OCT Guided Robotic Ophthalmic Microsurgery via Reinforcement Learning from Demonstration

Brenton Keller[1], Mark Draelos[1], Kevin Zhou[1], Ruobing Qian[1], Anthony Kuo[2], George Konidaris[3], Kris Hauser[4], and Joseph Izatt[1]

***Abstract*—Ophthalmic microsurgery is technically difficult because the scale of required surgical tool manipulations challenge the limits of the surgeon's visual acuity, sensory perception, and physical dexterity. Intraoperative optical coherence tomography (OCT) imaging with micrometer-scale resolution is increasingly being used to monitor and provide enhanced real-time visualization of ophthalmic surgical maneuvers, but surgeons still face physical limitations when manipulating instruments inside the eye. Autonomously controlled robots are one avenue for overcoming these physical limitations. We demonstrate the feasibility of using learning from demonstration and reinforcement learning with an industrial robot to perform OCT-guided corneal needle insertions in an ex vivo model of deep anterior lamellar keratoplasty (DALK) surgery. Our reinforcement learning agent trained on ex vivo human corneas, then outperformed surgical fellows in reaching a target needle insertion depth in mock corneal surgery trials. This work shows the combination of learning from demonstration and reinforcement learning is a viable option for performing OCT guided robotic ophthalmic surgery.**

***Index Terms*—Learning from Demonstration, Medical Robots and Systems, Deep Learning in Robotics and Automation, Microsurgery**

I. INTRODUCTION AND BACKGROUND

Ophthalmic microsurgeries are among the most commonly performed surgical procedures worldwide [1], [2]. These surgeries challenge the limits of surgeon's visual acuity, sensory perception, and physical dexterity because they require placement and manipulation of surgical tools with micrometer-scale precision and milli-Newton scale forces in delicate ocular tissues [3]. To visualize the operative field, ophthalmic microsurgical procedures are performed using a stereoscopic microscope suspended above the patient which provides a top-down view with limited depth perception provided by stereopsis. One particularly promising but difficult procedure is deep anterior lamellar keratoplasty (DALK), a novel form of partial thickness corneal transplantation (Fig. 1). In DALK microsurgery, the top three layers of the patient's cornea (epithelium, Bowman's layer, and stroma) are replaced, but the bottom two layers (Descemet's membrane (DM) and endothelium), which are still viable, are preserved [4]. Published studies have shown that successful DALK procedures have fewer comorbidities than conventional full-thickness corneal transplantation, which carries a 10-year risk of failure of 10–35 % [5]. However, DALK is technically very difficult to perform. This is because taken together, the DM and endothelial layers are only 20 micrometers thick, so manually dissecting the ∼500 micrometer-thick bulk of the overlying corneal layers [6] with sharp tools while not penetrating the DM and endothelial layer is exceedingly difficult. An improvement over manual dissection is the "big bubble" technique [7], in which the stroma and epithelium are separated from DM before resection by advancing a needle (or cannula) into the stroma and injecting air between the layers to achieve pneumodissection (Fig. 1). However, this technique still requires the ophthalmic microsurgeon to position the needle at a precise location deep in the cornea without puncturing the micrometers thick endothelium. Studies have shown that increased needle depth short of puncture, expressed as a percentage of corneal thickness, increased the likelihood of fully separating the layers [8], [9]. If the stroma and DM fail to fully separate, or if the needle perforates the endothelium, most surgeons abandon DALK and typically convert to conventional full-thickness corneal transplantation. Thus, tens of micrometers separate DALK failure from success. Because of this difficulty, stroma-DM separation failure rates in the literature have been reported as high as 44–54 % [10]–[12] and perforation rates reported up to 27.3 % [11]. One potential reason the failure rates are so high for this procedure is because of how difficult it is to determine needle depth in the cornea from the conventional view through the surgical microscope (Fig. 2A).

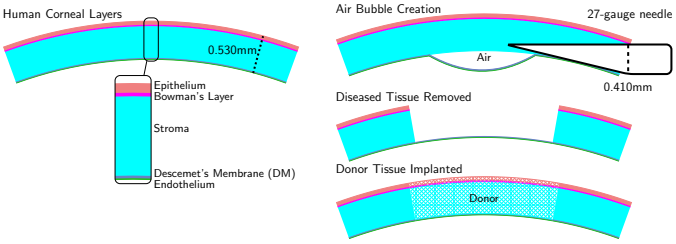

Fig. 1. Illustration of the "big bubble" DALK procedure. The surgeon inserts a needle as deep as possible into the stroma without puncturing Descemet's membrane (DM), then injects air to create a lamellar dissection of DM from the stroma. The top three layers are then removed and replaced with donor tissue.

One approach to augment the surgeon's view of the operating field is the use of intraoperative optical coherence tomography (OCT) in combination with the surgical microscope. OCT is a non-contact optical imaging modality that provides depth

*Support for this project was provided by National Institutes of Health (R01 EY023039) and the Duke Coulter Translational Partnership (2016-2018).

[1]B. Keller, M. Draelos, K. Zhou, R. Qian, and J. Izatt are with the Department of Biomedical Engineering, Duke University, Durham, NC, USA.

[2]A. Kuo is with the Department of Ophthalmology, Duke University Medical Center, Durham, NC, USA.

[3]G. Konidaris is with the Department of Computer Science Brown University, Providence, RI, USA.

[4]K. Hauser is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.

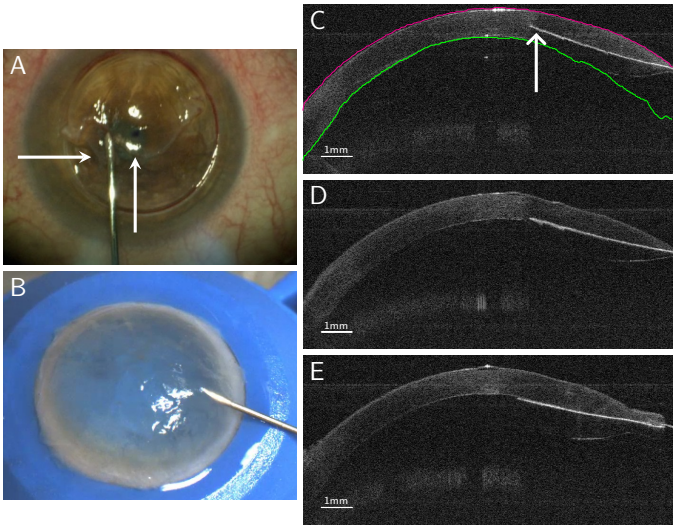

Fig. 2. Comparison of surgical field views for DALK. (A) Surgeon's view of DALK during live human surgery through a conventional operating microscope. White arrows denote the edge of a partial lamellar dissection. (B) Top down view of an ex vivo human cornea mounted onto an artificial anterior chamber used for studies in this report. Estimating the needle depth inside the cornea is very difficult from the views in (A) and (B). (C-E) OCT B-scan views along the needle's axis from ex vivo needle insertion trials performed by surgical fellows. The needle depth in the cornea from this view is much more apparent. (C) Unsuccessful insertion due to needle depth too shallow for successful layer separation. The white arrow points to the top surface of the needle, and green/magenta lines illustrate real-time automated segmentation of the epithelium and endothelium. (D) Successful insertion with the needle well positioned to inject air to separate stroma and DM. (E) Unsuccessful demonstration due to needle puncturing through both DM and endothelium.

resolved cross-sectional images in tissues [13]. Commercial [14]–[16] and research based [17]–[19] intraoperative OCT systems allow for live volumetric imaging of ophthalmic surgery. We have previously demonstrated the use of a custom-designed research intraoperative OCT system, which provides live volumetric imaging with micrometer-scale resolution [19], for real-time corneal surface segmentation and needle tracking [20] during simulated DALK procedures in ex vivo human corneas mounted in an artificial anterior chamber (Fig. 2B). Real-time reconstructed cross-sectional views of the DALK procedure along the needle axis (Fig. 2C-E) allow surgeons to visualize the top surface of the needle throughout the insertion procedure (the metal needle itself is opaque to OCT light, thus only the top surface is seen). While OCT is beneficial for visualization of procedures at the micrometer scale, surgeons still encounter physical limitations when attempting to perform needle manipulations at this scale. Microsurgeon hand tremor has been measured with RMS amplitude of 50–200 µm [21], [22], which is one reason robot manipulation could be beneficial for ophthalmic microsurgery.

Robots have been used to accomplish a wide variety of surgical tasks across several specialties including otolaryngology [23], urology [24], gastroenterology [25], and orthopedics [26]. Many robotic surgeries are performed laparoscopically with the da Vinci robot (Intuitive Surgical, Sunnyvale, CA), but this system has been found to be unsuitable for ophthalmic microsurgery due to poor visualization and lack of microsurgical tools [27]. To fill this need, both cooperatively controlled and teleoperated robots have been designed specifically for ophthalmic surgery [28]–[32] and have recently been used to perform vitreoretinal surgery in humans [33]. Cooperative and teleoperated systems can provide tremor reduction and/or motion scaling, thereby increasing a surgeon's ability to steadily hold and move tools to desired locations. However, these robots require input from a highly trained surgeon each time surgery is performed. Automating difficult or repetitive surgical tasks using robots could potentially increase the reliability of procedures, decrease operating time, and improve patient access to difficult but more beneficial procedures.

Extensive research has been conducted investigating needle insertions using flexible steerable needles, which can be controlled by the insertion and rotation speed of the needle [34] and monitored using ultrasound [35]. Sampling-based planners [36] and inverse kinematics [37] have been used to autonomously control flexible steerable needles. While flexible needle steering and DALK needle insertions have similar goals, advance a needle to a desired position, the relative scale of the needle to the tissue is vastly different. Flexible needle insertions are often performed on organs much larger than the needle such as the liver or breast. In DALK, the needle diameter is close to 80 % of the corneal thickness (Fig. 1), so changes in the needle position can have a large effect on the shape of the cornea.

Automating other common surgical procedures such as suturing and removal of undesirable tissue is another area of active research. Examples of automated tissue removal include simulated brain tumor removal using a cable driven robot [38], squamous cell carcinoma resection [39], and ablation of kidney tumors [40]. A semi-autonomous robotic system developed by Shademan *et al.* [41] used near-infrared markers and a commercial suturing tool to perform in vivo end-to-end anastomosis of porcine intestine. Others have automated these procedures using examples generated by humans performing the task in a process called learning from demonstration (LfD). Notable surgical LfD successes include tying knots faster than humans [42], suturing [43], cutting circles, and removing spherical tumors in tissue phantoms [44].

Learning from demonstration enables a robot to learn to perform a task without requiring an expert to explicitly program each action or movement. This is beneficial for complex tasks when an expert can perform the task, but not describe how to perform the task in sufficient detail for reproduction. However, LfD can result in suboptimal behavior if the demonstrations are low quality or the environment is substantially different from the one in which the demonstrations were collected. Improving beyond behaviors learned from demonstration is a common objective in LfD and has been accomplished using various methods [42], [45], including reinforcement learning (RL) [46]–[48]. Deep reinforcement learning has recently been used to exceed human performance in challenging game-playing domains such as Atari [47], [49] and Go [50], [51]. While most research on deep reinforcement learning has been conducted in simulation, some work has successfully applied these techniques to real world environments. Examples include robots learning to open doors [52], picking office supplies from a bin [53], and fitting shapes into matching holes [46], [54].

Using a real world model of the challenging domain of ophthalmic microsurgery, we present methods for ex vivo autonomous robotic DALK needle insertions under OCT guidance using deep deterministic policy gradients from demonstrations (DDPGfD) [46]. DDPGfD is an off-policy reinforcement learning algorithm which employs two neural networks, an actor network and a critic network. The actor produces a policy, which maps the state of the environment (the location of the needle relative to the corneal surfaces) to actions (how to move the needle), while the critic takes state and action as input and determines the value of taking the action in the state. We chose this method over on-policy RL techniques such as proximal policy optimization [55] or trust region policy optimization [56] because DDPGfD easily allows the use of expert demonstrations when learning a policy and has been shown to work in real world environments [46]. Intraoperative OCT provides volumetric imaging with micrometer resolution during the procedure, real-time corneal surface segmentation and needle tracking [20] quantitatively monitors the procedure, and an industrial robot enables precise positioning of the needle. We leverage expert demonstrations to seed the initial behavior and decrease the time required to learn while allowing for improvement and generalization beyond the demonstrations via reinforcement learning.

A secondary goal of this work is to demonstrate the feasibility of using deep reinforcement learning as a control method for robotic ophthalmic microsurgery. DDPGfD offers advantages over more traditional approaches for controlling the robot, e.g., a PID controller or motion planning. A PID controller, or simple linear motion, may be able to move the needle to the desired position in some instances, but could not compensate for corneal deformation, which can be large given the relative size of the needle to the cornea. Motion planning could account for corneal deformation, but to do so requires physical modeling of how the cornea will deform and how the needle will move inside the cornea. Obtaining models with the required accuracy to support motion planning would be difficult, and any new cornea or needle that deviates from the models could cause errors. By using reinforcement learning, we permit motions more complex than simple lines while removing the need to model the environment, and by learning from demonstrations we allow the surgeon to transfer their skill to the robot in a natural manner.

II. METHODS

In this section, we first review the theory behind deep deterministic policy gradients from demonstration. Next, we explain the state space, action space, reward function, and network architecture used in this work. We then describe the numerical simulation we performed to determine hyperparameter values which minimized the number of episodes required to learn the task. Finally, we describe our robotic ex vivo insertion experiment.

A. Reinforcement learning framework

We followed the DDPGfD framework [46], [57], [58] and modeled our environment as a Markov Decision Process (MDP) [59] with a continuous state space $s \in S$, continuous action space $a \in A$, transition function to determine the next state $s' = T(s, a)$, and reward function $r(s, a)$. The critic network, $Q(s, a)$, determined the value of a state-action pair and was parameterized by a weight vector θ. The actor network, $\mu(s)$, determined the action the agent took in state s and was parameterized by a weight vector ϕ. The goal of DDPGfD, as in Q-learning, is to find the optimal policy, $\mu^*(s)$, by finding the optimal Q-function, $Q^*(s, a)$. An optimal policy maximizes the total discounted reward $R = \sum_{i=0}^{\infty} \gamma^i r(s_i, a_i)$, where γ is a discount factor between zero and one. An optimal Q-function/policy is found by updating the weights of the actor and critic networks based on state, action, reward, next state transitions (s, a, r, s') experienced in the environment and stored in a replay buffer.

Both actor and critic networks utilized target networks [49], [60], $Q'(s, a)$ parameterized by θ' and $\mu'(s)$ parameterized by ϕ', respectively, when updating their weights. This helped stabilize learning. Initially, target networks were copies of the actor and critic networks and were periodically updated to match their non-target counterparts. Critic network updates minimized a loss based on the Bellman equation [61] using N transitions from the replay buffer and the target networks, as shown in (1) [58];

$$\begin{aligned} b_i &= r_i + \gamma Q'\big(s'_i, \mu'(s'_i)\big) \\ \underset{wrt\ \theta}{\text{minimize}}\ & \frac{1}{N} \sum_{i=0}^{N} \big(b_i - Q(s_i, a_i)\big)^2. \end{aligned} \tag{1}$$

Because the action space was continuous, it was not feasible to enumerate all possible actions to find the action that maximized $Q(s, a)$. Instead, the actor network was updated to minimize the negative value (i.e. maximize) the mean value of $Q(s, a)$ for the N sampled transitions, as shown in (2) [58];

$$\underset{wrt\ \phi}{\text{minimize}}\ -\frac{1}{N} \sum_{i=0}^{N} Q\big(s_i, \mu(s_i)\big). \tag{2}$$

We incorporated expert demonstrations by training the actor and critic networks on (s, a, r, s') demonstration tuples before performing online learning [62].

These tuples stayed in the replay buffer during online learning and have shown to accelerate learning [46], [47]. However, the demonstrations only covered a small subset of possible states and actions leaving most of the state/action space unexplored. Because of this, state action pairs that had not been visited could have received erroneously high Q-values when using only the loss functions in (1) and (2) to learn. To mitigate this problem, others have used a classification loss for the Q-function [47], or a behavior cloning (BC) loss for the actor network [63] when training from demonstrations. We used a behavior cloning loss in this work given by

$$L_{BC} = \frac{1}{N} \sum_{i=0}^{N} \big(\mu(s_i) - H(s_i)\big), \tag{3}$$

where $H(s_i)$ is the action the human took in state s_i. The final loss functions for the critic and actor networks used in

this work were a weighted sum of multiple individual losses. The critic and actor losses were

$$\begin{aligned} L_{critic} &= \lambda_{w_Q} L_{w_Q} + \lambda_Q L_Q \\ L_{actor} &= \lambda_{w_\mu} L_{w_\mu} + \lambda_{BC} L_{BC} + \lambda_\mu L_\mu, \end{aligned} \tag{4}$$

where L_Q is the loss from (1), L_{w_Q} is an L_2-norm loss on the critic network weights, L_μ is the loss from (2), L_{BC} is the behavior cloning loss on the expert demonstrations from (3), L_{w_μ} is an L_2-norm loss on the actor network weights, and the λ variables controlled the relative contribution of each loss.

B. State space, action space, reward function, and extracting demonstrations

1) State and Action Spaces: We kept track of two separate state spaces, one for the transition function of the MDP and one for learning. The transition function state space in the simulation was the x, y, z, pitch and, yaw of the needle tip. Our learning state space consisted of seven values, which are illustrated in Fig. 3. They were: Δx from goal, Δy from goal, Δyaw to face goal, Δpitch to face goal or avoid endothelium, needle percent depth, goal depth minus needle depth (Δdepth from goal), and a corneal deformation value d equal to the sum of squared residuals of a 2nd order polynomial fit to the epithelium of a B-scan along the needle. Indenting or deforming the cornea distorts the view of both embedded and inferior structures, such as surgical instruments and anterior segment anatomy. With the distorted view, the surgeon's understanding of anatomical relationships can be altered and affect successful performance of the procedure, which is why we included this deformation value in our state space.

The action space permitted yaw and pitch changes of $\pm 5°$ (between -20–$20°$ yaw and -5–$25°$ pitch) and movement of the needle between 10–250 µm in the direction it was facing at each time step. These limits where chosen to allow the agent to have precise control of the needle and to minimize deformation when deployed on real tissue.

Sparse reward functions are considerably easier to specify than shaped reward functions and require significantly less tuning. Additionally, when combined with demonstrations, sparse rewards have been shown to outperform shaped rewards [46]. Because of this, we used the following sparse reward function;

$$r(s) = \begin{cases} & \text{if } \left(\frac{\Delta x}{\rho_x}\right)^2 + \left(\frac{\Delta y}{\rho_y}\right)^2 \leq 1 \\ 5, & \text{and } \Delta\text{depth from goal} \leq \epsilon_\% \\ & \text{and } d \leq \epsilon_d \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where ρ_x and ρ_y form an ellipse around the goal in the x–y plane, $\epsilon_\%$ is a percentage depth threshold, and ϵ_d is a deformation threshold. We used an elliptical goal, as opposed to a circular goal, because the length of the insertion (in our setup along the x-axis) is more important than any displacement from the apex along the axis perpendicular to the insertion (in our setup along the y-axis). The corneal deformation threshold was set by a corneal surgeon after viewing multiple B-scans of needles in corneas and the associated deformation metric. If the agent received a positive reward, the episode terminated. However, there were five other conditions that also resulted in the termination of an episode: a percent depth greater than 100 (needle perforation), a percent depth less than 20, a Δx greater/less than zero (needle past the goal depending on insertion direction), a Δy greater than ϵ_y, or the number of steps in the episode exceeded a threshold. These additional termination conditions were added because they are either failure cases in DALK (needle perforation, needle too shallow, past goal), or to prevent unnecessarily long episodes. Thresholds for our reward and termination conditions were: $\rho_x = 0.35$ mm, $\rho_y = 0.5$ mm, $\epsilon_\% = 5\%$, $\epsilon_d = 6\,\text{mm}^2$, $\epsilon_y = 1.2$ mm, and a maximum of 25 steps per episode.

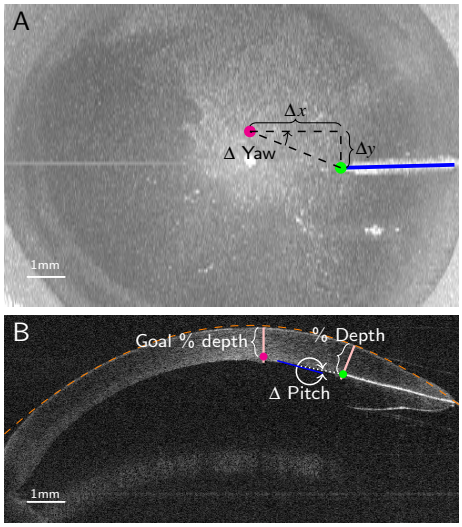

Fig. 3. Needle insertion state space. (A) En face OCT view of a needle in cornea. (B) OCT Cross sectional view along the needle's axis. The green dot represents the needle tip and the magenta dot represents the goal. The blue line represents the tangent line at the point where the needle would perforate the endothelium. The dashed orange line represents the 2nd order fit of the epithelium segmentation and was used to compute corneal deformation value d.

2) Extracting demonstrations: In our previous work [20], we recorded the volumetric OCT time series of ex vivo DALK needle insertions performed by corneal surgical fellows. To utilize this data as demonstrations for all of our experiments, we segmented the epithelial/endothelial surfaces and tracked the needle using our automatic algorithm for all volumes in each time series from this previous data. Then, we fit a cubic spline to the x, y, and z position of the tracked needle over time, thereby smoothing the surgeon's trajectories and enabling us to map the demonstrations onto our restricted action space. We divided the spline into segments no longer than 250 µm and for each segment endpoint we found a volume in the time series where the original needle position was closest to the spline fit. These needle position/volume pairs provided us with a smoothed trajectory that also approximated the deformation of the cornea. The last needle position in the smoothed trajectory was used to set the x–y goal location and the goal depth was set to 85% for each demonstration. Finally, we computed the (s, a, r, s') tuples for each trajectory. States and next states were computed from the needle/volume pairs.

To find the actions at each state, we computed the distance, change in yaw, and change in pitch, between the current needle tip and the needle tip of the next needle/volume pair. Rewards for state action pairs were computed using (5).

3) Network architecture and hyper-parameters: We utilized the network architecture described previously [58]. The critic network consisted of two hidden layers with 400, and 300 units with rectified linear activation functions. Actions in the critic network were included after the first hidden layer. The actor network used a similar architecture as the critic network, but had tanh activations for each of the outputs, bounding the actions between -1 and 1. States and actions were normalized between $[-1, 1]$ based on scan dimensions, maximum/minimum angles etc. before passing through a network.

We used an ADAM optimizer [64] for training with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-3}$, a 10^{-4} learning rate for the critic, and a 10^{-4} learning rate for the actor. After each step in online learning, we randomly sampled a batch of 256 transitions from the replay buffer, which included the expert demonstrations, and trained for 25 epochs [46], with a discount factor $\gamma = 0.95$. Target networks were updated every 50 epochs. We used $\lambda_{w_Q} = 5 \times 10^{-3}$, $\lambda_Q = 1$, $\lambda_{w_\mu} = 10^{-4}$, $\lambda_{BC} = 2$, and $\lambda_\mu = 1$ for the loss weights. The simulation and reinforcement learning code were written in Python 3.5 using Tensorflow [65].

4) Simulated needle insertions: Prior to performing ex vivo needle insertions, we ran numerically simulated needle insertions to test the feasibility of using DDPGfD for this task, and to determine hyper-parameter values that minimized the number of episodes required to learn the task.

Our simulated environment consisted of a static height map of two corneal surfaces (epithelium and endothelium) and a needle, modeled as a point. The simulated area was 12 mm (x) by 8 mm (y) by 6 mm (z). We encoded the goal state as an offset from the apex of the endothelial surface of the cornea and a desired depth expressed as a percentage of corneal thickness. In all experiments we used a goal offset of 0.75 mm in x and a goal depth of 85 %. An illustration of the simulation environment is shown in Fig. 4.

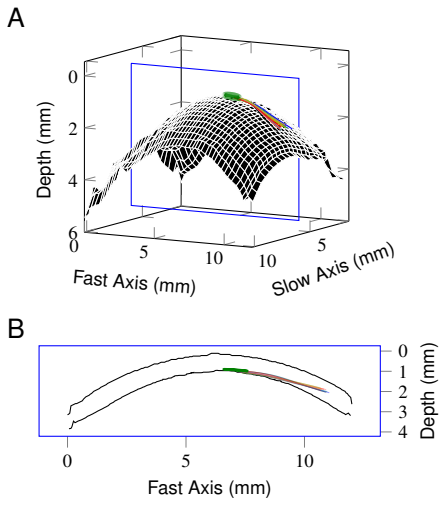

Fig. 4. Simulation environment with needle trajectories (colored lines) and goal area (green ellipse). (A) 3D view of the endothelium (black mesh) with plotted trajectories. The epithelium was used in the simulation, but is not shown here so the trajectories can be seen. (B) B-scan view of the simulated environment at the location denoted by the blue rectangle in (A). Trajectories were projected on to the slow axis before plotting.

Algorithm 1: DDPGfD for needle insertions.

Initialize Q, Q', μ, and μ' with weight vectors θ and ϕ
Fill replay buffer rb with (s, a, r, s') tuples from successful human demonstrations
Train Q, Q', μ, and μ' minimizing L_{critic} and L_{actor} for 500 epochs
for $i <$ *episodes* **do**
  Initialize environment
  $s \leftarrow$ random start state
  while *!terminate* **do**
    $a \leftarrow \mu(s)$
    if *rand(0,1)* $\leq$ *0.1* **then**
      $a \pm$ 10 %
    end
    $r, terminate \leftarrow R(s, a)$
    $s' \leftarrow T(s, a)$
    $s' \leftarrow s' +$ Noise
    Add (s, a, r, s') to rb
    Train Q, Q', μ, and μ' on minibatch from rb
    $s = s'$
    if *epochs mod* $50 = 0$ **then**
      $\theta' \leftarrow \theta$ and $\phi' \leftarrow \phi$
    end
  end
end

Prior to online learning, we trained the actor and critic networks with 20 successful human demonstrations provided by corneal surgical fellows for 500 epochs. After this pre-training, we ran 250 simulated insertion episodes. For all experiments, we assumed the needle had already been inserted into the cornea. The starting needle location for each episode was determined by first randomly choosing a point along a 20° arc 3.5–4.0 mm away from the goal. Then, a starting depth was randomly chosen between 40–60 % Finally, we randomly selected yaw and pitch angles that were within ±5° and ±2.5° of facing the goal. Following the methodologies presented in prior work [46], [58], we ran our simulation according to Algorithm 1. At each time step we computed the current state, ran an ϵ-greedy policy by adding uniform noise sampled from ±0.1 of the max action [63], executed the action, computed the next state with noise, and computed the reward/termination. We added $\mathcal{N}(0,\ 0.01\,\text{mm})$ to the needle's x, y, and z position, $\mathcal{N}(0,\ 0.1°)$ to the needle's pitch, $\mathcal{N}(0,\ 0.3°)$ to the needle's yaw, $\mathcal{N}(0,\ 0.026\,\text{mm})$ to the epithelium, and $\mathcal{N}(0,\ 0.032\,\text{mm})$ to the endothelium. These noise statistics were obtained from our prior work [20] and were an impediment to learning given the cornea is ∼500 µm thick. Every five episodes we froze the weights of the actor network and ran the current policy for 45 episodes on three simulated validation corneas (15 each) that had not been used in training or in demonstrations. We

ran the simulation ten times to account for different random initial network states.

C. Ex vivo human cornea robotic needle insertions

Following simulated needle insertions, DDPGfD was performed using an ex vivo model of DALK with human cadaver corneas and an industrial robot to position the needle.

1) System description: Our ex vivo system, pictured in Fig. 5A, consisted of an OCT system, an IRB 120 industrial robot (ABB, Zurich, Switzerland) with a custom 3D printed handle holding the needle, and a Barron artificial anterior chamber (Katena, Denville, NJ). Donor corneas, provided by Miracles in Sight Eye Bank (Winston-Salem, NC), were mounted on the artificial anterior chamber and pressurized with balanced salt solution. A 27-gauge needle was attached to the end of the custom handle and bent upward with the bevel facing down to prevent the handle from hitting the table when angling the needle toward the corneal apex. The use of ex vivo corneas for this study was approved by the Duke University Health System Institutional Review Board. The custom-built OCT system utilized a 100 kHz swept-source laser (Axsun, Technologies Inc., Billerica, MA) centered at 1060 nm. OCT imaging used a raster scan acquisition pattern with volume dimensions of 5.47 mm $\times$ 12.0 mm $\times$ 8.0 mm, consisting of 990 samples/A-scan, 688 A-scans/B-scan (500 usable A-scans/B-scan), and 96 B-scans/volume, which provided a volumetric update rate of 1.51 Hz.

2) System calibration: Because our action space required the robot to move and rotate about the needle tip, and we wanted to move the needle relative to the cornea to begin episodes, we needed to transform points back and forth between the OCT coordinate frame and the robot coordinate frame. Moving about the needle tip required us to find a 4×4 end-effector to needle tip transform, T_{N}, and moving the needle relative to the cornea required us to find the 4×4 world origin to OCT origin transform, T_{OCT}. To find these two transforms, we moved the robot to eight different points inside the OCT field of view with various pitch and yaw values. At each point we recorded the robot's end-effector position and the needle tip/base position in the OCT volume using our previously described needle tracking [20]. Once we collected the eight points, we simultaneously found T_{N} and T_{OCT} by minimizing

$$\sum_{i}^{8} \left\| T_{\text{OCT}}^{-1} T_{\text{EE}_i} T_{\text{N}} \begin{bmatrix} \vec{0} \\ 1 \end{bmatrix} - \begin{bmatrix} \mathbf{t_i} \\ 1 \end{bmatrix} \right\|_2^2 + \left\| T_{\text{OCT}}^{-1} T_{\text{EE}_i} T_{\text{N}} \begin{bmatrix} \mathbf{m} \\ 1 \end{bmatrix} - \begin{bmatrix} \mathbf{b_i} \\ 1 \end{bmatrix} \right\|_2^2, \tag{6}$$

where T_{EE} is the robot end-effector transform, $\mathbf{m}$ is the vector $[-12, 0, 0]^T$ representing the length in millimeters of the needle, $\mathbf{t}$ is the x, y, z needle tip position vector, and $\mathbf{b}$ is the x, y, z needle base position vector [66], [67].

3) Safety: While we used expert demonstrations to seed the robot's initial behavior, exploration of the state/action space was still necessary to improve the robot's success rate. Because of this need for exploration, the robot could potentially perform unreasonable or unsafe actions. To prevent any unsafe behavior, we implemented multiple safety constraints for this system utilizing software and human oversight. We set software Cartesian and angular velocity limits for the tool to $10\,\text{cm/s}$ and $\frac{\pi}{4}\,\text{rad/s}$. Each joint on the robot was set to have an angular velocity limit of $\frac{\pi}{2}\,\text{rad/s}$. We used 3D models of the environment and forward kinematics of the robot to prevent the robot from colliding with itself or other objects in the environment, such as the table and the microscope. If any part of the modeled robot was determined to be within 1 cm of another part of the modeled environment the robot stopped moving. Our reinforcement learning action space only permitted 250 µm movements at each step and the intended

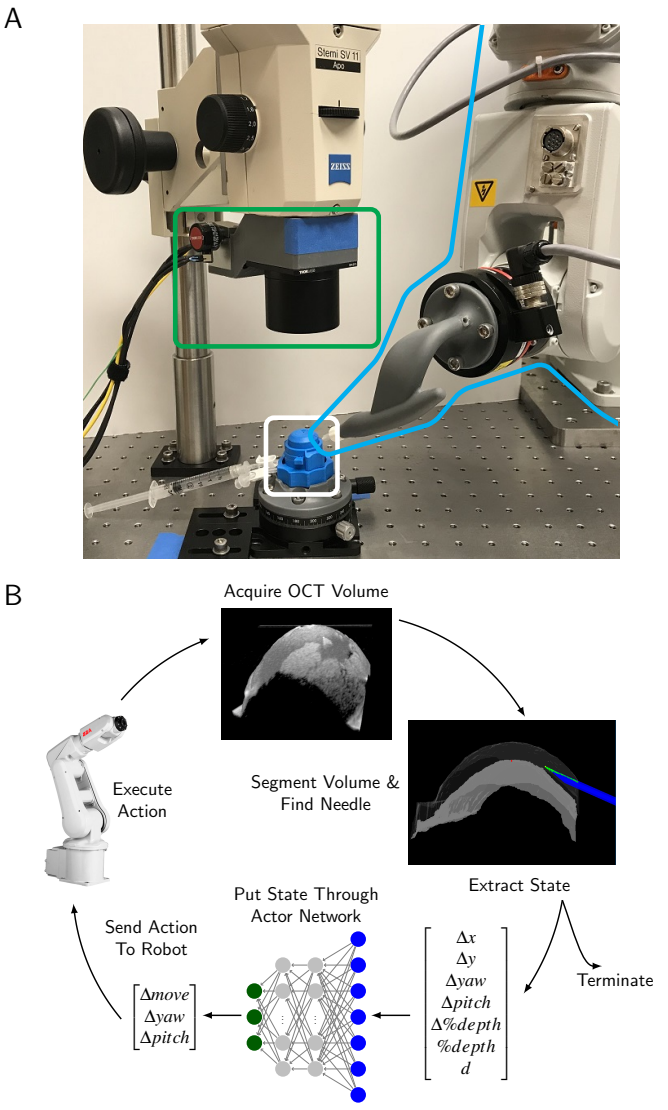

Fig. 5. Ex vivo human cornea robotic needle insertion system and flow diagram. (A) IRB 120 industrial robot with custom 3D printed handle holding a 27-gauge needle is outlined in cyan. The artificial anterior chamber is outlined in white. The OCT scanner coupled into the optics of a stereo zoom microscope is outlined in green. (B) Ex vivo episode flow diagram. The system acquired a volume, segmented it, and produced the state. The state and previous action were checked to see if the episode should terminate. If the episode continued, the state was passed through the actor network to obtain the action the robot should execute. After the robot executed the action, the process repeated.

action and anticipated location of the robot after performing the action was shown to an operator on a user interface. Finally, the robot could only move in response to the operator pressing a button on the user interface.

4) Line planner experiment: To provide a baseline with which to compare our reinforcement learning approach, we performed DALK needle insertions using a simple line planner. After inserting the needle in the cornea, the line planner advanced the needle in a straight line toward a goal position above the apex of the endothelium. In our experiments we used a goal depth of 90 % and performed 18 insertions using six corneas.

5) Reinforcement learning experiment: Each episode contained two phases; the initial insertion and needle advancement to the goal. Our state and action spaces were created for advancing the needle to the goal, which required us to program a semi-automatic initial insertion routine for the ex vivo trials. We began our semi-automatic insertion routine by randomly finding a starting point along a 20° arc 4.0–4.5 mm away from the goal. Then, we moved the robot outside the cornea facing the starting location at a 15° downward tilt. Next, we advanced the needle toward the starting location. At this point, the operator would inspect the OCT volume and optionally advance the needle further to ensure it was sufficiently embedded in the cornea. We then attempted to recreate the same starting conditions in the human corneas that we used in simulation (starting pitch ±2.5° of facing the goal, starting yaw ±5° from facing the goal, and depth between 40–60 %), but due to calibration errors, varying cornea stiffness, and varying anterior chamber pressures, the resulting starting yaws, pitches, and depths varied substantially more than in simulation.

After the completion of the initial insertion, the reinforcement learning policy was used to advance the needle toward the goal. The advancement toward the goal phase, depicted in Fig. 5B, started by collecting one OCT volume of the environment, which was then segmented to find the epithelial/endothelial surfaces and the needle position/orientation. The segmented surfaces and needle position/orientation were converted into our state space representation then passed through the actor network to produce an action, which was finally executed on the robot. After the robot completed the action, another volume was acquired and segmented to determine the reward. This process repeated until the episode terminated.

We trained the actor and critic networks for 500 epochs on 20 successful surgical fellow demonstrations and then trained the robot on 150 ex vivo insertion episodes. The learned policy from the simulation was not used in the ex vivo experiment. To minimize the number of corneas required, we performed approximately 10 insertions on each cornea. We evaluated the policy the agent had learned after a set number of training episodes (0, 50, 100, 125, and 150) by running 10 insertions on two corneas. We spread the 10 insertions over two different corneas (five each) at each of the evaluation checkpoints to test how well the learned policy generalized across corneas.

After 150 episodes, we ran 10 additional insertions using the final policy to obtain a total of 20 data points spread over three corneas. This data was compared to the performance achieved by corneal surgical fellows using a dataset obtained in a previously described experiment [20]. Briefly, three fellows were asked to perform 20 DALK needle insertions each on ex vivo human corneas mounted on an artificial anterior chamber. The fellows were instructed to position their needle to 80–90 % depth using only an operating microscope (10 insertions) or an operating microscope and OCT images displayed on a monitor next to the microscope (10 insertions). The OCT image provided was a tracked cross section along the axis of the needle labeled with the needle's calculated percent depth in the cornea.

We compared the mean and variance of the perforation-free final percent depth and the deformation between the line planer and final learned policy of RL. The variances were compared using Levine's test [68] and the means were compared using

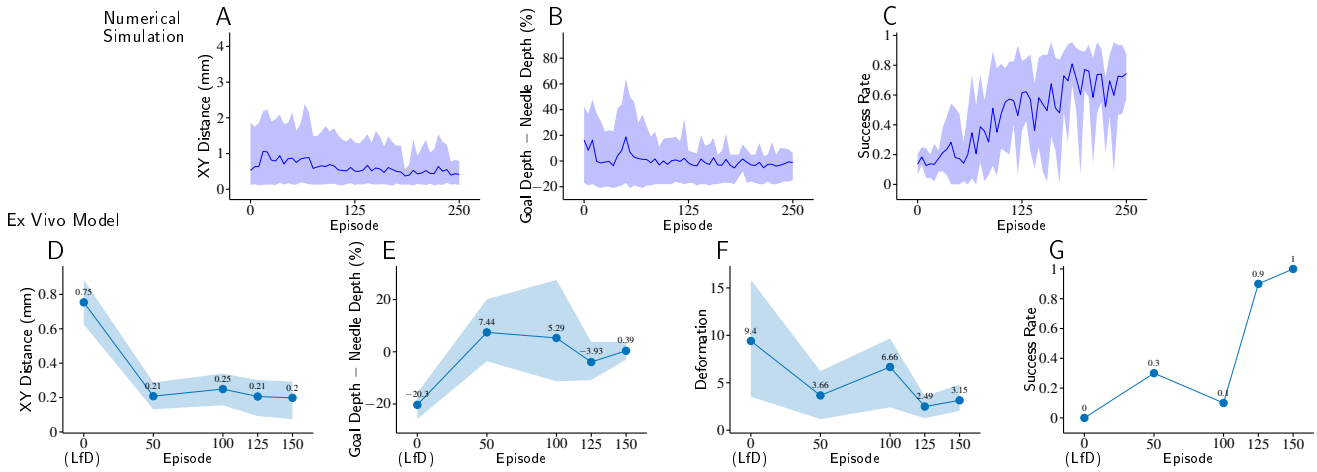

Fig. 6. Learning curves for simulated (A-C) and ex vivo (D-G) needle insertions. Means are denoted by solid lines and shaded regions represent the 90th and 10th percentiles. (A) The final distance between goal and the needle in the x–y plane. (B) The final difference between the goal depth and the needle depth expressed as a percentage of corneal thickness. (C) The success rate of the policy. (D) The final distance between goal and the needle in the x–y plane. (E) The final difference between the goal depth and the needle depth expressed as a percentage of corneal thickness. (F) The sum of squared residuals of a 2nd order polynomial fit to the epithelium, a measure of corneal deformation. (G) The success rate of the policy.

a two-sided T-test (deformation) and Welch's t-test (depth) [69]. The depths and deformation values were obtained via automatic segmentation. To compare the performance of the fellows to the robot using RL, we determined final perforation-free needle depths via manual segmentation. We used a two-sided T-test to determine if there was a significant difference in the mean between the final perforation-free percent depth obtained by the robot using RL compared to the surgical fellows using OCT, and we used Welch's t-test [69] to determine if there was a significant difference in the mean final depth between the robot using RL and the fellows who only used the microscope.

III. RESULTS

Learning curves for the numerically simulated needle insertions are shown in Fig. 6A-C, while learning curves for the ex vivo needle insertions are shown in Fig. 6D-G. The reported metrics (success, distance, depth, and deformation) for the ex vivo insertions were obtained from automatic segmentation and tool tracking during the episode. In the ex vivo experiments the agent had increasing success and reduced variability in percent depth and deformation as the number of training episodes increased. Due to the constraints and limits we imposed on the robot, there were no adverse safety events during the ex vivo learning process.

The line planner was able to reach the goal depth with an acceptable level of deformation eight times. In nine other trials either the deformation was too large, or the needle was not deep enough to be considered a successful trial. The line planner punctured the endothelium once. The mean perforation-free final percent depth and deformation (obtained via automatic segmentation) for the line planner was 80.34 % $\pm$ 7.56 % and 5.34 $\pm$ 2.86 ($N = 17$), compared to 84.16 % $\pm$ 4.01 % and 2.72 $\pm$ 1.03 ($N = 20$) for RL. The difference in mean for the final percent depth between RL and the line planner was not statistically significant ($p = 0.08$), but the difference in variance was ($p = 0.03$). The difference in mean deformation between RL and the line planner was statistically significant ($p = 0.001$), but the difference in variance was not ($p = 0.06$). Learning from demonstration did not have any perforation-free trials ($N = 10$) and had an average deformation of 9.41 $\pm$ 5.87.

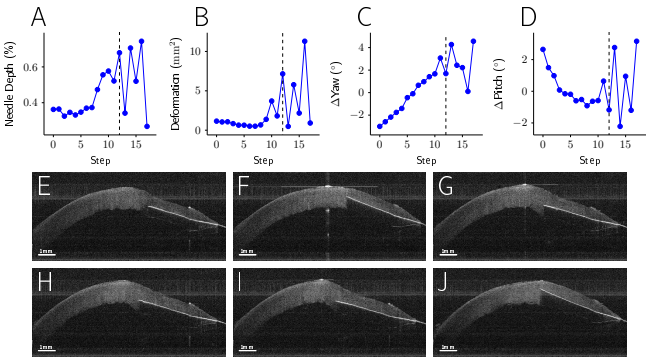

Fig. 7. Failure case after 100 training episodes. (A-D) Graphs of needle depth as a percentage of corneal thickness, corneal deformation, change in yaw, and change in pitch for each step in the episode. The dashed line in each graph represents where the images in (E-J) begin. (E-J) OCT cross sections along the axis of the needle for the final six steps in the episode. The large changes in pitch caused deformation of the cornea and a shallow final needle depth.

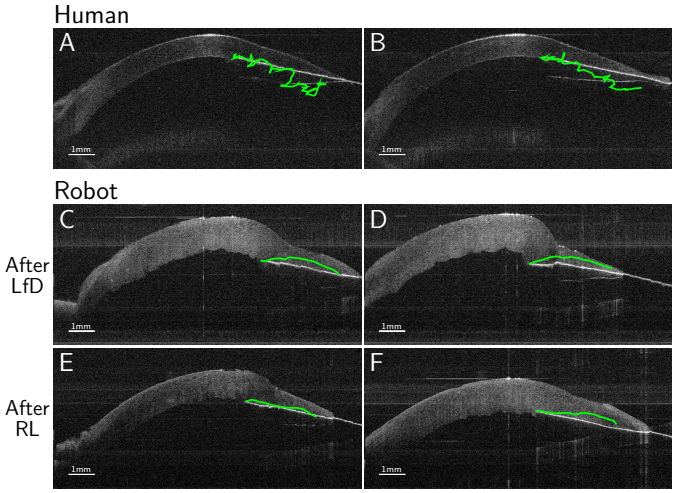

Fig. 8. Comparison of needle motions between surgical fellow and the robot. (A-F) OCT cross sections along the axis of the needle at the end of needle insertions. Green represents the path (projected onto 2D) of the needle tip throughout the insertion. (A-B) Insertions performed by surgical fellow. (C-D) Insertions performed by the robot after learning from demonstration. (E-F) Insertions performed by the robot after 150 reinforcement learning training episodes.

After succeeding in four trials at the 50 episode checkpoint, the agent's performance decreased at the 100 episode mark. A major reason for this decrease in performance was due to the agent's behavior near the goal. Once the needle approached the goal, the agent began to take actions with large changes in pitch and yaw. A large pitch change in one direction would then be followed by a large change in pitch in the opposite direction. This oscillation near the goal caused substantial corneal deformation, or the needle becoming too shallow in the cornea, resulting in a failed episode. This undesirable behavior is depicted in Fig. 7. The magnitude of swings in pitch near the goal decreased between episode 100 and 150. At the 125 training episode checkpoint only one episode failed due to perforation of the endothelium from a large negative change in pitch.

Fig. 8 shows OCT cross sections along the needle's axis at the end of insertions. Overlaid on top of these images is the path of the needle tip projected onto 2D. Fig. 8A-B shows two successful insertions performed by a surgical fellow used as training examples for LfD. Fig. 8C-D shows two insertions performed after learning from demonstration and Fig. 8E-F shows two insertions after 150 training episodes using the final learned policy. The policy before any online learning (LfD) advanced the needle toward a point below the endothelial apex, which caused significant unwanted deformation of the cornea. This deformation caused the automatic segmentation to prematurely and incorrectly report perforations in some trials. Our segmentation method assumed the epithelial and endothelial corneal surfaces would be smooth. When the cornea became too deformed the segmentation would underestimate the deformation, but needle tracking would accurately

report the needle's position causing the needle depth to be reported as greater than 100 %. Fortunately, these failures in segmentation did not substantially impact learning because the actions causing incorrect segmentation were undesirable anyway due to the deformation they caused. In contrast, using the final learned policy (after 150 training episodes) the robot was able to position the needle at the correct depth with minimal deformation.

The mean and standard deviation of the final perforation-free percent depth obtained by the robot using the learned policy was 84.75 % ± 4.91 % ($N = 20$) compared to 78.58 % ± 6.68 % for the surgical fellows using OCT ($N = 28$) and 61.38 % ± 17.18 % for the surgical fellows using only the operating microscope ($N = 15$). The difference in mean final percent depth between the robot and fellows using OCT was statistically significant ($p = 0.001$) as was the difference between the robot and fellows not using OCT ($p = 0.0001$). These statistics are illustrated in Fig. 9.

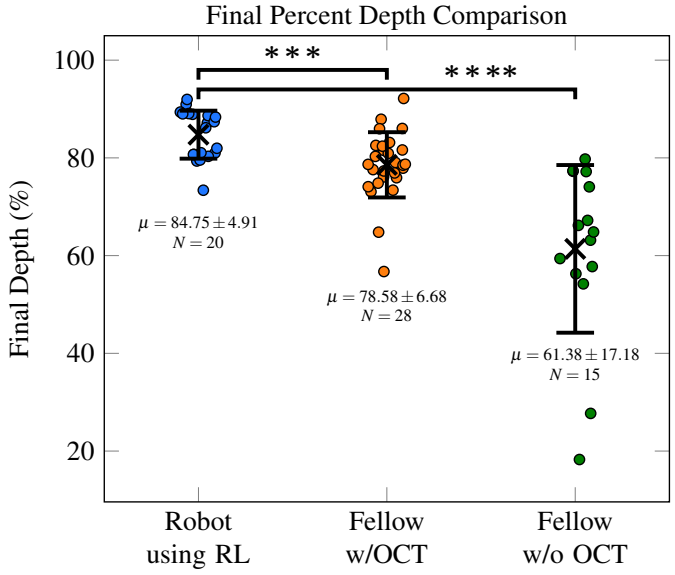

Fig. 9. Comparison of the final needle depth expressed as a percent of corneal thickness for perforation-free episodes between the robot using the final learned policy and corneal surgical fellows. A black X indicates the mean of the group and error bars denote one standard deviation.

IV. DISCUSSION

In this work our robot learned to insert a needle in tissue to a specific depth better than surgical fellows by using reinforcement learning from demonstration under OCT guidance. The learned policy achieved the desired results with minimal deformation to the tissue and generalized over multiple corneas. Behavior cloning expert demonstrations initialized the agent's behavior but did not prevent the agent from improving upon the demonstrations. Simpler methods for controlling the robot were either never successful (behavior cloning), or not as successful as reinforcement learning (line planner). While the line planner was able to successfully complete a trial in some cases, it was less consistent in achieving a target depth and had significantly more deformation than the reinforcement learning method.

We used a hand-crafted state space to reduce the number of episodes required to learn a suitable policy. Height maps of the two corneal surfaces and the tool could instead be used as state input to actor and critic networks removing the need to create a state space representation by hand. The most general approach would be to pass the entire OCT volume through a convolutional neural network which produces actions as the output. Combining perception and control into a single network has been shown to improve performance over separate perception and control networks [54]. However, this general approach would most likely require an excessive and infeasible number of corneal samples to learn a reasonable policy, even with expert demonstrations.

In this work we only used successful demonstrations from experts in the replay buffer and when applying a behavior cloning loss. Any set of demonstrations, successful or unsuccessful, can be used in the replay buffer for DDPGfD. To reduce the likelihood of unwanted behavior when learning, such as that exhibited after 100 trials (Fig. 7), one could put transitions from unsuccessful demonstrations in the replay buffer and/or add an additional loss function that penalizes the agent for taking the undesired actions in certain states.

The goal of "big bubble" DALK is to separate DM from the stroma by pneumodissecting those layers. Ideally, the reward from an insertion would be determined by whether pneumodissection occurred, rather than a proxy metric (needle depth) as used here. Using pneumodissection as a sparse reward might increase the probability of an agent learning a policy leading to bubble formation, but demonstrations would be more costly to obtain because one human cornea would be destroyed per episode. We drastically reduced the number of limited human corneas needed to learn a policy by using the final percent depth of the needle as a proxy for success, since it has been shown to be correlated with successful bubble formation [9].

An ex vivo model of the microsurgical procedure was used in this project because there are obvious medical-legal and ethical constraints to having a robot first learn a microsurgical procedure directly on patients. The presented data demonstrate the promise and potential of RL/LfD in this procedure, but transferring this directly to human surgery may require refinement of the ex vivo models. For example, the ex vivo model used in this work prevented the cornea from moving during the needle insertion, whereas the globe of the eye can move during surgery. A more realistic ex vivo model would allow for eyeball rotation and reproduce the physical constraints of surrounding anatomy (e.g., nose, eye socket). While an improved ex vivo model could never fully replicate in vivo surgery, the model free nature of DDPGfD removes the explicit assumption that the in vivo and ex vivo environments are identical, thus increasing the likelihood of an ex vivo trained agent succeeding in vivo. Improved and more complex eye models combined with RL/LfD could potentially be used to learn other ophthalmic microsurgical procedures such as retinal peels.

We envision that an autonomous DALK needle insertion system could be overseen by surgeons who understand the procedure and its benefits but have not accumulated the years

of experience required to perform the procedures themselves. However, using an autonomous robot in the operating room would require careful planning to ensure the safety of not only the patient, but of the surgeon and assistants as well. Additional safety features in conjunction with constraints on where the robot is allowed to move and how fast it can move may be necessary for use in human surgery. One such safety feature, utilized by Edwards *et al.* [33] was the ability to quickly retract an instrument away from the patient. If the robot is unable to complete the procedure or otherwise fails, the surgeon could intervene and revert the procedure to conventional full-thickness corneal transplantation.

ACKNOWLEDGMENT

The authors would like to thank Miracles in Sight (Winston-Salem, NC) for the use of research donor corneal tissue. We acknowledge research support from the Coulter Foundation Translational Partnership.

REFERENCES

[1] A. George and D. Larkin, "Corneal transplantation: the forgotten graft," *American Journal of Transplantation*, vol. 4, no. 5, pp. 678–685, 2004.

[2] E. Skiadaresi, C. McAlinden, K. Pesudovs, S. Polizzi, J. Khadka, and G. Ravalico, "Subjective quality of vision before and after cataract surgery," *Archives of Ophthalmology*, vol. 130, no. 11, pp. 1377–1382, 2012.

[3] P. K. Gupta, P. S. Jensen, and E. de Juan, "Surgical forces and tactile perception during retinal microsurgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 1999, pp. 1218–1225.

[4] E. Archila, "Deep lamellar keratoplasty dissection of host tissue with intrastromal air injection." *Cornea*, vol. 3, no. 3, pp. 217–218, 1984.

[5] S. P. Dunn, R. L. Gal, C. Kollman, D. Raghinaru, M. Dontchev, C. L. Blanton, E. J. Holland, J. H. Lass, K. R. Kenyon, M. J. Mannis, S. I. Mian, C. J. Rapuano, W. J. Stark, and R. W. Beck, "Corneal graft rejection ten years after penetrating keratoplasty in the cornea donor study," *Cornea*, vol. 33, no. 10, p. 1003, 2014.

[6] R. C. Wolfs, C. C. Klaver, J. R. Vingerling, D. E. Grobbee, A. Hofman, and P. T. de Jong, "Distribution of central corneal thickness and its association with intraocular pressure: The rotterdam study," *American Journal of Ophthalmology*, vol. 123, no. 6, pp. 767–772, 1997.

[7] M. Anwar and K. D. Teichmann, "Big-bubble technique to bare Descemet's membrane in anterior lamellar keratoplasty," *Journal of Cataract & Refractive Surgery*, vol. 28, no. 3, pp. 398–403, 2002.

[8] V. Scorcia, M. Busin, A. Lucisano, J. Beltz, A. Carta, and G. Scorcia, "Anterior segment optical coherence tomography–guided big-bubble technique," *Ophthalmology*, vol. 120, no. 3, pp. 471–476, 2013.

[9] N. D. Pasricha, C. Shieh, O. M. Carrasco-Zevallos, B. Keller, D. Cunefare, J. S. Mehta, S. Farsiu, J. A. Izatt, C. A. Toth, and A. N. Kuo, "Needle depth and big-bubble success in deep anterior lamellar keratoplasty: An ex vivo microscope-integrated OCT study," *Cornea*, vol. 35, no. 11, pp. 1471–1477, 2016.

[10] V. M. Borderie, O. Sandali, J. Bullet, T. Gaujoux, O. Touzeau, and L. Laroche, "Long-term results of deep anterior lamellar versus penetrating keratoplasty," *Ophthalmology*, vol. 119, no. 2, pp. 249–255, 2012.

[11] D. Smadja, J. Colin, R. R. Krueger, G. R. Mello, A. Gallois, B. Mortemousque, and D. Touboul, "Outcomes of deep anterior lamellar keratoplasty for keratoconus: learning curve and advantages of the big bubble technique," *Cornea*, vol. 31, no. 8, pp. 859–863, 2012.

[12] U. K. Bhatt, U. Fares, I. Rahman, D. G. Said, S. V. Maharajan, and H. S. Dua, "Outcomes of deep anterior lamellar keratoplasty following successful and failed 'big bubble'," *British Journal of Ophthalmology*, vol. 96, no. 4, pp. 564–569, 2012.

[13] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography," *Science*, vol. 254, no. 5035, p. 1178, 1991.

[14] HAAG-STREIT Group, "iOCT," 2018, accessed 1 October 2018, https://www.haag-streit.com/haag-streit-surgical/products/ophthalmology/ioct/.

[15] Carl Zeiss Meditec Inc., "OPMI LUMERA 700 - Surgical Microscopes - Retina - Medical Technology — ZEISS United States," 2018, accessed 1 October 2018, https://www.zeiss.com/meditec/us/products/ophthalmology-optometry/retina/therapy/surgical-microscopes/opmi-lumera-700.html.

[16] Leica Microsystems, "EnFocus - Product: Leica Microsystems," 2018, accessed 1 October 2018, https://www.leica-microsystems.com/products/optical-coherence-tomography-oct/details/product/enfocus/.

[17] Y. K. Tao, S. K. Srivastava, and J. P. Ehlers, "Microscope-integrated intraoperative OCT with electrically tunable focus and heads-up display for imaging of ophthalmic surgical maneuvers," *Biomedical Optics Express*, vol. 5, no. 6, pp. 1877–1885, 2014.

[18] S. Binder, C. I. Falkner-Radler, C. Hauger, H. Matz, and C. Glittenberg, "Feasibility of intrasurgical spectral-domain optical coherence tomography," *Retina*, vol. 31, no. 7, pp. 1332–1336, 2011.

[19] O. Carrasco-Zevallos, B. Keller, C. Viehland, L. Shen, G. Waterman, B. Todorich, C. Shieh, P. Hahn, S. Farsiu, A. Kuo, C. Toth, and J. Izatt, "Live volumetric (4D) visualization and guidance of in vivo human ophthalmic surgery with intraoperative optical coherence tomography," *Scientific Reports*, vol. 6, p. 31689, 2016.

[20] B. Keller, M. Draelos, G. Tang, S. Farsiu, A. N. Kuo, K. Hauser, and J. A. Izatt, "Real-time corneal segmentation and 3D needle tracking in intrasurgical OCT," *Biomedical Optics Express*, vol. 9, no. 6, pp. 2716–2732, 2018.

[21] L. F. Hotraphinyo and C. N. Riviere, "Three-dimensional accuracy assessment of eye surgeons," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, 2001, pp. 3458–3461.

[22] F. Peral-Gutierrez, A. Liao, and C. Riviere, "Static and dynamic accuracy of vitreoretinal surgeons," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, 2004, pp. 2734–2737.

[23] S. Weber, K. Gavaghan, W. Wimmer, T. Williamson, N. Gerber, J. Anso, B. Bell, A. Feldmann, C. Rathgeb, M. Matulic, M. Stebinger, D. Schneider, G. Mantokoudis, O. Scheidegger, F. Wagner, M. Kompis, and M. Caversaccio, "Instrument flight to the inner ear," *Science Robotics*, vol. 2, no. 4, p. eaal4916, 2017.

[24] M. Menon, A. Shrivastava, A. Tewari, R. Sarle, A. Hemal, J. O. Peabody, and G. Vallancien, "Laparoscopic and robot assisted radical prostatectomy: establishment of a structured program and preliminary analysis of outcomes," *The Journal of Urology*, vol. 168, no. 3, pp. 945–949, 2002.

[25] V. B. Kim, W. H. Chapman Iii, R. J. Albrecht, B. M. Bailey, J. A. Young, L. W. Nifong, and W. R. Chitwood Jr, "Early experience with telemanipulative robot-assisted laparoscopic cholecystectomy using da vinci," *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, vol. 12, no. 1, pp. 33–40, 2002.

[26] R. H. Taylor, B. D. Mittelstadt, H. A. Paul, W. Hanson, P. Kazanzides, J. F. Zuhars, B. Williamson, B. L. Musits, E. Glassman, and W. L. Bargar, "An image-directed robotic system for precise orthopaedic surgery," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 3, pp. 261–275, 1994.

[27] D. H. Bourla, J. P. Hubschman, M. Culjat, A. Tsirbas, A. Gupta, and S. D. Schwartz, "Feasibility study of intraocular robotic surgery with the da vinci surgical system," *Retina*, vol. 28, no. 1, pp. 154–158, 2008.

[28] X. He, J. Handa, P. Gehlbach, R. Taylor, and I. Iordachita, "A submillimetric 3-dof force sensing instrument with integrated fiber bragg grating for retinal microsurgery," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 2, pp. 522–534, 2014.

[29] H. Meenink, "Vitreo-retinal eye surgery robot: sustainable precision," Ph.D. dissertation, Technische Universiteit Eindhoven, 2011.

[30] R. A. MacLachlan, B. C. Becker, J. C. Tabarés, G. W. Podnar, L. A. Lobes Jr, and C. N. Riviere, "Micron: an actively stabilized handheld tool for microsurgery," *IEEE Transactions on Robotics*, vol. 28, no. 1, p. 195, 2012.

[31] C. Song, P. L. Gehlbach, and J. U. Kang, "Active tremor cancellation by a "smart" handheld vitreoretinal microsurgical tool using swept source optical coherence tomography," *Optics Express*, vol. 20, no. 21, pp. 23 414–23 421, 2012.

[32] H. Yu, J.-H. Shen, R. J. Shah, N. Simaan, and K. M. Joos, "Evaluation of microsurgical tasks with OCT-guided and/or robot-assisted ophthalmic forceps," *Biomedical Optics Express*, vol. 6, no. 2, pp. 457–472, 2015.

[33] T. Edwards, K. Xue, H. Meenink, M. Beelen, G. Naus, M. Simunovic, M. Latasiewicz, A. Farmery, M. de Smet, and R. MacLaren, "First-in-human study of the safety and viability of intraocular robotic surgery," *Nature Biomedical Engineering*, p. 1, 2018.

[34] R. J. Webster III, J. S. Kim, N. J. Cowan, G. S. Chirikjian, and A. M. Okamura, "Nonholonomic modeling of needle steering," *The International Journal of Robotics Research*, vol. 25, no. 5-6, pp. 509–525, 2006.

[35] T. K. Adebar, A. E. Fletcher, and A. M. Okamura, "3-d ultrasound-guided robotic needle steering in biological tissue," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 12, p. 2899, 2014.

[36] W. Sun, S. Patil, and R. Alterovitz, "High-frequency replanning under uncertainty using parallel sampling-based motion planning," *IEEE Transactions on Robotics*, vol. 31, no. 1, pp. 104–116, 2015.

[37] D. Glozman and M. Shoham, "Image-guided robotic flexible needle steering," *IEEE Transactions on Robotics*, vol. 23, no. 3, pp. 459–467, 2007.

[38] D. Hu, Y. Gong, B. Hannaford, and E. J. Seibel, "Semi-autonomous simulated brain tumor ablation with ravenii surgical robot using behavior tree," in *IEEE International Conference on Robotics and Automation*, 2015, pp. 3868–3875.

[39] J. D. Opfermann, S. Leonard, R. S. Decker, N. A. Uebele, C. E. Bayne, A. S. Joshi, and A. Krieger, "Semi-autonomous electrosurgery for tumor resection using a multi-degree of freedom electrosurgical tool and visual servoing," in *International Conference on Intelligent Robots and Systems*, 2017, pp. 3653–3660.

[40] F. Alambeigi, Z. Wang, Y.-h. Liu, R. H. Taylor, and M. Armand, "Toward semi-autonomous cryoablation of kidney tumors via model-independent deformable tissue manipulation technique," *Annals of Biomedical Engineering*, pp. 1–13, 2018.

[41] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science Translational Medicine*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[42] J. Van Den Berg, S. Miller, D. Duckworth, H. Hu, A. Wan, X.-Y. Fu, K. Goldberg, and P. Abbeel, "Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 2074–2081.

[43] C. E. Reiley, E. Plaku, and G. D. Hager, "Motion generation of robotic surgical tasks: Learning from expert demonstrations," in *IEEE Engineering in Medicine and Biology*, 2010, pp. 967–970.

[44] A. Murali, S. Sen, B. Kehoe, A. Garg, S. McFarland, S. Patil, W. D. Boyd, S. Lim, P. Abbeel, and K. Goldberg, "Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms," in *IEEE International Conference on Robotics and Automation*, 2015, pp. 1202–1209.

[45] C. G. Atkeson and S. Schaal, "Robot learning from demonstration," in *International Conference on Machine Learning*, vol. 97, 1997, pp. 12–20.

[46] M. Vecerík, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. A. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," *arXiv preprint arXiv:1707.08817*, 2017.

[47] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Z. Leibo, and A. Gruslys, "Deep q-learning from demonstrations," in *AAAI Conference on Artificial Intelligence*, 2018.

[48] W. Sun, J. A. Bagnell, and B. Boots, "Truncated horizon policy search: Combining reinforcement learning & imitation learning," in *International Conference on Learning Representations*, 2018.

[49] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[50] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[51] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[52] A. Yahya, A. Li, M. Kalakrishnan, Y. Chebotar, and S. Levine, "Collective robot reinforcement learning with distributed asynchronous guided policy search," in *International Conference on Intelligent Robots and Systems*, 2017, pp. 79–86.

[53] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[54] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

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

[56] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[57] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*, 2014, pp. 387–395.

[58] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[59] R. Bellman, "A markovian decision process," *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

[60] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning." in *AAAI Conference on Artificial Intelligence*, vol. 2, 2016, p. 5.

[61] R. Bellman, "On the theory of dynamic programming," *Proceedings of the National Academy of Sciences*, vol. 38, no. 8, pp. 716–719, 1952.

[62] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[63] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 6292–6299.

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

[65] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, https://www.tensorflow.org/.

[66] S. G. Johnson, "The NLopt nonlinear-optimization package," 2018, http://ab-initio.mit.edu/nlopt/.

[67] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

[68] H. Levene, "Contributions to probability and statistics," *Essays in honor of Harold Hotelling*, pp. 278–292, 1960.

[69] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.