控制与决策
Control and Decision

基于主动风险防御机制的多机器人强化学习协同对抗策略

孙辉辉, 胡春鹤, 张军国

引用本文:
孙辉辉,胡春鹤,张军国. 基于主动风险防御机制的多机器人强化学习协同对抗策略[J]. 控制与决策, 2023, 38(5): 1420−1429.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1375

基于主动风险防御机制的多机器人强化学习协同对抗策略

孙辉辉[1,2], 胡春鹤[1,3], 张军国[1,3†]

(1. 北京林业大学 工学院，北京 100083；2. 华北科技学院 机电工程学院，河北 廊坊 065201；
3. 国家林业和草原局林业装备与自动化重点实验室，北京 100083)

摘 要: 深度强化学习因其在多机器人系统中的高效表现,已经成为多机器人领域的研究热点. 然而,当遭遇连续时变、风险未知的非结构场景时,传统方法暴露出风险防御能力差、系统安全性能脆弱的问题,未知风险将以对抗攻击的形式给多机器人的状态空间带来非线性入侵. 针对这一问题,提出一种基于主动风险防御机制的多机器人强化学习方法(APMARL). 首先,基于局部可观察马尔可夫博弈模型,建立多机记忆池共享的风险判别机制,通过构建风险状态指数提前预测当前行为的安全性,并根据风险预测结果自适应执行与之匹配的风险处理模式;特别地,针对有风险侵入的非安全状态,提出基于增强型注意力机制的 Actor-Critic 主动防御网络架构,实现对重点信息的分级增强和危险信息的有效防御. 最后,通过广泛的多机协作对抗任务实验表明,具有主动风险防御机制的强化学习策略可以有效降低敌对信息的入侵风险,提高多机器人协同对抗任务的执行效率,增强策略的稳定性和安全性.

关键词: 深度强化学习；多机器人；风险防御；协同对抗；事件驱动

中图分类号: TP24 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1375

引用格式: 孙辉辉, 胡春鹤, 张军国. 基于主动风险防御机制的多机器人强化学习协同对抗策略 [J]. 控制与决策, 2023, 38(5): 1420-1429.

Cooperative countermeasure strategy based on active risk defense multi-agent reinforcement learning

SUN Hui-hui[1,2], *HU Chun-he*[1,3], *ZHANG Jun-guo*[1,3†]

(1. School of Technology，Beijing Forestry University，Beijing 100083，China；2. School of Mechanical and Electrical Engineering，North China Institute of Science and Technology，Langfang 065201，China；3. Key Lab of State Forestry and Grassland Administration for Forestry Equipment and Automation，Beijing 100083，China)

Abstract: Deep reinforcement learning (DRL) has become a hotspot in the field of multi-robot systems due to its efficient performance. However, when encountering unstructured environment with time-varying and unknown risks, the traditional DRL methods exposes the disadvantage of poor risk defense ability and fragile system security. The unknown risk will bring nonlinear intrusion to the state space of multi-robot systems in the form of anti attack, which will pose a serious threat to the estimation of robot motion strategy. To solve this problem, this paper proposes a multi-agent reinforcement learning method based on active risk defense mechanism (ARD-MARL). Firstly, based on the locally observable Markov game model, a risk discrimination mechanism with global communication information is established to predict the current behavior state. Secondly, in the strategy deployment stage, we build an event-triggered multi risk processing scheme to implement the matching security strategy for different levels of risk prediction. Then, aiming at the dangerous state with risk intrusion, an active defense Actor-Critic network architecture based on the enhanced attention mechanism is designed. Through magnifying the important information and restraining the threat information, a safer and more efficient motion strategy is generated. Finally, extensive experiments are carried out in multi-agent cooperative and confrontation tasks. The results show that the multi-robot reinforcement learning method with active security defense mechanism can effectively enhance the stability and anti risk ability, and improve the security of information transmissions.

Keywords: deep reinforcement learning；multiple robots；risk defense；coordinated confrontation；event-triggered

收稿日期: 2022-07-31；**录用日期:** 2022-12-01.
基金项目: 国家自然科学基金项目 (61703047)；河北省高等学校科学技术研究项目 (QN2021312).
责任编委: 刘涛.
†通讯作者. E-mail: zhangjunguo@bjfu.edu.cn.

0 引　言

随着大数据、计算机、人工智能等技术的飞速发展,多机器人系统已经在军事、民生、安全等领域取得重大突破[1-2],尤其是在多机器人协同侦察、对抗、围捕等方向,已经吸引了众多科研人员的广泛关注[3-4]. 但是,随着任务环境复杂度的不断提升和应用的不断深入,多机器人自主控制策略的设计也变得日益困难. 传统程式化的建模、规则化的设计已经无法满足多机器人复杂多变的协同对抗作业需求[5-6],以环境感知、数据处理、多机调度、自主决策为一体的端到端的多智能体强化学习系统逐渐成为当前多机器人快速发展的关键技术[7-8].

多智能体强化学习是研究多机器人协作与对抗的重要技术之一,在解决一系列序贯决策问题上展现出优异的性能[9-11]. 该方法仅通过与环境互动和反复的持续学习[12],就可以利用环境给予的反馈奖励自动学习到最佳行动策略[13]. 相比于传统的多机器人控制策略,多智能体强化学习系统避免了为被控对象建立的精确数学模型的过程,提高了对原始传感数据的信息处理能力,降低了传统控制器设计中累计误差的影响[14-16]. 特别是在处理多机器人协同对抗问题方面,多智能体强化学习能够在联合回报函数的指导下,利用数据交流与信息共享构建自主学习的协同控制策略,实现从原始多信息输入到联合动作输出的端到端的映射[17]. 近年来,在动态复杂的未知环境中,多智能体强化学习已经为多机器人的避障、导航、编队和分配调度等任务提供了基于数据驱动的通用性运动控制方案[18-20].

虽然多智能体强化学习方法给多机器人自主控制问题带来了新的突破,但是,当面对非安全状态下的未知环境时,当前强化学习系统仍存在抵御风险能力差的问题[21]. 传统的深度强化学习系统通过当前获取的状态进行训练和实时决策[22-23],可在复杂的未知环境中,一旦机器人的信息采集与传输系统受到对手的主动攻击或者被动干扰,即便是轻微的原始扰动,也有可能使机器人陷入混乱的状态[24]. 风险状态的机器人误以为自身处于某种特定的环境,从而容易做出失败的决策使自己陷入危险的情景. 特别是在一些高安全领域需求的真实世界中,比如协同导航、自动驾驶等任务,这种对于外部风险无法抵御的脆弱性在真实世界中将给机器人系统带来重大的危害,严重制约着强化学习方法在多机器人系统上的实际应用[25]. 为了减缓这些不安全动作带来的危害,一类安全强化学习方法被提出[26]. 它们通过设置代价函数或者惩罚负奖励限制机器人的极限动作空间. 当遇到危险动作或者极限状态,机器人将终止当前状态来避免进一步的风险. 这种被动式的风险防御方式通过限制机器人动作空间来降低机器人的风险概率[27-28],虽然从一定程度上解决了机器人非安全决策的影响,但是却使控制策略产生了动作过于保守问题. 面对这一痛点,一类提前风险预测与防御的强化学习方法被提出. 这类解决方法对机器人受到的干扰和攻击进行提前检测,并将状态空间划分为安全区域和非安全区域两部分[29]. 当遭遇到潜在危险的状态时,他们将把决策系统切换到一种防御类型的安全状态来抵御外界的干扰,使自己尽快从危险中恢复过来. 相较于被动式防御系统,这种主动式防御具有更加灵活的防御手段. 防御模式可根据自身所处状态信息进行自由切换,有安全稳定的学习效率. 但是,当前的此类方法仅对危险状态进行判别,缺乏对非安全区域的危险信息进行进一步地化解与处理[30-32].

本文在主动防御方法的基础上,针对多智能体强化学习在复杂环境中危险防御能力脆弱的问题,构建一种基于主动风险防御机制的多机器人强化学习协同对抗策略[33-35],实现对非己方干扰信息侵入的有效抵御,为多机器人协同作业提供一种更加安全稳定的自主运动控制方法[36-37]. 在多种多机器人合作与对抗环境中,可表现出良好的风险抵御性能.

1 问题描述

多机器人强化学习协同对抗是通过多个机器人与训练环境的试错交互, 自主地学习如何在对抗任务环境中取得最优决策. 为了表述多机器人在这类序贯决策问题中的动作空间、环境观测、系统奖赏和策略函数等对象转移关系,用马尔可夫博弈过程对机器人的决策过程进行描述,具体表示为1个6元组 $\Gamma \triangleq (S, A, R, O, P, \gamma)$. 其中:$S$为所有智能体的状态集合,$A$为智能体的动作空间集合,$R$为所有智能体的奖励函数,$O$为当前智能体的状态集合,$P$为状态转移概率分布,$\gamma$为未来奖励折扣因子. 所有机器人的联合策略可以表示为$\pi \triangleq [\pi_1, \pi_2, \ldots, \pi_n]$. 第$i$个智能体的动作策略与当前观察和当前动作的关系可以表示为

$$\pi_i = o_i \times a_i. \tag{1}$$

智能体i的目标是最大化自己的奖励,在联合策略的指导下,智能体i获得下一步动作后,根据状态转移概率分布P得到下一步状态,同时获得奖励r_i. 整个回合的最大奖励可以表示为所有奖赏累计折扣的期望值

$$v_\pi^i(s)=\sum_{t=0}^{T}\gamma^t E_{\pi,\rho}[r_i^t \mid s_0=s,\pi]. \tag{2}$$

其中: T 为回合结束时刻, t 为当前时刻, s_0 为所有智能体初始状态, r_i^t 为当前时刻的即时奖励, π 为机器人动作联合策略, ρ 为动作空间.

为了进一步考虑智能体动作空间对目标奖励的影响, 智能体 i 在联合策略 π 下的目标奖励可以用动作价值 Q 函数表示, 即

$$Q_\pi^i(s_t,a_t)=r_i^t+\gamma E_{s_{t+1}\sim p}E_{a_t\sim\pi}[Q_\pi^i(s_{t+1},a_t)]. \tag{3}$$

其中: s_{t+1} 为下一时刻的系统状态, a_t 为当前动作.

2 主动风险预测与防御

在多机器人系统任务场景中, 常常充斥着各类干扰和不确定信息, 加之敌方智能体的主动攻击, 给基于数据驱动的多智能体系统带来巨大的威胁. 机器人在状态迷雾中无法准确分辨自己所处的空间, 从而容易做出错误的决策导致任务失败. 特别是在真实环境中, 未知区域的风险威胁将会进一步上升, 严重影响多机器人系统的安全性. 为了解决这一问题, 本部分为多智能体强化学习网络框架增加设计主动风险防御机制来对抗外界风险的侵入, 保证多智能体自主决策的安全性和稳定性.

2.1 风险状态预测

为了解决多智能体强化学习在复杂环境中存在模型脆弱、安全性低的问题, 首先需要对机器人风险状态进行预测. 利用欧几里得距离对机器人当前状态与基线状态进行相似性度量, 并由此判断智能体是否处于非安全状态是强化学习领域风险预测的有效方法之一[38]. 文献[39]证明了该方法在强化学习的状态空间中可以有效地指导智能体避开危险区域, 并在之后广泛应用于杆平衡[40]、直升机悬停控制[41]和竞争模拟器[42]等任务.

在此研究基础上, 构建基于安全状态记忆池的风险判别方法, 具体结构如图 1 所示. 在策略部署阶段, 建立状态记忆池 F 存储安全状态下的基线策略样本, 为风险事件预测提供数据基准. 安全状态记忆池建立完成后, 通过新样本与已知安全样本的相似性判别确定当前状态是否处于危险区域.

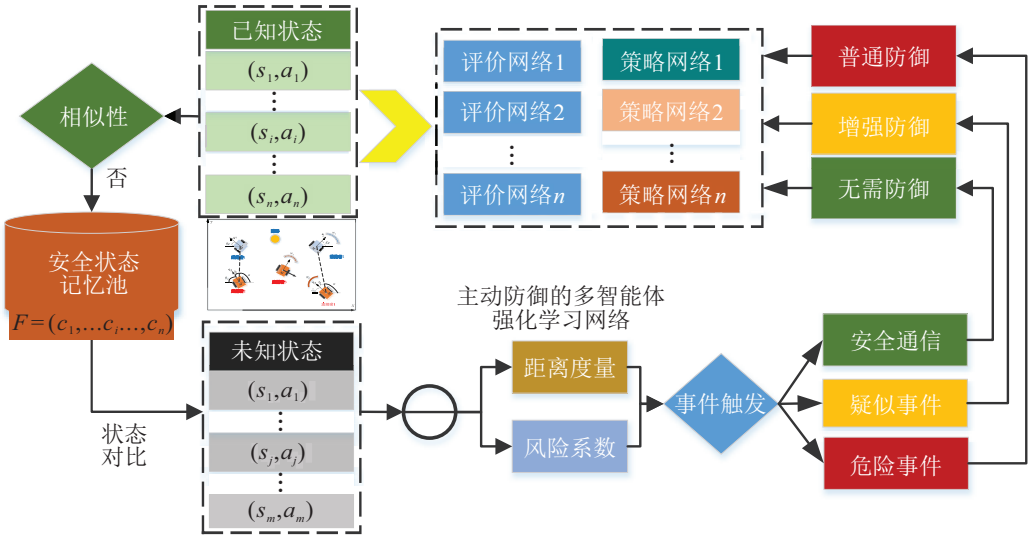


图 1 基于状态记忆池的风险预测

定义 1 状态记忆池 F 包含多组离散的状态样本 $F=(c_1,\dots c_i\dots,c_n)$. 每个离散状态样本 c_i 由机器人之前经历过的状态-动作对 (s_i,a_i) 构成.

样本池建立: 在机器人连续性的动作空间中, 经验样本的数量是接近于无限. 为了减轻样本池的容量压力, 以阈值过滤原则对机器人状态空间进行离散化. 机器人与环境完成一组交互后获取新的状态动作对 (s_i,a_i), 然后利用欧氏距离对当前状态动作样本进行距离度量

$$d(c_i,c_o)=\min(\|s_i-s_o\|_2+\|a_i-a_o\|_2). \tag{4}$$

如果一个新的状态动作对与已存样本的最近距离大于阈值距离 $d(c_i,c_0)>u$, 则当前新的状态动作对将被存入状态记忆样本池 $F'=F+1$.

阈值距离 μ 与记忆池中基线样本相关, 用状态记忆池中的平均距离确定阈值 μ. 具体可以表示为: 基线策略 π 在与环境互动中产生了一系列的状态动作对 $(c_1,c_2\dots,c_p)$, 并存于记忆池中. 根据记忆池样本可以得出阈值距离参数

$$u=\frac{d(c_1,c_2)+d(c_2,c_3)+\dots+d(c_{p-1},c_p)}{p-1}. \tag{5}$$

其中: c_i 为记忆池状态动作对, p 为状态样本数量.

状态记忆池随着策略的部署不断更新, 当记忆池

中的样本数据不再增加时,状态记忆池构建完成.

2.2　事件驱动触发机制

在部分通信的多机器人协同对抗环境中,本部分使用事件触发机制决定是否对机器人状态进行风险防御. 因此,马尔可夫博弈模型可重新用一个具有事件驱动机制的7元组来表示: $H \triangleq (S, A, R, O, P, \gamma, e)$,其中 e 为事件驱动机制的触发条件. 在对抗博弈过程中,如果事件 e 被触发,则表示机器人当前状态处于风险状态.

在策略执行阶段,以状态记忆池为基准,为机器人每个最新状态构建风险系数方程

$$b(c) = \frac{d(c_i, c_0)}{u}. \tag{6}$$

其中: $b(c)$ 为风险系数方程, d 为当前状态的最小距离度量.

在多机器人系统中,所有机器人的全局风险系数可以由向量表示,即

$$B(c) = (b_1(c), b_2(c), \ldots, b_i(c), b_n(c)). \tag{7}$$

然后,将全局智能体的风险系数的期望设定为事件驱动的触发条件,即

$$e(c) = E(b_i(c)) = \frac{1}{n}\sum_{i=1}^{n} b_i(c), \tag{8}$$

其中 n 为智能体的数量.

基于事件驱动触发条件,当智能体接收到一个新的状态 c_i 后,智能体通过计算新的状态和记忆池中离它最近的样本的欧氏距离,获得该状态所处环境的风险系数 $b(c_i)$.

风险系数是判断机器人所处状态的标准,根据危险程度的不同,状态空间被分为3种情形: 如果风险系数 $b(c_i) \in (0, 1]$,则表示机器人当前状态为安全状态,机器人将继续执行当前策略; 如果 $b(c_i) \in (1, e]$,则表示机器人处于安全缓冲区,机器人将进行保守策略决策,从记忆池中选取与当前状态相似性最强的状态动作样本作为下一步决策的参考; $b(c_i) \in (e, \infty)$ 的情况代表机器人处于危险状态,此时,事件驱动条件被触发,机器人将进入安全防御状态,应用主动防御策略对状态信息进行筛选和过滤处理,然后基于优化后的信息进行重新决策.

2.3　增强型自注意力机制的安全防御

智能体基于自身传感模块获取局部观测信息 o_i,然后应用策略网络生成动作 a_i,并且与环境作出互动. 为了方便信息的传输与共享,将每个智能体的观测感知进行编码

$$E = \Phi_\theta(O, A). \tag{9}$$

其中: E 为所有智能体观测和动作的表征, Φ_θ 为一层全连接感知器, O 为所有智能体的联合观察.

当机器人处于危险状态时,强化学习网络将主动切换为安全防御模式. 机器人将自己的观测和动作信息传入多层神经网络来获得表征 E,然后利用多头注意力机制对干扰和侵入信息进行过滤,提取出对当前表征更有价值的状态信息及权重. 因此,带有注意力机制的动作价值函数可以重新表示为

$$Q_i^\theta(o, a) = g_i(\varphi_{i,\theta}(o_i, a_i), x_i). \tag{10}$$

其中: g_i 为多层全连接感知器; $\varphi_{i,\theta} \in \Phi_\theta$ 为一层全连接编码,计算局部观测的动作价值; x_i 为除了当前智能体之外的所有其他智能体观测编码的权重信息,具体可以表示为

$$x_i = \sum_{j\neq i} \alpha_j v_j = \sum_{j\neq i} \alpha_j h(V\varphi_j(o_j, a_j)). \tag{11}$$

这里: V 为特征编码的共享矩阵, h 为非线性变换的激活函数, α_j 用来传递编码后的其他智能体注意力机制的相关性权重.

具体而言,在对智能体 i 的观测信息执行编码后,可以得到观测表征 e_i. 然后根据注意力机制模型计算每个 $j \neq i$ 智能体和智能体 i 的相关度,即注意力机制的权重

$$\alpha_j = e_j^{\mathrm{T}} W_k^{\mathrm{T}} W_q e_i. \tag{12}$$

其中: 参数矩阵 W_q 可以把 e_i 转化为查询关键字, W_k 参数矩阵把 e_j 映射为键值.

主动安全防御机制采用多头注意力模型. 每个注意力使用一组独立的参数 (W_k, W_q, V) 评估其他智能体对当前智能体的联合贡献,并以所有其余智能体的相对于当前智能体 i 的平均相关度构建相关度基准,即

$$B_i = \frac{1}{n-1}\sum_{j\neq i} \alpha_j. \tag{13}$$

接着,根据式(12)和(13),其他智能体与当前智能体之间的相关度指标可以定义为

$$R_j = \alpha_j - B_i. \tag{14}$$

在风险状态下,机器人的信息权重加强和削弱将分阶段处理. 注意力机制权重将会进一步增强. 增强系数 k 的相关取值分布如表1所示.

表1　注意力机制权重增强系数分布

R_j/B_j	$(-\infty, -2]$	$(-2, -1]$	$(-1, 0]$	$(0, 1]$	$(1, 2]$	$(2, \infty)$
k	$-\infty$	-4	-2	1	2	4

最后,通过一个softmax对注意力机制权重进行

归一化处理

$$\bar{\alpha}_j = \mathrm{softmax}(\tilde{\alpha}_j) = \frac{e^{\tilde{\alpha}_j}}{\sum_{n-1} e^{\tilde{\alpha}_j}}. \tag{15}$$

式(15)中:相关度较小的疑似危险信息将会被减小或者直接剔除;相关度较高关键信息将随相关程度分阶段加强.

2.4 网络框架更新

基于主动防御机制的多机器人强化学习运动规划策略由多个并行的策略网络 Actor 与评价网络 Critic 构成,每个智能体配备独立的 Actor-Critic 网络,具体结构如图2所示.

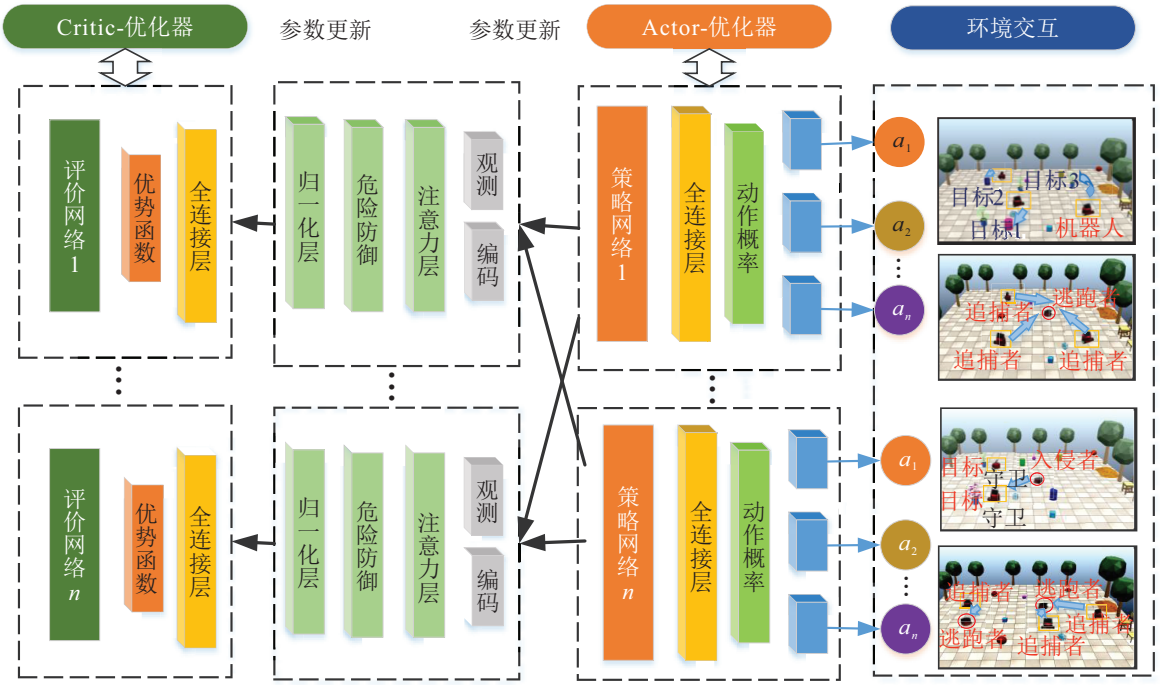


图 2 Actor和Critic网络结构

Actor网络和Critic网络均由多层感知器MLP构成. 评价网络Critic可以观测到所有其他智能体的动作状态并对其进行编码,通过注意力机制进行强化和筛选后,获取智能体 i 的优势函数 $A_i(o,a)$. 优势函数是根据智能体 i 的当前策略与环境互动获得的全局回报与平均回报的差值,具体可以表示为

$$A_i(o,a) = Q_i^{\pi}(o,a) - b(o,a_{\setminus i}), \tag{16}$$

其中 $b(o,a_{\setminus i})$ 为多智能体反事实的基线. 基线 $b(o,a_{\setminus i})$ 的期望可以通过在前向传播中输入智能体 i 的所有可能动作值分布来估计,其具体的计算过程为

$$b(o,a_{\setminus i}) = E_{a_i\sim\pi(o_i)}[Q_i^{\theta}(o,a)]. \tag{17}$$

多智能体系统的目标是通过最小化参数共享的联合回归损失来最大化动作价值 $Q_i^{\theta}(o,a)$. 为此,定义损失函数如下:

$$L_Q(\theta) = \sum_{i=1}^{n} E_{(o,a,r,o')D}[(y_i - Q_i^{\theta}(o,a))^2], \tag{18}$$

其中 y_i 为目标动作价值.

目标网络的网络参数不在训练时直接更新,而是每次从当前网络的网络参数中复制一部分,其参数优化仍然采用软更新的方式.

3 实验仿真与验证

为了测试本主动风险防御策略在多机器人任务环境中的性能表现,本部分在Gym环境、物理仿真环境和真实环境中分别开展了对比实验. 场景中涵盖了多机器人合作探索、避碰、追捕和防守等任务. 实验分别设置训练和测试两类环境. 首先使用多智能MPE场景作为策略的训练环境,环境从简单到复杂,分别为:合作导航、3对1追捕对抗、3对2追捕对抗和2对1侵占对抗4类. 在所有的环境中,训练场景不定时地发送干扰和威胁以误导机器人做出错误的决策,从而测试机器人的抗干扰和风险抵御能力.

3.1 模型训练

首先,分别基于本文提出的方法ARD-MARL、多智能体深度确定性策略梯度(MADPPG)[43]以及深度确定性策略梯度(DDPG)[44]对每个场景进行10 000个回合的策略训练,单个回合最大决策步数为100. 训练结束后,得到3种算法在各个场景下的平均回合奖励变化情况,如图3所示.

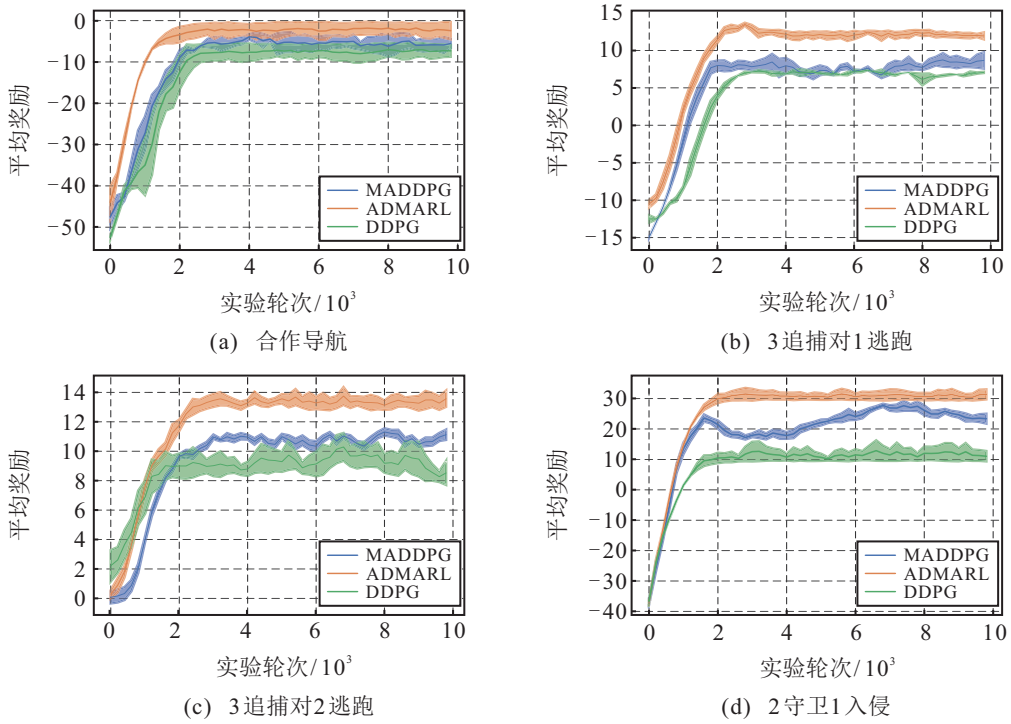


图 3　训练环境中多机器人奖励值对比

图3表示在4种场景中平均回合奖励变化趋势. 在经历了1 000轮训练之后,3种算法的奖励值都可以到达稳定的状态. 在4种训练环境中,本文所提ARD-MARL方法获得的平均奖励均高于MADDPG和DDPG方法的奖励值;特别是在3对2的追捕环境中,ARD-MARL方法的奖励值相比于其他两种方法分别提升了32.7%和51.6%. 综合奖励值的结果可以看出,本文所提具有主动防御机制的多智能体强化学习方法过滤掉了多数外界干扰和攻击,为机器人的有效决策提供更加充分的数据样本,保证决策的安全性和高回报性.

当面临具有风险入侵的环境时,为了验证不同风险状态的判定结果,统计不同危险系数下的系统奖励值情况,从奖励值的变化进一步分析智能体当前所处的状态情况,具体如图4所示.

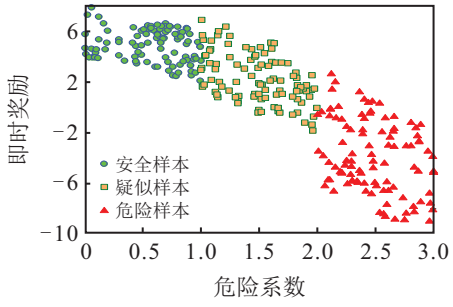


图 4　不同危险系数样本的即时奖励对比

从图4可以看出,随着危险系数的增加,系统获得的即时奖励值在逐渐地减少,这意味着处于疑似区域和危险区域的状态样本正在为机器人提供回报较低的非安全动作,低回报的动作策略将会使机器人偏离目标,甚至进入更加危险的状态.

3.2　模型测试

在MPE多智能体环境中训练完成之后,收敛后的模型将被迁移至更加接近真实环境的物理仿真器中来验证多智能体强化学习策略的泛化性能和稳定性. 基于机器人仿真器CoppeliaSim,同样建立4种多机器人仿真环境,分别为协同导航、合作追捕和侵占对抗环境,如图5所示.

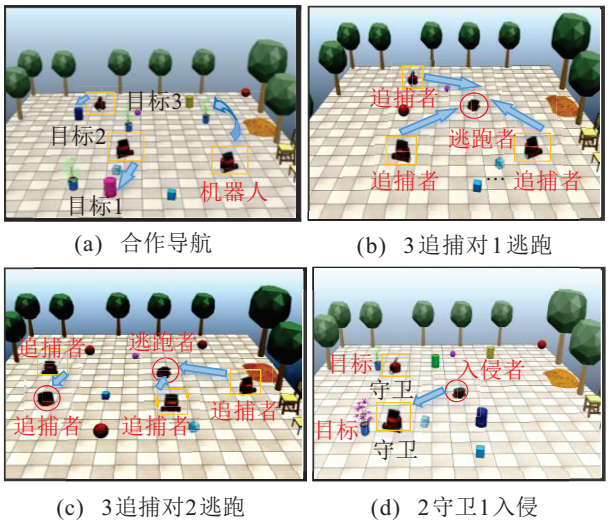


图 5　物理仿真测试环境

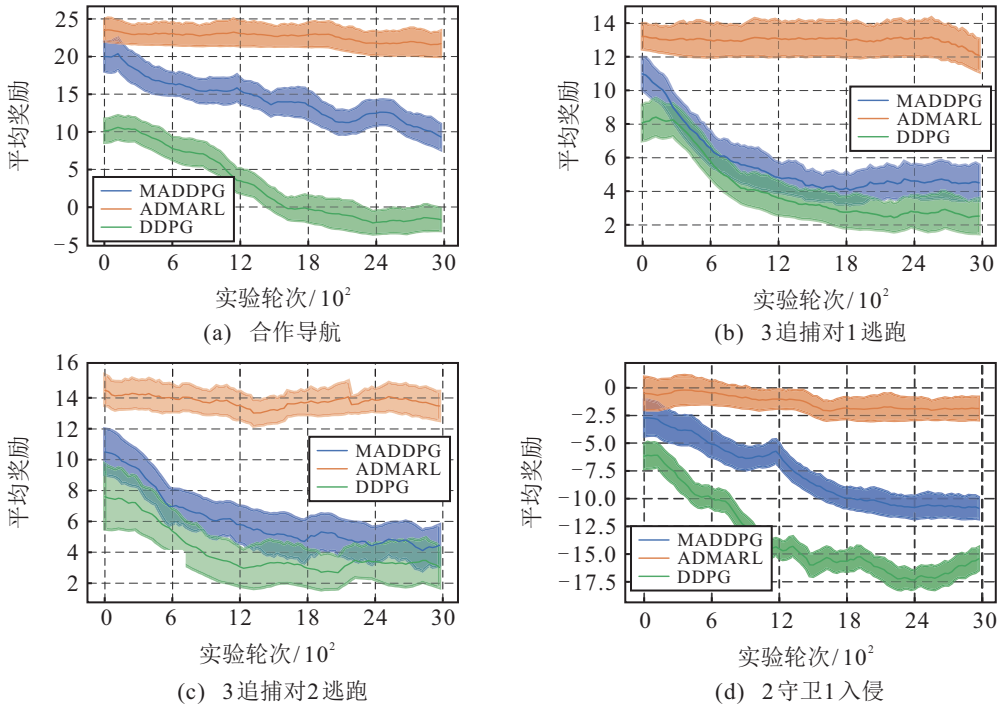
在4种场景中分别进行1 000轮次无风险侵入的测试实验,统计各种场景的任务成功率的变化,如表

2所示. 在无威胁的4种任务环境中, ARD-MARL训练完成后的平均任务成功率在95%以上, MADDPG算法的任务成功率在90%左右, 而DDPG平均任务成功率仅处于50%的水平.

进一步地, 在遭受危险入侵后的测试环境中, 对比不同任务场景中的奖励函数变化情况, 具体结果如图6所示.

表2 不同情况下的任务成功率对比 %

algorithms	任务成功率无威胁				任务成功率有威胁			
	场景1	场景2	场景3	场景4	场景1	场景2	场景3	场景4
ARD-MARL	98.4	96.4	95.4	96.5	92.4	93.8	90.6	90.5
DDPG	67	50.5	45.3	46.4	46.5	50.3	49.8	53.4
MADDPG	95.2	88.6	87.3	90.2	65.3	62.4	50.3	58.9



(a) 合作导航 (b) 3追捕对1逃跑

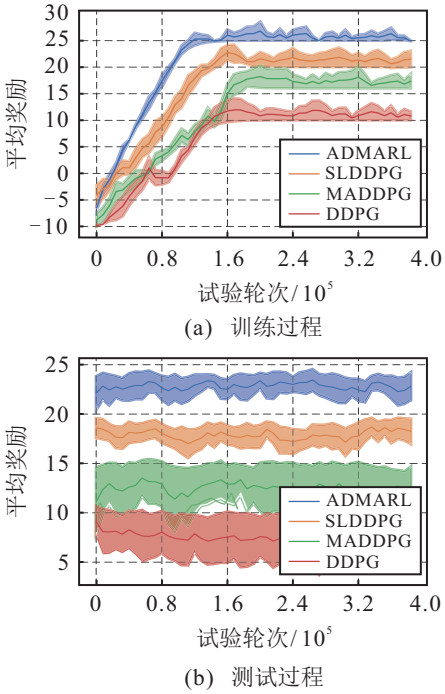
(c) 3追捕对2逃跑 (d) 2守卫1入侵

图6 危险侵入场景的奖励值对比

从图6中可以看出, 在增加了干扰和威胁信息后, 机器人的决策判断出现了偏差. 通过对比可以明显看出, 随着威胁信息的不断侵入, 不具备危险防御能力的MADDPG和DDPG算法的系统奖励值出现了明显的下降, 并且越复杂的场景表现越为明显. 不同的是, 由于主动防御机制的存在, 本文所提ARD-MARL方法在危险侵入后其奖励值没有受到太大的影响, 仅仅表现出了轻微的波动和5%以内的下降.

同时, 从表2中的有风险侵入的任务成功率对比可以发现, 在风险侵入后, 4种场景中的任务成功率都出现了明显的下降, 尤其是MADDPG算法下的任务成功率降低了30%左右, DDPG算法由于原始成功率较低的缘故, 在风险的侵入下降低了20%左右. 本文所提策略的任务成功率仅受到了轻微的影响, 主动风险防御机制能够成功地保护策略网络免受干扰和强攻击的威胁.

最后, 针对更加复杂的多机器人环境, 实验构建了带有多重阻隔和障碍的动态仿真环境. 在经历了各2 000轮次的训练和测试之后, 分别记录每种算法单回合奖励值的变化并绘制平均奖励曲线, 具体实验结果如图7所示.



(a) 训练过程

(b) 测试过程

图7 复杂场景动态场景中的奖励值

在动态场景实验中增添最新的李雅普诺夫安全深度确定性策略梯度(SLDDPG)强化学习方法作为对比方案[45]. 从图7(a)中可以看出,在训练2 000轮次之后,本文所提ARD-MARL方法的平均收敛值稳定在25左右,安全强化学习SLDDPG过分注重了安全性能而忽视了运动规划的最优解的探索,在追逐环境中容易造成追捕时间超时,平均奖励较低. 其他两种方法的奖励值在训练了800轮次之后分别稳定在17和11左右. 另外,通过图7(b)中的测试环境中奖励值变化可以看出,在新的环境中,虽然奖励值的波动幅度相比于训练场景有所增加,但整体获得的平均奖励值和原始环境中的收敛值处于同一水平,验证了本文所提方法在新的测试环境中良好的泛化性.

3.3　物理实验

为了测试所提方法在真实环境中的表现,本节将策略模型迁移至室内物理环境中进行实验. 由于物理设备条件的限制,仅对合作导航任务进行测试,具体场景布置如图8所示.

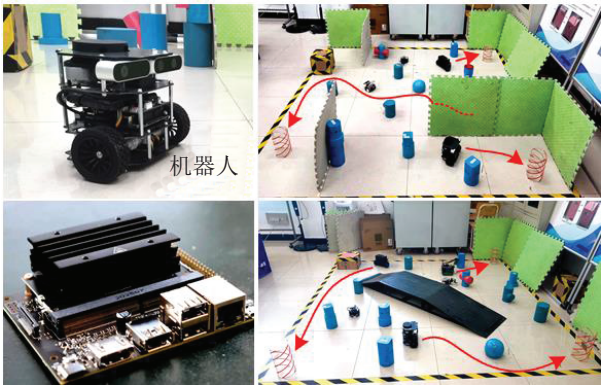
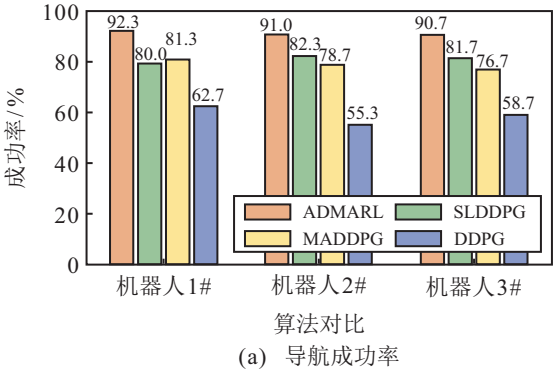


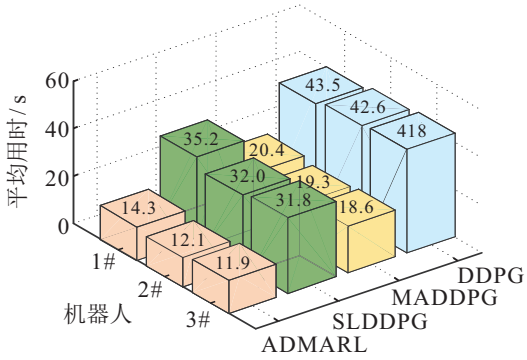
图 8　真实实验场景

实物场景中包括多种静态障碍物和动态障碍物. 多个机器人需要在场景中完成特定目标的自主导航任务. 在图8所示的实验环境中: 如果机器人未能在60 s内成功到达目标,则认为任务超时失败; 如果机器人被其他动态和静态障碍物阻拦,则被记录发生一次碰撞事件. 基于本文所提方法和其他3种对比算法,实验共进行了300轮次任务测试. 实验完成后,统计了每组机器人的导航成功率和平均导航时间,并且与其他方法进行了对比,具体结果如图9所示.

从图9(a)中可知,本文提出的方法ARD-MARL的导航成功率分别为92.3%、 91.0%和90.7%,比其在仿真环境中的测试结果略低. 但是相比于其他算法,在导航任务方面仍然具有最高的成功率. 从图9(b)的机器人单回合平均用时情况可以看出: 本文提出的方法ARD-MARL和多智能体强化学习方法MADDPG用时较少,单局平均用时均在20 s以下,而安全强化学习SLDDPG和深度确定性策略梯度DDPG的单回合用时达到了ARD-MARL的两倍以上. 分析其原因,安全强化学习方法由于采取了安全约束条件,机器人策略过于保守,机器人为了避开障碍物而选择了非最优路径,从而导致用时大量增加; DDPG方法容易让自己的运动策略进入局部最优,导致任务超时.



(a)　导航成功率



(b)　单回合平均用时

图 9　真实场景实验结果

综合真实场景中的实验结果,通过分析对比不同方法的任务成功率和平均用时的差异可以得出,本文提出的主动防御强化学习方法ARD-MARL在动态的真实环境中展现出较好的策略稳定性,机器人的导航成功率相对较高,具有良好的迁移性和泛化能力.

4　结　论

本文针对具有未知风险入侵的多机器人协同对抗环境,提出了一种具有主动风险防御机制的多机器人强化学习协同对抗策略(ARD-MARL). 相比于常规的强化学习策略,ARD-MARL主要从风险状态判别、风险事件处理和风险事件防御3个方面进行了研究: 1) 基于多机共享的安全记忆样本池构建了机器人风险判别机制,将机器人所处的状态空间划分为安全状态、疑似状态和风险状态3个区域; 2) 基于状态风险指数建立了事件驱动的多风险模式预处理机制,根据不同的风险状态自适应执行与之匹配的安全策略; 3) 针对特定的风险状态空间,通过对威胁信

息过滤与重要信息加强,大幅度提升策略的抗风险能力;4) 仿真与真实环境的实验结果表明,具有主动风险防御机制的多智能体强化学习协同对抗策略在未知风险的干扰和侵入情况下,表现出了良好的风险判别和抵御能力,在不同环境之间模型具有良好的可迁移性和泛化性.

虽然本文提出的方法降低了外界干扰和侵入风险带来的影响,但是却在一定程度上牺牲了对最优策略的探索性,尤其是在多机器人博弈环境中,策略的保守性将会降低机器人的对抗能力. 未来,如何在保证机器人安全性的同时,通过风险系数自主学习或最优策略激励等方式,进一步提高机器人的任务执行效率,是未来需要研究的课题之一. 另外,将本文算法成功部署到更加复杂的机器人对抗环境,努力克服真实世界场景带来的机械偏差、通讯迟滞和传感漂移等问题,是真实环境中的不确定性带来的另一挑战.

参考文献(References)

[1] Madridano Á, AI-kaH A, Martín D, et al. Trajectory planning for multi-robot systems: Methods and applications[J]. Expert Systems With Applications, 2021, 173: 114660.

[2] Blumenkamp J, Morad S, Gielis J, et al. A framework for real-world multi-robot systems running decentralized GNN-based policies[C]. 2022 International Conference on Robotics and Automation (ICRA). Philadelphia, 2022: 8772-8778.

[3] Samsani S S, Muhammad M S. Socially compliant robot navigation in crowded environment by human behavior resemblance using deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2021, 6(3): 5223-5230.

[4] 吴军, 徐昕, 王健, 等. 面向多机器人系统的增强学习研究进展综述[J]. 控制与决策, 2011, 26(11): 1601-1610.
(Wu J, Xu X, Wang J, et al. Recent advances of reinforcement learning in multi-robot systems: A survey[J]. Control and Decision, 2011, 26(11): 1601-1610.)

[5] 刘亚军, 訾斌, 王正雨, 等. 智能喷涂机器人关键技术研究现状及进展[J]. 机械工程学报, 2022, 58(7): 53-74.
(Liu Y J, Zi B, Wang Z Y, et al. Research progress and trend of key technology of intelligent spraying robot[J]. Journal of Mechanical Engineering, 2022, 58(7): 53-74.)

[6] Verma J K, Ranga V. Multi-robot coordination analysis, taxonomy, challenges and future scope[J]. Journal of Intelligent & Robotic Systems, 2021, 102(1): 1-36.

[7] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 3083-3102.
(Yan C, Xiang X J, Xu X, et al. A survey on scalability and transferability of multi-agent deep reinforcement learning[J]. Control and Decision, 2022, 37(12): 3083-3102.)

[8] 訾斌, 徐锋, 唐锴, 等. 基于机器视觉的喷涂机器人轨迹规划与涂装质量检测研究综述[J]. 控制与决策, 2023, 38(1): 1-21.
(Zi B, Xu F, Tang K, et al. Trajectory planning for spray-painting robot and quality detection of paint film based on machine vision: A review[J]. Control and Decision, 2023, 38(1): 1-21.)

[9] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey[J]. Artificial Intelligence Review, 2022, 55(2): 895-943.

[10] Boldrer M, Palopoli L, Fontanelli D. A unified Lloyd-based framework for multi-agent collective behaviours[J]. Robotics and Autonomous Systems, 2022, 156: 104207.

[11] Liu Y K, Liang H, Xiao Y, et al. Logistics-involved service composition in a dynamic cloud manufacturing environment: A DDPG-based approach[J]. Robotics and Computer-Integrated Manufacturing, 2022, 76: 102323.

[12] Xu J, Ma J, Gao X, et al. Adaptive progressive continual learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6715-6728.

[13] Zhang Y Z, Wu Z R, Ma Y H, et al. Research on autonomous formation of multi-UAV based on MADDPG algorithm[C]. 2022 IEEE 17th International Conference on Control & Automation. Naples, 2022: 249-254.

[14] Song X Y. MADDPG: An efficient multi-agent reinforcement learning algorithm[C]. International Conference on Advanced Algorithms and Neural Networks (AANN 2022). Zhuhai, 2022: 50-55.

[15] Qian S, Bao K L, Zi B, et al. Dynamic trajectory planning for a three degrees-of-freedom cable-driven parallel robot using quintic B-splines[J]. Journal of Mechanical Design, 2020, 142(7): 142.

[16] Chen Y, Han W, Zhu Q H, et al. Target-driven obstacle avoidance algorithm based on DDPG for connected autonomous vehicles[J]. EURASIP Journal on Advances in Signal Processing, 2022, 2022: 61.

[17] 鲁亚洲, 黄静, 阮晓钢, 等. 结合图神经网络和深度强化学习的机器人自主探索方法[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2021.1909.
(Lu Y Z, Huang J, Ruan X G , et al. A robot autonomous exploration method combining graph neural network and deep reinforcement learning[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2021.1909.)

[18] 董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展[J]. 控制与决策, 2022, 37(2): 278-292.
(Dong H, Yang J, Li S B, et al. Research progress of robot motion control based on deep reinforcement learning[J]. Control and Decision, 2022, 37(2): 278-292.)

[19] 邓小豪, 侯进, 谭光鸿, 等. 基于强化学习的多目标车辆跟随决策算法[J]. 控制与决策, 2021, 36(10): 2497-2503.
(Deng X H, Hou J, Tan G H, et al. Multi-objective vehicle following decision algorithm based on reinforcement

learning[J]. Control and Decision, 2021, 36(10): 2497-2503.)

[20] 郭方洪, 何通, 吴祥, 等. 基于分布式深度强化学习的微电网实时优化调度[J]. 控制理论与应用, 2022, 7(30): 1-8.
(Guo F H, He T, Wu X, et al. Real-time optimal scheduling for microgrid systems based on distributed deep reinforcement learning[J]. Control Theory & Applications, 2022, 7(30): 1-8.

[21] García J, Majadas R, Ferńandez F, et al. Learning adversarial attack policies through multi-objective reinforcement learning[J]. Engineering Applications of Artificial Intelligence, 2020, 96: 104021.

[22] 李保罗, 蔡明钰, 阚震. 线性时序逻辑引导的安全强化学习[J]. 控制与决策, 2022, 10(3): 1-9.
(Li B L, Cai M Y, Kan Z. Linear temporal logic guided safe reinforcement learning[J]. Control and Decision, 2022, 10(3): 1-9.)

[23] García J, Sagredo I. Instance-based defense against adversarial attacks in deep reinforcement learning[J]. Engineering Applications of Artificial Intelligence, 2022, 107: 104514.

[24] Xu C Y, Zhang B. Online adaptive optimal control algorithm of partial unknown system with adding experience replay and safety check[C]. Proceedings of 2022 Chinese Intelligent Systems Conference. Singapore, 2022: 681-696.

[25] Borkar V S. Q-learning for risk-sensitive control[J]. Mathematics of Operations Research, 2002, 27(2): 294-311.

[26] Li H P, He H B. Learning to operate distribution networks with safe deep reinforcement learning[J]. IEEE Transactions on Smart Grid, 2022, 13(3): 1860-1872.

[27] Liu Q, Liu Z, Xiong B, et al. Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function[J]. Advanced Engineering Informatics, 2021, 49: 101360.

[28] Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: A survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[29] Ishii H, Wang Y, Feng S, et al. An overview on multi-agent consensus under adversarial attacks[J]. Annual Reviews in Control, 2022, 53: 252-272.

[30] Mo K H, Tang W X, Li J, et al. Attacking deep reinforcement learning with decoupled adversarial policy[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 3566(99): 1.

[31] Zhou X, Zhou S, Mou X, et al. Multirobot collaborative pursuit target robot by improved MADDPG[J]. Computational Intelligence and Neuroscience, 2022, 2022: 4757394.

[32] Wan K F, Wu D W, Li B, et al. ME-MADDPG: An efficient learning-based motion planning method for multiple agents in complex environments[J]. International Journal of Intelligent Systems, 2022, 37(3): 2393-2427.

[33] Xu X Y, Hu H, Liu Y, et al. Moving target defense of routing randomization with deep reinforcement learning against eavesdropping attack[J]. Digital Communications and Networks, 2022, 8(3): 373-387.

[34] Li H R, Guo Y F, Huo S M, et al. Defensive deception framework against reconnaissance attacks in the cloud with deep reinforcement learning[J]. Science China Information Sciences, 2022, 65(7): 1-19.

[35] Xuan S Z, Ke L J. UAV swarm attack-defense confrontation based on multi-agent reinforcement learning[C]. Advances in Guidance, Navigation and Control. Singapore, 2022: 5599-5608.

[36] Huang L, Fu M, Qu H, et al. A deep reinforcement learning-based method applied for solving multi-agent defense and attack problems[J]. Expert Systems With Applications, 2021, 176: 114896.

[37] Xu J, Huang F, Wu D, et al. Deep reinforcement learning based multi-AUVs cooperative decision-making for attack-defense confrontation missions[J]. Ocean Engineering, 2021, 239: 109794.

[38] Garcia J, Fernandez F. Safe exploration of state and action spaces in reinforcement learning[J]. Journal of Artificial Intelligence Research, 2012, 45: 515-564.

[39] Santamaria J C, Sutton R S, Ram A. Experiments with reinforcement learning in problems with continuous state and action spaces[J]. Adaptive Behavior, 1997, 6(2): 163-217.

[40] Martin H J A, de Lope J. Exa: An effective algorithm for continuous actions reinforcement learning problems[C]. 2009 35th Annual Conference of IEEE Industrial Electronics. Portuga, 2009: 2063-2068.

[41] Jose A M H, Lope J. Learning autonomous helicopter flight with evolutionary reinforcement learning[C]. Las Palmas de Gran Canaria. Spain, 2011: 75-82.

[42] Borrajo F, Bueno Y, de Pablo I, et al. SIMBA: A simulator for business education and research[J]. Decision Support Systems. 2010, 48(3): 498-506.

[43] Lowe R, Wu Y I, Tamar A, et al. Multi-agentactor-critic for mixed cooperative-competitive environments[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-12.

[44] de Jesus J C, Bottega J A, de Souza Leite Cuadros M A, et al. Deep deterministic policy gradient for navigation of mobile robots[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(1): 349-361.

[45] Du B, Lin B, Zhang C, et al. Safe deep reinforcement learning-based adaptive control for USV interception mission[J]. Ocean Engineering, 2022, 246: 110477.

作者简介

孙辉辉 (1989–), 男, 博士生, 从事智能机器人及其控制方法的研究, E-mail: sunhuihui@bjfu.edu.cn;

胡春鹤 (1986–), 男, 副教授, 博士, 从事无人机自主控制、多无人机协同控制及其应用等研究, E-mail: huchunhe@bjfu.edu.cn;

张军国 (1978–), 男, 教授, 博士生导师, 从事智慧林业监测与信息处理、无人飞行器及林业特种机器人等研究, E-mail: zhangjunguo@bjfu.edu.cn.

(责任编辑: 闫 妍)