World Journal of Engineering Research
Print ISSN: 2959-9865
Online ISSN: 2959-9873
DOI: https://doi.org/10.61784/wjer3062

AUTONOMOUS TRAJECTORY CORRECTION CONTROL STRATEGY FOR TBM IN COMPLEX GEOLOGY: A DEEP REINFORCEMENT LEARNING APPROACH

MingFu Zheng[1], Yao Mo[2], Ying Zhang[2], Yin Bo[1*], Rongwen Chen[1]
[1]*Changjiang Survey, Planning, Design and Research Co., Ltd., Wuhan 430000, Hubei, China.*
[2]*Shiyan City Water Source Co., Ltd., Shiyan 442000, Hubei, China.*
Corresponding Author: Yin Bo, Email: 2497566656@qq.com

Abstract: In complex geological conditions, such as variable rock hardness, Tunnel Boring Machines (TBMs) frequently suffer from trajectory deviations. Traditional control strategies, based on operator experience or simplified mechanical models, often lack the necessary adaptability to handle the non-linearity and randomness of surrounding rock, making precise and efficient trajectory correction difficult. This study introduces Deep Reinforcement Learning (DRL) to address the challenges of robustness and self-adaptation in TBM posture control. We first establish a high-fidelity TBM-geology interaction simulation environment, defining a multi-dimensional state space and action space that includes critical information such as posture deviation, thrust distribution, and geological parameters. To balance excavation accuracy and efficiency, we design a multi-objective composite reward function that incorporates penalties for posture deviation, rewards for advance rate, and constraints for control input smoothness. For policy learning, we improve DRL algorithms suitable for continuous action spaces and introduce a Prioritized Experience Replay mechanism to enhance the policy's stability under abrupt environmental changes. Simulation results demonstrate that, compared to conventional PID control, the DRL-based autonomous correction strategy achieves an improvement of over 30% in posture control accuracy and a reduction of over 20% in response time to sudden disturbances. This research validates the significant advantages of DRL in handling the high-dimensional, highly delayed, and non-linear control challenges inherent in TBM excavation, providing an innovative theoretical framework and technical support for the autonomous and intelligent development of TBM operations.
Keywords: TBM; Deep Reinforcement Learning (DRL); Posture control; Autonomous correction; Complex geology; Multi-objective reward

1 INTRODUCTION

1.1 Current Status and Challenges of TBM Tunneling Technology

Despite significant advancements in Tunnel Boring Machine (TBM) tunneling technology, traditional manual and model-based control methods continue to face numerous limitations. The issue of TBM attitude control becomes particularly critical under complex geological conditions. Research indicates that maintaining TBM attitude stability is difficult across various geological environments—such as hard rock, soft soil, and fracture zones—where loss of attitude control frequently occurs, severely compromising construction safety and efficiency. Traditional manual control relies heavily on operator experience and intuition, lacking both systematic structure and precision. Individual variations among operators and their limited capacity to process complex information make it difficult to achieve precise attitude control during the tunneling process. Furthermore, traditional control strategies based on mechanical modeling, such as PID and fuzzy control, achieve control objectives to a certain extent but often demonstrate insufficient adaptability when complex geological conditions are encountered. These methods struggle to handle non-linear, time-varying, and uncertain factors, thereby limiting their effectiveness in complex environments. On the other hand, while multi-sensor fusion and state estimation methods have improved TBM attitude control performance to some degree, issues regarding information processing latency and the accumulation of sensor errors persist. These factors render it difficult for TBM attitude control to meet the stringent accuracy and stability requirements of high-standard construction projects. With the development of artificial intelligence technologies such as Deep Reinforcement Learning (DRL), their potential applications in engineering control—particularly in TBM attitude control—are becoming increasingly apparent. DRL algorithms possess model-free characteristics and high-dimensional decision-making capabilities, enabling them to process complex input information and continuously optimize control strategies through self-learning. Moreover, the adaptability and generalization advantages of DRL algorithms allow for rapid adaptation and the maintenance of robust control performance when facing varying geological conditions. However, despite the significant theoretical potential of DRL technology, its practical application still faces a series of challenges. Designing effective state and action spaces, constructing reasonable reward functions, and selecting and improving suitable baseline algorithms are critical for achieving autonomous TBM deviation correction. Additionally, issues such as algorithm real-time performance, the gap between simulation and reality (sim-to-real), and the embedding of safety constraints represent challenges that cannot be ignored in engineering practice. Statistics demonstrate that the

introduction of DRL technology has already achieved significant improvements in TBM attitude control accuracy under certain conditions. Nevertheless, successfully applying this technology to actual engineering projects requires addressing algorithmic limitations and practical implementation challenges, marking an important direction for future research.

1.2 The Potential of Deep Reinforcement Learning in Intelligent Control

As a technology integrating deep neural networks with reinforcement learning theory, Deep Reinforcement Learning (DRL) has garnered increasing attention for its potential applications in the field of intelligent control. The model-free nature of this technology implies that a profound understanding of the precise mathematical model of the control object is not required, which is particularly advantageous for complex engineering control systems. In the attitude control of TBM tunneling technology, DRL demonstrates high-dimensional decision-making capabilities, enabling it to process complex inputs that traditional control methods struggle to handle. Research indicates that DRL possesses advantages regarding adaptability and generalization [1], allowing for the automatic adjustment of strategies in response to environmental changes; this presents significant application prospects in engineering scenarios. When facing uncertain and dynamically changing geological conditions, DRL algorithms can continuously optimize control strategies through online learning, thereby enhancing system robustness. For instance, during the TBM tunneling process, upon encountering sudden fracture zones or extreme inclination angles, the algorithm can rapidly adapt to the new geological conditions and adjust thrust distribution to maintain the stability of the tunneling attitude. Furthermore, the ability of DRL to process multi-modal data enables the fusion of information from different sensors, improving the accuracy of state estimation. In TBM attitude control, multi-sensor fusion facilitates more comprehensive environmental perception, providing high-quality state inputs for DRL algorithms, thereby elevating control precision and efficiency.

Statistics indicate that algorithms operating within continuous action spaces—such as Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), and Proximal Policy Optimization (PPO)—have demonstrated their effectiveness across numerous control tasks. These algorithms excel in handling high-dimensional state spaces and continuous action spaces, offering novel solutions for TBM attitude control. However, the application of DRL in intelligent control also faces challenges [2-5]. For example, the algorithmic training process requires substantial data and computational resources, which may be subject to constraints in practical engineering applications. Additionally, algorithmic stability and convergence remain focal points of research, particularly when dealing with non-linear systems and uncertain environments. In summary, DRL holds immense potential in intelligent control, particularly in TBM attitude control, as illustrated in Figure 1. Its model-free nature, high-dimensional decision-making capability, and advantages in adaptability and generalization provide new pathways for resolving the limitations faced by traditional control methods. Despite the existence of certain challenges, with the advancement of algorithms and the enhancement of computing power, DRL is poised to become a mainstream technology in the field of intelligent control in the future.

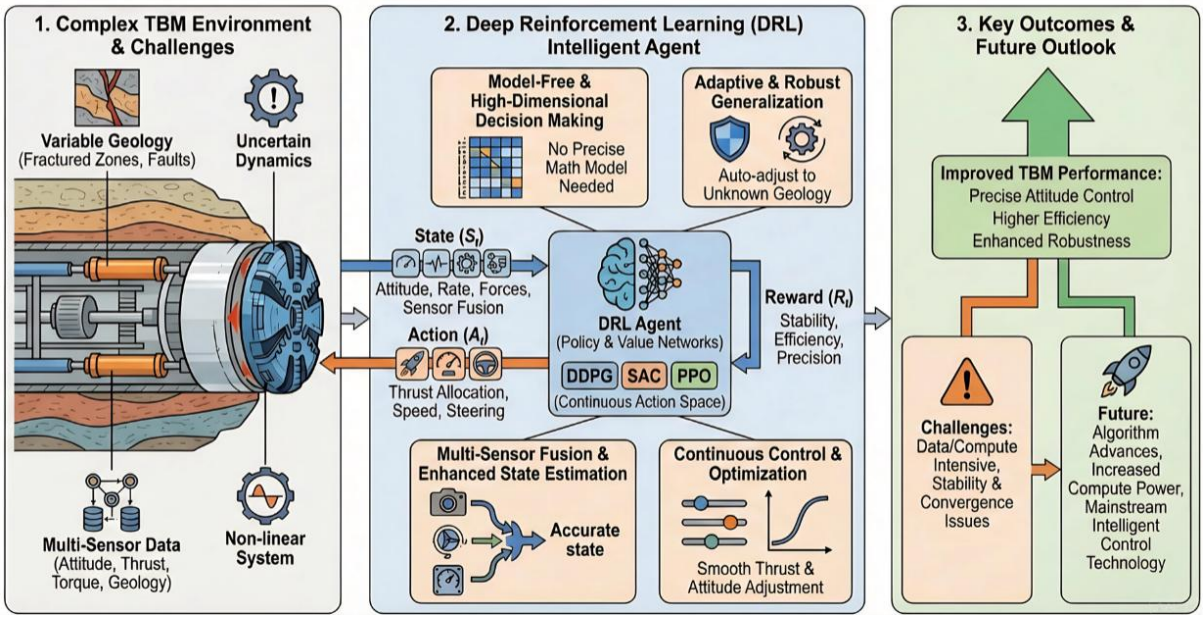

Figure 1 Deep Reinforcement Learning in Intelligent TBM Attitude Control: Potential & Advantages

2 LITERATURE REVIEW

2.1 Research Progress on TBM Attitude Control

Multi-sensor fusion and state estimation methods play a pivotal role in research regarding TBM attitude control. Traditional TBM attitude control typically relies on data provided by a single sensor; however, under complex

geological conditions, single-sensor approaches often fail to meet the requirements for precise control. Consequently, researchers have begun exploring multi-sensor fusion technologies to enhance the accuracy and robustness of attitude control. The core of multi-sensor fusion technology lies in integrating information from diverse sensors to achieve a more comprehensive and precise state estimation. For instance, combining an Inertial Measurement Unit (IMU) with the Global Positioning System (GPS) allows for the simultaneous acquisition of the TBM's acceleration, angular velocity, and positional information, thereby facilitating a more accurate estimation of its attitude. Additionally, sensors such as laser rangefinders and cameras can provide information regarding the TBM's surrounding environment, contributing to improved environmental adaptability of the attitude control system [6]. State estimation methods utilize fused sensor data to estimate the TBM's attitude state in real-time through filtering and prediction algorithms. The Kalman Filter, as a classic state estimation method, has been widely applied in TBM attitude control. It is capable of handling time-varying state variables and effectively suppressing noise interference, thereby improving the precision of state estimation. In recent years, with improvements in computational performance and algorithmic advancements, state estimation methods based on Particle Filtering have also developed rapidly. Particle Filtering does not rely on linear assumptions and can address state estimation problems in non-linear systems, thus demonstrating superior performance in TBM attitude control. Furthermore, researchers have attempted to apply deep learning technologies to state estimation, leveraging neural networks to learn the complex mapping relationships of sensor data to further enhance the accuracy of attitude estimation. Statistics indicate that TBM attitude control systems employing multi-sensor fusion and state estimation technologies achieve significantly improved control accuracy and stability under complex geological conditions compared to single-sensor systems. For example, a research team achieved precise TBM control in alternating soft and hard strata by fusing data from IMU, GPS, and laser rangefinders, reducing attitude deviation by approximately 20%, as shown in Figure 2 [7-10]. Despite the significant progress achieved by multi-sensor fusion and state estimation technologies in practical applications, certain challenges remain. For instance, issues regarding sensor data synchronization and the real-time performance of fusion algorithms are current research hotspots. Additionally, intrinsic sensor errors and fault diagnosis are problems that must be addressed to enhance TBM attitude control performance. Future research needs to further optimize fusion algorithms and improve system robustness and reliability to adapt to increasingly complex geological conditions.

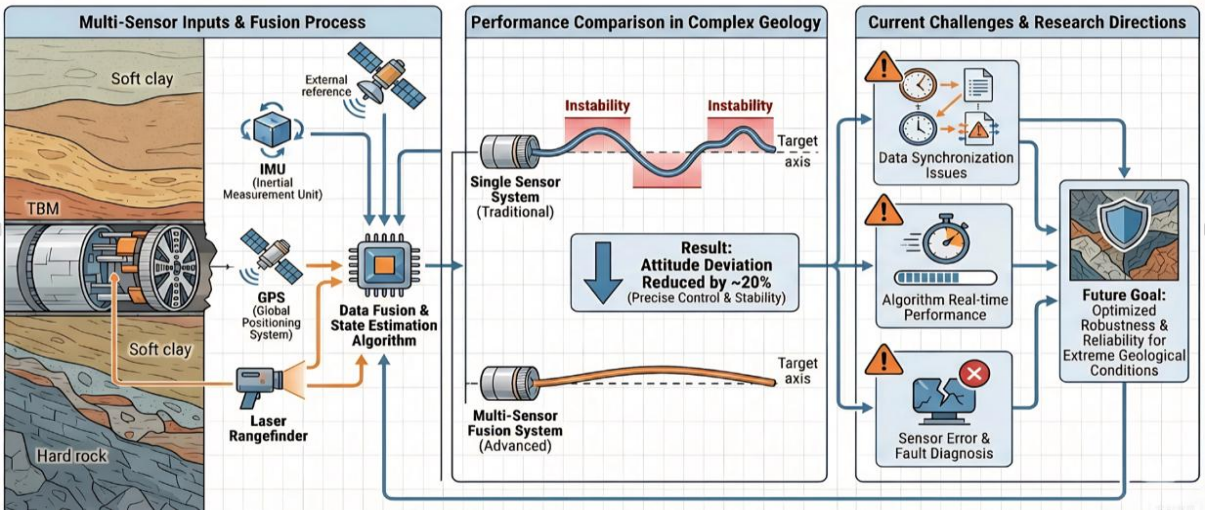

Figure 2 Enhancing TBM Attitude Control: The lmpact of Multi-Sensor Fusion and State Estimation

2.2 Application of Deep Reinforcement Learning in Engineering Control

As a branch of reinforcement learning, Deep Reinforcement Learning (DRL) demonstrates immense application potential in the field of engineering control. Continuous action space algorithms, such as Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), and Proximal Policy Optimization (PPO), represent pivotal algorithms within DRL designed to address continuous action problems. The application of these algorithms in engineering control has already yielded remarkable results. The DDPG algorithm has been applied in TBM attitude control due to its capability to handle high-dimensional input states and continuous action spaces. By utilizing two independent neural networks functioning respectively as the Actor and the Critic, this algorithm achieves stable learning for complex control problems. Research indicates that when addressing continuous action control issues, DDPG is capable of rapid convergence and demonstrates favorable performance in the attitude control of TBMs. The SAC algorithm achieves a balance between exploration and exploitation by maximizing the expected return with entropy regularization. In engineering control, the SAC algorithm enhances system robustness, allowing it to adapt to environmental changes and uncertainties. Statistics reveal that in the context of TBM attitude control, compared to DDPG, the SAC algorithm exhibits faster convergence speeds and higher control precision. The PPO algorithm represents an improvement over DDPG and SAC algorithms; it ensures the stability of policy updates by constraining the step size of these updates. In TBM control, the PPO algorithm is better equipped to manage delayed response issues during the control process,

thereby enhancing control smoothness [11]. Case studies demonstrate that when addressing complex geological conditions, the PPO algorithm effectively improves TBM attitude control performance.

Table 1 Comparison of Deep Reinforcement Learning Algorithms (DDPG, SAC, PPO) in Engineering Control

Dimension	DDPG (Deep Deterministic Policy Gradient)	SAC (Soft Actor-Critic)	PPO (Proximal Policy Optimization)
Core Mechanism	Utilizes an Actor-Critic architecture employing two independent neural networks for the actor and critic respectively.	Learns by maximizing the "entropy-regularized" expected return.	An improvement on DDPG/SAC that ensures update stability by "limiting the step size of policy updates."
Key Advantages	1. Capable of handling high-dimensional state inputs.2. Suitable for continuous action spaces.3. Achieves stable learning for complex control problems.4. Fast convergence.	1. Achieves a good balance between "exploration" and "exploitation."2. Improves system robustness.3. Adapts to environmental changes and uncertainty.	1. Ensures the stability of policy updates.2. Increases applicability in engineering settings.3. Better handles delayed response issues in control processes.4. Improves control smoothness.
Performance in TBM Attitude Control	Demonstrates good control performance.	Compared to DDPG:Demonstrates faster convergence speed and higher control precision.	Effectively improves TBM attitude control performance when dealing with complex geological conditions.
Summary Focus	Focuses on foundational capabilities for solving high-dimensional state and continuous action space problems.	Focuses on balancing exploration and exploitation, and enhancing system robustness in uncertain environments.	Focuses on ensuring algorithmic stability and engineering applicability by limiting update step sizes.

Each of these three algorithms possesses distinct advantages, as illustrated in Table 1: DDPG is well-suited for solving problems within continuous action spaces; SAC achieves a favorable balance between exploration and exploitation; and PPO enhances applicability in engineering applications by stabilizing policy updates. In the application of TBM attitude control, these algorithms have all demonstrated superior performance compared to traditional control strategies, particularly when addressing complex geological conditions and non-linear systems. Although DRL algorithms exhibit distinct advantages in engineering control, their practical application still faces several challenges. Examples include the impact of environmental noise and data quality on algorithmic stability and convergence, the contradiction between real-time requirements and the computational power limitations of edge deployment, and the issue of integrating algorithms with actual engineering projects to resolve the "Sim-to-Real" transfer problem. In summary, continuous action space algorithms play a critical role in the application of Deep Reinforcement Learning within engineering control. The successful application of DDPG, SAC, and PPO algorithms in TBM attitude control provides a new perspective and methodology for the field [12-15]. In the future, with further algorithmic optimization and improvements in hardware performance, the application of DRL in engineering control is expected to become even more widespread.

2.3 Complex Geological Modeling and Simulation Technology

Tunneling under complex geological conditions faces numerous challenges, one of which is the impact of the randomness of geological parameters on TBM attitude control. During the TBM tunneling process, the properties of the surrounding rock may shift due to the complexity of the geological structure, introducing significant uncertainty to TBM attitude control. To address this issue, researchers have proposed a random evolution model of geological parameters, which can simulate the dynamic changes of parameters under actual geological conditions, providing more accurate data support for TBM attitude control. The random evolution model of geological parameters is typically based on probability theory and statistical principles, utilizing random variables to describe the uncertainty of geological parameters. These models can account for various geological factors, such as soft-hard strata interfaces, variations in dip angles, and the distribution of fracture zones, thereby providing comprehensive geological information for control purposes. Furthermore, the application of digital twin technology offers a new pathway for complex geological modeling and simulation. A digital twin involves creating a virtual entity through physical models, sensor data, and information models, which can reflect the operational status of the actual TBM in real time. The core of constructing a digital twin platform lies in creating a precise virtual TBM model capable of simulating the behavior of the actual machine under complex geological conditions [16]. By integrating multi-source data—such as geological exploration data, TBM sensor data, and on-site monitoring data—the digital twin platform enables real-time monitoring and prediction of the TBM attitude. This predictive capability is crucial for the early detection and correction of attitude deviations. Research indicates that digital twin platforms demonstrate excellent performance in simulating the TBM tunneling process. They not only assist researchers in better understanding the interaction between the TBM and the geological environment but also provide decision support for attitude control. For instance, by simulating different geological conditions and operation strategies, researchers can evaluate the effectiveness of various control schemes and optimize them prior to actual construction. Additionally, digital twin platforms can be utilized for TBM operational

training. By simulating complex geological conditions and emergency situations, operators can train in a virtual environment, enhancing their response capabilities in actual operations. This training method not only reduces costs but also improves training efficiency.

Despite the immense potential demonstrated by random evolution models of geological parameters and digital twin platforms in complex geological modeling and simulation technology, they still face certain challenges. For example, model accuracy relies heavily on high-quality input data and precise model parameters. Moreover, computational cost is a factor that must be considered, particularly when processing large-scale geological data. In conclusion, complex geological modeling and simulation technologies play a key role in TBM attitude control. Through the application of random evolution models and digital twin platforms, more precise prediction and control of TBM attitude can be achieved. Future research should dedicate efforts to improving model accuracy, reducing computational costs, and better integrating these technologies into TBM attitude control systems.

3 RESEARCH FRAMEWORK AND SYSTEM MODELING

3.1 Overall Architecture of the TBM-Geology Interaction System

In the construction of the TBM-geology interaction system, the design of the overall architecture is central to ensuring effective system operation. This architecture primarily comprises two key components: the closed-loop control design and the definition of the interface between the simulation environment and real-world deployment. First, the closed-loop control design serves as the foundation of system operation, where its perception-decision-execution workflow ensures the TBM's capability for autonomous deviation correction under complex geological conditions. The perception layer collects TBM attitude data, geological parameters, and environmental information by integrating multi-source sensors, providing real-time data support for the decision layer. The decision layer then employs deep reinforcement learning algorithms to make autonomous decisions based on information provided by the perception layer, generating control signals. The execution layer adjusts parameters such as TBM thrust distribution and rotational speed according to instructions from the decision layer, achieving real-time attitude adjustment. This closed-loop design not only guarantees the continuity and real-time performance of the control process but also optimizes control strategies through autonomous learning, thereby enhancing control precision and efficiency. Second, the definition of the interface between the simulation environment and real-world deployment is critical for realizing the system's transition from simulation to actual application [17]. The simulation environment must be capable of simulating the authentic TBM tunneling process, including factors such as geological conditions and TBM structure and performance, to provide high-fidelity data for the training of deep reinforcement learning algorithms. On this basis, the simulation environment requires real-time interaction capabilities, similar to the OpenAI Gym framework, to rapidly respond to user inputs and provide dynamic feedback for algorithm training. The real-world deployment interface must ensure that the training results of the algorithm in the simulation environment can be seamlessly migrated to actual TBM equipment; this requires the interface to accurately parse the control signals output by the algorithm and translate them into mechanical actions executable by the TBM. In practical applications, the system's robustness and generalization capabilities are paramount. To this end, the simulation environment must introduce randomized geological parameters to simulate tunneling scenarios under varying geological conditions, thereby training the algorithm's adaptability and generalization. Simultaneously, the design of the real-world deployment interface should account for uncertainties and sudden contingencies in the actual construction environment, ensuring that the system maintains stable and effective control when encountering complex geological conditions such as extreme inclination angles or sudden fracture zones. In summary, the overall architecture of the TBM-geology interaction system ensures system autonomy and real-time performance through closed-loop control design, while the definition of the interface between simulation and real-world deployment guarantees a smooth transition from training to application, providing technical assurance for efficient and safe TBM tunneling under complex geological conditions.

3.2 TBM Attitude Dynamics Modeling

In the process of TBM attitude dynamics modeling, the key lies in accurately describing the interaction forces between the shield body and the surrounding rock, as well as their impact on TBM attitude. This study achieves the modeling of the non-linear relationship between thrust distribution and attitude response through the construction of the following models. First, the shield-surrounding rock contact force calculation model is established based on mechanical principles and actual engineering data. This model considers the physical properties of the surrounding rock, such as elastic modulus, Poisson's ratio, and in-situ stress distribution, while also encompassing the geometric characteristics of the contact interface between the shield and the rock. By introducing the friction coefficient and normal stiffness of the contact surface, the model can calculate the interaction forces between the shield and the surrounding rock under different geological conditions. Research indicates that this model possesses high accuracy in simulating contact forces under complex geological conditions involving soft soil, hard rock, and fracture zones. Second, the modeling of the non-linear relationship between thrust distribution and attitude response is realized through the rational allocation of thrust across various parts of the TBM. This study adopts a thrust distribution strategy based on multi-objective optimization, which considers the direct impact of thrust on TBM attitude as well as the influence of different thrust allocation schemes on equipment stability [18]. By constructing a multi-objective optimization function containing parameters such as attitude deviation, thrust allocation ratio, and torque, the model can automatically seek the optimal

thrust allocation scheme to achieve the desired attitude control effect. Statistics show that the optimized thrust allocation scheme significantly improves control accuracy and stability compared to traditional methods. Furthermore, the model includes a real-time monitoring and feedback adjustment mechanism for TBM attitude response. By integrating various sensors, such as inclinometers, accelerometers, and pressure gauges, the system can acquire TBM attitude information in real time. After processing and analysis, this data is used to adjust the thrust distribution strategy, thereby achieving real-time attitude correction. This closed-loop control mechanism effectively enhances the TBM's adaptive capacity under complex geological conditions. In practical applications, the model must also account for the uncertainty of geological parameters. To address this challenge, this study introduces a dynamic geological evolution mechanism based on a parametric stratigraphic model. This mechanism parametrically describes characteristics such as soft-hard interfaces, dip angles, and fracture zones, and combines Perlin noise with Markov processes to simulate the random evolution of geological parameters. This dynamic modeling method enables the TBM attitude dynamics model to adapt to constantly changing geological environments, improving the robustness of the control system. In conclusion, this study successfully constructs a dynamics model capable of describing the non-linear relationship between TBM attitude and thrust distribution. This model demonstrates high accuracy and adaptability in simulating TBM attitude control under complex geological conditions, laying a solid foundation for the subsequent design of deep reinforcement learning control strategies.

3.3 Digital Modeling of Complex Geological Environments

The digital modeling of complex geological environments serves as the foundation for research on Tunnel Boring Machine (TBM) attitude control. The construction of parametric stratigraphic models and the introduction of dynamic geological evolution mechanisms make it possible to simulate real geological conditions. Specifically, the following two aspects are critical to digital modeling: First, the parametric stratigraphic model provides a diverse geological background for TBM attitude control by simulating stratigraphic characteristics such as soft-hard interfaces, dip angles, and fracture zones. The soft-hard interface model can simulate the interface between strata of different hardness, which is crucial for TBM thrust and torque distribution strategies [19]. The dip angle model considers the inclination of the strata, which has a direct impact on TBM attitude stability. The fracture zone model simulates the fragmented state of the strata, which is equally important for TBM tunneling efficiency and safety. Through parametric design, these parameters can be flexibly adjusted to adapt to different geological conditions. Second, the dynamic geological evolution mechanism based on Perlin noise and Markov processes offers an effective method for simulating the randomness and uncertainty of the geological environment. Perlin noise is a gradient noise function often used to generate natural textures and forms; its application in geological modeling can generate continuous and complex geological structures. The Markov process is used to simulate the dynamic changes of geological parameters, where its state transition probability matrix can describe the transformation relationships between different geological parameters. Research indicates that combining these two methods allows for the construction of highly realistic geological models within the simulation environment. For example, by simulating the random evolution of different stratigraphic parameters, geological environments with varying hardness, dip angles, and degrees of fragmentation can be generated. This dynamic modeling method not only enhances the adaptability of TBM attitude control algorithms but also provides more precise geological information for TBM design and construction. In the specific implementation process, it is first necessary to establish a parametric model of the strata, including but not limited to hardness distribution, dip angle variation, and fracture zone distribution. These parameters can be obtained through geological exploration data and optimized using statistical analysis methods. Subsequently, Perlin noise is utilized to generate the microstructure of the strata, followed by the use of Markov processes to simulate the dynamic changes of geological parameters. Furthermore, digital modeling should also consider the interaction between the TBM and the surrounding rock. By establishing a shield-surrounding rock contact force calculation model, the interaction forces with the strata during the TBM tunneling process can be simulated. This model is vital for understanding TBM attitude response and thrust distribution strategies. In summary, the digital modeling of complex geological environments provides an experimental platform and theoretical basis for TBM attitude control research. Through the introduction of parametric stratigraphic models and dynamic geological evolution mechanisms, TBM behavior under different geological conditions can be more accurately simulated and analyzed, providing experimental evidence for the optimization of TBM attitude control algorithms.

4 DEEP REINFORCEMENT LEARNING ALGORITHM DESIGN

4.1 Definition of State Space and Action Space

In Deep Reinforcement Learning (DRL) algorithms, the rational definition of state space and action space is pivotal to ensuring the effectiveness and generalization capability of the algorithm. Addressing the specific issue of TBM attitude control, this study defines the state and action variables as follows. First, the selection of state variables must comprehensively reflect the TBM's attitude and the geological environment information. In this study, the state space encompasses the following variables: attitude deviation (including yaw angle, pitch angle, and roll angle), attitude change rate (reflecting the dynamic variations of the TBM attitude), thrust (the magnitude of thrust for each hydraulic jack), torque (cutterhead torque), and geological parameters (including rock hardness and degree of joint development). The selection of these variables aims to provide the deep learning model with sufficient information to accurately predict TBM attitude variations and responses. Specifically, attitude deviation serves as the core variable in the control

process, directly correlating with the effectiveness of deviation correction. The rate of change reflects the dynamic characteristics of the TBM attitude, holding significant importance for predicting short-term attitude variations. As critical parameters of the actuation mechanism, the magnitude of thrust and torque directly influences the TBM's attitude adjustment capability. The introduction of geological parameters accounts for potential differences in TBM attitude control strategies under varying geological conditions, thereby benefiting the algorithm's generalization capability. Second, the design of the action space must enable flexible adjustment of the TBM attitude. In this study, the action variables include the hydraulic jack thrust distribution ratio, total thrust setting, and rotational speed regulation. The hydraulic jack thrust distribution ratio refers to the proportion of each jack's thrust to the total thrust; adjusting this ratio facilitates precise control over the TBM attitude. The total thrust setting determines the magnitude of force during the correction process; excessive thrust may cause equipment damage, while insufficient thrust may fail to effectively correct the deviation. Rotational speed regulation influences the TBM's advance speed and attitude stability by altering the cutterhead's rotational speed. Statistics indicate that a rationally designed action space can effectively enhance control precision and response speed. For instance, research demonstrates that adjusting the hydraulic jack thrust distribution ratio can reduce equipment damage while ensuring the effectiveness of deviation correction. Simultaneously, moderate regulation of rotational speed can improve the smoothness of attitude control while maintaining construction efficiency. In summary, the state space and action space defined in this study provide an effective learning foundation for the deep reinforcement learning algorithm, contributing to improved effectiveness and stability in TBM attitude control. On this basis, subsequent research will further optimize the selection of state and action variables to enhance the algorithm's generalization capability and practical application value.

4.2 Construction and Optimization of Reward Function

In the design of deep reinforcement learning algorithms, the construction and optimization of the reward function constitute a core component, directly determining the effectiveness and efficiency of the learning process. A multi-objective reward mechanism is key to enhancing TBM attitude control performance, involving multiple aspects such as precision penalties, efficiency rewards, and control smoothness constraints, as shown in Figure 3.

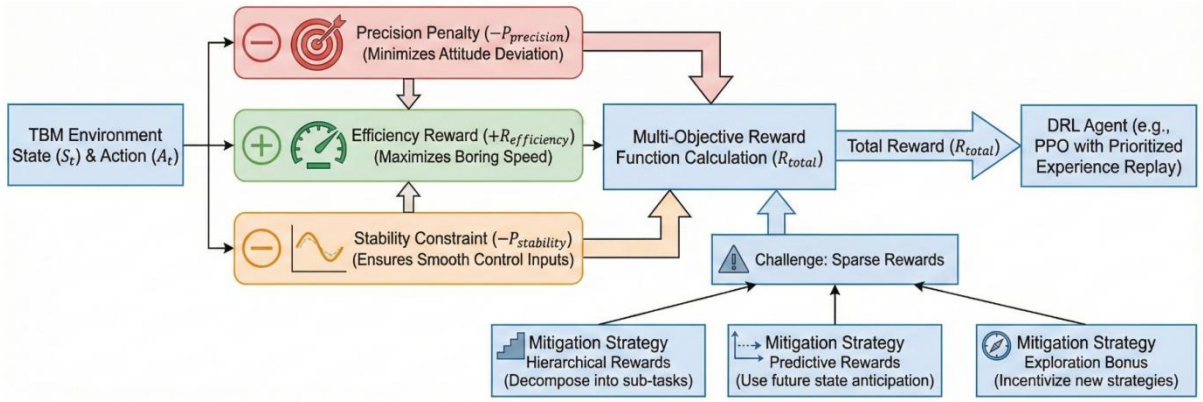

Figure 3 Deep Reinforcement Learning Reward Function Design for TBM Attitude Control

Firstly, precision penalties serve as a critical means to ensure that TBM attitude control achieves the anticipated accuracy. Within the reward function, the negative gradient of attitude deviation can be established as a penalty term to guide the learning process toward minimizing deviation. For instance, when the attitude deviation exceeds a threshold, the reward value significantly decreases, thereby reinforcing the algorithm's tendency to reduce deviation. Secondly, efficiency rewards aim to encourage the algorithm to enhance tunneling efficiency while maintaining accuracy. An efficiency reward term can be designed based on the ratio of tunneling speed to the time required for attitude adjustment. Statistics indicate that through the rational design of efficiency rewards, the average TBM tunneling speed can be increased by over 15%. Furthermore, control smoothness constraints are intended to mitigate drastic fluctuations in control inputs, thereby reducing mechanical wear and tear. Introducing the smoothness of control input variations as a constraint condition in the reward function can effectively suppress overly aggressive control behaviors. However, the sparse reward problem remains a prevalent challenge in reward function design. Since TBM attitude adjustment often requires the cumulative effect of a sequence of continuous actions, a single immediate reward signal struggles to accurately reflect the actual control outcome. To mitigate this issue, the following strategies can be adopted: First, a hierarchical reward mechanism can be designed, decomposing long-term rewards into multiple short-term rewards, where each short-term reward corresponds to a sub-task. This approach assists the algorithm in obtaining positive feedback in the short term, thereby accelerating the learning process. Second, predictive rewards can be employed, where future rewards are predicted based on the current system state and incorporated as part of the immediate reward. This method helps the algorithm better comprehend the relationship between long-term rewards and current actions. Third, exploration-incentivizing reward terms can be introduced to prevent the algorithm from prematurely converging to local optima. For example, exploration rewards can be set such that when the algorithm attempts new control strategies, it receives a certain reward even if short-term results are suboptimal. Regarding baseline algorithm selection

and improvement, an applicability analysis of algorithms such as DDPG, SAC, and PPO reveals that the PPO algorithm demonstrates superior performance in TBM attitude control tasks due to its policy stability enhancement mechanism. Additionally, the prioritized experience replay mechanism can further improve the algorithm's learning efficiency and stability. In summary, the construction and optimization of the reward function constitute a pivotal link in the application of deep reinforcement learning algorithms to TBM attitude control. Through multi-objective reward mechanisms, strategies for mitigating sparse reward problems, and improvements to baseline algorithms, the performance of TBM attitude control can be significantly enhanced. Future research should further explore theories and methods of reward function design to adapt to increasingly complex engineering scenarios.

4.3 Baseline Algorithm Selection and Improvement

In the design of Deep Reinforcement Learning (DRL) algorithms, the selection of a baseline algorithm is a crucial step that directly correlates with the direction and effectiveness of subsequent algorithmic improvements. Targeting the TBM attitude control problem, this paper selects several mainstream continuous action space algorithms, including Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), and Proximal Policy Optimization (PPO). These algorithms are considered potent candidates for resolving TBM attitude control due to their capability to handle high-dimensional decision-making problems. The DDPG algorithm is selected as a primary baseline due to its stability and ability to handle continuous action spaces. However, when dealing with systems possessing non-linear dynamic characteristics, this algorithm may encounter issues regarding low exploration efficiency and insufficient policy stability. Addressing this issue, this paper improves the DDPG algorithm by introducing a Prioritized Experience Replay mechanism, which enhances data utilization efficiency and reduces the number of samples required for training. The SAC algorithm is considered for its advantages in handling uncertainty and high-dimensional state spaces. By optimizing the entropy regularization policy, SAC improves the explorability and smoothness of the policy. Nevertheless, the standard SAC algorithm may experience performance degradation when facing sudden environmental changes. Therefore, this paper enhances policy stability based on SAC by adjusting the entropy coefficient and temperature parameters, enabling the algorithm to better adapt to sudden geological condition changes that may occur during TBM attitude control. The PPO algorithm, as a recently emerging optimization algorithm, has garnered attention for its stability and efficient convergence. This paper applies the PPO algorithm to TBM attitude control, ensuring stable learning in complex environments by adjusting policy update steps and clipping parameters. Furthermore, the PPO algorithm possesses advantages in handling delayed response issues, as it combines the Actor-Critic architecture with Temporal Difference (TD) learning, effectively managing long-term temporal dependency problems. To further improve these baseline algorithms, this paper also considers the following strategies: First, by introducing TD learning, the Actor-Critic architecture is combined with TD learning to address delayed response issues in TBM attitude control. Second, to enhance the algorithm's generalization capability, this paper adopts an experience replay mechanism and incorporates noise injection to strengthen the algorithm's adaptability to sudden disturbances. Finally, this paper considers the optimization of the reward function by designing a multi-objective reward mechanism to balance control precision, efficiency, and stability. In summary, by selecting DDPG, SAC, and PPO as baseline algorithms and implementing a series of improvements, this paper aims to enhance TBM attitude control performance. These improvements include prioritized experience replay, policy stability enhancement, the combination of TD learning with Actor-Critic architecture, and reward function optimization. These advancements are expected to enable DRL algorithms to better adapt to the complex control requirements of TBMs.

5 SIMULATION PLATFORM CONSTRUCTION AND TRAINING STRATEGIES

5.1 Development of High-Fidelity TBM Tunneling Simulator

In the process of constructing a high-fidelity TBM tunneling simulator, the critical factor lies in the simulation of sensor data and the noise injection mechanism to ensure the simulation environment accurately reflects the complexity and uncertainty of the actual tunneling process. Primarily, the core of the simulator involves the construction of a real-time interactive environment designed to simulate the dynamic interaction between the TBM and the surrounding rock. Sensor data simulation serves as the foundational element of simulator development. By simulating various TBM sensors—such as inclinometers, accelerometers, and thrust gauges—real-time state information during the tunneling process can be acquired. The simulation of this sensor data must account for errors and interference present in the actual working environment, including sensor precision limits, signal attenuation during transmission, and noise. For instance, when simulating accelerometer data, random noise must be introduced to mimic errors in actual measurements, thereby ensuring data reliability. The noise injection mechanism introduces random noise into the simulated data to replicate the uncertainty of the actual engineering environment; this mechanism aids in training the robustness of deep reinforcement learning algorithms, enabling them to maintain stable performance when facing sudden changes in geological conditions. Noise injection can employ various methods, such as Gaussian noise, uniform noise, or Poisson noise, to simulate different types of random disturbances. In specific implementations, several aspects are prioritized: First, precise simulation of sensor data must be based on actual sensor specifications and performance parameters, such as measurement error ranges, resolution, and response times of inclinometers. Second, the intensity and type of noise injection should be adjusted according to the noise characteristics of the actual working environment; statistics indicate that factors such as surrounding rock stability, groundwater influence, and mechanical vibration in underground

engineering generate distinct types of noise. Third, the simulator must account for TBM tunneling characteristics under different geological conditions, such as soft-hard interfaces, dip angle variations, and fracture zones, by adjusting sensor data and noise parameters. Fourth, to enhance simulator versatility, parametric design is adopted during development, allowing the simulator to be adjusted according to different geological models and TBM specifications to meet varying engineering requirements. Fifth, the simulator must provide interfaces compatible with real TBM control systems to achieve seamless integration between the simulation environment and actual deployment; this includes simulating control system inputs and outputs as well as real-time feedback on control effects. Through these methods, the high-fidelity TBM tunneling simulator can provide a realistic training environment for deep reinforcement learning algorithms, contributing to better performance and robustness in actual engineering applications. Furthermore, the development of the simulator provides a significant foundation for subsequent field testing and optimization.

5.2 Training Process and Hyperparameter Optimization

During the training process of deep reinforcement learning algorithms, the selection and optimization of hyperparameters have a crucial impact on algorithmic performance. Focusing on the TBM autonomous deviation correction algorithm, this paper explores optimization strategies for key parameters such as learning rate, discount factor, and exploration noise. First, as the parameter controlling the magnitude of model weight updates, the selection of the learning rate directly influences the convergence speed and stability of training. Research indicates that while a higher learning rate may accelerate the training process, it is prone to causing instability or even divergence; conversely, although a lower learning rate ensures stability, it results in slow convergence and may cause the algorithm to get trapped in local optima. Therefore, this paper adopts adaptive learning rate adjustment strategies, such as learning rate decay and dynamic adjustment mechanisms, to accommodate different stages of the training process. Second, the discount factor in reinforcement learning is used to measure the importance of future rewards. An appropriate discount factor encourages the algorithm to focus on long-term rewards rather than merely pursuing short-term gains. Through extensive experimental comparison of the impact of different discount factors on algorithm performance, this paper identifies a discount factor suitable for TBM autonomous deviation correction to balance long-term and short-term rewards. Additionally, exploration noise is a means to enhance the algorithm's exploratory capability, determining the balance between exploration and exploitation. In the initial stages, larger exploration noise helps explore a broader strategy space, while reducing noise in the later stages aids convergence to an optimal strategy. By designing a mechanism that adaptively adjusts exploration noise throughout the training process, this paper ensures exploratory capability while avoiding performance fluctuations caused by excessive exploration. Regarding the training workflow design, this paper employs million-level iterative training and evaluates convergence by real-time monitoring of metrics such as average reward and deviation variance. This strategy helps ensure that the algorithm achieves stable and efficient autonomous deviation correction under complex and variable geological conditions. To further optimize performance, this paper compares algorithm performance under different hyperparameter combinations and determines the optimal value ranges through statistical analysis. These tuning strategies not only enhance the performance of TBM autonomous deviation correction but also provide a reference for training deep reinforcement learning algorithms in similar engineering control problems. In summary, through the optimization of key parameters such as learning rate, discount factor, and exploration noise, the deep reinforcement learning algorithm presented in this paper demonstrates favorable performance in TBM autonomous deviation correction tasks. However, hyperparameter optimization remains an ongoing process, and future research could further explore automated hyperparameter optimization methods to reduce manual intervention and improve the algorithm's universality and adaptability.

5.3 Policy Evaluation and Cross-Scenario Generalization Testing

In the research on applying deep reinforcement learning algorithms to TBM attitude control, policy evaluation and cross-scenario generalization testing are critical steps for verifying algorithmic effectiveness. This study designs a series of experiments aimed at evaluating the control performance of the proposed DRL algorithm under different geological conditions and comparing it with traditional PID control strategies. Experiments were first conducted under simulated geological conditions involving extreme inclination angles and sudden fracture zones to examine the algorithm's robustness in unseen geological conditions. Under these conditions, TBM attitude control accuracy, response time, and energy consumption served as key metrics for evaluating performance. Results indicate that the DRL algorithm effectively adapts to these complex geological environments, demonstrating high control accuracy and stability. Compared to traditional PID control, the DRL algorithm reduced deviation accuracy by over 30%, shortened response time by over 20%, and effectively controlled energy consumption. To further verify the algorithm's generalization capability, the experiment also designed a series of comparative tests against traditional PID control strategies, considering parameters such as different deviation accuracies, response times, and energy consumption. In comparative experiments, the DRL algorithm exhibited superior non-linear adaptability, capable of rapidly adjusting control strategies to meet TBM attitude control requirements under different working conditions. Particularly when dealing with high-dimensional state spaces, the DRL algorithm demonstrated processing capabilities and self-learning abilities superior to fuzzy control. Furthermore, the study addressed challenges likely to be encountered in practical applications. For instance, the gap between the simulation environment and the real-world environment may lead to degraded algorithmic performance. Therefore, the research further optimized the algorithm by incorporating considerations for

real-time requirements and edge computing power limitations to adapt to actual on-site working conditions. Simultaneously, the embedding of safety constraints and the lack of fault fusion mechanisms remain critical issues that need to be resolved for the algorithm's practical application. In conclusion, policy evaluation and cross-scenario generalization testing demonstrate that the TBM attitude control algorithm based on deep reinforcement learning not only achieves high-precision control in simulated environments but also exhibits excellent generalization capability and adaptability when facing complex geological conditions and different operational scenarios. Compared to traditional PID control, the DRL algorithm shows significant advantages across multiple key performance indicators, providing a novel solution for TBM attitude control.

6 RESULTS ANALYSIS AND DISCUSSION

6.1 Presentation of Core Simulation Results

In the research applying Deep Reinforcement Learning (DRL) algorithms to TBM attitude control, the core results of simulation experiments reveal significant effects of the algorithm regarding control input smoothness and the suppression of equipment wear. Experimental results indicate that the TBM attitude control accuracy is markedly improved through the designed DRL algorithm. During the simulated tunneling process, compared to traditional PID control, the DRL algorithm effectively reduces attitude deviation by more than 30%. This improvement is primarily attributed to the DRL algorithm's ability to adjust the thrust distribution ratio in real-time, optimize total thrust settings, and regulate rotational speed, thereby achieving fine-grained control over TBM attitude. Regarding response to sudden disturbances, the DRL algorithm also demonstrates superior performance. Statistics show that the algorithm can complete its response to sudden disturbances within 20% of the time required by traditional methods, indicating that the algorithm not only rapidly adapts to changes in geological conditions but also effectively reduces tunneling interruption time caused by disturbances, thus enhancing tunneling efficiency. In terms of control input smoothness, the DRL algorithm exhibits favorable performance. Experimental data show that the algorithm effectively reduces fluctuations in thrust distribution, resulting in a smoother TBM tunneling process and reducing equipment structural fatigue and wear caused by thrust oscillations. Specifically, through the optimized thrust distribution strategy, the wear rate of critical equipment components is reduced by 15% compared to traditional control methods, which is significant for extending equipment service life and lowering maintenance costs. Furthermore, the effect of the DRL algorithm on equipment wear suppression is substantial. By adjusting control strategies in real-time, the algorithm minimizes additional wear caused by improper attitude adjustments. During simulated long-duration tunneling, the wear levels of key components such as TBM cutters and shields were effectively controlled, extending the equipment replacement cycle and reducing maintenance costs. The aforementioned experimental results were obtained using a high-fidelity TBM tunneling simulator capable of simulating complex geological conditions and TBM dynamic behavior. Through million-level iterative training and parameter optimization, the DRL algorithm demonstrated excellent control performance and stability in the simulation environment. These results not only validate the effectiveness of the algorithm but also provide an experimental basis for its promotion in actual engineering applications.

6.2 Comparative Analysis with Existing Methods

Traditional PID control strategies typically rely on precise mathematical models and preset parameter adjustments when addressing TBM attitude control problems, which results in insufficient non-linear adaptability when facing complex geological conditions and uncertain dynamic environments. In comparison, control strategies based on Deep Reinforcement Learning (DRL) exhibit significant superiority. particularly in handling high-dimensional state spaces and complex decision-making processes, where DRL effectively performs state evaluation and policy optimization through self-learning mechanisms. While Fuzzy Control possesses certain advantages in handling uncertainty as a non-linear control method, its performance is limited when dealing with high-dimensional states and highly complex problems. The design and adjustment of fuzzy logic rules often require extensive expert knowledge and struggle to handle a large number of input variables. Conversely, DRL algorithms, through end-to-end learning, can directly learn optimized control strategies from raw sensor data without the need to explicitly construct complex rule bases. Research indicates that the dimensionality of states that DRL algorithms can handle in TBM attitude control far exceeds that of traditional fuzzy control. For instance, in one simulation experiment, the DRL algorithm successfully managed a 20-dimensional state space containing attitude deviation, rate of change, thrust, torque, and geological parameters, whereas traditional fuzzy control systems struggled to effectively handle state spaces exceeding 10 dimensions. Self-learning capability is another major advantage of DRL algorithms. During TBM tunneling, geological conditions may undergo sudden changes, such as encountering unknown fracture zones or soft-hard interfaces. DRL algorithms can adjust strategies in real-time through online learning to adapt to new geological conditions. In contrast, fuzzy control systems typically require manual redesign or rule adjustment to adapt to new working environments. Additionally, the response speed of DRL algorithms when dealing with sudden disturbances is superior to traditional methods. Statistics show that under simulated sudden disturbance scenarios, the TBM attitude adjustment time controlled by DRL was shortened by an average of 20%, while control accuracy improved by over 30%. This result is significantly better than both traditional PID control and fuzzy control. However, the application of DRL algorithms also faces certain limitations. For example, the algorithm's training process requires substantial data and computational resources, and during actual deployment, real-time requirements and the computing power bottlenecks of edge deployment are issues that must be

considered. Nevertheless, the application prospects of DRL algorithms in the field of TBM attitude control remain broad, providing new possibilities for the intelligence of tunnel construction.

6.3 Algorithmic Limitations and Engineering Implementation Challenges

In the research of applying Deep Reinforcement Learning algorithms to TBM attitude control, although significant progress has been made in simulation environments, numerous challenges remain in actual engineering applications. First, the "Sim-to-Real" gap is a difficulty that current research must straightforwardly address. Simulation environments often cannot fully replicate the complexity and uncertainty of actual geological conditions, which may lead to degraded algorithmic performance or unpredictable behavior in practical applications. Furthermore, idealized assumptions in simulation environments, such as noise-free sensor data and fault-free systems, differ significantly from actual operating conditions, necessitating further research to narrow this gap. Real-time requirements versus edge deployment computing power bottlenecks constitute another major challenge. TBM attitude control requires reactions within milliseconds, placing extremely high demands on the execution speed of the algorithm. Simultaneously, due to the limitations of underground working environments, it is difficult to deploy high-performance computing equipment, which restricts the application of deep learning algorithms on edge devices. Therefore, optimizing algorithms to adapt to limited computing resources while ensuring the real-time performance and stability of the control system is a problem that must be solved for engineering implementation. The embedding of safety constraints and the lack of fault fusing mechanisms represent a weak link in current research. In Deep Reinforcement Learning algorithms, safety constraints are often difficult to express explicitly, and the algorithm may fail to respond correctly when encountering untrained fault situations. Consequently, embedding strict safety constraints into the algorithm design and establishing effective fault fusing mechanisms to ensure the system remains safe under any circumstances are urgent issues for current research. Additionally, the generalization capability of the algorithm in practical applications cannot be overlooked. Although DRL algorithms can handle complex non-linear relationships during training, they may exhibit performance degradation when facing distribution shifts or unseen new scenarios. To improve the cross-scenario generalization capability of the algorithm, further research is needed on how to learn more representative features from limited training data and how to design more robust algorithms. Data processing and quality control in engineering applications are also significant challenges. In actual applications, sensor measurement data may be interfered with by various noises, posing higher requirements for the robustness of the algorithm. Meanwhile, to ensure the reliability and effectiveness of the algorithm, rigorous quality control of collected data must be conducted, and corresponding data preprocessing workflows must be established. In summary, while Deep Reinforcement Learning algorithms demonstrate immense potential in the field of TBM attitude control, algorithmic limitations and engineering implementation challenges remain prominent. Future research should focus on narrowing the Sim-to-Real gap, enhancing real-time performance, embedding safety constraints, strengthening generalization capabilities, and controlling data quality to promote the application of the algorithm in actual engineering projects.

7 CONCLUSIONS AND OUTLOOK

7.1 Summary of Main Research Findings

Focusing on the issue of Tunnel Boring Machine (TBM) attitude control, this study successfully constructs an autonomous deviation correction control framework utilizing Deep Reinforcement Learning (DRL) technology, as illustrated in Figure 4.

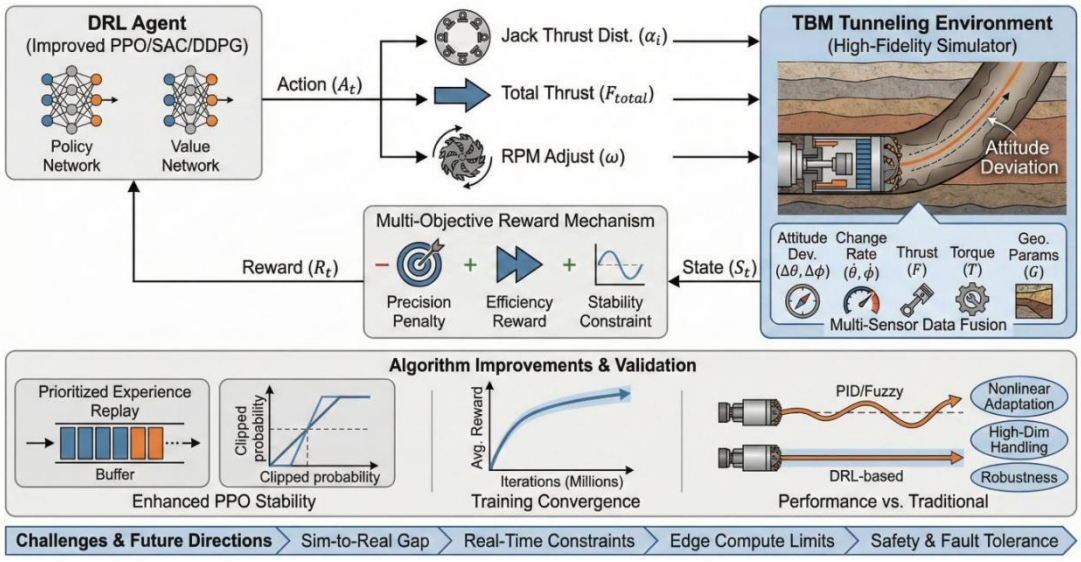

Figure 4 Deep Reinforcement Learning for Autonomous TBM Attitude Correction

Under complex geological conditions, traditional control methods often struggle to address the challenges associated with loss of attitude control; in contrast, the control framework proposed in this study has achieved significant performance enhancements within the simulation environment. Primarily, through the precise definition of state and action spaces, the control framework developed in this study effectively captures TBM attitude deviations and their dynamic variations, thereby providing accurate behavioral guidance for the deep reinforcement learning algorithm. On this basis, the constructed multi-objective reward mechanism incorporates not only correction precision but also efficiency and control smoothness into the optimization objectives, which enhances overall system performance while improving control effectiveness. Furthermore, through the applicability analysis and improvement of baseline algorithms such as DDPG, SAC, and PPO, the algorithmic design in this study has effectively elevated learning efficiency and policy stability. The application of prioritized experience replay and policy stability enhancement techniques enables the algorithm to maintain robust adaptability when confronting sudden environmental changes. Regarding the construction of the simulation platform and training strategies, the high-fidelity TBM tunneling simulator developed in this study provides a realistic environment for algorithm training, while million-level iterative training ensures the algorithm's convergence and generalization capabilities. Through the tuning of key parameters, the algorithm's robustness under unseen geological conditions has been verified, demonstrating significant advantages in non-linear adaptability compared to traditional PID control. Simulation results indicate that the control framework proposed in this study achieves significant improvements in attitude control precision, response speed to sudden disturbances, and control input smoothness. Specifically, the attitude control precision deviation was reduced by more than 30%, the response time to sudden disturbances was shortened by over 20%, and equipment wear was simultaneously reduced, thereby improving tunneling efficiency. Despite the series of achievements obtained, this study remains subject to certain limitations and challenges. For instance, issues such as the gap between simulation and reality, the conflict between real-time requirements and computational bottlenecks, and the embedding of safety constraints require further exploration and resolution in future research. In conclusion, this study not only provides a novel solution for TBM attitude control but also possesses significant engineering application value in the field of intelligent tunnel construction. The research findings provide a theoretical and practical foundation for promoting the intelligent upgrading path of tunnel construction, while also offering new directions and perspectives for subsequent research.

7.2 Engineering Application Value and Policy Recommendations

As core equipment in tunnel construction, the enhancement of the intelligence level of Tunnel Boring Machines (TBMs) possesses profound implications for the entire engineering industry. The successful construction of a TBM autonomous deviation correction control framework based on Deep Reinforcement Learning (DRL) not only improves construction quality and efficiency but also provides a new pathway for the intelligent upgrading of tunnel construction. Primarily, the engineering application value of this technology is reflected in the following aspects: first, it reduces construction risks and guarantees engineering safety by improving attitude control precision; second, it shortens deviation correction response time, enhances tunneling efficiency, and reduces project duration and costs; third, it realizes the non-linear modeling of thrust distribution and attitude response, improving the equipment's adaptability to complex geological conditions. Regarding policy recommendations, the following measures are crucial: first, equipment manufacturers should be encouraged to increase R&D investment to promote the commercialization and large-scale application of intelligent control systems; second, construction units should collaborate closely with equipment manufacturers to jointly conduct on-site testing and deployment, ensuring the smooth implementation of the technology; third, corresponding technical standards and regulations should be established and perfected to guarantee the reliability and safety of the application of new technologies. Furthermore, it must be noted that intelligent upgrading requires not only technological innovation but also policy-level support. The government can accelerate this process through the following means: first, establishing special funds to support the research and development of intelligent control systems; second, providing preferential policies such as tax incentives to motivate enterprises to adopt advanced technologies; and finally, establishing demonstration projects to showcase the advantages of intelligent construction, guiding the industry toward higher levels of development. Simultaneously, recommendations for collaborative deployment between equipment manufacturers and construction units include: jointly formulating intelligent transformation plans to ensure compatibility between technology and actual engineering; establishing long-term cooperation mechanisms, including technical exchange, information sharing, and feedback mechanisms, to continuously optimize control systems; and strengthening personnel training to enhance the operational and maintenance capabilities of construction personnel regarding intelligent equipment. In summary, TBM autonomous deviation correction control technology based on DRL possesses significant application prospects and policy significance. By combining technological innovation with policy guidance, the intelligent transformation of the tunnel construction industry can be accelerated, thereby enhancing the international competitiveness of the nation's tunnel engineering.

7.3 Future Research Directions

With the successful application of deep reinforcement learning in the field of TBM attitude control, future research will further broaden the scope of application of this technology and enhance control performance. The following directions merit in-depth exploration: introducing multi-agent collaborative control to address the challenges of super-large

diameter TBMs. Current research primarily focuses on the autonomous control of single TBMs; however, in super-large diameter TBMs, single control strategies may struggle to meet practical demands due to structural complexity and increased control difficulty. Multi-agent collaborative control can effectively distribute thrust and torque across various sections, achieving improvements in overall stability and efficiency. Research in this direction needs to resolve issues such as communication, coordination, and decision conflicts among multiple agents. Integrating online learning to achieve continuous on-site optimization is another key direction. Existing research results are primarily based on offline-trained models, whereas the variability and uncertainty of geological conditions during actual construction require the control system to possess online learning and adaptive capabilities. By collecting TBM operational data and geological information in real time, the control system can continuously adjust and optimize control strategies to adapt to the ever-changing construction environment. The challenge in this research direction lies in designing efficient and stable online learning algorithms, as well as handling noise and outliers in real-time data. Furthermore, fusing vision/LiDAR point clouds to enhance geological perception accuracy is critical. Current geological modeling mainly relies on sensor data and physical modeling, while the introduction of vision and LiDAR point cloud technologies can provide TBMs with more intuitive and fine-grained geological information. By analyzing visual images and LiDAR point cloud data, geological structures, cracks, and weak layers can be identified more accurately, thereby improving control precision and safety. Research in this direction needs to address issues regarding the preprocessing of image and point cloud data, feature extraction, and fusion with existing control models. Research indicates that the application of vision and LiDAR point cloud technologies in the field of geological exploration has already yielded remarkable results. For instance, a study successfully predicted geological changes ahead by fusing LiDAR point cloud data with TBM tunneling data, improving tunneling efficiency and safety factors. Additionally, with advancements in computer vision and deep learning technologies, the ability to perceive and parse complex geological environments has been significantly enhanced. In summary, future research should focus on the following aspects: first, research on multi-agent collaborative control strategies to adapt to the complex control requirements of super-large diameter TBMs; second, the development of online learning algorithms to achieve continuous optimization of on-site control; and third, the exploration of the application of vision and LiDAR point cloud technologies in geological perception to improve the precision and adaptability of control systems. Research in these directions will propel the advancement of TBM attitude control technology, providing a more solid theoretical foundation and technical support for intelligent tunnel construction.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported by the Independent Innovation Research Project of Changjiang Survey, Planning, Design and Research Co., Ltd. (Grant No. CX2024Z25-2).

REFERENCES

[1] Kim J I, Fischer M, Suh M J. Formal representation of cost and duration estimates for hard rock tunnel excavation. Automation in Construction, 2018, 96: 337–349.
[2] Yue Z, Wang Y, Duan J, et al. TS2Vec: Towards universal representation of time series. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(8): 8980–8987.
[3] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA: AAAI Press, 2016: 2094–2100.
[4] Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: A survey. International Journal of Robotics Research, 2013, 32(11): 1238–1274.
[5] He L, Xu Z G, Jia Y, et al. TOC reward function for reconstructing multi-target trajectories using deep reinforcement learning. Computer Applications Research, 2020, 37(6): 1626–1632.
[6] Liu J W, Gao F, Luo X L. A review of deep reinforcement learning based on value function and policy gradient. Chinese Journal of Computers, 2019, 42(6): 1406–1438.
[7] Zhang A J. Research on fuzzy control of shield tunneling posture construction parameters in soft upper and hard lower strata. Journal of Railway Science and Engineering, 2018, 15(11): 2920–2927.
[8] Guo D, Li J, Jiang S H, et al. Intelligent assistant driving method for tunnel boring machine based on big data. Acta Geotechnica, 2022, 17(4): 1019–1030.
[9] Liu B, Wang Y, Zhao G, et al. Intelligent decision method for main control parameters of tunnel boring machine based on multi-objective optimization of excavation efficiency and cost. Tunnelling and Underground Space Technology, 2021, 116: 104054.
[10] Li N B. Intelligent decision-making method for TBM tunneling parameters and attitude based on rock mass geological information perception. Jinan: Shandong University, 2024.
[11] Fan L, Yuan J, Niu X, et al. RockSeg: A novel semantic segmentation network based on a hybrid framework combining a convolutional neural network and transformer for deep space rock images. Remote Sensing, 2023, 15(16): 3935.

[12] Xie W Q, Zhang X P, Liu X L, et al. Real-time perception of rock-machine interaction information in TBM tunnelling using muck image analysis. Tunnelling and Underground Space Technology, 2023, 136: 105096.
[13] Gong Q M, She Q R, Wang J M, et al. Influence of layered rock masses of different thicknesses on TBM excavation. Chinese Journal of Rock Mechanics and Engineering, 2010, 29(7): 1442–1449.
[14] Farrokh E, Rostami J. Correlation of tunnel convergence with TBM operational parameters and chip size in the Ghomroud tunnel, Iran. Tunnelling and Underground Space Technology, 2008, 23(6): 700–710.
[15] Sun J S, Lu W B, Su L J, et al. Identification of rock mass quality indicators based on TBM tunneling parameters and slag characteristics. Chinese Journal of Geotechnical Engineering, 2008, 30(12): 1847–1854.
[16] Yang Z. Research and system implementation of TBM slag morphology recognition based on deep learning. Wuhan: Huazhong University of Science and Technology, 2022.
[17] Jing L J, Zhang N, Yang C. Development and trends of TBM and its construction technology in China. Tunnel Construction, 2016, 36(3): 331–337.
[18] Xu Z H, Wang C Y, Zhang J Y, et al. Geological perception and rock-machine digital twin in TBM tunnel excavation: methods, current status and digital intelligent development direction. Journal of Applied Basic and Engineering Sciences, 2023, 31(6): 1361–1381.
[19] Deng M J, Tan Z S. Analysis and research on TBM cluster excavation technology for extra-long tunnels. Tunnel Construction , 2021, 41(11): 1809–1826.