Furnari *et al. Art Int Surg.* 2025;5:103-15
DOI: 10.20517/ais.2024.32

Artificial Intelligence Surgery

Original Article

Open Access

Sequence-based imitation learning for surgical robot operations

Gabriele Furnari, Cristian Secchi, Federica Ferraguti

Department of Science and Methods for Engineering, University of Modena and Reggio Emilia, Reggio Emilia 42122, Italy.

Correspondence to: Dr. Gabriele Furnari, Department of Science and Methods for Engineering, University of Modena and Reggio Emilia, Tecnopolo, Piazzale Europa 1, Reggio Emilia 42122, Italy. E-mail: gabriele.furnari@unimore.it

How to cite this article: Furnari G, Secchi C, Ferraguti F. Sequence-based imitation learning for surgical robot operations. *Art Int Surg.* 2025;5:103-15. http://dx.doi.org/10.20517/ais.2024.32

Received: 28 May 2024 **First Decision:** 2 Aug 2024 **Revised:** 28 Aug 2024 **Accepted:** 30 Oct 2024 **Published:** 17 Feb 2025

Academic Editor: Peter Passias, Eyad Elyan, Andrew A. Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: This paper aims to advance autonomous surgical operations through imitation learning from video demonstrations.

Methods: To address this objective, we propose two main contributions: (1) We introduce a new dataset of virtual kidney tumor environments to train our model on. The dataset is composed of video demonstrations of tumor removal from the kidney, executed in a virtual environment, and kinematic data of the robot tools; (2) We employed an imitation learning architecture composed of vision transformers (ViT) to handle the frames extracted from the videos and of a long short-term memory (LSTM) structure to process surgical motion sequences with a sliding window mechanism. This model processes video frames and prior poses to predict the poses for both robotic arms. A self-generating sequence approach was implemented, where each predicted pose served as the latest element in the sequence, subsequently used as input for the next prediction together with the current frame of the video. The choice of architecture and methodology was guided by the need to effectively model the sequential nature of surgical operations.

Results: The model achieved promising results, exhibiting an average position error of 0.5 cm. The model was able to execute correctly 70% of the test tasks. This highlights the sequence-based approach's efficacy in capturing and predicting surgical trajectories.

Conclusion: Our study supports imitation learning's viability for acquiring task execution policies in surgical robotics.

The sequence-based model, combining ViT and LSTM architectures, successfully handles surgical trajectories.

Keywords: Imitation learning, robot-assisted surgery, artificial intelligence

INTRODUCTION

Significant strides have been made in the field of autonomous surgical operations, with researchers exploring various approaches to enhance the capabilities of surgical robots and improve patient outcomes[1–5]. However, integrating autonomous procedures into surgical practice remains a formidable challenge[6,7]. While supervised learning approaches have seen widespread adoption in domains such as image recognition and natural language processing, their application in surgical tasks remains limited. This limitation primarily stems from the scarcity of high-quality datasets, either simulated or real-world, necessary for training and validating machine learning models for surgical applications. In surgical robotics, where precision, safety, and reliability are of paramount importance, supervised learning poses unique challenges. Unlike tasks with well-defined goals and abundant training data, surgical procedures involve intricate spatial-temporal interactions and anatomical complexities. Training models solely on labeled data often fails to capture the nuances of surgical tasks, leading to sub-optimal performance and limited generalization capabilities. The majority of robot surgical datasets are made for phase detection of procedures[8,9] or object (such as organs or tools) detection and segmentation[10–12]. These datasets lack either kinematic data or task execution in environments with varying morphologies, making generalization across different scenarios difficult.

The JIGSAWS dataset[13] is probably the most known in surgical robotics. This dataset includes 76-dimensional kinematic data along with the video data of 101 trials of three elementary surgical tasks (suturing, knot-tying, and needle-passing) performed by six surgeons using the da Vinci Research Kit (dVRK), but it does not cover tasks performed in different scenarios. Similar considerations can be asserted for other works that use dVRK[14]. To address these challenges, this paper proposes a model to learn task execution in unseen environments. This necessitates learning a policy that dictates the agent's behavior based on its observed state. We achieve this by leveraging imitation learning[15], where agents learn by imitating expert demonstrations. This approach is particularly valuable in surgical robotics, where purely vision-based or geometric approaches for policy generalization across diverse environments are difficult, if not impossible.

By leveraging expert demonstrations, it empowers agents to learn complex tasks by mimicking skilled human behavior.

In domains such as robotic manipulation, imitation learning has proven particularly effective for tasks such as pick and place operations[16,17]. In pick and place tasks, the focus is often on learning the optimal picking and placing poses rather than the entire trajectory. Another prominent application of imitation learning is in the field of autonomous driving[18]. Here, agents learn to navigate complex road environments by imitating the driving behavior of human experts. Unlike continuous control tasks, actions in autonomous driving are often discrete, such as turning left or right, braking, or accelerating.

While imitation learning has proven successful in other domains, its application in surgery has been limited. Although it has the potential to teach agents to perform surgical tasks by observing expert behavior, this potential has not been fully realized, primarily due to the scarcity of high-quality datasets and the unique challenges posed by surgical environments. Although some datasets for surgical tasks do exist, they are mainly focused on segmentation and phase recognition. In the context of surgical robotics, imitation learning offers a promising avenue for teaching agents surgical procedures based on observed expert actions and circumvents the need for explicit programming or manual specification of task rules, allowing agents to acquire skills through demon-

stration and imitation. However, proposed models lack generalization across different scenarios[19] or use different approaches and very simple tasks[20–22].

The main contribution of this paper is:

- We address the scarcity of high-quality data in surgical robotics by introducing a novel dataset of synthetic kidney tumor environments. This dataset is generated by leveraging advanced 3D reconstruction techniques, based on real patients' computed tomography (CT) scans to create accurate kidney and tumor models. These models are then imported into a simulation environment, enabling the simulation of surgical tasks such as tumor removal. This approach provides a realistic and diverse set of environments for training and testing autonomous surgical systems.
- We propose a novel neural network architecture specifically designed for imitation learning in surgical robotics. This architecture combines the strengths of vision transformers (ViT) and long short-term memory (LSTM) networks. The ViT component effectively extracts spatial features from visual data, which is crucial for tasks such as tumor segmentation in the kidney environment. Meanwhile, the LSTM component excels at capturing temporal dependencies, which is essential for learning the sequential nature of surgical procedures. By combining these capabilities, our model learns to execute tumor removal tasks in new kidney tumor environments. Leveraging imitation learning techniques, the model learns from expert demonstrations in similar environments and adapts this knowledge to perform tasks autonomously in previously unseen scenarios. This demonstrates the efficacy of policy learning for autonomous surgical robotics.

METHODS

Creation of the kidney tumor dataset

In the realm of surgical robotics and automation, the development of algorithms capable of mimicking expert surgeon actions is paramount. One approach to achieve this is through imitation learning, where algorithms learn from human demonstrations. However, acquiring a dataset that accurately represents surgical scenarios poses significant challenges. In this section, we detail the creation of a dataset for imitation learning focused on tumor removal from the kidney, leveraging advanced simulation techniques and medical imaging technology. While our study focused on the specific task of kidney tumor removal, the potential applications of imitation learning in surgical robotics extend far beyond this single procedure. Indeed, its versatility and adaptability make our work well-suited for a wide range of surgical tasks across various anatomical structures and organs. In fact, we envision expanding our research to encompass a broader array of other organs and surgical procedures.

We utilized simulation open framework architecture (SOFA) (https://www.sofa-framework.org/) software because it can handle real-time deformation of soft objects, akin to human organs. The deformable nature of organs, such as the kidney, presents a complex challenge for surgical simulations. By applying empirically determined deformation parameters, including elasticity modulus, we could realistically simulate the behavior of soft tissues during surgical manipulation. By default, SOFA framework uses Young Module (E) and Poisson Coefficient (ν) as deformation parameters. Through empirical experimental tests[23], it has been demonstrated that the elasticity modules of the kidneys are E=180.32 kPa eG = 95.64 kPa (tangential elasticity modulus), which can be related by introducing the following formula:

$$G = \frac{E}{2(1+\nu)} \tag{1}$$

Therefore, we obtained $\nu = -0.057$. The mechanical characteristics of renal tumors, as highlighted in the research by Levillain *et al.*, may vary considerably based on the type of tumor, but they are generally softer than the renal tissue[24]. The following values were empirically identified, such as those that ensure biomechanical behavior as realistic as possible: E= 5 kPa eν=-0.8.

The goal is to faithfully replicate the physical stresses that occur during actual surgery, including gravitational forces, tissue deformations, and collisions between the instruments and the organs.

The procedure considered as a use case involves the use of an instrument to remove the tumor from the surrounding tissue and another tool to grab and remove it. We aim to reproduce every phase of this procedure promptly, ensuring that instruments behave realistically and that the tumor is removed precisely. This makes it possible to accurately emulate the removal of a renal tumor through the use of robotic surgery.

A fundamental aspect is the generation of a complete dataset that includes realistic variations in tumors and organ conditions. This dataset will be used to train an imitation learning model, which will learn from video recordings of demonstrations and tracking of kinematic data to autonomously perform the surgical procedure on new scenarios.

Generation of 3D environments from patient CT scans

The creation of realistic 3D environments from patients' CT scans is essential for accurate surgical simulations. In this section, we detail the process of transforming CT scans into anatomically accurate 3D models of the kidney and tumor, employing advanced AI methods and simulation software. The first step in the process involves segmenting the kidney and tumor regions from the CT scans. To accomplish this, we employed a state-of-the-art deep learning architecture known as U-Net[25]. The U-Net model is trained to perform semantic segmentation, accurately delineating organ boundaries within medical images. By training the U-Net on annotated CT scans, we obtained precise segmentation masks outlining the kidney and tumor regions[26]. With the segmented regions extracted, we proceeded to reconstruct 3D models of the kidney and tumor. Each segmented region was converted into a three-dimensional mesh representation using standard techniques. The mesh generation process involved interpolating the segmented contours to create a continuous surface, resulting in anatomically faithful 3D representations. Following mesh generation, we conducted refinement and optimization procedures to enhance the quality of the 3D models. This included smoothing the mesh surfaces to remove artifacts and irregularities introduced during the reconstruction process. Additionally, we optimized the mesh topology to ensure computational efficiency and stability during simulation. The 3D models are then refined and colorized, imported into the SOFA software environment, as shown in Figure 1. SOFA facilitated the integration of the 3D models into dynamic simulation scenarios, enabling real-time deformation and interaction with surgical instruments. The anatomically accurate 3D environments created from patient CT scans served as the foundation for realistic surgical simulations and were used to perform the task, i.e., tumor cutting and removal from the kidney. In our work, we generated a total of 53 distinct kidney tumor environments, each characterized by unique variations in tumor morphology and relative spatial arrangement since each of them corresponds to a different patient. To introduce diversity in surgical scenarios, we varied the shape and size of tumors across different environments. This variation in tumor morphology influenced the trajectories of surgical instruments, as surgeons navigated through different tissue structures and encountered varying degrees of resistance. Additionally, the size and shape variations ensured that each simulation presented unique challenges, reflecting the heterogeneity observed in clinical scenarios. Another factor contributing to the diversity of environments was the relative positioning of the tumor with respect to the kidney. By manipulating the spatial relationship between the tumor and the kidney, we translated the trajectories of surgical tools in space, simulating scenarios where tumors were located in different regions of the kidney. These differences were carefully chosen to introduce variability in the trajectories of surgical tools and to create visually distinct environments for simulation purposes, so that the ViT-LSTM model can learn the tumor-kidney spatial relationships and generalize across different scenarios. Despite these variations, the position of the kidney remained consistent across all environments, centered at the origin of the SOFA environment.

Figure 1. 3D reconstruction of the Kidney tumor environment imported into SOFA. SOFA: Simulation open framework architecture.

Video demonstrations execution

To capture surgical demonstrations for the imitation learning dataset, we employed a manual execution approach conducted by a single engineer, as our primary focus was testing the model across different environments rather than evaluating variability across different experts in the same environment. This approach enabled us to assess the model's adaptability and robustness in diverse scenarios. The demonstrations were made using specialized surgical tools within a simulated environment. The task of tumor removal from kidneys was executed using two distinct instruments: a cauter for the cutting phase and a grasper for grasping and removing the tumor post-incision. These instruments were controlled manually via a console controller, allowing for precise manipulation within the simulation environment. The integrated RGB-D camera is used as a stereoscope to capture videos of the demonstrations. The cauter, employed for the cutting phase, simulated the use of electro-cautery to incise tissue. This instrument was responsible for executing precise incisions along predetermined trajectories within the kidney tumor environment. Conversely, the grasper functioned as a versatile tool for grasping and removing the dissected tumor fragments from the surgical site. Together, these tools enabled the execution of a complete tumor removal procedure within the simulation. The execution of surgical demonstrations was facilitated by a modified version of the tissue dissection application developed within the SOFA_ENV framework, utilizing LapGym[27]. This application was tailored to our specific use case, providing a specialized environment for executing kidney tumor removal procedures. By leveraging the capabilities of LapGym, a laparoscopic surgery simulation module within SOFA_ENV, we were able to recreate realistic surgical scenarios while enabling manual control of surgical instruments. The cutting procedure starts from the left side of the tumor and proceeds clockwise along the visible perimeter. Once the cutting edge has completed more than half of its trajectory and passed the gripping tool, positioned on the right side, the latter approaches the tumor to grasp it and establish one secure grip, holding it in place for the final stage of the procedure. Once all contact points are removed, the tumor is completely separated from the kidney and must be excised from the surrounding healthy tissue: the task is considered complete when the tumor is positioned at a safe distance from the organ. Some of these steps are shown in Figure 2. It is very important to carry out all simulations, always respecting the same guidelines to ensure a certain consistency between the dataset demonstrations: not only is it essential to carry out the task correctly, but it is equally crucial to obtain a homogeneous dataset so that, during the training phase, the learning agent can recognize visual and sequential patterns in the actions performed across different environments.

Dataset composition

The dataset for our imitation learning project is composed of multi-modal data captured during the execution of surgical demonstrations within the simulated kidney tumor environments. Each demonstration is recorded using an RGB-D camera, providing both visual and depth information essential for understanding the surgical scene. The primary component of the dataset consists of videos captured through RGB-D cameras, with each demonstration recorded as a sequence of frames. These videos offer a detailed visual representation of

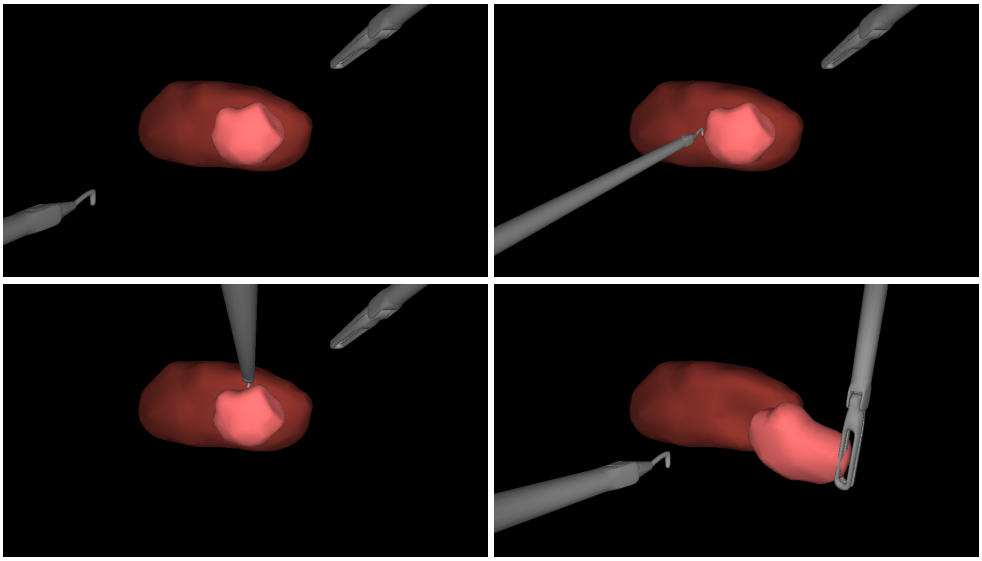

Figure 2. Some steps of the execution of tumor cutting and removal from the kidney executed by the expert. The cauter (tool on the left) cuts the tumor and the grasper (tool on the right) takes the tumor out of the kidney.

the surgical procedure, enabling the observation of tool manipulation and tissue interaction throughout the task execution. In addition to RGB video frames, depth information is also captured for every frame of each video using the same RGB-D camera. These depth data provide crucial spatial context, allowing for accurate estimation of the 3D geometry of the surgical scene. Depth information is invaluable for understanding the relative positions of surgical instruments, tissues, and anatomical structures within the environment. Alongside video and depth data, the dataset includes precise 3D pose information for both the cauter and the grasper instruments. The pose data comprises the XYZ position and unit quaternion for orientation of the tips of both tools at each time step of the trajectory. These tool poses serve as ground truth data for our imitation learning model, guiding the prediction of tool actions in subsequent time steps. Additionally, two more variables, one for each tool, are provided, describing the state of the tools: for the cauter, a boolean variable is considered with the value of 1 if the tool is cutting, 0 otherwise; for the grasper, the variable considered indicates the opening angle of the tool. In conclusion, the creation of the dataset described herein was necessitated by the absence of suitable resources for learning surgical tasks in the intended manner. Despite the growing interest in surgical automation and imitation learning, existing datasets fail to adequately address the specific requirements of our task. Real-world surgical datasets often lack the detail and precision required for effective imitation learning. Obtaining annotated data from actual surgical procedures is challenging due to ethical and logistical constraints. Additionally, real surgical environments may exhibit variability that is difficult to control and replicate consistently. A challenge for the future will be to have a good-quality dataset with real-world data. On the other side, while simulation platforms would offer a controlled and safe environment and a convenient, albeit preliminary, way for data collection, there are very few platforms where surgical environments can be simulated, and the existing datasets within these platforms often lack the nuanced details necessary for accurate imitation learning, such as realistic tissue deformation, instrument-tissue interaction dynamics, and the kinematics data of the tools exploited. Furthermore, the specific task of tumor removal from kidneys presents unique challenges that necessitate the creation of a tailored dataset.

Imitation learning model

Problem formulation

At its core, imitation learning seeks to train an agent to replicate expert actions based on observed demonstrations, without requiring explicit task specification or reward design. One common formulation of imitation learning is behavioral cloning from observation[28], where an agent learns to imitate expert behavior by di-

rectly mapping observed states to corresponding actions. In this paradigm, the agent aims to replicate the demonstrated behavior without explicitly understanding the underlying task dynamics. Instead, it learns a mapping from state observations to actions through supervised learning techniques. Central to the success of imitation learning is the notion of policy learning. A policy represents the strategy or decision-making function employed by an agent to select actions based on its observations of the environment. In our particular context, we employ behavioral cloning to train an agent for performing surgical tasks, such as tumor removal from kidneys. At each step of the task execution, the agent must decide on an action based on the current state of the environment and past states. This involves finding a policy - a function of the current visual observation and the state of the surgical tools (e.g., kinematic data) - that dictates the agent's actions to achieve the desired surgical outcome, i.e., the state the tools have to assume in the next time step. We can formulate our problem as follows: let s_t denote the current state of the environment at time step t, and a_t represent the action taken by the agent in response to s_t, and the agent learns a mapping from states to actions, represented by a policy $\pi(s)$. This mapping is learned from a dataset of expert demonstrations $\{(s_1, a_1), (s_2, a_2), \ldots, (s_N, a_N)\}$, where N is the number of demonstrations. The goal of policy learning is to find an optimal policy π^* that minimizes the discrepancy between the agent's actions and the expert actions. In the case of supervised learning, this discrepancy is typically measured using a loss function $L(\pi(s), a)$.

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{N} L(\pi(s_i), a_i) \tag{2}$$

where s_i and a_i are the states and actions from the expert demonstrations, respectively.

In our context, the agent's action at each time step is determined by a policy $\pi(s_t, \theta)$, parameterized by θ. The policy is learned to mimic the expert behavior observed in the demonstrations, where θ is optimized to minimize the loss between the agent's actions and the expert actions, as follows

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} L(\pi(s_i, \theta), a_i) \tag{3}$$

where N is the number of demonstrations. The state s_t is composed of the observation and the robot tools' state (their kinematic data), denoted as $s_t = (o_t, \text{tools}_t)$. Additionally, the parameter θ depends on past states, reflecting the sequential nature of policy learning in imitation learning tasks.

Let s_t denote the state of the environment at time step t, consisting of the observation and the predicted action at time t. Additionally, let w represent the size of the window of past states, including both past observations and past predicted actions.

The action at time step $t+1$, denoted as a_{t+1}, is obtained based on the application of the policy function $\pi(s_t, \theta)$ to the current state s_t, along with a window w of past states $s_{t-w+1}, s_{t-w+2}, \ldots, s_t$:

$$a_{t+1} = \pi(s_t, s_{t-1}, \ldots, s_{t-w}) \tag{4}$$

where $\pi(s_t)$ is the policy function that takes the current state s_t as input and predicts the action a_{t+1} to be taken at the next time step.

Each state s_t in the window is composed of both past observations and past predicted actions:

$$s_t = (o_t, a_t) \tag{5}$$

where o_t represents the observation at time step t, and a_t represents the predicted action at time step t.

Proposed architecture

The imitation learning model employed in our study leverages a hybrid architecture comprising a ViT [29], pre-trained on ImageNet, and a LSTM network as shown in Figure 3. This architecture is designed to effectively handle the multi-modal nature of the dataset, integrating both visual information from RGB-D frames and temporal sequences of tool poses. The model takes, as input, sequences of RGB-D frames along with past predictions iteratively to predict the actions that the surgical tools should take in the next time step, i.e., the poses every tool has to assume in the next step. By iteratively refining its predictions based on past observations and ground truth tool poses, the model learns to mimic the expert demonstrations and perform surgical tasks autonomously in new kidney tumor environments. The ViT serves as the backbone of our model, specifically configured with 16 patches and a resolution of 224 pixels, and is responsible for processing the visual information extracted from RGB-D frames. By employing self-attention mechanisms, the ViT can capture spatial relationships and long-range dependencies within the video data, enabling effective feature extraction from each frame. This allows the model to encode rich visual representations of the surgical scene.

In our model, the LSTM network is a sequence-to-one architecture, where it takes a sequence of input features and predicts a single output value. It plays a crucial role in processing sequences of features extracted by the ViT: it is responsible for capturing temporal dependencies and learning the dynamics of the surgical task over time. We consider the trajectories of surgical tools as sequences of steps, where the action at time step t is influenced by both the preceding steps and the current state of the environment. This approach enables us to effectively model the long-term dynamic of the trajectories in surgical operations. We opted for an LSTM network instead of a traditional deep network because LSTMs are specifically designed to handle sequential data, such as the trajectories in our study. The LSTM's ability to retain information over time and capture dependencies between time steps allows it to model the sequential nature of surgical operations more effectively. This approach resulted in better performance and more accurate predictions compared to using a standard deep network. Moreover, an attention mechanism is incorporated within the LSTM architecture to enhance the model's ability to focus on relevant parts of the input sequence. The LSTM layer processes the input sequence, which consists of features extracted by the ViT at time step t and the previous w features, along with the corresponding predicted actions of the robot tools. This sequence is fed into the LSTM, allowing the model to capture long-term dependencies and temporal dynamics.

Following the LSTM layer, an attention mechanism is employed to dynamically weigh the importance of each input feature in the sequence. The attention mechanism computes attention weights using a linear layer applied to the LSTM outputs. These weights reflect the relevance of each feature in the sequence, allowing the model to focus more on informative features while attenuating the impact of less relevant ones. The attention weights computed by the linear layer are used to compute a weighted sum of the LSTM outputs, yielding an attended input representation. This attended input is then passed through another linear layer to produce the final output, which represents the action the robot tools are expected to take at the next time step. The integration of the attention mechanism within the LSTM architecture enhances the model's performance by enabling it to selectively focus on relevant information in the input sequence, thereby improving its predictive capabilities.

Training of the model

During each training cycle, the model processes a single RGB-D frame and corresponding data from previous time steps to predict the actions of the surgical tool. The process unfolds as follows: (1) The RGB-D frame at time t is divided into patches and passed through the ViT and a multilayer perceptron (MLP) to extract high-dimensional features capturing spatial and temporal information; (2) The features extracted at time t are combined with features from previous time steps $t - 1$ to $t - w$, where w is the window size of past states considered in the input sequence. Additionally, past predicted actions from time t to $t - w$ are included in the fusion process. This temporal fusion mechanism enables the model to leverage historical information

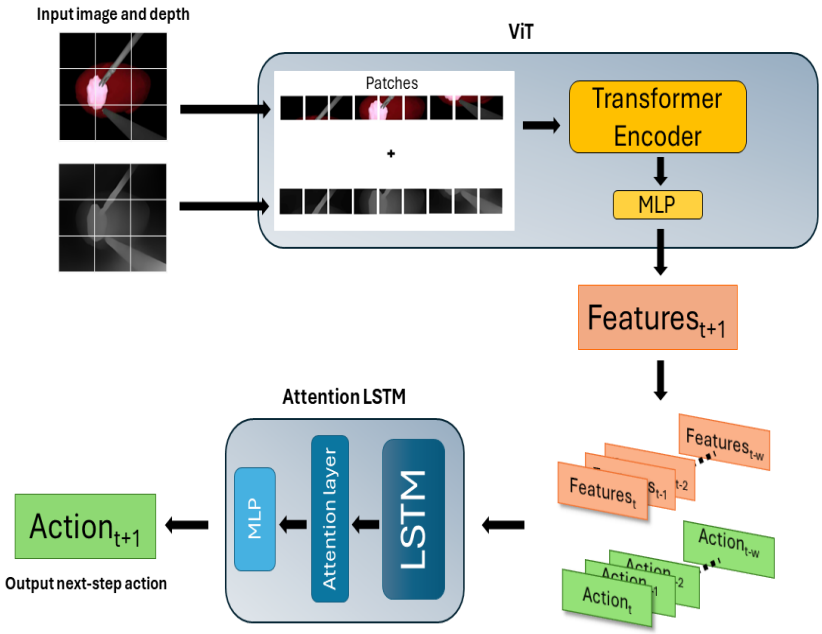

Figure 3. Cycle to predict an action for a single step of the trajectory.

and past predictions for context-aware action prediction; (3) The temporally fused features are passed into an Attention LSTM block, which comprises three main components: an LSTM layer, an attention mechanism, and a MLP. The LSTM layer processes the temporally fused features, capturing temporal dependencies and dynamics across the input sequence.The attention mechanism computes attention weights using a linear layer applied to the LSTM outputs. These weights reflect the relevance of each feature in the sequence, allowing the model to focus more on informative features while attenuating the impact of less relevant ones. The output of the attention mechanism is then passed through the MLP to produce the final prediction. The MLP outputs a 16-element vector comprising the xyz positions, unit quaternions of the two surgical tools, the cutting/not-cutting boolean variable for the cauter, and the opening angle of the grasper. This process is repeated for each point in the trajectory and for every trajectory in the dataset (i.e., every video of the dataset), allowing the model to learn and generalize across diverse surgical scenarios.

For training the model, we employed Adam optimizer, with an initial learning rate of 10^{-5}, which is reduced if the loss does not improve for 5 consecutive epochs (best results obtained to 10^{-6}). This adaptive learning rate schedule helps the model to converge more effectively and avoid getting stuck in local minima. We empirically determined that a window parameter of 10 elements in the input sequence provides a good balance between capturing temporal dependencies and computational efficiency. This window size allows the model to consider relevant past states and actions while avoiding excessive computational overhead. The model was trained for 50 epochs. Through experimentation, we observed that training beyond this point did not lead to significant improvements in performance, indicating that the model had converged to a stable solution. A custom loss function, which we call Temporal Loss, is introduced to address the challenge of mispredictions in later time steps. This loss function is designed to penalize errors more heavily as the predicted actions deviate further from the ground truth over time. By assigning higher weights to errors in later time steps, the Temporal Loss encourages the model to prioritize accurate predictions for actions in the immediate future while still considering the overall trajectory of the surgical task, since, due to the complex and dynamic nature of surgical procedures, the model may encounter difficulties in accurately predicting actions in later stages of the task. This is often exacerbated by accumulated errors and uncertainties as the task progresses. The dataset was split into

training and testing sets using an 85/15 ratio. This division ensures that the model is trained on a sufficiently large portion of the data while still reserving a separate subset for evaluating its generalization performance on unseen examples.

RESULTS

In our study, we utilized two tools to predict trajectories: a grasper and a cauter. The grasper presented a relatively straightforward challenge, as both the trajectory and the task were simple and easy to model. For this reason, the successful rate was 100% for the grasping part of the task. However, the cauter posed a more significant challenge due to the complexity of its trajectory and the precision required for the task. Predicting accurate trajectories for the cauter is crucial, as it directly impacts the effectiveness and safety of the surgical procedure. Consequently, the results section primarily focuses on the performance of the model with the cauter, as it represents the more demanding aspect of our work.

During training, we monitored the Mean Absolute Error on positions, achieving an average result of 0.5 cm. This metric quantifies the average discrepancy between the predicted and ground truth positions of the robot tools, providing insight into the model's accuracy in mimicking expert behavior.

Following the training part, we evaluated the model's performance in previously unseen environments to assess its generalization capabilities and effectiveness in simulated scenarios. The trained model demonstrated a commendable success rate, correctly executing the surgical task in 70% of new environments (7 out of 10 total tests). This achievement underscores the model's ability to generalize from training data to unseen test environments and effectively replicate expert behavior in diverse surgical settings.

In some scenarios, the model generated trajectories that surpassed the quality of the target trajectories observed in expert demonstrations, as shown in the example in Figure 4. These trajectories exhibited smoother and more continuous motion, effectively avoiding uncertainties, shakings, and corrections typically associated with manual execution. This indicates the model's potential to enhance the precision and efficiency of surgical procedures by minimizing unnecessary movements and ensuring consistent tool manipulation.

The results of our study indicate that there are no significant differences in performance between right and left tumors. Additionally, the translation of the tumor in space - meaning the position of the tumor relative to the kidney - does not have a substantial impact on the outcomes. This suggests that our model is robust to variations in tumor location and side, maintaining consistent performance regardless of these factors.

Despite its overall success, the model encountered difficulties in environments where the tumor shape and size deviated significantly from the norm. In such cases, characterized by unusually shaped tumors or exceptionally large sizes, the model exhibited reduced performance and struggled to execute the task accurately, i.e., the trajectory generated by the model deviates significantly from the target trajectory. As a result, the tumor may not be fully removed from the kidney, or the surgical margins may be excessive, leading to the removal of more kidney tissue than necessary. These outliers present unique challenges for the model, highlighting the need for further refinement of the model to ensure precise and accurate surgical outcomes and the importance of robustness and adaptability in real-world applications. The accompanying video clip shows the results in the virtual environment.

DISCUSSION

The project described in this work demonstrates the potential of imitation learning for automating surgical tasks. By leveraging synthetic environments and expert demonstrations, the trained model exhibits promising

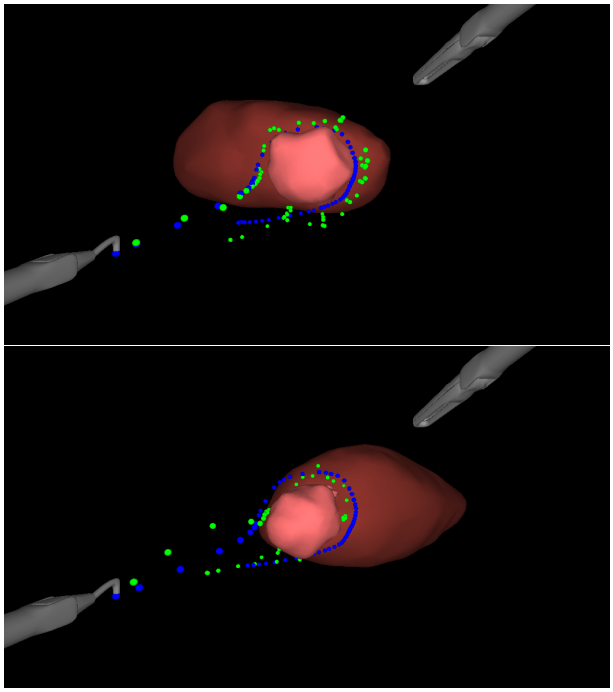

Figure 4. Two examples of successful tasks executed generating the trajectory using our model. In the picture, the target trajectory is highlighted in green while the generated trajectory is in blue. In the first case, we can notice that the generated trajectory is smoother and more accurate than the target trajectory.

capabilities in executing complex surgical procedures autonomously.

While direct comparisons with other works are challenging due to the differences in tasks, methods, and environments, we can still highlight some key distinctions. Unlike our study, which emphasizes the model's performance across different scenarios and environments, most existing works focus on variations in expert execution within the same environment. Furthermore, the tasks addressed in other studies often differ in complexity and objectives, making it difficult to draw straightforward parallels. Despite these differences, our work contributes uniquely by exploring the adaptability of the model across varying conditions, which is less commonly addressed in the current literature. In particular, the creation of the dataset and the development of the model represent significant contributions to the field of surgical robotics and automation. The creation of the dataset described in this paper fills a critical gap in the field of surgical imitation learning by providing a comprehensive training resource. Successful imitation learning relies on exposure to diverse environments. By creating a dataset encompassing a variety of kidney tumor scenarios, we ensure that our model learns to adapt to different anatomical configurations and surgical challenges. While the results are promising, challenges remain in handling outlier scenarios, since, when the tumor shape or size is significantly different from the norm, the model is not always able to execute the task, emphasizing the need for continued refinement and adaptation of the model to ensure robust performance in diverse surgical settings. Additionally, one of the primary challenges moving forward is the development of more realistic surgical environments. Current synthetic environments, while useful for training and initial testing, may lack the complexity and variability of real-world surgical settings. Future efforts should focus on enhancing the fidelity of simulated environments by incorporating more realistic texture of the tissues, adding other organs to vary the executed trajectories, and including elements such as blood and fat in the scene.

Furthermore, the tasks we modeled were simplified and not fully reflective of the complexity of real surgical procedures. While they are useful for initial experimentation, they do not capture the full range of challenges

encountered in real-world surgical tasks.

Transitioning from synthetic to real-world environments, both in-silico or real organs, presents a significant challenge in autonomous surgical systems. Real-world environments introduce a multitude of complexities, including variability in patient anatomy, tissue characteristics, intraoperative conditions, and robot control, which can lead to less accurate execution. Acquiring high-quality real-world datasets that capture this variability is essential for training models that can generalize effectively across diverse surgical scenarios.

While our research has primarily focused on automating specific surgical tasks such as tumor removal from kidneys so far, there is a need to expand the scope to encompass a broader range of surgical procedures, such as suturing and blood suction.

As the field of surgical robotics and automation continues to advance, addressing these future challenges will be paramount to the development of robust and versatile autonomous systems. By leveraging advances in machine learning, computer vision, and robotics, researchers can overcome these obstacles and pave the way for safer, more efficient, and more accessible surgical interventions.

DECLARATIONS

Acknowledgments
This work is part of the project "Robotic-assisted percutaneous nephrolithotomy with ultrasound guidance and 3D reconstruction superimposition" funded by the European Union - NextGenerationEU.

Authors' contributions
Main contribution, data collection, software development: Furnari G
Supervision, conceptual idea contribution, and paper review: Secchi C, Ferraguti F

Availability of data and materials
The dataset can be downloaded at the following link: Dataset link: https://drive.google.com/drive/folders/1vIstV6Y0zFjCd8smQi7ZwWHW4ahqQwTH?usp=drive_link. The supplementary material of this work can be found at the following link: Video link: https://drive.google.com/file/d/1W4DrWhKJeNbxEHUFLbh9aEEGMCK0KP9v/view?usp=drive_link.

Financial support and sponsorship
This work is part of the project "Robotic-assisted percutaneous nephrolithotomy with ultrasound guidance and 3D reconstruction superimposition" funded by the European Union - NextGenerationEU.

Conflicts of interest
All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate
All the CT scans used in this work are obtained anonymously from the Ospedale Civile Sant'Agostino - Estense hospital (Baggiovara, MO, Italy) and from the public database Cancer Imaging Archive.

Consent for publication
Not applicable.

REFERENCES

1. Richter F, Shen S, Liu F, et al. Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection. *IEEE Rob Autom Lett* 2021;6:1383-90. DOI
2. Saeidi H, Opfermann JD, Kam M, et al. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Sci Robit* 2022;7:abj2908. DOI
3. Su B, Yu S, Li X, et al. Autonomous robot for removing superficial traumatic blood. *IEEE J Transl Eng Health Med* 2021;9:1-9. DOI
4. Saeidi H, Le HND, Opfermann JD, et al. Autonomous laparoscopic robotic suturing with a novel actuated suturing tool and 3D endoscope. In: 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20-24; Montreal, Canada. IEEE; 2019. pp. 1541-7. DOI
5. Ginesi M, Meli D, Roberti A, Sansonetto N, Fiorini P. Autonomous task planning and situation awareness in robotic surgery. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24 - 2021 Jan 24; Las Vegas, USA. IEEE; 2021. pp. 3144-50. DOI
6. Attanasio A, Scaglioni B, De Momi E, Fiorini P, Valdastri P. Autonomy in surgical robotics. *Annu Rev Control Robot Auton Syst* 2021;4:651-79. DOI
7. Han J, Davids J, Ashrafian H, Darzi A, Elson DS, Sodergren M. A systematic review of robotic surgery: from supervised paradigms to fully autonomous robotic approaches. *Int J Med Robot Comp* 2022;18:e2358. DOI
8. Bawa VS, Singh G, KapingA F, et al. The SARAS endoscopic surgeon action detection (ESAD) dataset: challenges and methods. arXiv. [Preprint.] Apr 7, 2021 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2104.03178.
9. Bawa VS, Singh G, KapingA F, et al. ESAD: endoscopic surgeon action detection dataset. arXiv. [Preprint.] Jun 12, 2020 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2006.07164.
10. Yoon J, Hong S, Hong S, et al. Surgical scene segmentation using semantic image synthesis with a virtual surgery environment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022. pp. 551-61. DOI
11. Carstens M, Rinner FM, Bodenstedt S, et al. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Sci Data* 2023;10:3. DOI
12. Colleoni E, Edwards P, Stoyanov D. Synthetic and real inputs for tool segmentation in robotic surgery. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 700-10. DOI
13. Gao Y, Vedula SS, Reiley CE, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. Available from: https://cirl.lcsr.jhu.edu/wp-content/uploads/2015/11/JIGSAWS.pdf. [Last accessed on 13 Nov 2024]
14. Rivas-Blanco I, Pérez-del-Pulgar CJ, Mariani A, Tortora G, Reina AJ. A surgical dataset from the da Vinci Research Kit for task automation and recognition arXiv. [Preprint.] Jun 29, 2023 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2102.03643.
15. Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: a survey of learning methods. *ACM Comput Surveys* 2017;50:1-35. DOI
16. Zeng Andy, Florence P, Tompson J, et al. Transporter networks: rearranging the visual world for robotic manipulation. arXiv. [Preprint.] Jan 5, 2022 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2010.14406.
17. Yang S, Zhang W, Lu W, Wang H, Li Y. Cross-context visual imitation learning from demonstrations. In: 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31 - Aug 31; Paris, France. IEEE; 2020. pp. 5467-73. DOI
18. Eraqi HM, Moustafa MN, Honer J. End-to-end deep learning for steering autonomous vehicles considering temporal dependencies. arXiv. [Preprint.] Nov 22, 2017 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.1710.03804.
19. Tanwani AK, Sermanet P, Yan A, Anand R, Phielipp M, Goldberg K. Motion2Vec: semi-supervised representation learning from surgical videos. arXiv. [Preprint.] May 31, 2020 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2006.00545.
20. Murali A, Sen S, Kehoe B, et al. Learning by observation for surgical subtasks: multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms. In: 2015 IEEE International Conference on Robotics and Automation (ICRA); 2015 May 26-30; Seattle, USA. IEEE; 2015. pp. 1202-9. DOI
21. Zhao TZ, Kumar V, Levine S, Finn C. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv. [Preprint.] Apr 23, 2023 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2304.13705.
22. Kim JW, Schmidgall S, Krieger A, Kobilarov M. Learning a library of surgical manipulation skills for robotic surgery. In: Bridging the Gap between Cognitive Science and Robot Learning in the Real World: Progresses and New Directions. 2024. Available from: https://openreview.net/forum?id=fYRlaylCI3. [Last accessed on 13 Nov 2024]
23. Karimi A, Shojaei A. Measurement of the mechanical properties of the human kidney. *IRBM* 2017;38:292-7. DOI
24. Levillain A, Confavreux CB, Decaussin-Petrucci M, et al. Mechanical properties of breast, kidney, and thyroid tumours measured by AFM: relationship with tissue structure. *Materialia* 2022;25:101555. DOI
25. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv. [Preprint.] May 18, 2015 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.1505.04597.
26. Wasserthal J, Breit HC, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell* 2023;5:230024. DOI
27. Scheikl PM, Gyenes B, Younis R, et al. LapGym - an open source framework for reinforcement learning in robot-assisted laparoscopic surgery. *J Mach Learn Res* 2023;24:1-42. Available from: https://www.jmlr.org/papers/volume24/23-0207/23-0207.pdf. [Last accessed on 13 Nov 2024]
28. Torabi F, Warnell G, Stone P. Behavioral cloning from observation. arXiv. [Preprint.] May 11, 2018 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.1805.01954.
29. Dosovitskiy A, Beyer A, Kolesnikov A, et al. An image is worth 16× 16 words: transformers for image recognition at scale. arXiv. [Preprint.] Jun 3, 2021 [accessed 2024 Nov 13]. Available from: https://doi.org/10.48550/arXiv.2010.11929.