Toward Human-Out-of-the-Loop Endoscope Navigation Based on Context Awareness for Enhanced Autonomy in Robotic Surgery

Ziyang Chen, *Graduate Student Member, IEEE*, Ke Fan, Laura Cruciani, Matteo Fontana, Lorenzo Muraglia, Francesco Ceci, Laura Travaini, Giancarlo Ferrigno, *Senior Member, IEEE*, and Elena De Momi, *Senior Member, IEEE*

***Abstract*—Although the da Vinci surgical system enhances manipulation dexterity and restores 3D vision in robotic surgery, it requires surgeons to asynchronously control surgical instruments and the endoscope, which hinders a smooth operation. Surgeons frequently position the endoscope to maintain a good field of view during operation, potentially increasing surgical time and workload. In this paper, a Human-Out-Of-The-Loop (HOOTL) endoscope navigation control with the assistance of context awareness is proposed to enhance surgical autonomy. A comprehensive comparison study using 8 state-of-the-art networks was conducted to find out the best model for surgical phase recognition. Ten human subjects were invited to participate in a classic ring transferring task based on three different endoscope navigation pipelines on a da Vinci research kit platform, including standard endoscope navigation, semi-autonomous endoscope navigation with manual pedal control, and HOOTL endoscope navigation supported by vision-based phase recognition. The experimental results showed that the proposed endoscope navigation approach releases the operation need of controlling the pedals, and it significantly reduces the execution time compared to the other two navigation pipelines. The result of the NASA Task Load Index (NASA-TLX) questionnaire indicates that the proposed endoscope navigation can reduce the physical and mental load for the users.**

***Index Terms*—Robotic surgery, endoscope navigation control, context awareness, surgical autonomy, vision-based phase recognition.**

I. Introduction

ROBOT-ASSISTED Minimally Invasive Surgery (RAMIS) has been widely adopted in medical practice because it shows the potential to reduce intra-operative bleeding and tissue trauma to patients, and shorten the postoperative hospital stay compared to traditional open or laparoscopic surgery. Also, it provides surgeons a 3D view of the surgical field and increases dexterity with the instruments [1], [2]. The da Vinci Surgical System (dVSS, Intuitive Surgical, Sunnyvale, CA, USA) is a representative among surgical robots thanks to its commercialization success [3], [4]. It has been utilized in various minimally invasive surgeries in hospitals today, such as cholecystectomy [5], prostatectomy [6] and nephrectomy [7]. DVSS overcomes a major limitation of laparoscopic surgery, which is that the surgeon manipulates the surgical instruments while an assistant manipulates the endoscope, requiring a high degree of cooperation between them. In comparison, dVSS allows the surgeon to control the instruments or endoscope independently without requiring the involvement of an assistant. Nevertheless, the da Vinci robot does not support the surgeon in simultaneously operating the surgical instruments and the endoscope. To position the endoscope, the surgeon needs to interrupt the manipulation of the surgical instruments, and move to operate the foot switches on the pedal tray and the manipulators on the console, which may affect the smoothness of the operation and prolong the operating time [8].

Autonomy in medical robots is a promising but challenging direction that has attracted much attention in many research laboratories [9]. Furthering autonomy in medical robots has the potential to increase the accuracy of operations and reduce the workload of surgeons [10]. However, according to the definition of autonomy in [11], dVSS does not yet have autonomy as it is under the full control of surgeons during surgery. To increase the autonomy of dVSS, one of the popular topics is autonomous endoscopic navigation. It offers the possibility of freeing the control of the endoscope so that surgeons can concentrate on controlling the surgical instruments, which may release the fatigue of surgeons and improve surgical performance. Some works have been done to explore this field, for instance, the authors in [8] proposed a camera autonomous navigation approach based on the da Vinci robot. The camera can track the surgical tool tips by utilizing

Manuscript received 23 March 2024; revised 25 May 2024; accepted 23 June 2024. Date of publication 3 July 2024; date of current version 5 August 2024. This article was recommended for publication by Associate Editor H. D. Taghirad and Editor P. Dario upon evaluation of the reviewers' comments. This work was supported by the China Scholarship Council. *(Corresponding author: Ke Fan.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Politecnico di Milano Ethics Committee under Applicant No. 30/2023.

Ziyang Chen, Ke Fan, Laura Cruciani, and Giancarlo Ferrigno are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy (e-mail: ke.fan@polimi.it).

Matteo Fontana is with the Department of Urology, European Institute of Oncology, IRCCS, 20141 Milan, Italy.

Lorenzo Muraglia, Francesco Ceci, and Laura Travaini are with the Department of Nuclear Medicine, European Institute of Oncology, IRCCS, 20141 Milan, Italy.

Elena De Momi is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy, and also with the Department of Urology, European Institute of Oncology, IRCCS, 20141 Milan, Italy.

Digital Object Identifier 10.1109/TMRB.2024.3422618

the kinematic data, and users can determine the tracking modes (including tracking the single tool tip or the middle point between the two tools) by depressing the foot switch. The experimental results based on a ring transferring task showed that the designed navigation system promoted better operation performance for the users than the standard setup. As an extension, the authors of [12], [13] adopted the autonomous camera navigation approach to perform an ex vivo neobladder reconstruction in a dry lab. Ten urologists were invited to conduct this operation, and the results showed that the camera navigation method can boost the system usability and reduce the operation time compared to the standard camera control. Similarly, the authors in [14] implemented an autonomous endoscope navigation approach by tracking the middle points of the instruments in a virtual reality simulator for surgical skill training. Referencing a time-accuracy metric and a camera-related metric, they found that the novices can obtain better skill improvement with the assistance of autonomous endoscope navigation than the novices who achieved training in manual endoscope control.

The above human-in-the-loop endoscope navigation strategies require users to manually switch different camera tracking modalities. To advance the autonomy in the endoscope motion, the authors in [15] proposed an online gesture recognition based navigation approach. They exploited the kinematic data containing 17 dimensional features to introduce the situation awareness for endoscope motion mode switching without the involvement of humans. However, the prediction accuracy of 0.84 reported by the authors remains to be improved. Recognizing the surgical situation based on the kinematic data may not be reliable. Surgeons may have different manipulation gestures even when performing the same operation, resulting in variable kinematic data such as the pose and velocity of end-effectors. Furthermore, background information is also a critical resource to be exploited in context awareness except for the information of surgical instruments [16]. It can be noticed that surgeons perceive the surgical context by relying on the direct input of surgical video streams instead of kinematic data in RAMIS.

Vision-based surgical context recognition gradually becomes a mainstream direction, because the rapid development of deep learning technology promotes promising recognition performance in this field [17], [18], [19]. EndoNet is a well-known neural network proposed in [20] to recognize the surgical phases and tool labels using videos of cholecystectomy surgeries. It relies on the AlexNet architecture [21] with a customized feature processing module to perform a multi-task prediction. After that, the authors in [22] also proposed a multi-task recurrent convolutional network to predict phase labels and tool labels simultaneously. They designed an end-to-end network integrating residual units [23] and Long Short-Term Memory (LSTM) [24] modules to process features of video clips, and obtained high accuracy in both tool and phase recognition. Similarly, the authors in [25] proposed a unified framework integrating deep learning and knowledge representation to predict the surgical phases, steps and tool labels using 9 videos of robot-assisted partial nephrectomy, but the non-end-to-end predictive property hinders its online deployment in practice. Next, the authors in [26] built a memory bank module to model global features of long-range video clips, and another branch containing ResNet and LSTM was built to extract high-level feature representation of short-range video clips, then different-scale features were processed by an attention module for final phase classification. The authors in [27] also introduced the memory bank module and transformer [28] to perform an accurate phase recognition. Global feature modeling may heighten prediction performance, while it hinders real-time phase recognition.

Online context recognition provides the possibility to make surgical decisions automatically. In [29], the authors integrated context awareness into imitation learning to implement a human-robot shared control. The surgical instruments can perform adaptive movement either under the control of users or following the trajectory generated by imitation learning, which is determined by three recognized phases based on video streams. The experiments were conducted in a ring transferring task, and the results showed that the assistance of context awareness can perform autonomous surgical decisions and promote the manipulation performance of users.

A context awareness based endoscope navigation approach is proposed in this paper to implement a Human-Out-Of-The-Loop (HOOTL) endoscope control. It can be defined as task autonomy according to the definition of autonomy in [11], since the endoscope control modes and the surgical context are predefined based on specific tasks. This differs from the work of [15], where kinematic data was used for online gesture recognition, we adopted vision (i.e., images) to recognize the surgical context, which aims to perform online phase recognition rapidly and with high accuracy. Vision-based context recognition has high generalization since it extracts the features of both surgical instruments and background scenes compared to kinematics-based methods. To the best of our knowledge, this is the first work to utilize vision-based phase recognition to further the autonomy of endoscope navigation. We introduced a classic ring transferring task for the comparison of (1) the standard manual endoscope navigation (2) the semi-autonomous endoscope navigation [8] (3) the proposed HOOTL endoscope navigation, to explore the influence of different endoscopic autonomy by analyzing the data captured from ten human subjects.

In summary, this paper has the following contributions:

(a) It is the first time that vision-based surgical phase recognition is integrated into endoscope navigation to implement a HOOTL endoscope control.

(b) A comprehensive comparison study based on 8 state-of-the-art phase recognition networks involving images and kinematics was performed.

(c) A user study based on ten participants was achieved to explore the feasibility of the proposed HOOTL endoscope navigation.

II. Methods

Fig. 1 presents the three endoscope navigation pipelines, including the standard endoscope navigation template using

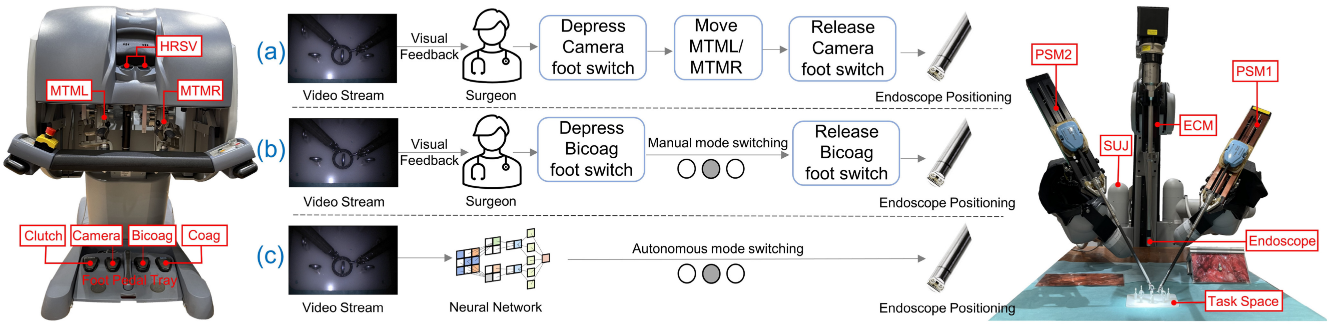

Fig. 1. A demonstration of the dVRK components and the three different endoscope navigation pipelines. (a) shows the standard pipeline for manipulating the endoscope, which requires a combination of the foot switch and manipulators; (b) presents the semi-autonomous endoscope navigation pipeline [8], and the users manually select the endoscope tracking modes by controlling a foot switch; (C) shows the autonomous endoscope navigation strategy to perform a HOOTL control based on context awareness.

both a foot switch and two manipulators, the semi-autonomous navigation by using a foot switch, and the HOOTL autonomous endoscope navigation with the assistance of context awareness. It also shows the main components of a da Vinci Research Kit (dVRK) in Leonardo Robotics Lab, Politecnico di Milano, Italy. The specific system description is as follows.

A. System Setup

DVRK (Intuitive Surgical Inc., US, and Johns Hopkins University) is a well-known surgical robot research platform derived from the first generation of dVSS, integrating customized software and electronics. It contains the master side and the patient side: the surgeons can control the Master-Tool Manipulators (MTMs) to remotely operate surgical instruments, and use the High-Resolution Stereo Viewer (HRSV) to observe surgical scenes captured by a stereo endoscope. The surgical instruments are mounted on the Patient Side Manipulators (PSMs), and the endoscope is mounted on the Endoscopic Camera Manipulator (ECM). The mechanical structure allows these arms to meet Remote Center of Motion (RCM) constraints during operation, to avoid conflicts with the skin entry point of the abdomen, and these RCM points can be repositioned by the passive Set Up Joints (SUJs) [30]. There is a foot pedal tray involving four pedals at the master sole: the Clutch pedal is used to extend the operating space of surgical instruments, the Camera pedal is used to adjust the viewpoint of the endoscope, while the Bicoag and Coag pedals are not activated in the setup.

Considering that the endoscope is driven by the ECM, we provide a description about the kinematics of the ECM arm. The ECM is a 4 Degrees of Freedom (DoFs) actuated arm that can move the endoscope around the RCM point following a RRPR sequence, where R is the revolute joint and P is the prismatic joint. Fig. 2 shows the kinematics of the ECM. The ECM end-effector pose can be described by a homogeneous transformation matrix T_{EE}^{RCM} between the end-effector frame F_{EE} and the RCM frame F_{RCM}, and it can be calculated by applying the standard Denavit-Hartenberg (DH) convention to the kinematic chain. Table I gives the DH parameters of the ECM arm [31], [32].

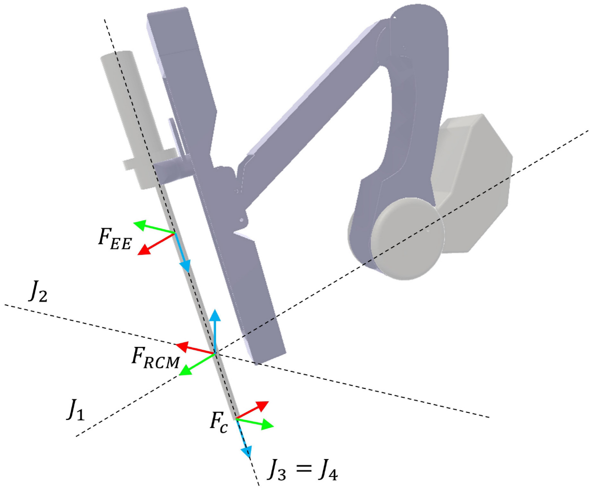

Fig. 2. ECM kinematic description. $J_1, \ldots, J_4$ are the actuation axes of the ECM arm. F_{EE}: ECM end-effector frame, F_{RCM}: RCM frame, F_C: camera frame.

TABLE I
DH PARAMETERS OF THE ECM ARM. $q_1, \ldots, q_4$ ARE THE JOINT VALUES

link	joint	a_i	α_i	d_i	θ_i	offsets
1	R	0	$\pi/2$	–	q_1	$\pi/2$
2	R	0	$-\pi/2$	–	q_2	$-\pi/2$
3	P	0	$\pi/2$	q_3	–	-0.3822m
4	R	0	0	–	q_4	0.3829m

B. Task Description

The ring transferring task belongs to one of the Fundamentals of Laparoscopic Surgery (FLS) tasks for surgical skill assessment [33], and it is a common scenario to measure the manipulation level of users on the dVRK platform [8], [14], [29], [34]. We adopted this task to evaluate the proposed endoscope navigation strategy, as demonstrated in Fig. 3. The task can be divided into three phases:

- LH: Operate the left instrument to pick up the black ring on Peg A, transfer it to Peg B, place it down and then pick it up again.
- BM: Operate both left and right instruments to swap the ring from the left instrument to the right instrument on the top of Peg C, then transfer it to Peg D.

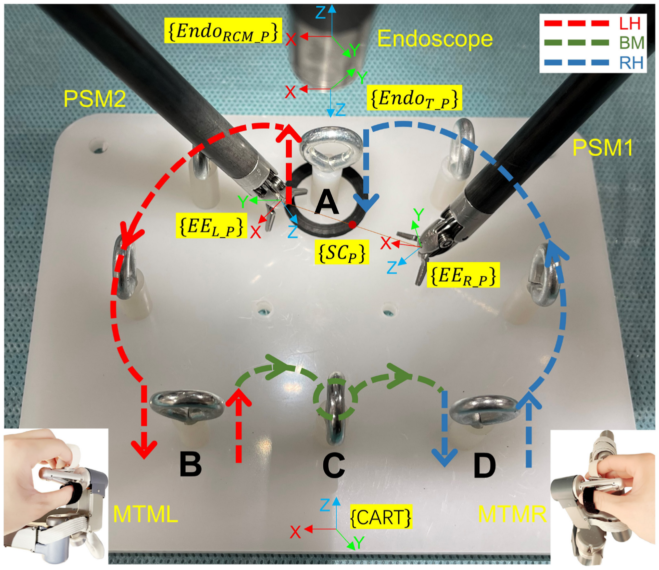

Fig. 3. A demonstration of the ring transferring task based on the dVRK platform. The red path represents “LH”, where the user needs to operate the left instrument to move the ring from Peg A to Peg B; the green path denotes “BM”, where the user performs a bimanual operation to swap the ring and transfer it to peg D; and the blue path demonstrates “RH”, where the user uses the right instrument to move the ring from peg D back to peg A. It also illustrates the definition of the reference frames, and the base frame is positioned at the base of the patient cart.

- RH: Operate the right instrument to place the ring down and pick it up again on Peg D, then transfer it to Peg A and place it down.

The diameter of the circular pegs is slightly smaller than that of the black ring, and the users need to avoid collisions with the pegs when manipulating the ring, so the users need a good Field Of View (FOV) using the endoscope during operation.

C. Endoscope Control

Instead of image-based endoscope navigation in the 2D image plane, position-based endoscope navigation is introduced to allow the endoscope motion in the 3D surgical space [35], and the positions of surgical instruments are considered as tracking objects for the endoscope. Three tracking modes based on the instruments are introduced, including the left end-effector, right end-effector and a variable-scale center point referencing both the left and right end-effectors [8]. The base frame is positioned at the base of the dVRK patient cart $\{CART\}$, and all subsequent position coordinates refer to the same base frame.

- Mode 1: The endoscope keeps tracking the end-effector position of the left instrument, and the position of the endoscope tip $Endo_{T\_P}$ can be formulated as:

$$Endo_{T\_P} = EE_{L\_P} - z_l \times \frac{EE_{L\_P} - Endo_{RCM\_P}}{\left\| EE_{L\_P} - Endo_{RCM\_P} \right\|} \tag{1}$$

where EE_{L_P} is the end-effector position of the left instrument, and $Endo_{RCM\_P}$ represents the RCM position of the endoscope. z_l is the zooming factor and it is set to 0.1 in this work. $\|,\|$ represents the Euclidean norm.

- Mode 2: The endoscope considers tracking the center point position SC_P between the left and right end-effectors in a variable scale manner, and the equation of the endoscope tip position $Endo_{T\_P}$ is written as:

$$SC_P = \frac{EE_{L\_P} + EE_{R\_P}}{2} \tag{2}$$

$$Endo_{T\_P} = SC_P - \left(z_m + s_d \left\| EE_{L\_P} - EE_{R\_P} \right\|\right) \times \frac{SC_P - Endo_{RCM\_P}}{\left\| SC_P - Endo_{RCM\_P} \right\|} \tag{3}$$

where EE_{L_P} and EE_{R_P} are the positions of left and right end-effectors, respectively. z_m is the zooming factor and it is equal to 0.08. s_d is the scale factor to control the variable zooming based on the distance between the left and right end-effectors, and it is set to 0.4. In this way, it implements a variable zooming effect influenced by both the center point and the distance between the two instruments.

- Mode 3: The calculation of the position of the endoscope tip is similar to mode 1, but the endoscope keeps tracking the end-effector of the right instrument,

$$Endo_{T\_P} = EE_{R\_P} - z_r \times \frac{EE_{R\_P} - Endo_{RCM\_P}}{\left\| EE_{R\_P} - Endo_{RCM\_P} \right\|} \tag{4}$$

where the zooming factor z_r is equal to 0.1 as well.

D. Context Awareness Based Tracking Mode Switching

Vision-based context awareness is introduced to enhance the autonomy of endoscope navigation by predicting three different operation phases in a ring transferring task, as shown in Fig. 3. Specifically, the endoscope will adopt Mode 1 (tracking the left end-effector) for the navigation when the recognized context is “LH”, the navigation will switch to Mode 2 (tracking the center point of the two end-effectors) when the phase is recognized to be “BM”, and the endoscope will perform the navigation of Mode 3 (tracking the right end-effector) with a prediction of “RH”. Using this navigation strategy can ensure that the endoscope continuously provides a good FOV for users during operation.

A series of neural networks [20], [22], [26], [29], [36] were compared to find out the best configuration to predict specific phases based on images. We also reproduced the kinematics-based model proposed in [15], which utilized 17 dimensional kinematic features to recognize the phases during operation. To construct the database for model training, we captured seven videos (containing 10763 images with the corresponding kinematic data) by performing the ring transferring task based on the dVRK platform. Following the image preprocessing approach in [26], we resized the images from 1920×1080 to 250×250 to reduce the inference time of the networks for real-time phase recognition. The phase labels of the images were manually annotated using the open source software Anvil [37]. It supports importing a single video, providing custom tracks for frame-by-frame phase labelling, and outputting a XML file containing the specific time and labels of all frames after annotation. The models [22], [26] rely on the input of video clips, and the model [15] utilizes the input of kinematics with consecutive frames, while other networks require a single image frame as the input. To promote the best performance of those models that use temporal features, we consider different downsampling rates when constructing the input of video clips and kinematics, including 1, 2, 5, 10, 15, 20 and 25. Here, the downsampling rate of 1 means that the frames consisting of the input are continuous, while a downsampling rate of 25 means that the number of frames between two sampled frames is 24 when making the video clip and kinematic input. The other parameters are standard configurations for their respective models. It should be noted that the downsampling operation was only used to construct video clips and kinematics for those models that consider inter-frame information as input, while all the collected images (for the vision-based models) and kinematic data (for the kinematics-based model) were used during the training and testing stages to maximize the utilization of the dataset and enhance the generalization.

To conduct a quantitative evaluation of the models, we consider six common metrics, including DICE coefficient $\left(\frac{2TP}{2TP+FP+FN}\right)$, precision $\left(\frac{TP}{TP+FP}\right)$, recall $\left(\frac{TP}{TP+FN}\right)$, specificity $\left(\frac{TN}{TN+FP}\right)$, accuracy $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$ and the inference time. Similar to the evaluation strategy in [22], [26], the accuracy was calculated at video-level by considering correct classifications in the entire video, while other metrics were calculated at phase-level and then we averaged the values of all phases to get the results of the entire video. 7-fold cross validation was adopted to ensure the reliability of the evaluation results. The models were trained on a server with an NVIDIA A100 GPU (40 GB), and evaluated on a local PC with an NVIDIA RTX 3080 GPU (16 GB).

E. User Study

Ten non-expert human subjects (22 to 29 years old, 6 males and 4 females, only one left-handed) were invited to perform the ring transferring task based on the dVRK platform in a dry lab. Written informed consent was obtained from all the subjects included in the study (Institutional Review Board (IRB) approval number: 30/2023). Three endoscope navigation pipelines were adopted for the participants:

- Control (C): the users perform the task based on a standard endoscope navigation pipeline by controlling both the Camera foot switch and MTMs.
- Semi-autonomous experiment (E1): the users perform the task using a semi-autonomous way [8] by controlling the Bicoag pedal to manually select different endoscope tracking modes.
- Autonomous experiment (E2): the users perform the task using a HOOTL endoscope navigation way, i.e., the neural network keeps recognizing the specific phase and switching different endoscope tracking modes automatically.

Before the experiments, the participants were provided around 30 minutes to become familiar with the da Vinci robot as well as the three different experiments. Once the users were ready to operate the robot, the formal experiments began and were repeated for three rounds. In each round, the three endoscope navigation experiments were performed randomly. The initial distance between the endoscope and the surgical instrument was set to 0.1m, so the users need to actively position the endoscope to maintain a good FOV. Furthermore, the function of the Bicoag foot switch was retained in the autonomous experiment (E2), in case the users want to switch the endoscope tracking modes by themselves caused of the wrong phase prediction of the network.

F. Performance Metrics

To quantitatively evaluate the proposed endoscope navigation system, we adopted four common performance metrics to analyze the data captured from the users, including:

1) The depressing number related to the endoscope foot switches N_{ENDO}. In the control experiment (C), the users need to depress the Camera pedal to position the endoscope, while the Bicoag pedal is utilized in the semi-autonomous experiment (E1), as well as the autonomous experiment (E2) if necessary, so it can be expressed by a unified formula as,

$$N_{ENDO} = \sum_{i=1}^{n}\{(1 \mid \mathrm{S}_{CAM} = 1) + (1 \mid \mathrm{S}_{BIC} = 1)\} \tag{5}$$

where n is the number of the collected data points in the whole task. $S_{CAM} = 1$ means that the depressing signal of the Camera pedal was detected as on, and $S_{BIC} = 1$ denotes that the user depressed the Bicoag pedal.

2) The movement path of the endoscope tip P_{ENDO} during operation, and it can be formulated as,

$$P_{ENDO} = \sum_{k=2}^{n}\left\| Endo_{T\_P}^{k} - Endo_{T\_P}^{k-1}\right\| \tag{6}$$

where $Endo_{T\_P}^{k}$ represents the k-th 3D position of the endoscope tip.

3) The movement path of MTMs P_{MTM} to perform the whole task, and it can be defined as,

$$P_{MTM} = \sum_{k=2}^{n}\left(\left\| M_{L\_P}^{k} - M_{L\_P}^{k-1}\right\| + \left\| M_{R\_P}^{k} - M_{R\_P}^{k-1}\right\|\right) \tag{7}$$

where $M_{L\_P}^{k}$ and $M_{R\_P}^{k}$ represent the end-effector positions of the left and right MTMs in the k-th data point, respectively.

4) The execution time T_{exe} to achieve the entire task.

The Wilcoxon signed-rank test ($p<0.05$) was adopted to check if there are significant differences among different endoscope navigation experiments. After the experiments, the users filled out a NASA-TLX questionnaire [38] by giving scores on six specific questions including Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. It helps understand the subjective workload concerning three different endoscope navigation pipelines.

III. Results

A. Comparison Study of the Networks for Context Awareness

Considering that some models require the input of video clips [22], [26] or kinematics [15] with consecutive frames, a preliminary comparison was made using different downsampling rates to generate the input of video clips and kinematics, and the result was shown in Fig. 4. Here, the TMRNet model contains three different architectures, and we tested their

TABLE II
QUANTITATIVE EVALUATION OF THE NETWORKS BASED ON 7-FOLD CROSS VALIDATION. THE NUMBER AFTER THE MODEL NAME IS THE BEST DOWNSAMPLING RATE CONSISTING OF VIDEO CLIP AND KINEMATIC INPUT, AND THE INPUT IS THE SINGLE FRAME WHEN THE NUMBER IS 0

Models	Accuracy	DICE	Precision	Recall	Specificity	Time (ms)
ResNeSt (0) / Ref.[36]	0.99±0.01	0.98±0.02	0.98±0.03	0.98±0.02	0.99±0.01	9.36±0.83
MTRCNet-CL (25) / Ref.[22]	0.99±0.01	0.98±0.02	0.98±0.02	0.98±0.03	0.99±0.01	9.22±0.11
TMRNet-S2 (5) / Ref.[26]	0.98±0.01	0.98±0.03	0.98±0.03	0.98±0.03	0.99±0.01	21.31±0.68
TMRNet-S1 (5) / Ref.[26]	0.98±0.02	0.97±0.04	0.98±0.03	0.97±0.07	0.99±0.02	21.10±0.86
NN-ICRA (0) / Ref.[29]	0.98±0.01	0.97±0.03	0.97±0.03	0.97±0.05	0.99±0.01	0.60±0.06
TMRNet-S3 (5) / Ref.[26]	0.98±0.01	0.97±0.03	0.97±0.03	0.96±0.06	0.99±0.01	56.05±2.00
PhaseNet (0) / Ref.[20]	0.90±0.21	0.87±0.30	0.86±0.31	0.89±0.29	0.95±0.21	1.19±0.05
Kinematics (15) / Ref.[15]	0.79±0.04	0.62±0.34	0.83±0.18	0.65±0.40	0.88±0.12	1.80±0.15

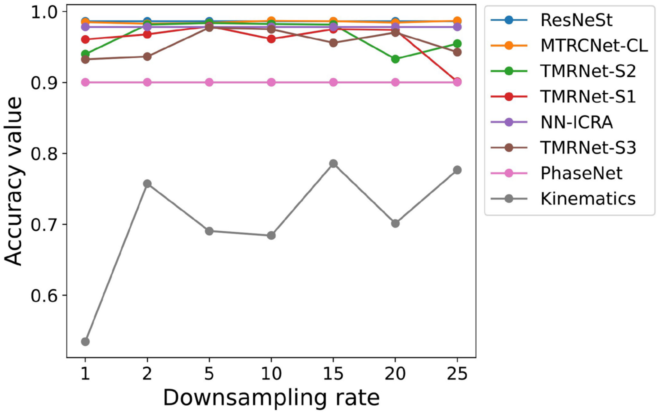

Fig. 4. The evaluation result using different downsampling rates to consist of video clips and kinematics with consecutive frames as input to the models. For the single frame input models, different downsampling rates do not affect the performance. The y-axis shows the values related to the accuracy metric based on 7-fold cross validation.

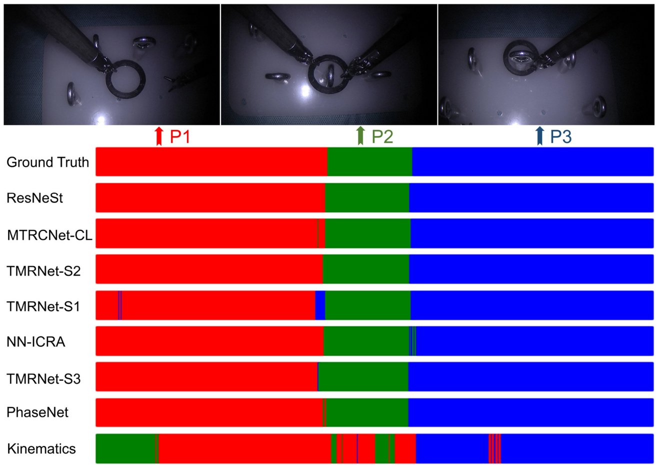

Fig. 5. A qualitative comparison of the phase recognition using the advanced models based on a ring transferring task. The color-coded ribbon illustrates the three defined phases along the temporal direction.

performance respectively from S1: the simplest structure to S3: the most complex structure. Then, we referenced the video-based accuracy metric to select the models with the best performance at the appropriate downsampling rate, and the quantitative result can be seen in Table II. The vision-based models ResNeSt [36] (single frame input) and MTRCNet-CL [22] (video clip input) got the highest recognition accuracy of 0.99, while the kinematics-based model [15] has the worst accuracy of 0.79 in predicting the phases. Also, we provided a qualitative evaluation of the phase recognition using a test video, as shown in Fig. 5. Then, we integrated the two models with the highest accuracy, ResNeSt and MTRCNet-CL, into our dVRK platform. After some practical tests by the participants, we observed that the model ResNeSt has higher generalization compared to the model MTRCNet-CL, so ResNeSt was adopted for the following user study. It can be explained that MTRCNet-CL uses the temporal features for phase recognition, however, the inter-frame information is variable under the manipulation of different users caused by some factors such as speed, and a fixed downsampling rate may not be able to handle such variability. On the contrary, ResNeSt obtained a reliable prediction because it requires a single frame as the input, i.e., it is not affected by the inter-frame variability, which ensures the generalization with the manipulation of different participants. The mean time of ResNeSt takes 9.36 ms (about 106 FPS), so it can provide real-time phase recognition during operation.

B. System Usability Evaluation

Fig. 6 shows the data distribution of ten participants who performed the ring transferring task in three endoscope navigation pipelines. Referencing the metric of N_{ENDO}, the result using the proposed HOOTL autonomous endoscope navigation (E2) has significant differences compared to both the standard navigation (C) and the semi-autonomous navigation pipelines (E1). Nobody utilized the Bicoag pedal in the autonomous endoscope navigation, though this foot switch was retained in this modality. It means that the selected vision-based model can perform an accurate and reliable phase prediction under the manipulation of different users, which implements a HOOTL endoscope control. Also, there are significant differences between E2 and the other two endoscope navigation experiments C and E1 when referencing T_{exe}. In the standard endoscope control, the users need to manually move the endoscope or the surgical instruments separately, which is a time-consuming operation. Similarly, in the semi-autonomous endoscope navigation pipeline, the users need to suspend their manipulation of the instruments and manually control the Bicoag pedal to switch different endoscope tracking modes. As a comparison, the users can concentrate

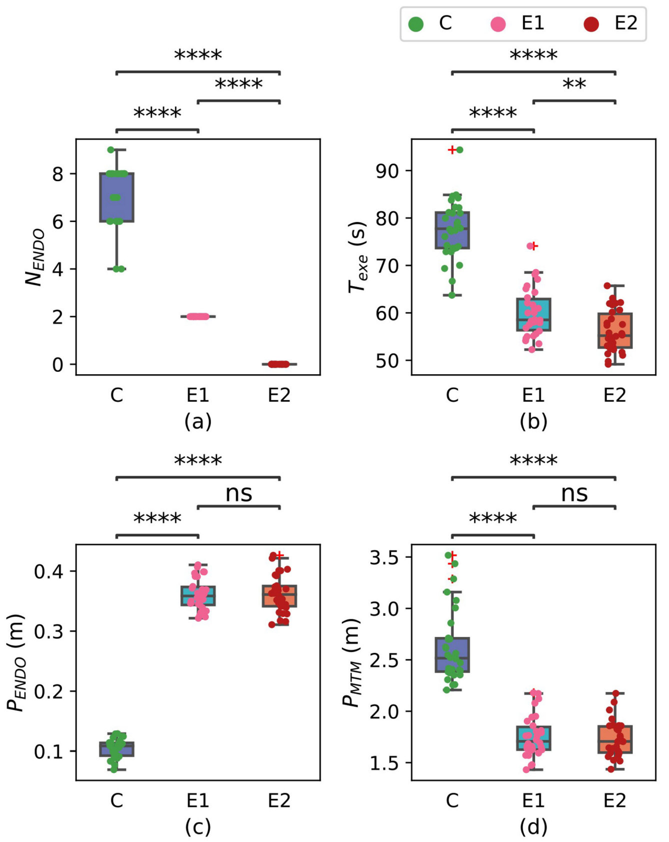

Fig. 6. The data captured from ten human subjects based on three endoscope navigation pipelines. "C" is the manual endoscope navigation, "E1" is the semi-autonomous endoscope navigation and "E2" represents the HOOTL endoscope navigation. The result of the statistical test is shown as ns:$0.05 < p \leq 1$, $\ast : 0.01 < p \leq 0.05$, $\ast\ast : 0.001 < p \leq 0.01$, $\ast\ast\ast : 0.0001 < p \leq 0.001$, and $\ast\ast\ast\ \ast : p \leq 0.0001$.

on the manipulation of surgical instruments in the HOOTL autonomous endoscope navigation, which can implement a smooth operation. When considering the other two metrics P_{ENDO} and P_{MTM}, there is no significant difference between E1 and E2, while the result of C is significantly different from the other two navigation pipelines. Specifically, the movement path of standard endoscope navigation is statistically smaller than the other two pipelines since manual navigation can not maintain a good FOV during the task. Furthermore, the path of MTMs is significantly greater than the semi-autonomous and HOOTL autonomous pipelines, since the users need to operate MTMs for both the surgical instrument positioning and the endoscope positioning in the standard navigation pipeline, which potentially introduces a high physical load.

The specific mean values and standard deviations are provided in Table III. With the assistance of context awareness based HOOTL endoscope navigation, the execution time T_{exe} is significantly reduced by 24.67% compared to the standard setup, and reduced by 6.57% compared to the semi-autonomous endoscope navigation. More importantly, it releases the need to control the pedals for adjusting the position of the endoscope during operation, while the mean number of depressing the pedals is 7.2 in the standard control (C) and 2 in the semi-autonomous modality (E1).

TABLE III
THE MEAN VALUES AND STANDARD DEVIATIONS FROM THE USERS

	C	E1	E2
N_{ENDO}	7.2000±1.2220	2.00±0.00	0.00±0.00
T_{exe} (s)	77.5410±6.0826	60.0551±5.0559	56.1079±4.3386
P_{ENDO} (m)	0.1049±0.0154	0.3611±0.0252	0.3617±0.0299
P_{MTM} (m)	2.6275±0.3517	1.7533±0.1971	1.7322±0.1721

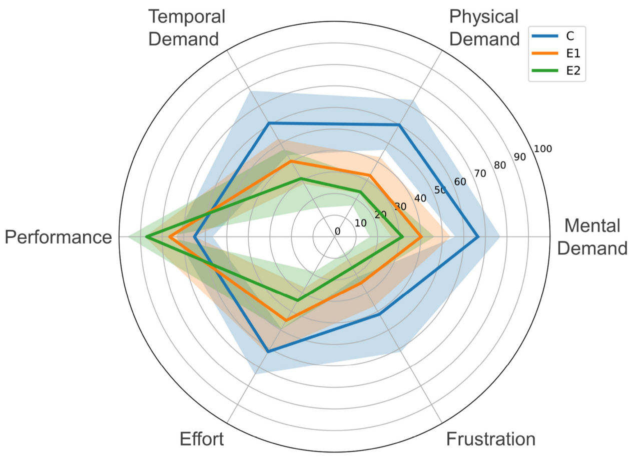

Fig. 7. The scores of the NASA-TLX questionnaire provided by the participants based on six specific questions. The mean scores based on the three endoscope navigation pipelines are shown as solid lines, and their standard deviations form the respective semi-transparent areas.

Fig. 7 presents the results of the NASA-TLX questionnaire filled out by the ten participants. The HOOTL autonomous endoscope navigation gets the highest mean score when referencing the metric of Performance, while the standard endoscope navigation gets the highest mean scores in the other five aspects. The users intuitively think that they have the best performance to achieve the task in the HOOTL experiment, while the standard navigation with manual endoscope manipulation introduces the highest workload in both physical and mental aspects. The statistical test shows that the weighted scores of the HOOTL autonomous endoscope navigation are significantly different from those of the standard navigation and the semi-autonomous navigation ($p = 0.0020$ and 0.0039).

IV. DISCUSSION

Autonomy in robotic surgery is an important research direction to reduce the operating burden for surgeons and increase the efficiency of surgery [11], and it can be mainly divided into two specific aspects: autonomy in surgical instruments and autonomy in endoscope. Endoscope autonomy is foreseeable and more likely to be deployed in existing medical robots than surgical instrument autonomy, because endoscopes do not directly contact soft tissues and organs during operation, while surgical instruments may damage delicate structures if accidents occur during autonomous operation. In this paper, we proposed a HOOTL endoscope navigation based on context awareness. Context awareness can adapt to the current surgical situation in the operating room and provide intelligent intra-operative decisions [39], which motivates us to integrate it into endoscope navigation to perform a HOOTL control, aiming at higher surgical benefits.

Deep learning has gradually taken a dominant position in various fields of medical image processing [40], [41] including surgical context recognition, so we introduced the vision-based neural networks to recognize specific surgical phases, which were used to automatically switch the endoscope tracking modes. After a comprehensive comparison study and physical tests on the dVRK platform, we found that ResNeSt can provide satisfactory phase recognition in terms of accuracy and speed, so it was adopted to perform the user study in the ring transferring task. Nevertheless, it should be noted that the models using temporal features may have better performance if they can adaptively process the inter-frame difference rather than adopting a fixed downsampling rate for the video clips. As more and more network architectures are proposed, the temporal feature based models may have higher generalization as the temporal feature may be a key information to be exploited in the surgical context recognition.

To evaluate the difference between the three endoscope navigation pipelines in Fig. 1, we invited ten participants to perform a classic ring transferring task. The result in Fig. 6 showed that the context awareness based HOOTL endoscope navigation approach allows the users to concentrate on the manipulation of surgical instruments, and relieve the manipulation burden of the endoscope, i.e., the users can perform a smooth operation using the HOOTL autonomous endoscope navigation, which helps shorten the operation time. According to the feedback of the participants after the experiments, the standard endoscope navigation pipeline needs improvement for them since they have to control both the foot switch and MTMs while suspending the operation of instruments when they want to position the endoscope, which makes their manipulation intermittent. As a comparison, the HOOTL endoscope navigation is preferred as they do not need to control either the pedal or the MTMs to position the endoscope, and they can concentrate on the surgical instruments to complete the task. The questionnaire scores also indicate that the HOOTL endoscope navigation can reduce the workload and improve performance from a subjective perspective.

A possible limitation comes from context awareness. Although the selected model showed strong generalization to the unseen users in our task, the complex clinical environment may affect the accuracy of context recognition. To address this issue, there are some possible solutions: (1) A foot pedal can be retained to allow users to manually switch endoscope tracking modes to address network errors if necessary. (2) Some authors mentioned that the integration of vision and kinematics may improve the context recognition ability in their future work [15], [16]. The integration of kinematic data may improve the generalization of the model, but it may also weaken the performance, which requires further research to verify. In our user study, the vision-based model showed a satisfactory recognition performance. (3) The emerging big models showed high generalization, such as Segment Anything (SAM) [42] in the field of image segmentation. Better context recognition models are expected in the near future.

Furthermore, we adopted a fixed orientation in the endoscope navigation, i.e., the endoscope maintains a horizontal shooting angle, as in some other works [15], [43]. Although the endoscope usually maintains a horizontal angle during surgery, surgeons sometimes manually rotate the angle of the endoscope for an optimal FOV. How to implement adaptive orientation adjustment in autonomous endoscope navigation remains a problem to be solved. Finally, the joint limits avoidance issue in autonomous endoscope navigation also needs to be further optimized. A promising solution proposed in [31] can be introduced to optimally constrain the motion of the ECM joints.

V. Conclusion

In this paper, a context-aware HOOTL endoscope navigation is proposed for augmented autonomy of robotic surgery. A comprehensive comparison study was conducted to find the best network among a number of state-of-the-art models for phase recognition. To evaluate the proposed HOOTL endoscope navigation, a user study involving ten participants was conducted based on a classic ring transferring task using three different endoscope navigation pipelines, and the experimental results showed that the proposed strategy can maintain a good FOV of the endoscope by automatically switching different endoscope tracking modes based on the recognized phases. It reduces the workload of the users, and significantly shortens the operation time, showing the possibility and potential to be integrated into clinical practice.

To increase the practicality of the proposed strategy in clinical use, we consider introducing more endoscope tracking modes and more elaborate surgical tasks involving more phases. In addition, some surgeons who have experience in robotic surgery can be invited to expand our user base.

References

[1] H. Alfalahi, F. Renda, and C. Stefanini, "Concentric tube robots for minimally invasive surgery: Current applications and future opportunities," *IEEE Trans. Med. Robot. Bion.*, vol. 2, no. 3, pp. 410–424, Aug. 2020.

[2] E. De Momi and A. Segato, "Autonomous robotic surgery makes light work of anastomosis," *Sci. Robot.*, vol. 7, no. 62, 2022, Art. no. eabn6522.

[3] G. G. Muscolo and P. Fiorini, "Force-torque sensors for minimally invasive surgery robotic tools: An overview," *IEEE Trans. Med. Robot. Bion.*, vol. 5, no. 3, pp. 458–471, Aug. 2023.

[4] Z. Chen et al., "FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery," *Comput. Biol. Med.*, vol. 163, Sep. 2023, Art. no. 107121.

[5] K. H. Sheetz, J. Claflin, and J. B. Dimick, "Trends in the adoption of robotic surgery for common surgical procedures," *JAMA Netw. Open*, vol. 3, no. 1, 2020, Art. no. e1918911.

[6] B. Rocco et al., "Robot-assisted radical prostatectomy with the Versius robotic surgical system: First description of a clinical case," *Eur. Urol. Open Sci.*, vol. 48, pp. 82–83, Feb. 2023.

[7] F. Porpiglia et al., "Three-dimensional augmented reality robot-assisted partial nephrectomy in case of complex Tumours (PADUA≥ 10): A new intraoperative tool overcoming the ultrasound guidance," *Eur. Urol.*, vol. 78, no. 2, pp. 229–238, 2020.

[8] T. Da Col, A. Mariani, A. Deguet, A. Menciassi, P. Kazanzides, and E. De Momi, "SCAN: System for camera autonomous navigation in robotic-assisted surgery," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2020, pp. 2996–3002.

[9] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annu. Rev. Control, Robot., Auton. Syst.*, vol. 4, no. 1, pp. 651–679, 2021.

[10] P. Fiorini, K. Y. Goldberg, Y. Liu, and R. H. Taylor, "Concepts and trends in autonomy for robot-assisted surgery," *Proc. IEEE*, vol. 110, no. 7, pp. 993–1011, Jul. 2022.

[11] G.-Z. Yang et al., "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.*, vol. 2, Mar. 2017, Art. no. eaam8638.

[12] T. Da Col et al., "Automating endoscope motion in robotic surgery: A usability study on da vinci-assisted ex vivo neobladder reconstruction," *Front. Robot. AI*, vol. 8, Nov. 2021, Art. no. 707704.

[13] Z. Chen et al., "Robot-assisted ex vivo neobladder reconstruction: Preliminary results of surgical skill evaluation," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 17, no. 12, pp. 2315–2323, 2022.

[14] A. Mariani et al., "An experimental comparison towards autonomous camera navigation to optimize training in robot assisted surgery," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1461–1467, Apr. 2020.

[15] 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 Robot. Autom. Lett.*, vol. 8, no. 12, pp. 8263–8270, Dec. 2023.

[16] G. Menegozzo, D. Dall'Alba, C. Zandona, and P. Fiorini, "Surgical gesture recognition with time delay neural network based on kinematic data," in *Proc. Int. Symp. Med. Robot. (ISMR)*, 2019, pp. 1–7.

[17] M. Wagner et al., "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102770.

[18] K. C. Demir et al., "Deep learning in surgical workflow analysis: A review of phase and step recognition," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 11, pp. 5405–5417, Nov. 2023.

[19] B. Zhang, A. Ghanem, A. Simes, H. Choi, A. Yoo, and A. Min, "SWNet: Surgical workflow recognition with deep convolutional network," in *Proc. 4th Med. Imag. Deep Learn.*, 2021, pp. 855–869.

[20] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.

[22] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101572.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] H. Nakawala, R. Bianchi, L. E. Pescatori, O. De Cobelli, G. Ferrigno, and E. De Momi, "'Deep-onto' network for surgical workflow and context recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, pp. 685–696, Apr. 2019.

[26] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng, "Temporal memory relation network for workflow recognition from surgical video," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1911–1923, Jul. 2021.

[27] X. Gao, Y. Jin, Y. Long, Q. Dou, and P.-A. Heng, "Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Interve. (MICCAI)*, 2021, pp. 593–603.

[28] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[29] D. Zhang et al., "Human-robot shared control for surgical robot based on context-aware sim-to-real adaptation," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 7694–7700.

[30] Z. Chen et al., "Towards safer robot-assisted surgery: A markerless augmented reality framework," 2023, *arXiv:2309.07693*.

[31] R. Moccia and F. Ficuciello, "Autonomous endoscope control algorithm with visibility and joint limits avoidance constraints for da vinci research kit robot," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 776–781.

[32] G. A. Fontanelli, M. Selvaggio, M. Ferro, F. Ficuciello, M. Vendittelli, and B. Siciliano, "A V-REP simulator for the da vinci research kit robotic platform," in *Proc. 7th IEEE Int. Conf. Biomed. Robot. Biomechatron. (Biorob)*, 2018, pp. 1056–1061.

[33] R. A. Joseph et al., "'Chopstick' surgery: A novel technique improves surgeon performance and eliminates arm collision in robotic single-incision laparoscopic surgery," *Surg. Endosc.*, vol. 24, pp. 1331–1335, Jun. 2010.

[34] Y. Long, J. Cao, A. Deguet, R. H. Taylor, and Q. Dou, "Integrating artificial intelligence and augmented reality in robotic surgery: An initial dVRK study using a surgical education scenario," in *Proc. Int. Symp. Med. Robot. (ISMR)*, 2022, pp. 1–8.

[35] Y.-C. Peng, D. Jivani, R. J. Radke, and J. Wen, "Comparing position-and image-based visual servoing for robotic assembly of large structures," in *Proc. IEEE 16th Int. Conf. Autom. Sci. Eng. (CASE)*, 2020, pp. 1608–1613.

[36] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2736–2746.

[37] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 1–4.

[38] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Jan. 1988.

[39] J. Neumann et al., "Ontology-based surgical workflow recognition and prediction," *J. Biomed. Inform.*, vol. 136, Dec. 2022, Art. no. 104240.

[40] X. Chen et al., "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102444.

[41] P. Celard, E. Iglesias, J. Sorribes-Fdez, R. Romero, A. S. Vieira, and L. Borrajo, "A survey on deep learning applied to medical images: From simple artificial neural networks to generative models," *Neural Comput. Appl.*, vol. 35, no. 3, pp. 2291–2323, 2023.

[42] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[43] E. Iovene et al., "Towards exoscope automation in neurosurgery: A markerless visual-servoing approach," *IEEE Trans. Med. Robot. Bion.*, vol. 5, no. 2, pp. 411–420, May 2023.

Open Access funding provided by 'Politecnico di Milano' within the CRUI CARE Agreement