2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11247411

Continuous renal calculi tracking for autonomous robotic ureteroscopic lithotripsy

Quan Zhang[1], Xiaohang Nie[1], Jingqian Sun[1], Ruichao Tang[1], Yichao Tang[1,2*]
1 Tongji University, Shanghai, China
2 Shanghai Innovation Institute, Shanghai, China
*Corresponding author. Email: tangyichao@tongji.edu.cn

***Abstract*— Renal calculi, while not inherently life-threatening, can induce excruciating pain during acute episodes. The predominant clinical treatment – ureteroscopic lithotripsy (URS) –currently faces challenges including restricted maneuverability, frequent manual adjustments during dynamic calculi movement, and tissue damage risk, highlighting the need for robotic assistance. This study proposes an autonomous lithotripsy system through three integrated technological advancements: 1) a robotic ureteroscope with sub-millimeter-scale positioning accuracy; 2) a concatenated Quenching-net semantic segmentation visual processing framework based on convolutional neural networks, achieving a segmentation accuracy of 98.5% (validated on 12,768 endoscopic images); 3) a control strategy developed through collaborative deep reinforcement learning(DRL), enabling a 93% success rate in tracking randomly moving calculi. This system's autonomous calculi localization capability reduces operator fatigue and may mitigate cognitive bias in calculi targeting. It demonstrates how embodied AI enhances medical procedural precision while preserving human oversight in critical decisions.**

I. INTRODUCTION

The global incidence of urolithiasis is between 5% and 15%, showing an increasing trend, with renal calculi accounting for 40% to 50% [1]. Calculi can cause severe pain and complications such as renal dysfunction [2]. The treatment of renal calculi can be categorized into pharmacological interventions and non-pharmacological approaches. Pharmacological therapy is suitable for calculi less than 6 mm in diameter with no urethral obstruction, but the recurrence rate is relatively high. Non-pharmacological treatments include extracorporeal shock wave lithotripsy (ESWL), percutaneous nephrolithotomy (PNL), and ureteroscopic lithotripsy (URS) [3]. ESWL is suitable for renal calculi with a diameter of 5 to 20 mm without incision. However, this modality necessitates multiple treatments and demonstrates limited efficacy for calculi with high density or those located in anatomically challenging positions. PNL is suitable for larger calculi but requires open surgery.

Compared with the aforementioned approaches, URS is accessed through the natural lumen of the human body and is characterized by minimally invasiveness and rapid recovery. However, success heavily relies on surgeon skill, demanding continuous fine-tuning of the ureteroscope during dynamic calculi movement, accelerating physician fatigue. Current ureteroscopes also suffer from limited bending angles, hindering access to difficult regions [4], [5], [6]. Developing a robotic system with a high automation level and high deformation capability is the solution to address these issues by enabling high-precision positioning and stable operation, reducing the workload of surgeons, and lowering surgical risks [7], [8].

The first set of ureteroscope-assisted robot systems was developed in 2008 [9]. Surgeons operated the ureteroscope system through a remote master-slave control interface. They determined the operational strategies according to the feedback obtained from the ureteroscope imaging. The field has since undergone significant technological iterations. The platform presented by the CoFlex system realized accurate operation in the animal model of bladder cancer through the modular integration of the six-degree-of-freedom robotic arm and the digital ureteroscope, and its master-slave control architecture based on force feedback can greatly shorten the duration of a single lithotripsy [10]. Xie et al. from Shanghai Jiao Tong University designed a master-slave robot prototype using magnetic field to locate and control the ureteral endoscope with a pressure sensor mounted at the catheter tip to monitor intrarenal pressure [11].

While multimodal sensing integration enhances robotic automation capabilities through enriched data inputs, the incorporation of redundant sensors inevitably exacerbates size constraints, ultimately compromising the system's maneuverability within confined luminal structures. This elucidates the predominant reliance on endoscopic imaging for guiding the endoscopic robots, mirroring the approach used in manual surgeries. In such imaging-guided surgeries, lesion recognition and localization are the keys to robotic navigation.

The easyUretero system has demonstrated excellent ergonomic performance during flexible ureteroscopy and laser lithotripsy procedures through its segmentation algorithm based on endoscopic medical imaging [12]. The SmartEye system uses a deep learning framework to process real-time endoscopic and multi-modal imaging data (such as MRI and CT) [13]. Through automatic feature learning and extraction, the synchronous processing and fusion of different image data are realized to solve the occlusion and interference issues. This provides a more comprehensive and optically clear field of view, and thus more accurate lesion identification [14]. To address the issues of multi-scene adaptation and dynamic disturbance, Wang et al. proposed a context-aware network for object detection under occlusion and interference, which is

especially suitable for the detection of moving kidney calculi [15]. The major challenges of ureteroscopy segmentation include dynamic interference due to the human body movement, and bubble interference generated by lithotripsy. Through introducing convolutional neural networks, it shows excellent performance in processing such medical images. For example, Zekun et al. proposed a self-supervised scene adaptation convolutional neural network to improve the performance of object detectors through the pseudo-label generation and cross-teaching mechanisms. This method has achieved an enhanced detection rate of kidney calculi [16].

In addition, depth estimation techniques play a crucial role in tracking the exact spatial coordinates of the kidney calculi. Jing et al. [17] used a multi-level fusion method combining RGB images and pseudo-lidar data to improve the depth estimation accuracy. They developed a ranging function to capture the distance between the endoscope and the interactive target. Yu et al. combined depth estimation with the YOLO framework to achieve efficient object detection and depth estimation from monocular camera images [18]. Xu et al. developed an integrated network that can accurately detect and estimate the depth of objects in monocular RGB images while ensuring high-speed processing[19].

Nevertheless, surgical robots still need to be improved in multi-scenario adaptation and dynamic disturbance handling [20], [21]. Future research focuses on enhancing automation using deep reinforcement learning (DRL)[22],[23] and embodied intelligence [24],[25], enabling autonomous image analysis and execution, alongside refined designs and control strategies for precision and safety.

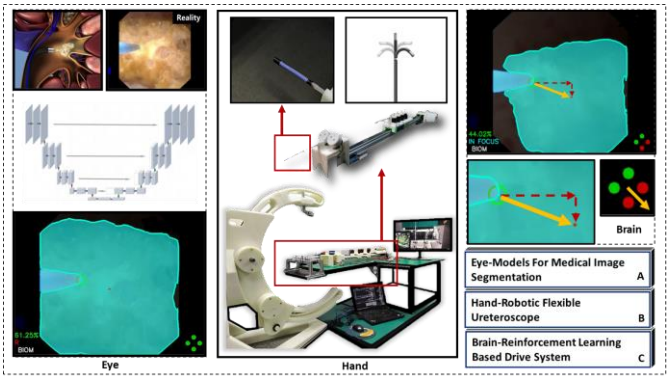

Figure 1. System design for autonomous ureteroscopic lithotripsy robot. (A) U-net model for medical image segmentation. (B) Robotic ureteroscope. (C) DRL-based control system architecture.

This study is dedicated to tackling the persistent challenge of manual fine-tuning of ureteral endoscope orientation during calculi removal in ureteroscopic surgery (URS). We present a novel ureteroscopic lithotripsy robotic system (depicted in Fig. 1) designed to autonomously and continuously track calculi targets, maintaining their central alignment within the field of view even amidst random calculi movements. The experimental results demonstrate that our robotic system achieves a segmentation accuracy of over 98% for calculi lesions and a center-alignment-control success rate of 93%. Such robust performance lays a solid foundation for fully autonomous ureteroscopy lithotripsy. The closed-loop control logic governing agent states (S_t), action selections (A_t), and reward feedback (R_t) is formalized in Fig. 2, establishing the foundation for autonomous target tracking.

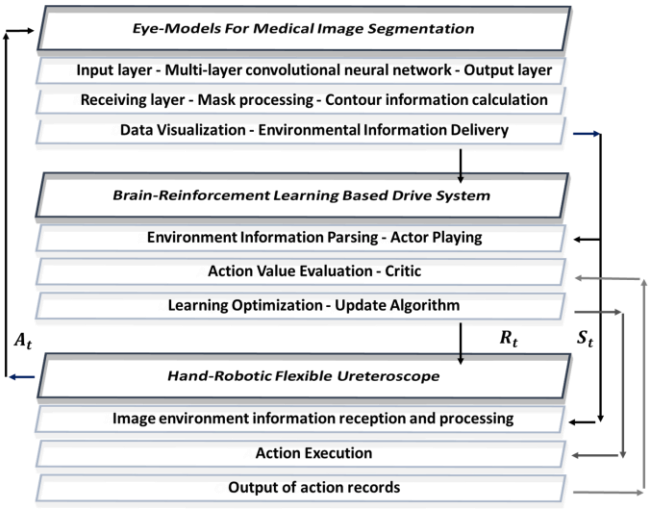

Figure 2.The overall system operation logic. S_t: The state of the agent at time step t. It contains all the information needed for the agent to decide the next action. A_t: The action chosen by the agent at time step t. It is the behavior taken by the agent based on the current state of St to interact with the environment. R_t: The reward received by the agent at time step t. It is the feedback from the environment to the agent's action, that guides the agent to learn a better policy.

II. Results

A. Autonomous Robotic Ureteroscopic Lithotripsy System Design

The system workflow (Fig. 3) integrates visual perception and motion control. The perception module processes endoscopic images and extracts critical information through three stages. Initially, the input layer applies adaptive histogram equalization to optimize image contrast while mitigating noise interference. Then utilizes an enhanced Quenching-net architecture that extracts high-level semantic features via downsampling while preserving anatomical details through upsampling operations, ultimately producing precise calculi segmentation masks. In the output phase, the system integrates geometric centroid computation with depth estimation algorithms to precisely localize calculi and execute spatial coordinate transformations through established projection relationships.

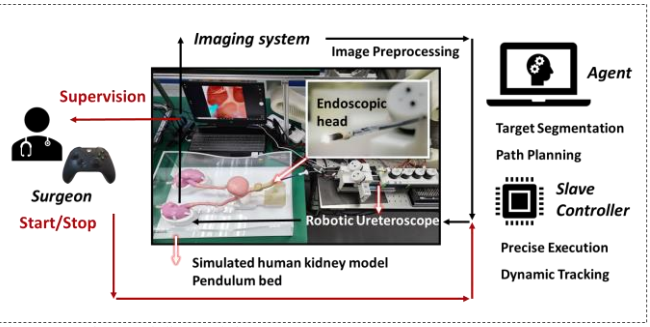

Figure 3. The workflow of the system.

The control module executes path planning based on positional deviations between target and current reference points, implemented via ROS2. A DRL layer (DDPG

algorithm) enables autonomous path correction. The execution layer achieves 0.1mm accuracy through cascaded PID control (position, velocity, current loops). A dual safety mechanism combines algorithmic monitoring (visual positioning error threshold: 2mm; control latency threshold: 50ms) with physical safeguards; exceeding thresholds triggers emergency stops, dissipating end-effector kinetic energy to <0.5 J within 150ms. This hybrid design balances algorithm robustness with mechanical safety, resolving the trade-off between speed and safety.

B. Robotic Ureteroscope

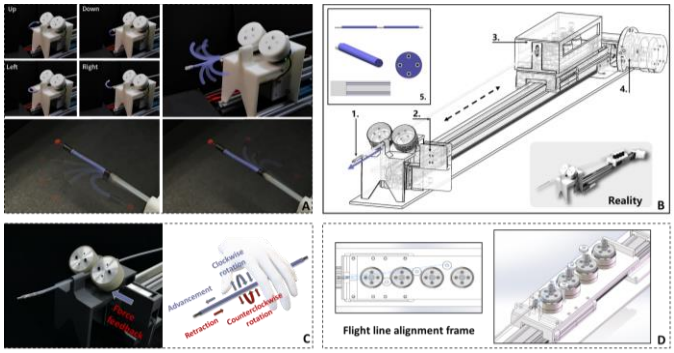

Figure 4. Robotic Ureteroscope. (A)Flexibility demonstration (360° bending). (B) Overall component diagram. The markings in the figure are explained as follows: 1. Robotic Ureteroscope; 2. Rotating Platform; 3. Steering Platform; 4. Pushing Platform; 5.Structural Diagram of the Main Body of the Robotic Ureteroscope. (C) Biomimetic friction wheel with force sensor. (D) Wire alignment frame preventing interference.

The core functional unit is a 4.0 mm diameter steerable head (Fig. 4) featuring a hybrid rigid-flexible modular design with graded stiffness silicone tubes. Four circumferentially symmetric 0.2-mm SMA wire arrays enable 3D controllable bending (±180° deflection, Fig. 4-A). The Nitinol-based actuation system combines axial stiffness with radial compliance, ensuring torsional stability during both linear advancement and directional steering. An integrated self-recoiling mechanism permits agile navigation through tortuous luminal pathways while maintaining rapid shape recovery post-deflection.

As illustrated in Fig. 4-B, the robotic control platform integrates a high-precision servo motor array with a biomimetic friction-wheel transmission system. A precision-engineered synchronous belt slider enables bidirectional actuation, achieving propulsion accuracy of 0.5 mm within a 0–10 N thrust range. This configuration permits simultaneous execution of rotational, axial, and articulatory motions essential for navigating constrained anatomical spaces.

A notable innovation is the biomimetic surgeon-finger friction wheel assembly (Fig. 4-C), featuring an angled motor mount that imparts bidirectional torsional forces to the catheter. The integrated micro-electrocylinder dynamically regulates wheel-endoscope contact pressure (0–15 N range), ensuring consistent rotational drive across variable loading conditions. The flywire management platform (Fig. 4-D) eliminates mechanical interference between steering and traction components, while an integrated tool channel supports multifunctional surgical instrumentation.

By advancing traditional endoscopy from single-DOF motion to spatial multi-DOF operation, the robotic ureteroscope achieves synergistic axial propulsion, circumferential rotation, and curvature adaptation within narrow luminal structures. This paradigm reduces tissue abrasion risks through two complementary strategies: (1) stability via closed-loop PID control of both propulsion force (regulated by drive motor torque) and steering angles (controlled by steering motor position), and (2) real-time trajectory compensation during complex navigation. Experimental validation in simulated calyx models confirms sub-millimeter positioning accuracy ($\leqslant$0.8 mm).

C. CNN-based semantic segmentation model for medical calculi image

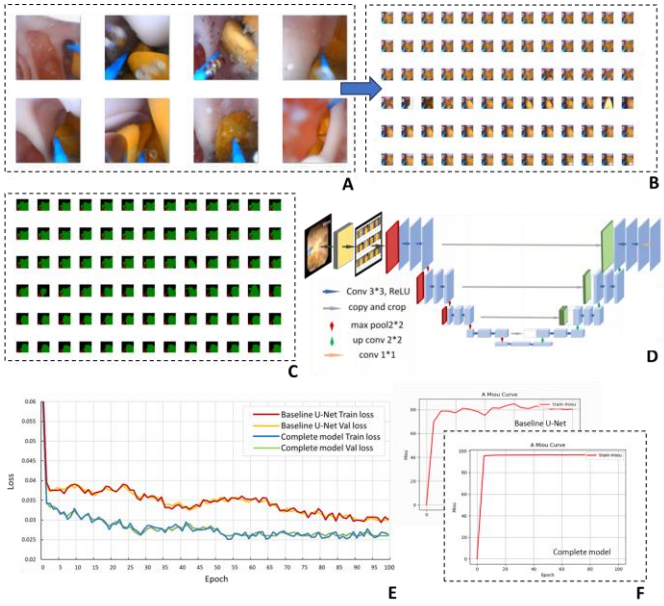

Figure 5. Visual semantic segmentation experiment. (A)Original video data, containing six different types of calculi, bubbles, and interference from calculi fragments. (B) JPEG format dataset. (C) Annotated dataset with semantic labels in PNG format. (D) Schematic diagram of the Quenching-net network structure. (E) Training comparison between Baseline Quenching-net and the complete model. (F) Miou metric comparison between Baseline Quenching-net and the complete model.

Accurate target semantic segmentation and depth estimation provide crucial data support for robotic control. To train our model, we collected endoscopic video data from a simulated human model featuring six distinct calculi types, with some sequences incorporating bubble interference and stone powder (Fig. 5-A). We then constructed the training dataset by randomly sampling frames from these videos (Fig. 5-B) and performing pixel-wise semantic annotation (Fig. 5-C), which was subsequently validated by clinical surgeons. Addressing key challenges in endoscopic urologic imaging—such as dynamic scene changes and complex anatomical interference—we developed an enhanced Quenching-net architecture (Fig. 5-D). This network integrates multi-scale feature fusion, spatial attention mechanisms, adaptive preprocessing, and depth estimation to achieve precise pixel-level calculi segmentation. Performance comparisons against the baseline model are shown in Fig. 5-E and Fig. 5-F.

In terms of model architecture improvement, this study has implemented three key innovations based on the Quenching-net framework. First, an adaptive preprocessing module was introduced, employing the contrast limited adaptive histogram equalization algorithm to dynamically adjust image contrast [26]:

$$T(I)(x,y) = \sum_{i=0}^{I(x,y)} \frac{p_i}{1+\lambda \cdot g(i)} \quad (1)$$

Among them, λ compensates for tissue-specific light absorption/scattering variations (tissue transmittance coefficient), and g(i) is the local gray-level distribution function enabling context-aware contrast enhancement. Experiments demonstrated this module suppresses overexposed areas and boosts calculi boundary contrast by 41%, significantly improving image clarity.

The multi-scale attention fusion mechanism is another major innovation in this study. The double-path attention module is embedded in the skip connection. The channel attention path adopts the most critical feature channel of the SENet structure for calculi recognition [27], and the expression of important features is enhanced by learning the channel weight. The spatial attention path captures contextual information through a 3×3 zero-padding convolution, taking into account spatial relationships between different locations in the image to better understand the position and shape of the calculi. The experimental results show that the IoU of the model is increased by 5.3% on the same data set, which effectively improves the accuracy of calculi segmentation.

Regarding the optimization of training strategies, this study employed a series of advanced data augmentation schemes and hybrid supervised training methods. The data augmentation schemes included physical simulation enhancement, motion blur synthesis, and adversarial sample generation. Physical simulation enhancement uses a light transmission model to simulate different tissue illumination scenarios, enhancing the model's adaptability to endoscopic images under various lighting conditions. Motion blur synthesis realistically simulates dynamic disturbances during surgery, improving the model's ability to handle actual surgical images. Adversarial sample generation uses CycleGAN for cross-domain style transfer to create diversified samples [28], further enriching training data and improving model robustness. Hierarchical loss function is used for hybrid supervised training:

$$L_{total} = \alpha L_{Dice} + \beta L_{Boundary} \quad (2)$$

Dice loss term (primary segmentation supervision):

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i p_i + \varepsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} p_i + \varepsilon} \quad (3)$$

$y_i \in 0,1$: True labeling of pixels (calculi/background); $p_i \in [0,1]$: Predicted probability of pixel; $\varepsilon = le^{-5}$: Numerical stability constant.

Boundary-aware loss term (edge enhancement):

$$L_{Boundary} = -\frac{1}{|\Omega|}\sum_{p\in\Omega}\left[w_p \cdot y_p log p + (1-y_p) log(1-p_p)\right] \quad (4)$$

The boundary loss term, derived through Sobel operator-based edge detection matrices, enhances model guidance in discerning calculi boundary features while optimizing segmentation precision.

These architectural refinements address critical challenges inherent in ureteroscopic renal calculi tracking: (1) Dynamic illumination fluctuations in the narrow urinary lumen can obscure calculus boundaries; the adaptive preprocessing module mitigates this by enhancing contrast and suppressing noise, ensuring consistent target visibility necessary for continuous tracking. (2) Scale variance among calculi fragments and frequent anatomical occlusions (e.g., by tissue folds or debris) challenge robust segmentation; multi-scale attention fusion maintains accuracy across scales and mitigates partial occlusions, enabling reliable path planning. (3) Motion artifacts from rapid scope movement or calculi drift blur target contours; the boundary-aware loss explicitly sharpens edge delineation, preventing tracking drift caused by ambiguous boundaries. Collectively, these modifications ensure segmentation performance remains robust under the demanding visual conditions of ureteroscopy, which is fundamental for autonomous tracking.Our validation protocol implemented five iterative optimization tiers, systematically enhancing model performance through hierarchical refinement (Fig. 6). Benchmark evaluation utilized a clinically annotated dataset of 12,768 endoscopic images encompassing seven calculi variants, demonstrating a Dice coefficient of 96.1% ± 2.3% and real-time inference capability (25.4–36.2 fps on RTX 4090 hardware). These quantitative outcomes confirm substantial advancements in both segmentation accuracy (±1.8% improvement over baseline) and operational efficiency, achieving clinical-grade performance thresholds.

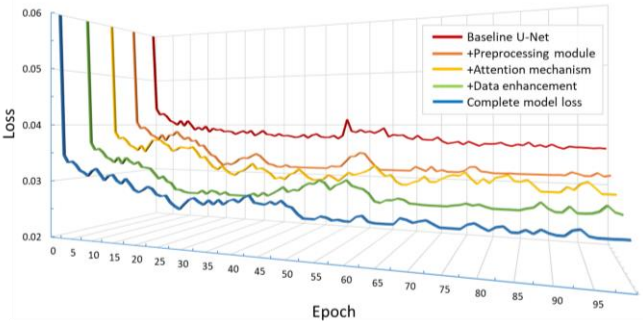

Figure 6. Comparison of loss results in the ablation experiment. The X-axis represents the epoch of the training: 0-100, the Z-axis represents the L_{total} value of the training, and the Y-axis is followed by five-fold lines distributed on five parallel planes, representing the training results from the baseline Quenching-net optimization to the complete model in this study from the far to the near.

Table 1. Ablation Experiment

Block Assembly	Loss (%)	FP Rate (%)	MIoU (%)
Baseline Quenching-net	3.1	4.1	92
+Preprocessing module	3.2	3.5	92.8
+Attention mechanism	2.9	3.6	94.2
+Data enhancement	2.8	2.4	96.9
Complete model	**2.6**	**1.9**	**98.5**

To systematically evaluate module contributions, we performed comprehensive ablation studies across iterative model configurations (Table 1). The baseline Quenching-net exhibited a validation loss of 3.1% with a 4.1% false-positive rate (FPR). Sequential integration of core components

demonstrated progressive performance gains: Preprocessing module reduced FPR by 14.6% (4.1% → 3.5%); Attention mechanisms decreased validation loss by 6.5% (3.1% → 2.9%); Data augmentation further optimized both metrics (Loss: 2.8%, FPR: 2.4%). The full integrated architecture achieved peak performance (Loss: 2.6%, FPR: 1.9%), reflecting a 16.1% total loss reduction and 53.7% FPR improvement over baseline. These results demonstrate compelling evidence for both individual module efficacy and their synergistic interplay, particularly in mitigating false-positive artifacts while maintaining convergence stability.

D. Control Algorithm and Drive System Optimization Based on Deep reinforcement learning (DRL)

For clinical implementation, we deployed the segmentation network on a custom-developed robotic ureteroscope platform featuring a ROS2-integrated real-time processing pipeline [29]. The image acquisition layer supports 1080p@30fps video stream input and can collect high-definition surgical images in real-time. The inference acceleration layer adopts TensorRT FP16 quantization technology [30], which significantly improves the inference speed of the model. An adaptive PID closed loop is established between the control interface layer and the robot controller to realize the precise control of surgical instruments. The communication frequency is stable at 30fps, and the end-to-end delay is controlled within 30ms, providing reliable technical support for its application in actual surgery. Accurate tracking of dynamic targets and efficient control of drive systems are the core challenges in autonomous calculi. In visual tracking tasks, although the cascade U-shaped network can segment the target contour and calculate the geometric center, the alignment of the robot head with the target center is the key to accurate tracking. For this purpose, we develop a tracking path planning algorithm based on DRL. By training Deep Q-Network (DQN) in a simulated environment to optimize the tracking effect [31], the algorithm accepts inputs such as the relative coordinate difference between the target and the robot reference point, the position of the robot head and the Angle of the steering motor, and outputs specific robot actions.

To enable surgical robots to autonomously learn and adapt to dynamically changing surgical environments, we built a DRL framework based on the Proximal Policy Optimization (PPO) algorithm [32] and designed a hierarchical reward mechanism. The basic reward term is $R_{base} = \frac{1}{1+\|\Delta p\|^2}$, where $\| \Delta p \|$ denotes the Euclidean distance between the robot's end-effector and target calculus in 3D space; the stability reward term is $R_{stab} = exp(-\frac{\|\Delta v\|}{v_{max}})$, where Δv is the magnitude of velocity change between consecutive timesteps and v_{max} is the maximum allowed velocity. Through continuous interaction with the environment, the agent learns action policies maximizing cumulative reward, forming a closed-loop optimization cycle. Policy training occurred in a PyBullet physics simulator featuring anatomically accurate renal phantoms with physiological disturbances (respiratory motion, irrigation flow). Domain randomization varied lighting and tissue properties. The virtual policy achieving 99% success was transferred to the physical system, attaining a 93% success rate on physical models, indicating manageable sim-to-real gap.

Policy transfer employed an adaptive residual layer, compensating for sim-to-real gaps via continuous PID refinement. Fig. 7 depicts the integrated DRL workflow: state inputs (including segmentation/depth data) inform action selection through DQN path planning and PPO policy generation. Executed actions modify the environment, yielding new states and rewards that continuously refine the control strategy via policy updates, forming a closed learning loop.

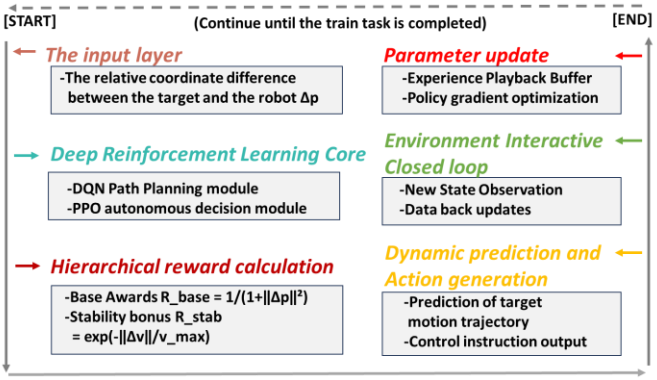

Figure 7. Flow chart of DRL for control algorithm and driving system.

To verify the effectiveness of the control algorithm and drive system optimization based on DRL, we designed and conducted in vitro model experiments. The results show that the proposed method is superior to the traditional PID control regulating end-effector position and visual servo control in tracking accuracy, response delay, and motion stability. The root means square error (RMSE) decreased from 3±1 mm to 1±0.5 mm, a 66.7% improvement; and movement stability increased by 43%, with a velocity variance $\sigma_v^2 < 0.02\text{mm}^2/\text{s}^2$. Additionally, the method achieved a 37.2% improvement in trajectory matching within a 3-second prediction window, demonstrating its superiority in dynamic target tracking. The algorithm's computation speed reaches above 25fps, with a processing time of $\leqslant$40 ms/frame, meeting real-time imaging processing requirements and indicating its feasibility for clinical application.

E. Tracking test experiment

We systematically assessed the robotic ureteroscope's tracking performance through phased experiments under controlled and anatomically realistic conditions. The platform was initialized prior to testing, with simulated calculi positioned at designated starting points and the robot configured to a standardized initial posture to maintain experimental consistency. Motion trajectories for calculi were programmed to include linear, curved, and randomized paths at velocities ranging from 1 to 10 mm/s, replicating the dynamic movement patterns observed in clinical settings. A visual monitoring system continuously recorded positional data of both target calculi and robotic components, enabling subsequent analysis of critical metrics including tracking error, operational speed, and success rate. The experimental configurations—from initial position calibration (Fig. 8-A) and in vitro algorithm performance (Fig. 8-B) to schematic representations of in vitro tracking (Fig. 8-C), simulated

human model tests (Fig. 8-D), and multi-scene complex condition validations (Fig. 8-E). Experimental results urinary tracking success rate (57/60 successful lock-ons), outperforming conventional algorithms' 86.7% success rate (52/60). To evaluate performance in complex environments, a physiologically accurate urinary system model was developed using silicone materials derived from human CT data, incorporating anatomical structures, bubble generation, fluid dynamics, and motion simulation. This environment replicated surgical challenges such as respiratory-induced tissue movement and intraoperative bubble interference. Under these rigorous conditions, the system maintained a 91.7% tracking success rate (55/60) with control latency consistently below 0.4 seconds.

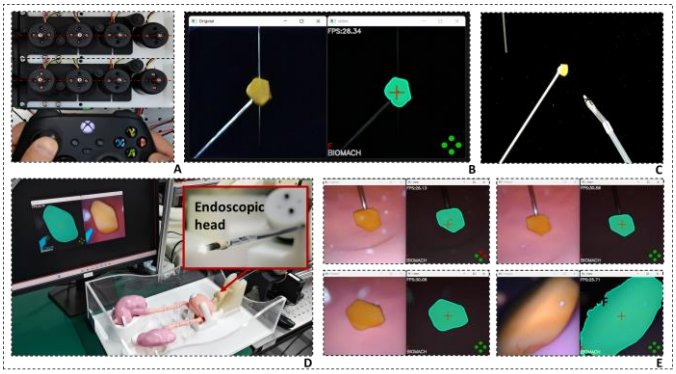

Figure 8. Target Tracking Experiments.(A) Initial position setting and Angle starting point calibration. (B) Comparison before and after processing of in vitro tracking test algorithm (C) Schematic diagram of in vitro tracking test (D) Schematic diagram of tracking test in simulated human model (E) Comparison before and after processing of multi-scene complex conditions tracking test algorithm in simulated human model.

III. DISCUSSION AND CONCLUSION

This study successfully developed and validated an autonomous robotic ureteroscopic system for continuous renal calculi tracking during lithotripsy procedures. The system integrates three key technological advancements: a high-precision robotic ureteroscope capable of sub-millimeter positioning accuracy (±0.8 mm) and flexible navigation (±180° bending); a robust visual processing framework using an enhanced Quenching-net CNN, achieving state-of-the-art calculi segmentation accuracy; and a deep reinforcement learning (DRL) control strategy enabling 93% success in tracking randomly moving calculi. Experimental validation across *in vitro* models and anatomically realistic simulations with physiological disturbances demonstrated the system's ability to maintain calculi alignment within the visual field, significantly reducing manual intervention. This work establishes a critical foundation for autonomous ureteroscopic lithotripsy, with the potential to mitigate surgeon fatigue, minimize cognitive targeting bias, and enhance procedural precision through embodied AI—all while preserving essential human oversight for critical decisions.

Notwithstanding these achievements, several limitations warrant attention. The current system lacks functional integration with lithotripsy lasers, restricting its capability to target tracking rather than autonomous stone fragmentation. Furthermore, while the center-tracking control strategy effectively maintains calculi visibility, and it falls short of the demands for actual lithotripsy. A more sophisticated approach is required—one that incorporates pre-scanning of calculi morphology to determine optimal laser fragmentation trajectories, maximizing energy efficiency and minimizing procedural time. Additionally, insufficient accuracy in 3D depth measurement poses challenges for precise laser focusing on calculi surfaces, potentially reducing energy utilization efficiency and increasing risks of collateral tissue injury.

The system's robustness, though validated under simulated physiological disturbances (e.g., bubbles, respiratory motion), requires further evaluation against unstandardized clinical variations observed in real patient anatomies. Unexpected tissue morphology, bleeding, or irrigation turbidity may impact performance, necessitating domain adaptation techniques and real-time uncertainty quantification modules for reliable clinical deployment.

To address these limitations, future research will prioritize several critical directions. Firstly, the integration of marker less tissue tracking technology will be explored. This advancement is anticipated to enhance the robot's capability for tissue recognition and tracking, facilitating more precise surgical maneuvers without the dependency on external markers[33]. Secondly, the development of fully autonomous lithotripsy functionality will be pursued. Through integration of lithotripsy lasers with the robotic system and the refinement of lithotripsy algorithms, the robot could autonomously determine the optimal lithotripsy strategy by analyzing calculi characteristics such as size, shape, and location [34]. Thirdly, the development of a 3D vision system with a more comprehensive and spatially accurate field of view would be investigated. These advancements will collectively bridge the gap between autonomous tracking and clinically viable autonomous lithotripsy.

IV. MATERIALS AND METHODS

A. Imaging System

A CMOS endoscopic camera with a diameter of 2 mm, resolution of 1920×1080, frame rate of 30fps, and a field of view of 120° was selected for this study. It can clearly capture the fine structures within the lumen. The integrated six-channel white light LED ring light source, with a stable color temperature of 5500K and illuminance adjustable between 50-5000lux, provides shadow-free lighting through fiber optic bundles, ensuring uniform and sufficient illumination in complex lumen environments.

B. High-level Control Strategy

At the beginning of the procedure, the surgeon inserts the head of the robotic ureteroscope into the entrance of the natural lumen. The flexible head and propulsion/steering enable the robot to reach its target position smoothly. Intelligent vision systems constantly detect objects in their field of view. Once a target is detected, the robot immediately activates its deep learning-based target locking mechanism. The imaging system uploads the captured image data to the robot vision system at a high rate. It can quickly segment the target contour and calculate the geometric center, calculate the relative displacement deviation between the robot head and the geometric center of the calculi, and provide key data support for the robot motion control. According to the relative displacement deviation data provided by the vision system, the

optimal motion control scheme is quickly planned, and the motor is driven to realize the accurate tracking of the target by the robot ureteroscope. The position and attitude adjustments of the robotic ureteroscope during the tracking process are fed back to the imaging system in real time, forming an efficient closed-loop tracking loop to ensure that the target remains within the precise control range of the robot. To ensure surgical safety, the system has multiple redundant emergency stop mechanisms. The surgeon or monitor can quickly terminate or resume robotic surgery at any stage via a console or panic button, significantly reducing surgical risks.

C. Low-level Control Logic

When the target is successfully detected and locked, the robot's tracking system is activated to drive the robot head to maintain relative positioning with the target. During the tracking process, the robot system analyzes the dynamic changes of the target in the complex environment and makes continuous and fine adjustments to the motor. The system collects motor current, voltage, position and other feedback signals in real time to ensure the stability of tracking. The system's preset emergency stop mechanism features both hardware and software redundancy, allowing surgeons or monitors to quickly interrupt or resume the automated surgical process at any moment based on actual conditions via hardware buttons or software commands. This provides comprehensive and reliable surgical safety.

D. Statistical Analysis

We implemented a rigorous systematic data collection and statistical analysis plan. Data collection covered several key performance indicators, including tracking success rate, operating time, frequency of robot movement parameter adjustments.

ACKNOWLEDGMENT

This work is supported in part by National Natural Science Foundation of China (62203333 and 52475309) and Shanghai Pilot Program for Basic Research. The authors sincerely thank Yutong Sun, Huiwen Duan, Yan Li, and Yuanzhi Xu for their contributions to data collection, as well as the model debugging and experiments.

REFERENCES

[1] G. Zeng *et al.*, "Prevalence of kidney stones in China: An ultrasonography based cross-sectional study," *BJU International*, vol. 120, no. 1, pp. 109–116, Jul. 2017, doi: 10.1111/bju.13828.

[2] W. Zhu *et al.*, "Dietary vinegar prevents kidney stone recurrence via epigenetic regulations," *eBioMedicine*, vol. 45, pp. 231–250, Jul. 2019, doi: 10.1016/j.ebiom.2019.06.004.

[3] J. K. Kim *et al.*, "Recent surgical treatments for urinary stone disease in a korean population: National population-based study," *Int J Urol*, vol. 26, no. 5, pp. 558–564, May 2019, doi: 10.1111/iju.13928.

[4] S. Doizi and O. Traxer, "Flexible ureteroscopy: Technique, tips and tricks," *Urolithiasis*, vol. 46, no. 1, pp. 47–58, Feb. 2018, doi: 10.1007/s00240-017-1030-x.

[5] K. A. Healy, R. W. Pak, R. C. Cleary, A. Colon-Herdman, and D. H. Bagley, "Hand problems among endourologists," *J Endourol*, vol. 25, no. 12, pp. 1915–1920, Dec. 2011, doi: 10.1089/end.2011.0128.

[6] W. W. Ludwig, G. Lee, J. B. Ziemba, J. S. Ko, and B. R. Matlaga, "Evaluating the ergonomics of flexible ureteroscopy," *J Endourol*, vol. 31, no. 10, pp. 1062–1066, Oct. 2017, doi: 10.1089/end.2017.0378.

[7] J. Y. Lee and S. H. Jeon, "Robotic flexible ureteroscopy: A new challenge in endourology," *Investig Clin Urol*, vol. 63, no. 5, pp. 483–485, Sep. 2022, doi: 10.4111/icu.20220256.

[8] J. Rassweiler, M. Fiedler, N. Charalampogiannis, A. S. Kabakci, R. Saglam, and J.-T. Klein, "Robot-assisted flexible ureteroscopy: An update," *Urolithiasis*, vol. 46, no. 1, pp. 69–77, Feb. 2018, doi: 10.1007/s00240-017-1024-8.

[9] P. F. Müller, D. Schlager, S. Hein, C. Bach, A. Miernik, and D. S. Schoeb, "Robotic stone surgery – current state and future prospects: A systematic review," *Arab Journal of Urology*, vol. 16, no. 3, pp. 357–364, Sep. 2018, doi: 10.1016/j.aju.2017.09.004.

[10] C. Schlenk *et al.*, "A robotic system for solo surgery in flexible ureteroscopy: Development and evaluation with clinical users," *Int J CARS*, vol. 18, no. 9, pp. 1559–1569, Apr. 2023, doi: 10.1007/s11548-023-02883-5.

[11] J. Li *et al.*, "The design of ureteral renal interventional robot for diagnosis and treatment," in *Advances in Neural Networks – ISNN 2018*, T. Huang, J. Lv, C. Sun, and A. V. Tuzikov, Eds., Cham: Springer International Publishing, 2018, pp. 711–718. doi: 10.1007/978-3-319-92537-0_81.

[12] J. Park *et al.*, "The usefulness and ergonomics of a new robotic system for flexible ureteroscopy and laser lithotripsy for treating renal stones," *Investig Clin Urol*, vol. 63, no. 6, pp. 647–655, Nov. 2022, doi: 10.4111/icu.20220237.

[13] B. H. Kann, A. Hosny, and H. J. W. L. Aerts, "Artificial intelligence for clinical oncology," *Cancer Cell*, vol. 39, no. 7, pp. 916–927, Jul. 2021, doi: 10.1016/j.ccell.2021.04.002.

[14] Z. A. Stoebner, D. Lu, S. H. Hong, N. L. Kavoussi, and I. Oguz, "Segmentation of kidney stones in endoscopic video feeds," in *Medical Imaging 2022: Image Processing*, I. Išgum and O. Colliot, Eds., San Diego, United States: SPIE, Apr. 2022, p. 121. doi: 10.1117/12.2613274.

[15] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust object detection under occlusion with context-aware CompositionalNets," 2020, *arXiv*. doi: 10.48550/ARXIV.2005.11643.

[16] Z. Zhang and M. Hoai, "Object detection with self-supervised scene adaptation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 21589–21599. doi: 10.1109/CVPR52729.2023.02068.

[17] L. Jing *et al.*, "Depth estimation matters most: Improving per-object depth estimation for monocular 3D detection and tracking," in *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA: IEEE, May 2022, pp. 366–373. doi: 10.1109/ICRA46639.2022.9811749.

[18] J. Yu and H. Choi, "YOLO MDE: Object detection with monocular depth estimation," *Electronics*, vol. 11, no. 1, p. 76, Dec. 2021, doi: 10.3390/electronics11010076.

[19] Z. Xu and Y. Jia, "Detection and depth estimation for objects from single monocular image," in *Proceedings of 2020 Chinese Intelligent Systems Conference*, vol. 705, Y. Jia, W. Zhang, and Y. Fu, Eds., in Lecture Notes in Electrical Engineering, vol. 705. , Singapore: Springer Singapore, 2021, pp. 27–35. doi: 10.1007/978-981-15-8450-3_4.

[20] G. Srimathveeravalli, T. Kesavadas, and X. Li, "Design and fabrication of a robotic mechanism for remote steering and positioning of interventional devices," *Int J Med Robot*, vol. 6, no. 2, pp. 160–170, Jun. 2010, doi: 10.1002/rcs.301.

[21] Q. Zhang, X. Nie, J. Peng, Y. Wei, M. Li, and Y. Tang, "Addressing the 'world map problem' for submillimeter medical robots requires advancing local sensing beyond external imaging dependence," *Device*, p. 100839, Jun. 2025, doi: 10.1016/j.device.2025.100839.

[22] X.-H. Zhou *et al.*, "A multilayer and multimodal-fusion architecture for simultaneous recognition of endovascular manipulations and assessment of technical skills," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2565–2577, Apr. 2022, doi: 10.1109/TCYB.2020.3004653.

[23] X. Ma, J. Zhou, X. Zhang, Y. Qi, and X. Huang, "Design of a new catheter operating system for the surgical robot," *Appl Bionics Biomech*, vol. 2021, p. 8898311, 2021, doi: 10.1155/2021/8898311.

[24] O. M. Omisore, T. Akinyemi, W. Duan, W. Du, and L. Wang, "A novel sample-efficient deep reinforcement learning with episodic

policy transfer for PID-based control in cardiac catheterization robots," Oct. 28, 2021, *arXiv*: arXiv:2110.14941. doi: 10.48550/arXiv.2110.14941.

[25] L. Karstensen *et al.*, "Learning-based autonomous vascular guidewire navigation without human demonstration in the venous system of a porcine liver," *Int J Comput Assist Radiol Surg*, vol. 17, no. 11, pp. 2033–2040, Nov. 2022, doi: 10.1007/s11548-022-02646-8.

[26] S. Gangwar, R. Devi, and N. A. Mat Isa, "Optimized exposer region-based modified adaptive histogram equalization method for contrast enhancement in CXR imaging," *Sci Rep*, vol. 15, no. 1, p. 6693, Feb. 2025, doi: 10.1038/s41598-025-90876-6.

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv*. doi: 10.48550/ARXIV.1709.01507.

[28] Z. Yang, J. Shao, and Y. Yang, "An improved CycleGAN for data augmentation in person re-identification," *Big Data Research*, vol. 34, p. 100409, Nov. 2023, doi: 10.1016/j.bdr.2023.100409.

[29] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Sci. Robot.*, vol. 7, no. 66, p. eabm6074, May 2022, doi: 10.1126/scirobotics.abm6074.

[30] Z. Xu, Y. Wang, and T. Wang, "Research on lightweight object detection algorithm for hand detection," in *Proceedings of the 2024 6th International Conference on Telecommunications and Communication Engineering*, Chengdu China: ACM, Nov. 2024, pp. 40–44. doi: 10.1145/3705391.3705398.

[31] H. Qin, T. Meng, K. Chen, and Z. Li, "A comparative study of DQN and D3QN for HVAC system optimization control," *Energy*, vol. 307, p. 132740, Oct. 2024, doi: 10.1016/j.energy.2024.132740.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv*. doi: 10.48550/ARXIV.1707.06347.

[33] W. Chi *et al.*, "Trajectory optimization of robot-assisted endovascular catheterization with reinforcement learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid: IEEE, Oct. 2018, pp. 3875–3881. doi: 10.1109/IROS.2018.8593421.

[34] "Deep reinforcement learning for guidewire navigation in coronary artery phantom | IEEE journals & magazine | IEEE xplore." Accessed: Aug. 28, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9648308