第 46 卷 第 2 期：459-471 高电压技术 Vol.46, No.2: 459-471
2020 年 2 月 29 日 High Voltage Engineering February 29, 2020
DOI: 10.13336/j.1003-6520.hve.20200131010

基于层次深度强化学习的带电作业机械臂控制技术

闫 冬[1]，陈 盛[1]，彭国政[1]，谈元鹏[1]，张玉天[1]，吴 凯[2]
（1. 中国电力科学研究院有限公司，北京 100192；
2. 国网安徽省电力有限公司电力科学研究院，合肥 230601）

摘 要：为了实现配电网带电作业机械臂的自主避障和自主导航，提出了一种基于子任务的层次深度强化学习算法的机械臂智能控制方法，设计了其对应的环境状态空间、动作策略和奖励函数并开展仿真实验进行效果验证。实验结果表明：全局随机障碍下单一模型成功率平均不足 35%，层次模型比单一模型在全空间的决策训练中更易收敛，在跨线作业和设备作业两种场景下避障导航成功率分别可提升至 90%和 71.01%；同时，提出的安全路径引导奖惩机制可有效提升考虑安全距离下的机械臂作业路径寻优效率。层次深度强化学习模型在应对不同目标及障碍时具有更强的泛化性能，可为实现全自主带电作业提供理论和技术参考。

关键词：带电作业；自主避障；自主导航；层次强化学习；深度强化学习；智能控制

Live Working Manipulator Control Technology Based on Hierarchical Deep Reinforcement Learning

YAN Dong[1], CHEN Sheng[1], PENG Guozheng[1], TAN Yuanpeng[1], ZHANG Yutian[1], WU Kai[2]
(1. China Electric Power Research Institute, Beijing 100192, China;
2. State Grid Anhui Electric Power Research Institute, Hefei 230601, China)

Abstract：To realize the autonomous obstacle avoidance and navigation of live working manipulator, we proposed a kind of subtask hierarchical deep reinforcement learning (HDRL) based on manipulator intelligent control model, and designed the corresponding environment state space, action strategy, and reward function. Meanwhile, we performed simulation experiments to verify the effect. The simulation results show that the success rate of obstacle avoidance and navigation model trained by single DRL algorithm is less than 35%. While the HDRL model not only gets better training convergence, but also has significant improvements in success rate which are 90% in wire stage and 71.01% in device stage. Moreover, the proposed safe operation rewards function can effectively guide an optimal route for manipulator considering the safety distance. The HDRL model keeps generalization performance in stage with different targets, providing theoretical and technical references for autonomous operation of live working manipulator.

Key words：live working; autonomous obstacle avoidance; autonomous navigation; hierarchical reinforcement learning; deep reinforcement learning; intelligent control

0 引言

配电线路网络是直接面向用户的电力基础设施，其覆盖面大、网络复杂，运行安全直接关系着供电系统的稳定与可靠，是保障电力供应安全和居民用电可靠的关键环节。为进一步减少停电时间，提高居民用电满意度，需要频繁开展带电作业操作。当前配电线路带电作业仍主要依赖人工攀爬 10 kV 电线杆，然后借助绝缘操作杆或绝缘斗臂车进行作业[1]，作业人员均处于高空、高电压的工作环境，存在安全风险，且对作业人员的技术水平要求较高。配电网带电作业操作环境危险复杂，作业任务标准、规范，适宜应用机器人替代人类开展带电作业[2-3]。自 80 年代起，美国、加拿大、西班牙、法国、日本等发达国家都先后开展了带电作业机器人的研究，根据其自动化程度主要可划分为人工遥控、远程遥操作和自动作业这 3 代机器人。90 年代末开始，国内的一些高校和研究机构开始了带电作业及巡检机器人的研究[4]。其中，山东电科院和鲁能智能共同开发了四代主从控制式带电作业机器人，但仍主要依赖人工操控[5-6]。

基金资助项目：国家重点研发计划(2018YFB1307400)。
Project supported by National Key R&D Program of China (2018YFB1307400).

2016 年，随着“AlphaGo”打败李世石，人工智能技术迎来新一轮发展热潮，深度强化学习是其中最受期待的关键技术之一，利用具有强大拟合能力的深度神经网络建立策略函数，将传统的值迭代计算转化为对神经网络的训练，打破了经典强化学习对于状态维度和离散动作的限制。虽然存在奖励稀疏、样本采集困难、稳定性较差等问题，但已在游戏、机器人控制、对话系统、自动驾驶、节能、电力系统和复合能流优化等领域发挥了重要作用[7-10]。

研究人员开展了深度强化学习在工业机器人抓取、定位和避障等领域的探索研究，并取得了一定成效[11-14]。文献[15]基于深度强化学习，通过工具在不同作业任务的表现效果，评估作业工具最优抓取点和抓取稳定性，实现机械臂根据不同任务类型，稳定抓取不同形状的工具完成无障碍地导航并完成作业。文献[16]将强化学习方法用于机械臂的避障导航任务，实现了机械臂无碰障路径规划。但是模型仅针对单一目标点进行训练和测试，没有实现全作业空间决策。文献[17]提出将机械臂 3 段臂视为 3 个不同智能体，限制 3 段臂始终处于同平面且分别设置策略，实现了机械臂在全作业空间决策。通过降低自由度降低了探索复杂度，但固定的姿态无法适应复杂环境带电作业对安全距离的要求[18]。分析以上研究成果可以看出，机械臂的避障导航过程存在不同状态—策略分布差异大和正向反馈奖励稀疏两大核心问题。仅使用单一的强化学习模型不能完整的探索空间，易造成策略陷入局部极小，且不能通过长时间多轮次的训练得出全空间机械臂避障导航控制策略。

本文通过深入分析机械臂的带电作业过程和动作次序，提出采用层次强化学习模型，建立分段决策并设置中间奖励，解决策略链过长及反馈奖励稀疏导致的空间探索和策略学习困难的问题，实现机械臂的避障导航，并保证足够的安全作业距离，为进一步实现全自主带电作业奠定基础。

1 深度强化学习技术

假设在时间 t，环境的状态空间量为 s_t，智能体获得来自环境的反馈奖励 r_t，r_t 是状态量 s_t 的函数，可表示为 $r_t=g(s_t)$。为了得到 t 时刻的动作，智能体通过计算策略分布函数 $\pi(\cdot|s)$计算或者采样得到 $a_t \in A$，其中 A 是智能体可能动作的集合。智能体通过执行 a_t，使环境状态量 s_t 转移到 s_{t+1}，其中 $s_{t+1}=f(s_t, a_t)$。重复执行直到完成任务目标或者到达上限 T，得到可应用于强化学习训练的策略轨迹，表示为 τ，见式(1)。

$$\tau=\{(s_t,a_t,r_t)\,|\,t=1\cdots T\} \tag{1}$$

为了评价各个状态 s_t 的价值，强化学习定义了状态价值函数 $V^{\pi}(s_t)$及状态–动作价值函数 $Q^{\pi}(s, a)$，用于评估处于各状态及执行相应动作时未来可能获得的奖励期望，见式(2)和式(3)：

$$V^{\pi}(s)=E\left(R_{t:\infty}\,|\,s_t=s,\pi\right) \tag{2}$$

$$Q^{\pi}(s,a)=E\left(R_{t:\infty}\,|\,s_t=s,a_t=a,\pi\right) \tag{3}$$

式中，$R_{t:\infty}$为累积折扣奖励见式(4)。

$$R_{t:\infty}=\sum_{i=0}^{\infty}\gamma^i r_{t+i} \tag{4}$$

式中γ为奖励折扣因子。

由此可定义优势函数 $A^{\pi}(s, a)$，用于表征状态 s 下，动作 a 相对于动作集合 A 中其他元素的优劣，见式(5)。

$$A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s) \tag{5}$$

1.1 深度 Q 网络

深度 Q 网络(deep Q network, DQN)强化学习算法模型是经典的以值函数为基础的算法，是在 Q-Learning 算法基础上发展而来。以深度神经网络替代 Q-Value 表，以最大化 Q 值为目标的一种强化学习算法，若定义 i 时刻神经网络模型参数为 θ_i，其损失函数 L^{DQN} 见式(6)。

$$L^{\mathrm{DQN}}(\theta_i)=E((y_i-Q(s,a;\theta_i))^2\,|\,s,a\sim\pi) \tag{6}$$

$$y_i=E(r+\gamma\max_{a'}Q(s',a';\theta_{i-1})\,|\,s'\sim f(s,a);s,a\sim\pi) \tag{7}$$

1.2 异步优势行动者–评论家

异步优势行动者–评论家(asynchronous advantage actor-critic, A3C)方法基于策略梯度，使用 Actor-Critic 框架并引入分布式训练，在提升算法性能和训练速度。不同于 DQN，A3C 算法输出策略概率分布 $\pi(\cdot|s)$，通过采样得到动作，属于不确定性输出算法。若定义 Actor-Critic 模型参数为 θ 和 θ'，网络参数更新梯度见式(8)。

$$\begin{cases}\theta\leftarrow\theta+\alpha\nabla_{\theta}\log\pi(a_t\,|\,s_t;\theta)A(s_t,a_t)\\ \theta'\leftarrow\theta'-\alpha\nabla_{\theta'}(R_{t:t+n-1}+\gamma^n V(s_{t+n};\theta'^{-})-V(s_t;\theta'))^2/2\end{cases} \tag{8}$$

$$R_{t:t+n-1}=\sum_{i=0}^{n-1}\gamma^i r_{t+i} \tag{9}$$

式中：$R_{t:t+n-1}$ 为 n-step 累积折扣奖励；α 为步长系数；θ'^{-} 为 Critic 网络异步参数。

1.3 深度确定性策略梯度

深度确定性策略梯度(deep deterministic policy

gradient, DDPG)方法则是结合了值函数法和策略梯度两类方法的特点。它基于 Actor-Critic 框架，但是使用 Q 函数作为评价模型，Actor 网络则计算期望 Q 值最高的动作输出。其目标函数为：

$$L_{\omega}^{\mathrm{DDPG}}=\min_{\omega}E_{\pi}((r_t+\gamma Q^{\omega}(s_{t+1},a_{t+1})-Q^{\omega}(s_t,a_t))^2)/2 \tag{10}$$

$$L_{\theta}^{\mathrm{DDPG}}=\max_{\theta}E_{\pi}(Q^{\omega}(s_t,\mu_{\theta}(s_t))) \tag{11}$$

式中：ω表示 Critic 模型参数；μ表示策略动作。

1.4　分布式近端策略优化(DPPO)

分布式近端策略优化(distributed proximal policy optimization, DPPO)深度强化学习算法是基于 A3C 算法和置信域策略优化算法(trust region policy optimization, TRPO)改进而来的策略单调提升学习方法，是一种训练效率更高且鲁棒性更强的方法[19-20]。其损失函数有 2 种形式，分别为：

$$\begin{aligned}L_t^{\mathrm{CLIP}}(\theta)=E(&\min(\eta A_{\pi_{\theta\mathrm{old}}}(s_t,a_t),\\&clip(\eta,1-\varepsilon,1+\varepsilon)A_{\pi_{\theta\mathrm{old}}}(s_t,a_t)))\end{aligned} \tag{12}$$

$$\begin{aligned}L_t^{\mathrm{KLPEN}}(\theta)=E(&\eta A_{\pi_{\theta\mathrm{old}}}(s_t,a_t)\\&-\beta_{\mathrm{KL}}(\pi_{\theta_{\mathrm{old}}}(\cdot|s_t)|\pi_{\theta}(\cdot|s_t)))\end{aligned} \tag{13}$$

新旧策略概率比为

$$\eta=\pi_{\theta_m}(a_t|s_t)/\pi_{\theta_{m\mathrm{old}}}(a_t|s_t) \tag{14}$$

式中：A_{π}表示优势函数；ε 为η的步长限制系数；β 为 KL 散度的惩罚系数。

DPPO 和 A3C 一样，作为不确定性输出算法输出为策略分布，作为 on-policy 算法依赖完整策略来计算目标函数。本文重点以 DPPO 模型中的 L^{CLIP} 损失函数为基础算法展开研究。

2　基于分层深度强化学习的机械臂自主作业算法模型

2.1　仿真训练环境

配电网典型机器人带电作业类型包括带电拆接引、导线清障、更换绝缘子等。考虑作业中可能出现的障碍物及限制，基于 python 建立两类抽象仿真模型，分别为跨线作业和设备作业，如图 1 所示，图中 *xyz* 为作业空间坐标系，机械臂自由度及参考系设定见附录。

2.1.1　跨线作业场景

跨线作业如图 1(a)所示，由于在线路延伸方向上的障碍无限长，只有机械臂在以线路延伸方向为法线的平面内做出跨越动作才能实现有效避障，故可以将动作限制在该平面内。在该模型中，机械臂的第 1、2 段臂与目标点共平面，则机械臂的有效自由度限制为第 2、3、5 自由度。

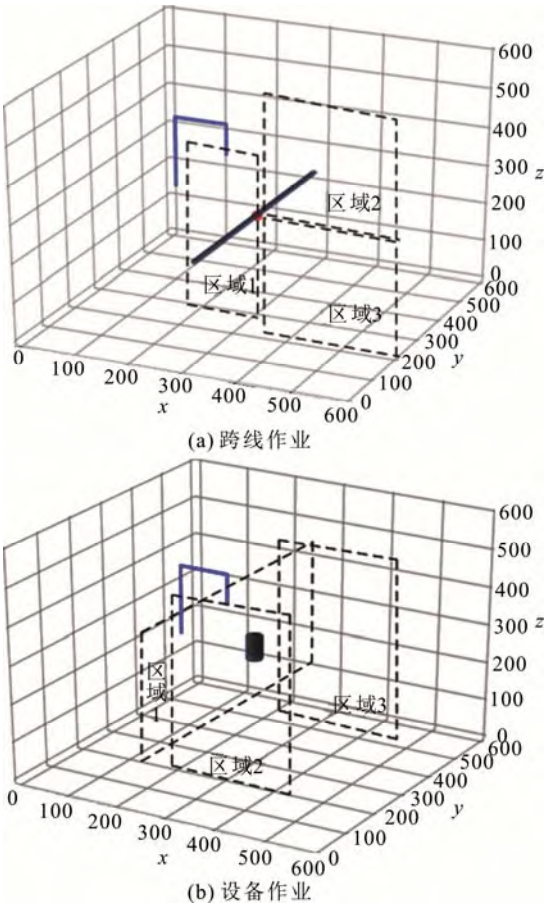


图 1　典型配电网带电作业抽象仿真模型

Fig.1　Simulation model for typical live working of distribution power grid

将跨线作业区域依照相对位置分为 3 部分，分别对应障碍前方区域 1，障碍后方影响较大的区域 2 和障碍后方影响较小的区域 3，用于比较不同区域模型的避障导航成功率。

2.1.2　绝缘子等设备作业场景

在对绝缘子等设备作业时，可能的障碍物为其他绝缘设备，如图 1(b)所示。由于障碍物在空间分布有限，机械臂在全空间避障行为有效，无法限制运动自由度，故建立一般 3 维模型进行仿真。以绝缘子障碍为例，由于带电作业安全距离的限制，实际作业中这些绝缘子的外观细节对于机器人路线规划不会造成影响，可以将绝缘子看作圆柱体型障碍物。由于机械臂第 6 自由度主要用于调节专用工具的姿态，实际无法改变第 3 段臂的空间位置，故在模型中不考虑该自由度。

将设备作业区域同样分为 3 部分，分别为机械臂与障碍的中间区域 1，右方区域 2，左方区域 3。

用于比较不同区域模型的避障导航成功率。

2.2 分层深度强化学习算法模型

自主作业目标是使机械臂在存在障碍的环境内，能够在不碰障碍的前提下，到达作业范围内任意目标点。考虑到机械臂要同时完成避障和导航 2 个任务，在训练阶段初期，机械臂会多次触碰障碍而获得环境反馈的惩罚，却很难探索到使机械臂末端到达目标点的动作来获得奖励，这造成了强化学习中常见的稀疏回报问题。稀疏回报使得训练严重放缓，且更容易使智能体策略陷入局部极小，无法达到最优解，在机械臂避障导航模型中表现为机械臂可以避障却无法完成导航任务。

本文提出一种层次深度强化学习算法模型，在处理交互轨迹长、奖励反馈稀疏的决策问题上，将任务目的分为数个阶段性的子任务，智能体在完成前一个子任务的基础上再继续执行后续任务[21]。将机械臂的避障导航任务转化为先完成避障再进行导航的顺序性任务，模型结构如图 2 所示。

其中，避障需要保证在机械臂运动过程中 3 段臂均不触碰障碍，直至第 1 段臂到达相对安全且保证后两段臂能够完成导航的位置。选择 on-policy 方法中高效稳定的 DPPO 算法训练模型避障性能。

导航过程要保证避障稳定，采用确定性输出的深度强化学习模型如 DQN、DDPG 算法，避免从策略分布中采样出错误动作。之后固定第 1 段臂，仅使用后 2 段臂完成导航。在跨线作业中由于状态空间维度低使用 DQN 算法处理；在设备作业中，由于状态空间维度高，更适用 DDPG 算法。

2.3 奖励函数设计

2.3.1 人工势场奖励函数

对于避障任务，使用人工势场对障碍物设置排斥势，如图 3(a)所示。对机械臂的每段臂而言，进入障碍物危险区时会获得反比于障碍到该臂最短距离平方的惩罚；当进入接触障碍范围时会得到更高的定值惩罚 c_2。惩罚函数 r_{obstacle} 为

$$r_{\text{obstacle}} = \sum_{i=1}^{3} r_i \tag{15}$$

其中

$$r_i = \begin{cases} 0 & , \ d_{i\min} > l_{\text{danger}} \\ -c_1 / d_{i\min}^2 & , \ l_{\text{touch}} < d_{i\min} \le l_{\text{danger}} \\ -c_2 & , \ d_{i\min} \le l_{\text{touch}} \end{cases} \tag{16}$$

式中：l_{danger} 为危险区半径；l_{touch} 为接触半径；d_{imin} 为障碍到臂最短距离；c_1—c_{10} 为奖励函数系数，大小为正。

为使机械臂接近目标点，对目标点设置吸引势，如图 3(b)所示。奖励值与末端到目标的距离平方呈反比。当接触目标时，会得到更高的定值奖励 c_4。奖励函数 r_{target} 为

$$r_{\text{target}} = \begin{cases} c_3 / d_{\text{target}}^2, & d_{\text{target}} > l_{\text{get}} \\ c_4, & d_{\text{target}} \le l_{\text{get}} \end{cases} \tag{17}$$

式中：l_{get} 为目标接触半径；d_{target} 为末端到目标的距离。

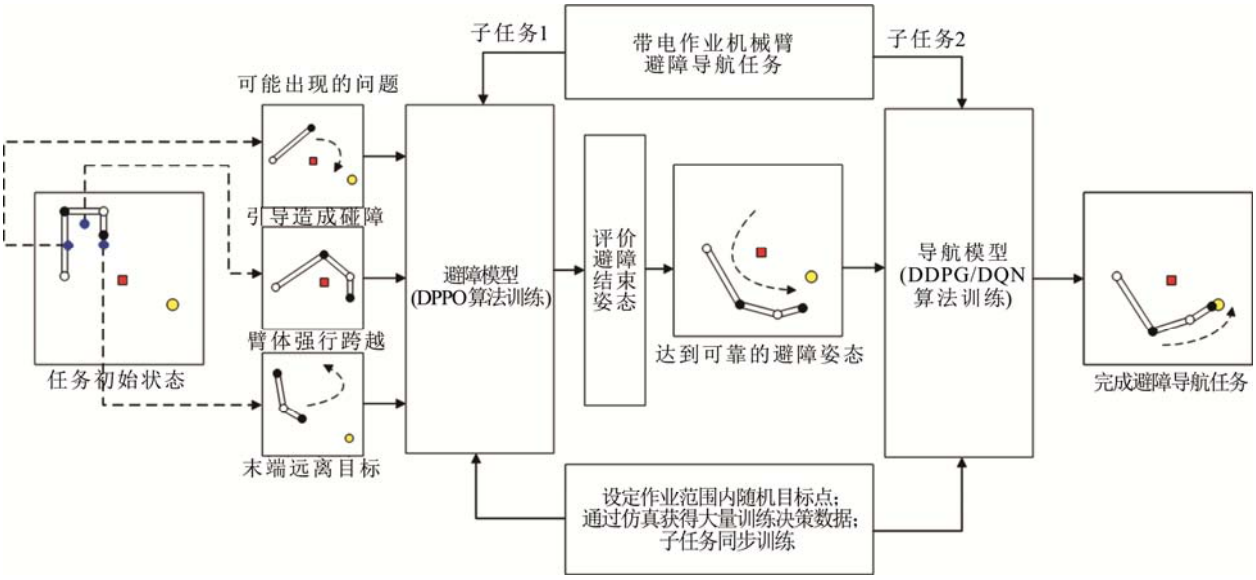


图 2 层次深度强化学习模型

Fig.2 Hierarchical deep reinforcement learning model

2.3.2　危险姿态奖励函数

机械臂危险姿态为障碍距离某段机械臂过近以及各段机械臂对障碍形成环绕。前者可以使用排斥势予以规避，后者需要判断危险姿态给与惩罚来进行限制，危险姿态及其对应几何关系见图 4。

在 3 维抽象模型中，机械臂 4 节点形成空间四面体，前 3 节点 *OAB* 和后 3 节点 *ABC* 可以确定 2 个平面。以 *OAB* 所在平面为例，作剩余节点 *C* 和障碍点 *W* 的投影，若在该平面内，障碍投影点 *W*′ 位于 4 个节点组成的四边形内 *OABC*′且障碍点到该平面的距离 $d_{\mathrm{TT'}}$过小，则视为危险状态，平面 *ABC* 同理；在 2 维抽象模型中，所有点处于同一平面，直接判断障碍是否处于 *OABC* 四边形中。若空间四边形 *OABC* 表面点集合为 Ψ，则奖励函数 r_{quad} 为

$$r_{\mathrm{quad}}=\begin{cases}-c_5, & W_{\mathrm{proj}}\bigcap\psi\neq\varnothing \\ & \text{且}\ \exists P_{\mathrm{proj}}\in W_{\mathrm{proj}},|WP_{\mathrm{proj}}|<d_{\mathrm{lim}} \\ 0, & \text{其他}\end{cases} \tag{18}$$

其中

$$\begin{aligned}W_{\mathrm{proj}}=&\{WW'\bigcap OABC'\,|\,WW'\perp OABC'\}\bigcup\\&\{WW''\bigcap O'ABC\,|\,WW''\perp O'ABC\}\end{aligned} \tag{19}$$

式中：Ψ 为空间四边形 *OABC* 表面点集合；d_{lim} 为距离 $d_{\mathrm{TT'}}$的阈值。

2.3.3　快速引导奖励函数

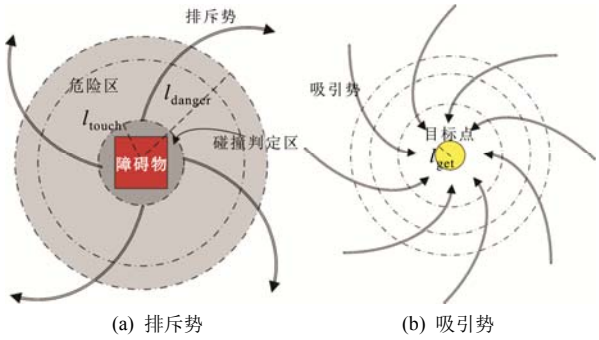
为了更快速地引导机械臂绕过障碍，需对第 1 段臂加以引导。由于第 1、2 段臂始终同平面，且以坐标轴 *z* 为轴旋转，3 维导航模型投影在 *xy* 平面内将简化为 2 维平面模型，因此第 1、2 段臂所构成平面位置决定了避障性能。设原点到一关节向量、原点到障碍点向量、原点到目标点向量在 *xy* 平面的投影分别为向量 ***OA***、***OW***、***OT***。

当 ***OA***、***OT*** 相对 ***OW*** 同侧时，以向量 ***OA***、***OT*** 的夹角为依据引导机械臂移动到与目标点同侧位置，如图 5(a)所示。在跨线场景中，由于自由度限制，不能通过直接引导 1 段臂跨跃障碍。本实验通过引导 1 段臂使用“绕远路”的方式到达安全侧，当 ***OA***、***OT*** 处于 ***OW*** 异侧时，给予反比于夹角的惩罚项，如图 5(b)所示。则引导奖励函数 r_{3Dguide} 及 r_{2Dguide} 形式为：

$$r_{\mathrm{3Dguide}}=\begin{cases}-c_6\theta_{\mathrm{TA}}/\pi & (h(\boldsymbol{OT},\boldsymbol{OA})<0)\\ 0 & (h(\boldsymbol{OT},\boldsymbol{OA})\geq 0)\end{cases} \tag{20}$$

$$r_{\mathrm{2Dguide}}=\begin{cases}-c_7(1-\theta_{\mathrm{WA}}/\pi) & (h(\boldsymbol{OT},\boldsymbol{OA})<0)\\ 0 & (h(\boldsymbol{OT},\boldsymbol{OA})\geq 0)\end{cases} \tag{21}$$

式中，$h(\boldsymbol{OT},\boldsymbol{OA})=(\boldsymbol{OW}\times\boldsymbol{OT})\cdot(\boldsymbol{OW}\times\boldsymbol{OA})$，$\theta_{\mathrm{TA}}$、$\theta_{\mathrm{WA}}$ 夹角如图 5 所示。



(a) 排斥势　　(b) 吸引势

图 3　人工势场法示意图

Fig.3　Artificial potential field diagram

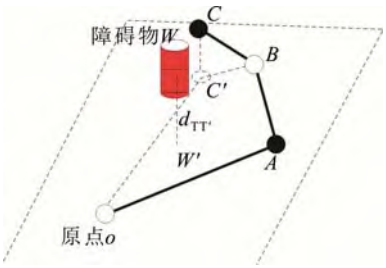


图 4　危险姿态示意图

Fig.4　Dangerous posture diagram

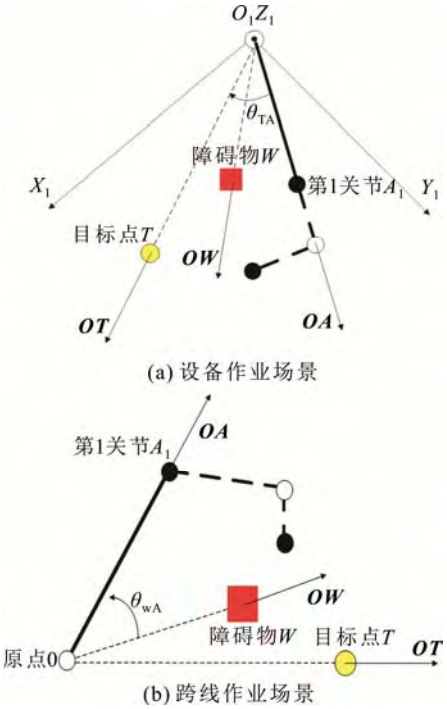


图 5　快速引导奖励函数示意图

Fig.5　Efficient-guidance function diagram

当危险姿态和引导规则奖励函数取值为零且第 1 段臂末端点所处位置满足 2、3 段臂完成导航任务时，判定为完成避障任务并给与相应奖励值 c_8。

对于分层模型中的无障碍导航任务，设定末端

引导奖励为负，大小正比于机械臂末端到目标点距离，正向奖励设置为到达目标点的奖励。奖励函数 r_{navi} 为

$$r_{\text{navi}}=\begin{cases}-d_{\text{target}}/c_9 & (d_{\text{target}}>l_{\text{get}})\\ c_{10} & (d_{\text{target}}\le l_{\text{get}})\end{cases} \tag{22}$$

2.4 网络构建和训练策略

机械臂避障导航层次深度强化学习模型关系如图 6 所示，图中 FC 表示全连接神经网络。

2.4.1 状态空间编码模块

状态空间编码模块负责将机械臂和环境的交互信息整理为输入到神经网络模型的状态空间数组。用于避障训练的状态空间 s_{avoid} 包含目标点向量、障碍物向量、末端到目标点向量、臂关节障碍物的向量、障碍到各段臂的最短距离，是否到达目标点及是否处在安全姿态；用于导航训练的状态空间 s_{navi} 包含目标点向量、末端到目标点向量和是否到达目标点指示量。

2.4.2 避障学习模块

避障学习模块负责执行 DPPO 算法训练机械臂完成避障任务。建立以 Actor-Critic 框架为基础的神经网络模型，设置子网络用于分布式训练。

2.4.3 导航学习模块

导航学习模块执行 DQN 或 DDPG 算法训练机械臂的第 2、3 段臂完成导航任务。跨线模型使用 Double DQN 方法；设备作业模型使用 DDPG 算法。

2.4.4 主控切换模块

主控模块判定环境当前状态是否满足避障任务达成条件，同时用于调整状态空间编码器输出的状态空间形式。若机械臂第 1 段臂末端到目标点的距离小于第 2、3 段臂之和，且式(17)、式(18)和(19)中定义的奖励函数取值为 0，则以当前状态执行导航训练。

3 实验结果分析对比

3.1 设定

跨线作业中，设置单一目标点和全空间随机目标点 2 个实验环境。单一目标的实验环境中验证危险动作判定和引导奖励函数的有效性，比较各类深度强化学习算法性能；全空间随机目标点测试不同模型全空间动作效果。

在设备作业仿真模型中，设计全空间随机目标点的实验环境，利用验证集验证模型在高维空间同样具备有效性。实验环境由于目标点不同导致初始奖励不同，故无法使用累计奖励判断训练效果。

3.2 跨线作业场景单一目标点

本文在跨线作业场景模拟环境中，对 DQN、DDPG 和 DPPO 算法进行仿真训练，仿真结果见图 7，训练过程累积奖励曲线见图 8。

如图 7 所示，经过训练机械臂均可实现绕过线状障碍并到达目标点。结合图 8 中各算法训练累积奖励曲线和表 1 中奖励曲线趋近收敛时的训练轮数和累积奖励大小均值可以看出，DQN、DDPG 和 DPPO 这 3 种算法的训练均可到达收敛值，此时的网络模型可以驱动机械臂达成设定的避障导航任

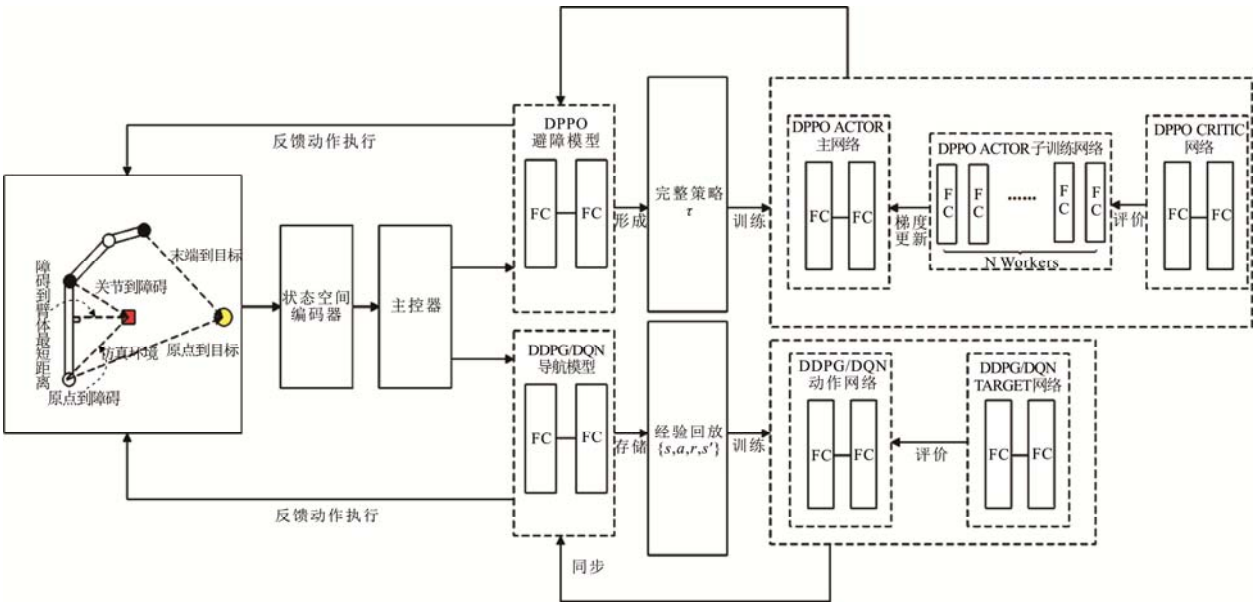


图 6 机械臂避障导航层次深度强化学习模型结构图

Fig.6 Hierarchical deep reinforcement learning model for obstacle-avoidance and target-guidance of working manipulator

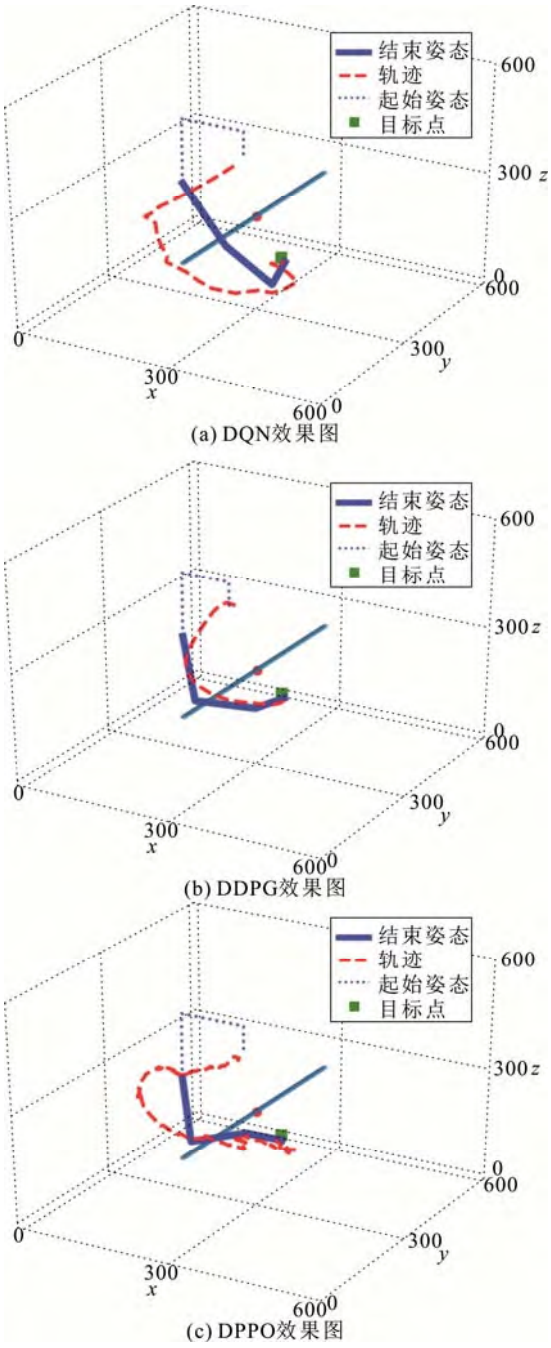


图 7　各算法仿真训练效果图

Fig.7　Simulation effect pictures of different algorithm training processes

务。但各算法之间训练性能不尽相同。可以看出 DPPO 算法收敛更快、获得奖励更高。DDPG 和 DQN 的训练也能逼近收敛但相对震荡更大，终止训练的时机对模型性能有更多影响。

危险动作和引导奖励效果对比轨迹如图 9 所示。当目标点与 1 段臂相对障碍处于异侧时，未加入四边形和引导奖励策略的训练更容易使机械臂以寻求最短路径的方式到达目标。相比之下，加入额

表 1　单一目标点下各强化学习算法训练性能比较

Table 1　Performance comparison between different RL algorithms under single target situation

算法	累积奖励	训练轮数
DQN	52.79±6.47	774±37
DDPG	103.41±9.72	553±41
DPPO	109.26±3.85	326±34

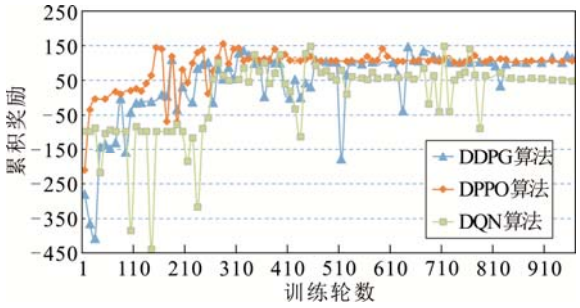


图 8　各算法奖励函数曲线对比图

Fig.8　Comparison of accumulate rewards from different algorithm training processes

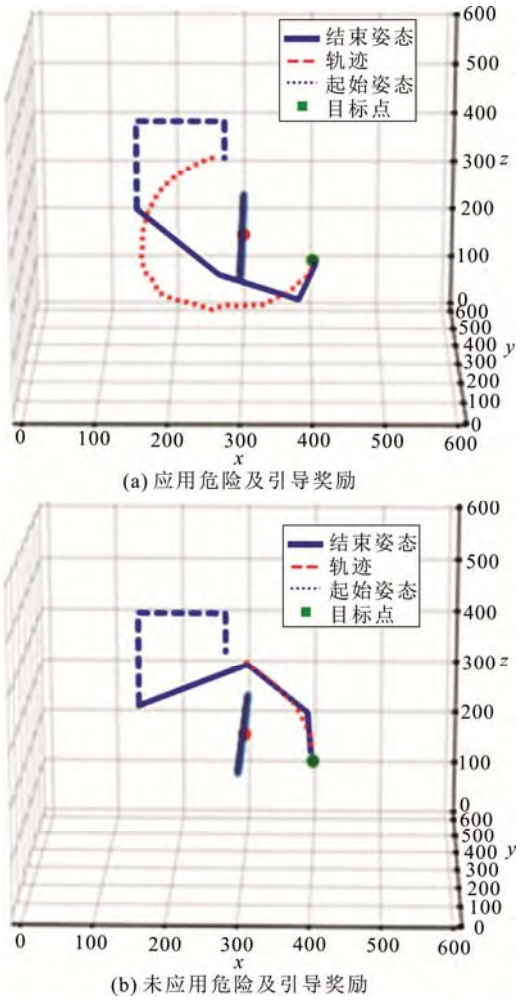


图 9　同目标点不同奖励数轨迹比较图

Fig.9　Comparison of the trajectories produced by different reward functions under the same target point

外奖励让机械臂学会绕开障碍点的路线，避免了接近障碍带来的损失。

3.3 跨线模型随机目标点场景

由于目标点不同造成任务之间的固有奖励不同，策略不再满足独立同分布，使得训练奖励也不会趋于稳定收敛值，单一模型难以学到统一避障导航方法。各类算法训练得到的机械臂运动轨迹如图10所示。从图10(a)—图10(c)中各模型轨迹来看，其可以完成避障。但当目标较远或位置难以到达时，各模型便无法顺利抵达目标点，在无法执行有效策略时的触发混乱。各算法对应的成功率、碰障率和未完成率如表2所示，可以看出，单一模型对全空间目标点任务成功率都较低，难以可靠完成多目标点任务。

相对于单模型，层次强化学习模型在任务完成率上有了很大的提高。图10(d)—图10(f)中展示了其运用于不同区域内目标点的导航运动情况。可以看出模型在危险姿态和引导奖励中学到相对安全的

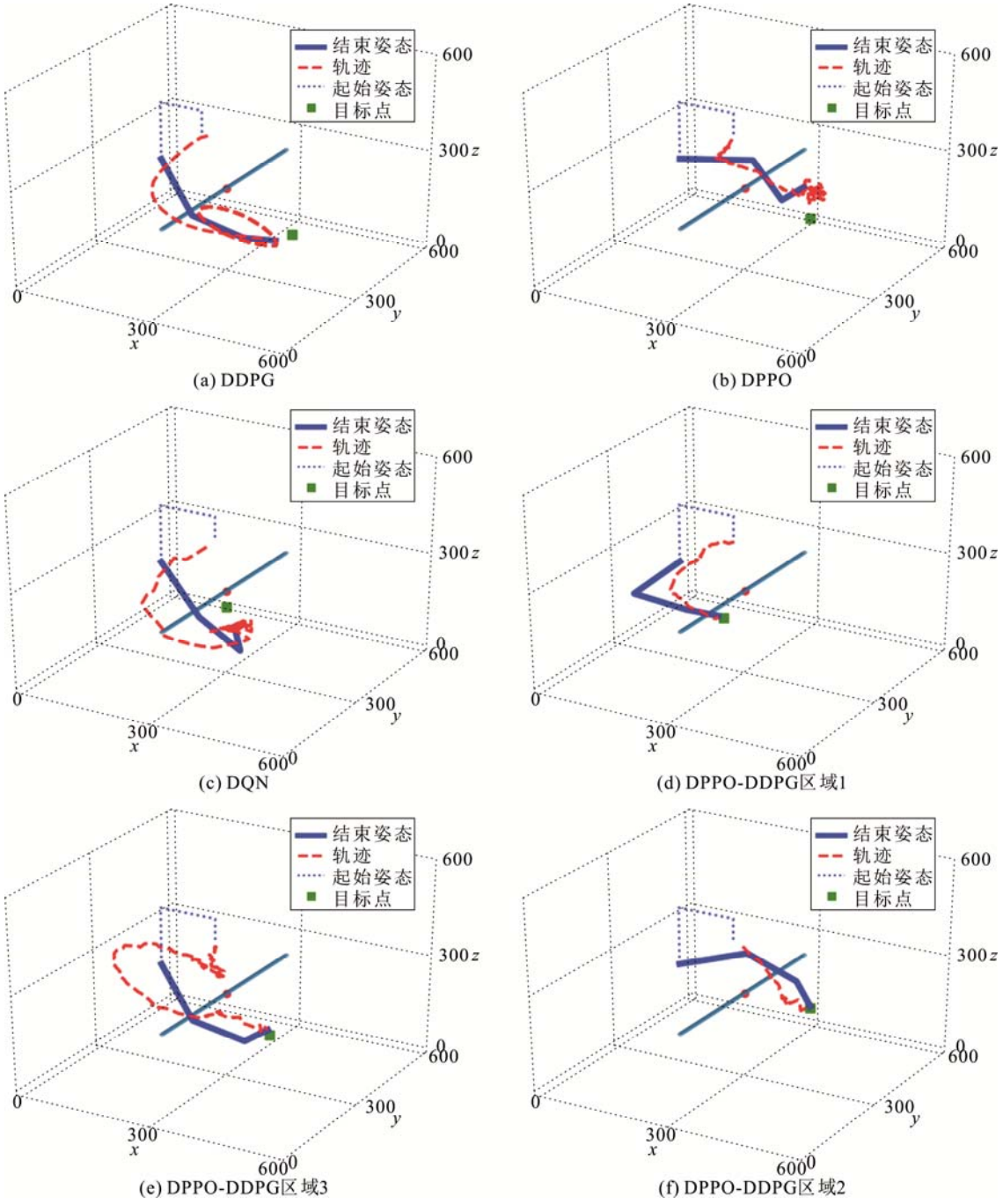


图10 跨线模型随机目标点各算法测试效果图

Fig.10 Effect pictures of different reinforcement learning models in line stage with random targets

路径，能够稳定地到达目标点完成任务。层次强化学习模型中各子任务的训练任务完成率及累积奖励值见图 11。从累积奖励的收敛趋势来看，任务分解使训练变得稳定；从任务成功率提升来看，其趋于平稳且保持高水平。子任务的高完成率保证其组合模型的任务高成功率，如表 2 所示。其中，DQN 导航训练的高成功率使模型未完成的概率明显降低。

层次强化学习算法各区域目标点测试结果如表 3 所示，对比各区域的成功率还可以看到，触碰障碍主要集中于区域 2，未完成主要发生于区域 1。未完成的主要原因为区域 1 取样样本少， DPPO 输出训练输出策略分布的方差大，容易采样到不合理动作。碰障的发生集中在难度最高的区域，根本在于模型本身存在失误率，DPPO-DQN 层次强化模型的避障成功率趋近于 91%，网络参数需要在图 11 中所示 DPPO 避障稳定区进行优选。由于模型整体失败概率低且不存在死点，可以通过重复决策来找到完成任务的路径。

3.4　设备作业场景随机目标点

为验证层次模型在高维的处理能力，使用相同结构(仅改变输入输出维度以适应算法需要)的 DPPO-DPPG 层次深度强化模型开展训练，并在图 1(b)中的验证集中进行测试，选取各方位的数个仿真结果展示于图 12 可以看到，DPPO 算法训练的避障部分依旧有效表现相对稳定，针对不同方位的目标点可以准确找到满足引导策略的位置，从而顺利动作第 1 段臂。分别在 3 个区域的验证集中统计测试模型，结果如表 4 所示。设备作业模型成功率相比较跨线模型有下降，未完成率有提升，原因是动作自由度和状态空间维度提升使网络处理能力下降，需要根据空间维度提升量增加网络节点数及深度来补偿。

和跨线作业场景相同，设备作业场景中层次模型的子任务训练趋于稳定，如图 13 所示。分析避障

表 2　跨线场景各强化学习算法随机目标点测试结果比较

Table 2　Test results comparison between different RL algorithms in line stage with random targets

算法	成功率/%	碰障率/%	未完成率/%
DQN	30.79	11.26	57.95
DDPG	28.03	14.45	57.52
DPPO	32.89	12.48	55.67
DPPO+DQN	90.00	7.89	2.11

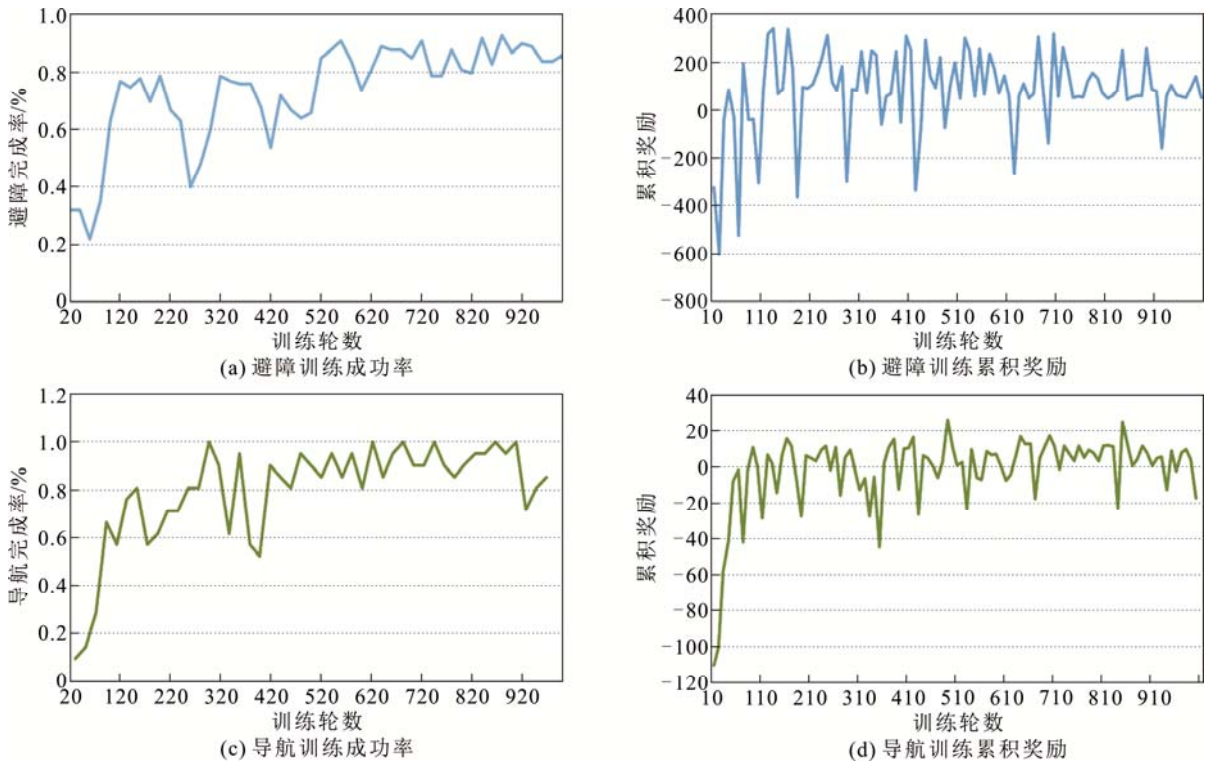


图 11　跨线作业场景层次深度强化模型训练完成率及累积奖励曲线

Fig.11　Completion rate and accumulate reward curve in training process for hierarchical deep reinforcement learning model in line stage

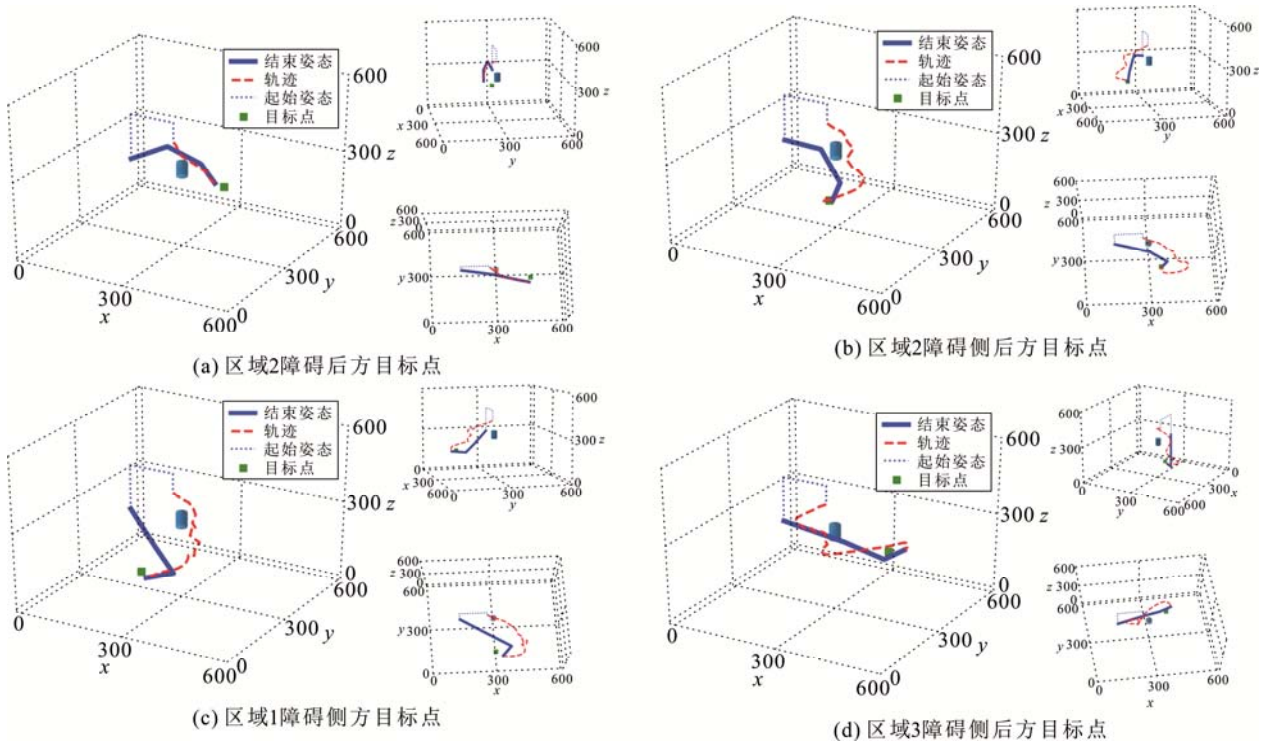


图 12 在设备作业场景下层次深度强化模型训练得到的机械臂运动轨迹和姿态三维空间效果图

Fig.12 Posture and trajectories of Hierarchical deep reinforcement learning model for random targets in device stage

表 3 跨线场景层次强化学习算法全空间目标点测试结果

Table 3 Test results of hierarchical deep reinforcement learning model in line stage

位置	成功率/%	碰障率/%	未完成率/%
区域 1	91.56	0.44	8.00
区域 2	83.37	15.42	1.20
区域 3	94.01	5.81	0.17

表 4 设备作业场景层次强化学习算法全空间目标点测试结果

Table 4 Test results of hierarchical deep reinforcement learning model in device stage

位置	成功率/%	碰障率/%	未完成率/%
区域 1	73.33	0	26.67
区域 2	74.25	5.18	20.57
区域 3	66.25	13.75	20.00
全部	71.01	6.88	22.03

训练曲线，可以看出状态空间复杂度的上升，对训练稳定造成影响，其任务完成率和累积奖励震荡范围明显加大，但成功率仍保持相对较高水平。

从图 13 中 DDPG 训练曲线中可以看到，其导航完成率收敛值趋近于 1，且奖励值接近收敛于最大值，表示其可以完成绝大部分目标点及多臂姿的导航任务，具有较高的任务成功率，且不易出现危险姿态，表现出算法在高维空间导航问题的强学习能力。

实验中机械臂末端的轨迹不是最优化的，是由于避障结束时后 2 段臂姿态无法限制，并且 DPPO 网络训练效率高，造成可供后续应对初态变化大的泛化训练样本减少。同时维度的扩大增加了状态量 s 的探索难度，即要求更充分地探索训练才可以提高此路线规划精度。由于以上原因限制了训练的多样性，间接影响了模型的泛化性能，可以通过预训练 DDPG 网络模型来解决。

对比图 13 中子任务训练成功率曲线稳定区域均值和表 3 整体成功率看到模型避障导航成功率低于 2 个子任务结合应有的成功率。原因为 2 个网络模型不能同时达到最佳，或因评价完成率地随机取点分布中心差异大导致 2 个模型分别在不同方位表现出最佳性能；DDPG 导航的离线随机取样训练方式也容易出现过量取样造成过拟合，使避障后的末态臂姿对决策的影响增加，造成了未完成率的提高和成功率的下降，可以进一步调整网络参数、训练轮数及组合不同批次训练模型来提升测试效果。

4 结论

1）提出的层次深度强化学习模型适宜解决带

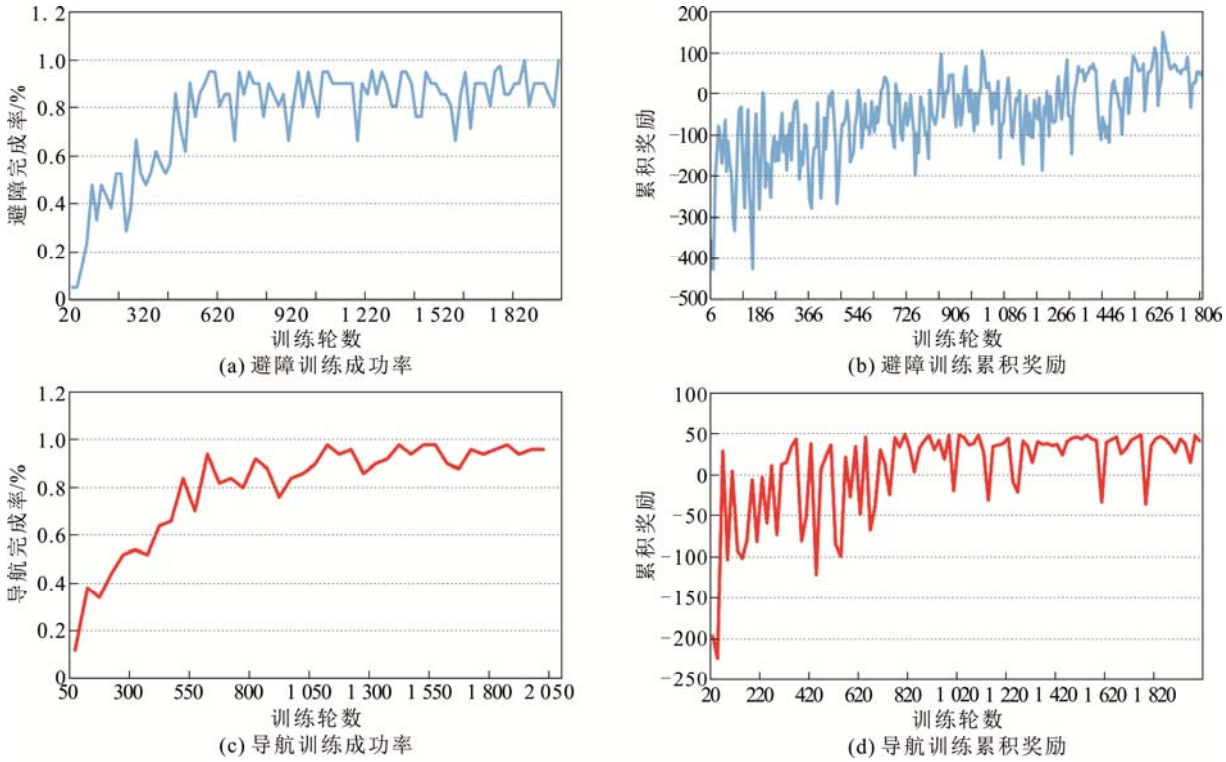


图 13　设备作业场景层次深度强化模型训练完成率和累积奖励曲线

Fig.13　Completion rate and accumulate reward curve in training process for hierarchical deep reinforcement learning model in device stage

电作业机器人作业避障导航问题，是实现带电作业机械臂自主控制的有效解决方案。

2）对比分析了各类深度强化学习算法模型在带电作业机械臂避障导航问题中的实验结果和表现。离散策略模型中，DDPG、DQN 适宜解决不依赖具体环境的无障碍导航问题；连续策略模型 DPPO 适宜解决连续动作避障问题，具有较强的鲁棒性，且其在线学习的特点使其更适合针对不同环境作特殊训练，与避障任务更契合。

3）该方法后续可进一步解决计算机视觉实现环境状态参数的读取和训练，最终实现带电作业机器人的自主导航作业。

附录见本刊网络版(http://hve.epri.sgcc.com.cn/CN/volumn/current.shtml)。

参考文献　References

[1] 胡　毅，刘　凯，彭　勇，等. 带电作业关键技术研究进展与趋势[J]. 高电压技术，2014，40(7)：1921-1931.
HU Yi, LIU Kai, PENG Yong, et al. Research status and development trend of live working key technology[J]. High Voltage Engineering, 2014, 40(7): 1921-1931.

[2] 刘　庭，唐　盼，周炳凌，等. 500 kV 线路绝缘斗臂车带电作业安全距离试验[J]. 高电压技术，2016，42(7)：2315-2321.
LIU Ting, TANG Pan, ZHOU Bingling, et al. Experiment research of minimum approach distance for live working used of 500 kV insulated aerial vehicles[J]. High Voltage Engineering, 2016, 42(7): 2315-2321.

[3] 刘存根，鲁守银，孙丽萍，等. 高压带电作业机械臂姿态监测系统研究[J]. 高电压技术，2015，41(3)：931-936.
LIU Cungen, LU Shouyin, SUN Liping, et al. Attitude monitoring system for high voltage electric power live line working manipulator[J]. High Voltage Engineering, 2015, 41(3): 931-936.

[4] 腾　云，陈　双，邓洁清，等. 智能巡检机器人系统在苏通 GIL 综合管廊工程中的应用[J]. 高电压技术，2019，45(2)：393-401.
TENG Yun, CHEN Shuang, DENG Jieqing, et al. Application of intelligent inspection robot system in Sutong GIL utility tunnel project[J]. High Voltage Engineering, 2019, 45(2): 393-401.

[5] 胡　毅. 输配电线路带电作业技术的研究与发展[J]. 高电压技术，2006，32(11)：1-10.
HU Yi. Research and development of live working technology on transmission and distribution lines[J]. High Voltage Engineering, 2006, 32(11): 1-10.

[6] 戚　晖，厉秉强，顾载新. 高压带电作业机器人绝缘防护技术研究[J]. 高电压技术，2003，29(5)：19-20.
QI Hui, LI Bingqiang, GU Zaixin. The research on insulation technology for HV hot line robot[J]. High Voltage Engineering, 2003, 29(5): 19-20.

[7] 陈艺璇，张孝顺，郭乐欣，等. 基于多智能体迁移强化学习算法的电力系统最优碳–能复合流求解[J]. 高电压技术，2019，45(3)：863-872.
CHEN Yixuan, ZHANG Xiaoshun, GUO Lexin, et al. Optimal carbon-energy combined flow in power system based on multi-agent

transfer reinforcement learning[J]. High Voltage Engineering, 2019, 45(3): 863-872.

[8] 万里鹏，兰旭光，张翰博，等. 深度强化学习理论及其应用综述[J]. 模式识别与人工智能，2019，1(1)：67-79.
WAN Lipeng, LAN Xuguang, ZHANG Hanbo, et al. A review of deep reinforcement learning theory and application[J]. Pattern Recognition and Artificial Intelligence, 2019, 1(1): 67-79.

[9] ZHANG D X, HAN X Q, DENG C Y. Review on the research and practice of deep learning and reinforcement learning in smart grids[J]. CSEE Journal of Power and Energy Systems, 2018, 4(3): 362-370.

[10] 刘建伟，高 峰，罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报，2019，42(6)：1406-1438.
LIU Jianwei, GAO Feng, LUO Xionglin. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.

[11] GU S, HOLLY E, LILLICRAP T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). Marina Bay Sands, Singapore: IEEE, 2017: 1-9

[12] YURI B, HARRI E, DEEPAK P, et al. Large-scale study of curiosity-driven learning[C] // 7th International Conference on Learning Representations. New Orleans, LA, United states: ICLR, 2019: 1-15

[13] ERIC J, COLINE D, VINCENT V, et al. Grasp2Vec: learning object representations from self-supervised grasping[EB/OL]. New York, USA: Cornell University, 2018[2019-08-22]. https: //arXiv. org/abs/1811.06964.

[14] LUO W H, SUN P, ZHONG F W, et al. End-to-end active object tracking and its real-world deployment via reinforcement learning[J]. Journal of Latex Class Files, 2015, 14(8): 1-16.

[15] FANG K, ZHU Y K, GARG A, et al. Learning task-oriented grasping for tool manipulation from simulated self-supervision[EB/OL]. New York, USA: Cornell University, 2018[2019-08-22]. https: //arXiv. org/abs/1806.09266.

[16] 李广创，程良伦. 基于深度强化学习的机械臂避障路径规划研究[J]. 软件工程，2019，22(3)：12-15.
LI Guangchuang, CHENG Lianglun. Research on obstacle avoidance path planning of the mechanical arm based on deep reinforcement learning[J]. Software Engineering, 2019, 22(3): 12-15.

[17] 徐 帷，卢 山. 基于 Sarsa(λ)强化学习的空间机械臂路径规划研究[J]. 宇航学报，2019，40(4)：435-443.
XU Wei, LU Shan. Analysis of space manipulator route planning based on Sarsa(λ) reinforcement learning[J]. Journal of Astronautics, 2019, 22(3): 12-15.

[18] 张秋实，王力农，方雅琪，等. 带电作业组合间隙放电特性仿真分析方法[J]. 高电压技术，2018，44(4)：1292-1301.
ZHANG Qiushi, WANG Linong, FANG Yaqi, et al. Simulation method for discharge characteristics of live working complex gap[J]. High Voltage Engineering, 2018, 44(4): 1292-1301.

[19] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. New York, USA: Cornell University, 2017[2019-08-22]. https: //arxiv. org/abs/1707.06347.

[20] HEESS N, DHRUVA T B, SRIRAM S, et al. Emergence of locomotion behaviors in rich environments[EB/OL]. New York, USA: Cornell University, 2017[2019-08-22]. https: //arxiv. org/abs/1707.02286.

[21] DILOKTHANAKUL N, KAPLANIS C, PAWLOWSKI N, et al. Feature control as intrinsic motivation for hierarchical reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3409-3418.

闫 冬
1993—，男，硕士，工程师
主要从事电力机器人智能控制和深度强化学习等人工智能技术研究工作
E-mail: yandong@epri.sgcc.com.cn

YAN Dong

陈 盛(通信作者)
1989—，男，博士生，工程师
主要从事电力机器人智能控制和深度强化学习等人工智能技术研究工作
E-mail: chensheng@epri.sgcc.com.cn

CHEN Sheng
Ph.D. candidate
Corresponding author

收稿日期 2019-08-25 修回日期 2019-12-21 编辑 余洋洋

附录 Appendix

1　带电作业机械臂模型及运动学分析

考虑带电作业末端执行器的重量基本大于 5 kg，本文以 UR10 机械臂为研究对象，进一步开展其抽象建模和运动学分析，主要研究 UR10 在笛卡尔坐标系下的 l_1、l_2、l_3 这 3 段臂的运动状态和控制策略，具体涉及对其转角控制和位姿评价。

1.1　笛卡尔坐标系

在笛卡尔坐标系中，可用(x, y, z)坐标表示 3 维空间中任意位置信息，即可通过以$\{o_1, o_2, o_3, o_4\}$原点构建基本方向相同的多个笛卡尔空间坐标系，以累加各段增量的方式计算机械臂各个末端点在 3 维空间的首末端位置，如图 A1 所示。

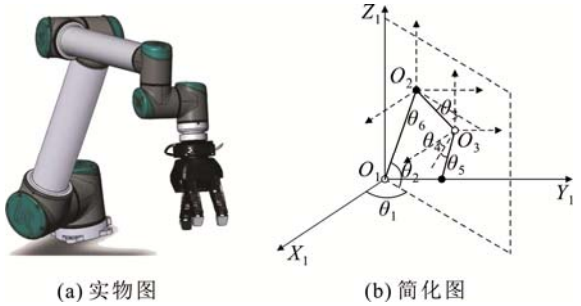


图 A1　UR10 实物图及其在笛卡尔坐标系中的简化模型

Fig.A1　UR10 model and simplified structure in Cartesian coordinates

1.2　运动自由度分析

本文使用的机械臂运动学分析主要为正向运算问题。正向运算问题用于在给定 3 段臂的臂长 l_1、l_2、l_3 和转角 θ_1、θ_2、θ_3、θ_4、θ_5、θ_6 信息的情况下，求解 3 维空间中的机械臂末端坐标。以 z 坐标轴为转轴，第 1 段臂端点为原点，构建机械臂模型，定义 6 个转角为 6 个自由度。θ_1 和 θ_2 控制第 1 段臂，其中 θ_1 表示 l_1 绕 z 轴旋转的角度，θ_2 表示 l_1 与 x–y 平面的角度。l_1、l_2 和 z 轴保持同平面。θ_3 控制 l_2，表示其与 l_1 的夹角。为便于计算，用第 2 段臂与 o_2 坐标系 x–y 平面夹角表示。θ_4—θ_6 控制 l_3，其中 θ_4、θ_5 的作用与 θ_1、θ_2 相似，使 l_3 在 o_3 坐标系内以任意角度运动，θ_6 控制 l_3 围绕自身臂轴旋转，不影响末端点坐标。机械臂各节点坐标及运算为

$$\begin{cases} \Delta l_1 = (l_1\cos\theta_2\cos\theta_1, l_1\cos\theta_2\sin\theta_1, l_1\sin\theta_2) \\ \Delta l_2 = (l_2\cos\theta_3\cos\theta_1, l_2\cos\theta_3\sin\theta_1, l_2\sin\theta_3) \\ \Delta l_3 = (l_3\cos\theta_5\cos\theta_4, l_3\cos\theta_5\sin\theta_4, l_3\sin\theta_5) \\ o_2 = o_1 + \Delta l_1 \\ o_3 = o_2 + \Delta l_2 \\ o_4 = o_3 + \Delta l_3 \end{cases} \tag{1}$$