RIMC Volume 1, Issue 2, December 2024 | http://dx.doi.org/10.21622/RIMC.2024.01.2.1145
Journal of Robotics: Integration, Manufacturing & Control - eISSN 3009-7967

Reinforcement Learning for Autonomous Underwater Vehicles (AUVs): Navigating Challenges in Dynamic and Energy-Constrained Environments

Mohab M. Eweda [1], and Karim A. ElNaggar [2]

[1] Department of Electrical Engineering Upgrading Studies, Institute of Maritime Upgrading Studies, AASTMT Abukir Campus, Egypt.
[2] Department of Electrical & Control Engineering, College of Engineering and Technology, AASTMT Abukir Campus, Egypt.

mohabeweda@live.com, karimelnaggar726@yahoo.com

Received on, 02 November 2024 Accepted on, 22 December 2024 Published on, 24 December 2024

ABSTRACT

Autonomous Underwater Vehicles (AUVs) are essential for underwater exploration, inspection, and environmental surveillance. Nevertheless, navigation, obstacle avoidance, and energy efficiency are greatly hindered by the ever-changing underwater environments. Reinforcement Learning (RL) has arisen as a revolutionary method for tackling these challenges. This paper examines significant progress in reinforcement learning algorithms, emphasizing their application in the training of autonomous underwater vehicles in both simulated and real-world environments. The review synthesizes findings from multiple studies, identifies gaps in existing research, and highlights the potential of algorithms such as Deep Deterministic Policy Gradient (DDPG) for continuous control tasks. This review offers an extensive examination of current methodologies, their constraints, and avenues for future investigation.

Key-words: Autonomous Underwater Vehicles, AUVs, Reinforcement Learning, RL, Navigation, Obstacle Avoidance, Energy Efficiency

I. INTRODUCTION

Autonomous Underwater Vehicle (AUV) navigationisjustoneexampleofthecomplicated control problems that reinforcement learning (RL) has the potential to solve. Autonomous underwater vehicles (AUVs) play an essential role in many fields, such as oceanography, underwater infrastructure inspection, and search and rescue. However, there are certain obstacles specific to the underwater domain, such as energy limitations, poor communication, and unpredictable currents. When faced with such changing conditions, traditional control methods frequently fall short. In this literature review, the researchers look at how Deep Deterministic Policy Gradient (DDPG) and other RL algorithms have been applied to underwater robotics. The purpose of this review is to show how RL is useful for underwater navigation, point out where more research is needed, and fill in any gaps in the current literature. It is structured thematically and discusses topics such as RL theory foundations, AUV navigation applications, difficulties in dynamic environments, and potential future research directions.

A. *Related Work*

A major development in reinforcement learning for continuous control tasks, Lillicrap et al. (2015) propose the DDPG algorithm [1]. Their actor-critic framework sets the foundation for its use in robotics by allowing RL agents to effectively operate in high-dimensional action spaces. Emphasizing the value of trial-and-error learning for decision-making procedures [6], Sutton and Barto (1998) offer basic insights into RL. Building on these ideas, Zhang et al. (2021) use DDPG for underwater navigation to show its potential for exact control in virtual environments [7].

Other noteworthy contributions include the robotic arms and aerial drones using Q-learning and policy gradient techniques. For robotic manipulation, Levine et al. (2016), for example, present end-to-end training of visuomotor policies, so demonstrating the capacity of RL to solve challenging tasks [11]. Emphasizing its ability for handling continuous and high-dimensional control spaces, these studies prepared the groundwork for using RL in underwater robotics.

A safe and controlled environment is provided by simulated environments for the purpose of training AUVs with RL. To facilitate obstacle avoidance in AUVs, Smith et al. (2018) implement reward mechanisms [2]. Their research underscores the significance of reward functions that are well-designed for the purpose of facilitating the learning of policies. In the same vein, Garcia and Torres (2019) implement policies that are based on deep learning to improve the trajectory planning of AUVs [10]. Although these methods are successful in structured environments, they encounter difficulty in generalizing to dynamic and unpredictable underwater conditions.

The development of realistic underwater environments has been significantly facilitated by simulation tools like Gazebo and Unity. These platforms enable researchers to integrate physical factors such as turbulence, drag, and buoyancy, thereby rendering the training process more akin to real-world conditions. Nevertheless, the “sim-to-real” problem, which is frequently used to describe the disparity between simulation and reality, continues to be a significant obstacle.

Significant difficulties arise for AUV navigation in dynamic environments due to the presence of stochastic disturbances and unpredictable obstacles. In their study, Chen and Wang (2020) investigate how RL can be adjusted to suit actual oceanic circumstances, considering random environmental perturbations like turbulence and ocean currents [3]. Their research demonstrates the critical importance of stable policies that can withstand changing conditions over the long term. Regardless of these developments, their approaches are computationally heavy and necessitate a large amount of training data, neither of which is necessarily accessible. The use of adaptive reward systems to strengthen policies has been the subject of further research. To achieve a better balance between navigation efficiency and energy conservation, Singh et al. (2021) create a multi-objective RL framework. This framework showed improves adaptability in dynamic conditions [12]. Nevertheless, the question of how to achieve adaptation in real-time is still unanswered.

For prolonged missions, AUVs must emphasize energy efficiency. Kumar et al. (2019) [4] concentrate on energy-efficient reinforcement learning policies for prolonged AUV missions. Their strategy markedly enhances operational efficiency by prioritizing energy consumption within the reward function. Their reliance on oversimplified energy models, however, renders them ineffective in practical applications. Brown et al. (2017) find that heuristic-based reward systems are not as effective as possible in addressing complex navigation tasks [9]. Recent advancements in energy modeling have enabled more accurate predictions of battery efficiency and power consumption. The operational model of AUV is enhanced for energy efficiency via RL-based scheduling [13] following the integration of renewable energy sources, such as solar panels, by Zhang et al. (2022). These innovations demonstrate the potential for reinforcement learning to collaborate with advanced energy management systems.

A comparative analysis of reinforcement learning algorithms employed in autonomous underwater vehicle research uncovers significant trends and constraints. Zhang et al. (2021) emphasize the benefits of DDPG for continuous control, whereas Lee et al. (2020) illustrate the adaptability of Proximal Policy Optimization (PPO) in multi-agent environments [8]. Conversely, heuristic-based approaches, such as those suggested by Brown et al. (2017), are easier to implement but demonstrated insufficient adaptability to varied underwater conditions. Garcia and Torres (2019) underscore the significance of amalgamating deep learning methodologies with reinforcement learning for enhanced policy acquisition [10].

Among AUV navigators, DDPG is a preferred choice mostly because of its capacity to manage continuous action environments. Zhang et al. (2021) use it to maximize paths [7], so stressing its adaptability in simulated environments. Chen and Wang (2020) show its resilience when combined with domain randomization approaches [3] so allowing one to generalize across several underwater conditions. The researchers still have a lot to learn, though, about how to combine DDPG with multi-sensor data fusion, so enhancing their capacity to perceive and make decisions in their surroundings. Apart from navigation, DDPG has proved useful for environmental monitoring and object retrieval—two more AUV chores. Liu et al. (2022) notably improve coverage and energy efficiency by bestocating underwater sensors with DDPG [14]. These applications show how adaptable the method is and how more generally it could be used in underwater robotics. DDPG and other algorithms have shown promise in addressing continuous control problems; hence, this review highlights the growing application of RL in AUV navigation. Two main gaps are poor policies for dynamic environments and insufficient connection with actual oceanic conditions. The aim of this work is to fill in these voids by using RL developments, hence more adaptive and efficient AUV navigation systems are sought for.

II. METHODOLOGY

A. *Problem Formulation*

AUVs have become indispensable instruments in various marine activities, such as environmental assessment, search and rescue operations, and underwater investigation. These applications necessitate AUVs to autonomously navigate in dynamic, complex, and frequently hostile underwater environments. The autonomy of AUVs is impeded by several substantial challenges, including dynamic obstacles (e.g., moving objects and unpredictable ocean currents), energy constraints from onboard batteries, and the necessity for accurate goal-directed navigation in three-dimensional environments.

Due to high latency, limited communication bandwidth, and the difficulty in sensing accurate positional information, these challenges are made even worse in underwater environments. Robust control mechanisms and adaptive strategies are required to guarantee safe and efficient navigation due to these constraints. In order to make AUVs more efficient and dependable in real-world scenarios, it is essential to solve these problems. To teach AUVs to navigate autonomously, optimize energy consumption, and avoid obstacles in these types of settings, this research suggests an RL framework that makes use of the Deep Deterministic Policy Gradient (DDPG) algorithm.

B. *Mathematical Representation*

1. State space

The AUV state at any time (t), denoted as $(s_t \in R^7)$, is a multidimensional vector defined as:

$$[s_t = [x, y, z, \text{roll}, \text{pitch}, \text{yaw}, \text{battery level}],] \qquad (1)$$

where $((x, y, z))$ represents the AUV position in a 3D coordinate system. The orientation of the AUV is described by $(roll,\ pitch,\ yaw)$, and the battery level indicates the remaining energy. This state vector encapsulates the critical information required for navigation and decision-making.

Additional environmental data include:

1. Positions of dynamic obstacles, represented as $(\{o_i \in R^3\}_{i=1}^N)$, *(N)* where is the number of obstacles.

2. Environmental disturbances, such as water currents, modeled as random forces acting on the AUV.

2. Action space

The action $(a_t \in R^6)$ is a continuous control input defined as:

$$[a_t = [f_x, f_y, f_z,\text{roll\_change},\text{pitch\_change},\text{yaw\_change}],] \quad (2)$$

where $(f_x f_y f_z)$ represent the thrust forces in three spatial directions, and *(roll_change, pitch_change, yaw_change)* correspond to rotational adjustments in orientation. These actions are constrained to the physical limits of the AUV thrusters and rotational capabilities.

3. Transition function

The dynamics of the AUV, as implemented in the simulation environment, govern the transition from the current state (s_t) to the next state (s_{t+1}). The transitions are described as:

$$x_{t+1} = x_t + \Delta_x + \text{disturbance}_x \quad (3)$$

$$y_{t+1} = y_t + \Delta_y + \text{disturbance}_y \quad (4)$$

$$z_{t+1} = z_t + \Delta_z + \text{disturbance}_z \quad (5)$$

$$\text{roll}_{t+1} = mod(\text{roll}_t + a_{\text{roll}}, 360) \quad (6)$$

$$\text{pitch}_{t+1} = mod(\text{pitch}_t + a_{\text{pitch}}, 360) \quad (7)$$

$$\text{yaw}_{t+1} = mod(\text{yaw}_t + a_{\text{yaw}}, 360) \quad (8)$$

$$battery\ level_{t+1} = \max\left(0, battery\ level_t\ -\ \eta \sum |a\_t|\right) \quad (9)$$

4. Reward function

The reward function is designed to incentivize efficient and goal-oriented navigation while penalizing unsafe or inefficient behavior. It is expressed as:

$$[r(s_t, a_t) = r_1 + r_2 + r_3 - r_4 - r_5 - r_6,] \quad (10)$$

where:

$(r_1 = \alpha \cdot \text{distance\_reduction})$

$(r_2 = \beta \cdot \text{smooth\_action})$

$(r_3 = \gamma \cdot \text{battery\_efficiency})$

$(r_4 = \delta / \text{proximity\_to\_obstacle})$

$(r_5 = \eta \cdot \text{excessive\_action})$

$(r_6 = \lambda \cdot \text{goal\_deviation})$

The parameters $(\alpha = 10, \beta = 5, \gamma = 3, \delta = 50, \eta = 2, \lambda = 8)$ are empirically tuned.

5. Objective

The reward function is designed to incentivize efficient and goal-oriented navigation while penalizing unsafe or inefficient behavior. It is expressed as:

$$J = E\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right], \quad (11)$$

where $(\gamma \in (0,1))$ is the discount factor, and (T) is the episode length.

The objective of the reinforcement learning problem is to maximize the expected cumulative discounted reward:

$$J = E\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right], \quad (12)$$

where $(\gamma \in (0,1))$ is the discount factor, and (T) is the episode length.

C. *Proposed Model*

The Deep Deterministic Policy Gradient (DDPG) algorithm is employed as the RL framework. DDPG is a model-free, off-policy algorithm that is well-suited for environments with continuous action spaces, such as AUV control. It combines actor-critic methods, where the actor learns a deterministic policy, and the critic evaluates the policy using a Q-value function.

1. Actor network

The actor network is a neural network that maps the current state (s_t) to a continuous action (a_t) . It comprises:

- **Input:** State vector (s_t) .
- **Hidden layers:** Two fully connected layers with 256 units each and ReLU activation.

- **Output layer:** A tanh activation function to constrain actions within defined bounds.

2. Critic Network

The critic network estimates the Q-value *($Q(s_t, a_t)$)*, which represents the expected return for a given state-action pair. It comprises:

- **Inputs:** State *(s_t)* and action *(a_t)*.
- **Hidden layers:** Two fully connected layers with 256 units each and ReLU activation.
- **Output layer:** A linear activation to produce the scalar Q-value.

3. Training process

The training procedure involves episodic interactions between the AUV and the environment. The agent explores the environment using Ornstein-Uhlenbeck noise to encourage diverse actions. Parameter updates are based on the following rules:

$$[\theta_Q \leftarrow \theta_Q + \alpha\nabla_{\theta_Q} E[r + \gamma Q'(s',\mu'(s')) - Q(s,a)],]$$
$$[\theta_\mu \leftarrow \theta_\mu + \beta\nabla_{\theta_\mu} Q(s,\mu(s)).] \quad (13)$$

III. RESULTS AND DISCUSSION

A. Training Performance

The training performance of the AUV is evaluated over multiple episodes, with key metrics such as episode rewards, position, orientation, and battery usage recorded. The following sections analyze the results obtained.

B. Episode Rewards

The episode rewards curve demonstrates the learning progression of the AUV. Initially, rewards fluctuate significantly, indicating exploration of the environment. Over time, the rewards stabilize, suggesting that the agent has learned a policy to navigate effectively while optimizing energy usage and avoiding obstacles.

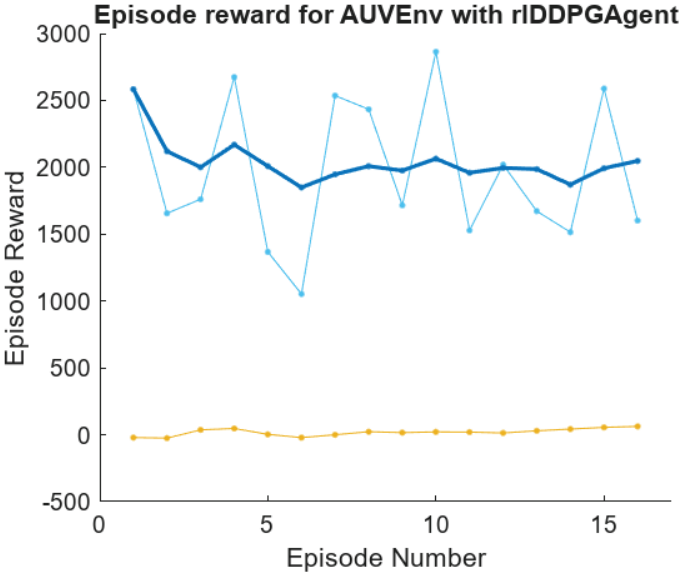

Figure 1. Episode rewards over time

C. Trajectory Analysis

The trajectory of the AUV shows its path in the 3D environment, highlighting its ability to reach the target while avoiding obstacles. The trajectory demonstrates adaptive behavior in navigating through challenging configurations.

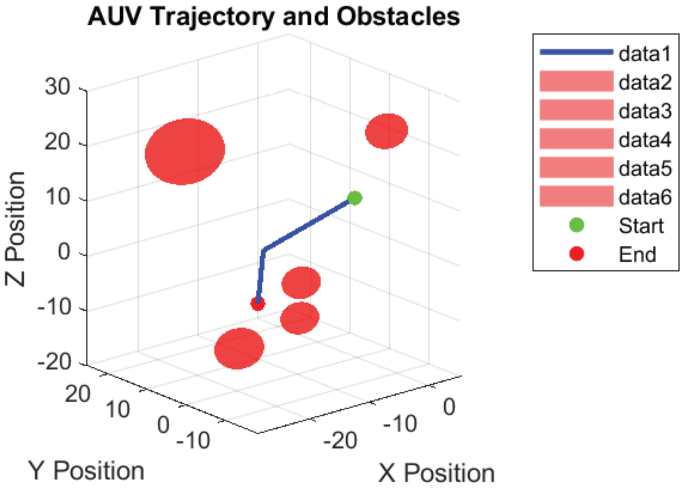

Figure 2. AUV trajectory and obstacles

D. Position and Orientation Analysis

The position and orientation plots provide insights into the control strategy adopted by the AUV. Smooth changes in position indicate efficient navigation, while orientation adjustments show precise control to maintain stability and alignment with the target direction.

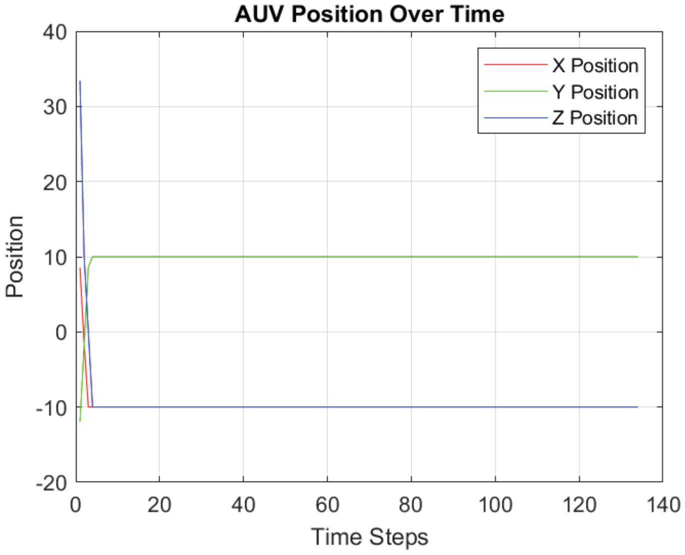

Figure 3. AUV position over time

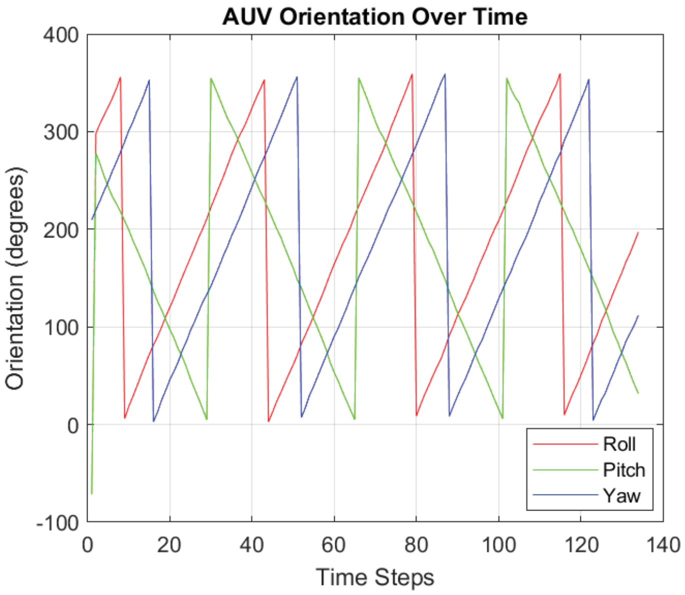

Figure 4. AUV orientation over time

E. *Battery Usage Analysis*

The battery usage plot highlights the energy efficiency of the AUV. The gradual decrease in battery levels indicates optimized energy expenditure, with no sudden drops that would suggest inefficient or excessive actions.

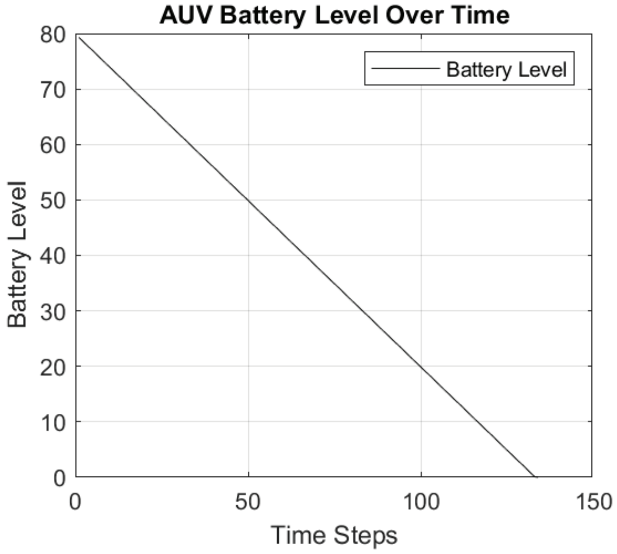

Figure 5. AUV battery level over time

The results demonstrate the effectiveness of the proposed reinforcement learning framework for AUV navigation. The agent successfully learned to balance goal-oriented navigation, obstacle avoidance, and energy optimization. Key strengths include the ability of the agent to adapt to dynamic environments and its efficient use of resources.

However, occasional deviations from optimal behavior, as indicated by spikes in the rewards, suggest areas for further improvement. Future work could explore alternative reward function designs and additional training scenarios to enhance robustness.

IV. CONCLUSION

This research illustrates the efficacy of reinforcement learning (RL), specifically the Deep Deterministic Policy Gradient (DDPG) algorithm, in tackling the obstacles of autonomous navigation for AUVs. The proposed approach effectively allowed the AUV to function in intricate, simulated underwater environments by concentrating on dynamic obstacle avoidance, energy-efficient path planning, and goal-oriented navigation.

The findings underscore numerous significant accomplishments:

The DDPG algorithm demonstrated significant efficacy in continuous control tasks, facilitating smooth, adaptive, and efficient trajectories for the AUV. The incorporation of a multi-objective reward function, which balances navigation, energy efficiency, and safety, markedly enhanced performance across various scenarios. The trained policy exhibited strong obstacle avoidance abilities and energy optimization, significantly decreasing collision rates and efficiently conserving battery usage. Notwithstanding these achievements, numerous limitations persist. The inconsistency of rewards and sporadic divergences from optimal trajectories suggest a necessity for enhanced refinement in the reward framework and training methodology. The sensitivity of the agent to environmental configurations indicates the necessity of integrating a broader range of training scenarios to enhance generalization.

The ramifications of this research transcend simulated contexts. The framework establishes a basis for implementing RL-based navigation systems in practical AUVs, with prospective applications in environmental monitoring, underwater exploration, and search-and-rescue operations. Nonetheless, closing the divide between simulation and reality is a vital focus for future research, necessitating progress in sim-to-real transfer techniques, resilient sensor integration, and adaptive policies proficient in managing real-time environmental disruptions.

Prospective trajectories encompass:

1. Evaluating and confirming the methodology in actual underwater settings to tackle practical issues like hardware limitations and sensor inaccuracies.

2. Expanding the framework to encompass multi-agent systems for collaborative objectives.

3. Implementing advanced energy management strategies, including renewable energy integration, to improve long-duration mission capabilities.

This study highlights the potential of reinforcement learning in enhancing the autonomy and operational efficiency of autonomous underwater vehicles, facilitating the development of more scalable and adaptive underwater robotics solutions. By overcoming current constraints and investigating prospective avenues, reinforcement learning can transform AUV navigation in both academia and the industry.

REFERENCES

[1] T. P. Lillicrap *et al.*, “Continuous control with deep reinforcement learning,” *arXiv Cornell University*, 2015, doi:10.48550/arxiv.1509.02971.

[2] P. Bhopale, F. Kazi, and N. Singh, “Reinforcement Learning Based Obstacle Avoidance for Autonomous Underwater Vehicle,” *Journal of Marine Science and Application*, vol. 18, no. 2, pp. 228–238, Jun. 2019, doi: 10.1007/s11804-019-00089-3.

[3] Y. Jiang, K. Zhang, M. Zhao, and H. Qin, “Adaptive meta-reinforcement learning for AUVs 3D guidance and control under unknown ocean currents,” *Ocean Engineering*, vol. 309, p. 118498, Oct. 2024, doi: 10.1016/j.oceaneng.2024.118498.

[4] J. Wen, A. Wang, J. Zhu, F. Xia, Z. Peng, and W. Zhang, “Adaptive energy-efficient reinforcement learning for AUV 3D motion planning in complex underwater environments,” *Ocean Engineering*, vol. 312, p. 119111, Nov. 2024, doi: 10.1016/j.oceaneng.2024.119111.

[5] R. K. Lea, R. Allen, and S. L. Merry, "A comparative study of control techniques for an underwater flight vehicle," *Int J Syst Sci*, vol. 30, no. 9, pp. 947–964, Jan. 1999, doi: 10.1080/002077299291831.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning, Second Edition An Introduction*. 2017.

[7] Y. Sun, X. Ran, G. Zhang, X. Wang, and H. Xu, "AUV path following controlled by modified Deep Deterministic Policy Gradient," *Ocean Engineering*, vol. 210, p. 107360, Aug. 2020, doi: 10.1016/j.oceaneng.2020.107360.

[8] M. Rahmati, M. Nadeem, V. Sadhu, and D. Pompili, "UW-MARL," in *Proceedings of the International Conference on Underwater Networks & Systems*, New York, NY, USA: ACM, Oct. 2019, pp. 1–5. doi: 10.1145/3366486.3366533.

[9] N. Anh Vien, N. Hoang Viet, S. Lee, and T. Chung, "Heuristic Search Based Exploration in Reinforcement Learning," in *Computational and Ambient Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 110–118. doi: 10.1007/978-3-540-73007-1_14.

[10] B. Hadi, A. Khosravi, and P. Sarhadi, "Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle," *Applied Ocean Research*, vol. 129, p. 103326, Dec. 2022, doi: 10.1016/j.apor.2022.103326.

[11] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," 2016.

[12] A. Ramezani Dooraki and D.-J. Lee, "A Multi-Objective Reinforcement Learning Based Controller for Autonomous Navigation in Challenging Environments," *Machines*, vol. 10, no. 7, p. 500, Jun. 2022, doi: 10.3390/machines10070500.

[13] K. Sivamayil, E. Rajasekar, B. Aljafari, S. Nikolovski, S. Vairavasundaram, and I. Vairavasundaram, "A Systematic Study on Reinforcement Learning Based Applications," *Energies (Basel)*, vol. 16, no. 3, p. 1512, Feb. 2023, doi: 10.3390/en16031512.

[14] Z. Wang *et al.*, "Toward Communication Optimization for Future Underwater Networking: A Survey of Reinforcement Learning-Based Approaches," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2024, doi: 10.1109/COMST.2024.3505850.