
Human-Robot Cooperation for Teleoperation in Robotic Surgery

Author Zhaoyang Jacopo Hu **Supervisors** Prof. F. Rodriguez y Baena¹
Submitted May 23rd 2025 Prof. E. Burdet²
Dr. K. Bhattacharyya³

¹Department of Mechanical Engineering, Imperial College London, UK.

²Department of Bioengineering, Imperial College London, UK.

³Intuitive Surgical Operations, Inc., Norcross GA, US.

A thesis submitted in fulfilment of the requirements for the degree of *Philosophiae Doctor*

Mechatronics in Medicine Laboratory
The Hamlyn Centre, Institute of Global Health Innovation
Department of Mechanical Engineering
Imperial College London

Mechatronics in Medicine Lab,
Department of Mechanical Engineering,
Imperial College London,
Exhibition Road, South Kensington,
SW7 2AZ

Zhaoyang Jacopo Hu © 2025

晚餐吃了吗？

Casa

Declaration

I hereby declare that the content of this thesis is original work of my own. That which is not, is acknowledged and referenced accordingly.

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

I would like to thank my supervisor, Professor Ferdinando Rodriguez y Baena, for being everything I needed during my PhD: a mentor, a guide, and a friend. A huge thank you to Professor Etienne Burdet for always being there with useful expertise and helpful advice, insightful feedback and guidance. A heartfelt thank you to Dr Kiran Bhattacharyya, for encouraging my curiosity, inspiring me and whose guidance has been a pivotal part of my journey.

I am deeply grateful to those who have guided and supported me with invaluable advices. My sincere thanks go to Ziwei and Ke for their mentorship and friendship.

I would like to thank the MiM lab and those who have been close during my PhD journey: Emilia, Dani, Ivet, Varell, Alex, Nicoletta, Sofia, Korn, and Ana. A special thank you to my friends on the 5th floor: Haozheng, Chi, Yihang, Lily, Yingwei, Jinshi, Angelo, and Rema. I would also like to thank the friends I made during my time in the United States: David, Rogerio, Sreeram, Connor, and Max. I would also like to acknowledge the students for which I had the pleasure to personally supervise: Longrui, Haoyi, and Sion.

I'm deeply grateful for the support from my fantastic sisters Giada and Alessia.

Lastly, I thank my 妈咪 e 爸比, to whom I owe everything and dedicate everything.

Abstract

This thesis investigates data-driven methods for implementing human-robot interaction in the da Vinci Research Kit (dVRK) platform, with a particular focus on skill modeling, integration of autonomy in the surgical workflow, and real-time performance assessment. As robotic surgery becomes increasingly prevalent, there is a growing need for intelligent systems that can both assist human operators and also understand and adapt to their actions in real-time. This work addresses this challenge by developing a high performing control framework through a sequence of studies that span through the different types of controllers in the spectrum of teleoperative human-robot cooperation and skill assessment: shared control, supervisory control, and assistive control.

Early chapters focus on shared control and supervisory control frameworks, introducing novel methods for learning bimanual surgical trajectories from demonstrations, enabling robots to generalise motion patterns from expert behavior. The subsequent work explores control paradigms that modulate robotic assistance in a data-driven approach, balancing autonomy and human input to achieve both higher performances and lower perceived workload. Objective evaluations demonstrate that adaptive assistance can benefit novice users without significantly impeding expert performance.

Building upon these foundations, the last chapter shifts focus to real-time surgical skill assessment using deep learning models trained on multimodal inputs. These systems are designed to provide frame-level predictions of technical skill, enabling continuous feedback during task execution.

Across all studies, this thesis emphasises the integration of expert data, task structure, and real-time capabilities to build responsive surgical systems that can fit into the surgical workflow. The contributions are validated with user studies and supported by extensive experimentation on datasets. User studies with participants show improved performance and perceived workload throughout all experiments when using the proposed control systems. Together, these contributions advance the state-of-the-art in human-robot cooperation and provide a foundation for intelligent systems in surgical assistance.

Contents

1	Introduction	21
1.1	Motivation	22
1.2	Research Objectives	23
1.3	Original Contributions	23
1.4	Publications	24
1.5	Technical and Clinical Acknowledgements	25
1.6	Thesis Overview	25
2	Current technologies for robot-assisted surgery	28
2.1	Introduction	29
2.2	Challenges and Limitations in Robotic-Assisted Surgery	31
2.3	Support during Robot-Assisted Surgery	35
2.3.1	Human-Robot Cooperation	35
2.3.2	Artificial Intelligence	40
2.4	Conclusion	44
3	Shared Control in Peg Transfer	45
3.1	Overview	46
3.2	Working Principle	47
3.2.1	Shared Control during Peg Transfer Task	47
3.2.2	Control Framework	48
3.3	Autonomous Agent	50
3.3.1	EG MaxEnt IRL Algorithm	50
3.3.2	Agent Training and HRI Environment	50
3.4	Experiment	52
3.4.1	Experiment Setup	52
3.4.1.1	Simulation Platform	52
3.4.1.2	dVRK Platform	52
3.4.2	Task and Control Modes	53
3.4.3	Experimental Protocol	54
3.4.3.1	Surgical Simulation	54
3.4.3.2	dVRK Platform	54
3.4.4	User Performance Evaluation	54

3.5	Results	54
3.5.1	Performance Comparison between MT and SC	55
3.5.2	Mental Effort Evaluation of MT and SC using Gaze Data	55
3.5.3	dVRK Platform Results	57
3.6	Discussion	60
3.7	Conclusion	61
4	Supervisory Control in Peg Transfer	63
4.1	Introduction	64
4.2	Methodology	66
4.2.1	Transverse Generative Adversarial Imitation Learning	66
4.2.2	Priority-changing Control	67
4.2.3	Control Methods	69
4.2.4	Simulation Experiment Results	70
4.3	Experiments	71
4.3.1	User study on dVRK	71
4.4	Results	73
4.4.1	Generalisation Assessment	73
4.4.2	Real-world Experiment	73
4.5	Discussion	74
4.6	Conclusions	75
5	Intent detection in Robotic Suturing	76
5.1	Introduction	77
5.2	Compact Holder for Enhanced Needle Alignment	80
5.3	Kinematic Gesture Recognition	81
5.3.1	Suturing Dataset	81
5.3.2	Gesture Recognition Algorithm	82
5.3.3	Suturing Gesture Classification	82
5.4	Confidence-based IAC	86
5.4.1	Intention Assimilation Controller	86
5.4.2	Confidence with Bayesian Inference	87
5.4.3	Shared Control Paradigms	89
5.4.3.1	Positioning	89
5.4.3.2	Push	90
5.4.3.3	Pull	90
5.4.3.4	Hand-off	91
5.4.3.5	Other	91
5.5	Experimental Setting	92

5.6	Robot-Camera Calibration	93
5.6.1	Sigma.7 to dVRK	93
5.6.2	Hand-Eye Calibration	94
5.6.2.1	Vision	94
5.6.2.2	Proprioception	94
5.7	Experiments	95
5.7.1	Target Reaching User Study	96
5.7.2	Suturing Task User Study	96
5.8	Results	97
5.8.1	Transformer for Two Classification Strategies	97
5.8.2	STITCHES Dataset Validation	98
5.8.3	Needle Pose Estimation Error	98
5.8.4	Target Reaching User Study	99
5.8.5	Suturing Task User Study	99
5.9	Discussion	101
5.10	Conclusion	103
6	Skill Assessment in Robot-Assisted Surgery	105
6.1	Introduction	106
6.2	Related Work	107
6.3	Data Pre-processing	108
6.3.1	Kinematic data	108
6.3.2	Video data	109
6.3.3	OSATS rating for skill evaluation	109
6.4	Models	110
6.4.1	CNN-LSTM-V (Vision only)	111
6.4.2	LSTM-K (Kinematic only)	111
6.4.3	DualLSTM-F (Fusion)	111
6.4.4	CNN-V (Vision only)	112
6.4.5	Transformer-V (Vision only)	112
6.4.6	Transformer-K (Kinematic only)	112
6.4.7	DualTransformer-F (Fusion)	113
6.5	Results and Discussion	113
6.5.1	Mean ρ and Mean $RMSE$	114
6.5.2	Real-time OSATS predictions	116
6.5.3	LOSI analysis	117
6.6	Conclusions and Future Work	120
7	Conclusions	122

7.1	Summary of Thesis Achievements	123
7.2	Limitations	124
7.3	Further Work	125
References		128
Appendix A Ethics		137
Appendix B Image Reuse Permissions		154

List of Figures

2.1	Senhance robot. (©2018 reproduced from Rao <i>et al.</i> [122] with permission from Springer Nature)	30
2.2	REVO-I robot. (©2018 reproduced from Rao <i>et al.</i> [122] with permission from Springer Nature)	31
2.3	Versius robot. (©2018 reproduced from Peters <i>et al.</i> [116] with permission from Springer Nature)	31
2.4	Group performance comparison for surgical robotic skill acquisition with control group as reference (task time and errors) on three surgical platforms: dV-Trainer, da Vinci Simulation System (DVSS), and Dome [134]. Symbols represent estimated time (top) and error (bottom) difference from control group in seconds in all tasks adjusted for baseline differences among groups. The control group was offered the Intuitive da Vinci curriculum. Error bars represent standard error of the estimates. Control group performance has been set as 0 in the Y-axis. P-values refer to comparison of each group's performance with control on the same tasks. (©2020 reproduced from Satava <i>et al.</i> [134] with permission from Wolters Kluwer Health, Inc.).	34
2.5	Research areas based on dVRK literature. (©2021 reproduced from d'Ettore <i>et al.</i> [27] with permission from IEEE)	36
2.6	Spectrum of human-robot teleoperation strategies. Each category—Manual Surgery, Assisted Control, Shared Control, Supervisory Control, and Full Automation—represents a distinct balance of autonomy and interaction between the human and the robot. The lower matrix shows which agent, Human (H) or Robot (R), is responsible for Sensing (S), Planning (P), and Acting (A) in each category. Taxonomy reproduced based on [34], [105], [12], [11], [96].	37
2.7	Histogram on data usage based on papers involving the dVRK. RI: raw image; KD: kinematic data; DD: dynamic data; SD: system data; ED: external data. (©2021 reproduced from d'Ettore <i>et al.</i> [27] with permission from IEEE)	41
3.1	Segmentation of the bimanual peg transfer task using shared control. The segments A and C are controlled by the autonomous agent. B and D are carried out by the human operator and E by a human assistant.	48
3.2	Process diagram for the training and testing of the bimanual peg transfer using shared control.	49

3.3	Experimental setup using surgical simulation and physical dVRK platform. (a) Participant performing the task on the same surgical simulation environment on which the autonomous agent was trained. (b) Users with experience on the surgical simulator validate the controllers on the physical dVRK platform. (c) Bimanual framework for the unified real-time simulator and reinforcement learning environment in a local-remote setup.	53
3.4	(a) Agent trajectory with different starting states during target approach. (b) Bimanual peg transfer user trajectories achieved after the participants became familiar with the task.	58
3.5	Experiment results with the laparoscopic simulator platform. (a) Average number of success trials. (b) Average completion time. (c) Subjective rating of questionnaire. Asterisks denote significant effects at $*p<0.05$, $**p<0.01$, $***p<0.001$	59
3.6	Experiment results with the dVRK system. (a) Average completion time of SC and MT. (b) Completion time with 10 trials for each control mode per subject. *Subjects that used console and endoscope during the experiment.	59
4.1	Control system and experimental setup. The agent is trained from the demonstration of an expert (i.e. a user with more than 300 hours of experience with the control interface). The priority-changing strategy incorporates the input from both operator and agent to control the robot.	65
4.2	Training comparison between GAIL, Random GAIL and tGAIL, with same initial settings. Random GAIL starts training at random points and tGAIL starts at points with the most failure rates.	67
4.3	Performance and questionnaire results from an experiment with participants. (a-b) Completion time and success rate of each participant using the three controllers: supervision (SUP), segmented with supervision (SEG-S) and manual control (FC). (c) Subjective rating of NASA TLX questionnaire.	68
4.4	Generalisation ability test of EG MaxEnt IRL (left), GAIL (middle) and tGAIL (right). The three algorithms are trained with the same ten expert trajectories (marked with colored lines), and the successful trials' starting points are marked as initial states.	71
4.5	Obstacle in the peg transfer task. The obstacle features a soft top and hard base. When transferring to a close peg, it serves as a visual barrier; when transferring to a far peg, it becomes a physical impediment.	72
4.6	Performance and questionnaire results from real-world experiment. (a-c) Completion time, human input and clutch usage of each participant using the three controllers. (d) Subject rating to questionnaire using NASA TLX. The terminology co, cw, fo, and fw stand for close obstacle, close without obstacle, far with obstacle, and far without obstacle, respectively.	72

5.1	Bimanual robotic suturing can be described as the composition of four surges: (1) Needle tip positioned at the entry point on the phantom surface; (2) Needle pushed inside the phantom reaching the exit point; (3) Needle pulled out from the exit point; (4) Needle passed back to the starting manipulator. During the suturing, a da Vinci Si camera tracks the relative pose of the elements in the surgical setting using markers.	78
5.2	High and low level inference of task actions. At the high level, the system comprises of a kinematic gesture recognition for multi-action tasks. At the low level, the system uses a target detection to understand the control of the human, which is regulated by a dynamic confidence parameter.	79
5.3	Design of the CHENA. (a-b) The structure comprises of walls that create a concavity where the needle is attracted and maintained fixed by a magnet. (c-d) The needle can self-align using two independent acting forces: the jaw pushing and the magnet pulling the needle towards the cavity.	81
5.4	Architecture of the proposed real-time gesture recognition system. The kinematic data from the dVRK captured in a moving window is fed to the classification network composed of N Transformer encoder blocks followed by fully connected layers and a softmax activation function. Classification probabilities are then averaged with an EMA window. Gestures are only updated when they have an averaged probability over the predefined threshold.	83
5.5	Four surges for robotic suturing. We use an additional <i>Other</i> class to refer to gestures fundamentally different from any of the surges in (a-d).	84
5.6	Comparison between a) traditional and b) C-IAC teleoperation of the dVRK. While traditional teleoperation has inherent delay to reach the current relative position of the user's hand, C-IAC allows to control the robot to move directly towards the predicted position. A confidence component λ is used in the shared control scheme to adjust the assistance from the robot. In cases of high confidence ($\lambda \approx 1$), the teleoperation delay can be minimised by allowing the robot to follow its own target.	87
5.7	Task frame position on the tissue surface using a ChArUco marker. The x-axis is parallel to the wound, the y-axis is perpendicular to the wound, and the z-axis is perpendicular to the tissue surface. During the suturing task, the robot refers to the task frame Cartesian coordinates for knowledge of the environment's objects pose, such as wound orientation and insertion point.	90

5.8	Bimanual suturing control scheme using C-IAC and Transformer for gesture classification. The markers and robot kinematics constitute the sensory information for the robot. The quality of the sensory information is processed through a Bayesian inference which provides the confidence parameter λ to the IAC. The kinematics are instead used by Transformer to decide the correct target τ_r according to the gesture that the user is performing. In IAC, the user input u_h is used to predict the human's target τ_h which, together with τ_r , determines the target for the robot's manipulator.	92
5.9	Experimental dVRK setup with sigma.7 devices and additional da Vinci Si Camera. The phantom is marked with an ChArUco marker and the end-effector is equipped with a CHENA and keydot marker. While the user has a 2D visual feedback from the surgical setting, the robot uses the da Vinci Si camera to obtain three-dimensional information from the markers.	93
5.10	Visual feedback during target reaching user study. The users perform <i>Positioning</i> and <i>Push</i> at each entry point. For each target reaching movement, the manipulator is reset to the same starting position.	97
5.11	Use of Transformer model and STITCHES dataset validation. (a-b) Transformer gesture classification using labelling strategy 1 and strategy 2. (c) 5-fold cross validation average accuracy of STITCHES dataset. The overall accuracy is 81.19% using the proposed labelling for the suturing task. (d) Real-time sample comparison between ground truth and predicted gesture. (e) Real-time suturing throw sample with use of EMA and probability threshold.	98
5.12	Novice gesture recognition online performance during the suturing task with C-IAC.	100
5.13	NASA TLX results from suturing task participants (i.e. novices, intermediate, and expert) for traditional teleoperation and C-IAC. In the spider charts, we first compare the two controllers using the average answers of the novices and the intermediate and expert's individual scores to provide a perspective from participants with different suturing skills. As all participants are inexperienced with the surgical robot, we then average all participants scores for Traditional vs C-IAC. In the box charts, we present a detailed view of the scores given by the novice participants ($n = 7$) during this user study.	101
6.1	Real-time prediction of OSATS scores for two representative sample subjects. Each subplot corresponds to a different OSATS ¹⁻³ component, where the x-axis denotes time and the y-axis reflects the prediction scores. The lines represent the ground truth label, and the predictions from the kinematic-only-based model, vision-based-only model, and the fusion model. The background is segmented with colored bands indicating the active gesture label G as provided by JIGSAWS, providing contextual cues from the temporal evolution of surgical gestures. . . .	117

6.2	Gesture G6 (Pulling) in JIGSAWS's trial. As the needle is pulled out, the tissue is visibly torn (highlighted in red).	118
6.3	Results from LOSI cross validation using kinematic-only model LSTM-K with testing on OSATS ¹⁻³ = [n, n, n].	119
6.4	Results from LOSI cross validation using vision-only model CNN-LSTM-V with testing on OSATS ¹⁻³ = [n, n, n].	120

List of Tables

2.1	Failures resulting in cessation of robotic surgery reported in Kim et al. [80]	32
2.2	Rating scale for the assessment of surgical skills across different domains. Each domain represents an OSATS score used to evaluate a surgeon's performance. Table reproduced from Gao et al. [42].	43
3.1	EG MaxEnt IRL hyperparameters.	51
3.2	Mean values across participants for gaze features associated with mental load during the peg transfers.	56
4.1	Controllers summary	70
5.1	Surgemes in JIGSAWS	85
5.2	Proposed surgeme labels (a-d) compared to jigsaws	85
5.3	Visual Information Performance Assignment	88
5.4	Needle Pose Estimation Error Results	99
5.5	Duration of Target Reaching (in seconds)	99
5.6	User Study Suturing Task Performance based on Participant Suturing Experience	99
6.1	Performance comparison for Suturing based on $\bar{\rho} \mid \overline{RSME}$, using different architectures and source data under LOSO and LOUO evaluation methods (highest performances in bold).	115
6.2	Performance comparison for Suturing based on $\rho \mid RSME$, using different architectures and source data under LOSO and LOUO evaluation methods (highest OSATS performances in bold). * $p < 0.05$	116

Acronyms

AI Artificial Intelligence. 25, 26, 40–42, 44, 124

BiLSTM Bidirectional Long Short-Term Memory. 107

C-IAC Confidence-based Intention Assimilation Controller. 15, 16, 76, 79, 80, 87, 89, 91–96, 99–104, 124, 126

CHENA Compact Holder for Enhanced Needle Alignment. 15, 16, 80, 81, 91–96, 98, 104

CNN Convolutional Neural Network. 42, 107, 111, 112

DNN Deep Neural Network. 42

DOF Degrees of Freedom. 29

DoF Degrees of Freedom. 52, 53, 89

dVRK da Vinci Research Kit. 13–16, 22, 23, 26, 29, 40, 41, 45, 47, 49–53, 57, 60, 61, 66, 69–71, 74–76, 80, 82–84, 87, 89, 92–94, 96, 103, 104, 122, 123

EG MaxEnt IRL Epsilon-Greedy Maximum Entropy Inverse Reinforcement Learning. 14, 26, 47, 50, 60, 61, 66, 69, 71, 73, 74, 123, 124

EMA esponential moving average. 15, 16, 82, 83, 98

FC Fully Manual Control. 14, 68–71, 73–75

FDA Food and Drug Administration. 29

GAIL generative adversarial imitation learning. 14, 66, 67, 71, 73, 74

HRI Human-Robot Interaction. 48, 50, 61

IAC Intention Assimilation Controller. 16, 79, 80, 86, 92, 103

IPA Index of Pupillary Activity. 55, 61

JIGSAWS JHU-ISI Gesture and Skill Assessment Working Set. 16, 26, 80, 81, 84, 93, 101, 105–110, 113, 117, 125

LOSI Leave-one-score-in. 110, 111, 113, 117, 118, 121

LOSO Leave-one-supertrial-out. 18, 110, 111, 114–117

LOUO Leave-one-user-out. 18, 110, 111, 114–117

LSTM Long Short-Term Memory. 110, 111

MaxEnt Maximum Entropy. 50

MaxEnt IRL Maximum Entropy Inverse Reinforcement Learning. 46, 51, 61, 62

MT Manual Teleoperation. 47, 53–55, 57, 60, 61

MTM Master Tool Manipulator. 29, 108

OSATS Objective Structured Assessment of Technical Skills. 16, 18, 26, 41, 42, 105, 107–109, 111, 113, 114, 116–119, 121, 124, 125

PSM Patient Side Manipulator. 29, 81, 82, 91–93, 95, 96, 108

- RNN** Recurrent Neural Network. 42, 107
- SC** Shared Control. 47, 53–55, 57, 60, 61
- SEG** Segmented Control. 69
- SEG-A** Autonomous Segmented Control. 69–71, 73–75
- SEG-S** Supervisory Segmented Control. 14, 68–70
- STAR** Smart Tissue Autonomous Robot. 64
- STITCHES** Suturing Task in Imperial’s Tracking Collection for Hamlyn’s Evaluation Set. 16, 82, 95, 98, 101–103
- SUP** Supervision Control. 14, 68–71, 73–75
- SurRoL** Surgical Robot Learning. 26, 47, 48, 50, 51, 60, 69
- tGAIL** transverse generative adversarial imitation learning. 14, 26, 63, 66, 67, 71, 73–75, 123

Chapter 1

Introduction

"The procedure was a blessing, being less invasive than the open method [...]. I opted for [robotic-assisted surgery] because of the recovery time, less blood loss, it seemed to be the logical way to go."

1.1 Motivation

The above quote comes from a patient in a 2018 study conducted by St John of God Subiaco Hospital, Austin Health, and the University of Notre Dame Fremantle in Australia, which explored the public's understanding of robotic surgery [124]. It encapsulates the perception and growing preference for robot-assisted surgical procedures, largely due to their minimally invasive nature, reduced blood loss, and shorter recovery times.

The da Vinci surgical system, developed by Intuitive Surgical (Sunnyvale, CA), is the leading platform in this domain. Through its advanced teleoperation capabilities, it enables surgeons to perform complex procedures with enhanced precision, flexibility, and control via a console-based interface. Its success has contributed to the global adoption of robotic systems across various surgical specialties. However, these advantages are accompanied by new challenges. One notable concern is the prolonged duration of robotic procedures. While the minimally invasive approach reduces recovery times for patients, the overall surgical time is often extended compared to traditional open or laparoscopic methods. This can have a range of implications. For patients, longer surgeries may require extended periods in physiologically stressful positions, such as the Trendelenburg position, which has been associated with risks including increased intracranial pressure, airway edema, and cardiovascular strain [104].

For surgeons, extended procedures raise concerns of cognitive and physical fatigue, especially during lengthy and complex operations. Studies have shown that prolonged exposure to surgical tasks, even in ergonomically designed teleoperation consoles like that of the da Vinci system, can impair performance, lead to mental overload, and potentially increase the risk of errors [163]. While the system provides an ergonomic advantage over traditional surgical postures, it does not eliminate the burden of extended concentration and decision-making demands. These observations highlight a critical opportunity: to augment existing teleoperative capabilities with intelligent assistance mechanisms that can support the surgeon throughout the procedure. By integrating intelligent human-robot cooperation frameworks into surgical robotics, the system can move from being a passive instrument under the surgeon's complete control to an active collaborator capable of providing context-aware support, adaptive motion assistance, and even predictive modeling of surgical tasks.

Thus, the motivation of this work lies in exploring how enhanced human-robot cooperation in robotic surgery within the context of teleoperation can reduce cognitive load, improve workflow efficiency, and ultimately contribute to more effective surgical outcomes. To investigate these possibilities, this thesis uses the da Vinci Research Kit (dVRK), an open-source research platform derived from the clinical da Vinci surgical system. The dVRK provides access to low-level control and sensor data, allowing the development of advanced control frameworks that go beyond conventional teleoperation.

1.2 Research Objectives

Using the dVRK as basis of the work conducted in this Thesis, the aim is to explore and evaluate novel control architectures for human-robot cooperation in teleoperative robotic surgery. The objectives of this thesis can be summarised as follows:

1. Novel control frameworks for robot-assisted systems.

To create a teleoperation framework which implements human-robot cooperation in the dVRK for surgical tasks, including training tasks, such as peg transfer, and common clinical procedures, such as suturing.

2. Perceived workload analyses.

To investigate and assess surgeon performance during robotic surgery tasks employing human-robot interaction frameworks. To find the key elements that can decrease surgeon's perceived mental workload.

3. Data analysis for skill assessment.

To investigate the feasibility and performance of machine learning models in assisting surgeons for surgical tasks. To validate the models in benchmark datasets widely used in the literature.

1.3 Original Contributions

The main contributions of this thesis include:

Contribution 1. *Development of novel shared control frameworks for peg transfer.*

This thesis presents shared and supervisory control approaches that segment the peg transfer task into human and robot subtasks, using learning from demonstration methods to train the autonomous agent. The framework improves task efficiency and user experience by enabling the robot to handle routine actions while maintaining user engagement.

Contribution 2. *Implementation of a novel assistive control framework for suturing.*

A new assistive control system was developed that combines real-time gesture recognition with confidence-based intent prediction to dynamically adapt robotic assistance during suturing. This approach enhances the user's performance and reduces cognitive load, particularly during complex maneuvers or in conditions with limited visibility.

Contribution 3. *Demonstration of human-robot cooperation improved performances in robotic surgery tasks.*

Extensive user studies were conducted to evaluate the impact of the proposed control frameworks across various tasks, confirming improvements in efficiency, accuracy, and user satisfaction through NASA TLX. These studies validate the practical benefits of integrating autonomous support into surgical workflows.

Contribution 4. *Construction of a real-time automatic surgical skill assessment pipeline.*

Deep learning-based models were developed to automatically assess suturing skills in real-time using kinematic and video suturing data. The study conducted provides insights that support objective and consistent evaluation of surgical tasks.

1.4 Publications

The work presented in this thesis has resulted in the following peer-reviewed conference papers and journal publications:

- Chen, L.*, **Hu, Z. J.***, Huang, Y., Burdet, E., Rodriguez y Baena, F. "Human Robot Shared Control in Surgery: A Performance Assessment". In: IEEE International Conference on Robotics and Automation (ICRA) (2024). *Co-first authors.
- **Hu, Z. J.**, Wang, Z., Huang, Y., Rodriguez y Baena, F., Burdet, E. "Towards human-robot collaborative surgery: Trajectory and strategy learning in bimanual peg transfer". In: IEEE Robotics and Automation Letters (2023), pp. 4553-4560.

Symposium papers, workshop abstracts and presentations arising from this thesis:

- **Hu, Z. J.**, Kim, S., Burdet, E., Rodriguez y Baena, F. "Gesture Recognition for Shared Control in Robotic Suturing with LSTM Networks and Transformers". In: Hamlyn Symposium 2024.
- **Hu, Z. J.**, Chen, L., Huang, Y., Burdet, E., Rodriguez y Baena, F. "Priority-Based Shared Control for Peg Transfer Task". In: Hamlyn Symposium 2023.
- **Hu, Z. J.**, Wang, Z., Huang, Y., Rodriguez y Baena, F., Burdet, E. "Towards Human-Robot Collaborative Surgery: Trajectory and Strategy Learning in Peg Transfer". In: International Conference on Intelligent Robots and Systems (IROS) 2022 Workshop.

Additional peer-reviewed surgical robotics-related journal and conference publications that have been co-authored:

- Zheng, H., **Hu, Z. J.**, Huang, Y., Cheng, X., Wang, Z., Burdet, E. "A User-Centered Shared Control Scheme with Learning from Demonstration for Robotic Surgery". In: IEEE International Conference on Robotics and Automation (ICRA) (2024).
- Zhang, H., Deng, K., **Hu, Z. J.**, Huang, B., Elson, D. S. "Hybrid Deep Reinforcement Learning for Radio Tracer Localisation in Robotic-assisted Radioguided Surgery". In: IEEE International

Conference on Robotics and Automation (ICRA) (2025).

The following manuscripts are under review:

- **Hu, Z. J.**, Xu, H., Kim, S., Li, Y., Rodriguez y Baena, F., Burdet, E. "Confidence-based Intent Prediction for Teleoperation in Bimanual Robotic Suturing". Under review in: IEEE Robotics and Automation Letters.
- **Hu, Z. J.**, Ranne, A., Abdelaal, A. E., Bhattacharyya, K., Burdet, E., Okamura, A. M., Rodriguez y Baena, F. "How are my stitches? Model selection and real-time skill assessment for robotic suturing". Under review in: IEEE Robotics and Automation Letters.
- Xu, H., **Hu, Z. J.**, Liu, Y., Brudfors, M., Kutter, O., Giannarou, S. "AI-assisted Robot Control with Markerless Visual Guidance". Under review in: IEEE International Conference on Intelligent Robots and Systems (IROS) (2025).

1.5 Technical and Clinical Acknowledgements

Chapter 2. Longrui Chen helped with the development of the algorithm and user study evaluation with participants.

Chapter 3. Giulio Anichini provided inputs and assistance for suturing techniques. Anton Deguet offered support in debugging and fixing of the da Vinci hardware.

The da Vinci platform and hand interfaces were provided by the Hamlyn Centre at Imperial College London.

The formatting of this thesis has been used from the template of Rob Robinson [128] and the link to the original template can be found here: <https://github.com/mlnotebook>.

1.6 Thesis Overview

Chapter 2 provides the background on technologies used to support surgeons in robot-assisted surgery, beginning with an overview of the da Vinci robot and other similar platforms currently available in the market. It discusses the strengths of these systems while also addressing their limitations. The chapter then introduces the concept of human-robot cooperation, emphasising the spectrum of teleoperative controllers, i.e. shared control, supervisory control, and assistive control, which are the focus of Chapters 3, 4, and 5, respectively. These different types of control mechanisms offer varying levels of autonomy to the robot, allowing for more flexible interaction between the human and robotic systems. Finally, the Chapter explores the application of Artificial Intelligence (AI) in automating robotic surgery and performing skill assessments, highlighting how

AI can enhance precision, reduce human error, and provide valuable feedback on surgical performance.

Chapter 3 describes the development of a human-robot cooperative framework, i.e. shared control, focusing on trajectory and strategy learning during a bimanual peg transfer. The chapter introduces a segmentation approach where control is alternated between the human operator and an autonomous agent, assigning different parts of the task based on their complexity and dexterity requirements. It presents the use of a modified SurRoL environment to train and test agents within the same platform, allowing seamless transitions from simulation to real-world robotic platforms like the dVRK. The training approach is based on an Epsilon-Greedy Maximum Entropy Inverse Reinforcement Learning (EG MaxEnt IRL) algorithm, designed to learn both trajectories and strategies from expert demonstrations while promoting robustness through enhanced exploration. The chapter details the setup for shared control, outlines the structure of the learning environment, and discusses the experimental protocols and user studies conducted to validate system with both a surgical simulator and the physical robot.

Chapter 4 presents an investigation into human-robot control in surgery, focusing on how a continuous collaborative strategy (unlike the previous segmented method) that can enhance surgical performance. It introduces a reinforcement learning-based approach called transverse generative adversarial imitation learning (tGAIL) to train robotic agents from expert demonstrations, aiming for improved learning consistency and robustness. The Chapter then proposes a priority-changing control method, where the control authority dynamically shifts between the human operator and the robot based on task context, specifically the deviation from the planned trajectory. Different control strategies, including continuous supervision, segmented control, and manual control, are described and implemented both in simulation and on a real dVRK system. A structured user study with novices is carried out to evaluate how these different control modes influence performance, workload, and interaction dynamics during a standardised peg transfer task.

Chapter 5 focuses in assisted control and presents a confidence-based control framework for bimanual robotic suturing using the dVRK. The system integrates high-level gesture recognition with low-level confidence-driven intention assimilation, allowing the robot to adaptively support the user's motion plans. At a high-level, a Transformer model is trained to classify surgemes in real-time, while at the low-level a Bayesian inference is used to estimate robot's confidence on its kinematic and visual data. This confidence informs a controller that blends user and robot trajectories. Two user studies demonstrate significant improvements in task performance compared to traditional teleoperation, validating the effectiveness of the proposed approach.

Chapter 6 investigates strategies to improve real-time surgical skill assessment by predicting OS-ATS scores from multimodal data. The work uses vision and kinematic data from the JIGSAWS dataset to predict continuous OSATS scores during surgical tasks. Seven unimodal and fusion deep learning models were developed, with the fusion architecture outperforming the baselines in

terms of alignment with ground-truth, as measured by Spearman's correlation coefficient. The results also reveal that models trained on high-skill demonstrations generalise better across different users, while those trained on low-skill data tend to underperform. This highlights the importance of expert-quality data in training robust assessment systems. This study contributes a data-driven approach that shifts surgical training from post-hoc evaluations to real-time feedback.

Chapter 7 summarises the contributions of the thesis, describes its limitations, and offers perspectives on future work directions to advance human-robot control frameworks in robot-assisted surgical systems.

Chapter 2

Current technologies for robot-assisted surgery

Robotics is a fundamental pillar of modern healthcare. With the advent of minimally invasive surgery as an alternative to traditional open surgery, the surgeon's loss of direct access to the internal anatomy of the patient has led to a greater effort on improving ergonomics and has focused on the integration of assistance and supportive systems in robotic telemanipulators [27]. While the da Vinci surgical system is by far the most popular and successful robot in the field of robot-assisted surgery, the expiration of many of its patents has opened the market to new surgical platforms, such as Senhance Surgical Robotic System, Revo-I, Versius, Aveta, Hinotori. This chapter presents the current state-of-the-art technologies used in teleoperative robot-assisted surgery, providing technical reviews and current challenges, following with an extensive analysis of the methods used to support the surgeons.

2.1 Introduction

When the first da Vinci surgical telemanipulator prototype was developed in the 90s, it quickly revealed its potential to “provide a technical solution to the intrinsic limitations of manual laparoscopic surgery” [10, 40]. By 2020, more than 5,000 da Vinci surgical systems were deployed worldwide, performing more than 7 million surgical procedures [27]. Robot-assisted surgery has revolutionised minimally invasive procedures by providing enhanced precision, dexterity, and visualisation compared to traditional laparoscopic techniques. Robotic arms offer a greater range of motion, reducing hand vibration and ergonomic strain on surgeons. Additionally, high-definition 3D imaging and motion scaling enable more immersive and complex procedures, which would be challenging with manual techniques. Then in 2024, Intuitive Surgical presented their latest generation of surgical robot: the da Vinci 5. The most noticeable improvement of model 5 is the inclusion of haptic feedback technology cleared by the FDA in the field of robotic surgery.

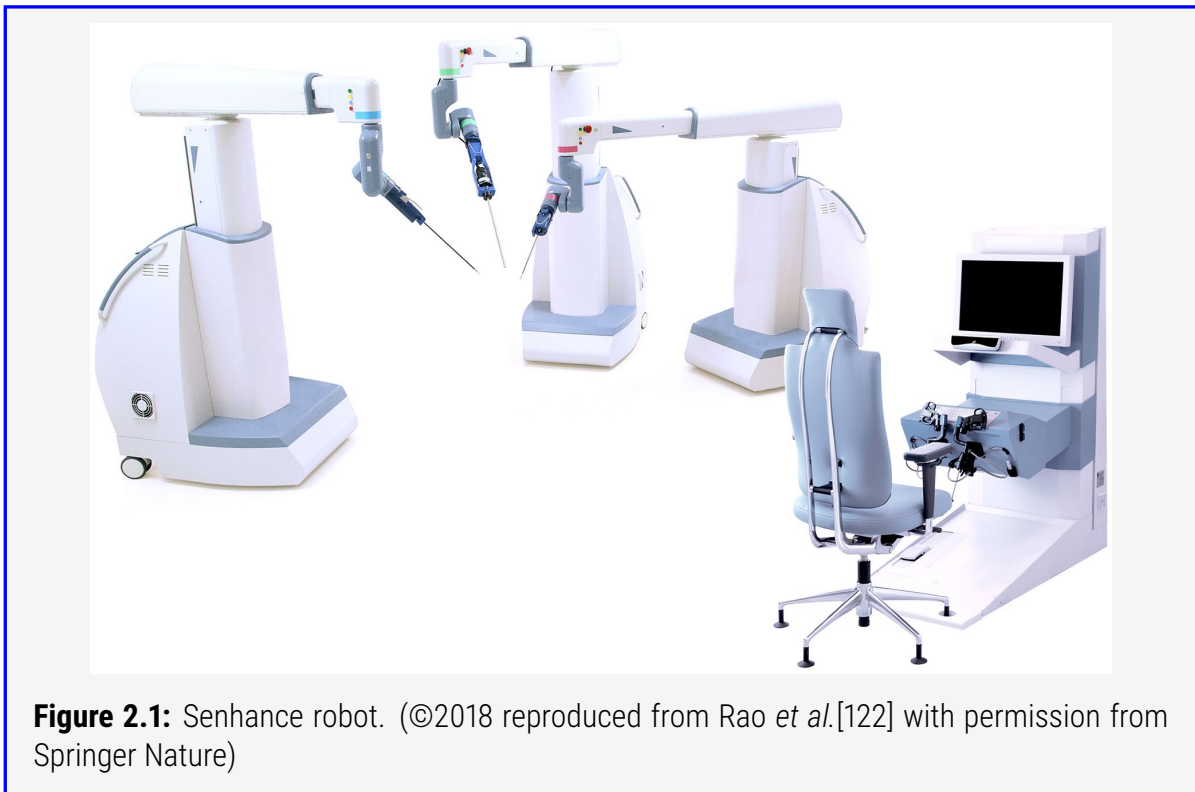
This growth in robot-assisted surgery field has also led to a large surgical robotics research community involving academic and industry collaborations. Part of this collaboration has culminated in the formation of the da Vinci Research Kit (dVRK) [76], a research platform created by repurposing decommissioned da Vinci surgical systems, which allows many researchers across more than 40 different institutions worldwide, one of which is Imperial College London, to keep innovating and addressing challenges in robot assisted surgery.

Much of the research work and scientific publications regarding the da Vinci surgical system is based on the dVRK [27]. The platform is based on leader-follower structure for the implementation of teleoperation. The surgeon can sit on a console and uses the Master Tool Manipulators (MTMs) to remotely control the Patient Side Manipulators (PSMs) composed of four robotic arms: three holding surgical instruments and one with an endoscopic camera. As haptic feedback was introduced only in the most recent version, the earlier da Vinci systems (including the dVRK) provide most of the surgical procedure’s information to the surgeon through visual feedback using a 3D endoscope positioned on the remote patient side. To control the tip of each instrument, the PSMs provide seven Degrees of Freedom (DOF) (three orientations, three translations, and grip), which are typically lost in conventional laparoscopy [40]. Effectively, the robotic platform allows the surgeons to regain 3D vision within the patients anatomy and regain the dexterity typically used in open surgery.

Since da Vinci surgical system first FDA approval in July of 2000, the robot has been the gold standard for robot-assisted minimally invasive surgery and dominated the field for almost 20 years. With the expiration of some da Vinci patents in 2019, many other competitors have entered the market with their robotic platform, allowing to stretch the innovation in healthcare robotics [82].

The Senhance surgical system (TransEnterix Surgical Inc, Morrisville, NC, USA) [5] was the first new multiport robotic system to receive FDA clearance in October 2017 after the success of the

da Vinci surgical system [82]. The Senhance platform focuses on applications for general surgery and gynecology procedures and differs from the da Vinci in three main aspects. Firstly, the system is composed of four independent and separate robotic arms that sit on their own carts. Secondly, the control of the camera is enabled through an infrared eye-tracking system on an open console [122]. Lastly, the hand interface, which is located at the console where the surgeons sit, is able to provide haptic feedback for an enhanced sensory feedback during procedures.



Another surgical robot is the REVO-I platform (Meere Company, South Korea), which is another follower-leader system that received the Korean regulatory approval in 2017 for operating in Korea [77]. Overall, the robot is similar to the da Vinci surgical system, as it also comprises a closed surgeon console, four manipulator arms on a single cart, and a vision cart [122]. The main difference from the da Vinci surgical system was the use of haptic feedback in the REVO-I, which now is incorporated in the da Vinci 5.

One further robotic platform that is worth noting is the Versius surgical system (Cambridge Medical Robotics Ltd, UK), which received the European CE Mark in March 2019 [82]. The system consists of the surgeon open console and individual robotic arms mounted on separate portable carts. Notable features include the surgeon interface, which requires polarised glasses for 3D vision and that allows to control the robotic arms through joystick handles and receive haptic feedback.

While many other robotic systems for robot-assisted surgery are currently commercially available, several challenges need to be addressed for the continuous development of these innovations. The next section will detail the specifics of these challenges.

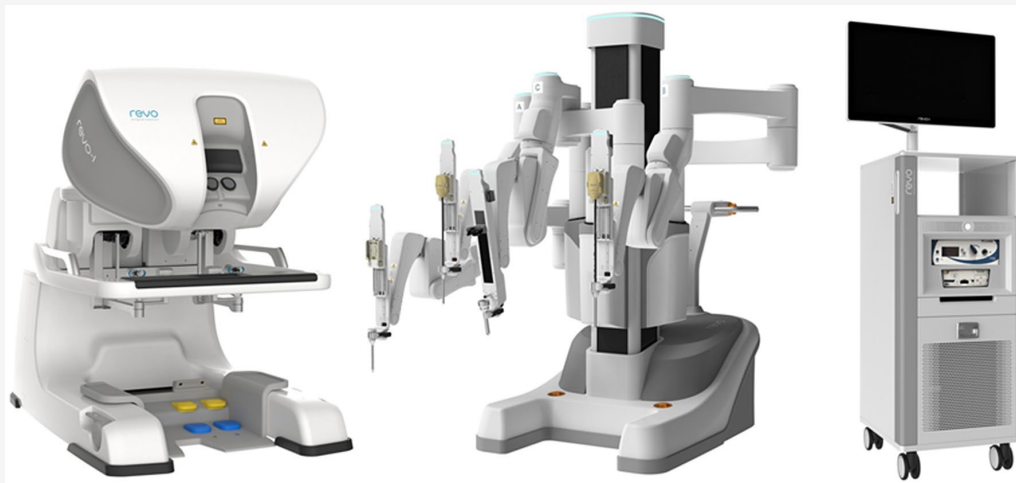


Figure 2.2: REVO-I robot. (©2018 reproduced from Rao *et al.*[122] with permission from Springer Nature)

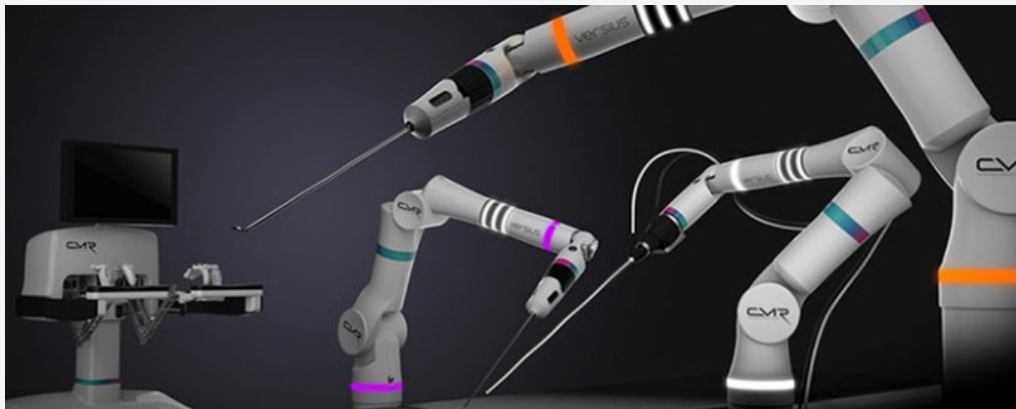


Figure 2.3: Versius robot. (©2018 reproduced from Peters *et al.*[116] with permission from Springer Nature)

2.2 Challenges and Limitations in Robotic-Assisted Surgery

The core of robotic-assisted surgery is the combination of human skill and robotic precision. Robot-assisted surgical systems are designed to augment a surgeon's capabilities, offering enhanced control over surgical tools and increased precision. While these systems have brought significant benefits to the field of healthcare engineering, they also came with a range of challenges and limitations that impact its broader adoption and utility. These challenges can be categorised into technical, financial, training-related, and ongoing advancements of the system.

From a technical perspective, robotic-assisted systems can occasionally malfunction, leading to delays or complications during surgery. These failures are well-documented in the literature, partic-

ularly in relation to the da Vinci surgical system [15, 75, 80]. A study by Borden et al. [15] highlights that, although rare, most of the reported malfunctions involve mechanical failures of various components, whose etiology included: set-up joint malfunction, arm malfunction, power error, monocular monitor loss, camera malfunction, breaking of surgeon's console hand piece, and software incompatibility. Kim et al. [80] reported some of these failures and their solutions as shown in Table 2.1. It is worth noting that the most common way to resolve the failure is the total or partial replacement of the system. Nayyar and Gupta [103] highlight that most of the malfunctions can be recognised before starting the surgery, underscoring the importance of preliminary checks to ensure the correct functioning of the robotic system before anesthesia.

Table 2.1: Failures resulting in cessation of robotic surgery reported in Kim et al. [80]

TYPE	NUMBER	MANAGEMENT
On/off failure	1	No use and change OR
Mechanical failure		
Console failure	5	1 Open conversion 1 Laparoscopic conversion 3 Repair by ourselves or intuitive staff
Arm failure	6	1 Laparoscopic conversion 5 Repair by ourselves or intuitive staff
Camera abnormal	2	Jack readjust
System error		
Out of range motion	10	Reset or repair by intuitive staff
Instrument failure		
Resano Forceps	1	Exchange for a new one
Curved Scissors	3	Exchange for a new one
Large needle driver	1	Exchange for a new one
Bipolar Maryland Forceps	8	Exchange for a new one
Permanent Cautery Hook	1	Exchange for a new one
Permanent Cautery Spatula	2	Exchange for a new one
PreCise Bipolar Forceps	1	Exchange for a new one
ProGrasp Forceps	2	Exchange for a new one

An additional challenge associated with robotic-assisted surgery is the financial commitment of acquiring and maintaining the equipment [102, 155]. Robotic surgical systems are often expensive and they require ongoing expenses for maintenance, software updates, and consumable instruments. These costs can strain healthcare budgets, particularly in smaller hospitals or in countries with limited healthcare funding. It is shown that generally robot-assistive surgery is more expensive than traditional laparoscopic procedures, due to higher initial costs and longer procedures, however Prewitt et al. [119] shows that utilising robots like the da Vinci surgical system can lead to

better postoperative patient outcomes and shorter recovery time, which increases the availability of hospital beds. Furthermore, Huettner et al. [65] observes that the acquisition of laparoscopic operating skills seems to enhance in trainees by practice in robotic systems, which would then justify the initial high investment [155]. Accelerated proficiency and performance improvement through these robotic platforms is extensively studied and debated in the literature [49, 55, 74], which can affect experienced and inexperienced surgeons differently.

Therefore, the learning curve associated with robotic-assisted surgery is another important challenge. Although the technology is designed to augment a surgeon's capabilities, mastering robotic systems requires specialised training and practice. Unlike traditional open laparoscopic surgery, where surgeons rely on direct haptic feedback, robotic platforms require adaptation to a visually dominant interface. This transition can be particularly challenging for experienced surgeons accustomed to conventional techniques [74, 95]. Surgeons must learn how to control robotic arms, navigate three-dimensional interfaces, and manipulate instruments in new ways. From published literature, it is unclear if the learning curve with robot-assisted surgery is faster or slower than conventional laparoscopy [145]. Generally, it is found that novice surgeons learn basic skills and perform tasks faster and with higher precision when using the robot than traditional laparoscopy [29, 64, 133], however a study from Heemskerk et al. [55] reported faster training during the latter. Moreover, the learning curve can vary significantly depending on the type of procedure being performed, the surgeon's prior experience, and the availability of structured training programs such as simulation-based learning. However, institutions with dedicated robotic surgery curricula apply their training programmes to all surgeons, regardless of their open, laparoscopic, or robotic surgical experience before actual independent practice with the robotic system [144]. To support the surgeons' training, the curricula often include simulation platforms that have shown to effectively promote robotic skill acquisition among residents, fellows, and attendings as well [134] (Figure 2.4).

Another critical challenge in robotic-assisted surgery relates to the effective incorporation of state-of-the-art technologies into the systems, particularly in terms of further supporting and assisting surgeons. As surgical robotic platforms continue to evolve, researchers are exploring ways to incorporate more intelligent assistance features, such as machine learning-based guidance [62], phase tracking classifiers [7, 156], and objective performance indicators [106]. These advancements aim to reduce the cognitive load and improve surgical efficiency. The current state of the technology in robotic-assisted surgery focuses in the development of passive tools that require constant surgeon input, such as the leader-follower structure, meaning that they rely entirely on the surgeon's input for every movement and decision during an operation. The robot's role is to replicate and enhance the surgeon's hand movements with greater precision and stability, but it does not make independent decisions or provide active guidance. The surgeon must control every aspect of the procedure, from tool placement to tissue manipulation, through direct input via a console. This reliance on full manual control contributes to fatigue during long procedures, highlighting the need

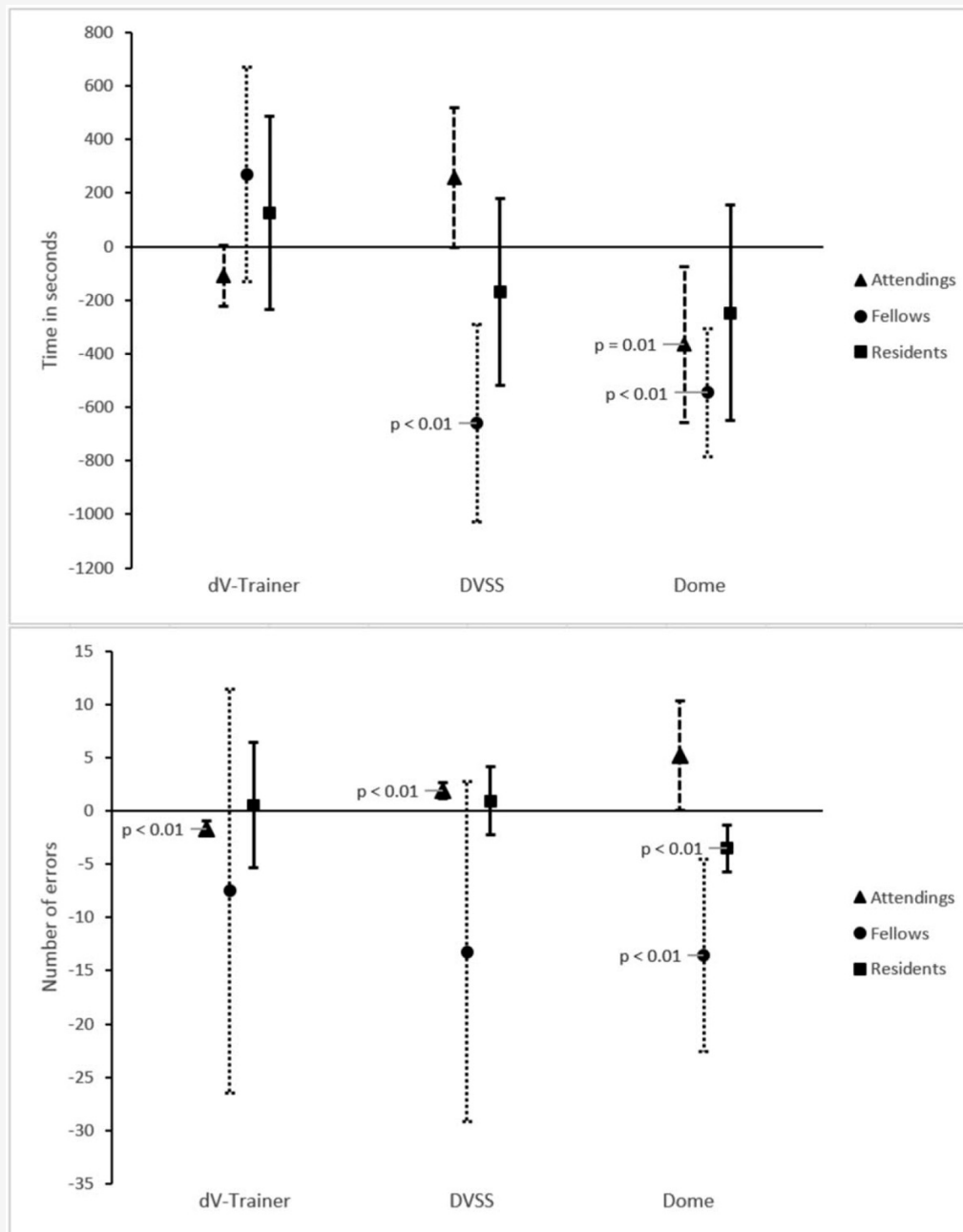


Figure 2.4: Group performance comparison for surgical robotic skill acquisition with control group as reference (task time and errors) on three surgical platforms: dV-Trainer, da Vinci Simulation System (DVSS), and Dome [134]. Symbols represent estimated time (top) and error (bottom) difference from control group in seconds in all tasks adjusted for baseline differences among groups. The control group was offered the Intuitive da Vinci curriculum. Error bars represent standard error of the estimates. Control group performance has been set as 0 in the Y-axis. P-values refer to comparison of each group's performance with control on the same tasks. (©2020 reproduced from Satava *et al.* [134] with permission from Wolters Kluwer Health, Inc.).

for support systems that can anticipate surgeon intentions and assist with repetitive or complex tasks. The traditional teleoperation approach limits the system's potential to offer real-time as-

sistance, such as suggesting optimal surgical paths or autonomously handling routine surgical tasks. Recent studies have explored the feasibility of semi-autonomous robotic functionalities, where AI-driven systems could provide adaptive assistance and feedback without removing surgeon oversight. Such developments could revolutionise robotic surgery by enabling more efficient and safer procedures. The work on this thesis will mainly focus on this aspect, which is extensively detailed in the following section.

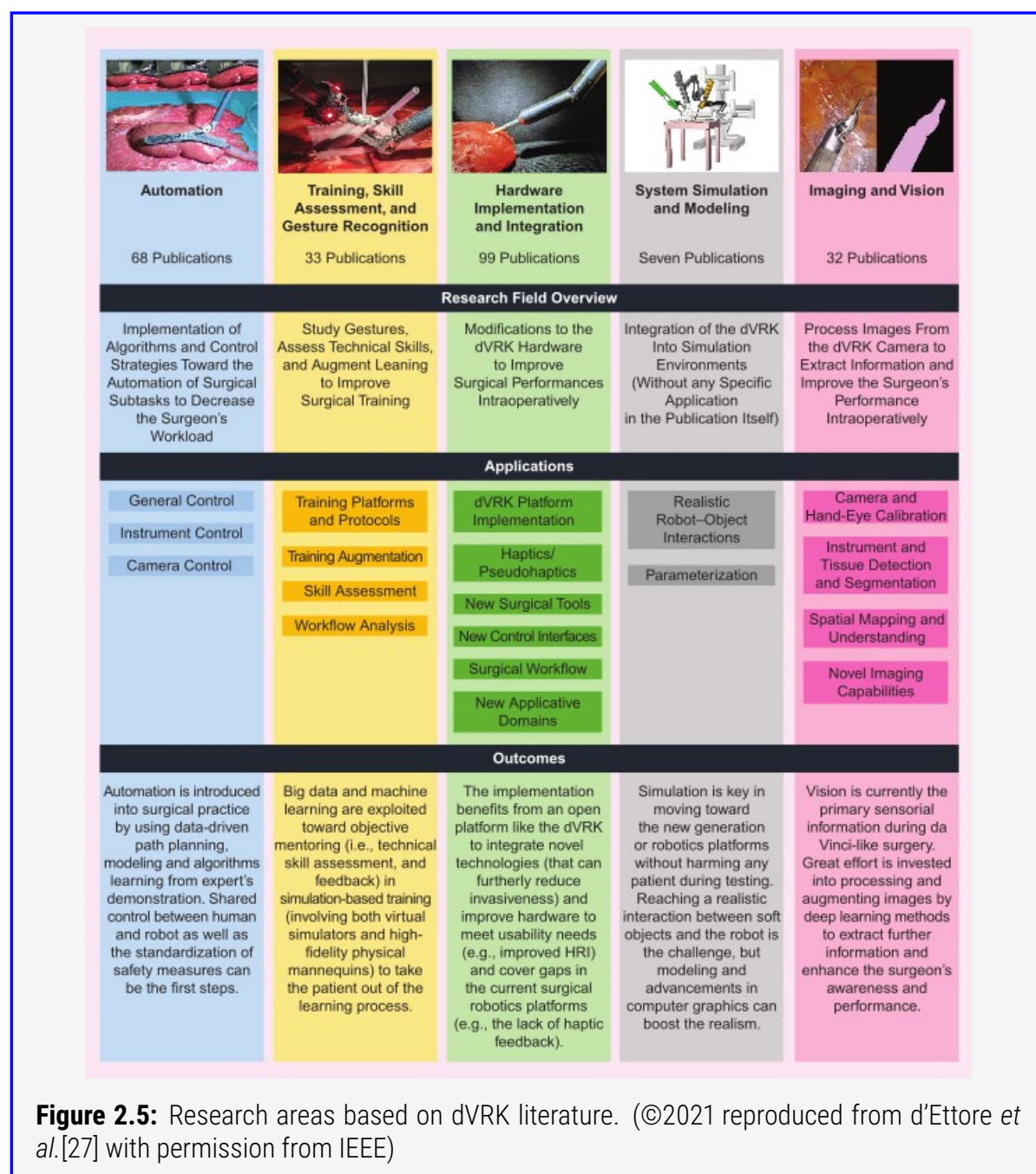
2.3 Support during Robot-Assisted Surgery

While current robotic platforms like the da Vinci surgical system have made significant steps in improving surgical precision and ergonomics, there is much potential for growth in terms of functionality, automation, and intelligence. As robotic systems continue to advance, integrating higher levels of autonomy and decision support could enhance surgical outcomes, reduce human error, and optimise procedural efficiency. Researchers are actively exploring ways to incorporate these technologies into these systems. D'Ettore et al. [27] has highlighted several research fields in the context of teleoperative robot-assisted surgery that can be categorised as: automation; training, skill assessment and gesture recognition; hardware implementation and integration; system simulation and modeling; and imaging and vision (Figure 2.5). Among these research fields, automation has seen increasing interest, particularly in developing robotic assistance features such as automated surgery, phase recognition, and task assistance to improve surgical outcomes. By using machine learning, it is possible to identify patterns in surgeon behavior, assess skill levels objectively, and refine robotic assistance to better adapt to the surgeon's needs. This approach not only enhances training and evaluation methods but also lays the foundation for future developments in semi-autonomous surgical systems.

This thesis will have a deeper focus in the role of automation and skill assessment through the use of control strategies and machine learning models, with the aim to reducing surgeon's perceived workload during surgical tasks.

2.3.1 Human-Robot Cooperation

While fully autonomous surgery remains an extremely complex challenge, there is potential for robotic systems to engage more actively in routine tasks during surgery, such as suturing or cutting tissues, through other supportive technologies. Recent advancements in motion planning, real-time gesture classifiers, and adaptive control strategies have demonstrated promising results in automating certain surgical subtasks with high precision [7, 79, 151]. This could reduce the cognitive and physical workload on surgeons during complex or lengthy procedures, allowing them to focus on more critical aspects of the operation. However, the introduction of automation involves technical hurdles as well. Ensuring an effective collaboration between the surgeon and



the robotic system requires intuitive human-machine interfaces, reliable safety mechanisms, and seamless adaptation to the surgeon's workflow. Surgeons must maintain a comfortable control over the procedure, and the transition from fully manual to semi-autonomous systems requires careful consideration of when and how automation is appropriate.

Figure 2.6 presents the key control methods of surgical tools in the spectrum of performing laparoscopic surgery. At the ends of this spectrum sit traditional manual laparoscopic surgery, where no robotic platform is used, and fully autonomous robotic surgery, where the human side is not needed. As the gap between these two milestones is very large, the literature is then rich with controls strategies to enable effective human-robot interactions. Among these there is human-

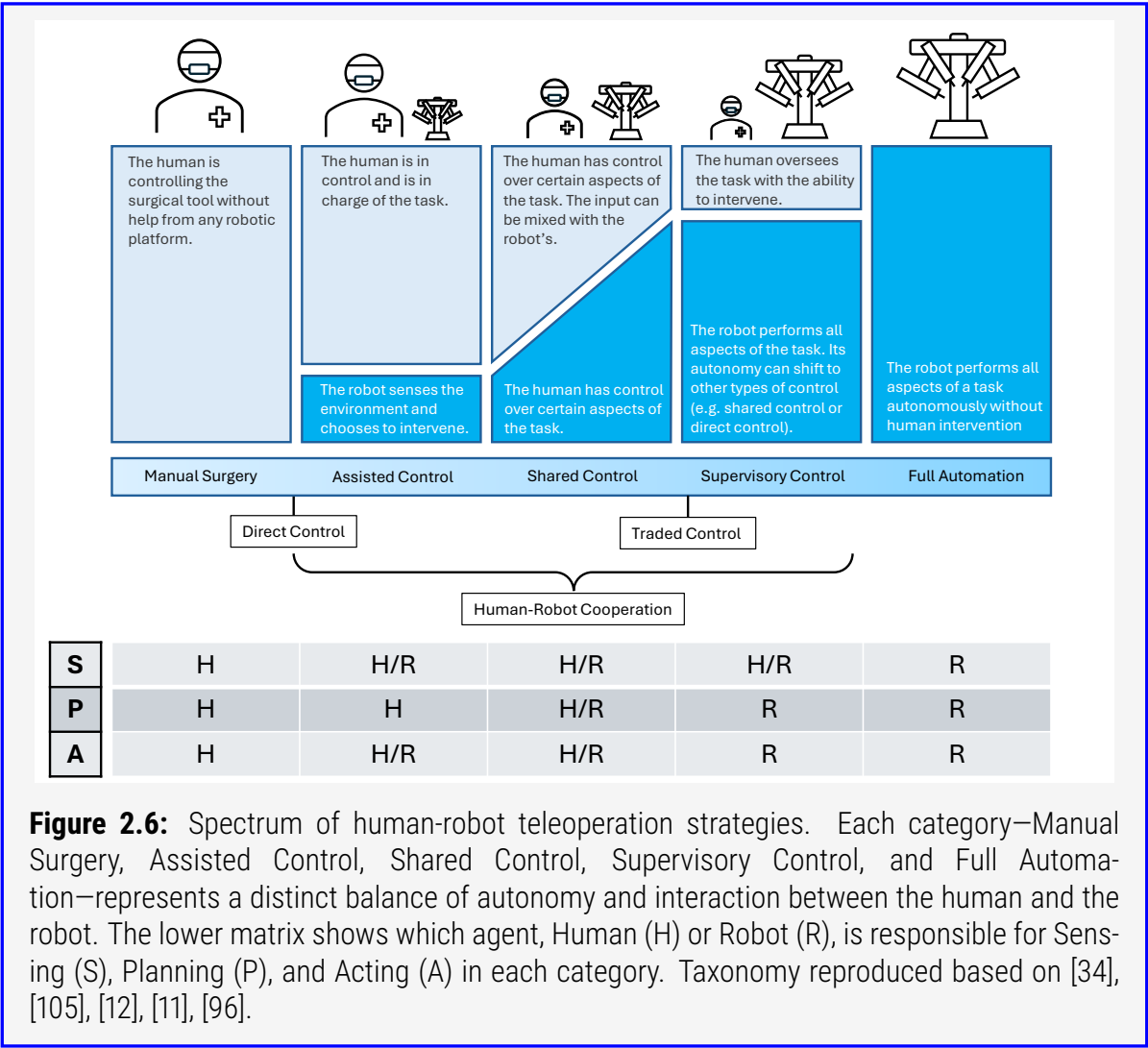


Figure 2.6: Spectrum of human-robot teleoperation strategies. Each category—Manual Surgery, Assisted Control, Shared Control, Supervisory Control, and Full Automation—represents a distinct balance of autonomy and interaction between the human and the robot. The lower matrix shows which agent, Human (H) or Robot (R), is responsible for Sensing (S), Planning (P), and Acting (A) in each category. Taxonomy reproduced based on [34], [105], [12], [11], [96].

machine cooperation, which refers to a control paradigm where the human and agent combine, alternate, share, or mix their control input of the physical robot [1, 37, 96]. Abbink et al. [1] identifies several challenges with this type of control, which have persisted to this day: trust, authority switching, skills degradation and situation awareness, excessive dependence, and inaccurate estimation of the robot reliability.

Figure 2.6 summarises the control methods used in teleoperative robotic surgery based on the commonly adopted techniques in the literature for laparoscopic surgery. While Manual Surgery and Full Automation represent the ends of this spectrum, it is possible to find three large fields in between: Assisted, Shared, and Supervisory Control.

Assisted control in teleoperative surgical robotics involves a human surgeon which maintains primary control over the procedure while the robotic system provides real-time intervention based on environmental sensing [99]. Unlike direct control, which simply scales and transfers the surgeon’s input from the interface to the manipulators, assisted control is designed to augment the surgeon’s dexterity with external inputs. The robotic system can continuously monitor intraoperative

conditions, such as tool position and force feedback [24], to detect potential dangers or errors and intervene if necessary by adjusting motion trajectories and restricting unsafe movements. A key advantage of assisted control lies in its ability to enhance surgical precision and reduce cognitive and physical fatigue with minimal inputs from the robotic system, which intervenes only when necessary. By incorporating real-time sensor data and intelligent algorithms, these systems can filter out physiological tremors, enforce virtual constraints, and provide motion scaling to refine complex maneuvers. As a result, robotic systems with assisted control mechanisms are increasingly explored in skill augmentation and are often integrated in other types of teleoperative controls. One of the fundamental challenges in developing effective assisted control strategies is ensuring seamless human-robot interaction. The system must be able to distinguish between intended movements and unintentional errors, requiring advanced sensor fusion techniques and machine learning models to interpret the surgeon's actions accurately. Achieving an optimal balance between autonomy and control remains a critical design consideration, as excessive intervention could lead to user frustration, whereas insufficient assistance may fail to prevent adverse events. The importance of incorporating varying levels of robotic assistance is crucial to addressing these challenges and is explored in shared controlled frameworks.

Within telerobotics, Niemeyer [105] defines shared control as

"... an architecture where some degree of autonomy or automated help is available to assist the user."

Instead, Abbink et al. [1], defines shared control as when

"Human(s) and robot(s) are interacting congruently in a perception-action cycle to perform a dynamic task that either the human or the robot could execute individually under ideal circumstances."

while Payne et al [112] defines it as

"A user-interface paradigm in surgical robotics by which a surgeon and robot share control of a surgical tool, so as to enhance surgical precision and allow the surgeon to retain executive control of the robot."

These definitions exclude systems that are fully autonomous or fully manually controlled, and refer to concepts of teleoperation and cooperative manipulation [1] with constant physical interaction. In teleoperation systems, shared control aims to combine the reliability and telepresence achievable during fully manual control with the efficacy and possible safety guarantees of autonomous control

[26, 53, 105]. The benefits of this combinations are well described by Abbott et al. [3]:

"A straight line drawn by a human with the help of a ruler is drawn faster and straighter than a line drawn freehand. Similarly, a robot can apply forces or positions to a human operator to help him or her draw a straight line."

Therefore, in shared control the surgeon and the robotic system collaborate on the same task at the same time. The robot can assist the surgeon by taking over certain low-level tasks while the human focuses on higher-level decision-making. This partnership allows for a more fluid and efficient interaction, reducing surgeon workload while maintaining safety and precision [34]. For example, during tasks like suturing or cutting, the robot could automatically stabilise the instrument or guide it along a predefined path, allowing the surgeon to focus on complex maneuvers [130]. Research in shared control in teleoperative robot-assisted surgery involves refining the human-robot interaction and improving the system's ability to predict, adapt, and respond to surgeon intent, with the aim of reducing cognitive load and improving performance during complex surgical procedures.

In the context of shared control in teleoperation surgical systems, numerous methods have been investigated. In [101], the shared control paradigm compensates for physiological motions during teleoperated cardiac surgery, which allows the surgeon, or leader, to operate on a static image of the heart while the remote manipulator, or follower, tracks and moves with the physiological motion. Abbink et al [1, 2] categorises this approach as a form of "mixed-input" shared control, since it involves a concurrent control but the activity is not continuously communicated to the human operator. Another shared control method, which overlaps with assisted control, involves the use of virtual fixtures to prevent the human to enter forbidden regions (passive assistance) or provide guidance through a path (active assistance) [1, 129]. The RIO Robotic Arm Interactive Orthopedic System (MAKO Surgical Corp., USA) [47] is an example of surgical system that employs virtual fixtures to provide assistance to the surgeon during the milling process. One further popular controller is the traded control, also called ON-OFF control or segmented control, which is part of the shared control category and has been a widespread approach for human-robot interaction, especially in scenarios where humans can serve as a backup system for robotic agents [1]. In this type of controller, the human and robot agent might work on the same task but alternate their control of the tool, therefore this type of controller represent an extreme case of shared control.

Lastly, supervisory control consists in a human operator that oversees the task execution while the robotic system performs all aspects of the procedure autonomously. This model of control takes advantage of the robot's precision and consistency while maintaining a safety net through human intervention capabilities. Unlike shared control, where both the human and robot contribute to task execution simultaneously, supervisory control allows the robot to operate independently under human oversight. The level of autonomy in this framework can vary dynamically, shifting, for example,

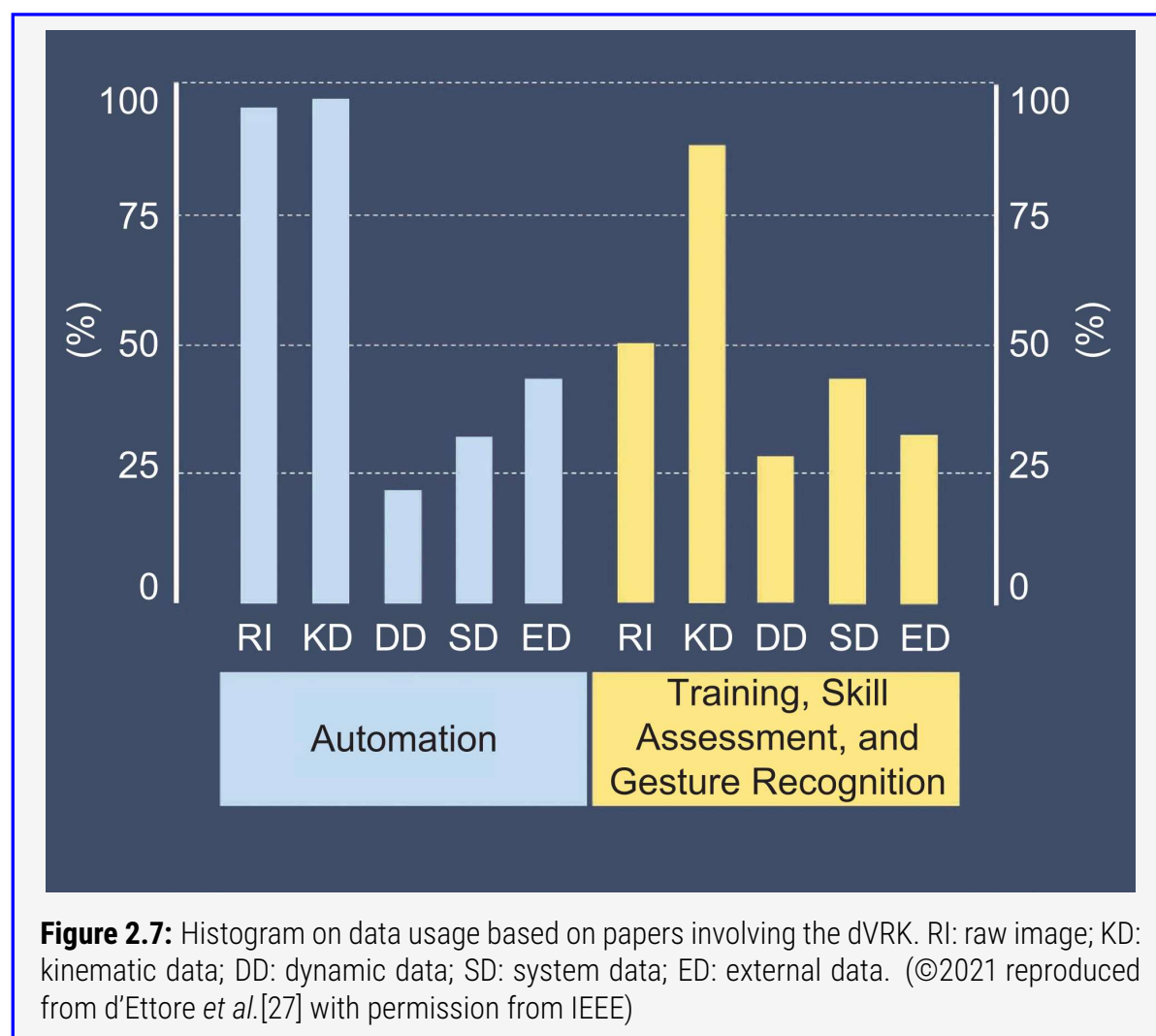
towards shared or direct control when necessary, depending on task complexity, surgical context, or unexpected intraoperative conditions. Noticeably, in supervisory control, the system can independently perform predefined tasks, such as suturing or tissue manipulation. Furthermore, incorporating real-time monitoring tools, such as computer vision-based tracking [141], enhances the operator's situational awareness, facilitating timely interventions when necessary. However, the effectiveness of supervisory control is contingent on robust sensory information, adaptive planning algorithms, and fail-safe mechanisms that ensure safe and effective execution. A critical advantage of supervisory control is its potential to reduce cognitive and physical workload for the surgeon while improving procedural efficiency. By delegating task execution to the robotic system, the human operator can focus on higher-level decision-making, such as assessing task needs or planning subsequent steps. However, challenges remain in ensuring the reliability of autonomous task execution, particularly in dynamic and unstructured surgical environments can affect robotic performance.

2.3.2 Artificial Intelligence

The use of Artificial Intelligence (AI) and machine learning models plays a crucial role in advancing healthcare engineering. In particular, robotic-assisted surgery is a field that can generate large datasets, which include event, video and kinematic data, that can be processed with machine learning architectures and used to push research in the areas of automation and skill assessment [100]. Figure 2.7 illustrate the most commonly used data for these two fields based on percentage of papers in the literature, highlighting the popularity of kinematic data, often due to its easy accessibility.

In automation, AI algorithms are being developed to analyse large amounts of surgical data, recognise patterns, generate models, and provide real-time feedback. D'Ettore et al. [27] describes how AI and machine learning is one of the most popular topics in dVRK research for improving surgical workflow and optimise task performance, with an enormous potential to enhance surgical precision, as AI can assist in identifying surgical procedures, generating path trajectories, and recommending optimal instrument movements.

As vision is the primary sensory information in teleoperative robot-assisted surgery, popular AI applications concern image processing solutions for detection of instruments, tissues, and spatial mapping [27, 142]. Several uses of AI in a supportive or complementary role also exist, which use its ability in creating paths and finding a solution to cooperate with a human operator in human-robot cooperation fashion [22, 25, 72, 158, 168]. As illustrated in Figure 2.6, automation is integrated in the human-robot cooperation framework by affecting how the system senses, plans, or acts. As the level of automation increases, the system moves from a low automation level to a higher one. One of the most common approaches for implementing human-robot cooperation in robotic-assisted surgery is using motion planning and trajectory optimisation, where AI-driven



models assist in generating smooth, collision-free movements for the robotic system. These automated planning techniques help improve precision, efficiency, and safety during procedures by minimising unnecessary movements. For instance, learning-based approaches such as imitation learning have been explored to enhance trajectory generation by using expert demonstrations and reward-based optimisation [117]. Additionally, reinforcement learning has also shown promise in generating and navigating paths that fit into the human workflow. Varier *et al.* [158] showed this by segmenting the suturing task into steps and using inverse reinforcement learning to allow the robot to learn the human's trajectories, and then divide the task's responsibilities between human and robot. Several simulator platforms have also been developed to support the development of reinforcement learning algorithms aimed at surgical robots [159, 165]. More recent innovations include the use of Transformer-based models [170] for the implementation of automated surgical procedures, but their use in human-robot cooperation have not been widely explored yet.

In skill assessment, AI is used to objectively evaluate a surgeon's performance by analysing video footage or kinematic data, and provide real-time or post-operative support. Many popular rating scales are available in the literature, such as the Objective Structured Assessment of Technical

Skills (OSATS) [83, 97], which assess criteria such as tissue handling, flow of operation, and dexterity. Gao et al. [42] describes a modified version used for robotic surgery, which is presented in Table 2.2. This approach allows standard and structured assessment of someone's surgical technical skill.

While post-analysis of surgical procedures with rating scales are often used for clinical studies, they are severely limited as they require an expert clinician to analyse and evaluate each piece of surgical data, leading to costly, time consuming, and biased results. By using artificial intelligence, it is possible to automate this process by training a model to empirically evaluate large volume of the data. Lam et al. [83] shows that, while the most popular methods were Hidden Markov Models and Support Vector Machines, in recent years there has been a shift towards the use of artificial neural networks and deep learning techniques. The authors argue that conventional machine learning methods require multiple stages of processing and feature extraction, and often an expert is necessary to select these component, while the end-to-end framework of deep neural networks can allow to directly capture the optimal features and high-level information to achieve high accuracy.

The most common deep learning architectures used for skill assessment include Deep Neural Networks (DNNs), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). These models have been extensively applied in surgical skill evaluation to analyse kinematic data, video recordings, and tool movement patterns. Jin et al. [70] conducted a study on general surgery, demonstrating that deep learning models can effectively extract tool usage patterns, movement range, and economy of motion as key performance metrics. Their approach involved analysing surgical tool trajectories and hand movements to assess proficiency levels, showing CNN's efficacy for spatial feature extraction, while RNNs helped capture temporal dependencies in motion data. Azari et al. [9] further validated the use of deep learning for skill assessment by confirming the scalability and objectivity of automated rating systems based on the OSATS framework. Their work highlighted how machine learning models could be trained on expert-annotated datasets to provide consistent and unbiased skill evaluations, reducing the variability found in traditional human assessments. Beyond these studies, recent works have explored combining attention mechanisms and Transformer-based models to enhance skill assessment by focusing on important motion features and critical surgical events [143, 170]. Such advancements indicate a growing trend towards using more sophisticated architectures to improve the precision and interpretability of AI-driven skill evaluation in robotic-assisted surgery.

However, the clinical adoption of AI technologies is still hindered by several constrains. Moglia et al. [100] list three main factors that limit the widespread use of AI in healthcare: reliance of AI on data properties, limited validation of datasets, and models interpretability. The authors explain that AI models are highly depended on data properties such as quality, quantity, and variety, leading to cases where conventional statistics perform better [66]. Furthermore, most research

Label No.	Domain	Rating Scale
1	Respect for Tissue	1—Very poor: Frequent or excessive pulling or sawing of tissue 3—Competent: Careful handling of tissue with occasional sawing or pulling 5—Clearly superior: Consistent atraumatic handling of tissue
2	Suture/Needle Handling	1—Awkward and unsure with repeated entanglement and poor knot tying 3—Majority of knots placed correctly with appropriate tension 5—Excellent suture control
3	Time and Motion	1—Made unnecessary moves 3—Efficient time/motion but some unnecessary moves 5—Clear economy of movement and maximum efficiency
4	Flow of Operation	1—Frequently interrupted flow to discuss the next move 3—Demonstrated some forward planning and reasonable procedure progression 5—Obviously planned course of operation with efficient transitions between moves
5	Overall performance	1—Very poor 3—Competent 5—Clearly superior
6	Quality of final product	1—Very poor 3—Competent 5—Clearly superior

Table 2.2: Rating scale for the assessment of surgical skills across different domains. Each domain represents an OSATS score used to evaluate a surgeon's performance. Table reproduced from Gao *et al.* [42].

studies mostly validate their models with their own datasets, making it harder to generalise and validate externally [19]. Lastly, the need of clear and transparent way to explain cause and effect in the system hinders the use of AI for non-experts, which surgeons even refer as unexplainable “black boxes” which impede trust [100, 125]. The authors also remind that the use of AI in clinical applications has challenges regarding privacy and confidentiality of patient’s data.

2.4 Conclusion

This Chapter has reviewed the background literature on robot-assisted surgery and the technologies platforms currently available in the market, with a focus on state-of-the-art robotic platforms such as the da Vinci surgical system, which exemplifies the current capabilities and clinical adoption of robotic teleoperation. The strengths of these systems, including enhanced precision, dexterity, and minimally invasive access, were highlighted to contextualise their widespread use.

However, current surgical robots remain fully manually teleoperated systems that depend entirely on the surgeon’s input, without offering autonomous decision-making or active support. This reliance can lead to operator fatigue and limits the potential for the system to contribute to complex or prolonged tasks. To address these limitations, the literature has introduced a spectrum of control systems ranging from fully manual teleoperation to fully autonomous robotic systems.

It is evident from the literature reviewed that there are three main intermediate forms within the spectrum to support the surgeon: assisted control, shared control (including traded control), and supervised control. These approaches aim to augment rather than replace the surgeon’s capabilities, offering support through stabilisation, motion guidance, or execution of tasks under human supervision. In parallel, artificial intelligence has emerged as a key enabler in these systems, capable of learning from surgical data to generate optimal tool trajectories, recognise surgical phases, and provide real-time skill assessment. This thesis will build upon these concepts to investigate the frameworks within the spectrum of cooperative robotic controls and ultimately design a system framework that can achieve high task performance and improved perceived workload (Chapters 3, 4, and 5). Lastly, the work in this thesis will present the development of real-time skill evaluation in surgical teleoperation that can assist in understanding surgeons’ performance in surgical tasks (Chapter 6).

Chapter 3

Shared Control in Peg Transfer

While the traditional control of surgical robots relies on fully manual teleoperation, human-robot cooperative systems promise to address issues such as workspace constraints and laborious tasks. The first control paradigm we will explore in this thesis is shared control, which can reduce the surgeon's workload and improve the overall surgical performance by supporting the surgeon effort during movements while keeping them in charge of complex control phases. In this Chapter, we propose a segmentation of the bi-manual peg transfer task that alternates manual and autonomous control based on task demands. The authority allocation in this shared control framework considers both the limitation of learning-based methods and the higher dexterity of humans during physical interaction. The motion and strategies are transferred from an expert human to a dVRK [76] using an epsilon-greedy on a maximum entropy inverse reinforcement learning algorithm. The model generated enables to train an intelligent agent that can skillfully collaborate with the human operator during the surgical task. The proposed control framework is verified both on a virtual platform and then on a real dVRK to assess its usability and robustness. The results show that, our method can achieve faster and more consistent peg transfers than compared to traditional teleoperation. An analysis of the participants' effort also reveals a significantly lower perception of the workload.

The content from this Chapter was published in IEEE Robotics and Automation Letters (RA-L) [60], June 2023, and was presented in IEEE International Conference of Robotics and Automation (ICRA), May 2024: Hu, Z. J., Wang, Z., Huang, Y., Rodriguez y Baena, F., Burdet, E. "Towards Human-Robot Collaborative Surgery: Trajectory and Strategy Learning in Peg Transfer".

3.1 Overview

Medical robots such as the da Vinci surgical system [40] offer surgeons increased dexterity and visual feedback during minimally invasive surgery. The study in [64] showed that, compared to traditional manual laparoscopic surgery, the da Vinci allows operators to perform quicker and more efficiently standardised exercises, making complex surgical procedures more accessible for people without surgical experience. However, the operator's performance can decrease over time when carrying out long surgical operations, which can lead to errors [73, 157]. This is exacerbated by repetitive or intensive actions during the procedure. For example, the frequent need to clutch (i.e. pressing the pedal to reposition the hand controllers in an ergonomic position without affecting the robot manipulators) is generally regarded as one source leading to longer completion times which can affect the surgical workflow and operator's cognitive load [28]. Recent work with fully automated agents [68] has also directed the attention towards the use of automation to improve surgical performance. These factors have motivated research in human-robot interaction to allow the surgeon to collaborate with an autonomous agent during the completion of a task [73].

In [72], different collaboration models were presented: supervised, shared, and full autonomy. The study reveals that a shared control, where the human and the robot sequentially alternate their control authority on the manipulators (i.e. traded control), leads to results with higher specificity and minimal effort to the participants. The works on shared control in surgery [110, 118, 158] corroborated the idea of a cooperative shared framework between surgeon and autonomous robot. However, while the segmentation of a shared control, or specifically traded control, has been widely explored, the interchange of the control of the surgical tool between operator and autonomous agent still needs an effective adaptive component to ensure a natural relationship that does not interrupt the surgical workflow [69]. The agent needs to fit in the surgeon's technique and strategy for completing a task with intuitive movements to the human operator and smooth transitions between the operator and agent [63]. For example in the past, methods to automate surgical tasks have been investigated through waypoints strategy [98], image-guided algorithms [81], and learning from demonstration [117]. With these methods the agent learns the trajectory necessary to complete a task, but not the strategy that ensures the natural surgical workflow and ergonomics of the surgeon. Furthermore, it is still problematic to correctly define all the different costs associated with the elements of the surgical environment. Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL), a method used to obtain the utility function of the environment, had previously been applied in [36, 87, 158, 172]. In [158], the authors implement MaxEnt IRL to generate the trajectory of a suturing motion, however without proposing a method to increase the robustness of the policy generated or analysing its effectiveness with human operators.

In this Chapter, we develop a collaborative control framework of the peg transfer task with task segmentation and sequential human-robot cooperation scheme. Using this framework, we implement

an EG MaxEnt IRL to learn both the trajectories and strategies from an expert user, maximising the generalisation of the trajectories with the minimum number of expert demonstrations.

We then perform a user study with ten participants to evaluate the benefits of our method by comparing the bimanual peg transfer task completed by Manual Teleoperation (MT) against Shared Control (SC). To validate and analyse our proposed setup design, learning algorithm and shared control framework, we implement the system to perform a bimanual peg transfer task, which is a fundamental skill of laparoscopic surgery [67, 68]. Several virtual platforms with standardised tasks for robotic surgery are available to execute the testing and validation of an agent, however simulators of the dVRK are designed either to simulate the actual surgical experience [33, 38] or to develop fully autonomous agents for surgical procedures [165]. Therefore, in this work we modify the SurRoL environment [165] to support a real-time simulator, which only requires Python. This modification will enable the development of a framework that integrates interfaces to control and interact with the autonomous agent in the same environment in which it was trained.

The Chapter is organised as follows: Section 3.2 describes the proposed segmentation of the bimanual peg transfer and the framework used for shared control. Section 3.3 describes adaptation of the simulated environment and the training algorithm of the agent. Section 3.4 details the experimental procedure on the simulated and real robotic platform, including protocols and evaluation methods. Section 3.5 analyses the results, and Section 3.6 discusses them.

3.2 Working Principle

3.2.1 Shared Control during Peg Transfer Task

Previous works [110, 158] have shown that the surgical motion can be effectively segmented to accommodate a shared control between human operator and agent. To ensure safety and take advantage of the robot dexterity, the segmentation of a task should account for surgical motions that require higher attention of the human operator, such as when in proximity of objects and physical interactions, in contrast to repetitive and laborious spatial relocation of the surgical tools. Physical interactions are particularly complicated as they require the use of an adaptive control to solve the instability of the system. Furthermore, physical interactions require to account for a larger domain of physical dimensions which can be challenging for reinforcement learning training algorithms. Therefore, to complete the peg transfer task, the control authority of the arm controlled by the participant is divided as illustrated in Figure 3.1, where each subtask is appropriately allocated to either the human or the agent skills and capacity. The proposed process begins by triggering the trained agent to autonomously position the surgical hand in proximity of the target block (A). Once the surgical tool is in a strategic position (i.e. the block can be successfully grabbed if approached from above), which is learned during training from expert demonstrations, the human operator is in control again and can continue the movement and easily grab the block (B). The task is followed

by the autonomous motion of the agent that brings the block closer to the other arm (C). The operator then performs a hand-off by physically interacting with the human assistant (D) who places it into the final peg position (E). This segmentation of the bimanual task is controlled by the human operator through a pedal and it allows to preserve the natural coordination between the two arms during hand-off while facilitating the control and workload of the human operator. In this work, we intend to represent the motions that a surgeon would perform alongside an assistant to clear the surgical space when removing debris or objects.

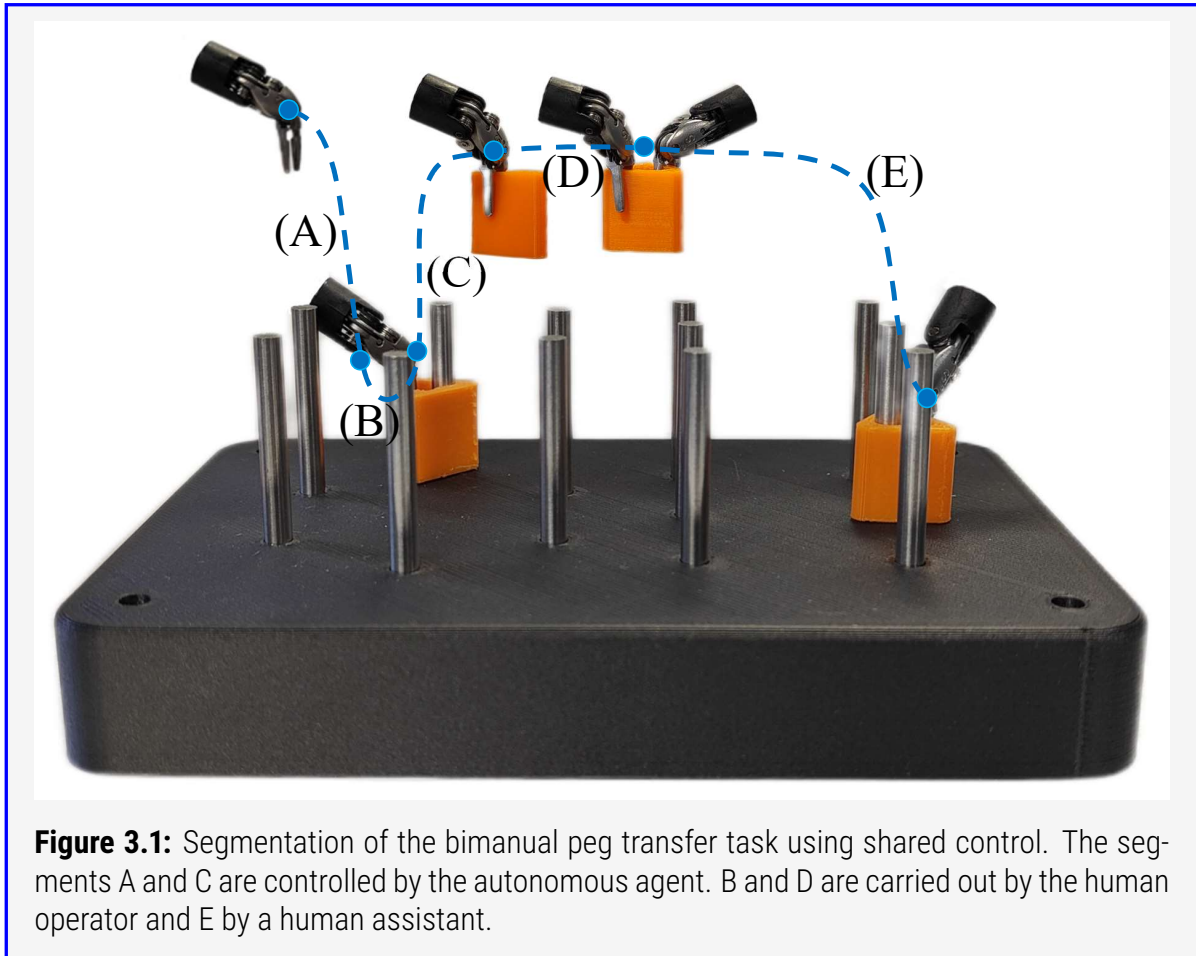


Figure 3.1: Segmentation of the bimanual peg transfer task using shared control. The segments A and C are controlled by the autonomous agent. B and D are carried out by the human operator and E by a human assistant.

3.2.2 Control Framework

The proposed control framework for training and Human-Robot Interaction (HRI) testing of the agent is shown in Figure 3.2. The demonstration of the task is provided by an expert user. In this work, an operator is considered expert if they have more than 300 hours of experience with the hand controllers.

To train the agent, the SurRoL environment is modified to integrate haptic controller interfaces (sigma.7 and omega.7 Force Dimension). The setup developed in Figure 3.2 allows to train the agent in the SurRoL platform and then subsequently quickly test it with a human assistant in the same environment. This allows to rapidly train and test the agent and it permits to assess the

model generated without alteration caused by the transition from sim-to-sim before implementing sim-to-real [114]. The features of the expert demonstrations are extracted and matched against the motions of the agent in the simulation using a gradient-based optimisation, which then updates the reward values of the environment. Once the training is completed, the model of the expert can be tested in either the same surgical simulator or the physical dVRK platform with a human operator by using a shared control framework.

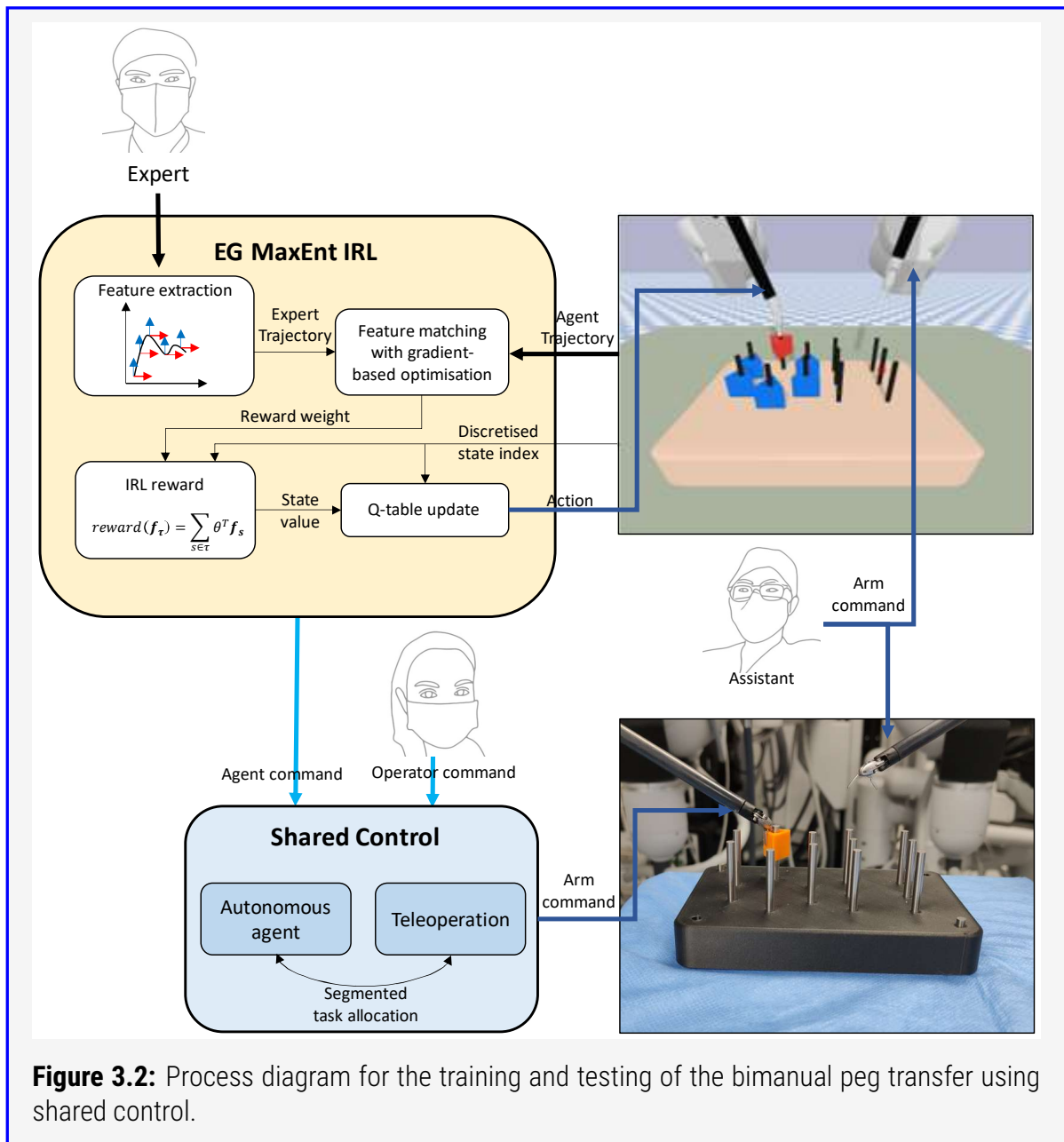


Figure 3.2: Process diagram for the training and testing of the bimanual peg transfer using shared control.

3.3 Autonomous Agent

3.3.1 EG MaxEnt IRL Algorithm

The EG MaxEnt IRL allows to retrieve the reward function of an unknown environment through expert demonstrations. The expert trajectory used in our model is recorded in the same SurRoL environment that will be used for training. The proposed algorithm uses Q-Learning to populate the Q-table with the expected rewards for a specific action at a certain state. To encourage exploration for a more robust policy, we use an epsilon-greedy method [153]. With this modification, the agent will randomly choose an action with ϵ probability instead of the action with maximum reward. Notice that epsilon-greedy is normally used as a tool to find a more optimal solution. However, the proposed EG MaxEnt IRL uses it to update the reward function on a larger state space. This allows the agent to generalise more the trajectory to the goal, making it more robust to states that have not been visited by the expert.

Using EG MaxEnt IRL, we train the agent to generate a reward function of the environment from one expert trajectory. During the training episode, the reward function is calculated according to the MaxEnt reward equation [172]:

$$\text{reward}(\mathbf{f}_\tau) = \sum_{s \in \tau} \theta^T \mathbf{f}_s \quad \theta \leftarrow \theta + \beta \nabla L(\theta) \quad (3.1)$$

where $\mathbf{f}_\tau$ is the feature count of the trajectory τ , $\mathbf{f}_s$ is the state feature count, and θ is the weight vector that is optimised at the end of each training episode using the Lagrangian gradient with β learning rate (Figure 3.2). The value function is then used to populate the Q-table and the policy $\pi(s, a)$ is generated using the Bellman equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} [Q(s', a')] - Q(s, a)] \quad (3.2)$$

where s is the current state, a is the current action, r is the reward value calculated from Equation (3.1), α is the learning rate, γ is the discount factor, and s' and a' is the next state and action. The algorithm implementation is described in Algorithm 1 using the hyperparameters in Table 3.1.

3.3.2 Agent Training and HRI Environment

The SurRoL environment was modified to enable both training and HRI in an unified platform. The models in the simulation consist of the dVRK robotic arms and the peg board with the blocks. Since the EG MaxEnt IRL algorithm requires state dimensions in discretised form, the state space of the SurRoL workspace is divided in a $22 \times 22 \times 22$ grid with 2.5 mm sensitivity in the x , y and z dimensions. The action space is simplified to movements along the three Cartesian dimensions, discretised to a 2.5 mm step per action. Notice that the agent step length coincides with the

Algorithm 1 EG MaxEnt IRL

```

1: Import expert trajectories dataset  $D = [\tau_1, \dots, \tau_T]$  and feature vectors  $f = [f_1, \dots, f_T]^T$ 
2: Initialise random reward weights  $\theta = [\theta_1, \dots, \theta_S]$ 
3: function AGENT TRAINING
4:   for 1 to N do
5:     while Episode not done do
6:       if  $n_{rand} < \epsilon$  then
7:          $a \leftarrow \text{random action}$ 
8:       else
9:          $a \leftarrow \max[Q(s, a)]$ 
10:      end if
11:       $\pi(a | s) \leftarrow Q\text{-Learning}$ 
12:       $\mathbf{f}_\tau \leftarrow \sum_{s \in \tau} \mathbf{f}_s$ 
13:    end while
14:     $\nabla L(\theta) \leftarrow \tilde{\mathbf{f}} - \sum_{s \in S} D_s \mathbf{f}_s$ 
15:     $\theta \leftarrow \theta + \beta \nabla L(\theta)$ 
16:  end for
17:  return EG MaxEnt IRL model
18: end function
19: while Simulation not done do
20:   EG MaxEnt IRL model  $\leftarrow$  state
21:   Update Simulation  $\leftarrow$  action
22: end while

```

Table 3.1: EG MaxEnt IRL hyperparameters.

Hyperparameter	Value
Weight θ learning rate	0.3
Input dimensions	x, y, z
Bellman equation learning rate	0.3
Total number of states	10,648
γ Discount factor	0.99
Action step size	2.5mm
ϵ Epsilon-Greedy	0-0.1

sensitivity, which allows us to assume a deterministic transition probability of the movements. With this approximation, the agent is represented by the dVRK end-effector which moves between the $2.5 \times 2.5 \times 2.5$ mm states. Since the action space for the expert demonstrator is continuous, we discretise its movements in the simulator. This ensures to both preserve the trajectory features and enable its implementation in SurRoL. It is worth noting that algorithms like the MaxEnt IRL method suffer from the curse of dimensionality. This is because the state of the agent along the trajectory in the surgical simulation is uniquely specified by flattening the 3-dimensional coordinates into a 1-dimensional index-array. This process implies that the numbers of possible states in which the robot can be increases exponentially according to $(gridsize)^3$ [158]. Therefore, the number of

states created in the agent training is 10,648.

3.4 Experiment

Two experiments were conducted to verify the proposed method with both the surgical simulator and the dVRK surgical robot. Ten subjects without motor impairment (6 males and 4 females, mean age = 26 ± 5 years, all right-handed and without previous surgical experience) participated in the study. All participants were informed about the experiment's purpose and protocol, and signed a consent form before starting. Ten subjects joined the simulation experiment and eight of them (4 males and 4 females, mean age = 25 ± 2 years) joined the physical robot test on the dVRK platform. The experiments were approved by the Research Ethics committee of Imperial College London (21IC7042).

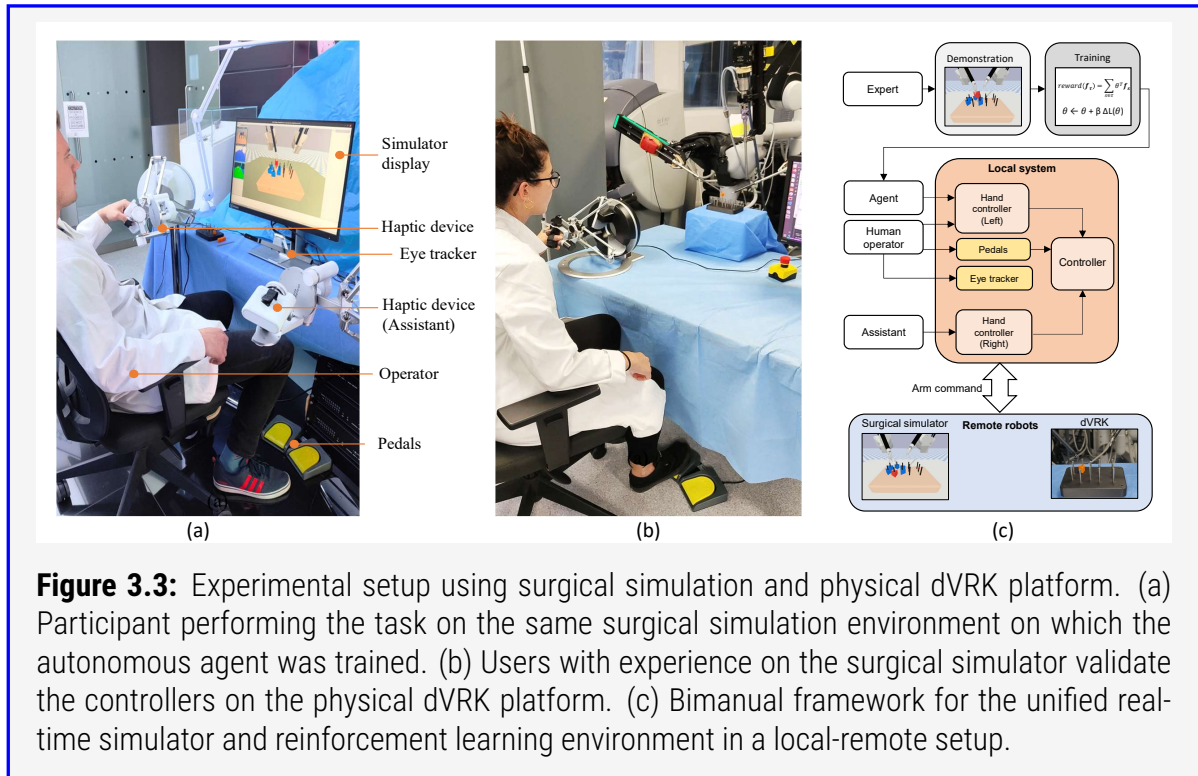
3.4.1 Experiment Setup

3.4.1.1 Simulation Platform

The first experiment was conducted with the surgical simulation platform, which is the same environment on which the agent was trained. The teleoperation system consists of local console and remote virtual laparoscopic instruments. The local console includes two hand controllers (sigma.7 Force Dimension [63]) and two foot pedals. To ensure safety, we monitor operator presence by only relinquishing control authority to the agent while the left pedal is being pressed – with control released back to the user when the pedal is released. The right pedal is used as clutch foot to uncouple the controller from the robotic arms to allow repositioning and better ergonomics [40]. Additionally, we add a constraint to Algorithm 1 to ensure that the agent movements are in the safe state space that was learned from the expert trajectory. The local console communicates with the remote virtual robot via TCP/IP to simulate the teleoperation delay between them. The hand controller can remotely control the virtual robot in seven DoF (position, orientation and grasping). The simulator monitor is equipped with Tobii Pro Fusion [4] to track the participant's eyes and observe eye related features which will be used to calculate concentration and stress.

3.4.1.2 dVRK Platform

The experiment was conducted with the dVRK platform as illustrated in Figure 3.3(b). Five participants performed the task with direct vision of the peg board, while three additional subjects were seated at the 3D dVRK interface console and tasked to perform the experiment with the 3D visual image coming from the endoscope. The experiment conducted with these two visual systems show that the condition of direct vision, which replicates and validates the setup used with the simulation platform, does not hinder the users during the task. We then formed a teleoperation with the hand controllers (sigma.7 Force Dimension [63]) as local console and patient side



manipulators of the dVRK system as the remote robots, without using the dVRK master tool manipulators. The hand controller could move the end-effector of the Endo-wrist in 6 DoF motion and grasping. The pedal interface is still used to allocate agent authority and activate the clutching function. The control of the robot works with the same modes and safety functions as in the virtual platform.

3.4.2 Task and Control Modes

The participants were seated at about 1 m distance from the peg board (about 0.5 m eyeline distance) and they could easily reach the hand controller and the pedal. During the peg transfer task, the handover requires the block to reach the robotic arm controlled by the assistant. The surgical scenario that we intend to represent is the manoeuvre that the surgeon performs to pass objects to an assistant to remove and clear the space of debris or gauzes. Two control modes were used in the experiments. In the MT mode, the participants use their left hand to control the remote virtual/physical left-side surgical instrument fully by themselves. In the SC mode, the intelligent agent performs subtasks A and C (detailed in Figure 3.1) and the human operator cooperates with the agent to perform subtasks B and D. In both control modes, the right-side remote robot is controlled by an expert assistant.

3.4.3 Experimental Protocol

3.4.3.1 Surgical Simulation

The participants were instructed on how to use the interfaces and how to complete the bimanual peg transfer task by controlling the left arm of the robot using both types of control. The experiment consisted of a total of 6 blocks. First, 2 blocks (MT and SC) of 5 minutes each to allow the participants to get familiar with the task and the interface. This is followed by 4 blocks made of 2 consecutive MT blocks and 2 consecutive SC blocks of 5 minutes each and the participants were instructed to complete as many peg transfers as possible. The order between MT and SC was randomised to disregard the learning during the experiment.

3.4.3.2 dVRK Platform

The experiment asked 8 participants that had previous experience from using the simulated platform to use again the two types of controllers in different order. However, unlike the surgical simulator, the real robot cannot be rapidly reset to its original position. Therefore, the experiment metrics were limited to measuring a total of 10 successful peg transfers for each control mode.

3.4.4 User Performance Evaluation

While conducting the experiments with each participant, the success rate and temporal performance of only the user controlled arm were measured. The evaluation of the participants' subjective workload was performed using a custom NASA TLX [51] to score the perceived Control, Mental, Temporal, Physical effort. The participants' mental load were also analytically calculated by processing the gaze data obtained from the Tobii Pro Fusion eye tracker.

3.5 Results

A comparison with the previous method applied in [158], which can be obtained by setting the epsilon-greedy variable to zero, determined that the addition of the epsilon-greedy component in the algorithm proposed allows to update a larger volume of the state space. The counting of the fully explored states (i.e. the reward values associated with all possible actions in that state have been updated) shows that the use of epsilon-greedy increases the fully explored states by 45% while it decreases the fully unexplored by 34%. The increased exploration of the states enables the agent to complete the task with initial starting states that are different from those provided by the expert. We provide evidence of this in Figure 3.4(a), which illustrates the agent reaching movement towards the target using deviated initial states. We then perform user demonstrations of the entire human-robot cooperative peg transfer using MT and SC to validate control framework developed. Typical trajectories of a trained user carried out at the end of the experiment are shown in Figure 3.4(b).

3.5.1 Performance Comparison between MT and SC

The successful peg transfer results of MT and SC are shown in Figure 3.5(a) and were analysed using statistical comparison tests [78] that determine significant differences between the two controllers. A peg transfer is defined as successful if the block is not dropped after the participant attempts to lift it. The data obtained was normally distributed, as assessed by Shapiro-Wilk's test [78] (all $p > 0.2$). A paired t-test [78] was conducted to determine the effect of different control schemes. When using MT mode, the participants could perform an average of 2 successful transfers within 5 minutes; while when robotic assistance is added, SC mode achieved about 6 successful transfers in the same time period, which is significantly higher than the MT mode ($p < 0.001$). As shown in Figure 3.5(b), the average completion time has the similar tendency, with the average completion time decreasing from using MT (134.9 ± 89.9 s) to SC (37.3 ± 5.69 s), (Wilcoxon Signed-Rank test, $p = 0.005$). Figure 3.5(c) illustrates the average responses and the standard deviation of the ten participants to the questionnaire. The rating results of mental effort and physical effort were normally distributed for both control modes ($p > 0.06$) and a paired t-test was used for normally distributed data and the Wilcoxon Signed-Rank test [78] for non-normally distributed data. It is interesting to observe that participants felt they had more *control* on the virtual environment when cooperating with intelligent agent rather than when they had full control of the virtual robot ($p = 0.008$). In addition, the participants felt SC also had less *Mental effort* ($p = 0.004$) and *Temporal effort* ($p = 0.02$) with respect to MT, but they did not felt significantly difference in *Physical effort* ($p = 0.06$) when using both control modes.

3.5.2 Mental Effort Evaluation of MT and SC using Gaze Data

In addition to the questionnaire in Figure 3.5(c), the gaze data collected from the eye tracking unit on the simulator platform can be analysed to assess mental load. Table 3.2 summarises the features extracted from the raw gaze data, which have been studied in prior works investigating mental load during teleoperation [32, 45, 92, 107, 135, 161]. These features can be divided into *eye movement*, *blink*, and *pupillary response*. The two main eye movement features are fixations (periods where the subject's gaze ceases moving and remains fixed on a point) and saccades (rapid movements of the eye between distant focus points). It can be seen that several of the eye movement features exhibit a shift as a result of changing from the MT to SC condition in a significant manner. The mean fixation duration and max fixation duration both are higher during MT, while fixation and saccade frequencies are also lower in MT. Within the blink features, it can be observed that the maximum blink duration increased for MT. Finally for pupillary response, the Index of Pupillary Activity (IPA) (IPA) is used to track the diameter of the pupil, and it has been linked to mental load responses, where higher values indicated higher task difficulty [32].

Table 3.2: Mean values across participants for gaze features associated with mental load during the peg transfers.

<i>Eye Movement Features</i>	Ratio MT/SC	p
Fixation Frequency	0.54	0.0001
Mean Fixation Duration	2.02	0.001
Max Fixation Duration	1.64	0.001
Saccade Frequency	0.52	0.0001
Mean Saccade Duration	1.06	0.72
Max Saccade Duration	0.73	0.51
Mean Saccade Speed	0.94	0.20
Max Saccade Speed	0.81	0.14
<i>Blink Features</i>		
Blink Frequency	0.60	0.0004
Mean Blink Duration	0.90	0.14
Max Blink Duration	0.74	0.01
<i>Pupillary Response Features</i>		
IPA Left	0.90	0.19
Mean Left Pupil Diameter	1.02	0.16
Left Pupil Deviation	0.93	0.29
Max Left Pupil Diameter	0.98	0.26
IPA Right	0.75	0.01
Mean Right Pupil Diameter	1.01	0.23
Right Pupil Deviation	0.91	0.23
Max Right Pupil Diameter	0.98	0.13

3.5.3 dVRK Platform Results

Eight subjects were randomly selected to test the controllers with the physical dVRK. Their temporal performance with the system is shown in Figure 3.6. The similar means and variations in the results indicate that the performances between subjects 1-5 and 6-8 are comparable despite the different visual system used. The participants required an average completion time of 29.77 ± 4.05 s using MT and 24.57 ± 1.43 s via SC without significant difference ($p = 0.15$), which may due to the practice with the simulator. However, we found that the performance between subjects varied considerably especially under the MT control mode. In addition, as shown in Figure 3.6(b), the control mode may have different effect to different subjects. To subject 2, 3, 6 and 8, SC control mode provides obvious assistance to complete the task efficiently than MT, but this tendency did not appear on subject 1, 5 or 7.

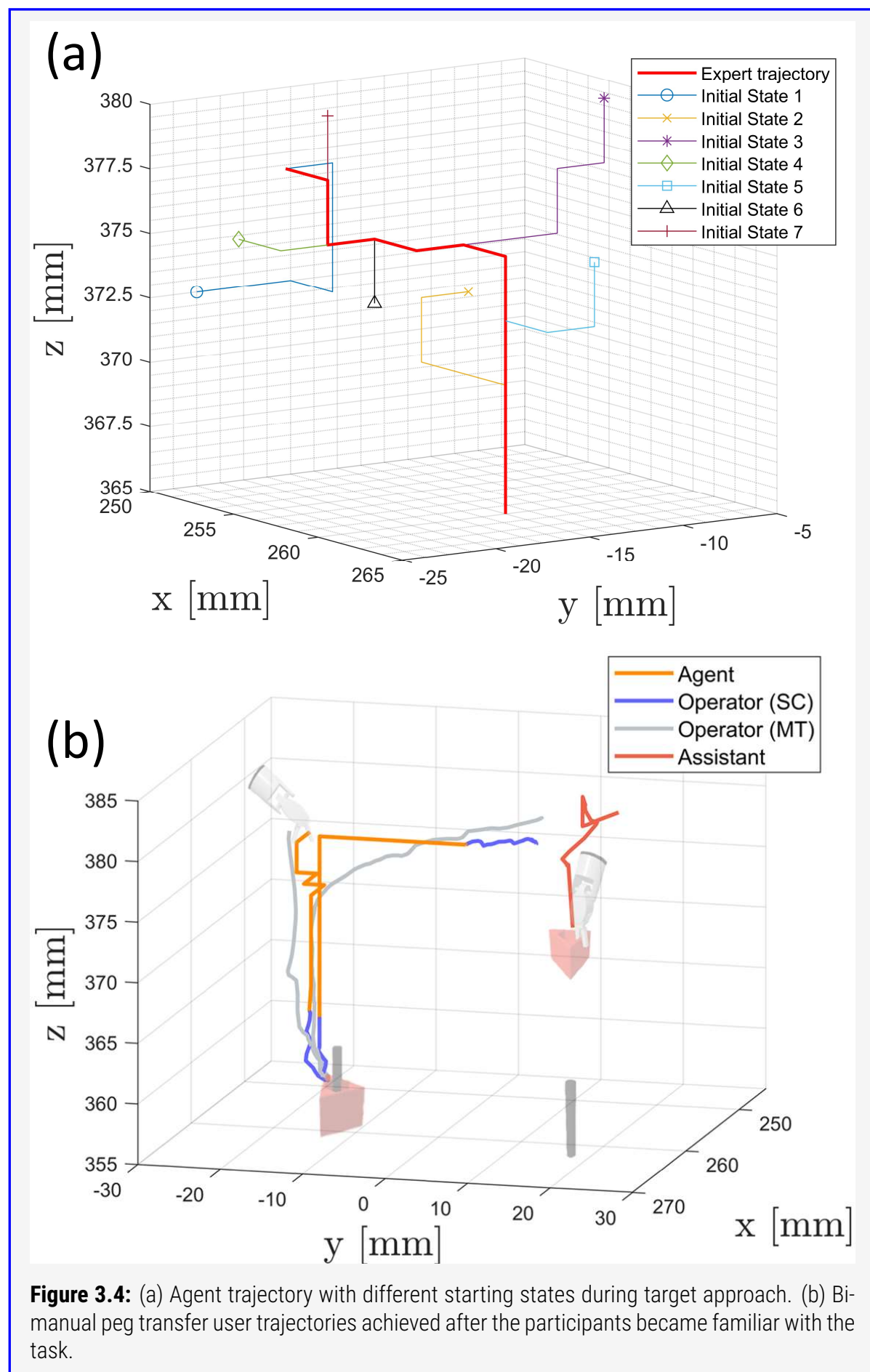
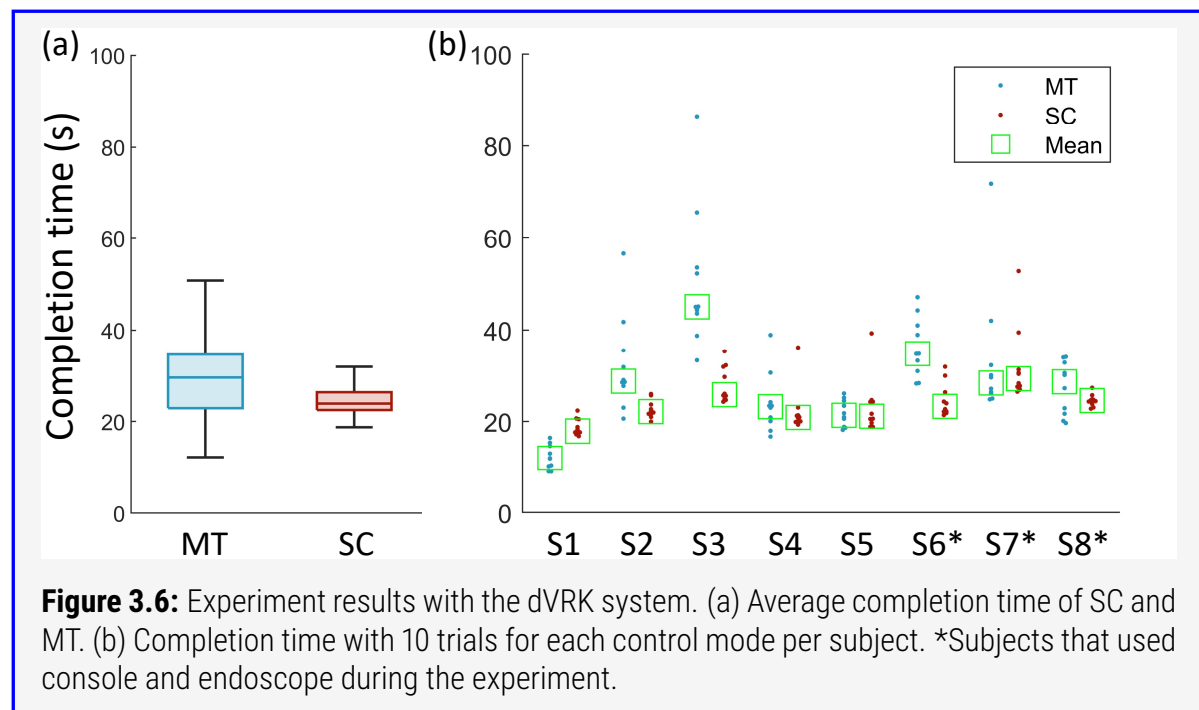
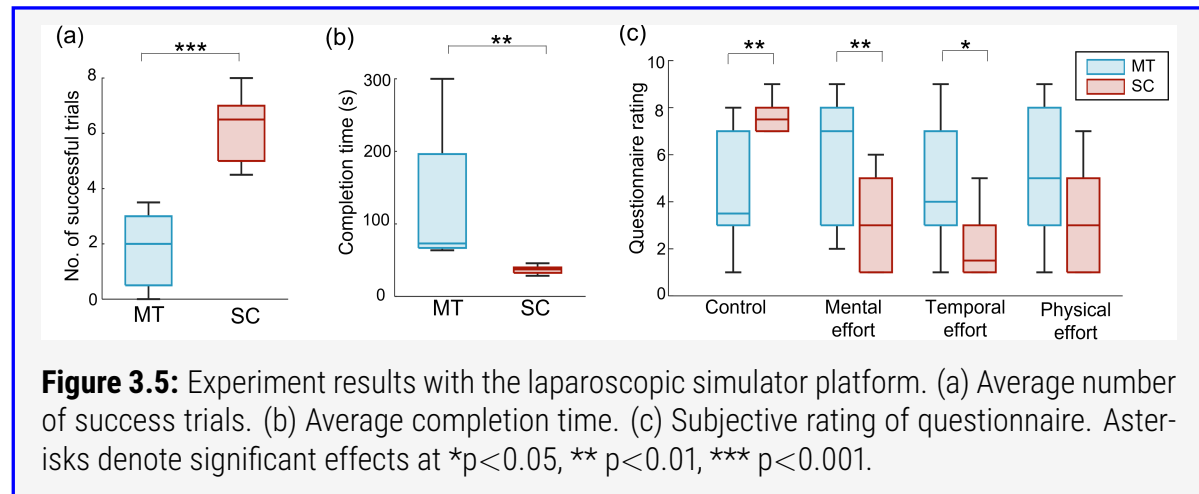


Figure 3.4: (a) Agent trajectory with different starting states during target approach. (b) Bi-manual peg transfer user trajectories achieved after the participants became familiar with the task.



3.6 Discussion

In this Chapter, EG MaxEnt IRL was used to create a model of the surgeon strategy and generate trajectories for the bimanual peg transfer task. The algorithm was successfully implemented in SurRoL for validation. The simulation environment was modified to directly integrate both a reinforcement learning training environment and a hand controller for remote teleoperation. Since epsilon-greedy promotes both exploration and exploitation, the reward values in the Q-table are not only updated near the states visited by the expert and the agent explores a larger volume of the workspace, allowing to obtain a more robust policy.

Noticeably from Figure 3.4, the agent learned the expert strategy to approach the block from above and not other directions. This shows the importance of using EG MaxEnt IRL: if the agent had only learned to end at the target, it would have reached it from a non-specific direction, interrupting the surgical workflow as the surgeon would have to manually reposition the end-effector above the block. Instead, since the agent learned the strategy, the surgeon can just resume the directional movement of the previous autonomous segment. This is clear in Figure 3.4(b) as the operator and the agent continue each others trajectory without abrupt change in direction.

The results obtained from the surgical simulator (Figure 3.5) show that participants without surgical experience can achieve a higher success rate for the bimanual peg transfer task when using SC. In accordance to the results of [68], which exhibited consistently higher success rates with a fully autonomous agent than through direct human operation, our results find a similar performance improvement while offering the possibility for the human to retain a certain level of control while still using an autonomous agent to improve their performance. The results also suggest that the learning curve for the participants when using SC is smaller than MT. The temporal performance indicates that the users can achieve a higher consistency in the speed, with small time difference between each successful attempt. The ratings in the questionnaire show that overall the users perceive a lower workload when using SC instead of MT, in the categories analysed. A common comment provided by the participants was that “the firm grip of the block was the most significant challenge” when completing the task. Interestingly, although the SC mode only assists in the trajectory and does not provide direct help in the grasping, the participants felt significantly more in control during SC as they can simply continue the motion set by the agent.

The algorithm is then successfully implemented on the real dVRK platform with eight trained participants. The results recorded in Figure 3.6 are comparable with those obtained on the simulation platform. When using SC, the participants still maintained a higher time consistency. When using MT, they required a shorter average time to complete the task if compared to the simulator experiments. This can be attributed to both the previous training with the simulator and to the more rich sensory information of the real surgical environment. As recommendation for future user studies, we suggest to record the time spent per segment, as it would allow a breakdown of the time spent

in each subtask across participants. This could potentially highlight the segments requiring the most support and guide future assistance strategies.

From the gaze results in Table 3.2, we observed higher fixation duration and max fixation duration in MT, along with lower fixation and saccade frequencies, when compared to SC. The increase in saccades for SC may be a result of the experiment fixed time-limit and the increased number of successful trials during this time limit. More successful grasps over the time span will result in the operator performing more action sequences, thus requiring them to shift their attention more times. Conversely, for MT, as fewer successful trials were accomplished within the time limit, the higher fixation duration may indicate that the operator spent much more time focusing on specific steps in their action sequences during a trial. While some prior works have shown that increased fixation duration has a negative correlation with mental load [45, 135], a recent study has highlighted differing fixation behaviours depending on whether the subject is experiencing a *perceptual* or *cognitive load* [92], where the fixation duration will either decrease or increase respectively. Given the visual scene does not rapidly change during the experiment, it can be argued the subjects are primarily experiencing a cognitive load and so the observed result is consistent with prior works. Additionally, the maximum blink duration is higher during SC, which is again consistent with prior studies that show a positive relationship between blink duration and mental load [45, 107, 161]. Furthermore, the right IPA is significantly higher during SC. While this indicates a higher load for SC, it is possible that, similar to the fixation and saccade frequency feature, the higher value may be linked to the greater number of successful trials the participant completes. The participants perform more work, and so their eye must engage in more movement. It is not clear why the IPA for the right eye indicates significance, while the left eye does not. While it is possible this is due to head movements, alternatively it may be linked to dominant eye effects recently observed during mental load studies considering pupil diameter sizes in response to challenging surgical tasks, where in many cases the right pupil was observed to be larger than the left [17]. This discrepancy in pupil behaviours may be used in future research for the effective assessment of mental load where left or right eye dominance varies between individuals. Overall, these gaze features agree with the subjective ratings, where the MT condition exhibited a higher mental workload.

3.7 Conclusion

The system demonstrated the benefits of training agents for shared control of surgical robots during the completion of fundamental movements in robotic surgery using EG MaxEnt IRL. It showed how shared control impacts mental load if compared to traditional manual teleoperation. We developed the algorithm in an unified simulator platform that can accelerate training and testing of HRI control frameworks and successfully implement it on both a virtual and real dVRK. We showed that our control framework can allow users to experience significant reduction of effort and mental load while improving performance. As explained in [87, 158], a significant limitation of MaxEnt IRL

is its susceptibility to large state space, which limits the number of dimensions that can be used during training. Therefore a further improvement of the algorithm is to investigate methods that can be adapted to directly handle continuous state space environments, [20, 90, 93]. In the next Chapters, we investigate other techniques part of the human-robot cooperation spectrum, namely supervisory control and assisted control, tackling the limitation found in MaxEnt IRL (i.e. Chapter 4) and implementing in more complex tasks such as suturing (i.e. Chapter 5).

Chapter 4

Supervisory Control in Peg Transfer

This Chapter extends the work from Chapter 3 by analysing the human-robot cooperative controllers in the surgical domain, focusing on both shared control and supervisory control paradigms. To address the previous limitations (i.e. discretised observation and action space), we propose a reinforcement learning algorithm, transverse generative adversarial imitation learning (tGAIL), which is employed to train the robot from the expert's demonstration. The algorithm can accommodate a continuous workspace and shows competitive generalisation ability compared to inverse reinforcement learning and conventional tGAIL. We then propose a priority-changing control method to effectively combine the surgeon and robot's strengths by dynamically adjusting control priority based on the deviation distance. We show that using this method in a supervision framework boosts the performance of the human operator when completing the peg transfer task. By learning from the expert and collaborating with the human during the task, the intelligent agent can help to reduce operation time by 31.7% and the human input by 60.5% compared to direct teleoperation.

The content from this Chapter was published in the proceedings of the IEEE International Conference on Robotics and Automation (ICRA), May 2024: Chen, L.*, Hu, Z. J.*, Huang, Y., Burdet, E., Rodriguez y Baena, F. "Human Robot Shared Control in Surgery: A Performance Assessment".

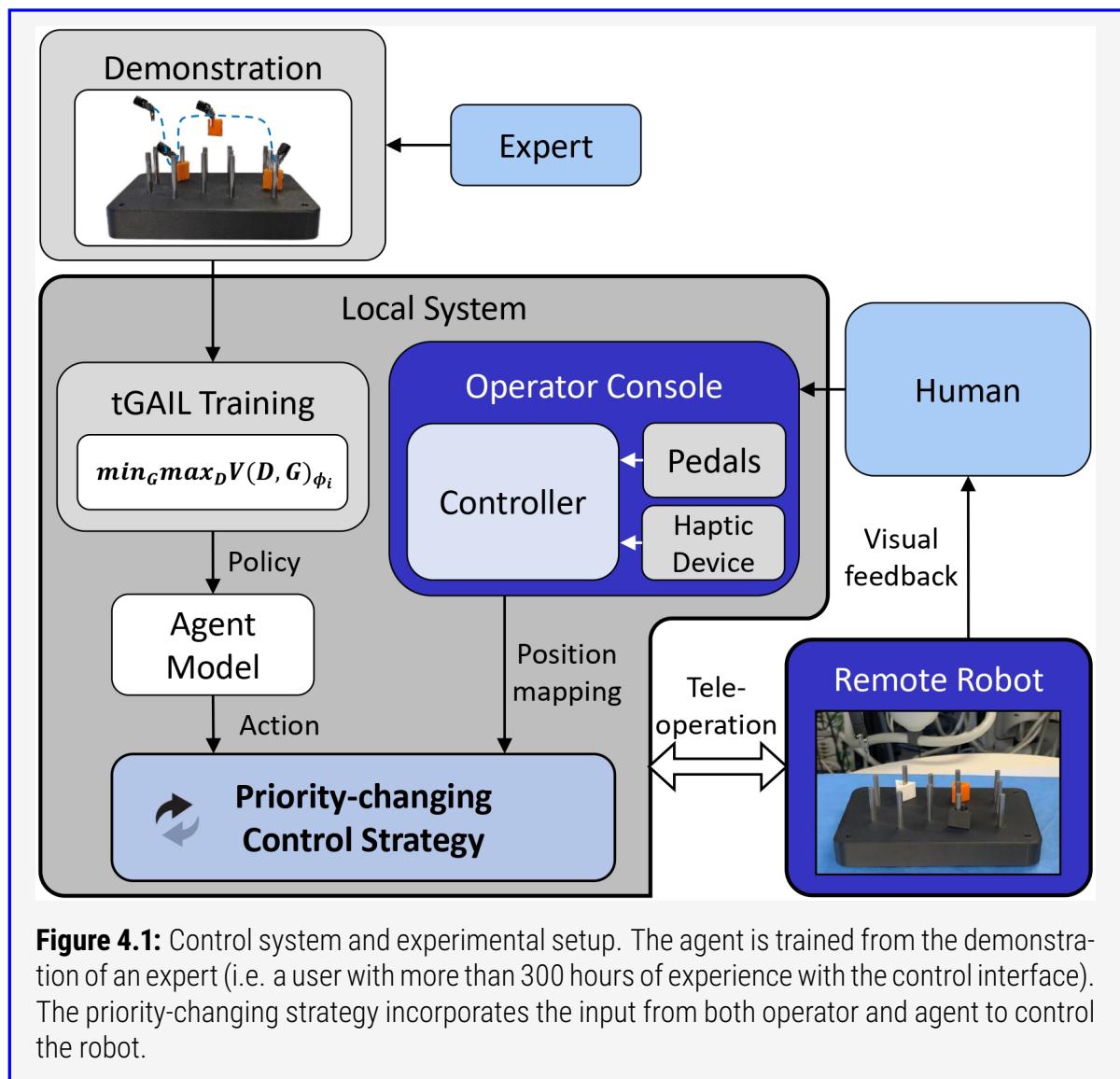
*Co-first authors.

4.1 Introduction

Robotic-assisted surgery employs robotic systems to assist a human operator during the surgical procedure [16, 85, 131, 164]. These systems are designed to provide surgeons with increased dexterity and precision, allowing them to perform complex procedures with greater accuracy and less effort than traditional surgical methods [64]. Despite these advantages, fully autonomous robotic surgeries remain unrealistic due to the inherent complexity of surgical decision-making, potential unpredictability, as well as ethical and legal implications. These issues have led to the concept of human-robot cooperation, which is gaining prevalence in the field [72, 73, 89].

Human-robot cooperative controllers attempt to combine the advantages of humans and robots to effectively complete the surgical task. This approach allows the surgeon to retain control over the procedure while also taking advantage of the robotic system's advanced precision and efficiency. Furthermore, the robotic system can also be programmed to perform certain tasks automatically, such as maintaining a specific surgical position [84] or performing repetitive motions [115]. Therefore, human-robot cooperation systems can help to reduce the surgeon's cognitive load and fatigue, which could lead to improved surgical outcomes. Several benchmarks [67] are available to evaluate the performance of these controllers, however, the operator can struggle with the task due to inexperience or limited environmental information. Therefore, a study involving novices who operate in an adequately challenging environment is necessary. Performance assessments have been used before to analyse different control methods in [72], however the analysis was not performed on standardised tasks of laparoscopic surgery [67], such as peg transfer.

Common techniques used for human-robot cooperative control include shared control [60] and supervisory [123] methods. Shared control methods, such as segmented control, have a level of pre-determined and fixed autonomy, meaning that the operator cannot obtain control authority unless the agent permits it or a specific switch event occurs. [60] and [158] implement a switch of the control authority between a human operator and a fully autonomous agent according to the task demands. However, to ensure the constant presence of the human surgeon, the fully autonomous component of the segmented control could be replaced by a supervisory type of control. Supervisory methods in shared control allow a dynamic adjustment of the level of autonomy in a surgical robot based on the current task demands and the capabilities of the human operator. An example is the Smart Tissue Autonomous Robot (STAR) [130], which is a robot that uses supervisory shared control to perform surgical procedures autonomously, under the supervision of a surgeon. The controller can achieve seamless switching by adopting real-time context awareness of the surgery. Another supervisory method called guidance priority adaptation, or priority-changing [123, 148], allows the human operator to provide both forces and high-level priority commands to the robot if errors or issues occur during completion of the task. Guidance priority adaptation allows for the reduction of the human operator's cognitive load, by providing the robot with the appropri-



ate level of autonomy while performing the surgical task, which can also be divided into different stages. In [14], the control during the peg transfer task is divided into a supervised autonomous phase during coarse operation and a manual control phase during precise operation. Compared to peg transfer, in suturing [158] the task can be broken down into more stages and the robot can perform needle pulling, translating, and reorienting. Simple or repetitive stages are usually allocated to the robot, while the human is responsible for the most important tasks.

The prevalent control methods involve authority switching between surgeon and trained agent, with the latter primarily handling course control. Therefore, a more focused study on the role of supervision-based control can help understand how performance and acceptance affect novices and inexperienced users of a surgical robotic system.

In this Chapter, we propose a control strategy based on guidance priority adaptation to increase the automation in surgical robots. The priority is determined by the distance between the agent's trajectory and the robot hand, and represents the surgeon's trust in the agent. We choose to imple-

ment this supervision control strategy on both a virtual and physical dVRK [60, 76, 165] and conduct experiments with two detailed user studies. The main contributions of the work in this Chapter are as follows:

- We introduce tGAIL, a novel reinforcement learning algorithm designed to learn from expert demonstrations. The algorithm offers improved robustness and consistency in learning.
- We implement a novel priority-changing control approach on the dVRK.
- We perform an evaluation study on supervision, segmented, and manual control in a real surgical robotic system in the peg transfer task.

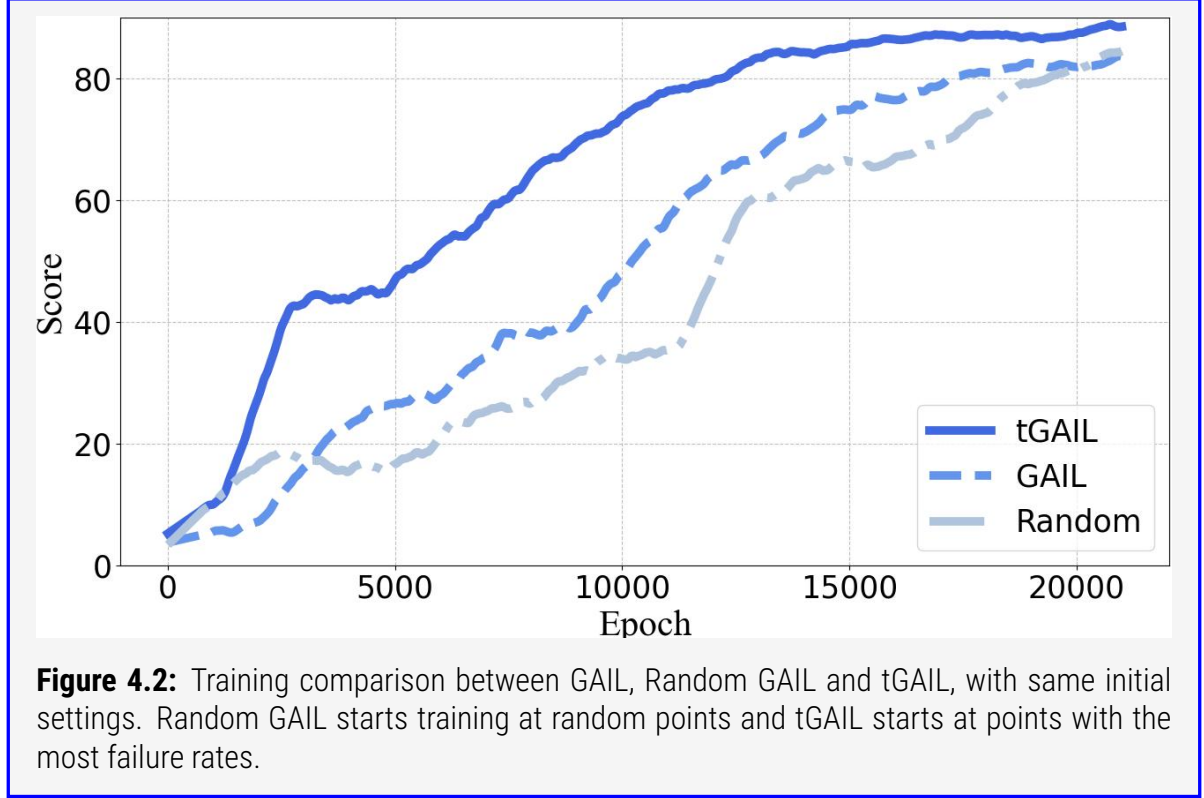
The Chapter is structured as follows: Section 4.2 introduces tGAIL, the priority-changing control, and the simulation user study findings. Section 4.3 details the real-world experiment. Section 4.4 elaborates on the results. Section 4.5 discusses the findings of the user study, ending with conclusions in Section 4.6.

4.2 Methodology

4.2.1 Transverse Generative Adversarial Imitation Learning

In the first experiment, we implement an Epsilon-Greedy Maximum Entropy Inverse Reinforcement Learning (EG MaxEnt IRL) algorithm to validate the feasibility of utilising our expert trajectories dataset for learning from demonstration in a peg transfer task, as in Chapter 3 [60]. As an EG MaxEnt IRL trained agent has already been established to be working effectively in [60], we aim to use the same agent and perform a first study in a simulated environment to determine the most effective role of supervision in a segmented framework. Subsequently, for the real dVRK study, we present a refined supervisory algorithm which we call transverse generative adversarial imitation learning (tGAIL), shown in Figure 4.1 and Algorithm 4.1. Built upon generative adversarial imitation learning (GAIL) [58], tGAIL traverses the expert's trajectory τ_E , encompassing the expert's policy π_E . Starting from the worst point ϕ_i responsible for most failures, tGAIL seeks to minimise the generator's score while maximising the expert's score. Unlike sequential learning, tGAIL covers all scenarios simultaneously, resulting in accelerated learning, as demonstrated in Figure 4.2.

The EG MaxEnt IRL agent's discretised observation space, governed by the q-table, encounters challenges in high-resolution observation spaces. In contrast, GAIL uses generative adversarial network techniques, accommodating a continuous workspace. By training both the generator and discriminator, the GAIL agent can potentially exhibit higher generalisation compared to the EG MaxEnt IRL agent. Both training methods utilise the same expert demonstrations and environment.



Algorithm 4.1 Transverse Generative Adversarial Imitation Learning

Require: Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator setting parameters

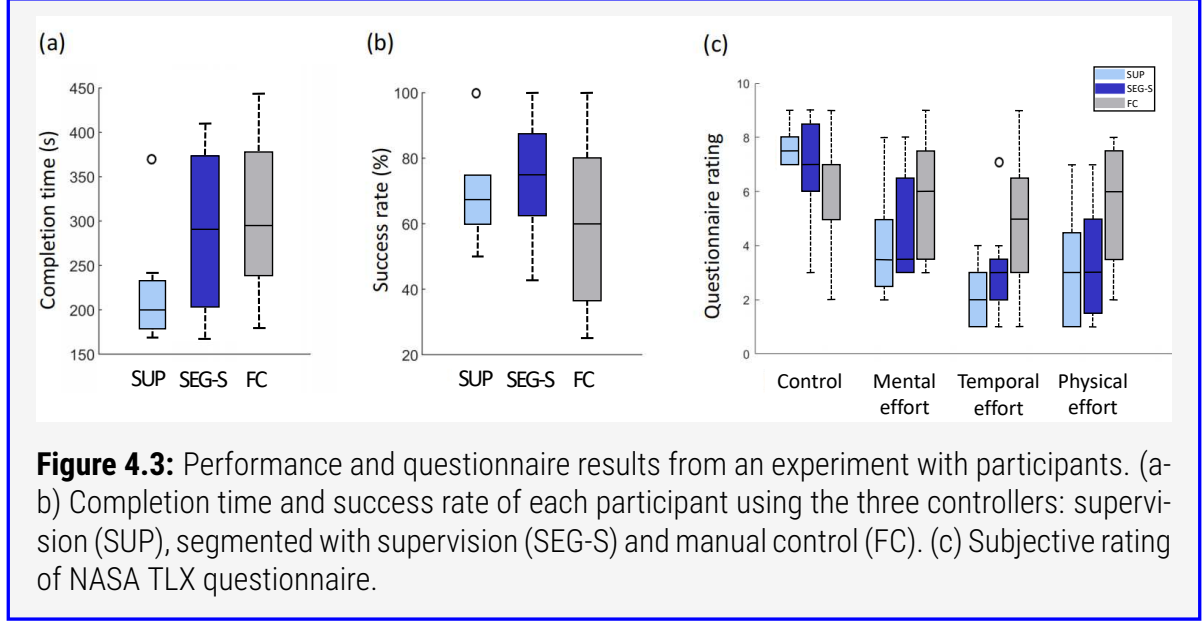
- 1: **for** $i = 1, 2, 3 \dots$ **do**
 - 2: Sample trajectories τ_i and find the worst starting point ϕ_i
 - 3: Update the discriminator D and the generator (policy) G with cost function $\min_G \max_D V(D, G)_{\phi_i}$
 - 4: **end for**
-

4.2.2 Priority-changing Control

Control priority governs the operator's influence relative to the agent's input. Dynamic control priority enables the surgeon to exert control as needed, saving effort when the robot operates effectively. The control strategy that switches control priority is defined as:

$$\Delta x(t) = (1 + k_\alpha)\Delta x_h(t) + (1 - k_\alpha)\Delta x_a(t) \quad (4.1)$$

where $\Delta x(t)$ is the overall command passed to the robot, $\Delta x_h(t)$ is the human input and $\Delta x_a(t)$ is the agent's input. The parameter k_α is the relative coefficient that achieves guidance priority adaptation, defined as:



$$k_{\alpha} = \begin{cases} 0, & D \leq D_{min} \\ \frac{1}{2} \left[1 - \cos \frac{\pi(D-D_{min})}{D_{max}-D_{min}} \right], & D_{min} < D \leq D_{max} \\ 1, & D_{max} < D \end{cases} \quad (4.2)$$

where D is the distance between the autonomous agent pattern and end-effector position in one time step, while D_{min} and D_{max} are the minimal and maximal interference distances, respectively.

The guidance priority is determined by the deviation of the robot's end-effector from the path generated by the agent. While the distance between these is less than the minimum interference distance D_{min} , slight trembling will not affect the agent's task. As the distance increases, the control priority is gradually transferred to the human. If the deviation becomes larger than the threshold of maximum interference D_{max} , the human receives full control priority.

The work in [123] uses force and deviation distance to achieve guidance priority adaptation, but the human needs to apply a relatively large force to get involved in decision-making when the end-effector follows the robot's desired trajectory. By allocating the same control priority at the start of the task, the human and the robot can work as partners. This allows the robot to perform its assigned task without scaling, which will later occur during the negotiation of the control authority after the switching point. Using our proposed method, the agent can conduct the task without external aid; if any correction is needed, the human can get involved in the task with partial priority.

4.2.3 Control Methods

Initially, we validate the supervised control method in simulation, integrating it with EG MaxEnt IRL, as it is well established for discrete observation space applications. Eight participants (4 females and 4 males, 22-27 years old, with little to no experience with teleoperation devices) used a local teleoperation console to operate a virtual environment with a simulation of the dVRK robot. We employed a real-time simulator, integrating the modified SurRoL environment [60] with a hand interface (omega.7, Force Dimension) to control a virtual dVRK. The participants were instructed on how to use the interfaces and how to complete the peg transfer task by using three types of control: supervision, segmented, and manual control. In this initial study, we aimed to use the established EG MaxEnt IRL to determine the overall user experience with the three different methods and investigate the effectiveness of the proposed supervisory control formulation. We also integrated the guidance priority adaptation with a collaborative framework [110] in order to allow more freedom and control to the human when using segmented control. The description of the three controllers are as follows:

1. *Supervision Control (SUP)*: The agent generates the trajectories required for completing the peg transfer while the human operator can provide high-level commands to the agent at any time in a shared control fashion.
2. *Segmented Control (SEG)*: The control priority is exchanged between human and agent. The task stages where the human operator has full control were determined according to [60]. During movements towards targets, the agent performs the task under supervision (SEG-S) or autonomously (SEG-A). This corresponds to the approach and repositioning of the block. When in close range to the target, the authority is switched to the human that performs the necessary physical interactions. This corresponds to picking up and placing down the block.
3. *Fully Manual Control (FC)*: The controller simply filters tremors from the human operator's hand but no agent is involved in controlling the surgical tool.

The controllers used to teleoperate in the surgical simulator are summarised in Table 4.1, where the operator is the human and the agent refers to the trained model.

During the experiment, successful peg transfers are defined as the relocation of the block from the starting peg to the goal peg without failures, which are events such as peg dropping or collisions with the environment. The participants were required to complete three successful peg transfers with each controller and the success rate was calculated as $success\ rate = \frac{successes}{failures + successes} * 100$. The entire experiment consisted of a total of four phases: a first training phase in which the participants would get familiar with the interface and simulator, followed by two phases for SUP and SEG-S, and a fourth phase for FC. The order of SUP and SEG-S was randomised to prevent bias due to learning. While conducting the experiments with each participant, the success rate, operator input, and temporal performance of the user were measured. The evaluation of the participants'

subjective workload was performed using a custom NASA TLX [51] to score the perceived Control, Mental, Temporal (i.e. perceived time spent on the task), and Physical effort experienced.

Table 4.1: Controllers summary

	Approach target	Interact with target
Supervision	Operator+Agent	Operator+Agent
Segmented	Operator+Agent	Operator
Manual	Operator	Operator

4.2.4 Simulation Experiment Results

The data was normally distributed as tested using the Shapiro-Wilk's test. We used one-way ANOVA to analyse the effect of the control modes and post-hoc comparison with Bonferroni correction between single pairs. The results from the study conducted with eight participants are shown in Figure 4.3(a-b). We compared the data of the single controllers with each other using the least significance difference test. We found that SUP is faster than FC ($p = 0.055$) and similar to SEG-S ($p = 0.118$), while SEG-S and FC resulted in a similar average and standard deviation, i.e. $(\sigma, \mu) = (289.2, 93.6)$ and $(305.9, 92.7)$, respectively. All three controllers exhibited similar success rates with the SEG-S (74.1, 20.3) showing moderately better performance than SUP (69.4, 15.5) and FC (59.7, 28.3). Overall, both newly proposed methods revealed a tendency to outperform the manual teleoperation.

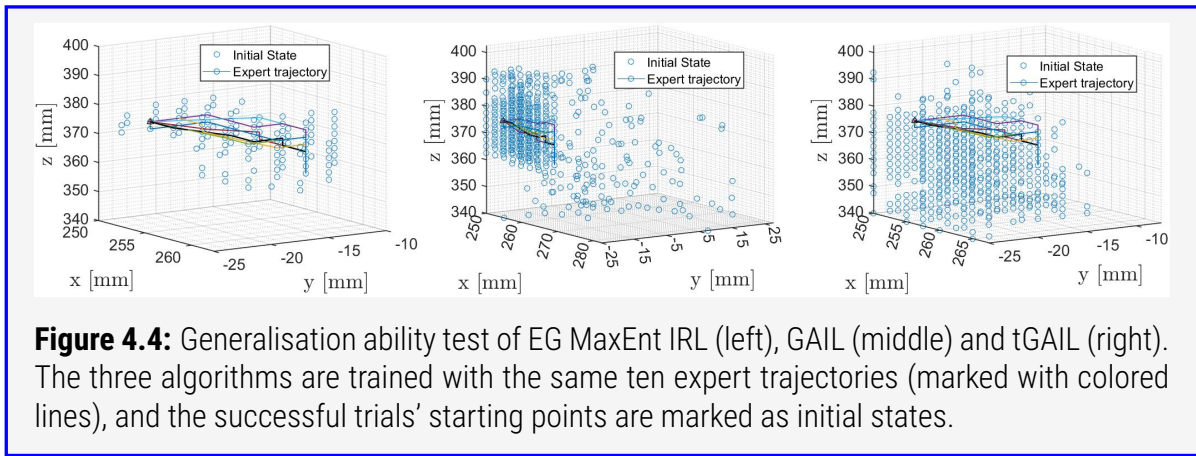
The responses obtained from the questionnaire are illustrated in Figure 4.3(c). The participants felt similar control confidence when using the three controllers ($p = 0.398$), with SUP being the highest rated (7.6, 0.74), followed by SEG-S (6.9, 1.9), and FC (6.1, 2.1), with the latter presenting the largest standard deviation of the three. The user's rating of mental effort shows the same pattern with a similar standard deviation for all three methods. The subjects' temporal effort was felt different among the three modes ($p = 0.028$). SUP demonstrates less effort than traditional FC ($p = 0.028$). This is followed by SEG-S which required less temporal effort compared to FC, but with no statistical significance ($p = 0.196$). The same result is reflected in the scoring of the participants' physical effort, with SUP (3.1, 2.2) and SEG-S (3.3, 2.1) having analogous ratings. From users' feedback, a common issue with SEG-S is the switching between manual and supervised control. The users found that a sudden and not voluntary switch between the two controllers initially caused confusion and it could only be fully learned towards the end of the block.

Overall, the integration of the supervision in the segmented framework does not seem to outperform SUP. Therefore, we decided to proceed to implement the same controllers in the user study on the dVRK, with the supervision component of the segmented control converted back to fully autonomous, i.e. SEG-A.

4.3 Experiments

A user study was conducted on the real dVRK with 20 participants (10 females and 10 males, 22-44 years old, all without previous experience with the surgical robot system). The experiments were approved by the College Research Ethics Committee of Imperial College London (21IC7042). Participants were informed about the experiments' purpose and protocol and signed a consent form before starting.

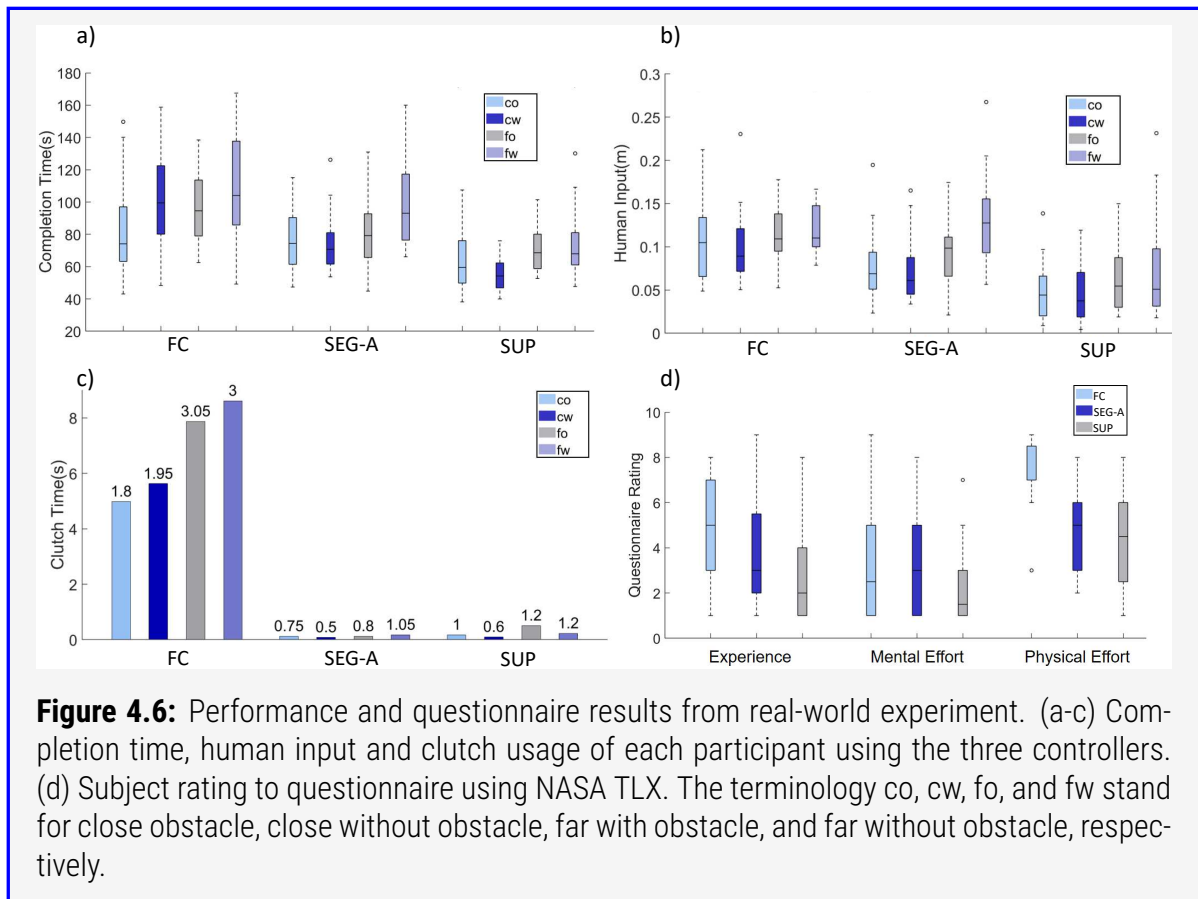
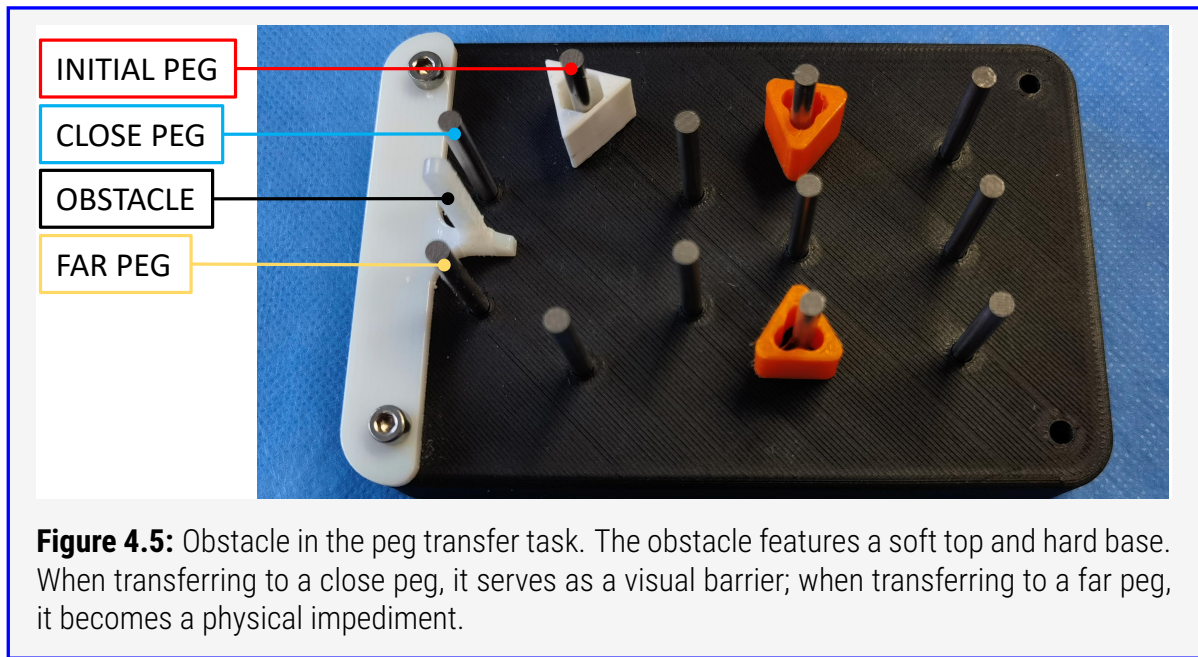
4.3.1 User study on dVRK



While the majority of the experimental settings remain consistent with the simulated study, the real-world experiment introduces an added layer of complexity by including a physical obstacle. This increases the difficulty of the task and tests how the control strategies affect the user's experience and performance when interacting. To ensure safety throughout the study, we allow the operator to switch back to manual by pressing a pedal. Another pedal is used as a clutch to uncouple the controller from the robotic arms and allow repositioning for better ergonomics [40].

The task commences from a random position and encompasses twelve distinct trials, representing different combinations of mode, target, and obstacle. The experiment includes three modes (SUP, SEG-A, and FC) and two target objectives: placing the block into either a close peg or a far peg, as depicted in Figure 4.5. As inexperienced participants are more prone to causing collisions between the pegs and the robot's arm, pegs on the right are excluded from the experiment to mitigate this risk. The trials are evenly divided, with six incorporating an obstacle and the other six without. The obstacle's function varies depending on the stage of the experiment. During the transfer of the peg to the closer pillar, the obstacle serves as a visual barrier, challenging the perception of the user and reliance on the agent. Conversely, when the peg is transferred to the more distant pillar, the obstacle takes on a physical role. The situation reflects challenges in surgery that the agent may struggle to overcome, such as unpredictable events not encountered in the training data.

Each participant, initially, is allocated a five-minute training to familiarise with the system and task.



Following this phase, the experiment commences with a randomised sequence of inclusion and exclusion of the obstacle. At each round, the experiment is then conducted across three different control modes, in a randomized sequence to ensure unbiased evaluations. Upon completion of all trials, participants are requested to complete a questionnaire.

4.4 Results

4.4.1 Generalisation Assessment

A test was conducted to assess the generalisation ability and performance between EG MaxEnt IRL and tGAIL. In this generalisation ability test (Figure 4.4), both methods were trained on the same dataset, which consists of ten expert trajectories, all starting and ending at the same locations. By providing a limited number of expert trajectories, the evaluation focuses on the algorithms' adaptability and robustness.

Following training, the agent's performance is assessed by starting it from random points. Each initial state is denoted by a blue circle in Figure 4.4, representing the agent's starting positions. In the generalisation ability test, EG MaxEnt IRL succeeded in 117 states, GAIL achieved success in 452 states, and tGAIL succeeded in 508 states. Despite the added complexity of training GAIL and tGAIL with discriminator and generator components, it demonstrated superior learning and adaptability compared to EG MaxEnt IRL. GAIL excels near expert trajectories and in reaching distant positions, however it lacked consistency and robustness. In contrast, tGAIL consistently maintained steady performance.

4.4.2 Real-world Experiment

In the user study, despite the presence of a far obstacle, SEG-A exhibited an 85% success rate, a notable decrease compared to other tasks, which all achieved near 100%. The increased success rate compared to the simulation can be attributed to gaining 3D perception. The likelihood of obstacle interaction varied across modes: it was highest in FC, reduced in SEG-A, and minimum in SUP.

The performance analysis from the real-world experiment is depicted in Figure 4.6 (a-b). The three control modes showed difference in efficiency ($p < 0.001$). SUP demonstrates to be faster compared to SEG-A ($p < 0.001$) and FC ($p = 0.001$), aligning with the simulation results. SEG-A and FC showed similar time performance ($p = 0.075$). The average and standard deviation (σ, μ) for FC, SEG-A, and SUP are (92.1, 26.0), (78.9, 19.3), and (62.9, 13.8), respectively. The human input results showed to be not normally distributed, therefore we used Friedman test which revealed a difference in the group ($p < 0.001$). SUP not only completes the task in the least time but also requires the least operator's input (0.047, 0.029) compared to FC ($p < 0.001$) and SEG-A ($p = 0.005$). FC exhibits the largest average and standard deviation (0.105, 0.033), and SEG-A falls in the middle with (0.084, 0.037). The experiment also examined clutch usage, as illustrated in Figure 4.6 (c). The clutch is used to adjust the hand pose when it reaches workspace limits, and its use costs time and diverts the surgeon's attention [35]. Figure 4.6 (c) represents the average time spent with the clutch enabled for each trial. Compared to FC, both SUP and SEG-A allow the operator to concentrate on the task rather than the workspace limits. This focus enhances

efficiency and safety by potentially reducing surgery time.

As shown in Figure 4.6 (d), SUP scored the lowest across all metrics, albeit with greater variability than SEG-A. Some participants exhibited conflicting decisions with the agent in SUP, a phenomenon absent in SEG-A. Both SUP and SEG-A were found to conserve participants' mental and physical effort, with SUP showing higher performance on average.

4.5 Discussion

tGAIL's strategy of training from the worst starting point improved the policy generalisability, when compared to EG MaxEnt IRL and GAIL. This approach improves on the algorithm's ability to overcome learning challenges and exhibits enhanced robustness compared the other methods.

In the simulation study, increased autonomy in the task reduced the completion time, enhanced and stabilized the success rate, and predominantly saved participants' effort. This suggests that inexperienced users have a preference for shifting the control authority to the agent in order to improve performance, possibly due to unfamiliarity with the task.

In the real dVRK user study, SUP surpassed SEG-A in completion time, human input, and questionnaire ratings. In contrast, FC consistently demonstrated higher standard deviations. As autonomy increases, efficiency rises, simplifying the task for new users by reducing the required effort and minimising variability. Although the increased complexity introduced visual and physical barriers to the operators, users' feedback after failing indicated that failures often arose from participant overconfidence as they became familiar with the task. Furthermore, the use of clutch significantly decreased in SEG-A and SUP compared to FC. This can be attributed to the fact that in FC the participants required more time and attention to adjust the controller. Conversely, in SUP and SEG-A, the clutch served as a pause to observe the environment or make quick adjustments.

Approximately 10% of users found SUP challenging due to conflicts with the agent's path selection, leading to varied user experiences. Users who effectively collaborate with the agent reported smoother experiences with minimal manual intervention. In contrast, 90% of users perceived FC as the most demanding mode, requiring full concentration and significant physical and mental effort. Among users, 45% preferred SEG-A for its reduced attention demands, while 55% favored SUP. When considering the user's perspective as potential patients, 45% expressed a preference for SUP, 30% for FC, and 25% for SEG-A. These preferences indicate the feasibility of implementing supervisory control methods in surgery that provide continuous support and assistance. It is important to note that all participants were new to the surgical system. Due to their inexperience, they are likely to have found SUP akin to autonomous control, which offered a sense of safety, while still having the ability to make adjustments to the robot. Furthermore, as Figure 4.6 shows, targeted support in specific task segments can reduce effort and time. This indicates that correctly identifying and supporting key segments is essential and a distributional analysis of task time by

segment and participant is recommended as future work.

The user study in the peg transfer task highlights SUP's advantages in reducing completion time, human input, and enhancing the overall user experience. Although obstacles were used to emulate the challenge of a physical or visual barriers during the task, it is essential to acknowledge that, due to the simplified nature of this standardised task, the results cannot be directly translated to real surgical scenarios and that further work is needed to study more complex tasks that can actively engage the user, such as suturing.

4.6 Conclusions

In this Chapter, we introduced and validated a novel priority-changing control method used with tGAIL, which demonstrated improved learning ability and robustness. The experiment conducted on the dVRK system demonstrated the feasibility of this control strategy.

The user study with the dVRK has further revealed the advantages of the various control approaches, with SUP showing notable performance improvements over SEG-A and FC. These findings, together with the higher acceptance of SUP than FC, highlight the feasibility and importance of incorporating autonomous assistance during surgical applications while maintaining user acceptance. Therefore, in Chapter 5 we focus on the last remaining method in the spectrum of teleoperative controls which can support and augment the surgeon's input in a continuous and adaptive manner: assistive control. Following the results from the user studies, we address problems such as adaptation to an unstructured environment and seamless switching between different type of controllers in the suturing task setting.

Chapter 5

Intent detection in Robotic Suturing

Robotic-assisted procedures offer enhanced precision, but while fully autonomous systems are limited in task knowledge, difficulties in modeling unstructured environments, and generalisation abilities, fully manual teleoperated systems also face challenges such as delay, stability, and reduced sensory information. To address this and the limitations discussed in the previous Chapters, we developed an interactive control strategy that assists the human operator by predicting their motion plan at both high and low levels, complementing the strengths of the human with the robot. At the high level, a surgeme recognition system is employed through a Transformer-based real-time gesture classification model to dynamically adapt to the operator's actions, while at the low level, a Confidence-based Intention Assimilation Controller (C-IAC) adjusts robot actions based on user intent and shared control paradigms. The system is built around a robotic suturing task on the dVRK, supported by sensors that capture the kinematics of the robot and task dynamics. Experiments with users of varying skill levels demonstrated the effectiveness of the proposed approach, showing statistical significant improvements in task completion time and user satisfaction compared to traditional teleoperation.

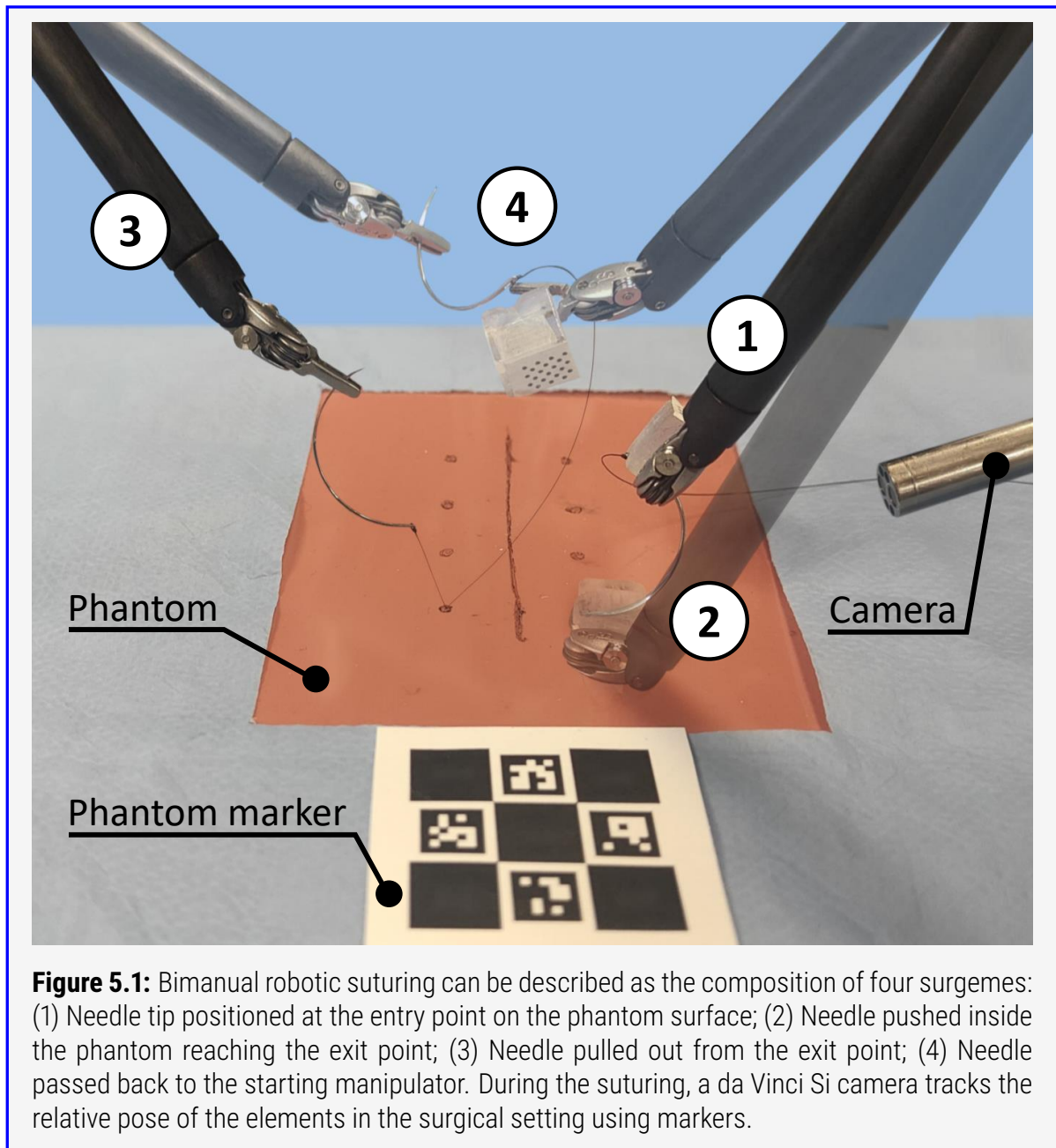
The content from this Chapter is currently under review in IEEE Robotics and Automation Letters (RA-L): Hu, Z. J., Xu, H., Kim, S., Li, Y., Rodriguez y Baena, F., Burdet, E. "Confidence-based Intent Prediction for Teleoperation in Bimanual Robotic Suturing".

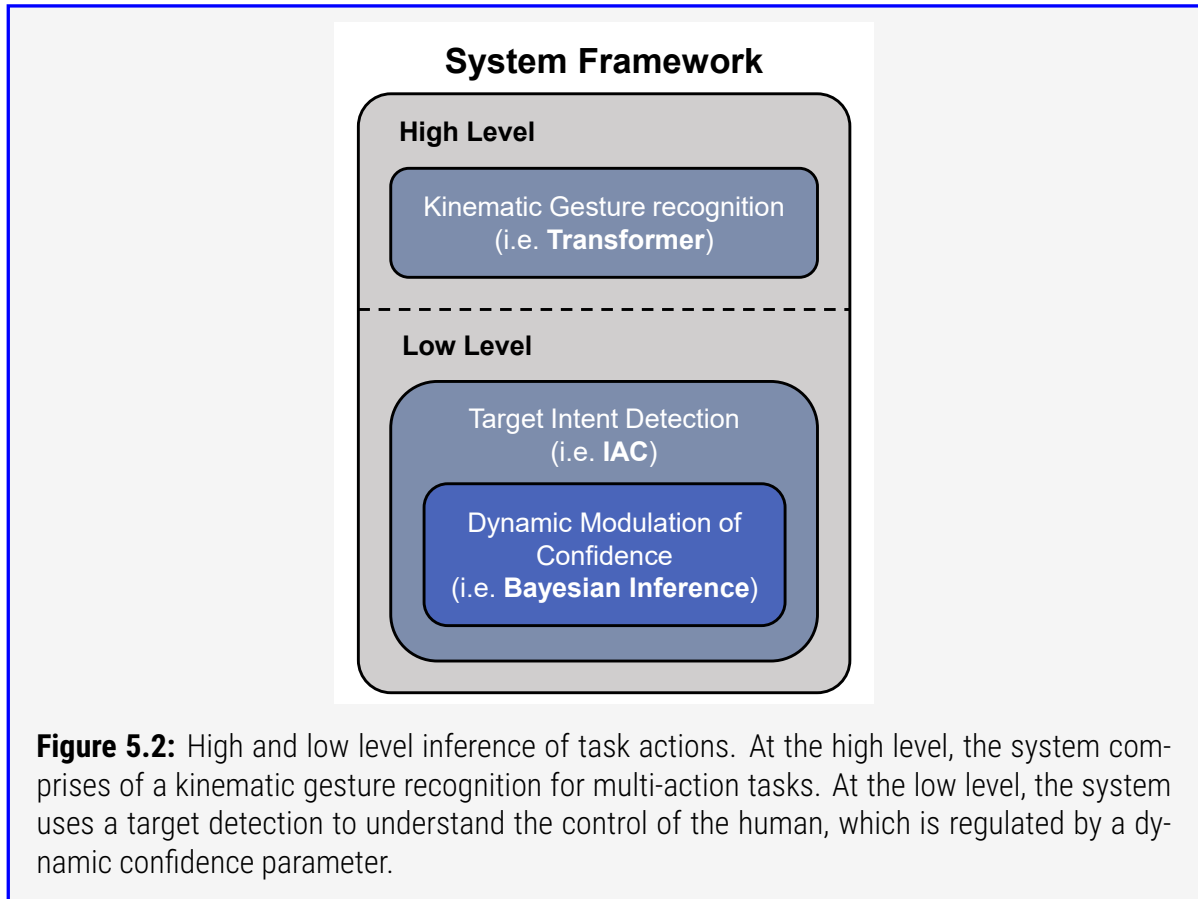
5.1 Introduction

Traditional teleoperation is based on a leader-follower relation, where the human operator has full control of the robot's movements [139]. While this approach offers clear roles and responsibilities, it does not fully utilise the capabilities of both the robot and the operator. Robots such as the da Vinci surgical system are equipped with sensors and models that can provide valuable local information not accessible to the human operator due to, for example, visual occlusions, or because they operate in different sensory modalities. To address such limitations, the concept of human-robot cooperation has emerged. By spanning across the spectrum between traditional fully manual teleoperation and full autonomy, human-robot cooperation uses the benefits of both to enhance teleoperation with the robot's sensory data and control [69]. Human-robot cooperation methods have previously been used to assist the surgeon during suturing [39, 136], however these methods overlook the effect on positional information and suturing motion by the unstructured nature of tissues, environmental unknowns, or instrument errors. For example, during robotic surgery it is common for the camera to be occluded by elements in the environment such as body material or parts of the robot's manipulators [154].

To address these shortcomings, in this Chapter we aim to develop a system framework and an interaction control strategy that implement human-robot cooperation on the da Vinci surgical system as an information exchange between a human operator and a robot agent [88], inspired by recent studies on collaborative human-human interaction [41, 150]. These studies revealed how the inconspicuous exchange of sensory information between two agents is beneficial to both, so that even the partner with superior sensorimotor control improves their performance from the interaction. Importantly, the partners share not just their current position but their motion plan, allowing each of them to continuously anticipate the other's actions and optimise their own movements accordingly [149].

To successfully interact with the human operator, the robot must predict their motion intention at different levels. At the task level, the robot first needs to identify the operator's desired high-level action. In robot-assisted surgery for instance, specific hand gestures, or *surgemes*, define interventions [7]. These surgemes can be used to infer the surgeon's task throughout the procedure (Figure 5.1). For example, in [111] online gesture classification was used to automate the camera motion. Similarly, in [120], a recurrent neural network was used to successfully switch between assistance modes based on general bimanual gestures. After recognising a surgeme, the robot can then estimate the operator's motion intention at a lower level to achieve seamless adaptation during collaborative tasks through a parameter, such as confidence [71, 88], which would allow to dynamically modulate the control over the manipulators, akin to the interaction observed between connected humans carrying out a simple tracking task [88, 149]. The works of [21, 60, 151, 169] suggest that incorporation of the human motion intentions and adaptation can minimise the operator's





effort. This system framework is summarised in Figure 5.2.

The *Intention Assimilation Controller (IAC)* algorithm of [151] can be used by the robot to identify the operator's motion plan and combine it with its own plan according to the task requirements and the robot's sensing, using a convex combination based on human and robot inputs. However, [151] did not specify the criteria for transitioning between these modes. This transition may be regulated based on the robot's confidence on its sensors [71, 88]. For instance, confidence scores can be used to avoid collision [23] or improve trajectory tracking [130], facilitating negotiation between human and robot based on information received. In this work, we will consider the camera data to be more accurate than forward kinematics for estimating position, but also more susceptible to occlusions and fluctuations.

This Chapter describes a human-robot cooperation framework with interaction control strategy integrating task-level prediction with Confidence-based Intention Assimilation Controller (C-IAC), implemented on the da Vinci surgical system. This four-armed robot with high-definition cameras offers to surgeons an immersive experience, but limited sensory information can make tasks such as suturing tedious and time consuming through repetitive actions and sustained attention to the dynamics of the environment [109]. We hypothesise that, by combining human and robot sensing and motion planning, one can obtain the benefits of sensory augmentation to enhance suturing performance [88] and comfort.

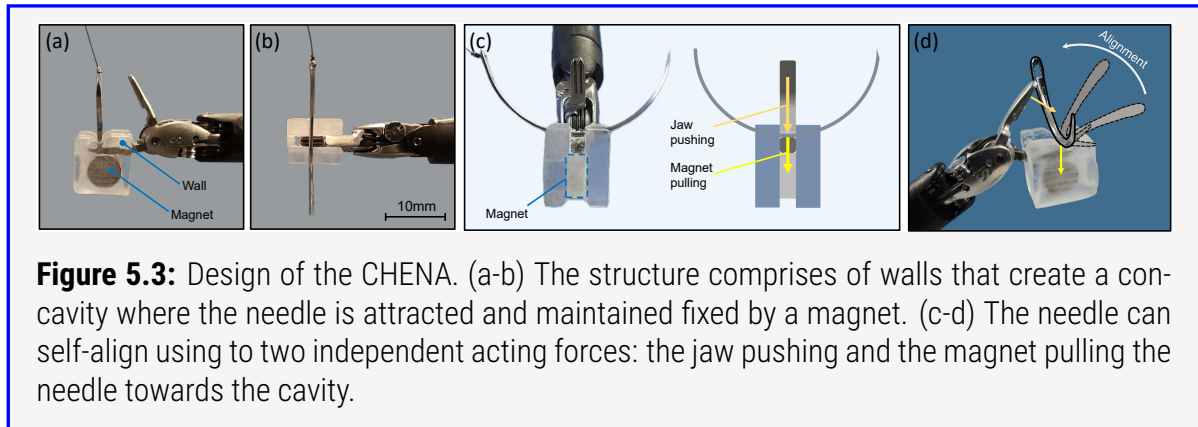
To test our hypothesis, we first developed a kinematic suturing dataset and trained a Transformer model to infer suturing activity in real-time on the dVRK by decomposing it into surgical gestures. We then designed an IAC-based controller with confidence level computed from Bayesian inference analysing kinematic and camera data. This controller was implemented to assist the operator in four key suturing surges. We also developed the Compact Holder for Enhanced Needle Alignment (CHENA), a mechanical device that can automatically and easily position the suturing needle orientation into a known position on the manipulator. Experiments with two groups of eight and nine users evaluated the system's performance and validated the interaction control approach. To our knowledge, this is the first implementation of human-robot cooperation for bimanual robotic suturing with use of real-time kinematic gesture recognition on a real dVRK using a Transformer-based model, which dynamically provides the correct assistance to manipulators using shared control paradigms.

The main novelties presented of this work are as follows:

- Robot-Camera calibration system for tracking of markers on a suturing task.
- Improved design for needle holders to facilitate needle grasping and reduce position uncertainty.
- Integration of human-robot cooperation with gesture recognition model for bimanual robotic suturing using the real dVRK.
- Implementation of a dynamic parameter to compute the confidence component based on Bayesian inference.
- Development of a confidence-based IAC controller used to teleoperate the da Vinci robot within a human-robot cooperation framework.
- User studies to demonstrate the performance of the C-IAC during the suturing task.

5.2 Compact Holder for Enhanced Needle Alignment

During suturing, operators often require multiple manipulations of the needle to achieve the correct and optimal orientation [43, 113]. This can be observed on the JIGSAWS dataset [42], where multiple hand-offs of the needle are performed before starting the insertion motion of suturing. Previous attempts in solving this problem include [137], which created a component whose geometry would allow the alignment of the needle and that could be easily attached to the existing surgical tool without requiring a complete replacement. However, the device can be difficult to use as it increases the size of the end-effector, making the grabbing of the needle awkward. Therefore, we propose a low-cost Compact Holder for Enhanced Needle Alignment (CHENA), which improves



on the previous designs in the literature [113, 137] by easing the grabbing and holding of the needle (Figure 5.3). The needle aligner in [137] was intended for autonomous suturing and required a large rear wall and catching area to guide the needle towards the grasping position. The necessity of these parts force the components to extrude over the opposite side of the surgical tool, limiting its maneuverability and maximum jaw open angle. In contrast, CHENA self-aligns and constrains the position of the needle without the need of large extrusion in the catching area through two possible mechanisms: the minimalistic geometry of the catching area and the use of a magnet to create a driving force for the needle to align even without closing the jaws. This design further improves on previous needle holders as it requires less mechanical effort to place the needle on the surgical tool regardless of the manipulator orientation.

5.3 Kinematic Gesture Recognition

Gesture recognition through kinematic data analysis has been explored using various machine learning models, including Long Short-Term Memory and Convolutional Neural Networks [6, 30, 156]. The Transformer model [160] improves upon existing models with its ability to process and analyse data in parallel with its novel attention mechanism. This architecture, which typically includes both an encoder and a decoder, has shown promising results in surgical gesture recognition [140]. Moreover, encoder-only configurations of the Transformer have proven effective in classification tasks [91, 167], suggesting their potential applicability in surgical gesture classification. Building on this foundation, we introduce a real-time gesture recognition system that uses the encoder-only Transformer model. This system integrates kinematic data from the two Patient Side Manipulators (PSMs) and two sigma.7 devices, processing a continuous window of real-time kinematic data to classify surgical gestures dynamically.

5.3.1 Suturing Dataset

A significant obstacle in the field of gesture recognition in robotic surgery is the lack of datasets with annotated demonstrations [7], with one of the few being JIGSAWS [42]. To accommodate our

specific surgical task setting, we have constructed and labelled a dataset of bimanual suturing motions made of kinematic data from the dVRK and hand interfaces on which we train our gesture recognition model. We refer to the dataset collected at Imperial College London in the Hamlyn Centre as *Suturing Task in Imperial's Tracking Collection for Hamlyn's Evaluation Set (STITCHES)*. The dataset includes continuous real-time kinematic data captured from the two sigma.7 devices and PSMs using the dVRK software. We record a kinematic dataset from a user performing suturing tasks on the dVRK along with video data at 20 Hz. The user has more than 900 hours of prior practical experience using the dVRK. The recorded kinematics captures movements from the two PSMs and two sigma.7s, documenting a total of 19 kinematic features per device. These features include the end effector position (3), rotation matrix (9), linear velocity (3), angular velocity (3), and the gripper angle (1). Gesture labels were manually added to the kinematic data using the video for reference, introducing an additional feature for a total of 77 features per time step. We collected 10 recordings in total with each recording having four throws. A throw, or suture, is defined as the combination of the surges to complete one stitch or loop on the wound.

5.3.2 Gesture Recognition Algorithm

The architecture of the proposed gesture recognition algorithm is shown in Figure 5.4. The kinematic window captures the linear velocity (3), rotational velocity (3), and gripper angle (1) from each of the four devices at a rate of 20 Hz, resulting in 28 features per time step. This window spans 60 time steps, equating to 3 seconds of data. The captured data is then processed through a classification network composed of two Transformer encoder blocks, each with their attention heads, followed by two fully connected layers of 64 units each and a softmax activation function for classification.

To improve stability in real-time, we incorporate two post-processing techniques: the application of an exponential moving average (EMA) window of size 10 over the output probabilities and a probability threshold of 0.8 applied to the averaged probabilities. The EMA window smooths fluctuations in probability values caused by noise and transient errors in gesture execution, leading to more stable and reliable gesture predictions. Additionally, the EMA gives more weight to the recent data points, making the system more responsive to new gestures compared to a regular moving average. The probability threshold ensures that only gestures with high-probability are recognised. To validate the model, we employed a 5-fold cross-validation approach, dividing the dataset of 10 recordings into five folds, each comprising a pair of recordings. The results of this validation are presented in Section 5.8.

5.3.3 Suturing Gesture Classification

Suturing is considered a fundamental task of laparoscopic surgery [127]. In [42], the authors distinguished 15 different gestures in robotic surgery (Table 5.1), 10 of which are used in robotic suturing. However, a simpler classification of suturing can be achieved with just 4 fundamental gestures [7,

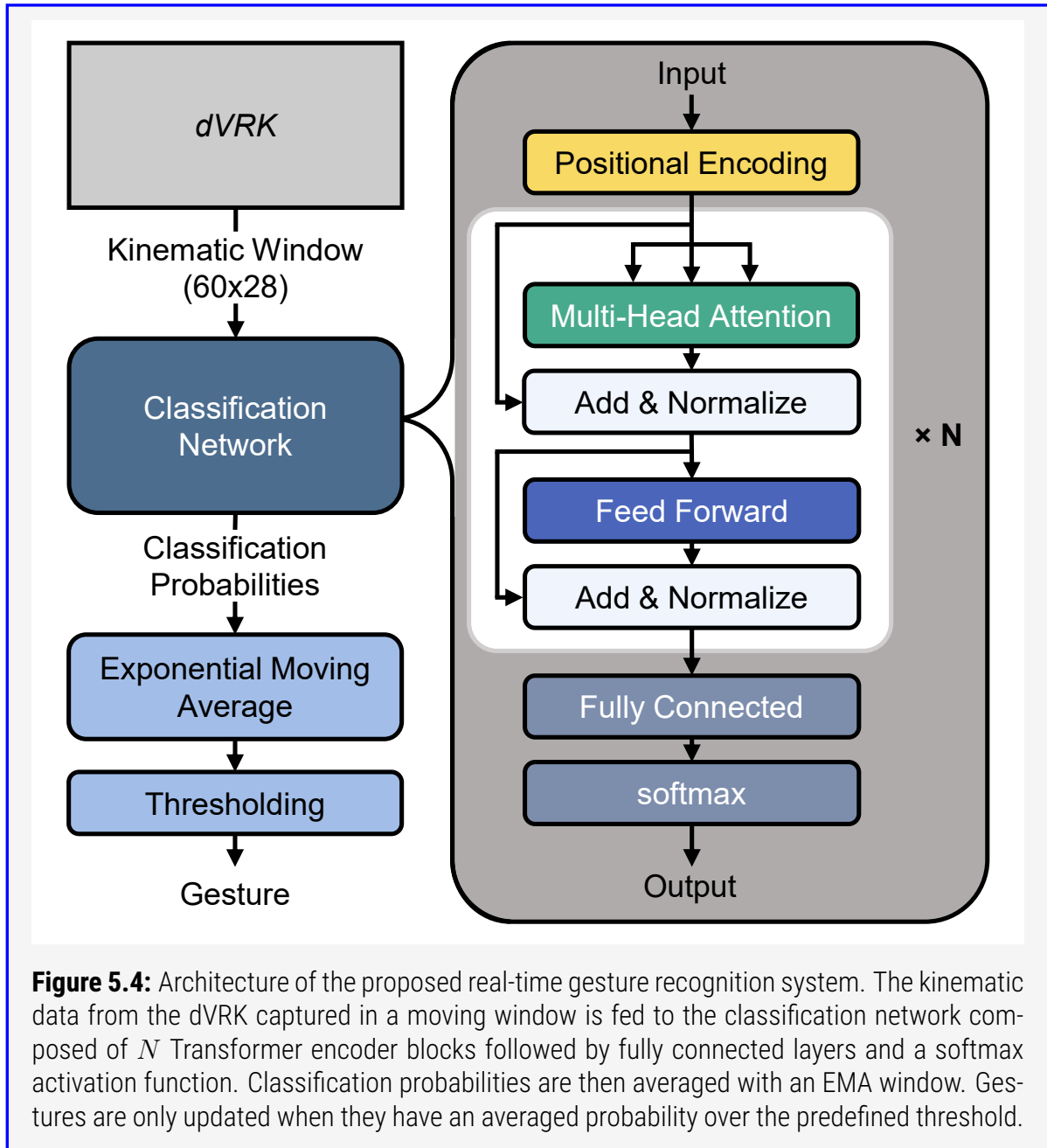
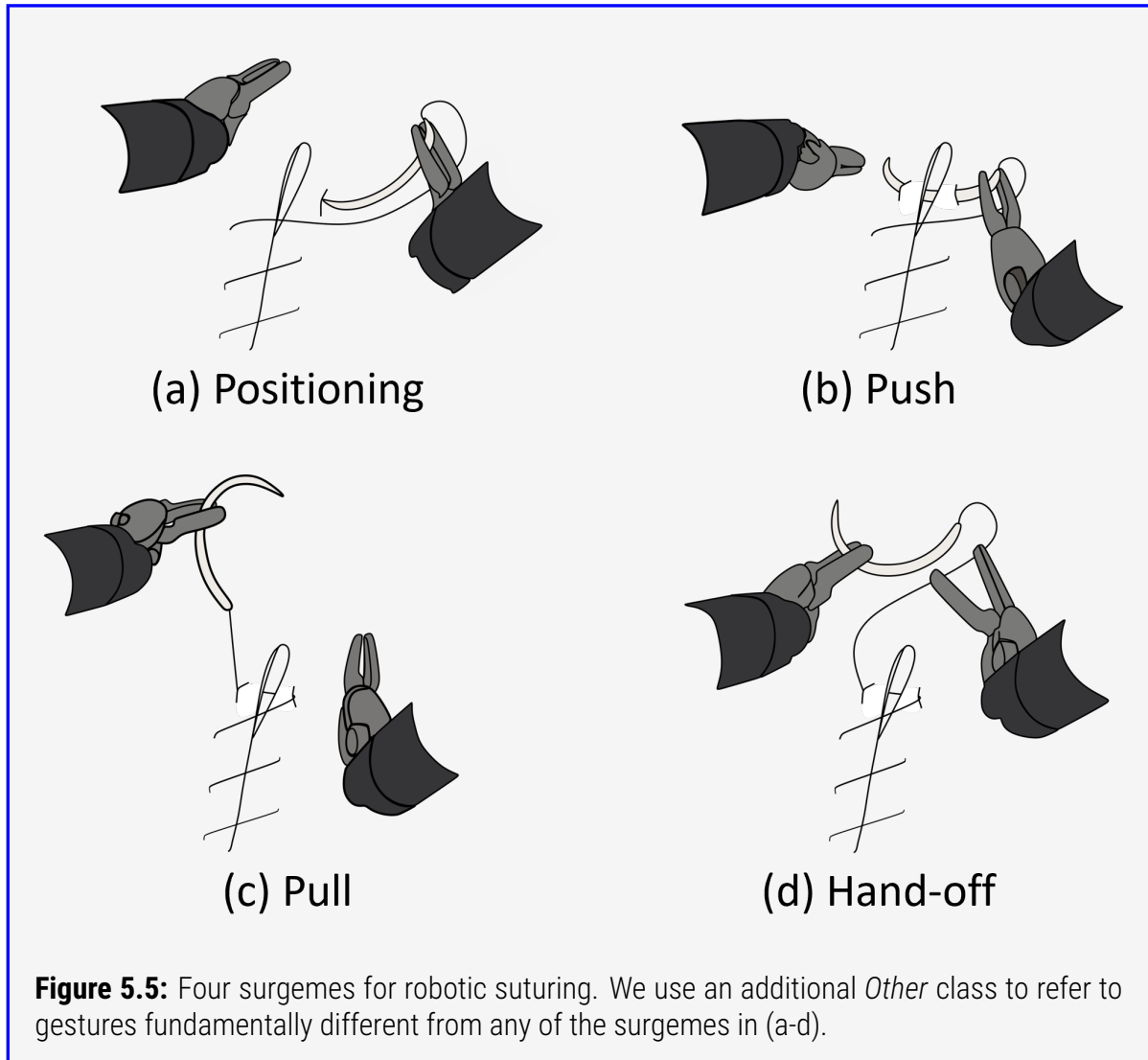


Figure 5.4: Architecture of the proposed real-time gesture recognition system. The kinematic data from the dVRK captured in a moving window is fed to the classification network composed of N Transformer encoder blocks followed by fully connected layers and a softmax activation function. Classification probabilities are then averaged with an EMA window. Gestures are only updated when they have an averaged probability over the predefined threshold.



137]. We show here that this classification can in fact be viable using the JIGSAWS dataset [42], which has been widely recognised as a benchmark dataset for kinematics based on the dVRK. We propose two different strategies for grouping the original gesture labels into 5 classes. The 5 classes include the fundamental gestures (a) *Positioning*, (b) *Push*, (c) *Pull*, and (d) *Hand-off*, as illustrated in Figure 5.5, and an additional *Other* class to group the remainder of the gestures in JIGSAWS.

Table 5.2 illustrates the two proposed classifications of the gesture labels G in JIGSAWS. For strategy 1, only the original gesture G is labelled as the corresponding surgeme with the rest labelled as *Other*. For strategy 2, gestures G that are deemed similar in motion to the four main gestures are grouped together and the rest is labelled as *Other*.

We argue that if a Transformer-based model can classify with similar accuracy using both strategies, then a simplification of the surgical task into four essential surgemes is in fact valid. We then can implement strategy 2 for the Transformer-based model that will be used in the human-robot cooperation system.

Table 5.1: Surgemes in JIGSAWS

Surgeme	Gesture Description
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to centre with needle in grip
G6	Pulling suture with left hand
G7	Pulling suture with right hand
G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Looseing more suture
G11	Dropping suture at end moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

Table 5.2: Proposed surgeme labels (a-d) compared to jigsaws

Surgeme	Strategy 1	Strategy 2
(a)	G2	G2, G5
(b)	G3	G3
(c)	G6	G6, G10
(d)	G4	G4, G8
Other	G1, G5, G8, G9, G10, G11	G1, G9, G11

5.4 Confidence-based IAC

5.4.1 Intention Assimilation Controller

The idea of the Intention Assimilation Controller (IAC) is for the robot to extract the human operator's motion plan from their interaction force, which can then be combined with the robot's motion plan [151]. Assuming that the robot has information about the force u_h applied by the operator and the system position x and velocity $\dot{x}$, IAC can predict the human operator's motion plan and induce a stable human-robot interaction [151] using the simple linear robot controller:

$$u_h = -L_1(x - \tau) - L_2\dot{x} \quad (5.1)$$

where L_1 , L_2 are the stiffness and viscosity matrices [59, 151], while the target τ is extracted from this equation during the movement by applying a Kalman filter to compute the predicted state estimate of the human target $\hat{\tau}_h$. Finally, the robot is controlled using a linear controller as defined in Equation 5.2. In this equation, the robot is using as target the convex combination of the estimated human target $\hat{\tau}_h$ and its own target τ_r , selected according to its task and sensing information:

$$\tau = \lambda\tau_r + (1 - \lambda)\hat{\tau}_h \quad (5.2)$$

where the confidence factor λ determines how much the robot relies on the information from the human relative to its own, affecting the assistance behaviour during the human robot interaction.

When $\lambda = 0$, the robot considers its own information unreliable and will follow completely the predicted human target position $\hat{\tau}_h$. Critically, $0 < \lambda < 1$ sets a cooperation between the human and robot control inputs, considering both of their targets. This configuration allows to benefit from the speed of reaching the predicted human target instead of the current human position and allows to adjust the confidence based on the sensory information detected by the robot. While when $\lambda = 1$, the robot disregards the human interaction and a coactive behaviour arises with the robot trying to achieve its own target. This assumes that the robot has all the necessary information, which allows it to optimally complete the task without being delayed by the teleoperation latency. When $\lambda > 1$, the robot can also display an antagonistic behaviour where it tries to counteract the human control input to achieve its own target. Since we aim to support the user, the system needs to ensure $0 \leq \lambda \leq 1$. The implementation of the confidence-based IAC compared to traditional teleoperation is illustrated in Figure 5.6. Being able to continuously and dynamically adjust λ is key to achieving a smooth and situation-specific interaction between user and robot. The proposed method allows to optimise the control authority in real-time according to the system's confidence on either the human or robot performance.

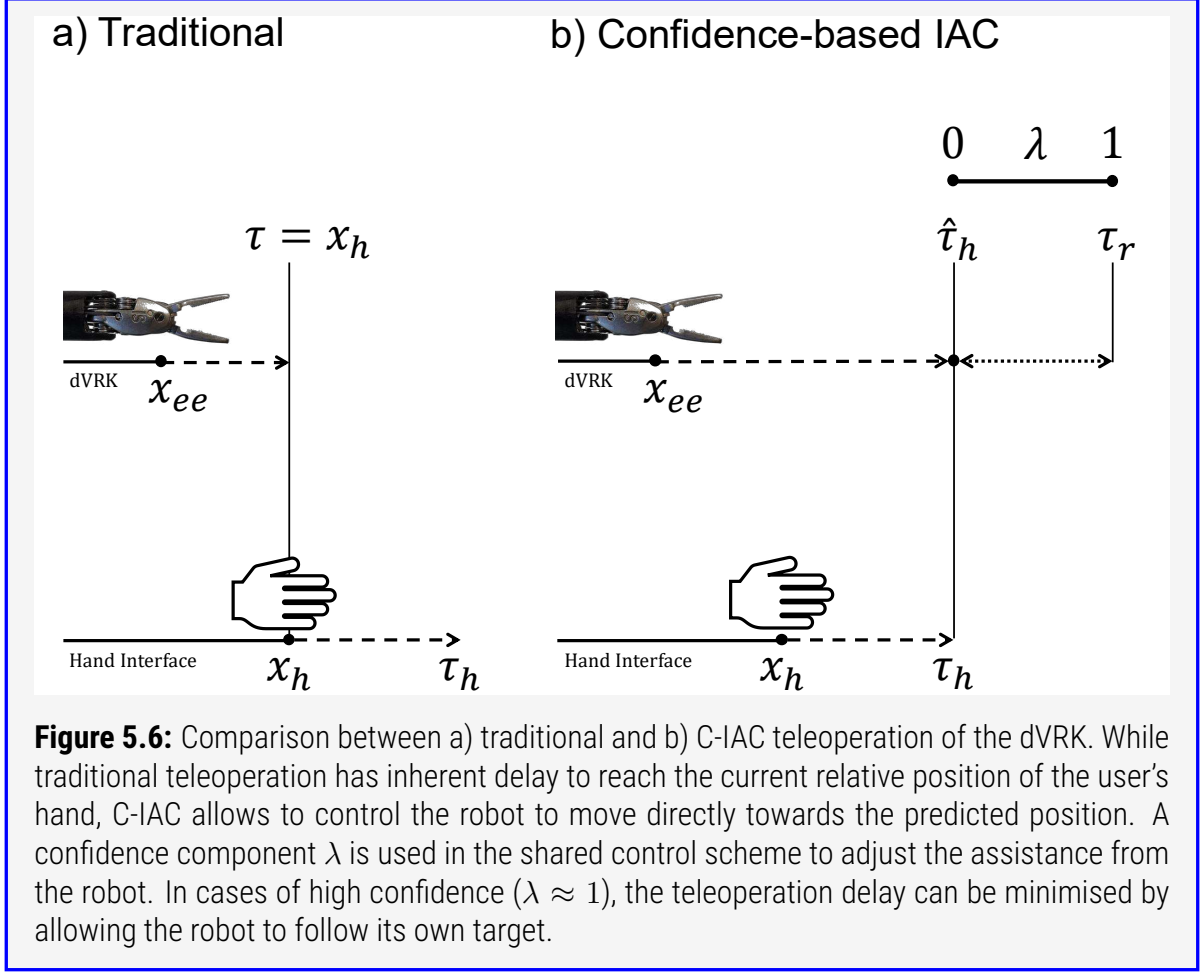


Figure 5.6: Comparison between a) traditional and b) C-IAC teleoperation of the dVRK. While traditional teleoperation has inherent delay to reach the current relative position of the user's hand, C-IAC allows to control the robot to move directly towards the predicted position. A confidence component λ is used in the shared control scheme to adjust the assistance from the robot. In cases of high confidence ($\lambda \approx 1$), the teleoperation delay can be minimised by allowing the robot to follow its own target.

5.4.2 Confidence with Bayesian Inference

We define two sensory information for the robot: vision and proprioception. Visual information is acquired by tracking the markers placed on the robot's manipulator and phantom, specifically one keydot kd and one ChArUco ch marker, respectively. Proprioception, on the other hand, is estimated using the robot joint angles. While proprioception information is consistently available, visual information can be prevented if obstructions occur, leading to loss of tracking.

In this work, we define confidence as the trust towards the sensory data exchanged between the different sensors of the system. The work in [46] demonstrated that Beta distributions are effective in inferring the human agent's trust in the robotic agent's performance by analysing the robot's performance history. We adopt Bayesian inference to calculate the trust in sensory data continuously, giving more importance to the most recent observations. Specifically, vision is modeled probabilistically using the Beta distribution, while proprioception is not probabilistically modeled as it's treated as always available but less reliable. The Beta distribution is represented with its shape parameters, denoted as $B(\alpha, \beta)$, and they are updated empirically over time and dynamically as new sensory signals are received. Bayesian inference fuses current visual performance data with prior confidence using the Beta posterior, with the Bayesian posterior mean acting as a soft reliability

weight between robot and human targets, integrating over time via recursive updates.

In this work, we use the Bayesian inference explored in [46] to estimate the confidence, or trust, of the robot agent in the visual information received. We define the performance on the i th time step as $p_i \in \{0, 1\}$, where $p_i = 1$ corresponds to a successful tracking of the markers, and $p_i = 0$ a failure. The system's performance history in a given window of length n is then denoted as $P = \{p_1, \dots, p_n\}$. Effectively, the visual information yields a boolean message for both the keydot marker or ChArUco that indicates if in a given time step the information is received or absent. This determines an arbitrarily assigned performance: 0 if the system does not have confidence, and 1 if the system has confidence (i.e. $kd \vee ch$, illustrated in Table 5.3).

Table 5.3: Visual Information Performance Assignment

kd	ch	Performance p_i	Gain w_{p_i}
✗	✗	0	w_0
✓	✗	1	w_1
✗	✓	1	
✓	✓	1	

The Beta distribution dynamics [46] are used to model the system's trust in the visual tracking for two main reasons:

1. The distribution is bounded within the the interval $[0, 1]$ which is consistent with the confidence boundaries to achieve the interaction behaviour needed (Figure 5.6).
2. The distribution dynamics enables the property of influencing the confidence at the current moment i by previous time steps $i - 1$ [86].

As the marker tracking performance is received, at every i th time step the system's trust dynamics is calculated as:

$$B_i(\alpha_i, \beta_i) \quad (5.3)$$

$$\begin{aligned} \alpha_i &= \begin{cases} \alpha_{i-1} + w_1, & \text{if } p_i = 1 \\ \alpha_{i-1}, & \text{if } p_i = 0 \end{cases} \\ \beta_i &= \begin{cases} \beta_{i-1} + w_0, & \text{if } p_i = 0 \\ \beta_{i-1}, & \text{if } p_i = 1 \end{cases} \end{aligned} \quad (5.4)$$

where α_i and β_i are the shape parameters of the distribution, and w_{p_i} are the corresponding gains. The posterior mean is then used to estimate the system's confidence:

$$\lambda = E(B_i) = \frac{\alpha_i}{\alpha_i + \beta_i} \quad (5.5)$$

Equation 5.5 ensures a continuous value within $[0, 1]$ to assess the confidence of the system in its own sensory information. The value can also be scaled to meet all interaction behaviours described in Section 5.4.1, depending on the task needs.

It is important to note that the confidence defined is highly dependent on the visual information and less on proprioception. In fact, we assume that, when available, visual information of the end-effector position is more reliable than proprioception after a robust hand-eye calibration of the camera. Since the dVRK is a cable-driven system, the manipulator is subject to large positioning error due to the elasticity of the cables and long kinematic chains [48, 113]. For our specific human-robot cooperation framework, we empirically tuned the gains w_{p_i} and assigned a positive performance whenever *kd* or *ch* is detected, and a zero when they are not, as shown in Table 5.3.

5.4.3 Shared Control Paradigms

We propose the following shared control paradigms based on the generalisation of the suturing task into the four fundamental surgemes illustrated in Figure 5.5. Note that the constraints and targets implemented in the shared control paradigms proposed can be chosen arbitrarily and ours is just one possible solution after consulting a neurosurgeon from Imperial College London for insights on suturing. We employ the following terminology to represent quantities measured relative to the task frame depicted in Figure 5.7 with respect to the manipulator tracked with the keydot marker:

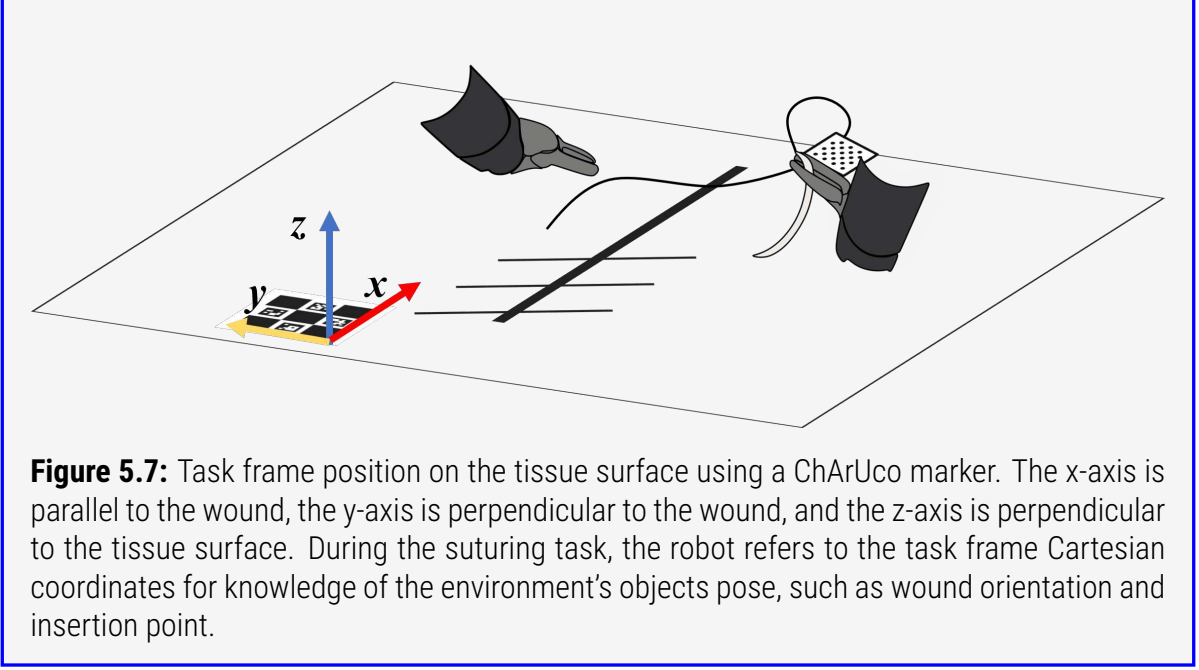
- $\tau = (x, y, z) \in \mathbb{R}^3$ is target position computed by the C-IAC.
- $\lambda = (x, y, z) \in \mathbb{R}^3$ is the confidence parameter applied on the end-effector.

5.4.3.1 Positioning

If we imagine the curved needle as a semicircle surface, during *Positioning* the user is trying to align the needle's plane to be perpendicular to the wound. During this gesture, the needle tip should be touching the insertion point at a fixed height from the tissue. The insertion point can be in different parts along the tissue near the wound. Therefore, the manipulator should be free to move along a plane parallel to the tissue surface and at a fixed target height.

$$\tau = (\tau_h^x, \tau_h^y, \tau_h^z), \quad \lambda = (0, 0, \lambda^z) \quad (5.6)$$

As the needle plane needs to be perpendicular to the wound for an effective insertion of the needle, we apply a further control option to allow the human to control the DoF for the orientation of the



end-effector autonomously through a pedal.

5.4.3.2 Push

During the *Push* surgeme, the needle plane should remain perpendicular to both the tissue surface and the wound direction. Movements along the wound direction are particularly dangerous as tissues can be lacerated and damaged. Therefore, we restrict movements of the manipulator along the wound axis, while allowing free motion in the other Cartesian coordinates. This allows to correctly make a circular motion that can permit the tip of the needle to emerge from the other side of the wound.

$$\tau = (\tau^{x_j}, \tau_h^y, \tau_h^z), \quad \lambda = (\lambda^x, 0, 0) \quad (5.7)$$

where x_j refers to the j th entry point position in x during the suturing task. To ensure that the needle is maintained perpendicular to both wound and tissue surface, we make the control for the end-effector orientation autonomous as done in the *Positioning* phase.

5.4.3.3 Pull

Rakita et al. [120] described that when two hands are working in close proximity, with one hand relatively stationary and the other with higher translational and rotational velocity, keeping the stationary hand stable can help the user to successfully complete the task. As this scenario is equivalent to pulling during suturing, we implement a similar type of constraint to the right hand by keeping it steady above the tissue surface. A steady right hand above the tissue can allow the user to con-

centrate on the pulling motion with the left hand. To ease the next throwing motion, we also set the robot target to the next entry point position.

$$\tau = (\tau^{x_{j+1}}, \tau^y, \tau^z), \quad \lambda = (\lambda^x, \lambda^y, \lambda^z) \quad (5.8)$$

5.4.3.4 Hand-off

The hand-off of the suturing needle usually requires a coordination between the two manipulators to allow a correct grasping. However, with the CHENA this is greatly simplified as the needle self-aligns to the end-effector. Therefore, we simply make sure that the *Hand-off* gesture occurs in proximity of the next entry point to facilitate the next *Positioning* gesture, but we allow the PSM1 manipulator to freely move towards the PSM2.

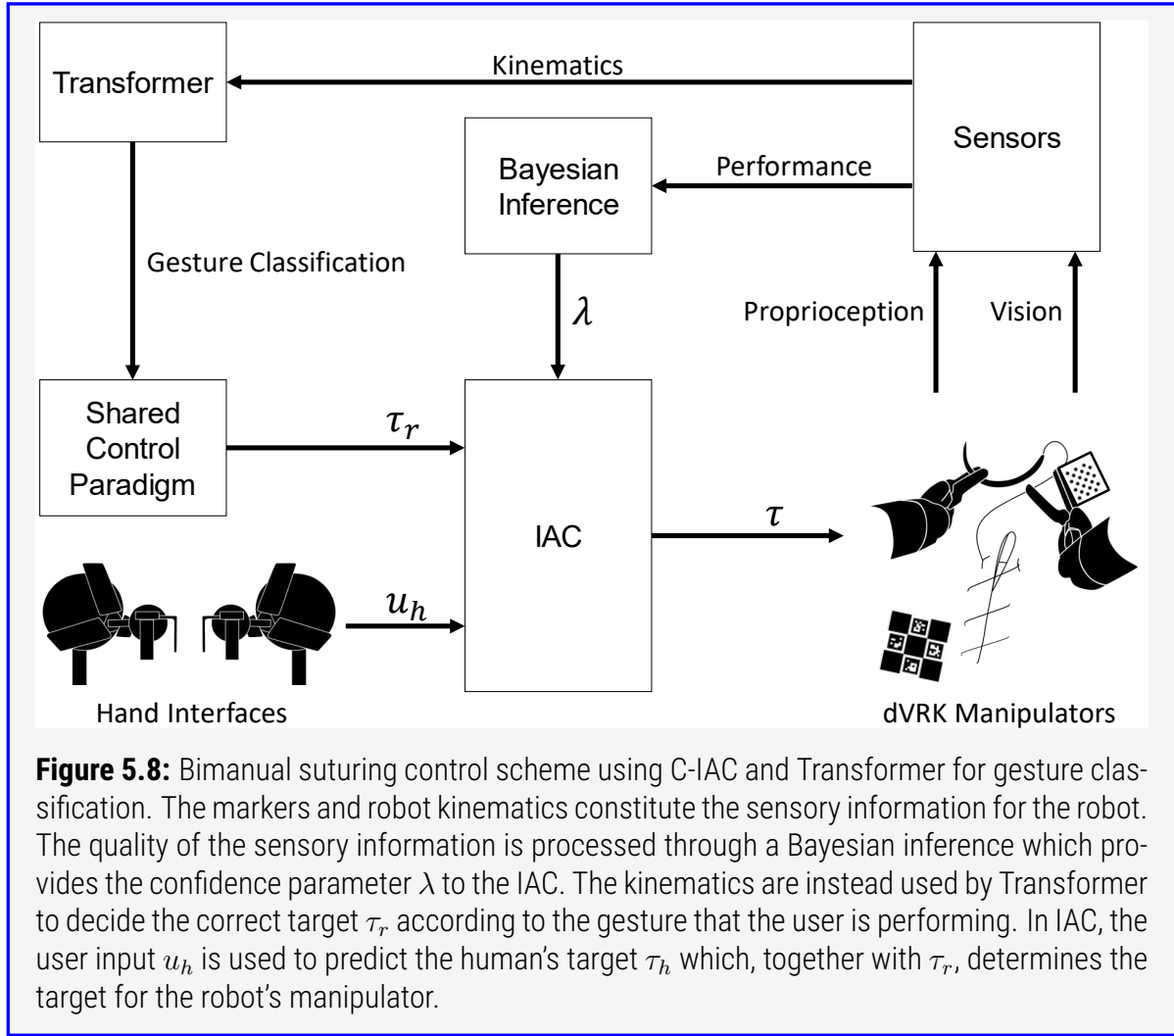
$$\tau = (\tau^{x_{j+1}}, \tau_h^y, \tau^z), \quad \lambda = (\lambda^x, 0, \lambda^z) \quad (5.9)$$

5.4.3.5 Other

In the classification of *Other* hand movements, the user is passing kinematic movements to the robot's PSMs that do not resemble any of the first four surges. To account for the unstructured and unpredictable nature of these movements, we set the robot to not apply any constraint on the manipulator, which corresponds to a the $\lambda = 0$ for all Cartesian coordinates.

$$\tau = (\tau_h^x, \tau_h^y, \tau_h^z), \quad \lambda = (0, 0, 0) \quad (5.10)$$

Notice that it would be possible to set hard constraints to impose forbidden movements to the user for each control paradigm, but interaction forces with the human user may disrupt the control. Instead, the C-IAC allows to create softer and more dynamic constraints by predicting the user intention and considering the system's confidence throughout the task. The complete system framework is shown in Figure 5.8.



5.5 Experimental Setting

For the experiments, we use the dVRK from Intuitive Surgical. The robot includes two PSMs: right PSM1 with DeBakey Forceps and left PSM2 with Large Needle Drivers. Only the PSM1 is equipped with a CHENA since we assume that the users are right handed. The needle used is a 1/2 circle, 24 mm chord length with triangular cutting point. The robot is controlled in Cartesian space using the dVRK software. Two sigma.7 hand interfaces are used to control the robot, using the TCP/IP protocol to communicate with the dVRK. Similarly to [64, 132], we renounce to the binocular vision system and provide instead a 2D high resolution display to the human user using a RealSense camera, with a 50 ms delay introduced by the teleoperation system to observe the effects on performance between different controllers [126]. One element that can contribute to increase the complexity and unpredictability of the task and surgeon's performance is latency, which is considered one of the biggest hurdle to enable telepresence [147]. [108, 126, 166] showed that a latency below 200 ms is ideal for telesurgery, but as little as 50 ms can already cause effects on the users performance. Our setting aims to highlight the difference in sensory information between human and

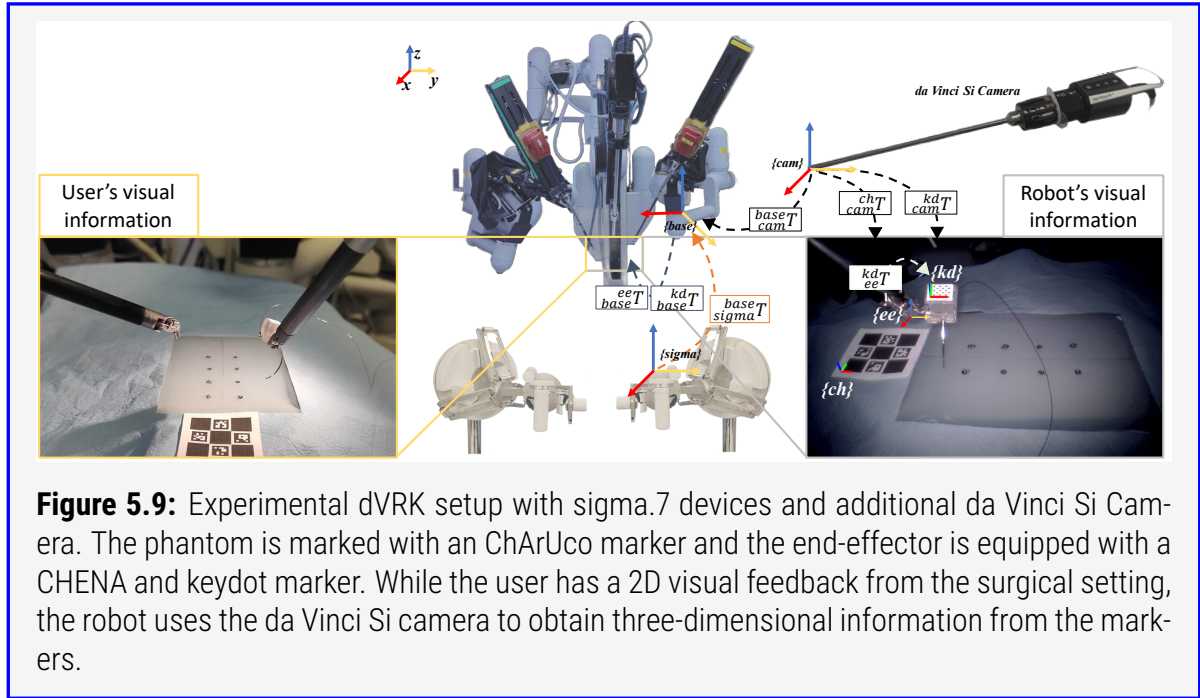


Figure 5.9: Experimental dVRK setup with sigma.7 devices and additional da Vinci Si Camera. The phantom is marked with an ChArUco marker and the end-effector is equipped with a CHENA and keydot marker. While the user has a 2D visual feedback from the surgical setting, the robot uses the da Vinci Si camera to obtain three-dimensional information from the markers.

robot, which will be reflected in the performance results. We maintain pedals options for functions such as clutching and control of the end-effector orientation.

For the suturing phantom, we create a layer of about 1 mm thickness using Ecoflex 00-50. The phantom layer emulates the setup used in JIGSAWS to ensure comparable conditions, which are well established in the literature for gesture recognition. Similarly, the wound and entry-points are manually marked, measured and recorded relative to the ChArUco marker fixed on the phantom. A keydot marker is also attached to the CHENA to track the PSM1 hand and needle pose. This setup implies that while C-IAC is used to teleoperate both PSMs, only the one with the keydot marker can employ confidence-based properties, since it is both tracked kinematically and visually. Both markers in this setting are tracked using a da Vinci Si endoscope, which provides high resolution images to the robot.

5.6 Robot-Camera Calibration

From the sigma.7s and the dVRK manipulators we can obtain the poses of the user's hands and end-effectors as ${}^{sigma}_{baseS}T$ and ${}^{ee}_{base}T$, respectively, where *sigma* is the hand interface, *ee* is the PSM's end-effector, and *baseS* and *base* are the sigma.7 and dVRK base coordinates.

5.6.1 Sigma.7 to dVRK

To map the pose from the hand interfaces to the PSMs, it is necessary to transform the position and orientation of both hands. This task can be challenging as the sigma.7 and the PSM have different joint limits [132]. We implement the following transformation to match the hand pose:

$${}_{baseS}^{sigma}T \cdot {}_{base}^{baseS}T \cdot {}_{base}^{ee}T^{-1} = {}_{ee}^{sigma}T \quad (5.11)$$

Notice that while in traditional teleoperation we scale the absolute position of the sigma.7 hand interface, when C-IAC is implemented, the translations in the transformation matrix are passed to the dVRK as depicted in Figure 5.6 (b).

5.6.2 Hand-Eye Calibration

The system comprises of a camera that tracks a marker on the robot's end-effector, as shown in Figure 5.9. To solve the robot-sensor calibration problem, we employ the hand-eye calibration method [138]:

$$AX = YB \quad (5.12)$$

where A is the transformation from the da Vinci Si Camera to the keydot marker (i.e. ${}_{cam}^{kd}T$), B is the transformation from global robot coordinates to end-effector (i.e. ${}_{base}^{ee}T$), while X and Y are the unknown fixed homogeneous transformations between marker-end-effector and camera-base, respectively.

To calculate the unknowns X and Y of Equation 5.12, which correspond to ${}_{kd}^{ee}T$ and ${}_{cam}^{base}T$, respectively, we record 50 poses of the robot arm with the keydot marker and use a separable solution [138]. We propose to use two sensory information channels to estimate the robot's tool position: vision, based on the data arriving from the markers tracking, and proprioception, derived from the kinematics of the robot to estimate the keydot pose. In the following paragraphs, we detail how the manipulators are brought to the desired pose.

5.6.2.1 Vision

The system uses an endoscope of a da Vinci Si to track a keydot marker on the CHENA and a ChArUco marker positioned on the phantom. From the continuous visual tracking of the keydot and ChArUco, we use a separable solution to estimate the following transformations:

$${}_{kd}^{cam}T, {}_{ch}^{cam}T \quad (5.13)$$

where kd and ch refer to the keydot and ChArUco markers, and cam denotes the frame for the da Vinci Si camera.

5.6.2.2 Proprioception

As the dVRK manipulators are cable-driven systems, their position estimates can become inaccurate over time [48, 113]. Therefore, the proprioception proposed relies on the update of the keydot pose whenever vision is available. To distinguish the keydot position estimated by the robot kine-

matics alone or camera alone, we name them $kd1$ and $kd2$, respectively. Assuming that visual information is more representative of the reality and less prone to errors than proprioception, we can calculate the offset compared to when visual tracking of the robot's marker is present to compensate for the uncertainties derived by the cable-driven system as:

$${}_{kd1}^{cam}T^{-1} \cdot {}_{kd2}^{cam}T = {}_{kd2}^{kd1}T \quad (5.14)$$

We then calculate the keydot position from proprioception:

$$(B \cdot Y)^{-1} \cdot X = ({}_{base}^{ee}T \cdot {}_{cam}^{base}T)^{-1} \cdot {}_{kd1}^{ee}T = {}_{kd1}^{cam}T \quad (5.15)$$

$${}_{kd1}^{cam}T \cdot {}_{kd2}^{kd1}T = {}_{kd2}^{cam}T \quad (5.16)$$

Notice that the final transformation matrix obtained uses $kd2$ which is the keydot estimated by the camera. However, ${}_{kd2}^{cam}T$ is solely reliant on the robot's kinematics and its $kd2$ can be considered as the robot's "memory" of the keydot pose when visual tracking ceased.

As the ChArUco is used to estimate the phantom's position for the task, we compute the offsets from the keydot marker to the desired pose for both vision and proprioception as:

$${}_{cam}^{kd}T \cdot {}_{ch}^{cam}T \cdot {}_d^{ch}T = {}_d^{kd}T \quad (5.17)$$

where d is the desired pose position and orientation relative to the camera. Therefore, this allows the robot to obtain the offsets to estimate its pose in the real-world environment.

5.7 Experiments

The experiments were approved by the College Research Ethics committee of Imperial College London (21IC7042). Participants were informed about the experiments' purpose and protocol, and signed a consent form before starting. A first set of experiments was carried out to validate our Transformer-based model, the STITCHES dataset, and to measure the needle pose estimation error. Following this, we conducted user study experiments to compare traditional teleoperation against C-IAC across two tasks: (i) target reaching, to systematically analyse the controller's performance, and (ii) four surgical throws, to represent a key surgical function. To ensure a fair comparison, the PSM1 was equipped with a CHENA throughout all experiments. We evaluated the controllers' performances by analysing completion time and user satisfaction, which was assessed using a NASA TLX questionnaire.

5.7.1 Target Reaching User Study

The suturing task involves a relatively complex and unstructured environment that can affect a detailed evaluation of the controllers used. To determine the efficacy of C-IAC compared to traditional teleoperation in a standard manner, we design a target reaching experiment that consists of *Positioning* and *Push* gestures. We recruited 8 participants (1 female and 7 males, all right handed, age 23-30) who had little to no experience in teleoperation or robotic suturing. After 10 minutes of familiarisation with the dVRK setup and suturing movements, the subjects are tasked to control the PSM1 from a fixed starting position and reach four different entry points in the phantom. At each entry point, the reaching movement is considered completed upon needle insertion. The same starting pose is used for each of the four reaching movements, with entry points positioned on a straight line at {15, 30, 45, 60}mm from the starting position (Figure 5.10), thus challenging the user's depth perception.

For this experiment, we compare two types of teleoperation: traditional and C-IAC. Each of the controllers is tested on four subjects. In order to set a consistent setting and obtain comparable results between the participants, the confidence component λ in the C-IAC is gradually increased from 0 to 0.8 using a linear function instead of the method described in Section 5.4.2. This allows us to ensure that the same type of assistance is implemented between the trials that use that type of controller. We use Equation 5.7 as control paradigm throughout the trials with C-IAC. Notice that control paradigm and the maximum value of λ are chosen arbitrarily based on the context of the experiment and to allow the human operator to always retain some control.

5.7.2 Suturing Task User Study

To evaluate the benefits of C-IAC among 9 users (all males, age 21-41, all right handed) with different suturing skill levels, we compared the proposed controller against traditional teleoperation using a full suturing task with four entry points that requires four suturing throws. For this experiment, we aim to both examine the performance of all users with the system and compare participants of different skill levels: novice, intermediate, and expert. The skill level refers to their experience in suturing, but none of the subjects had experience with surgical robots. The novices (7 subjects, age 21-29) do not have any medical experience. The intermediate (1 subject, age 23) is a fifth year medical student. The expert (1 subject, age 41) is an neurosurgeon with practice in the medical field. Each participant trained on the system for eight sessions, split between the two controllers, of 20 minutes each before carrying out the experiments using traditional teleoperation and C-IAC, in a randomised order.

A clinical criteria for a successful suturing includes maintaining perpendicularity of the needle plane to the wound and skin [43, 113]. Since the needle is kept perpendicular to the manipulator using the CHENA, we tracked the manipulator orientation through the keydot marker which allowed us to assess if the needle was maintained perpendicular during the task. Additionally, we also mea-

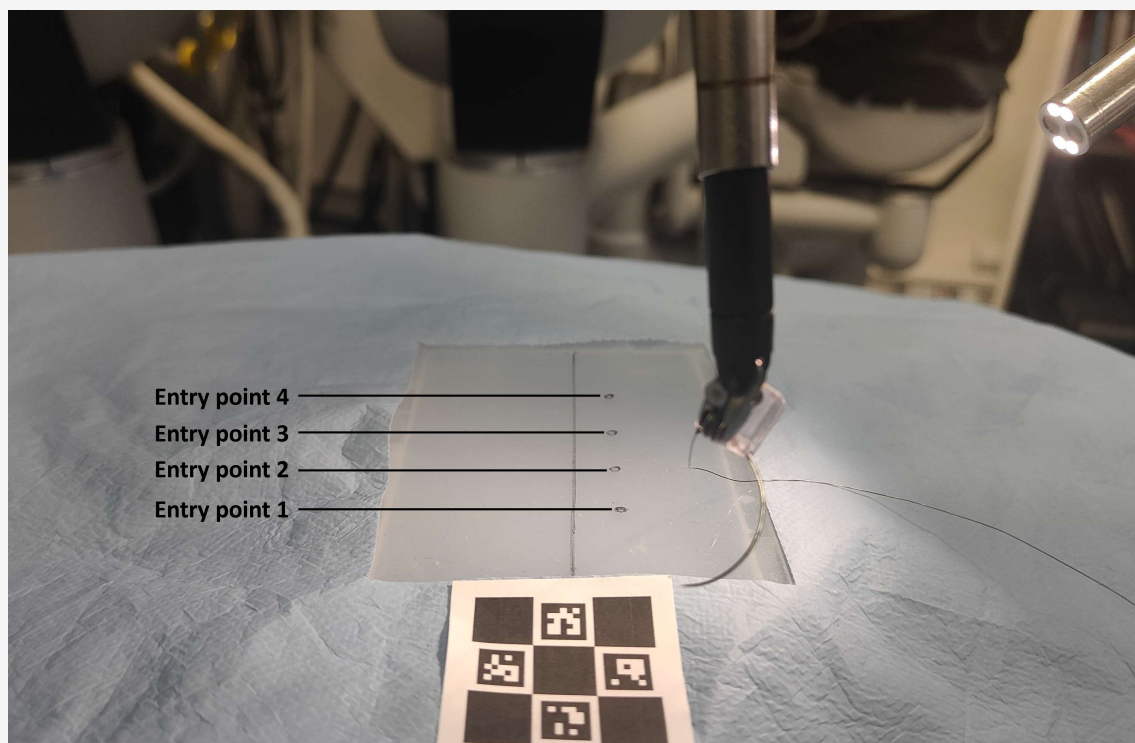


Figure 5.10: Visual feedback during target reaching user study. The users perform *Positioning* and *Push* at each entry point. For each target reaching movement, the manipulator is reset to the same starting position.

sured the average accuracy of the Transformer model throughout the task and the time required to complete it. At the end, the participants were asked to fill out a NASA TLX questionnaire to analyse perceived task load and their satisfaction with the system.

5.8 Results

5.8.1 Transformer for Two Classification Strategies

The Transformer model is applied on the two strategies described in Section 5.3.2 to demonstrate the feasibility of reducing the suturing task into four fundamental surgeme motions, along with one for general movements. Figure 5.11 (a-b) compares the confusion matrices for the two strategies. The Transformer-based model achieves an overall accuracy of 77.67% and 76.72% for strategy 1 and strategy 2, respectively. Furthermore, the confusion matrices exhibit a consistent trend across the two strategies, with some misclassifications between the surgemes *Positioning* and *Push*, which are both predominantly right-hand gestures. The class *Other* has the lowest accuracy in both strategies, while *Push* was the highest.

The similarities between the results obtained indicate that the robotic suturing task can indeed be generalised with only four fundamental surgemes as in strategy 2, which was thus used for

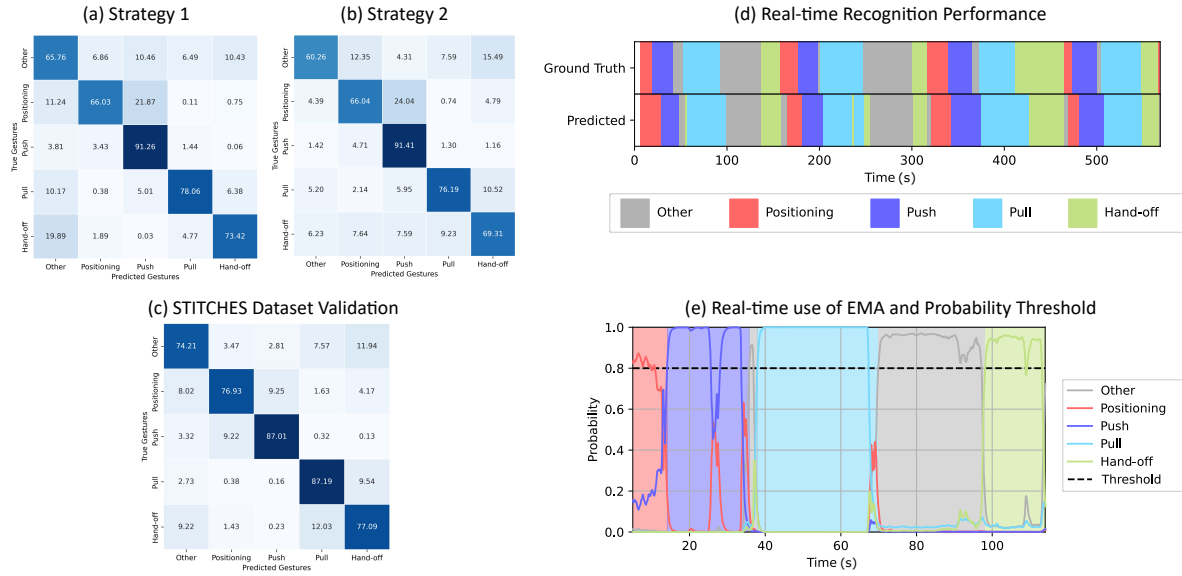


Figure 5.11: Use of Transformer model and STITCHES dataset validation. (a-b) Transformer gesture classification using labelling strategy 1 and strategy 2. (c) 5-fold cross validation average accuracy of STITCHES dataset. The overall accuracy is 81.19% using the proposed labelling for the suturing task. (d) Real-time sample comparison between ground truth and predicted gesture. (e) Real-time suturing throw sample with use of EMA and probability threshold.

subsequent experiments.

5.8.2 STITCHES Dataset Validation

Training, validation, and experiments of our Transformer-based model were conducted on the STITCHES dataset consisting of 10 recordings. With the implementation of the proposed post-processing techniques (i.e. EMA and probability threshold), the model demonstrated a classification accuracy of 81.19%. This result is shown in Figure 5.11 (c-d), which illustrates the confusion matrix obtained during the validation and the comparison between the actual gestures and the predicted in a sample. Figure 5.11 (e) provides a graphical representation of the post-processing technique applied in real-time during one suturing throw in one sample of the dataset.

5.8.3 Needle Pose Estimation Error

Due to the size and shape of the suturing needle, it is hard to place markers to establish a ground truth and estimate its pose through visual tracking [113, 137]. However, with the CHENA, the plane of the needle can be approximately considered in a fixed orientation perpendicular to the manipulator's gripper. Therefore, to compute the needle pose error, we compare the pose of the CHENA between the estimation received from visual tracking and the kinematics of the robot. In Table 5.4, we move the robot to 80 random positions and calculate the mean and standard deviation for the position and orientation between the two poses estimated.

Table 5.4: Needle Pose Estimation Error Results

DoF	Position (mm)			Orientation (deg)		
	x	y	z	Roll	Pitch	Yaw
Mean	0.65	0.93	0.84	2.51	2.30	1.24
Std.	1.00	0.87	0.68	2.73	1.83	1.38

5.8.4 Target Reaching User Study

As the reaching movements end with the insertion of the needle into the skin, we can evaluate the time required by the participants to perform accurate insertions. As the data for the traditional teleoperation were not normally distributed, a Wilcoxon rank-sum test was used to compare the two controllers, which revealed a difference between the two groups ($p = 0.004$). The average time required by the participants for each entry point and the total average time for each controller are reported in Table 5.5. For each entry point, the C-IAC decreased the average time to reach the target and perform a precise insertion, with a total average and standard deviation of $(\sigma, \mu) = (36.1, 15.4)$, and $(61.9, 39.6)$ for traditional control.

Table 5.5: Duration of Target Reaching (in seconds)

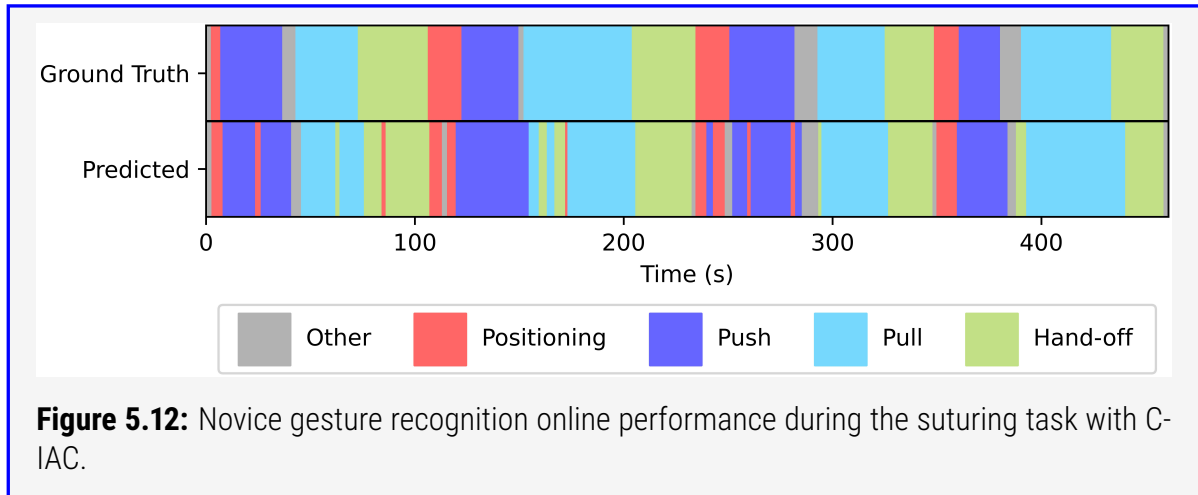
Controller	Traditional	C-IAC
Entry point 1	61 ± 6.8	25 ± 5.4
Entry point 2	61 ± 33.3	45 ± 24.4
Entry point 3	41 ± 10.6	39.5 ± 13.5
Entry point 4	84.8 ± 72.8	35 ± 9.8
Total	61.9 ± 39.6	36.1 ± 15.4

5.8.5 Suturing Task User Study

Table 5.6: User Study Suturing Task Performance based on Participant Suturing Experience

	Novices		Intermediate		Expert	
	Traditional	C-IAC	Traditional	C-IAC	Traditional	C-IAC
Avg. $\perp$ Error in Push (°)	28.83±10.11	7.27±5.26	21.98	4.25	34.47	2.19
Real-time Classification Accuracy (%)	—	65.50±15.25	—	67.37	—	67.17
Avg. Positioning Time (s)	26.31±18.99	19.94±9.23	24.82	22.87	36.95	32.93
Avg. Push Time (s)	43.07±18.97	38.05±21.95	42.00	47.83	47.33	40.07
Avg. Pull Time (s)	76.91±40.68	48.26±17.35	65.65	50.97	87.10	66.12
Avg. Hand-off Time (s)	54.09±19.21	35.29±14.17	36.86	27.68	74.84	49.66
Avg. Throw Time (s)	222.70±73.80	155.94±40.41	214.66	181.15	290.27	218.46
Total Suturing Time (s)	892.18±204.53	652.55±158.71	859.01	724.95	1161.51	873.87

During the user study with subjects of different suturing skill levels, the participants (i.e. novice,



intermediate, and expert) performed four consecutive suturing throws. Their performance is illustrated in Table 5.6. The results demonstrate the superiority of C-IAC, yielding a smaller perpendicular error and reduced duration in all surges composing the suturing task. Statistical analysis was carried out on the performances of the 7 novices. As the data is not normally distributed, we use Wilcoxon Signed Rank test to compare them, which revealed a significant improvement with C-IAC in the perpendicular error during *Push* ($p < 0.001$). Furthermore, an analysis of the singular throw time indicates that C-IAC yields faster performance than traditional teleoperation ($p < 0.001$), in particular during *Pull* ($p < 0.001$) and *Hand-off* ($p < 0.001$).

The responses to the NASA TLX questionnaire from the 7 novices are analysed in Figure 5.13, which shows that the workload is perceived to be smaller with C-IAC than traditional teleoperation. After ensuring the normality of the data, we performed t-test comparison to evaluate the scores given by the 7 novice subjects. The results are summarised in the box charts of Figure 5.13. The statistical analysis found that the participants felt to perform better with C-IAC ($p < 0.01$), while requiring less effort ($p < 0.05$). Similar tendencies are exhibited for temporal demand ($p = 0.079$) and frustration experienced ($p = 0.083$).

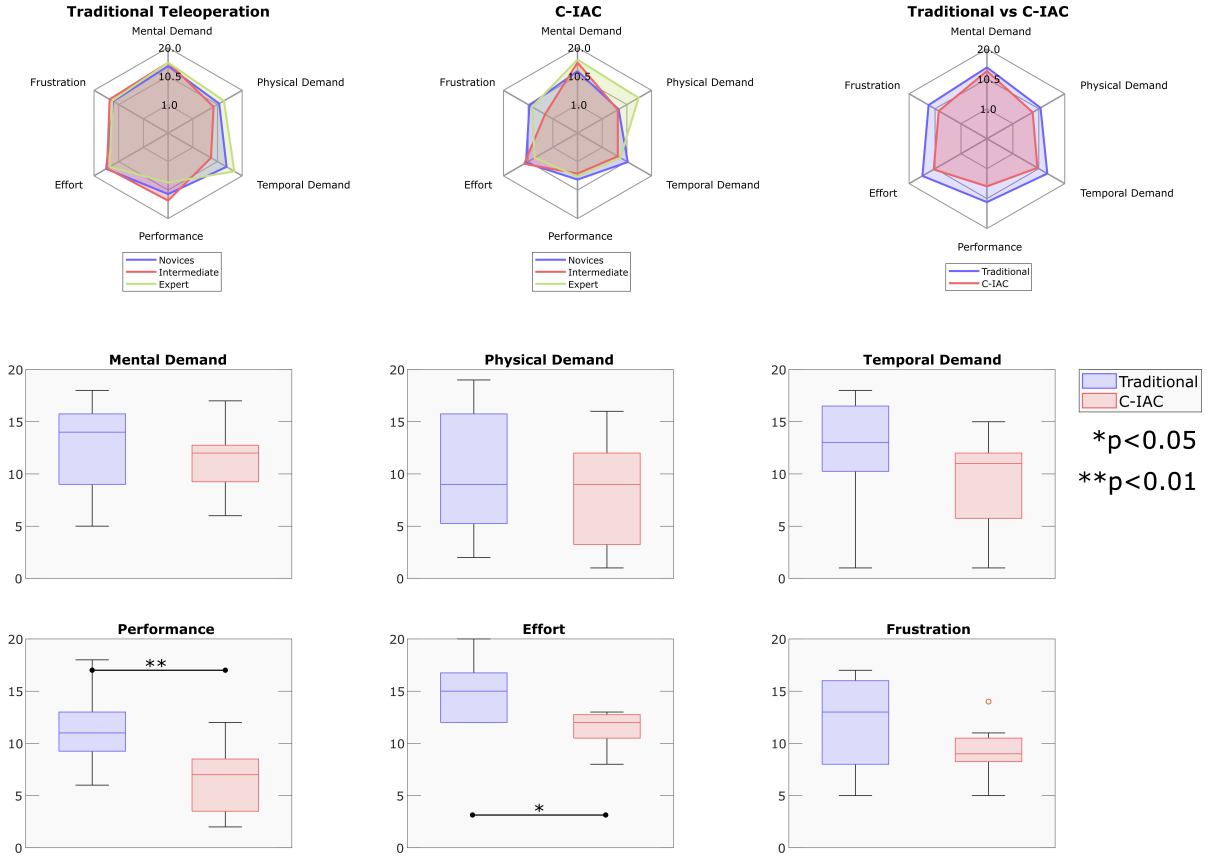


Figure 5.13: NASA TLX results from suturing task participants (i.e. novices, intermediate, and expert) for traditional teleoperation and C-IAC. In the spider charts, we first compare the two controllers using the average answers of the novices and the intermediate and expert’s individual scores to provide a perspective from participants with different suturing skills. As all participants are inexperienced with the surgical robot, we then average all participants scores for Traditional vs C-IAC. In the box charts, we present a detailed view of the scores given by the novice participants ($n = 7$) during this user study.

5.9 Discussion

The classification accuracy results obtained using a reduced number of surgemes confirmed the feasibility of simplifying the suturing gestures into four fundamental surgemes and a general one. While *Push* had the highest overall accuracy, *Other* received the lowest one, likely due to the aggregation of very diverse kinematic gestures in one class. To train the Transformer model, we created the STITCHES dataset and implemented strategy 2 (i.e. labelling of the data using five classes), which yielded a similar trend in accuracy as with JIGSAWS, as illustrated in Figure 5.11.

During the creation of STITCHES, we observed that for the first 1-2 throws, the user has to perform bimanual hand movements to pull out more thread from the tissue in order to have enough thread to perform the next suturing throws (i.e. G9 in Table 5.1). This can be observed in the samples illustrated in Figure 5.11 (d-e), where the *Other* gesture appears after the *Pull* gesture for the first few throws. This implies that although this movement is not a fundamental gesture for a single

suturing throw, it is necessary when performing a series of suturing motions. In [111], this characteristic of the suturing task was not detected due the limitation to two suturing throws, which makes our implementation of online gesture recognition for user assistance closer to real-world implementation.

From the results derived from the Robot-Camera calibration, needle pose estimation exhibited an error in position and orientation consistent with the literature [113, 137] and was sufficiently accurate to carry out the next experiments. As for the first user study, the target reaching experiment focused on speed and precision, it offers an insight on the behaviour of the C-IAC method. This study revealed difference between the traditional teleoperation and C-IAC. After a statistical analysis, C-IAC proved to be faster than the traditional method, also leading to a smaller standard deviation.

To better understand the system's applicability, we conducted an additional user study with four complete and consecutive throws. We tested on users with different skill levels in suturing (i.e. novice, intermediate, and expert), however none of them had previous experience in robotic suturing. Overall, the C-IAC lead to faster task completion than traditional teleoperation, both relative to average throw time and total task time. For the perpendicularity requirement during needle insertion (i.e. *Push* gesture), C-IAC showed to produce significantly less error, while maintaining similar average *Push* time between the two controllers. This noticeable performance was accompanied by the improvement in *Pull*, *Hand-off*, and throw time compared to the traditional method, indicating that the autonomous component was well integrated in the system's workflow and facilitated operation without disturbing.

For these experiments, the comparison between ground truth, which was obtained with manual labelling, and the real-time classification with the Transformer-based model showed an accuracy that is lower than the one obtained during the dataset validation. This was expected as the dataset was recorded on a different user and not on the participants, due to the controller's requirements. Additionally, as both the intermediate and expert had already developed their own style for suturing, their gesture recognition is more affected than the novice. Furthermore, the participants required more suturing time than the demonstrations in STITCHES, which can affect the accuracy of the trained model. Interestingly, the expert required the most time for completing the task despite having more experience in traditional suturing. This may be attributed to the strong bias of the surgeon for non-robotic laparoscopic suturing, which affects the learning curve [74, 95].

Figure 5.13 reports the scores to the NASA TLX questionnaire given by the novices. On average, C-IAC imposes less workload than the traditional controller. Specifically, C-IAC shows significant improvement in performance and less effort. The relatively small variability noticed in both Table 5.6 and Figure 5.13 for C-IAC across all participants also suggests that the proposed controller can lead to a more constrained variability between surgeons and novices. Notice that the suturing time for four throws across the participants is higher than previous works [42, 137, 146]. This is

attributed to the specific surgical setting, which is designed to challenge the user's sensory information and promote human-robot interaction. Limitations such as latency, 2D display, different hand interfaces, and inexperience with the surgical setup of the task settings are the main contributors to the overall higher suturing time.

A sample comparison between ground truth gestures and predicted gestures is shown in Figure 5.12. The different accuracy and misclassifications compared to Figure 5.11 (d) can be attributed to the nature of the C-IAC's Transformer model dataset. Since C-IAC requires a trained gesture recognition model, if the suturing dataset used from its training is constructed from C-IAC, this would require a trained model as well. Consequently, the suturing dataset used to train the C-IAC Transformer model was constructed from traditional teleoperation movements, and as a result, the C-IAC model is fundamentally based on these movements. This solution demonstrates the C-IAC ability to handle lower accuracy models by adjusting the confidence λ parameter. In our tests, we found that setting a maximum value for the confidence parameter to a value lower than 1 is preferable due to the sensors restriction of both the human and the robot. This suggests that a robotic system with richer sensory information and task knowledge, such as additional cameras, force sensors or complex path planning, can allow to increase the confidence to full automation (i.e. $\lambda = 1$). Such changes can potentially further improve the performances obtained in Figure 5.13.

5.10 Conclusion

At the two ends of the spectrum of control, we can observe traditional teleoperation, which is affected by delay and sensory limitation, and full automation, which is limited by sensory reading inaccuracies and lack of broad knowledge and understanding about the task. In this Chapter, we addressed these challenges by developing a unified human-robot system made of high and low level intent detection to improve suturing task performance, while relieving the cognitive workload of the human operator. To our knowledge, this represents the first implementation of real-time intent detection for the interaction control in bimanual robotic suturing, while accounting for confidence across multiple sensory modalities.

First, we provided tailored assistance to the human user by inferring the surgical actions performed in real-time using a Transformer-based model trained on a dataset we constructed from a dVRK setup. The dataset, called STITCHES, was validated and used in the control framework using kinematic data from the dVRK to classify four fundamental suturing gestures and a general one.

We presented C-IAC, a new interaction control framework that integrates IAC and a confidence system based on Bayesian inference for use in the teleoperated suturing task. We constructed a setting where the higher sensory information from the robot can be integrated with the human skills to achieve a higher task performance while lowering the cognitive load.

We gave visual information to the robot by placing markers on the phantom and needle alignment device (CHENA) with a high resolution da Vinci Si camera. A calibration method was provided for Robot-Camera calibration to enable the robot to estimate the needle position using both vision and proprioception. The complete system was tested and validated on the dVRK robot where a series of user studies are conducted to evaluate its use against traditional teleoperation. The user studies were conducted on participants with a wide range of skills, from novices to an expert surgeon, revealing that C-IAC provides faster time performances than the traditional method. Early results on the novice participants suggested a lower cognitive workload perceived throughout the experiment. A direct comparison with the surgeon's performance gave an indication that the method can reduce variability between surgeons and novices, which makes it desirable for real-world use. Overall, the teleoperation system proposed provides the possibility to assist the human operator to perform more efficient gestures with less cognitive workload. This system framework can be expanded beyond robot-assisted surgery as it can be easily transferred to other any field requiring remote teleoperation tasks.

We acknowledge that future work should consider larger datasets, as Transformer-based architectures are inherently data-intensive and their performance typically scales with the amount of available training data. In the future, a larger sample size will be acquired to analyse the use of the C-IAC under different parameters, such as different probability thresholds and λ limits. Furthermore, to enrich the environmental information, other data can be provided to the robot (e.g. contact forces, multi-arm tracking, markless methods, and path planning algorithms) and to the human (e.g. 3D vision and haptic feedback).

Chapter 6

Skill Assessment in Robot-Assisted Surgery

Automated feedback systems in robotic surgery remain a critical challenge in developing objective skill assessment for training and evaluation. In this Chapter, we investigate the key criteria to achieve higher performances for predicting the surgical skill level in real-time based on Objective Structured Assessment of Technical Skills (OSATS) scores. To accomplish this, we carry out three main analyses focusing on model architecture, real-time performance, and skill-level-based cross-validation. For the model architecture, we evaluate the effectiveness of multimodal deep learning models for predicting surgical skill levels using synchronised kinematic and vision data from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset. Our models include separate unimodal baselines and fusion architectures that integrate features from both modalities and are evaluated using mean Spearman's correlation coefficients, demonstrating that the fusion model consistently outperforms unimodal models for real-time predictions. For the real-time performance, we observe the prediction's trend over time and highlight correlation with the surgeon's gestures. For the skill-level-based cross-validation, we separately trained models on surgeons with different skill levels, which showed that high-skill demonstrations allow for better performance than those trained on low-skilled ones and generalise well to similarly skilled participants. Our findings highlight the potential of multimodal learning for fine-grained evaluation of surgical performance and underscore the value of expert-level training data for model generalisation.

The content from this Chapter is currently under review in IEEE Robotics and Automation Letters (RA-L): Hu, Z. J., Ranne, A., Abdelaal, A. E., Bhattacharyya, K., Burdet, E., Okamura, A. M., Rodriguez y Baena, F. "How are my stitches? Model selection and real-time skill assessment for robotic suturing".

6.1 Introduction

Skill assessment in robotic surgery is a critical component in improving surgical training and ensuring patient safety. With the increasing adoption of robotic systems like the da Vinci surgical system, there is a growing demand for reliable, objective, and efficient methods to evaluate surgical performance. Traditional approaches to skill assessment rely on post-procedure analysis, often involving subjective expert evaluations and manual review of recorded data. While these methods provide valuable insights, they lack immediacy and scalability, making it challenging to offer real-time feedback to surgeons during procedures.

To address this limitation, researchers have developed machine learning models capable of automatically assessing surgical skill. A widely used resource for training such models is the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [42]. JIGSAWS includes multimodal data captured during robotic surgical tasks, combining synchronised vision and kinematic data streams. The dataset is widely used as a common standard for investigating surgeon movements during robotic surgery. However, in the context of skill assessment, the annotations of technical skills in these kinds of datasets are severely limited due to using a single score per trial. Due to this constraint, most existing studies have focused on post-hoc analyses that assign a single score to an entire trial, in contrast to a more granular approach that can provide frame- or window-level assessments, thereby leaving a gap in the development of real-time evaluation methodologies. Additionally, the difficulty in recruiting expert surgeons often leads to datasets with participants of diverse surgical experience, casting doubts over the quality of the data and which users should be prioritised for training or evaluation [52]. Furthermore, a clear split exists in the literature between vision-only and kinematic-only based models, ignoring the potential benefits of multimodal models.

In this Chapter, we aim to address these gaps by investigating how high-performing skill assessment models can be achieved by conducting three analyses. To assess the model architectures, we first use different deep learning models commonly used in the literature that are capable of processing different data streams for predictions of skill levels during robotic surgical tasks, specifically robotic suturing. To explore the real-time performance, we then observe and highlight key findings that resulted from the real-time implementation through a manual visual observation of the models' predictions and actual surgeons' gestures. Lastly, to investigate how the users' skills in the dataset can affect the model performances, we implement skill-based cross-validation.

The contributions of this work are as follows: (1) We perform real-time granular skill assessment, which shifts the paradigm from post-hoc evaluations to real-time feedback mechanisms; (2) We present a comparative analysis of seven fundamental models using unimodal and multimodal approaches; (3) We investigate how the surgeon's skill level can affect the performance of the skill assessment models.

6.2 Related Work

Surgical skill assessment has commonly relied on an expert reviewing videos of the surgical task and manually scoring the performance of the surgeons. Zia et al. [171] highlighted the biased and time-consuming nature of manual assessments, emphasising the need for automation.

Several approaches have been explored to automate surgical skill assessment. Early methods used Hidden Markov Models for modeling surgical gestures and classifying skill levels [152]. More recent studies have applied deep learning architectures, such as convolutional and recurrent neural networks (CNNs, RNNs), to extract spatiotemporal features from kinematic and video data. For instance, Zia et al. [171] introduced a feature fusion approach using sequential motion texture, discrete Fourier transform, and discrete cosine transform to classify skill levels in the JIGSAWS dataset.

Several studies have also focused on categorical skill classification, grouping surgeons into novice, intermediate, or expert levels [42, 162]. However, these methods often lack the granularity needed for real-time, continuous assessment. One of the most widely adopted skill assessment frameworks is the Objective Structured Assessment of Technical Skills (OSATS), which evaluates surgical performance based on predefined criteria such as respect for tissue, time and motion, and overall performance which allows a granular study of the surgical skills. Benmansour et al. [13] proposed a Convolutional Neural Network and Bidirectional Long Short-Term Memory (CNN+BiLSTM) model for predicting OSATS scores directly from kinematic data, achieving state-of-the-art performance. However, the use of BiLSTM limits the architecture to only perform post-hoc analysis on the recorded data. Additionally, many surgical skill datasets provide only a single label for an entire trial, typically assigned after task completion. While useful for global assessment, this labeling approach limits the ability to evaluate skill progression throughout the procedure. However, it is possible to use these trial-level labels to supervise in a finer-grained manner, enabling models to infer skill continuously over time despite the coarse annotation [57].

This work extends prior research in surgical skill assessment by introducing three primary contributions:

- **Performance analysis with different input modalities:** We analyse the performances of models based on different data channels (i.e. kinematics and vision) using Spearman's correlation coefficient and Root Mean Square Error (*RMSE*).
- **Real-time OSATS Score Prediction:** Unlike studies that classify user skills into discrete categories (novice, intermediate, expert), our approach focuses on predicting continuous OSATS scores in real-time. This allows for a more nuanced evaluation of surgical performance throughout a procedure.
- **Comprehensive Dataset Analysis:** Existing studies often apply machine learning models

without extensive dataset analysis. We conduct a detailed study of the dataset, addressing potential limitations in the OSATS labels.

Our approach aims to enhance the precision, interpretability, and practical applicability of real-time surgical skill assessment in robotic surgery training.

6.3 Data Pre-processing

The JIGSAWS dataset is a collection of labeled dry lab tasks performed by 8 surgeons on the da Vinci surgical system. In this Chapter, we focus on the dataset's suturing data, as it is a widely studied and analysed task for state-of-the-art models [50, 61]. The dataset consists of three parts: kinematic data, video data, and manual annotations. Both the kinematic and video data are captured at 30 Hz and synchronised. The kinematic data are collected from the da Vinci surgical system while the video data captures the surgical scene from an endoscopic camera. The dataset also includes manual annotations for the gesture labels (i.e. the surgical activity performed by the surgeon in a segment of time) and skill assessments using the OSATS metrics.

6.3.1 Kinematic data

The dataset provides 76 kinematic variables (3 for Cartesian position, 9 for the rotation matrix, 3 for linear velocities, 3 for angular velocities, and 1 for gripper angle, for each manipulator) obtained from the manipulators of the da Vinci platform: the patient-side manipulators (PSM1 and PSM2) and surgeon-side tool manipulators (MTM1 and MTM2). These variables can then be used to construct the features. To represent the interaction between two hands, Rakita et al. [120] showed that it can be achieved using six features which can be computed as scalar-valued features to represent a single time point:

$$\mathbf{s}_t = \begin{bmatrix} \|\mathbf{p}_t^d - \mathbf{p}_t^n\| \\ \|\mathbf{p}_t^d - \mathbf{p}_t^n\| - \|\mathbf{p}_{t-1}^d - \mathbf{p}_{t-1}^n\| \\ \|\mathbf{p}_t^d - \mathbf{p}_{t-1}^d\| \\ \|\text{displacement}(\mathbf{q}_t^d, \mathbf{q}_{t-1}^d)\| \\ \|\mathbf{p}_t^n - \mathbf{p}_{t-1}^n\| \\ \|\text{displacement}(\mathbf{q}_t^n, \mathbf{q}_{t-1}^n)\| \end{bmatrix} \quad (6.1)$$

where $\mathbf{p}_t^d$, $\mathbf{q}_t^d$, $\mathbf{p}_t^n$, and $\mathbf{q}_t^n$ are the positions and orientations of the dominant and nondominant hands at time step t , respectively. The six scalar values calculated, as shown in Equation 6.1, produce relative coordinates (in contrast to absolute coordinate frames): (i) hand offset (distance between hands), (ii) hand offset velocity (rate of change of hand offset), (iii) translational and (iv) rotational velocities of the dominant hand, and (v) translational and (vi) rotational velocities of the

nondominant hand. To account for variations in the kinematic features, min-max normalisation was then applied to scale the features.

As in Rakita et al. [120], the features $\mathbf{s}_t$ only encode a single, discrete hand pose event. For the real-time implementation, we set up a window ω with many such events together in a long concatenated vector $\mathbf{f}_t$ to encode motion over time:

$$\mathbf{f}_t = \begin{bmatrix} \mathbf{s}_{t-\omega} \\ \mathbf{s}_{t-\omega+1} \\ \vdots \\ \mathbf{s}_t \end{bmatrix} \quad (6.2)$$

Each kinematic sequence and corresponding vision frame sequence were segmented using a sliding window approach. A fixed window size ω of 50 time steps with a stride of 10 was selected to maintain a fine-grained temporal resolution while ensuring overlapping contextual information. This approach enabled the model to capture temporal dependencies within both kinematic and visual data streams.

6.3.2 Video data

The video data was preprocessed to ensure consistency and facilitate model training. Each video was first converted into a sequence of frames extracted at a fixed 30 Hz frame rate to standardise temporal resolution across all samples.

Each frame was resized to 224×224 pixels to maintain consistency across samples. The images were converted into tensors and normalised using standard ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]), ensuring compatibility with pretrained models. Finally, as mentioned, the frames were organised into fixed-length sequences, with overlapping sliding windows used to preserve temporal continuity and provide context for the model.

6.3.3 OSATS rating for skill evaluation

In the JIGSAWS's suturing dataset, the surgical skills of the participants are annotated in a standard way by using a modified OSATS approach. These are six scores manually assigned by a cardiac surgeon for each trial ("Respect for tissue", "Suture/needle handling", "Time and motion", "Flow of operation", "Overall performance", and "Quality of final product"), and each rating is scored on a Likert scale of 1 (Worst) through 5 (Best). In this Chapter, we use the notation OSATS^n to denote the n -th score, and OSATS^{n-m} to indicate a range of scores from n to m , inclusive.

It is important to note that these scores are designed for post-hoc evaluation, which the literature was limited to due to the single score label per trial. As our aim is to implement a granular prediction that can be used in a real-time scoring framework, their interpretation must be adjusted to reflect

the temporal nature of the evaluation. For instance, while “Respect for Tissue” can naturally be assessed in real-time, “Flow of Operation” should require analysis over a longer time window to provide meaningful feedback. As such, in the context of real-time scoring, these scores will be interpreted with respect to the specific time window under consideration.

6.4 Models

In this work, we implement seven machine learning frameworks to investigate the usefulness of kinematic and video information in skills estimation tasks. Holistically, we have chosen two classes of architectures, both of which are well studied in the literature, and adapted these to fit the requirements set out (i.e. use of unimodal and multimodal models). The first is the Long Short-Term Memory (LSTM) approach, where we fed each frame of the kinematic/video sequence window one by one into the network, thus iteratively generating a score at each instance. The second is the Transformer approach, where the network predicts the score sequence at once with multi-headed attention. We will investigate the effect of using each modality individually, as well as the effect of fusing them together. Henceforth, vision-only and kinematic-only refer to unimodal training of the models, while fusion refers to the use of both kinematic and video data channels.

All models were trained with mean squared error loss and Adam optimiser, with a learning rate of 5×10^{-5} for 10 epochs, chosen empirically due to the limited dataset size. Furthermore, we implemented early stopping that triggers if the validation loss increases for more than 2 epochs. Both of these measures were designed to reduce overfitting.

In JIGSAWS and derived studies, the two cross-validation techniques used to evaluate the performance of models are:

- Leave-one-supertrial-out (LOSO): the model is trained on data from all subjects, but one trial from each subject is used for validation. The model's performance is evaluated across all participant ensuring the robustness of a method.
- Leave-one-user-out (LOUO): the model is trained on data from all subjects except one, which is used for testing. This process is repeated for each subject, ensuring that the model's performance is evaluated with a subject that has not previously been seen in the training dataset.

We also propose a third cross-validation technique:

- Leave-one-score-in (LOSI): the model is trained on data that includes at least one matching score among the first three OSATS scores (i.e. $\exists i \in \text{OSATS}^{1-3}$ s.t. $i = n$), namely “Respect for Tissue”, “Suture/Needle Handling”, and “Time and Motion”. For instance, if a trial has OSATS¹⁻³ equal to (1, 1, 2), it would be included in the training set for folds corresponding to scores 1 and 2 but excluded from folds for scores 3, 4, and 5. This approach ensures that

each fold contains trials with at least one overlapping score, allowing the model to generalise across similar skill levels.

In this work, we performed cross-validation in the LOSO, LOUO and LOSI manner, by dividing the data into windowed folds, leaving one fold out, then evaluating on the rest, and repeating for the next fold. Data augmentation was not implemented on the video data to avoid distorting or removing key information from within the frame, such as the suturing needle, the target suturing point, etc.

6.4.1 CNN-LSTM-V (Vision only)

We begin by performing initial feature extraction on the input tensor of shape (Batch, Channel, Window Time, Height, Width) via a ResNet-18 backbone, a powerful CNN that specialises in identifying features at different scales, as it uses square kernels of size 7 and 3 with a high count of learned filters [54]. This is followed by flattening the 2D feature maps into a single array (Batch, Window Time, CNN Channel $\times$ Height $\times$ Width), stacking it along a new dimension, then passing it into $3 \times$ LSTM blocks for the model to learn temporal dependencies between features. LSTM processes the features in a sequential manner, reading in the current time stamp and the previous recurrent state, before generating the current prediction. We picked the LSTM hidden dimension to be comparable in size to CNN Channel $\times$ Height $\times$ Width to capture the complexity of the data, but not larger to avoid making the model too large.

Dropout and batch normalisation were applied to the output of the LSTM to encourage the model to utilise its full capacity, and generalise better to unseen data. A bridging linear layer was introduced to translate the LSTM output to OSATS scores.

6.4.2 LSTM-K (Kinematic only)

In comparison to vision, kinematic recordings have fewer features (18 features for kinematics vs 224×224 pixel features for vision). Therefore, the need for complex feature extractors is not justified. As a result, we passed the windowed kinematic data of shape (Batch, Window Time, Features) into 3 LSTM blocks to extract temporal features with a hidden dimension size of 128. This dimension is intentionally larger to increase model complexity. Following the LSTM module, the output is passed into a linear layer with dropout and batch normalisation to generate the score predictions.

6.4.3 DualLSTM-F (Fusion)

LSTM focuses on refining outputs over time and distilling data via its gated architecture. Thus, we have opted to use two LSTMs, one for each type of data, to first generate two time-independent encoded feature arrays, stacking these along a new dimension, before performing a series of 2D

convolution with kernel size 2 to condense them further. Finally, a series of fully connected layers was used to generate the final score.

6.4.4 CNN-V (Vision only)

To benchmark the significance of incorporating temporal information into the model creation, we also designed a model consisting of CNN layers alone. Following feature extraction via ResNet, the feature maps were flattened, then averaged across time, before being passed to a fully connected layer. We hypothesise that average pooling would have a significant impact on the results, as the model loses the ability to identify variations, and the ability to highlight key frames/features and use these critical information to its advantage.

6.4.5 Transformer-V (Vision only)

Transformers capture global information from across the entire input array, even if key information is sparse. By combining them with CNN feature extractors, the model is able to capture both spatial and temporal features, using the advantages of both methods [31, 121].

As described in the original Transformer architecture [160], the input array is passed to the encoder, where self-attention takes place, allowing the model to look for correlations within the array elements and generate the attention map. Thereafter, cross-attention is performed between the map and the desired target array during training. While inferencing, the target is then set to zero, such that the model can generate it. Due to the large number of computations needed to perform attention, Transformers are known to be computationally intensive, demanding large datasets for training [18, 31].

In this case, given the absence of a time dependent target array, the Transformer will only consist of a ResNet-18 feature extractor, with 3 layers (hidden dimension 50) of self-attending encoder that attends across the time dimension. Two bridge layers, one after the ResNet, and another after the Transformer are included. The first reduces the size of the feature array in hopes of saving computational resource, while the second bridges the encoder embeddings with the final score predictions. In all of our Transformers, we opted for sinusoidal positional encoding to improve computational efficiency.

6.4.6 Transformer-K (Kinematic only)

The kinematic Transformer consists of 3 simple layers of self-attending layers with a hidden dimension of 50, equivalent to that of the sliding window length. Following embedding by self-attention, the attention map was flattened and linear layers were used to generate the final output score.

6.4.7 DualTransformer-F (Fusion)

In our fusion Transformer, we incorporate an additional decoder module that cross-attends between the kinematic and vision features. The intuition behind this decision is that the model may utilise key frames and features in time from kinematics, identified by two independent self-attention branches, to assist its decision making in the main branch, which focuses on the video data. As a result, we selected the query and keys to be the vision encoder output, while the values are drawn from the kinematic encoder output. The generated attention map from the decoder output is a sound representation of the key features. Using such information, we hypothesise that the model will outperform its counterparts that only rely on one data modality.

6.5 Results and Discussion

Using the neural networks architectures described in Section 6.4, we trained models for the robotic suturing task. A total of three separate analyses were conducted to understand how high performing skill assessment models can be achieved:

- *Model performance analysis thorough ρ and RMSE*: Using all seven models described in Section 6.4, the average of the mean Spearman's correlation coefficients and Root Mean Square Error of all OSATS in each trial is calculated to evaluate models with different input modalities and compare with the results in the literature.
- *Visual highlights of real-time OSATS predictions and correlation with surgical gestures*: We used the single OSATS score per trial provided by JIGSAWS to perform granular predictions for the entire trial and observed the correlation with surgical gestures. We performed real-time prediction of the single OSATS score as it is often overlooked in favor of post-hoc evaluation. By highlighting potential patterns learned across trials, we aim to extract additional information from the single label restriction and explore its use for real-time feedback.

Due to computational constraints, we focus on representative kinematic-, vision-, and fusion-based models (i.e. LSTM-K, CNN-LSTM-V, and DualLSTM-F).

- *Analysis of LOSI cross-validation*: We performed a validation with the LOSI setup using vision-only and kinematic-only models to observe learning performance across data with different user skills. This analysis aims at understanding the quality of datasets that contain different skilled users, highlighting the most useful data.

The experimental evaluation was conducted by calculating the *RMSE* values between the ground truth and predicted OSATS scores and using the Spearman's correlation coefficient ρ , which is an evaluation metric that also provides a check for statistical significance using the *p*-value. The value ρ can be within -1 and +1, with a value closer to +1 indicating a stronger correlation between the predicted and ground truth OSATS.

6.5.1 Mean ρ and Mean $RMSE$

By using the cross-validation methods LOSO and LOUO proposed by [42], we compared the performances of the models. Table 6.1 shows the results from averaging the mean Spearman's correlation coefficients and Root Mean Square Error ($\bar{\rho} \mid \overline{RMSE}$) of all OSATS in each trial for all the suturing dataset. Overall, it can be observed that using video data allows the models to perform better in both the LOSO and LOUO setups. In particular, in LOSO, models that use video data (i.e. CNN-LSTM-V, DualLSTM-F, CNN-V, Transformer-V, and DualTransformer-F) can achieve a $\rho > 0.8$ while keeping $RMSE \leq 0.6$. LOUO displays a similar trend, with all models trained on video data performing $\rho > 0.6$ and $RMSE \leq 0.8$. The best performing model in LOSO is CNN-V, while in LOUO is CNN-LSTM-V. Noticeably, CNN-LSTM-V is the second best performing model in LOSO, while CNN-V is sixth in LOUO. The worst performing models are those reliant solely on kinematic data (i.e. LSTM-K and Transformer-K), with the worst performing in both LOSO and LOUO being Transformer-K. Interestingly, while the non-Transformer fusion model (i.e. DualLSTM-F) did not show improvements with respect to the model trained on vision data only, the DualTransformer-F improved its performance with respect to the Transformer models trained on video or kinematic data only.

The results obtained can also be directly compared with those in the literature. Compared to the study from Zia et al. [171], deep learning models outperformed the use of holistic features in automated skill assessment, regardless of the cross-validation scheme used. In the LOSO setup, Benmansour et al. showed that their kinematic-only based state-of-the-art model obtained an overall ρ of 0.65 using CNNBiLSTM. Our results on fusion models then indicate that the adoption of both video and kinematic data can further improve this performance. Guo et al. [44], which used a vision-only Transformer model, also shows comparable results for LOSO and LOUO, with ρ of 0.87 and 0.75, respectively. Our results suggest that a fusion with kinematic data could further improve the performance, especially considering the lower computational demands of kinematic-based models.

In order to analyse each OSATS more accurately, Table 6.2 presents the model-wise comparison of performance across six OSATS criteria in LOSO and LOUO using the Spearman's correlation coefficient and Root Mean Square Error ($\rho \mid RMSE$) checking for statistical significance.

In the LOSO setup, CNN-V, which relies solely on visual input, achieved the highest overall performance, with ρ values exceeding 0.89 in all metrics and a remarkably low $RMSE$. This suggests that visual cues alone carry rich information regarding skill assessment in suturing. The CNN-LSTM-V model, which incorporates temporal modeling of visual features, also performed well, particularly in *Time and Motion* (0.894 $\mid$ 0.504) and *Suture Handling* (0.885 $\mid$ 0.539), showing that temporal modeling offers benefits over static visual features. Interestingly, the fusion model (DualLSTM-F), which uses both visual and kinematic data, demonstrated competitive and consistent performance across all metrics, indicating the advantage of multimodal inputs. All im-

Table 6.1: Performance comparison for Suturing based on $\bar{p} \mid \overline{RSME}$, using different architectures and source data under LOSO and LOUO evaluation methods (highest performances in bold).

	Suturing	
	LOSO	LOUO
CNN-LSTM-V	0.861 0.516	0.843 0.553
LSTM-K	0.642 0.887	0.612 0.881
DualLSTM-F	0.857 0.522	0.671 0.698
CNN-V	0.898 0.431	0.568 0.810
Transformer-V	0.821 0.592	0.635 0.825
Transformer-K	0.560 0.850	-0.0133 1.004
DualTransformer-F	0.808 0.606	0.654 0.758

provements were statistically significant ($p < 0.05$). In contrast, the LSTM-K model using only kinematic data showed comparatively lower correlation values across all metrics, suggesting that visual data plays a more pivotal role under LOSO. Transformer-based models showed mixed results. The Transformer-V model showed strong performance, e.g. (0.911 | 0.418) in *Quality of Final Product*, though slightly behind CNN-V and fusion models. However, the Transformer-K model again underperformed, consistent with other kinematic-only approaches.

The LOUO setup, which is generally more challenging due to the higher variance introduced by unseen users, resulted in lower overall performance across models. Nonetheless, several models remained robust. The CNN-LSTM-V model achieved strong performance in most metrics, notably in *Overall Performance* (0.925 | 0.415) and *Suture Handling* (0.888 | 0.536), confirming its generalisation capability. The fusion model (DualLSTM-F) retained moderate performance across metrics but did not outperform the vision-based CNN-LSTM-V. The DualTransformer-F model under LOUO performed relatively well in *Suture Handling* (0.848 | 0.5642) but struggled in *Quality of Final Product* (0.302 | 1.003), suggesting limitations in generalising fusion learning across users. The Transformer-K model failed to generalise under LOUO, with negative or near-zero correlation across most metrics (e.g., (-0.2028 | 0.9827) in *Quality of Final Product*), reinforcing the trend observed in the LOSO setup.

Overall, models utilising visual data, especially CNN-based architectures, demonstrated superior

performance across both LOSO and LOUO. Fusion models yielded competitive results, particularly under LOSO, indicating potential for a multimodal approach. Kinematic-only models consistently underperformed, and Transformer-based models showed promise but require further tuning and data to achieve better performance.

Table 6.2: Performance comparison for Suturing based on ρ | $RSME$, using different architectures and source data under LOSO and LOUO evaluation methods (highest OSATS performances in bold). * $p < 0.05$

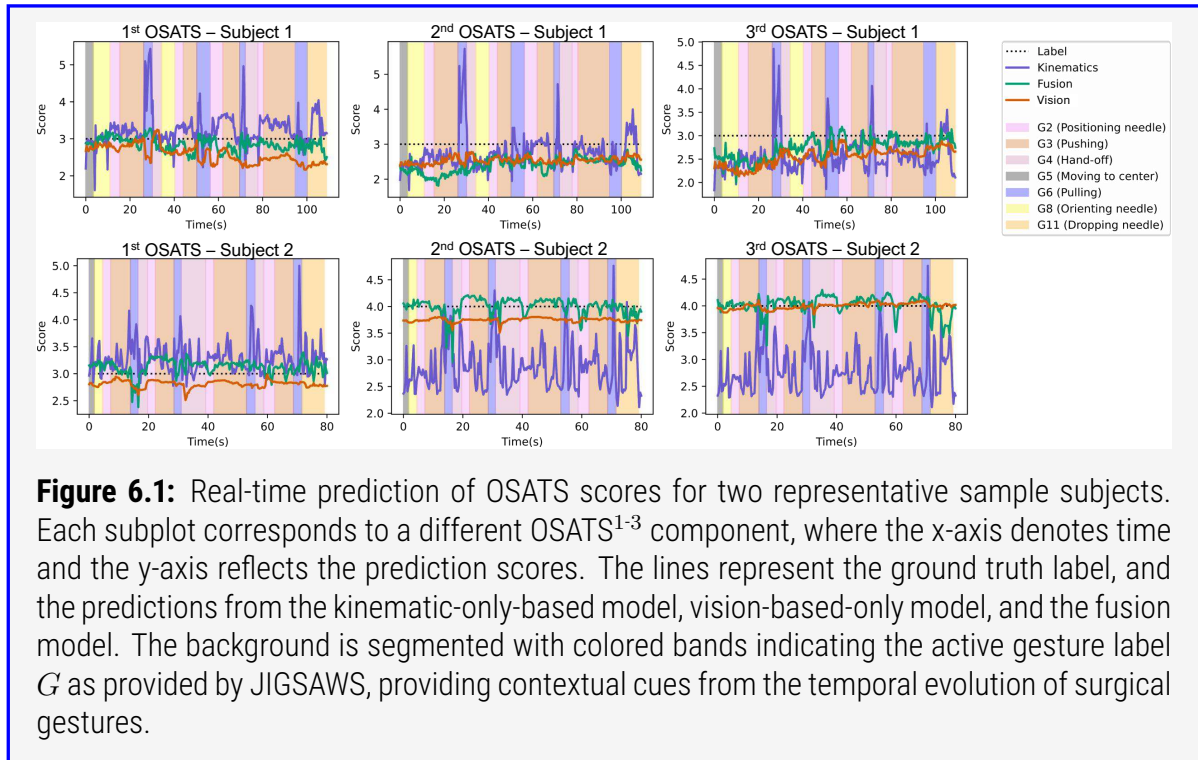
LOSO										
	Respect for tissue		Suture handling		Time and motion		Flow of operation		Overall performance	Quality of final product
CNN-LSTM-V	0.8118	0.515	0.885	0.539	0.894	0.504	0.870	0.494	0.872 0.508	0.836 0.534
LSTM-K	0.594*	0.837	0.612*	0.970	0.674*	0.971	0.699*	0.820	0.678* 0.890	0.597* 0.821
DualLSTM-F	0.852*	0.496	0.831*	0.612	0.854*	0.538	0.863*	0.483	0.845* 0.575	0.894* 0.403
CNN-V	0.893*	0.392	0.897*	0.425	0.899*	0.460	0.900*	0.423	0.867* 0.508	0.929* 0.363
Transformer-V	0.766*	0.583	0.825*	0.573	0.809*	0.699	0.791*	0.621	0.822* 0.624	0.911* 0.418
Transformer-K	0.558*	0.848	0.638*	0.893	0.559*	0.902	0.481*	0.811	0.521* 0.849	0.605* 0.791
DualTransformer-F	0.773	0.590	0.865*	0.540	0.770*	0.733	0.819*	0.574	0.785* 0.639	0.835* 0.539
LOUO										
	Respect for tissue		Suture handling		Time and motion		Flow of operation		Overall performance	Quality of final product
CNN-LSTM-V	0.729*	0.667	0.888*	0.536	0.848*	0.573	0.861*	0.534	0.925* 0.415	0.806* 0.565
LSTM-K	0.520*	0.881	0.645*	0.976	0.513*	0.964	0.638*	0.801	0.688* 0.895	0.660* 0.744
DualLSTM-F	0.814*	0.578	0.694*	0.741	0.619*	0.766	0.572*	0.718	0.700* 0.633	0.629* 0.735
CNN-V	0.713*	0.703	0.624*	0.854	0.500*	0.929	0.488*	0.810	0.668* 0.700	0.417* 0.837
Transformer-V	0.647*	0.712	0.705*	0.892	0.695*	0.730	0.617*	0.761	0.657* 0.730	0.492* 1.065
Transformer-K	-0.167	0.970	0.192	1.074	0.101	1.057	-0.034	0.940	0.031 0.996	-0.203 0.983
DualTransformer-F	0.731*	0.727	0.848*	0.564	0.679*	0.750	0.638*	0.757	0.725* 0.680	0.302 1.003

6.5.2 Real-time OSATS predictions

To further illustrate the typical temporal behavior of the models, we visualise the real-time prediction trends with two representative trials in Figure 6.1. Across both trials, the fusion model demonstrates consistently more stable and accurate predictions compared to unimodal baselines. In the first trial, the fusion outputs closely follow the ground truth labels, even in gesture-rich segments involving frequent transitions between G2, G3, and G6. While the kinematic model exhibits high variance and the vision model tends to underestimate, the fusion model smooths over these variations, suggesting it effectively utilises complementary cues from both modalities.

The benefits of fusion become even more evident in the second trial. Here, the kinematic model again produces noisy predictions, particularly during prolonged pulling (G6) and pushing (G3) phases, where the amplitude fluctuates. The vision model, while smoother, remains insensitive to certain transitions, leading to flattened predictions across gestures. In contrast, the fusion model maintains a higher alignment with the ground truth and adapts more sharply to gesture changes.

These qualitative results support our quantitative findings and highlight the robustness of the fusion model in dynamic surgical scenarios. The model's ability to maintain label alignment and smooth fluctuations in real time suggests strong potential for future applications in surgical skill monitoring and intraoperative feedback systems.



Additionally, we observed a recurring pattern, where the kinematic model exhibits sharp upward spikes while the vision model simultaneously shows prediction drops, especially during segments labeled as the pulling gesture (G6). This divergence can be attributed to the different sensitivities of each modality. From the kinematic perspective, the pulling motion is relatively smooth and may be interpreted as skillful, leading to higher predicted scores. However, the vision model captures more context: as the needle is pulled through tissue, the skin is often visibly displaced or stretched, sometimes excessively as shown in Figure 6.2, which introduces a perceived risk of laceration. The vision model appears to correctly associate this with a decline in performance quality, hence the drop in its prediction. This highlights the value of visual input in capturing subtle but clinically relevant indicators of technique that purely motion-based data might miss.

6.5.3 LOSI analysis

While LOSO and LOUO are used in the literature for intra-task generalisation [56], we proposed the LOSI cross-validation to analyse the learning of skill levels within the JIGSAWS suturing dataset. The motivation of this section lies in the significant limitation in surgical datasets concerning the diversity and imbalance of participant expertise within such data. Recruiting expert surgeons is difficult, often resulting in datasets that contain a broad range of skill levels. This raises concerns about the reliability of labels and the influence of user skill level on model performance [52]. To investigate this, we focus on skill-level-based cross-validation, examining how models trained on participants of varying expertise perform and generalize across different user groups. The LOSI approach allows the investigation of skill structures within the suturing and prioritises data from

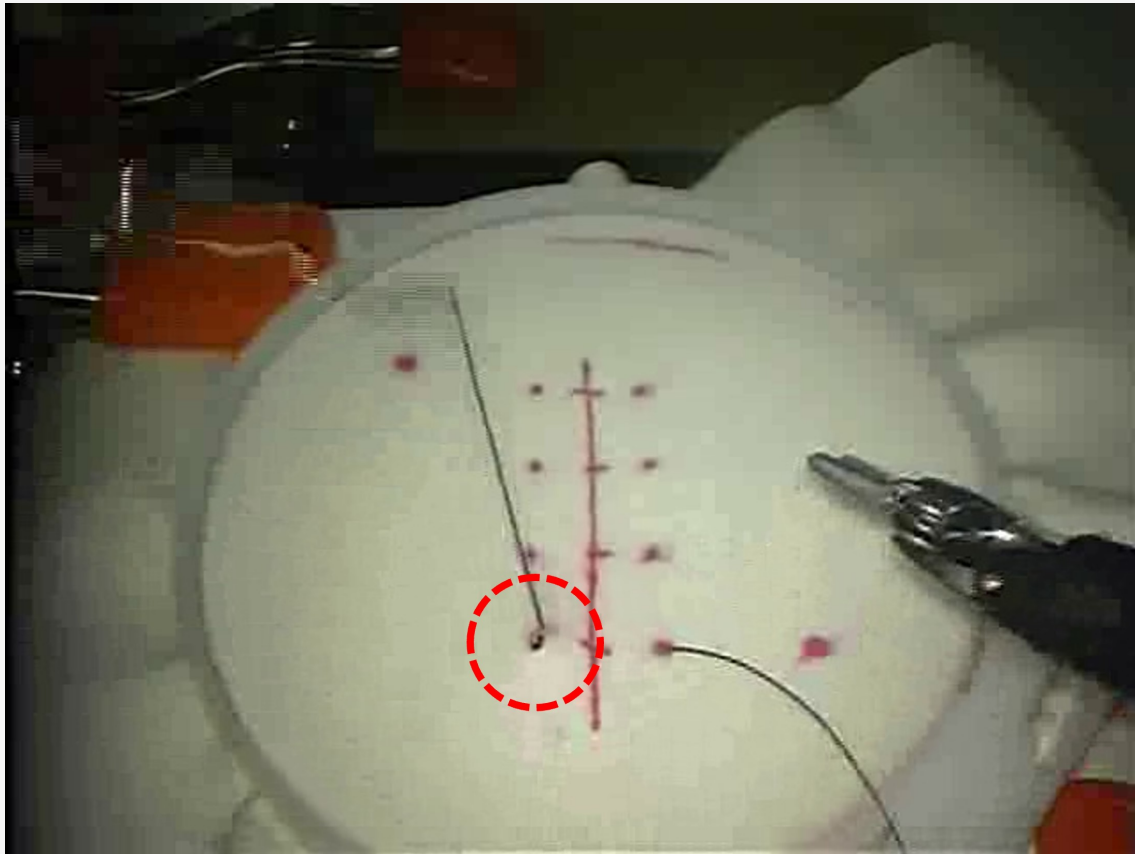
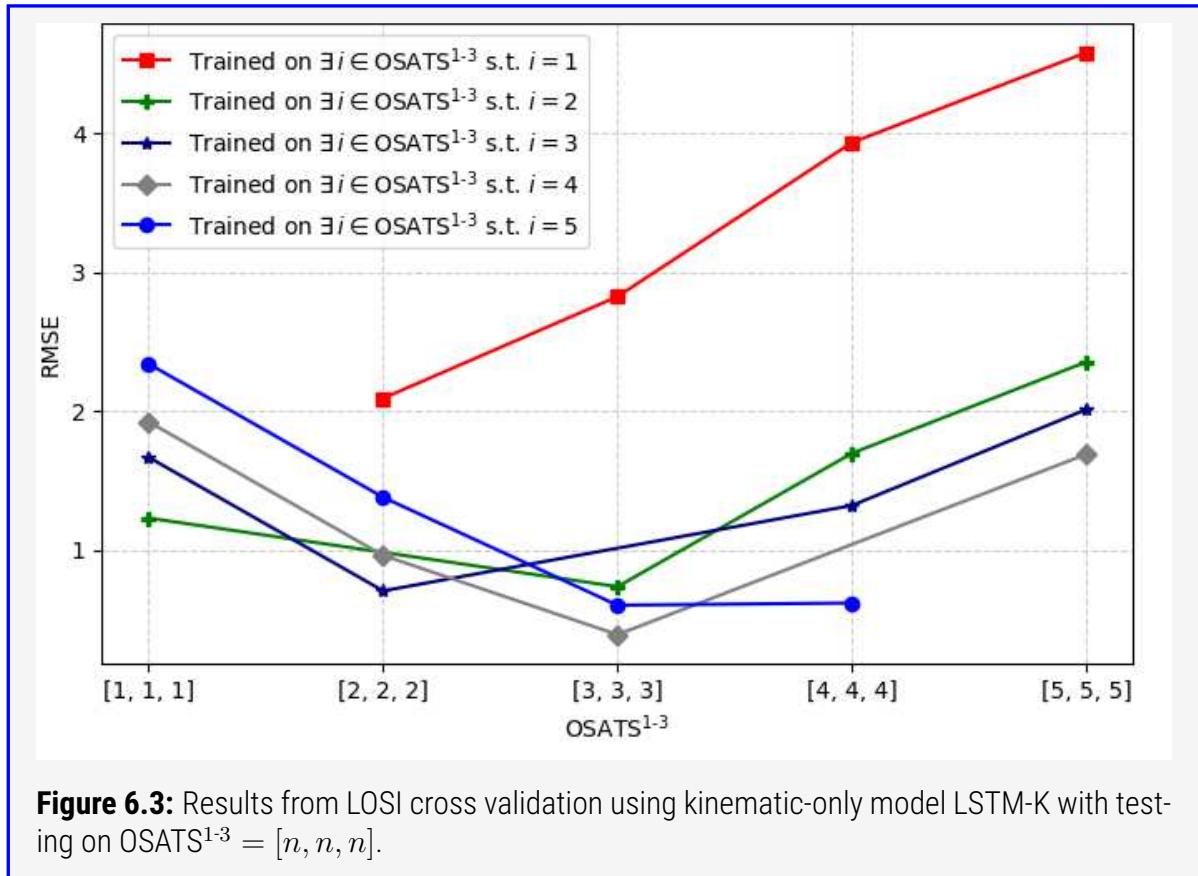


Figure 6.2: Gesture G6 (Pulling) in JIGSAWS's trial. As the needle is pulled out, the tissue is visibly torn (highlighted in red).

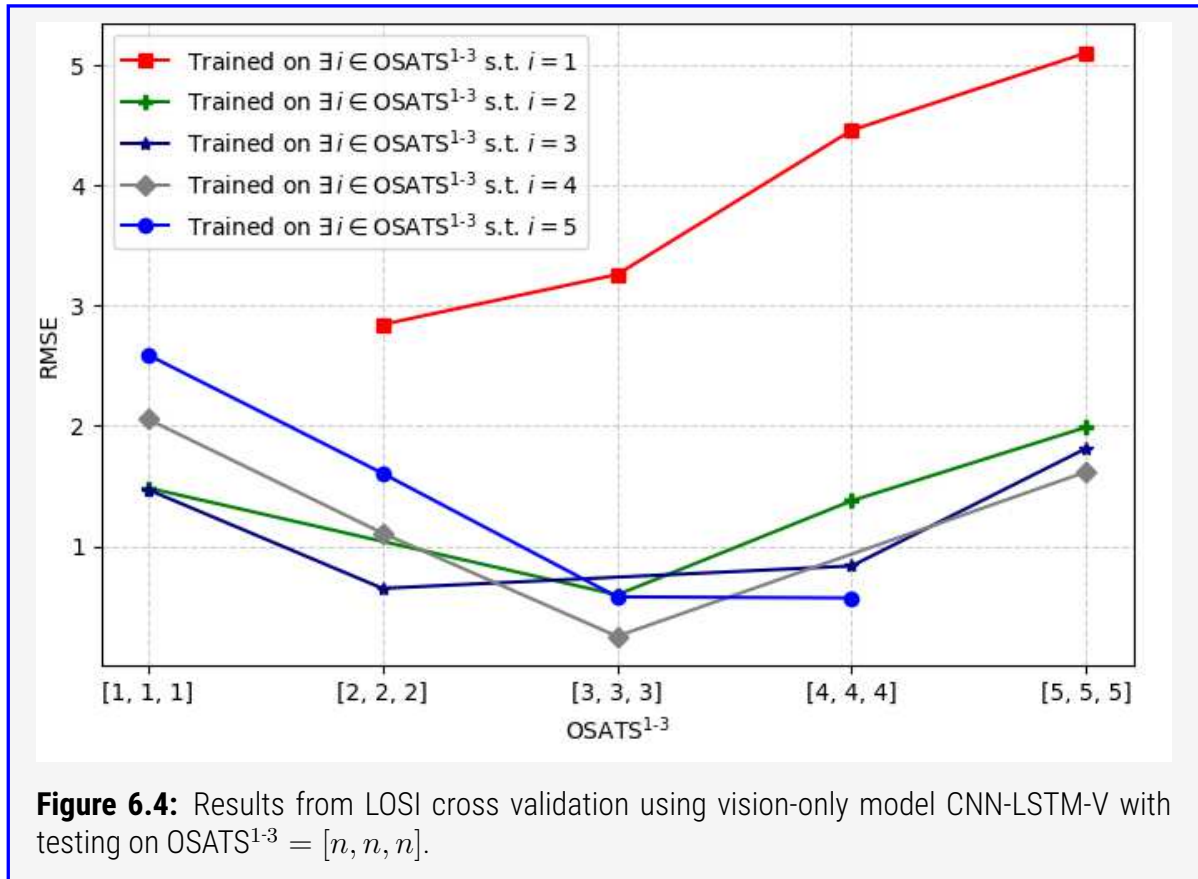
datasets during training for optimal model training. This analysis approach relies on the idea that datasets may be imbalanced, potentially reflecting the difficulty of recruiting expert surgeons or an over-representation of novice participants and that, as robotic surgery becomes more widespread, larger datasets will need prioritisation of the most useful data.

We apply LOSI cross-validation, where the model is trained on trials that share at least one common score among the first three OSATS dimensions (i.e. $OSATS^{1-3}$). Essentially, we train a model on trials that contain a given score in any of the first three OSATS parameters, and evaluate it on the remaining trials, while testing on trials that have all same values in $OSATS^{1-3}$. We plot the *RMSE* performance of the five resulting models across these conditions to assess generalisation across skill levels. Figure 6.3 and Figure 6.4 present the results obtained using the kinematic and vision-based models, respectively. The motivation for training on trials that contain a given score in any of $OSATS^{1-3}$, rather than requiring all three to match, is that scores typically cluster closely for a given participant. This setup helps to maximise training of participants' trials and account for potential reviewer bias, as evaluators often assign similar scores across categories.

The trends that can be extracted from the results indicate that models trained on trials with a score of 1 in $OSATS^{1-3}$ perform the worst, struggling to generalise to participants with higher skill levels.



Interestingly, this is not true for models trained with a score of 5 in the first three OSATS dimensions, which instead generalise effectively on participants with similar skill level, with the performance declining with participants of lower skill level. As the result is true for both models trained on kinematic and video data, this suggests that the skill level is detectable in both information channels and that data from higher skill level users have greater importance for skill level detection. This is likely due to the fact that robotic suturing requires a structured and consistent gesture pattern that is encoded in participants with a high skill level. This likely reflects the lack of consistent patterns in low-skill performances. However, as even low-skilled participants (except those with OSATS equal to 1) showed a similar trend, this indicated that even moderate skill levels exhibit structured gestures that models can learn from. An additional general trend shows that models trained on adjacent skill levels tend to better predict intermediate trials, with those trained on higher scores often outperforming those trained on lower ones, with the model trained on $OSATS^1$ always scoring the worst. For instance, in the kinematics result (Figure 6.3), for trials with $OSATS^{1-3}$ scores of 3-3-3, the model trained on score 4 yields better *RMSE* than the one trained on score 2, followed by models trained on 5 and then 1, resulting in a performance order of $4 > 2 > 5 > 1$.



6.6 Conclusions and Future Work

This Chapter explored real-time surgical skill assessment by developing unimodal and multimodal models that integrate kinematic and video data. To understand how performance can be enhanced for OSATS predictions in a real-time setup, we investigated: (1) the use of deep learning architectures with different data channels, (2) real-time prediction correlation with surgeons' gestures, and (3) cross-participant generalisation across different skill levels.

The results provide insight into the effectiveness of multimodal models for assessing surgical skill using both kinematic and video data. Quantitatively, the mean Spearman correlation coefficients demonstrate that fusion models are more consistent than unimodal architectures, suggesting that combining visual and motion cues provides a more robust representation of surgical performance. Notably, models trained on trials labeled with higher scores in OSATS^{1-3} generalise better across participants, while those trained on low-skill examples perform poorly, likely due to a lack of structured gesture patterns in novice executions.

Real-time prediction analysis further supports these trends. Visualisation of trials' performances shows that the fusion model yields smoother and more accurate temporal predictions, closely aligning with ground truth labels. These predictions are more stable and context-sensitive, particularly during gesture transitions. Interestingly, we observe gesture-specific behaviors: kinematic

predictions often spike during pulling gestures, while vision predictions dip, indicating that, while motion data may interpret pulling as smooth and high-skill behavior, visual input can capture risky tissue deformation missed by kinematics alone.

Our LOSI results show that models trained with the lowest score trials (i.e. score 1) perform poorly when predicting higher-skill trials, highlighting the limitations of learning from under-performing demonstrations alone. In contrast, models trained on any higher scores generalise better to neighboring and lower scores. This suggests that using data from participants with at least some degree of proficiency is more beneficial than relying solely on those with the lowest skill level, underscoring the importance of including skilled individuals in training datasets. We also observe that trials with certain scores are best predicted by models trained on adjacent scores, with models trained on higher scores generally outperforming those trained on lower ones. This suggests an asymmetry in generalisation: high-skill demonstrations offer richer learning signals for cross-level prediction than low-skill ones. These findings emphasise the importance of balanced training sets that include mid-to-high scoring trials, especially in automated skill assessment systems. Furthermore, these findings also help support the development of frameworks aimed at establishing paradigms for predicting surgical performance and patient outcome [94]. It is worth noting that some of these observations might be dataset-specific due to the dataset's size. The small dataset size restricts the model's capacity to generalize and may result in suboptimal performance or overfitting. Despite this limitation, we aimed to evaluate the model's feasibility and preliminary performance on the available data and future work will focus on larger datasets to fully leverage the representational power of Transformer architectures and achieve more robust and generalizable results.

There are several avenues that can be addressed in future works. The single-label-per-trial annotation setup limits temporal resolution and introduces potential biases from subjective expert scoring. Future work should involve more granular annotations, for instance, surgeon-provided scores for every 10 seconds window, to better evaluate the temporal validity of real-time predictions. Additionally, the suturing dataset size restricted our analysis primarily to the first three OS-ATS dimensions. As larger, more diverse datasets become available, future models should aim to predict all six OSATS dimensions and explore differences between annotators to mitigate label bias.

Chapter 7

Conclusions

This thesis has investigated advanced human-robot cooperation strategies for teleoperating the dVRK, with the goal of enhancing assistance for the human operator to achieve high performance and low perceived workload. A range of paradigms were explored, progressing from shared to supervisory and assisted control systems, each building on insights from the previous to design a high-performing system framework. Chapters 3, 4, and 5 addressed distinct control frameworks that progressed through the full spectrum of autonomy, from predefined subtask sharing to dynamic role switching and predictive assistance, highlighting key aspects for developing an effective controller, such as continuous adaptation, real-time assistance, and use of expert demonstrations. These studies demonstrated how autonomy can be utilised to improve surgical task performance, reduce user workload, and adapt in real-time to user intent and behavior. Chapter 6 expanded the scope from assistance to assessment, providing a deep dive into automated skill evaluation in robotic surgery using multimodal data and deep learning. Through this investigation, we identified the architectural, data-driven, and validation choices that most impact model performance and generalisation. Together, these Chapters present a comprehensive framework for both augmenting and evaluating surgical performance in robotic-assisted procedures. In this final Chapter, we summarise the main contributions of each chapter, reflect on the broader implications of this work, discuss remaining limitations, and propose directions for future research in cooperative and intelligent surgical systems.

7.1 Summary of Thesis Achievements

In Chapter 3, a shared control framework for human-robot cooperative surgery was developed using inverse reinforcement learning and evaluated in the bimanual peg transfer task. The task was segmented into subtasks assigned either to the human or the autonomous agent, allowing the robot to handle relocation movements while the human retained control over physical interactions. The agent was trained using EG MaxEnt IRL, enabling it to learn movement strategies from an expert demonstration. Experimental results demonstrated that shared control significantly outperformed manual teleoperation across multiple metrics. Participants achieved more successful peg transfers under shared control within the same time period. Task completion times were also significantly reduced and performance variability between trials decreased, indicating improved consistency. Subjective workload assessments showed reduced mental and temporal effort, and participants reported feeling more in control when collaborating with the autonomous agent. Eye-tracking data supported these findings, revealing shorter fixation durations and higher saccade frequencies under shared control, indicators of reduced cognitive demand. These results demonstrate the effectiveness of integrating human-robot cooperation frameworks to improve surgical performance and reducing user workload by maintaining surgeon engagement while offloading routine operations to the robot.

In Chapter 4, we explored a further type of control framework: supervisory control. A novel priority-changing control strategy was developed and evaluated to enhance human-robot cooperation during surgical tasks. This approach allowed dynamic adjustment of control authority between a human operator and an autonomous agent depending on the robot's deviation from the agent's planned trajectory. The new reinforcement learning method tGAIL was proposed to improve the agent's generalisation and robustness from 10 expert demonstrations. Experimental results demonstrated the effectiveness of this approach through both simulation and real-world user studies using the dVRK. In simulation, supervision control consistently outperformed segmented and fully manual control, achieving faster task completion and lower reported mental and physical workload. These trends were confirmed in real-world experiments, where supervision control required the least human input, minimised clutch usage, and delivered the highest user satisfaction. Users generally preferred supervised control for its reduced effort and improved performance, although some noted occasional conflicts with the agent's decision-making. Overall, Chapter 4 concludes that the proposed control strategies, especially the supervised mode, significantly enhance efficiency, reduce user effort, and are well-received by novice users. These findings reinforce the potential of human-robot cooperative teleoperative controllers that integrate an agent that can intervene with continuous actions in both the state and action spaces, and suggest that such methods can be further extended to more complex procedures like suturing. The subsequent Chapter builds directly on these findings.

In Chapter 5, we developed a system based on the assisted control framework. A novel control framework was introduced for bimanual robotic suturing that combines high-level gesture recognition with a low-level confidence-based intent prediction. The system utilises a Transformer model for real-time gesture classification and a C-IAC to dynamically combine the system's confidence in its own sensory inputs with the operator's predicted intentions. This dual-layered approach allows the robot to support the user by anticipating their actions and adjusting assistance accordingly, particularly during tasks affected by visual occlusion or teleoperation delays. Experimental evaluations demonstrated that C-IAC significantly improves suturing performance compared to traditional teleoperation. Across both novice and more experienced participants, the system led to reduced task completion times, lower perpendicularity errors during needle insertion, and decreased cognitive workload.

To further investigate the use of AI in robot-assisted surgery, in Chapter 6, a study was conducted to understand the key factors for high performing real-time surgical skill assessment (OSATS scores) using deep learning models. To achieve this, three main analyses were conducted. Firstly, deep learning frameworks used in the literature were compared to understand the performance differences according to architecture and data channel (i.e. kinematics, vision, or fusion). Then, the real-time predictions of the OSATS scores were displayed and visually analysed for correlation with performance and surgeon's gestures. Lastly, a user-skill-based cross-validation method was used to determine the most influential data in the dataset. The study demonstrated that multimodal models, particularly those trained on high-skill data, outperformed unimodal ones and generalised better to users of higher skill levels. These findings reinforce the value of expert demonstrations in training effective assessment models. The work shows that automated skill assessment benefits from both data fusion, particularly video data, and a dataset with high-skilled participants.

7.2 Limitations

The work on this thesis investigates the human-robot cooperative controllers starting from shared control, to supervisory control, and to assisted control in Chapters 3, 4, and 5, respectively. Each Chapter presented has specific limitations which are generally addressed in the next by building on prior findings.

One key limitation of the shared control framework presented in Chapter 3 is its reliance on fixed task segmentation. While this method effectively divides control between the human operator and the autonomous agent, it lacks adaptability to real-time context changes. Surgical procedures are inherently dynamic and may require fluid transitions in control authority based on the patient's condition, surgeon intent, or unexpected complications. The Chapters 4 and 5 tackle precisely this limitation, allowing the robot to dynamically interact with the surgeon. Another limitation concerns the autonomous agent used, which is based on the EG MaxEnt IRL approach. While this method

showed promising results in capturing strategy, its discrete actions and limited training data can hinder generalisation to alternative or more realistic trajectories. The framework also does not account for inter-user variability or incorporate mechanisms to adapt the agent to specific surgeon behaviors over time, which would be crucial in a clinical setting.

The approach described in Chapter 4 introduces a more flexible control scheme by allowing dynamic priority adjustment between human and agent. However, this flexibility also introduces new challenges. In particular, the supervision mode occasionally resulted in control conflicts, where the user and the agent issued competing commands. This conflict not only affected task fluency but also led to inconsistent experiences among participants. A further limitation lies in the participant pool used for the user studies. All participants were novices with no prior experience using surgical robots. While this choice helped highlight the accessibility and potential of human-robot cooperative frameworks for new users, it limits the applicability of findings to trained surgeons. Experienced users may interact with the system differently, and their feedback could surface more nuanced issues related to control authority, timing, and task execution. As such, Chapter 5 involved more surgical professionals to validate the system.

A primary limitation of the proposed system in Chapter 5 lies in its reliance on the pre-trained gesture recognition model to detect high-level operator gestures. While the Transformer-based classifier achieved good accuracy during validation, its real-time performance was lower due to domain shift. Specifically, the model was trained on data from a different user than those participating in the experiments. This suggests that classification accuracy may degrade when generalising to new users or variations in movement patterns. A second limitation regards the user studies, which were conducted in controlled settings with phantoms and a limited participant pool, mostly composed of novices. While this helped assess the usability and learning curve of the system, it limits the generalisability of the results to expert surgeons or real clinical settings.

Finally, the main limitation of the work in Chapter 6 is the reliance on trial-level OSATS scores for supervision, while the model operates at a frame level. This mismatch in granularity can affect the precision of training and evaluation, especially when short segments do not reflect the overall trial score. Additionally, some models, such as Transformers, require large amount of data to achieve high performance, while the dataset from JIGSAWS constitutes a relatively small sample size with 39 trials.

7.3 Further Work

The work in this thesis represents a step towards the implementation of human-robot cooperation in robotic surgery but several challenges need to be addressed for future clinical adoption.

Future research should focus on enhancing the adaptability of the control framework presented in Chapter 3 by incorporating real-time intent recognition and context-aware task segmentation.

This would allow the system to dynamically shift control between human and agent, generalising to more tasks. Another important direction is to validate the framework in more complex and clinically relevant tasks, such as suturing, similar to Varier et al. [158]. These tasks involve greater variability in movement and timing, presenting a more rigorous test of shared autonomy. Integrating multimodal sensing, such as vision and force, could also enable the system to operate in a richer perceptual space, improving its utility and safety in real surgical procedures.

Chapter 4 showed that controllers that can provide continuous assistance, as in supervision control, enable better performance than discrete approaches. Building on the promising results of the dynamic priority-changing control approach, future work should explore personalised control strategies that adapt to individual user behavior and skill level. Incorporating user modeling or adaptive autonomy could help resolve conflicts observed in supervision mode by adjusting control authority more intelligently. Additionally, integrating trust estimation or confidence metrics may enable smoother and more intuitive collaboration between human and agent, as done in Chapter 5. Further validation is needed with expert surgeons, expanding the evaluation beyond peg transfer which will provide more insights into the method's generalisation and effectiveness.

To extend the capabilities of the C-IAC framework presented in Chapter 5, future work should explore its application integrating richer sensory feedback, such as force sensing or 3D vision, which could improve the confidence estimation mechanism and allow the robot to operate with higher autonomy. Another critical direction involves personalising the system to individual users. By modeling user-specific intent patterns and learning adaptive thresholds for confidence λ , the control framework could offer more intuitive and seamless support. Expanding the dataset with a larger and more diverse group of users, including experienced surgeons, would also help refine gesture classification and better assess generalisation.

Finally, to address the granularity mismatch identified in Chapter 6, future work should focus on collecting or generating finer-grained labels, such as gesture-level or frame-level skill annotations. These would provide more accurate supervision signals aligned with the model's temporal resolution and improve the system's ability to offer precise, real-time feedback during surgical tasks. Future efforts should also explore larger-scale data collection and incorporate data augmentation or synthetic data generation. Architectures like Transformers would then be able to model complex temporal and multimodal patterns in the surgical data.

In conclusion, by spanning shared, supervisory, and assistive control systems, and integrating real-time skill assessment, the work in this thesis proposed intelligent systems in human-robot collaboration frameworks that can effectively support surgical tasks without compromising the surgical workflow. These findings not only advance the technical capabilities of surgical platforms but also underscore their potential to implement automation, reduce workload, and promote safer, more efficient procedures. The impact of this work can extend to the autonomous assistance in other safety-critical domains, most notably autonomous driving [8], which has already demonstrated

how adaptive autonomy, multimodal sensing, and continuous intent prediction can be deployed at scale. As robotic surgery continues to evolve, such systems lay a strong foundation for the next generation intelligent assistance in surgical settings.

References

- [1] Abbink, D. A., Carlson, T., Mulder, M., et al. "A topology of shared control systems—finding common ground in diversity". In: *IEEE Transactions on Human-Machine Systems* 48.5 (2018), pp. 509–525.
- [2] Abbink, D. A., Mulder, M., and Boer, E. R. "Haptic shared control: smoothly shifting control authority?" In: *Cognition, Technology & Work* 14 (2012), pp. 19–28.
- [3] Abbott, J. J., Marayong, P., and Okamura, A. M. "Haptic virtual fixtures for robot-assisted manipulation". In: *Robotics Research: Results of the 12th International Symposium ISRR*. Springer. 2007, pp. 49–64.
- [4] Akshay, S., Megha, Y., and Shetty, C. B. "Machine learning algorithm to identify eye movement metrics using raw eye tracking data". In: *2020 Third International Conference on Smart Systems and Inventive Technology*. IEEE. 2020, pp. 949–955.
- [5] Altobelli, E., Gidaro, S., Bove, A. M., et al. "1405 TELELAP ALF-X: a novel telesurgical system for the 21st century". In: *The Journal of Urology* 189.4S (2013), e575–e576.
- [6] Amsterdam, B. van, Clarkson, M. J., and Stoyanov, D. "Multi-task recurrent neural network for surgical gesture recognition and progress prediction". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 1380–1386.
- [7] Amsterdam, B. van, Clarkson, M. J., and Stoyanov, D. "Gesture recognition in robotic surgery: a review". In: *IEEE Transactions on Biomedical Engineering* 68.6 (2021).
- [8] Ansari, S., Naghdy, F., and Du, H. "Human-machine shared driving: Challenges and future directions". In: *IEEE Transactions on Intelligent Vehicles* 7.3 (2022), pp. 499–519.
- [9] Azari, D. P., Frasier, L. L., Quamme, S. R. P., et al. "Modeling surgical technical skill using expert assessment for automated computer rating". In: *Annals of surgery* 269.3 (2019), pp. 574–581.
- [10] Ballantyne, G. H. and Moll, F. "The da Vinci telerobotic surgical system: the virtual operative field and telepresence surgery". In: *Surgical Clinics* 83.6 (2003), pp. 1293–1304.
- [11] Basañez, L., Nuño, E., and Aldana, C. I. "Teleoperation and Level of Automation". In: *Springer Handbook of Automation*. Springer, 2023, pp. 457–482.
- [12] Beer, J. M., Fisk, A. D., and Rogers, W. A. "Toward a framework for levels of robot autonomy in human-robot interaction". In: *Journal of human-robot interaction* 3.2 (2014), p. 74.
- [13] Benmansour, M. et al. "Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria". In: *International Journal of Computer Assisted Radiology and Surgery* 18.5 (2023), pp. 929–937.
- [14] Berthet-Rayne, P., Power, M., King, H., et al. "Hubot: A three state human-robot collaborative framework for bimanual surgical tasks based on learned models". In: *2016 IEEE International conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 715–722.
- [15] Borden Jr, L. S., Kozlowski, P. M., Porter, C. R., et al. "Mechanical failure rate of da Vinci robotic system." In: *The Canadian journal of urology* 14.2 (2007), pp. 3499–3501.
- [16] Bowyer, S. A. and Baena, F. R. y. "Dissipative control for physical human–robot interaction". In: *IEEE Transactions on Robotics* 31.6 (2015), pp. 1281–1293.
- [17] Cagiltay, N. E. and Dalveren, G. G. M. "Are left-and right-eye pupil sizes always equal?" In: *Journal of Eye Movement Research* 12.2 (2019).

- [18] Cao, Y.-H., Yu, H., and Wu, J. "Training vision transformers with only 2040 images". In: *European Conference on Computer Vision*. Springer. 2022, pp. 220–237.
- [19] Chen, J., Remulla, D., Nguyen, J. H., et al. "Current status of artificial intelligence applications in urology and their potential to influence clinical practice". In: *BJU international* 124.4 (2019), pp. 567–577.
- [20] Chen, X.-l., Cao, L., Xu, Z.-x., et al. "A study of continuous maximum entropy deep inverse reinforcement learning". In: *Mathematical Problems in Engineering* 2019 (2019).
- [21] Chen, L., Hu, Z. J., Huang, Y., et al. "Human Robot Shared Control in Surgery: A Performance Assessment". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 15165–15171.
- [22] Chen, Z., Fan, K., Cruciani, L., et al. "Toward Human-Out-of-the Loop Endoscope Navigation Based on Context Awareness for Enhanced Autonomy in Robotic Surgery". In: *IEEE Transactions on Medical Robotics and Bionics* (2024).
- [23] Chernova, S. and Veloso, M. "Interactive policy learning through confidence-based autonomy". In: *Journal of Artificial Intelligence Research* 34 (2009), pp. 1–25.
- [24] Chua, Z. and Okamura, A. M. "Characterization of real-time haptic feedback from multimodal neural network-based force estimates during teleoperation". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 1471–1478.
- [25] Cimolino, G. and Graham, T. N. "Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–21.
- [26] Conway, L., Volz, R., and Walker, M. "Tele-autonomous systems: Methods and architectures for intermingling autonomous and telerobotic technology". In: *Proceedings. 1987 IEEE International Conference on Robotics and Automation*. Vol. 4. IEEE. 1987, pp. 1121–1130.
- [27] D’Ettorre, C., Mariani, A., Stilli, A., et al. "Accelerating surgical robotics research: A review of 10 years with the da vinci research kit". In: *IEEE Robotics & Automation Magazine* 28.4 (2021), pp. 56–78.
- [28] Da Col, T., Mariani, A., Deguet, A., et al. "Scan: System for camera autonomous navigation in robotic-assisted surgery". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2020, pp. 2996–3002.
- [29] Di Lorenzo, N., Coscarella, G., Faraci, L., et al. "Robotic systems and surgical education". In: *JSLS: Journal of the Society of Laparoendoscopic Surgeons* 9.1 (2005), p. 3.
- [30] DiPietro, R., Lea, C., Malpani, A., et al. "Recognizing surgical activities with recurrent neural networks". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part I* 19. Springer. 2016, pp. 551–558.
- [31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).
- [32] Duchowski, A. T., Krejtz, K., Krejtz, I., et al. "The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-Vis Task Difficulty with Pupil Oscillation". In: *2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. doi: 10.1145/3173574.3173856.
- [33] Enayati, N., Okamura, A. M., Mariani, A., et al. "Robotic assistance-as-needed for enhanced visuomotor learning in surgical robotics training: An experimental study". In: *2018 IEEE International Conference on Robotics and Automation*. 2018, pp. 6631–6636.
- [34] Endsley, M. R. and Kaber, D. B. "Level of automation effects on performance, situation awareness and workload in a dynamic control task". In: *Ergonomics* 42.3 (1999), pp. 462–492.

- [35] Eslamian, S., Reisner, L. A., and Pandya, A. K. "Development and evaluation of an autonomous camera control algorithm on the da Vinci Surgical System". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 16.2 (2020), e2036.
- [36] Finn, C., Levine, S., and Abbeel, P. "Guided cost learning: Deep inverse optimal control via policy optimization". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 49–58.
- [37] Flemisch, F., Abbink, D. A., Itoh, M., et al. "Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human–machine cooperation, cooperative guidance and control, shared, traded and supervisory control". In: *Cognition, Technology & Work* 21 (2019), pp. 555–568.
- [38] Fontanelli, G. A., Selvaggio, M., Ferro, M., et al. "A v-rep simulator for the da vinci research kit robotic platform". In: *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics*. 2018, pp. 1056–1061.
- [39] Fontanelli, G. A., Yang, G.-Z., and Siciliano, B. "A comparison of assistive methods for suturing in MIRS". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 4389–4395.
- [40] Freschi, C., Ferrari, V., Melfi, F., et al. "Technical review of the da Vinci surgical telemanipulator". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 9.4 (2013), pp. 396–406.
- [41] Ganesh, G., Takagi, A., Osu, R., et al. "Two is better than one: Physical interactions improve motor performance in humans". In: *Scientific reports* 4.1 (2014), pp. 1–7.
- [42] Gao, Y., Vedula, S. S., Reiley, C. E., et al. "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling". In: *MICCAI workshop: M2cai*. Vol. 3. 3. 2014.
- [43] Guni, A., Raison, N., Challacombe, B., et al. "Development of a technical checklist for the assessment of suturing in robotic surgery". In: *Surgical endoscopy* 32 (2018), pp. 4402–4407.
- [44] Guo, J., Han, S., and Liu, Y.-H. "MT-ViT: Multi-Task Video Transformer for Surgical Skill Assessment from Streaming Long-term Video". In: *IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2024.
- [45] Guo, Y., Freer, D., Deligianni, F., et al. "Eye-Tracking for Performance Evaluation and Workload Estimation in Space Telerobotic Training". In: *IEEE Transactions on Human-Machine Systems* 52.1 (2022), pp. 1–11. doi: 10.1109/THMS.2021.3107519.
- [46] Guo, Y. and Yang, X. J. "Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach". In: *International Journal of Social Robotics* 13.8 (2021), pp. 1899–1909.
- [47] Hagag, B., Abovitz, R., Kang, H., et al. "RIO: Robotic-arm interactive orthopedic system MAKOpasty: User interactive haptic orthopedic robotics". In: *Surgical Robotics: Systems Applications and Visions*. Springer, 2010, pp. 219–246.
- [48] Haghighipanah, M., Li, Y., Miyasaka, M., et al. "Improving position precision of a servo-controlled elastic cable driven surgical robot using unscented kalman filter". In: *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2015, pp. 2030–2036.
- [49] Hanly, E. J., Marohn, M. R., Bachman, S. L., et al. "Multiservice laparoscopic surgical training using the daVinci surgical system". In: *The American Journal of Surgery* 187.2 (2004), pp. 309–315.

- [50] Hari, K. et al. "STITCH: Augmented dexterity for suture throws including thread coordination and handoffs". In: *IEEE 2024 International Symposium on Medical Robotics (ISMR)*. 2024, pp. 1–7.
- [51] Hart, S. G. "NASA-task load index (NASA-TLX); 20 years later". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 50. 9. Sage CA: Los Angeles, CA. 2006, pp. 904–908.
- [52] Hashemi, N. et al. "Acquisition and usage of robotic surgical data for machine learning analysis". In: *Surgical Endoscopy* 37.8 (2023), pp. 6588–6601.
- [53] Hayati, S. and Venkataraman, S. "Design and implementation of a robot control system with traded and shared control capability". In: *1989 IEEE International Conference on Robotics and Automation*. IEEE Computer Society. 1989, pp. 1310–1311.
- [54] He, K., Zhang, X., Ren, S., et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [55] Heemskerk, J., Gemert, W. G. van, Vries, J. de, et al. "Learning curves of robot-assisted laparoscopic surgery compared with conventional laparoscopic surgery: an experimental study evaluating skill acquisition of robot-assisted laparoscopic tasks compared with conventional laparoscopic tasks in inexperienced users". In: *Surgical Laparoscopy Endoscopy & Percutaneous Techniques* 17.3 (2007), pp. 171–174.
- [56] Hendricks, A., Panoff, M., Xiao, K., et al. "Exploring the limitations and implications of the jigsaws dataset for robot-assisted surgery". In: *IEEE Robotics and Automation Letters* (2024).
- [57] Hira, S. et al. "Video-based assessment of intraoperative surgical skill". In: *International journal of computer assisted radiology and surgery* 17.10 (2022), pp. 1801–1811.
- [58] Ho, J. and Ermon, S. "Generative adversarial imitation learning". In: *Advances in neural information processing systems* 29 (2016).
- [59] Hogan, N. "Impedance control: An approach to manipulation: Part II—Implementation". In: (1985).
- [60] Hu, Z. J., Wang, Z., Huang, Y., et al. "Towards human-robot collaborative surgery: Trajectory and strategy learning in bimanual peg transfer". In: *IEEE Robotics and Automation Letters* 8.8 (2023), pp. 4553–4560.
- [61] Hu, Z. J., Xu, H., Kin, S., et al. "Confidence-based Intent Prediction for Teleoperation in Bimanual Robotic Suturing". In: *arXiv preprint arXiv:2504.20761* (2025).
- [62] Huang, T., Chen, K., Li, B., et al. "Guided reinforcement learning with efficient exploration for task automation of surgical robot". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 4640–4647.
- [63] Huang, Z., Wang, Z., Bai, W., et al. "A Novel Training and Collaboration Integrated Framework for Humanndash;Agent Teleoperation". In: *Sensors* 21.24 (2021).
- [64] Hubens, G., Coveliers, H., Balliu, L., et al. "A performance study comparing manual and robotically assisted laparoscopic surgery using the da Vinci system". In: *Surgical Endoscopy and other interventional techniques* 17 (2003), pp. 1595–1599.
- [65] Huettner, F., Dynda, D., Ryan, M., et al. "Robotic-assisted minimally invasive surgery; a useful tool in resident training—the Peoria experience, 2002–2009". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 6.4 (2010), pp. 386–393.
- [66] Hung, A. J. *Can machine-learning algorithms replace conventional statistics?* 2019.
- [67] Hur, H., Arden, D., Dodge, L. E., et al. "Fundamentals of laparoscopic surgery: a surgical skills assessment tool in gynecology". In: *JSLs: Journal of the Society of Laparoendoscopic Surgeons* 15.1 (2011), p. 21.

- [68] Hwang, M., Ichnowski, J., Thananjeyan, B., et al. "Automating Surgical Peg Transfer: Calibration With Deep Learning Can Exceed Speed, Accuracy, and Consistency of Humans". In: *IEEE Transactions on Automation Science and Engineering* (2022).
- [69] Jarrasse, N., Sanguineti, V., and Burdet, E. "Slaves no longer: review on role assignment for human-robot joint motor action". In: *Adaptive Behavior* 22.1 (2014), pp. 70–82.
- [70] Jin, A., Yeung, S., Jopling, J., et al. "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 691–699.
- [71] Kam, M., Saeidi, H., Hsieh, M. H., et al. "A confidence-based supervised-autonomous control strategy for robotic vaginal cuff closure". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 12261–12267.
- [72] Kaplan, K. E., Nichols, K. A., and Okamura, A. M. "Toward human-robot collaboration in surgery: Performance assessment of human and robotic agents in an inclusion segmentation task". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 723–729.
- [73] Kassahun, Y., Yu, B., Tibebu, A. T., et al. "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions". In: *International journal of computer assisted radiology and surgery* 11.4 (2016), pp. 553–568.
- [74] Kaul, S., Shah, N. L., and Menon, M. "Learning curve using robotic surgery". In: *Current urology reports* 7.2 (2006), pp. 125–129.
- [75] Kaushik, D., High, R., Clark, C. J., et al. "Malfunction of the Da Vinci robotic system during robot-assisted laparoscopic prostatectomy: an international survey". In: *Journal of endourology* 24.4 (2010), pp. 571–575.
- [76] Kazanzides, P., Chen, Z., Deguet, A., et al. "An open-source research kit for the da Vinci® Surgical System". In: *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2014, pp. 6434–6439.
- [77] Kim, D. K., Park, D. W., and Rha, K. H. "Robot-assisted partial nephrectomy with the REVO-I robot platform in porcine models". In: *European urology* 69.3 (2016), pp. 541–542.
- [78] Kim, H.-Y. "Statistical notes for clinical researchers: Nonparametric statistical methods: 1. Nonparametric methods for comparing two groups". In: *Restorative Dentistry and Endodontics* 39.3 (2014).
- [79] Kim, J. W., Zhao, T. Z., Schmidgall, S., et al. "Surgical robot transformer (srt): Imitation learning for surgical tasks". In: *arXiv preprint arXiv:2407.12998* (2024).
- [80] Kim, W. T., Ham, W. S., Jeong, W., et al. "Failure and malfunction of da Vinci Surgical systems during various robotic surgeries: experience from six departments at a single institute". In: *Urology* 74.6 (2009), pp. 1234–1237.
- [81] King, B. W., Reisner, L. A., Pandya, A. K., et al. "Towards an autonomous robot for camera control during laparoscopic surgery". In: *Journal of laparoendoscopic & advanced surgical techniques* 23.12 (2013), pp. 1027–1030.
- [82] Koukourikis, P. and Rha, K. H. "Robotic surgical systems in urology: What is currently available?" In: *Investigative and clinical urology* 62.1 (2021), p. 14.
- [83] Lam, K., Chen, J., Wang, Z., et al. "Machine learning for technical skill assessment in surgery: a systematic review". In: *NPJ digital medicine* 5.1 (2022), p. 24.
- [84] Lauretti, C., Cordella, F., Tamantini, C., et al. "A surgeon-robot shared control for ergonomic pedicle screw fixation". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 2554–2561.
- [85] Lawrie, T. A., Liu, H., Lu, D., et al. "Robot-assisted surgery in gynaecology". In: *Cochrane Database of Systematic Reviews* 4 (2019).

- [86] Lee, J. and Moray, N. "Trust, control strategies and allocation of function in human-machine systems". In: *Ergonomics* 35.10 (1992), pp. 1243–1270.
- [87] Li, K. and Burdick, J. W. "A function approximation method for model-based high-dimensional inverse reinforcement learning". In: *arXiv preprint arXiv:1708.07738* (2017).
- [88] Li, Y., Sena, A., Wang, Z., et al. "A review on interaction control for contact robots through intent detection". In: *Progress in Biomedical Engineering* 4.3 (2022), p. 032004.
- [89] Liang, J., Yang, Y., Wang, Y., et al. "Variable admittance control for human-robot collaboration in robot-assisted orthopedic surgery". In: *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2019, pp. 1544–1550.
- [90] Lillicrap, T. P., Hunt, J. J., Pritzel, A., et al. "Continuous control with deep reinforcement learning". In: *arXiv preprint arXiv:1509.02971* (2015).
- [91] Lin, T., Wang, Y., Liu, X., et al. "A survey of transformers". In: *AI open* 3 (2022), pp. 111–132.
- [92] Liu, J.-C., Li, K.-A., Yeh, S.-L., et al. "Assessing Perceptual Load and Cognitive Load by Fixation-Related Information of Eye Movements". In: *Sensors* 22.3 (2022). ISSN: 1424-8220. DOI: 10.3390/s22031187.
- [93] Long, Y., Wei, W., Huang, T., et al. "Human-in-the-loop Embodied Intelligence with Interactive Simulation Environment for Surgical Robot Learning". In: *arXiv preprint arXiv:2301.00452* (2023).
- [94] Ma, R., Ramaswamy, A., Xu, J., et al. "Surgical gestures as a method to quantify surgical performance and predict patient outcomes". In: *NPJ Digital Medicine* 5.1 (2022), p. 187.
- [95] Maniar, H. S., Council, M. L., Prasad, S. M., et al. "Comparison of skill training with robotic systems and traditional endoscopy: implications on training and adoption". In: *Journal of Surgical Research* 125.1 (2005), pp. 23–29.
- [96] Marcano, M., Díaz, S., Pérez, J., et al. "A review of shared control for automated vehicles: Theory and applications". In: *IEEE Transactions on Human-Machine Systems* 50.6 (2020), pp. 475–491.
- [97] Martin, J., Regehr, G., Reznick, R., et al. "Objective structured assessment of technical skill (OSATS) for surgical residents". In: *British journal of surgery* 84.2 (1997), pp. 273–278.
- [98] Minelli, M., Sozzi, A., De Rossi, G., et al. "Integrating model predictive control and dynamic waypoints generation for motion planning in surgical scenario". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2020, pp. 3157–3163.
- [99] Moccia, R., Iacono, C., Siciliano, B., et al. "Vision-based dynamic virtual fixtures for tools collision avoidance in robotic surgery". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1650–1655.
- [100] Moglia, A., Georgiou, K., Georgiou, E., et al. "A systematic review on artificial intelligence in robot-assisted surgery". In: *International Journal of Surgery* 95 (2021), p. 106151.
- [101] Moustris, G. P., Mantelos, A. I., and Tzafestas, C. S. "Shared control for motion compensation in robotic beating heart surgery". In: *2013 IEEE International Conference on Robotics and Automation*. IEEE. 2013, pp. 5819–5824.
- [102] Nakadi, I. E., Mélot, C., Closset, J., et al. "Evaluation of da Vinci Nissen fundoplication clinical results and cost minimization". In: *World journal of surgery* 30 (2006), pp. 1050–1054.
- [103] Nayyar, R. and Gupta, N. P. "Critical appraisal of technical problems with robotic urological surgery". In: *BJU international* 105.12 (2010), pp. 1710–1713.
- [104] Ngu, J., Tsang, C., and Koh, D. *The da Vinci Xi: a review of its capabilities, versatility, and potential role in robotic colorectal surgery*. *Robot Surg Res Rev*. 2017; 4: 77–85.
- [105] Niemeyer, G., Preusche, C., Stramigioli, S., et al. "Telerobotics". In: *Springer handbook of robotics* (2016), pp. 1085–1108.

- [106] Oh, D. S., Ershad, M., Wee, J. O., et al. "Comparison of Global Evaluative Assessment of Robotic Surgery with objective performance indicators for the assessment of skill during robotic-assisted thoracic surgery". In: *Surgery* 174.6 (2023), pp. 1349–1355.
- [107] Orden, K. F. V., Limbert, W., Makeig, S., et al. "Eye Activity Correlates of Workload during a Visuospatial Memory Task". In: *Human Factors* 43.1 (2001), pp. 111–121. doi: 10.1518/001872001775992570. eprint: <https://doi.org/10.1518/001872001775992570>.
- [108] Orosco, R. K., Lurie, B., Matsuzaki, T., et al. "Compensatory motion scaling for time-delayed robotic surgery". In: *Surgical endoscopy* 35 (2021), pp. 2613–2618.
- [109] Ostrander, B. T., Massillon, D., Meller, L., et al. "The current state of autonomous suturing: a systematic review". In: *Surgical Endoscopy* (2024), pp. 1–15.
- [110] Padoy, N. and Hager, G. D. "Human-machine collaborative surgery using learned models". In: *2011 IEEE International Conference on Robotics and Automation*. 2011, pp. 5285–5292.
- [111] Pasini, N., Mariani, A., Deguet, A., et al. "GRACE: Online Gesture Recognition for Autonomous Camera-Motion Enhancement in Robot-Assisted Surgery". In: *IEEE Robotics and Automation Letters* (2023).
- [112] Payne, C. J., Vyas, K., Bautista-Salinas, D., et al. "Shared-control robots". In: *Neurosurgical Robotics* (2021), pp. 63–79.
- [113] Pedram, S. A., Shin, C., Ferguson, P. W., et al. "Autonomous suturing framework and quantification using a cable-driven surgical robot". In: *IEEE Transactions on Robotics* 37.2 (2020), pp. 404–417.
- [114] Peng, X. B., Andrychowicz, M., Zaremba, W., et al. "Sim-to-real transfer of robotic control with dynamics randomization". In: *2018 IEEE International Conference on Robotics and Automation*. 2018, pp. 3803–3810.
- [115] Pervez, A., Latiffee, H., Ryu, J.-H., et al. "Motion encoding with asynchronous trajectories of repetitive teleoperation tasks and its extension to human-agent shared teleoperation". In: *Autonomous Robots* 43 (2019), pp. 2055–2069.
- [116] Peters, B. S., Armijo, P. R., Krause, C., et al. "Review of emerging surgical robotic technology". In: *Surgical endoscopy* 32 (2018), pp. 1636–1655.
- [117] Pore, A., Tagliabue, E., Piccinelli, M., et al. "Learning from demonstrations for autonomous soft-tissue retraction". In: *2021 International Symposium on Medical Robotics*. IEEE. 2021, pp. 1–7.
- [118] Power, M., Rafii-Tari, H., Bergeles, C., et al. "A cooperative control framework for haptic guidance of bimanual surgical tasks based on learning from demonstration". In: *2015 IEEE International Conference on Robotics and Automation*. 2015, pp. 5330–5337.
- [119] Prewitt, R., Bochkarev, V., McBride, C. L., et al. "The patterns and costs of the Da Vinci robotic surgery system in a large academic institution". In: *Journal of Robotic Surgery* 2 (2008), pp. 17–20.
- [120] Rakita, D., Mutlu, B., Gleicher, M., et al. "Shared control-based bimanual robot manipulation". In: *Science Robotics* 4.30 (2019), eaaw0955.
- [121] Ranne, A., Velikova, Y., Navab, N., et al. "Aiaresseg: Catheter detection and segmentation in interventional ultrasound using transformers". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2024, pp. 8187–8194.
- [122] Rao, P. P. "Robotic surgery: new robots and finally some real competition!" In: *World journal of urology* 36 (2018), pp. 537–541.
- [123] Ren, H., Li, Z., Wu, Q., et al. "Guidance Priority Adaptation in Human-Robot Shared Control". In: *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE. 2022, pp. 1634–1639.

- [124] Reynolds, B. R., Bulsara, C., Zeps, N., et al. "Exploring pathways towards improving patient experience of robot-assisted radical prostatectomy (RARP): assessing patient satisfaction and attitudes". In: *BJU international* 121 (2018), pp. 33–39.
- [125] Ribeiro, M. T., Singh, S., and Guestrin, C. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
- [126] Richter, F., Orosco, R. K., and Yip, M. C. "Motion scaling solutions for improved performance in high delay surgical teleoperation". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 1590–1595.
- [127] Ritter, E. M. and Scott, D. J. "Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery". In: *Surgical innovation* 14.2 (2007), pp. 107–112.
- [128] Robinson, R. D. "Reliable Machine Learning for Medical Imaging Data through Automated Quality Control and Data Harmonization". In: (July 2020).
- [129] Rosenberg, L. B. "Virtual fixtures: Perceptual tools for telerobotic manipulation". In: *Proceedings of IEEE virtual reality annual international symposium*. Ieee. 1993, pp. 76–82.
- [130] Saeidi, H., Opfermann, J. D., Kam, M., et al. "A confidence-based shared control strategy for the smart tissue autonomous robot (STAR)". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 1268–1275.
- [131] Salisbury, J. and Guthart, G. "The Intuitive™ telesurgery system: Overview and application". In: *Proc. IEEE International Conference on Robots and Automation*. Vol. 1. 2000, pp. 618–621.
- [132] Saracino, A., Deguet, A., Staderini, F., et al. "Haptic feedback in the da Vinci Research Kit (dVRK): A user study based on grasping, palpation, and incision tasks". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 15.4 (2019), e1999.
- [133] Sarle, R., Tewari, A., Shrivastava, A., et al. "Surgical robotics and laparoscopic training drills". In: *Journal of Endourology* 18.1 (2004), pp. 63–67.
- [134] Satava, R. M., Stefanidis, D., Levy, J. S., et al. "Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty, multi-institutional randomized control trial". In: *Annals of surgery* 272.2 (2020), pp. 384–392.
- [135] Schulz, C., Schneider, E., Fritz, L., et al. "Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment". In: *British Journal of Anaesthesia* 106.1 (2011), pp. 44–50. issn: 0007-0912. doi: <https://doi.org/10.1093/bja/aeq307>.
- [136] Selvaggio, M., Moccia, R., Ficuciello, F., et al. "Haptic-guided shared control for needle grasping optimization in minimally invasive robotic surgery". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 3617–3623.
- [137] Sen, S., Garg, A., Gealy, D. V., et al. "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization". In: *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 4178–4185.
- [138] Shah, M., Eastman, R. D., and Hong, T. "An overview of robot-sensor calibration methods for evaluation of perception systems". In: *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*. 2012, pp. 15–20.
- [139] Sheridan, T. B. *Telerobotics, automation, and human supervisory control*. MIT press, 1992.
- [140] Shi, C., Zheng, Y., and Fey, A. M. "Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 8017–8024.
- [141] Shinde, N. U., Chiu, Z.-Y., Richter, F., et al. "SURESTEP: An Uncertainty-Aware Trajectory Optimization Framework to Enhance Visual Tool Tracking for Robust Surgical Automation". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 6953–6960.

- [142] Shvets, A. A., Rakhlin, A., Kalinin, A. A., et al. "Automatic instrument segmentation in robot-assisted surgery using deep learning". In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2018, pp. 624–628.
- [143] Sirajudeen, N., Boal, M., Anastasiou, D., et al. "Deep learning prediction of error and skill in robotic prostatectomy suturing". In: *Surgical Endoscopy* (2024), pp. 1–9.
- [144] Sood, A., Jeong, W., Ahlawat, R., et al. "Robotic surgical skill acquisition: what one needs to know?". In: *Journal of Minimal Access Surgery* 11.1 (2015), pp. 10–15.
- [145] Soomro, N., Hashimoto, D., Porteous, A., et al. "Systematic review of learning curves in robot-assisted surgery". In: *BJS open* 4.1 (2020), pp. 27–44.
- [146] Stefanidis, D., Wang, F., Korndorffer, J. R., et al. "Robotic assistance improves intracorporeal suturing performance and safety in the operating room while decreasing operator workload". In: *Surgical endoscopy* 24 (2010), pp. 377–382.
- [147] Sung, G. T. and Gill, I. S. "Robotic laparoscopic surgery: a comparison of the da Vinci and Zeus systems". In: *Urology* 58.6 (2001), pp. 893–898.
- [148] Tahir, N. and Parasuraman, R. "Analog Twin Framework for Human and AI Supervisory Control and Teleoperation of Robots". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2022).
- [149] Takagi, A., Ganesh, G., Yoshioka, T., et al. "Physically interacting individuals estimate the partner's goal to enhance their movements". In: *Nature Human Behaviour* 1.3 (2017), p. 0054.
- [150] Takagi, A., Hirashima, M., Nozaki, D., et al. "Individuals physically interacting in a group rapidly coordinate their movement by estimating the collective goal". In: *Elife* 8 (2019), e41328.
- [151] Takagi, A., Li, Y., and Burdet, E. "Flexible Assimilation of Human's Target for Versatile Human-Robot Physical Interaction". In: *IEEE Transactions on Haptics* 14.2 (2020), pp. 421–431.
- [152] Tao, L. et al. "Sparse hidden markov models for surgical gesture classification and skill evaluation". In: *Information Processing in Computer-Assisted Interventions*. Springer. 2012, pp. 167–177.
- [153] Tokic, M. and Palm, G. "Value-difference based exploration: adaptive control between epsilon-greedy and softmax". In: *Annual Conference on Artificial Intelligence*. Springer. 2011, pp. 335–346.
- [154] Tukra, S., Marcus, H. J., and Giannarou, S. "See-through vision with unsupervised scene occlusion reconstruction". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.7 (2021), pp. 3779–3790.
- [155] Turchetti, G., Palla, I., Pierotti, F., et al. "Economic evaluation of da Vinci-assisted robotic surgery: a systematic review". In: *Surgical endoscopy* 26 (2012), pp. 598–606.
- [156] Van Amsterdam, B., Funke, I., Edwards, E., et al. "Gesture recognition in robotic surgery with multimodal attention". In: *IEEE Transactions on Medical Imaging* 41.7 (2022), pp. 1677–1687.
- [157] Van Den Berg, J., Miller, S., Duckworth, D., et al. "Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations". In: *2010 IEEE International Conference on Robotics and Automation*. 2010, pp. 2074–2081.
- [158] Varier, V. M., Rajamani, D. K., Goldfarb, N., et al. "Collaborative suturing: A reinforcement learning approach to automate hand-off task in suturing for surgical robots". In: *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE. 2020, pp. 1380–1386.
- [159] Varier, V. M., Rajamani, D. K., Tavakkolmoghaddam, F., et al. "Ambf-rl: A real-time simulation based reinforcement learning toolkit for medical robotics". In: *2022 International Symposium on Medical Robotics (ISMR)*. IEEE. 2022, pp. 1–8.

- [160] Vaswani, A., Shazeer, N., Parmar, N., et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
- [161] Volden, F., De Alwis Edirisinghe, V., and Fostervold, K.-I. "Human Gaze-Parameters as an Indicator of Mental Workload". In: *20th Congress of the International Ergonomics Association*. Cham: Springer, 2019. ISBN: 978-3-319-96059-3.
- [162] Wang, Z. and Majewicz Fey, A. "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 1959–1970.
- [163] Weber, J., Catchpole, K., Becker, A. J., et al. "Effects of Flow Disruptions on Mental Workload and Surgical Performance in Robotic-Assisted Surgery". In: *World journal of surgery* 42.11 (2018), pp. 3599–3607.
- [164] Wright, J. D. "Robotic-assisted surgery: balancing evidence and implementation". In: *Jama* 318.16 (2017), pp. 1545–1547.
- [165] Xu, J., Li, B., Lu, B., et al. "SurRoL: An Open-source Reinforcement Learning Centered and dVRK Compatible Platform for Surgical Robot Learning". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2021, pp. 1821–1828.
- [166] Xu, S., Perez, M., Yang, K., et al. "Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator". In: *Surgical endoscopy* 28 (2014), pp. 2569–2576.
- [167] Zabihi, S., Rahimian, E., Asif, A., et al. "Trahgr: Transformer for hand gesture recognition via electromyography". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).
- [168] Zhang, D., Wu, Z., Chen, J., et al. "Human-robot shared control for surgical robot based on context-aware sim-to-real adaptation". In: *2022 International conference on robotics and automation (ICRA)*. IEEE. 2022, pp. 7694–7700.
- [169] Zheng, H., Hu, Z. J., Huang, Y., et al. "A User-Centered Shared Control Scheme with Learning from Demonstration for Robotic Surgery". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 15195–15201.
- [170] Zheng, Y. and Majewicz-Fey, A. "Transformer-based automated skill assessment and interpretation in robot-assisted surgery". In: *2024 International Symposium on Medical Robotics (ISMR)*. IEEE. 2024, pp. 1–8.
- [171] Zia, A. and Essa, I. "Automated surgical skill assessment in RMIS training". In: *International journal of computer assisted radiology and surgery* 13 (2018), pp. 731–739.
- [172] Ziebart, B. D., Maas, A. L., Bagnell, J. A., et al. "Maximum entropy inverse reinforcement learning." In: *Aaai*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438.

Appendix A

Ethics

Ethical approval for the user study conducted in Chapter 3, 4, and 5.

ICREC - SETREC APPLICATION FORM

Part 1 (to be completed by all)

Section 1: Details of Principal Investigator

For all projects the Principal Investigator (PI) must be employed by Imperial College London or hold an honorary contract. For all student projects, the student's supervisor must be named as PI. Student, co-investigator and collaborator details must be added to Section 14.

1. Name (incl. title)	Etienne Burdet
2. Position (at Imperial College London)	Professor of Human Robotics
3. Faculty	Engineering
4. Division/ School/ Department	Bioengineering
5. Email <i>Imperial College email</i>	e.burdet@imperial.ac.uk
7. Summary of skills (experience relevant to the study and in any procedures to be used) (350 characters max)	The PI has 30 years of experience in psychophysical experiments with healthy subjects. Previous studies on human-human interaction, titled "Study of force and stiffness control of the human wrist" and "Physically coupled interaction in human-human and human-robot systems" passed ICREC approval and has yielded fruitful results. Extensive knowledge of human physiology, human sensorimotor control, modelling, electrical stimulation and robotics.

Section 2: Research type	1. Are you conducting research? Yes ☒ No ☐ 2. Are you conducting a service evaluation, audit or public involvement? Yes ☐ No ☒ 3. Does your study only involve analysis of secondary data which is publicly available, and permission is not required to access the data? Yes ☐ No ☒
If you answered no to question 1 and yes to questions 2 or 3, your study does not need ethics review and you will not need to complete this form.	

Section 3: Filter for ICREC and SETREC	1. Is the primary aim of the research answering a human health related question? Yes ☐ No ☒ 2. Is the primary aim of the research answering a non-health related science, social science, engineering or technology related question? Yes ☒ No ☐ 3. Is the primary aim of the research to answer an educational question? Yes ☐ No ☒
---	--

Section 4: Risk level categorisation

This section determines the research risk level and if the application requires full committee review.

- a) Does the research involve drugs/medication? *If yes, please attach the SmPc.*
Yes ☐ No ☒
- b) Does the research involve genetically modified materials? *If yes, please also complete [appendix two](#) and attach the GM Safety Committee letter.*
Yes ☐ No ☒
- c) Will you be recruiting vulnerable participants? i.e. children (15 years or younger), adults (16 years or over) who are unable to consent, people in care, the mentally ill or individuals with learning difficulties?
Yes ☐ No ☒
- d) Will participants take part in the study without their explicit consent? i.e. studies involving deception.
Yes ☐ No ☒
- e) Will you be recruiting prisoners or young offenders?
Yes ☐ No ☒
- f) Is there any aspect of the proposed research which could potentially cause harm to the reputation of the College? i.e. could the research be considered controversial or prejudiced?
Yes ☐ No ☒
- g) Could participants disclose any illegal or harmful activity due to the nature of the research?
Yes ☐ No ☒
- h) Will personally sensitive subjects be discussed that have the potential to induce stress, anxiety or negative consequences for the participant?
Yes ☐ No ☒
- i) Will the researcher be in a position of influence or authority over the participants that could give rise to a perceived pressure to participate? i.e. lecturers/teachers and students.
Yes ☐ No ☒

Section 4: Continued Risk level categorisation This section determines the research risk level and if the application requires full committee review. Meeting dates and submission deadlines ICREC/SETREC.	j) Does the study involve physically intrusive procedures, administration of substances, use of bodily fluids, tissues, DNA or RNA? <i>Use of relevant material must be registered with Imperial College Tissue Bank under the College HTA license.</i> Yes ☐ No ☒ k) Does the study involve ultrasound or sources of non-ionizing radiation? i.e. radiation, MRI, or fMRI. Yes ☐ No ☒ l) Are there any potential conflicts of interest, or what could be perceived by an outside observer as conflicts of interest? Yes ☐ No ☒ m) Will undue incentives for participants be offered? Incentives should be proportionate to the burden imposed and justified by the benefits. Yes ☐ No ☒ n) Are you using any medical device in the UK that is CE/UKCA marked but is being used outside its product limitation? Or are you using any non-CE/non-UKCA marked product(s)? <i>For more information on regulating medical devices.</i> Yes ☐ No ☒ o) Does the proposed research raise any ethical issues that are not covered above? Yes ☐ No ☒
--	---

If you answered **YES TO ANY** of the questions, your study is considered high risk and you must complete the entire application, parts 2, 3 and 4 of this form.

If you answered **NO TO ALL** the questions above, your study is considered low risk. Complete parts 2 and 4, skipping part 3.

Part 2 (to be completed by all)

Section 5: Project Description

1. Full title of study	Tri-manipulation with hand-free interfaces for surgical application
2. P code or cost code and study funder (only if applicable to study)	N/A
3. Lead organisation (who has overall responsibility for the study)	Imperial College London
4. List of location(s) where study will be conducted	Room B436 and 523, Bessemer Building, South Kensington Campus, London SW7 2AZ; Room 415, Sir Michael Uren Hub, White city campus, 86 Wood Ln, London, W12 0BZ
5. Proposed start date <i>From start of advertising and/or recruitment</i>	01/08/2021
6. Proposed end date <i>To end of data collection</i>	01/08/2024

Section 6: Project Summary

Provide a summary of the project in **lay terms**: a brief description of reasons for doing the study, the aims, how data will be disseminated and any expected benefits to the participant, researchers or others. (500 words max)

Our main aim is to understand if/how humans are able to make use of different non-invasive interfaces for the purpose of controlling an external robotic device with and without robotic assistance. By using signal processing and adaptive algorithms we aim to construct natural and intuitive control schemes that can be learnt by an untrained person to control multiple limbs efficiently. The knowledge gained will also be applied to explore the improvement of human and robot intelligence in a shared control and develop a new generation of surgeon training systems for medical teleoperation scenarios.

Recent robotic devices, such as a robotic third arm, have enabled human subjects to augment their capabilities. However, we are still limited in our understanding of how to control these devices and train humans to use them. Uncovering the interfaces best suited for supernumerary limb control opens the possibility for individuals to perform tasks independently even if the task requires more than two hands.

Data Dissemination:

This study is part of the EU H2020 NIMA project. Anonymized data from the participants of the study might be shared with members of these projects. This data will be shared through a protected and restricted cloud storage, where access will only be granted to specific individuals within the research groups. The data shared will be strictly confidential and will not include any identifiable information or contact details. The results will further be published in conference or journal papers, in an anonymized way, ensuring strict confidentiality of the disseminated data.

Benefits

(1) Participants: Participants, after the study, will be told about the impact of the study and where its potential applications lie. They will also be given copies of conference/journal papers if they wish. They will also interact with novel robotics algorithms that we have developed that cannot be experienced elsewhere, which will be of significant interest to the participants.

(2) Researchers: The study will greatly increase our understanding of how to augment human capability using robotic devices. The new insights will lead to further development in medical robotics and supernumerary robotic limb technologies.

(3) Others: The outcomes of the study are foreseen to be substantial; conference and journal papers in high-impact journals will follow the outcome of this study. Connections between the institutions involved in the TRIMANUAL and NIMA projects should have a large impact on further studies and funding applications for EU grants.

Section 7: Research Methods

What methods will you be using in this study? Briefly describe in lay terms: what will happen, the number of times and any data collection techniques. (500 words max)

Experiments will be conducted comparing the effectiveness of different multiple-manual /agent coordination strategies.

- A multiple degree of freedom foot-controlled interface using the pose/force of the subject's foot with/without haptic feedback to foot.*
- Two hand interfaces using the pose of the subject's hands with/without haptic feedback to hands.*
- Robotic systems as supernumerary robotic limb or teleoperated robots, this includes the Kinova robotic arms, Da Vinci surgical robotic systems (No animal or human in-vivo surgical test will be involved, the surgical task operation is limited to simulator or ex-vivo test using three tools.)*

The experiments will be conducted with the subject placed in front of a screen and/or VR interface in which they can observe the external robot(s). Depending upon the subject's experimental group, they will then be connected to the command interfaces. The subjects will sit comfortably, tele-manipulate three or four virtual or realistic robotic arms using hands and feet. We will compare solo manipulation, robot-assisted operation and two person's cooperation. The participants will need to:

- Read and sign an informed consent form and fill a demographic questionnaire.*
- Dominant hand & foot test: the subjects will be asked to fill a standard handed questionnaire and perform a ball-kick test, the preferred hand and leg will be recorded. The duration for this phase last around 5 minutes for each subject.*
- Calibration and habituation phase: The participant will be given 15 to 20 minutes to familiarize with the experiment setup, e.g., how to use the foot, hand interfaces and teleoperate (virtual) robotic arms. mask*
- Training and task phase: The participants will need to move their foot or/and hands to control the robotic arms to perform a serial of multiple-handed operation tasks either by themselves or cooperate with a partner. Between each trial, an appropriate time to rest is allocated.*
- At the end of the experiment, the investigator provides a questionnaire to evaluate the subject's perception of the different conditions. The subject is also asked if they have any questions about the purpose of the study and the results so far (if the participant is interested).*

Depending on the experiment group subjects will be in either a one-session or a multiple-session (conducted for up to ten days) study. The practice and/or test phase will be repeated and last for 30 to 60 minutes each day.

Data Dissemination:

This study is part of the EU H2020 NIMA research projects. The data about human & robot performance will be collected. No identifiable information will be requested or collected from the participants. Non-identifiable and anonymized personal details such as age, gender or

hand/arm preference will be collected through a questionnaire. The data will be pseudo-anonymized. The documentation linking the name with their subject ID (that is used subsequently as anonymized data) will be stored separately on the Imperial cloud storage, will be password protected, and will only be accessible to the lead and co investigators.

Data collected from the experiment will consist of timestamps and data from the neuromechanical and physiological sensing and tracking devices.

This data will be shared through a protected and restricted cloud storage, where access will only be granted to specific individuals within the research groups. The data shared will be strictly confidential and will not include any identifiable information or contact details. The results will further be published in conference or journal papers, in an anonymized way, ensuring strict confidentiality of the disseminated data.

Data Location:

All the raw data is expected to be shared or stored only within the EEA.

Confidentiality: Only demographic information relevant to the study will be recorded (i.e. age, gender and hand/arm preference). Identifying data will only be recorded in consent and will not be published. This will be stored separately of all other data. All other data will only refer to participant IDs.

The data collected from the experiments will be directly stored in the Imperial College cloud storage and will be saved onto a password restricted access shared drive, specifically setup by Imperial ICT. The drive is connected to the Imperial College account database and the restricted access will allow only Imperial College researchers involved in this project to access the data.

Section 8: Participant Recruitment

Provide details of methods of recruitment, participant inclusion and exclusion criteria and the number of participants you are aiming to recruit. Include details of any incentives (such as financial reimbursement). (500 words max)

Attach as separate documents (if applicable):

- Recruitment and advertising material (email, poster, social media advert)
- Oral information scripts

Up to 200 participants may partake in our study as the project encompasses a number of experiments in different operation conditions and using virtual reality or robotic systems. All participants must be:

- 18-65 years of age
- have no motor and/or cognitive impairments.
- have no difficulty in understanding and expressing in English.

Participants outside of these criteria will be excluded.

Participants will be mainly recruited through flyers placed around the Imperial College South Kensington Campus, subject to approval.

The advertising email will be of the form:

“Are you interested in robotics and human-machine interaction? We are looking for participants, with no known neuromuscular disorders or chronic injuries, to take part in a study to further our knowledge of human augmentation, such as controlling a robotic third arm. Participants will use hand and foot interfaces to control multiple robot arms. The experimental protocol is in accordance with the rules established by the Imperial College Research Ethics Committee.

The study takes roughly 40 minutes each day and will last at most ten days and will take place in the Department of Bioengineering at Imperial College, South Kingston campus. £10 /hour will be provided for your valuable contribution. Please email yanpei.huang@imperial.ac.uk for further details.

*Regards,
Yanpei Huang
Human Robotics Group
Department of Bioengineering
Imperial College London”*

The subject compensation for experiments can be reimbursed by the EU H2020 NIMA project.

Section 9: Informed Consent

Include details of how you will be obtaining consent.

- i. Detail the process for ensuring informed consent of all research participants.
- ii. The withdrawal process(es).
- iii. If vulnerable persons are to be used in the study, give separate specific information on how you will ensure consent.
- iv. If participants whose first language is not English are to be recruited, state clearly how the details of the study will be explained, and the consent processed.

i: The study will be explained to each subject by one of the investigators, and they will then be given an information sheet detailing the study aims and methodology. Each subject will then have one day to read and process the information, and then given the opportunity to ask questions to the investigators. Following this process, they will then be asked to fill in and sign a written consent form (of which they will then receive a copy).

ii: Subjects are free to refuse participation or withdraw from the study. This can be done either at the time of recruitment, during data collection, or within 1 year from data acquisition. We will provide each subject with the contact details (both email and telephone) of the principal investigator, Professor Etienne Burdet, on the participant information sheet for this purpose. If the publication of the study results is expected or happens to fall prior to the expiration of the 1-year deadline for withdrawal, the principal investigator will contact the subjects, informing them about their right to withdraw before publication of results. Once participants take the decision to withdraw from the study, all data and records of their participation will be removed and destroyed.

iii: There is no vulnerable person involved in this study.

iv: If the participant has difficulty understanding or expressing themselves in English, they will be excluded from the study as part of the inclusion criteria.

Section 10: Ethical Summary

Has any part of this proposal received prior ethics approval?

Yes ☐ No ☒

Is this study subject to local ethics approval?

Yes ☐ No ☒

If yes, list all local approvals required.

If yes or if rejected, please give details and attach any relevant documents.
(150 words max)

N.A.

Provide details of what you consider to be the ethical issues surrounding this project: your own physical safety, COVID-19 safety measures, data protection/ confidentiality and how you have addressed this. Include details if you will inform participants of the results. If the study is of a sensitive nature include information regarding signposting to relevant support groups.

If you answered yes to any questions in section 3, please provide specific information on those ethical issues and how they will be mitigated. Detail any PPI undertaken as part of

study set up or design.
(500 words max)

As our study participants are healthy adult subjects, the technologies are non-invasive, the experiment consists of common robot teleoperation scenarios, the participants will be supervised by a researcher that has been trained to use the specific devices, and power to the devices can be turned off instantly there are no serious ethical and safety issues. The remote robotic systems are located far from the participants' reach, which will not bring any risk to the operator. Furthermore, subjects will be free to withdraw from the experiment at any moment. Redundant safety measures have been implemented to protect both subjects and researchers, which include software limits, electrical circuit breakers as well as an emergency button.

The subjects and experimenter will follow the current sanitary requirements. For the current covid related measures, they will wear masks during the whole experiment and keep social distancing. If the experimenter/subjects has to be within 1m, they will wear both face shield and mask. Both experimenter and subjects need to have Covid test before the experiment and if they are tested positive, they are not allowed to come to do the experiment.

All data will be anonymised and information capable of identifying participants will not be stored or recorded. Test participants will not be identifiable in any publications. All participant data will be stored on password-protected Imperial College London PCs only accessible to the research team. Subjects will not be told the results of the study in the general instance. If requested it will however be made available.

Section 11: Documentation checklist

Mark as either Yes/
No/ In process

- a. Do Imperial College' insurers need to be notified about your project?

If your project is running abroad and is not qualitative or data only, or if your project is interventional and involves pregnant women, children under 5 or more than 5000 participants you may need additional insurance cover. [Insurance for studies](#), email the [insurance team](#) with any insurance enquiries.

If yes, please provide confirmation that insurance cover has been agreed.

Yes ☐ No ☒ In process ☐

- b. Has your research project been independently peer reviewed?

This can be organised by the [Peer Review Office](#) (within the RGIT). If you answered yes to any questions in section 3, you may be asked to ensure the study is peer reviewed. However, the study does not have to use the RGIT's office for peer review.

	Yes ☐ No ☒ In process ☐
	c. Are you developing a mobile app? See the mobile app webpage for more information. Yes ☐ No ☒ In process ☐
	d. Have you had a Disclosure and Barring Service (DBS) check carried out? If yes, when (add date). For more information about DBS, check government guidance and the College website. Yes ☐ No ☒ In process ☐
	e. Do you need a contractual agreement in place? For further information, please contact your faculty research service. Yes ☐ No ☒ In process ☐
	f. Do you have permissions to use the data in your study? This may be required if you are looking at secondary data. Yes ☒ No ☐ In process ☐
	g. Has Imperial College's Risk Assessment procedure been followed? Contact your departmental administrator for further information. Yes ☐ No ☒ In process ☐

Section 12: Confidentiality and management of personal and other research data	a. I understand it is the responsibility of the researcher to ensure all research data is securely stored during and after the study in accordance with College Guidelines, Codes of Practice, Policies and Procedures. Yes ☒ No ☐
	b. I confirm that all the processing of personal information related to the study will be in full compliance with the GDPR. Including but not limited to, the creation of all necessary documentation (PIS, Data Protection Impact Assessments, Consent forms etc.) Yes ☒ No ☐

Part 3 (only to be completed if **yes** was answered to any question in section 4)

Section 13: Mitigation of Risks and Safeguarding

Explain the precautions taken to protect the health and safety of researchers, participants and others associated with the project.

You need to safeguard the wellbeing and safety of children and adults at risk involved in research activities. Safeguarding means taking all reasonable steps to prevent harm, exploitation, and abuse from occurring; protecting people, especially adults 'at risk' and children, from that harm; and responding appropriately when harm does occur.

- Explain what information you have on the potential harms this research can address or exacerbate for researchers, participants and wider communities.
- Explain how you are building the rights of potential or actual victims/ survivors of safeguarding incidents into the research design, including questions and methodology, to ensure respect, dignity and safety.

Visit the website for more information on [safeguarding for research](#). (500 words max)

N.A.

- I will ensure that access to (community-based – for studies outside of the UK) complaint mechanisms to raise safeguarding concerns are built into the programme design and are discussed and explained with participants.
Yes ☐ No ☐
- I am willing to modify or even cancel planned research if potential harm to researchers, participants or communities is too great.
Yes ☐ No ☐
- I will ensure that we and our research partners reach a shared understanding of safeguarding
Yes ☐ No ☐

Part 4 (to be completed by all)

Section 14: Co-investigators/ Collaborators

If there are more than four co-investigators, please use a separate sheet and follow the format below.

1. Name	Yanpei Huang
2. Position <i>Incl. organisation, company, institution</i>	Research Assistant
3. Role in the study <i>(what contributions you will make and relevant experience)</i>	Interface design, experimental practicalities, data analysis.
4. Email <i>Work not personal</i>	yanpei.huang@imperial.ac.uk

1. Name	Jonathan Eden
2. Position <i>Incl. organisation, company, institution</i>	Marie Sklodowska-Curie Individual Fellow
3. Role in the study <i>(what contributions you will make and relevant experience)</i>	Experimental practicalities, data analysis, and modelling of the data.
4. Email <i>Work not personal</i>	j.eden@imperial.ac.uk

1. Name	Ziwei Wang
2. Position <i>Incl. organisation, company, institution</i>	Research Associate
3. Role in the study <i>(what contributions you will make and relevant experience)</i>	Robot control and modelling, data analysis.
4. Email <i>Work not personal</i>	ziwei.wang@imperial.ac.uk

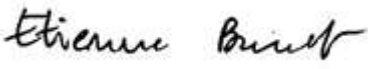
1. Name	
2. Position <i>Incl. organisation, company, institution</i>	
3. Role in the study <i>(what contributions you will make and relevant experience)</i>	

4. Email <i>Work not personal</i>	
--------------------------------------	--

Signatures Page - PI Declaration

I declare that:

- I undertake to abide by the ethical principles underlying the Declaration of Helsinki (1964) and subsequent amendments and good practice guidelines on the proper conduct of research.
- I undertake to abide by the Data Protection Act 2018 and General Data Protection Regulation (Europe) and any applicable local laws.
- I undertake to abide by all local laws and regulations for non-UK research.
- I will report any adverse or unforeseen events or protocol violations and deviations which occur to the Ethics and Research Governance Co-ordinator within 24 hours.
- I will provide an [annual progress report](#) of the project until the end of the study.
- If I register my study on a public database, i.e. ClinicalTrials.gov, I will report results on that database within one year of study completion.
- I will provide [notification of the end or early termination of](#) the research project.
- I will provide [notification of amendment](#) to ICREC/SETREC if there are any changes to the research protocol or personnel which affect the ethical aspects of the project.
- I will assist ICREC/SETREC in any continuing review of the project deemed necessary by the Committee or Faculty Members.
- All information on this form is correct.

PI Name	Etienne Burdet	
PI Signature		Date 12/07/2021
If full committee review is required would you be willing to attend the ICREC/SETREC meeting to answer any questions about your proposal?		Yes

Any attendance must be by the PI named in section four. Attendance at the meeting will give you the opportunity to answer any ethics questions raised by the committee.

Head of Department (please indicate below your decision and the reasons for it)			
Decision	Approved	Referral to Committee	Not required
Reason	All ethical issues have been addressed. This is an experienced group		
Signature		Date 13 th July 2021	
Name	Anthony M J Bull		

Appendix B

Image Reuse Permissions

Springer Nature Images

All images replicated in this thesis from Springer Nature publications have been referenced according to the following format: (©[year of publication] reproduced from [authors] [reference] with permission from Springer Nature). Individual licenses for each reused image have been obtained from Springer Nature:

- Figure 2.1. Copyright Clearance Center Order License ID: 1612952-1
- Figure 2.2. Copyright Clearance Center Order License ID: 1612952-1
- Figure 2.3. Copyright Clearance Center Order License ID: 1607888-1

Wolters Kluwer Health, Inc.

All images replicated in this thesis from Wolters Kluwer Health, Inc. publications have been referenced according to the following format: (©[year of publication] reproduced from [authors] [reference] with permission from Wolters Kluwer Health, Inc.). Individual licenses for each reused image have been obtained from Wolters Kluwer Health, Inc.:

- Figure 2.4. Wolters Kluwer Health, Inc. License Number: 6027681233158

The Institute of Electrical and Electronics Engineers, Incorporated (IEEE)

All images replicated in this thesis from IEEE publications have been referenced according to the following format: (©[year of publication] reproduced from [authors] [reference] with permission from IEEE). Individual licenses for each reused image have been obtained from IEEE:

- Figure 2.5. Copyright Clearance Center Order License ID: 1607899-1
- Figure 2.7. Copyright Clearance Center Order License ID: 1607899-1