

Article

Robust Adaptive Dynamic Positioning: An Asynchronous Actor and Critic Approach with Meta-Driven Radial Function Network

Wanjin Huang ^{1,2}, Jiqiang Li ^{1,2} and Guoqing Zhang ^{1,2,*}¹ Navigation College, Dalian Maritime University, Dalian 11626, China; hwj_495@163.com (W.H.); lijqliang2018@163.com (J.L.)² State Key Laboratory of Maritime Technology and Safety, Dalian Maritime University, Dalian 116026, China

* Correspondence: zgq_dlm@163.com

Abstract

Dynamic Positioning systems are crucial for modern marine vessels to maintain positions or track trajectories under environmental disturbances. Traditional model-based and neural network control schemes often suffer from heavy computational burdens, low-velocity nonlinearities, and chattering near decision boundaries during waypoint transitions, which can trigger actuator saturation. To address these challenges, this paper proposes an enhancing robust adaptive control algorithm. Specifically, a model-free control framework is developed by employing an asynchronous deep Actor–Critic neural network with multi-layer perceptron for high-precision policy approximation in continuous spaces. To accelerate convergence, an online meta-driven radial basis function network is proposed for adaptive reward shaping, optimized by the Adam scheme. Furthermore, at the guidance level, a hysteresis state machine and an adaptive damping reference model are designed to decouple wave-induced high-frequency chattering and eliminate thrust saturation. By applying dynamic surface control, the proposed scheme avoids complex thrust allocation calculations. The proposed method enhances system autonomy and ensures smooth transient behavior while maintaining compatibility with standard marine hardware.

Keywords: meta-driven reward shaping; adaptive damping reference model; hysteresis state machine; asynchronous Actor–Critic

Academic Editor: Xinqiang Chen

Received: 30 June 2026

Revised: 27 July 2026

Accepted: 30 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Dynamic Positioning (DP) systems represent a core technology in modern marine engineering [1]. These systems enable fully actuated vessels to maintain a predefined position or track desired trajectories. By actively generating control forces, they counteract environmental disturbances without relying on mooring lines [2]. In recent years, artificial intelligence has introduced new paradigms for autonomous control in marine engineering [3,4]. To address model uncertainties and unpredictable external disturbances, researchers have utilized neural networks and Takagi Sugeno fuzzy systems [5,6]. However, these methods typically involve a large number of computational parameters. Advanced control schemes, such as event-triggered control [7] and time-delay compensation [8], have optimized resource utilization. Nevertheless, traditional model-based algorithms

still show limited adaptability when dealing with complex hydrodynamic interactions [9] and low-velocity nonlinear properties in dynamic marine environments [10].

To overcome these limitations, Reinforcement Learning (RL) has emerged as a promising model-free paradigm for dynamic positioning systems [11–13]. Unlike traditional methods [14] that suffer from the curse of dimensionality, we employ an asynchronous deep Actor–Critic neural network with Multi-Layer Perceptron (MLP) to achieve high-precision policy approximation in continuous spaces.

At the same time, to accelerate convergence, an online meta-driven Radial Basis Function Network is proposed for adaptive reward shaping [15,16], which dynamically refines reward signals based on real-time tracking performance [17]. To ensure stable parameter updates, the Adam optimization scheme is incorporated to scale gradient steps. However, the exploration of RL can still induce abrupt state jumps during large waypoint transitions, triggering actuator saturation [18,19].

Consequently, ensuring smooth transient behavior requires systematic trajectory planning at the guidance level [20]. This planning must work in tandem with the low-level control loop. Particularly in multi-target positioning tasks [21], the smoothness of waypoint transitions determines the overall performance of the system. Conventional switching strategies based on acceptance circles often suffer from discrete chattering near decision boundaries. This chattering is primarily caused by wave-induced high-frequency motions. Although second-order critically damped reference models are widely used to filter reference trajectories, static parameter configurations present a trade-off. They fail to simultaneously suppress overshoot under large deviations and maintain response sensitivity under small deviations [22]. Therefore, designing a mechanism to suppress transient shocks by dynamically adapting reference parameters remains a critical bottleneck.

Motivated by these observations, this paper proposes a robust adaptive dynamic positioning control scheme based on an asynchronous and a hybrid reward fusion network. This scheme enhances the overall control performance of ship dynamic positioning from two aspects: optimization of the reinforcement learning algorithm and dynamic trajectory smoothing of the guidance layer. The major contributions of this article are twofold:

(1) An asynchronous reinforcement learning framework is developed for model-free vessel dynamic positioning, incorporating a meta-driven radial basis function network-based adaptive reward shaping strategy, is developed to dynamically optimize reward signals according to real-time tracking performance combined with Adam gradient updating.

(2) A hysteresis state machine and an adaptive damping reference model are established to refine the guidance trajectory at the planning level. Time-varying damping regulation balances dynamic response and steady-state smoothness, which effectively suppresses wave-induced high-frequency chattering during waypoint switching.

2. Issues Regarding the Formulation and Preliminaries

2.1. Preliminaries

Throughout this paper, $\text{diag}\{y_1, \dots, y_n\}$ denotes the main diagonal matrix, with the diagonal elements being $y_1, \dots, y_n$. For a given matrix $X = [x_{i,j}] \in \mathbb{R}^{m \times n}$ where $\mathbb{R}^{m \times n}$ represents the set of all $m \times n$ matrices consisting of real numbers, m and n represent the numbers of rows and columns of the matrix, respectively. $|\cdot|$ indicates the matrices of the absolute operator of a scalar. $\|\cdot\|$ is the Euclidean norm of a vector and $\|\cdot\|_F$ denotes the Frobenius norm. Specially, $\|X\|_F^2 = \text{tr}\{X^T X\} = \sum_{i=1}^m \sum_{j=1}^n x_{i,j}^2$ refers to the

square of the Frobenius norm of the matrix, which is the sum of the squares of all elements. $\text{tr}\{X\}$ is the sum of the main diagonal elements of X . The notation $(\cdot)$ acts as a general placeholder for arbitrary variables or parameters. For any quantity denoted by $(\cdot)$, $\widehat{(\cdot)}$ stands for its estimated value, and the corresponding estimation error is defined as $\widetilde{(\cdot)} = (\cdot) - \widehat{(\cdot)}$.

The operator $\odot$ denotes the Hadamard product, which performs element-by-element multiplication between vectors or matrices of the same dimension. Given an n -dimensional vector $z \in \mathbb{R}^n$, along with its component-wise lower bound vector $a \in \mathbb{R}^n$ and upper bound vector $b \in \mathbb{R}^n$ satisfying $a_i < b_i$ for all $i=1,2,\dots,n$, the smooth projection operator $\text{Proj}\{z,a,b\} = [\text{Proj}\{z_1,a_1,b_1\}, \dots, \text{Proj}\{z_n,a_n,b_n\}]^T$ is introduced to constrain each component of z inside the compact interval $[a_i, b_i]$. This operator remains continuously differentiable and satisfies $\text{Proj}\{z_i,a_i,b_i\} = z_i$ if $z_i \in (a_i, b_i)$.

2.2. Dynamic Model of Marine Vessels

To achieve precise dynamic positioning (DP) and satisfy the high stability requirements of offshore operations, a 3-degree-of-freedom (DOF) marine vessel model is considered. According to seakeeping and maneuvering theory, the kinematic and dynamic equations are expressed as:

$$\dot{\eta} = R(\psi)\mu \quad (1)$$

$$M\dot{\mu} + N(\mu)\mu = \tau + \tau_w \quad (2)$$

$$\tau = T(\beta)\kappa(n)\mu_p \quad (3)$$

where the vector $\eta = [x, y, \psi]^T \in \mathbb{R}^3$ describes the position coordinates and heading angle relative to the earth-fixed frame, $\mu = [u, v, r]^T \in \mathbb{R}^3$ represents the linear and angular velocity vector within the body-fixed frame, and the rotational matrix is expressed as Equation (4). Note that $R^T(\psi) = R^{-1}(\psi)$, $\|R\| = 1$.

$$R(\psi) = \begin{bmatrix} \cos(\psi) & \sin(\psi) & 0 \\ -\sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

In Equation (2), $M \in \mathbb{R}^{3 \times 3}$ and $N(\mu)\mu$ represent the system inertia mass matrix and the nonlinear hydrodynamic terms, respectively, and $N(\mu)\mu$ includes Coriolis forces and centripetal forces as well as nonlinear damping forces. $\tau = [\tau_u, \tau_v, \tau_r]^T$ represents the controlling force and torque provided by the driving device, and Equation (1) expresses detailed distribution models are given. $\tau_w = [\tau_{wu}, \tau_{wv}, \tau_{wr}]^T$ is a bounded disturbance caused by the disturbances of the wind, wave, and current environment.

$$\mathbf{M} = \begin{bmatrix} m - X_{\dot{u}} & 0 & 0 \\ 0 & m - Y_{\dot{v}} & m x_G - Y_{\dot{r}} \\ 0 & m x_G - Y_{\dot{r}} & I_z - N_{\dot{r}} \end{bmatrix} \quad (5)$$

$$\mathbf{N}(\boldsymbol{\mu})\boldsymbol{\mu} = \begin{bmatrix} -X_u - X_{|u|u}|u|u + Y_v|v|r + Y_r r r \\ -X_{\dot{u}} u r - Y_v v - Y_r r - Y_{|v|v}|v|v - Y_{|v|r}|v|r \\ (X_{\dot{u}} - Y_{\dot{v}}) u v - Y_{\dot{r}} u r - N_v v - N_r r - N_{|v|v}|v|v - N_{|v|r}|v|r \end{bmatrix} \quad (6)$$

I_z represents the moment of inertia along the body-fixed z_b axis. $Y_{\dot{v}}$, $Y_{\dot{r}}$ and $N_{\dot{r}}$ characterize added mass and inertia moment. Around $\mathbf{N}(\boldsymbol{\mu})\boldsymbol{\mu}$, X_u , $X_{|u|u}$, Y_v , etc., are all hydrodynamic force derivatives.

In Equation (3), $\mathbf{T}(\boldsymbol{\beta}) \in \mathbb{R}^{3 \times q}$ denotes the thrust allocation configuration matrix, where the positive integer q represents the total number of independent propeller thrusters equipped on the vessel; $\boldsymbol{\kappa}(n) = \text{diag}\{\kappa_1(n_1), \dots, \kappa_q(n_q)\} \in \mathbb{R}^{q \times q}$ is the unknown speed-dependent thrust coefficient matrix and each element $\kappa_i(n_i)$ is the time-varying thrust gain of the i -th thruster related to propeller rotational speed n_i ; $\boldsymbol{\mu}_p = [p_1, p_2, \dots, p_q]^T$ is the actual pitch ratio input vector with $p_i \in [-1, 1]$, $i = 1, \dots, q$. β is the bearing angle especially for the rotatable thruster [23].

Assumption 1. As the inertial mass matrix, $\mathbf{M}$ is positive definite and invertible. Due to the inherent port starboard symmetry and approximate fore aft symmetry of surface vessels, this assumption is typically satisfied.

Assumption 2. In Equation (2), $\boldsymbol{\tau}_w$ represents a bounded disturbance, i.e., there exists a positive constant vector $\boldsymbol{\tau}_w = [\tau_{wu}, \tau_{wv}, \tau_{wr}]^T > 0$. Note that $\boldsymbol{\tau}_w$ is an unknown vector solely for the purposes of analysis.

Assumption 3. The gradient vectors are bounded on the compact sets Ω_i , such that $\|\mathbf{G}_i(t)\| \leq \xi_i$ for all $t \geq 0$, where $\xi_i > 0$ is positive constants.

Lemma 1. In practical engineering, the thrust generated by a propulsor is always limited. Therefore, for a given vessel, the force coefficient of the propulsor is a constant that must satisfy the condition $0 < \underline{\kappa}_i \leq \kappa_i(n_i) \leq \overline{\kappa}_i$, where $i = 1, \dots, q$. Both $\underline{\kappa}_i$ and $\overline{\kappa}_i$ are not explicitly known and are introduced exclusively for the purpose of stability proof.

Lemma 2. For any $\boldsymbol{\Theta} \in \Omega$ and its optimal target $\boldsymbol{\Theta}^* \in \Omega$, the projection operator satisfies the following inequality [24]:

$$(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T \text{proj}(\boldsymbol{\Theta}, Y) \leq (\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T Y \quad (7)$$

2.3. Meta-Driven Reward Shaping Mechanism and Adam Meta Gradient Optimization

To address the slow convergence of reinforcement learning caused by sparse tracking rewards, a baseline tracking penalty reward is defined as:

$$r_0 = -c_1 \sqrt{x_e^2 + y_e^2} - c_2 |\psi_e| - c_3 \|\boldsymbol{\mu}_e\| - c_4 \|\mathbf{u}_{rl}\| \quad (8)$$

where $c_1, c_2, c_3, c_4 > 0$ are fixed weight hyperparameters for position error, heading error, velocity error and control effort penalty.

Meanwhile, an auxiliary Reward Shaping Network is introduced. The corrective feedback reward is queried directly from the activated grid indices:

$$r_{\text{rfn}} = f_{\text{rfn}}(s_t) \quad (9)$$

where $f_{\text{rfn}}(\cdot)$ denotes the forward mapping of the RFN.

Utilizing a dynamic blending mechanism, the final reward r_t is formulated as:

$$r_t = r_0 + \vartheta(t) \cdot (r_{\text{rfn}} - r_0) - \mathbb{I}_s \cdot c_s \quad (10)$$

where $\vartheta(t) \in [0, \vartheta_{\max}]$ is a warm-up coefficient scaling factor, and $\mathbb{I}_s$ indicates safety filter activation, and $c_s > 0$ is the safety penalty constant for suppressing saturation risk.

The parameters of the Meta-Reward Shaping network are updated using a meta-performance index computed over a sliding horizon of length Δ . And a rolling buffers for tracking errors and control effort are defined as:

$$\begin{aligned} \text{pos}_{\text{buf}}^{(i)} &= \left\{ \sqrt{x_e(j)^2 + y_e(j)^2} \right\}_{j=i-\Delta+1}^i \\ \text{psi}_{\text{buf}}^{(i)} &= \left\{ |\psi_e(j)| \right\}_{j=i-\Delta+1}^i \\ \text{ctrl}_{\text{rms}}^{(i)} &= \sqrt{\frac{1}{\Delta} \sum_{j=i-\Delta+1}^i \|\mathbf{u}(j)\|^2} \end{aligned} \quad (11)$$

where $\text{pos}_{\text{buf}}^{(i)}$ and $\text{psi}_{\text{buf}}^{(i)}$ are a sliding window storage of a horizontal position Euclidean error and absolute heading tracking error over recent Δ time steps. $\text{ctrl}_{\text{rms}}^{(i)}$ is the root-mean-square value of control input $\mathbf{u}$ within the sliding window horizon.

Through a weighted combination, the expression of the meta-performance index $J_{\text{meta}}^{(i)}$ can be further obtained as follows:

$$J_{\text{meta}}^{(i)} = R_J \left(\text{pos}_{\text{buf}}^{(i)}, \text{ctrl}_{\text{rms}}^{(i)}, \text{psi}_{\text{buf}}^{(i)} \right) \quad (12)$$

Finally, the parameter set is updated via the following gradient scheme:

$$\begin{aligned} \mathbf{G}_i(t) &= \boldsymbol{\Phi}_i(x) \left((\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*)^T \boldsymbol{\Phi}_i(x) + E_i \right) \\ \boldsymbol{\Theta}_i^{(ii+1)} &= \boldsymbol{\Theta}_i^{(ii)} - \alpha_i \mathbf{G}_{\boldsymbol{\Theta}_i} \end{aligned} \quad (13)$$

where α_i denotes the positive meta-learning rate. $\boldsymbol{\Phi}_i(x)$ is the bounded vector, and E_i represents the bounded reconstruction/Bellman residual error.

For any parameter, $\boldsymbol{\Theta}_t$ is updated using the Adam optimizer rules, which prevent gradient explosion:

$$\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1 - \beta_1}, \hat{v}_t = \frac{v_t}{1 - \beta_2} \\
\theta_t &= \theta_{t-1} - \rho_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}}
\end{aligned} \quad (14)$$

where $m_t, \dots, v_t$ denote an intermediate calculation quantity in Adam [25,26].

3. Robust Adaptive Neural Cooperative Controller

This section is established by introducing an Actor–Critic-based reinforcement learning strategy to construct an adaptive control scheme for vessel dynamic positioning (DP) systems [27]. By iteratively learning and optimizing control policies, the proposed controller is capable of adapting to complex and time-varying ocean disturbances. Subject to Assumptions 1 and 2, we devise a robust neural adaptive control framework grounded on dynamic surface control theory for the DP dynamics formulated in Equations (1) and (2). The overall structural sketch and detailed module flowchart of the developed DP control architecture are visualized in Figure 1, which expresses the complete structure and logical flow of the entire control system [28–30].

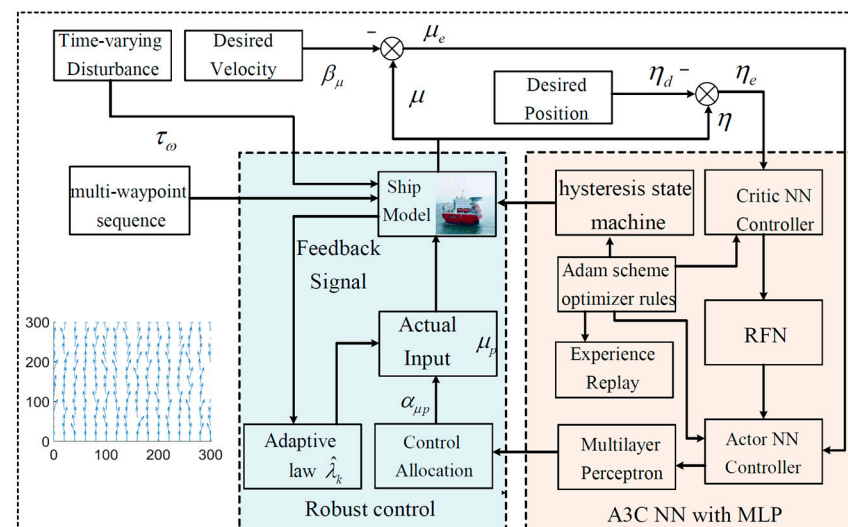


Figure 1. Overall schematic framework of the proposed reinforcement learning-based adaptive control system.

3.1. Actor-Critic State-Value and Policy Generation

Define the closed-loop system tracking error state vector of the vessel as $s_t = [\eta_e, \mu_e]^T = [x_e, y_e, \psi_e, u_e, v_e, r_e]^T \in \mathbb{R}^6$, where the position η_e and velocity errors μ_e are defined from Equations (27) and (32), respectively. The critic network approximates the state value function $V(s_t)$ to evaluate the performance of the current policy [31–33]. Given the state vector s_t , the forward pass propagation is defined as:

$$h_c = \max(0, \theta_{c1} s_t + b_{c1}) \quad (15)$$

$$V_t = \theta_{c2} h_c + b_{c2} \quad (16)$$

where Θ_{c1}, b_{c1} and Θ_{c2}, b_{c2} are the weights and biases of the hidden layer and the output layer, respectively.

To prevent parameter drift and control signal windup, a projection operator is applied to the parameter update laws of the actor networks $\Pi \in \{\mu, \sigma\}$. For the parameter element at index $[i, c]$, the update law is defined as:

$$\dot{\Theta}_{\Pi}[i, c] = \text{Proj}[\alpha_{\pi} \nabla_{\Pi}^{(c)}, -\theta_{\Pi, \max}, \theta_{\Pi, \max}] \quad (17)$$

where α_{π} denotes the learning rate, $\nabla_{\Pi}^{(c)}$ is the policy gradient, and $\theta_{\Pi, \max}$ represents the predefined parameter boundary to enforce stability.

The actor network parameterized by Θ generates the mean μ_t and standard deviation σ_t of the action distribution. Sharing the feature extractor representation h_a , the forward pass is modeled as:

$$h_a = \max(0, \Theta_{a1} s_t + b_{a1}) \quad (18)$$

$$z_{\mu} = \Theta_{\mu} h_a + b_{\mu}, \mu_t = U_{\max} \tanh(z_{\mu}) \quad (19)$$

$$z_{\sigma} = \Theta_{\sigma} h_a + b_{\sigma}, \sigma_t = \ln(1 + e^{z_{\sigma}}) \quad (20)$$

where $U_{\max}$ scales the maximum control authority. The mean action μ_t is bounded by a hyperbolic tangent function, and the standard deviation σ_t is constrained to be positive using Soft plus activation. h_a is a hidden layer feature output activated by a ReLU function $\max(0, \cdot)$. Θ_{a1}, b_{a1} , Θ_{μ}, b_{μ} and $\Theta_{\sigma}, b_{\sigma}$ are a weight matrix and bias vector of a Actor hidden layer MLP, a output layer weights/biases for action mean μ_t and a action standard deviation σ_t , respectively.

To ensure smooth control output and reduce high-frequency mechanical wear [34], exploration actions sampled from the Gaussian distribution are filtered:

$$u_{t1} = \mu_t + \sigma_t \odot \xi, \xi \sim \mathcal{N}(0, I) \quad (21)$$

$$u_{t1}^{(ii)} = \begin{cases} u_{t1}, & ii=1 \\ \alpha_a u_{t1}^{(ii-1)} + (1-\alpha_a) u_{t1}, & ii>1 \end{cases}$$

where ξ is the zero-mean Gaussian noise vector, and α_a is the exponential smoothing factor.

In the next step, the Critic target value R_t and the generalized advantage estimate A_t are calculated as:

$$R_t = r_{\text{final}} + \gamma V_{\text{prev}}, A_t = R_t - V_t \quad (22)$$

To account for physical latency in real-world environments, the state vector history with an observation delay of Δw is utilized, denoted as $s_{t-\Delta w} = [x_e(i-\Delta w), \dots, r_e(i-\Delta w)]^T$. Consequently, the target value of the previous state, V_{prev} is reconstructed as:

$$V_{\text{prev}} = \begin{cases} \Theta_{c2} \max(0, \Theta_{c1} s_{t-\Delta w} + b_{c1}) + b_{c2}, & t \geq t-1 \\ V_t, & \text{otherwise} \end{cases} \quad (23)$$

The policy log-likelihood, policy entropy, and the loss functions for the Critic and Actor are formulated as:

$$\log \pi = -\frac{1}{2} \sum \left(\frac{(u_i - \mu_i)^2}{\sigma_i^2} + 2 \ln \sigma_i + \ln(2\pi) \right) \quad (24)$$

$$H_\pi = \frac{1}{2} \sum \ln(2\pi e \sigma_i^2) \quad (25)$$

$$L_c = \frac{1}{2} A_t^2, L_a = -(\log \pi A_t + \beta_a H_\pi) \quad (26)$$

where β_a is the entropy coefficient. Let $\boldsymbol{\theta}_c = \{\boldsymbol{\theta}_{c1}, \boldsymbol{\theta}_{c2}\}$ and $\boldsymbol{\theta}_a = \{\boldsymbol{\theta}_{a1}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma\}$ denote the parameters of the critic and actor networks, respectively.

Remark 1. Due to the localized approximation property of tile coding, the optimal (ideal) parameter tables are defined as V_{prev} . The parameter estimation errors are defined as: $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i^* - \boldsymbol{\theta}_i$. Owing to the application of saturation clipping, the parameter estimation errors are inherently bounded within a compact set, yielding $\|\tilde{\boldsymbol{\theta}}_i\| \leq \Theta_{i,\max}$.

3.2. Control Design

Define the geodetic frame position and heading tracking error vector between actual vessel state $\boldsymbol{\eta} = [x, y, \psi]^T$ and desired reference trajectory $\boldsymbol{\eta}_d = [x_d, y_d, \psi_d]^T$ as:

$$\boldsymbol{\eta}_e = \boldsymbol{\eta} - \boldsymbol{\eta}_d \quad (27)$$

Taking the time derivative of Equation (9), one has obtained as:

$$\dot{\boldsymbol{\eta}}_e = \mathbf{R}(\psi) \boldsymbol{\mu} - \dot{\boldsymbol{\eta}}_d \quad (28)$$

Utilizing the position error dynamics in Equation (10), the immediate control $\boldsymbol{\alpha}_\mu \in \mathbb{R}^3$ can be directly selected as

$$\boldsymbol{\alpha}_\mu = \mathbf{R}^{-1}(\psi) (-\mathbf{K}_\eta \boldsymbol{\eta}_e + \dot{\boldsymbol{\eta}}_d), \quad (29)$$

where $\mathbf{K}_\eta$ denotes a positive definite diagonal matrix. Nevertheless, analytically obtaining the derivative of $\boldsymbol{\alpha}_\mu$ is computationally prohibitive due to its high complexity. To circumvent the “explosion of complexity” issue, the dynamic surface control technique [14,27] is employed by introducing a first-order low-pass filter with a time constant matrix $\mathbf{t}_\mu = \text{diag}\{t_u, t_v, t_r\}$.

$$\boldsymbol{\alpha}_\mu - \boldsymbol{\beta}_\mu = \mathbf{t}_\mu \dot{\boldsymbol{\beta}}_\mu, \boldsymbol{\alpha}_\mu(0) = \boldsymbol{\beta}_\mu(0) \quad (30)$$

The filtered output vector $\boldsymbol{\beta}_\mu$ serves as the command signal for the velocity $\boldsymbol{\mu}$. Meanwhile, define the error vector as:

$$\mathbf{q}_\mu = \boldsymbol{\alpha}_\mu - \boldsymbol{\beta}_\mu, \quad (31)$$

and the velocity tracking vector as:

$$\boldsymbol{\mu}_e = \boldsymbol{\mu} - \boldsymbol{\beta}_\mu. \quad (32)$$

the derivative of $\mathbf{q}_\mu$ can be derived by Equations (13) and (14):

$$\begin{aligned} \dot{\mathbf{q}}_\mu &= \dot{\boldsymbol{\alpha}}_\mu - \dot{\boldsymbol{\beta}}_\mu \\ &= \boldsymbol{\Omega}_\mu(\boldsymbol{\psi}, r, \boldsymbol{\eta}_e, \dot{\boldsymbol{\eta}}_e) - \boldsymbol{\Gamma}_\mu^{-1} \mathbf{q}_\mu \end{aligned} \quad (33)$$

where $\boldsymbol{\Omega}_\mu(\cdot) = [\boldsymbol{\Omega}_u(\cdot), \boldsymbol{\Omega}_v(\cdot), \boldsymbol{\Omega}_r(\cdot)]^T$ contains bounded continuous function elements. Subsequently, combining Equations (28) and (33) yields the derivative of the position tracking error:

$$\dot{\boldsymbol{\eta}}_e = \mathbf{R}(\boldsymbol{\psi})(\boldsymbol{\alpha}_\mu - \mathbf{q}_\mu + \boldsymbol{\mu}_e) - \dot{\boldsymbol{\eta}}_d \quad (34)$$

where the primary control variable $\boldsymbol{\alpha}_{\mu p} = [\alpha_{up}, \alpha_{vp}, \alpha_{rp}]^T$ in the proposed design is chosen as Equation (38) for the desired thrusts term $\mathbf{T}(\boldsymbol{\beta})\boldsymbol{\kappa}(n)\boldsymbol{\mu}_p$.

Propeller thrust servo systems are highly susceptible to significant uncertainties arising from intricate system structures and varying operating environments. These uncertainties, such as parameter perturbations, external environmental loads, and load variations [25], directly lead to fluctuations in the thrust coefficient, thereby threatening propulsion system stability [28,35]. To mitigate these issues, this study introduces a robust adaptive control scheme capable of online parameter self-tuning to accommodate variations in vessel states and propeller dynamics [29]. Specifically, the parameter $1/\kappa_i(n_i)$ is estimated via $\hat{\lambda}_k$ [30,35]. Based on Equations (17) and (18), the control input vector is synthesized as $\mathbf{p} = [p_1, \dots, p_q]^T$, which parameters are updated by the adaptive law in Equation (19). Consequently, the proposed method guarantees superior stability and efficiency under dynamic environments.

$$\mathbf{p} = \text{sgn}(\boldsymbol{\mu}_p) \times \sqrt{|\boldsymbol{\mu}_p|} \quad (35)$$

$$\boldsymbol{\mu}_p = \text{diag}\{\hat{\lambda}_1, \dots, \hat{\lambda}_q\} \mathbf{T}^\dagger(\cdot) \boldsymbol{\alpha}_{\mu p} \quad (36)$$

$$\dot{\hat{\lambda}}_k = -\chi_k \left(\sum_{i=u,v,r} \sum_{j=u,v,r} T_{k,j}^\dagger T_{i,k} i_e \alpha_{jp} \right) + \chi_k \sigma_k (\hat{\lambda}_k - \lambda_k(0)), k=1, \dots, q \quad (37)$$

$$\boldsymbol{\alpha}_{\mu p} = -\mathbf{K}_\mu \boldsymbol{\mu}_e + \boldsymbol{\mu}_t + \dot{\boldsymbol{\beta}}_\mu - \mathbf{R}(\boldsymbol{\psi}) \boldsymbol{\eta}_e \quad (38)$$

where $\mathbf{K}_\mu, \chi_k = [\chi_1, \dots, \chi_q]^T$ and $\boldsymbol{\sigma}_k = [\sigma_1, \dots, \sigma_q]^T$ are the positive parameters, especially $\sigma_k (\hat{\lambda}_k - \lambda_k(0))$ could protect the estimation $\hat{\lambda}_k$ from the drifting divergence. $\mathbf{T}^\dagger$ denotes the pseudo inverse of $\mathbf{T}$. According to the adaptive law equation (Equation (37)), the gain correlation coefficient is added to compensate for the influence of model uncertainty and interference [31].

3.3. Error-Threshold Adaptive Reference Model

In route optimization for marine vessels, direct connections between discrete waypoints typically generate piece-wise straight-line paths with sharp corners. This non-smooth geometric profile poses significant challenges for trajectory generation, as the sudden changes in path tangents violate the vessel's dynamic feasibility constraints and often lead to numerical convergence difficulties in optimization algorithms [36]. To suppress discrete chattering near waypoint transitions, the multi-waypoint sequence is defined as:

$$\mathcal{W} = \{\mathbf{w}_k = [x_d^k, y_d^k, \psi_d^k]^T \in \mathbb{R}^3 \mid k=1, 2, \dots, N_{wp}\} \quad (39)$$

Let $I_{wp}(t) \in \{1, 2, \dots, N_{wp}\}$ be the active waypoint index. The transition distance to the target is calculated as:

$$d_{wp}(t) = \sqrt{(x_d^{I_{wp}(t)} - x(t))^2 + (y_d^{I_{wp}(t)} - y(t))^2} \quad (40)$$

To prevent frequent switching triggered by environmental disturbances, a discrete freeze counter $C_f(t) \in \mathbb{N}$ is introduced. The active waypoint updates only when the acceptance threshold is met $d_{wp}(t) < d_{reach}$ and the freeze period has expired $C_f(t) = 0$, as governed by:

$$I_{wp}(t + \Delta t) = \begin{cases} I_{wp}(t) + 1, & \text{if } d_{wp}(t) < d_{reach} \text{ and } C_f(t) = 0 \text{ and } I_{wp}(t) < N_{wp} \\ I_{wp}(t), & \text{otherwise} \end{cases} \quad (41)$$

$$C_f(t + \Delta t) = \begin{cases} N_f, & \text{if } I_{wp}(t + \Delta t) \neq I_{wp}(t) \\ \max(0, C_f(t) - 1), & \text{otherwise} \end{cases} \quad (42)$$

where the lock duration is T_f , yielding $N_f = \text{round}(T_f / \Delta t)$. This mechanism decouples waypoint transitions from high-frequency wave-induced disturbances, stabilizing the guidance system.

To smooth step control commands during waypoint transitions, a second-order non-linear reference model is designed. Let $\boldsymbol{\eta}_d$ be the reference state vector, satisfying:

$$\dot{\boldsymbol{\eta}}_d = -2\mathbf{Z}(t)\boldsymbol{\Omega}(t)\boldsymbol{\beta}_\mu - \boldsymbol{\Omega}^2(t)(\boldsymbol{\eta}_d - \mathbf{w}_{I_{wp}(t)}) \quad (43)$$

The natural frequency $\boldsymbol{\Omega}(t) = \text{diag}(\omega_{n,x}, \omega_{n,y}, \omega_{n,\psi})$ and damping ratio $\mathbf{Z}(t) = \text{diag}(\zeta_x, \zeta_y, \zeta_\psi)$ adapt dynamically to the global tracking error $e_p(t) = \sqrt{(x_d - x)^2 + (y_d - y)^2}$ as follows:

$$\omega_{n,i}(t) = \omega_{n,i}^0 \cdot [1 - \alpha_\omega \cdot \tanh(\gamma_\omega \cdot e_p(t))] \quad (44)$$

$$\zeta_i(t) = \zeta_i^0 \cdot [1 + \alpha_\zeta \cdot \tanh(\gamma_\zeta \cdot e_p(t))] \quad (45)$$

where positive tuning parameters $\alpha_\omega \in (0, 1)$ and $\alpha_\zeta > 0$ scale the adaptation bounds, while $\gamma_\omega, \gamma_\zeta > 0$ determine the sensitivity to tracking errors.

Remark 2. The adaptive law exhibits the following behaviors depending on the tracking error. When a large transition error occurs $e_p \gg 0$, the damping ratio increases to eliminate overshoot.

Simultaneously, the natural frequency $\omega_{n,i}(t)$ decreases, slowing down the transition rate of η^d and preventing control input saturation in the actor network. As the tracking error e_p converges to 0, the parameters smoothly return to their nominal values, recovering system sensitivity and tracking bandwidth.

3.4. Stability Analysis

Theorem 1. Consider a closed-loop system composed of vessel dynamics, a robust neural control law, an Actor–Critic NN, and adaptive laws. Assume that the initial conditions are satisfied:

$$\eta_e^T(0)\eta_e(0) + \mu_e^T(0)\mu_e(0) + q_e^T(0)q_e(0) + \sum \tilde{\Theta}_i^T(0)\tilde{\Theta}_i(0) \leq 0, \text{ and there exist adjustable}$$

parameters $K_\eta, K_\mu, t_\mu, \lambda_k, \chi_k, \gamma_c, \gamma_a$. The stability of the closed-loop system, which comprises the vessel dynamics in Equations (1) and (2), the control strategies in Equations (29) and (38), and the adaptation law in Equation (19), is rigorously guaranteed. Specifically, the introduced RL-integrated adaptive NN control scheme ensures that the following properties hold under arbitrary initial states:

(1) Semi-global Uniform Ultimate Boundedness: All state trajectories within the closed-loop control system are bounded in a semi-global and uniformly ultimate manner.

(2) Strictly Regulated Tracking and Learning Errors: The tracking position and velocity error η_e , μ_e , and the parameter estimation errors $\tilde{\Theta}_i$ of the reinforcement learning process are constrained within a narrow vicinity of the origin.

(3) Optimal Control Approximation: The optimal policy of the RL framework is derived by estimating the ideal but unknown neural weights. An online estimator is developed for both the actor and critic weights, enabling parallel recursive updates of the neural networks.

To verify these properties, a candidate Lyapunov function is constructed below based on the physical characteristics of vessel DP operations and the controller design:

$$V = \frac{1}{2}\eta_e^T\eta_e + \frac{1}{2}\mu_e^T M \mu_e + \frac{1}{2}q_e^T q_e + \frac{1}{2} \sum_{k=1}^q \frac{\kappa(n_k) \tilde{\lambda}_k^2}{\chi_k} + V_{RL} \quad (46)$$

The time derivative of Equation (25) is as:

$$\dot{V} = \eta_e^T \dot{\eta}_e + \mu_e^T M \dot{\mu}_e + q_e^T \dot{q}_e + \sum_{k=1}^q \frac{\kappa(n_k) \tilde{\lambda}_k \dot{\tilde{\lambda}}_k}{\chi_k} + \dot{V}_{RL} \quad (47)$$

Similarly with the former, we further have:

$$\eta_e^T \dot{\eta}_e = \eta_e^T (R(\psi) \mu - \dot{\eta}_d) \leq -K_\eta \eta_e^T \eta_e + \eta_e^T R(\psi) \mu_e + \|\eta_e\|_F^2 + \frac{1}{4} \|q_e\|_F^2 \quad (48)$$

According to Equations (4) and (14) and Young's inequality:

$$\mu_e^T M \dot{\mu}_e = \mu_e^T [\tau + \tau_\omega - N(\mu) \mu] - \mu_e^T q_e t_\mu^{-1} M \quad (49)$$

where $N(\mu) \mu - \tau_\omega$ is treated with an actor parameter. According to Equation (28):

$$\begin{aligned} \mu_e^T M \dot{\mu}_e &= \mu_e^T [\tau - \mu_t - \varepsilon_a - q_e t_\mu^{-1} M] \\ &= -\mu_e^T [K_\mu \mu_e + \varepsilon_a - T \kappa(\cdot) \tilde{\lambda}_k T^T \alpha_{\mu p}] \\ &\leq -\left(K_\mu - \frac{1}{4} I\right) \mu_e^T \mu_e + \mu_e^T T \kappa(\cdot) \tilde{\lambda}_k T^T \alpha_{\mu p} + \|\varepsilon_a\|_F^2 \end{aligned} \quad (50)$$

Besides, we also have:

$$\mathbf{q}_\mu^T \dot{\mathbf{q}}_\mu \leq -\mathbf{q}_\mu^T \mathbf{t}_\mu^{-1} \mathbf{q}_\mu + \sum_{i=u,v,r} \frac{K_{\Omega_i}^2 q_i^2}{2b} + \frac{3}{2} \quad (51)$$

Through the designed adaptive control law in Equation (19) is further obtained by:

$$\sum_{k=1}^q \frac{\kappa(n_k) \tilde{\lambda}_k \dot{\lambda}_k}{\chi_k} = \sum_{k=1}^q \kappa(\cdot) \tilde{\lambda}_k \left[\sum_{i=u,v,r} \sum_{j=u,v,r} T_{k,j}^\dagger T_{i,k} \alpha_{jp} - \sigma_k \left(\hat{\lambda}_k - \dot{\lambda}_k(0) \right) \right] \quad (52)$$

The first term cancels out one of $\mathbf{T}\kappa(\cdot) \tilde{\lambda}_k \mathbf{T}^\dagger \alpha_{\mu p} \boldsymbol{\mu}_e$, and the second term is determined by Young's inequality:

$$\mathbf{T}\kappa(\cdot) \tilde{\lambda}_k \mathbf{T}^\dagger \alpha_{\mu p} \boldsymbol{\mu}_e + \sum_{k=1}^q \frac{\kappa(n_k) \tilde{\lambda}_k \dot{\lambda}_k}{\chi_k} \leq -\frac{1}{2} \sum_{k=1}^q \kappa(n_k) \sigma_k \left\{ \tilde{\lambda}_k^2 - \left[\hat{\lambda}_k - \dot{\lambda}_k(0) \right] \right\} \quad (53)$$

Finally, for each network $i \in \{r, a, c\}$, define a positive-definite Lyapunov function

candidate as: $V_i(t) = \frac{1}{2} \|\boldsymbol{\theta}_i(t) - \boldsymbol{\theta}_i^*\|^2$. Taking the time derivative of $V_i(t)$ and applying **Lemma 2** yields:

$$\begin{aligned} \dot{V}_i(t) &\equiv (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)^T \dot{\boldsymbol{\theta}}_i \equiv (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)^T \text{proj}(\boldsymbol{\theta}_i, -\rho_i \mathbf{G}_i^A) \\ &\leq -\rho_i' \tilde{\boldsymbol{\theta}}_i^T \boldsymbol{\Phi}_i \left(\tilde{\boldsymbol{\theta}}_i^T \boldsymbol{\Phi}_i + E_i \right) \leq - \sum_{i \in \{r, a, c\}} \left[\frac{\rho_i'}{2} \tilde{\boldsymbol{\theta}}_i^T \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T \tilde{\boldsymbol{\theta}}_i - \frac{\rho_i'}{2} E_i^2 \right] \end{aligned} \quad (54)$$

where $\rho_i' > 0$ is a positive scaling constant related to the Adam step bounding. Assuming the activation functions $\boldsymbol{\Phi}_i(x)$ satisfy the Persistence of Excitation condition such that the average of $\boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T \geq \beta_i \mathbf{I}$ with $\beta_i > 0$.

To sum up, the final form is obtained by combining Equation (48) from Equation (54):

$$\begin{aligned} \dot{V} &\leq -(\mathbf{K}_\eta - \mathbf{I}) \boldsymbol{\eta}_e^T \boldsymbol{\eta}_e - \left(\mathbf{K}_\mu - \frac{5}{4} \mathbf{I} \right) \boldsymbol{\mu}_e^T \boldsymbol{\mu}_e - \sum_{i=u,v,r} \left(t_\mu^{-1} - \frac{1}{4} - \frac{K_{\Omega_i}^2}{2b} \right) q_i^2 - \sum_{i \in \{r, a, c\}} \frac{\rho_i' \beta_i}{2} \|\tilde{\boldsymbol{\theta}}_i\|^2 \\ &\quad + \frac{3}{2} b + \varepsilon_a^2 + \frac{1}{2} \sum_{k=1}^q \kappa(n_k) \sigma_k \left[\hat{\lambda}_k - \dot{\lambda}_k(0) \right]^2 + \sum_{i \in \{r, a, c\}} \frac{\rho_i' E_i^2}{2} - \frac{1}{2} \sum_{k=1}^q \kappa(n_k) \sigma_k \tilde{\lambda}_k^2 \\ &\leq -(\lambda_{\min} \{\mathbf{K}_\eta\} - 1) \boldsymbol{\eta}_e^T \boldsymbol{\eta}_e - \left(\lambda_{\min} \{\mathbf{K}_\mu\} - \frac{5}{4} \right) \boldsymbol{\mu}_e^T \boldsymbol{\mu}_e - \sum_{i=u,v,r} \left(t_\mu^{-1} - \frac{1}{4} - \frac{K_{\Omega_i}^2}{2b} \right) q_i^2 \\ &\quad - \frac{1}{2} \sum_{k=1}^q \kappa(n_k) \sigma_k \tilde{\lambda}_k^2 + \varrho \end{aligned} \quad (55)$$

$$\begin{aligned} a_1 &\leq \lambda_{\min} \{\mathbf{K}_\eta\} - 1, a_2 \leq \lambda_{\min} \{\mathbf{K}_\mu\} - \frac{5}{4}, \\ a_3 &\leq \sum_{i=u,v,r} \left(t_\mu^{-1} - \frac{1}{4} - \frac{K_{\Omega_i}^2}{2b} \right), a_4 \leq \sum_{i \in \{r, a, c\}} \frac{\rho_i' \beta_i}{2} \end{aligned} \quad (56)$$

Based on the above analysis, the Lyapunov parameter has been selected as:

$$a = \min\{a_1, a_2, a_3, a_4\}$$

$$\varrho = \frac{3}{2}b + \varepsilon_a^{-2} + \frac{1}{2} \sum_{k=1}^q \kappa(n_k) \sigma_k \left[\hat{\lambda}_k - \hat{\lambda}_k(0) \right]^2 + \sum_{i \in \{r, a, c\}} \frac{\rho'_i}{2} \bar{E}_i^2 \quad (57)$$

By adjusting the following parameter matrix, all tracking errors of the closed-loop control system are converged to a compact set. Further, by adjusting the parameter $K_\eta, K_\mu, t_\mu^{-1}, \gamma$, the following conditions can be obtained. Integrating two sides of Equation (55), one can get $V(t) \leq \varrho/2a + (V(0) - \varrho/2a)\exp(-2at)$. The $V(t)$ can converge into $\varrho/2a$ with $t \rightarrow \infty$. Therefore, all the state variables in the closed-loop control system are SGUUB.

4. Numerical Simulation

In this section, two representative case studies are compared to illustrate the effectiveness and merits of the control scheme, i.e., the comparative experiment with the result in [2]; to verify the DP adaptive control method based on integral RL, MATLAB 2025b is used as the simulation platform to conduct experiments on the marine environment model. The DP system model used in the experiment includes a thruster, servo system, control system and external disturbance model. And the environmental disturbances are as: The wind is at level 6, the number of wave frequency segments is 10, and the wave velocity is $10 \times 1852/3600$.

With the desired position and heading are chosen as $\eta_d = [0\text{m}, 0\text{m}, 140^\circ]^\text{T}$, and the initial states are in order $\eta(0) = [100\text{m}, 100\text{m}, 156^\circ]^\text{T}$ and $\mu(0) = [0\text{m/s}, 0\text{m/s}, 0^\circ/\text{s}]^\text{T}$.

Comparative simulation outcomes are presented in Figures 2–6. As observed in Figure 2, the vessel trajectory generated by the developed scheme possesses a narrower error bound and smoother curve, enabling the vessel position to converge steadily within the target attitude region. As the simulation proceeds, the system achieves favorable control accuracy. The presented algorithm achieves superior performance regarding control quality and positioning precision. The trajectory of the comparative algorithm deviates significantly to the left, resulting in a much longer path. In contrast, the trajectory of the proposed algorithm is highly close to that of the algorithm without trajectory optimization, with only minor deviations in the initial phase. Both yield paths are substantially shorter than the comparative algorithm and align closely with the straight line from the start to the end, validating the efficiency of the proposed scheme.

With the assistance of the embedded trajectory optimization mechanism, the vessel trajectory of the proposed algorithm exhibits a far narrower error range and smoother curve profile. During the initial phase prone to position overshoot and drift for the other methods, the proposed strategy suppresses sudden lateral deviations to a negligible level, which attain fast convergence while avoiding remarkable trajectory deformation. As the tracking process evolves, the vessel gradually settles within the desired tight positioning region, achieving better positioning and tracking accuracy compared with the contrast approach.

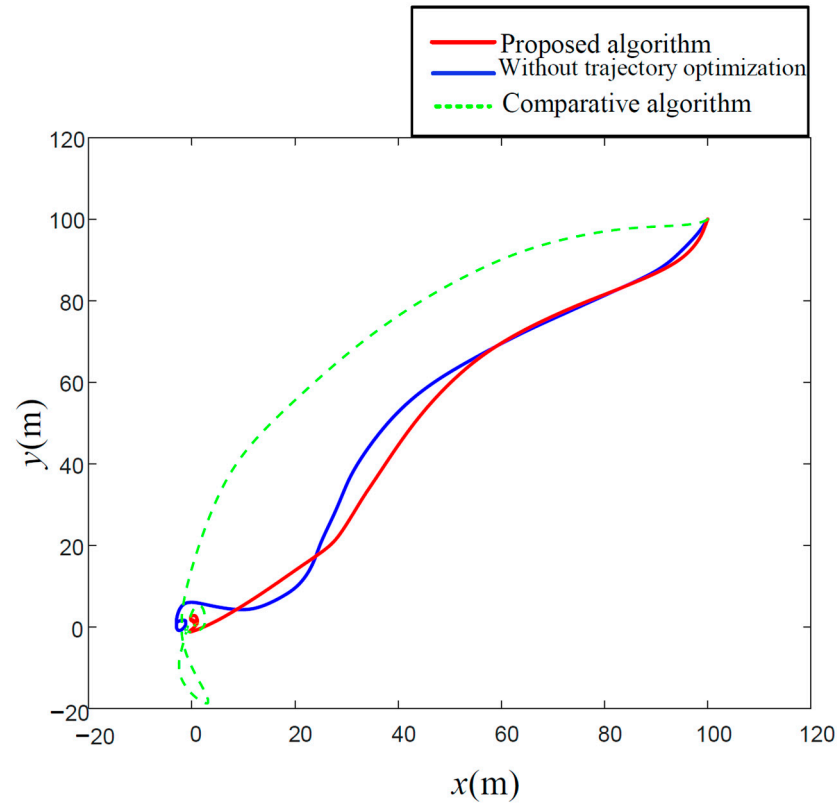


Figure 2. Position tracking performance comparison among different algorithms.

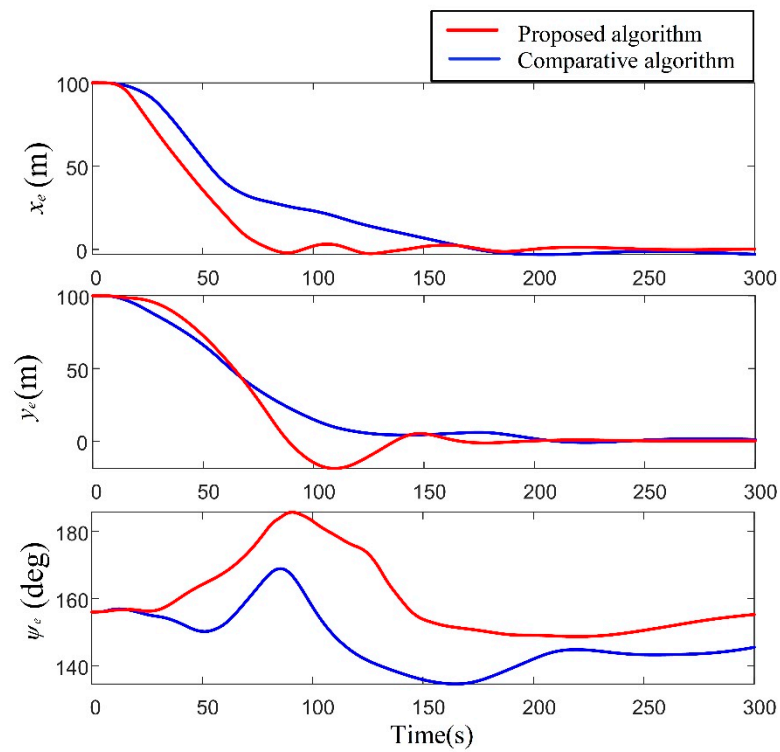


Figure 3. Comparison of the vessel position vector variables.

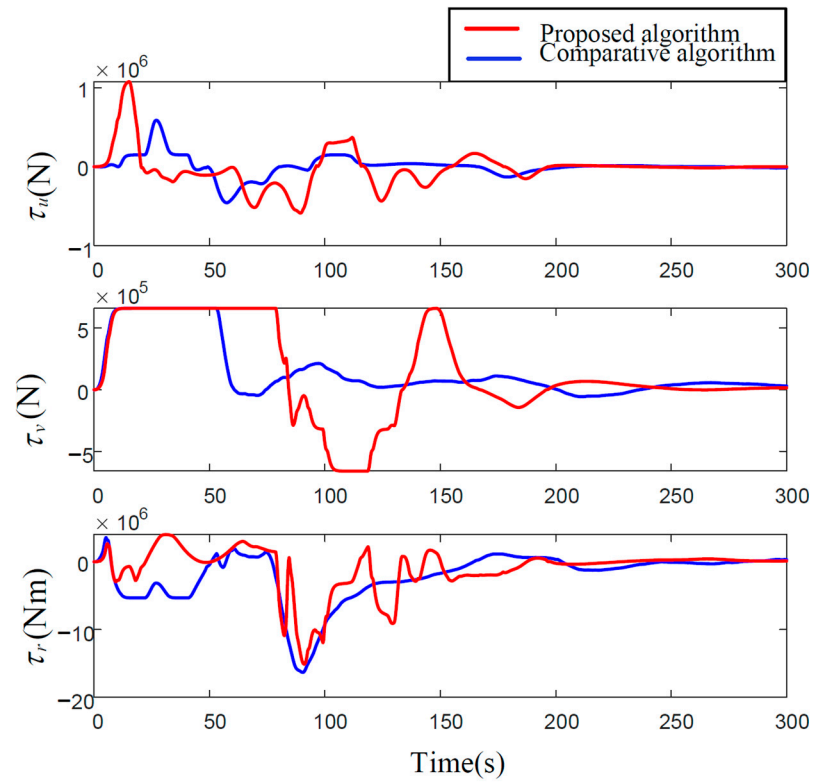


Figure 4. Comparison of the vessel control inputs (τ_u, τ_v, τ_r) .

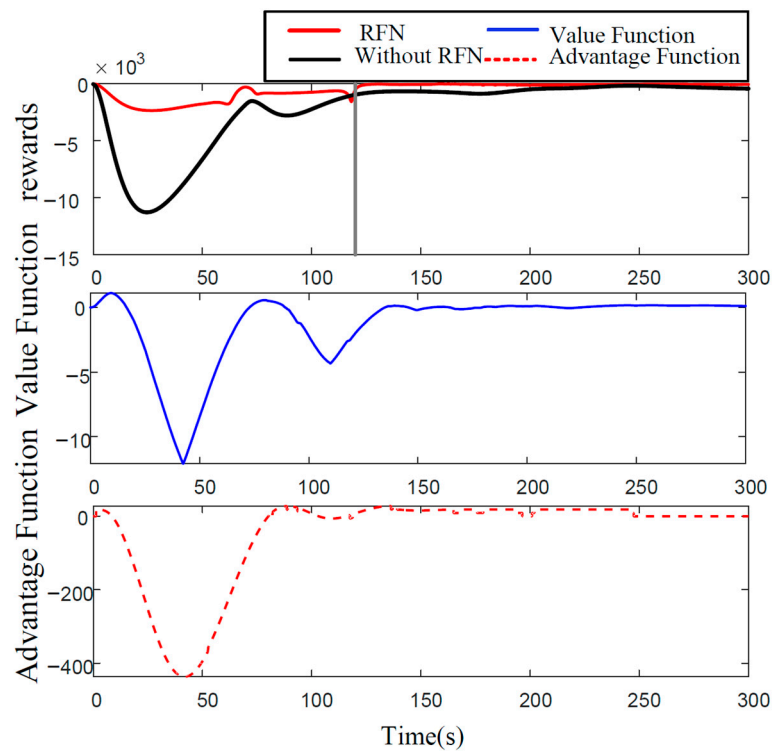


Figure 5. Convergence Analysis of RFN Reward and Value and Advantage Functions.

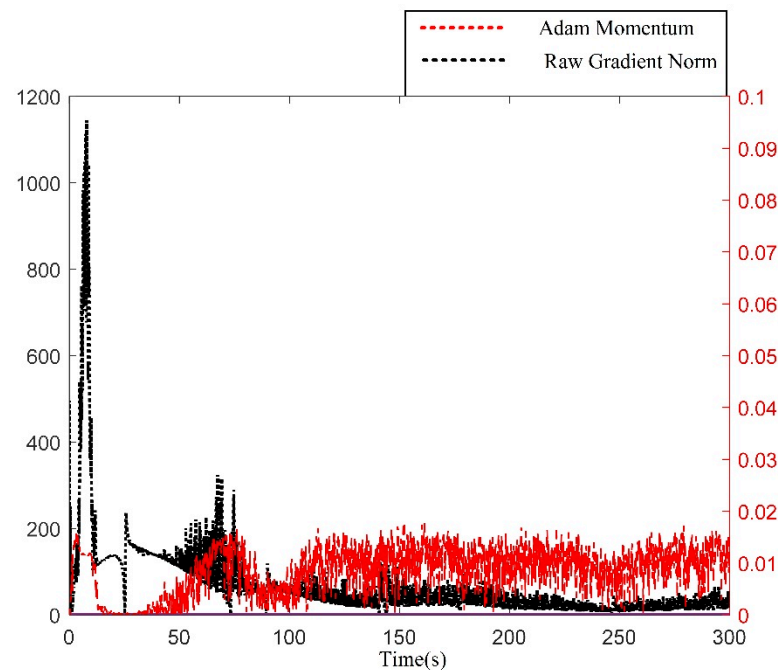


Figure 6. Adam Momentum vs. Raw Gradient Norm.

As illustrated by the specific state parameters (x, y, ψ) in Figure 3, rapid convergence to the target area is achieved by both methods. Nevertheless, the scheme in [2] suffers from pronounced oscillations in velocity and attitude tracking errors, exhibits significant fluctuations in attitude and velocity errors, whereas the proposed algorithm maintains smoother variations, achieving precise targeting of the desired attitude. In the late convergence phase when errors are close to the target tolerance band, persistent slight fluctuations can still be detected on the curve of the comparative algorithm, indicating slow stabilization of vessel planar states. However, after rapid error reduction, the curve of the proposed algorithm closely adheres to the zero-error reference with negligible residual oscillations. The whole transient process of the proposed scheme contains nearly no obvious overshoot, delivering superior transient smoothness and higher steady-state positioning accuracy in horizontal plane control.

The temporal evolution of the control forces and moments is presented in Figure 4. In the surge loop, the proposed algorithm demands a higher thrust in the initial phase compared to the comparative algorithm, which is necessary to drive rapid error elimination. In the sway loop, the proposed controller maintains a high force output early on with minor fluctuations, whereas the comparative algorithm backs off earlier but exhibits longer settling times. The yaw moments of both algorithms show a similar maximum deflection during the turning phase. As the vessel approaches the target, all control actions of the proposed scheme smoothly settle to near-zero levels. This demonstrates that the algorithm achieves high precision without causing actuator saturation or excessive high-frequency control chattering. Figure 5 displays the training indicators of the reinforcement learning process, comprising the system rewards, the Value function, and the Advantage function. The training curve with the Reward Function Network consistently maintains a high reward level throughout the run. Conversely, the scheme without RFN experiences a severe performance drop in the initial training phase before gradually recovering, illustrating the critical role of RFN in preventing reward-vanishing issues. Correspondingly, both the Value and Advantage functions experience early drops due to transient tracking errors but subsequently converge smoothly to a stable, near-zero steady state. These synchronized trends demonstrate that the proposed meta-reward shaping strategy stabilizes

the Actor–Critic networks, leading to faster value estimation and consistent policy improvement.

Figure 6 illustrates the profiles of the Raw Gradient Norm corresponding to the left axis, and black dashed line and the Adam Momentum corresponding to the right axis, and red dashed line. In the initial phase, the raw gradient norm spikes sharply, followed by an oscillatory decay, eventually stabilizing near 0 after 100 s. Conversely, the Adam momentum remains very low initially and stabilizes within the range of 0 to 0.02 with high-frequency, small-amplitude oscillations after 100 s. This demonstrates that the Adam optimizer effectively smooths out drastic gradient fluctuations in the early stage and maintains policy parameter stability via adaptive adjustments in the later stage. During the simulation process of reinforcement learning, the agent continuously interacts with the environment and updates randomly by sampling from the Replay Buffer. This introduces continuous random noise. As long as the network is still undergoing fine-tuning and parameter updates, the momentum mean will continue to fluctuate at a high frequency within this small range, and will not settle down to 0 like the gradient does.

In summary, the reinforcement learning training process and control performance verify the efficacy and superiority of the proposed framework. During the offline training phase, the integration of the reward function network successfully mitigates the issue of vanishing rewards inherent in early-stage exploration, preventing drastic declines in performance. This is complemented by the Adam optimization mechanism, which effectively dampens gradient steps to ensure the numerical stability and parameter convergence of the Actor–Critic networks. During the online execution phase, the agent successfully balances control precision and energy efficiency. By adapting to system dynamics and trajectory constraints, the proposed controller achieves faster error convergence and higher path efficiency compared to conventional methods. It generates smooth, non-oscillatory control inputs that prevent actuator saturation, demonstrating strong robustness and adaptability for high-performance motion control.

5. Conclusions

In this paper, an enhancing robust adaptive dynamic positioning control algorithm has been successfully developed for marine vessels. By integrating a tile-coded adaptive reinforcement learning algorithm with an adaptive damping reference model, the proposed framework effectively addresses the challenges of model uncertainties, environmental disturbances, and actuator saturation. The asynchronous deep Actor–Critic network with multi-layer perceptron achieves precise policy approximation in continuous spaces, while the online meta-driven radial basis function network accelerates convergence through dynamic reward shaping. Crucially, the introduction of a hysteresis state machine and an adaptive reference model successfully decouples high-frequency wave chattering and eliminates transient thrust saturation during large waypoint transitions. This integrated scheme provides a promising, hardware-compatible solution for high-performance and smooth autonomous marine control.

Author Contributions: Conceptualization, J.L. and W.H.; methodology, W.H.; software, W.H.; validation, J.L.; formal analysis, J.L.; investigation, J.L.; writing—original draft preparation, W.H.; writing—review and editing, J.L. and G.Z.; supervision, J.L.; project administration, G.Z.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: The paper is partially supported by the China Postdoctoral Science Foundation (2025M771628), the Postdoctoral Fellowship Program of CPSF (GZC20251126), the Natural Science Foundation of Liaoning Province (2026-BS-0216), and the Teacher Development Project of Dalian Maritime University (JF2025Z02).

Data Availability Statement: Data available on request due to restrictions. The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The authors would like to thank the anonymous reviewers for their valuable comments.

References

1. Zhai, M.; Wu, W.; Tsai, S. The Effects of Artificial Intelligence Orientation on Inefficient Investment: Firm-Level Evidence from China's Energy Enterprises. *Energy Econ.* **2025**, *141*, 108048.
2. Alagili, O.; Fernando, E.; Ahmed, S.; Imtiaz, S.; Murrant, K.; Gash, B.; Islam, M.; Zaman, H. Experimental Investigations of An Energy-Efficient Dynamic Positioning Controller for Different Sea Conditions. *Ocean Eng.* **2024**, *299*, 117297.
3. Ma, D.; Chen, X.; Ma, W.; Zheng, H.; Qu, F. Neural network Model Based Reinforcement Learning Control for AUV 3-D Path Following. *IEEE Trans. Intell. Veh.* **2014**, *9*, 893–904.
4. Shen, H.; Wu, J.; Wang, Y.; Wang, J. Reinforcement Learning-Based Robust Tracking Control for Unknown Markov Jump Systems and Its Application. *IEEE Trans. Circuits Syst. II-Express Briefs* **2024**, *71*, 1211–1215.
5. Gao, Y.; Su, S.; Zong, Y.; Zhang, L.; Guo, X. Adaptive Fuzzy Gain-Scheduling Robust Control for Stability of Quadrotors. *Appl. Math. Model.* **2025**, *138*, 115816.
6. Xie, W.; Liang, J.; Wang, Z.; Yang, J. Non-Monotonic Lyapunov Function Based Membership Function Dependent Robust Control of Takagi-Sugeno Fuzzy Systems. *Eng. Appl. Artif. Intell.* **2025**, *152*, 110785.
7. Ning, J.; Ma, Y.; Chen, C. Event-Triggered-Based Distributed Formation Cooperative Tracking Control of Under-Actuated Unmanned Surface Vehicles with Input and State Quantization. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 7081–7097.
8. Ruan, Y.; Tang, T. An optimized Youla-Kucera parametrization with time-delay compensation in multirate parallel control for disturbance rejection up to Nyquist frequency. *Control Eng. Pract.* **2025**, *158*, 106285.
9. Li, J.; Zhang, G.; Zhang, W.; Chang, T. Propeller Thrust-Saving Adaptive Control for Rotor-Assisted Vehicle via the Triggering Guidance. *IEEE Trans. Transp. Electrification* **2025**, *11*, 8538–8547.
10. Song, W.; Zuo, Y.; Tong, S. Fuzzy Optimal Event-Triggered Control for Dynamic Positioning of Unmanned Surface Vehicle. *IEEE Trans. Syst. Man Cybern.-Syst.* **2025**, *55*, 2302–2311.
11. Li, J.; Zhu, M.; Zhang, G. Integral dynamic event-triggered control for surface-underwater vehicles via an improved LVS guidance. *Ocean Eng.* **2026**, *359*, 125750.
12. Zheng, Z.; Ruan, L.; Zhu, M.; Guo, X. Robust Optimal Control of Point-Foot Biped Robots Using a Reinforcement Learning Approach. *IEEE Trans. Veh. Technol.* **2020**, *399*, 479–490.
13. Zhang, G.; Sun, Z.; Li, J.; Huang, J.; Bin, Q. Iterative Learning Control for Path-following of ASV with the Ice Floes Auto-select Avoidance Mechanism. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 13927–13938. <https://doi.org/10.1109/TITS.2025.3564279>.
14. Qu, C.; Cheng, L.; Ga Gao, S.; Huang, X. Experience Replay Enhances Excitation Condition of Neural-Network Adaptive Control Learning. *J. Guid. Control Dyn.* **2025**, *48*, 496–507.
15. Xu, X.; Huang, X.; Wang, L. Empty container repositioning problem using a reinforcement learning framework with multi-weight adaptive reward function. *Marit. Policy Manag.* **2024**, *51*, 1742–1763.
16. Zhou, X.; Han, G.; Zhou, G.; Xue, Y.; Lv, M.; Chen, A. Hybrid DQN-Based Low-Computational Reinforcement Learning Object Detection with Adaptive Dynamic Reward Function and ROI Align-Based Bounding Box Regression. *IEEE Trans. Image Process.* **2025**, *34*, 1712–1725.
17. He, Y.; Liu, Y.; Yang, L.; Qu, X. Deep Adaptive Control: Deep Reinforcement Learning-Based Adaptive Vehicle Trajectory Control Algorithms for Different Risk Levels. *IEEE Trans. Intell. Veh.* **2024**, *9*, 1654–1666.
18. Ahmad, I.; Narmeen, R.; Tabassum, H.; Becvar, Z. Intelligent Localization-Based DDPG for UAV-Borne RIS in Vehicular Communication. *IEEE Trans. Veh. Technol.* **2025**, *74*, 18168–18173.
19. Chen, X.Z.; Yin, S.H.; Li, Y.J.; Xiang, Z.R. Dynamic path planning for multi-USV in complex ocean environments with limited perception via proximal policy optimization. *Ocean Eng.* **2025**, *326*, 120907. <https://doi.org/10.1016/j.oceaneng.2025.120907>.
20. Abreu, M.; Reis, L.; Lau, N. Addressing Imperfect Symmetry: A Novel Symmetry-Learning Actor-Critic Extension. *Neurocomputing* **2025**, *614*, 128771.
21. Tagliaferri, F.; Viola, I. A Real-Time Strategy Decision Program for Sailing Yacht Races. *Ocean Eng.* **2017**, *134*, 129–139.
22. Yang, S.; Yang, L.; Wu, T.; Shi, S.; Ma, L.; Zhu, J. Multi-target automatic positioning based on angle and distance parallel measurement. *Opt. Express* **2023**, *31*, 42595–42612.

23. Zhang, G.; Yin, S.; Li, J.; Zhang, W.; Zhang, W. Game-Based Event-Triggered Control for Unmanned Surface Vehicle: Algorithm Design and Harbor Experiment. *IEEE Trans. Cybern.* **2025**, *55*, 2729–2741. <https://doi.org/10.1109/TCYB.2025.3556042>.
24. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101.
25. Emmanuel, C.; Gerald, C.; Dobri, A. Emotive Facial Recognition Using Vision Transformers for Affective Computing Applications. In Proceedings of the 2026 18th International Conference on Knowledge and Smart Technology, Pattaya, Thailand, 21–24 January 2026; pp. 182–187.
26. Guan, D.; Li, W. Enhancing pixel-level interactive editing with dense-UNet and Adam point update in diffusion models. *Appl. Intell.* **2025**, *55*, 797.
27. Rezaei, M.; Mhamdi, M.; Allili, M. Adaptive Bit Selection via Deep Reinforcement Learning for Large-Scale Image Hashing. *Electronics* **2026**, *15*, 1735.
28. Liu, A.; Wang, D.; Qiao, J. An Advanced Robust Integral Reinforcement Learning Scheme with The Fuzzy Inference System. *Int. J. Robust Nonlinear Control* **2024**, *34*, 11745–11759.
29. Wei, J.; Peng, B.; Lee, X.; Palpanas, T. DET-LSH: A Locality-Sensitive Hashing Scheme with Dynamic Encoding Tree for Approximate Nearest Neighbor Search. *Proc. VLDB Endow.* **2024**, *17*, 2241–2254.
30. Liu, H.; Zhou, W.; Wu, Z.; Zhang, S.; Li, G.; Li, X. Refining Codes for Locality Sensitive Hashing. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 1274–1284.
31. Chen, Y.; Ding, J.; Chen, Y.; Yan, D. Nonlinear Robust Adaptive Control of Universal Manipulators Based on Desired Trajectory. *Appl. Sci.* **2024**, *15*, 2219. <https://doi.org/10.3390/app14052219>.
32. Zheng, L.; Fiez, T.; Alumbaugh, Z.; Chasnov, B.; Ratliff, L. Stackelberg Actor-Critic: Game-Theoretic Reinforcement Learning Algorithms. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *36*, 9217–9224.
33. Zhang, G.; Li, Z.; Li, J.; Zhang, W.; Bin, Q. Prescribed Performance Path Following Control for Rotor-Assisted Vehicles via an Improved Reinforcement Learning Mechanism. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 17395–17405.
34. Qin, C.; Zhu, T.; Jiang, K.; Wu, Y. Integral Reinforcement Learning-Based Dynamic Event-Triggered Safety Control for Multi-player Stackelberg-Nash Games with Time-Varying State Constraints. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108317.
35. Xu, B.; Yang, C.; Shi, Z. Reinforcement Learning Output Feedback NN Control Using Deterministic Learning Technique. *IEEE Trans. N Neural Netw. Learn. Syst.* **2014**, *25*, 635–641.
36. Zhang, G.; Xing, Y.; Zhang, W.; Li, J. Prescribed Performance Control for USV-UAV via a Robust Bounded Compensating Technique. *IEEE Trans. Control Netw. Syst.* **2025**, *12*, 2289–2299. <https://doi.org/10.1109/TCNS.2025.3566320>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.