大化的策略参数θ。对于PPO中的损失函数，也有不同的定义方法，如无裁剪或惩罚、带裁剪、带KL惩罚等，从MuJoCo实验[31]来看，带裁剪的PPO实现简单，而且效果更好。因此，本文中采取的是带裁剪的PPO，算法具体内容如下。

算法1　PPO算法

初始化策略参数θ，θ_{old}

重复每轮更新

　　重复每个Actor

　　重复T步

　　每步使用旧的策略参数θ_{old}产生决策

　　计算每一步中的优势估计A

迭代K步

　　求解累积期望回报函数的策略梯度，每次使用小批量数据

　　用策略梯度更新θ策略参数

更新新的策略参数至θ_{old}

算法1中的θ_{old}与θ分别指的是策略近似函数的旧参数与新参数，也可描述为更新前的策略函数与当前新的策略函数。此算法的累积期望回报目标函数为

$$L_t(\theta)=\min(r_t(\theta)A_t,\text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)A_t) \quad (7)$$

式中：$r_t(\theta)$为采用新旧策略函数概率的比值；ε为裁剪系数。本研究使用的PPO中的裁剪系数$\varepsilon=0.2$，学习率为10^{-4}，批尺寸为5 120，神经网络中隐藏层单元数分别为128和256。当$r_t(\theta)\notin[1-\varepsilon,1+\varepsilon]$时，优势函数$A_t$被裁剪，使得在旧策略函数基础上进行多次更新，同时避免更新后的策略函数偏离原来的策略函数过大。

5　实验与结果

5.1 数字战场仿真环境

5.1.1 对抗场景设置

为了验证本文方法的可行性和优越性，以大规模防空任务为例，红方单位包括6个远程火力单元和6个近程火力单元以及预警机1架。需要保卫的要地为1个指挥所和1个机场。其中，远程火力单元由1个远程传感器和8个远程拦截器组成，近程火力单元由1个近程传感器和3个近程拦截器组成。

蓝方设置18枚巡航弹，20架无人机，12架战斗机，4架轰炸机和2架干扰机，分批对红方进攻。蓝方编队规模、作战任务是固定的，但各批次的突防路线和到达时间是随机的。

第1个批次由18枚巡航导弹分为2条突防路线攻击指挥所及机场，巡航弹飞行高度100 m进行超低空突防，红方必须合理规划资源，在拦截的前提下让弹药资源消耗最小；第2批次为由20架无人机2~3 km高度突防、12架战斗机飞行高度100 m超低空突防，并且摧毁第1批次进攻后暴露的火力单元；第3个批次为由4架轰炸机突防轰炸要地。

5.1.2 对抗准则设置

远程传感器最大探测距离为200 km，扇区为120°，近程传感器最大探测距离为60 km，扇区为

DOI: 10.16182/j.issn1004731x.joss.22-0432

360°；制导过程传感器需要全程开机，开机时会暴露自身位置；防空导弹拦截远界为160 km(远程)、40 km(近程)，针对无人机、战斗机、轰炸机、反辐射导弹、空对地导弹在杀伤区的高杀伤概率为75%，低杀伤概率为55%，针对巡航导弹在杀伤区的高杀伤概率为45%，低杀伤概率为35%；反辐射导弹射程为110 km，命中率为80%；空对地导弹射程为60 km，命中率为80%；蓝方干扰机干扰扇区为15°，红方传感器受到干扰后，根据干扰等级，相应降低杀伤概率。

当红方指挥所受到3次攻击时红方失败；当蓝方轰炸机与红方指挥所之间的距离小于10 km时，红方失败；当红方传感器损失超过60%时，红方失败；当蓝方损失的战斗机超过30%时，红方胜利。

5.2 实验硬件配置

仿真环境配置：CPU Intel Xeon E5-2678V3，88核，256 G内存；训练环境配置：GPU*2型号NVIDIA GeForce 1080Ti，72核，11 G显存。

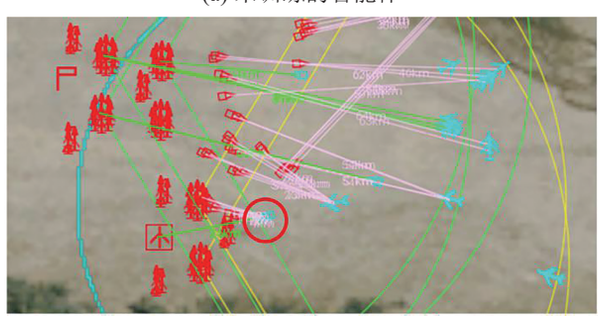
5.3 实验1：可行性验证

不同于分配策略优化算法，本研究利用DRL方法对MDP进行动态求解，但在此之前需要通过训练来优化神经网络参数。因此，需要验证训练后的智能体是否可以学习到有效的任务分配策略，成功保卫要地。在实验1中，将智能体按照第5节中的DRL方法训练100 000次，智能体在对抗中获得的奖励值可以达到65左右，胜率达到55%左右。

将未训练过的智能体与训练100 000次的智能体分别在环境中进行推演，行为对比如图6所示。

从行为对比可以看出，未训练过的智能体采用随机策略，没有拦截正在攻击要地的目标，几乎无法取胜；经过训练的智能体可以学习到有效策略，保卫要地的同时拦截高价值目标。

(a) 未训练的智能体



(b) 训练后的智能体

图6　相同想定的行为对比
Fig. 6　Behavior comparison of same scenarios

5.4 实验2：任务分配优势验证

在大规模复杂场景下，任务分配模式比目标分配模式更具优势，分配结果更加灵活。为了验证这一点，在本实验中，任务分配智能体的MDP建模如2.3节所示。如2.1节所描述，本实验中，任务分配模式下，火力单元中的传感器和拦截器只要未被摧毁，就可以继续分配，且各个火力单元的传感器与拦截器可以自由组合。对于目标分配智能体而言，采用火力单元–目标的模式，一个火力单元内的拦截器只能由该火力单元的传感器指挥，且当传感器遭到攻击损坏后该火力单元即丧失作战能力，不再参加目标分配。

5.4.1 训练数据对比分析

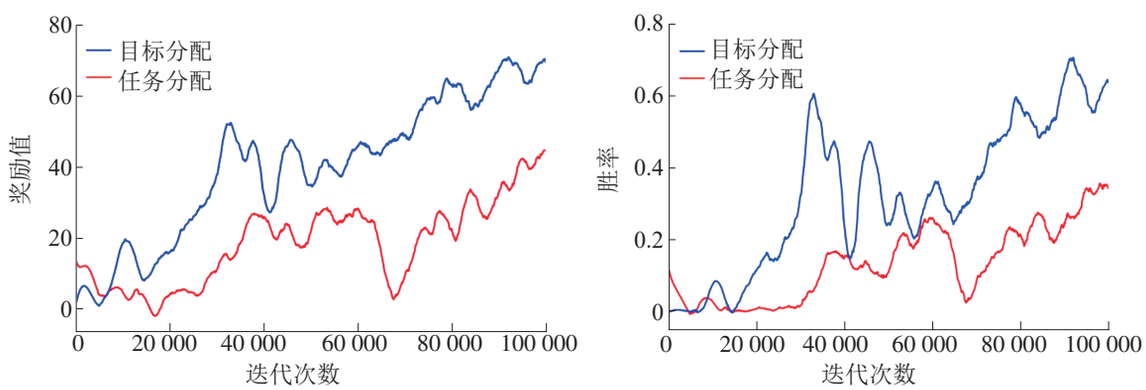
将2个智能体在5.1节的想定中迭代训练100 000次，对比结果如图7所示。

可以看出，通过训练，任务分配智能体和目标分配智能体的决策水平均得到了提升，与目标分配智能体相比，任务分配智能体可以在相同时间步内获得更高的胜率和奖励值。

5.4.2 数字战场单局结果对比

在本实验中，用专家规则库的方法求解MDP模型，作为传统方法与智能体模型进行对比。在训练结束后，将2个智能体模型以及传统方法模型分别放入数字战场进行离线推演，对战结束前一刻输出的战斗结果示意图如图8所示。对战过程中的战况统计如图9所示。

可以看出，根据专家规则库求解的传统方法并不能抵挡住第1批次的进攻，因传感器损失过多而失败。目标分配智能体和任务分配智能体均能够抵挡住第1批次的巡航弹进攻，但第2批次对于红方而言，防御压力很大，既要拦截无人机和战斗机，也要拦截所有作战飞机发射的大量空对地导弹和反辐射弹。由于目标分配模式下多个火力单元间的传感器和拦截器不能灵活组合，在第2批次进攻时，因攻击范围和资源饱和等问题，大部分资源用于拦截无人机和空对地导弹，大多处于被动防御状态，最后因轰炸机距离太近而失败。相比之下，任务分配智能体可以在第1批次时精准拦截，节省更多弹药资源，并且在第2批次进攻时，利用灵活协同、可实施性强的优势，迅速打击蓝方战斗机，从而赢得胜利。



(a) 奖励值对比　　(b) 胜率对比

图7　训练结果对比

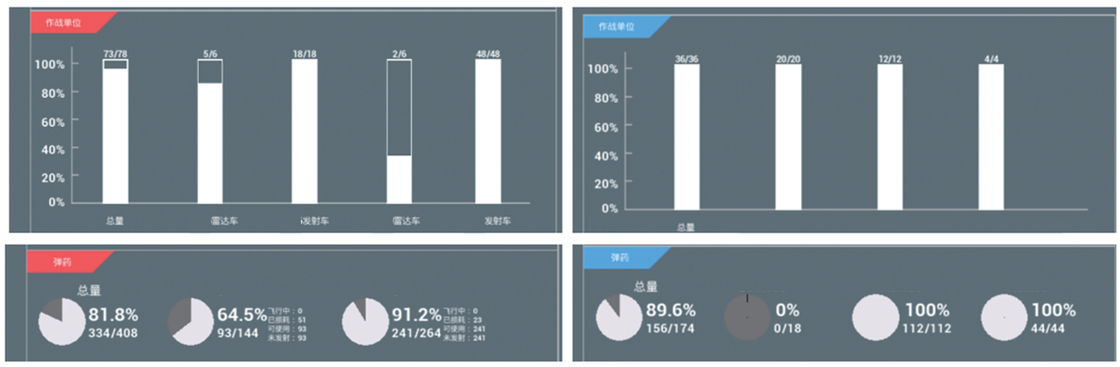
Fig. 7　Comparison of training results

	发射车#01	发射车#02	发射车#03	发射车#04	发射车#05	发射车#06	发射车#07
巡航导弹#01	无	无	无	无	无	无	无
巡航导弹#08	无	无	无	无	无	无	无
巡航导弹#15	射击1次，未杀伤	射击1次，未杀伤	无	射击1次，杀伤	无	无	无
巡航导弹#03	无	无	无	无	无	无	无
巡航导弹#11	无	无	无	无	无	无	无
巡航导弹#06	无	无	无	无	无	无	无
巡航导弹#02	无	无	无	无	无	无	射击1次，未杀伤
巡航导弹#04	无	无	无	无	无	无	无
巡航导弹#18	无	无	无	射击1次，未杀伤	无	无	无
巡航导弹#10	无	无	无	无	无	无	无
巡航导弹#16	无	射击2次，未杀伤	无	无	无	射击2次，杀伤	无
巡航导弹#17	无	无	射击2次，杀伤	无	无	无	射击1次，未杀伤
巡航导弹#07	无	无	无	无	无	射击1次，未杀伤	无
巡航导弹#05	无	无	无	无	射击2次，杀伤	无	无
巡航导弹#14	射击1次，未杀伤	无	无	无	射击1次，未杀伤	无	无
巡航导弹#09	无	无	射击1次，未杀伤	无	无	无	射击1次，未杀伤
巡航导弹#12	射击1次，未杀伤	无	无	无	无	无	无
巡航导弹#13	射击1次，杀伤	射击1次，未杀伤	射击1次，未杀伤	射击2次，未杀伤	无	无	无

图8　对抗推演输出结果示意图

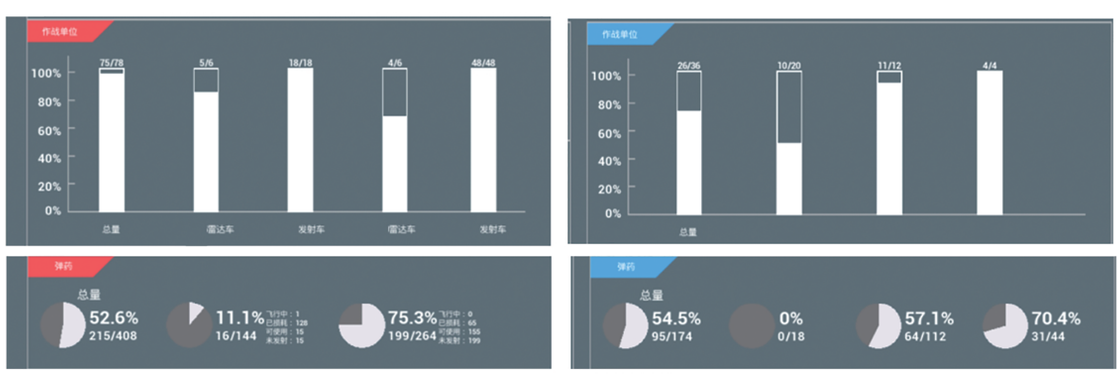
Fig. 8　Schematic diagram of adversarial inference output resultst

DOI: 10.16182/j.issn1004731x.joss.22-0432



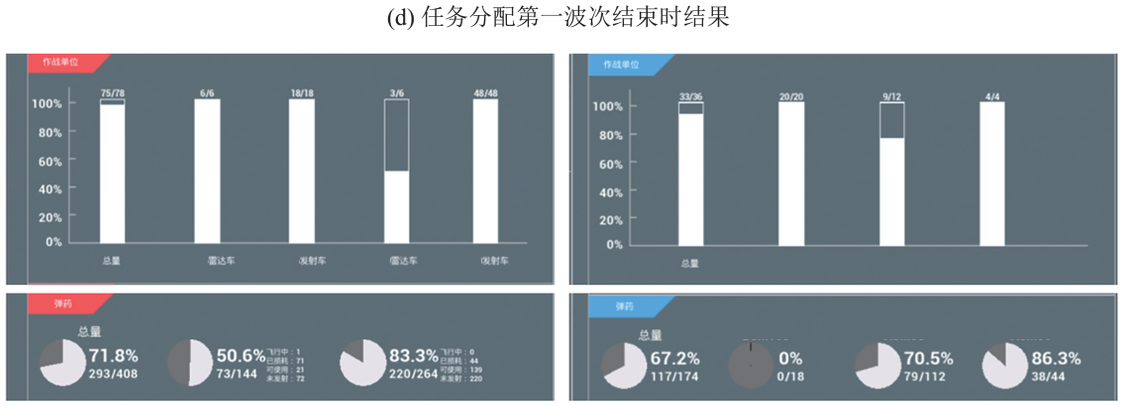
(a) 传统方法对抗结束时结果

(b) 目标分配第一波次结束时结果



(c) 目标分配对抗结束时结果

(d) 任务分配第一波次结束时结果

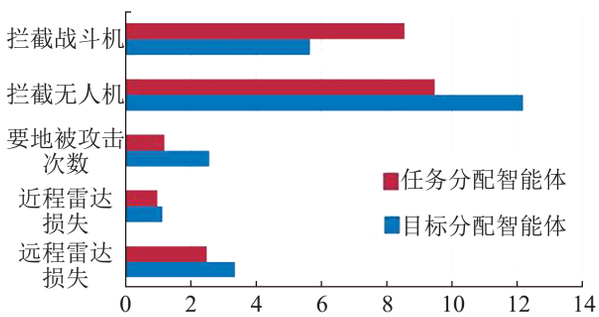


(e) 任务分配对抗结束时结果

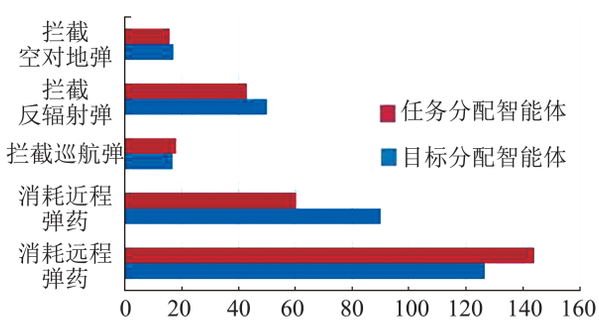
图9　对抗过程战况统计
Fig. 9　Battle statistics during rivalry

5.4.3 数字战场统计结果对比

为进一步对比目标分配模式与任务分配模式的区别，在训练结束后，将 2 个智能体模型放入数字战场进行离线推演 100 局并统计对抗结果，如图 10 所示，智能体部分行为对比如图 11 所示。



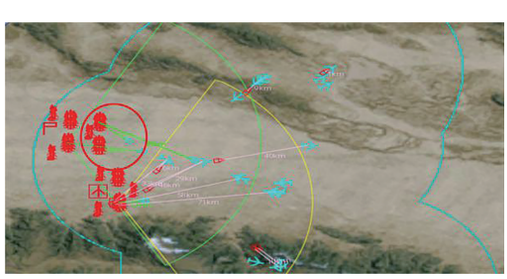
(a) 对抗统计结果第1部分



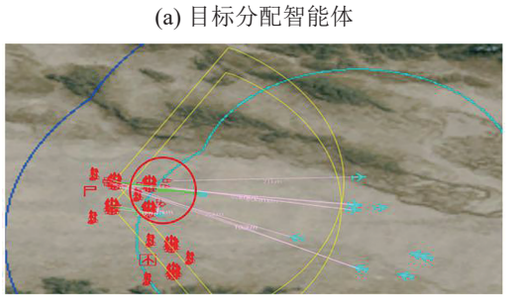
(b) 对抗统计结果第2部分

图 10 对抗结果对比

Fig. 10 Comparison of results of confrontation



(a) 目标分配智能体



(b) 任务分配智能体

图 11 智能体的行为对比

Fig. 11 Behavioral comparison of agents

从战损对比结果可以看出，目标分配智能体在对抗中损失较多，且近程弹药消耗较多，处于被动防御状态；任务分配智能体面对相同场景时资源分配更合理，己方损失更少。在拦截高威胁目标的同时尚有能力将资源用于拦截更多的高价值目标。从图 11 的行为对比可以看出，目标分配模式下当火力单元的传感器被攻击后，该火力单元不再参与拦截目标；任务分配模式下面对类似情况，该火力单元的拦截器依然可以对目标进行拦截。

6 结论

针对在大规模场景中的动态任务分配策略最优算法求解速度不足的问题，将其与 DRL 结合，用于求解大规模防空任务分配问题。基于分配策略最优算法的思想，将任务分配问题建模为 MDP，设计了合理的状态空间、动作空间及奖励函数；设计 DRL 训练框架并构建数字战场交互环境，对该问题进行求解；在数字战场中对 DRL 方法的有效性和任务分配模式的优越性进行了验证。实验结果表明：在大规模场景中基于分配策略最优思想的 DRL 任务分配方法可以有效提升求解问题的速度，能够有效应对态势的改变迅速做出决策，资源运用也更加合理灵活。本研究为动态任务分配提供了新的思路，下一步的研究内容包括：①改进 DRL 训练框架，进一步提升任务分配效率；②改进深度神经网络结构，进一步提升智能体训练效率。

参考文献：

[1] Zhang Jiandong, Chen Yuyang, Yang Qiming, et al. Dynamic Task Allocation of Multiple UAVs Based on Improved A-QCDPSO[J]. Electronics, 2022, 11(7): 1028.

[2] Wang Yao, Shi Yongkang, Liu Yunhui. Research on Improved Genetic Simulated Annealing Algorithm for Multi-UAV Cooperative Task Allocation[J]. Journal of Physics: Conference Series, 2022, 2246(1): 012081.

[3] Ma Yingying, Wang Guoqiang, Hu Xiaoxuan, et al. Two-stage Hybrid Heuristic Search Algorithm for Novel Weapon Target Assignment Problems[J]. Computers & Industrial Engineering, 2021, 162: 107717.

DOI: 10.16182/j.issn1004731x.joss.22-0432

[4] Kong Lingren, Wang Jianzhong, Zhao Peng. Solving the Dynamic Weapon Target Assignment Problem by an Improved Multiobjective Particle Swarm Optimization Algorithm[J]. Applied Sciences, 2021, 11(19): 9254.

[5] 王幸运, 田野, 强晓明, 等. 基于协同效能的反导作战任务分配模型[J]. 空军工程大学学报(自然科学版), 2013, 14(4): 27-31.
Wang Xingyun, Tian Ye, Qiang Xiaoming, et al. Mission Assignment Model for Anti-missile Combat Based on Cooperative Efficiency[J]. Journal of Air Force Engineering University(Natural Science Edition), 2013, 14(4): 27-31.

[6] 姜欢, 陈万春. 防空作战动静态武器目标分配初步研究[J]. 飞行力学, 2007, 25(4): 90-93.
Jiang Huan, Chen Wanchun. Dynamic and Static Weapon-target Assignments of Air Defense[J]. Flight Dynamics, 2007, 25(4): 90-93.

[7] Hosein P, Walton J, Athansm. Dynamic Weapon Target Assignment Problems With Vulnerable C2 Nodes[J]. Proceedings of the Command & Control Symposium, 1988, 1: 1-10.

[8] Hosein P, Athans M. Preferential Defense Strategies, Part 1: The Static Case[R]. MITLaboratory for Information and Decision Systemswith Partial Support, Cambridge, MA, Tech. Rep, 1990.

[9] 韩松臣, 秦俊奇, 韩品尧, 等. 马尔可夫决策过程在目标分配中的应用[J]. 哈尔滨工业大学学报, 1996, 28(2): 32-36.
Han Songchen, Qin Junqi, Han Pinyao, et al. An Application of the Markov Decision Process to Target Assignment[J]. Journal of Harbin Institute of Technology, 1996, 28(2): 32-36.

[10] 杨进帅, 李进, 王毅. 武器–目标分配问题研究[J]. 火力与指挥控制, 2019, 44(5): 6-11.
Yang Jinshuai, Li Jin, Wang Yi. Study of Weapon Target Assignment Problem[J]. Fire Control & Command Control, 2019, 44(5): 6-11.

[11] Zhou Wenhong, Liu Zhihong, Li Jie, et al. Multi-target Tracking for Unmanned Aerial Vehicle Swarms Using Deep Reinforcement Learning[J]. Neurocomputing, 2021, 466: 285-297.

[12] He Lei, Aouf N, Song Bifeng. Explainable Deep Reinforcement Learning for UAV Autonomous Path Planning[J]. Aerospace Science and Technology, 2021, 118: 107052.

[13] 刘传波, 邱志明, 吴玲, 等. 动态武器目标分配问题的研究现状与展望[J]. 电光与控制, 2010, 17(11): 43-48.
Liu Chuanbo, Qiu Zhiming, Wu Ling, et al. Review on Current Status and Prospect of Researches on Dynamic Weapon Target Assignment[J]. Electronics Optics & Control, 2010, 17(11): 43-48.

[14] 邱鸿泽. 基于自适应大邻域搜索算法的武器-目标分配问题研究[D]. 长沙: 国防科技大学, 2018.
Qiu Hongze. Weapon Target Assignment Research Based on Adaptive Large Neighborhood Search[D]. Changsha: National University of Defense Technology, 2018.

[15] 韩松臣. 导弹武器系统效能分析的随机理论方法[M]. 北京: 国防工业出版社, 2001.
Han Songchen. Stochastic Theory and Method for Effectiveness Analysis of Missile Weapon Systems[M]. Beijing: National Defense Industry Press, 2001.

[16] 陈英武, 蔡怀平, 邢立宁. 动态武器目标分配问题中策略优化的改进算法[J]. 系统工程理论与实践, 2007, 27(7): 160-165.
Chen Yingwu, Cai Huaiping, Xing Lining. An Improved Algorithm of Policies Optimization of Dynamic Weapon Target Assignment Problem[J]. Systems Engineering-Theory & Practice, 2007, 27(7): 160-165.

[17] 何鹏, 周德云, 王谦. 多UCAV任务分配有限阶段MDP方法和算法[J]. 火力与指挥控制, 2012, 37(10): 99-101, 104.
He Peng, Zhou Deyun, Wang Qian. Finite Stage MDP for Task Allocation in UCAVs Cooperative Control[J]. Fire Control & Command Control, 2012, 37(10): 99-101, 104.

[18] Ma Qiaoyun, Liu Tongsheng. Modeling Task Allocation in MAS With MDP[C]//The 15th International Conference on Industrial Engineering and Engineering Management (IE&EM'2008). Zhengzhou: Chinese Mechanical Engineering Society, 2008: 581-585.

[19] Esa Hyytiä, Righter R, Aalto S. Task Assignment in a Heterogeneous Server Farm With Switching Delays and General Energy-aware Cost Structure[J]. Performance Evaluation, 2014, 75-76: 17-35.

[20] Girard J, Reza Emami M. Concurrent Markov Decision Processes for Robot Team Learning[J]. Engineering Applications of Artificial Intelligence, 2015, 39: 223-234.

[21] 韦刚, 高嘉乐, 孙文. 多目标–多武器系统目标分配模型与算法研究[J]. 飞航导弹, 2016(5): 77-82.

[22] Yue Longfei, Yang Rennong, Zhang Ying, et al. Deep Reinforcement Learning for UAV Intelligent Mission Planning[J]. Complexity, 2022, 2022: 3551508.

[23] Shi Yuchun, Zheng Hao, Li Kang. Data-driven Joint Beam Selection and Power Allocation for Multiple Target Tracking[J]. Remote Sensing, 2022, 14(7): 1674.

[24] Park J H, Farkhodov K, Lee S H, et al. Deep Reinforcement Learning-based DQN Agent Algorithm for Visual Object Tracking in a Virtual Environmental Simulation[J]. Applied Sciences, 2022, 12(7): 3220.

[25] Sasaki H, Horiuchi T, Kato S. Experimental Study on

Behavior Acquisition of Mobile Robot by Deep Q-network [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2017, 21(5): 840-848.

[26] Yang Yang, Li Juntao, Peng Lingling. Multi-robot Path Planning Based on a Deep Reinforcement Learning DQN Algorithm[J]. CAAI Transactions on Intelligence Technology, 2020, 5(3): 177-183.

[27] Seliman S M S, Sadek A W, He Qing. Automated Vehicle Control at Freeway Lane-drops: A Deep Reinforcement Learning Approach[J]. Journal of Big Data Analytics in Transportation, 2020, 2(2): 147-166.

[28] Huang Hongji, Yang Yuchun, Wang Hong, et al. Deep Reinforcement Learning for UAV Navigation Through Massive MIMO Technique[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1): 1117-1121.

[29] 杨晨, 张少卿, 孟光磊. 多无人机协同任务规划研究[J]. 指挥与控制学报, 2018, 4(3): 234-248.
Yang Chen, Zhang Shaoqing, Meng Guanglei. Multi-UAV Cooperative Mission Planning[J]. Journal of Command and Control, 2018, 4(3): 234-248.

[30] 付强, 王刚, 鲁伟超, 等. 防空反导智能指控探索与实践[C]//第八届中国指挥控制大会论文集. 北京: 兵器工业出版社, 2020: 44-49.
Fu Qiang, Wang Gang, Lu Weichao, et al. Exploration and Practice of Intelligent Command and Control of Air Defense and Antimissile[C]//Proceedings of the 8th Chinese Conference on Command and Control. Beijing: Weapon Industry Press, 2020: 44-49.

[31] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2022-03-23]. http://arxiv.org/abs/arXiv:1707.06347.

DOI: 10.16182/j.issn1004731x.joss.22-0432