本文引用格式：黄兆军,张彦佳,左晓雯,等.基于监督式 DDPG 算法的小型 ROV 运动控制方法[J].自动化与信息工程,2025,46(3):23-29.

HUANG Zhaojun, ZHANG Yanjia, ZUO Xiaowen, et al. Motion control method for small ROV based on supervised DDPG algorithm[J]. Automation & Information Engineering, 2025,46(3):23-29.

基于监督式 DDPG 算法的小型 ROV 运动控制方法*

黄兆军　张彦佳　左晓雯　陈泽汛

（珠海城市职业技术学院，广东 珠海 519090）

摘要：针对深度确定性策略梯度（DDPG）算法用于无人遥控有缆水下机器人（ROV）运动控制时，存在学习时间长且难以收敛等问题，提出基于监督式 DDPG 算法的小型 ROV 运动控制方法。在 DDPG 算法的初始学习阶段引入监督学习算法，通过专家经验引导，加快神经网络收敛速度，缩短学习时间。仿真试验结果表明，监督式 DDPG 算法比 DDPG 算法的控制效果更好。

关键词：监督式 DDPG；小型 ROV；运动控制；专家经验；强化学习

中图分类号：TP242.3　　**文献标志码：**A　　**文章编号：**1674-2605(2025)03-0004-07

DOI：10.12475/aie.20250304　　开放获取

Motion Control Method for Small ROV Based on Supervised DDPG Algorithm

HUANG Zhaojun　ZHANG Yanjia　ZUO Xiaowen　CHEN Zexun

(Zhuhai City Polytechnic, Zhuhai 519090, China)

Abstract: To address the issues of prolonged learning time and difficulty in convergence when using the Deep Deterministic Policy Gradient (DDPG) algorithm for motion control of remotely operated tethered underwater vehicles (ROVs), this paper proposes a supervised DDPG-based motion control method for small ROVs. During the initial learning phase of the DDPG algorithm, a supervised learning approach is introduced to accelerate neural network convergence and reduce learning time by leveraging expert experience. Simulation results demonstrate that the supervised DDPG algorithm achieves superior control performance compared to the standard DDPG algorithm.

Keywords: supervised DDPG; small ROV; motion control; expert experience; reinforcement learning

0　引言

无人遥控有缆水下机器人（remotely operated vehicle, ROV）因在开发难度、研制周期、资金投入和产业化等方面具有优势，成为水下机器人领域的研究重点，并已广泛应用于海水养殖、海洋勘探、水下搜救和石油开发等领域。运动控制是 ROV 的核心技术之一，包括 PID 控制、模糊控制、滑膜变结构控制和 S 面控制等方法。但这些方法均存在一定的局限性，如 PID 控制在处理非线性复杂耦合系统时面临挑战；模糊控制受限于规则库的完备性和规则结构的合理性，当系统复杂度较高时易出现“规则爆炸”问题；滑膜变结构控制和 S 面控制存在实现困难且易产生抖动等问题[1]，制约了 ROV 的产业化进程。

近年来，随着人工智能技术的快速发展，智能控制算法逐渐应用于 ROV 运动控制领域[2]。其中，深度确定性策略梯度（deep deterministic policy gradi- ent, DDPG）算法作为机器学习的一种深度强化学习算法，无需精确的数学模型，通过智能体与环境的交互即可

* 基金项目：2023 年度广东省普通高校特色创新项目“海水养殖专用低成本小型便携式巡检 ROV 关键技术的研究与实现”（2023GKTSCX088）

实现控制策略的优化，具有环境自适应性，适用于连续、实时决策且不确定性较高的水下环境，成为当前 ROV 运动控制领域的重要研究方向。但 DDPG 算法存在学习时间长、虚实迁移效果差和收敛难等问题，导致其在 ROV 运动控制中的实际应用效果并不理想，目前多数研究仍停留在仿真实验阶段。

为此，本文对 DDPG 算法进行改进，提出基于监督式 DDPG 算法的小型 ROV 运动控制方法，旨在改善算法的收敛性和稳定性。

1 DDPG 算法理论

DDPG 算法是一种为解决连续控制问题而提出的深度强化学习算法[3]。该算法采用 Actor-Critic 架构，结合策略神经网络和价值神经网络，对输入的高维数据进行拟合处理和决策，实现端对端的策略优化和控制，在连续状态空间下，输出一个确定的动作[4]。

DDPG 算法可分为采样、训练、参数更新 3 个流程[5]，如图 1 所示。

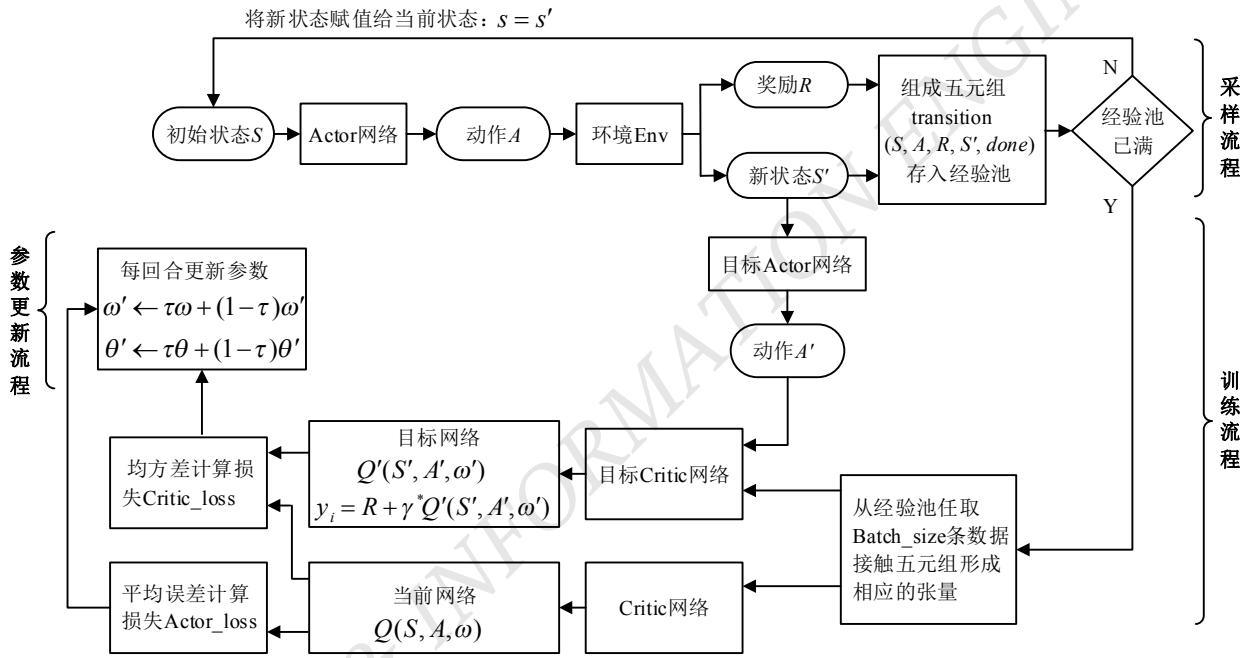

图 1 DDPG 算法流程

采样流程是智能体与环境交互以收集经验数据的过程。Actor 网络根据智能体的初始状态 S 输出一个连续动作 A，并作用于环境 Env；环境变化影响智能体，使其转移到下一个新状态 S'，并同步反馈一个奖励信号 R；将这一完整的交互经验元组（当前状态 S, 动作 A, 奖励 R, 新状态 S', 终止标志 *done*）存储到经验池中，并将新状态 S'返回赋值为新的初始状态，持续循环上述流程，直到经验池填满，进入训练流程。

训练流程独立于智能体与环境的实时交互，利用经验池中存储的历史数据进行离线学习，更新网络参数以优化控制策略。首先，随机从经验池中提取规定数量（Batch_size）的经验元组；然后，Critic 网络计算当前状态 S 和执行动作 A 对应的预测回报值 Q，并基于平均误差计算 Actor_loss 损失；同时，目标 Critic 网络基于新状态 S'和目标动作 A'计算目标回报值 Q'（考虑折扣因子 γ *和奖励 R）；最后，基于 Q 和 Q' 的均方差计算 Critic_loss 损失。

参数更新流程基于训练流程计算的 Actor_loss 损失和 Critic_loss 损失来调整神经网络的参数。Critic 网络通过梯度下降法最小化 Critic_loss 损失更新参数 ω；Actor 网络通过最大化 Critic 网络评估的 Q 值方向更新参数 θ。但为了稳定训练过程，防止剧烈波动，

上述更新参数并非直接复制到相应的网络中，而是采用“软更新”的方式间接进行。即通过指数平滑的方式将 Critic 网络参数和 Actor 网络参数缓慢混合到对应的目标Critic网络参数ω'和目标Actor网络参数θ'中，以确保目标网络参数的变化是渐进且稳定的，从而提升 DDPG 算法学习的稳定性。

DDPG 算法是以累积奖励最大化为目标，使智能体在与环境的交互中不断学习最优策略的一种非监督式机器学习算法。因缺乏监督，该算法策略学习过程和动作训练均是随机的，导致学习时间较长且神经网络难以收敛，造成大量数据浪费，存储开销增加，神经网络的泛化能力降低，实用性较差[6]。

2 监督式 DDPG 算法

为改善上述问题，本文在 DDPG 算法的基础上引入监督学习算法，提出一种监督式 DDPG 算法。该算法通过监督学习算法的专家经验来指导 ROV 的策略学习[7]，使最优策略的探索和学习具有一定的方向性和目的性[8]，从而缩短学习时间，加快神经网络收敛。监督式 DDPG 算法的原理如图 2 所示。

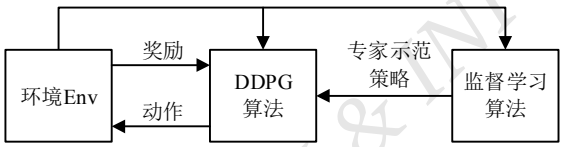

图 2 监督式 DDPG 算法原理图

设 DDPG 算法对 ROV 的选择动作为a_R，则监督式 DDPG 算法对 ROV 的选择动作为

$$a = ka_R + (1-k)a_s \tag{1}$$

式中：a_s为监督学习算法提供的指导动作；k为 DDPG 算法与监督学习算法的融合度权重系数，取值范围为$[0,1]$。

监督式 DDPG 算法不修改 DDPG 算法的策略。但监督学习算法的介入时长和性能占比通常需要人为设定，这可能导致 DDPG 算法在已学习到比监督学习算法更优的策略时，监督学习算法产生阻碍作用[10-12]。本文利用融合度权重系数k分阶段自动调整监督学习算法[9]，即随着监督式 DDPG 算法逐步逼近最优策略，监督学习算法逐步退出，以免影响 DDPG 算法的性能。

1） 当$k \in \{0,1\}$时，$k=0$，表示监督式 DDPG 算法在初始学习阶段，智能体在监督学习算法下进行动作选择和训练；$k=1$，表示监督式 DDPG 算法在训练阶段后期，监督学习算法完全退出，智能体在 DDPG 算法下进行运动控制。

2） 当$k \in (0,1)$时，表示监督学习算法和 DDPG 算法同时存在，若 DDPG 算法没有向最优策略逼近，则k需选择较小的值，使监督学习算法占主导地位；随着 DDPG 算法不断向最优策略逼近，需逐渐增加k值，使 DDPG 算法逐步占主导地位。

2.1 监督采样

监督式 DDPG 算法根据 DDPG 算法选择动作a_R和监督学习算法指导动作a_s的误差梯度进行参数更新，即 Actor 网络参数更新引入了 DDPG 算法和监督学习算法的误差，使策略神经网络向监督学习算法的专家示范策略π_s逼近。Actor 网络的参数更新公式为

$$\theta \leftarrow \theta + k\Delta\theta_R + (1-k)\Delta\theta_s \tag{2}$$

式中：θ为 Actor 网络参数，$\Delta\theta_R$和$\Delta\theta_s$分别为 DDPG 算法和监督学习算法的网络参数变化量。

考虑到策略学习时，Q值是从采样数据中泛化训练得到的，因此可以利用监督学习算法得到的监督数据来提升策略神经网络的收敛速度，从而加快 DDPG 算法的学习进程。借助监督数据对动作加以指导，这相当于减小了包含最优动作的动作集和需要处理的状态数量，因此监督式 DDPG 算法可以更快地进行Q值估计和最优$Q_{\max}(s',a')$逼近。

在监督式 DDPG 算法强化学习的过程中，采样数据中包含监督数据的概率较大，尤其在训练阶段的前期，此时损失函数的计算公式为

$$L_{\text{all}}(\omega)=(1-\lambda)L_R(\omega)+\lambda L_s(\omega\,|\,P_s) \tag{3}$$

式中：λ 为监督学习采样数据占总采样数据的比例，L_R 为强化学习数据误差，L_S 为监督学习数据误差。

在监督式 DDPG 算法策略优化的过程中，专家示范策略引导网络参数加快收敛[13]。在向目标值逼近的过程中，Critic 网络的$\hat{Q}=\pi_\omega(s,\hat{a})$、融合监督学习后的$\hat{Q}(s',a')$均比无监督信号的$Q(s,a)$大，因此引入监督学习后更容易逼近目标值。

通过 Critic 网络和目标 Critic 网络进行网络参数更新时，策略梯度变为

$$r_t+\lambda\max Q_\pi(s',a')-\hat{Q}(s,a') \tag{4}$$

因为$\hat{Q}$值更大，所以更逼近目标值，更新梯度也向正方向优化，这使 Critic 网络计算的Q值更准确：

$$\theta\leftarrow\theta-a_a\text{mean}(\hat{Q}(s,a)) \tag{5}$$

基于上述更新策略，Actor 网络在策略学习过程中更快地向最优策略逼近，缩短了强化学习的时间。

2.2 行为克隆

在 DDPG 算法中，深度神经网络的隐藏层较多且神经节点连接复杂，本文利用反向传播（back propagation, BP）算法进行网络参数更新。在训练样本充足的情况下，以监督学习算法的专家示范策略为指导标签来训练策略神经网络，可实现专家示范策略的克隆[12]。

基于监督学习算法的专家示范策略，将状态s的估计误差e_s作为输入，采用监督回归的方法引导策略神经网络学习专家控制器输出的控制轨迹：

$$\{\tau_1,\tau_1,\cdots,\tau_n\} \tag{6}$$

每条专家控制轨迹样本都基于同样的状态–动作空间：

$$\tau_i=\left\{S_1^i,S_2^i,\cdots,S_n^i\right\} \tag{7}$$

从专家控制轨迹样本中随机采样，以每条控制轨迹所包含的状态–动作对作为监督学习算法采样的数据集，通过监督学习从该数据集采样并进行训练：

$$D=\left\{(s_1.a_1),(s_2.a_2),\cdots,(s_n.a_n)\right\} \tag{8}$$

为有效解决同一状态下连续采样导致的神经网络泛化能力下降的问题，在对监督学习采样数据集进行回归拟合时，需采样多条不同的控制轨迹样本。以控制动作为指导标签，状态误差为特征，通过 DDPG 神经网络进行回归学习来拟合逼近最优策略。需要注意的是，如果状态空间的样本数据较多，仅基于监督学习算法更新网络参数会带来时间累积误差，因此需要合理控制监督学习算法的介入程度。

2.3 监督式 DDPG 算法的 Actor 网络参数更新

DDPG 算法是以 Actor 网络最终的学习策略为最优策略，基于梯度下降原理进行网络参数更新：

$$\theta_{t+1}=\theta_t+\alpha\nabla_\theta\pi_\theta(s_t)\nabla_a Q_\omega(s_t,a_t)\big|_{a=\pi'_\theta(s)} \tag{9}$$

式中：α 为梯度更新步长。

Critic 网络的 Q 估计值基于动作 a 求导数获得梯度，并与 Actor 网络对其他参数的导数相乘得到更新梯度。

融合监督学习算法后，通过调整损失函数使策略更新向专家示范策略方向逼近，从而完成专家示范策略的克隆。在 Critic 网络参数更新时，引入边界函数，增加当前状态s下 Actor 网络的选择动作与专家示范策略下选择动作之间的差值，其损失函数也会同步增大。用$\pi_s(s)$表示专家示范策略函数，示教差距定义为

$$\left|\pi_\theta(s)-\pi_s(s)\right|=\Delta\pi_s \tag{10}$$

以监督学习数据为指导标签的 DDPG 神经网络参数更新可以表示为

$$\Delta\theta^s\leftarrow\alpha\delta(a_s,a_\pi)\nabla_\theta\pi_\theta(s) \tag{11}$$

式中：δ 为监督学习算法下的 TD-error，可用示教差距来计算：

$$\delta(a_s,a_\pi)=E(s) \tag{12}$$

以监督学习算法的每个采样状态误差来计算损失函数，使神经网络输出向专家示范策略方向逼近：

$$E(s)=\frac{1}{2}\Delta\pi(s)^2 \tag{13}$$

与 DDPG 算法原有的损失函数合并后，监督式 DDPG 神经网络的损失函数为

$$J=J_Q+J_s=\frac{1}{M}\sum\beta(r+rQ'-Q)^2+\frac{1}{2}(1-\beta)E(s) \tag{14}$$

采用最速梯度下降法对 DDPG 神经网络参数进行启发式更新，通过损失函数对误差的负梯度进行调整：

$$\Delta\theta_s=-\alpha\nabla_\theta E_\theta(s) \tag{15}$$

基于链式法则将公式(15)展开，并将同一状态 s 下的专家示范动作替换为采样动作，则监督学习部分的神经网络参数梯度可表示为

$$\begin{aligned}\Delta\theta_s&=\alpha_\theta(a_s\big|_{s\in S}-\pi_\theta(s))\nabla_\theta\pi_\theta(s)\Delta\theta_s\\&=\alpha_\theta(a_s\big|_{s\in S}-\pi_\theta(s))\nabla_\theta\pi_\theta(s)\end{aligned} \tag{16}$$

最终，监督式 DDPG 算法归一化处理 Q 值梯度和监督误差梯度的合并值，从而得到包含监督学习算法的目标 Actor 网络参数更新公式：

$$\theta\leftarrow\theta+\alpha\left[\beta\Delta_a Q_\theta(s,a)\big|_{a=\pi'(s)}+(1-\beta)\nabla E(s)\right]\nabla\theta\pi_\theta \tag{17}$$

式中：β 为权重系数，可通过人为调整 β 值来调整 DDPG 算法和监督学习算法的梯度比例。

在融合应用 DDPG 算法和监督学习算法时，需将 DDPG 控制器和监督学习控制器串联使用。两种算法的控制策略均要求智能体与环境进行交互，通过经验回放机制存储状态信息和专家示范策略信息，在损失函数中引入监督误差，使参数更新过程中在探索更大奖励值的同时向专家示范策略逼近。

3 监督式 DDPG 控制器的设计

基于本文提出的监督式 DDPG 算法设计监督式 DDPG 控制器，其框架如图 3 所示。

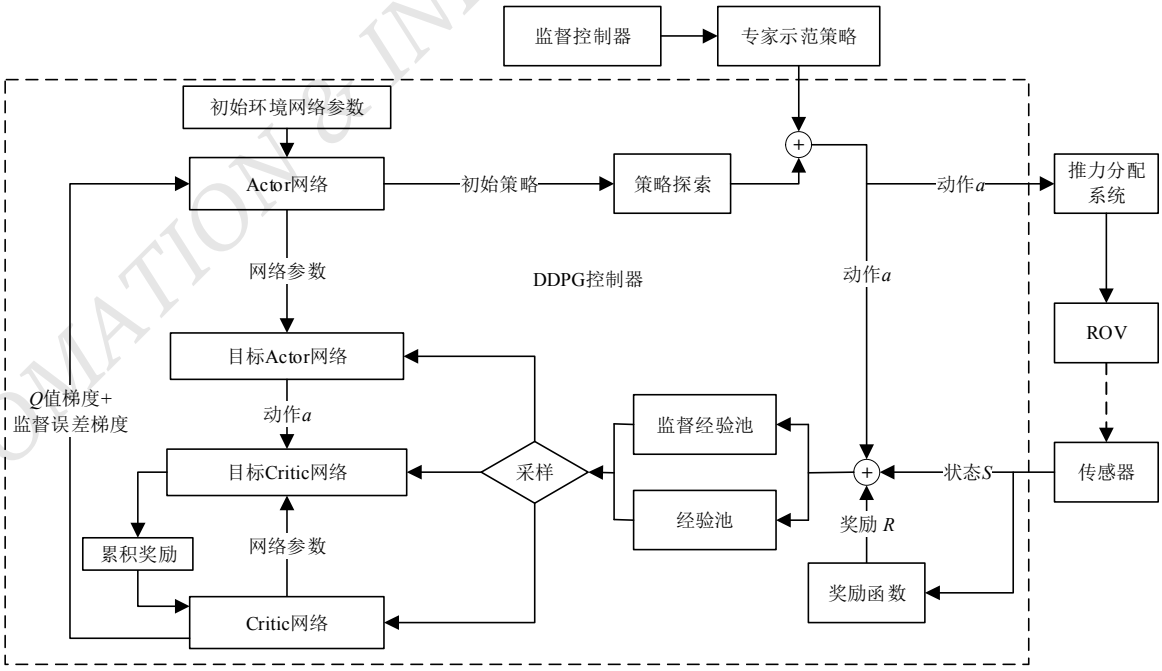

图 3 监督式 DDPG 控制器框架

监督式 DDPG 控制器的工作原理如下：在训练初期，利用预训练的监督控制器提供专家示范策略，主导决策以确保策略安全探索；同时，将 DDPG 控制器的 Actor-Critic 机制（Critic 网络评估动作价值，Actor 网络生成初始策略）与监督控制器提供的专家示范策略融合，通过调整融合度权重系数实现控制权从模仿到自主决策的平滑过渡。在此过程中，Actor 网络持续受到专家示范策略与奖励的双重引导，使 DDPG 控制器在保障安全性的前提下渐进优化策略，最终形成适应复杂动态环境的自主控制能力。

DDPG 算法采用在线探索和离线策略结合的方式进行训练和学习。为了更好地发挥监督学习算法的引导作用，设计了 2 个独立的经验池：一个用于存放 DDPG 算法的经验数据；另一个用于存放监督学习算法的监督示范经验数据。在训练阶段，从这 2 个经验池中并行采样，并利用监督示范经验数据引导 DDPG 神经网络的训练方向。

监督式 DDPG 算法融合度权重系数 k 按以下方式自动调节：

1） 训练初始阶段（$k=1$），此时监督控制器完全主导控制运动过程，每完成一个训练回合，k 减小为原来的 90%（即 $k \leftarrow k \times 0.9$），直到 $k=0.5$；

2） 当 $k=0.5$ 时，经验池已填满，DDPG 神经网络开始更新网络参数，此时监督学习算法和 DDPG 算法并行发挥作用，每完成一次网络参数更新，k 值减小为原来的 90%（即 $k \leftarrow k \times 0.9$），直到 $k=0$，监督控制器完成引导使命，完全退出，DDPG 控制器完全主导控制运动过程。

4 仿真试验与分析

通过仿真试验验证本文提出的监督式 DDPG 算法对小型 ROV 的控制效果。设定 Actor 网络的更新速率和 Critic 网络的学习速率均为 0.002，折扣因子 $\gamma=0.9$。为便于仿真，当前网络和目标网络均采用软更新的方式，更新速率为 0.01，经验池训练样本的容量为 2 000 个，仿真步长为 0.01 s，试验周期为 600 个回合，每回合步数为 500 步。

以 ROV 偏航角姿态定位控制为例进行仿真验证。设 ROV 的初始偏航角 $\psi_0=0°$，经过训练后，ROV 能够在 2 节航速下维持 $\psi_t=60°$；同时，在定航运行时，在第 20～25 s 期间引入幅度为 2°，均值为 0 的随机扰动。DDPG 算法和监督式 DDPG 算法的奖励值变化趋势对比如图 4 所示，监督式 DDPG 算法下的偏航角镇定过程如图 5 所示。

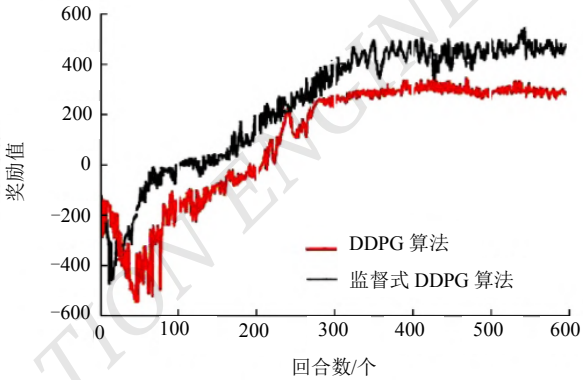

图 4 DDPG 算法和监督式 DDPG 算法奖励值变化趋势对比

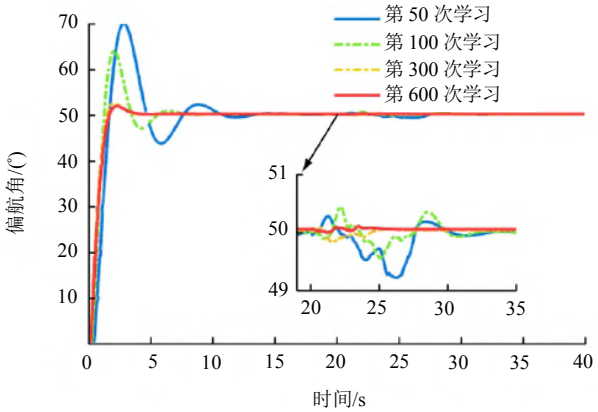

图 5 监督式 DDPG 算法下的偏航角镇定过程

由图 4 可知，监督式 DDPG 算法能更快地学习到具有更大奖励值的动作，且融合监督学习算法后，稳定后的奖励平均值由约 300 增加到 400 多，至少增加了 33%，证明了监督式 DDPG 算法的有效性。

由图 5 可知：不同学习次数的学习效果存在差异，学习次数越多，控制性能越优异。在第 300 次学习后，监督式 DDPG 算法对偏航角的控制性能基本上

达到了预期要求；对比第 50 次学习和第 600 次学习的情况，偏航角超调量由 30%下降到 3%，镇定耗时由 15 s 缩短至 4 s，说明融合监督学习算法后的 DDPG 算法学习效果有明显提升。

5 结论

本文在 DDPG 算法中引入监督学习算法，提出了监督式 DDPG 算法。从仿真试验结果可知，本文提出的监督式 DDPG 算法与 DDPG 算法相比，控制效果有明显提升。但将该算法应用于小型 ROV 运动控制时，仍然存在虚实迁移效果差的问题，后续仍需对该问题进行深入研究。

参考文献

[1] 李若霆.基于深度强化学习的视觉导航算法研究[D].太原:中北大学,2023.

[2] 蔡军,苟文耀,刘颜.基于 actor-critic 框架的在线积分强化学习算法研究[J].电子测量与仪器学报,2023,37(3):194-201.

[3] 张严心,孔涵,殷辰堃,等.一类基于概率优先经验回放机制的分布式多智能体软行动-评论者算法[J].北京工业大学学报,2023,49(4):459-466.

[4] 陈恺丰,田博睿,李和清,等.基于 DDPG 算法的双轮腿机器人运动控制研究[J].系统工程与电子技术,2023,45(4):1144-1151.

[5] 李凌霄,王伟明,贺佳飞,等.基于 DDPG 的自主水下机器人角度控制研究[J].计算机仿真,2023,40(4):422-426;503.

[6] 王鹃,张冲,龚家新,等.基于机器学习的模糊测试研究综述[J].信息网络安全,2023,23(8):1-16.

[7] 江铃燚,郑艺峰,陈澈,等.有监督深度学习的优化方法研究综述[J].中国图象图形学报,2023,2(4):963-983.

[8] Uc-cetina V. Supervised reinforce learning using behavior models[C]//Sixth International Conference on Learning and Applications(ICMLA 2007).IEEE, 2007:336-341.

[9] 杨辉,王禹,李中奇,等.专家监督的 SAC 强化学习重载列车运行优化控制[J].控制理论与应用,2022,39(5):799-808.

[10] 苏萌韬,曾碧.基于渐进式神经网络的多任务强化学习算法[J].机电工程技术,2022,51 (11):21-25.

[11] 曾纪钧,梁哲恒.监督式强化学习在路径规划中的应用研究[J].计算机应用与软件,2018,35(10):185-188.

[12] 王亦晨,刘雪梅.基于冲突搜索增强深度强化学习的多 AGV 路径规划方法[J].机电工程技术,2024,53(8):23-27;88.

[13] ARGALL B D, CHERNOVA S, VELOSO M, et al. A survey of robot learning from demonstration[J]. Robotics and Autonomous Systems, 2009,57(5):469-483).

[14] ROSENSTEIN MT, BARTO AG, SI J, et al. Supervised actor-critic reinforcement learning[J]. Learning and Approximate Dynamic Programming: Scaling Up to the Real WORLD, 2004:359-380.

作者简介：

黄兆军，男，1982 年生，硕士研究生，高级工程师，主要研究方向：智能控制。E-mail: hzj4735@126.com

张彦佳，女，2003 年生，专科，主要研究方向：电气自动化技术。

左晓雯，女，2002 年生，专科，主要研究方向：电气自动化技术。

陈泽汛，男，2002 年生，专科，主要研究方向：电气自动化技术。