Autonomous Navigation of an Ultrasound Probe Towards Standard Scan Planes with Deep Reinforcement Learning

Keyu Li, Jian Wang, Yangxin Xu, Hao Qin, Dongsheng Liu, Li Liu*, and Max Q.-H. Meng*, *Fellow*, *IEEE*

***Abstract*— Autonomous ultrasound (US) acquisition is an important yet challenging task, as it involves interpretation of the highly complex and variable images and their spatial relationships. In this work, we propose a deep reinforcement learning framework to autonomously control the 6-D pose of a virtual US probe based on real-time image feedback to navigate towards the standard scan planes under the restrictions in real-world US scans. Furthermore, we propose a confidence-based approach to encode the optimization of image quality in the learning process. We validate our method in a simulation environment built with real-world data collected in the US imaging of the spine. Experimental results demonstrate that our method can perform reproducible US probe navigation towards the standard scan plane with an accuracy of $4.91mm/4.65°$ in the intra-patient setting, and accomplish the task in the intra- and inter-patient settings with a success rate of 92% and 46%, respectively. The results also show that the introduction of image quality optimization in our method can effectively improve the navigation performance.**

***Index Terms*— Autonomous Ultrasound Acquisition, Deep Reinforcement Learning, Image Quality Optimization.**

I. INTRODUCTION

Due to the advantages of portability, non-invasiveness, low cost and real-time capabilities over other imaging techniques, ultrasound (US) imaging has been widely accepted as both a diagnostic and a therapeutic tool in various medical disciplines [1]. However, US acquisition in current clinical procedures requires specialized personnel to manually navigate the probe towards the correct imaging plane, which is very time-consuming and the imaging quality is highly dependent on the sonographer. Moreover, the heavy workload has been exposing the sonographers to health risks such as work-related musculoskeletal disorders [2][3]. Therefore, an automation of the US scanning process holds great promise for reducing the heavy workload of sonographers, shortening the examination time, yielding high-quality, standardized and operator-independent imaging results, and improving access to care in remote or rural communities.

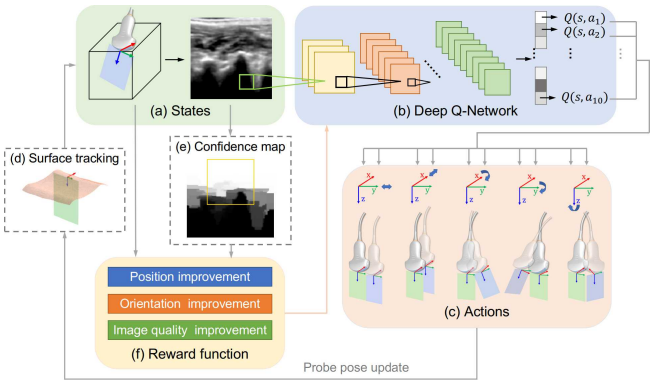

Fig. 1. An overview of the presented method for autonomous navigation of a US probe. At each time step, a 2D image is acquired with the current probe pose (a) and serves as the input of the deep Q-network (b). The optimal movement action is selected from 10 actions associated with the 5-DOF pose of the probe (c), and 1 DOF is used to track the patient surface (d). The confidence map (e) is computed from the US image to estimate the image quality, and the reward function (f) encourages the improvement of position, orientation and image quality during the navigation.

Over the past few decades, the advances in robotic US acquisitions have demonstrated the potential of using robots to automatically cover a region of interest in the patient [4][5][6][7]. However, these methods ignore the interpretation of the images and the precise positioning of the probe for visualization of the desired imaging planes, i.e., the standard scan planes, which are important in clinical US examinations since they can clearly show the anatomical structures and contain valuable information for identifying abnormalities or performimg biometric measurements [8][9]. Autonomous probe navigation towards the standard scan planes remains a challenging task because it involves interpretation of the highly complex and variable US images acquired during the scan and their spatial relationships.

In recent years, Reinforcement Learning (RL) has achieved superior performance in sequential decision-making problems in many real-life applications such as mobile robot navigation [10]. Inspired by the latest developments in RL-based medical image analysis [11][12], in this work, we propose a deep RL framework for the autonomous US acquisition task, where an agent continuously controls the 6-D pose of a virtual US probe based on the acquired images to navigate towards the standard scan plane, as shown in Fig. 1.

This work was partially supported by Hong Kong RGC GRF grant #14210117, Hong Kong RGC TRS grant T42-409/18-R and Hong Kong RGC GRF grant #14211420 awarded to Max Q.-H. Meng.

K. Li, L. Liu, and Y. Xu are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: kyli@link.cuhk.edu.hk; liliu@cuhk.edu.hk).

J. Wang is with the School of Biomedical Engineering, Shenzhen University, Shenzhen, China.

H. Qin is with Sonoscape Medical Corp., Shenzhen, China.

D. Liu is with the Department of pain, Peking University Shenzhen Hospital, Shenzhen, China.

Max Q.-H. Meng is with the Department of Electronic and Electrical Engineering of the Southern University of Science and Technology in Shenzhen, China, on leave from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute of the Chinese University of Hong Kong in Shenzhen, China (e-mail: max.meng@ieee.org).

*Corresponding authors.

Moreover, we encode the optimization of image quality in the RL framework with a confidence-based approach to improve the navigation performance. The video demonstration can be found at `https://youtu.be/_jcmyAA0WxM`. The main contributions of this work are as follows:

- We present the first deep RL framework to control the 6-D pose of a virtual US probe based on real-time image feedback, in order to mimic the decision-making process of sonographers to autonomously navigate towards the standard scan planes.
- Furthermore, based on the *ultrasound confidence map* [13], we propose a confidence-based method to introduce the optimization of image quality in the learning of navigation strategy. We empirically demonstrate in the experiments that our method can improve the navigation performance.
- A simulation environment is delicately designed and built with real-world US data to simulate the US probe navigation scenario. We take into account several practical limitations in real-world US scans to make the learned navigation policy easier to generalize to real-world applications.

II. RELATED WORK

A number of methods have been proposed to detect the standard scan planes from 3D US volumes. Some methods used CNNs to regress the transformation from the current plane to the standard plane [14][15], but this kind of prediction may cause abrupt changes in position rather than gradual and continuous changes, which is undesirable for the probe navigation task. Alansary et al. [11] parameterized a plane as $ax + by + cz + d = 0$, and customized an RL agent to learn step-by-step adjustment of the plane parameters to find the standard plane in 3D MRI scans. Dou et al. [16] extended this method for US standard plane detection. However, these methods focus on detecting the planes in the pre-acquired and processed 3D volumes and cannot be directly used to navigate a probe.

Some researchers addressed the probe navigation problem with imitation learning methods such as inverse reinforcement learning [17] and behavioral cloning [18] to learn from expert demonstrations. However, complete and accurate expert demonstrations can be intractable or expensive to obtain in the clinical US scans. Jarosik et al. [19] customized an RL agent to move a virtual probe in a simple and static toy environment, but the real-world probe navigation task is much more complicated and challenging due to the highly variable anatomy among patients. In [20], the researchers used RL to learn cardiac US probe navigation in a simulation environment built with spatially tracked US frames acquired by a sonographer on a grid covering the chest of the patient. Similarly, [21] used 2D images acquired on a grid covering the lower back to train an RL agent to find the sacrum with 2-DOF actions. Since the simulation environments in these grid-based methods are built with 2D US images acquired by a limited number of probe poses, the actions learned by these methods are restricted to the collected data. Also, the pose of the patient relative to the probe is assumed to be static, which is difficult to achieve in real US scanning scenarios.

In addition, the prior work [17]–[21] only controls a part of the probe pose based on the image feedback instead of the complete 6-DOF probe pose, and none of these methods consider the influence of image quality on the navigation performance. In this work, we build a simulation environment using 3D US volumes reconstructed from real-world US data covering the region of interest in the patient. Therefore, the agent can arbitarily change the 6-DOF pose of the virtual probe and the resulting US image can be sampled in the volume. Besides, we adopt a probe-centric action parameterization to relax the requirements for the patient's pose during the scan, thereby making the learned probe navigation policy easier to generalize to real-world applications. Moreover, we propose a method to take into account the image quality optimization in the learning of navigation strategy.

III. METHODS

A. Reinforcement Learning for Probe Navigation

In this work, we consider the autonomous US acquisition task where a US probe is automatically navigated to obtain an image of the standard scan plane. This problem can be formulated as a sequential decision-making problem in the RL framework, where the agent, in this case a virtual US probe, interacts with the environment, in this case the virtual patient, in a sequence of observations, actions and rewards. We define the RL framework for probe navigation as the following.

1) States: In our environment, the virtual patient is a reconstructed 3D US volume V that covers the region of interest in the patient. We set the virtual US probe as a commonly used 2D probe with a field-of-view of $h \times w$. At time step t, the 6-D pose of the probe $\{P\}$ with respect to the world coordinate frame $\{W\}$ can be described by a spatial transformation matrix ${}^{W}_{P}T_t$, which uniquely determines the probe's position $\boldsymbol{p}_t = [p_x, p_y, p_z]$ and orientation $\boldsymbol{q}_t = [q_x, q_y, q_z, q_w]$ (represented with a quaternion). As shown in Fig. 1 (a), assuming that the yz plane of the probe is aligned with the image plane, then the 2D US image of size $h \times w$ acquired with the current probe pose can be sampled from the patient $I_t \leftarrow sample(V, \boldsymbol{p}_t, \boldsymbol{q}_t)$. The goal probe pose is $(\boldsymbol{p}_g, \boldsymbol{q}_g)$, corresponding to the standard plane image $I_g \leftarrow sample(V, \boldsymbol{p}_g, \boldsymbol{q}_g)$. We consider the visual navigation task as patially observed, where the goal and the probe pose in the world coordinate system are unobservable, and the agent can only observe the acquired US images. A sequence of m recent images are stacked together as the state $s_t := [I_{t-m+1}, \cdots, I_t]$ to take into account the dynamic information [22].

2) Actions: Based on the observation at time step t, the agent takes an action selected by its policy $\pi\colon s_t \mapsto a_t$. Basically, we define the navigation action as a transform operator in the probe frame $a_t := {}^{P}T_t$ that transforms the current probe pose to a new pose as

$$ {}^{W}_{P}T_{t+1} = {}^{W}_{P}T_t \cdot {}^{P}T_t \tag{1} $$

Different from [20][21] which represent the actions in the world coordinate frame, we adopt a probe-centric action parameterization to relax the restrictions on the patient's actual pose in the world. We only require the coronal plane of the patient to be roughly parallel to the horizontal plane (xy plane of $\{W\}$). As shown in Fig. 1 (c), 10 discrete actions related with 5 DOFs of the probe are used, namely, 4 actions to translate a certain distance $\pm d_{step}$ along the probe's x, y axes and 6 actions to rotate a certain angle $\pm\theta_{step}$ around the probe's x, y, z axes, respectively. Since we use the height of the probe to track the patient surface (which will be explained in *3)*), we slightly modify the 4 translational movement actions to translation along the projections of the probe's x, y axes on the horizontal plane x', y'.

Similar to [11][12], we use hierarchical action steps to search for the plane in a coarse-to-fine manner. Specifically, a total of 5 step sizes are used. The action step is initialized as $d_{step} = 5mm$ and $\theta_{step} = 5°$, and a buffer is used to store 30 historical poses $[(\boldsymbol{p}_{t-29}, \boldsymbol{q}_{t-29}), \cdots, (\boldsymbol{p}_t, \boldsymbol{q}_t)]$. If 3 pairwise Euclidean distances between the historical poses are less than a threshold 0.01, the agent is considered to have converged to a pose and the action step will be reduced by 1 unit until it becomes zero.

3) State transition under restrictions: If there are no restrictions, the probe pose can be updated according to the selected action as in the previous work [19][20][21]. Here, we instead consider two requirements for the probe pose in real-world US scans: i) the contact between the probe and the patient surface should be maintained to ensure sufficient acoustic coupling, and ii) the tilt angle of the probe should be limited to ensure the comfort and safety of the patient.

To this end, we first update the horizontal position of the probe (p_x, p_y) using (1), and use the z-coordinate of the probe to track the patient surface $p_z \leftarrow surface(p_x, p_y)$, as shown in Fig. 1 (d). In order to extract the surface equation $z = surface(x, y)$, for each pair of (x, y) in the volume V, we approximate the surface point as the point with the largest z-coordinate whose gray value is not zero. Note that this intensity-based method is only used to estimate the patient surface in our simulation. In real-world applications, the patient surface can be extracted in real time based on data obtained with external sensors such as an RGB-D camera [5][6]. Second, after the new probe orientation is given by (1), the resulting tilt angle of the probe (i.e., angle between the z-axis of the probe $\hat{\boldsymbol{z}}_p$ and the $-z$ direction of $\{W\}$) is $\beta = \arccos(\hat{\boldsymbol{z}}_p, [0, 0, -1]^T) = \arccos(-{}^W_P T_{t+1}(3, 3))$. We restrict the tilt angle to be smaller than 30°. If $\beta > 30°$, the probe orientation will not be updated. After the new probe pose $\boldsymbol{p}_{t+1}, \boldsymbol{q}_{t+1}$ is obtained under the above restrictions, a new US image I_{t+1} can be acquired, and the observation is updated to s_{t+1}.

4) Reward function: In our probe navigation task, the reward function should be designed to encourage the agent to move towards the goal. Instead of simply classifying the results of actions as moving closer to or further away from the goal and assigning corresponding rewards [11][20][21], we design the reward function to be proportional to the amount of pose improvement. At time step t, the distance-to-goal can be measured in position and orientation, respectively:

$$d_t = \left\|\boldsymbol{p}_t - \boldsymbol{p}_g\right\|_2, \ \theta_t = 2\arccos(|\langle \boldsymbol{q}_t, \boldsymbol{q}_g \rangle|), \tag{2}$$

where d_t is the Euclidean distance between the current positions of the probe and the goal, and θ_t is the minimum angle required to rotate from the current probe orientation to the goal orientation. Then, the pose improvement after taking action a_t normalized with the step size is

$$\Delta d_t = \frac{d_t - d_{t+1}}{d_{step}}, \ \Delta\theta_t = \frac{\theta_t - \theta_{t+1}}{\theta_{step}}, \tag{3}$$

where $\Delta d_t, \Delta\theta_t \in [-1, 1]$.

In addition, we assign a high reward $(+10)$ for task accomplishment ($d_t \leq 1mm$ and $\theta_t \leq 1°$) and add some penalties based on the restrictions of the environment. If the action causes the tilt angle of the probe $\beta > 30°$, the agent will receive a penalty of -0.5. If the probe moves outside the patient (the proportion of pixels with non-zero gray value in I_t is less than 30%), the agent will get a penalty of -1. In summary, the reward function (without confidence improvement) is defined as

$$r_t = \begin{cases} -1, & \text{if moving out of } V; \\ -0.5, & \text{if } \beta > 30°; \\ 10, & \text{if reaching goal;} \\ \Delta d_t + \Delta\theta_t, & \text{otherwise.} \end{cases} \tag{4}$$

5) Termination conditions: During training, we terminate an episode when: a) the goal is reached, or b) the number of steps exceeds the maximum limit, or c) the action step is reduced to zero, or d) the probe moves out of the patient. During testing, only the termination conditions b,c,d are used due to the absence of the goal's true location.

B. Confidence-aware Agent

In clinical US examinations, the sonographer will continuously adjust the probe to obtain clear images while searching for the correct imaging plane, and avoid positions that may cause poor image quality. This motivates us to take into consideration the impact of image quality on the agent's navigation performance. Similar to [6][23], we evaluate the image quality using the *ultrasound confidence map* [13], which estimates the pixel-wise confidence in the image based on a random walks framework, as shown in Fig. 1 (e). At time step t, the confidence map $C_t \leftarrow \text{confMap}(I_t)$ is computed from the US image, $C_t(i, j) \in [0, 1]$. Let S denote the region of interest (ROI) in the image, the quality of the image I_t can be represented by the average ROI confidence

$$c_t = \frac{1}{|S|} \sum_{(i,j)\in S} C_t(i, j) \tag{5}$$

The improvement of image quality after taking action a_t can be represented by $\Delta c_t = c_{t+1} - c_t$. We hypothesize that encouraging the improvement of image quality in addition to reducing the distance-to-goal can help the agent learn a better navigation strategy. This has been empirically verified by the

Algorithm 1: SonoRL

Input: patient dataset D
Output: Q-network

1 Initialize experience replay memory E with demonstration;
2 Pre-train the network Q with E;
3 Initialize target network $\widehat{Q} \leftarrow Q$;
4 Initialize steps $n = 0$;
5 **while** *steps* $n < N$ **do**
6 Sample a patient US volume V and goal probe pose $\boldsymbol{p}_g, \boldsymbol{q}_g$ from D;
7 Extract patient surface $z = surface(x, y)$;
8 Initialize time $t = 0$, action step d_{step}, θ_{step};
9 Initialize probe pose $\boldsymbol{p}_t, \boldsymbol{q}_t$ randomly, observe an image $I_t \leftarrow \text{sample}(V, \boldsymbol{p}_t, \boldsymbol{q}_t)$ and initialize observation s_t;
10 Calculate ROI confidence c_t;
11 **repeat**
12 Select a random action a_t with probability ε, otherwise select $a_t = \arg\max_a Q(s_t, a)$;
13 Update probe pose under restrictions $\boldsymbol{p}_{t+1}, \boldsymbol{q}_{t+1} \leftarrow \text{updateProbePose}(\boldsymbol{p}_t, \boldsymbol{q}_t, a_t, surface)$;
14 Observe new image $I_{t+1} \leftarrow \text{sample}(V, \boldsymbol{p}_{t+1}, \boldsymbol{q}_{t+1})$ and update observation s_{t+1} ;
15 Calculate ROI confidence c_t;
16 Calculate reward r_t by (4) or (6);
17 Store transition (s_t, a_t, r_t, s_{t+1}) in E;
18 Update action step d_{step}, θ_{step} if needed;
19 $t \leftarrow t + 1$;
20 $n \leftarrow n + 1$;
21 **if** $(n \mod K) = 0$ **then**
22 Sample a minibatch data from E;
23 Update network Q by gradient descent;
24 **end**
25 **if** $(n \mod C) = 0$ **then**
26 Update target network $\widehat{Q} \leftarrow Q$;
27 **end**
28 **until** *Termination condition satisfied*;
29 **end**
30 **return** Q

experimental results in Section IV-C. Therefore, we introduce a confidence-aware auxiliary term in the reward function to encode the optimization of image quality in the learning process. As shown in Fig. 1 (f), the modified reward function for the confidence-aware agent takes into consideration the improvement of position, orientation and image quality:

$$r_t = \begin{cases} -1, & \text{if moving out of } V; \\ -0.5, & \text{if } \beta > 30°; \\ 10, & \text{if reaching goal;} \\ \Delta d_t + \Delta\theta_t + \Delta c_t, & \text{otherwise.} \end{cases} \quad (6)$$

C. Deep Q-Network Training

1) Deep Q-Learning Algorithm: In the RL framework, the agent learns to maximize the discounted sum of future rewards $G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-t-1} r_T$ where $\gamma \in (0, 1)$ is a discount factor and T is the time step when the episode is terminated. The optimal policy $\pi^*: s_t \mapsto a_t$ is to maximize the expected return

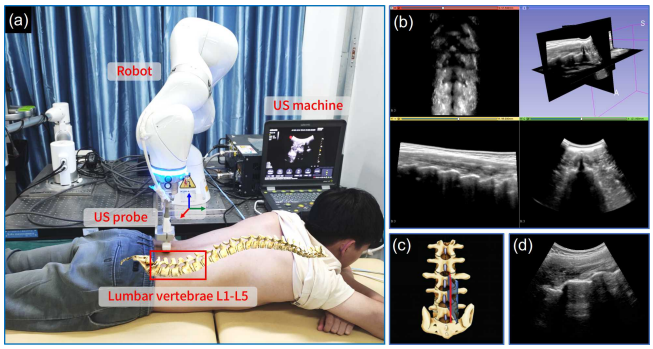

Fig. 2. Illustration of the data acquisition. (a) A robotic US system is used to acquire B-mode US images of the L1-L5 lumbar vertebrae of the volunteers for 3D volume reconstruction. (b) is the visualization of an exemplary data volume in 3D Slicer [25]. (c) illustrates the paramedian sagittal oblique scan (PMSOS) [9], and (d) shows an image of the PMSOS plane acquired by a clinician.

$$\begin{aligned} \pi^* &= \arg\max_a Q^*(s, a), \\ Q^*(s, a) &= \mathbb{E}_{s'}[r(s, a) + \gamma \max_{a'} Q^*(s', a')] \end{aligned} \quad (7)$$

where the optimal state-action value function $Q^*(s, a)$ is defined as the maximum expected return following any policy π: $Q^*(s, a) = \max_\pi \mathbb{E}_\pi[G_t | s_t = s, a_t = a]$.

The deep Q-learning algorithm [22] uses a deep neural network to approximate the Q-function, and train the network by the temporal-difference method with experience replay and target network techniques. We implement the deep Q-learning algorithm with minor modifications targeted at our application. We refer to our DQN-based RL framework for US probe navigation as *SonoRL*, which is outlined in Algorithm 1. The Q-network is pre-trained on some demonstration trajectories generated with an expert policy which selects actions to maximize the one-step pose improvement $\Delta d + \Delta\theta$ (line 1-3). Subsequently, the network is trained with self-generated experiences by interacting with the environment with an ε-greedy policy (line 4-29). Two different agents are used, i.e., *SonoRL w/ conf* and *SonoRL w/o conf*, which use different reward functions (line 16).

2) Implementation Details: We adopt the SonoNet-16 [8] architecture initially proposed for US standard plane detection as our Q-network model and remove the final softmax layer (see Fig. 1 (b)). The network is trained every 10 interaction steps with a batch size of 32 using Adam optimizer [24], and the target nerwork is updated every $1k$ training steps. The discount factor γ is 0.9. The exploration rate ε decays linearly from 0.5 to 0.1 in the first $100k$ interaction steps and stays unchanged for the remaining steps. The experience replay memory has a capacity of $100k$ and is initialized with $5k$ demonstration data. During the pre-training phase, $100k$ demonstration experiences are collected and the network is updated for $10k$ steps with a learning rate of 0.01. During reinforcement learning, the learning rate is set to 0.01 for the first $40k$ training steps, 0.001 for the next $40k$ steps, 5e-4 for the next $30k$ steps, and 1e-4 for the remaining steps.

TABLE I

QUANTITATIVE RESULTS OF US PROBE NAVIGATION TOWARDS THE PMSOS PLANE

Methods		$\Delta d + \Delta\theta$	Position error (mm)	Orientation error $(^\circ)$	SSIM	Success rate	Average number of steps
Intra-observer errors		–	4.30 ± 1.53	3.34 ± 1.60	0.68 ± 0.14	–	–
Intra-patient	SonoRL w/o conf	0.34 ± 0.15	5.85 ± 6.77	13.00 ± 32.20	0.66 ± 0.18	79%	61
	SonoRL w/ conf	$\mathbf{0.35 \pm 0.13}$	$\mathbf{4.91 \pm 4.44}$	$\mathbf{4.65 \pm 2.61}$	$\mathbf{0.67 \pm 0.21}$	**92%**	66
Inter-patient	SonoRL w/o conf	0.16 ± 0.17	21.87 ± 15.30	39.93 ± 61.04	0.43 ± 0.16	21%	63
	SonoRL w/ conf	$\mathbf{0.18 \pm 0.17}$	$\mathbf{13.93 \pm 11.58}$	$\mathbf{29.12 \pm 51.46}$	$\mathbf{0.50 \pm 0.23}$	**46%**	76

IV. EXPERIMENTS

In order to demonstrate the effectiveness of our proposed RL framework, we apply it to the task of US imaging of the spine, to autonomously navigate the probe towards one of the standard scan planes – the paramedian sagittal oblique scan plane (PMSOS) [9], as illustrated in Fig. 2 (c)(d).

A. Data Acquisition

A total of 41 3D US volumes of the spine that cover the L1-L5 lumbar vertebrae are acquired from 17 healthy male volunteers aged 20 to 26. The average volume size of our dataset is $350 \times 397 \times 274$ and each voxel is $0.5 \times 0.5 \times 0.5mm^3$. In order to acquire the data volumes, we built a robotic US system using a KUKA LBR iiwa 7 R800 (KUKA Roboter GmbH, Augsburg, Germany) with a C5-1B convex US transducer mounted at its end-effector and connected to a Wisonic Clover diagnostic US machine (Shenzhen Wisonic Medical Technology Co., Ltd, China), as shown in Fig. 2 (a). The volunteers were in a prone position on an examination bed during the scan. Before acquisition, a clinician selected the scanning parameters, applied the coupling gel on the surface of the volunteers and specified the start and end points of the scan. During acquisition, the robot linearly moved the probe from the start point to the end point under Cartesian impedance control. A high stiffness value $(2000N/m)$ was set in the xy plane and a low stiffness value $(50N/m)$ was set along the z-axis, and an additional force of $5N$ was applied in the downward direction to ensure contact between the probe and the patient. The B-mode images and the probe pose measured by the robot were transmitted to a computer for volume reconstruction using a squared distance weighted approach [26]. An exemplary data volume is shown in Fig. 2 (b). In addition, the clinician manually navigated the probe towards the PMSOS plane (see Fig. 2 (d)) and the corresponding probe pose was recorded.

B. Simulation Setup

We built a simulation environment in Python for US probe navigation, and implemented the *SonoRL* algorithm based on the Tensorpack [27] interface. At each time step, the agent observes an image of size 150×150 and stacks 4 recent frames as the state. The selected ROI is of size 110×90. In each episode, the horizontal position of the probe is randomly initialized in the center region $\{(x, y) : x \sim \mathcal{U}(0.3W, 0.7W), y \sim \mathcal{U}(0.2L, 0.8L)\}$, where L, W are the length and width of the data volume, and the initial z-coordinate of the probe is determined by the extracted surface. The initial z-axis of the probe is aligned with the $-z$ direction of the world frame, and the probe is randomly rotated $\eta \sim \mathcal{U}(0, 360^\circ)\}$ around its z-axis. The maximum number of steps in each episode is limited to 120.

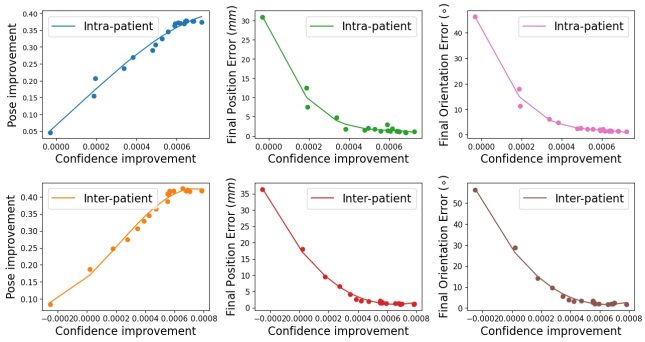

Fig. 3. The navigation performance against the image quality improvement of the *SonoRL w/o conf* agent during training in the intra- and inter-patient settings. As the average confidence improvement per step increases, the average pose improvement per step increases and the final position and orientation errors decrease, showing that the navigation performance is positively correlated with the improvement of image quality.

C. Quantitative Evaluation

In order to fully evaluate the effectiveness of our method, we consider two different settings: *intra-patient* and *inter-patient*. In the intra-patient setting, the model is trained with 33 data volumes obtained from 17 subjects and tested with 8 unseen data volumes obtained from 8 of these subjects. The design of this setting is motivated by the real-world scenario where more than one US acquisition of the same patient is required, such as pre- and post-operative ultrasonography. In the inter-patient setting, the model is trained with 33 data volumes obtained from 14 subjects and tested with 8 data volumes obtained from 3 unseen subjects. This task is more difficult since it requires the learned policy to be generalized to patients with highly variable anatomical structures. The *SonoRL* agents with and without confidence optimization are refered to as *SonoRL w/ conf* and *SonoRL w/o conf*, respectively. Each model is optimized for $160k$ and $200k$ iterations in the intra- and inter-patient settings respectively to achieve stable performance.

We first studied the relationship between the *SonoRL w/o conf* agent's navigation performance and the improvement of image quality during training, as shown in Fig. 3. In both the intra- and inter-patient settings, as the confidence improvement in each step becomes greater, the pose im-

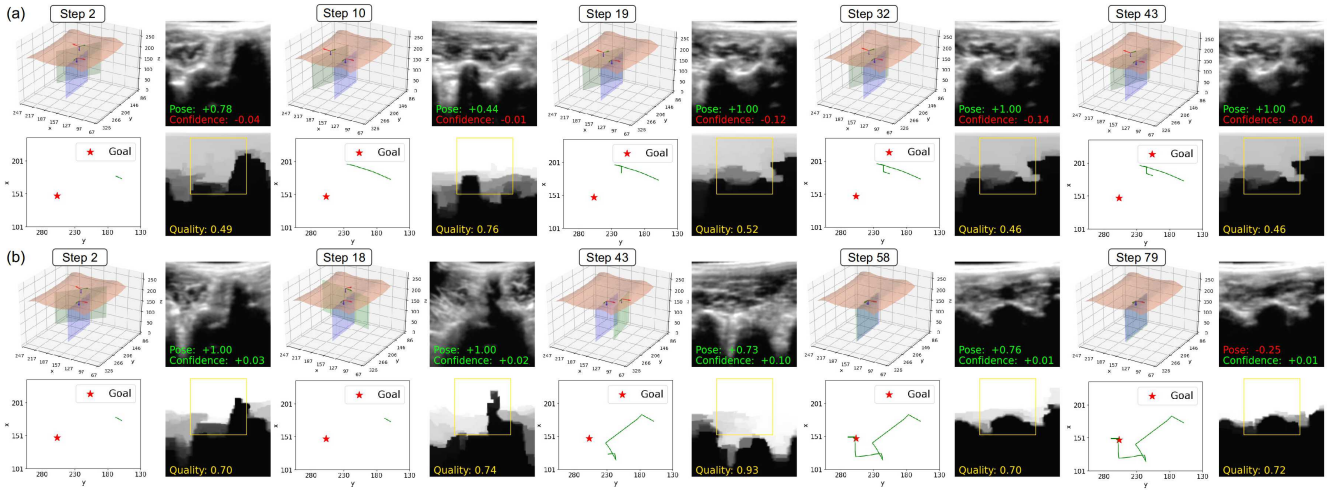

Fig. 4. Snapshots of the trajectories of the (a) *SonoRL w/o conf* and (b) *SonoRL w/ conf* agents in an intra-patient test case. The 3D plot shows the current plane (green), the goal plane (blue), the patient surface (salmon) and the poses of the current probe and the goal. Below it shows the top-view trajectory of the agent on the xy plane (green), and the goal position is indicated by a red star. The current plane image is displayed on the right side of each 3D plot, with the pose and confidence improvement of the current step marked in green (positive) or red (negative). The confidence map below it shows the pixel-wise confidence in the current image with the average confidence in the ROI (yellow rectangle).

provement in each step increases and the final pose error decreases. This indicates that the navigation performance is positively correlated with the improvement of image quality, as predicted in Section III.B. Therefore, although only $\Delta d + \Delta\theta$ is used in the reward function (4), the agent also implicitly learns to increase Δc as it gradually learns to correctly navigate the probe.

The quantitative evaluation of the two agents is carried out on 24 random test cases for each setting, including 3 navigation tests on each virtual patient. We compare the performance of the two agents and the intra-observer errors of a human expert in Table I. The metrics include the average pose improvement per step ($\Delta d + \Delta\theta$), the final pose errors, the structural similarity (SSIM) [28] between the final plane image and the goal image, success rate and the average number of steps. The navigation is considered successful if the final pose error is less than $10mm/10°$.

In the intra-patient setting, the *SonoRL* agents can successfully accomplish most of the tasks, and the final pose error of the *SonoRL w/ conf* agent is similar to intra-observer errors, indicating that our method is able to perform reproducible probe navigation on familiar patients. In both settings, as the *SonoRL w/ conf* agent tends to avoid locations with poor image quality, it takes more navigation steps than the *SonoRL w/o conf* agent, but achieves higher pose improvement and has a greater chance of successfully reaching the goal. Both agents show a degraded performance in the inter-patient setting, mainly because the task is more difficult and challenging. Nevertheless, the introduction of confidence optimization improves the navigation performance by a large margin, which indicates that the optimization of image quality improves the generalization of the learned policy to highly variable patient anatomy.

D. Qualitative Evaluation

We further compare our methods with and without confidence optimization through qualitative analysis of an intra-patient test case, where the *SonoRL w/o conf* agent fails to navigate to the goal but the *SonoRL w/ conf* agent successfully reaches the goal with an accuracy of $1.04mm/2.18°$. The trajectories of the two agents are visualized in Fig. 4.

As shown in Fig. 4 (a), the *SonoRL w/o conf* agent greatly reduces the position error in the first 10 steps. However, since the agent is not aware of the degradation of image quality, it blindly chooses actions to reduce the pose error and navigates to a location with poor image quality (step 19, $c = 0.52$). In the remaining steps, the agent is stuck in this place until the end of the navigation (step 43) and fails to reach the goal.

In contrast, the *SonoRL w/ conf* agent adjusts its orientation in the first 18 steps to improve both the pose and the image quality. Then, the agent gradually approaches the goal in steps 19 to 58 based on the high-quality images, and further fine-tunes its pose in the remaining 21 steps. Since the confidence-aware agent considers both the distance-to-goal and the image quality when making decisions, it may not choose the most aggresive actions to achieve the goal, but will carefully avoid the bad acoustic windows during the navigation. Therefore, it ends up with a relatively longer navigation time and the trajectory appears circuitous (see Fig. 4 (b)). However, this strategy allows the agent to approach the goal more stably because the acquired images are clear and contain key information of the anatomical structures to guide the navigation.

V. CONCLUSIONS

In this paper, we propose a deep RL framework to navigate a virtual US probe towards the standard scan planes, taking into account several practical limitations in real-world US scans. The optimization of image quality is encoded in the learning process through a confidence-based method. Experimental results in a simulation environment built with real US data validate the effectiveness of the proposed method and demonstrate that the introduction of confidence

optimization improves the navigation performance. As a next step, the methods will be extended to investigate more state-of-the-art RL algorithms (e.g., [29][30][31]), and integrated with existing robotic US systems for real-world applications.

REFERENCES

[1] K. K. Shung, "Diagnostic ultrasound: Past, present, and future," *J Med Biol Eng*, vol. 31, no. 6, pp. 371–4, 2011.

[2] G. Brown, "Work related musculoskeletal disorders in sonographers," *BMUS Bulletin*, vol. 11, no. 3, pp. 6–13, 2003.

[3] M. Muir, P. Hrynkow, R. Chase, D. Boyce, and D. Mclean, "The nature, cause, and extent of occupational musculoskeletal injuries among sonographers: recommendations for treatment and prevention," *Journal of Diagnostic Medical Sonography*, vol. 20, no. 5, pp. 317–325, 2004.

[4] A. S. B. Mustafa, T. Ishii, Y. Matsunaga, R. Nakadate, H. Ishii, K. Ogawa, A. Saito, M. Sugawara, K. Niki, and A. Takanishi, "Development of robotic system for autonomous liver screening using ultrasound scanning device," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2013, pp. 804–809.

[5] C. Hennersperger, B. Fuerst, S. Virga, O. Zettinig, B. Frisch, T. Neff, and N. Navab, "Towards mri-based autonomous robotic us acquisitions: a first feasibility study," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 538–548, 2016.

[6] S. Virga, O. Zettinig, M. Esposito, K. Pfister, B. Frisch, T. Neff, N. Navab, and C. Hennersperger, "Automatic force-compliant robotic ultrasound screening of abdominal aortic aneurysms," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 508–513.

[7] Q. Huang, J. Lan, and X. Li, "Robotic arm based automatic ultrasound scanning for three-dimensional imaging," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1173–1182, 2018.

[8] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, and D. Rueckert, "Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2204–2215, 2017.

[9] M. K. Karmakar and K. J. Chin, *Spinal Sonography and Applications of Ultrasound for Central Neuraxial Blocks*. New York, NY: McGraw-Hill Education, 2017. [Online]. Available: accessanesthesiology.mhmedical.com/content.aspx?aid=1141735352

[10] K. Li, Y. Xu, J. Wang, and M. Q.-H. Meng, "Sarl: Deep reinforcement learning based human-aware navigation for mobile robot in indoor environments," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 688–694.

[11] A. Alansary, L. Le Folgoc, G. Vaillant, O. Oktay, Y. Li, W. Bai, J. Passerat-Palmbach, R. Guerrero, K. Kamnitsas, B. Hou *et al.*, "Automatic view planning with multi-scale deep reinforcement learning agents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 277–285.

[12] A. Alansary, O. Oktay, Y. Li, L. Le Folgoc, B. Hou, G. Vaillant, K. Kamnitsas, A. Vlontzos, B. Glocker, B. Kainz *et al.*, "Evaluating reinforcement learning agents for anatomical landmark detection," *Medical image analysis*, vol. 53, pp. 156–164, 2019.

[13] A. Karamalis, W. Wein, T. Klein, and N. Navab, "Ultrasound confidence maps using random walks," *Medical image analysis*, vol. 16, no. 6, pp. 1101–1112, 2012.

[14] A. Schmidt-Richberg, N. Schadewaldt, T. Klinder, M. Lenga, R. Trahms, E. Canfield, D. Roundhill, and C. Lorenz, "Offset regression networks for view plane estimation in 3d fetal ultrasound," in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 109493K.

[15] Y. Li, B. Khanal, B. Hou, A. Alansary, J. J. Cerrolaza, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz *et al.*, "Standard plane detection in 3d fetal ultrasound using an iterative transformation network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 392–400.

[16] H. Dou, X. Yang, J. Qian, W. Xue, H. Qin, X. Wang, L. Yu, S. Wang, Y. Xiong, P.-A. Heng *et al.*, "Agent with warm start and active termination for plane localization in 3d ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 290–298.

[17] M. Burke, K. Lu, D. Angelov, A. Straižys, C. Innes, K. Subr, and S. Ramamoorthy, "Learning robotic ultrasound scanning using probabilistic temporal ranking," *arXiv preprint arXiv:2002.01240*, 2020.

[18] R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automatic probe movement guidance for freehand obstetric ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 583–592.

[19] P. Jarosik and M. Lewandowski, "Automatic ultrasound guidance based on deep reinforcement learning," in *2019 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2019, pp. 475–478.

[20] F. Milletari, V. Birodkar, and M. Sofka, "Straight to the point: reinforcement learning for user guidance in ultrasound," in *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*. Springer, 2019, pp. 3–10.

[21] H. Hase, M. F. Azampour, M. Tirindelli, M. Paschali, W. Simson, E. Fatemizadeh, and N. Navab, "Ultrasound-guided robotic navigation with deep reinforcement learning," *arXiv preprint arXiv:2003.13321*, 2020.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[23] P. Chatelain, A. Krupa, and N. Navab, "Confidence-driven control of an ultrasound probe," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1410–1424, 2017.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

[26] Q.-H. Huang, Y.-P. Zheng, M.-H. Lu, and Z. Chi, "Development of a portable 3d ultrasound imaging system for musculoskeletal tissues," *Ultrasonics*, vol. 43, no. 3, pp. 153–163, 2005.

[27] Y. Wu *et al.*, "Tensorpack," https://github.com/tensorpack/, 2016.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," *arXiv preprint arXiv:1710.02298*, 2017.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[31] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.