Semantic-based autonomous context tracking in cognitive robotic surgery

Andrea Roberti, Marco Bombieri, Daniele Meli, Riccardo Muradore

***Abstract*—Autonomous robotic surgery can benefit from advances in artificial intelligence to improve the outcome of surgical procedures, enhance situation awareness, and optimize the user experience of surgeons. In this paper, we focus on the important problem of autonomous context tracking in robotic surgery, aiming at tracking the relevant items in the surgical scene with the endoscopic camera arm (ECM) of the da Vinci Research Kit (dVRK) robot. We propose *SemTrack*, a novel method to track both instruments and anatomical parts of interest, and overcome the lack of interpretability and trustworthiness of recent deep learning solutions. We leverage natural language processing to extract a symbolic task formalization from surgical notes and texts. We then use this interpretable formalization to track the flow of the phases in the operation, including both inter-phase transitions and intra-phase target anatomies and instruments. In the context of the tumor removal of phantom-based lateral partial nephrectomy, we validate the feasibility and accuracy of our methodology at tracking relevant scene items for enhanced situation awareness. Moreover, our approach has better performance also in terms of task duration, even with moving targets, than static and teleoperated ECM.**

***Index Terms*—Autonomous robotic surgery, endoscope control, situation awareness, visual servoing, natural language processing, semantic segmentation**

I. Introduction

AUTONOMY is the frontier challenge of robotic surgery, with the potential to significantly improve the outcome of the surgical procedure, enhancing faster, precise motion, and the surgeons' experience, minimizing their fatigue and cognitive load [1], [2]. Following the latest international regulations for safe intelligent systems[1], researchers have recently focused on designing reliable and trustworthy autonomous surgical robotic systems, e.g., by ensuring safe motion and interaction [3], [4]. Attempts have also been made to achieve conditional (i.e., human-supervised) autonomy [5], where task-level cognition based on formal representation and reasoning [6], [7], [8] can guarantee explainability and enhance user trust [9]. A crucial requirement for cognitive surgical robots is the autonomous operation of the endoscope, in order to adjust the field of view of the surgical camera and focus on relevant anatomical parts or instruments. This has the potential to increase the attention and minimize the effort of the surgeon in conditional autonomy, and to enhance perception capabilities of fully autonomous systems [2], [10]. However, most existing approaches focus only on instrument tracking [11], [12], while the focus on anatomical parts and a broader context is still an open challenge [10].

In this paper, we propose *SemTrack* a novel methodology for autonomous tracking of relevant surgical entities (*context*), including both instruments and anatomic structures. We start from a formal representation of the procedure extracted from expert-written textual descriptions, in the form of a Finite State Machine (FSM), triggering transitions between low-level steps when specific clinical (anatomical) conditions occur. Each state in the FSM, models a low-level step and refers to an anatomical target to be tracked via image-based visual servoing, while the surgical instruments (either autonomous or teleoperated) execute the task. Moreover, the relative positions of anatomical parts and instruments are used to determine the transitioning conditions between states of the FSM, triggering changes in the tracking targets. Anatomical parts are identified via semantic segmentation, a state-of-the-art data-efficient solution for image classification in surgery [13]. Finally, our solution automatically adjusts the pose of the endoscope (hence, the field of view of the surgical camera), following the part of interest for each step of the surgical procedure via image-based visual servoing.

Our method presents the following innovations with respect to the state-of-the-art:

- Differently from the most popular approaches for autonomous instrument tracking [14], we also track relevant anatomical parts, according to the combination of semantic segmentation and automatic state recognition. This has the potential to provide greater benefits to surgeons letting then focus their attention on salient clinical sub-scenes [10].
- Most existing approaches for autonomous endoscopes focus on anatomical targets and rely on black-box deep learning techniques to identify surgical steps and track relevant anatomies accordingly [10]. This may introduce inaccuracies in procedural representation, heavily relying on the availability of large training datasets of real executions, which are often unavailable in the surgical domain. On the contrary, we employ natural language processing based on a surgical language model [15], [16] to extract an interpretable formal representation of the task at hand, starting from available surgical notes.
- We extensively validate the performance of our methodology on the teleoperated tumor removal phase of Lateral Partial Nephrectomy (LPN), a crucial task in robot-

Andrea Roberti and Riccardo Muradore are with the Department of Engineering for Innovation Medicine, University of Verona, Italy, e-mail: {andrea.roberti; riccardo.muradore}@univr.it

Marco Bombieri was with the Department of Foreign Languages and Literatures, University of Verona, Italy, when this work was carried out, and is currently with the Department of Information Engineering and Computer Science, University of Trento, Italy, e-mail: marco.bombieri@unitn.it

Daniele Meli is with the Department of Computer Science, University of Verona, Italy, e-mail: daniele.meli@univr.it

[1]EU Artificial Intelligence Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

assisted minimally invasive surgery (RA-MIS) [17]. In particular, we investigate not only the accuracy of our methodology but also several aspects of the user experience, such as task performance (e.g., task duration), comparison with camera teleoperation, and the differences between discretized and continuous control.

The paper is organized as follows. In Section II, we first present the recent advances in autonomous endoscope control for robotic surgery. We then describe our methodology in detail in Section III, and validate it experimentally for the Lateral Partial Nephrectomy procedure in Section IV. Finally, in Sections V-VI we discuss our results, and present limitations and future developments.

II. Related works

Autonomous endoscope control is a fundamental capability of current and next-generation surgical robots, reducing surgeons' fatigue and cognitive workload while enhancing operational performance [10], [2]. More broadly, autonomous perception and camera management represent key building blocks towards higher levels of surgical robotic autonomy [1]. Most control algorithms for surgical cameras focus on tracking robotic instruments to guarantee procedural safety through dynamic adjustment of the field of view [2], [18], [19], with increasing evidence of effectiveness from the clinical community [20]. These approaches can successfully support novice training [21] and have been combined with virtual and mixed reality technologies to enhance surgeons' situational awareness [22], [23]. To improve software-based tracking performance, dedicated hardware and sensing solutions have also been developed [24]. Recent works have further explored surgeon-preference-guided camera control [14] and gesture-aware camera assistance [25], improving the adaptability of autonomous visual systems to the surgical context.

With the recent uptake of machine and deep learning techniques for surgical machine vision [26], [27], richer situation awareness can be achieved, paving the way towards the automation of increasingly complex surgical procedures [28]. This evolution has reshaped the field of autonomous endoscope control, where visual perception and understanding of the complete surgical scene are progressively assisting and replacing the mere tracking of surgical instruments [12]. For instance, [11] exploited instrument shape recognition from images for more accurate zoom control. In [29], an attention mechanism was developed to guide image focus in multiple simulated surgical scenarios, accurately matching human visual attention. [30] proposed a multi-stage image-guided strategy for autonomous endoscope navigation based on contour recognition in intestinal environments. Homography-based endoscope visual servoing for surgical manipulation was explored in [31], providing camera control capabilities independent of depth information. More recently, deep homography prediction has been proposed to imitate expert camera motions directly from endoscopic images, further demonstrating the potential of learning-based visual servoing for autonomous camera control [32].

Beyond low-level visual tracking, recent research has increasingly investigated semantic understanding of the surgical scene and workflow awareness. Surgical workflow analysis techniques can identify phases, steps, and activities from intra-operative data, providing contextual information that can support autonomous robotic behaviours [26]. Structured representations of procedures and surgical activities have been proposed to model the temporal evolution of interventions and the interactions among surgical agents [33], [34], [28]. Similarly, context-aware endoscope navigation systems have recently been developed to adapt camera behaviour according to the recognized procedural situation rather than solely to instrument positions [10]. These advances indicate a progressive shift from reactive camera control towards cognitive visual systems capable of exploiting semantic information about the ongoing surgical task.

The above methodologies either focus exclusively on autonomous endoscope navigation or address isolated stages of surgical procedures, often neglecting the explicit representation of the overall procedural workflow, which is central to the present work and to higher levels of surgical robotic autonomy. Recently, more sophisticated cognitive architectures for autonomous endoscope control have been developed, both for surgical training exercises [10] and real procedures such as rectal resection [35]. These solutions rely on recordings of previously executed procedures, which are processed through machine or deep learning algorithms to retrieve procedural models [34]. As highlighted in [35], this limits applicability in real surgical robotic scenarios. In fact, learned models often reproduce surgeon-specific skills and strategies rather than capturing a general and explicit representation of the procedure. Moreover, the adoption of learning-based autonomous systems in surgery remains challenging because of their limited interpretability and the difficulty of assessing their trustworthiness and reliability [36]. In addition, deep learning methods typically require large training datasets, which are often unavailable in surgical settings [2].

To address these limitations, recent works have explored explicit and symbolic representations of surgical knowledge for autonomous decision making. Logic-based frameworks have been proposed for surgical task planning and situation awareness [8], [9], while deliberative architectures have been developed to reason about procedural progression and anatomical uncertainty during surgery [6]. In parallel, natural language processing techniques have recently been employed to formalize procedural surgical knowledge from textual descriptions [16], [15], providing interpretable representations that can complement data-driven perception systems.

In contrast to existing approaches, our methodology combines semantic procedure understanding, interpretable task modeling, and autonomous context-aware visual tracking within a unified framework. Starting from textual descriptions of LPN procedures provided by surgeons, we follow [16] to derive a task formalization based on logical specifications and first-order predicates. The resulting representation explicitly encodes surgical actions, instruments, anatomical targets, and clinically relevant aspects of the procedure, which are combined through logical specifications to model both phase progression and task execution. Unlike conventional autonomous camera systems that predominantly rely on low-

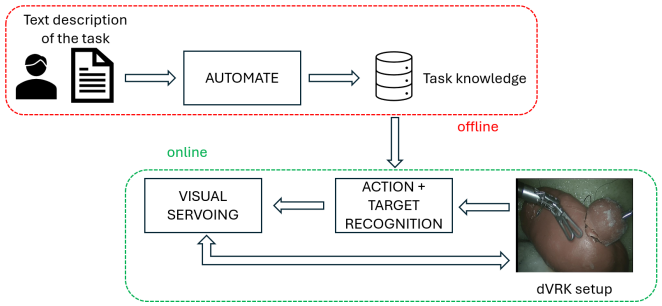


Fig. 1: Overview of the proposed methodology architecture, comprising an offline phase for automatic extraction of surgical task knowledge from textual descriptions, and an online phase for real-time target and action recognition with image-based visual servoing during procedure execution

level visual cues, instrument tracking, or data-driven imitation of expert behaviour, our framework exploits symbolic procedural knowledge to dynamically identify and track the scene entities that are relevant to the current surgical context. In this way, semantic information about the procedure directly guides visual attention and camera behaviour. Moreover, in contrast to workflow-recognition systems that provide contextual information without actively controlling the visual process, our approach closes the loop between procedure understanding and autonomous scene tracking. Finally, unlike purely learning-based cognitive architectures, the proposed methodology provides an explicit and human-interpretable representation of the surgical task, improving transparency, explainability, and traceability of the generated decisions. Therefore, this work contributes a semantic-based approach for autonomous context tracking that jointly exploits surgical knowledge representation, workflow awareness, and machine vision towards trustworthy cognitive robotic surgery.

III. Methodology

In this section we describe the main stages and components of our methodology (Figure 1). The architecture starts with an *offline* phase for the *automatic extraction of surgical task knowledge*, where the procedural flow involving sequences of steps, agents, and anatomical targets for each action is automatically retrieved from a specialized text description written by surgeons (e.g., procedural notes or books). Then, during the *online* phase, while the procedure is executed (either in teleoperation or autonomously), the simultaneous *target and action recognition* is performed from the kinematics and endoscopic camera reading, and an image-based *visual servoing* control strategy is implemented for *target tracking*.

A. Automatic task knowledge extraction from text

In the offline phase of our methodology, we want to automatically construct a formal representation of the surgical task as an FSM $\langle S, \Sigma, T, s_0, s_f \rangle$, where S is the set of states (in our case, actions); Σ is a set of symbols (in our case, conditions or events); $T : S \times \Sigma \rightarrow S$ is the transition map; s_0, s_f are the initial and final state(s), respectively. Moreover, we want to retrieve additional semantic information about the states / actions, representing them as predicates $s(a, t)$ to identify also the agent (arm) a executing the action and the anatomical target t of the action (e.g., `dissect(scissor, fat)`).

The input to our pipeline is a specialized text description $\mathcal{T}$. We use AUTOMATE [16] to derive the FSM and relevant information about the task. $\mathcal{T}$ must satisfy syntactic and semantic requirements described in [16], i.e., the use of present or imperative tense for verbs, both in active and passive forms; and the use of auxiliary verbs and implied subjects / objects.

For each sentence $\mathcal{S} \in \mathcal{T}$, AUTOMATE follows this pipeline.

1) Action identification: Standard Part-of-Speech (PoS) tagging [37] is used to identify the main verb $\mathcal{V}$, which corresponds to the main action to be executed.

2) Semantic analysis: Semantic Role Labeling (SRL) is used to identify semantic roles $\mathcal{R}$ of other parts of $\mathcal{S}$ associated with $\mathcal{V}$. We perform SRL leveraging the available PropBank linguistic and semantic dataset [38]. More specifically, SRL labels parts of text as *arguments*. The labels can either be *"Arg0"* for the subject, i.e., who performs the action; *"Arg1"* for the object, i.e, the entity affected by the action; *ArgM-ADV* for adverbial modifiers (including *if / else* conditions); *ArgM-TMP* for temporal markers; *ArgM-LOC* for locations; *ArgM-DIR* for indicating the direction of motion; and *ArgM-MNR* for modalities, i.e., describing how an action is executed [39].

As an example, consider the following sentence $\mathcal{S}$:

> The surgeon dissects the tumor in a medial-to-lateral direction, using cold Endoshears.

PoS tagging returns *dissect* as $\mathcal{V}$; then, SRL identifies the following roles $\mathcal{R}$:

> dissect(the surgeon (*Arg0*); the tumor (*Arg1*); in a medial-to-lateral direction (*ArgM-DIR*); cold Endoshears (*ArgM-MNR*)).

3) FSM construction: Given SRL annotations, we convert them to predicates in the form $\mathcal{V}$(*ArgM-MNR*, *Arg1*), where the subject is the robotic instrument (tagged as *ArgM-MNR)* and the anatomical target is the object (tagged as *Arg1*)[2]. Since we are interested in a formal description of the robotic task, we omit predicates not corresponding to robotic actions, e.g.,

> The surgeon identifies the renal vein.

Such predicates can be easily isolated, because they have as *Arg0* an human or because they do not contain any *ArgM-MNR* roles encoding specific surgical instruments[3].

The selected predicates correspond to states $S = \{s(a, t)\}$ in the FSM, and we assume that a transition condition (symbol) exists between consecutive predicates occurring in the text. The specific transitioning event $e_{1,2} \in \Sigma$ between two consecutive actions (states) $s_1(a_1, t_1) \xrightarrow{e_{1,2}} s_2(a_2, t_2)$ is then identified from *ArgM-ADV* or *ArgM-TMP* roles.

For instance, from the sentence

> The fat is dissected with the bipolar scissors. When the kidney is exposed, the tumor is removed with the scissors.

[2]In the passive verbal form, *Arg1* is replaced with *Arg0*, i.e., the subject is inverted with the object.

[3]As explained in [16] and confirmed by surgeons, the setup with the surgical instruments is usually described at the beginning of reports and procedural descriptions.

we identify `dissect(scissors, fat)` as s_1, `remove(scissors, tumor)` as s_2, and *the kidney is exposed* (*ArgM-TMP*) as $e_{1,2}$. If no *ArgM-ADV* nor *ArgM-TMP* is specified, we assume the disappearance from the scene of either the agent a_1 or the target t_1 as the transitioning event to s_2.

B. Action and target recognition

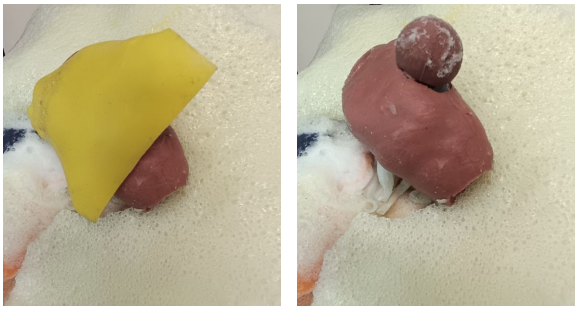
During the online phase of our methodology, while the surgical procedure is executed (either autonomously or in teleoperation), we need to recognize the current action $s_c(a_c, t_c)$ (state of the FSM) and the associated target t_c to be tracked via visual servoing. For action recognition, starting from the first action in the FSM retrieved offline, we monitor the occurrence of specific transitioning events in Σ, in order to trigger action change. To this aim, the inputs of our recognition module are the FSM description and the stream of RGB images from the endoscopic camera. Since the target and the semantics of the transitioning events depend on the surgical context, i.e., the anatomies and the instruments (e.g., the visibility of the kidney in the previous example), we employ a *semantic segmentation module* to recognize the regions and centroids corresponding to the different items at run time. The semantic segmentation module is based on the Generative Adversarial Network (GAN) model [40]. Unlike other models, this network requires only a small number of RGB samples for pre-training and has fast inference times (we achieve ≈ 15 Hz end-to-end control rate), thus representing a convenient and scalable choice for surgical scenarios. Specifically, the model employs a conditional GAN architecture, where the generator produces segmentation maps conditioned on input images, and the discriminator guides the generator toward more accurate outputs by distinguishing real from synthetic segmentations. This adversarial training strategy not only improves segmentation accuracy but also helps capture fine anatomical structures, making it particularly suitable for data-constrained medical environments. We perform data augmentation and cross-validation to prevent overfitting. We refine the output of the GAN via morphological operations to remove noise. Moreover, we compute the centroid $\bar{t}_c$ of the mask of the target t_c region in image space, serving as the key feature for the endoscopic arm (ECM) to track.

C. Visual servoing

Given $\bar{t}_c$, the centroid of the target of the current action, the visual servoing module controls the motion of the endoscopic camera to center on it. More specifically, the goal of the ECM controller is to minimize at each time step the error

$$\boldsymbol{e}(t) = \boldsymbol{\phi}(\bar{t}_c(t), \eta) - \boldsymbol{\phi}^*, \tag{1}$$

where $\boldsymbol{\phi}(\bar{t}_c(t), \eta)$ is a vector of k visual features, in which η is a set of parameters that represent potential additional knowledge about the vision and controlled system (in our case, the camera intrinsic parameters). The vector $\boldsymbol{\phi}^*$ contains the desired values of the features. In this work, the control scheme used is Image-Based Visual Servoing (IBVS) [41], which minimizes the error between the current and desired feature positions in the image plane, ensuring smooth and accurate tracking.



(a) (b)

Fig. 2: The phantom setup for LPN a) with and b) without covering fat.

The relationship between the feature velocity and the ECM velocity is given by

$$\dot{\boldsymbol{\phi}} = \boldsymbol{L}_{\boldsymbol{\phi}} \boldsymbol{v}, \tag{2}$$

where $\boldsymbol{v}$ is the twist of the ECM and $\boldsymbol{L}_{\boldsymbol{\phi}}$ is the interaction matrix (or image Jacobian).

For a generic image point with pixel coordinates (u, v), the corresponding interaction matrix in the image (pixel) domain is defined as

$$\mathbf{L}_{\phi} = \begin{bmatrix} -\frac{f}{z} & 0 & \frac{u}{z} & \frac{uv}{f} & -\frac{f^2+u^2}{f} & v \\ 0 & -\frac{f}{z} & \frac{v}{z} & \frac{f^2+v^2}{f} & -\frac{uv}{f} & -u \end{bmatrix}, \tag{3}$$

where f denotes the focal length in pixels and z is the depth of the point in the camera frame. This formulation is equivalent to the standard interaction matrix expressed in normalized image coordinates, scaled by the focal length to operate directly in pixel space. In the implementation, the depth was set to a constant value ($z = 1$), as commonly done in IBVS when reliable depth measurements are unavailable. This allows stable image-space control without explicit 3D reconstruction, and remains effective under limited depth variations, although the motion scale is only approximate.

The control law for IBVS is then defined as

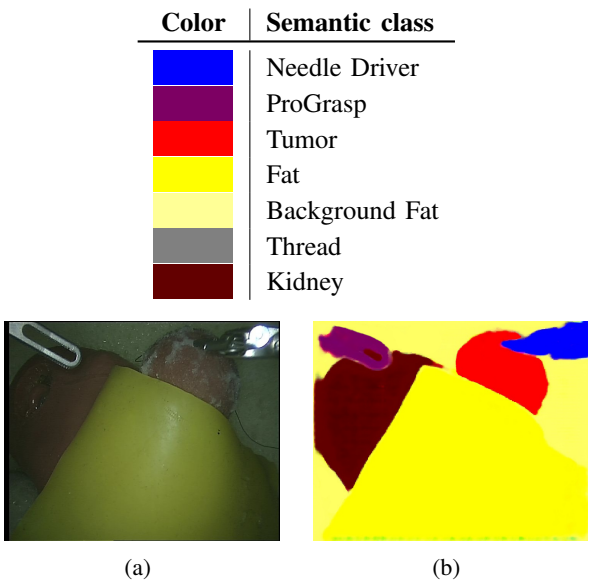
$$\boldsymbol{v} = -\lambda \boldsymbol{L}_{\boldsymbol{\phi}}^{+} (\boldsymbol{\phi} - \boldsymbol{\phi}^*) \tag{4}$$

where $\lambda > 0$ is a positive gain affecting convergence speed, and $\boldsymbol{L}_{\boldsymbol{\phi}}^{+}$ is the pseudo-inverse of the interaction matrix. Unsafe or non-smooth movements are prevented from inner dVRK constraints (e.g., RCM), low-pass filtering on image features and bounded Cartesian and angular velocities (see Table I for further details).

Since IBVS directly relies on image-space features rather than an explicit 3D reconstruction, it is particularly effective in dynamic and constrained environments like minimally invasive surgery, where accurate and stable tracking is critical.

IV. Experiments

We validate our methodology on a simplified version of the tumor removal phase of the lateral partial nephrectomy (LPN), a challenging benchmark for robot-assisted surgery



(a) (b)

Fig. 3: Top: semantic segmentation color encoding. Bottom: example of segmentation.

[17]. The procedure is executed on the custom-built phantom in Figures 2a-2b, in teleoperation by an expert (non-clinical) user of the dVRK. The GAN architecture used for semantic segmentation is the one from [40] and trained with 100 images[4], annotated by four independent non-clinical users of the dVRK platform, containing the manual annotations of the anatomical entities. The semantic classes are reported in Figure 3, together with an example of semantic segmentation for our task. GAN training was conducted on a workstation equipped with an NVIDIA GPU (2 NVIDIA GeForce RTX 2080 Ti with 11264MiB available each), using PyTorch and CUDA acceleration.

We design the experiments to answer the following research questions (RQs):

RQ1 What is the accuracy of tracking the target anatomy with our IBVS control scheme?

RQ2 Does IBVS improve the accuracy of semantic segmentation, useful for enhanced situation awareness and a potential extension to autonomous procedure execution?

RQ3 How does the discretization of the control action computed via IBVS affect the performance of our pipeline?

RQ4 Can semantic segmentation efficiently and accurately detect action transitions in the FSM of the task?

RQ3 addresses the insights by recent literature [25], highlighting the potentially bad impact of continuously tracking specific context targets (especially moving ones), in terms of motion sickness for the surgeon and consequent degradation of the overall task performance.

We compare the following approaches:

SemTrack The methodology described in Section III;

Discr. The methodology described in Section III, but with discretized IBVS control. Specifically, if the module of the linear (respectively, angular) control action in Equation (4) is less than 1 $\mathrm{cm/s}$ (respectively, 1 $\mathrm{rad/s}$), then zero velocity is commanded;

Static The ECM is not actuated;

Teleop The ECM is teleoperated by the surgeon via the console, activated through the use of the camera control pedal.

For each approach, the dVRK operator performs the task 5 times, in order to make a statistical analysis of the results.

The main parameters for IBVS are reported in Table I.

[4]Reducing the dataset size up to 50 images did not reduce the segmentation performance significantly.

TABLE I: Main IBVS parameters.

Parameter	Symbol / Value	Description
Control gain	$\lambda = 0.3$–0.6	IBVS proportional gain (empirically tuned)
Cartesian velocity limit	$v_{\max} = 5\ \mathrm{mm/s}$	Translational velocity saturation
Angular velocity limit	$\omega_{\max} = 2\ \mathrm{deg/s}$	Rotational velocity saturation
Depth normalization	$z = 1$	Constant-depth assumption for IBVS Jacobian
Low-pass filter cutoff	2 Hz	Reduces oscillations under noise/occlusion
Control frequency	50 Hz	IBVS update and dVRK servo rate

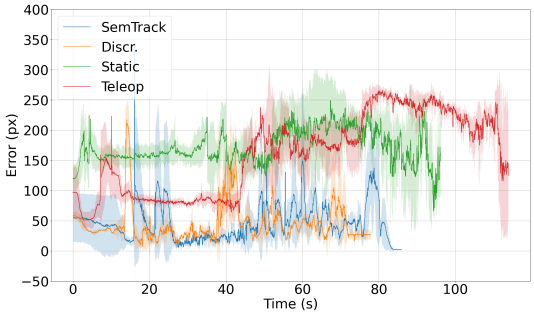


Fig. 4: Tracking error of the target anatomy (mean ± std dev.).

A. Task knowledge

The input to AUTOMATE is a textual description of the LPN provided by San Raffaele hospital in Milan (Italy) [33] respecting the language constraints reported in [16]. The output FSM with states, transitions and targets is reported in Table II. Given the limited features of our phantom and dVRK setup (e.g., absence of the renal vein and only ProGrasp and Needle Driver instruments available), we only consider the red states in the table as the final FSM for our LPN.

B. RQ1: tracking accuracy

We measure the tracking accuracy as the pixel error (hence, in the image coordinates) between the centroid of the image (where the ECM is actually pointing) and the centroid of the target anatomy of the current action $\bar{t}_c$. Figure 4 shows the results for the approaches, averaging over five executions. Our methodology (blue curve) keeps the lowest tracking error along the full task[5], even with respect to the teleoperation

[5]The peak at $\approx$ 17 s corresponds to the end of the `defat` action, and is due to the abrupt removal of the fat tissue which triggers a high velocity command from IBVS.

State / action	Next transition
`defat(ProGrasp, fat)`[a]	the fat is not present
`clamp(Satinsky, renal_vein)`	bleeding stopped
`suction(aspirator, blood)`	no blood visible
`dissect(Needle Driver, tumor)`[b]	the tumor is removed
`suction(aspirator, blood)`	no blood visible
`apply(Needle Driver, suture needle)`[c]	sutured wound
`remove(Satinsky, clamp)`	N/A

[a] In the surgical text, the scissors are also required to cut the fat out. Given the limited hardware available, we assume that the fat is already cut and must only be removed.
[b] According to the surgical text, the scissors should be used. In our simplified setup, we use the tip of the Needle Driver to cut the tumor out.
[c] Given the very small size of the needle and its inaccurate semantic segmentation, for this action, we consider the agent Needle Driver (holding the needle) as the target to be tracked.

TABLE II: FSM states (actions) for tumor removal in LPN in the form $s(a,t)$ denoting the arm a and the target t. **Red** states are actually considered in our task.

of the ECM (red curve), which is inherently imprecise. In addition, the teleoperation of the ECM only reduces the tracking error in the beginning of the task (up to $\approx$ 45 s), where either the static fat or the tumor must be tracked; on the other hand, it does not significantly improve the tracking accuracy with respect to the static ECM (green curve) when the moving Needle Driver must be tracked.

Furthermore, **Teleop** requires the largest amount of time ($\approx$ 30 s more than our method) for task completion, representing an additional and potentially disturbing sub-task for the operator. This mainly depends on the last and longest suturing action (61 ± 8 s duration on average for **Teleop** vs. 29 ± 4 s for **SemTrack**), which requires time and effort from the operator to track the moving target, still achieving a high pixel error.

On aggregate, the approaches achieve the following tracking errors (95% confidence intervals[6] are reported):

SemTrack $46 - 101$ px;

Static $173 - 116$ px;

Teleop $161 - 80$ px.

C. RQ2: semantic segmentation accuracy

We measure the semantic accuracy, i.e., the quality of the recognition of action-specific targets, using the standard measures of precision, recall, intersection over union (IoU) and Dice index (reporting the 95% confidence interval, CI), in order to compare the blobs identified by the semantic segmentation with their ground truth. To clarify, the underlying GAN semantic segmentation architecture remains identical across all four test conditions. The performance variations reported in III isolate the impact of the camera control strategy on perception. The continuous and discretized autonomous control methods maintain the target anatomy in the optimal region of the camera's field of view, thereby generating higher-quality, stabilized viewpoints that inherently improve the semantic segmentation output compared to the less optimal viewpoints generated by manual teleoperation or a static camera. From Table III[7], our method (either continuous, **SemTrack**, or discretized, **Discr**) always attains the best or the second best performance on both IoU and Dice. Specifically, aggregating over all target anatomies, the **Discr** version outperforms **Teleop** by at least 10% on average, both in terms of IoU and Dice, showing the clear advantage of automatic camera motion. The continuous IBVS control (**SemTrack**) is the second-best method, though with more moderate improvement.

Looking at the single action targets, the impact of camera motion is crucial for the smallest anatomies (i.e., tumor and Needle Driver), as evident from the drop in performance of the **Static** approach. On the other hand, all methods perform quite well for the fat identification. While **Teleop** is the best-performing method for the fat and the Needle Driver, the second-best performing method (**Discr** and **SemTrack**, respectively) attains always comparable IoU and Dice. On the other hand, the advantage of **Discr** vs. **Teleop** for tumor identification is testified by 6% improvement for IoU and 10% improvement for Dice. In general, the worst performance attained by all methods for the tumor identification are probably due to its color similarity with the kidney (see Figure 2b).

D. RQ3: continuous vs. discretized control

To balance the potential motion sickness and undesired bad effects of continuously tracking moving targets in the surgical scene (e.g., the Needle Driver in the suturing step), we assess the tracking error and semantic accuracy[8] of the discretized version (**Discr**) of our methodology vs. the continuous IBVS (**SemTrack**).

[6]Here and throughout all the paper, the confidence interval is based on the t-distribution, as of the standard `scipy` library.

[7]The number of samples to compute the metrics differs among different executions and baselines. On average, the first state `defat` (where the target is the fat) lasts $\approx$ 600 samples, while the two other states `dissect` (tumor) and `apply` (needle driver) last $\approx$ 800 samples.

[8]We do not evaluate the accuracy in the detection of FSM transitions, since it does not depend on the tracking control action computed by IBVS.

Target	Metric	SemTrack	Discr.	Static	Teleop
fat	IoU	$0.822 - 0.008$	$\underline{0.828 - 0.006}$	$0.762 - 0.027$	$\mathbf{0.886 - 0.013}$
	Dice	$0.894 - 0.006$	$\underline{0.900 - 0.005}$	$0.839 - 0.026$	$\mathbf{0.928 - 0.013}$
tumor	IoU	$0.079 - 0.012$	$\mathbf{0.465 - 0.012}$	$0.206 - 0.006$	$\underline{0.404 - 0.010}$
	Dice	$0.365 - 0.013$	$\mathbf{0.596 - 0.014}$	$0.293 - 0.009$	$\underline{0.492 - 0.009}$
Needle Driver	IoU	$\underline{0.648 - 0.006}$	$0.619 - 0.003$	$0.516 - 0.006$	$\mathbf{0.658 - 0.005}$
	Dice	$\underline{0.777 - 0.005}$	$0.751 - 0.003$	$0.693 - 0.005$	$\mathbf{0.780 - 0.003}$
Aggregate	IoU	$\underline{0.562 - 0.009}$	$\mathbf{0.633 - 0.004}$	$0.285 - 0.006$	$0.533 - 0.007$
	Dice	$\underline{0.643 - 0.008}$	$\mathbf{0.755 - 0.004}$	$0.424 - 0.007$	$0.632 - 0.006$

TABLE III: Semantic accuracy (mean 95% CI) measured using IoU and Dice. **Best** and second-best results. **SemTrack**: proposed method; **Discr.**: proposed method with discretized IBVS; **Static**: fixed ECM; **Teleop**: surgeon-controlled ECM.

Figure 4 shows that the tracking error of **Discr** is still significantly lower than **Static** and **Teleop** approaches. Indeed, it achieves an average error of 46–52 px (95% CI), i.e., very close to the continuous control (**SemTrack**), but with lower variance due to the fewer oscillations. More interestingly, Table III shows that the discretized version of our method attains the best IoU and Dice scores for the semantic segmentation, aggregating over all anatomies, outperforming **Teleop** and **SemTrack** by $\approx 10\%$ and with the lowest 95% CI. The advantage of **SemTrack** is most probably due to the more stable camera motion, which is particularly useful to accurately identify the most uncertain tumor anatomy (similar to the kidney and subject to instrument occlusion during cutting). Moreover, Figure 4 shows that **Discr** requires less time for task completion (≈ 8 s less than **SemTrack**), possibly evidencing the improved usability by the dVRK operator.

E. RQ4: FSM transition detection

To measure the accuracy in the detection of the FSM transitions, in Table IV we report the error (in seconds) between the actual time of step transition (manually annotated) and the time of automatic detection of the same transition (**Next transition** column in Table II). The error is at most 0.23 s on average, which is compatible with a real surgical application, thanks to the fast inference time of our GAN architecture (≈ 15 Hz end-to-end control rate). Interestingly, the continuous control method (**SemTrack**) gets a higher time error than the discretized version (**Discr**). This is probably due to the continuous tracking of the anatomies, which leads to abrupt camera motion when the target exits the scene (e.g., *the fat is not present* for the action `defat`); this introduces inaccuracies in the semantic segmentation, hence the detection of FSM transitions as well.

Transition	SemTrack	Discr.
defat → dissect	0.23 ± 0.07	0.15 ± 0.10
dissect → apply	0.14 ± 0.13	0.09 ± 0.03
Average	0.19 ± 0.11	0.12 ± 0.09

TABLE IV: Temporal error [s] in FSM transition detection (mean $\pm$ std deviation). **Best** and second best results.

V. Discussion and Limitations

Our experiments have evidenced the feasibility and advantage of autonomous context tracking in surgical tasks via semantic IBVS on the ECM, even versus the teleoperation of the ECM. Specifically, our methodology allows to accurately track the targets of specific actions of the procedure, both anatomies and instruments (**RQ1**). This results in the more accurate segmentation of the targets (**RQ2**), even when they are moving (e.g., the Needle Driver in the suturing action) or hard to identify due to instrument occlusions and similarity with other scene elements (e.g., the tumor to be dissected). The improved accuracy can support the autonomous execution of the surgical task, and possibly reduces the effort for task execution by the dVRK operator, especially lowering the time for completing the steps with moving actions (e.g., suturing with the Needle Driver). Also, our IBVS control strategy can be discretized to avoid motion sickness for the operator, still achieving very good performance (**RQ3**). Finally, the semantic segmentation allows the detection of step transitions with sufficient accuracy for a real surgical task (**RQ4**).

We believe that our work serves as an initial step towards effective and scalable autonomous context tracking in robotic surgery. Nevertheless, we highlight the following limitations and rooms for improvement and future studies:

- **Constant-depth assumption in IBVS.** The IBVS controller assumes constant depth ($z = 1$), yielding pixel-space errors that do not directly reflect metric accuracy. This may limit transferability to settings with significant depth variations, where depth-aware IBVS would be required. However, pixel errors remain consistent for relative comparison under identical conditions.
- **More sophisticated knowledge extraction strategies**: we are currently relying on the existing AUTOMATE pipeline [16] to extract the FSM task description from text. While we preliminary assessed the robustness of this methodology against syntactic variants of the surgical text (see supplementary material), the use of more specific surgical lexicons [42], more powerful language models and approaches [43], or the combination of visual and language models [44] will be investigated in future works. This will potentially allow to scale to more complex tasks and the challenges of in-vivo perception.

- **Comparison to learning-based solutions**: while our methodology requires significantly less training time and ensures higher interpretability and trustworthiness than previous deep learning techniques [10], [35], a comparison with such models would be helpful to effectively compare the benefits and drawbacks of both approaches. We remark that this empirical analysis has not been possible, given the lack of publicly available implementations.
- **User study**: the findings of our work would benefit from an extended user study. For instance, this would allow to better investigate the advantages of discretized IBVS control in terms of motion sickness and surgical user experience. In this regard, we remark however that this has already been claimed by recent research [25]. Moreover, standard objective [45] and subjective [46] user studies performed on dVRK operators with different expertise could further highlight the potential of our methodology for real surgery.
- **Translation to in-vivo environments**: The current experimental validation is constrained to a phantom setup. While this provides a necessary, controlled baseline to evaluate the coupled perception-control architecture, it significantly simplifies the visual and physical conditions of real-world surgery. In an actual clinical scenario, the perception module will have to contend with dynamic lighting, specular reflections on wet tissues, visual occlusions from blood and smoke, and non-rigid tissue deformations. These factors will likely reduce the accuracy of the semantic segmentation and, consequently, the stability of the tracking. Furthermore, the rigid nature of the extracted FSM might struggle with unpredictable intraoperative events (e.g., sudden bleeding requiring immediate, off-script intervention). Transitioning this framework to the real world will require training the GAN on robust, in-vivo datasets, replacing the constant-depth IBVS assumption with active depth estimation (e.g., via stereo vision), and implementing dynamic error-recovery states within the FSM to handle deviations from the standard surgical flow.

VI. Conclusion

In this paper we proposed *SemTrack*, a novel methodology for autonomous context tracking in robotic surgery, combining interpretable automatic task knowledge extraction from surgical notes and text via natural language processing, and semantic segmentation for task flow monitoring and image-based visual servoing for controlling the ECM. In the context of a paradigmatic phantom-based tumor removal phase of lateral partial nephrectomy, our methodology outperforms the teleoperated and static ECM in terms of tracking accuracy, recognition accuracy, and overall task performance (durations of the full procedure and steps with moving targets).

In the future, we will extend this first step towards autonomy-enhanced surgical situation awareness, by testing our methodology on more complex and diverse surgical tasks, comparing to other public implementations when available, and conducting a rigorous user study.

Acknowledgments

The authors would like to thank the surgical team at San Raffaele Hospital, Milan, for providing procedural data and clinical insights used in this work.

This work was supported by the PNRR and EU Funding acknowledgment: Finanziato a valere sulle risorse del Piano Nazionale Ripresa E Resilienza (PNRR) Missione 4, "Istruzione e Ricerca" - Componente 2, "dalla ricerca all'impresa" - linea di investimento 1.3, finanziato dall'Unione Europea – Nextgenerationeu", progetto "Future Artificial Intelligence – Fair" Spoke 2 2 - PE0000013, CUP C63C22000770006.

We would like to thank Intuitive Foundation for their support with the da Vinci Research Kit (dVRK). Further details are available at https://www.intuitive-foundation.org/dvrk/

References

[1] T. Haidegger, "Autonomy for surgical robots: Concepts and paradigms," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 2, pp. 65–76, 2019.

[2] C. D'Ettorre, A. Mariani, A. Stilli, F. R. y Baena, P. Valdastri, A. Deguet, P. Kazanzides, R. H. Taylor, G. S. Fischer, S. P. DiMaio, *et al.*, "Accelerating surgical robotics research: A review of 10 years with the da vinci research kit," *IEEE Robotics & Automation Magazine*, vol. 28, no. 4, pp. 56–78, 2021.

[3] Y. Sun and T. C. Lueth, "Safe manipulation in robotic surgery using compliant constant-force mechanism," *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 3, pp. 486–495, 2023.

[4] A. Pore, D. Corsi, E. Marchesini, D. Dall'Alba, A. Casals, A. Farinelli, and P. Fiorini, "Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4025–4031, IEEE, 2021.

[5] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," 2017.

[6] E. Tagliabue, D. Meli, D. Dall'Alba, and P. Fiorini, "Deliberation in autonomous robotic surgery: a framework for handling anatomical uncertainty," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11080–11086, IEEE, 2022.

[7] D. Meli, E. Tagliabue, D. Dall'Alba, and P. Fiorini, "Autonomous tissue retraction with a biomechanically informed logic based framework," in *2021 International Symposium on Medical Robotics (ISMR)*, pp. 1–7, IEEE, 2021.

[8] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini, "Autonomous task planning and situation awareness in robotic surgery," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3144–3150, IEEE, 2020.

[9] D. Meli, H. Nakawala, and P. Fiorini, "Logic programming for deliberative robotic task planning," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 9011–9049, 2023.

[10] Z. Chen, K. Fan, L. Cruciani, M. Fontana, L. Muraglia, F. Ceci, L. Travaini, G. Ferrigno, and E. De Momi, "Toward human-out-of-the loop endoscope navigation based on context awareness for enhanced autonomy in robotic surgery," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[11] Y.-J. Yang, A. N. Vadivelu, C. H. Pilgrim, D. Kulić, and E. Abdi, "A novel perception framework for automatic laparoscope zoom factor control using tool geometry," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4700–4704, IEEE, 2021.

[12] J. Peng, C. Zhang, L. Kang, and G. Feng, "Endoscope fov autonomous tracking method for robot-assisted surgery considering pose control, hand–eye coordination, and image definition," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–16, 2022.

[13] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial intelligence review*, vol. 54, pp. 137–178, 2021.

[14] Y. Huang, J. Li, X. Zhang, K. Xie, J. Li, Y. Liu, C. S. H. Ng, P. W. Y. Chiu, and Z. Li, "A surgeon preference-guided autonomous instrument tracking method with a robotic flexible endoscope based on dvrk platform," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2250–2257, 2022.

[15] M. Bombieri, M. Rospocher, S. P. Ponzetto, and P. Fiorini, "Surgicberta: a pre-trained language model for procedural surgical language," *International Journal of Data Science and Analytics*, vol. 18, no. 1, pp. 69–81, 2024.

[16] M. Bombieri, D. Meli, D. Dall'Alba, M. Rospocher, and P. Fiorini, "Mapping natural language procedures descriptions to linear temporal logic templates: an application in the surgical robotic domain," *Applied Intelligence*, vol. 53, no. 22, pp. 26351–26363, 2023.

[17] G. E. Cacciamani, L. G. Medina, T. Gill, A. Abreu, R. Sotelo, W. Artibani, and I. S. Gill, "Impact of surgical factors on robotic partial nephrectomy outcomes: comprehensive systematic review and meta-analysis," *The Journal of urology*, vol. 200, no. 2, pp. 258–274, 2018.

[18] R. Moccia and F. Ficuciello, "Autonomous endoscope control algorithm with visibility and joint limits avoidance constraints for da vinci research kit robot," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 776–781, IEEE, 2023.

[19] H. Kossowsky and I. Nisky, "Predicting the timing of camera movements from the kinematics of instruments in robotic-assisted surgery using artificial neural networks," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 2, pp. 391–402, 2022.

[20] C. Song, X. Ma, X. Xia, P. W. Y. Chiu, C. C. N. Chong, and Z. Li, "A robotic flexible endoscope with shared autonomy: a study of mockup cholecystectomy," *Surgical endoscopy*, vol. 34, pp. 2730–2741, 2020.

[21] A. Mariani, G. Colaci, T. Da Col, N. Sanna, E. Vendrame, A. Menciassi, and E. De Momi, "An experimental comparison towards autonomous camera navigation to optimize training in robot assisted surgery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1461–1467, 2020.

[22] K. Zinchenko and K.-T. Song, "Autonomous endoscope robot positioning using instrument segmentation with virtual reality visualization," *IEEE Access*, vol. 9, pp. 72614–72623, 2021.

[23] X. Ma, C. Song, L. Qian, W. Liu, P. W. Chiu, and Z. Li, "Augmented reality-assisted autonomous view adjustment of a 6-dof robotic stereo flexible endoscope," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 2, pp. 356–367, 2022.

[24] T. Cheng, W. Li, W. Y. Ng, Y. Huang, J. Li, C. S. H. Ng, P. W. Y. Chiu, and Z. Li, "Deep learning assisted robotic magnetic anchored and guided endoscope for real-time instrument tracking," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3979–3986, 2021.

[25] N. Pasini, A. Mariani, A. Deguet, P. Kazanzides, and E. De Momi, "Grace: Online gesture recognition for autonomous camera-motion enhancement in robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8263–8270, 2023.

[26] K. C. Demir, H. Schieber, T. WeiseRoth, M. May, A. Maier, and S. H. Yang, "Deep learning in surgical workflow analysis: a review of phase and step recognition," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[27] A. Roberti, D. Meli, G. De Rossi, R. Muradore, and P. Fiorini, "Semantic monocular surgical slam: Intra-operative 3d reconstruction and pre-operative registration in dynamic environments," in *2023 21st International Conference on Advanced Robotics (ICAR)*, pp. 486–491, IEEE, 2023.

[28] F. Falezza, N. Piccinelli, G. De Rossi, A. Roberti, G. Kronreif, F. Setti, P. Fiorini, and R. Muradore, "Modeling of surgical procedures using statecharts for semi-autonomous robotic surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 4, pp. 888–899, 2021.

[29] H. Gao, W. Fan, L. Qiu, X. Yang, Z. Li, X. Zuo, Y. Li, M. Q.-H. Meng, and H. Ren, "Savanet: Surgical action-driven visual attention network for autonomous endoscope control," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 4, pp. 2655–2667, 2022.

[30] M. Li, B. Wang, J. Yang, J. Cao, C. You, Y. Sun, J. Wang, and D. Wu, "Multistage adaptive control strategy based on image contour data for autonomous endoscope navigation," *Computers in Biology and Medicine*, vol. 149, p. 105946, 2022.

[31] M. Huber, J. B. Mitchell, R. Henry, S. Ourselin, T. Vercauteren, and C. Bergeles, "Homography-based visual servoing with remote center of motion for semi-autonomous robotic endoscope manipulation," in *2021 International Symposium on Medical Robotics (ISMR)*, pp. 1–7, IEEE, 2021.

[32] M. Huber, S. Ourselin, C. Bergeles, and T. Vercauteren, "Deep homography prediction for endoscopic camera motion imitation learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 217–226, Springer, 2023.

[33] E. Oleari, A. Leporini, D. Trojaniello, A. Sanna, U. Capitanio, F. Dehó, A. Larcher, F. Montorsi, A. Salonia, F. Setti, and R. Muradore, "Enhancing surgical process modeling for artificial intelligence development in robotics: the saras case study for minimally invasive procedures," in *International Symposium on Medical Information and Communication Technology (ISMICT)*, 2019.

[34] R. Younis, A. Yamlahi, S. Bodenstedt, P. Scheikl, A. Kisilenko, M. Daum, A. Schulze, P. Wise, F. Nickel, F. Mathis-Ullrich, *et al.*, "A surgical activity model of laparoscopic cholecystectomy for co-operation with collaborative robots," *Surgical Endoscopy*, pp. 1–13, 2024.

[35] M. Wagner, A. Bihlmaier, H. G. Kenngott, P. Mietkowski, P. M. Scheikl, S. Bodenstedt, A. Schiepe-Tiska, J. Vetter, F. Nickel, S. Speidel, *et al.*, "A learning robot for cognitive camera control in minimally invasive surgery," *Surgical Endoscopy*, vol. 35, no. 9, pp. 5365–5374, 2021.

[36] C. Zhang, M. S. Hallbeck, H. Salehinejad, and C. Thiels, "The integration of artificial intelligence in robotic surgery: A narrative review," *Surgery*, vol. 176, no. 3, pp. 552–557, 2024.

[37] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 69–72, 2006.

[38] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," *CoRR*, vol. abs/1904.05255, 2019.

[39] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[41] K. Hashimoto, T. Kimoto, T. Ebine, and H. Kimura, "Manipulator control with image-based visual servo," in *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, pp. 2267–2268, IEEE Computer Society, 1991.

[42] M. Bombieri, M. Rospocher, S. P. Ponzetto, and P. Fiorini, "The robotic-surgery propositional bank," *Lang. Resour. Evaluation*, vol. 58, no. 3, pp. 1043–1071, 2024.

[43] A. Moglia, K. Georgiou, P. Cerveri, L. Mainardi, R. M. Satava, and A. Cuschieri, "Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test," *Artificial Intelligence Review*, vol. 57, no. 9, p. 231, 2024.

[44] S. Zargarzadeh, M. Mirzaei, Y. Ou, and M. Tavakoli, "From decision to action in surgical autonomy: Multi-modal large language models for robot-assisted blood suction," *IEEE Robotics and Automation Letters*, 2025.

[45] S. Walldén, E. Mäkinen, and R. Raisamo, "A review on objective measurement of usage in technology acceptance studies," *Universal Access in the Information Society*, vol. 15, no. 4, pp. 713–726, 2016.

[46] Y. G. Kim, J. W. Shim, G. Gimm, S. Kang, W. Rhee, J. H. Lee, B. S. Kim, D. Yoon, M. Kim, M. Cho, *et al.*, "Speech-mediated manipulation of da vinci surgical system for continuous surgical flow," *Biomedical Engineering Letters*, vol. 15, no. 1, pp. 117–129, 2025.