Mechatronics 118 (2026) 103536

Contents lists available at ScienceDirect

Mechatronics

journal homepage: www.elsevier.com/locate/mechatronics

A reinforcement learning-based path-planning method for cable-driven hyper-redundant robots in unknown environments☆

Zhenpu Zhu, Zhanxuan Peng, Yu Rong, Guoying Gu *

State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
Shanghai Key Laboratory of Intelligent Robotics, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Keywords:
Hyper-redundant robots
Path planning
Soft robot applications
Deep reinforcement learning
Experience replay algorithms

ABSTRACT

Cable-driven hyper-redundant robots (CDHRRs) show great potential for confined-space detection due to their dexterity and adaptability. However, the redundant degrees of freedom (DoFs) present significant challenges for CDHRRs in planning feasible paths with stable configurations. To address this challenge, this paper presents a path-planning algorithm to improve the stability and applicability of generated paths. Firstly, the Soft Actor–Critic (SAC) algorithm is used to efficiently plan paths for CDHRR. To address the issue of unstable configurations, a path smoothness index and a reward function are designed to evaluate and enhance the feasibility of generated paths. Then, the Hindsight Experience Replay (HER) and Prioritized Experience Replay (PER) algorithms are applied to solve the sparse-reward problem of CDHRRs with corresponding scenarios. Finally, the proposed method is validated via both simulation and real-world experiments. Simulation results show that, compared to the Rapidly-exploring Random Tree Star (RRT*) and Arc-segment RRT (As-RRT), the computation time of the presented algorithm is reduced by 96.56% and 97.95%, respectively. Moreover, this deep reinforcement learning (DRL)-based algorithm generates smoother paths with a more stable configuration. In real-world experiments, the proposed algorithm demonstrates a 14.9% and 10% improvement in path smoothness and success rate over As-RRT.

1. Introduction

Cable-driven hyper-redundant robots (CDHRRs) have gained increasing attention because of their high flexibility and slim body size [1–3]. These advantages contribute to their extensive applications in diverse fields, such as minimally invasive surgery [4–6], aerospace [7,8], and nuclear industries [9–11]. In these safety-critical scenarios, the robot is often required to navigate cluttered and partially observed workspaces while maintaining mechanically feasible and stable configurations. Therefore, path-planning algorithms are crucial for improving task efficiency and success rates [12].

In this context, recent reviews on CDHRRs have summarized state-of-the-art (SOTA) modeling and control methods. These reviews specifically highlight the existing gap between geometric feasibility and mechanically stable execution in such confined, safety-critical workspaces [13–15]. As for CDHRRs' path-planning algorithms, they are typically implemented in either the configuration space (C-space) or the workspace (W-space). In C-space, algorithms like A* and Rapidly Exploring Random Tree (RRT) are widely applied [16,17]. However, the A* algorithm requires high computational complexity in continuous spaces and dynamic environments [18]. Besides, RRT suffers from suboptimal non-smooth paths and inefficient exploration–exploitation balance [19,20]. More importantly, such search-based planners do not explicitly encode configuration stability for large-scale CDHRRs. Consequently, the resulting paths may be difficult to execute safely under strong antagonistic forces. Inverse kinematics-based W-space algorithms utilize Follow-The-Leader (FTL) algorithm to solve the multi-solution problem [21]. Nevertheless, these algorithms typically lead to suboptimal paths and unstable configurations [22]. To address these problems, Specialized Rapidly-exploring Random Tree (Sp-RRT) [23] and Arc-segment (As)-RRT [24] are proposed. However, Sp-RRT requires complete information, which is unrealistic for most cases. Meanwhile, As-RRT also faces the issue of relying on manual teleoperation with visual guidance and an unsatisfactory task success rate [24]. Moreover, human intervention is indispensable since CDHRRs' instability frequently results in low success rates. Overall, conventional model-based planners are often inadequate for large-scale CDHRR path planning in unknown environments. They struggle to jointly handle high-dimensional continuous control, partial observability, and configuration stability constraints.

☆ This paper was recommended for publication by Associate Editor Oliver Sawodny.

* Corresponding author.
E-mail address: guguoying@sjtu.edu.cn (G. Gu).

https://doi.org/10.1016/j.mechatronics.2026.103536
Received 12 December 2025; Received in revised form 27 April 2026; Accepted 28 April 2026
Available online 5 May 2026

These observations reveal a key gap in large-scale CDHRR planning for unknown, confined environment. The planner should not only find collision-free solutions, but also explicitly evaluate and discourage mechanically unstable configurations, while preserving path smoothness and kinematic feasibility. Meanwhile, practical tasks such as obstacle avoidance and target grasping often provide sparse and delayed feedback, which calls for a learning paradigm with improved exploration and sample efficiency.

To address these limitations, Deep Reinforcement Learning (DRL) algorithms have been explored in path-planning. Proximal Policy Optimization (PPO) [25,26] introduces a trust region, avoiding performance instability caused by excessive policy updates. However, it has limitations such as slow strategy updates and unsatisfactory convergence speed in continuous action spaces. Deep Deterministic Policy Gradient (DDPG) [27,28] addresses the limitations by employing an actor–critic architecture to handle high-dimensional continuous action spaces efficiently. Nonetheless, it is sensitive to hyperparameters and prone to local optima due to random noise. Soft Actor–Critic (SAC) [29] incorporates a maximum entropy framework to balance exploration and exploitation with a temperature parameter. This feature allows SAC to effectively plan a path and enhance stability [30]. Additionally, SAC achieves higher average rewards in robotic path-planning tasks, which is crucial for efficiently finding optimal paths [31]. However, directly applying SAC to large-scale CDHRRs is insufficient, because it does not explicitly account for the robot-specific challenges that dominate feasibility and safety in unknown environments. Furthermore, some algorithms fail to adequately account for the unique characteristics of continuum robots [32–34]. This omission not only hampers training efficiency but also hinders practical applications such as multi-target and FTL tracking.

Moreover, to avoid instability, many controllers are only tested in small-scale environments [27]. They are unable to reflect real-world scenarios where antagonistic forces and convoluted exploration paths are common [35,36]. Since robots have no prior knowledge of the unknown environment, the configuration during exploration requires careful consideration. Specifically, for large-scale continuum robots, such as the SJTU-Snake III with 24 redundant degrees of freedom (DoFs), instability configuration is highly likely to be induced. If the robot explores the environment in a haphazard manner, an unstable W-shaped configuration is highly likely to occur, owing to antagonistic forces [37]. This concern is consistent with prior force-closure analysis of cable-driven open chains, which shows that equilibrium feasibility depends on the existence of positive cable tensions under a given configuration and cable-routing condition [38]. In CDHRRs, cable-hole friction, stick–slip transitions, and loading history can further affect the evolution of driving forces during trajectory tracking [39]. Therefore, abrupt or highly non-uniform curvature variations may indicate potentially unfavorable tension distributions or mechanically less stable configurations.

Furthermore, this instability is not merely a performance issue. It can lead to poor repeatability, abrupt motion, and increased risk of mechanical overload during interaction. SAC also faces challenges in addressing sparse-reward tasks, which are prevalent in path-planning applications [40]. From a broader robotics perspective, robot DRL surveys have identified a mainstream SOTA direction for sample-efficient learning in continuous control. This direction corresponds to off-policy actor–critic methods with replay-based data reuse [41]. When it comes to CDHRR path planning, the problem becomes more intricate due to unstable configurations. Consequently, the key open challenge is to develop a learning-based planner for large-scale CDHRRs in unknown environments. Such a planner needs to improve exploration efficiency under sparse rewards, while explicitly steering the policy toward smooth and stable configurations.

To tackle these difficulties, we propose a DRL-based path-planning method for CDHRRs to enhance their applicability and stability. Firstly, the SAC algorithm is adapted for CDHRRs. To explicitly address unstable configurations, we design an instability-aware reward shaping strategy and introduce a path smoothness index to evaluate the feasibility and configuration stability of the generated paths. Furthermore, the HER and PER algorithms are respectively applied in corresponding scenarios to handle the sparse-reward problem. Experimental results highlight the method's computational efficiency and the feasibility of the planned paths. Additionally, the effectiveness of our approach is validated through obstacle-avoidance and target-grasping tasks.

In summary, the main novelties and contributions of this work can be summarized as follows:

(1) We propose an instability-aware SAC-based path-planning framework for CDHRRs. Specifically, an instability-aware reward shaping strategy and a path smoothness index are introduced to explicitly evaluate and penalize mechanically unstable configurations, thereby improving path feasibility and execution stability in unknown environments.
(2) We develop a learning strategy to enhance sample efficiency under sparse-reward conditions. By integrating HER and PER algorithm into the SAC framework for different task scenarios, the proposed method improves exploration efficiency and accelerates convergence in multi-obstacle and multi-target planning tasks.
(3) We validate the proposed framework through both simulation and real-world experiments on large-scale CDHRRs. The results demonstrate improved path smoothness, higher task success rates, and reduced reliance on human intervention, confirming the effectiveness and robustness of the proposed method in practical applications.

The rest of this article is organized as follows. The CDHRR applied in this work is introduced in Section 2. The path-planning method of CDHRR is presented in Section 3, including the network structure, reward function design, and experience replay algorithm. In Section 4, the accuracy and effectiveness of the proposed methods are verified in both simulated and experimental platforms. Finally, conclusions and suggestions for future work are discussed in Section 5.

2. System description

As depicted in Fig. 1(a), the applied hyper-redundant robot, SJTU-Snake III, comprises a motioning platform, a linear-feeding platform, a drive box, a cable-guiding mechanism, and a robotic manipulator. The linear-feeding platform is positioned atop the bottom motion platform with the drive box mounted above it. The cable-guiding mechanism connects the drive box to the robotic manipulator by twelve 2-DoF parallel platforms. Each platform is linked in series via gimbal joints and controlled by three driving cables. The maximum bending angle between adjacent sections is 50.7°. The other detailed geometric parameters are listed in Table 1. The manipulator's end is equipped with a gripper. As shown in Fig. 1(b), the gripper features a binocular camera, and its clamping part is specifically designed to enhance gripping performance.

Table 1
Geometric parameters of SJTU-Snake III.

Parameters	Value
Outer diameter	55 mm
Total arm length	2130 mm
Aspect ratio	38.73
Total mass of arm	0.91 kg
End load	1.5 kg

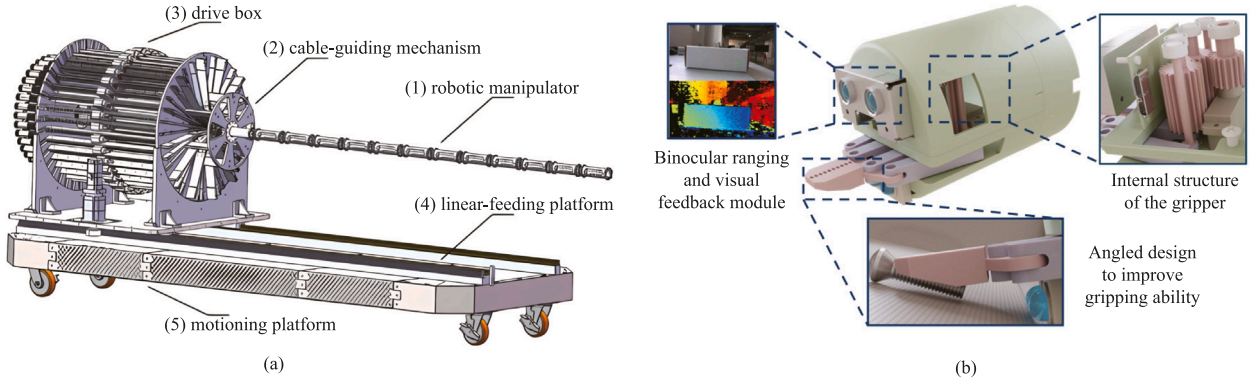

Fig. 1. Structure and end-effector of the CDHRR. The gripper is equipped with a binocular ranging and visual feedback module.

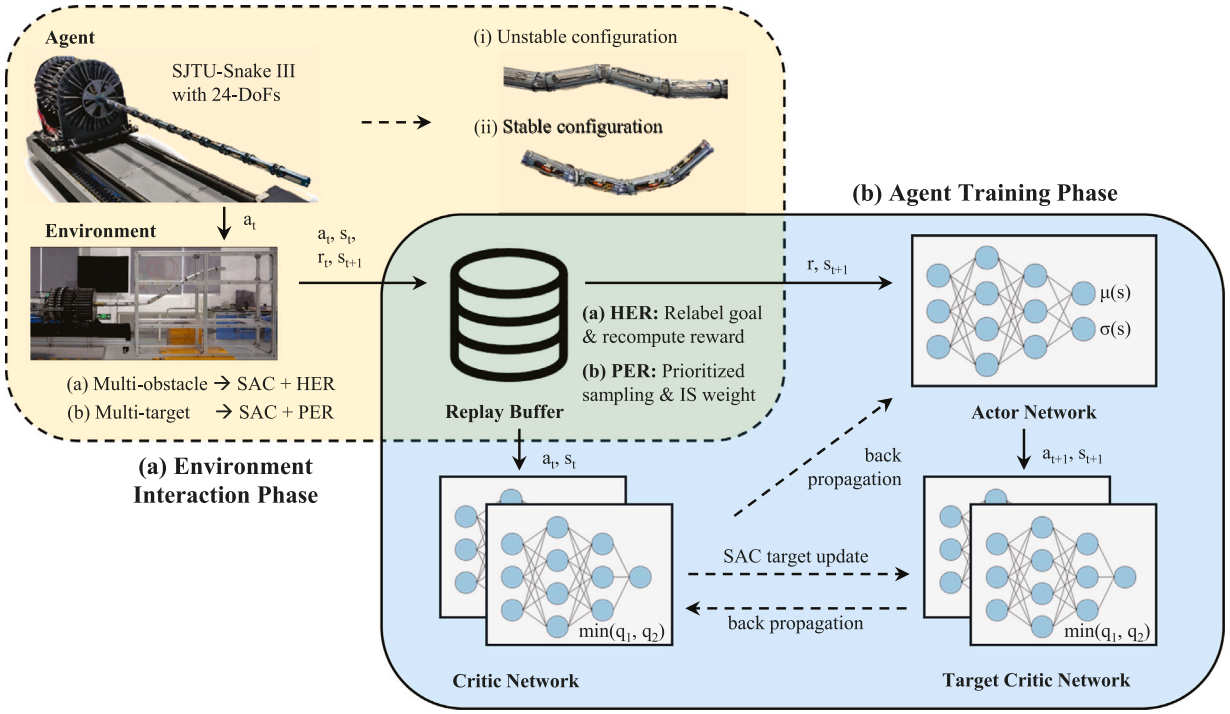

Fig. 2. Diagram of the DRL process of the CDHRR and the neural-network training pipeline. (a) During environment interaction, large-scale CDHRRs are prone to instability. Random exploration may yield antagonistic-force-induced W-shaped collapse, degrading control accuracy, shrinking the reachable workspace, and risking irreversible damage. (b) Scenario-dependent agent training phase based on SAC. The actor network outputs a stochastic action distribution, and two critics estimate Q-values for SAC updates. In multi-obstacle environments, HER is incorporated with SAC at the replay-buffer level by relabeling goals and recomputing rewards. In multi-target environments, PER is incorporated with SAC at the sampling/update level by drawing mini-batches according to priorities and applying importance-sampling weights to the critic loss. Note that HER and PER are enabled in different scenarios, while the target-network soft update remains a SAC-specific step.

3. Path-planning method

This section presents the instability-aware DRL-based path-planning framework for large-scale CDHRRs. The framework is developed based on SAC. It aims to achieve obstacle avoidance and multi-target approaching tasks. It also guarantees smooth and mechanically feasible motions. The solution includes three key components. The first is a task-oriented DRL formulation for path planning. The second is an instability-aware reward shaping strategy with a path smoothness index. The third is scenario-dependent experience replay to improve learning efficiency under sparse rewards. This section first introduces the DRL formulation and SAC-based framework. It then explains the reward design and path smoothness index. Finally, the experience replay strategy and its scenario-specific use are described.

3.1. Definition and framework of the DRL algorithm

In this study, the agent is a CDHRR designed for target-approaching and obstacle-avoidance tasks. It is equipped with sensors for environmental perception and actuators for movement execution. The environment refers to the operation space, comprising both target locations and obstacles. The agent navigates within a randomly generated, completely unknown environment. The FTL algorithm enables each segment of the CDHRR to dynamically align with the motion of the end-effector, thereby streamlining the computational representation of the system. This simplification enhances simulation efficiency and supports more fluid iterative workflows in robotic modeling. The target point is labeled only when it appears within the agent's field of view. The desired end-effector pose is defined as control goal $g \in \mathbb{R}^6$, comprising three translations and three rotations. The goal is reached by following a sequence of states for $t = 0, \ldots, T$, where each step t corresponds to a quasi-static state and T denotes the total number of states. The control problem of the targeted robotic system is then formulated as a Markov Decision Process (MDP), comprising the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ depicted in Fig. 2:

- $\mathcal{S}$ is the *state space*, where each state is explicitly defined as a vector $s_t = [x_t, y_t, z_t]^T \in \mathcal{S}$. Here, x_t, y_t, and z_t denote the 3D Cartesian coordinates of the CDHRR end-effector along the X, Y, and Z axes in the base frame, respectively. The dimensionality of the state vector is fixed as $|s_t| = 3$. This end-effector-centric representation is sufficient for characterizing the robotic system state due to the FTL control strategy. Once the end-effector position is known, the joint angles of the entire robot can be uniquely determined via the Denavit-Hartenberg (DH) method, thus avoiding redundant state variables related to individual joints. To maintain consistency with this end-effector-based formulation, the goal g is treated as a conditioning variable rather than being concatenated into the state input.
- $\mathcal{A}$ is the *action space*, where action $a_t \in \mathcal{A}$ is generated by the DRL agent according to the current state s_t and its own policy. In simulation, $\mathcal{A}$ represents the robot's next movement direction. In the real-world system, it corresponds to the motor drive quantities obtained through inverse kinematics calculation.
- $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ is the *transition probability* from state s_t to s_{t+1} when action a_t is taken.
- r is the *reward function* that assigns a scalar reward value to each transition (s_t, a_t, s_{t+1}, g). We specifically design a CDHRR-based reward function to perform corresponding tasks.
- $\gamma \in [0, 1)$ is the *discount factor*, discounting future rewards over time. It is set to 0.99 in this work.

Neural network architecture is visually depicted in Fig. 2. The actor network receives a 3-dimensional state vector as input, which is succeeded by two hidden layers and an output layer. Here, $\mu(s)$ and $\sigma(s)$ represent the mean and the logarithm of the action's standard deviation. Each hidden layer contains 512 nodes and uses ReLU activation. The critic network takes in state and action vectors, outputting estimated Q-values q_1 and q_2. The network is composed of four fully connected layers. The two hidden layers each consist of 512 nodes and use ReLU activation functions. The specific values of the parameters will be introduced in Section 4. The entire dataset collected during training is used to update the network parameters, validate the policy performance, and test the generalization ability of the trained SAC agent.

SAC adopts a maximum entropy framework, balancing exploration and exploitation via a temperature parameter to effectively plan paths and enhance the operational stability of CDHRRs. To tackle complex environment navigation challenges, a SAC-based DRL framework is developed for efficient CDHRR path-planning. Based on the environment type, the training method is selected accordingly. For multi-obstacle environments with sparse rewards, SAC combined with HER is adopted. For multi-target environments, SAC paired with PER is employed. These two mechanisms are enabled in different scenarios rather than being used simultaneously. The critic and actor losses are computed as

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{B}}\left[\left(\hat{y} - Q_\theta(s,a)\right)^2\right], \tag{1}$$

$$\hat{y} = r + \gamma \cdot \mathbb{E}_{a'\sim\pi_\phi(s')}\left[Q_{\theta^-}(s',a')\right], \tag{2}$$

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_{s\sim\mathcal{B},a\sim\pi_\phi(s)}\left[Q_\theta(s,a) - \alpha \cdot \log \pi_\phi(a|s)\right], \tag{3}$$

where $\mathcal{B}$ is the experience replay buffer, θ is the critic network parameters, $Q_\theta(s,a)$ is the current Q-value predicted by the critic network, $\hat{y}$ is the target Q-value, r is the immediate reward, γ is the discount factor, π_ϕ is the actor network, ϕ is the actor network parameters, θ^- is the target critic network parameters, $\mathbb{E}_{a'\sim\pi_\phi(s')}$ is the expectation over actions sampled from the policy in the next state, α is the temperature parameter, and $\log \pi_\phi(a|s)$ is the logarithm probability of the policy distribution.

3.2. Design of reward function

To enable safe multi-target tracking and obstacle avoidance in unknown environments, the reward function must encourage target-approaching and penalize unstable configurations. The reward function R_{total} includes target-approaching reward R_{target}, arc-segment rewards R_{local} and R_{global}, and collision penalty R_{collide}.

$$R_{\text{total}} = R_{\text{target}} + R_{\text{local}} + R_{\text{global}} + R_{\text{collide}}. \tag{4}$$

The target-approaching reward R_{target} is concerned with whether the agent detects and approaches targets. A substantial positive reward R_s is granted when the Euclidean norm of the scaled error $\|\tilde{e}\|_2$ drops below a predefined threshold Θ, signifying the successful achievement of the goal. Conversely, if this condition is not met, a penalty proportional to $\|\tilde{e}\|_2$ and scaled by a factor ϵ is imposed. Additionally, R_d plays a dual role in penalizing redundant attempts and enhancing the efficiency of the exploration process.

$$R_{\text{target}} = \begin{cases} R_s, & \text{if goal reached.} \\ -\epsilon\|\tilde{e}\|_2, & \text{if goal detected but unreached.} \\ -R_d, & \text{if goal undetected.} \end{cases} \tag{5}$$

Local consistency reward encourages the current action to be similar to the previous action, promoting stability in the agent's behavior. Let $\mathbf{a}$ denote the current action and $\mathbf{a}'$ denote the previous action. The local consistency reward is calculated as

$$R_{\text{local}} = -\frac{1}{3}\sum_{i=1}^{3} |a_i - a_i'|. \tag{6}$$

Global smoothness reward incentivizes the agent to maintain a smooth trajectory by penalizing abrupt changes in the path's curvature. Let $\mathbf{v}_1, \mathbf{v}_2$ be the vectors between consecutive points. The reward is calculated based on the change of curvature between consecutive segments as

$$R_{\text{global}} = -\left(1 - \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \cdot \|\mathbf{v}_2\| + 0.001}\right). \tag{7}$$

Given that the environment is unknown and obstacle-laden, a negative reward R_{collide} is developed. A penalty R_c is imposed when collisions occur, thereby discouraging unsafe actions as

$$R_{\text{collide}} = -R_c, \text{ if agent collides.} \tag{8}$$

The proposed reward function is tailored to the unique physical characteristics of CDHRRs, enhancing the system's operational stability while effectively balancing exploration and exploitation to improve path quality. Notably, mechanical stability is not modeled explicitly via analytical formulations. Instead, it is implicitly encoded into the learn-

ing process through carefully designed reward components. Specifically, a local consistency penalty discourages abrupt action changes that would induce sudden cable tension fluctuations, while a global smoothness term penalizes sharp path curvature variations that lead to unstable configurations such as antagonistic-force-induced W-shaped structures. This comprehensive reward structure enables the agent to learn optimal paths through environmental interaction and adapt its strategy dynamically via reward signals, providing an indirect yet effective approximation of mechanical stability without requiring a full analytical system model, and thus demonstrating strong suitability for real-world applications.

3.3. Experience replay algorithm for sparse-reward condition

In this work, the agent is under a sparse reward condition, receiving rewards only upon reaching the target point. This makes it hard for the agent to gain effective feedback from the environment, thereby leading to slow or failed learning. For diverse environments with multiple obstacles and targets, corresponding experience replay functions need to be introduced. In multi-obstacle environments, SAC-HER is adopted to mitigate sparse rewards by augmenting the replay buffer with goal-relabeled transitions. In the multi-target environment, SAC-PER is adopted to improve sample efficiency by prioritizing transitions with large TD errors.

HER transforms unsuccessful episodes into informative training data by relabeling the desired goal with a hindsight goal and recomputing the corresponding reward. For a collected trajectory

$$\tau = \{(s_h, a_h, s_{h+1})\}_{h=0}^{H-1}, \tag{9}$$

the achieved goal is defined as

$$g_h^{\text{ach}} = f(s_{h+1}), \tag{10}$$

where $f(\cdot)$ extracts the task-relevant goal representation (in this work, the end-effector 3D position). For each step h, the original transition $(s_h, a_h, r_h, s_{h+1}, g)$ is stored into the replay buffer $\mathcal{B}$ with

$$r_h = \text{compute\_reward}\big(f(s_{h+1}),\, g\big). \tag{11}$$

With a hindsight replay ratio η, an additional relabeled transition is further generated using the future strategy. Specifically, a future index is sampled as

$$k \sim \mathcal{U}\{1, \ldots, H-h\}, \tag{12}$$

and the hindsight goal is set as

$$g_h^{\text{her}} \leftarrow f(s_{h+k}). \tag{13}$$

The relabeled reward is recomputed by

$$r_h^{\text{her}} = \text{compute\_reward}\big(f(s_{h+1}),\, g_h^{\text{her}}\big), \tag{14}$$

and the additional HER transition $(s_h, a_h, r_h^{\text{her}}, s_{h+1}, g_h^{\text{her}})$ is stored into $\mathcal{B}$. This procedure increases the number of goal-consistent transitions and accelerates learning in sparse-reward multi-obstacle navigation.

In the multi-target environment, PER is adopted to improve sample efficiency by prioritizing transitions with large Temporal-Difference (TD) errors. Each transition i in $\mathcal{B}$ is associated with a positive priority scalar $p_i > 0$. Transitions are sampled according to the probability

$$P(i) = \frac{p_i^\alpha}{\sum_j p_j^\alpha}, \tag{15}$$

where α is the prioritization exponent controlling how strongly prioritization is used. To correct the sampling bias introduced by prioritized sampling, the importance-sampling (IS) weight is computed as

$$w_i = \left(\frac{1}{|\mathcal{B}|\, P(i)}\right)^\beta, \tag{16}$$

where β is the IS exponent, and $|\mathcal{B}|$ denotes the current number of transitions in the buffer. The IS weights can be normalized within each mini-batch for numerical stability.

Moreover, priorities are updated using TD errors computed from the SAC target. For a sampled transition (s_i, a_i, r_i, s_i'), the next action $a_i' \sim \pi_\phi(\cdot|s_i')$ is sampled and the entropy-regularized SAC target is computed as

$$\hat{y}_i = r_i + \gamma\Big(Q_{\theta^-}(s_i', a_i') - \alpha_{\text{SAC}} \log \pi_\phi(a_i'|s_i')\Big), \tag{17}$$

where Q_{θ^-} denotes the target critic network, and α_{SAC} is the SAC temperature parameter. The TD error is then computed as

$$\delta_i = \hat{y}_i - Q_\theta(s_i, a_i), \tag{18}$$

and the priority scalar is updated by

$$p_i \leftarrow |\delta_i| + \epsilon, \tag{19}$$

where $\epsilon > 0$ avoids zero priorities.

In summary, for the multi-obstacle navigation setting with sparse rewards, HER augments the replay buffer by adding goal-relabeled transitions with recomputed rewards to enrich goal-consistent experience, and mini-batches for SAC updates are sampled uniformly from $\mathcal{B}$. For the multi-target setting, PER modifies the sampling distribution and applies IS weights to boost sample efficiency by prioritizing transitions with large TD errors, while the actor update follows the standard SAC objective. In both scenarios, SAC acts as a common off-policy actor–critic backbone, and HER or PER provides the corresponding scenario-specific replay mechanism. Algorithms 1–3 shows the scenario-dependent training procedure.

Algorithm 1 Scenario-dependent training framework

Input: Scenario type EnvType.
1: **if** EnvType = multi-obstacle **then**
2: Run Algorithm 2 (SAC-HER).
3: **end if**
4: **if** EnvType = multi-target **then**
5: Run Algorithm 3 (SAC-PER).
6: **end if**

Algorithm 2 SAC-HER for multi-obstacle environments

Input: Batch size B, training steps T, trajectory length H, HER ratio η, target update rate τ.
1: Initialize replay buffer $\mathcal{B}$ with capacity N.
2: Initialize actor π_ϕ, critics $Q_{\theta_1}, Q_{\theta_2}$, and target critics.
3: **for** $t = 1$ to T **do**
4: Collect trajectory $\tau = \{(s_h, a_h, s_{h+1})\}_{h=0}^{H-1}$ using $a_h \sim \pi_\phi(\cdot|s_h)$.
5: **for** $h = 0$ to $H-1$ **do**
6: Compute original reward $r_h = r(s_h, a_h, s_{h+1}, g)$; store $(s_h, a_h, r_h, s_{h+1}, g)$ into $\mathcal{B}$.
7: With probability η, apply **HER**:
8: Set hindsight goal $g_h^{\text{her}} \leftarrow f(s_{h+1})$ (achieved goal).
9: Recompute $r_h^{\text{her}} = r(s_h, a_h, s_{h+1}, g_h^{\text{her}})$.
10: Store additional transition $(s_h, a_h, r_h^{\text{her}}, s_{h+1}, g_h^{\text{her}})$ into $\mathcal{B}$.
11: **end for**
12: **if** $|\mathcal{B}| \geq B$ **then**
13: Sample a mini-batch of B transitions **uniformly** from $\mathcal{B}$.
14: Update critics and actor using standard SAC losses; update target networks with τ.
15: **end if**
16: **end for**
17: **return** Trained π_ϕ and $Q_{\theta_1}, Q_{\theta_2}$.

Algorithm 3 SAC-PER for multi-target environments

Input: Batch size B, training steps T, trajectory length H, PER exponents α, β, ϵ, target update rate τ.

Initialize prioritized replay buffer $\mathcal{B}$ with capacity N; assign $p_i = 1$ for newly added samples.
Initialize actor π_ϕ, critics $Q_{\theta_1}, Q_{\theta_2}$, and target critics.
for $t = 1$ to T **do**
  Collect transitions and store (s, a, r, s', g) into $\mathcal{B}$ with initial priority $p = 1$.
  if $|\mathcal{B}| \geq B$ **then**
    Compute sampling probabilities $P(i) = p_i^\alpha / \sum_j p_j^\alpha$.
    Sample a mini-batch $\mathcal{I}$ of B indices according to $P(i)$.
    Compute IS weights $w_i = (|\mathcal{B}|P(i))^{-\beta}$ and normalize within the mini-batch.
    Compute SAC TD errors δ_i for $i \in \mathcal{I}$.
    Update priorities: $p_i \leftarrow |\delta_i| + \epsilon$.
    Update critics using weighted loss $\mathcal{L}_{critic} = \left(\sum_{i \in \mathcal{I}} w_i \delta_i^2\right) / B$; update actor with standard SAC loss.
    Update target networks with τ.
  end if
end for
return Trained π_ϕ and $Q_{\theta_1}, Q_{\theta_2}$.

4. Simulation and experimental validation

This section validates the performance of the proposed SAC-based planning framework and its instability-aware design. The evaluation follows the core objectives introduced earlier. These objectives include efficient exploration in unknown environments, smooth and feasible path generation, and better task success under sparse rewards. Tests are carried out in both simulation and real-world conditions. In simulation, path length, computation time, the path smoothness index, and exploration ability are measured. These metrics reflect planning efficiency and path feasibility in complex scenes. In real-world experiments, the path smoothness index and task success rate are emphasized. These indicators directly affect the safety and repeatability of large-scale CDHRR operations. This section first presents the experimental setup and different scenarios. It then shows the results obtained from simulation tests. Finally, the results from real-world experiments are provided and analyzed.

4.1. Experimental setup

For simulations, the CDHRR is tested in fuel tanks. As shown on the left side of Fig. 3(a), the robot enters a multi-obstacle environment (1800 mm × 1400 mm × 400 mm) to execute detection and grasping tasks. The interior consists of six interconnected, distinct chambers. These chambers are connected by oval holes with a 500 mm major axis and a 300 mm minor axis. Partitions and outer frame-like obstacles are arranged for structural support. The robot is placed at a random initial position to detect and grasp debris targets. Additionally, in simulation, the end-effector position, namely the state $s_t = [x_t, y_t, z_t]^T$, is directly retrieved from the simulator's internal state recorder. For the real-world CDHRR system, s_t is computed in real-time by combining joint angle encoder data with the DH method, ensuring accurate end-effector localization without direct Cartesian coordinate measurement.

In practical CDHRR tasks such as inspection and retrieval inside enclosed tanks, the robot generally only has coarse prior data about the environment. This data includes overall dimensions, compartment structure, and passage locations of the tank. The exact distribution of internal clutter and the positions of targets remain unknown. To replicate this practical condition, the coarse geometric characteristics of the tank are fixed. These characteristics include the tank's dimensions, chamber connectivity, and hole layout. Fine-level elements are randomized for each episode. Specifically, three elements are randomly sampled for each episode: the robot's initial position, the obstacle placements, and the positions of debris targets. This randomized setting models the uncertainty of internal conditions under a known coarse structure.

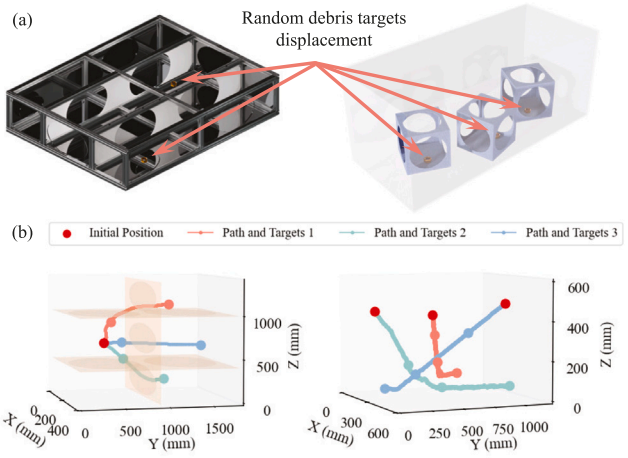

Fig. 3. (a) CDHRR's experimental fuel tank platforms, corresponding to a multi-obstacle environment and a multi-target environment, respectively. The tank geometry provides coarse prior structure, while obstacle placements, entrance positions, and debris target locations are randomized across episodes to model unknown internal conditions. (b) Simulation results of CDHRR end-effector paths and targets in a multi-obstacle environment and a multi-target environment. The red dot marks the initial position of the end-effector. The orange, black, and green dots represent intermediate target points, with the corresponding lines denoting the planned trajectories. For quantitative comparisons, all methods are evaluated using the same instantiated episodes. These episodes apply identical entrance configurations, obstacle arrangements, target positions, and constraints to ensure consistent evaluation.

The right side of Fig. 3(a) shows a multi-target environment with a rectangular oil tank (1200 mm × 600 mm × 600 mm). This environment contains cubic regions with a 300 mm side length and 280 mm diameter holes. Debris targets for retrieval are placed inside these cubic regions. Target positions are randomized across episodes. All target placements meet collision-free constraints.

A fair and interpretable comparison between methods is ensured by evaluating the proposed method and all baseline approaches on the same set of instantiated scenarios. These baseline approaches include existing methods related to CDHRR. For each reported experiment, the random seed is fixed. Identical episode instances are used for all methods as well. These instances include the same initial position, the same obstacle configuration, and the same target placement. All methods follow the same constraints and termination criteria. These criteria cover workspace boundaries, collision checking, and the same maximum step or time budget. The success or failure of each method is judged by the same rules.

Simulations were implemented in Python using NumPy, PyTorch, OpenAI Gym, and Matplotlib. Neural networks were trained on a 2.1 GHz 16-core i7-13700 CPU. An NVIDIA RTX A2000 GPU was also used for the training process.

The implementation details and hyperparameter settings of the method are listed as follows. The agent is trained for 1000 episodes, with a maximum horizon of $H = 1000$ steps per episode. The SAC updates use a mini-batch size of $B = 128$. Policy evaluation is conducted every 100 episodes under the same instantiated episode set for all compared methods. The replay buffer capacity is $N = 10^6$. SAC uses

Table 2
Comparison between RRT*, Sp-RRT, As-RRT, SAC-HER, and SAC-PER generated path in corresponding scenarios.

Scenario	Algorithm	Path length (mm)	Computation time (s)	Path smooth index	Exploration
Multi-obstacle	RRT*	1525.47 ± 1.89	35.48 ± 0.32	0.0647 ± 0.0011	Limited
	Sp-RRT	1821.65 ± 2.15	0.300 ± 0.01	0.4393 ± 0.0024	Limited
	As-RRT	1764.89 ± 6.45	59.45 ± 1.28	0.4721 ± 0.0091	Effective
	SAC-HER	1693.23 ± 5.87	1.220 ± 0.05	0.5259 ± 0.0088	Effective
Multi-target	RRT*	947.935 ± 1.24	18.79 ± 0.24	0.0222 ± 0.0007	Limited
	Sp-RRT	1111.36 ± 1.83	0.174 ± 0.01	0.0351 ± 0.0011	Limited
	As-RRT	1003.56 ± 4.62	46.37 ± 0.97	0.1228 ± 0.0049	Effective
	SAC-PER	997.208 ± 4.08	0.723 ± 0.04	0.1231 ± 0.0045	Effective

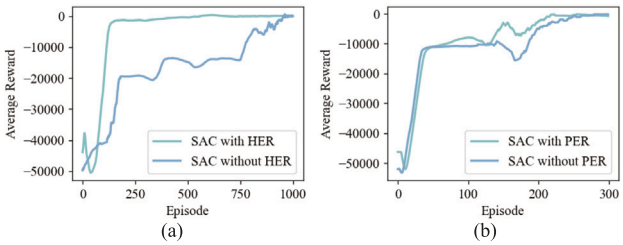

Fig. 4. Ablation study for SAC algorithm with and without (a) HER in multi-obstacle environments (b) PER in multi-obstacle environments. The SAC algorithm, incorporating experience replay, exhibits faster convergence compared to its counterpart without experience replay.

the discount factor $\gamma = 0.99$, the soft target update rate $\tau = 10^{-3}$, and Adam optimizer with learning rate 2×10^{-4} for actor, critics, and temperature optimization. The temperature is automatically tuned with target entropy $-\dim(\mathcal{A})$. During training, an ϵ-greedy action replacement is applied with $\epsilon_0 = 1.0$, $\epsilon_{\min} = 0.01$, and decay factor 0.999 per step. For the sparse-reward multi-obstacle setting, HER is enabled with the hindsight replay ratio $\eta = 0.8$. For the multi-target setting, PER is adopted with a standard configuration $\alpha = 0.6$, β linearly annealed from 0.4 to 1.0, and $\epsilon = 10^{-6}$ for priority stabilization.

Regarding hyperparameter trends, increasing η or α generally accelerates early learning by providing more informative replay samples but may increase replay bias if set excessively high. Moreover, a larger β improves bias correction at the cost of higher variance in gradient updates. Similarly, larger B typically reduces update variance, but increases compute per update, and larger τ adapts targets faster but can introduce training oscillations. The chosen replay configurations are supported by the ablation results in Fig. 4, where enabling HER/PER consistently improves convergence over the corresponding SAC baselines.

4.2. Simulation results and discussion

In the simulation experiment part, the designed reward function can guide the agent to complete complex tasks. As depicted in Fig. 3(b), the agent demonstrates the ability to execute a range of tasks successfully. The agent navigates around obstacles and moves toward debris targets, and identifies and approaches multiple debris targets. This figure shows that the generated paths exhibit favorable curvature characteristics, making them suitable for CDHRR operations.

To validate the effectiveness of the introduced experience replay algorithm, ablation experiments are conducted. As depicted in Fig. 4, in multi-obstacle scenarios, the SAC algorithm augmented with HER demonstrates a convergence speed that is 2.88 times faster than SAC in the absence of HER. Similarly, within multi-target settings, SAC equipped with PER exhibits a convergence rate 1.27 times quicker than SAC without PER. SAC-PER yields no significant performance gain over the baseline SAC, which stems from the intrinsic mechanism of PER. While PER improves sample efficiency by prioritizing transitions with high TD errors, it fails to explicitly address the sparse-reward structure and multi-solution nature inherent to multi-target path-planning tasks. In contrast, HER directly alleviates reward sparsity by relabeling failed trajectories as successful experiences for alternative goals, and performs exceptionally well in multi-obstacle environments where feasible paths are difficult to discover. Accordingly, SAC-HER achieves far more substantial performance improvements than SAC-PER. Furthermore, for large-scale CDHRRs with high-dimensional continuous action spaces and stability constraints, the exploration bottleneck lies more in feasible trajectory discovery than sample reuse alone. Thus, although PER benefits training stability, its impact on overall planning performance remains limited in this context.

To evaluate the suitability of the generated path for CDHRR, we proposed a path smoothness index as

$$S = \left(1 - \frac{\kappa_{\max}}{\kappa_{\mathrm{UL}}}\right) \times \exp\left(-\left(\frac{\kappa_{\max} - \kappa_{\mathrm{avg}}}{\kappa_{\mathrm{avg}}}\right)^2\right). \tag{20}$$

The exponential term and squared term in the formula are designed to penalize deviations in curvature. The squared term $\left[\left(\kappa_{\max} - \kappa_{\mathrm{avg}}\right) / \kappa_{\mathrm{avg}}\right]^2$ magnifies the difference between the maximum curvature $\kappa_{\max}$ and the average curvature κ_{avg}, while the exponential function $\exp(-x)$ further intensifies this penalty, especially for large deviations. This ensures that only when the curvature distribution is uniform, the maximum curvature does not exceed the safety upper limit κ_{UL}, and the maximum curvature is close to the average curvature, the path will be considered smooth and suitable.

When the path satisfies the condition that the maximum curvature $\kappa_{\max}$ does not exceed the upper limit κ_{UL}, and the maximum curvature and average curvature κ_{avg} are relatively close, the index will be high. Such a path is considered suitable for CDHRR and can obtain a large balance value. Conversely, if the maximum curvature is too large or the maximum curvature and average curvature differ greatly, the balance will decrease, reflecting that the path is not smooth.

Although the proposed smoothness index is not a direct physical measure of mechanical stability, it serves as a practical proxy. In CDHRRs, excessive curvature and uneven curvature distribution are often associated with unstable configurations, increased antagonistic forces, and poor repeatability. By penalizing large curvature and non-uniform curvature distribution, the smoothness index indirectly reflects the feasibility and stability of the robot configuration. Nevertheless, it remains a qualitative metric rather than a rigorous stability criterion.

Table 2 presents a comparison of our proposed algorithm and other algorithms in terms of path length, computation time, and path smoothness. The compared methods cover representative SOTA baselines for CDHRR planning, including sampling-based planners (RRT*) and CDHRR-oriented planners (Sp-RRT and As-RRT). It reports the quantitative comparison under the multi-obstacle navigation setting. To

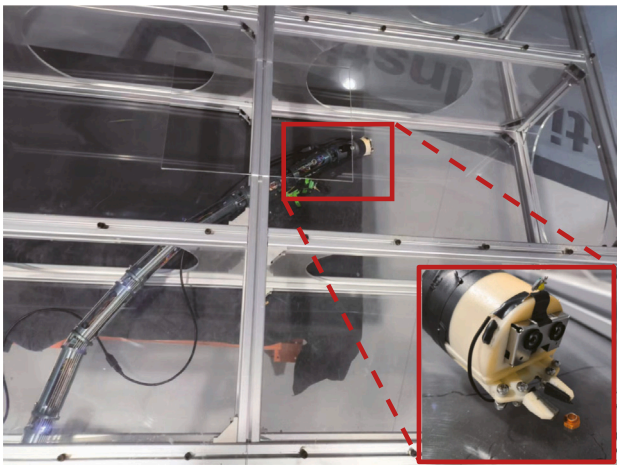

Fig. 5. Real-world experiments with SJTU-Snake III in multi-obstacle environments using the SAC-HER algorithm. These experiments involve path generation and target tracking tasks. The robot moves through chambers to carry out detection tasks. The generated path is smooth, enabling the robot to perform tasks effectively.

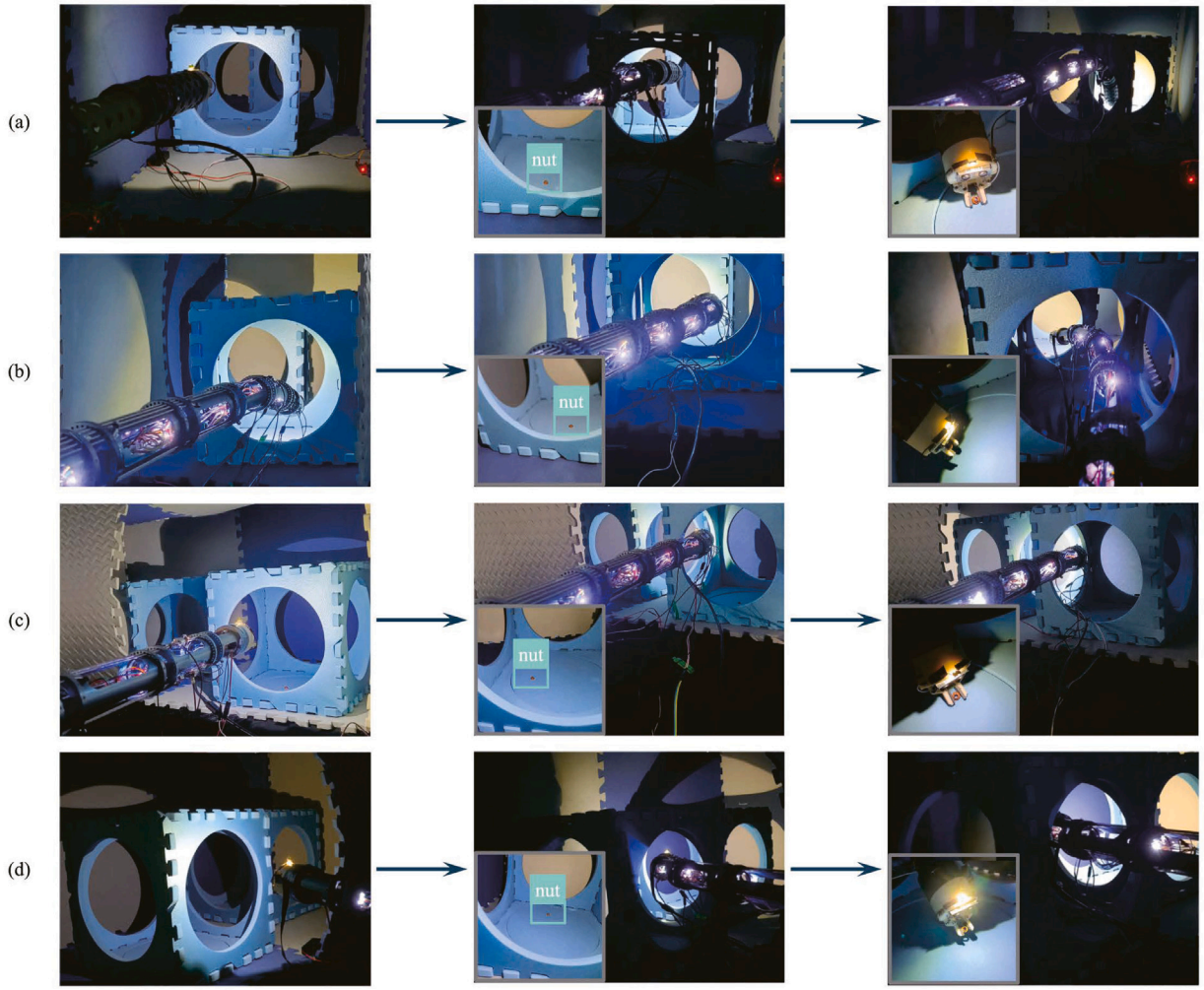

Fig. 6. Real-world experiments with SJTU-Snake III in multi-target environments using the SAC-PER algorithm. The real-world test encompasses the path generation and target tracking task. (a)–(d) Four distinct internal tank environments were constructed to simulate unknown settings.

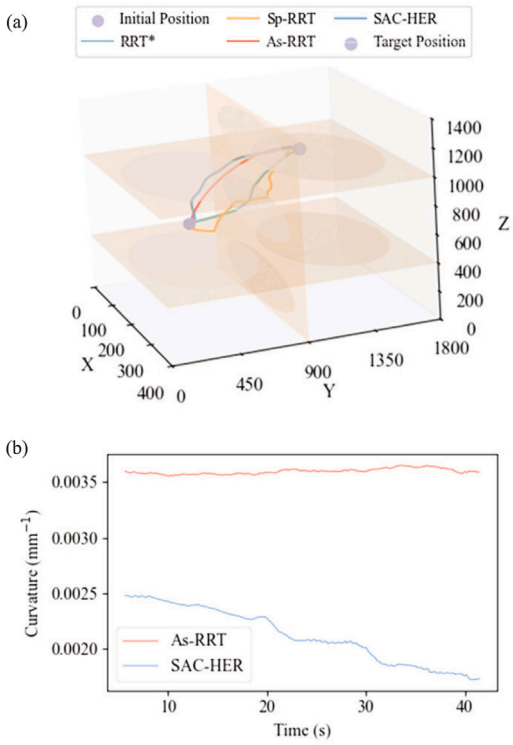

Fig. 7. (a) Path generated by the proposed method and all baseline algorithms. (b) Curvature of paths generated by SAC-HER and As-RRT algorithm. The path generated by As-RRT exhibits a maximum curvature of 0.0036 mm^{-1} and an average curvature of 0.0036 mm^{-1}. In comparison, the path generated by SAC-HER has a maximum curvature of 0.0025 mm^{-1} and an average curvature of 0.0021 mm^{-1}.

ensure a fair and interpretable comparison, all methods are evaluated with the same environment generation, observation/action definitions, and training budget. Since this setting suffers from sparse rewards caused by obstacle-induced failures, HER is activated for the SAC baseline in Table 2. All results are reported as the mean ± standard deviation from 10 independent experimental runs. To statistically validate the observed performance improvements, significance testing was performed on results from 10 independent runs. Normality was evaluated via the Shapiro–Wilk test, and variance homogeneity was examined using Levene's test. With the equal-variance assumption violated, Welch's ANOVA was applied to test for overall group differences across each evaluated metric. Games-Howell post-hoc tests were then conducted between the proposed method and each baseline, an approach well-suited for heteroscedastic data that also corrects for multiple comparisons. Full statistical results are detailed in Appendix A.

As shown in Table 2, the SAC-HER algorithm demonstrates a 7.05% reduction in path length compared with Sp-RRT and a 4.06% decrease relative to As-RRT. It also manifests a 96.56% enhancement in computational efficiency over RRT* and a 97.95% improvement compared with As-RRT. In terms of path smoothness, our algorithm outperforms existing methods. It achieves a 712.83% improvement over RRT*, a 19.71% gain versus Sp-RRT, and a 36.21% betterment than As-RRT. In the multi-target scenario, SAC-PER maintains competitive path length while significantly improving computational efficiency and path smoothness. Although RRT* attains the shortest path, SAC-PER reduces the path length by 10.27% compared with Sp-RRT and by 0.63% relative to As-RRT. More notably, SAC-PER requires substantially less computation time, achieving efficiency gains of 96.15% over RRT* and 98.44% over As-RRT. Meanwhile, SAC-PER produces the smoothest trajectory, with the path smoothness index improved by 454.50%, 250.41%, and 0.24% compared with RRT*, Sp-RRT, and As-RRT, respectively. Therefore, SAC-PER offers a favorable trade-off among path quality, planning speed, and trajectory smoothness in multi-target path planning. Welch's ANOVA indicates significant differences among methods for all three metrics ($p \ll 0.001$), and Games-Howell post-hoc tests confirm that the proposed method differs significantly from each baseline on every metric (all adjusted $p < 0.001$). Therefore, this method has proved to be suitable for CDHRR path planning tasks in simulation environments.

4.3. Experimental results and discussion

Due to physical hardware security and wear risks, training was not directly deployed on physical systems. Instead, the practical feasibility of the proposed path planning method is validated by reproducing simulation-generated optimal paths on real-world platforms. Real-world experiments focus on the planning module, as the task involves entering a fuel tank simulator to identify internal debris. The perception module employs a stereo camera on the CDHRR's end effector for real-environment target recognition, with tank internal obstacles and target points randomly generated. The planning module generates feasible paths accordingly.

As shown in Fig. 5, once inside the fuel tank, the CDHRR's highly redundant mechanical structure offers distinct advantages. Its dexterity structure allows flexible movement between chambers. The manipulator navigates around fixed partitions and moves through connecting holes between chambers. Following the planned path, it advances step by step toward the deep target location within the fuel tank. Throughout this process, the manipulator demonstrates an intelligent

and orderly exploration posture. This highlights its strong adaptability and operability in complex spatial environments, providing a reliable guarantee for accomplishing the internal tank exploration mission.

Fig. 6 demonstrates the effectiveness of the CDHRR within another test fuel tank. To simulate unknown and varied environments, four distinct internal tank settings were meticulously constructed. During the experimental process, a stereo camera accurately detects debris. Following this detection, our proposed algorithm generates feasible paths for navigating the debris. Within these configurations, the CDHRR autonomously followed the generated paths to successfully identify and grasp debris targets, thereby validating its capability in complex, unfamiliar scenarios.

Fig. 7 offers a visual comparison of the path and its curvatures generated by SAC-HER and other algorithms. The path smoothness index for As-RRT stands at 0.3452, whereas SAC-HER achieves an index that is 14.9% higher. Furthermore, the experimental results demonstrate that our algorithm attains an 80% success rate, surpassing all previous algorithms. Previous algorithms, including As-RRT, exhibited lower success rates, primarily attributed to the relatively high curvature of the generated paths. Although the maximum curvature of As-RRT is close to its average curvature, the relatively high curvature still impacts the stability of CDHRR operations. In contrast, the proposed algorithm generates paths with lower curvature while effectively completing tasks. This advancement ensures the feasibility and stability of CDHRR operations, thereby enhancing the overall success rate of the task.

5. Conclusion

This study presents an instability-aware DRL path-planning framework for CDHRRs operating in confined and unknown workspaces. Instead of relying on complete environmental information or manual teleoperation, the proposed planner integrates a task-driven learning policy with a smoothness-based executability metric. It enables the robot to generate mechanically feasible and stable paths with minimal human intervention.

Quantitatively, the proposed planner achieves substantially higher planning efficiency than SOTA sampling-based baselines. SAC-HER reduces computation time by 96.56% compared with RRT* and by 97.95% compared with As-RRT. Meanwhile, it yields smoother paths, with a 712.83% improvement over RRT*, 19.71% over Sp-RRT, and 36.21% over As-RRT. SAC-PER improves computational efficiency by 96.15% over RRT* and 98.44% over As-RRT in the multi-target scenario, while improving the path smoothness index by 454.50% and 250.41% compared with RRT* and Sp-RRT, respectively. In real-world validations conducted on SJTU-Snake III robot platform, the learned planner delivers a 14.9% improvement in path smoothness and a 10% increase in task success rate relative to As-RRT. These results demonstrate that the generated trajectories are not only collision-free but also more executable for large-scale CDHRR operations. Moreover, under sparse-reward multi-obstacle training, the incorporation of HER accelerates SAC convergence by 2.88× over the SAC baseline without HER. It verifies the effectiveness of replay-based data augmentation in improving learning efficiency. It is also observed that the performance improvement brought by PER is relatively limited compared with HER in the considered scenarios (only 1.27×). This finding suggests that, for CDHRR path planning under strong stability constraints, enhancing the policy's exploration capability is more critical than prioritized sample reuse.

Overall, these results validate that the proposed method can efficiently generate smooth and stable paths for CDHRRs in unknown environments and reduce human intervention during exploration and execution. Notably, mechanical stability is only approximated qualitatively via path smoothness in the reward function, without a rigorous quantitative physical definition or explicit system model. Looking forward, several research directions can further enhance the practicality and generality of instability-aware DRL planning for large-scale CDHRRs in confined unknown workspaces. First, we will focus on risk-sensitive and safety-constrained learning by integrating configuration stability margins into policy optimization, treating stability indicators as safety objectives instead of implicit side effects. Second, we will enhance planner's robustness to respond to newly revealed obstacles and targets while ensuring smooth executable motions. Finally, we will investigate task-specific experience replay strategies that are better aligned with the inherent characteristics of continuum robot path planning.

CRediT authorship contribution statement

Zhenpu Zhu: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Zhanxuan Peng:** Writing – review & editing, Visualization, Conceptualization. **Yu Rong:** Writing – review & editing, Formal analysis. **Guoying Gu:** Supervision, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China [grant number 52025057], the Science and Technology Commission of Shanghai Municipality, China [grant number 24511103400] and Xplore Prize.

Appendix A. Statistical validation details

To address the statistical validation request, hypothesis testing was performed on the 10-run results reported in Table 2. Normality was first assessed using the Shapiro–Wilk test for each method and metric, and variance homogeneity was subsequently examined using Levene's test. As heteroscedasticity was confirmed, Welch's ANOVA was used to evaluate overall group differences for each metric. Finally, Games-Howell post-hoc comparisons were performed between SAC-HER and each baseline, with adjusted p-values reported to properly account for multiple comparisons (see Table A.3).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mechatronics.2026.103536.

Data availability

Data will be made available on request.

Table A.3
Statistical validation for Table 2.

Metric	Analysis	Group/Comparison	Statistic	df	Adjusted p-value	Sig.
Path length	Shapiro–Wilk	RRT*	0.9769	9	9.47×10^{-1}	****
	Shapiro–Wilk	Sp-RRT	0.9204	9	3.61×10^{-1}	****
	Shapiro–Wilk	As-RRT	0.9589	9	7.73×10^{-1}	****
	Shapiro–Wilk	SAC-HER	0.9274	9	4.23×10^{-1}	****
	Levene's test	All	4.0642	(3,36)	1.38×10^{-2}	*
	Welch's ANOVA	Overall	34 259.69	(3,18.60)	$<1 \times 10^{-35}$	***
	Games-Howell	SAC-HER vs. As-RRT	25.98	17.84	$<1 \times 10^{-15}$	***
	Games-Howell	SAC-HER vs. RRT*	−86.03	10.85	$<1 \times 10^{-15}$	***
	Games-Howell	SAC-HER vs. Sp-RRT	−64.96	11.37	$<1 \times 10^{-15}$	***
Computation time	Shapiro–Wilk	RRT*	0.9569	9	7.50×10^{-1}	****
	Shapiro–Wilk	Sp-RRT	0.9727	9	9.15×10^{-1}	****
	Shapiro–Wilk	As-RRT	0.9544	9	7.21×10^{-1}	****
	Shapiro–Wilk	SAC-HER	0.9591	9	7.76×10^{-1}	****
	Levene's test	All	24.6223	(3,36)	7.69×10^{-9}	***
	Welch's ANOVA	Overall	44 500.86	(3,15.39)	$<1 \times 10^{-30}$	***
	Games-Howell	SAC-HER vs. As-RRT	143.75	9.03	$<1 \times 10^{-15}$	***
	Games-Howell	SAC-HER vs. RRT*	334.51	9.44	$<1 \times 10^{-15}$	***
	Games-Howell	SAC-HER vs. Sp-RRT	57.06	9.72	$<1 \times 10^{-13}$	***
Path smooth index	Shapiro–Wilk	RRT*	0.9538	9	7.14×10^{-1}	****
	Shapiro–Wilk	Sp-RRT	0.9007	9	2.23×10^{-1}	****
	Shapiro–Wilk	As-RRT	0.9456	9	6.17×10^{-1}	****
	Shapiro–Wilk	SAC-HER	0.8896	9	1.68×10^{-1}	****
	Levene's test	All	5.3739	(3,36)	3.67×10^{-3}	**
	Welch's ANOVA	Overall	72 781.64	(3,16.77)	$<1 \times 10^{-34}$	***
	Games-Howell	SAC-HER vs. As-RRT	−13.43	17.98	$<1 \times 10^{-10}$	***
	Games-Howell	SAC-HER vs. RRT*	−164.29	9.28	$<1 \times 10^{-15}$	***
	Games-Howell	SAC-HER vs. Sp-RRT	29.99	10.33	$<1 \times 10^{-10}$	***

Significance codes:
* $0.01 \leq p < 0.05$.
** $0.001 \leq p < 0.01$.
*** $p < 0.001$.
**** $p \geq 0.05$.

References

[1] Ma S, Liang B, Wang T. Dynamic analysis of a hyper-redundant space manipulator with a complex rope network. Aerosp Sci Technol 2020;100:105768. http://dx.doi.org/10.1016/j.ast.2020.105768.

[2] Mu Z, Liu T, Xu W, Lou Y, Liang B. A hybrid obstacle-avoidance method of spatial hyper-redundant manipulators for servicing in confined space. Robotica 2019;37(6):998–1019. http://dx.doi.org/10.1017/S0263574718001406.

[3] Zhang Z, Zheng L, Yu J, Li Y, Yu Z. Three recurrent neural networks and three numerical methods for solving a repetitive motion planning scheme of redundant robot manipulators. IEEE/ASME Trans Mechatronics 2017;22(3):1423–34. http://dx.doi.org/10.1109/TMECH.2017.2683561.

[4] Coemert S, Gao A, Carey JP, Traeger MF, Taylor RH, Lueth TC, Armand M. Development of a snake-like dexterous manipulator for skull base surgery. In: 2016 38th annu. int. conf. IEEE eng. med. biol. soc.. EMBC, 2016, p. 5087–90. http://dx.doi.org/10.1109/EMBC.2016.7591871.

[5] Hillel AT, Kapoor A, Simaan N, Taylor RH, Flint P. Applications of robotics for laryngeal surgery. Otolaryngol Clin North Am 2008;41(4):781–91. http://dx.doi.org/10.1016/j.otc.2008.01.021, Laryngeal Cancer.

[6] Ikuta K, Hasegawa T, Daifu S. Hyper redundant miniature manipulator "Hyper Finger" for remote minimally invasive surgery in deep area. In: Proc. 2003 IEEE int. conf. robot. autom.. ICRA, Vol. 1, 2003, p. 1098–102. http://dx.doi.org/10.1109/ROBOT.2003.1241739.

[7] Alatorre D, Nasser B, Rabani A, Nagy-Sochacki A, Dong X, Axinte D, Kell J. Teleoperated, in situ repair of an aeroengine: Overcoming the internet latency hurdle. IEEE Robot Autom Mag 2019;26(1):10–20. http://dx.doi.org/10.1109/MRA.2018.2881977.

[8] Tonapi MM, Godage IS, Vijaykumar AM, Walker ID. A novel continuum robotic cable aimed at applications in space. Adv Robot 2015;29(13):861–75. http://dx.doi.org/10.1080/01691864.2015.1036772.

[9] Qin G, Cheng Y, Pan H, Zhao W, Shi S, Ji A, Wu H. Systematic design of snake arm maintainer in nuclear industry. Fusion Eng Des 2022;176:113049. http://dx.doi.org/10.1016/j.fusengdes.2022.113049.

[10] Buckingham R, Graham A. Nuclear snake-arm robots. Ind Robot 2012;39(1):6–11. http://dx.doi.org/10.1108/01439911211192448.

[11] Ma S, Hirose S, Yoshinada H. Development of a hyper-redundant multijoint manipulator for maintenance of nuclear reactors. Adv Robot 1994;9(3):281–300. http://dx.doi.org/10.1163/156855395X00201.

[12] Su H, Liu M, Liu H, Huo J, Gou S, Su Q. Path planning of hyper-redundant manipulators for narrow spaces. IET Cyber-Syst Robot 2022;4(3):251–63. http://dx.doi.org/10.1049/csy2.12055.

[13] Rao P, Peyron Q, Lilge S, Burgner-Kahrs J. How to model tendon-driven continuum robots and benchmark modelling performance. Front Robot AI 2020;7. http://dx.doi.org/10.3389/frobt.2020.630245.

[14] Amanov E, Nguyen T-D, Burgner-Kahrs J. Tendon-driven continuum robots with extensible sections—A model-based evaluation of path-following motions. Int J Robot Res 2021;40(1):7–23. http://dx.doi.org/10.1177/0278364919886047.

[15] Grassmann R, Rao P, Peyron Q, Burgner-Kahrs J. FAS—A fully actuated segment for tendon-driven continuum robots. Front Robot AI 2022;9. http://dx.doi.org/10.3389/frobt.2022.873446.

[16] Du Z-c, Ouyang G-Y, Xue J, Yao Y-b. A review on kinematic, workspace, trajectory planning and path planning of hyper-redundant manipulators. In: 2020 10th IEEE int. conf. cyber technol. autom. control intell. syst.. CYBER, 2020, p. 444–9. http://dx.doi.org/10.1109/CYBER50695.2020.9279171.

[17] Seleem IA, El-Hussieny H, Ishii H. Recent developments of actuation mechanisms for continuum robots: A review. Int J Control Autom Syst 2023;21(5):1592–609. http://dx.doi.org/10.1007/s12555-022-0159-8.

[18] Zhao Q, Liu H, Zhang Y, Wang J. Path planning fusion algorithm based on improved A-star and adaptive dynamic window approach for mobile robot. Int J Ind Eng 2023;30(5):1078–89. http://dx.doi.org/10.23055/ijietap.2023.30.5.8713.

[19] Ji H, Xie H, Wang C, Yang H. E-RRT*: Path planning for hyper-redundant manipulators. IEEE Robot Autom Lett 2023;8(12):8128–35. http://dx.doi.org/10.1109/LRA.2023.3325716.

[20] Karaman S, Frazzoli E. Sampling-based algorithms for optimal motion planning. Int J Robot Res 2011;30(7):846–94. http://dx.doi.org/10.1177/0278364911406761.

[21] Tang L, Wang J, Zheng Y, Gu G, Zhu L, Zhu X. Design of a cable-driven hyper-redundant robot with experimental validation. Int J Adv Robot Syst 2017;14(5):172988141773445. http://dx.doi.org/10.1177/1729881417734458.

[22] Zhang W, Shan L, Chang L, Dai Y. SVF-RRT*: A stream-based VF-RRT* for USVs path planning considering ocean currents. IEEE Robot Autom Lett 2023;8(4):2413–20. http://dx.doi.org/10.1109/LRA.2023.3245409.

[23] Wei H, Zheng Y, Gu G. RRT-based path planning for follow-the-leader motion of hyper-redundant manipulators. In: 2021 IEEE/RSJ int. conf. intell. robots syst.. IROS, IEEE; 2021, p. 3198–204. http://dx.doi.org/10.1109/IROS51168.2021.9635876.

[24] Deng J, Li Z, Zheng Y, Gu G. An RRT-based motion planning method for hyper-redundant manipulators in confined spaces. In: 2022 IEEE int. conf. robot. biomimet.. ROBIO, IEEE; 2022, p. 1067–73. http://dx.doi.org/10.1109/ROBIO55434.2022.10011650.

[25] Hirade K, Niiyama R. Curriculum reinforcement learning for obstacle avoidance postures for a hyper-redundant manipulator. In: 2025 IEEE/SICE int. symp. syst. integr.. SII, 2025, p. 199–204. http://dx.doi.org/10.1109/SII59315.2025.10871128.
[26] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, http://dx.doi.org/10.48550/arXiv.1707.06347, arXiv.
[27] Zhang D, Ju R, Cao Z. DDPG-based path planning for cable-driven manipulators in multi-obstacle environments. Robotica 2024;42(8):2677–89. http://dx.doi.org/10.1017/S0263574724001048.
[28] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. 2015, http://dx.doi.org/10.48550/arXiv.1509.02971, arXiv.
[29] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proc. 35th int. conf. mach. learn.. PMLR; 2018, p. 1861–70. http://dx.doi.org/10.48550/arXiv.1801.01290.
[30] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P, Levine S. Soft actor-critic algorithms and applications. 2019, http://dx.doi.org/10.48550/arXiv.1812.05905.
[31] Sun D, Wen J, Wang J, Yang X, Hu Y. A path planning method based on deep reinforcement learning with improved prioritized experience replay for human-robot collaboration. In: Kurosu M, Hashizume A, editors. Human-comput. interact.. Springer Nature Switzerland; 2024, p. 196–206. http://dx.doi.org/10.1007/978-3-031-60412-6_15.
[32] Xu H, Xue C, Chen Q, Yang J, Liang B. Continuous multi-target approaching control of hyper-redundant manipulators based on reinforcement learning. Mathematics 2024;12(23):3822. http://dx.doi.org/10.3390/math12233822.
[33] Singh B, Kumar R, Singh VP. Reinforcement learning in robotic applications: a comprehensive survey. Artif Intell Rev 2022;55(2):945–90. http://dx.doi.org/10.1007/s10462-021-09997-9.
[34] Lilge S, Nuelle K, Childs JA, Wen K, Rucker DC, Burgner-Kahrs J. Parallel-continuum robots: A survey. IEEE Trans Robot 2024;40:3252–70. http://dx.doi.org/10.1109/TRO.2024.3415230.
[35] Shi X, Guo Y, Chen X, Chen Z, Yang Z. Kinematics and singularity analysis of a 7-DOF redundant manipulator. Sensors 2021;21(21):7257. http://dx.doi.org/10.3390/s21217257.
[36] Rao P, Pogue C, Peyron Q, Diller E, Burgner-Kahrs J. Modeling and analysis of tendon-driven continuum robots for rod-based locking. IEEE Robot Autom Lett 2023;8(6):3126–33. http://dx.doi.org/10.1109/LRA.2023.3264869.
[37] Zhu Z, Li Z, Peng Z, Liu C, Gu G. Hybrid tension and configuration control of cable-driven hyper-redundant robots for high accuracy and stability. IEEE Robot Autom Lett 2025;10(6):5601–8. http://dx.doi.org/10.1109/LRA.2025.3559829.
[38] Mustafa SK, Agrawal SK. On the force-closure analysis of n-DOF cable-driven open chains based on reciprocal screw theory. IEEE Trans Robot 2012;28(1):22–31. http://dx.doi.org/10.1109/TRO.2011.2168170.
[39] Zhang S, Zhao J, Zhang X, Bi H, Yao W, Chen F, Peng H, Liu C. Quasi-static modeling of a cable-driven continuum manipulator considering non-smooth cable-hole friction and experimental verification. Mech Mach Theory 2024;204:105856. http://dx.doi.org/10.1016/j.mechmachtheory.2024.105856.
[40] Zhang Y, Chen P. Path planning of a mobile robot for a dynamic indoor environment based on an SAC-LSTM algorithm. Sensors 2023;23(24):9802. http://dx.doi.org/10.3390/s23249802.
[41] Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: A survey. Int J Robot Res 2013;32(11):1238–74. http://dx.doi.org/10.1177/0278364913495721.

Zhenpu Zhu is currently pursuing the Ph.D. degree in mechanical engineering at Shanghai Jiao Tong University, Shanghai, China. His research interests include modeling and control of cable-driven hyper-redundant robots.

Zhanxuan Peng is currently pursuing the Ph.D. degree in mechanical engineering at Shanghai Jiao Tong University, Shanghai, China.

His research interests include end effector design and motion control of hyper-redundant manipulators.

Yu Rong received the B.E. degree in mechatronic engineering from Northwestern Polytechnical University, Xi'an, China, in 2019. He is currently working toward the Ph.D. degree in mechatronic engineering with Shanghai Jiao Tong University, Shanghai, China.

His research interests include modeling and control of soft continuum robots.

Guoying Gu received the B.E. degree (Hons.) in Electronic Science and Technology and Ph.D. degree (Hons.) in Mechatronic Engineering from Shanghai Jiao Tong University in 2006 and 2012 respectively. He is a former Humboldt Fellow with visiting experience at top institutions in the US, Singapore and Canada, he is now a Distinguished Professor at Shanghai Jiao Tong University. He has published over 150 academic papers, focuses on soft robotics, bioinspired/wearable robots and smart materials, won the National Science Fund for Distinguished Young Scholars and XPLORER Prize, and serves as associate editor for multiple top international journals.