AUV Waypoint-based Guidance using Deep Reinforcement Learning

Mostafa Ahmed Atef Kotb
Egypt-Japan University of Science and Technology, Alexandria, Egypt.
mostafa.kotb@ejust.edu.eg

***Abstract*—In today's dynamic landscape, Remotely Operated Underwater Vehicles (ROVs) have emerged as irreplaceable tools, contributing significantly to diverse fields such as marine research, oil and gas energy exploration, and underwater networks. However, the manual control of ROVs poses challenges due to the unpredictable nature of human behavior and the demand for high skill levels. This challenge has catalyzed the introduction of Autonomous Underwater Vehicles (AUVs) as viable alternatives for specific tasks. This paper presents a solution aimed at enhancing the efficiency of underwater operations by proposing a waypoint-based guidance system for AUVs through deep reinforcement learning. The primary methodology employed in this study is the Proximal Policy Optimization (PPO) algorithm. To simulate the underwater environment accurately, the HoloOcean Simulator was utilized as a robust platform for training and evaluation.**

The focal point of the research involves training the AUV to comprehend its kinematic model, navigate effectively through the underwater terrain, and accomplish designated tasks autonomously. The trained model exhibits adeptness in avoiding collisions and near-misses while sequentially reaching predefined waypoints. This innovative approach not only minimizes the need for human intervention but also paves the way for increased contributions across various domains. The proposed model not only showcases the capabilities of AUVs but also exemplifies the power of deep reinforcement learning in navigating underwater environments. This paper contributes to the evolving landscape of autonomous underwater systems, opening avenues for future advancements in underwater robotics and exploration.

***Index Terms*—AUV, Waypoint Guidance, PPO, HoloOcean, Policy**

I. Introduction

In the vast and dynamic realm of underwater exploration, ROVs have evolved into indispensable instruments, playing crucial roles across diverse fields such as marine research, oil and gas energy exploration, and underwater infrastructure maintenance. Their versatility in conducting tasks such as data collection, inspection, and repair of submerged structures has made them essential assets in the exploration and utilization of the ocean's depths. However, the intricate and challenging nature of underwater environments demands precise control and maneuverability, a requirement that poses challenges when relying on manual operation.

The reliance on manual control for ROVs introduces a set of challenges stemming from the unpredictable nature of human behavior and the demand for high skill levels, hindering the efficiency and reliability of underwater operations. In environments where precision and adaptability are most important, the limitations of manual operation become particularly inconvenient. The quest for a solution to these challenges has given rise to the emergence of AUVs as viable alternatives. AUVs offer the advantage of autonomy, freeing them from the constraints of manual control and enabling them to execute specific tasks with enhanced efficiency.

The transition from manual to autonomous underwater vehicles is driven by the imperative to overcome the limitations imposed by human-operated systems, unlocking new possibilities in underwater exploration and applications. However, the advent of AUVs introduces its own set of challenges, primarily centered around the need to train these vehicles to operate autonomously in complex underwater terrains.

To address this challenge, this paper proposes a novel waypoint-based guidance system for AUVs, leveraging the power of deep reinforcement learning. The use of reinforcement learning allows the AUV to learn and adapt its behavior based on interactions with the environment, reducing the need for explicit programming and manual intervention.

This represents a significant shift towards enhancing the autonomy and adaptability of AUVs, ultimately contributing to more efficient and reliable underwater operations.

The primary methodology employed in this study is the Proximal Policy Optimization algorithm, a state-of-the-art reinforcement learning algorithm known for its stability and efficiency in training complex models. PPO strikes a delicate balance between exploration and exploitation, making it well-suited for training AUVs to navigate through unpredictable underwater environments. By utilizing PPO, this research aims to enhance the AUV's ability to comprehend its kinematic model, navigate effectively, and autonomously accomplish designated tasks, ultimately reducing the dependency on manual control.

To facilitate accurate simulation of the underwater environment for training and evaluation, the HoloOcean Simulator is employed as a robust platform. This simulator enables the realistic representation of underwater conditions, providing a controlled yet challenging environment for the AUV to learn and refine its navigation capabilities. The incorporation of this simulation platform ensures that the AUV is exposed to a diverse range of scenarios, preparing it for real-world challenges in complex underwater terrains.

The core focus of this research lies in training the AUV to autonomously navigate underwater terrains, avoiding collisions and near-misses, while sequentially reaching predefined waypoints. This innovative approach not only addresses the challenges posed by manual control but also opens avenues for increased efficiency and contributions across various domains, from marine research to industrial applications.

In conclusion, this paper makes a significant contribution to the evolving landscape of autonomous underwater systems, showcasing the capabilities of AUVs and exemplifying the power of deep reinforcement learning in navigating challenging underwater environments. By introducing a waypoint-based guidance system and utilizing the PPO algorithm, this research paves the way for future advancements in underwater robotics and exploration, promising increased autonomy and reliability in underwater operations. The integration of autonomous systems in underwater exploration holds the potential to revolutionize how we interact with and understand the vast and mysterious depths of our oceans.

II. Related Work

The integration of deep reinforcement learning (DRL) to enhance AUV navigation and autonomy has garnered considerable attention, with numerous studies delving into diverse algorithms and training methodologies. The exploration primarily focuses on addressing challenges related to waypoint-based navigation and obstacle avoidance in intricate underwater environments. This section provides a comprehensive overview of related work, categorized into waypoint navigation and obstacle avoidance with DRL, underscoring the pivotal role of simulation platforms for effective training.

A. Waypoint Navigation with DRL

A pioneering study by Zhang et al. [1] introduce a hybrid DRL approach for Unmanned Aerial Vehicles (UAVs), combining global waypoints with real-time obstacle avoidance through the Soft Actor-Critic (SAC) algorithm. The proposed model exhibits adaptability to complex environments, showcasing superior collision avoidance compared to traditional methods. The architecture, depicted in the figure below, highlights the interaction between the human operator, the SAC policy network, and the environment.

Another noteworthy contribution comes from Zhang et al. [2], employing the Double Deep Q-learning (DDQL) algorithm for achieving waypoint navigation in aerial vehicles. Emphasizing efficient training through prioritized experience replay, the method demonstrates successful navigation in both nominal and disturbed flight conditions.

Building on this, Himanshu et al. [3] explore the application of the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm for high-level and low-level control of quadrotor UAVs during waypoint navigation. Results exhibit successful navigation and improved performance compared to conventional control methods, particularly in the presence of disturbances.

B. Obstacle Avoidance with DRL

In the realm of obstacle avoidance, Sola et al. [4] implement the SAC algorithm to control and guide

an AUV for waypoint tracking tasks, resulting in improved energy efficiency and reduced information requirements compared to traditional Proportional-Integral-Derivative (PID) controllers.

Additionally, Havenstrøm et al. [5] propose a DRL framework for obstacle avoidance and task completion in underwater environments, employing the PPO algorithm. The designed reward function encourages goal achievement and collision prevention, leading to successful navigation and task completion in simulated scenarios with dynamic obstacles.

C. Simulators for DRL Training

Researchers, recognizing the significance of effective training platforms, have utilized various simulators to enhance DRL-based AUV control. Chu et al. [6] utilize the MuJoCo simulator to train an AUV for motion control using a deep imitation algorithm, showcasing successful navigation in simulated environments.

In another study, Xie et al. [7] employ the MORSE simulator to train an UAV for tracking and landing using a DDPG algorithm. The research underscores the efficacy of DRL in handling complex scenarios, emphasizing the flexibility of using various simulation platforms tailored to specific environments.

These examples underscore the extensive research conducted on DRL-based AUV navigation and obstacle avoidance. Leveraging the power of DRL and effective simulation platforms, researchers are advancing the capabilities of AUVs, making them more adept at handling challenging underwater tasks in real-world applications.

III. Methodology

This section outlines the methodology employed in the research, encompassing the preparation of the simulation environment, the design of the reward function, and the training of the agent. The entirety of this work is accessible on the GitHub repository: ROV-Autonomous-Waypoint-Guidance.

A. HoloOcean

The research primarily utilized the HoloOcean Simulator, a realistic underwater robotics simulator based on Holodeck [8]. HoloOcean features multi-agent missions, a variety of underwater sensors, including an innovative imaging sonar sensor implementation, easy installation, and user-friendly operation. It offers extensive customization options with various agents, sensors, and scenarios, allowing the design of custom environments for multiple experiments. Table 1 provides a comparison between different underwater simulations, highlighting HoloOcean's superiority.

Designing the experiment

For the waypoint guidance experiment, designated points were spawned for the agent to follow in a predefined order. Square targets simulated waypoints, while spherical probes represented obstacles to be avoided. Target and obstacle positions were randomly generated within a 100x100x100 cube. Refer to Figure 1 for a visualization of the experiment environment.

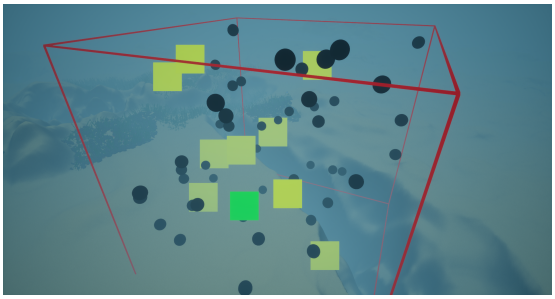

Fig. 1: Experiment Environment

Customizing the Agent

HoloOcean has a wide selection of agents, like TurtleAgent, Torpedo, etc. The HoveringAUV agent, equipped with an 8-thruster configuration enabling movement in 6 Degrees of Freedom (DoFs), was selected for the experiment. Custom sensors were added to the agent to suit the guidance mission:

1) Pose Sensor: Provides the forward, right, and up vectors, along with the agent's XYZ location.
2) Velocity Sensor: Measures x, y, and z velocity in the global frame.
3) Rotation Sensor: Obtains the agent's rotation in the world (Roll, Pitch, Yaw) in degrees.
4) Range Finder Sensors: Return distances to nearest collisions in specified directions, enabling the agent to perceive its surroundings. Multiple sensors were added for comprehensive coverage.

Simulator	Multi-Agent	Communications	Sonar	Open Source	Small Dependencies	Maintained	Simple Mission Setup
HoloOcean	✓	✓	✓	✓	✓	✓	✓
UUV Simulator	✓	×	⋆	✓	×	⋆	✓
UWSim	✓	×	⋆	✓	×	×	×
MarineSim	✓	×	×	×	×	×	⋆

TABLE I: Comparison of common underwater simulators. ✓denotes that the simulator has a feature, ⋆ that it has a limited implementation or is unknown, and × that it does not. There are many other common robotics simulators that are not listed here, but most don't have support for underwater robotics.

Refer to Figure 2 for the HoveringAUV agent with distributed range finder sensors.

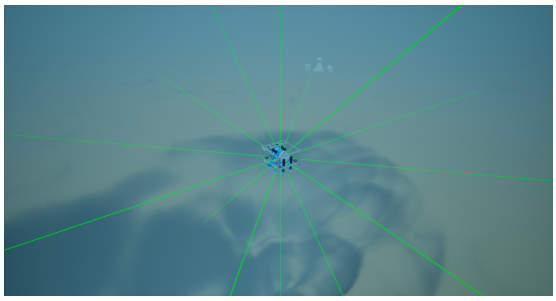

Fig. 2: Agent with range finder sensors

B. Designing the reward function

The reward function defines how the agent will learn by defining when will the agent be rewarded and when will it be penalized. The first edition of reward function was:

$$\begin{aligned}\text{Reward} = &(-30)\cdot\text{Collision}\\ &+(-5)\cdot\text{Near-miss}\\ &+(-1)\cdot\text{Incline}\\ &+(1)\cdot\text{Towards\_target}\\ &+(30)\cdot\text{Reach\_target}\end{aligned} \tag{1}$$

- Collision: colliding with obstacle.
- Near-miss: distance between the agent and the obstacle is less than 1.
- Incline: degree of inclination in roll or pitch.
- Towards_target: moving in the direction of the target.
- Reach_target: reaching the actual target.

The previous reward function was defective as the agent would break it by either getting outside the designated area or staying in its place without moving. The final reward function was:

$$\begin{aligned}\text{Reward} = &(-100)\cdot\text{Outside\_box}\\ &+(-30)\cdot\text{Collision}\\ &+(-5)\cdot\text{Near-miss}\\ &+(-1)\cdot\text{Incline}\\ &+(-1)\cdot\text{Static}\\ &+(1)\cdot\text{Towards\_target}\\ &+(30)\cdot\text{Reach\_target}\\ &+(1000)\cdot\text{Complete\_game}\end{aligned} \tag{2}$$

- Outside_box: getting outside the pre-defined box.
- Static: staying in the same position for 50 ticks.
- Complete_game: reaching all targets.

C. Training the Agent

The training of the AUV agent for waypoint-based guidance in this study is performed using the Proximal Policy Optimization algorithm. PPO is selected for its suitability in addressing the challenges posed by continuous action spaces, a common characteristic of AUV navigation tasks.

PPO is a state-of-the-art reinforcement learning algorithm that strikes a balance between sample efficiency and stability during training [9]. Its policy optimization approach involves iteratively collecting experiences from the environment, updating the policy to maximize expected rewards, and incorporating a clipped objective function to prevent large policy updates. This property makes PPO well-suited for continuous control tasks, such as waypoint-based guidance, where the agent needs to navigate through a three-dimensional space.

The choice of PPO for AUV guidance is underpinned by its ability to handle high-dimensional state and action spaces, making it suitable for the complex and dynamic underwater environments.

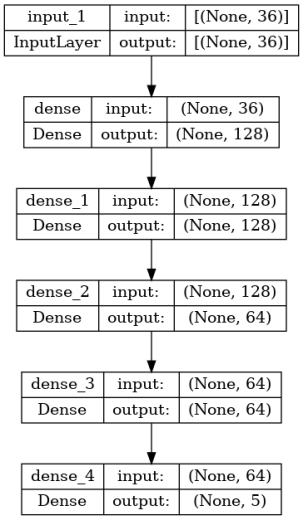

Fig. 3: Policy Network Structure

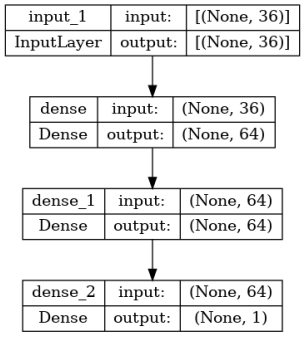

Fig. 4: Value Network Structure

Additionally, PPO's stability ensures smoother convergence during training, mitigating the risk of policy oscillations that could impede the learning process.

During training, the AUV interacts with the environment, where the agent receives observations about its state from the mentioned sensors. The designed reward function guides the learning process. The PPO agent's policy and value networks are updated based on advantages calculated from the observed rewards and predicted values. The structure of the networks is shown in Figures 3 and 4.

IV. Results

The model's performance was constrained by limited computational resources available for training. The training duration spanned 700 episodes using the latest iteration of the reward function. Each episode was restricted to 10,000 steps, with a reward threshold set at -1000. Figure 5 provides insights into policy and value losses per episode, along with the total gained reward per episode.

Despite the resource limitations, the model's performance appears promising. Over the 700 iterations, the model demonstrates a trend of accumulating positive rewards across multiple episodes. This suggests that the model has begun to grasp the controls of the AUV and comprehend the environmental features. While the results are encouraging, further training with increased resources could potentially yield more refined and robust outcomes.

V. Discussion

In this section, we will analyze the previous results and discuss the development process leading to the final outcomes.

A. Using DPPG

The application of the Deep Deterministic Policy Gradients (DDPG) algorithm in training the AUV for the experiment faced significant challenges. Despite DDPG's suitability for continuous action spaces, the complexities of the continuous observation and action spaces in the AUV scenario hindered effective learning.

The AUV's extensive sensor input alongside the need for precise navigation and obstacle avoidance, led to a high-dimensional observation space. Similarly, controlling multiple thrusters translated into a complex continuous action space.

DDPG struggled to converge efficiently in this intricate underwater environment, highlighting the difficulties associated with applying deep reinforcement learning in such continuous and dynamic spaces. Challenges may arise from issues like exploration-exploitation balance, hyperparameter tuning, or the need for more specialized algorithms tailored to specific environmental dynamics.

The limitations observed suggest avenues for future research, including exploring advanced algorithms, incorporating domain knowledge, or investigating ensemble techniques. These approaches aim to address the challenges posed by the continuous nature of the AUV guidance problem, opening doors for more effective learning in complex robotics scenarios.

B. Using PPO with the first reward function

The first reward function design aimed to motivate the AUV to navigate successfully through the waypoints. However, during the training with the PPO algorithm, a notable challenge emerged. The model learned to exploit the reward structure by finding suboptimal strategies that led to accumulating rewards without achieving the primary task.

Specifically, the model discovered that it could maximize rewards by deviating from the desired

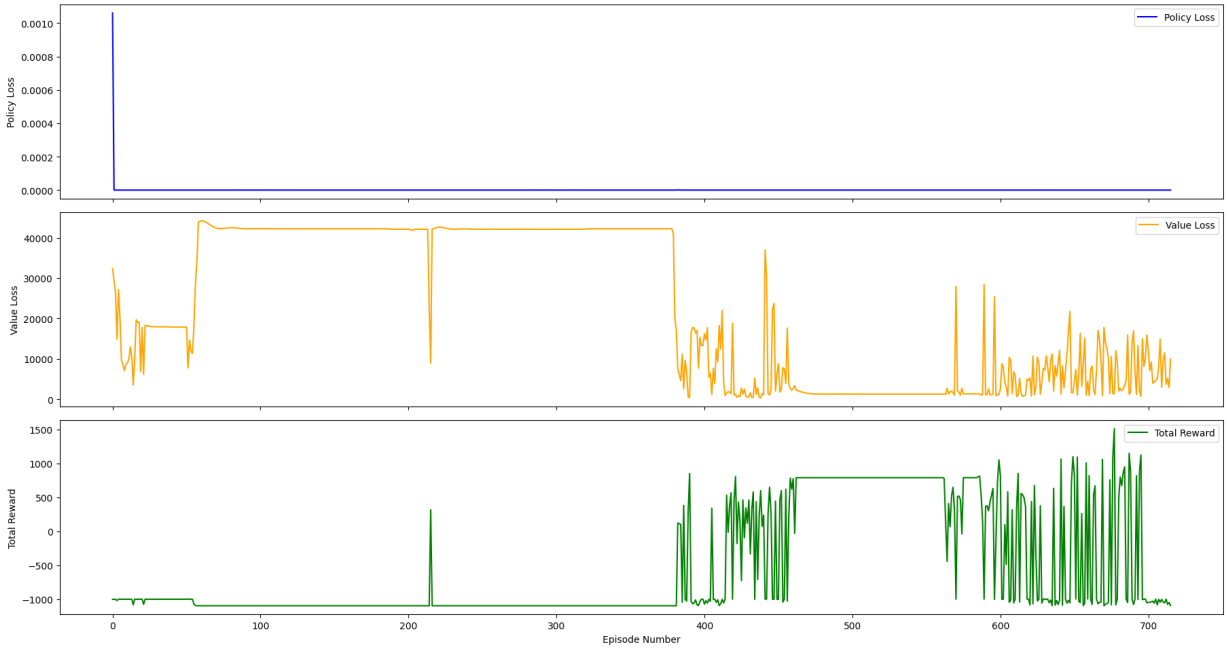

Fig. 5: Policy Loss, Value Loss, Total Reward per episode

behavior. It learned to exploit the system by either going outside the predefined box or remaining static, both of which resulted in positive rewards or at least no penalties. This unintended behavior posed a critical issue, as the agent prioritized obtaining rewards through shortcuts rather than completing the waypoint-based guidance task as intended.

This learning quirk reflects the sensitivity of reinforcement learning algorithms to reward functions. The model, driven by the goal of maximizing cumulative rewards, found alternative routes to success that were inconsistent with the desired behavior.

C. Using PPO with the second reward function

Recognizing the challenges posed by the initial reward function, the second reward function was developed to rectify the observed exploitation issues. This revised reward structure incorporated penalties for undesired behaviors and adjusted reward magnitudes to better guide the learning process:

1) Penalty for Outside the Box: To discourage the AUV from deviating outside the predefined operational area, a substantial penalty was introduced for such actions. This penalty aimed to emphasize the importance of staying within the designated boundaries, aligning the model's behavior with the task requirements.
2) Penalty for Static Behavior: Another critical modification involved introducing a penalty for static behavior. By penalizing the AUV for remaining stationary for a specified duration (50 ticks), the reward function discouraged the model from exploiting the system by staying in one place without actively engaging in the guidance task.
3) Enhanced Target-Reaching Reward: The reward for successfully reaching the target was significantly increased to reinforce the importance of completing the navigation task. In addition, the reward of completing the game was added. This adjustment aimed to counterbalance the unintended shortcuts discovered by the model in the initial reward function.

These modifications had a profound impact on the learning dynamics. By penalizing undesired actions and reinforcing task-specific accomplishments, the model learned a more faithful representation of the desired behavior. The refined reward function successfully steered the training process away from unintended shortcuts, fostering a more robust and task-oriented learning experience.

VI. Conclusion

In conclusion, this paper presents a comprehensive approach to enhancing the efficiency of underwater operations through a novel waypoint-based

guidance system for AUVs using deep reinforcement learning. The primary methodology employed in this study is the PPO algorithm, showcasing its suitability for training AUVs to navigate complex underwater terrains.

The research addresses the challenges associated with manual control of ROVs, emphasizing the unpredictable nature of human behavior and the demand for high skill levels. The transition to AUVs offers autonomy and adaptability, freeing them from the constraints of manual control and enabling them to execute specific tasks with enhanced efficiency.

The proposed waypoint-based guidance system leverages the power of deep reinforcement learning, allowing the AUV to learn and adapt its behavior based on interactions with the environment. The HoloOcean Simulator serves as a robust platform for training and evaluation, providing a realistic representation of underwater conditions.

The core focus of the research involves training the AUV to autonomously navigate underwater terrains, avoid collisions and near-misses, and sequentially reach predefined waypoints. The introduced reward function proves effective in guiding the learning process and steering the model away from unintended shortcuts.

The paper outlines challenges in DDPG algorithm development, including limitations and reward function refinement for exploitation concerns. These insights guide future research, encouraging exploration of advanced algorithms, domain knowledge incorporation, and reward structure refinement. The findings emphasize deep reinforcement learning's potential in navigating underwater environments, contributing to autonomous underwater systems.

Finally, this research makes a significant contribution to the field of autonomous underwater systems, showcasing the capabilities of AUVs and exemplifying the power of deep reinforcement learning in addressing complex navigation tasks. The proposed waypoint-based guidance system opens avenues for future advancements in underwater robotics and exploration, promising increased autonomy and reliability in underwater operations.

VII. References

References

[1] S. Zhang, Y. Li, F. Ye, X. Geng, Z. Zhou, T. Shi, "A Hybrid Human-in-the-Loop Deep Reinforcement Learning Method for UAV Motion Planning for Long Trajectories with Unpredictable Obstacles," *Drones*, vol. 7, no. 5, p. 311, 2023, doi: 10.3390/drones7050311.

[2] Y. Zhang, Y. Zhang, and Z. Yu, "Path following control for UAV using deep reinforcement learning approach," *Guid. Navig. Control*, vol. 1, no. 1, pp. 2150005, 2021. doi: 10.1142/S2737480721500059

[3] K. Himanshu, Hari Kumar, Jinraj V Pushpangathan, "Waypoint Navigation of Quadrotor using Deep Reinforcement Learning," *IFAC-PapersOnLine*, vol. 55, no. 22, pp. 281-286, 2022, ISSN 2405-8963, doi: 10.1016/j.ifacol.2023.03.047.

[4] Y. Sola, G. Le Chenadec, B. Clement, "Simultaneous Control and Guidance of an AUV Based on Soft Actor–Critic," *Sensors*, vol. 22, pp. 6072, 2022, doi: 10.3390/s22166072.

[5] S. T. Havenstrøm, A. Rasheed, O. San, "Deep Reinforcement Learning Controller for 3D Path Following and Collision Avoidance by Autonomous Underwater Vehicles," *Frontiers in Robotics and AI*, vol. 7, 2021. doi: 10.3389/frobt.2020.566037.

[6] Z. Chu, B. Sun, D. Zhu, M. Zhang, and C. Luo, "Motion control of unmanned underwater vehicles via deep imitation reinforcement learning algorithm," *IET Intelligent Transportation Systems*, vol. 14, pp. 764-774, 2020. doi: 10.1049/iet-its.2019.0273.

[7] J. Xie, X. Peng, H. Wang, W. Niu, and X. Zheng, "UAV Autonomous Tracking and Landing Based on Deep Reinforcement Learning Strategy," *Sensors*, vol. 20, pp. 5630, 2020. doi: 10.3390/s20195630.

[8] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "HoloOcean: An Underwater Robotics Simulator," *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 3040-3046, doi: 10.1109/ICRA46639.2022.9812353.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," 2017, arXiv preprint, arXiv:1707.06347. [Online]. Available: https://arxiv.org/abs/1707.06347