Learning to Navigate Endoscopic Capsule Robots

Mehmet Turan, Yasin Almalioglu, Hunter B. Gilbert, Faisal Mahmood, Nicholas J. Durr, Helder Araujo, Alp Eren Sarı, Anurag Ajay, and Metin Sitti

***Abstract*—Deep reinforcement learning (DRL) techniques have been successful in several domains, such as physical simulations, computer games, and simulated robotic tasks, yet the transfer of these successful learning concepts from simulations into the real world scenarios remains still a challenge. In this letter, a DRL approach is proposed to learn the continuous control of a magnetically actuated soft capsule endoscope (MASCE). Proposed controller approach can alleviate the need for tedious modeling of complex and highly nonlinear physical phenomena, such as magnetic interactions, robot body dynamics and tissue-robot interactions. Experiments performed in real *ex-vivo* porcine stomachs prove the successful control of the MASCE with trajectory tracking errors on the order of millimeter.**

***Index Terms*—Deep reinforcement learning, model-free control learning, endoscopic capsule robot, actor-critic.**

I. Introduction

WIRELESS capsule endoscopes (WCEs) and other smart swallowable capsules are revolutionizing gastroenterology screening, and are frequently used in lieu of esophagogastroduodenoscopy and colonoscopy for routine diagnostic imaging of the gastrointestinal (GI) tract. WCEs offer patients a non-invasive, substantially less painful and stressful experience than other GI screening procedures involving the use of flexible endoscopes and catheters [1], [2]. Beyond imaging, some WCEs have the ability to provide information from a wide array of sensors, such as pH, temperature, pressure, and chemical concentrations [3]. Several designs also facilitate interaction with the surrounding tissues, and include mechanisms for biopsy, drug delivery, and more complex therapeutic interventions [4]. To take full advantage of the new capabilities offered by WCEs, the ability to precisely navigate them through the desired region of the GI tract is crucial. An accurate and precise position and orientation control of the WCE will enable the physician to bring the robot to the relevant region where biopsies of millimeter accuracy will be performed, drugs to specific sites will be delivered or physiological data with high spatiotemporal resolution will be measured. In addition to these, diagnostic imaging may also benefit from improved control significantly. As an example, second-generation WCEs fail to detect polyps in approximately 14% of patients [5]. Gaining complete control over the capsule robot pose may ameliorate the rate of false negatives, and more importantly, active capsule robots can facilitate multiple angles of view, improving the amount of information available to the physician when pathological sites of interest are discovered.

Remote magnetic actuation which is a type of external wireless capsule actuation has a unique advantage that it does not require on-board actuators. This greatly reduces the electromechanical complexity of the device and removes the need for extra energy and space. Figure 1 demonstrates possible application scenario of an actively steerable capsule endoscope at hospitals, where the doctor navigates the robot to the desired organ regions

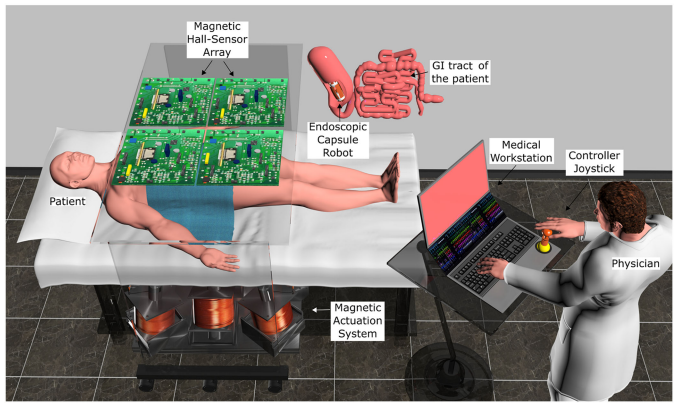

Fig. 1. Demonstration of the magnetically actuated WCE platform designed to visualize the GI tract in a minimally invasive manner. MASCE is made up of an RGB camera, battery and a permanent magnet. The electromagnetic coils based actuation unit below the patient table exerts forces and torques to realize the desired motion. Physician operates the screening process in real-time using the live video stream onto the medical workstation and the joystick to control the endoscopic capsule to the desired position/orientation. Magnetic hall sensors are placed on the top to localize the robot.

Manuscript received January 20, 2019; accepted June 7, 2019. Date of publication June 24, 2019; date of current version July 3, 2019. This letter was recommended for publication by Associate Editor M. Rentschler and Editor P. Valdastri upon evaluation of the reviewers' comments. This work was supported by the Max Planck ETH Center for Learning Systems. M. Turan and Y. Almalioglu equally contributed to this letter. *(Corresponding author: Mehmet Turan.)*

M. Turan and M. Sitti are with the Physical Intelligence Department, Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany (e-mail: turan@is.mpg.de; sitti@is.mpg.de).

Y. Almalioglu is with the Computer Science Department, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: yasin.almalioglu@cs.ox.ac.uk).

H. B. Gilbert is with the Department of Mechanical and Industrial Engineering, Louisiana State University, Baton Rouge, LA 70803 USA (e-mail: hbgilbert@lsu.edu).

F. Mahmood and N. J. Durr are with the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: faisalm@jhu.edu; ndurr@jhu.edu).

H. Araujo is with the Institute for Systems and Robotics, University of Coimbra, 3004-531 Coimbra, Portugal (e-mail: helder@isr.uc.pt).

A. E. Sarı is with the Department of Electrical and Electronics Engineering, Middle East Technical University, 06800 Ankara, Turkey (e-mail: asari@metu.edu.tr).

A. Ajay is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: aajay@mit.edu).

Digital Object Identifier 10.1109/LRA.2019.2924846

for monitoring, drug delivery and/or biopsy-like operations. Permanent magnets or electromagnets may be located outside of the patients body to impose the desired force and torque on the capsule that is located inside the body [6], [7]. Depending on the environment the capsule locomotion may take different forms, including sliding on surfaces, translating in fluids, or rolling [6], [8], [9]. Control techniques that rely on the physician to close the loop are possible in some instances due to the stable attraction between the external magnet and the capsule [6], but a fully automatic controller is necessary to achieve high-precision navigation. Closing the loop requires accurate measurement of the capsules pose, and several non-line-of-sight techniques exist for this purpose [10]–[12]. However, closed-loop remote magnetic manipulation of WCEs still remains challenging because it typically requires accurate models of the magnetic fields produced by the actuators so that the dependence of the robots position and orientation on the field can be appropriately predicted [9]. Furthermore, these techniques often require calibration of the field models [13] and models of the robot-tissue interaction.

In terms of magnetic actuation for capsule robots, many different approaches have been proposed in recent years. Keller *et al.* [15] use magnetic coils to actuate WCEs inside a water-filled stomach for screening purposes, which provides 10 basic motion types. Unlike our work, Keller *et al.* do not provide a closed loop control and indicate large drifts in the upper parts of the stomach. Moreover, this control technique cannot be used for biopsy and drug delivery because it does not appear to offer enough precision based on the presented results. For propulsion, a single rotating magnetic dipole that can generate screw-like motion for WCEs has been proposed by Popek *et al.* [16]. Although, such magnetic manipulation is shown to be practical in lumen-like surfaces the technique has not been demonstrated in a relatively inflatable biological space such as the insufflated stomach. Moreover, Popek *et al.* use phantoms, whereas we use real porcine stomachs which makes explicit comparisons tedious. Finally and most importantly, our method is aimed not only at propulsion, which is the goal of Popek *et al.* paper, but also at fine-scale position and orientation control. Their motion type is limited to screw propulsion rather than a 5-DoF motion. Locomotion on solid surfaces can be achieved through open-loop orientation control if the magnetic force is closed-loop controlled. Closed-loop control techniques which may be based on standard controllers such as proportional-integral and proportional-integral-derivative [9], [17] are excessively dependent on the accuracy of the model used to describe the interaction between the robot and the external magnetic field. Complex and non-linear nature of the imposed external magnetic fields makes it difficult to close the loop with such traditional model-dependent controllers. In addition to uncertainty in the magnetic field, there are uncertainties in the operation environment which further complicate the control task: mechanical properties and geometry vary from one location to another, and unknown disturbances such as peristalsis, robot-tissue interactions, abnormalities on the tissue (such as existing tumors) or nearby magnetic materials may have significant influences on the overall system dynamics.

The idea that human beings learn by interacting with the environment is probably the first to occur to us when we think about the nature of learning. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals [18]. Like humans excelling at solving a wide variety of challenging problems from low-level motor control through to high-level cognitive tasks by interacting with the environment, recently proposed artificial agents can construct and learn their own knowledge directly from raw inputs by receiving rewards or punishments in a reinforcement learning paradigm [19]–[21]. Motivated by this recent success of deep reinforcement learning techniques in robotic control, in particular actor-critic deep Q-learning methods [19], we propose a solution for controlling complex medical devices, in particular WCEs. The novelties of the proposed controller are listed below:

- The proposed learning-based, data-driven control system does not require complex physics-based modelling of the system. This eliminates the need for expert knowledge and tedious hand-engineering work for complex modelling which is particularly useful for a controller designed to navigate in the challenging environments such as GI tract with peristaltic motions.
- The proposed control system is capable of adapting to different patients, and organs within the GI tract. We provide empirical evidence of such domain adaptation and transfer learning via extensive tests inside multiple ex-vivo porcine stomach instances.

The rest of the letter is organized as follows: Section II introduces the proposed learning based control method and gives its mathematical and theoretical background. Section III demonstrates the experimental setup, describes the details of the training and testing protocol and gives quantitative and comparative results for the proposed controller. Section IV and V discuss the shortcomings and limitations of the approach. Finally, conclusion summarizes the letter and gives some future directions.

II. Method

The main idea of reinforcement learning is that an artificial agent may learn to optimize its behavior for a state transition by interacting with the environment and justifying the quality of the taken action by interpreting received reward scores. This approach applies in principle to any type of sequential decision-making problem relying on past experience. The environment may be stochastic, the agent may only observe partial information about the current state, the observations may be high-dimensional (e.g., frames and time series), the agent may freely gather experience in the environment or, on the contrary, the data may be constrained (e.g., no access to an accurate simulator or limited data due to high costs or data privacy etc.). Deep reinforcement learning (DRL) which combines the reasoning capability of reinforcement learning (RL) with the representation power of deep learning, has led to very successful agents in recent years that are able to address more challenging non-linear sequential decision-making problems. DRL is most useful

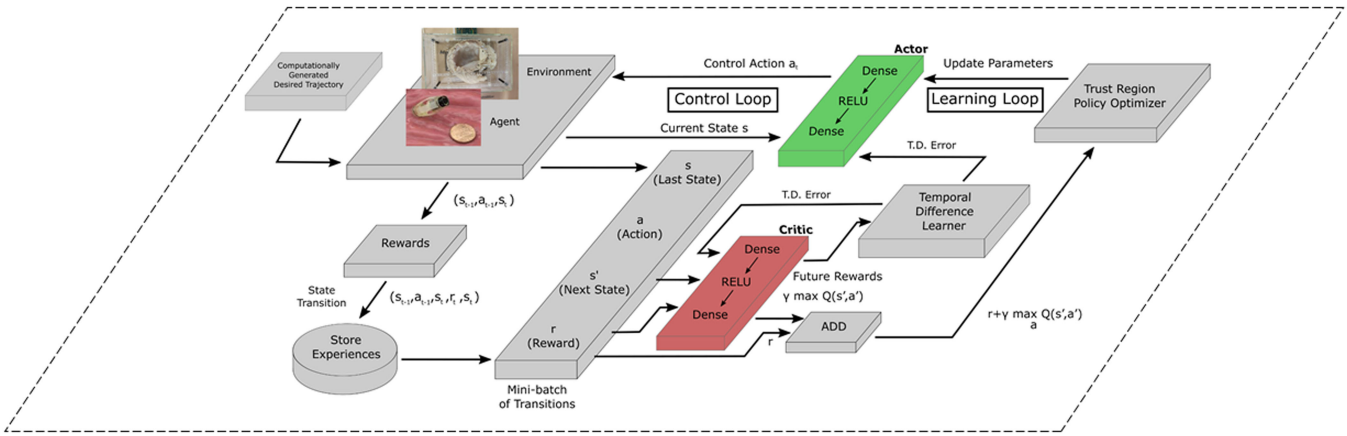

Fig. 2. **Architecture of the proposed DRL approach.** The MASCE interacts with the real environment and is driven by magnetic fields produced by electromagnets. Optitrack measures the 6 DoF pose of the capsule robot. The reward function is calculated based on the observed motion and velocity. Each action and state transition pair results in a new experience which is stored in the experience replay memory. The critic adds current reward to the predicted future reward score and sends the sum to the trust region policy optimizer [14]. After each action, the critic evaluates the new state to determine whether the reward is greater or smaller than expected. That evaluation is performed by the temporal difference (TD) error. If the TD error is positive, the critic suggests that the tendency to select this action in future should be strengthened, whereas if the TD error is negative, the critic suggests that the tendency of choosing this action for that state transition should be weakened. The parameters of both the actor and critic neural networks are updated accordingly.

in high dimensional state-space problems and complicated tasks with lower prior knowledge thanks to its ability to learn different levels of abstractions from data.

A. Deep Reinforcement Learning Architecture

Our control system is designed to make the capsule robot track a desired pose, which could be specified by either a teleoperator or an automatic path planner. Figure 2 illustrates the overall system architecture diagram for the proposed deep reinforcement learning approach. The controller receives a measurement of the MASCE state s, which comprises a 6 DoF pose and a 6 DoF rigid body velocity. The controller then computes the action a, which describes the electromagnetic coil currents to realize the dictated motion. After the action is executed and one time step from time t to time $t+1$ has elapsed, a new measurement of the state is acquired and the agent calculates a reward score $r(s_t, a_t)$ associated with taking action a_t in state s_t. The goal of the agent is to find the optimal stochastic policy $\pi(s \mid a)$ such that the value function $V^{\pi}(s)$, which describes the cumulative reward of following the current policy π from the starting state s, is maximized. The definition of V^{π} is

$$V^{\pi}(s_t) = \sum_{i \geq 0}^{\infty} \gamma^i r(s_{t+1}, a_{t+i}) \ , \tag{1}$$

where γ is the discount rate that prioritizes immediate rewards over the predicted future rewards.

In the derivation of the optimal policy, a Q function, $Q(s, a)$, is associated with each state-action pair that approximates the expected discount cumulative reward (i.e. value function) of taking action a at state s following the current policy π. The optimal policy is the action choice that maximizes $Q(s, a)$ given the current state: $\pi(s) = \arg\max_{a'} Q(s, a')$.

The actor network $\mu(s_k)$ that learns to approximate the optimal policy comprises a dense layer with an output tensor (size 32) followed by a RELU layer and a dense layer with an output vector of length 9 that represents the current values. Following the dense layer output, a final *soft-max* layer creates the probability distribution of the output current values. The critic network, on the other hand, is made up of a dense layer with an output tensor (size 32) followed by a RELU layer and a dense layer with a scalar Q-value output, that is $\gamma \max_{a'} Q(s', a')$ for the next state s'. The experience memory D, with capacity N_D, is used to store the state transitions and Q values to obtain independently and identically distributed samples in mini-batches. Upon execution of a selected action a_t, the observed total reward $r(s_t, a_t)$ is used to update weights of the critic network via the temporal difference error, current reward, and the last action taken by the agent. The policy network is updated relative to the temporal difference error, state s_k and $r + \gamma \max_{a'} Q(s', a')$ using trust region optimization [14] where expectations are replaced by sample averages and the Q value is replaced by an empirical estimate following the estimation procedure presented in [21].

B. Reward Function

The choice of reward function is important in terms of achieving a convergence of the controller's weights in a reasonable time. In our control task, one wants the capsule to track a desired 6-DoF pose without exceeding the pre-defined safe velocity threshold. The dominant term of the reward is proportional to the Euclidean distance between the desired and achieved pose. A velocity penalty penalizes velocities greater than the upper limit of $\mathbf{v}_{safe}$. An agent that takes an action in effect, is assigned a reward score r_t, which is defined by:

$$\begin{aligned} r_t = -(&||\mathbf{p}_t - \mathbf{p'}_t|| + \alpha||\mathbf{w}_t - \mathbf{w'}_t|| \\ &+ \beta||\mathbf{p}_t - \mathbf{p}_{t-1}|| + \theta 1[\mathbf{v}_t \geq \mathbf{v}_{safe}]) \end{aligned} \tag{2}$$

where $\mathbf{p}_t$ and $\mathbf{w}_t$ are the achieved translational and rotational parameters, and $\mathbf{p'}_t$ and $\mathbf{w'}_t$ are the desired translational and

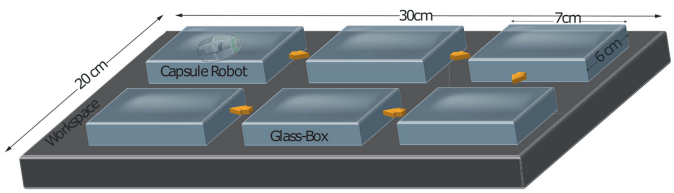

Fig. 3. **Demonstration of the sliding container technique.** The MASCE is operated inside a glass container which is slid into different locations in the work space to collect equal amount of data from each region.

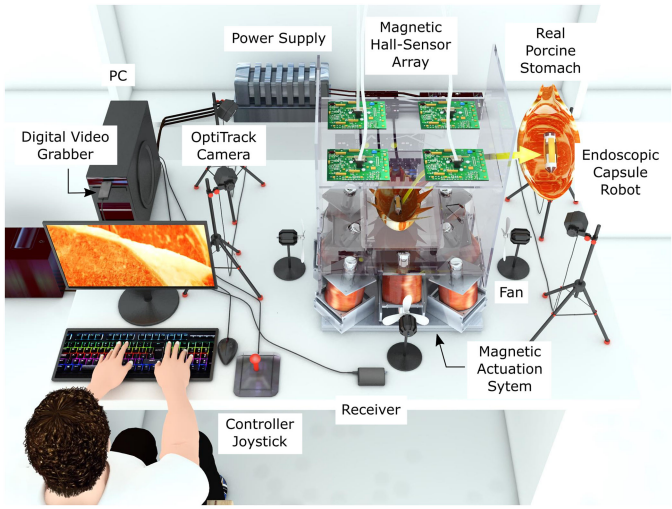

Fig. 4. **The experimental setup.** The setup consists of a wireless MASCE steered using a joystick, an actuation system of 9 electromagnetic coils, cooling fans, Prime13 cameras and a magnetic hall sensor array to localize the MASCE, a receiver to capture the transmitted image signal, a digital video grabber, and a real porcine stomach in a glass bowl.

rotational motion parameters at time index t, respectively. 1 is indicator function. v_{safe} is 2.3 cm/s and θ is 5.0, which is an empirically determined value using log-linear variations with logarithmic jumps.

III. Experiments and Results

Three assessments were performed to study training and testing performance of the proposed DRL-based controller. The assessments are (1) a 36 hour study of the learning performance in a guided training protocol using multiple real porcine stomachs; (2) a comparative assessment with the state-of-the-art DRL methods Actor Critic using Kronecker-Factored Trust Region (ACKTR) [22], Advantage Actor Critic (A2C) [23], and Deep Deterministic Policy Gradient (DDPG) [19] methods; (3) tests using multiple real porcine stomachs which were not seen by the agent during the training.

A. Experimental Setup

The experiments took place in one of the robotics labs at the Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Stuttgart. Details of the experimental setup are illustrated in Fig. 4. The MASCE includes a permanent ring magnet to enable a magnetic actuation. The electromagnetic coil array generates fields that cause a controllable force and torque to be applied to the MASCE. Sixty-four three-axis magnetic hall sensors are placed at the top of the workspace and four Prime13 Optitrack infrared cameras are placed on the corners of the workspace in order to provide feedback to the controller and evaluate the accuracy of the controller. Due to rapid saturation issue of hall-effect sensor array during long training sessions, Optitrack sensor system was employed to provide pose feedback to the agent. During testing however, pose feedback was received from the hall-effect sensor array, whereas the Optitrack sensor system was used to evaluate the translational and rotational accuracies of the controller. Cooling fans were placed around the coil array to prevent overheating.

The environment of the MASCE is an ex-vivo porcine stomach (see Fig. 2), resembling the physical properties that would be found in the human GI tract. Five porcine stomachs were used during training, and additional twelve stomachs were used for validation tests and were not included in the training data. By adopting this configuration, any overfitting of the properties of specific stomach is avoided and adaptability of the controller is validated. The implementation of the DRL-based method is derived from OpenAI Baselines libraries [24]. The proposed controller is capable of achieving a 20 Hz sampling rate on an operator PC with an Intel i7 processor. The final system implementation was operated at a 5 Hz sample rate. Commanded current values were saturated at 5 A to prevent overheating of the coils, and the command slew rate was artificially limited to 0.5 A/s to prevent the application of rapid alternating currents near the peak values.

B. Assessment of Learning

For the proposed method, training protocol had a termination criteria of achieving a mean translational error of 3.0 mm and orientational error of 0.05 rad for the most recent 100 executed motions by the DRL based controller. The time intervals required by the proposed controller to achieve these accuracies were used as the duration of the training sessions for ACKTR, DDPG and A2C to have a common benchmark. The training assessment took place in two phases. In the first phase of training, the DRL-based algorithm is trained in a simple rigid environment to reduce the complexity of the problem and is guided to improve workspace exploration (see Fig. 3). The goal of this strategy is to promote the following outcomes: (1) complete workspace coverage; and (2) a good initial rate of convergence as a result of a simplified environment. In the second phase of training, the algorithm is trained in multiple porcine stomachs. The desired motions are generated randomly during the training and testing session. Maximum velocity limits for x and/or y translation ($\pm$2.30 cm/s) and x and/or y rotation ($\pm$0.52rad/s) were imposed.

The first training phase took place in a small glass container. The container of size 7 cm $\times$ 6 cm was placed within the total workspace of dimensions 30 cm $\times$ 20 cm. The container was placed in six uniformly distributed positions within the workspace in an effort to ensure the agent trained in the entire

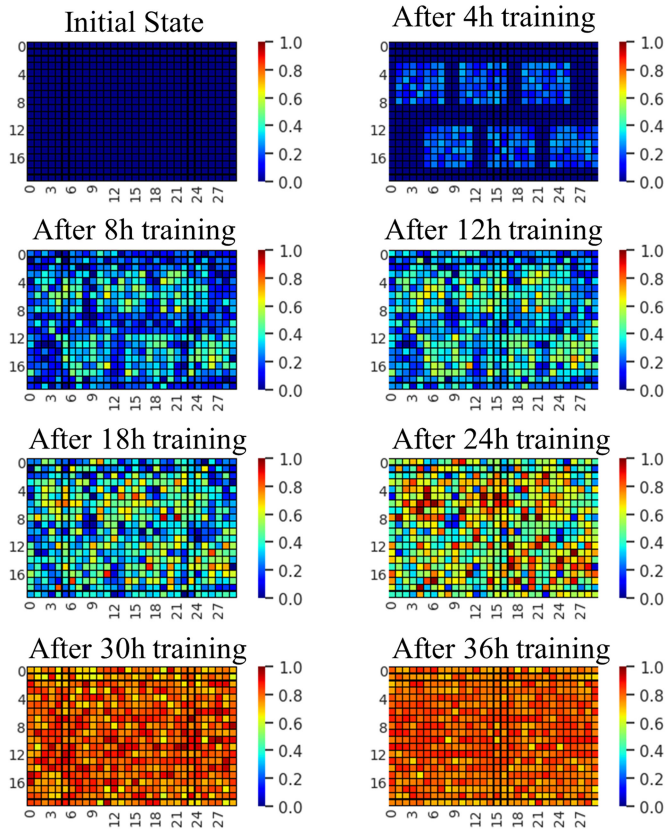

Fig. 5. Heatmaps of normalized reward scores for various time intervals of training. The axis dimensions represent the operation area of 30 cm × 20 cm, corresponding to the porcine stomach container (see Fig. 3).

workspace. The MASCE was allowed to operate for 20 minutes in each location of the container. Then, the container was moved to the next workspace position. In total, the agent was trained for a period of four hours in the six alternate workspace positions.

In the second phase of training, the capsule was placed in five different porcine stomachs. The stomachs were replaced at the following intervals of total training time: 8 h, 12 h, 18 h, 24 h, and 30 h. An additional six hours of training was performed in the fifth stomach to assess whether the reward diverges when operating for extended times in the same environment.

Figure 5 shows normalized reward scores after each interval of training. The workspace is divided into 600 subregions in order to facilitate visualization of the spatial dependency reward score. The color bar to the right depicts the reward score from blue (0) to red (1). Heatmap 5 b illustrates the normalized reward scores after the initial 4-hour glass-box training phase. Heatmaps 5 c 5 h represent the reward scores obtained after 8 h, 12 h, 18 h, 24 h, 30 h and 36 h training periods respectively. These periods represent the five different real-porcine-stomach alternate positions that the glassbox was placed after the initial 4h training session. Figure 5 illustrates a significant increase in the normalized reward scores after 24 h. After 36 h training the controller achieves high reward scores covering the whole operation area, which implies that the agent was successful in generating its control strategy over the entire workspace by the end of the training protocol.

The maximum achieved velocity during the training was 2.75 cm/s. The mean and standard deviation of the velocity were 0.72 cm/s and 0.28 cm/s respectively. The maximum achieved angular velocity during the training was 0.86 rad/s. The mean and standard deviation of the angular velocity were 0.42 rad/s and 0.29 rad/s respectively.

C. Comparative Assessment of Learning with ACKTR, A2C, and DDPG

Two different training strategies were employed and compared: (1) the guided training (GT) and (2) curriculum training (CT). To compare our method against other DRL-based control methods, the mean reward scores for our controller, the ACKTR method, the A2C method, and the DDPG method were assessed. Each controller was trained five times with random training trajectories to assess repeatability of the training process. First we applied GT, following the protocol details already explained in Section III-B. Figure 6 shows the results of GT for the proposed DRL method compared to those of the other tested DRL methods, implying shorter training times for the proposed approach.

To achieve a better sample efficiency, we applied CT for each DRL-based controller in which we divided the complex capsule motion learning task into three sessions with increasing complexity. In the first training session where the controller only learns rotational motions in xy-plane, after the initial four hour glass-box training the capsule interacts with five different stomachs sequentially each lasting for 1h 6min making a total training time of 9.5h for each controller. In the second session, it learns only translational motions in the xy plane, and in the last session it learns how to combine gained knowledge from the previous two stages to achieve more complex motions types (xy-rotation and translation), each lasting for 57min (see Fig. 6).

Figure 7 shows the reward scores during training session. As seen in this figure, the training time for each controller decreases significantly in case of CT compared to GT method. Moreover, it is apparent in both Fig. 6 and Fig. 7 that rotational motions in general are learned faster than the translational motions. This finding is consistent with the underlying physics pertaining to the dynamics of the system; a magnetic MASCE design with one cylindrical permanent magnet is deemed more appropriate for rolling and rotational motions in comparison to translations. In CT (see Fig. 7), the elapsed time of the trajectories is 28.75 hours with a maximum achieved velocity of 2.23 cm/s. The mean and standard deviation of the velocity are 0.81 cm/s and 0.2 6cm/s respectively. Maximum achieved angular velocity during the training is 0.66 rad/s. The mean and standard deviation of the angular velocity are 0.31 rad/s and 0.18 rad/s respectively.

D. Comparative Assessment of Control Performance

Figure 8 illustrates sample user-defined, ground truth trajectories and the corresponding trajectories realized by the DRL-based control. Figure 9 shows a plot of Euclidean distance vs. time and orientation error vs. time for trajectory 8a. One can see the successful realization of the desired trajectories with minimal deviations on the order of millimeter scale in both figures.

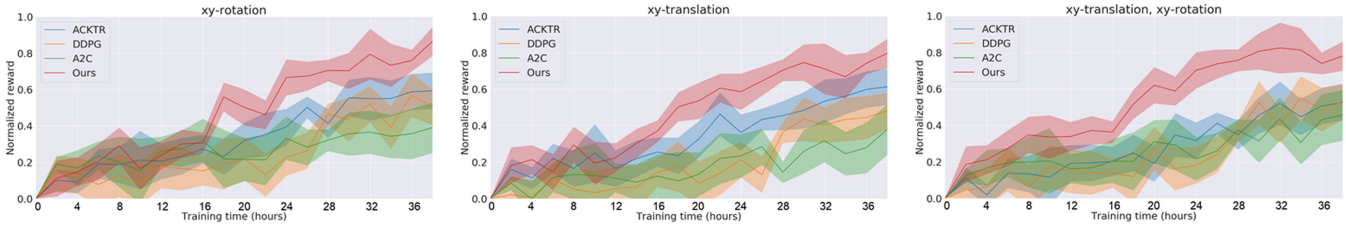

Fig. 6. Experimental results for the guided training consisting of three independent sessions: xy-rotation, xy-translation and joint xy-translation & rotation. The bold lines represent mean of achieved rewards whereas the bands represent the standard deviations.

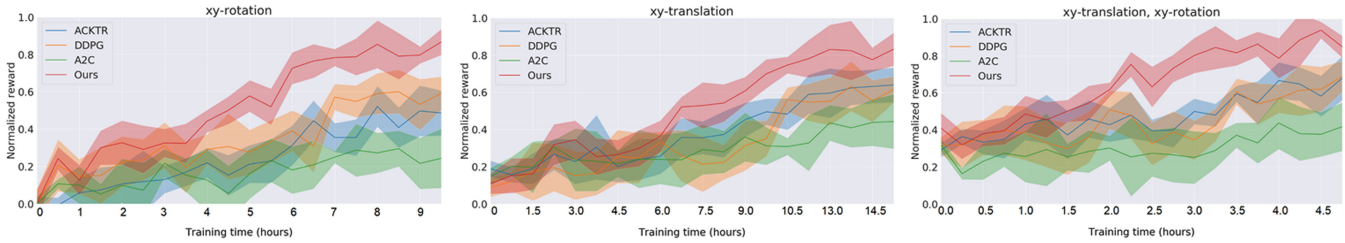

Fig. 7. Experimental results for the hierarchical training consisting of three sequential sessions: xy-rotation, xy-translation and joint xy-translation & rotation. The bold lines represent mean of achieved rewards whereas the bands represent the standard deviations.

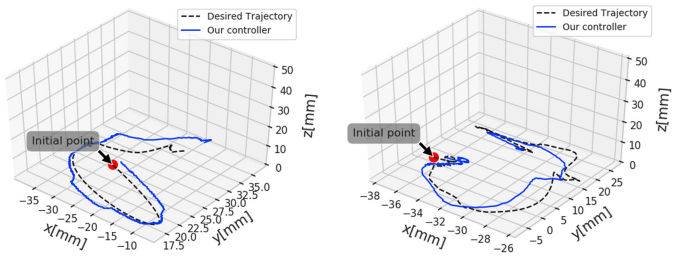

Fig. 8. **Trajectory tracking results.** Sample user-defined trajectories vs. achieved trajectories by the proposed controller.

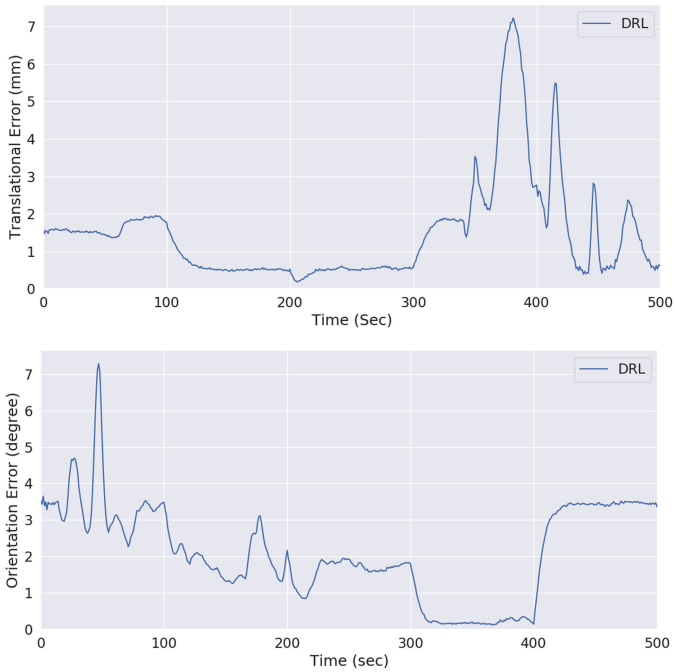

Fig. 9. Translational and rotational errors vs time for the trajectory of Fig. 8a.

Deviations from the desired trajectories are caused by under-representation of complex and loopy motions in the training data. Figure 10 demonstrates the performance of the proposed controller when exposed to new stomachs not seen during the training session. Overall, the DRL-based control exhibits very good performance in terms of both achieved reward scores and physical accuracies. A substantial learning period takes place at the beginning of each exposure to a new environment. In general, when first exposed to the new environment, the DRL-based controller exhibits lower rewards, however it quickly adapts to the new environment and starts to receive higher reward scores in a few minutes. Moreover, the adaptation time decreases substantially across the testing session (see Fig. 9) implying that the agent makes successful use of pre-gained knowledge to learn adaptation. As an example, the adaptation time from stomach#1 to stomach#2 lasts around 15 minutes, whereas the adaptation time from stomach#11 to stomach#12 is around 3 minutes.

IV. Discussion

3D trajectories shown in Fig. 8, the Euclidean distance and orientation error shown in Fig. 9, the adaptation capability shown in Fig. 10, and different types of motions realized by the MASCE system shown in Fig. 11 indicate that the proposed controller stays more-or-less close to the user-defined trajectories even when exposed to a completely new stomach instance. The precision in general remains on the order of millimeter scale which is enough for tasks such as controlled GI tract monitoring and targeted drug delivery. However, as seen between time interval $[350 - 450]$ seconds in Fig. 9a, the accuracy slightly decreases up to few millimeters. The main reasons for this decrease

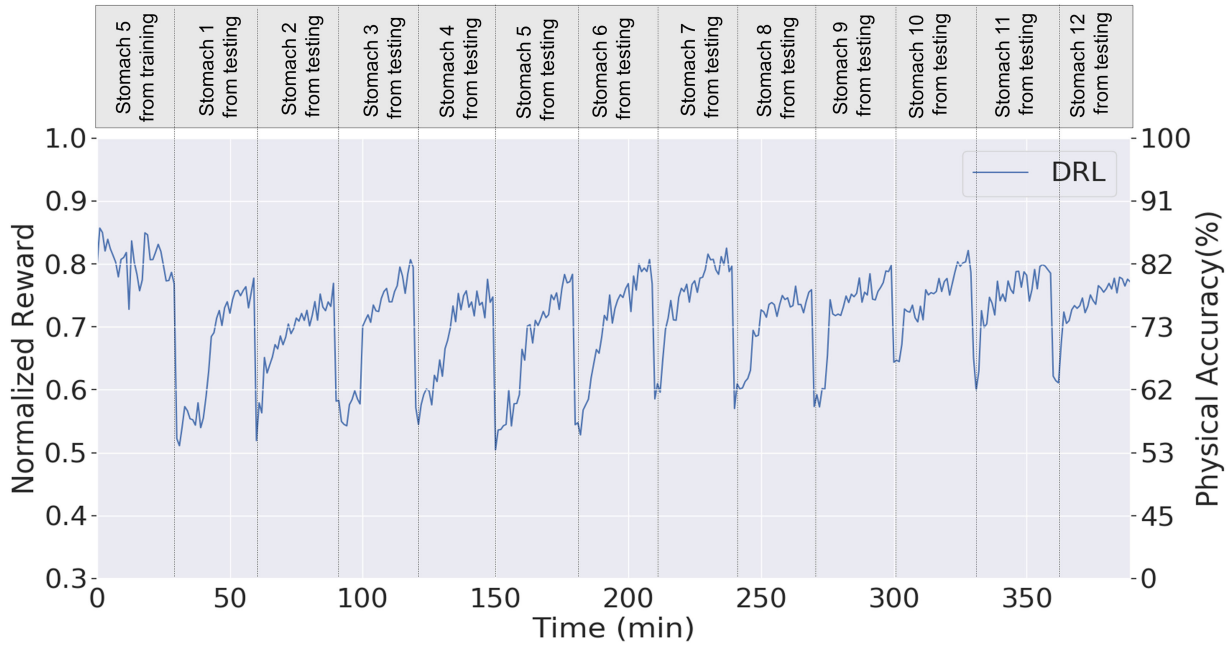

Fig. 10. **Reward scores and physical accuracies for the controller in new environments.** A substantial learning and adaptation effort is shown by the DRL-based control method. Performance degradation occurs when the controller agent is first exposed to the new environment, but it improves the reward score and accuracy over very short time intervals.

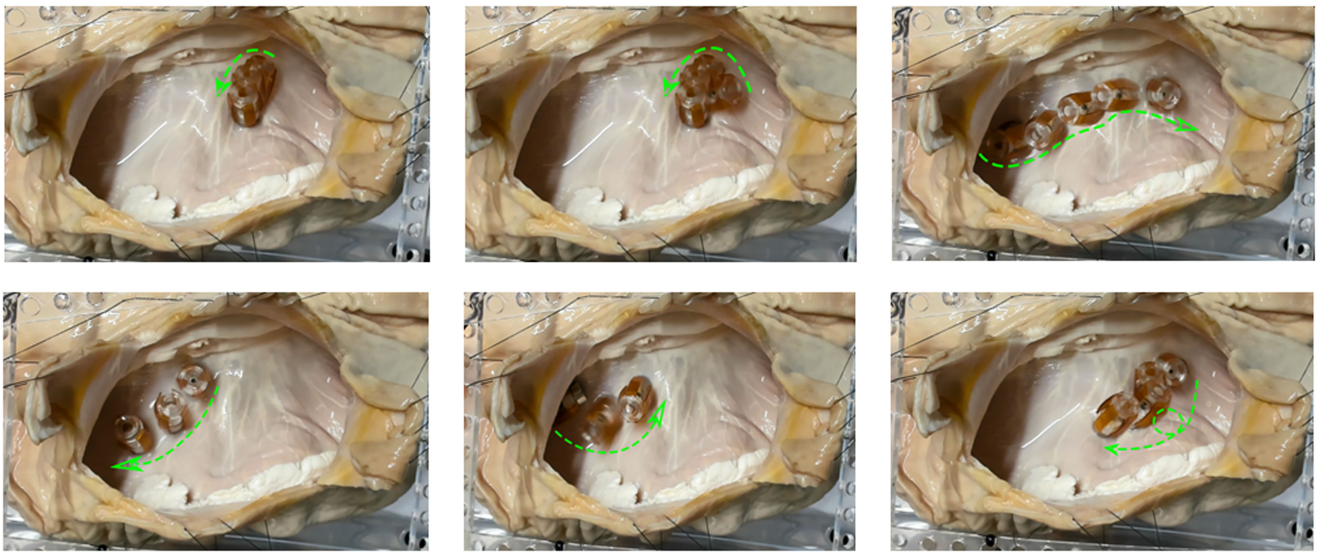

Fig. 11. Sample trajectories realized by MASCE using the proposed control algorithm. The desired trajectories are shown approximately as green dotted arrows.

is user-defined complex motions, which are not easy to realize due to capsule robot dynamics and obstacles caused by surface topography. More representation of such complex motions in the training data can solve this issue and increase even the overall controller accuracy to sub-millimeter scale making the approach relevant for biopsy-like operations [25]. On the other hand, one of the main drawbacks of the proposed method is the long training sessions taking up to 30 hours. This data inefficiency issue is a general problem of the state-of-the-art DRL methods in literature (see Fig. 7). A more sample efficient learning algorithm may be of great advantage for such a tedious real world tasks leading to significant amount of decrease in the overall training and adaptation time.

V. Shortcomings & Limitations

Coil Operating Frequency: A high coil operating frequency on the order of 10-20 Hz is desirable for achieving continuous and smooth capsule robot motions inside GI tract. Our current coil system operates at 5 Hz and is limited by the response time

of the coils, fast current alterations which lead to heating-up of the coils and current arise times. We argue that these limitations can be removed by replacing our current coil system with a more sophisticated system that can safely operate at a higher frequency.

Adaptation: Our proposed set-up is currently trained for the stomach and may not generalize to the entire GI tract. However, we hypothesize that a system developed for the stomach is likely to adapt to other regions of the GI tract with slight training given the similarity in the properties of the lumen [25].

Gastric Motility: In the current set of experiments, we do not test on a deformable stomach. Our intuition is that since a controller learned in a controlled static environment works well in a similar static environment, axiomatically, a controller learned in a deformable environment is likely to work well in such a situation. Our future work will focus on training the agent in a controlled but more realistic and deformable environment with gastric motility.

VI. Conclusion

In conclusion, we have presented a DRL-based method for learning the continuous control of a magnetically actuated capsule robot, and we have demonstrated the successful training and testing of this controller entirely in the real ex-vivo porcine stomach sets. The method completely alleviates the need for a complex system model, and is able to adapt to new environments that are similar to the training examples within about 2-3 minutes. DRL-based control holds great promise for control of complex medical systems. A future step towards real hospital conditions will be training and testing the controller in a more deformable environment. Finally, real human tests can follow to prove the effectiveness in real medical operation conditions. The project source code is available at the Github, https://github.com/yasinalm/DRL-CapsuleEndoscopeControl and a video demonstration of the proposed control is availed at https://youtu.be/xZx_uy5D3yw.

Acknowledgment

H. B. Gilbert would like to thank the Alexander von Humboldt foundation for support.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, 2000, Art. no. 417.

[2] J. G. Albert *et al.*, "Diagnosis of small bowel Crohn's disease: A prospective comparison of capsule endoscopy with magnetic resonance imaging and fluoroscopic enteroclysis," *Gut*, vol. 54, no. 12, pp. 1721–1727, 2005.

[3] M. Sitti *et al.*, "Biomedical applications of untethered mobile milli/microrobots," *Proc. IEEE*, vol. 103, no. 2, pp. 205–224, Feb. 2015.

[4] G. Ciuti *et al.*, "Frontiers of robotic endoscopic capsules: A review," *J. Micro-Bio Robot.*, vol. 11, nos. 1–4, pp. 1–18, 2016.

[5] C. Spada *et al.*, "Accuracy of first- and second-generation colon capsules in endoscopic detection of colorectal polyps: A systematic review and meta-analysis," *Clin. Gastroenterol. Hepatol.*, vol. 14, no. 11, pp. 1533–1543, 2016.

[6] G. Ciuti *et al.*, "Robotic versus manual control in magnetic steering of an endoscopic capsule," *Endoscopy*, vol. 42, no. 2, pp. 148–152, 2010.

[7] A. J. Petruska and J. J. Abbott, "An omnidirectional electromagnet for remote manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 822–827.

[8] A. W. Mahoney and J. J. Abbott, "Generating rotating magnetic fields with a single permanent magnet for propulsion of untethered magnetic devices in a lumen," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 411–420, Apr. 2014.

[9] A. W. Mahoney and J. J. Abbott, "Five-degree-of-freedom manipulation of an untethered magnetic device in fluid using a single permanent magnet with application in stomach capsule endoscopy," *Int. J. Robot. Res.*, vol. 35, nos. 1–3, pp. 129–147, 2016.

[10] C. Di Natali, M. Beccani, N. Simaan, and P. Valdastri, "Jacobian-based iterative method for magnetic localization in robotic capsule endoscopy," *IEEE Trans. Robot.*, vol. 32, no. 2, pp. 327–338, Apr. 2016.

[11] C. Hu, W. Yang, D. Chen, M. Q.-H. Meng, and H. Dai, "An improved magnetic localization and orientation algorithm for wireless capsule endoscope," in *Proc. IEEE 30th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2008, pp. 2055–2058.

[12] T. D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules." *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2387–2399, Sep. 2012.

[13] A. J. Petruska, J. Edelmann, and B. J. Nelson, "Model-based calibration for magnetic manipulation," *IEEE Trans. Magn.*, vol. 53, no. 7, pp. 1–6, Jul. 2017.

[14] M. P. Deisenroth *et al.*, "A survey on policy search for robotics," *Found. Trends Robot.*, vol. 2, nos. 1–2, pp. 1–142, 2013.

[15] H. Keller *et al.*, "Method for navigation and control of a magnetically guided capsule endoscope in the human stomach," in *Proc. 4th IEEE RAS EMBS Int. Conf. Biomed. Robot. Biomechatronics*, 2012, pp. 859–865.

[16] K. M. Popek, T. Hermans, and J. J. Abbott, "First demonstration of simultaneous localization and propulsion of a magnetic capsule in a lumen using a single rotating magnet," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1154–1160.

[17] K. M. Miller, A. W. Mahoney, T. Schmid, and J. J. Abbott, "Proprioceptive magnetic-field sensing for closed-loop control of magnetic capsule endoscopes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2012, pp. 1994–1999.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[19] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, arXiv:1509.02971.

[20] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[21] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.

[22] Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, and J. Ba, "Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2017, pp. 5279–5288.

[23] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[24] P. Dhariwal *et al.*, "Openai baselines," 2017. [Online]. Available: https://github.com/openai/baselines

[25] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.