文章编号:1671-4598(2025)07-0313-08 DOI:10.16526/j.cnki.11-4762/tp.2025.07.038 中图分类号:TP181 文献标识码:A

基于深度强化学习的堤坝巡检机器人避障方法

张鹏程[1]，胡宁宁[1]，王 宵[1]，左登勇[2]，袁冬莉[2]
(1. 上海勘测设计研究院有限公司 勘察检测研究院，上海 200434；
2. 西北工业大学 自动化学院，西安 710129)

摘要：为解决复杂动态环境下的堤坝巡检机器人的自主避障问题，提出一种基于改进深度确定性策略梯度算法的巡检路径避障方法；为提高算法的收敛速度和稳定性，引入优先经验回放机制处理训练样本，并加入随机噪声提高算法对环境的探索能力；为解决算法对环境缺乏先验知识，易于陷入局部迭代的问题，引入了人工势场法提高算法初始阶段的学习效率及快速收敛性；完成了改进算法的状态动作空间、奖励函数和避障训练流程设计；仿真结果表明，改进算法收敛速度快、规划的路径短，使巡检机器人具有更优越的避障能力；经实验验证算法实现了复杂环境下的自主动态避障。

关键词：巡检机器人；深度强化学习；动态避障；堤坝巡检；改进 DDPG 算法

Obstacle Avoidance Method for Dam Inspection Robots Based on Deep Reinforcement Learning

ZHANG Pengcheng[1], HU Ningning[1], WANG Xiao[1], ZUO Dengyong[2], YUAN Dongli[2]
(1. Investigation & Testing Research Institute, Shanghai Investigation, Design & Research Institute Co., Ltd., Shanghai 200434, China;
2. School of Automation, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: In order to solve autonomous obstacle avoidance of dam inspection robots in complex dynamic environments, an obstacle avoidance method based on improved deep deterministic policy gradient (I-DDPG) algorithm is proposed. To improve the convergence speed and stability of the algorithm, the prioritized experience replay mechanism is introduced to process training samples, and also random noise is added to improve the exploration ability of the algorithm on environments. To solve the problems of the algorithm lacking prior knowledge of the environment and being prone to local iteration, the gravitational field function method is introduced to improve the fast convergence and learning efficiency of the algorithm in the initial stage, and designs the state action space, reward function and obstacle avoidance training process of the improved algorithm. Simulation results show that the improved algorithm has the advantages of rapid convergence speed, short planing path, and better obstacle avoidance performance for inspection robots. Practical application verifies autonomous dynamic obstacle avoidance of the algorithm in complex environments.

Keywords: inspection robot; deep reinforcement learning; dynamic obstacle avoidance; dam inspection; I-DDPG algorithm

0 引言

堤坝巡检机器人因其自动化程度高等特点逐渐成为堤坝土工膜渗漏自主检测领域[1]的专用机器人[2]。在实际工程应用中，由于堤坝场区布局柔性多变、存在行人闯入等问题，巡检机器人往往面临的是一个复杂动态的环境，要求其必须具备复杂环境下的自适应能力，能够不断地感知信息做出合理的动态避障决策[3]。

传统路径规划算法如 D* 算法[4]、蚁群算法[5]等实现原理简单且已有改进较为成熟，但通常需依托于相对

收稿日期:2024-06-05； **修回日期**:2024-07-17。
基金项目:国家自然科学基金(61473229)；航空科学基金(20181353013)；教育部中国高校产学研联合基金(2021ZYA03006)。
作者简介:张鹏程(1980-)，男，硕士，高级工程师。
通讯作者:袁冬莉(1966-)，女，副教授。
引用格式:张鹏程，胡宁宁，王 宵，等. 基于深度强化学习的堤坝巡检机器人避障方法[J]. 计算机测量与控制，2025，33(7)：313-320.

完整的环境信息作为先验知识[6-7]，建立环境模型，对环境的自适应能力差，普遍存在收敛速度不佳、易于陷入局部极值等问题，难以在未知的动态环境中不断感知信息实现移动机器人的实时行为决策[8-10]。近年来，随着 Deep Mind 团队研制出的围棋博弈系统 Alpha Go 以较大优势战胜人类顶尖围棋高手，国内外学者开始聚焦于使用具有自主学习能力的强化学习方法，实现移动机器人在未知动态环境下的自主规避[11-13]。文献［14］将深度 Q-Learning 算法与经验回放和启发式知识相结合减小了移动机器人动作选取的随机性，提高了神经网络的训练效率。文献［15］结合改进人工势场法提出改进 Q-Learning 算法，机器人能够在缺乏先验环境信息的情况下跳出局部最优解。Q-Learning 算法固然能在简单环境下取得一定成效，但在复杂环境中将产生高维的状态和动作空间使运算量急剧增加[16-17]。

2013 年，DeepMind 团队创新性地将神经网络替代 Q 表融合进 Q-Learning 算法中形成深度 Q 网络（Deep Q-Network，DQN），后于 2015 年提出改进版本[18]，自此深度强化学习算法受到广泛关注，通过将深度学习的感知能力和强化学习的决策能力相结合实现端到端的控制[19-21]。文献［22］通过引入长短期记忆网络保存样本数据并设置经验回放池阈值，提高了算法的快速收敛性，但离散的动作空间仍难以模拟机器人实际运动情况。在连续控制领域，经典的深度确定性策略梯度（DDPG，deep deterministic policy gradient）算法[23]借鉴了 DQN 的经验回放及目标网络的优势，基于 Actor-Critic 架构，成为连续动作空间下实现移动机器人复杂动态环境下的自主避障研究热点[24]。文献［25］针对采用深度强化学习算法在动态避障过程中存在的规划轨迹长等问题，提出在奖励函数设计中引入距离梯度和角度梯度引导机器人朝终点移动，加速算法收敛。文献［26］融合人工势场法优化 DDPG 算法的动作选择策略，有效缩减了规划的路径长度。然而在实际应用中，复杂动态环境下的基于深度强化学习的巡检机器人避障问题普遍存在着奖励稀疏、探索周期长以及收敛稳定性不佳[27]等问题。

为解决复杂动态环境下的巡检机器人自主避障问题，提出一种人工势场法引导的改进 DDPG（I-DDPG，improved DDPG）算法。为提高样本利用率进而提升算法的快速收敛性和稳定性，引入优先经验回放机制（PER，prioritized experience replay）处理训练样本，并加入随机噪声增加对环境的探索能力；由于 DDPG 算法对环境缺乏先验知识，训练初期随机选取动作导致训练时间较长且容易陷入局部迭代的问题，引入人工势场法的引力场函数提高在训练初期的学习效率及算法的快速收敛性。仿真结果表明，I-DDPG 算法能够实现巡检机器人在复杂环境下的自主规避动态障碍物，在训练前期有着更快的收敛速度和更优越的稳定性，从而验证算法的有效性。经实际应用实现了复杂环境下的自主动态避障。

1 DDPG 算法及改进 DDPG 算法

1.1 DDPG 算法

DDPG 算法结合了异策略的 Actor-Critic 框架和深度学习技术。Actor 网络用于与环境进行交互并根据当前状态输出智能体的策略参数以生成相应的动作。Critic 网络则根据 Actor 网络生成的动作和当前状态计算出该动作的 Q 值，并通过反向传播更新 Actor 网络的参数，以使得生成的动作最大化该状态下的 Q 值。同时，Critic 网络通过最小化预测的 Q 值和实际的 Q 值之间的误差来更新自身的参数 w，从而提高对状态价值的估计能力。Actor-Critic 算法的近似策略梯度如式（1）和式（2）所示：

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[\nabla_\theta \log\pi_\theta(s,a) Q_w(s,a)] \tag{1}$$

$$\Delta\theta = \alpha\nabla_\theta \log\pi_\theta(s,a) Q_w(s,a) \tag{2}$$

DDPG 算法在设计上考虑到了单个神经网络的训练不稳定的问题，汲取了 DQN 算法延时更新目标网络的方法，将 Actor 网络和 Critic 网络在保持神经网络结构不变的情况下，通过设置不同的网络参数分别拆分为在线网络和目标网络两个子网络。在线网络在每个时间步都以最新的网络参数进行更新实现实时学习。目标网络的参数则是定期从在线网络中复制得到，并且这些参数不会被更新，仅用于计算 TD 目标值，在线网络与目标网络参数的不同切断了两者之间的相关性。因此，通过周期性地更新目标网络的参数，可以使得目标值更加稳定，从而提高训练的效率和稳定性。

对于 Critic 网络的更新，利用目标 Actor 网络计算出状态 s_{t+1} 下动作 $a' = \mu'(s_{t+1} \mid \theta^{\mu'})$，设置目标网络计算 TD 目标：

$$y_i = r_i + \gamma Q'[s_{t+1}, \mu'(s_{t+1} \mid \theta^{\mu'}) \mid \theta^{Q}] \tag{3}$$

在网络的更新中，采用梯度下降法以学习率 l 最小化评估误差 L_c：

$$L_c = \frac{1}{M}\sum_i^M [y_i - Q(s_i, a_i \mid \theta^Q)]^2 \tag{4}$$

对于 Actor 网络的更新，其损失函数可以通过计算确定策略梯度：

$$\nabla_{\theta^\mu} J_\beta(\mu) = \frac{1}{M}\sum_i^M [\nabla_a Q(s,a \mid \theta^Q)\mid_{s=s_i, a=\mu(s_i)} \cdot \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\mid_{s=s_i}] \tag{5}$$

DDPG 算法在训练过程采用软更新的方式，将在线网络的参数与目标网络的参数进行插值平均，以此来更新在线网络的参数。DDPG 算法的框架如图 1 所示。

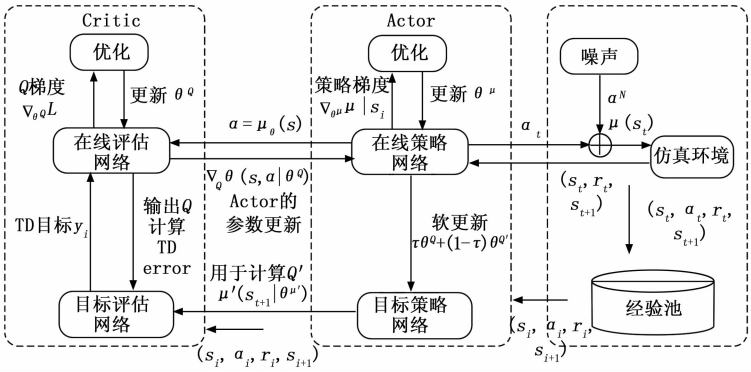

图 1　DDPG 算法框架

1.2　改进 DDPG 算法

1.2.1　优先经验回放

优先经验回放机制主要针对经验的抽取方式提出改进，通过设置经验优先级提高神经网络的更新效率，并在 Q 网络上实现了 TD 误差的绝对值最大化。这种机制有效打破了采集样本之间的相关性，并在奖励稀疏的情况下提高了学习效率。PER 通过使用 TD-error 衡量历史经验样本的价值并相应赋予其优先级，TD-error 表示当前动作的输出价值与目标价值的差值，TD-error 越大意味着历史经验更需要被学习。PER 机制的关键在于采样时优先抽取优先级高的样本数据来弥补随机采样的缺陷。同时为避免网络的过拟合，PER 采用概率方式抽取样本，并确保即便是 TD-error 为 0 的样本也具有一定的抽取概率。

PER 机制赋予每个样本 i 一个优先级 p_i。在从经验池中抽取样本时优先抽取最具有价值的历史经验，并将经验样本根据经验数据的优先级存入二叉树中。

第 i 个样本被抽取的概率为：

$$P(i)=\frac{p_i}{\sum_k p_i} \tag{6}$$

优先级 p_i 表示为：

$$p_i=|\delta_i|+\varepsilon_{PER} \tag{7}$$

$$\delta_i=r_{t+1}+\gamma\max_{a_{t+1}}Q_w(s_{t+1},a_{t+1})-Q(s_t,a_t) \tag{8}$$

其中：δ_i 为 TD-error，ε_{PER} 是很小的正数，确保低优先级的经验被抽取的概率大于 0。

经验回放机制导致的采样误差，DDPG 的估计更加偏向 TD-error 比较大的样本，无法对所有的状态－动作价值函数做正确的估计。为了解决此问题，在优先经验回放算法中引入了重要性采样权重 w_i 为：

$$w_i=\left(\frac{1}{N}\cdot\frac{1}{P(i)}\right)^\beta \tag{9}$$

式中，N 表示经验池的大小；β 为超参数，用来决定以多大的程度抵消 PER 对收敛结果的影响。$\beta\to 1$ 表示权重会消除前面采样概率对于模型的偏置，但也可能减慢模型的收敛速度，实验中一般取 $\beta\in(0,1)$。

1.2.2　人工势场法引导

对于深度强化学习的巡检机器人自主动态避障问题，只有在智能体与障碍物发生碰撞或者到达目标位置的情况下才能获得即时的奖励。奖励函数的稀疏性导致智能体在训练初期的学习效率低、迭代次数多，尤其是在未知训练环境中会产生过多的无效搜索严重浪费计算资源。因此，根据目标点和障碍物点的位置信息构造人工势场法引导的 DDPG 算法。此时，势场中各个状态的势场值 $U(s_i)$ 在状态 s_i 时采取最优动作获得最大累积收益 $V(s_i)$ 为：

$$V(s_i)=|U(s_i)| \tag{10}$$

研究基于未知环境，障碍物位置无法确定，为了在训练初始阶段就能够有效引导智能体朝向目标点运动，因而在 Q 函数初始化的过程中只考虑引入引力场函数 $U_a(q)$，表示为：

$$U_a(q)=\frac{1}{2}k_a\rho_g^2(s) \tag{11}$$

其中：k_a 为引力场系数，$\rho_g(s)$ 为巡检机器人的位置与目标点之间的距离。

通过增加接近目标位置的状态值提高智能体在初始阶段的学习效率，加快算法收敛：

$$Q(s_i,a)=r+\gamma V(s_{i+1}) \tag{12}$$

1.2.3　状态及动作空间设计[10]

在 I-DDPG 算法的规避训练策略中，巡检机器人朝目标位置移动并同时规避动态障碍物，在设计状态空间时着重考虑巡检机器人与动态障碍物及目标点的位置关系，因此设计的状态空间如下：

$$S=\begin{cases}(x,y)k\\ \theta/2\pi\\ d_{\text{obs}}/k\\ [(x-x_{\text{obs}}),(y-y_{\text{obs}})]k\\ d_{tar}/k\\ [(x-x_{tar}),(y-y_{tar})]/k\end{cases} \tag{13}$$

式中，(x,y) 和 θ 分别表示巡检机器人的当前位置与运动朝向，k 为标准化系数；d_{obs} 和 d_{tar} 表示移动机器人与最近障碍物和目标点的距离；$[(x-x_{\text{obs}}),\ y-y_{\text{obs}})]$ 和 $[(x-x_{\text{tar}}),\ (y-y_{\text{tar}})]$ 分别表示巡检机器人与最近障碍物和目标点的距离信息。

巡检机器人的动作空间需要根据实际运动过程中的线速度和角速度进行设定。由于设置的巡检机器人观测范围为前方，所以将其动作空间设定为前进控制和转向控制。

1.2.4　奖励函数设计

奖励函数设置为：

$$R = r_1 + r_2 \tag{14}$$

当巡检机器人抵达终点位置时获得固定的奖励，而和障碍物碰撞时得到惩罚。当巡检机器人尚未抵达目标或碰到障碍物时，奖励值包含两部分：一是巡检机器人与最近障碍物的距离信息的奖励值；二是巡检机器人和目标点距离信息的奖励值。两部分奖励值之和是巡检机器人执行每一次动作后获得的最终的奖励值。

巡检机器人与最近障碍物的距离信息的奖励值 r_1 表示为：

$$r_1 = \begin{cases} -1 & d_{obs} \leqslant r_{obs} \\ -0.3 + \dfrac{d_{obs} - d_{sec}}{d_{sec}} * \alpha, & r_{obs} < d_{obs} \leqslant d_{sec} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

式中，r_{obs} 为设定的障碍物半径，d_{sec} 为障碍物威胁区域。固定奖励 $r_1^{n_1} = -1$ 是碰到障碍物后得到的较大惩罚，$r_1^{n_2} = -0.3$ 是进入障碍物威胁区域后受到的惩罚；设定威胁区域梯度 $\alpha = 1$，考虑威胁区域内距离障碍物越近威胁越大，设置渐进惩罚，随着威胁区域的不断深入，惩罚也随之加大。

巡检机器人与目标点距离信息的奖励值 r_2 可表示为：

$$r_2 = \begin{cases} 3, d_{tar} \leqslant d_0 \\ -\dfrac{d_{tar}}{d_2} * \beta, d_{tar} > d_0 \end{cases} \tag{16}$$

式中，d_2 表示从传感器探测到最近障碍物并开始做出决策的起始点到目标路径点的距离，d_0 为考虑巡检机器人实际情况设立的目标区域范围。设定 $r_2^n = -\dfrac{d_{tar}}{d_2}$ 是到达目标位置前执行每一步决策获得的渐进惩罚，用于引导巡检机器人尽快到达目标位置完成规避决策，设定目标位置梯度 $\beta = 1$；当完成规避时给予固定奖励 $r_2^{p_1} = 3$。

采用 I-DDPG 算法对巡检机器人的避障决策的训练流程如图 2 所示。

1）随机初始化 Actor 网络和 Critic 网络，初始化奖励折扣因子 γ、网络更新率 τ 以及学习率 l；

2）复制 Actor 网络和 Critic 网络的参数，初始化目标网络 $Q'(s,a \mid \theta^{Q'})$ 和 $\mu'(s \mid \theta^{\mu'})$，$\theta^{Q'} \leftarrow \theta^{Q}$，$\theta^{\mu'} \leftarrow \theta^{\mu}$；

3）设置训练回合数 E，初始化样本容量为 P 的经验池，并设置抽取的样本数量 M（$M \leqslant P$）；

4）每个训练回合 E：

（1）初始化探索噪声 a^N，且随着训练过程的进行每回合以衰减率 υ 减小噪声大小；

（2）机器人在预定航迹上探测到障碍物，开始规避，设定当前时刻 $t = 0$，接受传感器反馈的障碍物的位置信息、以及机器人的自身状态信息；

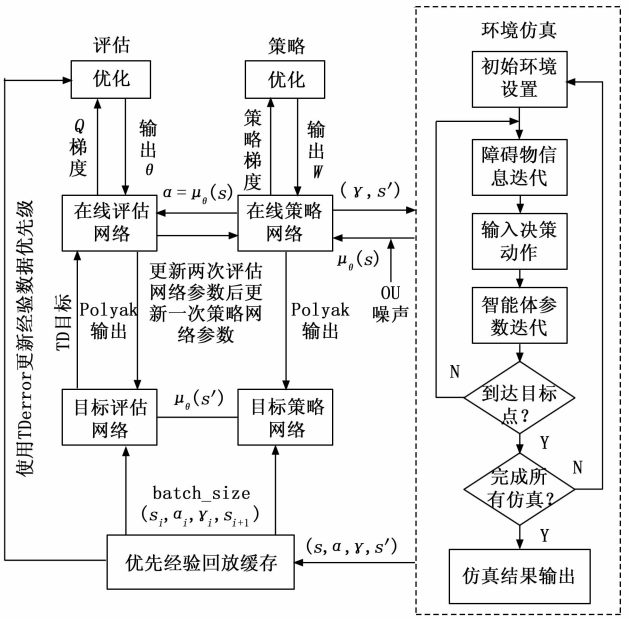

图 2 规避决策训练流程

（3）由 Critic 网络 $\mu(s \mid \theta^{\mu'})$ 生成动作指令 $a_t = \mu(s_t \mid \theta^{\mu}) + a^N$ 并执行，更新环境状态，下一个状态 s_{t+1}，得到的奖励 r，并判断是否完成避障 Done；

（4）将 $(s, a_t, r, s_{t+1}, \text{Done})$ 存入经验回放池中，并设置优先级 p_i；

（5）根据优先级采样经验样本并计算相应的重要性采样权重 w_j，计算 TD 误差 δ_j 并以此更新经验样本的优先级 p_j；

（6）计算 TD 目标 y_i；

（7）采用梯度下降方法以学习率 l 最小化评估误差 L_c，从而更新 Critic 网络；

（8）采用梯度上升方法以学习率 l 最大化价值；

（9）采用软更新方式更新目标网络；

（10）时间推进 $t = t + 1$，重复步骤 3）至 9）直至完成规避决策，结束当前回合。

5）训练结束。

由于 DDPG 算法学习的策略是确定性策略，在进行探索时为避免陷入局部最优，可以自定义添加噪声，从而增加探索其他动作的可能性。

在每个规避决策训练回合中，策略网络中将状态 s 作为输入，在输出控制指令 a_t 时引入噪声 a^N，经底层控制策略约束优化后由智能体执行，获得环境反馈 r 并迁移至下一状态 s_{t+1}。同时，判断智能体是否完成对障碍物的规避并抵达终点位置。执行完成的每一步，都将其作为经验 $(s, a_t, r, s_{t+1}, \text{Done})$ 储存在经验回放池中，通过优先经验回放机制进行采样，以此更新神经网络参数。

2　仿真与实验

算法中设置的相关参数如表 1 所示。

表 1　算法参数设置

参数	参数值
γ（奖励折扣因子）	0.99
τ（网络更新率）	0.005
l（学习率）	0.001
a^N（初始化噪声）	0.3
υ（衰减率）	0.999 5
P（经验池容量）	100 000
M（单次抽样数量）	128
E（训练回合数）	500

仿真环境基于 Python3.8 编程实现，结合 Pytorch1.7 开发包设计 I-DDPG 算法用于巡检机器人避障决策训练，硬件配置主要包括 RTX2060 显卡、16 G 内存和 AMD Ryzen7-4800 处理器。

根据所设计的算法流程，在搭建的仿真环境中对巡检机器人进行避障决策训练，不同环境下每个训练回合的奖励值及运行步长如图 3 所示。

在训练初始阶段，由于巡检机器人刚开始与环境交互，漫无目的的探索导致碰撞障碍物或者远离目标点，获得较大的惩罚。通过不断地学习在训练后期奖励值及运行步长逐渐呈现平稳趋势，表明算法已经收敛。后期的训练过程存在小幅度波动是由动作探索策略造成的，算法仍以小探索度 $a^N = 0.1$ 进行动作探索。

图 4 展示了巡检机器人避障训练的效果。以图左侧黑色圆形表示环境中的静态障碍物，图右侧圆形为环境中的动态障碍物。可以看出，巡检机器人在未知环境中能够自主规避存在的动态障碍物，避障完成后恢复原有航迹运行从而到达目标点，从而进一步验证了改进算法的有效性。

仿真验证改进算法的可行性后，为进一步测试算法性能，分别从奖励曲线、运行步长、规划时长和路径长度 4 个方面进行对比分析，如图 5 所示。

如图 5（a）所示，在训练初始阶段，由于 DDPG 算法无法区分经验样本的重要性，只能不断探索并尝试

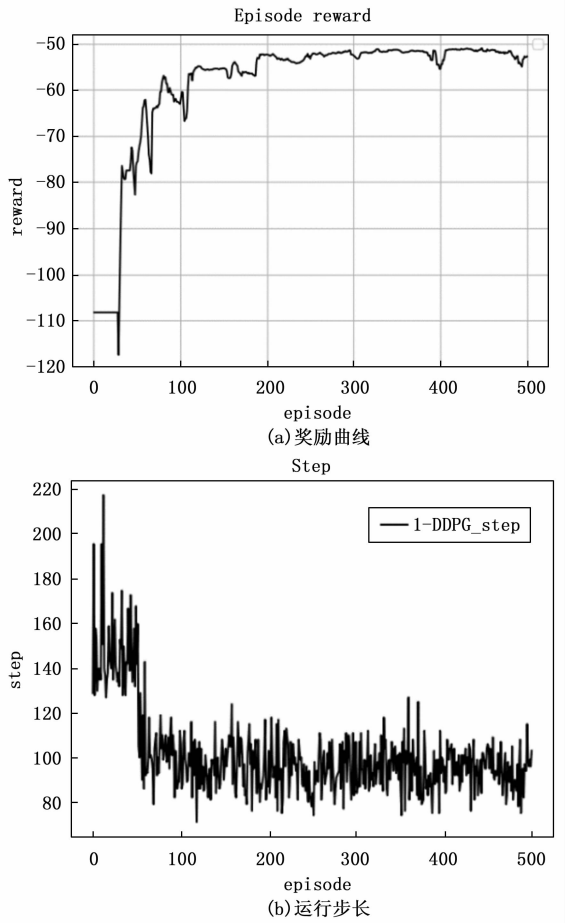

图 3　改进算法奖励曲线和运行步长

学习，奖励曲线的波动较大。而 I-DDPG 算法采用了优先经验回放机制，优先学习重要性采样权重大的经验，相比于 DDPG 算法降低了曲线的波动，且更早地开始加速收敛。因此，在相同训练条件下的场景中，I-DDPG 算法收敛更快且具有较好的稳定性。根据算法的运行步长及规划的路径长度，相较于 DDPG 算法，在训练前期 I-DDPG 算法曲线波动较小，更早且更快地加速收敛，进一步验证了算法的稳定性和快速收敛性。

I-DDPG 和 DDPG 算法的平均训练结果对比如表 2 所示，结合图 5（c）可知，在同样的目标终点下 I-DDPG 算法的平均规划时间稍长，原因在于前者在采样时样本需要在二叉树中搜索，且优先经验回放采用 TD-

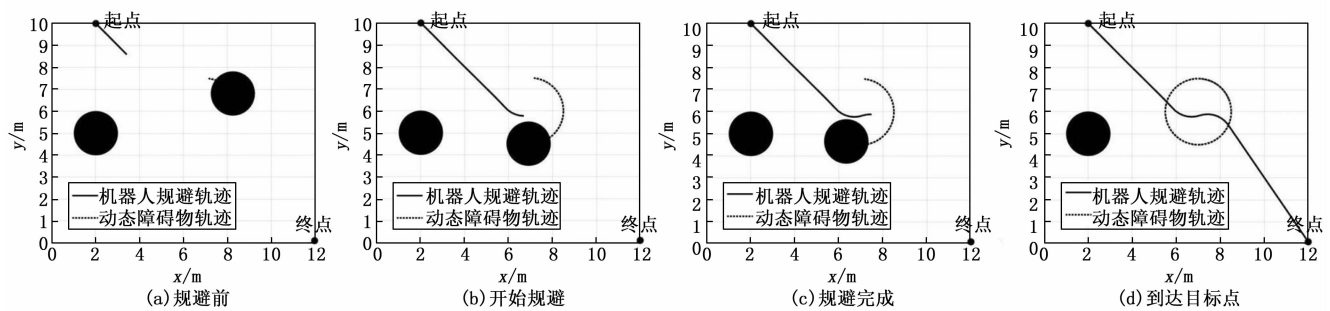

图 4　避障决策训练效果

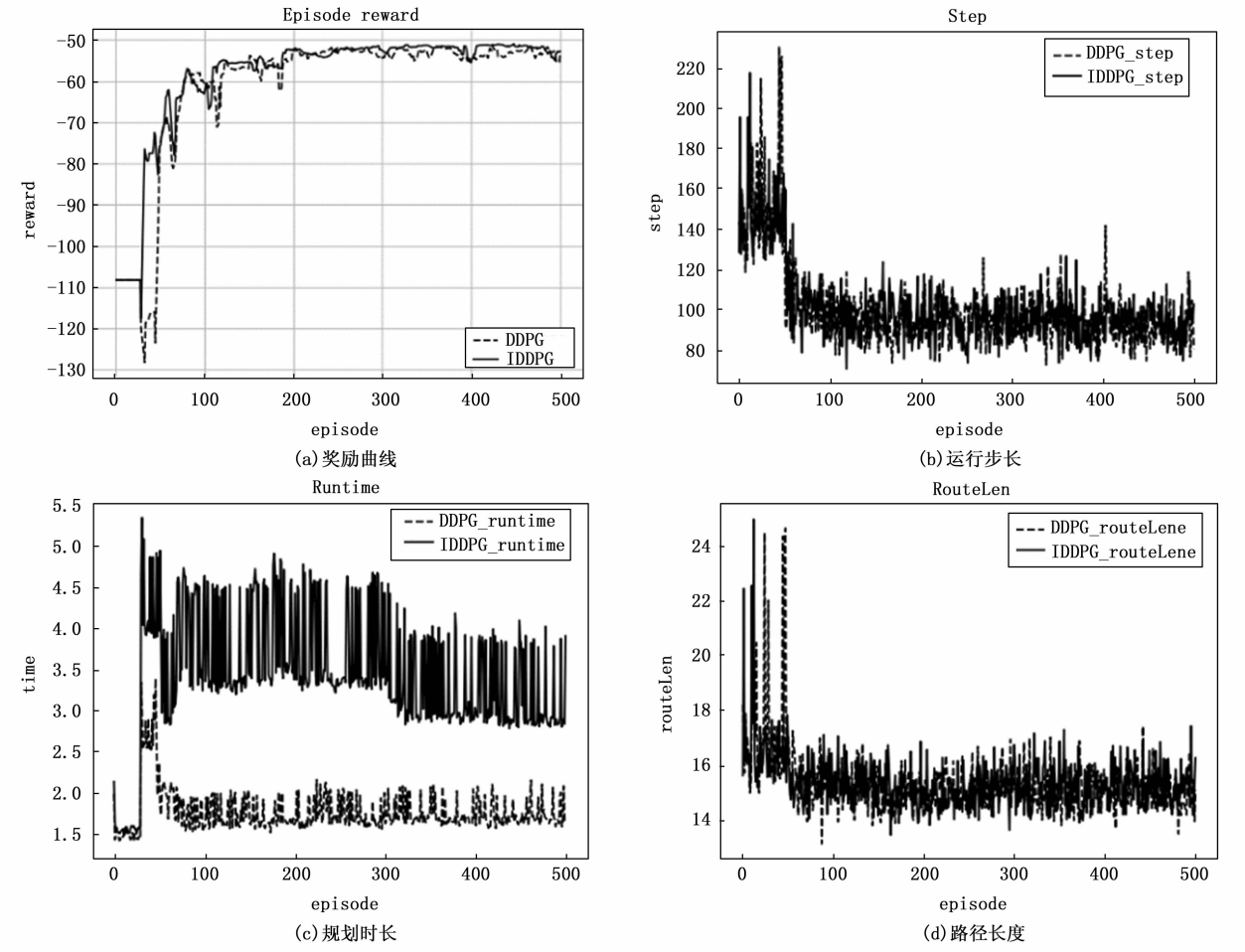

图 5 算法性能对比

error 衡量经验样本的价值并相应赋予其优先级，提高了时间成本。但极大提高了样本数据的利用率，在稀疏奖励的场景下提高了算法的收敛速度和算法的稳定性。

表 2 算法性能对比

	运行步长	规划时间	路径长度
DDPG	101	1.8 s	15.3 m
I-DDPG	100	3.4 s	15.1 m

为充分验证本文堤坝巡检机器人路径规划算法的实用性和工程落地效果，基于搭建的巡检机器人硬件平台并面向 ROS Melodic 设计自主规划软件架构，对本文提出的改进 DDPG 算法进行实验验证。实验场地选取两侧地势较高且存在一定坡度的小道模拟堤坝场景，位于空旷地带便于接收 GPS 差分信号。

将区域坐标信息发送至机器人端并在 Rviz 中生成语义地图，设置巡检机器人的检测半径为 0.3 m，通过激光雷达信息获得局部环境信息。实验中采用的二维激光雷达仅能接收到与自身同处一个水平面的障碍物信息。为避免巡检机器人与环境中的动态障碍物发生碰撞，需要确保其形状高度高于巡检机器人上激光雷达的安装高度以便于被激光雷达探测。实验中假定的动态障碍物，通过在小推车上摆放纸箱并轻推小推车使其运动实现模拟。初始状态时设定动态障碍物位于机器人相向运动的正前方约 4 m 处，给定目标位置设定在机器人需绕行区域外的 5 m 处。

图 6 展示了动态障碍物下的避障实际效果，从图中可以看出机器人在距离障碍物约 1.5 m 处开始做出避障决策，左拐以避开迎面而来的障碍物，规避完成后右拐直行到达目标点。实际路径效果如图 7 所示。

表 3 记录了动态障碍物规避实验中所提算法与原算法的数据对比结果。在运动时长及路径长度方面，本文所提算法相较于原算法均有明显提升。在避障成功率方面，相较于原算法的 50%，本文算法有着近 90%的避障成功率。避障失败主要是由于实验中的动态障碍物是通过对小推车施加推力使其运动，难以确保多次实验中小推车的运动速度及其他条件都能保持一致。经过多次实验的结果表明，在动态障碍物的运动速度低于 1 m/s 的情况下，巡检机器人能够及时做出规避决策并成功避开障碍物，验证了所提算法在相同条件下具有更稳定的避障成功率。

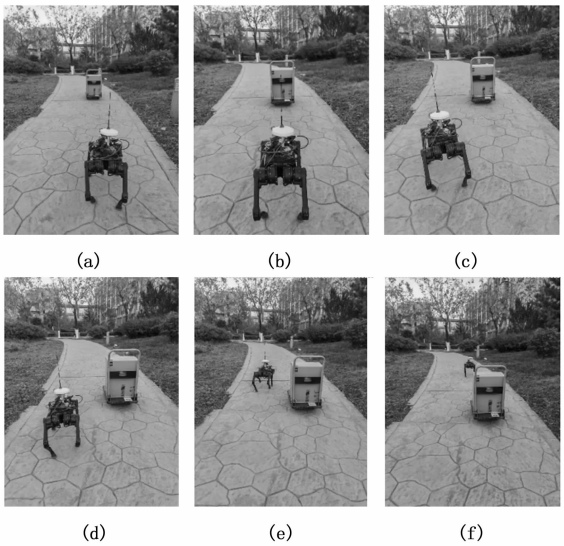

图 6　动态规避效果

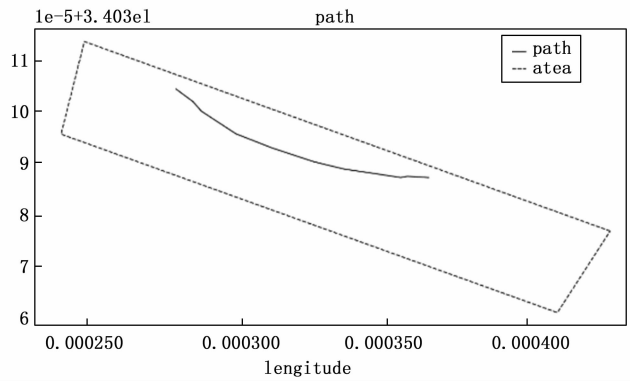

图 7　实际路径效果图

表 3　实验结果对比

	运行时长	轨迹长度	成功率
原算法	7.9 s	7.0 m	50%
所提算法	9.4 s	7.8 m	90%

改进算法在具有较高避障成功率的前提下，相较于原算法在运行时长及实际轨迹长度等方面具有更为优越的评价指标，进一步验证所提算法的实用性与有效性。

3　结束语

针对在复杂环境下采用 DDPG 算法实现堤坝巡检机器人自主规避动态障碍物时，主要存在的训练时间较长、样本利用率低且在训练初始阶段易于陷入局部迭代的问题，提出相应的改进措施。为提高样本利用率从而提升算法的快速收敛性和稳定性，引入 PER 机制处理训练样本；并加入随机噪声增加算法在环境中的探索能力；针对 DDPG 算法对环境没有先验知识，训练初期随机选取动作导致训练时间较长且容易陷入局部迭代，引入人工势场法的引力场函数以提高算法初始阶段的学习效率，提升算法的快速收敛性。完成了状态动作空间、奖励函数和规划决策训练流程设计，经过仿真验证，改进算法收敛速度快、规划的路径短，使巡检机器人具有更优越的避障能力，从而验证算法的有效性。经实际应用实现了复杂环境下的自主动态避障，从而验证算法的可行性。在巡检策略方面，仅针对单一智能体的路径规划问题进行了研究，在未来的研究中可以进一步探讨多智能体之间的协同路径规划。

参考文献：

[1] GILSON-BECK A. Controlling leakage through installed geomembranes using electrical leak location [J]. Geotextiles and Geomembranes, 2019, 47 (5): 697 - 710.

[2] ANKUSH VERMA, AYUSH KAIWART, NIKHIL DHAR DUBEY, et al. A review on various types of in-pipe inspection robot [J]. Materials Today, 2022, 50 (5): 1425 - 1434.

[3] 高春艳，陶　渊，吕晓玲，等，非结构化环境下巡检机器人环境感知技术研究综述 [J]. 传感器与微系统，2023，42 (4)：10 - 13.

[4] STENTZ A. Optimal and efficient path planning for partially-known environments proceedings of the proceedings of the 1994 IEEE International Conference on Robotics and Automation, F 8 - 13 May 1994 [C] //1994, 1 (4): 3310.

[5] 张　恒，何　丽，袁　亮，等，基于改进双层蚁群算法的移动机器人路径规划 [J]. 控制与决策，2022，37 (2)：303 - 13.

[6] 翟　丽，张雪莹，张　闲，等，基于势场法的无人车局部动态避障路径规划算法 [J]. 北京理工大学学报，2022，42 (7)：696 - 705.

[7] SHAHER ALSHAMMREI, SAHBI BOUBAKER, LIOUA KOLSI. Improved dijkstra algorithm for mobile robot path planning and obstacle avoidance [J]. Computers Materials&Continua, 2022, 72 (3): 4433 - 4453.

[8] ZHOU C, HUANG B, FRANTI P. A review of motion planning algorithms for intelligent robots [J]. Journal of Intelligent Manufacturing, 2022, 33 (2): 387 - 424.

[9] ALSHAMMREI S, BOUBAKER S, KOLSI L. Improved dijkstra algorithm for mobile robot path planning and obstacle avoidance [J]. CMC-Comput Mat Contin, 2022, 72 (3): 5939 - 5954.

[10] WANG Y, BAI P, LIANG X, et al. Reconnaissance mission conducted by UAV swarms based on distributed PSO path planning algorithms [J]. IEEE Access, 2019, 7: 105086 - 105099.

[11] GAO X, YAN L, LI Z, et al. Improved deep deterministic policy gradient for dynamic obstacle avoidance of mobile robot [J]. IEEE Transactions on Systems Man Cybernetics-Systems, 2023, 53 (6): 1 - 8.

[12] 杨秀霞，高恒杰，刘　伟，等，RVO-DDPG 算法在多

UAV 集结航路规划的应用 [J]. 计算机工程与应用，2023，59 (1)：308－316.
[13] ZHU P M，DAI W，YAO W J，et al. Multi-robot flocking control based on deep reinforcement learning [J]. IEEE Access，2020，8：150397－150406.
[14] JIANG L，HUANG H，DING Z. Path planning for intelligent robots based on deep Q-learning with experience replay and heuristic knowledge [J]. IEEE-Caa Journal of Automatica Sinica，2020，7 (4)：1179－8.
[15] YAO Q，ZHENG Z，QI L，et al. Path planning method with improved artificial potential field-a reinforcement learning perspective [J]. IEEE Access，2020，8：135513－135523.
[16] 齐昊罡. 基于深度强化学习的移动机器人路径规划研究 [D]. 吉林：吉林大学，2021.
[17] 田晓航，霍 鑫，周典乐，等，基于蚁群信息素辅助的 Q 学习路径规划算法 [J]. 控制与决策，2023，38 (12)：3345－3353.
[18] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning [J]. Nature，2015，518 (7540)：529－33.
[19] ZINDLER，FRIEDEMANN，LUCCHI，et al. Towards dynamic obstacle avoidance for robot manipulators with deep reinforcement learning [J]. Mechanisms and Machine Science，2022，120：89－96.
[20] WANG X，GU Y，CHENG Y，et al. Approximate policy-based accelerated deep reinforcement learning [J]. IEEE Transactions on Neural Networks and Learning Systems，2020，31 (6)：1820－1830.
[21] 范鑫磊，李 栋，张 尉，等，基于深度强化学习的导弹规避决策训练研究 [J]. 电光与控制，2021，28 (1)：81－85.
[22] 周治国，余思雨，于家宝，等，面向无人艇的 T-DQN 智能避障算法研究 [J]. 自动化学报，2023，49 (8)：1645－1655.
[23] 张柏鑫，杨毅镔，朱华中，等. 基于深度强化学习的移动机器人动态路径规划算法 [J]. 计算机测量与控制，2023，31 (1)：153－159.
[24] 蔡 泽，胡耀光，闻敬谦，等，复杂动态环境下基于深度强化学习的 AGV 避障方法 [J]. 计算机集成制造系统，2023，29 (1)：236－245.
[25] 徐建华，邵康康，王佳惠，等，基于改进强化学习的移动机器人动态避障方法 [J]. 中国惯性技术学报，2023，31 (1)：92－99.
[26] 张 瀚，解明扬，张 民，等，融合 DDPG 算法的移动机器人路径规划研究 [J]. 控制工程，2021，28 (11)：2136－42.
[27] CHU Z，WANG F，LEI T，et al. Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance [J]. IEEE Transactions on Intelligent Vehicles，2022，8 (1)：108－120.

(上接第 303 页)

[8] 漆 琪，杨吉祥，丁 汉. 面向航空发动机复杂机匣深腔测量的机器人位姿优化及关节路径平滑方法 [J]. 机械工程学报，2023，59 (15)：17－27.
[9] 俞青松，徐向荣，刘胤真. 基于 Transformer 的机器人像素级抓取位姿检测 [J]. 工程设计学报，2024，31 (2)：238－247.
[10] 陈 钧，宋 薇，周 洋. 一种多模块神经网络与遗传算法相结合的单目位姿估计方法 [J]. 机器人，2023，45 (2)：187－196.
[11] 李贵虎，高贵军，李军霞，等. 基于扩展卡尔曼滤波的清仓机器人位姿识别方法 [J]. 工矿自动化，2024，50 (5)：99－106.
[12] 李鸣昊，王 辉，伍思欢，等. 考虑执行器最优分配的空间站舱内机器人位姿耦合控制 [J]. 宇航学报，2023，44 (7)：1073－1083.
[13] 丁 江，崔家旭，左启阳，等. 基于无迹卡尔曼滤波算法的喷涂机器人末端位姿补偿系统 [J]. 机械传动，2024，48 (1)：8－13.
[14] 高文斌，杜 涛，罗瑞卿. 一种利用位置误差模型的机器人位姿标定方法 [J]. 机械科学与技术，2023，42 (11)：1852－1859.
[15] 游 刚，李世芸，仇隽挺，等. 基于改进 AMCL 与点云匹配校正的落布机器人定位分析 [J]. 机械设计与研究，2024，40 (1)：56－62.
[16] 梁春苗，姚宁平，姚亚峰，等. 煤矿用钻孔机器人钻臂定位误差补偿研究 [J]. 煤田地质与勘探，2024，52 (3)：153－163.
[17] 邓忠元，刘 冉，曹志强，等. 基于单 UWB 融合里程计的多机器人相对定位方法 [J]. 计算机应用研究，2023，40 (3)：839－844.
[18] 沈 跃，肖鑫桦，刘 慧，等. 果园机器人 LiDAR/IMU 紧耦合实时定位与建图方法 [J]. 农业机械学报，2023，54 (11)：20－28.
[19] 刘超凡，于文涛，曲 枫，等. 基于视觉辅助与自适应粒子滤波的机器人自定位 [J]. 郑州大学学报：理学版，2023，55 (5)：73－80.
[20] 李 港，赖钦伟，蒋 林，等. 基于改进 SIFT 的室内机器人惯性视觉定位系统 [J]. 组合机床与自动化加工技术，2023 (2)：155－159.
[21] 崔玉明，刘送永，吕振礼，等. 基于级联优化和强度特征的地下退化环境机器人自主精准定位 [J]. 仪器仪表学报，2023，44 (12)：208－216.
[22] 刘 铁，段 勇. 融合 CNN 和 Transformer 的机器人室内场景识别 [J]. 电子测量与仪器学报，2023，37 (5)：223－229.