IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

# Sim-To-Real Transfer for Visual Reinforcement Learning of Deformable Object Manipulation for Robot-Assisted Surgery

Paul Maria Scheikl[1], Eleonora Tagliabue[2], Balázs Gyenes[1], Martin Wagner[3], Diego Dall'Alba[2], Paolo Fiorini[2], and Franziska Mathis-Ullrich[1]

***Abstract*—Automation holds the potential to assist surgeons in robotic interventions, shifting their mental work load from visuomotor control to high level decision making. Reinforcement learning has shown promising results in learning complex visuomotor policies, especially in simulation environments where many samples can be collected at low cost. A core challenge is learning policies in simulation that can be deployed in the real world, thereby overcoming the sim-to-real gap.**

**In this work, we bridge the visual sim-to-real gap with an image-based reinforcement learning pipeline based on pixel-level domain adaptation and demonstrate its effectiveness on an image-based task in deformable object manipulation. We choose a tissue retraction task because of its importance in clinical reality of precise cancer surgery. After training in simulation on domain-translated images, our policy requires no retraining to perform tissue retraction with a $50\%$ success rate on the real robotic system using raw RGB images. Furthermore, our sim-to-real transfer method makes no assumptions on the task itself and requires no paired images. This work introduces the first successful application of visual sim-to-real transfer for robotic manipulation of deformable objects in the surgical field, which represents a notable step towards the clinical translation of cognitive surgical robotics.**

***Index Terms*—Surgical Robotics: Laparoscopy; Reinforcement Learning; Computer Vision for Medical Robotics**

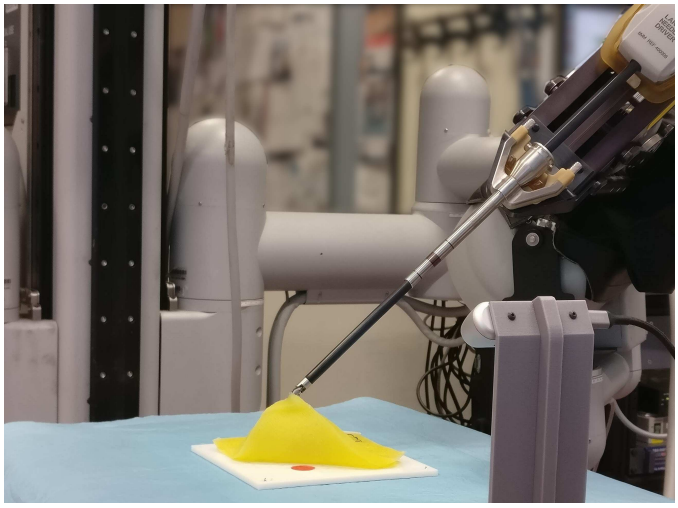


Fig. 1: Experimental setup for tissue retraction. We combine an Intel RealSense camera and the da Vinci Research Kit with a ProGrasp instrument grasping a yellow sheet of silicone attached to a board.

## I. INTRODUCTION

LEARNING behaviors in simulation and transferring them to real robotic systems (sim-to-real) is a prominent topic of research in robot-assisted surgery since learning on a real surgical robotic system is often infeasible [1]–[4]. Training in simulation enables end-to-end Reinforcement Learning (RL) of complex tasks in a safe and controlled environment, without requiring direct access to the real surgical robotic system. Training on real surgical robotic systems is impractical as RL algorithms often require millions of environment interactions to train a policy and unsafe behavior during training may damage the system [3].

Existing works on sim-to-real policy transfer for robot-assisted surgery utilize low-dimensional observations such as the state of the robot and known goal positions, so that the inputs to the policy are the same during training in simulation and execution on the real robotic system [1], [2], [5]. The success of state-based policies heavily relies on accurate extraction of task-relevant information from the scene. This makes it highly challenging to exploit state-based methods in surgical applications involving the manipulation of deformable tissues. The large configuration space of deformable objects cannot be fully extracted from data provided by standard surgical sensors (*e.g.* endoscopic camera) and is difficult to describe in a compact state. Image-based approaches, on the other hand, can learn task relevant features directly from sensor data. In this way, they are able to infer relevant information that is otherwise inaccessible to state-based approaches. However, image-based approaches have never been demonstrated in robot-assisted surgery due to the difficulty of transferring learned policies to real systems across the large visual domain

Manuscript received: August 10, 2022; Revised November 16, 2022; Accepted November 27, 2022.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Karlsruhe House of Young Scientists (KHYS), the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health", and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 742671 (ARS).

[1] P. M. Scheikl, B. Gyenes, and F. Mathis-Ullrich are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

Corresponding author: franziska.ullrich@kit.edu

[2] E. Tagliabue, D. Dall'Alba, and P. Fiorini are with the Department of Computer Science, University of Verona, 37134 Verona, Italy.

[3] M. Wagner is with the Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, 69120 Heidelberg, Germany.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

gap between simulated and real images [5], [6].

Several methodologies for transferring image-based policies across the visual domain gap between simulation and reality are currently investigated. Domain Randomization (DR) addresses the visual sim-to-real gap by randomly augmenting visual parameters of the simulation, *e.g.* texture and lighting, such that the trained policy learns to extract generalized, task-relevant visual features. DR approaches have been shown to translate well into reality [7]–[9] but they can be highly task specific and hard to tune [10], [11]. In contrast, pixel-level Domain Adaptation (DA) addresses the visual sim-to-real gap by directly transforming the images from one domain into the other. Recent related works employ Generative Adversarial Networks (GANs) to transform images for robotic grasping tasks [12]–[15]. These works share the limitation that large amounts of real world data are required to train the GANs. Furthermore, application-specific auxiliary tasks must be defined in order to stabilize GAN training to avoid mode collapse and hallucinated objects in the translated images [13]. This restricts their use to scenarios where these auxiliary tasks are applicable and may require additional data generation. Ho et al. [12] collect data from $135\,000$ task executions for GAN training. As an auxiliary task, they enforce predicting consistent bounding boxes by an object detection model for original and translated images. James et al. [14] invert the problem by training a GAN to translate images from domain-randomized simulations into a canonical simulation and show that the GAN is also able to translate images from reality into the canonical simulation. Most of these works employ CycleGANs [16] as their DA model. As an alternative to CycleGAN, CUT [17] and DCL [18] maximize mutual information between patches of the original and translated image through a contrastive learning approach. In robotics research, CUT is employed to generate synthetically labelled data for computer vision tasks such as semantic segmentation [19], [20] but has not been investigated for DA in RL.

In this work, we propose a pipeline for image-based RL of a surgical robotic task. Sim-to-real transfer is achieved through pixel-level DA using a contrastive GAN.

The contributions of this work are two-fold:
1) In the field of surgical robotics, we present the first successful sim-to-real transfer of an image-based RL policy to a real surgical robotic system. A visuomotor policy is trained in simulation and evaluated in reality without retraining. The approach is shown for Tissue Retraction (TR) (see Fig. 1), a common long-horizon task in deformable object manipulation. By relying on image inputs, the presented approach may be applied to other image-based tasks in robotic control, without the need of designing hand-crafted task-specific policies and is thus not limited to TR.
2) In the field of DA, we demonstrate that contrastive GANs are capable of performing unpaired image-to-image translation with less data than previous methods and do not require additional application-specific auxiliary tasks to stabilize training. The contrastive loss serves as an auxiliary task to retain useful visual features during image translation. Unlike previous work, the contrastive loss is independent of the RL task and can thus be considered a task-agnostic approach. This is the first successful application of pixel-level DA in RL for deformable object manipulation.

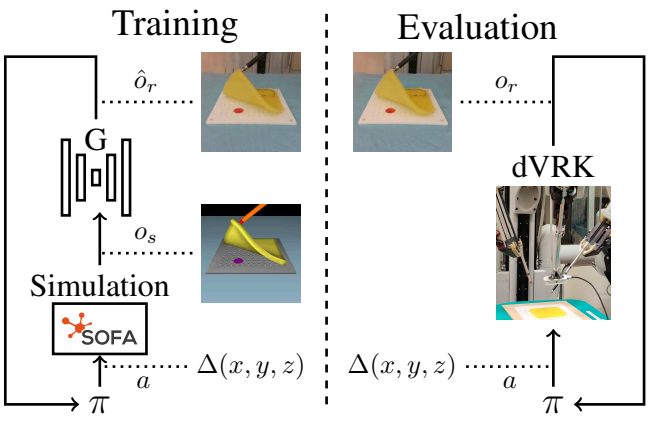



Fig. 2: Overview of training and evaluation settings. A policy $\pi$ is trained in simulation on translated observations $\hat{o}_r = G(o_s)$. During evaluation on the robotic system, the policy receives real image observations $o_r$. The actions $a$ of the policy are deltas in the gripper's Cartesian coordinates to solve a tissue retraction task.

## II. METHODS

The goal of this work is to train a visuomotor policy in a simulation of robot-assisted surgery on visual observations, such that the policy can be deployed on the real system without retraining. An overview of training and evaluation settings is given in Fig. 2. During policy learning, images from simulation $o_s$ are translated into the image domain of the real system $o_r$ before being passed to the policy $\pi$ that generates actions $a$ which are applied to the simulation. The required translation model $G$ is learned from unpaired examples of both the real and simulation domain and frozen during policy learning. Compared to related work, this presents a more general approach for sim-to-real transfer of visuomotor policies in deformable object manipulation as it does not require application-specific auxiliary tasks that stabilize training at the cost of restricting their transferability to other tasks.

### A. Domain Adaptation

A policy $\pi_s$ trained on simulated images is unsuitable for deployment in the real world because of the visual domain gap between observations in simulation and reality, even though the underlying dynamics and reward structure are similar. Sequential decision making problems are frequently formalized as Markov Decision Processes (MDP). MDPs can be generalized to Partially Observable Markov Decision Processes (POMDP) [21] that assume that the dynamics of the process are determined by an MDP, while the true state of the process is hidden and cannot directly be observed. As the true state of the deformable object is not accessible, and we can only observe the scene through image observations, we formally interpret the TR tasks in reality and simulation as two POMDPs that differ in their transition $T$ (physical

domain gap) and observation $O$ (visual domain gap) functions. Hidden states of the processes are assumed to be the same in simulation and reality.

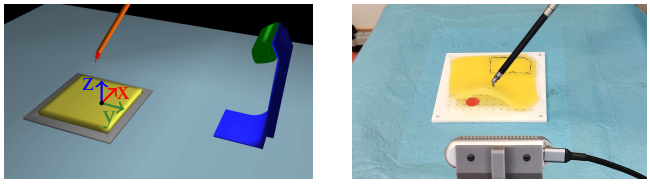
The goal of pixel-level DA is bridging the visual domain gap by changing the appearance of an image while task-relevant information is preserved. To this effect, this work investigates training an Unpaired Image-To-Image (UI2I) translation model $G : O_s \to O_r$ that translates images $o_s$ from the simulation's observation function $O_s$, such that the translated images $\hat{o}_r = G(o_s)$ are indistinguishable from samples $o_r \sim O_r(o \mid \hat{z})$ taken from the real observation function $O_r$, where $\hat{z}$ is the hidden state. UI2I methods utilize unpaired image data, making it feasible to train $G$ without access to hidden states of either environment.

$G$ is trained on unpaired image data collected in simulation and on the real system such that different lighting conditions are represented in the dataset, and then frozen for policy training. A policy $\pi_g$ is then trained in simulation on translated observations $\hat{o}_r$. Thus, the policy $\pi_g$ can be directly deployed on raw images $o_r$ from the real camera during execution on the real robotic system. In order to encourage retaining features of real images, $G$ also learns an identity mapping of real images by minimizing the distance between real images $o_r$ and their translation $\hat{o}'_r = G(o_r)$. In contrast to previous work, this approach requires knowledge about the Cartesian positions of the robot's gripper, grasping point, and end point only during training in simulation and not during execution on the real system.

Here, we investigate two methods for training contrastive UI2I models, CUT [17] and DCL [18], with CycleGAN [16] as a baseline. Contrastive learning learns to associate a *query* to a *positive* example and to contrast the *query* to *negative* examples. In the case of CUT and DCL, the *query* is a patch [17] of the translated image, the *positive* is the patch at the same location in the original image, and the *negatives* are patches from the original image at other locations. Image fidelity is compared manually, with only the best method used for policy learning as translation model $G$. During policy learning, the instance of $G$ used at each time step is sampled uniformly from an ensemble of 7 models from different training runs to compensate for possible bias of the individual GAN instances toward a specific lighting condition in the dataset. Intuitively, this is a form of visual domain randomization since the appearance of the translated images is slightly different for each specific instance of $G$ but the content of the images is the same.

### B. Reinforcement Learning

*1) Tissue Retraction Task:* TR is an elementary surgical task that consists of grasping and pulling deformable tissue in order to expose a target area to the endoscopic camera. Furthermore, it is important to provide an appropriate force to tension the tissue for dissection without ripping tissue apart. TR is thus of utmost importance, especially in cancer surgery, in order to dissect cancerous lymphatic tissue off of major blood vessels. Due to its ubiquity in surgical procedures, autonomous execution of TR has previously been investigated in the literature [2], [3], [22], [23]. In this work, we implement TR of a rectangular soft tissue [24]. The position of the target, as well as the attachment points to the surrounding environment, are assumed to be known from pre-operative data to limit complexity, even if in reality they will change during surgical dissection. The task is executed on the da Vinci Research Kit (dVRK) using a single Patient Side Manipulator (PSM), as illustrated in Fig. 3.




(a) (b)

Fig. 3: (a) Tissue retraction scene implemented in SOFA with illustrated coordinate system and (b) experimental setup on the real robotic system.

*2) Learning Environment:* The TR task is implemented in the Simulation Open Framework Architecture (SOFA) [25] framework (see Fig. 3 (a)), relying on a finite element method. The tissue is modeled as a mesh of tetrahedral elements with its top right part fixed to simulate the attachment to the board in the real environment. The robot is modeled as the distal part of a dVRK PSM including the gripper and instrument shaft respecting a fixed remote center of motion. Motion of the robot is constrained to a $83\,\mathrm{mm}$ high, $180\,\mathrm{mm}$ deep, and $140\,\mathrm{mm}$ wide workspace box above the tissue similar to the typical operative space of the PSM. The gripper starting position in each episode is uniformly sampled from the workspace with a minimum height of $40\,\mathrm{mm}$.

The TR task is divided into grasping and retracting phases. The positions of the grasping and end points, as well as the target, are fixed for training and evaluation. The grasping point was selected above the target area, while the end point was chosen to achieve good visibility of the target area in the image observations, similar to previous work [2].

In the grasping phase, the gripper can move freely in the workspace. If the distance between the gripper and the desired grasping point decreases below $3\,\mathrm{mm}$ and the gripper is below the tissue surface, the tissue is automatically grasped and the retraction phase is entered. In the retraction phase, the tissue is attached to the gripper. The episode is completed successfully when the distance between the gripper and the desired end point reduces to $3\,\mathrm{mm}$ or less. Each episode is limited to $1\,000$ steps before the environment is automatically reset. Collisions are detected between the gripper jaw tips and the tissue, based on whether the gripper is within a bounding box around the tissue in its non-deformed state.

*3) Reward, Observation and Action Space:* The simulation environment adheres to the OpenAI gym standard [26]. The policy outputs the parameters of a Gaussian distribution, from which actions consisting of task space velocities are sampled. The continuous action is clipped to the interval $[-1, 1]$, scaled such that the limits of the action space correspond to a maximum robot velocity of $3\,\mathrm{mm\,s^{-1}}$ in each direction, and an action repeat of 3 is applied. The environment generates a

$256 \times 256$ RGB image from a static camera perspective, which is then translated by the image-to-image translation model $G$. The four most recent images are concatenated and scaled down to a resolution of $84 \times 84$, resulting in observations of size $84 \times 84 \times 12$ that are subsequently passed to the policy. Observation space and network architecture are based on [27].

The reward function is split into grasping $r_g$ (Eq. 4) and retraction $r_r$ (Eq. 5) phases to match the two phases of the TR task. In the grasping phase, the agent receives a negative reward proportional to the distance between the current gripper position $\boldsymbol{x}_t$ and the grasping point $\boldsymbol{x}_g$, normalized by the absolute size of the workspace:

$$d_g = -\frac{||\boldsymbol{x}_g - \boldsymbol{x}_t||_2}{||\boldsymbol{x}_{max} - \boldsymbol{x}_{min}||_2} \tag{1}$$

where $\boldsymbol{x}_{max}$ and $\boldsymbol{x}_{min}$ are the corner points of the workspace bounding box. It further receives a constant negative reward equivalent to the reward seen at the beginning of the retraction phase based on grasping point $\boldsymbol{x}_g$ and end point $\boldsymbol{x}_e$:

$$c_g = -\alpha * \frac{||\boldsymbol{x}_g - \boldsymbol{x}_e||_2}{||\boldsymbol{x}_{max} - \boldsymbol{x}_{min}||_2}. \tag{2}$$

A weight of $\alpha = 1.2$ ensures that the agent is incentivized to transition to the retraction phase. In order to enforce safe policy behaviors on the real robot, an additional term $p_c$ penalizes collisions between gripper and tissue proportional to the distance to the grasping point, and a further term $p_w$ penalizes actions that would violate the workspace boundaries. Finally, the reward function assigns a one-time positive reward of $e_g = 1$ to a successful grasp. In the retraction phase, the agent receives a negative reward $d_e$ proportional to its distance to the end point, normalized by the absolute size of the workspace:

$$d_e = -\frac{||\boldsymbol{x}_e - \boldsymbol{x}_t||_2}{||\boldsymbol{x}_{max} - \boldsymbol{x}_{min}||_2}. \tag{3}$$

Additionally, a one-time positive reward of $e_r = 1$ is awarded when the episode is successfully completed. Thus, the overall reward function for grasping and retraction is

$$r_g = d_g + c_g + p_c + p_w + e_g, \tag{4}$$
$$r_r = d_e + e_r. \tag{5}$$

*4) Proximal Policy Optimization:* Stable Baselines 3 [28] and its implementation of Proximal Policy Optimization (PPO) [29] are utilized to train the agent. The agent is trained with a discount of $\gamma = 0.995$ and $\lambda_{GAE} = 0.95$ for a total of $10^7$ environment steps over $8$ parallel environments. The policy is updated after every $8 \times 128$ environment steps with a minibatch size of $256$ and $4$ epochs per PPO iteration. Learning rate and clip ratio follow a linearly decreasing schedule starting at $0.1$ and $2.5 \cdot 10^{-4}$, respectively. PPO's ratio clip is set to $0.2$, the value and entropy loss coefficients are $0.5$ and $0.001$, respectively, and gradients are clipped to a maximum norm of $0.5$.

*5) Agent Architecture:* The agent is split into policy and value estimation networks that do not share learnable parameters but are similar in architecture. Both networks consist of three convolutional layers with square kernel sizes $8, 4, 3$ and strides of $4, 2, 1$, followed by a fully connected layer with $512$ neurons. The policy network has a head with $3$ neurons for predicting the mean of a Gaussian distribution corresponding to the task-space velocities, and the value network has a head with $1$ neuron. The policy network also contains $3$ learnable log standard deviation parameters that do not depend on the input. ReLU non-linearities are applied after all layers except the final layer.

*6) Curriculum Learning:* Curriculum learning [30] is an approach to gradually increase the difficulty of a learning environment in order to simplify learning complex tasks. This work employs curriculum learning to tune the weight of the collision and workspace violation terms of the reward function during grasping $r_g$ (Eq. 4). The curriculum of the terms $p_c$ and $p_w$ of $r_g$ are defined as

$$p_c = -w_c \frac{||\boldsymbol{x}_g - \boldsymbol{x}_t||_2}{||\boldsymbol{x}_{max} - \boldsymbol{x}_{min}||_2}, \text{ if in collision} \tag{6}$$
$$p_w = -w_w, \text{ if action violates workspace} \tag{7}$$

with factors $w_c$ and $w_w$ linearly increasing from $0.0$ to $10$ and $0.2$, respectively, over $10^7$ environment steps.

## III. Experimental Evaluation

The experimental setup consists of tissue represented by a rectangular silicone sheet attached to a board at a fixed set of attachment points, following the approach in [24]. The target is represented by a red circular marker on the board. The PSM on the dVRK is equipped with a ProGrasp instrument, as illustrated in Fig. 3. An Intel RealSense D435 camera is used to capture monocular RGB images constituting the visual observations.

### A. Unpaired Image-To-Image Translation

The UI2I models CUT, DCL, and CycleGAN are each trained with 7 random seeds. Each training run is executed for 48 hours on four NVIDIA A100-40 GPUs with a minibatch size of $4$. The unpaired image dataset is created by performing grasping and retracting with different starting, grasping, and end points $82$ times under $3$ uncontrolled lighting conditions, for a total of $246$ TR executions. RGB image observations are captured at $30\,\text{Hz}$. Data is captured on the real system and in simulation with a resolution of $256 \times 256$. This dataset is split randomly into $90\%$ for training and $5\%$ each for validation and testing.

### B. Sim-To-Real Evaluation Scenarios

The experimental goal is to evaluate how well a policy that is trained on translated images in simulation performs on the real robotic setup. This is done without manually accounting for physical inaccuracies, such as modelling the robot dynamics, or environment conditions, such as changes in lighting, posing a more realistic and challenging task.

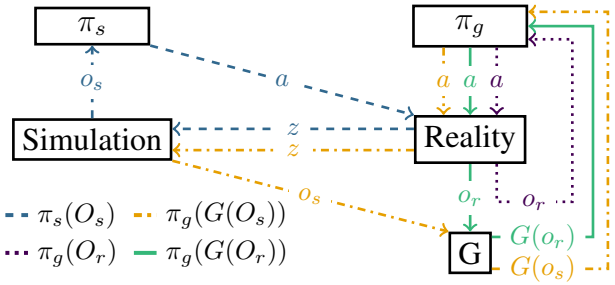



Fig. 4: Illustrated data flow for the four sim-to-real evaluation scenarios. All policy actions $a$ are executed on the real robot. For scenarios $\pi_s(O_s)$ and $\pi_g(G(O_s))$ the simulation is updated with the real robot states $s$ to generate observations $o_s$ from simulation. Scenarios $\pi_g(O_r)$ and $\pi_g(G(O_r))$ receive observations $o_r$ from the real system.

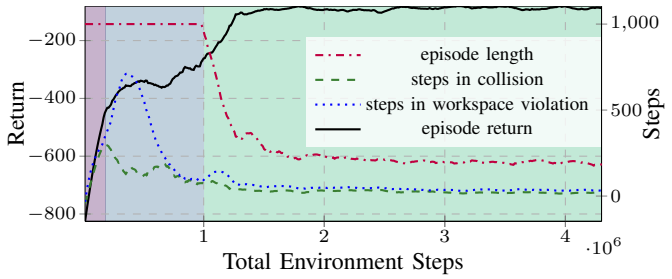



Fig. 5: Smoothed learning curves for a training run of $\pi_g$. Average episode return, episode length, and steps in collision and workspace violation are shown over total steps in the learning environment. Three phases are identifiable during learning: when the agent is predominantly learning to grasp (purple), when it is learning to retract (blue), and when it is mainly optimizing to reduce episode length and collisions (green).

The proposed approach is evaluated in four different scenarios as illustrated in Fig. 4. The **first scenario**, denoted $\pi_s(O_s)$, evaluates the physical domain gap between simulation and reality by conditioning on observations $o_s$ from the simulation under state transitions $T_r$ from the real system. Policy $\pi_s$ is trained in simulation without image translation and then executed on the real system in a digital twin approach, where the gripper position in simulation is set to match the gripper position on the real system. The subscript $s$ indicates that this policy is conditioned on observations $o_s$ from simulation. The policy receives observations from simulation but predicts actions that are applied to the real system instead of the simulation. The **second scenario**, denoted $\pi_g(G(O_s))$, follows the same approach, but additionally translates simulation images during training and execution. Policy $\pi_g$ is trained on translated images $\hat{o}_r = G(o_s)$ in simulation as indicated by the subscript $g$. The **third scenario**, denoted $\pi_g(O_r)$, is the target scenario of the sim-to-real transfer. The same policy $\pi_g$ receives real images $o_r$ during execution. The **fourth scenario**, denoted $\pi_g(G(O_r))$, evaluates whether the image translation model can mitigate the influence of changes in lighting conditions on the real system. Images $o_r$ from the real system are translated by $G$ before being passed to the policy, exploiting the learned identity mapping of real images as described in Section II-A.

Each scenario is executed from 32 different starting positions. The starting positions are determined by dividing the robot's planar workspace into a $4 \times 4$ grid to generate 16 starting positions on the XY-plane. Each starting position is executed on two starting heights ($z = 60\,\text{mm}$ and $z = 70\,\text{mm}$) to sample different heights within the allowed workspace.

### C. Evaluation Metrics

Policy performance is evaluated based on trajectory outcome and quality metrics. This work considers four different trajectory outcomes: (1) *success* if the policy completes both grasping and retracting phases, exposing the visual target as shown in Fig. 3; (2) *partial success* if the policy completes the grasping phase but does not expose the visual target; (3) *tissue stress*, where execution is aborted prematurely if the gripper applies excessive stress on the tissue through collisions before grasping; and (4) *time out* if the policy fails to reach the grasping point within the time limit of 1 000 steps. The termination criterion of outcome (3) is triggered if the gripper moves more than $8\,\text{mm}$ laterally in collision. The evaluated trajectory quality metrics are the number of steps in collision and the absolute path length.

## IV. RESULTS

### A. Unpaired Image-To-Image Translation

The unpaired image dataset contains 278 735 images from the real system collected over 2.58 hours and 108 994 images generated in simulation. Figure 6 (a) illustrates that both CUT and DCL successfully learn the image translation task, while CycleGAN does not, producing inconsistent visual features and spurious features. On visual assessment, however, DCL produces more consistent results for images from early steps in the retraction phase and is thus used as the image translation model $G$ for RL. Figure 6 (b) illustrates a case where DCL successfully translates the simulation image, but CUT fails to correctly change the appearance of the gripper for the closed grasp. The manual inspection of 600 randomly chosen images showed that 60% for CycleGAN, 30.5% for CUT, and 6.5% for DCL of the translated images showed inconsistent or spurious visual features.

### B. Training in Simulation

Learning TR on translated images in simulation requires roughly 2 million environment steps as illustrated in Fig. 5. The policy learns to grasp successfully after roughly 200 000 environment steps, much faster than learning the retraction part of the task. The impact of curriculum learning can be seen in the number of environment steps spent in collision and workspace violation. The task is learned quickly, yet with many safety violations. After learning the task and progressively increasing the importance of safe actions, the unsafe parts of the trajectories decrease while task success continues to improve. The learning can be roughly separated into three phases as indicated in Fig. 5: the agent learning to grasp the tissue, the agent learning to retract the tissue, and the agent optimizing its learned behaviour to reduce episode length, collisions, and workspace violations. Both policies $\pi_s$

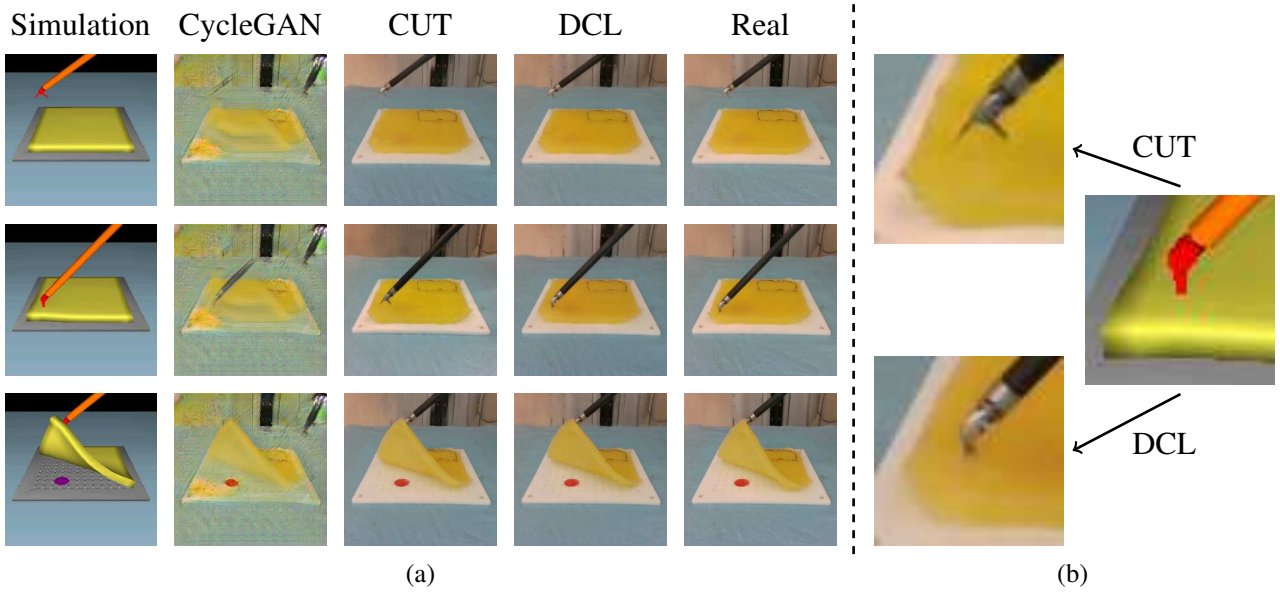



Fig. 6: Images from simulation and their translations by CycleGAN, CUT, and DCL, as well as real images (a) and a magnified view (b) on Simulation, CUT, and DCL images from the second row of (a).

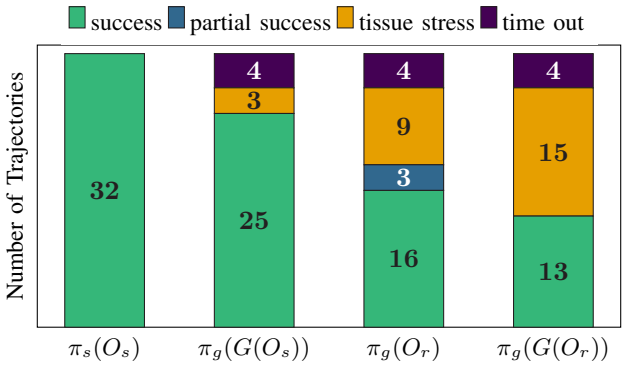



Fig. 7: Trajectory results for all four experiment cases split into their outcomes. Possible trajectory outcomes are task failure due to time out or excessive tissue stress, partial task success if grasping is successful but retraction failed, and success if the target was visible at trajectory end.

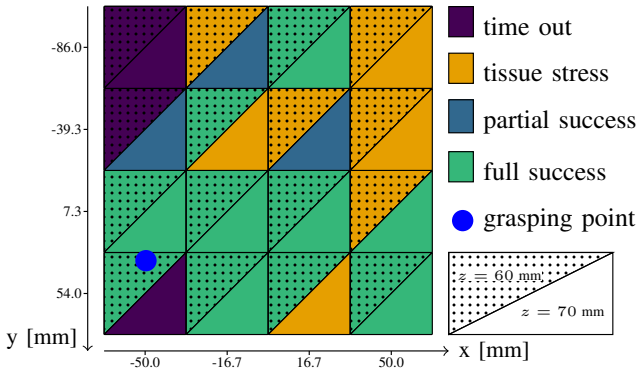



Fig. 8: Trajectory outcome of case $\pi_g(O_r)$ for 16 evaluated starting positions on the XY-plane. Each starting position was executed on two starting heights ($z = 60\,\text{mm}$ and $z = 70\,\text{mm}$).

TABLE I: Evaluation results for the scenarios defined in Section III-B with regard to the metrics defined in Section III-C.

| Scenario | Success Rate | Path Length | Collisions |
|---|---|---|---|
| $\pi_s(O_s)$ | 32/32 | 210 mm | 2.93 steps |
| $\pi_g(G(O_s))$ | 25/32 | 235 mm | 12.16 steps |
| $\pi_g(O_r)$ | 16/32 | 219 mm | 14.56 steps |
| $\pi_g(G(O_r))$ | 13/32 | 210 mm | 18.08 steps |

and $\pi_g$ achieve a task success rate of 100% in simulation by the end of training.

### C. Sim-To-Real Evaluation

The trajectory outcomes and quality metrics as described in Section III-C are presented in Fig. 7 and Tab. I. Scenario $\pi_s(O_s)$, where the baseline policy $\pi_s$ is executed in simulation with the digital twin in the loop, was successful from all starting positions. Obtained results for scenario $\pi_g(G(O_s))$, where policy $\pi_g$ is executed on translated images with the digital twin in the loop, include three trajectories that were aborted due to excessive collisions, and four timed out without solving the grasping task. When the same policy $\pi_g$ receives real images for scenario $\pi_g(O_r)$, the observed success rate was further reduced to 16/32. In contrast to the two previous experiments, three trajectories solved the grasping task, but were not successful in exposing the visual target. The final scenario $\pi_g(G(O_r))$, with policy $\pi_g$ executed on real images passed through the image translation model $G$, reaches a success rate of 13/32. The amount of steps in collision increased over all four scenarios leading to an increase in the number of trajectories terminated due to excessive tissue stress.

Figure 8 illustrates the relation between starting positions and trajectory outcome for scenario $\pi_g(O_r)$ in more detail. Starting positions near the rear left corner $(x, y) = (-50.0, -86.0)$ mm of the tissue tend to time out without grasping, while starting positions near the rear right corner

$(x, y) = (50.0, -86.0)$ mm tend to terminate due to excessive collisions.

## V. DISCUSSION

The results of scenario $\pi_g(O_r)$ show that the policy trained in simulation on the transformed observation function $G(O_s)$ is able to perform the TR task directly from raw camera images from the real observation function $O_r$. This indicates that the trained model $G$ is able to bridge the visual domain gap. The intuition for scenario $\pi_g(G(O_r))$ was that $G$ may mitigate the influence of changes in illumination by translating real images from different lighting conditions into a consistent appearance. Surprisingly, the policy performs worse in this scenario than the policy on real images directly. This also supports the claim that the image-to-image translation results in the loss of some image information necessary to compensate for the novel dynamics on the real robot.

On the real evaluation setup, starting positions near the rear right corner collide excessively with the tissue. These starting positions were furthest away from the grasping point and thus also lead to the longest trajectories. Longer trajectories were observed to have more collisions, since the simplified collision model in simulation is somewhat more permissive than reality. Future work will investigate additional reward terms that encourage safe policies that are robust to changes in environment dynamics. The trajectory outcomes of evaluation scenario $\pi_s(O_s)$ show that policy $\pi_s$ is robust to the physical domain gap between simulation and the real system even though the simulation neither models measurement inaccuracies nor the dynamic behavior of the cable driven mechanism of the PSM. Both $\pi_s(O_s)$ and $\pi_g(G(O_s))$ are evaluated on observations from the same distribution as they were trained on, but under different dynamics through the digital twin. The drop in performance when comparing $\pi_s(O_s)$ to $\pi_g(G(O_s))$ shows that policy $\pi_g$ learned a behavior that does not translate well to the additional safety constraints, *i.e.* trajectory terminated on excessive tissue stress, as described in Section III-C. This result indicates that learning on translated images results in policies less robust to changes in dynamics. This may be due to the relative visual complexity of the translated images compared to the ones from simulation, which may make it more difficult for the policy to infer the environment's state from the observation. Preliminary tests with policies trained on a single GAN showed much lower success rates than the proposed method that used an ensemble of GANs. This strengthens the idea that using an ensemble of GANs may be interpreted as a form of domain randomization. Future work may further investigate how employing an ensemble of GANs compares to classical visual domain randomization.

The GAN used in this work was trained with a total of $246$ TR trajectories and occasional random motions in the environment. This is substantially less data than state-of-the-art methods that achieve comparable task success in the real world, but require $10\,000$ task executions [12] and additional task specific auxiliary tasks or real world labels. Data collection for TR task execution was straightforward since the required motion could be planned from predefined start and end points of the motion. Tasks where motion planning has to consider the elastic behavior and dynamics of the deformable object may complicate data collection. Simulation and evaluation setup were calibrated to have the same camera perspective and object positions. This choice was made to limit the image-to-image translation task to changes in image appearance [12], [13].

The imbalance of required training steps for learning the two-phased task may indicate that the initially learned features relevant for solving the grasp phase contradict the features relevant for the retraction phase, and that relearning features that can be generalized to both phases requires prolonged learning time. Additional challenges were encountered over the course of this work that are not explicitly described in this contribution, but should be mentioned to aid future work in the field. (1) Learning success in simulation was highly dependent on the camera perspective. Multiple different camera positions were evaluated and some, that often did not substantially differ to the human eye, were so detrimental to training, that the task was not learnable by the agent. (2) Reward normalization was essential for fast and repeatable training success. (3) Normalizing the real observations with running mean and standard deviation to mitigate the effects of changes in $O_r$ caused by changing lighting conditions is unlikely to yield better results as changes in lighting change more than just the brightness of an image. Different lighting conditions may also change the overall hue of an image as well as the presence of shadows.

The achieved task success rate of $50\%$ is aligned with results obtained in state-of-the-art works for image-based sim-to-real transfer in robotics [10], although we expect that task success could be further improved by including more sophisticated methods such as hierarchical RL to learn a policy per phase, pretraining the agent with behavioral cloning, or including additional DR as proposed in [14]. We do not include these methods here to focus on DA rather than the absolute task success. Preliminary experiments with the same approach and hyperparameters on a different surgical robotic task, a Tissue Manipulation setup similar to [4], yield comparable results and a $63\%$ task success rate. In contrast to TR, Tissue Manipulation is a robotic control task in which a visual marking on a deformable object must be aligned with an overlayed target location on the image observation. The manipulation is indirect because the deformable object is manipulated at a grasping point that does not coincide with the visual marking of interest. Thus, the policy must learn to model the deformation in order to align the marking with the target location. In comparison to [4], our method does not rely on an image-processing pipeline to generate state-observations and does not require training the policy on the real robotic system. We plan to extend these preliminary experiments to show that the approach can be transferred to other image-based tasks without changes to the pipeline. Moreover, initial tests were performed with DR for the TR task but did not yield noticeable benefits over naive sim-to-real transfer without DA and was thus not included as a separate experiment.

Although the proposed pipeline is able to perform image-based sim-to-real transfer of a soft tissue manipulation task,

some limitations remain that will be addressed in future work. In the context of surgical task learning, the method may be transferred to other tasks without modification and creation of auxiliary tasks for stabilization. The learned weights of the image translation model, however, are unlikely to perform successfully on new tasks. Despite requiring significantly less data compared to related works [12], [13], it is necessary to record unpaired image data to train the translation models on a new task. From an image-to-image translation perspective however, the greatest limiting factor for practical application is that changes in camera perspective introduce additional changes in image content, not only in appearance. The presented approach thus relies on a precise calibration of camera positions between real system and simulation. Achieving view independence would increase the practical applicability of the approach for more realistic surgical scenarios where camera perspectives change over the course of the task and calibration between real setup and simulation is challenging.

Building on the achieved results, future work will investigate the approach on setups with increased realism that include complex visual features, a broader physical domain gap, and non-calibrated camera perspectives.

## VI. CONCLUSION

This work is the first successful demonstration of sim-to-real transfer of a visuomotor policy in the context of robotic surgery. A data efficient approach for sim-to-real transfer through DA utilizing a contrastive GAN is presented. The policy is trained in a soft-body simulation on image observations and then transferred to a real robotic system without retraining the policy. Furthermore, the contrastive GAN approach does not require task specific auxiliary tasks, and requires much less data to train compared to most examples from literature. Evaluation of the policy on the real system demonstrates successful sim-to-real transfer for a TR task and points out several critical challenges that must be addressed in future work. The successful transfer of an image based policy from simulation to a real robotic system for deformable object manipulation paves the way towards making RL viable for robotic surgery and is a sizeable leap towards cognitive surgical robots.

## REFERENCES


[1] B. Thananjeyan, A. Garg, S. Krishnan, C. Chen, L. Miller, and K. Goldberg, "Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 2371–2378.

[2] E. Tagliabue, A. Pore, D. Dall'Alba, E. Magnabosco, M. Piccinelli, and P. Fiorini, "Soft tissue simulation environment to learn manipulation tasks in autonomous robotic surgery," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2020, pp. 3261–3266.

[3] A. Pore, E. Tagliabue, M. Piccinelli, D. Dall'Alba, A. Casals, and P. Fiorini, "Learning from demonstrations for autonomous soft-tissue retraction," in *Int. Symp. Med. Robot. (ISMR)*, 2021, pp. 1–7.

[4] C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, and J. Rosen, "Autonomous Tissue Manipulation via Surgical Robot Using Learning Based Model Predictive Control," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3875–3881.

[5] C. D'Ettorre, S. Zirino, N. N. Dei, A. Stilli, E. De Momi, and D. Stoyanov, "Learning intraoperative organ manipulation with context-based reinforcement learning," *Int. J. Comput. Assist. Radiol. Surg.*, 2022.

[6] P. M. Scheikl, *et al.*, "Cooperative Assistance in Robotic Surgery through Multi-Agent Reinforcement Learning," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2021, pp. 1859–1864.

[7] 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 *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2017, pp. 23–30.

[8] S. Grün, S. Höninger, P. M. Scheikl, B. Hein, and T. Kröger, "Evaluation of Domain Randomization Techniques for Transfer Learning," in *Int. Conf. Adv. Robot. (ICAR)*, 2019, pp. 481–486.

[9] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 3803–3810.

[10] O. M. Andrychowicz, *et al.*, "Learning dexterous in-hand manipulation," *Int. J. Rob. Res.*, vol. 39, no. 1, pp. 3–20, 2020.

[11] J. Matas, S. James, and A. J. Davison, "Sim-to-Real Reinforcement Learning for Deformable Object Manipulation," in *Conf. Robot. Learn. (CORL)*, 2018, pp. 734–743.

[12] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai, "Retinagan: An object-aware approach to sim-to-real transfer," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 10 920–10 926.

[13] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari, "Rl-cyclegan: Reinforcement learning aware simulation-to-real," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11 157–11 166.

[14] S. James, *et al.*, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12 627–12 637.

[15] K. Bousmalis, *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 4243–4250.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[17] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Europ. Conf. Comp. Vis. (ECCV)*, 2020.

[18] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, "Dual contrastive learning for unsupervised image-to-image translation," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, 2021.

[19] B. T. Imbusch, M. Schwarz, and S. Behnke, "Synthetic-to-real domain adaptation using contrastive unpaired translation," in *IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, 2022, pp. 595–602.

[20] G. Narasimhan, K. Zhang, B. Eisner, X. Lin, and D. Held, "Self-supervised transparent liquid segmentation for robotic pouring," in *IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2022, pp. 4555–4561.

[21] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman, "Acting optimally in partially observable stochastic domains," in *Aaai*, vol. 94, 1994, pp. 1023–1028.

[22] A. Pore, *et al.*, "Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2021, pp. 4025–4031.

[23] A. Attanasio, *et al.*, "Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6528–6535, 2020.

[24] E. Tagliabue, D. Meli, D. Dall'Alba, and P. Fiorini, "Deliberation in autonomous robotic surgery: a framework for handling anatomical uncertainty," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 11 080–11 086.

[25] F. Faure, *et al.*, "SOFA: A multi-model framework for interactive physical simulation," in *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, 2012, vol. 11, pp. 283–321.

[26] G. Brockman, *et al.*, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

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

[28] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[30] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *J. Mach. Learn. Res.*, vol. 21, no. 1, 2022.