•智能交叉科学与工程•

DOI:10.12454/j.jsuese.202301015

本刊网刊

基于改进强化学习的复杂海况下船舶升沉补偿控制研究

张　琴,周静宜,王星月,胡　雄*

(上海海事大学 物流工程学院,上海 201306)

摘　要:受多变海况下风浪涌影响,剧烈的船舶随机运动威胁着海上风机吊载安装的安全性,并对海上作业和人员换乘等造成不好的影响。为提高海上作业的可靠性、安全性和稳定性,针对船舶升沉运动补偿中面临的环境多变和建模不准确的问题,提出基于改进强化学习的复杂海况下船舶升沉补偿控制方法。首先,对由伺服驱动器、伺服电机及编码器和电动缸构成的补偿系统进行机理法建模,作为强化学习训练的环境。其次,通过马尔可夫决策过程描述智能体的策略和奖励,同时采用双延迟深度确定策略梯度算法(TD3)强化学习算法作为控制策略,将Actor网络中的输出层TanH层的幅值扩大两倍,使TD3算法能更好地应对复杂海况,并通过主网络和目标网络的Actor和Critic 6个网络的迭代更新,得到自学习的最优控制动作输出。最后,添加OU动作噪声到目标策略中,能够更好地探索状态空间,并通过组合奖励函数改善智能体的学习和决策能力,使智能体可以适应复杂海况下的强化学习任务,在不同海况等级下利用已知的信息找到最优解,从而避免陷入局部最优,以提高船舶升沉运动的补偿精度。仿真结果表明,在恶劣的复杂海况下改进TD3算法有较好的补偿控制效果;将训练好的模型用于3～6级以及历时更长的变海况环境下的船舶升沉运动补偿,其补偿效率最高可达到99.95%,优于PSO优化的反步法控制和传统TD3强化学习方法,体现了良好的泛化性。

关键词:复杂海况环境;船舶升沉运动;补偿控制系统;TD3强化学习

中图分类号:P751　　**文献标志码:**A　　**文章编号:**2096-3246(2025)04-0123-15

在低碳发展的大前提下,可再生能源的开发与利用已经成为今后的发展方向[1]。由于海上风力资源稳定、发电功率大,因此海上风力发电作为新能源的重要推手在全球范围内迅速发展[2]。但风机的安装运维在特殊的海洋环境中也伴随着更大的挑战[3-4]。

多变海况环境的起伏影响使得风电运维船舶呈现复杂的运动状态,这给需要通过舷梯换乘到风机平台进行维护的海上作业人员带来了安全隐患[5]。其中船舶的升沉运动对舷梯换乘影响最大,而通过控制搭载舷梯的补偿平台,减少舷梯和靠船船桩之间的相对升沉位移,使其与风电塔保持相对静止,可提高人员转移的安全性[6]。因此,开展对补偿平台的建模和控制方法研究,实现多变海况环境下船舶的运动补偿,对安全高效地维护风机吊载具有重要意义。

补偿平台控制由伺服驱动器、伺服电机及编码器和电动缸等组成,稳定性控制的前提是搭建系统模型,常用建模方法有实验测试法和机理建模法。实验测试法根据被控对象的输入和相应输出信号关系,运用数学模型系统建模。Salah–Eddine等[7]利用多项式模型对输入输出数据处理建模,实现了对伺服电机的模型辨识,最终得到带外部干扰项的误差模型,可用于模拟或实际应用,但精度有限,不适用多变环境。机理分析法在研究系统运行机理的基础上,通过理论推导得到系统模型。翟富刚等[8]通过机理分析法构建永磁同步伺服电机数学模型和动力学模型,对相关机械参量进行折算并在传动系统等效力学模型中加入伺服电动缸的动态刚度模型,建立完整的伺服电动缸数学模型,但在复杂过程中,难以对模型中的参数进行测量、辨识,导致模型不准确。

基于船舶运动补偿系统模型,常用的控制方法

收稿日期:2023－12－11　　**修回日期:**2024－04－23　　**网络出版日期:**2024－06－03

基金项目:国家自然科学基金项目(NSFC52105466)

作者简介:张　琴(1982—),女,博士,副教授. 研究方向:深度强化学习的稳定性补偿控制. E-mail:qinzhang@shmtu.edu.cn

* **通信作者:**胡　雄,教授,E-mail:huxiong@shmtu.edu.cn

有：比例积分微分控制（PID）、反步法控制（BSC）、模型预测控制（MPC）和滑模控制（SMC）[9−10]。PID 通过控制被控对象实现控制需求，使控制系统响应向控制目标靠拢[11]。Mei 等[12]以伺服液压电机为研究对象，提出一种针对船舶升沉运动的变参数 PID 控制，通过调整控制参数补偿船舶的升沉运动，实现补给船在稳定状态下释放，但由于 PID 参数调节需大量时间，难以实现实时跟踪。模型预测控制以模型为基础，计算有限时间内的最优控制率，通过反馈的测量值和预测值误差校正整个控制，但它对模型准确度要求高，泛化性不强。Woodacre 等[13]研究了一种并行 MPC–PID 主动升沉补偿系统的控制方法，将控制误差分别输入到 PID 和 MPC 控制中，并将两个控制信号求和后输给被控对象以实现对系统的控制，但此方法过于依赖模型，不适合复杂多变的研究对象。反步法将复杂的非线性系统分解成阶数小的子系统，利用李雅普诺夫稳定性定理推导控制速率，提高闭环系统的稳定性。马长李等[14]在 3 级海况下推导船舶升沉运动轨迹及其电液系统的非线性模型，验证了自适应反步补偿策略的稳定性，但反步法控制率公式中含有的未知参数（k_1、k_2）会影响跟踪效果，甚至会出现失控现象，可采用 PSO 算法优化反步法控制中的参数。Zhang 等[15]设计一种反步控制方法来跟踪船舶的升沉运动，并采用粒子群算法对控制参数进行优化，补偿效率有所提高。滑模控制是以现代控制理论为基础，以李雅普诺夫函数为主要数学核心的一种控制理论。Cai 等[16]为舰载 Stewart 平台设计了 SMC 方案，并提出一种新的速度前馈补偿器改善其控制性能。但在实际应用中，SMC 难以跟踪滑块运动过程，可能使系统产生振荡。以上控制方法均基于模型，而船舶运动随复杂海况时变，当被控对象模型发生变化时，需重新设计控制率。因此，探究无模型的强化学习控制算法用于船舶运动补偿系统，对提高补偿系统的泛化性和精度有着重要意义。

目前，强化学习在控制中已经有了广泛的应用。Yang 等[17]采用了深度确定性策略梯度算法（DDPG）生成空战机动动作值，并将优化动作作为初始样本添加到 DDPG 回放缓冲区中，有效避免了随机生成初始动作导致的学习效率低下，但确定性策略不利于动作的探索，容易陷入局部最优解。Zinage 等[18]采用了一种无模型的在线强化学习 DDPG 算法来获取训练实验中的经验，但是 DDPG 方法在学习时会出现估计值函数过大的问题。双延迟深度确定策略梯度算法（TD3）选取最小的函数值作为更新目标，抑制了持续过高估计。Chu 等[19]基于 TD3 算法，采用监督学习的方法，从闭环控制数据中克隆行为策略，加快了强化学习算法的训练收敛速度，但 TD3 算法中值函数的单步更新会产生积累误差。Zhang 等[20]提出 PID−Guide TD3 算法，通过选择 PID 动作和 TD3 网络输出中评价函数较高值作为动作值，从而加快 TD3 算法的训练速度，其中评价函数是训练难点。

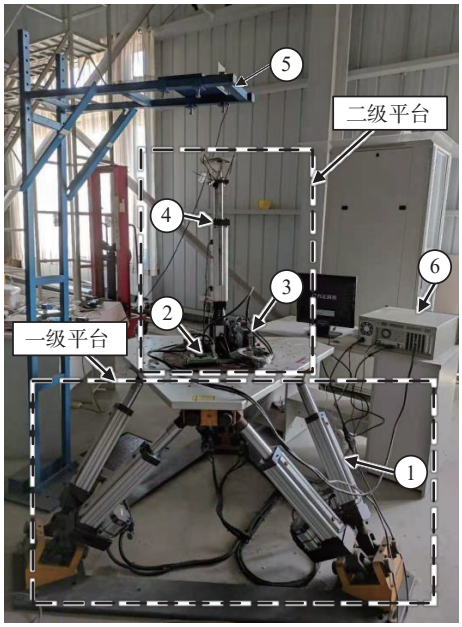
综上所述，少有将强化学习算法[21]应用于船舶运动补偿系统中的成果，需对补偿系统搭建合适的网络及环境以验证强化学习对多海况下的船舶运动的补偿效果和泛化性。尽管有学者采用 TD3 算法研究多变海况，但补偿系统的训练速度有待提高。故本文开展在多变海况环境下船舶升沉运动补偿控制系统运行机理的研究，并通过改进 TD3 算法对补偿系统进行控制，使补偿系统的泛化性得到更大的改善。

1 船舶升沉运动补偿模型

1.1 问题描述

受风、浪等自然环境作用影响，船舶会产生不规则运动。海况等级越高，船舶升沉运动越复杂，而船舶升沉运动往往会影响海上作业的安全性和稳定性。本文主要研究多海况下船舶升沉运动，因此，需要建立船舶升沉运动补偿模型，进而进行补偿控制。

船舶升沉运动补偿系统由一级平台和二级平台构成，图 1 为实验室搭建的船舶升沉补偿实验系统。



1.Stewart平台; 2.伺服驱动器; 3.伺服电机及编码器; 4.电动缸; 5.激光传感器; 6.工控机及运动控制卡。

图 1 船舶升沉补偿实验系统

Fig. 1 Experimental system for ship heave and sink compensation

Stewart 平台[22–23]称为一级平台，用来模拟船舶的运动；伺服驱动器、伺服电机及编码器和电动缸构成的整体为二级平台，即为（船舶升沉运动）补偿系统，通过控制伺服电动缸的运动来实现船舶运动的补偿；激光传感器的功能是将二级平台的位置情况反馈给工控机，实现控制的闭环。

1.1.1　伺服电机模型

为实现对补偿系统的控制，首先要对执行机构建模。补偿系统由伺服驱动器驱动伺服电机[24–25]转动，然后由电机带动电动缸内部的滚珠丝杠转动，从而使电动缸做伸缩运动，因此需要对伺服电机和电动缸建模[26–28]。

伺服电机模型反映实际电机动作的幅值和延时时间，通过给定电机输入电压时的输出转矩来驱动负载，因此本文在确保满足此条件的前提下，用机理法建立两相交流电机的模型，得到伺服电机的传递函数模型 G_1，如式（1）所示：

$$G_1(s)=\frac{\theta_{\mathrm{m}}(s)}{U_{\mathrm{a}}(s)}=\frac{C_{\mathrm{m}}}{s(J_{\mathrm{m}}s+(f_{\mathrm{m}}+C_{\omega}))} \tag{1}$$

式中，θ_{m} 为电机转子角位移，U_{a} 为输入电压，J_{m} 和 f_{m} 分别为折算到电动机上的总转动惯量和总黏性摩擦系数，C_{m} 为常系数，C_{ω} 为阻尼系数。

根据本文的电机额定电压、额定转矩、额定输出功率等相关参数，可以得到 C_{m}=0.013 N·m/V，$C_{\omega}=2.94\times10^{-2}$ N·m/(r·s^{-1})，$f_{\mathrm{m}}=2.54\times10^{-2}$ N·m/(r·s^{-1})，$J_{\mathrm{m}}=0.678\times10^{-4}$ kg·m^2，代入式（1）可得本文使用的伺服电机的传递函数：

$$G_1(s)=\frac{13}{0.0678s^2+54.8s} \tag{2}$$

1.1.2　电动缸模型

电动缸是通过高强度伺服同步带将电机和滚珠丝杠进行模块化设计，将电机的旋转运动转化为滚珠丝杠螺母的直线运动，进而转化为电动缸内伸缩杆的直线运动。理想情况下，电机转动角度 θ_{m} 和伸缩杆位移 X_{L} 之间为线性关系，但在实际应用中，在电机输出的角位移通过传动机构转化为伸缩杆的位移时存在延时，故在传递函数中加入惯性时间常数为 T_{d} 的 1 阶惯性环节。因此，电动缸运动的数学模型为：

$$G_2(s)=\frac{X_{\mathrm{L}}(s)}{\theta_{\mathrm{m}}(s)}=\frac{P_{\mathrm{n}}}{2\pi i(T_{\mathrm{d}}s+1)} \tag{3}$$

式中：i 为减速比，i=1；P_{n} 为滚珠丝杠的导程，根据丝杠参数，可得 P_{n}=5 mm，代入式（3），电动缸的传递函数为：

$$G_2(s)=\frac{5}{3.14s+6.28} \tag{4}$$

由于本文的运动控制卡采用差动方式控制伺服电动缸，其模型可以简化为一个微分环节，综合式（2）和（4），整个伺服电动缸执行系统的传递函数为：

$$G_3(s)=\frac{X_{\mathrm{L}}(s)}{U_{\mathrm{a}}(s)}=\frac{1\,729.53}{s^2+810.23s+1\,616.44} \tag{5}$$

基于得到的伺服电动缸执行系统模型，需要强化学习算法控制以实现船舶升沉补偿的目的，首先要通过马尔可夫决策过程描述强化学习的状态转移具体过程。

1.2　船舶升沉运动补偿的 MDP 模型

马尔可夫决策过程（MDP）是一种序贯决策的数学模型，可在具有马尔科夫特性的系统中来模拟智能体的策略和奖励。马尔可夫决策过程由状态空间 S、动作空间 A、奖励函数 R、状态转移概率 P 和折扣因子 γ 组成。在船舶升沉运动补偿系统中的 MDP 模型描述如下。

1）状态空间

状态空间表示实际运动位移和实际运动速度，智能体在获取观测到的环境状态空间后选择动作值执行，然后环境空间会发生变化。本文研究船舶升沉方向的运动补偿控制，因此需要把船舶实际运动位移和实际运动速度作为状态空间中的量。本文选取的状态空间如式（6）所示：

$$S=(x_{1\_\mathrm{actual}},x_{2\_\mathrm{actual}},v_{1\_\mathrm{actual}},v_{2\_\mathrm{actual}}) \tag{6}$$

式中：x_{1_actual}、x_{2_actual} 分别为图 1 中一、二级平台的实际运动位移，能直观地监测补偿效果；v_{1_actual}、v_{2_actual} 分别为一、二级平台的实际运动速度，能检测平台运动的快慢，进而使船舶升沉运动补偿更好地进行。

2）动作空间

动作空间表示智能体输出的动作，用来补偿船舶升沉运动系统。本文对船舶升沉方向的运动进行补偿控制，由于船舶升沉运动补偿系统的控制采用的是位置控制，因此智能体输出的动作值为二级平台的输入 x_{2_control}：

$$A=x_{2\_\mathrm{control}} \tag{7}$$

3）奖励函数

MDP 神经网络依据奖励值的引导不断迭代更新才能得到满足船舶升沉运动补偿的网络参数。对于船舶升沉运动补偿系统，智能体的目标是通过控制二级平台的位移来更好地补偿一级平台的位移量，即二级平台的位置偏差越小越好，以下称之为（位移）补偿误差。因此设计具有评价此变量能力的奖励函数 R_{position}：

$$R_{\mathrm{position}}=-\left|x_{1\_\mathrm{actual}}+x_{2\_\mathrm{actual}}\right| \tag{8}$$

式（8）表明，当环境中的一级平台和二级平台间位移之和的绝对值越小，所获得的奖励越大，明确地向智能体传递了位移补偿误差越小越好的目标。

训练过程中，在满足位移标准的同时，二级平台升沉运动的速度也需要和一级平台保持一致（称之为速度补偿误差），这样才能保证电动缸动作的平稳性。此时奖励函数 R_{velocity} 为：

$$R_{\text{velocity}}=-\left|v_{1\_\text{actual}}+v_{2\_\text{actual}}\right| \tag{9}$$

综合式（8）和（9），将两个奖励函数按照不同的权重整合为总体的奖励函数 R_{total}：

$$R_{\text{total}}=\alpha R_{\text{position}}+(1-\alpha)R_{\text{velocity}} \tag{10}$$

式中，α 为位移补偿误差对应的奖励 R_{position} 所占权重。由于位移补偿误差奖励 R_{position} 要比速度补偿误差奖励 R_{velocity} 更能反映补偿效果的好坏，因此将前者的权重设置得较大一些。

4）状态转移概率

智能体执行某个动作之后，当前状态会以某种概率转到另一个状态。假设在 t 时刻，当状态为 S 时，智能体执行动作 A 后，在 t+1 时刻，状态变为 S' 的概率为状态转移概率，即：

$$P_{ss'}=PS_{t+1}=S'|(S_t=S,A_t=A) \tag{11}$$

式中，$P_{ss'}$ 为智能体从状态 S 转到另一个状态 S' 的概率。本文中，动作 A 为二级平台的输入 x_{2_control}，状态 S 为式（6）中的状态空间。

5）折扣因子

由于智能体的每一步行动对后续都是有影响的，除了设定环境奖励 R_{t+1}，还可以将后续的奖励也累加起来，形成总体奖励 G_t，即：

$$G_t=R_{t+1}+\gamma R_{t+2}+\cdots=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \tag{12}$$

式中，γ 表示奖励衰减因子，其取值在[0,1]之间。当 γ 为 0 时，智能体只考虑当下的奖励，不依赖于未来获得的奖励，称为贪婪算法。当 γ 为 1 时，表示对当前和未来的奖励都平等对待，而不进行衰减处理。一般情况下，γ 取 0 到 1 之间的数字，表示当前奖励的权重比未来奖励的权重大。

综上所述，船舶升沉运动补偿的 MDP 模型如图 2 所示。智能体获取观测到的环境状态空间 S 后，由对应算法的智能体选择一个动作值 A 进行执行。根据变化后的环境状态与目标之间的关系设计奖励函数 R，然后将奖励值也反馈给智能体，经过迭代智能体得到最大化的累计奖励，从而实现自主决策和控制。基于多变海况下船舶升沉运动，要实现对船舶高效率的补偿控制，不仅要通过 MDP 模型描述 TD3 算法中智能体的策略[29]和奖励，还需要对 TD3 算法进行网络结构、奖励函数、动作噪声等 3 个方面的改进。

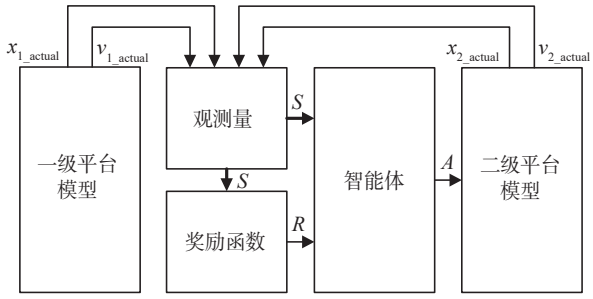


图 2 船舶升沉运动补偿的 MDP 模型

Fig. 2 MDP model for compensation of ship heave motion

2 改进 TD3 补偿船舶升沉运动

2.1 TD3 算法基本结构

为了对模型准确地进行船舶升沉运动补偿，本文采用无模型的双延迟深度确定性策略梯度（TD3）算法。为了使网络表示的值函数更稳定，TD3 算法的网络结构分为在线网络和目标网络两个部分，在线网络产生当前的 Q 值，目标网络产生下一步的 Q 值，Q 值指当前状态下执行动作的价值。在线网络和目标网络分别包含一个 Actor 网络和两个 Critic 网络。Actor 网络是根据策略输出要执行的动作，Critic 网络是根据价值函数对动作的好坏进行评价。

1）两个 Critic 网络

两个 Critic 网络分为在线 Critic 网络和目标 Critic 网络。在线 Critic 网络负责计算当前 Q 值，同时更新自身的网络参数；目标 Critic 网络负责计算目标 Q 值，其网络参数定期由式（13）进行更新。假设观测到真实值之后，样本中存在随机噪声 ε，由于存在这样的一个误差，所以估计的最大值一般会比真实的最大值还要大，如式（14）所示：

$$Q_{\theta'}=R+\gamma\max_{a'}Q(S',A') \tag{13}$$

$$E_{\varepsilon}(\max_{A'}(Q(S',A')+\varepsilon))\geq\max_{A'}Q(S',A') \tag{14}$$

式（13）～（14）中，R 为每一步获取的奖励，S' 为下一步智能体的状态，A' 为下一步智能体将要选择的动作，E_{ε} 为噪声均值，$Q_{\theta'}$ 为目标 Critic 网络的网络参数。

在每一次迭代学习过程中，估计的是当前价值最大动作的 Q 值，大于真正与环境交互动作的 Q 值，因此会造成 Q 值估计过高的问题。

针对这一问题，TD3 算法采用两个 Critic 网络分别对 Q 值进行估计，根据估计值的大小，选择较小的 Q 值作为目标值，有助于抑制价值函数的高估问题，如式（15）所示：

$$y=R+\gamma\min_{i=1,2}Q_{\theta'_i}(S',A') \tag{15}$$

式中，y为目标Q值，用于更新Critic网络的参数，使Q值估计更准确。

TD3算法采用两个Critic网络，其网络结构如图3所示。

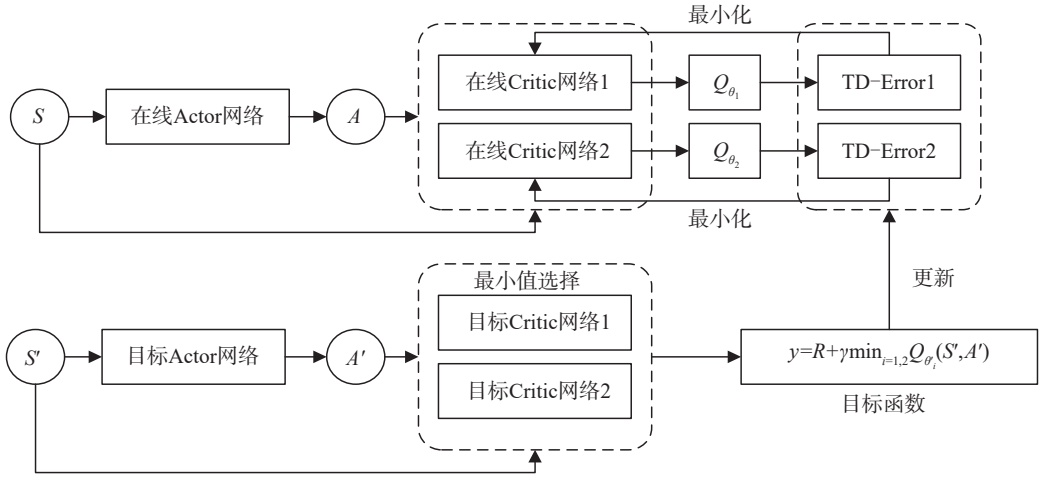


图3　TD3算法网络结构图

Fig. 3　Network structure diagram of TD3 algorithm

图3中，在线Actor网络和目标Actor网络的网络参数分别为φ和φ'，在线Critic网络和目标Critic网络的网络参数分别为$\theta_{i=1,2}$和$\theta'_{i=1,2}$。在线Actor网络的输入是环境状态S，输出是动作A；目标Actor网络的输入是下一刻的环境状态S'，输出为该状态下的动作A'。两个在线Critic网络根据状态和动作，分别输出对应的Q值。为避免Q值被高估，选择两个目标Critic网络中输出值最小的一个对目标进行更新。

2）延迟Actor网络更新

为了使网络的训练更加稳定，TD3算法采用目标网络软更新的方式，即在当前网络更新数次之后，再进行目标网络的更新。除此之外，TD3算法还对Actor网络采用延迟更新的思想，智能体先按照Actor网络的输出执行动作，在Critic网络更新H次之后再进行Actor网络的更新：

$$\varphi(k+h)\leftarrow\varphi(k)+\Delta\varphi(k+h),h>H \tag{16}$$

式中，h为延迟更新的步数间隔。

3）目标策略平滑正则化

TD3算法加入了正则化，即在由策略输出的动作中加入噪声ζ，如式（17）所示，这样对于相似的动作，对应的Q值也相似。因此，减小了函数逼近时误差产生的影响，使Critic网络计算更加准确，动作选择更加平滑。

$$A'\leftarrow\pi_{\varphi'}(S')+\zeta,\zeta\sim\mathrm{clip}(N(0,\sigma),-c,c) \tag{17}$$

式中，智能体通过确定性函数，根据输入的状态S'输出动作A'，$\pi_{\varphi'}$为目标Actor网络的网络参数φ'对应的策略，ζ为服从正态分布N的噪声，σ为噪声方差。为了避免在小概率中抽样出很大或很小的值，对噪声做裁剪处理，使噪声处在[$-c$,c]的范围。

因此，TD3算法在解决了过估计问题的基础上，采用了延迟Actor网络更新和正则化的思想，增强了算法的稳定性。但是，在采用此算法解决具体问题时，还需对其环境状态、动作等要素进行合理设计，才能使网络结构能够满足多海况的补偿要求，奖励函数能更好地引导强化学习算法进行探索并收敛，从而增强算法的泛化性。

2.2　TD3算法用于船舶升沉运动补偿控制的难点及改进

2.2.1　船舶升沉运动补偿控制的难点问题

将TD3算法应用到船舶升沉运动补偿控制系统中存在难点问题，需要进行以下改进：1）通过对网络结构的改进，达到多海况补偿控制的需求；2）对奖励函数进行改进，以更好地引导强化学习算法进行探索并收敛；3）对动作噪声进行改进，使动作噪声更适用于本文的船舶升沉运动补偿系统。

2.2.2　Actor网络结构设计

Actor网络特征输入层是将环境状态S或动作信息A等特征作为网络的输入，通过全连接层把提取到的特征综合起来，经由非线性的ReLU层激活非线性映射，再由TanH激活函数得到输出。TanH函数是双曲正切函数，该函数相当于是一种阈值函数，将输出限制在−1～1 m之间。由于本文实验中涉及的海况（6级海况）范围超过−1～1 m，因此，Actor网络的输出层采用TanH′激活函数，即幅值为原TanH激活函数的两倍，如式（18）所示：

$$\mathrm{TanH}'(x)=2\mathrm{TanH}(x)=2\frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}} \tag{18}$$

2.2.3　奖励函数设计

对于船舶升沉运动补偿系统来说，评价控制策略

好坏的依据是对整个控制过程进行评判，即对一段时间内实际控制后运动的位移进行测量，总的补偿误差越小越好。因此，对于船舶升沉运动补偿系统强化学习的奖励函数设计，应遵循补偿后的误差越小则反馈奖励值越大的规则。考虑到补偿误差较大时需避免奖励变化过大造成动作变化过快，在误差小一些时需增大奖励值来加快训练，并且奖励值为0时也需要将误差大小明显区分开，以及在不同误差值下智能体有不同的探索速度等问题，本文结合正态分布和线性函数进行奖励函数的设计，使奖励反馈值更好地引导智能体进行训练，如图4所示。

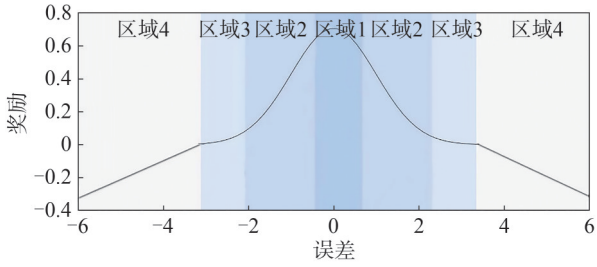


图4 正态分布和线性函数组合的奖励函数示意图

Fig. 4 Schematic diagram of the reward function composed of the normal distribution and the linear functio

由图4可知：在补偿误差值小的情况下，正态分布为主导；在补偿误差值较大时（区域3），智能体按奖励值较大的方向寻找训练的目标。当补偿误差值稍小时（区域2），奖励函数的斜率变大，奖励值变大，能够加快算法的收敛。当补偿误差值很小时（区域1），奖励函数的斜率减小，这表明当误差在一定的范围内时，奖励值相差不大，可稳定输出动作。误差极大时（区域4），线性函数为主导，智能体在所有误差值下的探索速度不变，能够引导智能体最终收敛。式（19）为组合奖励函数R_{total}的数学公式：

$$R_{\text{total}}=0.9\times R(e_{\text{x}})+0.1\times R(e_{\text{v}}),$$

$$R(e_{\text{x}})=\begin{cases}-5\times|e_{\text{x}}|, e_{\text{x}}>0.01;\\ \mathrm{e}^{-\frac{e_{\text{x}}^2}{9\times10^{-6}}}, e_{\text{x}}\le 0.01\end{cases}$$

$$R(e_{\text{v}})=\begin{cases}-5\times|e_{\text{v}}|, e_{\text{v}}>0.01;\\ \mathrm{e}^{-\frac{e_{\text{v}}^2}{9\times10^{-6}}}, e_{\text{v}}\le 0.01\end{cases}$$

$$e_{\text{x}}=x_{1\_\text{actual}}+x_{2\_\text{actual}},\ e_{\text{v}}=v_{1\_\text{actual}}+v_{2\_\text{actual}} \tag{19}$$

式中，e_{x}为一、二级平台之间的位移之和，e_{v}为一、二级平台之间的速度之和，$R(e_{\text{x}})$为位移补偿误差对应的奖励，$R(e_{\text{v}})$为速度补偿误差对应的奖励，R_{total}为按照不同权重整合为总体的奖励函数。由于奖励函数的设计遵循规则为补偿后的误差越小反馈的奖励越大，因此将线性奖励函数中的斜率设为−5。位移补偿误差对应的奖励要比速度补偿误差对应的奖励更能反映补偿效果的好坏，前者权重设置更大，因此将位移补偿误差对应的奖励所占权重设为0.9，速度补偿误差对应的奖励所占的权重设为0.1。通过组合奖励函数可知，随着平台之间的位移之和与速度之和增大，智能体在前期获得的线性奖励上升速度较快，后期获得的正态奖励上升速度也较快。因此，组合奖励函数能够使智能体朝着奖励值更高的方向训练。

2.2.4 动作噪声设计

式（14）中强化学习的动作噪声决定了智能体的探索能力，对惯性系统来说，从发出控制信号到实际物体产生速度或位移的动态过程是一个积分过程，而积分过程是典型的低通滤波，需要变化较平缓的OU噪声信号。同时，当不断地向实际硬件设备输入不同大小的控制信号，若按照前后控制信号差距很大的方式进行频繁地探索，无疑会使硬件设备损坏得更快。因此，为延长硬件设备的使用寿命，也可采用前后变化较小的OU噪声进行探索。

从式（14）可以看出，Critic网络在更新网络参数时，样本中存在的随机噪声ε会使估计的最大值增大，此时采用强化学习中的OU动作噪声来提高算法的训练效果。在每个采样时间步长k处，使用以下公式更新噪声值$N(k)$：

$$N(k+1)=N(k)+c_{\text{mac}}(N_{\text{m}}-N(k))T_{\text{s}}+N_{\text{sd}}(k)\cdot \text{randn}\left(\text{size}\left(N_{\text{m}}\right)\right)\cdot\text{sqrt}(T_{\text{s}}) \tag{20}$$

式中，T_{s}为智能体的采样时间，c_{mac}为噪声模型常数，N_{m}为噪声模型均值，randn表示用于生成一组符合正态分布的随机数的函数。为了使OU噪声在均值附近做出线性负反馈，在每个采样时间步中，OU噪声的标准差衰减公式如下所示：

$$D_{\text{sd}}(k+1)=\max\left(N_{\text{sd}}(k)(1-N_{\text{sdd}}),N_{\text{sdmin}}\right) \tag{21}$$

式中，N_{sd}为标准差，N_{sdd}为标准差的衰减率。对于连续动作信号，适当地设置噪声标准差以鼓励探索非常重要。通常将$N_{\text{sd}}\cdot\text{sqrt}(T_{\text{s}})$设置为动作范围的1%～10%之间的值。如果强化学习中的智能体收敛太快，导致陷入局部最优解难以进一步寻优，可以通过增加噪声来促进代理探索，也就是增加标准差N_{sd}。此外，为了增加探索，可降低标准差的衰减率N_{sdd}。

综上，本文针对船舶升沉补偿系统的控制问题，在强化学习TD3算法框架下进行了3个关键方面的改进：将TD3算法的网络结构中Actor网络的输出层采用自定义TanH′激活函数，即幅值为原TanH激活函数的两倍，在本文船舶升沉运动位移超过−1～1 m，使用TanH′激活函数时的补偿误差明显降低；TD3算法中的奖励函数采用线性函数和正态分布函数相结合的

方法，同时具备了前期和后期较快的收敛速度；TD3算法中的动作噪声选用OU噪声进行动作探索，经过智能体的训练后，OU噪声最终收敛时获得的奖励要大于使用高斯噪声训练时的奖励。

改进后的TD3算法要通过船舶升沉运动补偿平台进行训练，如图5所示。

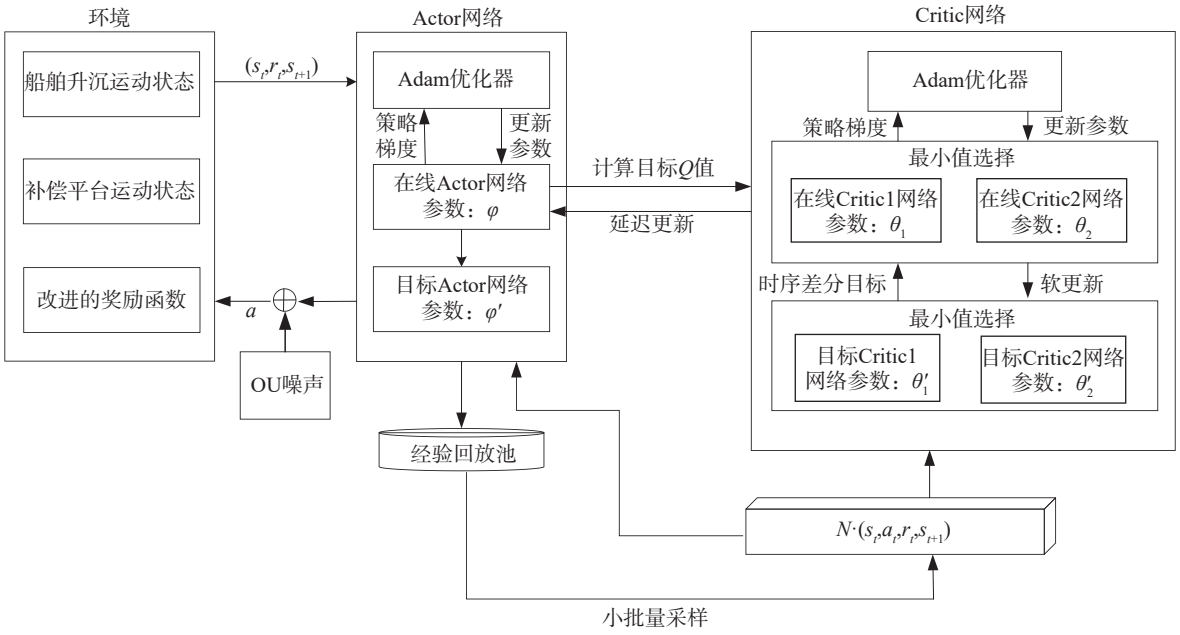


图5　基于船舶升沉运动补偿的改进TD3控制方法示意图

Fig. 5　Schematic diagram of the improved TD3 control method based on ship heave motion compensation

本文中船舶升沉运动补偿的强化学习训练过程分为4个部分。首先，将运动数据直接输入一级平台来模拟船舶运动；其次，强化学习智能体通过Actor网络与环境交互，在t时刻接收状态s_t后输出动作a_t、随后环境返回奖励r与下一时刻的状态s_{t+1}，加入改进的OU噪声后作用于环境中的二级平台，得到补偿平台的运动数据，即下一时刻的状态s_{t+1}和奖励r；然后，通过从经验回放池里取出状态(s_t,a_t,r,s_{t+1})来更新Critic网络和Actor网络，并每隔一段时间更新目标Actor和目标Critic网络；最后，经过不断地迭代更新，网络策略朝着奖励值更高的方向训练，从而提高强化学习的训练和控制效果，得到控制平台高精度地补偿船舶升沉运动的目的。

3　实验结果与分析

强化学习算法作为一种复杂的机器学习算法，对仿真的硬件设备有着较高的要求，本文使用的工作站具体配置为：操作系统（Windows 11），处理器（I9–10980XE @ 3.00 GHz），内存（32.0 GB），显卡（NVIDIA GeForce RTX 3070），显存（8 GB），编程语言（MATLAB 2021b）。

3.1　船舶运动解析

在风、浪、涌的复杂环境下，海面上存在不同等级的海况，海况等级越高则波浪幅值越大，周期越长。通过描述不同海况等级下船舶升沉运动轨迹，验证所设计的控制算法的适用性。本文船舶升沉运动数据由ANSYS软件中的AQWA模块生成，该模块能够计算各种结构流体动力学特性和波浪响应，可以生成船舶6个自由度下的位移、速度和加速度。本文所采用工程船模型的参数如下：长109.0 m，宽24.0 m，深6.5 m，质量7 292.96 t，吃水5.7 m，船的行驶速度为0，浪向角180°。由于6级海况下的波浪幅值最大，严重影响风机安装，补偿控制的挑战也最大，故本文重点研究6级海况下的补偿控制效果。

3.2　改进TD3算法在6级海况下的补偿控制效果

3.2.1　网络结构设计前后仿真对比

分别将TanH函数和TanH′函数作为Actor网络的输出层激活函数进行训练，将6级海况以及变海况下的船舶运动作为测试数据，结果发现，使用TanH′激活函数的TD3算法补偿效果比使用TanH激活函数好，两者的船舶升沉运动补偿误差对比如图6、7所示。

从图6中可以看出，使用TanH函数和使用TanH′函数的补偿误差都接近于0，但是在30～40 s内，TanH函数的补偿误差出现峰值，而TanH函数仍然保持较低的补偿误差，在70 s附近也出现同样的现象。总的来说，使用TanH′函数的补偿效果更优。从图7来看，在前160 s，两种激活函数的补偿误差相差不大，但是随着船舶运动更剧烈，船舶升沉运动的幅值超过1 m时，使用TanH函数的补偿误差出现了多次尖峰，补偿效果差，而使用TanH′函数的补偿误差一直能保持在0附近，补偿效果更稳定，补偿精度更高。经计算，使用

TanH′激活函数时均方根误差(RMSE)分别为0.001 9、0.129 6,正规化均方根误差(NRMSE)分别为0.000 8、0.035 4,补偿效率η分别为99.65%、98.05%,均高于使用TanH激活函数的补偿效率。因此,为获得更好的补偿效果,Actor网络的输出层选用更适合船舶升沉运动补偿系统的TanH′激活函数。

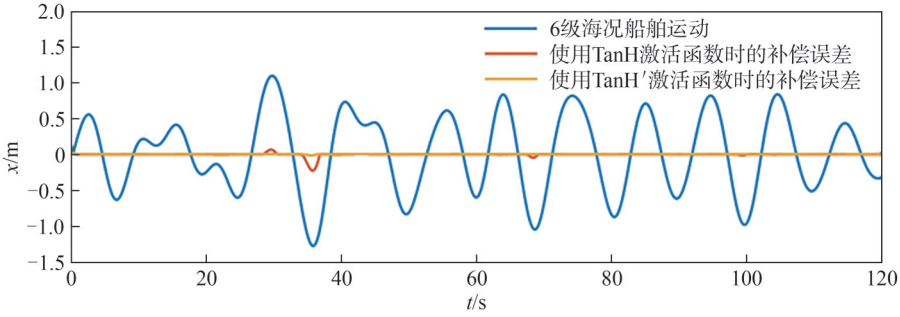


图6 6级海况下不同Actor网络激活函数的补偿误差

Fig. 6 Compensation error curves of different Actor network activation functions under sea conditions of level 6

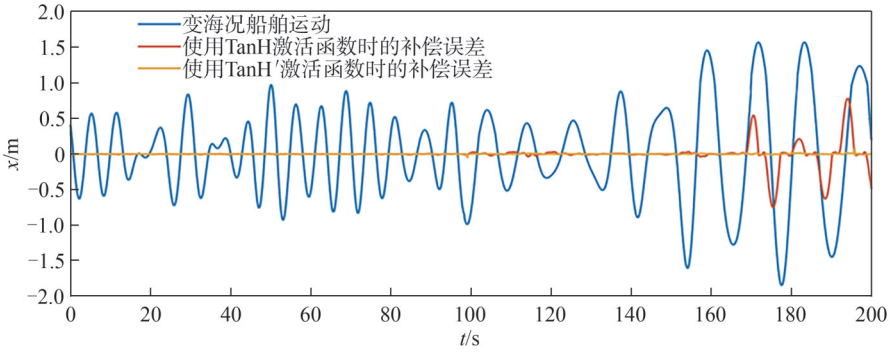


图7 变海况下不同Actor网络激活函数的补偿误差曲线

Fig. 7 Compensation error curves of activation functions of different Actor networks under variable sea conditions

3.2.2 *奖励函数设计前后仿真对比*

分别以线性奖励函数、正态奖励函数和组合奖励函数作为TD3算法船舶升沉运动补偿控制算法的奖励函数,对6级海况及变海况下的船舶升沉运动进行训练,结果分别如图8、9所示。

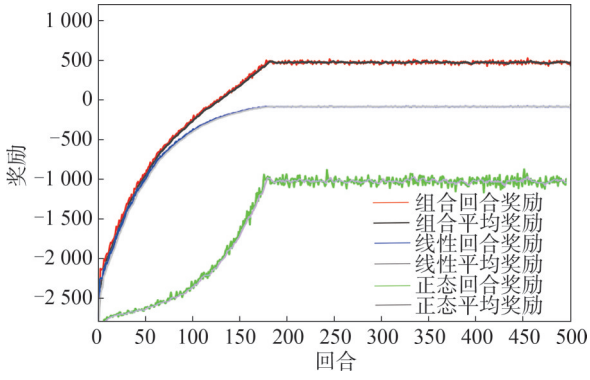


图8 6级海况下奖励函数的训练过程比较

Fig. 8 Comparison of the training process of the reward function under level 6 sea conditions

从图8可以看出:使用线性函数作为奖励函数时,前期(50回合以前)获得的奖励上升速度较快,后期(100回合之后)较慢;使用正态分布函数作为奖励函数时,后期(100回合之后)获得的奖励上升速度较快,前期较慢(50回合以前)。分析原理和训练过程发现,使用线性函数作为奖励函数时,训练后期由于斜率为定值,故训练速度下降;而正态分布奖励函数的前期误差巨大时,奖励值接近于0,并且斜率很小不利于区分动作的优劣,因此前期训练速度较慢;而使用组合奖励函数的训练过程既具备了线性奖励的前期较快收敛速度,也具备了正态分布奖励的后期收敛速度快的优点,并且能搜索到更优的动作值。从图9可以看出,使用正态函数作为奖励函数时,在前500个回合,奖励值小于−500,训练过程不稳定,并且

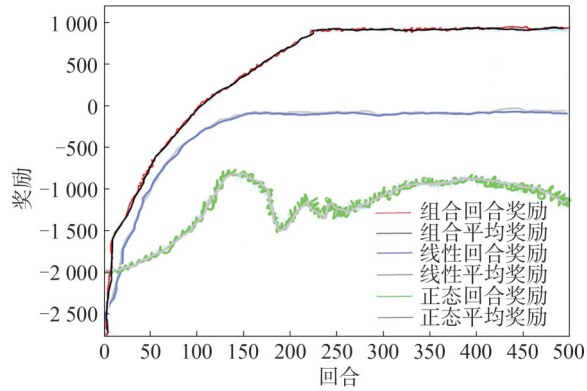


图9 变海况下奖励函数的训练过程比较

Fig. 9 Comparison of the training process of the reward function under variable sea conditions

没有收敛。使用线性奖励函数和组合奖励函数时，获得的奖励都有上升，都存在收敛的现象，但是组合奖励函数的上升速度更快，最终收敛值更高，因此，组合奖励函数的训练效果是 3 种奖励函数中效果最好的一种。3 种奖励函数在 6 级海况以及变海况下智能体训练后的测试效果如图 10、11 所示。从图 10 可以看出：在幅值更高、海况更恶劣的 6 级海况下，线性奖励函数补偿误差的幅值明显高于其他两种奖励函数，补偿效果不佳；对于其他两种奖励函数，都能够良好地实现控制跟踪，但组合奖励函数的补偿误差的幅值更小，补偿效果更优。因此，组合奖励函数更有利于 TD3 算法的补偿控制。从图 11 可看出，采用组合奖励函数的 TD3 算法补偿误差更接近于 0，远远大于线性奖励函数和正态奖励函数的补偿精度，能够应对变海况的复杂情况，并进一步改善在恶劣海况下的补偿。

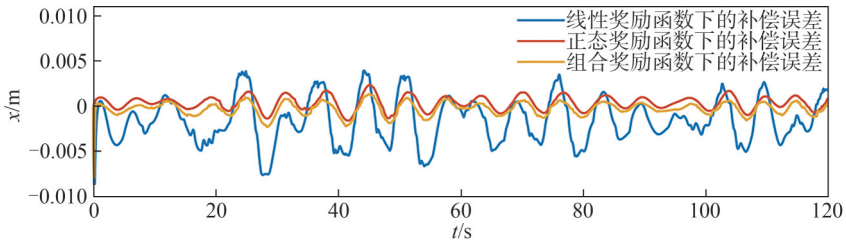


图 10　6 级海况下 3 种奖励函数训练之后的补偿误差

Fig. 10　Compensation errors after training of three reward functions under sea conditions of level 6

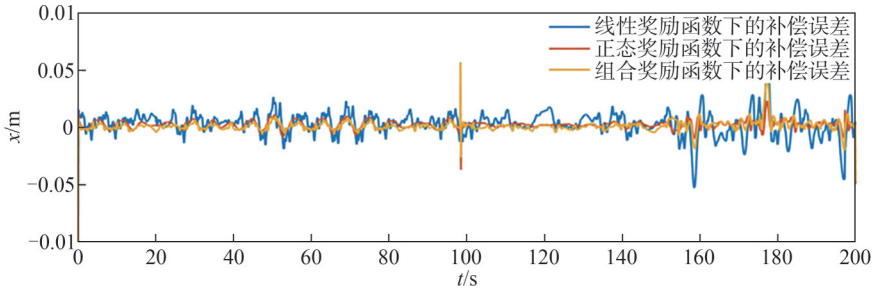


图 11　变海况下 3 种奖励函数训练之后的补偿误差

Fig. 11　Compensation errors after training of three reward functions under variable sea conditions

经计算，在 6 级海况下，由线性奖励函数进行训练的测试结果分别为：RMSE 为 0.003 0，NRMSE 为 0.002 8，补偿效率 η 高达为 98.66%；由正态奖励函数进行训练的测试结果为：RMSE 为 0.000 9，NRMSE 为 0.000 8，补偿效率 η 为 99.61%；由组合奖励函数进行训练的测试结果为：RMSE 为 0.000 8，NRMSE 为 0.000 7，补偿效率 η 为 99.64%。在变海况下，由线性奖励函数进行训练测试结果分别为：RMSE 为 0.014 2，NRMSE 为 0.003 9，补偿效率 η 为 99.78%；由正态奖励函数进行训练的测试结果为：RMSE 为 0.010 3，NRMSE 为 0.002 8，补偿效率 η 为 99.90%；由组合奖励函数进行训练的测试结果为：RMSE 为 0.010 2，NRMSE 为 0.002 8，补偿效率 η 为 99.92%。

从上述指标中可以看出：在 6 级海况及变海况下，线性奖励函数的 RMSE、NRMSE 是最高的，补偿效率最低；组合奖励函数的各项指标均优于另外两种奖励函数，其中，RMSE 分别降低 0.002 3、0.003 9，NRMSE 分别降低了 0.002 1、0.001 1，补偿效率分别提升了 0.98%、0.14%，证实了所设计控制算法的有效性。因此，采用该奖励函数进行多海况环境下船舶升沉运动补偿系统的强化学习训练。

3.2.3　动作噪声设计前后的仿真对比

将高斯噪声、OU 噪声的噪声模型标准差和标准差的衰减率分别设为 0.6 和 1.0×10^{-5} 进行训练比较。图 12、13 分别为使用高斯噪声和 OU 噪声对 6 级海况以及变海况下的船舶运动进行训练后的补偿误差。由图 12、13 可见，在同样的奖励函数和噪声参数下，使用 OU 噪声训练时，最终收敛时获得的奖励要大于使用高斯噪声时因此，验证了上文所说的 OU 噪声可以相对探索地更远，且适用于本文的船舶升沉运动补偿系统。

在 6 级海况及变海况下，计算了两种动作噪声下的各种指标，由此来评估补偿效果。由高斯噪声作为动作噪声进行训练的测试结果分别为：RMSE 为 0.000 8、0.013 1，NRMSE 为 0.000 7、0.003 6，补偿效率 η 为 99.64%、99.82%；由 OU 噪声作为动作噪声进行训练的测试结果为：RMSE 为 0.000 6、0.001 3，NRMSE 为 0.000 6、0.003 6，补偿效率 η 为 99.73%、99.83%。从上述指标来看，OU 噪声的 RMSE 和 NRMSE 更低，补偿效率更高。因此，采用 OU 噪声作为动作噪声进行船舶升沉运动补偿系统的强化学习训练。

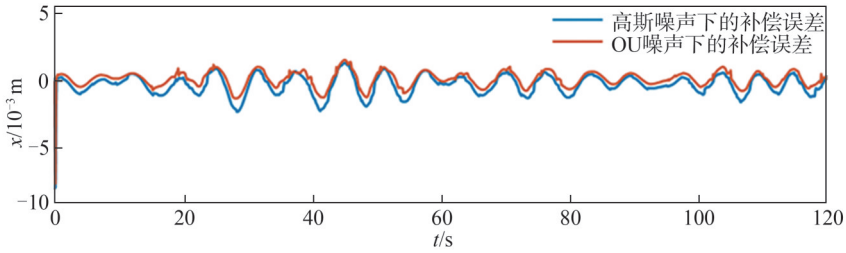


图 12 6级海况下高斯噪声和OU噪声下训练之后的补偿误差

Fig. 12 Compensation errors after training under Gaussian noise and OU noise in sea conditions of level 6

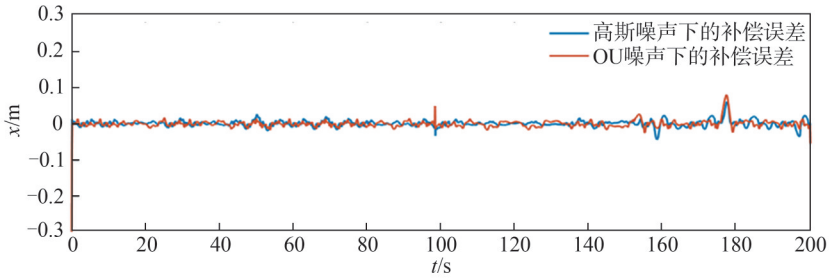


图 13 变海况下高斯噪声和OU噪声下训练之后的补偿误差

Fig. 13 Compensation errors after training under Gaussian noise and OU noise under variable sea conditions

3.3 改进的TD3算法泛化性测试

基于以上分析，在TD3算法的基础上，改进了网络结构，并采用组合奖励函数和OU动作噪声进行多海况下船舶升沉运动补偿控制。为了验证所设计算法的适用性，描述多海况环境下船舶升沉运动，如图 14 所示，随着海况等级的增大，船舶运动曲线的幅值也随之增大，周期明显增长，船舶运动更剧烈。变海况下的船舶升沉运动波形如图 15 所示。由图 15 可以看出：在 0～100 s 时，船舶运动的波形为 5 级海况，其波形幅值更小，最高仅为 1 m；在 100 s 时，由 5 级海况变化到 6 级海况，随着海况等级的增大，船舶升沉运动的幅值也随之增大，且频率也有所降低。不规则波下的船舶升沉运动变化更剧烈，海况等级不稳定，同时也增加了补偿控制的难度。因此，要用多种海况等级下的船舶运动验证改进的 TD3 算法的泛化性。

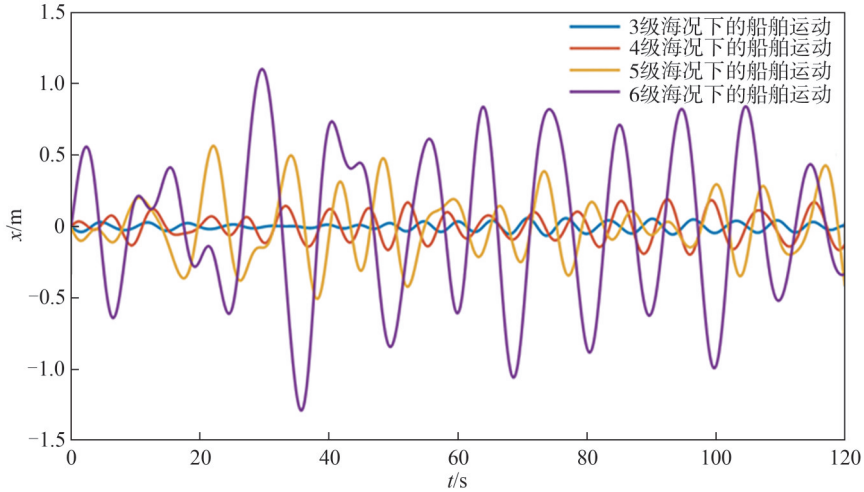


图 14 3～6级多海况不规则波下的船舶升沉运动

Fig. 14 Heave motion of ship under irregular waves with multiple sea conditions of levels 3 to 6

为测试所设计的改进的 TD3 算法的泛化性，将训练好的强化学习智能体分别对 3～6 级海况以及变海况下的船舶升沉运动进行补偿控制，并与传统的 TD3 算法下的船舶升沉运动补偿控制效果进行对比。多海况下船舶运动的补偿仿真实验结果如图 16 所示。从图 16 可以看出，对强化学习 TD3 算法进行合理设计后的智能体经训练后，对 3～6 级海况以及变海况下的船舶运动都能够进行较好地补偿。对于 3～6 级海况以及变海况下的船舶运动，改进的 TD3 算法的补偿精度要高于传统 TD3 算法，可见改进的奖励函数和 OU 动作噪声在多海况下的补偿有效性。其中：虽然 6 级海况的幅值超过 −1～1 m 的范

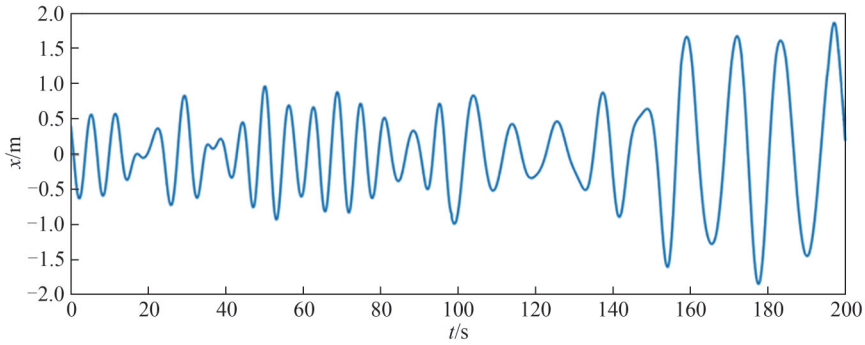
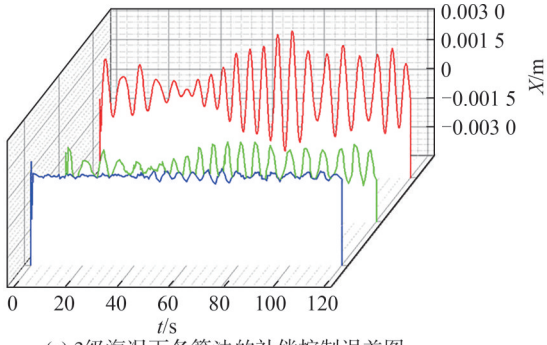


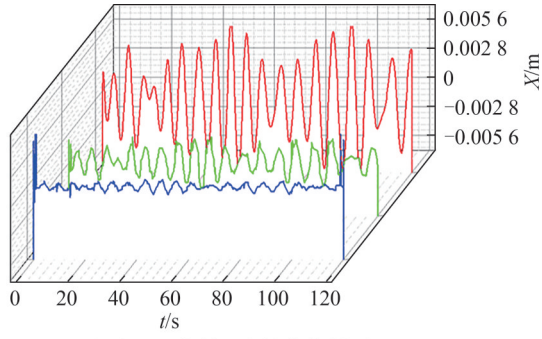
图 15　变海况下的船舶升沉运动

Fig. 15　Heave motion of ship under variable sea conditions

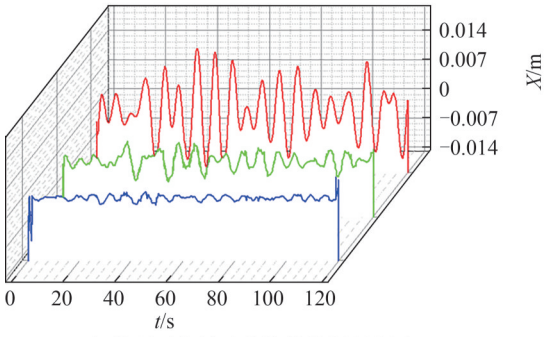
围，但在改进的激活函数作用下，其补偿精度远大于传统 TD3 算法，进一步改善了在恶劣海况下的补偿效果；从图形上看，和基于 PSO 优化反步法[30–31]控制仿真实验相比，强化学习控制仿真实验的补偿误差更小。对多海况下强化学习控制仿真和 PSO 优化反步法控制仿真的结果进行数据分析，计算均方根误差 RMSE、正规化的均方根误差 NRMSE 和补偿效率 η，结果如表 1 所示。



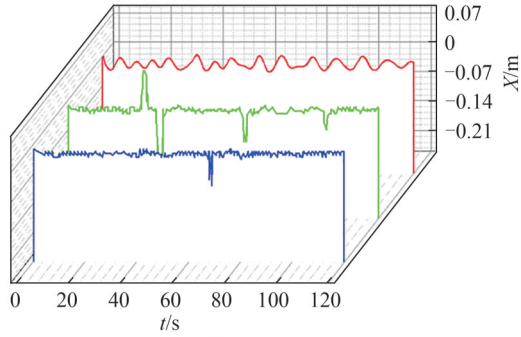
(a) 3级海况下各算法的补偿控制误差图



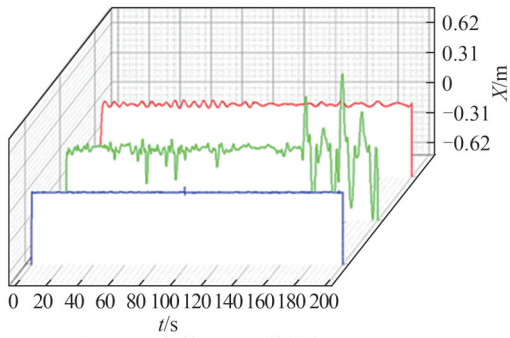
(b) 4级海况下各算法的补偿控制误差图



(c) 5级海况下各算法的补偿控制误差图



(d) 6级海况下各算法的补偿控制误差图



(e) 变海况下各算法的补偿控制误差图

PSO优化反步法控制的补偿误差　　传统的TD3算法的补偿误差　　改进的TD3算法的补偿误差

图 16　基于强化学习的多海况不规则波下船舶升沉运动的补偿仿真实验结果

Fig. 16　Compensation simulation experiment results of ship heave motion under irregular waves in multiple sea conditions based on reinforcement learning

表1 基于强化学习的多海况不规则波下船舶升沉运动的补偿仿真实验评价

Tab. 1 Evaluation of compensation simulation experiments for ship heave motion under irregular waves in multiple sea conditions based on reinforcement learning

海况等级	PSO优化反步法控制			传统TD3算法控制			改进的TD3算法控制		
	RMSE	NRMSE	η/%	RMSE	NRMSE	η/%	RMSE	NRMSE	η/%
3	0.001 3	0.010 3	95.20	0.001 7	0.013 8	93.55	0.000 2	0.001 8	99.18
4	0.003 5	0.009 2	96.17	0.002 0	0.005 2	97.84	0.000 3	0.000 9	99.64
5	0.006 2	0.005 8	97.19	0.002 4	0.002 2	98.91	0.000 6	0.000 6	99.73
6	0.009 9	0.004 1	98.17	0.030 0	0.012 6	94.43	0.001 0	0.000 4	99.81
变海况	0.026 1	0.007 1	96.25	0.172 6	0.047 1	97.68	0.009 5	0.002 6	99.95

从表1可以看出，在3～6级多海况下，改进的TD3算法控制的RMSE和NRMSE值均最小，与传统TD3算法控制相比补偿效率提高约0.82%～5.80%，与PSO优化反步法控制相比补偿效率提高约1.64%～5.99%。因此，在3～6级多海况以及变海况下，改进的TD3算法对船舶升沉运动补偿的补偿效果有显著提高，具有很好的泛化性。

4 结 论

针对复杂海况环境下的船舶升沉运动情况，采用强化学习算法设计适合船舶升沉运动补偿控制策略。首先，搭建了船舶升沉运动补偿系统的强化学习环境；然后，通过合理设计网络结构、奖励函数和动作噪声来提升TD3算法的训练和测试效率。仿真结果显示，采用改进的TD3算法强化学习训练过程，在3～6级海况以及变海况下，补偿效率均达到99.18%以上。最后，通过将训练好的模型用于3～6级多海况下的船舶运动仿真实验，证明了所设计方法在船舶升沉运动补偿控制的补偿效果良好，其3种评价指标均优于基于PSO优化反步法控制和传统TD3算法控制，泛化性能良好。

本文虽然取得了一定的成果，但将方法应用于实际平台还存在困难，硬件设备的耦合关系需要进行分析和验证，对此将进行进一步的应用研究。

参考文献：

[1] Song Yuzhang. Analysis on the present situation and prospect of offshore wind power generation[J]. China Plant Engineering, 2021(21): 258–259. [宋育章. 浅析海上风力发电的现状及展望[J]. 中国设备工程, 2021(21): 258–259.]

[2] Luo Chengxian. Current status of offshore wind power in the world[J]. Sino–Global Energy, 2019, 24(2): 22–27. [罗承先. 世界海上风力发电现状[J]. 中外能源, 2019, 24(2): 22–27.]

[3] Li Shizhu, Chen Zhijie, Zhuang Jiemin, et al. Contents and maintenance and repair of offshore wind power[J]. Electric Engineering, 2022(6): 61–63. [黎石竹, 陈智杰, 庄杰敏, 等. 海上风电维护检修的内容及维检[J]. 电工技术, 2022(6): 61–63.]

[4] Qi Lei, Mei Song, Chen Shigao, et al. Application of offshore wind power in offshore oil field and research on risk factors of offshore wind power project development[J]. Modern Chemical Research, 2023(5): 122–124. [祁雷, 梅嵩, 陈诗高, 等. 海上风电在海上石油领域的应用及海上风电项目开发风险因素应对研究[J]. 当代化工研究, 2023(5): 122–124.]

[5] Wang Xin, Cui Yakun, Xue Haibo, et al. Overview of offshore wind power platform operation and maintenance boarding system at home and abroad[J]. Science and Technology & Innovation, 2019(20): 55–58. [王鑫, 崔亚昆, 薛海波, 等. 国内外海上风电平台运维登靠系统概况[J]. 科技与创新, 2019(20): 55–58.]

[6] Yin Li, Qiao Dongsheng, Li Binbin, et al. Modeling and controller design of an offshore wind service operation vessel with parallel active motion compensated gangway[J]. Ocean Engineering, 2022, 266: 112999.

[7] Salah–Eddine M, Sadki S, Bensassi B. Microcontroller based data acquisition and system identification of a DC servo motor using ARX, ARMAX, OE, and BJ models[J]. Advances in Science, Technology and Engineering Systems Journal, 2020, 5(6): 507–513.

[8] Zhai Fugang, Yin Yanbin, Li Chao, et al. Stiffness modeling and feedforward control of servo electric cylinder drive system[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(2): 442–449. [翟富刚, 尹燕斌, 李超, 等. 伺服电动缸传动系统刚度建模与前馈控制[J]. 吉林大学学报(工学版), 2021, 51(2): 442–449.]

[9] Lyu Congxin, Wang Bo, Chen Jingbo, et al. Review and prospect of control strategies for permanent magnet synchronous motors[J]. Electric Drive Automation, 2022, 44(4): 1–10. [吕从鑫, 汪波, 陈静波, 等. 永磁同步电机控制策略综述与展望[J]. 电气传动自动化, 2022, 44(4): 1–10.]

[10] Kang Teng. Research on longitudinal PID control of intelli-

gent driving vehicle based on genetic algorithm[J]. Automotive Digest,2022(10):52–56. [康腾. 基于遗传算法的智能驾驶车辆纵向PID控制研究[J]. 汽车文摘,2022(10):52–56.]

[11] Wang Lufeng. Study of PID control system based on DC traction motor speed control system for electric vehicle[J]. Automobile Applied Technology,2020,45(10):106–108.[王露峰. 基于PID控制的电动汽车直流驱动电机调速研究[J]. 汽车实用技术,2020,45(10):106–108.]

[12] Mei Tianxiang, Yang Yi, Chen Jianbo, et al. Design of heave compensation control system based on variable parameter PID algorithm[C]//Proceedings of the 2018 Chinese Control and Decision Conference(CCDC). Shenyang: IEEE, 2018:825–829.

[13] Woodacre J K, Bauer R J, Irani R. Hydraulic valve–based active-heave compensation using a model–predictive controller with non–linear valve compensations[J]. Ocean Engineering,2018,152:47–56.

[14] Ma Changli, Liu Cong, Ma Ben. Research on wave heave simulation and adaptive compensation strategy based on disturbance observer[J]. Chinese Journal of Engineering Design,2019,26(6):728–735. [马长李,刘聪,马奔. 基于干扰观测器的波浪升沉模拟及自适应补偿策略研究[J]. 工程设计学报,2019,26(6):728–735.]

[15] Zhang Qin, Wang Xingyue, Zhang Zhengzhong, et al. Wave heave compensation based on an optimized backstepping control method[J]. China Ocean Engineering, 2022, 36(6): 959–968.

[16] Cai Yunfei, Zheng Shutao, Liu Weitian, et al. Sliding–mode control of ship-mounted Stewart platforms for wave compensation using velocity feedforward[J]. Ocean Engineering,2021,236:109477.

[17] Yang Qiming, Zhu Yan, Zhang Jiandong, et al. UAV air combat autonomous maneuver decision based on DDPG algorithm[C]//Proceedings of the 2019 IEEE 15th International Conference on Control and Automation(ICCA). Edinburgh:IEEE,2019:37–42.

[18] Zinage S, Somayajula A. Deep reinforcement learning based controller for active heave compensation[J]. IFAC-PapersOnLine,2021,54(16):161–167.

[19] Chu Zhenzhong, Sun Bo, Zhu Daqi, et al. Motion control of unmanned underwater vehicles *via* deep imitation reinforcement learning algorithm[J]. IET Intelligent Transport Systems,2020,14(7):764–774.

[20] Zhang Zhibin, Li Xinhong, An Jiping, et al. Model-free attitude control of spacecraft based on PID–guide TD3 algorithm[J]. International Journal of Aerospace Engineering, 2020,2020(1):8874619.

[21] Qin Zhihui, Li Ning, Liu Xiaotong, et al. Overview of research on model-free reinforcement learning[J]. Computer Science,2021,48(3):180–187. [秦智慧,李宁,刘晓彤,等. 无模型强化学习研究综述[J]. 计算机科学,2021,48(3):180–187.]

[22] Kong Jiaxiang, Wang Boyang, Liu Zhuyong, et al. Rigid-flexible dynamic modeling and simulation of Stewart platform spacecraft[J]. Journal of Dynamics and Control,2022, 20(6):76–84. [孔嘉祥,王博洋,刘铸永,等. 带Stewart平台的航天器刚柔耦合动力学建模与仿真分析[J]. 动力学与控制学报,2022,20(6):76–84.]

[23] Cai Yunfei, Zheng Shutao, Liu Weitian, et al. Sliding-mode control of ship–mounted Stewart platforms for wave compensation using velocity feedforward[J]. Ocean Engineering,2021,236:109477.

[24] Teethi T I, Lu Hu, Min Huan, et al. An improved reinforcement learning method for drone avoidance decision control[J]. Journal of Detection & Control,2022,44(3):68–73. [Tajmihir Islam Teethi,卢虎,闵欢,等. 基于改进强化学习的无人机规避决策控制算法[J]. 探测与控制学报,2022,44(3):68–73.]

[25] Chen Lingyu, Zheng Jieji, He Aihua, et al. Design of safety integrated servo motor drive module[J]. Optics and Precision Engineering,2023,31(1):42–56. [陈凌宇,郑杰基,何爱华,等. 安全集成伺服电机驱动模块设计[J]. 光学 精密工程,2023,31(1):42–56.]

[26] Zhu Mengmei, Zhou Guangxu, Song Ningran, et al. Servo motor position and speed detection based on serial absolute encoder[J]. Micromotors,2021,54(1):63–67. [朱孟美,周广旭,宋宁冉,等. 基于串行绝对式编码器的伺服电机位置及转速检测[J]. 微电机,2021,54(1):63–67.]

[27] Zhang Qingbo. Study on the stiffness modeling and control method of planetary roller screw servo-electric cylinder system[D]. Harbin: Northeast Forestry University,2022. [张庆博. 行星滚柱丝杠伺服电动缸系统刚度建模及控制方法研究[D]. 哈尔滨:东北林业大学,2022.]

[28] Chen Chao, Zhao Shengdun, Cui Minchao, et al. Study status and developing trend of electric cylinder[J]. Journal of Mechanical Transmission,2015,39(3):181–186.[陈超,赵升吨,崔敏超,等. 电动缸的研究现状与发展趋势[J]. 机械传动,2015,39(3):181–186.]

[29] Joshi T, Makker S, Kodamana H, et al. Twin actor twin de-

layed deep deterministic policy gradient (TATD3) learning for batch process control[J]. Computers & Chemical Engineering,2021,155:107527.

[30] Geng Zhiwei,Wang Shuang,Yang Shuangyi.Tracking control of wheeled mobile robot based on backstepping and hierarchical sliding mode control[J].Manufacturing Automation,2022,44(6):109–112.[耿志伟,王爽,杨双义.基于反步法和分层滑模控制的轮式移动机器人轨迹跟踪[J].制造业自动化,2022,44(6):109–112.]

[31] Lou Peng,Song Jianqiao,Yin Peiwu,et al.Control rate design of high-altitude vehicle power motor based on ADRC method [J].Small & Special Electrical Machines,2020,48(10):51–53.[娄鹏,宋剑桥,殷佩舞,等.基于自抗扰反步法的高空飞行器动力电机控制率设计[J].微特电机,2020,48(10):51–53.]

Research on Ship Heave Motion Compensation Control Under Complex Sea State Environment Based on Improved Reinforcement Learning

ZHANG Qin, *ZHOU Jingyi*, *WANG Xingyue*, *HU Xiong*[*]

(School of Logistics Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract:

Objective In the vast expanse of the boundless sea, the capricious and ever−shifting interplay of wind and waves often presents unpredictable challenges to maritime operations. Particularly in the midst of an ever−changing marine environment, ships frequently encounter powerful gusts and tumultuous swells, whose restless and complex movements not only pose a significant threat to the secure installation of offshore wind turbine units but also introduce considerable uncertainty to maritime operations and personnel transfers. These destabilizing elements result in operational delays, equipment damage, or even harm to personnel, necessitating utmost emphasis on dependability, safety, and stability in offshore operations. Thus, in the quest to address these concerns and bolster the efficiency and safety of maritime endeavors, researchers actively explore and pioneer diverse techniques aimed at compensating for the vertical motion of vessels. The underlying objective of these techniques lies in precisely governing vessel movements and counteracting heave provoked by wind and waves, ensuring the steadfastness and security of offshore operations. However, despite the immense potential and value that this technology holds, it encounters significant challenges in practical application. The inherent complexity and inscrutability of vessel systems introduce obstacles in modeling and control. In addition, the ability to swiftly and accurately adjust compensation strategies during actual operations to accommodate ever-changing oceanic conditions remains an exigent conundrum in need of resolution. Therefore, this study presents a compensation control method for ship heave under complex sea conditions using an improved reinforcement learning approach.

Methods This novel method imparts fresh insights into addressing heavy compensation in offshore operations and heralds a new trajectory for the evolution of future offshore operation technologies. The study employs principles of mechanics to furnish a comprehensive model of the wave compensation system, encompassing servo drives, servo motors, encoders, and hydraulic cylinders. This model serves a dual purpose: it simulates various performance indicators of the vessel heave compensation system and functions as the training environment for reinforcement learning. With the mechanical model of the vessel heave compensation system firmly established, the study applies the Markov decision process to determine the agent's strategy and reward mechanism. Within this process, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm assumes a central role as the core control strategy. The TD3 algorithm approximates the value function and policy by harnessing deep neural networks, equipping it to tackle complex and nonlinear sea condition challenges. Aligned with the uncertainty and complexity entailed by the maritime milieu, this study specifically fine–tunes the output layer of the Actor network by amplifying the amplitude of the TanH function. This adjustment endows the Actor network with the ability to generate more versatile and extensive control actions, adeptly adapting to the capriciousness of the sea. During the training process, the study employs two independent network structures, the main network and the target network, each comprising an Actor and a Critic network, amounting to a total of six networks. Through iterative updates of these networks, the system continually learns and optimizes its control strategies, culminating in the generation of self-learning optimal control actions. The study incorporates Ornstein–Uhlenbeck (OU) action noise into the target policy to enhance the adaptability of the agent amidst complex sea conditions. OU noise is a specialized form of stochastic process that engenders smooth and correlated random oscillations over continuous time, making it particularly suited for exploration within continuous state spaces. In reinforcement learning environments, the inclusion of OU noise aids the agent in broader exploration during the nascent stages of training, facilitating the discovery of potentially advantageous state-action pairs that augment task completion. In addition, the study devises a reward function that integrates linear and Gaussian components to guide the agent's learning and decision-making

processes. This composite reward function not only reflects the quality of current state–action pairs but also incorporates predictions and evaluations of future states. This design augments the agent's understanding of task objectives and enables it to formulate effective strategies during protracted learning processes. By adopting this approach, the agent gradually adapts to the demands of reinforcement learning tasks amid dynamically shifting sea conditions, evading the pitfalls of local optima. Even in the face of variable and complicated sea conditions, the agent continually optimizes its compensation strategies through self–learning and adaptive adjustments, heightening accuracy in compensation and assuring the secure installation of offshore wind turbine units. In turn, this fortifies the bastion of offshore operations and safeguards personnel transfers.

Results and Discussions Simulation experiments demonstrate the outstanding effectiveness of the improved TD3 algorithm in compensation control when confronted with adverse and complex sea conditions. The study applies the trained model to a simulated vessel heave compensation system, subjecting it to a range of complex sea conditions, spanning sea states classified from level three to level six, as well as varying marine environments. In these diversified test scenarios, the improved TD3 algorithm exhibits remarkable adaptability and stability. Particularly noteworthy is its exceptional compensation efficiency, attaining a maximum of 99.95%. This accomplishment highlights the algorithm's superb compensation control capabilities, furnishing a high degree of safety to the installation of offshore wind turbine units. This algorithm surpasses step control methods optimized through particle swarm optimization and outperforms traditional TD3 reinforcement learning methodologies. In addition, the improved TD3 algorithm boasts favorable generalization capabilities, indicative of its capacity to swiftly adapt and generate effective compensation control strategies even in untrained and novel sea conditions.

Conclusions Therefore, the improved TD3 algorithm opens up vast potential and application value in the field of vessel heave compensation, furnishing robust technical support to the installation of offshore wind turbine units and the safety of offshore operations. Through its complex melding of mechanics, reinforcement learning, and innovative control strategies, this study advances the creation of an advanced and dependable system for maritime operations, bound to reshape the landscape of offshore endeavors.

Key words: complex sea conditions; heave motion of ship; compensation control system; TD3 reinforcement learning

（编辑　张　琼）

引用格式: Zhang Qin,Zhou Jingyi,Wang Xingyue,et al.Research on ship heave motion compensation control under complex sea state environment based on improved reinforcement learning[J].Advanced Engineering Sciences,2025,57(4):123–137.
[张琴,周静宜,王星月,等.基于改进强化学习的复杂海况下船舶升沉补偿控制研究[J].工程科学与技术,2025,57(4):123–137.]