Journal of Physics:
Conference Series

PAPER • OPEN ACCESS

TD3 compensated control for uncertain AUV trajectory tracking

To cite this article: Zhiqun Liu *et al* 2025 *J. Phys.: Conf. Ser.* **2999** 012016

TD3 compensated control for uncertain AUV trajectory tracking

Zhiqun Liu, Chaoyang Chen*, Yi Huang, Dan Yang

Hunan University of Science and Technology, Taoyuan Road, Yuhu District, Xiangtan City,Hunan Province, China

E-mail: cychen@hnust.edu.cn

Abstract. This paper investigates the trajectory tracking problem of autonomous underwa-ter vehicles (AUV).In order to solve the problems of underdrive,non-completeness,strong cou-pling and modeling errors with random perturbations that may be encountered by underwater autonomous vehicles,in real operations. The ideal model, i.e.,the nominal model,is firstly es-tablished based on the non-completely underdriven strongly coupled AUV dynamics,and the nominal model control rate is designed based on the backstepping method and Lyapunov function.Deep reinforcement learning techniques are then utilized to deal with modeling errors and stochastic perturbations of the system model in real operations,and a compensation method based on the Twin Delayed Deep Deterministic policy gradient(TD3) is proposed.Finally,numerical simulation is carried out,and the simulation results show that the control method proposedin this paper effectively reduces the effects of modeling errors and random perturbations,on the accurate trajectory tracking of AUV.
Keywords: TD3,autonomous underwater vehicl,trajectory tracking,virtual control ra-tio,compensating control

1. Introduction

The development of marine resources is increasingly becoming the focus of world attention[1,2].Autonomous underwater vehicles (AUVs) have garnered significant attention from scholars in fields such as nautical charting, energy exploration,military reconnaissance, and marine environment detection, as they serve as an emerging tool for ocean exploration[3-6].The study of trajectory tracking of AUVs is indispensable to the above fields and is one of the hot issues in AUV control research in recent years[7].In the real trajectory tracking engineering applications of AUVs,there are problems such as underdriving,non-completeness, strong coupling, and modeling errors and random perturbations[8,9],which bring great challenges to the research work,and many scholars have studied for this purpose.[10] explored the trajectory tracking problem of robots, emphasizing the development of tracking controllers

Journal of Physics: Conference Series **2999** (2025) 012016 doi:10.1088/1742-6596/2999/1/012016

that achieve global asymptotic stabilization. Building on this foundation,[11] proposed an adaptive controller for guiding a unicycle type mobile robot during trajectory tracking considering disturbances under different loads.Considering the physical limitations of the actuator,[12] Solving the input-constrained trajectory tracking problem for incompletely mobile robots.However,all of the above literatures studied are trajectory tracking for general-purpose mobile robots,which are difficult to be directly applied to AUVs.Consequently,[13,14] proposed a Lyapunov-based MPC framework to improve AUV trajectory tracking performance.[15] introduced a Liapunov-based backstepping control framework to mitigate the control complexity associated with autonomous underwater vehicles (AUVs). To address the uncertainties inherent in AUV dynamics,[16-18] used the sliding mode method,which can effectively accomplish the tracking task under the inaccuracy of the system model parameters with random perturbations,but the control rate is not smooth enough for practical application.

Deep reinforcement learning (DRL) has gained significant traction in intelligent robot control[19-23].[24] developed a neural network-based reinforcement learning algorithm for underactuated AUV trajectory tracking.[25,26] successfully implemented policy gradient-based RL for underwater cable tracking,[27.28] demonstrated superior AUV path tracking performance using DRL compared to traditional methods. Nevertheless, these approaches face challenges, including high training costs and convergence difficulties.

In summary,this paper presents an advanced control strategy for AUV that integrates nominal model control with the Twin Delayed Deep Deterministic policy gradient (TD3) compensation mechanism.This hybrid approach is specifically designed to address the complex and uncertain disturbances commonly encountered in practical AUV control engineering applications,such as underactuation,incompleteness,strong coupling,modeling errors,and random perturbations.The proposed control method substantially improves the trajectory tracking performance of AUVs by effectively counteracting the adverse effects of disturbances.In contrast to existing literature,the primary contributions of this paper are as follows:

1)To address the challenges posed by modeling inaccuracies and stochastic disturbances in AUV systems,this study employs the TD3 algorithm to enhance trajectory tracking control.The proposed approach effectively compensates for the adverse effects of both systematic modeling errors and random environmental perturbations,thereby significantly improving the precision and robustness of AUV trajectory tracking.By leveraging the TD3 algorithm's dual-critic architecture and delayed policy updates, the control system achieves superior performance in handling uncertainties and maintaining stability under dynamic underwater conditions.This advancement provides a reliable solution for precise AUV navigation in complex marine environments.

2)To address the inherent challenges of underactuation,system incompleteness,and strong coupling in AUV dynamics,this study proposes a hierarchical control strategy based on virtual control rates.The designed approach systematically decouples the position and velocity control loops,thereby significantly reducing the computational complexity of the control model while maintaining system stability.By implementing this hierarchical control architecture,the proposed method effectively resolves the difficulties associated with the nonlinear and coupled nature of AUV systems,enabling more efficient and precise motion control.This innovative strategy not only simplifies the controller design but also enhances the overall system performance in practical underwater applications.

2. System Modeling and Problem Formulation

AUV are usually regarded as rigid bodies with six degrees of freedom in three-dimensional space,and this motion has strong coupling properties.To simplify this coupling property, researchers usually decompose the 3D motion of an AUV into two 3-degree-of-freedom motions in the horizontal and vertical planes.Research has usually focused on the horizontal plane motion of underdriven AUVs, and the specific form of motion is shown in Figure 1.

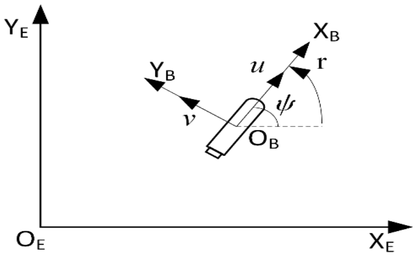

Figure 1. Horizontal coordinate system of AUV

Where $X_EO_EY_E$ is the reference coordinate system and $X_BO_BY_B$ is the body coordinate system,The dynamics of AUV motion in the horizontal plane is modeled as follows:

$$\begin{cases} \dot{\eta} = J(\eta)v \\ M\dot{v} + C(v)v + D(v)v = \tau + d \end{cases} \tag{1}$$

where $\eta = [x, y, \psi]^T$ is the position information,$v = [u, v, r]^T$ is the position information,$J(\eta)$ is the rotation operator,$M = \mathrm{diag}(M_x, M_y, M_\psi)$ is the inertia matrix,$C(v)$ is the Coriolis force matrix, $D(v) = diag(X_u, Y_v, N_r) + diag(D_u|u|, D_v|v|, D_r|r)$ is the water damping term matrix, $\tau = [F_u, F_v, F_r]^T$ is the control input and $d = [d_u, d_v, d_r]^T$ is the environmental disturbance force,included among these:

$$J(\eta) = \begin{bmatrix} cos\,\psi & -sin\,\psi & 0 \\ sin\,\psi & cos\,\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$$C(v) = \begin{bmatrix} 0 & 0 & -M_y v \\ 0 & 0 & M_x u \\ M_y v & -M_x u & 0 \end{bmatrix} \tag{3}$$

For the above parameter matrices there are $\| J(\eta) \| < \bar{J}, \bar{J} > 0$ and $J(\eta)$ nonsingular,$M > 0$ and $\| M \|$bounded.

In practical scenarios, the exact dynamics of AUVs are often not precisely known.Consequently, the parameter matrix in the AUV dynamics model(1) can be decomposed into two components:

$$\begin{cases} M = \widehat{M} + \delta M \\ D(v) = \widehat{D}(v) + \delta D(v) \\ C(v) = \hat{C}(v) + \delta C(v) \end{cases} \tag{4}$$

Remark 1:The developed AUV dynamics model comprehensively captures the intricate physical characteristics inherent to autonomous underwater vehicles, including underactuation, system incompleteness, and strong nonlinear coupling. Furthermore, the model incorporates a robust framework that explicitly accounts for both systematic modeling inaccuracies and stochastic environmental perturbations encountered during actual AUV operations. This comprehensive modeling approach not only preserves the essential dynamic properties of the AUV system but also provides a realistic representation of the challenges faced in practical underwater environments, thereby establishing a reliable foundation for subsequent controller design and performance evaluation.

3. Controller Design

Suppose the desired trajectory is $\eta_d = [x_d, y_d, \psi_d]^T$, the real trajectory is $\boldsymbol{\eta}$, the tracking error is defined as $\tilde{\eta}(t) = \eta_d(t) - \eta(t)$, and the control objective is:

$$\lim_{t\to\infty}\tilde{\eta}(t) = 0 \tag{5}$$

3.1. Nominal model controller design and stability proof

In order to reduce the control complexity,for the above AUV dynamics model,uncertainty disturbances such as modeling errors and random disturbances are ignored, thus defining the nominal model:

$$\begin{cases} \dot{\hat{\eta}} = \hat{J}(\hat{\eta})\hat{v} \\ \widehat{M}\dot{\hat{v}} + \widehat{C}(\hat{v})\hat{v} + \widehat{D}(\hat{v})\hat{v} = \hat{\tau} \end{cases} \tag{6}$$

Considering the strong coupling characteristics of its model,the virtual control rate is designed for (6),and the virtual control rate is used as the tracking target of (7) to realize the speed and position hierarchical control of AUV and reduce the control complexity.Define the virtual control quantity as:

$$\hat{v}_c = \hat{J}^{-1}(\hat{\eta})\dot{\eta}_d - \hat{J}^{-1}(\hat{\eta})\left(k_1\tilde{\hat{\eta}} + k_2\int_0^t \tilde{\hat{\eta}}dt\right) \tag{7}$$

Define the nominal control error as:

$$\tilde{\hat{\eta}}(t) = \hat{\eta}(t) - \eta_d(t) \tag{8}$$

$$\tilde{\hat{v}}(t) = \hat{v}(t) - \hat{v}_c(t) \tag{9}$$

The resulting design control volume is:

$$\hat{\tau} = \widehat{M}\left(\hat{v}_c - \hat{v} + \dot{\hat{v}}_c\right) + \widehat{C}(\hat{v})\,\hat{v} + \widehat{D}(\hat{v})\,\hat{v} \tag{10}$$

Theorem 1:Using the virtual control quantity $\hat{v}_c$ designed in (8) and the real control quantity $\hat{\tau}$ designed in (9) acting on the position system (6) and the velocity system (7),respectively,the nominal model trajectory tracking error masymptotically converges to 0,i.e. $\lim_{t\to\infty}\tilde{\hat{\eta}}(t) = 0, \lim_{t\to\infty}\tilde{\hat{v}}(t) = 0$.

Proof:Consider the Liapunov function for $V_1(t) = \frac{1}{2}\|\tilde{\hat{\eta}}(t)\|^2, V_2(t) = \frac{1}{2}\|\tilde{\hat{v}}(t)\|^2$.Derivation with respect to time yields:

$$\dot{V}_1(t) = \left(\hat{\eta}(t) - \eta_d(t)\right)\left(\hat{J}(\hat{\eta}(t))\,\hat{v}(t) - \dot{\eta}_d(t)\right)$$

$$\dot{V}_2(t) = \left(\hat{v}(t) - \hat{v}_c(t)\right)\left(\dot{\hat{v}}(t) - \dot{\hat{v}}_c(t)\right)$$

$$= (\hat{v}(t) - \hat{v}_c(t))\left(\widehat{M}^{-1}\left(\hat{\tau} - \widehat{C}(\hat{v})\hat{v} - \widehat{D}(\hat{v})\hat{v}\right) - \dot{\hat{v}}_c(t)\right)$$

$$= (\hat{v}(t) - \hat{v}_c(t))\left(\hat{v}_c(t) - \hat{v}(t) + \dot{\hat{v}}_c(t) - \dot{\hat{v}}_c(t)\right) = -(\hat{v}(t) - \hat{v}_c(t))^2 < 0$$

Since the above equation shows that $\lim_{t\to\infty}\tilde{\hat{v}}(t) = 0$, for any small $\lambda(T) > 0$ satisfies $t_1 > T$ when there is $\hat{v}(t_1) - \hat{v}_c(t_1) < \lambda(T)$,so let $t > T$, rewrite $\dot{V}_1(t)$ as:

$$\begin{aligned}\dot{V}_1(t) &= \tilde{\eta}(t)\cdot\left(\hat{J}(\hat{\eta}(t))(\hat{v}_c(t)+\lambda(T))-\dot{\eta}_d(t)\right)\\ &= \tilde{\eta}(t)\cdot\left(\dot{\eta}_d(t)-k_1\tilde{\eta}-k_2\int_0^t\tilde{\eta}dt+\lambda(T)-\dot{\eta}_d(t)\right)\\ &\le -k_1\|\tilde{\eta}\|^2+\|\lambda(T)\hat{J}(\hat{\eta}(t))\|+\|k_2\tilde{\eta}(t)\int_0^t\tilde{\eta}(t)dt\|\\ &\le -k_1\|\tilde{\eta}\|^2+\|\lambda(T)\|\|\hat{J}(\hat{\eta}(t))\|+\|k_2\|\|\tilde{\eta}(t)\int_0^t\tilde{\eta}(t)dt\|\end{aligned}$$

Taking $\| k_1 \|$ sufficiently large and $\| k_2 \|$, $\| \lambda(T) \|$ sufficiently small,we can derive $\dot{V}_1(t) < 0$,so the nominal control error of the nominal model controller converges asymptotically.This proves the theorem.

Remark 2: The proposed controller demonstrates robust stability, enabling the system to effectively execute tracking tasks while inherently compensating for modeling inaccuracies and environmental disturbances. This stability characteristic not only ensures reliable performance under ideal conditions but also addresses the fundamental control challenges associated with AUV dynamics, including underactuation, system incompleteness, and strong nonlinear coupling. The controller's inherent robustness effectively mitigates the impact of these complex dynamic characteristics, thereby providing a comprehensive solution to the persistent control difficulties in AUV operations. This achievement represents a significant advancement in underwater vehicle control, as it simultaneously resolves multiple technical challenges while maintaining system stability and tracking precision.

3.2. TD3 Compensator Design

To alleviate the influence of modeling errors and noise interference on the controller's performance, the design of a control compensator is incorporated when the controller is applied to the actual AUV model.TD3, a model-less reinforcement learning algorithm, requires only output data to train the intelligent body to adapt to a complex environment. Consequently,TD3 can be utilized to train the reinforcement learning intelligence as the AUV trajectory tracking control compensator,thereby mitigating the impact of modeling error and noise interference on the actual control effect.

Define the $s(k)$ and $s(k+1)$ are the state values of the AUVs at moments k and k+1, $a(k)$ is the input value of the TD3 compensator to the AUV system,$r(k)$ is the reward value of the AUV system at moment k under the action of $a(k)$,and define the difference between the two adjacent moments to be the algorithmic sampling time T.Where $s = \{xd, yd, \psi_d, x, y, \psi, u, v, r\}$ and $a = \delta\tau$ and $r = \tilde{\eta}^T Q\tilde{\eta} + a^T Ra$ .Define the action value function $Q(s(k), a(k)) = \sum_{l=k}^{\infty}\gamma^l r(l)$ with determining the action strategy $a(k) = \mu(s)$.Critic network(CN) is used to estimate $Q(s,a)$ as$Q(s,a,w)$,Action network(AN) is used to estimate $\mu(s)$ as$\mu(s_k,\theta_k)$,the specific algorithm flow is as follows:

(1) Arbitrarily select sequences from the empirical playback(s(k),a(k),r(k),s(k+1)),define the estimated output as $\hat{a}'(k+1) = \mu'(s_{k+1},\theta_k) + \text{noise}$,where noise is AN exploration noise.

(2) Let the outputs of the target CN 1 and the target CN 2 be $\hat{Q}_1'(s,a,w_{1'})$ and $\hat{Q}_2'(s,a,w_2')$,respectively,and similarly let the outputs of the CN 1 and the CN 2 be $\hat{Q}_1(s,a,w_1)$ and$\hat{Q}_2(s,a,w_2)$,respectively,and the output of the target AN be $\mu'(s,\theta)$ The TD error is thus defined as,where i is the number of iterative learning:

$$\begin{cases} \delta_{1,i} & = r(k) + \gamma \min_{j=1 \text{ or } 2} \left\{ \hat{Q}'_j\left(s(k+1), \hat{a}'(k+1), w'_j(i)\right) \right\} - \hat{Q}_1\left(s(k), \hat{a}'(k), w_1(i)\right) \\ \delta_{2,i} & = r(k) + \gamma \min_{j=1 \text{ or } 2} \left\{ \hat{Q}'_j\left(s(k+1), \hat{a}'(k+1), w'_j(i)\right) \right\} - \hat{Q}_2\left(s(k), \hat{a}'(k), w_2(i)\right) \end{cases} \tag{11}$$

(3) This results in parameter updates for CN 1 with CN 2 and AN,where $w_i^{'}$ is the parameter corresponding to the smaller of$\hat{Q}_1$,$\hat{Q}_2$:

$$\begin{cases} w_{1,i+1} & = w_{1,i} + \alpha_w \delta_{1,i} \dfrac{\partial \hat{Q}_1(s_k, \mu(s_k, \theta_k), w_{1,i})}{\partial w} \\ w_{2,i+1} & = w_{2,i} + \alpha_w \delta_{2,i} \dfrac{\partial \hat{Q}_2(s_k, \mu(s_k, \theta_k), w_{2,i})}{\partial w} \\ \theta_{i+1} & = \theta_i + \alpha_\theta \dfrac{\partial \mu(s_k, \theta_i)}{\partial \theta} \left(\dfrac{\partial \hat{Q}(s_k, \mu(s_k, \theta_k), w'_i)}{\partial \mu} \right) \end{cases} \tag{12}$$

(4) The target CN and target AN parameters are updated through a soft update mechanism, which can be mathematically expressed as follows:

$$\begin{cases} w'_{1,i+1} & = \xi w_{1,i} + (1-\xi) w'_{1,i} \\ w'_{2,i+1} & = \xi w_{2,i} + (1-\xi) w'_{2,i} \\ \theta'_{i+1} & = \xi \theta_i + (1-\xi) \theta'_i \end{cases} \tag{13}$$

(5) Increment i by 1 and return to step (1) until a predetermined number of iterations is reached or a specified stopping condition is satisfied. The specific flow of the aforementioned TD3 algorithm is illustrated in Figure 2.

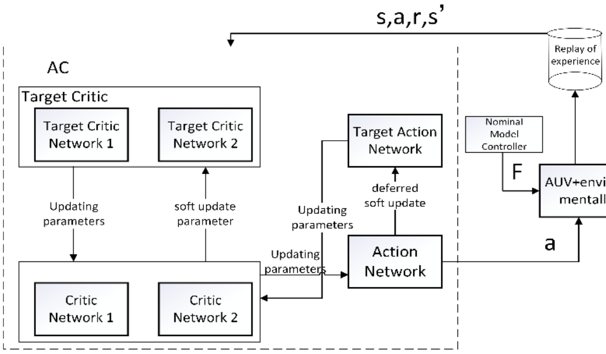

Figure 2. TD3 Compensation Control Algorithm Flowchart

Remark 3: Leveraging the inherent convergence properties of the TD3 algorithm,the compensator developed in this study is theoretically guaranteed to achieve enhanced control performance through systematic iterative refinement.In practical operational scenarios, the compensator engages in continuous cycles of data acquisition,processing,and parameter optimization.The TD3 algorithm's dual-network architecture, coupled with its delayed policy update mechanism and target policy smoothing regularization,ensures robust learning from accumulated operational data while effectively mitigating overestimation bias in value function approximation.

4. Simulation experiment and result analysis

Numerical simulations were conducted utilizing MATLAB,and the parameters of the nominal model are set to be $M_x = 216.6, M_y = 593.2, M_\psi = 28.7$, $X_u = 26.9$, $Y_v = 35.8$, $N_r = 3.5$, $N_r = 3.5$, $D_u = 100.3$, $D_v = 503.8$, $D_r = 76.9$, $d = [200\,sin(t), 600\,cos(t), 10\,sin(t) + 10\,cos(t)]^T$ .Set $\eta(0) = [0,0,0]^T$, $v(0) = [0,0,0]^T$, $\eta_d(t) = [20\,cos(0.1t), 20\,sin(0.1t), 0.1t]^T$ .Set the controller parameters as $k_1 = diag([1,1,1])$, $k_2 = diag([0.01,0.01,0.01])$;Critic network parameter learning rate $\alpha_w = 0.001$,Action network parameter learning rate α_θ=0.005.

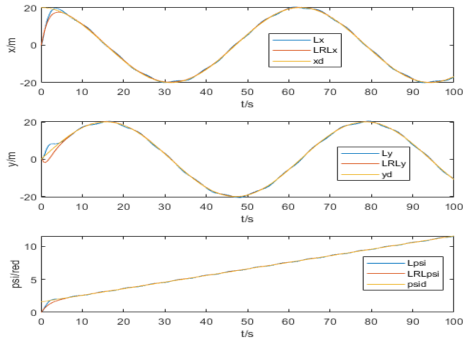

Figure 3. AUV trajectory tracking component plot

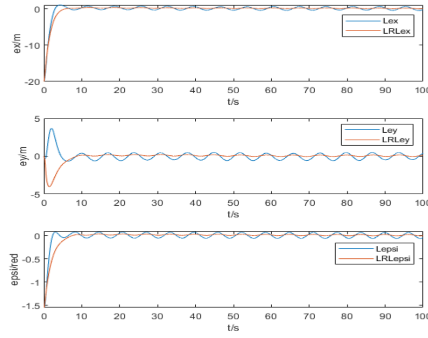

Figure 4. AUV trajectory tracking error plot

This study proposes an anti-jamming control method for AUV trajectory tracking,combining backstepping control based on a nominal model with DRL. The method tackles the complex characteristics of AUV, including underactuation, nonholonomic constraints, strong coupling, as well as the modeling errors and random perturbations that are prevalent in practical control applications. The proposed control method's effectiveness is rigorously validated through theoretical analysis and extensive simulations. Results indicate that the method significantly reduces the impact of modeling errors and random perturbations on the precise trajectory tracking of AUV, thereby enhancing their performance in complex underwater environments. This approach is shown to have substantial practical application value for improving the robustness and accuracy of AUV trajectory tracking.

Table 1. Comparison of mean square error in each direction.

Controller Type	x	y	ψ
L	0.0053	0.0049	0.0006
LRL	0.0024	0.0014	0.0002

5. Conclusions

Taking the trajectory tracking problem of AUV as the research background, this paper proposes an anti-jamming control method based on the backstepping control of nominal model and the DRL,taking into account the problems of underdriving,non-integrity,strong coupling as well as the modeling error and random perturbation encountered in the practical control engineering application of AUV.After theoretical demonstration and simulation experiments,it can be seen that the anti-jamming control method designed in this paper can effectively reduce the influence of modeling error and random perturbation on the accurate trajectory tracking of AUV under the consideration of the complex physical characteristics of underdrive,non-integrity and strong coupling of AUVs,which is of high practical application value.

6. Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62222306.

References

[1] Shiming H, Liwei K, Yanjun L and Xiang J 2020 IEEE Transactions on Industrial

Electronics 68 8464–8473
[2] Braginsky B, Baruch A and Guterman H 2020 Ocean Engineering 197 106868
[3] Yang T, Yu S and Yan Y 2019 Applied Ocean Research 82 109–116
[4] Gao Z and Guo G 2020 IEEE/CAA Journal of Automatica Sinica 7 539–545
[5] Gu N, Wang D, Peng Z and Liu L 2019 IEEE Transactions on Systems, Man, and Cybernetics: Systems 51 5105–5115
[6] Ling J, Xingqun Z, Zhiqiang H, Yiyi Y, Liming L and Shihong B 2023 Information and control 52 129–141
[7] Yanikoglu I, Gorissen B L and den Hertog D 2018 European Journal of Operational Research 277 799–813
[8] Yuqing L and Jiapeng L 2022 Control Engineering of China 29 867–872
[9] Jincun L, Junzhi Y and Ziye Z 2021 Information and Control 129–140
[10] Weiguo W, Huitang C and Yuejuan W 2001 Journal of Automation 27 326–331
[11] Martins F N, Celeste W C, Carelli R, Sarcinelli-Filho M and Bastos-Filho T F 2008 Control Engineering Practice 16 1354–1363
[12] Li Z, Deng J, Lu R, Xu Y, Bai J and Su C Y 2015 IEEE Transactions on Systems, Man, and Cybernetics: Systems 46 740–749
[13] Shen C, Shi Y and Buckham B 2017 IEEE Transactions on Industrial Electronics 65 5796–5805
[14] Hao L Y, Wang R Z, Shen C and Shi Y 2024 IEEE Transactions on Industrial Electronics 20
[15] Repoulias F and Papadopoulos E 2007 Ocean engineering 34 1650–1667
[16] Elmokadem T, Zribi M and Youcef-Toumi K 2017 Ocean Engineering 129 613–625
[17] Qiao L and Zhang W 2019 IEEE Transactions on Industrial Informatics 16 1248–1258
[18] Karkoub M, Wu H M and Hwang C L 2017 Ocean Engineering 145 188–198
[19] Mousavi S S, Schukat M and Howley E 2018 Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016: Volume 2 (Springer) pp 426–440
[20] Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland
A K, Ostrovski G et al. 2015 nature 518 529–533
[21] Cruz F, Magg S, Weber C and Wermter S 2016 IEEE Transactions on Cognitive and Developmental Systems 8 271–284
[22] Huo Y, Li Y and Feng X 2018 IOP Conference Series: Materials Science and Engineering vol 428 (IOp Publishing) p 012063
[23] Chu Z, Sun B, Zhu D, Zhang M and Luo C 2020 IET Intelligent Transport Systems 14 764–774
[24] Yuh J 1994 IEEE Control Systems Magazine 14 39–46
[25] El-Fakdi A and Carreras M 2008 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE) pp 3635–3640
[26] Carreras M, El-fakdi A and Ridao P 2013 Marine Robot Autonomy 287–328
[27] Yu R, Shi Z, Huang C, Li T and Ma Q 2017 2017 36th Chinese control conference (CCC) (IEEE) pp 4958–4965
[28] Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G et al. 2015 nature 518 529–533