

# ***Obstacle Avoidance Path Planning for Autonomous Underwater Vehicles Based on TRPO and PPO Algorithms: A Review***

**Zhihao Yan**

*School of Mechanical Engineering, Hubei Engineering University, Xiaogan, China  
023301634214yzh@stu.hbeu.edu.cn*

**Abstract.** The Autonomous Underwater Vehicle (AUV) has become a key tool for underwater exploration and resource investigation. However, in the complex underwater environment, the existing AUV still depends on manual operation, and the automatic obstacle avoidance capability and real-time decision-making performance are insufficient. This paper addresses the obstacle avoidance problem of autonomous underwater vehicles, reviews the research progress of reinforcement learning, and highlights the differences in the use of Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) in path planning. In view of AUV underwater nonlinear dynamics and complex environmental constraints, the general methods of the two methods in state space modeling, action space design and reward function construction are summarized, and their performance in strategy stability, training efficiency and environmental adaptability are compared and analyzed. Research shows that TRPO guarantees the stability and monotonic improvement of the policy iteration through a KL divergence constraint on a policy update mechanism, but the computational overhead is large, which is not conducive to embedded applications. Studies indicate that, in terms of policy update mechanisms, TRPO ensures the stability and monotonic improvement of policy iteration through a KL divergence constraint; nevertheless, its high computational cost limits applicability in embedded systems. In contrast, PPO streamlines optimization via objective clipping, improving training efficiency and practical use, though policy oscillations can occur in complex, obstacle-rich environments. Consequently, this paper further discusses potential research directions, including lightweight algorithm improvements, multi-agent cooperative obstacle avoidance, and Sim-to-Real transfer.

**Keywords:** Autonomous Underwater Vehicle, Trust Region Policy Optimization, Proximal Policy Optimization, obstacle Avoidance Algorithm

## **1. Introduction**

The Autonomous Underwater Vehicle (AUV) has been widely employed in water quality monitoring, seabed mapping, and underwater exploration due to its small size, unmanned operation, and flexible characteristics. It has become an important research object in the field of marine

engineering and intelligent control. However, it is significantly influenced by nonlinear, coupled, and underactuated dynamics in underwater environments, and precise hydrodynamic parameters are hard to acquire. Therefore, model-based control methods, such as PID, sliding mode, and model predictive control, show reduced performance under ocean current disturbances and sensor noise [1]. And this highlights the challenge of achieving robust and safe autonomous control and obstacle avoidance in uncertain environments [2]. At the method level, traditional obstacle avoidance algorithms such as A\*, the artificial potential field method, the rapidly exploring random tree (RRT), and the dynamic window approach, achieve high computational efficiency in structured or static environments. These methods generally rely on environmental models and face challenges such as limited generalization, infeasible paths, and difficulty handling multi-objective constraints in unknown dynamic underwater scenarios. In contrast, Deep Reinforcement Learning (DRL) enables end-to-end policy learning via interaction with the environment. It reduces reliance on explicit models and offers a new approach for complex underwater obstacle avoidance. However, existing strategy optimization methods based on TRPO and PPO still exhibit limited stability and adaptability under strong disturbances, depth constraints, and three-dimensional coupled control conditions. Based on this, this paper explores the use of TRPO and PPO in three-dimensional AUV obstacle avoidance tasks, and investigates their performance differences and applicable boundaries under different obstacle environments and control constraints. It may provide a reference for the related research of underwater autonomous obstacle avoidance control methods.

## 2. Motion characteristics and obstacle avoidance basis of autonomous underwater vehicle

### 2.1. Underwater motion characteristics and control

Under the underactuated constraint in the depth dimension, AUV motion exhibits strong coupling between attitude and velocity, which differs from that of two-dimensional motion systems [3]. In the state description, the AUV usually uses the inertial coordinate system to represent the spatial position  $(x, y, z)$  and attitude angle  $(\varphi, \theta, \psi)$ , and the body coordinate system to represent the linear velocity  $(u, v, w)$  and angular velocity  $(p, q, r)$ . Due to the underactuation of the vertical channel, the three-dimensional motion needs to coordinate the control allocation of the horizontal and vertical directions at the same time. When obstacles are distributed across different depth layers, the system adjusts the pitch angle  $\theta$  to achieve diving and surfacing indirectly, as direct control of the vertical velocity  $w$  is not possible. Thus, depth control exhibits inherent indirectness and constraints. At the dynamic level, the vertical motion is primarily influenced by buoyancy and fluid damping, and its dynamics can be expressed as:

$$M\dot{v} + C(v)v + D(v)v + g(\eta) = \tau \quad (1)$$

The restoring force term ensures attitude stability, while the nonlinear velocity-dependent damping effects, such as the  $|w|w$  term, impose constraints on vertical acceleration. Regarding control coupling, interactions exist between the control surfaces and the propulsion system. Deflection of the control surfaces influences heading and pitch, while variations in thruster speed affect forward velocity and vertical moment, reflecting the coupling between horizontal and vertical motion.

## 2.2. Obstacle avoidance tasks and environmental challenges

The complexity of underwater obstacle avoidance task mainly comes from the partial observability and dynamic uncertainty of environmental state [4]. Due to the limited underwater visibility, AUV usually relies on sonar to obtain sparse and low-resolution distance information. Meanwhile, optical sensors cannot work effectively, thus making the environmental perception obviously incomplete. At the dynamic level, ocean current shear continuously alters the relative motion between the target and the carrier, reducing the effectiveness of static path-planning methods in underwater environments. In terms of spatial constraints, obstacles may be distributed in different depth layers, while AUV cannot directly control the vertical velocity  $w$ , and can only adjust the depth indirectly through the pitch angle  $\theta$ . At the same time, it is limited by buoyancy, damping and thruster thrust, resulting in weaker vertical maneuverability than that of horizontal motion, thus increasing the difficulty of three-dimensional obstacle avoidance. In addition, in terms of perception and environmental coupling, sonar has the problem of time delay and blind area, and close-range obstacles may be missed. Multi-path acoustic reflection and background noise further amplify the observation uncertainty and make the state information more sparse. Therefore, this kind of task requires the control strategy to realize online decision-making under the condition of lack of complete environment model, and to coordinate horizontal and vertical motion at the same time to meet the safety requirements of obstacle avoidance in three-dimensional space.

## 2.3. Limitations and problems of obstacle avoidance methods

In complex underwater environments, the motion constraints and environmental uncertainty of AUV make it challenging for traditional hierarchical obstacle avoidance methods to ensure feasible and safe paths [4, 5]. Due to the serious attenuation of optical signals by underwater media, AUV mainly relies on sonar to obtain environmental information. However, sonar observation usually has the problems of low resolution, sparse sampling and time delay, which makes it difficult to identify the location and boundary of obstacles stably. Besides, water flow disturbance and ocean current shear will cause the relative velocity change between the target and the carrier, so that the environmental state shows obvious time-varying characteristics, which further increases the uncertainty of state prediction. On this basis, AUV itself is also affected by the coupling of nonholonomic constraints and control input. For example, the vertical motion needs to be adjusted indirectly through the pitch angle, and the propeller and the rudder surface are coupled to each other, which easily leads to deviation between the expected trajectory and the actual execution. When the environmental disturbance is enhanced or the depth changes rapidly, this deviation will be further amplified, thus affecting the enforceability and stability of the path. Thus, the key bottleneck of the existing methods lies in the challenge of reliably representing the environmental state under partial observation and in keeping planned paths consistent with actual control under dynamic constraints.

## 3. Improvements and performance of reinforcement learning algorithms in AUV obstacle avoidance

### 3.1. Task requirements and control constraints

In AUV underwater path-planning tasks, the unstructured environment, sparse observations, and time-varying flow disturbances demand strong online decision-making and adaptive capabilities to handle uncertainty from incomplete state information and rapid environmental changes. On this

basis, through continuous interaction with the environment, DRL learns the mapping relationship from state to action without explicit environmental modeling, and uses reward feedback to gradually optimize the strategy, so as to adapt to the dynamic changes and perceived noise effects in complex underwater scenes. Therefore, under continuous action space and multiple constraints, DRL can improve the real-time adjustment ability and robustness of the strategy to a certain extent, making it more suitable for AUV obstacle avoidance and path planning tasks in complex underwater environments [6, 7].

### 3.2. Algorithm principle and implementation mechanism

Under the framework of KL divergence constraint, TRPO controls each policy update within the trust region by limiting the distribution difference between the old and new policies. Moreover, the stable iterative optimization of the strategy is realized by maximizing the surrogate objective based on the generalized advantage function. Since the surrogate objective only provides a local performance improvement guarantee, its solution usually adopts a first-order linear approximation and a second-order approximation of the KL constraint, hence transforming the original problem into a constrained optimization form [8]. Its standard form is as follows:

$$\max_{\theta} E_t \left[ \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_t \right], \text{ s.t. } E_t [D_{KL}(\pi_{\theta_{old}}(\cdot|s_t) \parallel \pi_{\theta}(\cdot|s_t))] \leq \delta \quad (2)$$

where  $\pi_{\theta}$  is the current policy,  $\pi_{\theta_{old}}$  denotes the old policy, the policy ratio characterizes the change in action probabilities, and  $A_t$  refers to the advantage function, which measures the relative value of an action. Further, KL divergence is used to characterize the difference in policy distribution, which is defined as:

$$D_{KL}(\pi_{\theta_{old}} \parallel \pi_{\theta}) = E_{t \sim \pi_{\theta_{old}}} \left[ \log \frac{\pi_{\theta_{old}}(s)}{\pi_{\theta}(s)} \right] \quad (3)$$

When the KL divergence is 0, the two strategies are exactly the same. As the value increases, the degree of strategy offset increases. Therefore, TRPO constrains policy updates via KL divergence to ensure stable iteration. At the implementation level, TRPO usually combines conjugate gradient and linear search to approximate the constrained optimization problem. Due to the dependence on second-order information, the calculation process is complex and inefficient. In this regard, existing studies reformulate the original problem using dual optimization, improving solution efficiency and boosting policy return by modifying objectives and constraints [9]. In addition, in the AUV three-dimensional obstacle avoidance task, due to the continuous state and strong disturbance characteristics of the underwater environment, unconstrained strategy updates can easily lead to sudden changes in control output. The TRPO constraint mechanism improves control continuity but its high computational cost limits real-time application.

Building on this foundation, PPO reduces computational complexity by replacing the second-order constraint with a first-order optimization process while preserving the stability of policy updates. Specifically, its methods are primarily categorized into KL-penalty-based and clipping-based types. In the KL-penalty approach, a soft constraint is implemented by introducing a KL divergence penalty term into the objective function, which takes the following form:

$$\theta_{k+1} = \arg \max_{\theta} (L(\theta) - \beta_k D_{KL}(\pi_{\theta} \parallel \pi_{\theta_k})) \quad (4)$$

where the penalty coefficient  $\beta$  is used to adjust the constraint strength, so as to balance between exploration and stability.

In addition, PPO-clip performs strategy updates by clipping the probability ratio, and its objective function is as follows:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t [\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)] \quad (5)$$

where  $r_t(\theta)$  represents the probability ratio between the new and old policies, and the clip function restricts it to the interval  $[1 - \epsilon, 1 + \epsilon]$ , thereby limiting the magnitude of a single update step. This mechanism enables stable optimization without explicitly introducing high-order constraints.

At the application level, compared with TRPO, PPO has higher training efficiency and is more suitable for high-frequency interactive scenes in AUV three-dimensional obstacle avoidance. Due to still relying on a large amount of environmental sampling, the training cost is high, and there may still be some strategic fluctuations in the obstacle-intensive areas [9, 10]. To address this, existing methods introduce adaptive learning rates and clipping coefficients to reduce reliance on manual tuning and enhance cross-environment adaptability. By introducing a policy constraint factor  $\tau$ , the upper bound of KL divergence is adjusted to achieve a compromise between the strict constraint of TRPO and the clipping mechanism of PPO, thereby improving the efficiency and stability of policy update [10].

### 3.3. Enhancement approach and performance evaluation

The improvements of TRPO and PPO mainly target two goals: reducing optimization complexity and enhancing policy update efficiency. While maintaining policy stability, optimization proceeds along two directions: simplifying second-order solutions and reconstructing constraint mechanisms. For TRPO, improvements target the computational burden of second-order optimization. By applying dual optimization, the original KL-constrained problem is reformulated, boosting solution efficiency while maintaining the trust region, and enhancing policy update convergence [9]. However, since it still relies on second-order information and the constrained optimization framework, computational complexity remains high, making it difficult to meet the high-frequency update requirements of real-time control systems. For PPO, improvements focus on simplifying and engineering the constraint mechanisms. In the KL-penalty approach, a soft constraint is applied by introducing and dynamically adjusting the penalty coefficient  $\beta$ . In the clipping approach, policy updates are directly limited by truncating the probability ratio, avoiding complex constraint-solving. Further, the introduction of a policy constraint factor  $\tau$  adjusts the effective update interval, improving policy update efficiency while reducing over-clipping [10].

## 4. Conclusion

This paper reviews the application status and challenges of two DRL algorithms, TRPO and PPO, in the obstacle avoidance path optimization of AUV. The results show that the DRL method, with its powerful self-learning and generalization capabilities, has significant advantages over traditional methods in dealing with complex, dynamic and lack of accurate models of underwater environment obstacle avoidance problems, especially in the rapid decision-making of unknown waters. However, this paper has several limitations. First, no experimental simulations were conducted, which prevents quantitative assessment of the performance differences and operational boundaries between

TRPO and PPO in AUV obstacle avoidance, as well as detailed evaluation in real underwater environments. Second, the algorithms require high sample complexity and their exploration mechanisms are limited, which can reduce learning efficiency in sparse-reward scenarios. Third, as model-free methods, the algorithms do not explicitly model the environment, which limits forward-looking planning and makes the learned policies less interpretable. In view of the above challenges, future research should enhance sample efficiency, design adaptive learning to reduce hyperparameter tuning, and integrate offline and online reinforcement learning for safe and effective trajectory planning in unfamiliar underwater environments.

## References

- [1] Sun, Y. (2025). Research on intelligent control methods of AUV based on PPO algorithm. Shandong University. <https://doi.org/10.27272/d.cnki.gshdu.2024.006619>
- [2] Li, J. (2026). Research on AUV path planning and trajectory tracking methods based on SAC algorithm. Shenyang Jianzhu University. <https://doi.org/10.27809/d.cnki.gsjgc.2025.000865>
- [3] Yang, G. (2018). Research on predictive control strategies of autonomous underwater vehicles near the water surface. Harbin Engineering University.
- [4] Wan, Y., Zhu, Z., Zhong, C., et al. (2025). Dynamic path planning of robotic arms based on improved PPO algorithm. *Journal of System Simulation*, 37(6), 1462-1473. <https://doi.org/10.16182/j.issn1004731x.joss.24-0122>
- [5] Kumar, K. A. (2025). Modified giant armadillo optimization-assisted deep reinforcement learning-based controllers for trajectory tracking and obstacle avoidance of autonomous underwater vehicle. *International Journal of Image and Graphics*. <https://doi.org/10.1142/s0219467827500501>
- [6] Sawada, R., Sato, K., & Majima, T. (2020). Automatic ship collision avoidance using deep reinforcement learning with LSTM in continuous action spaces. *Journal of Marine Science and Technology*. <https://doi.org/10.1007/s00773-020-00755-0>
- [7] Chang, M., Wang, Z., Qu, X., & Li, H. (2025). Research on path planning algorithms for unmanned underwater vehicles. *Ship Electrical Technology*, 45(12), 90-95. <https://doi.org/10.13632/j.meee.2025.12.014>
- [8] Li, L., Luo, X., & Song, X. (2026). Mathematical guarantees for trust region policy optimization. *Information Sciences*, 740, 123190. <https://doi.org/10.1016/j.ins.2026.123190>
- [9] Xu, H., Xuan, J., Zhang, G., & Lu, J. (2025). Twin trust region policy optimization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
- [10] Cui, Z. (2025). Collision avoidance decision-making method for multi-surface autonomous vessels based on improved PPO algorithm. Dalian Maritime University. <https://doi.org/10.26989/d.cnki.gdlhu.2024.002445>