

Reinforcement Learning Applications in Unmanned Underwater Vehicles: A Review

Zinuo Tian^{1,2}, Sai Deng^{1,2}, Zhengxing Wu^{1,2} and Chao Zhou^{1,2}

Received: 07 February 2025 / Accepted: 08 May 2025

© Harbin Engineering University and Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract

The ocean, which is a key component of Earth's ecosystem, requires advanced technologies for deep and highly comprehensive exploration. Unmanned underwater vehicles (UUVs) play an important role in this task, but their development encounters great challenges due to the complex and dynamic underwater environment. Reinforcement learning (RL) has recently emerged as a promising method to improve the capabilities of UUVs. This study comprehensively reviews the implementations of RL in UUVs, with a focus on key tasks such as motion planning, navigation and control, and multiagent coordination. We investigate current difficulties and emerging trends, as illustrated by a case study. This review aims to provide a foundation for RL-based control and decision-making in UUVs and offer actionable insights for advancing studies in this rapidly evolving domain.

Keywords Unmanned underwater vehicles; Reinforcement learning; Motion planning; Navigation control; Multiagent collaboration

1 Introduction

Oceans cover nearly 360 000 000 km², which accounts for two-thirds of the Earth's surface area of 510 000 000 km² (Yuh, 2000). Furthermore, around 37% of the world's population resides within 100 km of the coastline (Cohen et al., 1997). Unmanned underwater vehicles (UUVs), as intelli-

gent equipment capable of conducting tasks underwater, are vital in substituting for humans while allowing precise data acquisition and expanding research scopes (Wynn et al., 2014). For instance, UUVs can be utilized for deep-sea terrain mapping, hydrological parameter measurement, and biodiversity surveys in marine scientific research (Yoerger et al., 2007). In ocean engineering, they can perform tasks including underwater structure inspection, pipelinelaying assistance, and subsea cable maintenance (Mai et al., 2016). In national defense, they can be adopted for underwater reconnaissance, anti-submarine warfare, and maintenance (Terracciano et al., 2020). Figure 1 displays the typical application of UUVs in different tasks.

However, the underwater environment is complex, dynamic, and full of uncertainties; this condition introduces

Article Highlights

- This study the transformative potential of reinforcement learning (RL) in advancing intelligent control and decision-making for unmanned underwater vehicles (UUVs), specifically in handling difficulties induced by complex and dynamic underwater environments.
- RL implementations in UUVs are comprehensively reviewed, including motion planning, navigation and control, and multiagent collaboration.
- A comprehensive analysis of state-of-the-art RL algorithms is conducted, such as Q-learning, Deep Q-Networks, deep deterministic policy gradient (DDPG), proximal policy optimization, and their variants, with a focus on their adaptations for underwater tasks.
- This study features a practical case study demonstrating the application of a DDPG-based RL framework for target tracking control in a biomimetic robotic shark, which the potential of RL in real-world underwater scenarios.

✉ Sai Deng
sai.deng@ia.ac.cn

¹ The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

² The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

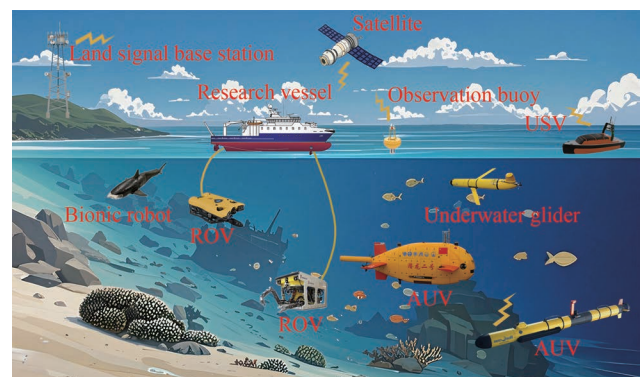


Figure 1 Typical applications of UUVs in different tasks (Background: AI-generated via Doubao (<https://www.doubao.com/>). UUV images: Reference designs from DEEPINFAR, FET ROVs, and SIACAS Qianlong-II AUV.)

great challenges to the navigation, perception, path planning, and control capabilities of UUVs (Wong et al., 2018). Traditional approaches, such as proportional–integral–derivative (PID) control (Borase et al., 2021) and model predictive control (MPC) (Fernandez and Hollinger, 2017), typically encounter complexities in parameter tuning, lack adaptability, or depend on precise environmental models. In real-world underwater scenarios, developing accurate models is frequently difficult due to water current disturbances, visibility changes, and sensor noise (Huang et al., 2020). Moreover, traditional modeling and control methods present limitations in robustness and adaptability when confronted with complex, dynamic, and unknown scenarios (Zhang et al., 2022). Thus, they fail to meet the demands for effective and safe operation. These restrictions reveal a critical research gap and underscore the urgent need for advanced control and decision-making frameworks that can efficiently manage the complexities and uncertainties of the underwater environment.

In response to these restraints, reinforcement learning (RL), which is an emerging branch of artificial intelligence, provides a promising paradigm. As presented in Figure 2, RL involves an agent that learns optimal behaviors by interacting with an environment and maximizing cumulative rewards through trial-and-error approaches (Sutton and Barto, 2018). Compared with traditional methods, the key advantages of RL are highly relevant to UUVs: it can learn efficient policies without explicit environmental models, handle high-dimensional state and action spaces inherent in complex robotic systems, and adapt to varying conditions (Christensen et al., 2022). Therefore, utilizing RL improves the autonomy, robustness, and efficiency of UUVs in challenging underwater operations.

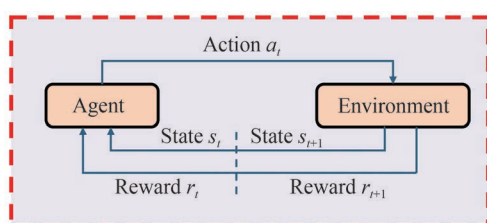


Figure 2 Agent–environment interaction

The application of RL to UUVs is a rapidly evolving field, which is marked by diverse methods and accumulating successes. However, this implementation also encounters specific underwater hurdles, including data scarcity and the sim-to-real gap. Such rapid progress, which is coupled with unique complexities, requires a comprehensive and critical synthesis of the current state-of-the-art techniques. This review aims to fulfill this need by systematically investigating the implementations of RL across key UUV tasks—motion planning, navigation and control, and multiagent collaboration. We assess prominent methods, pinpoint critical challenges, and discuss promising future research directions. The following key questions are specifically addressed:

1. Which RL algorithms exhibit optimal efficacy for specific operational tasks of UUVs?
2. What fundamental practical obstacles hinder the deployment of RL on the operational platforms of UUVs?
3. Which research priorities are most important for bridging the gap between the theoretical potential of RL and the robust real-world performance of UUVs?

Ultimately, this review aims to provide a robust foundation and actionable insights for advancing the autonomy of UUVs based on RL.

To guarantee a comprehensive and relevant overview, we systematically queried major academic databases, such as IEEE Xplore, ScienceDirect, Google Scholar and Web of Science. The search used combinations of keywords encompassing “Unmanned Underwater Vehicle”, “Autonomous Underwater Vehicle”, “Remotely Operated Vehicle”, “Reinforcement Learning”, “Deep Reinforcement Learning”, and their abbreviations, which were coupled with task-specific terms including “motion planning”, “path planning”, “navigation”, “guidance”, “multiagent”, and “collaboration”.

The literature selection selected peer-reviewed journal articles and conference proceedings published between 2020 and 2025. Seminal earlier works constructing foundational concepts or algorithms directly relevant to the applications of UUVs were also included. Papers were chosen based on their relevance to the implementation of RL to solve specific operational challenges of UUVs, specifically in motion planning, navigation and control, and multiagent systems. Works focusing purely on theoretical RL developments without UUV application, RL implementations in unrelated domains, or non-peer-reviewed sources were excluded unless they presented highly significant contributions. This systematic approach guarantees that the review synthesizes pertinent and contemporary research while acknowledging foundational contributions.

The remainder of the paper is organized as follows. Section 2 offers an overview of the fundamental theories in the realms of UUVs and RL, with the aim of providing readers with a basic framework of the two research areas. Section 3 elaborates on the current application status of RL in three key tasks of UUVs: motion planning, navigation and control, and multiagent collaboration. Section 4 chooses a biomimetic robotic shark as a typical case study for analysis. Section 5 reviews the main challenges and proposes future research directions. Section 6 concludes the study.

2 Overview of UUVs and RL

2.1 Overview of UUVs

2.1.1 Classification and application scenarios of UUVs

UUVs are autonomous or semi-autonomous systems that are particularly designed to perform specific tasks in under-

water environments. They are extensively applied in marine science, industry, military, and environmental protection. According to their autonomy and operation mode, UUVs can be primarily classified into the following categories (Yuh, 2000):

1. **Autonomous Underwater Vehicles (AUVs):** As displayed in Figure 3(a), AUVs are untethered UUVs with autonomous navigation, decision-making, and control capabilities. They can conduct pre-programmed tasks or autonomously plan actions depending on their sensor perception of the environment without requiring real-time control through a physical connection (Li and Du, 2021). The high degree of autonomy of AUVs makes them suitable for large-scale and long-duration underwater missions, including ocean floor mapping, hydrological data collection, and target search (Paull et al., 2014).

2. **Remotely Operated Vehicles (ROVs):** As illustrated in Figure 3(b), ROVs are UUVs connected to a surface vessel through an umbilical cable and are remotely controlled by an operator. ROVs are mainly equipped with cameras, manipulators, and other equipment; they can be employed for underwater observation, sampling, and maintenance tasks (Macreadie et al., 2018). The operational range of ROVs is limited due to the need for an umbilical cable for power and communication. However, they provide real-time video feedback and precise control, making them advantageous in tasks needing fine manipulation and high-resolution observation, including deep-sea pipeline inspection, shipwreck salvage, and underwater archaeological excavation (Baharin and Kamarudin, 2025).

3. **Hybrid Underwater Vehicles (HUVs):** As presented in Figure 3(c), HUVs integrate the features of AUVs and ROVs. They can be remotely controlled through an umbilical cable or autonomously operated in an untethered mode (Bao et al., 2023). This design enables HUVs to utilize the large-scale operational capabilities of AUVs and the precise control capabilities of ROVs. Thus, HUVs provide greater flexibility and adaptability in complex underwater missions.

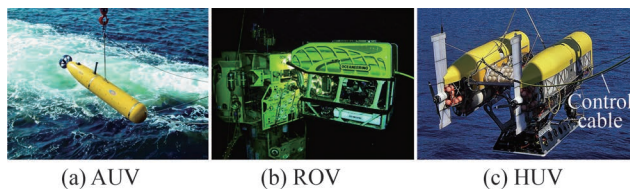


Figure 3 Classification of UUVs

2.1.2 Complexities of underwater tasks

UUVs operate in a highly complex and uncertain environment, which is mainly reflected in the following aspects:

1. **Dynamic and Uncertain Underwater Environment:** The underwater environment is affected by different factors, such as ocean currents, tides, internal waves, and tur-

bulence; this condition results in strong dynamics and unpredictability (Thorpe, 2007). The aforementioned factors can significantly and adversely affect the motion control and navigation of UUVs. For instance, strong currents can lead to deviations from the planned trajectory, and turbulence can result in unstable attitudes (Fossen, 2011). In addition, underwater biological activities, water quality changes, and reduced visibility can interfere with the sensor data and task execution of the robot (Cong et al., 2021).

2. **Limited Sensor Types and Noisy Data:** UUVs mainly depend on acoustic and visual sensors for environmental perception and localization due to the strong absorption of electromagnetic waves by water (Huy et al., 2023). However, acoustic sensors are vulnerable to multipath effects and environmental noise, which leads to significant noise and uncertainty in the data (Coates, 1990). Visual sensors such as cameras are also influenced by light attenuation, scattering, and absorption in the underwater environment, which results in degraded image quality and a very limited effective detection range (Schettini and Corchs, 2010). Electromagnetic sensors are occasionally used but suffer from severe absorption and interference, which further limit their applicability in underwater environments. Moreover, although biological sensors offer unique detection capabilities, they are highly sensitive to temperature variations and prone to biofouling, making calibration and long-term deployment difficult. Furthermore, obtaining high-quality underwater sensor data typically needs costly equipment and complex data processing techniques; this requirement further increases the difficulty of underwater tasks (Dunbabin and Marques, 2012).

3. **Underwater Communication and Energy Challenges:** Underwater wireless communication primarily depends on acoustic waves, which have limited bandwidth and high latency and are prone to environmental interference. This restriction introduces challenges for the remote control and data transmission of UUVs (Morozs et al., 2020). Meanwhile, the strict constraints of UUVs on space and weight further restrict energy storage capacity, which necessitates researchers to carefully consider the energy consumption of task planning and control strategies to maximize endurance and operational efficiency (Kelasidi et al., 2015).

The specific diagram of the complexity factors of the underwater tasks is displayed in Figure 4. Overall, the field of UUVs encounters numerous challenges. Exploring more intelligent, efficient, and robust control and optimization methods is essential for advancing the development of UUV technology.

2.2 Overview of RL

2.2.1 Origin and development of RL

Machine learning includes supervised learning, unsupervised learning, and RL (Singh et al., 2022), as compared in Figure 5. RL, as inspired by behavioral psychology, pro-

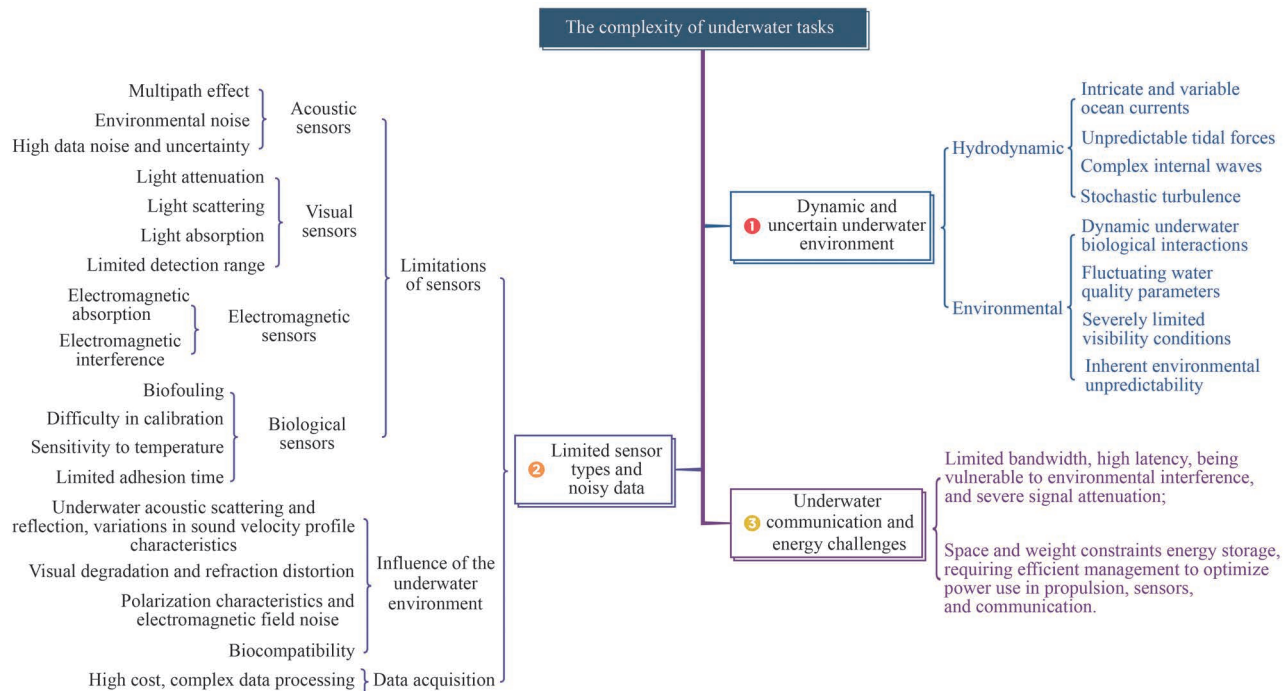


Figure 4 Complexities of underwater tasks

vides a framework for agents to learn optimal sequential decisions through trial-and-error interaction with the environment (Sutton and Barto, 2018). It is theoretically based on dynamic programming (Bellman, 1966) and Markov decision processes (Howard, 1960).

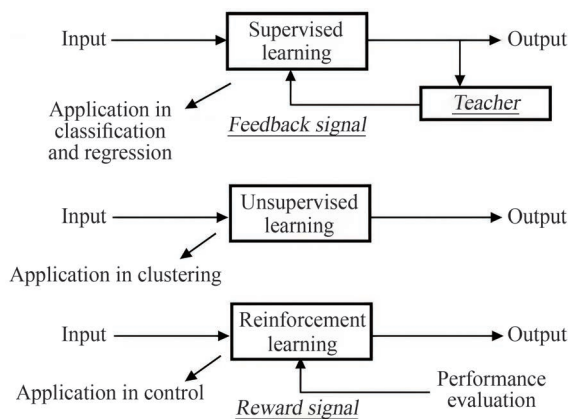


Figure 5 Framework of machine learning (Singh et al., 2022)

Early RL methods were typically tabular, which limits their application in the large and continuous state space typical in UUVs (Bertsekas, 1996). A key advancement was Q-learning (Watkins, 1989), which is a model-free algorithm that learns action values. Around the same time, Sutton (1988) proposed the temporal difference learning algorithm, which provides important theoretical support for the development of Williams (1992) further proposed the REINFORCE algorithm. Unlike value function-based

methods, this algorithm is a policy gradient-based RL algorithm that directly parameterizes the policy and optimizes the parameters through a gradient ascent.

In recent years, the integration of deep neural networks (DNNs) has marked the era of deep RL (DRL) (LeCun et al., 2015). DNNs serve as powerful function approximators, which enable RL to handle the high-dimensional sensory inputs and complex dynamics encountered by UUVs. Landmark algorithms such as Deep Q-Networks (DQN) (Mnih, 2013) demonstrate superhuman performance in complex tasks by integrating Q-learning with deep learning. Subsequent developments focus on enhancing stability and performance, encompassing variations such as Double DQN (Van Hasselt et al., 2016) and Dueling DQN (Wang et al., 2016). Policy gradient methods, including trust region policy optimization (Schulman, 2015) and proximal policy optimization (PPO) (Schulman et al., 2017), offer stable ways to optimize policies directly, specifically for continuous actions. Algorithms such as deep deterministic policy gradient (DDPG) (Lillicrap, 2015) specifically address continuous action spaces inherent in robotics control. The success of AlphaGo (Silver et al., 2016) and AlphaZero (Silver et al., 2018) further emphasizes the potential of DRL for complex decision-making, which transforms DRL to new heights.

2.2.2 Main RL algorithms

According to learning methods, RL algorithms can be divided into the following categories:

1. Value-Based Algorithms: These algorithms indirectly learn the optimal policy by learning the state-value func-

tion $V(s)$ or the action-value function $Q(s, a)$.

1) Q-learning (Watkins, 1989): It is an off-policy algorithm that iteratively updates the Q -function. The update rule for a transition from state s_t to s_{t+1} after taking action a_t and receiving reward r_{t+1} is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \cdot \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right] \quad (1)$$

where α refers to the learning rate, γ represents the discount factor, and the term $\max$ denotes the maximum estimated future value achievable from state s_{t+1} , which reflects the greedy strategy.

2) SARSA (Rummery and Niranjan, 1994): It is an on-policy algorithm that uses the actual next action a_{t+1} to update Q :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \cdot Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \quad (2)$$

Comparing (1) and (2) indicates that the update of SARSA depends on the tuple $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$, which is the basis for its name. This condition makes its learning more conservative and dependent on the current policy.

3) Deep Q-Networks (Mnih, 2013): DQN extends Q-learning using a DNN parameterized by θ to approximate the action-value function $Q(s, a; \theta) \approx Q^\pi(s, a)$. It also incorporates experience replay and target networks to stabilize training and improve sample efficiency when dealing with high-dimensional state representations, such as images.

2. Policy Gradient Algorithms: These algorithms directly parameterize the policy $\pi(a|s; \theta)$ (stochastic) or $\mu(s; \theta)$ (deterministic) with parameters θ via gradient ascent on the expected cumulative reward $J(\theta)$.

1) REINFORCE (Williams, 1992): It is a Monte Carlo policy gradient method that updates policy parameters θ based on the gradient.

2) PPO (Schulman et al., 2017): It is an advanced actor-critic algorithm that enhances training stability by adopting a clipped surrogate objective function or an adaptive KL penalty to restrict large policy updates.

3. Actor-Critic Algorithms: These methods integrate value-based and policy gradient approaches. They use an actor network to represent the policy, and a critic network to estimate a value function, which helps evaluate the actions of the actor and reduce variance in policy updates (Konda and Tsitsiklis, 1999). DDPG and Soft Actor-Critic (SAC) are widely adopted examples, which are specifically effective for the continuous control of UUVs.

4. Model-Based RL: Unlike the previous three model-free methods, these algorithms attempt to learn a model of the dynamics of the environment $p(s_{t+1}, r_{t+1}|s_t, a_t)$. This

learned model can then be utilized for planning or generating simulated experience, which potentially improves sample efficiency (Feinberg et al., 2018; Luo et al., 2021; Sutton, 1990). However, precisely modeling the complex and unpredictable underwater environment is still highly challenging for the applications of UUVs.

Each category exhibits its own trade-offs and is best suited for particular scenarios. Value-based methods can be sample-efficient but struggle with continuous actions. Policy gradient methods manage continuous actions well but are less stable. Actor-critic methods strike a balance between the two. Model-based methods promise efficiency but rely on model accuracy, which is a major concern in underwater environments. The choice of algorithm depends on the task requirements of UUVs. The chosen algorithm should be capable of balancing factors such as action space continuity, required sample efficiency, and robustness to environmental uncertainty. A summary of classical RL algorithms is presented in Table 1.

After the fundamental principles of UUVs are established, the complex operating environment is delineated, and the core algorithms of RL are reviewed, the next section pivots to the central focus of this study: the concrete application of these RL techniques to address the specific challenges in UUV operations. The previously discussed constraints of traditional methods, along with the demonstrated potential of RL, require a deeper examination of the employment of RL across key functionalities of UUVs.

3 Applications of RL in UUVs

RL, as a data-driven intelligent decision-making method, offers a new approach to solving robot control and decision-making issues in complex, dynamic, and unknown environments. Owing to the complexity of the underwater environment, such as ocean currents, unpredictable obstacles, and limited visibility, along with the noise and uncertainty of sensor data and the highly nonlinear and coupled dynamics of UUVs, traditional control and planning methods frequently struggle to achieve optimal performance. In particular, these approaches have difficulty handling real-time adaptability to dynamic environmental changes, robustness against sensor noise and measurement errors, and efficacy in handling high-dimensional and nonlinear decision-making problems common in UUV operations. In recent years, DRL, with its learning capability in high-dimensional state spaces, complex action spaces, and uncertain environments, has emerged as a powerful paradigm for improving the autonomy of UUVs (Christensen et al., 2022).

This chapter discusses the research status of RL in UUVs from three aspects: RL-based motion planning, navigation and control, and multiagent collaboration.

Table 1 Typical RL algorithms

Items	Policy learning method	Off-policy	Policy representation	Applicable action space
Temporal difference	Value Function	Yes (Except SARSA)	Table/Function Approximation	Discrete
TD(λ)	Value Function	Yes (Except SARSA)	Table/Function Approximation	Discrete
Q-learning	Value Function	Yes	Table/Function Approximation	Discrete
SARSA	Value Function	No	Table/Function Approximation	Discrete
DQN	Value Function	Yes	Neural Network	Discrete
Double DQN	Value Function	Yes	Neural Network	Discrete
REINFORCE	Direct Policy	Yes	Neural Network/Function Approximation	Both
PPO	Direct Policy	Yes	Neural Network	Continuous
Actor-Critic	Combination of Policy and Value Function	Both	Neural Network + Function approximation	Both
A2C	Combination of Policy and Value Function	Yes	Neural Network + Function Approximation	Both
TRPO	Direct Policy	Yes	Neural Network	Continuous
DDPG	Combination of Policy and Value Function	Yes	Neural Network	Continuous
SAC	Combination of Policy and Value Function	Yes	Neural Network	Continuous
Dyna	Value Function	Both	Table/Function Approximation	Discrete
MVE	Value Function	Both	Table/Function Approximation	Discrete
SLBO	Combination of Policy and Value Function	Yes	Neural Network	Both

3.1 Motion planning

Motion planning is vital for the autonomous operation of UUVs, and it aims to generate safe and efficient trajectories based on task requirements and environmental information. Traditional approaches typically depend on accurate environmental models and manually designed rules, which causes difficulty in their adaption to complex and dynamic underwater environments. RL-based techniques often model the motion planning problem as a Markov decision process and learn the optimal motion strategy through interaction. According to the task hierarchy, motion planning of UUVs can be divided into path planning (finding a collision-free route), trajectory planning (generating a smooth and dynamically feasible path), and task planning. This section primarily focuses on the first two aspects, while task planning will be introduced in Section 3.3 on multi-agent collaboration.

1. **Early Approaches and Static Environments:** Initial RL implementations commonly focused on path planning in static or well-defined environments. For example, Yoo and Kim (2016) integrated Q-Learning with a smoothing algorithm, which considers the kinematics and dynamics of UUVs to produce feasible trajectories. While demonstrating basic feasibility, tabular Q-learning methods suffer from the curse of dimensionality, which limits their applicability to complex and continuous state spaces inherent in UUVs. To enhance convergence in such tabular settings, Zhu et al. (2019) proposed an adaptive Q-Learning algo-

rithm that dynamically modifies the exploration rate based on environmental complexity. This algorithm demonstrates faster convergence than standard Q-learning but is still confined to discretized state spaces. Wang et al. (2024b) aimed to enhance the convergence and path quality of Q-learning using a neural network-smoothed approach (NSFQ) with Bezier curve smoothing. This approach improves heading stability but retains the core limitations of value-based methods in high-dimensional continuous domains.

2. **Deep RL for Complex Planning:** The advent of DRL has allowed addressing high-dimensional problems. Sun et al. (2020) used hierarchical DQN with prioritized experience replay for 3D path planning, which efficiently reduces state space dimensionality. Although designing optimal hierarchies can be difficult, this hierarchical structure represents a useful technique. Yao et al. (2021) addressed time-optimal path planning in dynamic current fields using a two-layer interval optimization scheme, which explicitly handles uncertainty in current predictions—a critical factor often simplified in other studies.

3. **Actor–Critic Methods for Dynamic Environments and Continuous Control:** A clear trend involves the adoption of actor–critic algorithms such as DDPG and PPO, which are suitable for continuous action spaces in UUV control. Lan et al. (2022) adopted DDPG for path planning, which adapts to varying currents by reformulating the MDP. Lyu et al. (2023) constructed an end-to-end perception–planning–execution method using a two-layer DDPG with hindsight experience replay, which enhances

efficacy in dynamic environments. DDPG demonstrates effectiveness but exhibits sensitivity to hyperparameters and potential instability during training. This limitation requires careful tuning in practice. For this purpose, Wang et al. (2025) proposed a gated recurrent unit (GRU)–PPO model leveraging local visual input. By using CNNs and GRUs to process sensor data, this approach integrates perception more directly into planning, which enables local path optimization and obstacle avoidance. However, similar to many vision-based RL approaches, its performance heavily depends on the quality and realism of simulated visual data, which highlights the sim-to-real transfer challenge.

4. Hybrid and Novel Approaches: Some studies investigate hybrid methods in complex environments. Yan et al. (2023) proposed a novel RL-based tuna swarm optimization, where RL guides agents in selecting strategies within a bio-inspired optimization framework, and the optimization process is smoothed by B-splines. This hybridization provides an interesting direction but needs careful balancing between the exploration mechanisms of RL and the exploitation focus of the optimizer.

Overall, RL for motion planning has advanced from tabular methods in static environments to DRL capable of managing continuous spaces, dynamic conditions, and even integrating perception. Table 2 offers a comparative overview of these works. However, challenges remain, such as hyperparameter sensitivity, the need for robust perception, and the persistent sim-to-real gap, especially for approaches depending heavily on simulated sensor data.

3.2 Navigation and control

Navigation (ascertaining pose and velocity) and control (performing maneuvers) are core capabilities of UUVs, where RL exhibits considerable promise. Traditional model-based controllers, such as PID and MPC, typically struggle with uncertain hydrodynamics, environmental disturbances, and sensor noise. The data-driven and model-free nature of

RL provides potential for enhanced adaptability and robustness. As displayed in Figure 6, RL can be integrated with traditional controllers or leveraged for end-to-end learning (Tong et al., 2023).

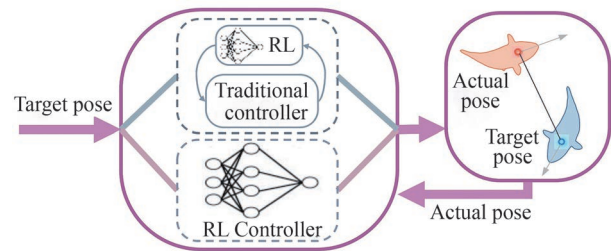


Figure 6 Two application methods of reinforcement learning in UUV control (Tong et al., 2023)

1. Foundational and Tracking Control: Early works, including Gaskett et al. (1999), established the feasibility of model-free RL for basic thruster control with continuous states and actions. El-Fakdi and Carreras (2013) validated a two-step sim-to-real transfer using natural actor–critic for vision-based cable tracking, which is a valuable early example of bridging the reality gap. Yu et al. (2017) adopted DDPG for trajectory tracking, which exhibits improved accuracy over PID in complex scenarios. However, the base DDPG algorithm can result in instability, which motivates further refinements. For example, Sun et al. (2021) augmented DDPG with potential fields for safety and SumTree prioritized replay for efficient learning in deep-sea navigation, although their validation was limited to 2D. Wang et al. (2024c) realized robust depth tracking for underactuated AUVs using an adversarial DRL approach with LSTM state prediction. This method demonstrates superior performance over PID in tank experiments by explicitly training for robustness against state transition variations—a key advantage over techniques assuming perfect state knowledge. Similarly, Wang et al. (2024a) improved DDPG with a twin-critic design and an integral compensator to

Table 2 Comparison of representative RL-based motion planning methods for UUVs

Tasks	Algorithm	Main contribution	Validation
Trajectory Planning based on Kinodynamics (Yoo and Kim, 2016)	Q-Learning+Smoothing	Consider kinematics, dynamics, and currents, and use smoothing for feasible trajectory	Simulation with Currents
Path Planning in Static Env. (Zhu et al., 2019)	Adaptive Q-Learning	Dynamically adjust greedy rate based on environmental complexity model	Simulation
3D Path Planning (Sun et al., 2020)	Hierarchical DQN+PER	Hierarchical training reduces state space dimensionality; prioritized replay improves efficiency	Simulation
Time-Optimal Path Planning in Dynamic Env. (Yao et al., 2021)	Interval Optimization	Two-layer scheme handles uncertain dynamic currents using interval ocean model	Simulation with Dynamic Currents
Path Planning in Complex Env. (Yan et al., 2023)	RL+Tuna Swarm Opt.	Hybrid approach using RL for agent strategy selection within swarm optimization and B-spline for smoothing	Simulation
Local Trajectory Planning in Dynamic Env. (Lyu et al., 2023)	Two-Layer DDPG+HER	End-to-end perception–planning–execution; hindsight replay improves dynamic efficiency	Simulation in Dynamic Env.
Local Path Optimization in Dynamic Env. (Wang et al., 2025)	GRU-PPO (Local Vision)	Uses CNN+GRU for visual feature extraction and state mining; PPO for local optimization	Vision-based Simulation

enhance the path following accuracy and robustness of USVs, as demonstrated in experiments.

2. Addressing Model Uncertainty and Spatio–Temporal Changes: Ma et al. (2024) introduced an actor–model–critic architecture for 3D path following. By learning a state transition model alongside the policy and value function, this method seeks to better understand and adapt to spatio-temporal environmental changes. The approach potentially improves sample efficiency over purely model-free methods, although the accuracy of the learned model itself is still problematic.

3. Complex Maneuvers and End-to-End Learning: RL is increasingly adopted for complex and high-stakes maneuvers. Pereira and Pinto (2024) constructed an end-to-end RL approach for the docking of autonomous surface vehicles in unknown environments. This method introduces the concept of illegal actions within an actor–critic framework to accelerate learning. This work is notable for its real-world verification on a critical task, although the learned policy might be highly task-specific.

4. Biomimetic UUV Control: In recent years, biomimetic UUVs have gathered increasing attention due to their capability to mimic the swimming patterns of underwater creatures; thus, they can achieve longer endurance, better low-speed maneuverability, and significantly reduced operating noise (You et al., 2020). A distinct trend encompasses adopting RL to control biomimetic UUVs. The unconventional propulsion mechanisms of these UUVs introduce unique challenges for traditional modeling. Su et al. (2019) integrated RL with active disturbance rejection control for biomimetic glider attitude control. Zhang et al. (2023) used Q-learning for manta ray depth control, while Vu et al. (2023) utilized DDPG to optimize central pattern generator parameters for robotic fish locomotion. Sun et al. (2023) used DRL and imitation learning for the formation control of schooling fish robots within high-fidelity CFD simulations. These works emphasize the potential of RL for controlling complex and hard-to-model systems, but they typi-

cally depend heavily on simulation due to the experimental cost and complexity of biomimetic prototypes.

5. Leveraging Demonstrations and Vision: Imitation learning is gaining considerable attention. Chen et al. (2024) proposed an importance-reweighted GAIL algorithm that allows AUVs to learn control strategies from expert demonstrations of varying quality, which enhances generalization. Politi et al. (2024) realized efficient 3D navigation using DRL integrated with real-time visual sensor data for obstacle avoidance in unknown environments. These techniques mitigate the data sparsity and exploration challenges by utilizing prior knowledge (demonstrations) or richer sensor modalities (vision). However, they introduce dependencies on demonstration quality or sensor performance and sim-to-real transfer.

6. Energy Efficiency and 3D Navigation: Wen et al. (2024) tackled the crucial aspect of energy efficiency in 3D AUV navigation using a policy feature-based SAC variant. This approach incorporates an energy consumption model and an adaptive reward function to balance energy use and travel distance. This consideration is critical for long-duration missions and is commonly ignored in purely performance-driven RL objectives. Hadi et al. (2024) employed DRL for formation motion planning and control, with a focus on collision avoidance strategies.

In navigation and control, the trend is clearly toward sophisticated DRL algorithms capable of managing continuous control, complex dynamics, and specific maneuvers. Table 3 summarizes key studies. Integrating techniques such as imitation learning, adversarial training, and learned models shows promise for enhancing robustness and sample efficiency. Moreover, applications to biomimetic UUVs are expanding. However, real-world verification remains limited compared with simulation studies. Key shortcomings stem from the sample inefficiency of RL and the difficulty in reward function design to capture complex objectives such as safety and efficiency simultaneously.

Table 3 Comparison of representative RL-based navigation and control methods for UUVs

Objectives	Algorithm	Main contribution	Validation
Vision-based Cable Tracking (El-Fakdi and Carreras, 2013)	Natural Actor–Critic	Two-step sim-to-real transfer for autonomous cable tracking	Simulation and Real-world
Bionic Glider Attitude Control (Su et al., 2019)	RL+ADRC	Combines RL with active disturbance rejection control to handle currents	Simulation
Safe Navigation in Deep Sea (Sun et al., 2021)	DDPG+Potential Fields+SumTree	Improves DDPG with potential fields for safety and SumTree for prioritized replay	2D Simulation
Formation Control of Schooling Fish Robots (Sun et al., 2023)	DRL+Imitation Learning	High-fidelity CFD environment for training; composite reward; handles complex flows	Simulation with CFD
AUV Control Strategy Learning (Chen et al., 2024)	IR-GAIL (Imitation Learning)	Importance reweighted GAIL learns from variable quality demonstrations; better generalization	Simulation
Underactuated Depth Tracking (Wang et al., 2024c)	Adversarial DRL+LSTM	LSTM for state prediction; adversary enhances robustness against disturbances and dynamics	Towing Tank Experiment
End-to-End ASV Docking (Pereira and Pinto, 2024)	Actor-Critic+illegal Actions	Introduces illegal actions to accelerate learning for complex docking maneuvers	Real ASVs

3.3 MultiAgent collaboration

The limitations of single-UUV systems motivate the use of multi-UUV systems for improved efficiency, robustness, and coverage. However, this utilization introduces complexities in coordination, communication, and decentralized decision-making. Avoiding resource conflicts and waste while guaranteeing maximum efficiency of formation task execution is the main direction of multiagent collaboration research (Zhang et al., 2024).

1. Task Allocation: Traditional allocation methods, including the Hungarian algorithm and auction algorithm, are commonly based on optimization algorithms in operations research, which require global information and high computation. Meanwhile, RL offers adaptive and decentralized alternatives. The principles of task allocation are presented in Figure 7. Hou et al. (2023) utilized asynchronous advantage actor–critic for multi-AUV task allocation. The independent learning approach demonstrates feasibility but may struggle to capture complex inter-agent dependencies effectively. To this end, Zhang et al. (2024) used value decomposition networks within a centralized training with decentralized execution approach for joint task allocation and collision avoidance, which improves efficiency. Yu and Zhang (2024) adopted multiagent twin delayed DDPG for dynamic collaborative task decision-making. These methods usually achieve better coordination than independent learners but encounter scalability issues as the number of agents increases. They also frequently assume reliable communication during training, which contrasts with the limitations of underwater communication in reality. Xu et al. (2024) later addressed scalability by transforming the NP-hard multiagent problem into a finite-time Markov game. They also proposed an efficient offline RL training framework, which leverages previously collected data but is potentially sensitive to dataset quality.

2. Formation Control and Maneuvering: Several differ-

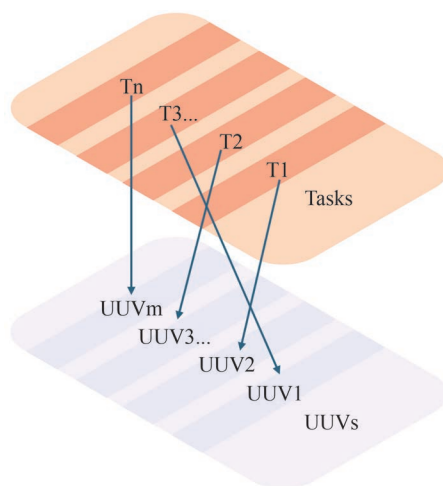


Figure 7 Multi-UUVs' task allocation principle

ent formation patterns are presented in Figure 8. Traditional formation approaches typically depend on predefined rules or accurate models, while MARL enables adaptive behavior adjustment. Wei et al. (2023b) proposed a goal-based hierarchical RL (HRL) algorithm for the path planning of multi-USV formation, which was validated through experiments. HRL is a promising approach for managing complexity in MARL. It simplifies learning by operating at different levels of abstraction, although designing the hierarchy itself is nontrivial. Han et al. (2023) presented a distributed DRL approach using independent DDPG with adaptive RBFNs to approximate dynamics for USV formation control, which highlights efforts to handle system dynamics within the learning framework. Zhu et al. (2024) introduced a hierarchical software-defined networking MARL framework with an advantage–attention actor–critic architecture for large-scale AUV cluster control. This method demonstrates enhanced convergence, tracking, and scalability. Thus, it directly tackles the crucial scalability limitation of many MARL algorithms.

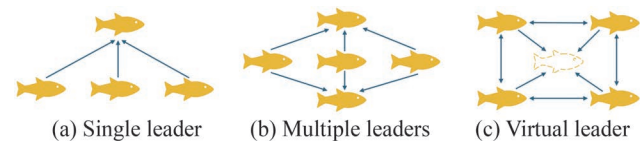


Figure 8 Several different formation patterns

3. Complex Coordination Tasks such as Hunting and Confrontation: MARL is also being implemented for more demanding coordination tasks. Xu et al. (2021) first adopted DRL to confrontational tasks under limited perception and communication by using a novel encoding network. On this basis, Xia et al. (2023) realized multitarget hunting with USV fleets using distributed PPO to handle partial observability. Wang et al. (2022) established a collaborative target capture scheme with dynamic formations and adversarial training for robustness. Wei et al. (2023a) analyzed a hunting system in turbulent and adversarial environments by introducing differential game theory into MARL while considering environmental factors and communication delays. These works facilitate the use of MARL in complex, dynamic, and uncertain scenarios. However, they typically require sophisticated algorithm design and careful reward shaping.

The application of MARL to UUVs is a rapidly growing area, which shifts from basic coordination to complex tasks such as dynamic allocation, intricate formations, and adversarial engagements. Table 4 contrasts representative works on the application of RL and MARL to coordinate multiple UUVs. The trend transitions toward more sophisticated MARL paradigms to overcome the constraints of independent learning and handle scalability and communication restraints. Partial observability, robustness to environmental dynamics, and effective decentralized execution remain key directions. The reliance on simulation is remarkably

Table 4 Comparison of representative RL-based multi-agent collaboration methods for UUVs

Tasks	Algorithm	Main contribution
Collaborative Target Capture (Wang et al., 2022)	MADRL+Adversarial	Dynamic formation methods, adaptive allocation; adversarial training for robust surrounding
Multi-USV Formation Path Planning (Wei et al., 2023b)	Goal-Based Hierarchical RL+APF	Hierarchical structure simplifies planning; improved APF aids training; composite reward
Joint Task Allocation and Collision Avoidance (Zhang et al., 2024)	Value Decomposition Networks	Improves efficiency and autonomous avoidance using VDN for centralized training
Dynamic Collaborative Task Decision (Yu and Zhang, 2024)	Multi-Agent Twin Delayed DDPG	Addresses long planning time, large data needs; potential scalability issues noted
Multi-Agent NP-Hard Problem (Xu et al., 2024)	Offline RL Framework	Transforms problem to finite-time Markov game; offline RL improves scalability and generalization
Large-Scale AUV Cluster Control (Zhu et al., 2024)	Hierarchical SDN MARL+A2C	Hierarchical framework with advantage-attention actor-critic; improves speed, accuracy, and scale

pronounced in MARL due to the compounded complexity and cost of real-world multirobot experiments, which further emphasizes the sim-to-real challenge.

4 Case study: application of RL in the target tracking control of a biomimetic shark

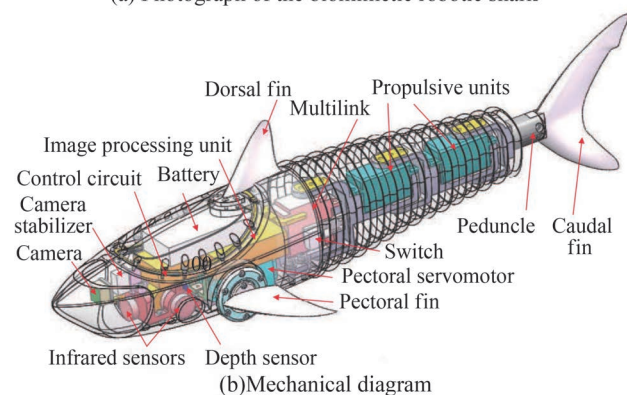
In this chapter, we present a case study on target tracking control within a 2D plane. This study uses the biomimetic robotic shark developed by the Institute of Automation, Chinese Academy of Sciences, as displayed in Figure 9 (Yang et al., 2016). The key parameters of the biomimetic robotic shark, such as size and mass, are summarized in Table 5. This case study is based on the research framework of Yu et al. (2021).

Traditional methods of target tracking control for UUVs usually rely on dynamic models. However, the hydrodynamic model of fish swimming is complex and hard to establish accurately, especially for biomimetic robotic fish, where the propulsion method introduces more prominent model uncertainty. In addition, the complex and dynamic underwater environment, with its numerous interference factors, further complicates the target tracking task. RL provides a model-free approach by enabling agents to interact with their environment and modify behavior through reward maximization. This chapter performs a case study based on the control characteristics and task requirements of this platform. This analysis serves as an example to show readers the basic ideas of using RL to solve related problems of UUVs.

In this case study, the target tracking control problem is analyzed for a biomimetic robotic fish with multiple joints that are connected in series. Given that heading control of the biomimetic robotic shark involves continuous action spaces, the DDPG algorithm is used. DDPG, as a model-



(a) Photograph of the biomimetic robotic shark



(b) Mechanical diagram

Figure 9 Biomimetic robotic shark of the institute of automation, chinese academy of sciences**Table 5** Technical specification of the robot

Items	Characteristics
Size ($L \times W \times H$)	Approx. 483 mm $\times$ 208 mm $\times$ 125 mm
Total mass	Approx. 1.3 kg
Joint configuration	Posterior body: 3; Pectoral fins: 2
On-board sensors	Infrared detectors, pressure sensor, IMU, camera
Controller	ARM Cortex-M4
Power supply	7.4 V rechargeable Li-Polymer batteries

free RL method, is specifically suitable for continuous control tasks. By integrating the strengths of deterministic policy gradient methods and DQN, DDPG enhances learning efficiency and stability through mechanisms including experience replay and target networks. These features make it an effective choice for overcoming the challenges of target tracking in underwater environments.

First, the state space is analyzed. To represent the motion state of the biomimetic robotic shark and the relative direction of the target, we define the state space as a 2D vector:

$$s = (\Psi, \omega) \quad (3)$$

where Ψ refers to the relative bearing angle of the target with respect to the biomimetic robotic shark, and ω signifies the yaw angular velocity of the biomimetic robotic shark. This state space definition satisfies the Markov property, i.e., the current state contains sufficient historical information to predict the state of the next moment and can be used to construct a Markov decision process.

Second, the action space is explored. The heading control of the biomimetic robotic shark is realized by the swing of the tail. With the assumption that the motion control of the biomimetic robotic shark is controlled by a CPG network, the offset parameter β of the CPG network determines the steering behavior of the biomimetic robotic shark. Therefore, we define the action space as a 1D continuous variable:

$$a = \beta' \quad (4)$$

where β' represents the normalized CPG offset parameter used to optimize the steering amplitude of the biomimetic robotic shark. In practical applications, β' is mapped to the actual control parameter β of the CPG network through the following formula:

$$\beta = k_p \cdot a \quad (5)$$

where k_p denotes a preset action mapping coefficient.

Third, the reward function is designed. Correctly designing the reward function is important to the learning direction of the RL algorithm. In this case, we expect that the biomimetic robotic shark can track the target quickly and smoothly. Thus, the reward function should consider the tracking error and the smoothness of the control action. Therefore, a piecewise composite reward function is designed, which involves two parts: tracking error reward and angular velocity penalty:

1. Tracking error reward r_Ψ : We want the target bearing angle Ψ to be as close to zero as possible. Therefore, when $|\Psi|$ is small, a larger reward is given; when $|\Psi|$ is large, a smaller reward or even a penalty is provided. The piece-

wise function form is given as follows:

$$r_\Psi = \begin{cases} a_1 - b_1 \cdot |\Psi|, & \text{if } |\Psi| \leq \Psi_{th} \\ c_1, & \text{if } |\Psi| > \Psi_{th} \end{cases} \quad (6)$$

where a_1 , b_1 , and c_1 are adjustable parameters, and Ψ_{th} is a preset bearing angle threshold.

2. Angular velocity penalty r_ω : To prevent the robotic shark from making sharp turns, we penalize large angular velocities ω . The following linear function can be adopted:

$$r_\omega = -k_\omega \cdot |\omega| \quad (7)$$

where k_ω represents an adjustable penalty coefficient.

The final reward function is the weighted sum of the above two parts and is normalized to balance the dimensions and numerical ranges of different reward items:

$$r = \frac{w_\Psi \cdot r_\Psi + w_\omega \cdot r_\omega}{w_\Psi + w_\omega} \quad (8)$$

where w_Ψ and w_ω denote the weight coefficients of the tracking error reward and the angular velocity penalty, respectively, which are employed to adjust their relative importance in the composite reward function.

Then, the DDPG algorithm is trained according to the following steps:

1) Initialization: The network parameters θ_μ and θ_Q of the actor network $\mu(s; \theta_\mu)$ and the critic network $Q(s, a; \theta_Q)$ are initialized, along with the parameters θ'_μ and θ'_Q of the corresponding target networks $\mu(s; \theta'_\mu)$ and $Q(s, a; \theta'_Q)$. The experience replay buffer R is also initialized.

2) Environment Interaction: At each time step t , the biomimetic robotic shark chooses an action a_t based on the current state s_t and the policy $\mu(s_t; \theta_\mu)$ output by the actor network. In addition, a certain amount of exploration noise is added to guarantee sufficient exploration of the state space.

3) Action Execution: The biomimetic robotic shark executes the action a_t , interacts with the environment, transitions to the next state s_{t+1} , and receives the corresponding reward r_t .

4) Experience Storage: The state transition tuple (s_t, a_t, r_t, s_{t+1}) is stored in the experience replay buffer R .

5) Network Update: When the number of samples in the experience replay buffer reaches a certain threshold, a batch of samples is randomly sampled to update the actor network and the critic network:

a) Critic network update: The target network is used to calculate the target Q value y_i :

$$y_i = r_i + \gamma \cdot Q(s_{i+1}, \mu(s_{i+1}; \theta'_\mu); \theta'_Q) \quad (9)$$

where γ is the discount factor, which is used to balance the importance of current and future rewards. Then, the critic network parameters θ'_Q are updated by minimizing the loss function:

$$L(\theta_Q) = \frac{1}{N} \sum_{i=1}^N \left(y_i - Q(s_i, a_i; \theta_Q) \right)^2 \quad (10)$$

where N denotes the size of the mini-batch.

b) Actor network update: According to the deterministic policy gradient theorem, the actor network parameters θ_μ are updated using the following formula, where J represents the expected cumulative reward objective function:

$$\nabla_{\theta_\mu} J(\theta_\mu) \approx \frac{1}{N} \sum_{i=1}^N \left[\nabla_a Q(s, a; \theta_Q) \Big|_{s=s_i, a=\mu(s_i; \theta_\mu)} \cdot \nabla_{\theta_\mu} \mu(s; \theta_\mu) \Big|_{s=s_i} \right] \quad (11)$$

c) Soft update of target networks: For improving training stability, soft update is utilized to update the target network parameters:

$$\theta'_Q \leftarrow \tau \cdot \theta_Q + (1 - \tau) \cdot \theta'_Q \quad (12)$$

$$\theta'_\mu \leftarrow \tau \cdot \theta_\mu + (1 - \tau) \cdot \theta'_\mu \quad (13)$$

where τ signifies the soft update coefficient, which usually has a small value.

6) Steps 2–5 are repeated for iterative learning until the DDPG algorithm converges to an optimal policy. In this way, the biomimetic robotic shark can choose the best action according to the current state to maximize the cumulative reward.

Through the RL training based on the DDPG algorithm, the biomimetic robotic shark is expected to learn an efficient control strategy for autonomously tracking moving targets within a 2D plane. During the testing phase, the biomimetic robotic shark should rapidly modify its heading based on the bearing information of the target, which ensures that the target remains near the center of the field of view of the camera. Compared with traditional model-based control methods, the RL-based approach presents greater robustness and adaptability, which make it highly suitable for addressing the complexity and uncertainty of underwater environments.

Notably, this case study offers only a foundational framework for resolving the target tracking problem of UUVs using RL. In practical applications, the definition of the state space and action space, the design of the reward function, and the selection of algorithm parameters must be tailored and adjusted according to specific task requirements, the performance characteristics of the biomimetic robotic shark, and the unique attributes of the underwater environment to obtain the desired control performance.

Overall, this case study demonstrates the application of RL algorithms to the 2D target tracking control problem of a biomimetic robotic shark. By reasonably designing the state space, action space, and reward function, as well as utilizing the learning capabilities of the DDPG algorithm, the biomimetic robotic shark can potentially acquire efficient control strategies through interactions with the environment, which eliminates the need for an accurate dynamic model. This capability highlights the feasibility of achieving autonomous target tracking in underwater settings. Moreover, this case study offers a conceptual framework and practical guidance for applying RL to UUV control tasks. It serves as a reference for future research into highly complex and challenging scenarios.

5 Challenges and future trends

5.1 Challenges

RL demonstrates great potential for UUVs, but its application to real complex underwater environments still encounters many challenges, as discussed below.

5.1.1 Data sparsity and inefficient exploration

RL algorithms require a large amount of data, but acquiring interaction data underwater is exceptionally challenging (Griffiths, 2002). UUV deployment in real-world scenarios is costly, time-consuming, and constrained by energy limitations, as well as carries the risk of loss or damage. These factors severely restrict the obtainable volume of data compared with terrestrial robotics. In addition, the quality of data is typically compromised by noise and partial observability. Early-stage exploration, which is essential for discovering effective behaviors, is also hampered. Random exploration strategies common in RL are commonly infeasible due to safety risks and the sheer vastness and featurelessness of many underwater domains (e.g., deep-sea mapping or target search). The high dimensionality of the state and action spaces of UUVs further exacerbates the curse of dimensionality. Addressing this problem demands an exponential increase in data, which are fundamentally scarce. Consequently, learning frequently remains slow, and the cost and risk associated with collecting sufficient data can be prohibitive.

5.1.2 Complexity of state–action spaces and reward function design

The performance of an RL algorithm largely relies on the representation of the state and action spaces, i.e., the position, velocity, thrust and rudder angle control commands, and other information of UUVs. Selecting the appropriate representation and discretizing or continuously processing the state and action space have an important effect on the performance and convergence speed of the algorithm. For exam-

ple, in an underwater navigation task, if the state space only comprises the position of the robot and ignores the attitude, the algorithm may fail to learn an effective strategy (Politi et al., 2024).

The reward function, as a bridge connecting environmental feedback and agent behavior, also directly influences the learning goal and final performance. In the applications of UUVs, the reward function needs to be designed based on specific task requirements. For instance, in a target tracking task, the reward function can be designed according to the distance and relative speed (Yu et al., 2017); and in a path planning task, the reward function can be designed based on the length, smoothness, and safety of the path (Sun et al., 2020). Designing a reward function that can efficiently guide agent learning and reflect task requirements is challenging.

5.1.3 Optimization and adaptation of RL algorithms

The operating environment of UUVs generally exhibits high dimensionality and continuous state and action spaces, which introduce higher requirements for the convergence speed, computational resource consumption, and robustness to noise of RL algorithms. Furthermore, UUVs are commonly powered by batteries with limited energy (Kelasidi et al., 2015), and communication bandwidth is constrained and possesses high latency (Morozs et al., 2020), which also introduce issues to the RL algorithms. Some works have optimized the RL algorithm to improve algorithm efficiency (Sun et al., 2021; 2020; Wang et al., 2021) and adaptability (Wen et al., 2024; Zhu et al., 2019). However, constructing RL algorithms that are simultaneously sample-efficient, computationally frugal, robust, and capable of real-time adaptation still remains an active research area.

5.1.4 Sim-to-real transfer challenges

Training RL policies entirely in the real world is commonly impractical due to the data sparsity issue described in Section 5.1.1. Simulation generates vast amounts of data quickly and safely but struggles to meet the reality gap between simulation and the real complex underwater environment, which poses a major obstacle. Precisely modeling underwater hydrodynamics, such as nonlinear drag, added mass, lift forces, turbulence effects, and interactions with the seabed and surface, is extremely complex (Thorpe, 2007). Simulating realistic sensor behavior, especially for multipath effects, sonar speckle noise, scattering, absorption, and vision-related turbidity, is also equally difficult.

Moreover, environmental disturbances, including ocean currents, are unpredictable and random, which further intensifies the problem of model mismatch. Therefore, when the RL strategy trained in the simulation environment is directly transferred to the real underwater environment, its performance generally drops significantly, or it even fails entirely. For instance, an obstacle avoidance strategy that performs effectively in a simulation environment may fail because it

cannot precisely perceive the position and velocity of actual obstacles (Wang et al., 2025).

5.1.5 Limited onboard computing resources

UUVs, specifically smaller ones, commonly operate under strict size, weight, power, and cost constraints, which limit the available onboard computational resources. Many state-of-the-art DRL algorithms, which rely on DNNs, require significant processing power and memory for training and inference (LeCun et al., 2015). Deploying these complex models directly into resource-constrained embedded systems is challenging. Model compression and lightweight network architectures provide partial solutions. However, realizing efficient and real-time onboard RL inference is still a major technical hurdle.

5.2 Future trends

Despite the challenges, the application prospects of RL in UUVs are still broad. In response to the challenges, future research can be conducted in the following aspects:

5.2.1 Enhancing data utilization and exploration efficiency

To address the limitation of data sparsity and inefficient exploration identified in Section 5.1.1, promising avenues include the following:

1) Transfer Learning: Knowledge from simulations, related tasks, or prior missions can be utilized to warm-start learning, which reduces the need for extensive de novo data collection.

2) Model-Based RL: Approximate models of dynamics and the environment of UUVs, even if imperfect, can be developed. These models can be used for planning or generating simulated experiences offline, which significantly minimizes reliance on costly real-world interactions. Techniques must focus on addressing uncertainty effectively, which is the key limitation noted in current model-based attempts.

3) Offline RL: Methods capable of learning effective policies purely from previously collected datasets can be established, which is an emerging trend demonstrating potential to leverage existing UUV logs without risky online interaction.

4) Few-Shot/Meta-Learning: Algorithms that can adapt rapidly to new tasks or variations in the environment using only minimal new data while leveraging experience across related prior tasks need to be designed.

5) Efficient Exploration Strategies: Naive exploration methods, such as epsilon-greedy, can be replaced by more sophisticated methods driven by curiosity, intrinsic motivation, or uncertainty estimation. This shift guides UUVs to explore informative regions of the state–action space more intelligently and safely. Techniques such as hindsight experience replay can also help optimally balance exploration with safety constraints in unpredictable underwater environments.

5.2.2 Optimizing state and action space and reward function design

To tackle the complexities of state–action spaces and reward function design, further research is needed in the following areas:

1) Representation Learning: Techniques such as autoencoders or self-supervised learning need to be utilized to automatically learn compact and informative state representations from high-dimensional sensor data, including sonar images and camera feeds, which potentially reduces the dimensionality challenge.

2) HRL: Complex tasks can be decomposed into simpler subtasks or layers of abstraction. HRL can simplify learning by operating over temporally extended actions or goal-based sub-policies, which supports the management of complex spaces and facilitates long-term planning. At present, modifying the structure and learning process within HRL for UUV tasks is an active research area.

3) Automated Reward Design: Techniques including inverse RL can be explored to infer reward functions from expert demonstrations, which reduces the burden of manual reward engineering. Multiobjective RL frameworks can also help systematically design rewards that robustly capture complex, which potentially conflicts objectives in dynamic underwater scenarios.

5.2.3 Robust and resource-aware RL algorithms

Responding to the need for algorithms that can address operational constraints of UUVs, development should focus on the following aspects:

1) Robust RL: Algorithms that are inherently less sensitive to noise, disturbances, and model inaccuracies can be developed. These algorithms might involve adversarial training, risk-sensitive optimization criteria, or methods that explicitly account for uncertainty.

2) Adaptive RL: Algorithms that can continuously adapt online to changing environmental conditions or unexpected events, such as component failure during a mission, can be considered.

3) Algorithm Optimization for Embedded Systems: Lightweight RL architectures and efficient inference techniques suitable for deployment on resource-constrained UUV hardware can be constructed. Moreover, neuromorphic computing or specialized hardware accelerators can be explored, which represents a critical intersection of RL and embedded system research.

5.2.4 Facilitating efficient simulation-to-reality transfer

Bridging the substantial sim-to-real gap remains a critical priority for future research. By establishing a more realistic underwater environment simulation model while considering the effects of hydrodynamics, sensor noise, and environmental disturbances, the gap between the simulation environment and the real environment can be narrowed.

1) High-Fidelity Simulation: Simulators can be improved

by integrating highly accurate hydrodynamic models, advanced sensor models such as ray tracing for sonar and physically based rendering for vision, and realistic environmental dynamics including stochastic current models and seabed interactions.

2) Domain Randomization: Simulation parameters such as mass, drag coefficients, current speeds, and sensor noise levels can be systematically randomized during training, which forces the policy to learn behaviors robustly across various conditions. This capability increases the likelihood of successful transfer.

3) Domain Adaptation: Techniques that explicitly align the feature distributions or dynamics between the simulation and real-world domains by using limited real-world data can be employed.

4) System Identification: Data collected from the real UUVs can be used to fine-tune simulator parameters or learn a residual model that corrects the predictions of the simulator.

5) Grounded Simulation: Real sensor data or hardware-in-the-loop components can be integrated directly into the simulation loop for highly realistic training feedback.

6) Safe RL for Transfer: Methods that ensure safety constraints (e.g., collision avoidance) and maintain operational limits during the transfer process and real-world fine-tuning can be developed, which is a critical enabler for practical deployment.

5.2.5 Enhanced multiagent collaboration

To address the challenges specific to multi-UUV systems, future research should explore the following directions:

1) Scalable MARL: Multiagent RL algorithms that scale effectively to larger numbers of UUVs can be developed to address the nonstationarity and coordination challenges inherent in multirobot systems.

2) Communication-Efficient MARL: Algorithms that are specifically tailored for low-bandwidth and high-latency acoustic communication can be designed, which enables effective collaboration despite communication constraints. Investigating novel communication protocols informed by MARL needs is also an important interdisciplinary direction, where graph neural networks may show promise.

Overall, although significant challenges remain, targeted research in these areas, which directly addresses the gaps and limitations identified in this review and combines RL, robotics, control theory, and ocean engineering, is crucial. Overcoming these hurdles will reveal the transformative potential of RL, which enables more autonomous, capable, and resilient UUVs for exploring the underwater environment.

6 Conclusions

This study presents a comprehensive review of current research on the application of RL in UUVs. The critical

role of UUVs in ocean exploration and development is first introduced, as well as a detailed discussion of the limitations of traditional methods in complex underwater environments. These constraints establish a clear research gap and highlight the need for adopting highly adaptive and intelligent approaches such as RL. Subsequently, an overview of the UUVs and RL is provided. On this basis, the core of this review elaborates the research status of RL across three key task areas: motion planning, navigation and control, and multiagent collaboration. Through a comprehensive analysis of the main methods, representative studies, and existing challenges, the review synthesizes evidence demonstrating the use of RL to tackle issues such as model uncertainty, high-dimensional state and action spaces, and the need for real-time adaptation. These areas are domains where conventional methods frequently prove insufficient. Then, a practical case study employing a biomimetic robotic shark platform further illustrates the tangible application of RL for addressing specific UUV control problems, which offers a valuable reference. Finally, the challenges and future development trends are discussed.

In summary, RL, as a data-driven intelligent decision-making methodology, has presented unique advantages and remarkable potential in addressing control and decision-making problems for UUVs in complex, dynamic, and uncertain environments. Numerous challenges remain, but these obstacles are expected to be progressively resolved as research deepens and interdisciplinary integration advances. In the future, RL is expected to play an increasingly vital role in enhancing the intelligence and autonomy of UUVs, which provides robust technical support for human exploration and utilization of the ocean. With continued progress, RL is anticipated to provide further breakthroughs in the UUV domain. These innovations ultimately enable the widespread application of UUVs across diverse and complex tasks.

Funding Supported by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant No. 2023142) and in part by the National Natural Science Foundation of China (Grant No. 62033013).

Acknowledgement The authors acknowledge the Doubao multimodal AI model for assistance with illustrative elements in Figure 1 and reference imagery of UUVs from DEEPINFAR, Forum Energy Technologies, and the Shenyang Institute of Automation, Chinese Academy of Sciences.

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

- Baharin UNA, Kamarudin KR (2025) Using remote-controlled vehicles (ROV) as tools for sea cucumber conservation: A review. *Maritime Technology and Research* 7: 274346-274346. <https://doi.org/10.33175/mtr.2025.274346>
- Bao H, Zhang Y, Song M, Kong Q, Hu X, An X (2023) A review of underwater vehicle motion stability. *Ocean Engineering* 287: 115735. <https://doi.org/10.1016/j.oceaneng.2023.115735>
- Bellman R (1966) Dynamic programming. *Science* 153: 34-37. <https://doi.org/10.1126/science.153.3731.34>
- Bertsekas DP (1996) Neuro-dynamic programming. Athena Scientific
- Borase RP, Maghade DK, Sondkar SY, Pawar SN (2021) A review of PID control, tuning methods and applications. *Int. J. Dynam. Control* 9: 818-827. <https://doi.org/10.1007/s40435-020-00665-4>
- Chen T, Zhang Z, Fang Z, Jiang D, Li G (2024) Imitation learning from imperfect demonstrations for AUV path tracking and obstacle avoidance. *Ocean Engineering* 298: 117287. <https://doi.org/10.1016/j.oceaneng.2024.117287>
- Christensen L, De Gea Fernández J, Hildebrandt M, Koch CES, Wehbe B (2022) Recent advances in AI for navigation and control of underwater robots. *Curr Robot Rep* 3: 165-175. <https://doi.org/10.1007/s43154-022-00088-3>
- Coates RFW (1990) Underwater acoustic systems. Macmillan Education UK, London. <https://doi.org/10.1007/978-1-349-20508-0>
- Cohen JE, Small C, Mellinger A, Gallup J, Sachs J (1997) Estimates of coastal populations. *Science* 278: 1209-1213. <https://doi.org/10.1126/science.278.5341.1209c>
- Cong Y, Gu C, Zhang T, Gao Y (2021) Underwater robot sensing technology: A survey. *Fundamental Research* 1: 337-345. <https://doi.org/10.1016/j.fmre.2021.03.002>
- Dunbabin M, Marques L (2012) Robots for environmental monitoring: significant advancements and applications. *IEEE Robotics & Automation Magazine* 19: 24-39. <https://doi.org/10.1109/MRA.2011.2181683>
- El-Fakdi A, Carreras M (2013) Two-step gradient-based reinforcement learning for underwater robotics behavior learning. *Robotics and Autonomous Systems* 61: 271-282. <https://doi.org/10.1016/j.robot.2012.11.009>
- Feinberg V, Wan A, Stoica I, Jordan MI, Gonzalez JE, Levine S (2018) Model-based value estimation for efficient model-free reinforcement Learning. <https://doi.org/10.48550/arXiv.1803.00101>
- Fernandez DC, Hollinger, GA (2017) Model predictive control for underwater robots in ocean waves. *IEEE Robot. Autom. Lett.* 2: 88-95. <https://doi.org/10.1109/LRA.2016.2531792>
- Fossen TI (2011) Handbook of marine craft hydrodynamics and motion control. John Wiley & Sons Ltd
- Gaskett C, Wettergreen D, Zelinsky A (1999) Reinforcement learning applied to the control of an autonomous underwater vehicle. *Proceedings of the Australasian Conference on Robotics and Automation (AuCRA)*
- Griffiths G (2002) Technology and applications of autonomous underwater vehicles. CRC Press
- Hadi B, Khosravi A, Sarhadi P (2024) Adaptive formation motion planning and control of autonomous underwater vehicles using deep reinforcement learning. *IEEE J. Oceanic Eng.* 49: 311-328. <https://doi.org/10.1109/JOE.2023.3278290>
- Han Z, Wang Y, Sun Q (2023) Straight-path following and formation control of USVs using distributed deep reinforcement learning and adaptive neural network. *IEEE/CAA J. Autom. Sinica* 10: 572-574. <https://doi.org/10.1109/JAS.2023.123255>
- Hou X, Wang J, Bai T, Deng Y, Ren Y, Hanzo L (2023) Environment-aware AUV trajectory design and resource management for multi-tier underwater computing. *IEEE J. Select. Areas Commun.* 41: 474-490. <https://doi.org/10.1109/JSAC.2022.3227103>
- Howard RA (1960) Dynamic programming and markov processes.

- MIT Press 2: 39-47
- Huang H, Tang Q, Li J, Zhang W, Bao X, Zhu H, Wang G (2020) A review on underwater autonomous environmental perception and target grasp, the challenge of robotic organism capture. *Ocean Engineering* 195: 106644. <https://doi.org/10.1016/j.oceaneng.2019.106644>
- Huy DQ, Sadjoli N, Azam AB, Elhadidi B, Cai Y, Seet G (2023) Object perception in underwater environments: a survey on sensors and sensing methodologies. *Ocean Engineering* 267: 113202
- Kelasidi E, Pettersen KY, Gravdahl JT (2015) Energy efficiency of underwater robots. 10th IFAC Conference on Manoeuvring and Control of Marine Craft MCMC 48: 152-159. <https://doi.org/10.1016/j.ifacol.2015.10.273>
- Konda V, Tsitsiklis J (1999) Actor-critic algorithms. *Advances in neural information processing systems* 12: 1008-1014
- Lan W, Jin X, Chang X, Wang T, Zhou H, Tian W, Zhou L (2022) Path planning for underwater gliders in time-varying ocean current using deep reinforcement learning. *Ocean Engineering* 262: 112226. <https://doi.org/10.1016/j.oceaneng.2022.112226>
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521: 436-444
- Li D, Du L (2021) AUV trajectory tracking models and control strategies: a review. *Journal of Marine Science and Engineering* 9: 1020. <https://doi.org/10.3390/jmse9091020>
- Lillicrap TP (2015) Continuous control with deep reinforcement learning. arXiv preprint arXiv: 1509.02971
- Luo Y, Xu H, Li Y, Tian Y, Darrell T, Ma T (2021) Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. <https://doi.org/10.48550/arXiv.1807.03858>
- Lyu X, Sun Y, Wang L, Tan J, Zhang L (2023) End-to-end AUV local motion planning method based on deep reinforcement learning. *Journal of Marine Science and Engineering* 11: 1796. <https://doi.org/10.3390/jmse11091796>
- Ma D, Chen X, Ma W, Zheng H, Qu F (2024) Neural network model-based reinforcement learning control for AUV 3-D path following. *IEEE Trans. Intell. Veh.* 9: 893-904. <https://doi.org/10.1109/TIV.2023.3282681>
- Macreadie PI, McLean DL, Thomson PG, Partridge JC, Jones DOB, Gates AR, Benfield MC, Collin SP, Booth DJ, Smith LL, Techera E, Skropeta D, Horton T, Pattiaratchi C, Bond T, Fowler AM (2018) Eyes in the sea: unlocking the mysteries of the ocean using industrial, remotely operated vehicles (ROVs). *Science of The Total Environment* 634: 1077-1091. <https://doi.org/10.1016/j.scitotenv.2018.04.049>
- Mai C, Pedersen S, Hansen L, Jepsen KL, Yang Z (2016) Subsea infrastructure inspection: A review study. *IEEE International Conference on Underwater System Technology: Theory and Applications (USYS)*, Penang, Malaysia, 71-76. <https://doi.org/10.1109/USYS.2016.7893928>
- Mnih V (2013) Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602
- Morozs N, Gorma W, Henson BT, Shen L, Mitchell PD, Zakharov YV (2020) Channel modeling for underwater acoustic network simulation. *IEEE Access* 8: 136151-136175. <https://doi.org/10.1109/ACCESS.2020.3011620>
- Paull L, Saeedi S, Seto M, Li H (2014) AUV navigation and localization: a review. *IEEE Journal of Oceanic Engineering* 39: 131-149. <https://doi.org/10.1109/JOE.2013.2278891>
- Pereira MI, Pinto AM (2024) Reinforcement learning based robot navigation using illegal actions for autonomous docking of surface vehicles in unknown environments. *Engineering Applications of Artificial Intelligence* 133: 108506. <https://doi.org/10.1016/j.engappai.2024.108506>
- Politi E, Stefanidou A, Chronis C, Dimitrakopoulos G, Iraklis Varlamis (2024) Adaptive deep reinforcement learning for efficient 3D navigation of autonomous underwater vehicles. *IEEE Access* 12: 178209-178221. <https://doi.org/10.1109/ACCESS.2024.3508031>
- Rummery GA, Niranjan M (1994) On-line Q-learning using connectionist systems. University of Cambridge, Department of Engineering Cambridge, Cambridge, UK
- Schettini R, Corchs S (2010) Underwater image processing: state of the art of restoration and image enhancement methods. *EURASIP J. Adv. Signal Process.* 746052. <https://doi.org/10.1155/2010/746052>
- Schulman J (2015) Trust region policy optimization. arXiv preprint arXiv:1502.05477
- Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. <https://doi.org/10.48550/arXiv.1707.06347>
- Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529: 484-489
- Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362: 1140-1144. <https://doi.org/10.1126/science.aar6404>
- Singh B, Kumar R, Singh VP (2022) Reinforcement learning in robotic applications: a comprehensive survey. *Artif Intell Rev* 55: 945-990. <https://doi.org/10.1007/s10462-021-09997-9>
- Su Z, Zhou M, Han F, Zhu Y, Song D, Guo T (2019) Attitude control of underwater glider combined reinforcement learning with active disturbance rejection control. *J Mar Sci Technol* 24: 686-704. <https://doi.org/10.1007/s00773-018-0582-y>
- Sun Y, Luo X, Ran X, Zhang G (2021) A 2D optimal path planning algorithm for autonomous underwater vehicle driving in unknown underwater canyons. *JMSE* 9: 252. <https://doi.org/10.3390/jmse9030252>
- Sun Y, Ran X, Zhang G, Xu H, Wang X (2020) AUV 3D path planning based on the improved hierarchical deep Q network. *Journal of Marine Science and Engineering* 8: 145. <https://doi.org/10.3390/jmse8020145>
- Sun Y, Yan C, Xiang X, Zhou H, Tang D, Zhu Y (2023) Towards end-to-end formation control for robotic fish via deep reinforcement learning with non-expert imitation. *Ocean Engineering* 271: 113811. <https://doi.org/10.1016/j.oceaneng.2023.113811>
- Sutton RS (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Machine Learning Proceedings*, Elsevier. 216-224
- Sutton RS (1988) Learning to predict by the methods of temporal differences. *Mach Learn* 3: 9-44. <https://doi.org/10.1007/BF00115009>
- Sutton RS, Barto AG (2018) Reinforcement learning. Second edition: an introduction. MIT Press
- Terracciano DS, Bazzarello L, Caiti A, Costanzi R, Manzari V (2020) Marine robots for underwater surveillance. *Curr Robot Rep* 1: 159-167. <https://doi.org/10.1007/s43154-020-00028-z>
- Thorpe SA (2007) An introduction to ocean turbulence. Cambridge University Press, Cambridge
- Tong R, Feng Y, Wang J, Wu Z, Tan M, Yu J (2023) A survey on reinforcement learning methods in bionic underwater robots. *Biomimetics* 8: 168. <https://doi.org/10.3390/biomimetics8020168>
- Van Hasselt H, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. *Proceedings of the AAAI Conference on*

- Artificial Intelligence, Phoenix, AZ, 2094-2100
- Vu QT, Pham MH, Nguyen VD, Duong VT, Nguyen HH, Nguyen TT (2023) Optimization of central pattern generator-based locomotion controller for fish robot using deep deterministic policy gradient. In: Nguyen DC, Vu NP, Long BT, Puta H, Sattler KU (Eds.), *Advances in engineering research and application, lecture notes in networks and systems*. Springer International Publishing, Cham, 764-770. https://doi.org/10.1007/978-3-031-22200-9_80
- Wang Y, Cao J, Sun J, Zou X, Sun C (2024a) Path following control for unmanned surface vehicles: a reinforcement learning-based method with experimental validation. *IEEE Trans. Neural Netw. Learning Syst.* 35: 18237-18250. <https://doi.org/10.1109/TNNLS.2023.3313312>
- Wang Y, Lu C, Wu P, Zhang X (2024b) Path planning for unmanned surface vehicle based on improved Q-Learning algorithm. *Ocean Engineering* 292: 116510. <https://doi.org/10.1016/j.oceaneng.2023.116510>
- Wang Y, Xi M, Weng Y (2025) Intelligent path planning algorithm of autonomous underwater vehicle based on vision under ocean current. *Expert Systems* 42: e13399. <https://doi.org/10.1111/exsy.13399>
- Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N (2016) Dueling network architectures for deep reinforcement learning. *International Conference on Machine Learning*. PMLR, 1995-2003
- Wang Z, Wen Z, Xia Q, Cai W (2022) Deep reinforcement learning based multi-UUV cooperative control for target capturing. *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCoM/CyberSciTech)*, Falerma, Italy, 1-6. <https://doi.org/10.1109/DASC/PiCom/CBDCoM/Cy55231.2022.9927810>
- Wang Z, Xiang X, Duan Y, Yang S (2024c) Adversarial deep reinforcement learning based robust depth tracking control for underactuated autonomous underwater vehicle. *Engineering Applications of Artificial Intelligence* 130: 107728. <https://doi.org/10.1016/j.engappai.2023.107728>
- Wang Z, Zhang S, Feng X, Sui Y (2021) Autonomous underwater vehicle path planning based on actor-multi-critic reinforcement learning. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 235: 1787-1796. <https://doi.org/10.1177/0959651820937085>
- Watkins CJCH (1989) Learning from delayed rewards
- Wei W, Wang J, Du J, Fang Z, Ren Y, Chen CLP (2023a) Differential game-based deep reinforcement learning in underwater target hunting task. *IEEE Transactions on Neural Networks and Learning Systems* 36(1): 462-474. <https://doi.org/10.1109/TNNLS.2023.3325580>
- Wei X, Wang H, Tang Y (2023b) Deep hierarchical reinforcement learning based formation planning for multiple unmanned surface vehicles with experimental results. *Ocean Engineering* 286: 115577. <https://doi.org/10.1016/j.oceaneng.2023.115577>
- Wen J, Wang A, Zhu J, Xia F, Peng Z, Zhang W (2024) Adaptive energy-efficient reinforcement learning for AUV 3D motion planning in complex underwater environments. *Ocean Engineering* 312: 119111. <https://doi.org/10.1016/j.oceaneng.2024.119111>
- Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn* 8: 229-256. <https://doi.org/10.1007/BF00992696>
- Wong C, Yang E, Yan XT, Gu D (2018) Autonomous robots for harsh environments: a holistic overview of current solutions and ongoing challenges. *Systems Science & Control Engineering* 6: 213-219. <https://doi.org/10.1080/21642583.2018.1477634>
- Wynn RB, Huvenne VAI, Le Bas TP, Murton BJ, Connelly DP, Bett BJ, Ruhl HA, Morris KJ, Peakall J, Parsons DR, Sumner EJ, Darby SE, Dorrell RM, Hunt JE (2014) Autonomous underwater vehicles (AUVs): their past, present and future contributions to the advancement of marine geoscience. *Marine Geology* 352: 451-468. <https://doi.org/10.1016/j.margeo.2014.03.012>
- Xia J, Luo Y, Liu Z, Zhang Y, Shi H, Liu Z (2023) Cooperative multi-target hunting by unmanned surface vehicles based on multi-agent reinforcement learning. *Defence Technology* 29: 80-94. <https://doi.org/10.1016/j.dt.2022.09.014>
- Xu J, Huang F, Wu D, Cui Y, Yan Z, Zhang K (2021) Deep reinforcement learning based multi-AUVs cooperative decision-making for attack-defense confrontation missions. *Ocean Engineering* 239: 109794. <https://doi.org/10.1016/j.oceaneng.2021.109794>
- Xu J, Zhang Z, Wang J, Han Z, Ren Y (2024) Multi-AUV pursuit-evasion game in the internet of underwater things: an efficient training framework via offline reinforcement learning. *IEEE Internet of Things Journal* 11: 31273-31286. <https://doi.org/10.1109/JIOT.2024.3416616>
- Yan Z, Yan J, Wu Y, Cai S, Wang H (2023) A novel reinforcement learning based tuna swarm optimization algorithm for autonomous underwater vehicle path planning. *Mathematics and Computers in Simulation* 209: 55-86. <https://doi.org/10.1016/j.matcom.2023.02.003>
- Yang X, Wu Z, Yu J (2016) Design and implementation of a robotic shark with a novel embedded vision system. *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Qingdao, China, 841-846. <https://doi.org/10.1109/ROBIO.2016.7866428>
- Yao X, Wang F, Yuan C, Wang J, Wang X (2021) Path planning for autonomous underwater vehicles based on interval optimization in uncertain flow fields. *Ocean Engineering* 234: 108675. <https://doi.org/10.1016/j.oceaneng.2021.108675>
- Yoerger DR, Jakuba M, Bradley AM, Bingham B (2007) Techniques for deep sea near bottom survey using an autonomous underwater vehicle. In: Thrun S, Brooks R, Durrant-Whyte H (Eds.), *Robotics research, springer tracts in advanced robotics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 416-429. https://doi.org/10.1007/978-3-540-48113-3_36
- Yoo B, Kim J (2016) Path optimization for marine vehicles in ocean currents using reinforcement learning. *J Mar Sci Technol* 21: 334-343. <https://doi.org/10.1007/s00773-015-0355-9>
- You K, Wu H, Hsu Y, Song S (2020) A selected review of reinforcement learning-based control for autonomous underwater vehicles. *Sci. Sin. -Inf.* 50: 1798. <https://doi.org/10.1360/SSI-2020-0065>
- Yu H, Zhang S (2024) Collaborative task decision-making of multi-UUV in dynamic environments based on deep reinforcement learning. *Ships and Offshore Structures*: 1-11. <https://doi.org/10.1080/17445302.2024.2391788>
- Yu J, Wu Z, Yang X, Yang Y, Zhang P (2021) Underwater target tracking control of an untethered robotic fish with a camera stabilizer. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51: 6523-6534. <https://doi.org/10.1109/TSMC.2019.2963246>
- Yu R, Shi Z, Huang C, Li T, Ma Q (2017) Deep reinforcement learning based optimal trajectory tracking control of autonomous underwater vehicle. *36th Chinese Control Conference (CCC)*, IEEE, Dalian, China, 4958-4965. <https://doi.org/10.23919/ChiCC.2017.8028138>
- Yuh J (2000) Design and control of autonomous underwater robots: a

- survey. *Autonomous Robots* 8: 7-24. <https://doi.org/10.1023/A:1008984701078>
- Zhang D, Pan G, Cao Y, Huang Q, Cao Y (2023) Depth control of a biomimetic manta robot via reinforcement learning. In: Sun F, Cangelosi A, Zhang J, Yu Y, Liu H, Fang B (Eds.), *Cognitive Systems and Information Processing, Communications in Computer and Information Science*. Springer Nature Singapore, Singapore, 59-69. https://doi.org/10.1007/978-981-99-0617-8_5
- Zhang J, Ren J, Cui Y, Fu D, Cong J (2024) Multi-USV task planning method based on improved deep reinforcement learning. *IEEE Internet of Things Journal* 11: 18549-18567. <https://doi.org/10.1109/JIOT.2024.3363044>
- Zhang T, Tian R, Yang H, Wang C, Sun J, Zhang S, Xie G (2022) From simulation to reality: a learning framework for fish-like robots to perform control tasks. *IEEE Trans. Robot.* 38: 3861-3878. <https://doi.org/10.1109/TRO.2022.3181014>
- Zhu S, Han G, Lin C, Tao Q (2024) Underwater target tracking based on hierarchical software-defined multi-AUV reinforcement learning: a multi-AUV advantage-attention actor-critic approach. *IEEE Trans. on Mobile Comput.* 23: 13639-13653. <https://doi.org/10.1109/TMC.2024.3437376>
- Zhu Y, Dai C, Xia J, Wu H, An L (2019) Adaptive Q-learning algorithm for AUV route planning. *IOP Conf. Ser.: Mater. Sci. Eng.* 569: 052029. <https://doi.org/10.1088/1757-899X/569/5/052029>