



Autonomic Robotic Ultrasound Imaging System Based on Reinforcement Learning

Guochen Ning, Xinran Zhang , and Hongen Liao , *Senior Member, IEEE*

Abstract—Objective: In this paper, we introduce an autonomous robotic ultrasound (US) imaging system based on reinforcement learning (RL). The proposed system and framework are committed to controlling the US probe to perform fully autonomous imaging of a soft, moving and marker-less target based only on single RGB images of the scene. **Methods:** We propose several different approaches and methods to achieve the following objectives: real-time US probe controlling, soft surface constant force tracking and automatic imaging. First, to express the state of the robotic US imaging task, we proposed a state representation model to reduce the dimensionality of the imaging state and encode the force and US information into the scene image space. Then, an RL agent is trained by a policy gradient theorem based RL model with the single RGB image as the only observation. To achieve adaptable constant force tracking between the US probe and the soft moving target, we propose a force-to-displacement control method based on an admittance controller. **Results:** In the simulation experiment, we verified the feasibility of the integrated method. Furthermore, we evaluated the proposed force-to-displacement method to demonstrate the safety and effectiveness of adaptable constant force tracking. Finally, we conducted phantom and volunteer experiments to verify the feasibility of the method on a real system. **Conclusion:** The experiments indicated that our approaches were stable and feasible in the autonomic and accurate control of the US probe. **Significance:** The proposed system has potential application value in the image-guided surgery and robotic surgery.

Index Terms—Automatic ultrasound imaging, deep reinforcement learning, robotic ultrasound system.

I. INTRODUCTION

MEDICAL robots have been developed and applied in many fields such as minimally invasive surgery, image-guided surgery and rehabilitation [1], [2]. In general, clinical ultrasound scanning relies on manual navigation of the probe

Manuscript received September 2, 2020; revised November 24, 2020 and January 14, 2021; accepted January 21, 2021. Date of publication January 26, 2021; date of current version August 20, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 82027807 and 81771940 and in part by the National Key Research and Development Program of China under Grant 2017YFC0108000 (Corresponding author: Hongen Liao.)

Guochen Ning and Xinran Zhang are with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, China.

Hongen Liao is with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China (e-mail: liao@tsinghua.edu.cn).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TBME.2021.3054413>, provided by the authors.

Digital Object Identifier 10.1109/TBME.2021.3054413

like ultrasound (US)-guided endovascular navigation [3] and abdominal aorta imaging [4]. Operator experience is the important factor in affecting the quality and efficiency of manually operated ultrasound imaging systems and may lead to high inter-operator variability and potential instability [5]. In contrast, robotic ultrasound systems have better stability and flexibility, especially in a long-term ultrasound scanning process [6], [7]. Medical robots also depend on multiform high-precision information. Among medical robotic instruments, automatic US imaging systems have been widely studied. Ultrasonic imaging is a popular and fast noninvasive imaging method, and its automated operation is of great help to improve the imaging efficiency. Especially in robotic surgery, accurate and automatic control over the movement of the US probe is the key point of the automatic ultrasound system. Generally, the accuracy of the US probe' position and stable ultrasound probe control are important factors of image quality. After the target position is obtained, an appropriate contact force between the US probe and the target affect both the imaging quality and the safety of the system [8]. In addition, the movement of the target and the surface deformation also bring challenges to the accurate control of the ultrasound probe. All these demands depend on high-precision information acquisition devices, such as 3D cameras and precise force sensors.

At present, multiple control models have been used in some clinical robotic US systems. Most of these methods are based on waypoints and marker-based automatic control methods [9], [10]. 3D scene reconstructed-based path planning and marker-based object pose estimation are typical methods. Based on this mode, previous robot ultrasonic systems used the 3D camera to collect and reconstruct the surface of the scene and planned the robot movement path according to the analyzed surface [6]–[11]. Some researches combined force sensors and adjusted the contact force from the movement of the US probe [12], [13]. These control methods have several associated parts and the relevance of each part needs to match a specific task very well. In addition to the intraoperative information, preoperative information has also been integrated into robot control methods. In the researches discussed in [14] and [15], preoperative high-precision images like magnetic resonance imaging (MRI) and computerized tomography (CT) were registered with intraoperative US images and visual markers to guide the movement of the US probe and the robot. Also based on tracking implementation, Chatelain *et al.* used a robotic arm and a 3 degree of freedom (DOF) visual serving to keep the needle in the center the of US image [16]. These studies analyze the reconstructed environment

or visual markers and automatically extract artificially defined features from the 3D visual information. However, the scene-reconstruction-based approach has a general occlusion problem, especially in marker-based method. Meanwhile, the accuracy and efficiency of these methods depend significantly on the accuracy and speed of 3D imaging equipment [17]. Therefore, the accuracy of the acquisition device has become one of the limitations of the accuracy of the robotic ultrasound system. Also, the contact force and the movement of the target have not been fully considered.

The clinical US imaging generally faces the problem of interference from target movement. To compensate for the baseline drift caused by target movement, some studies have realized the tracking of moving surfaces or organs by designing special mechanical structures and control methods. Sasaki *et al.* proposed a compact portable US diagnostic robot for home healthcare based on image matching and robot control [18]. Similarly, Bowthorpe *et al.* proposed a predictive feedback control scheme for image-guided beating-heart surgery based on a Smith predictor [19]. Seo *et al.* proposed a slave robot based on the rotary Stewart platform to perform real-time US imaging [20]. There are also some researches that controlled the US probe based on the ultrasound image information [21]. These works compensated for the errors caused by the target movement at some level, but application scenarios are targeted and imitated. A robust, autonomous, and safe controlling method is important for robotic US imaging.

The interaction between the robot and the environment is complex because of the rich contact between the US probe and the skin. The force control of the ultrasound probe affects not only the ultrasound imaging results, but also the safety of the robotic ultrasound system. The conventional force control methods include hybrid position and force control [22], impedance control [23] and admittance control [24]. These methods integrate force control and position control by analyzing the dynamic relationship between the end of the robotic arm and the environment and use the same strategy to achieve force control and position control [22]. Thus, prior knowledge of the environment is important in the controller's parameters defining to properly match the environment and the task [25]. Existing researches have proposed many controller designs for overcoming the unknown information of the environment by adjust the parameters of the controller such as fuzzy control [26], adaptive variable control [27] and reinforcement learning [28]. On the contrary, we are committed to adjusting the contact force of the robot directly through interaction with observation rather than adjusting the stiffness of the robot.

Deep reinforcement learning (DRL) combines deep learning (DL) and reinforcement learning (RL) and has proven the feasibility of this approach in complex decision-making tasks [29]. Further studies have shown the potential of RL to learn controlling directly from observations [30]. In the research of mechanical device control, RL has been applied in some basic scenes including motor skill [31], vehicle control [32] and planar arm movement [33]. In complex decision-making tasks like door-opening [34] and cap twisting [35], RL-based methods demonstrated excellent control effects. Kumar *et al.* proposed DRL algorithms that achieved success in robotic grasping tasks

by learning manipulation of real robots from simulation [36]. In the medical field, the US of RL-based methods to control medical devices has already been considered and reviewed. Yu *et al.* proposed a robotic auto-focus system based on a deep Q-learning model [37]. Tan *et al.* proposed robot-assisted training in laparoscopy based on DRL [38]. These studies showed that complex robot control can be realized by DRL based on simple and efficient state. Compared with the artificially defined feature extraction based on traditional or deep learning methods in robotics US imaging control, RL avoids the uncertainty of manually defined states. Therefore, RL is considered a potential solution for an automatic ultrasound imaging task.

This article introduces a fully autonomous robotic US imaging system to improve US imaging efficiency and stability. First, an RL agent-based proximal policy optimization (PPO) method is utilized to generate robot commands based on predefined reward functions and scenes [39]. The proposed robotic US imaging method is different from the robotic US imaging methods of conventional systems that attempt to estimate surface poses or plan action path. We use a single RGB image of the scene as the only observation for the system. The force state and US image state are also important factors when describing the state of the system that cannot be observed from the RGB camera. Also, for further reducing the dimension of the scene, the proposed method comprises a state representation learning (SRL) model to reduce the invalid information in the image and encode the invisible information into the latent space [40], [41]. In the part of the US probe adaptable constant force tracking and accurate force controlling, we put forward the idea of force-to-displacement (F2D) method. The principle of F2D is to set the contact force as the action space of the end-effector and map the contact force to the lowest level robot command. The main contributions of this article are:

- 1) We propose an RL-based robotic US imaging system that does not require target pose estimation and path planning. The proposed method directly outputs the action of the ultrasound probe based on a single RGB image via the state representation model and RL model. The purpose of the system is to make automatic ultrasound imaging a flexible, marker-less and randomly located target in a disturbing environment. Feasibility evaluation and comparison with conventional scene reconstruction methods will be performed.
- 2) To obtain a stable US image on a soft target, we propose the F2D method to control the US probe constantly contacting the target with a safe and proper force. The purpose of the method is to realize adaptable constant force tracking of the US probe on an unstable and soft target. We will compare the effectiveness between the F2D and the conventional methods especially in such contact-rich scenes.

II. METHODS

The robotic ultrasound imaging system aims to improve the efficiency of the ultrasound imaging process. With reference to manual ultrasonic imaging operation, the system needs to

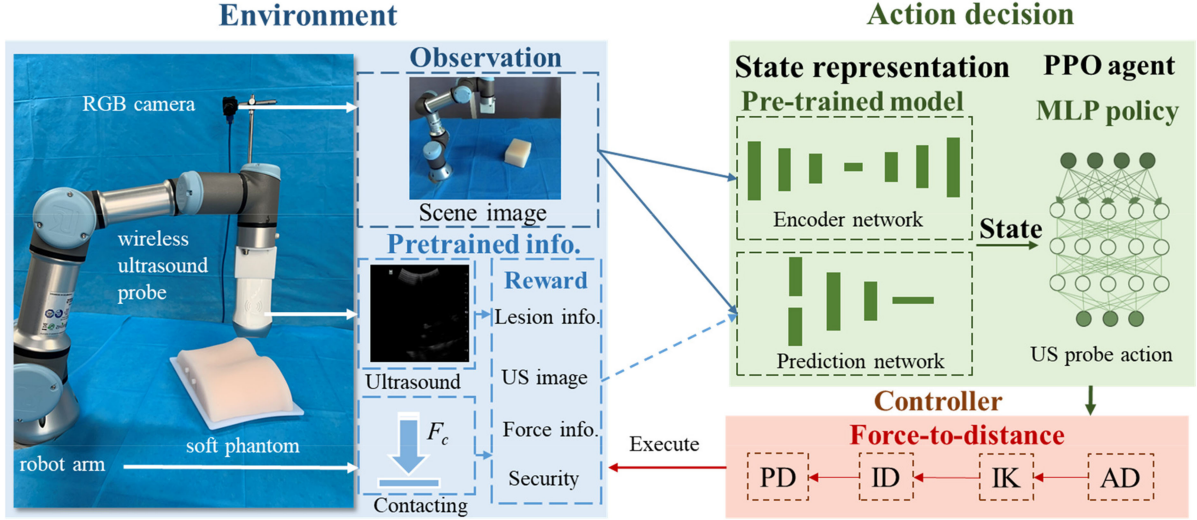


Fig. 1. The proposed methods and framework for automatic ultrasound imaging system including three parts: environment (blue), action decision model (green) and robot controller (red). The environment including all hardware and information including RGB images, ultrasound images and force information. Before the input of the observation for the reinforcement learning model to generate action commands, the ultrasound images and force information are encoded into the RGB image space by the state representation model and the RGB image is set as the only observation. The action commands are converted by the proposed force-to-distance controller and input to the robot to execute the ultrasound imaging actions.

perform real-time action decision-making and control the US probe fully automatic. In the manual operation process, the state of the ultrasound image is the criterion for evaluating the ultrasound imaging task rather than the target coordinates or contact force.

To realize the autonomous robotic ultrasound imaging system, the focus of our work is on the action decision-making and the control method of the US probe. The proposed robot US imaging system consist of three parts, as illustrated in Fig. 1. The environment part contains all the hardware devices including an RGB camera, a US probe, an integrated force sensor and a lightweight robot. The method part contains the reinforcement learning model and the state representation model for action decision making. The controller part consist of four command mapping processes to achieve adaptable constant force tracking of the US probe. Different from the conventional methods, no coordinate points are collected in the system based on this idea, which means the dependence of on high-precision collection equipment is avoided in our system. In this section, the proposed robot US imaging method will be introduced and explained in the following order: Ultrasound imaging environment and reinforcement learning, state representation learning for invisible information encoding, force-to-displacement adaptable constant force tracking method and training details in simulator and real system.

A. Robotic Ultrasound Imaging System and Reinforcement Learning

As mentioned above, the environment is defined as the interaction between the robotic ultrasound device and the target. A standard RL setup is posed in the framework of the Markov decision process (MDP), which can be defined by \mathcal{S} , \mathcal{A} , p , \mathcal{R} and γ [42]. \mathcal{S} is the sets of states, \mathcal{A} is the continuous action

space, $p(s_{t+1}|s_t, a_t)$ is the stochastic dynamics between states for a given action, $\mathcal{R}(s, a) = r \in \mathbb{R}$ is the reward function and $\gamma \in [0, 1]$ is the discount factor. The principle of RL is to train an agent following the policy $\pi(a|s)$ which can maximize the expected reward r by selected an action a_t based on the interaction with current observations o_t at time step t . The policy is parameterized by θ and defined as $\pi_\theta(a|s)$. Herein, we recognize the robotic ultrasound imaging task as a discrete-time continuous MDP. The aim of the task is to control the US probe and obtain the US images automatically from a random located target. At time step t , the robot makes an action based on current observation to obtain a higher reward value. Following this setup, a policy gradient theorem is:

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_s p_\pi(s) \sum_a Q_{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a|s) \\ &= \mathbb{E}_\pi [\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s, a)] \end{aligned} \quad (1)$$

where $p_\pi(s) = \sum_t \gamma^t p(s_t = s | s_0, \pi)$ and $Q_{\pi_\theta}(s, a)$ is the action-value function. Based on this theorem, θ is optimized to maximize the expected return in a policy gradient algorithm:

$$\theta^* = \arg \max_\theta E \left[\sum_t \gamma^t r_t [s_0, \pi] \right] \quad (2)$$

Considering the limitations of the dataset in the environment, a policy gradient algorithm-based algorithmic solution, PPO agent, is trained to maximize the loss function [39]:

$$J_t^{CLIP'}(\theta, \theta') = \mathbb{E} [J_t^{CLIP}(\theta) - c_1 J_t^{VE}(\theta') + c_2 H(\pi_\theta | s_t)] \quad (3)$$

where c_1 , c_2 are weights of loss, and $H(\pi_\theta | s_t)$ is the entropy term. $J_t^{VE}(\theta')$ is the error term on the value estimation with discount factor γ and target value function to encourage

sufficient exploration with $H(\pi_\theta|s_t)$:

$$J^{VE}(\theta') = (V_{\theta'}(s_t) - (r(s_t, a_t) + \gamma V_{\theta_{target}}(s_{t+1})))^2 \quad (4)$$

$J_t^{CLIP}(\theta)$ is the loss that is limited by a clipped ratio ϵ to stabilize the update procedure:

$$J_t^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t(s, a), \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t(s, a) \right) \right] \quad (5)$$

where $\hat{A}_t(s, a)$ is the advantage estimator and $r_t(\theta) = \pi_\theta(a|s)/\pi_{\theta_{old}}(a|s)$ is the probability ratio between current policy and old policy. In the environment, the observation is the visual information that is directly composed of a $256 \times 256 \times 3$ RGB image. The high-dimensional information of the scene can be contained from the scene image, including robot posture, target position and end-effector position. Moreover, the combination of US images and force information represents effective information related directly to the task. Based on this information, the agent will maximize the reward in training according to a parameterized policy. Since the dimensions of each state are different, we built a multi-layer perceptron model as the policy model with three hidden layers.

The task of reaching and scanning a target with a US probe has several related factors. Referring to the manual operation steps in the US imaging, we incorporate position approaching, contact maintaining, and US images into the reward function design. The reward function is the sum of each term with weights. First, the US probe needs to be moved towards the target in space. Therefore, the dense distance reward $\mathcal{R}_{distance}$ is defined as the Euclidean distance between the US probe and the target:

$$\mathcal{R}_{distance} = -\omega_1 \|P_{probe} - P_{target}\| \quad (6)$$

After the US probe is close enough to the target, the main reward component is the sum of \mathcal{R}_{US} and \mathcal{R}_{leison} . \mathcal{R}_{US} is a sparse reward for detecting the existence of US images. Once the US image appears, we reward the policy about the similarity index of the US image I_t^U and the template of the physiological target I_t^O that is evaluated by the mutual information score.

$$\mathcal{R}_{US} = \begin{cases} \omega_2, & \text{if image exist} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

$$\mathcal{R}_{leison} = \omega_3 \sum_{I_t^U} \sum_{I_t^O} p(I_t^U, I_t^O) \log \left(\frac{p(I_t^U, I_t^O)}{p(I_t^U)p(I_t^O)} \right) \quad (8)$$

where p represents the joint probability distribution. \mathcal{R}_{US} and \mathcal{R}_{leison} encourage the US probe to contact the physiological structure related position as correctly as possible.

In the last part of the positive reward, we add an additional reward $\mathcal{R}_{maintain}$ for encouraging continued scanning.

$$\mathcal{R}_{maintain} = \omega_4 \sum_{i=1}^n \mathcal{R}_{USi} \quad (9)$$

where n represent the number of steps in one episode and \mathcal{R}_{USi} represent the value of the \mathcal{R}_{US} in every step.

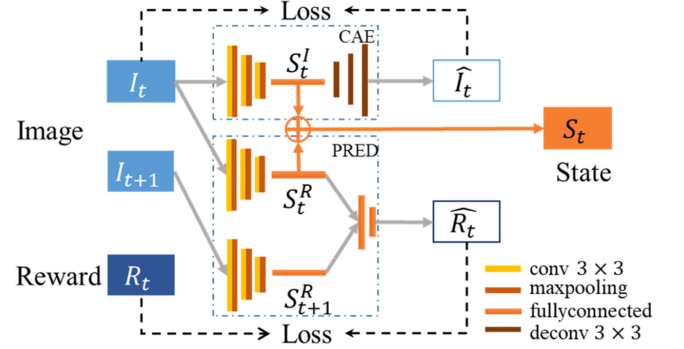


Fig. 2. The framework of the state representation model. In the pre-trained model, the RGB image I_t is encoded by a convolution auto encoder network and the latent space of the model is set as the image state. In another part, a reward prediction network that has the same structure is used to encode the reward-related information. The latent space of the prediction model is set as the rest of the state.

Considering the safety requirements in medical applications, we also constrain the robotic US imaging task from the reward settings. We harshly penalize the agent for applying forces over the threshold force and reaching workspace limits. The episode ends and restarts with these collisions occurring.

B. State Representation and Multiform Information Encoding

As mentioned in the previous section, multiform information in different dimensions is required to describe the state of the task. The RGB image of the scene provides position-related information, and the US image provides task-related information. These high-dimensional states also contain irrelevant information that may cause interference. Transforming the high-dimensional image into the low-dimensional state through the state representation model can effectively reduce the search space and provide more efficient expression information for RL [41].

In our environment, the RGB image is the main high-dimensional state. However, the force information and US information cannot be obtained from RGB images. Moreover, the sparsity of these two parts is different. Since US image state and force state are correspond directly to the positive value of the reward function, we assume these two states can be represented through the combination of the reward function and the RGB image. To effectively reduce the dimension of these three states we combined image auto encoding and the reward prediction model as the state representation model. As shown in Fig. 2, our state representing part is an end-to-end model.

For image representation, we use a convolutional auto encoder (CAE), which consists of an encoding-decoding model for the RGB image reconstruction. The encoder consists of three layers of convolutional neural network and each layer is combined with an activation layer and a max-pooling layer. The latent space output by the encoder is composed of a vector with a dimension of 128. This latent space S_t^I will be reconstructed into the reconstructed image by the decoder. The loss function is cross-entropy and computed between the reconstructed image

\hat{I}_t and the real image I_t . After the training is finished, the latent space S_t^I is set as a part of the current state and the decoding layers are abandoned.

Similar to image encoding process, we use the same encoding model to encode the current image I_t and the next image I_{t+1} , and output two vectors S_t^R and S_{t+1}^R with a dimension of 128. These two vectors will be input into the reward prediction model which consists of three fully connected layers and an active function. The output of the reward prediction model is defined as the predicted value of the reward. Similarly, the error is computed between the predicted reward \hat{R}_t and the real reward R_t . Finally, the image state S_t^I and the reward state S_t^R are concatenated as the current state S_t and the invisible information is represented by the visible RGB image. Before training the state representation model, we will collect data through the RL model training process and the data set contains all the kinds of information mentioned before. After the training completed, the SR model will be integrated as the previous part of the end-to-end RL process.

C. Force-To-Displacement Control Method for Ultrasound Imaging Task

The adaptable constant force tracking of the US probe is paramount in our US imaging method. In some recent researches, there are some common robot controllers like direct torque control, joint space control, impedance control and admittance control. Like the reward function design in the previous section, the robotic ultrasound imaging is associated with both task accomplishment and safety. Especially considering the environment in the contacting between the end-effector (US probe) and soft target, we propose an admittance controller-based force-to-distance control method that is safer and more efficient to such rich-contact and force-based tasks [28]–[43].

The admittance controller dynamically adjusts the characteristics of the end-effector by imitating a spring-damper system. When the robot is in contact with the environment, the admittance is used to describe the characteristics of the robot. By adjusting the three parameters, B and K of the admittance controller change the force or position between the end-effector and the environment:

$$-F_c = M(\ddot{x}_{des} - \ddot{x}) + B(\dot{x}_{des} - \dot{x}) + K(x_{des} - x) \quad (10)$$

where M is the inertial matrix, B is the damping matrix, K is the stiffness matrix, x is the position vector, x_{des} is the desired position and $F_c \in \mathbb{R}^3$ is the desired contact force in three directions. Due to the stiffness and the characteristic of the environment is well known, all parameters of the controller have adjusted manually to match the US imaging task.

Assuming the force signal F_c can be gathered by the force sensor from the interaction of the end-effector and the soft surface, according to the formula above, the 3-D position correction of the end-effector $x_c = x - x_{des}$ can be transformed from the contact force F_c . In this way, the desired contact force can be converted to position correction. To obtain a lower level robot command, the desired position is mapped into the joint coordinates by the inverse kinematics (IK) and the inverse

dynamic (ID) method [44]:

$$\dot{q} = J^T \dot{x} \quad (11)$$

$$\tau = ID(\ddot{q}_{des}) = D(q)\ddot{q}_{des} + C(q, \dot{q})\dot{q}_{des} + G(q) \quad (12)$$

where J is the end effector Jacobian, D is the mass matrix, C is the centrifugal and G is the gravity term. The IK and ID model of the UR3 robot is referenced from [45]. Therefore, we convert the F_c to the command of the motor. Finally, the command is completed by a proportional derivative (PD) controller. We can map the lowest robot command, joints torque ($u = \tau \in \mathcal{T}$), to the desired contact force F_c :

$$u = ID(K_d(\dot{q} - \dot{q}_{des}) + K_p(q - q_{des})) \quad (13)$$

In general, the contact force is generated from the movement of the end-effector. The action space is defined in spatial position. However, the accuracy of the contact force depends on the accurate target position and the characteristics of the target. In our method, the positive reward is related mainly to the contacting force and US images and the position reward is still needed for encouraging movement towards the target. In addition, the correspondence between the movement and the force is not clear for the unknown characteristics of the target. Thus, we put forward force to the displacement (F2D) method which integrates the admittance controller and the force action space. In the iteration of the RL training, the $F_c \in \mathbb{R}^3$ is set as the action space and output by the agent, which means the agent will output a desired contact force between the end-effector and the target. The specific definition of the desired contact force is a vector with a dimension of 3×1 . Each value in the vector represents the desired force in the three directions from -15 to 15 (an empirical value of the contact force), and the symbol of the value indicates the direction of the force. When the end-effector is not in contact with the environment ($F_c' = 0$), the robot performs a moving action ($F_c = m$), as if the end-effector is pulled by an “invisible hand” and moved in space. When the end-effector is contacted by the target ($F_c' \neq 0$), this “external force” will become the contact force and generate a positive reward. Finally, the actual contact force will be equal to the desired contact force ($F_c' = F_c$).

D. Simulation and Hardware System

Considering the huge time-consuming effort of training a real robot, the platform of the system includes the simulator and real hardware. The simulator is used for training models and verifying the proposed methods and processes. The feasibility of the entire framework is verified on the real hardware.

The proposed robotic US imaging system consists of a lightweight robot, an RGB camera, a wireless US scanning device and a built-in force sensor. The US probe (Wireless UProbe ultrasonic device, HengTeng, China) is directly mounted onto the end-effector of the robot and outputs real-time US images to the computer with 20FPS and the resolution of $256 \times 256 \times 3$. The RGB camera is located at a proper position to gather the entire scene. In the system, the coordinates between the camera and the robot are without calibration or registration, which means the position of the camera has no need to be specific as long as the

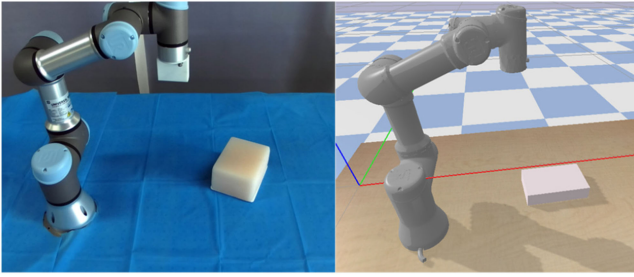


Fig. 3. The simulation environment (right) is built according to the real system (left) to reduce the time consumption in training and evaluation.

robot and the target can be gathered by the camera. To control the US probe, we use a 6-DoF UR3 robot (UR3, Universal robot, Denmark). For robotic US system, since the robotic US platform needs to meet the requirements of clinical usage, the UR3 robot subsystem has been certified for human-machine collaboration due to compliance with functional safety standards. Therefore, it is considered safe to interact with the skin directly. The computing platform includes a CPU (Intel (R) Core (TM) i7-4790K) and a GPU (NVIDIA GeForce 1080 Ti). The robot responds and realizes the commands at the frequency of the output actions of the agent at 120Hz.

In the simulator, we built an environment with the real hardware setup based on the physical simulator PyBullet [46]. As shown in Fig. 3, the virtual camera is located at the same related position as the real hardware setup. To simulate the force sensor which is integrated at the end of the robot, we set the joint torque of the end-effector as the contact force between the end-effector and the environment. The target moves on the desktop randomly in training, and the surface is modeled as a soft material. Due to the surface of the target needs to be coated with ultrasonic couplant, we set the sliding coefficient of friction of the surface as a small value. Because of the internal structures of the target and because the ultrasound images are difficult to model in the simulator, we instead use the relative position of the ultrasound probe to represent the current state of the ultrasound image in the simulator, which is consistent with the objective situation in the real environment. The ultrasound-related reward function in the simulation is represented by the distance between the ultrasound probe and the internal structure of the target. The distance was based on the prior knowledge of the real phantom and the US scanner. The model of the real environment is transferred from the model of the simulator and further trained for US-related optimization.

III. EXPERIMENTS AND RESULTS

In the experiment, we performed quantitative and qualitative evaluation for the automatic US robot imaging system. In the quantitative evaluation, the feasibility experiment for the entire system and framework was evaluated by the success rate and the runtime efficiency of the imaging task. A comparison experiment with a scene reconstruction method was performed to indicate the difference of our method in control accuracy and robustness. Furthermore, the impact of each process and method

in the framework was evaluated including the impact of the state representation on the system training and the stability of the F2D controller for the adaptable constant force tracking of the soft surface. In the qualitative assessment section, the feasibility of the proposed system through the phantom experiments and the volunteer experiment was verified.

A. Feasibility of Autonomous Ultrasound Image Acquisition System

The purpose of the proposed system and methods was to control the US probe to autonomously image a soft target in a changing environment. The RL model was pre-trained with two million steps in the simulator to reduce the training time consumption and performed an additional 100 episodes of training in the real system to further optimize the US-related parameters.

To prove the feasibility of the proposed automatic US imaging system, we evaluated US imaging success rate and efficiency in the real system. In the experiment, three randomly placed soft phantom with different shapes were used as the target. The policy used in different phantom experiments was the same as the policy trained with Phantom 1. When a stable US image of the target was obtained before termination, the task was considered successful. If the US image was not obtained before the end of the movement of the robot or the US probe was out of the workspace or safe force limitation, the task was considered as a failure. In the success rate verification experiment, each marker-less phantom was statically placed in 10 different positions.

To compare with another typical marker-less method-scene reconstruction, we used a state-of-art high-precision stereo camera (Ensensio-N35 stereo camera, IDS, Germany) to perform the conventional US probe-guiding process.¹ The N35 camera was accurately calibrated and registered into the robot's coordinate space. The reconstructed target surface's coordinate was resolved and transformed to the robot for movement. In addition, we added the same anthropogenic occlusion as interference (I) in the two methods to compare the robustness of the framework. A total of 18 sets of results were collected. The detailed comparison from one trial is illustrated in Fig. 4 to indicate the difference intuitively. The ultrasound image of the target was acquired both in clear and interfered environment. As shown in Figure 4, because of the proper relative position and contact force between the US probe and the phantom, the internal structure and the position of the phantom acquired by our method was complete and proper in the US image. In contrast, the improper contact between the US probe and the phantom resulted in lack of completion or even missing of the internal structure in the US image acquired by the N35 system.

To further verify the feasibility and robustness of the proposed framework, we randomly moved the phantom by hand in different directions. The movement range of the phantom was limited in the workspace. One process of the experiment is shown in Movie 1. During the experiment, the phantom was first placed statically on the table. In the beginning of the experiment, the robot moved toward to the phantom and contacted with the

¹[Online]. Available: <https://en.ids-imaging.com/ensensio-n35.html>

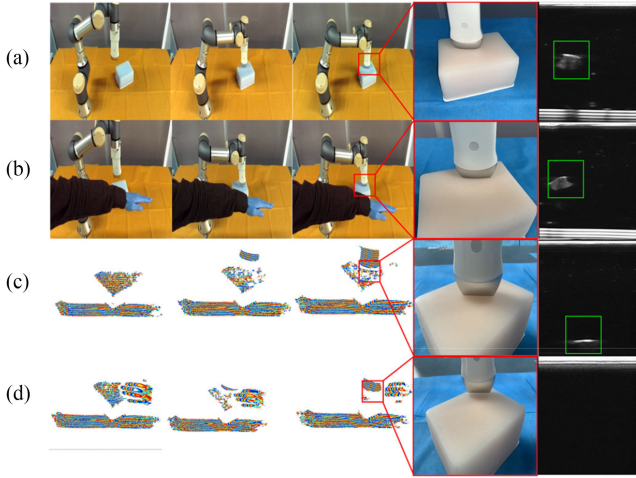


Fig. 4. The comparisons of our method and conventional method in ultrasound imaging task. In the proposed method, the single RGB image was set as the observation. The US probe was controlled by the robot to image the phantom in the environment with (a) and without interference (b). In the same environments (c) (d), the N35 camera was used to obtain the spatial coordinates of the target and guide the movement of the ultrasound probe. The ultrasound probe was moved to the obtained coordinates to perform imaging task.

TABLE I
SUCCESS RATE OF THE ULTRASOUND IMAGING TASK IN
DIFFERENT ENVIRONMENTS

Environment	Success rate		
	Phantom 1	Phantom 2	Phantom 3
Our-Static	100%	90%	90%
Our-I-Static	80%	90%	70%
N35-I-Static	60%	50%	60%
N35-Static	80%	70%	70%
Our-Moved	80%	80%	70%
Our-I-Moved	70%	60%	60%

phantom successfully. Next, the probe was moved on the surface to acquire the US image of the lesion. Then, the phantom was moved by hand and the probe was still in stable contact with the surface. After the US image of the lesion was obtained again, we moved the robot to interfere the system. As shown in the Movie 1, the probe was controlled back to the surface by the robot, and the US image of the lesion was acquired again.

We also performed experiments on the success rate with moving targets. Three different phantoms moved in the same direction at different positions. The criterion to judge the US imaging task was same as for the static experiment and the results of all the success rate experiment are shown in Table I. The success rates of US imaging with and without interruption were evaluated separately, indicating the basic feasibility of the proposed method. Furthermore, in the moving phantom experiments, the success rate was decreased but still maintained a satisfactory value. The main reason for the decrease in the success rate was the separation of the probe from the phantom

surface due to the smooth surface. Fig. 4 showed that the US probe in our method accurately contacted with the target based on the RGB image both with and without interference. The US image of the target was also obtained. The movement speed of the robot was limited within 3 cm/s for safety. In the three sets of static experiments, the time consumption of the automatic ultrasound imaging process was similar to the freehand processes (<20 s).

In contrast, the target coordinates obtained by the scene reconstruction method were largely accurate in the absence of interference. But the contact force control based on the 3D reconstructed method was difficult to be performed accurately, which result in insufficient US imaging quality. In the experiment with interference, the control accuracy of the US probe was decreased because the accuracy and integrity of the reconstruction scene was significantly affected. The ultrasound image could not be obtained. The success rate results indicated that the N35 system could have satisfactory performance only in static interference-free environment. However, in the disturbed environment, the success rate was significantly reduced due to the missing structure of the two overlapping edges at different depths.

B. Evaluation of State Representation Model

In the method evaluation experiment, we were committed to verifying the following two aspects, the efficiency of our method in the US imaging task and whether the SR model provided the agent with task-related information without inputting US state and force as the observation. Each experiment in this section was performed in the simulator. In each part of the experiment, a total of 100 trials were performed and the position of the target was random and recorded. Every model in the experiment performed 1 million steps of training with the same hyperparameters.

To evaluate the efficiency of our method, the ratio between we defined the path of the US probe and the absolute distance between the US probe and the target as the index. The path of the US probe was the sum of the displacement in each step before termination. The targets were placed in different positions in 100 valid trials. The average efficiency of our system was $93.7 \pm 1.9\%$, indicating that the actions and paths output by the agent before completing the US imaging task were efficient.

To evaluate the impact of the SR model, we set up another three pipelines for comparison. **Our method:** the method with complete SR model and the combined latent space was input as the observation. **Pipeline 1:** the method without combining the SR model, and the RGB image were directly inputted into the agent as the observation. **Pipeline 2:** the method had direct force state, US state and the encoded image state as the observation. **Pipeline 3:** the method with only the encoded image state as the observation. Fig. 5 and Table II illustrated the learning curve and success rate achieved by each pipeline in 100 trials. From Fig. 5, the policy with direct image input obtained positive rewards with difficulty, which verified the lack of task-related information reducing the success rate of the rich-contact tasks. Although the learning curve of the policy with the image encoding model had improvement, the positive reward value was still not effectively

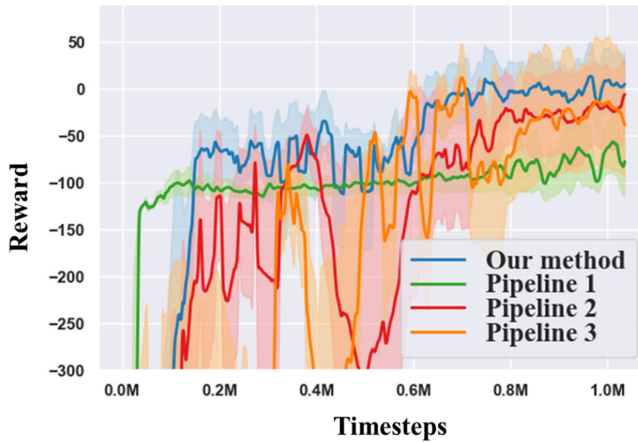


Fig. 5. Comparisons of learning curves for different pipelines in ultrasound imaging task.

TABLE II
SUCCESS RATES OF ULTRASOUND IMAGING TASK WITH DIFFERENT
PIPELINES IN SIMULATION ENVIRONMENT

Method	Success rate
Pipeline 1	63%
Pipeline 2	83%
Pipeline 3	85%
Our method	93%

obtained. Compared with the previous pipeline, Pipeline 2 had not been significantly improved by combining direct force and ultrasound information. Although Pipeline 2 contained direct US state and force state, but these two states were sparse especially at the time of the US probe was in the air. In contrast, our method significantly improved the learning curve and the success rate. Even in the absence of force and US information as direct observations, our policy could achieve a 93% success rate with only RGB images as the observation.

C. Evaluation of Force-to-Distance Method

For US imaging tasks, one of the most important parts of our method is the adaptable constant force tracking of the US probe. In this section, we evaluated the effectiveness and feasibility of the proposed F2D method by comparing the conventional displacement-to-force way. In addition, the stability of the method and the security of the system were evaluated on the real hardware.

Like the framework introduced before, the proposed method converted the output value of the action space as the desired contact force and mapped it to the lowest-level robot commands. A conventional displacement-to-force framework was performed to compare the difference between the proposed action space and the end-effector's position action space in US imaging tasks. In the conventional displacement-to-force framework, the output value of the action space was used as the end-effector's position of the US probe and the actions were directly accomplished

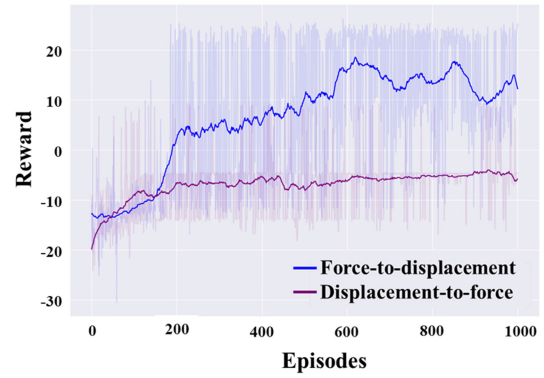


Fig. 6. Comparison of learning curves of force-to-displacement method and displacement-to-force method.

by an inverse kinematics controller. Other conditions of the environment were unchanged.

Figure 6 illustrates the learning curve of two the methods in the real system training process. Because positive rewards were associated with adaptive constant force and US images, stable contact force produced higher policy rewards. The learning curve indicated that the F2D method had better performance in obtaining positive rewards. Since the characteristics of the target were usually uncertain, the mapping relationship from distance to contact force was not clear. This also leads to more time consumption in learning the relationship. In contrast, the proposed method avoided the derivation or acquisition of this relationship, and achieved a better contact effect directly via outputting the contact force.

To quantitative evaluate the performance of the F2D method in adaptable constant force tracking, as shown in Fig. 7. We recorded the force values by the force sensor in the Movie 1.² During the process of imaging, the average value and the standard deviation of the contact force in the z-direction were 11.9 N and 1.7 N. As illuminated in the curve of the force, the contact force between the US probe and the surface was properly stable. Moreover, there was no overshoot of the contact force at each moment of contact. During the moving process in the x, y-direction, the force in z-direction remained unchanged and the force in the x, y-direction indicated that the displacement of the US probe was not driven by friction but was controlled by the robot. During the interruption process, the contact force continued to maintain a safe and stable value after the US probe was back. The force in all three directions remained stable throughout the process even with a slippery soft phantom. The experimental results indicated that the proposed control method has satisfactory performance in the control of the ultrasonic probe in a soft environment.

D. Ultrasound Imaging Acquisition in Volunteer Experiment

To verify the feasibility of the proposed system and framework in human ultrasound imaging tasks, we performed volunteer

²Supplementary movie is available in the supporting documents /Movie 1

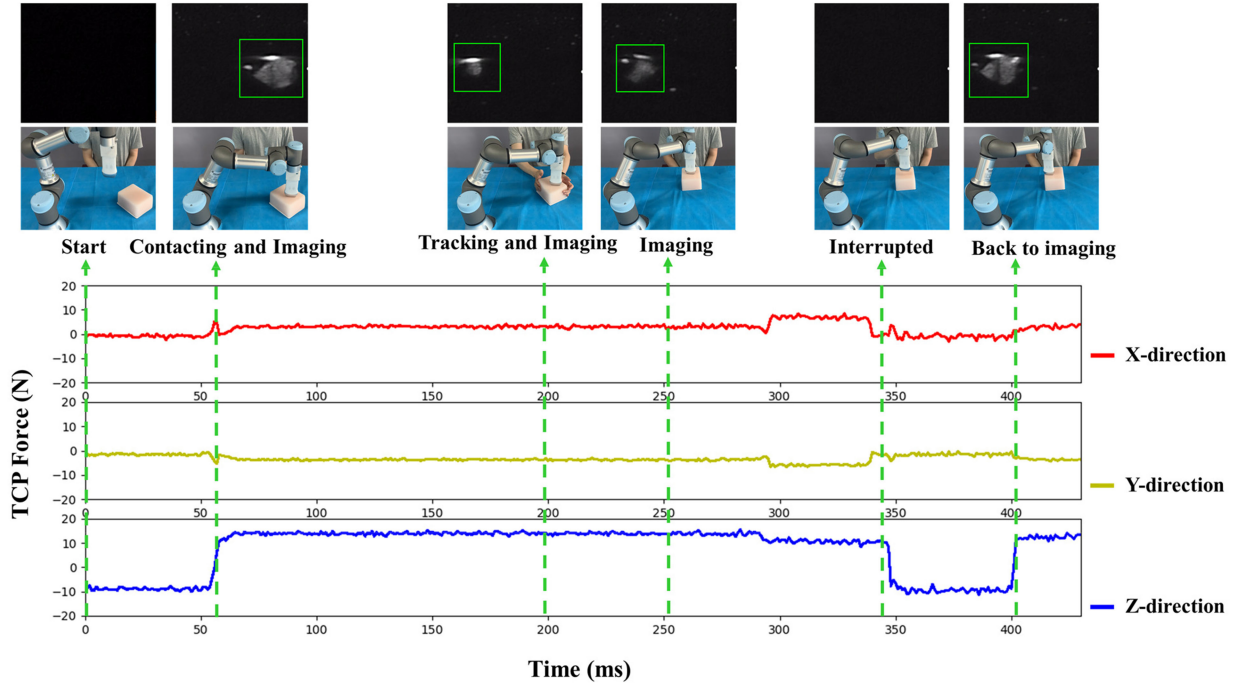


Fig. 7. The curve of contact force in Movie 1. Throughout the ultrasound imaging process, the force remained stable and proper in three directions. In the process where the phantom was moved, the ultrasound probe contacted the surface with a stable z-direction force. The small change value of force in the x, y-direction indicated that the movement of the ultrasound probe was autonomous but not driven by friction.

experiments with the real system. During ultrasound imaging of the human body, the baseline drift caused by breathing was the main factor that affected the control accuracy of the system. Therefore, the abdomen area of the volunteer was set as the testing area for the floating on the surface during breathing. Volunteers lay in unspecified positions and the imaging area was within the workspace of the system. The parameter settings of the US probe and ultrasonic couplant were same as in the phantom experiment. To evaluate the control effect of the probe in the human body imaging process more obviously, volunteers took deep breaths during the experiment to make the abdominal surface change. If the force exceeded the set safety value, the process was terminated and the robot returned to a safe posture.

During the volunteer experiment, there was no large x, y-displacement of the body, so we recorded the change of the position and force of the ultrasound probe in the z-direction. The scene and results of the volunteer experiment are shown in Fig. 8. As shown in the results, irregular movement of the ultrasound probe in the z-direction was caused by the abdominal movement. The average value and standard deviation value of the contact force in the z-direction was 11.3 N and 4.8 N in the US imaging process. The larger standard deviation value indicated the influence of breathing on the contact force. The value of the contact force was temporarily decreased or increased due to the movement of the abdomen caused by the inhalation and exhalation of the volunteer but still maintained a suitable value for US image acquisition. Overall, the results indicated that our robotic ultrasound system could perform stable ultrasound imaging when the volunteer was breathing or moving slightly. To quantitatively evaluate the movement accuracy of the ultrasound

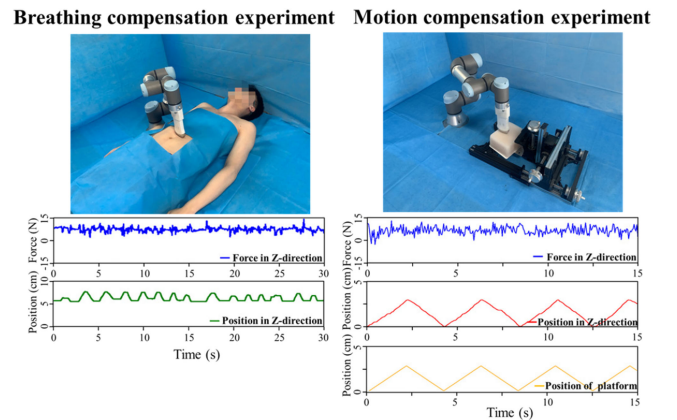


Fig. 8. Scenes and results of the respiratory compensation experiment. The force and position of the ultrasonic probe in the z-direction were recorded. The contact force remained stable in both the breathing and the motion compensation experiment.

probe in the z-direction, the soft phantom was fixed on a lifting platform and moved up and down. As shown in Fig. 8, the average contact force of the ultrasound probe in the z-direction was 12.1 N.

Furthermore, the lumbar spine of the volunteer was set as the target for the automatic ultrasound imaging test. The preoperative lumbar spine image was manually segmented and used as the template for \mathcal{R}_{lesion} to replace the template of the phantom. In the experiment, the RL model was without further training because the imaging area of the volunteer was similar to the phantom used before. When the arbitrary lumbar vertebrae of

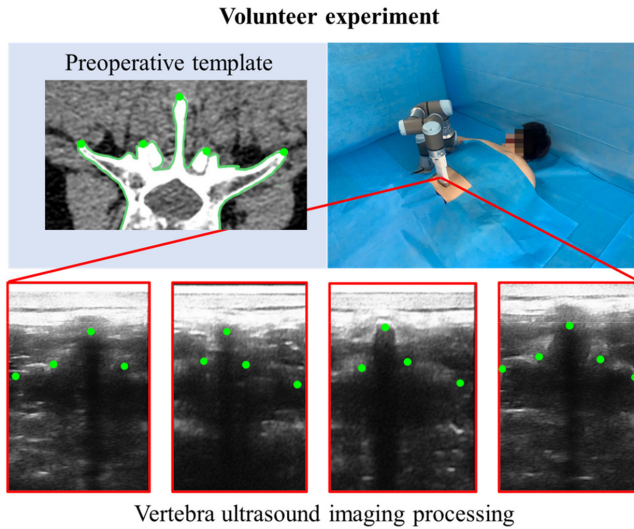


Fig. 9. Scenes and results of the volunteer experiment. The volunteer lumbar spine was set as the ultrasound imaging target. The preoperative lumbar spine image was segmented and used as the template to calculate \mathcal{R}_{lesion} .

the volunteer were imaged before termination, the task was considered successful. The experimental results and scenarios are shown in Fig. 9. As shown in Figure 9, the lumbar spine of the volunteer was in the proper position of the US image in the results of the volunteer experiment. The main structural landmarks match the template and could be clearly identified from the US image. This demonstrated that the ultrasound probe had moved into the correct position and was in contact with the skin with the appropriate contact force

IV. DISCUSSION

We have presented a novel system to achieve automatic ultrasound imaging for a moving and soft target. The system proposed in this article combines the reinforcement learning and state representation models to realize a robust and accurate action output based on a single RGB camera. In terms of the US imaging control method, contrary to the conventional feature extraction method based on traditional segmentation or deep learning that relied on large amounts of existing data and supervised algorithms, the proposed RL based method can learn the control strategy based on exploring the simultaneous feedbacks and maximizing rewards without the requirement of a current data set. In terms of system configuration, the proposed method avoids the reliance on high-precision acquisition equipment in scene reconstruction and pose estimation. Even in the tasks with interference, the proposed method can accurately control the US probe to reach the target position and the accuracy of the US probe control is better than the conventional method to some extent. Moreover, the proposed F2D control method achieves a stable force control of the US probe, especially for the soft target. In the case where the characteristic of the target is unknown, the proposed method outputs a proper adaptable constant contact force for a flexible surface instead of outputting a pre-set force or distance increment.

In the experimental part, the success rate experiment performed has verified the feasibility of the proposed system and method. Compared with the conventional manual defined feature extraction methods, our method obtained 20% improvement in success rate experiment. The qualitative experiment illustrated the accuracy and robustness of the entire framework. Volunteer experiments indicated the potential of clinical application of our system. To date, our work has been dedicated to proposing a new set of processes and methods to achieve a more intelligent automatic robot US imaging system. In future work, we will focus on typical clinical applications and add more clinical information into consideration.

First, we have proved that the proposed framework and system is feasible in principle. In the clinical environment, more specific requirements and details need to be considered. In some US imaging environments, the ultrasound probe needs to contact the surface at a specific posture. One of our future studies is to add more dimensions such as orientation control to the action space to deal with more complex environment.

In this paper, the proposed ultrasound system was presented and validated in a generic ultrasound imaging application. As a machine learning approach to solve process decision-making problems, RL has a more direct task representation capability compared to standard methods. In contrast, the significance of standard methods and fully supervised learning methods is to improve system accuracy by increasing the recognition accuracy of targets, rather than directly establishing the relationship between the output and the task. Therefore, RL has potential applications in complex medical scenarios due to its characteristics in building the relationship of tasks and actions. We will focus on the application and performance of RL-based ultrasound robots in different environments and compare them with existing methods to further investigate the feasibility of this approach in the future work.

On the other hand, a targeted medical background also brings more complex environment modeling and longer training time. Meanwhile, datasets for medical scenarios are difficult to gather. This means the transformation of RL models from simulation to real scenarios is very important especially in medical backgrounds. Therefore, domain adaptation-based scalability and model transfer of our ultrasound robotic systems will also be a major topic in the future work. Moreover, this article discusses mainly the method of controlling the ultrasound probe to image the soft target. In the clinical environment, the ultrasound image itself also has information to guide the movement of the ultrasound probe. For example, the location information and movement of targets in the body may need to be considered. In future work, the real-time ultrasound image processing, especially the automatic identification and tracking of the internal non-rigid target, is one of our main interests.

V. CONCLUSION

In this article, we propose a fully automatic ultrasound imaging system that combines multiple processes and methods for ultrasound imaging of soft, marker-less and moving targets. In this system, the combination of the reinforcement learning

model and the state representation model provides spatial action instructions for the probe under the guidance of the single RGB image. The proposed force-to-displacement control method realizes the adaptable constant force tracking of the soft target by the ultrasound probe. Quantitative and qualitative experiments prove the feasibility and robustness of our system. This system and framework provide a solution for clinical low-cost, autonomous ultrasound imaging systems.

REFERENCES

- [1] A. Priester, S. Natarajan, and M. O. Culjat, "Robotic ultrasound systems in medicine," *IEEE Trans. Ultrason., Ferroelect.,* vol. 60, no. 3, pp. 507–523, Mar. 2013.
- [2] K. Lau *et al.*, "A flexible surgical robotic system for removal of early-stage gastrointestinal cancers by endoscopic submucosal dissection," *IEEE Trans. Ind. Informat.*, vol. 12, no. 6, pp. 2365–2374, Dec. 2016.
- [3] F. Chen, J. Liu, and H. Liao, "3D Catheter shape determination for endovascular navigation using a two-step particle filter and ultrasound scanning," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 685–695, Mar. 2017.
- [4] L. Beales *et al.*, "Reproducibility of ultrasound measurement of the abdominal aorta," *Brit. J. Surg.*, vol. 98, pp. 1517–1525, 2011.
- [5] H. N. Cardinal, J. D. Gill, and A. Fenster, "Analysis of geometrical distortion and statistical variance in length, area, and volume in a linearly scanned 3-D ultrasound image," *IEEE Trans. Med. Imag.*, vol. 19, no. 6, pp. 632–651, Jun. 2000.
- [6] G. T. Sung and I. S. Gill, "Robotic laparoscopic surgery: A comparison of the da vinci and zeus systems," *Urology*, vol. 58, no. 6, pp. 893–898, 2001.
- [7] Q. Huang, J. Lan, and X. Li, "Robotic arm based automatic ultrasound scanning for three-dimensional imaging," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1173–1182, Feb. 2019.
- [8] A. M. Priester, S. Natarajan, and M. O. Culjat, "Robotic ultrasound systems in medicine," *IEEE Trans. Ultrason., Ferroelect., Freq. Control*, vol. 60, no. 3, pp. 507–523, Mar. 2013.
- [9] J. Rosen, "Surgical robotics," in *Medical Devices: Surgical and Image-Guided Technologies*, Hoboken, NJ, USA: Wiley, 2013, pp. 63–98.
- [10] Z. Pan *et al.*, "Comparison of medical image 3D reconstruction rendering methods for robot assisted surgery," in *Proc. Int. Conf. Adv. Robot. Mechatronics*, Hefei, China, 2018, pp. 94–99.
- [11] S. Merouche *et al.*, "A robotic ultrasound scanner for automatic vessel tracking and three-dimensional reconstruction of B-mode images," *IEEE Trans. Ultrason. Ferroelect.*, vol. 63, no. 1, pp. 35–46, Jan. 2016.
- [12] K. Mathiassen *et al.*, "An ultrasound robotic system using the commercial robot UR5," *Front. Robot. AI*, vol. 3, no. 1, pp. 1–16, 2016.
- [13] C. Hennesperger *et al.*, "Towards MRI-Based autonomous robotic US acquisitions: A first feasibility study," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 538–548, Feb. 2017.
- [14] M. L. Balter *et al.*, "Adaptive kinematic control of a robotic venipuncture device based on stereo vision, ultrasound, and force guidance," *IEEE Trans. Ind. Electron.*, vol. 64, no. 2, pp. 1626–1635, Feb. 2017.
- [15] I. Kuhlemann *et al.*, "Patient localization for robotized ultrasound-guided radiation therapy," in *Proc. Imag. Comput. Assistance Radiat. Ther.*, 2015, pp. 105–112.
- [16] P. Chatelain *et al.*, "Real-time needle detection and tracking using a visually servoed 3D ultrasound probe," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 1676–1681.
- [17] B. Meng *et al.*, "Robot-assisted mirror ultrasound scanning for deep venous thrombosis detection using RGB-D sensor," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14247–14261, 2016.
- [18] Y. Sasaki *et al.*, "Development of compact portable ultrasound robot for home healthcare," *J. Eng.*, vol. 2019, no. 14, pp. 495–499, 2019.
- [19] M. Bowthorpe *et al.*, "Smith predictor-based robot control for ultrasound-guided teleoperated beating-heart surgery," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 157–166, Jan. 2014.
- [20] J. Seo *et al.*, "Development of prototype system for robot-assisted ultrasound diagnosis," in *Proc. 15th Int. Conf. Control, Automat. Syst.*, 2015, pp. 1285–1288.
- [21] G. Ning *et al.*, "Real-time and multimodality image-guided intelligent HIFU therapy for uterine fibroid," *Thno*, vol. 10, no. 10, pp. 4676–4693, Mar. 2020.
- [22] T. Yoshikawa, "Dynamic hybrid position/force control of robot manipulators—Description of hand constraints and calculation of joint driving force," *IEEE J. Robot. Autom.*, vol. 3, no. 5, pp. 386–392, Oct. 1987.
- [23] H. Seraji and R. Colbaugh, "Force tracking in impedance control," *Int. J. Robot. Res.*, vol. 16, no. 1, pp. 97–117, 1993.
- [24] V. Gullapalli *et al.*, "Learning reactive admittance control," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 1992, pp. 1475–1480.
- [25] H. Bruyninckx and J. De Schutter, "Specification of force-controlled actions in the 'task frame formalism'—A synthesis," *IEEE Trans. Robot. Autom.*, vol. 12, no. 4, pp. 581–589, Aug. 1996.
- [26] B. Baigzadehnoe *et al.*, "On position/force tracking control problem of cooperative robot manipulators using adaptive fuzzy backstepping approach," *ISA Trans.*, vol. 70, pp. 432–446, 2017.
- [27] J. Duan *et al.*, "Adaptive variable impedance control for dynamic contact force tracking in uncertain environment," *Robot. Auton. Syst.*, vol. 102, pp. 54–65, 2018.
- [28] R. Martín-Martín *et al.*, "Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Macau, China, 2019, pp. 1010–1017.
- [29] K. Arulkumaran *et al.*, "Deep reinforcement learning: A brief survey," *IEEE Signal. Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [30] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [31] E. Theodorou *et al.*, "Reinforcement learning of motor skills in high dimensions: A path integral approach," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2010, pp. 2397–2403.
- [32] P. Abbeel *et al.*, "An application of reinforcement learning to aerobatic helicopter flight," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1–8.
- [33] K. M. Jagodnik *et al.*, "Human-like rewards to train a reinforcement learning controller for planar arm movement," *IEEE Trans. Hum. Mach. Syst.*, vol. 46, no. 5, pp. 723–733, Oct. 2016.
- [34] S. Gu *et al.*, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 3389–3396.
- [35] S. Levine *et al.*, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2016.
- [36] V. C. Kumar *et al.*, "Contextual reinforcement learning of visuo-tactile multi-fingered grasping policies," 2019, *arXiv:1911.09233v2*.
- [37] X. Yu *et al.*, "A robotic auto-focus system based on deep reinforcement learning," in *Proc. 15th Int. Conf. Control, Automat., Robot. Vis.*, 2018, pp. 204–209.
- [38] X. Tan *et al.*, "Robot-assisted training in laparoscopy using deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 485–492, Apr. 2019.
- [39] J. Schulman *et al.*, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
- [40] T. Lesort *et al.*, "State representation learning for control: An overview," *Neural Netw.*, vol. 108, pp. 379–392, 2018.
- [41] A. Raffin *et al.*, "Decoupling feature extraction from policy learning: Assessing benefits of state representation learning in goal based robotics" 2019, *arXiv: 1901.08651*.
- [42] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 9, no. 5, pp. 1054–1054, Sep. 1998.
- [43] P. Varin *et al.*, "A comparison of action spaces for learning manipulation tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 6015–6021.
- [44] P. Falco and C. Natale, "On the stability of closed-loop inverse kinematics algorithms for redundant robots," *IEEE Trans. Robot.*, vol. 27, no. 4, pp. 780–784, Aug. 2011.
- [45] M. Felix *et al.*, "ROS-Industrial universal robot meta-package," *Universal Robot*, 2012. [Online]. Available: https://github.com/ros-industrial/universal_robot
- [46] C. Erwin and B. Yunfei, "PyBullet, a python module for physics simulation for games," *PyBullet*, 2016. [Online]. Available: <http://pybullet.org/>