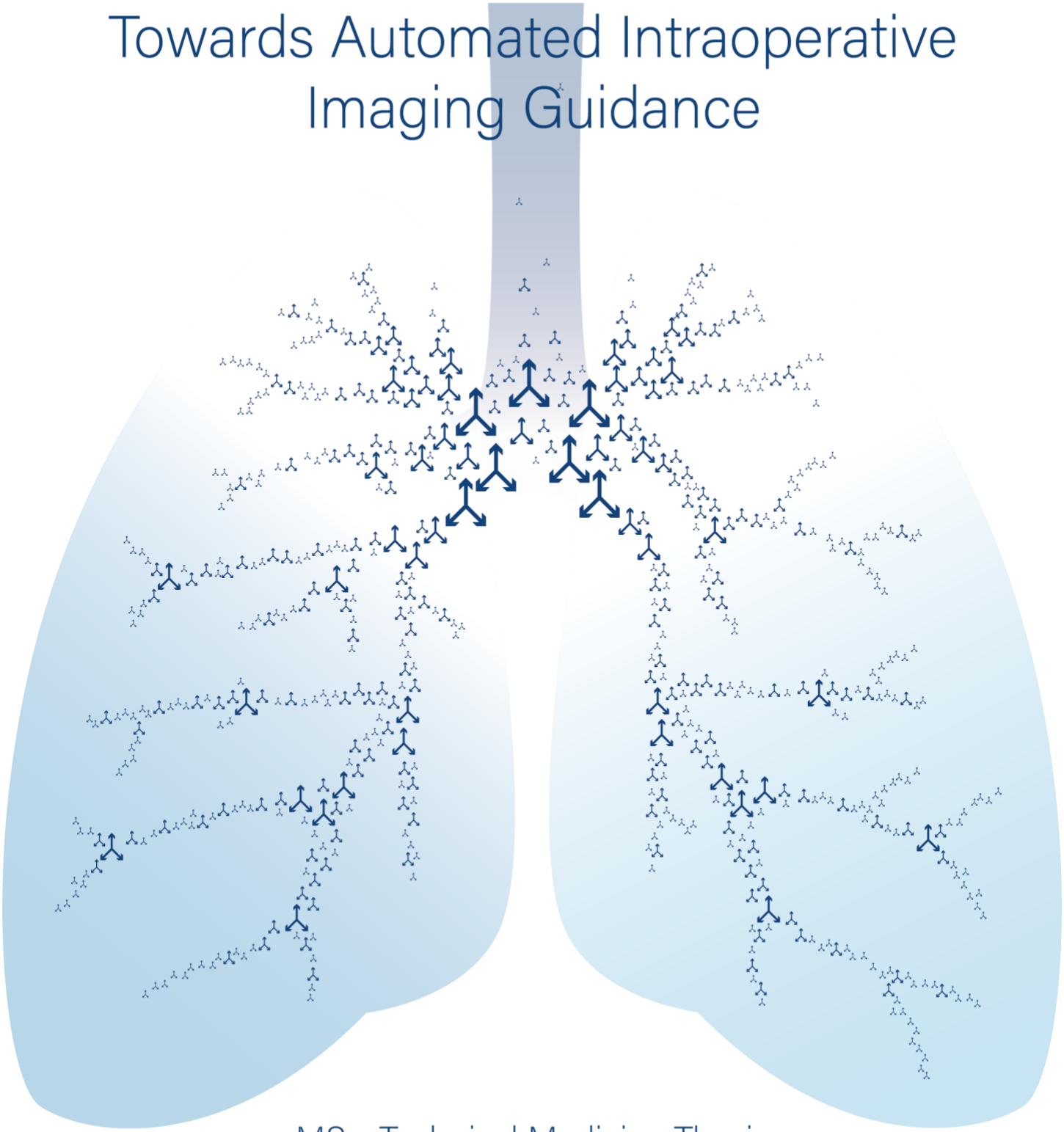
Phase Recognition for Pulmonary Orientation Detection:

Towards Automated Intraoperative Imaging Guidance



MSc. Technical Medicine Thesis

M.C.J. Doornbos

Leids Universitair Medisch Centrum

TU Delft

Erasmus MC
University Medical Center Rotterdam

Phase Recognition For Pulmonary Orientation Detection: Towards Automated Intraoperative Imaging Guidance

by

Marie-Claire Doornbos

Student number: 4492951

[16 01 2024]

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Department of Cardiothoracic Surgery, Erasmus MC

July 2023 – January 2024

Supervisor(s):

Dr. Amir H. Sadeghi, MD, PhD, Erasmus MC

Dr. Bart Cornelissen, KT, PhD, Erasmus MC

Thesis committee members:

Prof. Dr. John J. van den Dobbelsteen, TU Delft, Erasmus MC & LUMC

Dr. Amir H. Sadeghi, MD, PhD, Erasmus MC & UMC Utrecht

Dr. Bart Cornelissen, KT, PhD, Erasmus MC

An electronic version of this thesis is available at http://repository.tudelft.nl

Leids Universitair Medisch Centrum

TU Delft

Erasmus MC University Medical Center Rotterdam

Preface

This graduation project presents the final work I have delivered within the study program of Technical Medicine. Reflecting on these eventful years, I am happy with choosing this path, combining my interest in technology with a passion for healthcare. The importance of our profession has become increasingly clear to me and I have enjoyed experiencing this in daily practice. While challenging myself on a technical level, I got the unique opportunity of independently running pre-operative out-patient clinics for heart and lung surgery. I'm very grateful to have been given this responsibility and experience, as preparation for my upcoming studies in Medicine.

I extent my gratitude to my supervisors, Amir and Bart, for their guidance throughout my (slightly longer than usual) research journey. Their flexibility and dedication over 20 months, allowed me to conclude my Technical Medicine journey, while making a smooth transition to my upcoming medical career. The weekly meetings with Bart provided a platform to resonate all my questions and thoughts, keeping me focused and aligned with the end goal. Thank you for your comprehensive feedback and chats. Meanwhile, Amir offered me hands-on clinical experience in the operating room, challenging me to fully understand the cardiothoracic anatomy and enthusing me with his far-reaching interest in medical technology. The many iterations and feedback sessions with Amir for the submission of my literature study challenged me to remain critical on my own work and encouraged me to philosophize about innovative applications of this research. Special thanks to Quinten, who started off as a close study friend, gradually growing into his role as my daily supervisor during his PhD. He always managed to cheer me up if needed, as both a colleague and friend. I would like to extend my thanks to Yiping, Yasmina and Ronald, who pitched in when it was most needed. Your enthusiasm about my research topic was contagious and I couldn't have done it without your generous and indispensable technical support.

Finally, I want to express my appreciation to my family, Philip, roommates and friends, for their unconditional support during the past years. Additionally, I would like to thank all my colleagues and friends from RG6 for always being available for a coffee, chat, or fun activity proving the much-needed balance to of our hard work.

List of Abbreviations

3D	Three-Dimensional
AI	Artificial Intelligence
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
EMC	Erasmus Medical Centre
LLL	Left Lower Lobe
LUL	Left Upper Lobe
MIS	Minimally Invasive Surgery
MS-TCN	Multi-Stage Temporal Convolutional Neural network
NSCLC	Non-Small Cell Lung Cancer
PPV	Positive Predictive Value
RATS	Robot-Assisted Thoracic Surgery
RLL	Right Lower Lobe
RML	Right Middle Lobe
RUL	Right Upper Lobe
TeCNO	Temporal Convolutional Networks for the Operating room
TPR	True Positive Rate
VATS	Video-Assisted Thoracic Surgery

Table of Contents

Marie-Claire J. Doornbos[1,2], Quinten J. Mank[1], Bart M.W. Cornelissen[3], Amir H. Sadeghi[4]

1. Department of Cardiothoracic surgery, Thoraxcenter, Erasmus MC, Rotterdam, The Netherlands
2. Educational program Technical Medicine; Leiden University Medical Center, Delft University of Technology & Erasmus University Medical Center Rotterdam, The Netherlands.
4. Department of Cranio-Maxillofacial surgery, Erasmus MC, Rotterdam, The Netherlands
4. Department of Cardiothoracic surgery, University Medical Center Utrecht, Utrecht, The Netherlands

Abstract

Objective: This study introduces a novel deep-learning-based orientation recognition approach for detecting intraoperative lung orientation during robot-assisted anatomical resections, including lobectomy and segmentectomy. This method can potentially aid in anatomical structure identification, facilitate training and education, improve procedural efficiency, and enhance intraoperative imaging navigation.

Methods: We developed a unique dataset encompassing various pulmonary procedures, being the first to report on recognition of intraoperative orientation. The TeCNO model, initially developed for laparoscopic cholecystectomies, was adapted for this study. Model performance was evaluated using accuracy, precision, recall, and F1-score, and we explored the influence of dataset composition, intraoperative factors such as 3D model presence, and visual impairments.

Results: The model achieved an overall accuracy of 70%, indicating potential in recognizing lung orientation. High performance was achieved in recognizing non-surgical sequences, 'Fissure', and 'Inferior' views. 'Posterior' and 'Anterior' views showed inferior performance. Variability in performance was attributed to the heterogeneity of orientation transitions and increased complexity compared to more standardized procedures. The limited dataset size and imbalances in label distribution potentially impacted model performance.

Conclusion: This study demonstrates the feasibility of applying phase recognition to detect orientation of the lung and exploring how the unique characteristics of our dataset affect model performance opposed to surgical phase recognition. The results suggest promising applications for intraoperative imaging guidance and automated adjustment of 3D models, particularly for complex orientations like the interlobar 'Fissure view'. Future research should focus on enhancing model performance and assessing its clinical implementation in diverse surgical settings.

Keywords: Artificial Intelligence, Deep-Learning, Phase Recognition, Orientation Recognition, Lung Lobectomy, Lung Segmentectomy

I. Introduction

Non-small cell lung cancer (NSCLC) is one of the most common types of cancer, leading to cancer-related death worldwide, since advanced-stage diagnosis often results in limited treatment options [1, 2]. Earlier diagnosis permits surgical resection which has become the standard of care. Within pulmonary surgery, minimally invasive surgery (MIS) procedures are procedures are widely performed and both video- and robot-assisted thoracic surgery (VATS/RATS) show improved surgical outcomes compared to the equivalent open procedure [3]. MIS provides a three-dimensional (3D) operating view, assisting surgeons in identifying essential anatomical structures [2, 3]. Nevertheless, MIS presents challenges, such as limited control over intraoperative bleeding and restricted flexibility of surgical instruments. The introduction of RATS has addressed some of these limitations, reducing surgeons' discomfort, increasing precision through wristed instrumentation, and providing improved 3D depth perception [2, 4]. While RATS has shown improvement regarding safety, cost-effectiveness and surgical outcomes, the significant learning curve remains a limiting factor [2, 5, 6]. In addition, proximity of the thoracoscope to target structures can impose limited visibility, presenting additional challenges in recognizing anatomical structures [7].

Given patient-specific anatomical variations of bronchovascular anatomy, lobectomy and segmentectomy procedures can be complex, further limiting intraoperative recognition of anatomy [8]. This emphasizes the importance of pre-operative planning, for which computed tomography (CT) imaging currently is the gold standard [9, 10]. Artificial Intelligence (AI)-enhanced surgical planning tools are emerging that use CT images to automatically generate 3D reconstructions and enable visualization through extended reality methods [9, 11]. These tools create a pre- and intraoperative environment that provides surgeons with 3D anatomical understanding [8, 11, 12].

Hence, MIS lobectomy and segmentectomy procedures can be guided by patient specific 3D models, providing improved spatial orientation and insight to the complex bronchovascular anatomy, including intrathoracic vessels, bronchi and the tumor [10, 13]. Currently, the orientation of the 3D lung model is manually adjusted to correspond to the intraoperative situation. Lack of associative connection between the 3D model and intraoperative surgical view requires manual input from either the surgeon or an assistant, relying on their anatomical and technical expertise. This remains a major limitation [14]. To improve intraoperative guidance, there is a need to develop a method to automatically adapt the orientation of the patient-specific 3D models. Achieving this requires identification and detection of intraoperative anatomy and orientation during lobectomy and segmentectomy procedures [15]. An imaging analysis method to do so is phase recognition [16].

Automated phase recognition uses AI to identify different surgical phases, by employing deep-learning (DL) algorithms that have been trained on annotated datasets of surgical videos [16, 17]. Current algorithms are designed to match video segments to specific surgical procedure steps. Previous work typically utilized a two-stage modelling methodology including the feature extraction capability of a Convolutional Neural Network (CNN) combined with a model leveraging the temporal relationship between current and prior/future video frame [18, 19]. To effectively train these models and perform research on the best approaches for phase recognition, access to high-quality annotated datasets is essential. Due to the time-intensive and expertise-dependent nature of surgical

video annotation and the privacy concerns regarding medical data, there is a scarcity of open-source datasets [20]. To the best of our knowledge, no open-source datasets are available for MIS lobectomy or segmentectomy procedures.

I.I. Goals & Objectives

This study aims to implement a DL surgical phase recognition algorithm for automatic detection of the intraoperative orientation of the lung during RATS, hereafter referred to as orientation recognition. Our objective is to understand how the unique characteristics of our dataset impact the performance of this approach.

II. Methods

II.I Study Design

A retrospective single-centre cohort study was conducted at the Cardiothoracic Surgery Department of the Erasmus Medical Centre (EMC), the Netherlands. We curated a dataset of 27 available RATS videos, collected between December 2022 and December 2023. All patients underwent a robot-assisted lobectomy or segmentectomy procedure for NSCLC. Patients were included after obtaining informed consent, approved by the Institutional Medical Ethical Committee (MEC-2023-008/MEC-2023-0397) and all data was anonymized and handled according to the EMC privacy guidelines. Procedures were performed by two experienced cardiothoracic surgeons and one senior cardiothoracic resident.

II.II Orientation Definition

Distinct orientations were defined based on intraoperative views encountered during lobectomy/segmentectomy procedures. Throughout these procedures, lung orientation varies across actions and phases and differs between left and right-sided resection of the lungs. In collaboration with a cardiothoracic surgeon, five orientations of the lung parenchyma or pulmonary hilar/arterial structures were identified as crucial for 3D model orientation (Fig 1.).

1. Anterior view: the lung parenchyma is retracted posteriorly. Left-sided resections include visibility of pericardial tissue and the thoracic wall on the left side, and the lung parenchyma visible on the right. Right-sided resections are characterized by the appearance of the lung parenchyma on the left, with the pericardial tissue often visible on the right.
2. Posterior view: the lung parenchyma is retracted anteriorly. For resections of the left lung, the aorta is prominently visible on the right and the lung parenchyma on the left. For right lung resections the thoracic wall is apparent on the left, and the lung parenchyma on the right. The pericardial tissue may become visible on the left/inferior aspect of the video view.
3. Inferior view: the lung parenchyma is retracted superiorly, with the lung parenchyma visible at the top of the video view. This lung orientation is commonly related to the surgical release of the pulmonary ligament.

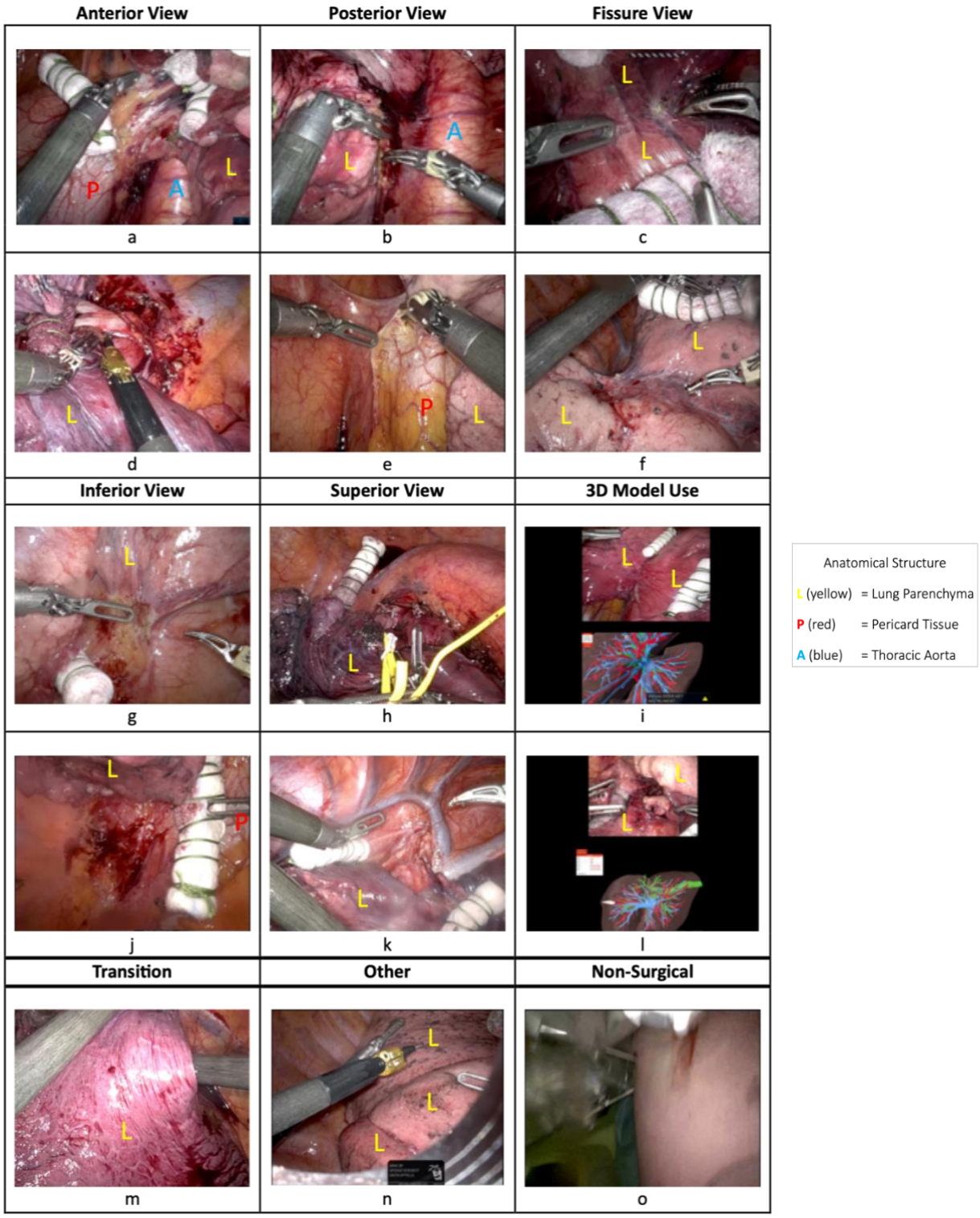


Figure 1. Examples of Pulmonary Orientation Definitions. a. Anterior View (left lower lobe resection. b. Posterior View (left lower lobe resection). c. Fissure View (left lower lobe resection). d. Anterior View (right upper lobe resection) e. Posterior View (right lower lobe resection). f. Fissure View (right lower lobe resection). g. Inferior View (left lower lobe resection). h. Superior View (left upper lobe resection). i. 3D Model use during Fissure View j. Inferior View (right lower lobe resection). k. Superior View (right lower lobe resection). l. 3D Model use during fissure view. m. Transition between two orientation n. View that is not recognized as one of the predefined orientations and labelled as "Other". o. The thoracoscope is retracted from the patient, labelled as "Non-Surgical".

4. Superior View: the lung parenchyma is retracted inferiorly, visualizing the lung parenchyma in the bottom of the video view. Left-sided resections may specifically include visibility of the vessels of the aortic arch.
5. Fissure view: characterized by the separation of two lung lobes, revealing the interlobar fissure.

Three supplementary categories are defined: 'Transition', 'Other' and 'Non-surgical'. 'Transition' involves all video sequences with an extended transition between orientations, defined by the moment when an instrument starts manipulating the lung parenchyma into a new orientation. 'Transition' concludes when a new orientation can be distinctly identified. Video sequences labelled as 'Other' are more heterogeneous, encompassing blurred vision preventing orientation identification or any orientation of the lung parenchyma that defies classification within the predefined orientations. 'Non-surgical' pertains to any video sequence where no clear surgical perspective is identifiable due to the retraction of the thoracoscopic video scope outside of the body. The start and end of the Non-Surgical' is defined by the instrument port (increasingly) becoming (in)visible.

II.III Pre-processing

The surgical videos were recorded with an 8mm robotic endoscope camera (Da Vinci Xi plus, 30-degree angle) with a frame rate of 60 frames per second (fps) and a resolution of 1440x900 or 1280x1024 pixels. Pre-processing of the video data consisted of shortening the videos by defining new start and end-times, converting the frame rate to 25 fps and resizing the videoframes to 1125x900 pixels. Detailed steps are provided in Appendix A.

II.IV Labelling

Labelling was performed through the Anvil video annotation tool [18]. Annotations included framewise annotation of the intraoperative orientation of the lung. 3D model presence, using the TilePro extension of the DaVinci Robot, and instrument port visibility, was annotated to analyze intraoperative use of the 3D model and potential visual impairment. To guarantee the standardization and reproducibility of annotations, and to mitigate potential sources of error, explicit rules for annotation were established in discussion with a cardiothoracic surgeon, which are provided in Appendix B.

A MSc Technical Medicine thesis candidate (MCD) of the Department of Cardiothoracic Surgery at the EMC, the Netherlands, labelled videos of lobectomy and segmentectomy procedures of the varying lung lobes as instructed. Three feedback sessions (after 40%, 80% and 100%) with a Cardiothoracic Surgeon (AS) and a Technical Physician (QM), both experienced in performing RATS lobectomy and segmentectomy procedures, were conducted for revision of the labelling method on five randomly selected videos. An inter-rater orientation transition agreement and a Cohen's Kappa score [21] per category, similar to annotation revision in the HeiChole challenge [20], expressed the variability between annotator and experts. Intra-observer variability was investigated for three videos, one of each revision session, after a period of 4-6 weeks, using Cohen's Kappa score. Following the completion of all feedback sessions, a thorough error correction was carried out across all videos. Concurrently, the experts were consulted for consensus on complex or challenging video sequences.

II.V Orientation Recognition Model

The Temporal Convolutional Networks for the Operating room (TeCNO) [22] was adopted as a two-stage modelling approach for surgical orientation recognition. This approach utilizes a CNN to extract features from individual frames, without any temporal context, followed by the implementation of a Multi-Stage Temporal Convolutional Neural network (MS-TCN), that captures sequential dynamics [18].

Stage 1: CNN Feature Extraction

For visual feature extraction, without temporal context, the deep residual convolutional neural network architecture of ResNet50 [23] was employed as a single-task network for orientation recognition [22]. For each frame, the model estimated probability distributions across all possible orientations, to indicate the probability of the frame belonging to each catagory [18].

Stage 2: MS-TCN Temporal Aggregation

Frames before and after a given frame may provide useful information to predict the likelihood of that frame belonging to a certain orientation [24]. A MS-TCN can establish the relationship between current and prior/future frames of a surgical video [18, 22]. We have utilized the MS-TCN of TeCNO, since this approach has recently achieved state-of-the art results for surgical phase recognition in the Cholec80 dataset [22, 24] and showed best performance compared to Trans-SVNet (transformer-based architecture) on real-world data [18, 25]. The output of the MS-TCN is an orientation prediction for each frame in the input sequence [24]. TeCNO only relies on future frames, enabling potential intraoperative use. Furthermore, TeCNO allows for a significant reduction in computational cost compared to other models [22]. The complete TeCNO model for our dataset can be found in figure 2.

II.VI Model Training

Our dataset was split in a train-validation-test ratio of 50:20:30, following a previously described method [26], ensuring representation of both types of procedures (segmentectomy/lobectomy) and the different sides (left/right) in each group. All hyperparameters were tuned during experimental runs to select the model that performed best on the test set. We implemented our method in PyTorch [27], training our models on a NVIDIA Quadro RTX 6000 24GB GPU.

II.VII Evaluation Methods

To comprehensively measure the performance of the model on our specific dataset, we employ various evaluation metrics common for surgical phase recognition, including Accuracy, Precision, Recall and F1 score [26]. Accuracy is defined as the proportion of correct predictions in each video. The overall accuracy is determined by computing the average across all orientations and videos, to ensure each orientation and video is weighted equally, independent of the occurrence rate of the orientation or video length [24]. Precision is the positive predictive value (PPV) and checks if the orientation is recognized incorrectly. Recall is the true positive rate (TPR) which checks whether parts of an orientation sequence are missed. Subsequently, a F1 score is employed to measure both how accurately and comprehensively an orientation is recognized. The F1 score can be computed by taking the mean of the mean precision and mean recall, considering false positives and negatives

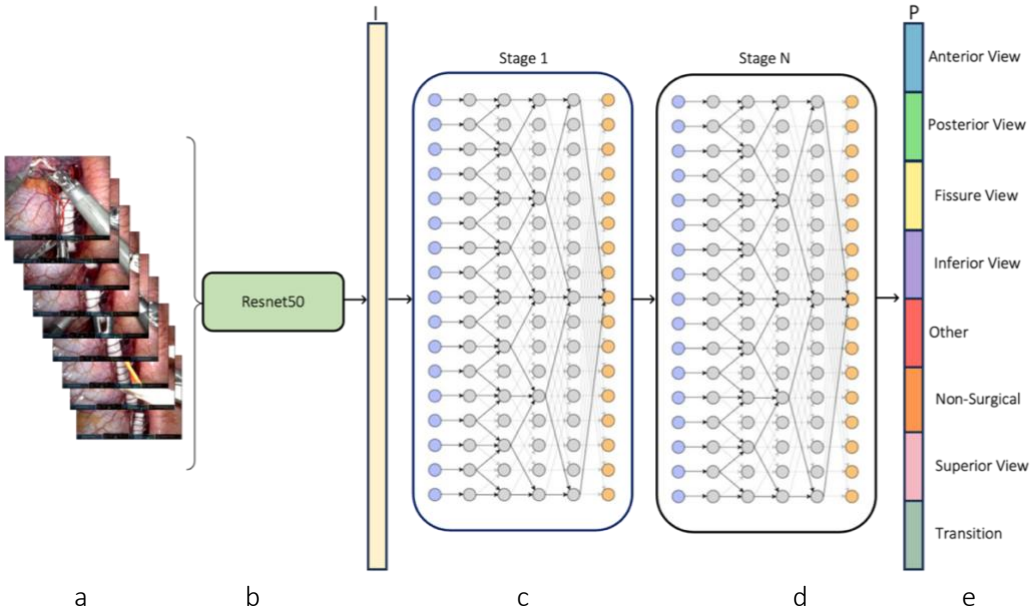


Figure 2. Overview of the proposed neural network, using a Multi-Stage Temporal Convolution Network [28] a. The lobectomy and segmentectomy videos are split into video frames using a frame rate of 25 frames per second (fps). b. Feature extraction is performed using a deep convolutional neural network (CNN), ResNet50 [23]. For each frame the predicted feature vector, expressing the visual information content of the video frame, is compared to the original frame. c. All feature vectors are combined into a sequence of feature vectors, serving as input to the MS-TCN model. d. The MS-TCN captures the temporal information in between video frames. The multiple stages of TCN's allows refinement of previous stage predictions at every step [22, 29]. e. The MS-TCN model provides orientation phase predictions for each video frame.

[26]. Precision and Recall are computed per orientation per video and subsequently averaged over all videos. Final scores are obtained by averaging the orientation-wise scores.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} * 100$$

III. Results

III.I Annotation

A total of 27 procedures were annotated. Inter-rater agreement (between student and expert) and intra-rater agreement, evaluated across three revision sessions, is presented in Appendix C and figure 3. For inter-rater agreement, the orientation transition agreement score varied from 62.5% for a video in the first revision session to 94.74% for a video of the final revision session. The Cohen's Kappa Scores ranged from 0.81 to 1, indicating a strong to nearly perfect level of agreement Intra-rater agreement, evaluated with a Cohen's Kappa score, varied from 0.75 (moderate) for a video from the first revision session to 0.96 (almost perfect) for a video of the second revision session.

III.II Dataset

The annotated dataset, consisting of 27 annotated RATS lobectomy and segmentectomy procedures, is visualized in Figure 4. The distribution of the dataset split can be found in Appendix D. The videos in our dataset had a median duration of 127 minutes. The shortest video in our dataset was 30 minutes and the longest video was 285 minutes. These durations consider the adjusted start and end times.

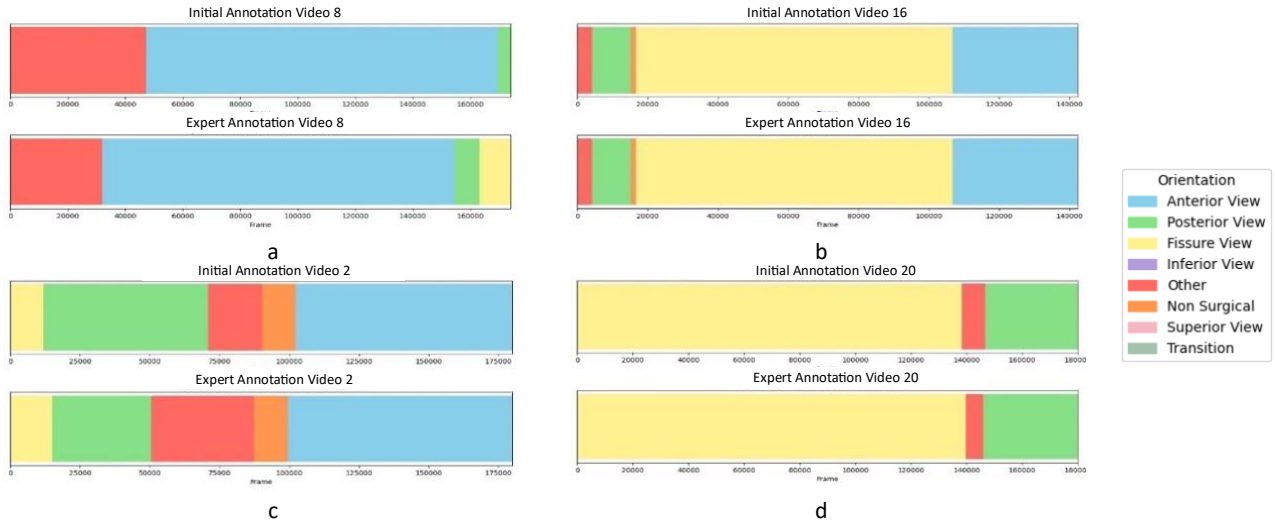


Figure 3. Timeline plots of inter- and intra-rater agreement, illustrating orientation transition errors. Each phase is encoded by a different color explained in the legend. a. Video 08 (revision session 2): lowest inter-rater agreement. b. Video 16 (revision session 3): highest inter-rater agreement. c. Video 02 (revision session 1): lowest intra-rater agreement. b. Video 20 (revision session 2): highest intra-rater agreement.

Figure 5a. shows how the orientation annotations are distributed across the complete dataset. The exact number of frames and percentages are incorporated in Appendix E. The orientation with a 'Fissure view' appeared most in our data as 41.8% of all video frames were annotated with this label. 'Posterior view' and 'Anterior view' follow at 25.4% and 15.4%. Of the five identified orientations, 'Inferior view' appeared least with 3.2% of all video frames annotated with this label. 12.0% of the dataset is annotated as 'Other' and 2.2% as 'Non-Surgical'.

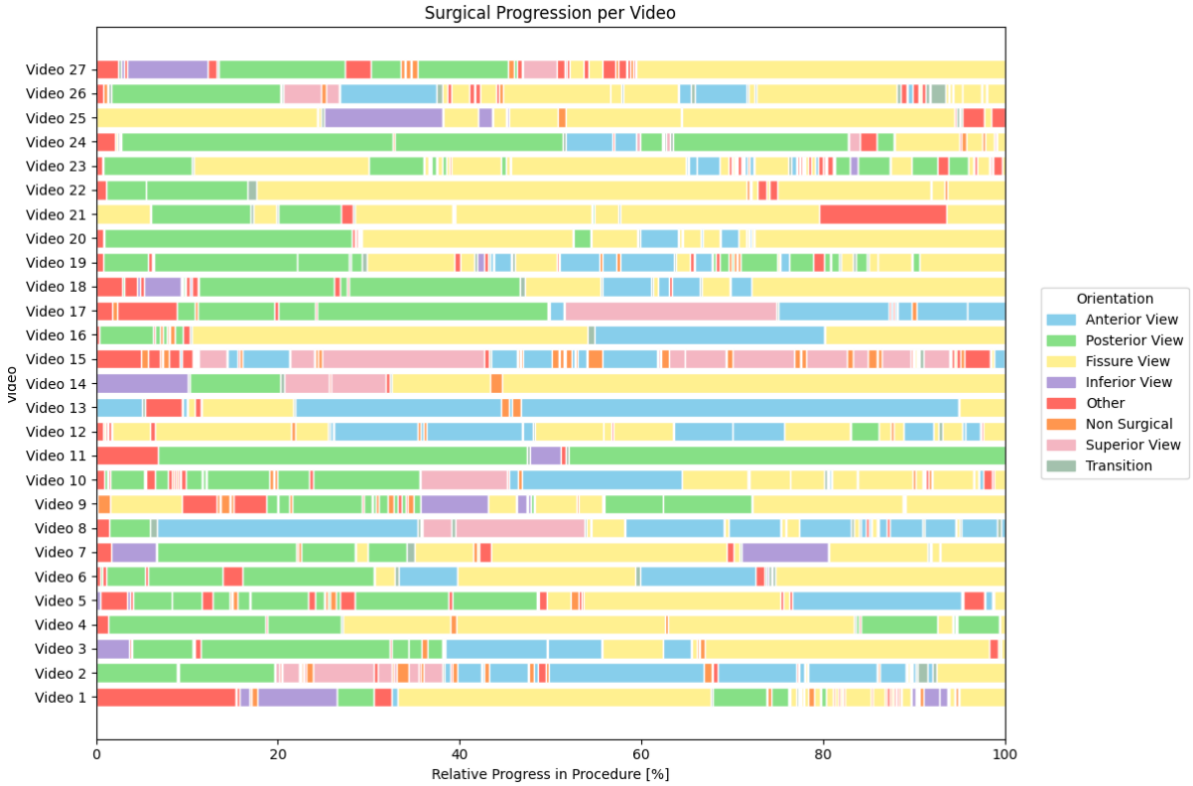


Figure 4. Annotated orientation per video plotted over the relative progress in the procedure in %

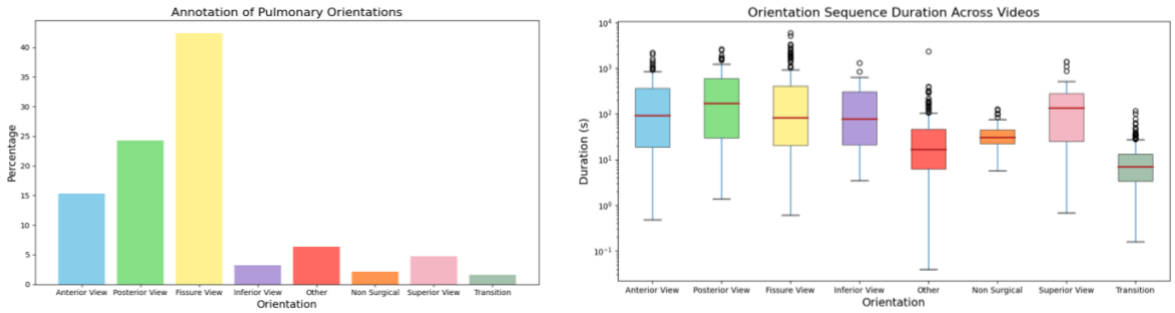


Figure 5. Orientation & Duration throughout the complete dataset a. Distribution of Orientation [%]. b. Distribution of orientation sequence duration across videos (in second) throughout the complete dataset. Duration (y-axis) is presented in log-scale

The variability in the lengths and occurrences of orientation are shown in Figure 5b. The 'Anterior view' orientation had the longest median duration in the dataset at 2.84 (0.02–43.27) minutes, followed by 'Superior view' at 2.27 (0.01–23.35) minutes, 'Posterior view' at 1.39 (0.01–36.54) minutes, 'Fissure view' at 1.29 (0.06–21.71) minutes, and 'Inferior view' at 0.28 (0.00–38.43) minutes. The median duration for the 'Other' label was 1.57 (0.01–36.54) minutes, while 'Non-Surgical' and 'Transition' had median durations below 1 minute, with 0.52 (0.10–2.12) minutes and 0.11 (0.00–1.95) minutes, respectively. The 3D model was used in 4.3% of all annotated frames, primarily (52.4%) during the 'Fissure view'. The duration of the 3D model usages was a minimum of 0.1 minutes and a maximum of 27 minutes (median of 0.13 minutes) (Appendix E).

Finally, we analyzed the orientation transitions in our dataset. Table 1. shows in what proportion a transition is observed from one orientation to another.

	Next Phase [%]							
Labels	*Anterior View*	*Posterior View*	*Fissure View*	*Inferior View*	*Superior View*	*Transition*	*Other*	*Non-Surgical*
Anterior View	-	0	0.85	2.56	0	**47.01**	41.03	8.55
Posterior View	0	-	0	0	0	34.88	**48.84**	16.28
Fissure View	0.47	0.47	-	1.42	0	**55.45**	31.75	10.43
Inferior View	0	0	0	-	0	43.33	**50.00**	6.67
Superior View	0	0	0	0	-	34.21	**47.37**	18.42
Transition	23.26	16.28	**42.86**	4.65	4.98	-	7.97	0
Other	8.09	21.32	19.49	3.31	4.41	21.32	-	**22.06**
Non-Surgical	18.03	17.21	23.77	3.28	9.02	0	**28.69**	-

Table 1. Phase transitions in our dataset. The numbers indicate the relative (in percent) transitions from one phase (row-axis) to another (column-axis).

III.III Orientation Recognition

The complete distribution of the dataset split, corresponding feature visualisations and distribution of cross-validation are provided in Appendix D and an elaborate description of the hyperparameter tuning and eventual selection is presented is presented in Appendix F.

Table 2. provides the performance of feature extraction and the three-stage temporal approach of TeCNO on our dataset. Accuracy, Precision, Recall and F1-score are provided for the final used dataset and averaged over all folds of cross validation. Additionally, accuracies for different subsets of our dataset were extracted, displaying the influence of certain characteristics of the dataset on model performance.

Final Dataset

	Accuracy	Precision	Recall	F1-Score
Without TCN (TeCNO)	65.73 ± 23.69	42.79 ± 14.69	52.47 ± 12.28	35.98 ± 16.31
TeCNO Stage I	66.71 ± 16.83	38.41 ± 38.62	47.48 ± 41.04	29.91 ± 12.95
TeCNO Stage II	37.25 ± 13.99	51.26 ± 45.35	19.81 ± 31.99	16.32 ± 08.25
TeCNO Stage III	35.91 ± 28.52	45.39 ± 30.61	15.94 ± 6.5	9.25 ± 6.46

a

Cross-Validation

	Accuracy	Precision	Recall	F1-Score
Without TCN (TeCNO)	63.22 ± 20.13	44.23 ± 13.59	59.36 ± 11.32	38.53 ± 13.76
TeCNO Stage I	67.38 ± 18.25	56.08 ±38.64	33.2 ± 39.84	26.76 ± 9.34
TeCNO Stage II	55.34 ± 29.24	45.65 ± 34.53	25.12 ± 41.16	17.57 ± 9.26
TeCNO Stage III	54.77 ± 28.18	47.67 ± 24.98	24.34 ± 7.16	9.26 ± 9

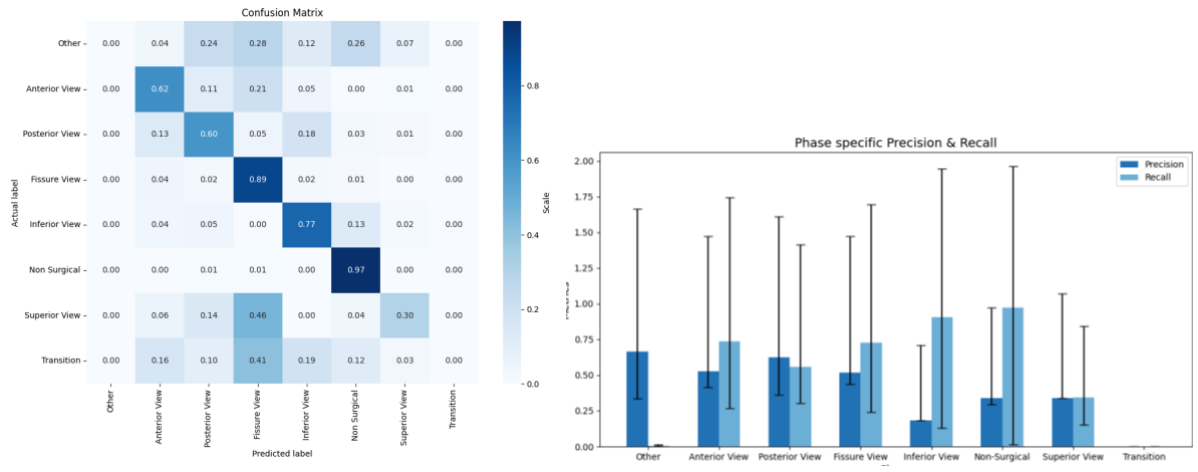
b

	Final Dataset		**Cross- Validation**	
	Accuracy CNN	Accuracy TCN	Accuracy CNN	Accuracy TCN
Left-Sided Resection	75.52 ± 14.01	79.08 ± 9.54	72.13 ± 15.87	65.45 ± 18.20
Rights-Sided Resections	52.71 ± 25.71	58.68 ± 23.77	57.71 ± 18.81	60.50 ± 23.99
Upper/Middle Lobe Resections	58.51 ± 30.31	63.01 ± 27.97	61.94 ± 25.61	50.34 ± 24.30
Lower Lobe Resections	64.02 ± 19.24	69.65 ± 16.20	64.29 ± 10.50	74.37 ± 5.33

c

Table 2. Model performance (mean ± std) using the evaluation metrics Accuracy, Recall, Precision and F1-score on our dataset. The std for accuracy is computed across all videos. The std for precision, recall and F1-score is computed across all orientation. a. Model performance of the final selected dataset. b. Model performance averaged over all 5 folds of cross-validations. c. Model performance for subsets of the dataset to interpreted performance for Left- vs. Right-sided lung restions and Upper/Middle Lobe vs. Lower Lobe Resections.

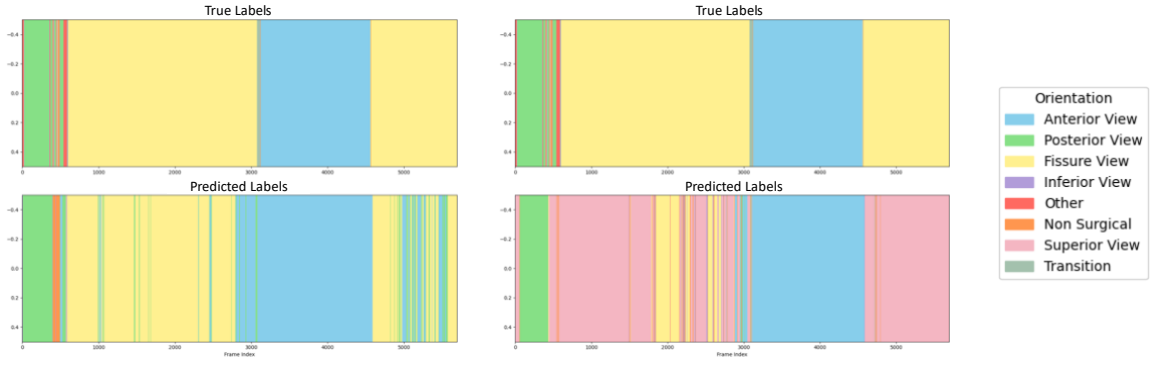
Figure 6. provides evaluation of the comparative analyses of the per-orientation performance reached by the two-stage TCN model, through a confusion matrix and variations in key performance metrics precision and recall. The diagonal line of the confusion matrix indicates TPR, ranging from 0.00 for the label 'Other' and 'Transition' to 0.97 for the 'Non-Surgical' category. The TPR for orientations ranges from 0.00 for the prediction of the 'Posterior view' to 0.89 for the 'Fissure view'.



a b

Figure 6. Comparative orientation phase-specific model performance analysis. a. Normalized confusion matrix (NCM), showing phase-specific accuracy. Rows in the NCM correspond to the annotated actual phase label (ground truth), whereas columns correspond to the predicted phase labels. The diagonal elements of the NCM present the proportion of correct predictions per phase. The color of the heat map indicates the proportion of frames that is allocated to each phase label, dark blue signifying a high proportion and light blue a low proportion. b. Orientation phase-specific Precision and Recall across our test set.

To visualise the predictive accuracy of our approach, we provide the models predictions for the videos that show best and worst performance, compared to the initial annotations that are considered ground truth (figure 7).



a b

Figure 7. Timeplots of ground truth (top) and predictions (bottom), illustrating qualitative results for phase recognition with the TeCNO model on our own dataset. a. Video 16 for which the model achieved the highest accuracy of 89.23%. b. Video for which the model achieved the lowest accuracy of 16.39%.

IV. Discussion

Our study introduced a proof-of-concept DL-based orientation recognition approach [22], to automatically detect the intraoperative orientation of the lung during left- and right-sided Robot-Assisted lobectomy and segmentectomy resections. These orientations are closely related to the surgical approach and the bronchovasculature encountered throughout different phases of the procedure. Automatic orientation recognition could aid in identifying anatomical structures, serve to train and educate, and provide insights into different surgical approaches to potentially improve procedural efficiency. Ultimately, this might enhance intraoperative imaging navigation by facilitating the automated orientation of 3D models.

We have developed a unique dataset for orientation recognition, the first comprising pulmonary procedures and intraoperative orientation. Our model achieved an overall accuracy of 70% using two-stage TCN, indicating a reasonable ability to classify orientations correctly. However, the low precision, recall and F1-score indicate considerable inconsistency in its performance. Additionally, this study shows a steep learning curve for annotation, suggesting that relatively inexperienced raters can be adequately trained to perform orientation annotations.

The TeCNO model [22] was initially developed for surgical phase recognition of laparoscopic cholecystectomies, using the Cholec80 [29] dataset (80 laparoscopic cholecystectomy videos). This is the most frequently referenced resource for surgical phase recognition, showing an accuracy of 88.56% for two TCN stages [22]. The inferior performance of our dataset compared to Cholec80 was expected given the dissimilarity between intraoperative phases and orientation, the completely different procedure compared to a Cholecystectomy procedure and our small dataset.

Surgical phase recognition models are trained to recognize the consecutive surgical steps of a procedure, such as preparation, dissection and cutting [16]. Given the predominantly linear workflow, cholecystectomy procedures exhibit highly deterministic phase transitions, allowing models to effectively utilize the temporal information between sequences [18]. Our dataset displays highly heterogeneous orientation transitions, due to orientations being present across various surgical phases (figure 4). Most lung orientations are succeeded by a ‘Transition’ or ‘Other’

sequence, but no clear pattern can be distinguished for the orientation thereafter. Thus, the model can exploit less distinctive temporal information from our dataset, contributing to lower model performance. 'Transition' label can provide information on the ease of transitioning between orientations, potentially indicating the surgeon's expertise and the labeler-specific decision making regarding the initiating and conclusion of orientations [17]. However, the model has the tendency to predict 'Transition' sequences as 'Other', suggesting that combining these labels might enhance the predictive power.

Pulmonary procedures are complex due to the granular subdivision of pulmonary anatomy [30]. Differences in right and left pulmonary anatomy and variation in surgical approach, e.g., anterior to posterior, fissure-first or -last, can add to the complexity of recognition [31]. Most current research on surgical phase recognition is primarily focused on more standardized general surgery procedures, such as laparoscopic cholecystectomy [17, 32]. A similar procedure is a thoracic robot-assisted minimally invasive esophagectomy procedure, for which recent research has achieved an accuracy of 84% using the TeCNO model, suggesting opportunity for performance improvement using our dataset [32].

The varying results for subsets of our dataset imply high influence of the composition of the dataset on model performance. For example, feature extraction performs better for left-sided resections, possibly aided by prominent anatomical landmarks such as the thoracic aorta and pericardial tissue, while the temporal model excels slightly for right-lung resection. Interestingly, lower lobe resections show notably higher performance than upper/middle lobe resections, which potentially correlates to the higher risk of intraoperative technical challenges and inconsistency of subsequent surgical steps [31, 33]. Furthermore, temporal model performance excels for lower lobe resections. These resections are more anatomically straightforward and therefore may entail a more standardized surgical approach [31].

Furthermore, discrepancy in performance is observed between different orientations. The model shows high performance in detection of the non-surgical category, the 'Fissure' and the 'Inferior' view. More difficulty is observed in predicting the 'Posterior' and the 'Anterior' view. This is potentially related to our mixed dataset of right and left lung resection. Interestingly, this correlates to our analyses of two videos with persistent low accuracies (video 8 and 11), that show an overrepresentation of certain orientations, 'Anterior view' and 'Posterior view', and limited presence of the 'Fissure view'. The categories 'Superior view' and 'Transition' are rarely classified correctly. Given the limited size of our dataset, imbalances in frequencies of labels potentially hinders the learning capability of the model [17].

While the potential application of this orientation recognition approach seems promising, this study was subject to several limitations. The dataset utilized in this research is sourced from a single medical center, raising potential challenges in generalizing our findings. Our model is trained on a very small and heterogeneous dataset that presents a large variety in performance based on the specific dataset split. Bar et al. [34], show that increasing the amount of data highly influences the performance of recognition models. Thus, to investigate the full potential of orientation recognition, a larger dataset, spanning multiple institutions is needed to ensure adequate variability for training.

Despite clear annotation guidelines, annotation is time-consuming, and dependent on the annotator's expertise [17]. Involving expert annotators is necessary to enhance the quality of the dataset and approaches to reduce the burden of annotation can be considered, such as un-supervised [35], self-supervised learning [36, 37] or federated learning [38], soft tissue tracking [35] and more extensive data augmentation [37].

The considerable accuracy standard deviation suggests a big performance gap between videos. Such high variability of accuracy across videos might suggest overfitting [17]. Especially the performance of the temporal model is highly influenced by the composition of the dataset and may even decrease the models predictive power opposed to results from feature extraction. Our study uses a TCN as temporal model, however, while a particular model may be effective for one purpose or type of procedure, it may not be suitable for another [16, 17]. Other temporal methods that may be employed are e.g., dynamic time warping [39, 40], hidden Markov models [41], Long Short-Term Memories (LSTM) networks [42, 43] or transformers [25].

Furthermore, our method consists of an independent spatial and temporal stage. Future work could explore integrated models to preserve temporal information during feature extraction and enhance context aggregation [44]. Performance might also benefit from further hyperparameter optimization [37]. Latest innovations including dynamically adaptable weights to handle class imbalance, a Moment Loss function that penalizes undesirable transitions to prevent overfitting and a dual dilated layer that combines different receptive fields to improve model performance could be investigated to improve model performance [28, 45, 46]. Finally, future work could study the influence of 3D model presence, visibility of the instrument port or presence of a significant amount of blood or smoke on predictive performance.

Intraoperative orientation recognition can have diverse applications. It can offer a valuable tool for training and education in understanding the diverse surgical approaches, aid surgeons in refining surgical techniques and enhance intraoperative decision making and procedural efficiency. Additionally, it may support students and young residents to recognize the correlation between lung orientation, surgical steps and essential anatomical landmarks. Hence, it may contribute to automatic intraoperative anatomy recognition [15]. A persistent challenge in this area remains detection restriction to exposed anatomical structures, lacking information on the underlying anatomy. Leveraging knowledge of intraoperative orientation can inform on the probability of the presence of concealed structures being present. Similarly, it can complement the recognition of surgical steps. These advancements can be implemented concurrently, mutually reinforcing one another by providing supplementary information for decision-making.

Ultimately, orientation detection could enhance intraoperative imaging navigation by enabling automatic adjustment of intraoperative 3D models. This can be specifically relevant in the ‘Fissure view’, where our data shows frequent use of 3D model assistance. Interestingly, we observed highest occurrence and model performance in the ‘Fissure view’ as well. Automatic ‘Fissure view’ orientation of a dynamic 3D model like PulmoSR (MedicalVR, Amsterdam, The Netherlands) [12] is specifically relevant, as interlobar 3D simulation involves substantial adjustments, beyond simple translation and rotation. Automation could eliminate the need for manual input by surgeons or remote assistants

and improve intraoperative workflow. Detailed steps to enable automated 3D model orientation are outlined in Appendix G. Finally, automatic orientation detection could contribute to the growing research topic of Augmented Reality registration [47] in surgery, facilitating initial 3D model positioning to simplify subsequent registration steps.

V. Conclusion

In conclusion, our study introduced a proof-of-concept orientation recognition approach to automatically identify intraoperative lung orientation during RATS lobectomy/segmentectomy procedures. We developed and utilized a unique dataset with intraoperative orientations of the lung and evaluated the performance of an existing surgical phase recognition model, TeCNO, in this context. Despite the challenges of adapting this model to our distinctive dataset, which is small and features less predictable orientation transitions, results show potential to enhance intraoperative guidance and 3D model alignment, particularly in the most often used interlobar 'Fissure view'. Further developments are necessary to improve performance of orientation recognition and prospective studies should be performed to assess the clinical implementation.

VI. References

1. Organization, W.H. *Lung Cancer*. 2023 [cited 2023 19-07-2023]; Available from: https://www.who.int/news-room/fact-sheets/detail/lung-cancer#:~:text=Lung%20cancer%20is%20the%20leading,when%20treatment%20options%20are%20limited.
2. Catelli, C., et al., *RoboticAssisted (RATS) versus Video-Assisted (VATS) lobectomy: A monocentric prospective randomized trial.* Eur J Surg Oncol, 2023. **49**(12): p. 107256.
3. Aiolfi, A., et al., *Pulmonary lobectomy for cancer: Systematic review and network meta-analysis comparing open, video-assisted thoracic surgery, and robotic approach.* Surgery, 2021. **169**(2): p. 436-446.
4. Zhang, L. and S. Gao, *Robot-assisted thoracic surgery versus open thoracic surgery for lung cancer: a system review and meta-analysis.* Int J Clin Exp Med, 2015. **8**(10): p. 17804-10.
5. Wilson-Smith, A.R., et al., *The learning curve of the robotic-assisted lobectomy-a systematic review and meta-analysis.* Ann Cardiothorac Surg, 2023. **12**(1): p. 1-8.
6. Erwin, P.A., et al., *Consensus for Thoracoscopic Lower Lobectomy: Essential Components and Targets for Simulation.* The Annals of Thoracic Surgery, 2022. **114**(5): p. 1895-1901.
7. Kim, D., et al., *The Uncomfortable Truth: Open Thoracotomy versus Minimally Invasive Surgery in Lung Cancer: A Systematic Review and Meta-Analysis.* Cancers, 2023. **15**(9): p. 2630.
8. Bakhuis, W., et al., *Essential Surgical Plan Modifications After Virtual Reality Planning in 50 Consecutive Segmentectomies.* Annals of Thoracic Surgery, 2022. **115**(5): p. 1247-1255.
9. Sadeghi, A.H., et al., *Virtual reality and artificial intelligence for 3-dimensional planning of lung segmentectomies.* JTCVS Techniques, 2021. **7**: p. 309-321.
10. Li, C., et al., *Augmented Reality and 3-Dimensional Printing Technologies for Guiding Complex Thoracoscopic Surgery.* The Annals of thoracic surgery, 2021. **112**(5): p. 1624-1631.
11. Vervoorn, M.T., et al., *Application of three-dimensional computed tomography imaging and reconstructive techniques in lung surgery: A mini-review.* Frontiers in Surgery, 2022. **9**.
12. Bakhuis, W., et al., *Video-assisted thoracic surgery S7 segmentectomy: use of virtual reality surgical planning and simulated reality intraoperative modelling.* Multimedia manual of cardiothoracic surgery : MMCTS, 2023. **2023**.
13. Tokuno, J., et al., *Resection Process Map: A novel dynamic simulation system for pulmonary resection.* The Journal of Thoracic and Cardiovascular Surgery, 2020. **159**(3): p. 1130-1138.
14. Bakhuis, W., et al., *Video-assisted thoracic surgery S7 segmentectomy: use of virtual reality surgical planning and simulated reality intraoperative modelling.* Multimed Man Cardiothorac Surg, 2023. **2023**.
15. den Boer, R.B., et al., *Computer-aided anatomy recognition in intrathoracic and -abdominal surgery: a systematic review.* Surgical Endoscopy, 2022. **36**(12): p. 8737-8752.
16. Garrow, C.R., et al., *Machine Learning for Surgical Phase Recognition: A Systematic Review.* Ann Surg, 2021. **273**(4): p. 684-693.
17. Demir, K.C., et al., *Deep Learning in Surgical Workflow Analysis: A Review of Phase and Step Recognition.* IEEE J Biomed Health Inform, 2023. **27**(11): p. 5405-5417.
18. Kirtac, K., et al., *Surgical Phase Recognition: From Public Datasets to Real-World Data.* Applied Sciences, 2022. **12**(17): p. 8746.
19. Jin, Y., et al., *SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network.* IEEE Trans Med Imaging, 2018. **37**(5): p. 1114-1126.
20. Wagner, M., et al., *Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark.* Medical Image Analysis, 2023. **86**: p. 102770.
21. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochem Med (Zagreb), 2012. **22**(3): p. 276-82.

22. Czempiel, T., et al. *TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks*. 2020. Cham: Springer International Publishing.
23. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
24. Golany, T., et al., *Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy.* Surgical Endoscopy, 2022. **36**(12): p. 9215-9223.
25. Gao, X., et al. *Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer*. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. 2021. Cham: Springer International Publishing.
26. Funke, I., D. Rivoir, and S. Speidel, *Metrics Matter in Surgical Phase Recognition.* arXiv preprint arXiv:2305.13961, 2023.
27. al., W.F.e. *PyTorch Lightning Documentation*. 2023 12-01-2024]; Available from: https://pytorch-lightning.readthedocs.io/en/1.0.8/.
28. Li, S., et al., *MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation.* IEEE Trans Pattern Anal Mach Intell, 2023. **45**(6): p. 6647-6658.
29. Twinanda, A.P., et al., *EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos.* IEEE Transactions on Medical Imaging, 2017. **36**(1): p. 86-97.
30. Rea, G. and M. Rudrappa, *Lobectomy*, in *StatPearls*. 2023, StatPearls Publishing
Copyright © 2023, StatPearls Publishing LLC.: Treasure Island (FL).
31. DeArmond, D.T., et al., *Lung lobectomy surgical approach and resource utilization differ by anatomic lobe in a statewide discharge registry.* J Thorac Dis, 2022. **14**(8): p. 2791-2801.
32. Takeuchi, M., et al., *Automated Surgical-Phase Recognition for Robot-Assisted Minimally Invasive Esophagectomy Using Artificial Intelligence.* Ann Surg Oncol, 2022. **29**(11): p. 6847-6855.
33. Bryan, D.S., et al., *Consensus for Thoracoscopic Left Upper Lobectomy—Essential Components and Targets for Simulation.* The Annals of Thoracic Surgery, 2021. **112**(2): p. 436-442.
34. Bar, O., et al., *Impact of data on generalization of AI for surgical intelligence applications.* Sci Rep, 2020. **10**(1): p. 22208.
35. Cartucho, J., et al., *SurgT challenge: Benchmark of soft-tissue trackers for robotic surgery.* Med Image Anal, 2024. **91**: p. 102985.
36. Paysan, D., et al., *Self-supervised representation learning for surgical activity recognition.* Int J Comput Assist Radiol Surg, 2021. **16**(11): p. 2037-2044.
37. Ramesh, S., et al., *Dissecting self-supervised learning methods for surgical computer vision.* Med Image Anal, 2023. **88**: p. 102844.
38. Kassem, H., et al., *Federated Cycling (FedCy): Semi-Supervised Federated Learning of Surgical Phases.* IEEE Trans Med Imaging, 2023. **42**(7): p. 1920-1931.
39. Ahmadi, S.A., et al., *Recovery of surgical workflow without explicit models.* Med Image Comput Comput Assist Interv, 2006. **9**(Pt 1): p. 420-8.
40. Padoy, N., et al., *Statistical modeling and recognition of surgical workflow.* Medical Image Analysis, 2012. **16**(3): p. 632-641.
41. Blum, T., et al., *Modeling and online recognition of surgical phases using Hidden Markov Models.* Med Image Comput Comput Assist Interv, 2008. **11**(Pt 2): p. 627-35.
42. Jin, Y., et al., *Temporal Memory Relation Network for Workflow Recognition From Surgical Video.* IEEE Trans Med Imaging, 2021. **40**(7): p. 1911-1923.
43. Ban, Y., et al., *Aggregating Long-Term Context for Learning Laparoscopic and Robot-Assisted Surgical Workflows.* 2021 IEEE International Conference on Robotics and Automation (ICRA), 2020: p. 14531-14538.
44. Chen, Z., et al., *Surgical Temporal Action-aware Network with Sequence Regularization for Phase Recognition.* ArXiv, 2023. **abs/2311.12603**.

45. Park, M., et al., *Multi-Stage Temporal Convolutional Network with Moment Loss and Positional Encoding for Surgical Phase Recognition.* Diagnostics, 2023. **13**(1): p. 107.
46. Fernando, K.R.M. and C.P. Tsokos, *Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks.* IEEE Transactions on Neural Networks and Learning Systems, 2022. **33**(7): p. 2940-2951.
47. Sadeghi, A.H., et al., *Current and Future Applications of Virtual, Augmented, and Mixed Reality in Cardiothoracic Surgery.* Ann Thorac Surg, 2022. **113**(2): p. 681-691.
48. Mintz, Y. and R. Brodie, *Introduction to artificial intelligence in medicine.* https://doi.org/10.1080/13645706.2019.1575882, 2019. **28**(2): p. 73-81.
49. Bini, S.A., *Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?* The Journal of Arthroplasty, 2018. **33**(8): p. 2358-2361.
50. Amisha, et al., *Overview of artificial intelligence in medicine.* Journal of Family Medicine and Primary Care, 2019. **8**(7): p. 2328-2328.
51. Uddin, S., et al., *Comparing different supervised machine learning algorithms for disease prediction.* BMC Medical Informatics and Decision Making, 2019. **19**(1): p. 1-16.
52. Kawka, M., et al., *Intraoperative video analysis and machine learning models will change the future of surgical training.* Intelligent Surgery, 2022. **1**: p. 13-15.
53. Project, F. *FFmpeg Documentation*. 2023; Available from: https://ffmpeg.org/ffmpeg.html.
54. Kipp, M., *420ANVIL: The Video Annotation Research Tool*, in *The Oxford Handbook of Corpus Phonology*, J. Durand, U. Gut, and G. Kristoffersen, Editors. 2014, Oxford University Press. p. 0.
55. AI, S. *Annotation*. 2023; Available from: https://doc.superannotate.com/docs/pixel-annotation.

VII. Supplementary Files

Appendix A. Technical Details

Short Technical Introduction

AI is a broad-based field of computer science that provides the ability to imitate intelligent human behaviour [16, 48]. It can handle and optimize highly complex systems consisting of very complex data sets, through the application of algorithms [49]. Within AI, Machine Learning (ML) algorithms are capable of automatically learning from experience and therefore, modify and improve processing upon newly acquired information. Given enough data, a ML algorithm is able to extract complex patterns that are invisible to humans and accurately classify unseen data [50-52]. A subset within ML is Deep Learning (DL), which can imitate human brain processing through a Convolutional Neural Network (CNN), a mathematical model that is inspired by the neural networks of the human brain, considering multiple datasets simultaneously throughout different layers [48, 50, 51]. In the review of Garrow et al., [16] it became apparent that most ML algorithms used for surgical phase recognition are based on supervised learning [17]. A supervised learning algorithm is a prediction model for unlabelled data developed through analysis of a labelled dataset [51].

Pre-processing

To address variations in video frame dimensions and aspect ratios of the video in our dataset, a preprocessing methodology utilizing the FFmpeg multimedia processing tool [53] was used, considering both spatial and temporal transformation. Spatial transformation involved a series of operation including cropping, padding, and scaling. A spatial cropping operation was employed to standardize the spatial dimensions of the video frames. The frames were cropped to a resolution of 1125x900 pixels, ensuring consistency in the region of interest (ROI). The choice of 900 pixels as the height was determined by selecting the shortest pixel height encountered in the dataset. To maintain the aspect ratio observed in the 1280x1024 pixel video frames (0.8 ratio), the width was adjusted accordingly. For frames originally sized at 1440x900 pixels, the presence of black edges allowed for the safe cropping of the width to 1125 pixels without resulting in information loss.Subsequently, a symmetric padding operation was executed to preserve the original aspect ratio. This step ensures that subsequent analyses are not biased by variations in the original frame dimensions. Additionally, scaling was performed, reducing the frames to a width of 375x300 pixels. This adjustment proved essential for efficient data handling during the process of splitting videos in individual frames, minimizing storage demands that could potentially impede or slow down the overall process.

A temporal transformation was applied by adjusting the frame rate to 25 fps. This modification was crucial for both the annotation process, automatically facilitated in Anvil [54], and the subsequent splitting of video frames, optimizing the computational efficiency of these processes.Videos were shortened by defining a new start and end time. The start times were manually selected after the insertion of all instruments upon the first movement. For lobectomy procedures the end time of the video was determined upon the complete detachment of the to be resected lung lobe from the remaining part of the lung. Regarding segmentectomy procedures the end time of the video was determined by the initiation of the lung parenchyma stapling, given the anatomically disrupted view that follows this resection.

Appendix B. Orientation Definition and Annotation Rules

Orientation Definition

Multiple methods were scrutinized to get familiarized with the surgical steps of a lobectomy and segmentectomy RATS procedure. Initially, various surgeries were observed to understand lung orientation and identify key structures like pulmonary veins, arteries, and bronchi. Theoretical research on lung anatomy, particularly bronchovasculature, was conducted online. Additionally, four videos of surgeries were thoroughly analyzed, focusing on lung orientation, surgical steps, and corresponding time stamps. This highlighted the importance of pulmonary orientation during lobectomy and segmentectomy surgery.

Annotation Rules

In the following, additional details on the annotation protocol for Orientation, 3D Model and Port Visibility Annotation as described in the 'Methods' section of the main article are presented.

Orientation Annotation

The orientation annotation encompasses eight labels, including five orientations of the lung: 'Anterior View' (P1), 'Posterior View' (P2), 'Fissure View' (P3), 'Inferior View' (P4), 'Superior View' (P5). Additionally, three labels – 'Transition' (P6), 'Other' (P7) and 'Non-Surgical' (P8), are included to categorize all video frames/sequences not fitting the defined orientations. The orientations do not necessarily occur in a fixed order. Identified orientations (P1 – P5) are typically preceded or followed by a 'Transition' or 'Other' sequence. The 'Transition' label includes sequences where the lung parenchyma is being reoriented, beginning with the movement of tissue by surgical tools and ending when a new orientation is clearly established. 'Other' covers sequences not identifiable as specific orientations or clear transitions, including frames obscured by blood, smoke, or with the lung parenchyma completely out of view.

3D Model Annotation

The 3D model, displayed beneath intraoperative video footage using the Tilepro extension of the Da Vinci robot, is annotated irrespective of the orientation. The Tilepro functionality is manually activated by the operating surgeon and can be active across all types of orientations. The annotation starts and ends with frames showing any activation of the Tilepro functionality, even if the 3D model itself is not immediately visible. It is not a default label for every video frame; thus, frames without port-visibility are marked as 0, while those with a visible instrument port are marked as 1.

Port-Visibility Annotation

This annotation applies to all frames where the instrument port is even marginally visible, during one of the five orientations (P1 – P5) or within the 'Other' label. The 'Port-visibility' label is always annotated concurrently with another orientation and is represented by a distinct label. Again, this is not a default label for every video frame, assigned with 0 or 1. Port visibility and the 'Non-Surgical' (P6) label are mutually exclusive. The transition between 'Port-Visibility' and 'Non-Surgical' is defined by the frame in which the instrument port becomes more visible compared to preceding frames.

Annotation Platform

Several annotation methods were considered and explored. Initially three videos of lobectomy procedures of the lower left lobe were analyzed through the SuperAnnotate software [55] provided by Orsi Academy (Ghent, Belgium). An example of the annotation platform is visualized in figure 8. This software allows for guided annotation through the magic polygon option. With more easily identifiable structures, placing a considered number of points towards the edges of the target structure provides the software with enough information to identify the complete structure. Subsequently, each polygon can manually be adjusted. For these three videos annotation of all instruments was performed, to explore the potential added value of instrument detection for phase, orientation or anatomy recognition. Furthermore, three major anatomical structures were annotated: the pericardial tissue (heart), the lung parenchyma and the aorta. These structures were selected since they can be identified relatively easily and appear in most of the video frames. Annotation of these structures was performed to familiarize myself with the pulmonary anatomy and assess the possibility of structure annotation for potential anatomic structure detection. Additionally, these three videos were thoroughly analyzed regarding pulmonary orientation and surgical steps, to better understand the lobectomy procedure and anatomy of the left lung. In general, The SuperAnnotate software provides an intuitive platform for anatomical structure annotation, however, remains labor intensive due to the limited quality of automation and lack of in between frame interpolation. In addition, this platform was not open source, disqualifying this platform for further use.

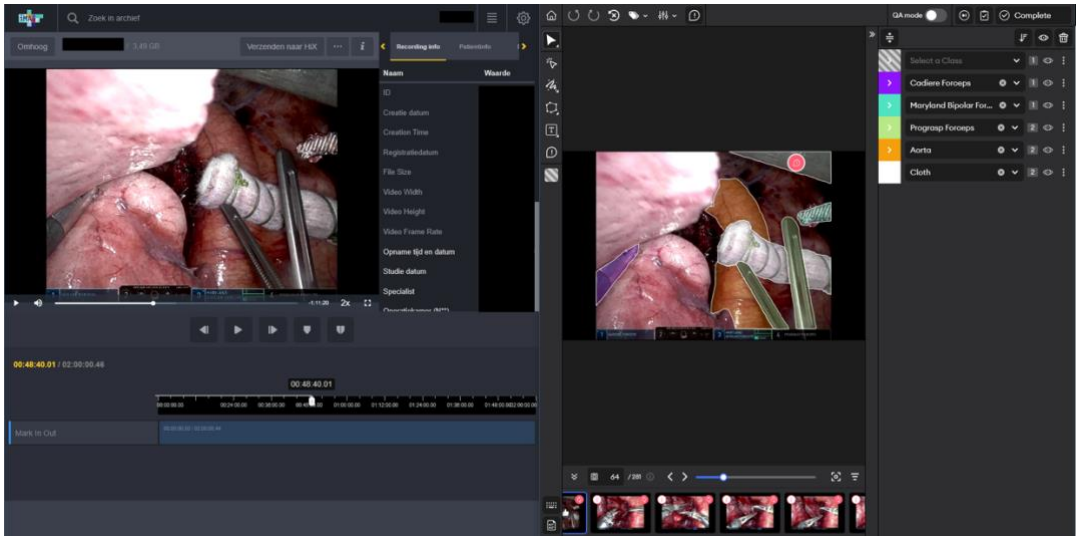
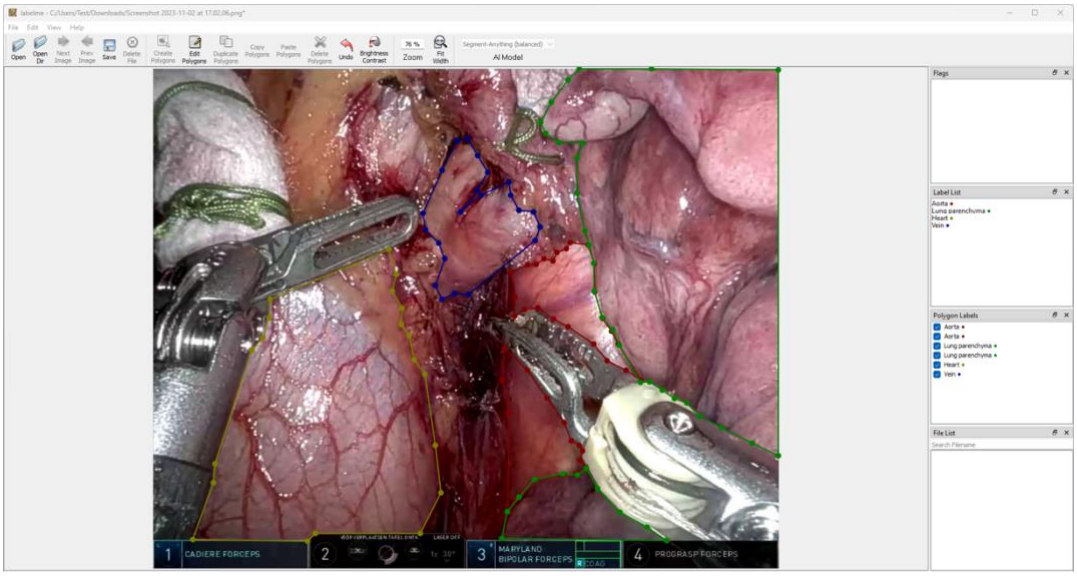


Figure 8. Left: IMAS video platform to access and visualize intraoperative videos obtained in the Erasmus Medical Center. Right: SuperAnnotate annotation platform interface

Consequently, various other open-source annotation platforms were considered, including LabelMe [8], CVAT [9] and ANVIL (Annotation of Video and Language Data) [10]. Both LabelMe and CVAT provide user-friendly interfaces for the annotation of anatomical structures, with the option for a local download to enable data privacy. An example of both annotation platform interfaces is provided in figure 9. In LabelMe, users have the ability to draw polygons around objects within images, facilitating the creation of labeled datasets for object detection and segmentation tasks. However, LabelMe does not support frame-by-frame labeling, a feature vital for tasks such as orientation labeling. CVAT is designed for annotating both images and videos for computer vision

projects, supporting the annotation on video sequences. This platform provides enhanced features like shape interpolation between video frames and automatic annotation using deep learning models. Despite these capabilities, the local installation of CVAT lacked some of these functions, resulting in a more labor-intensive process for frame-to-frame labeling.

a.



b.

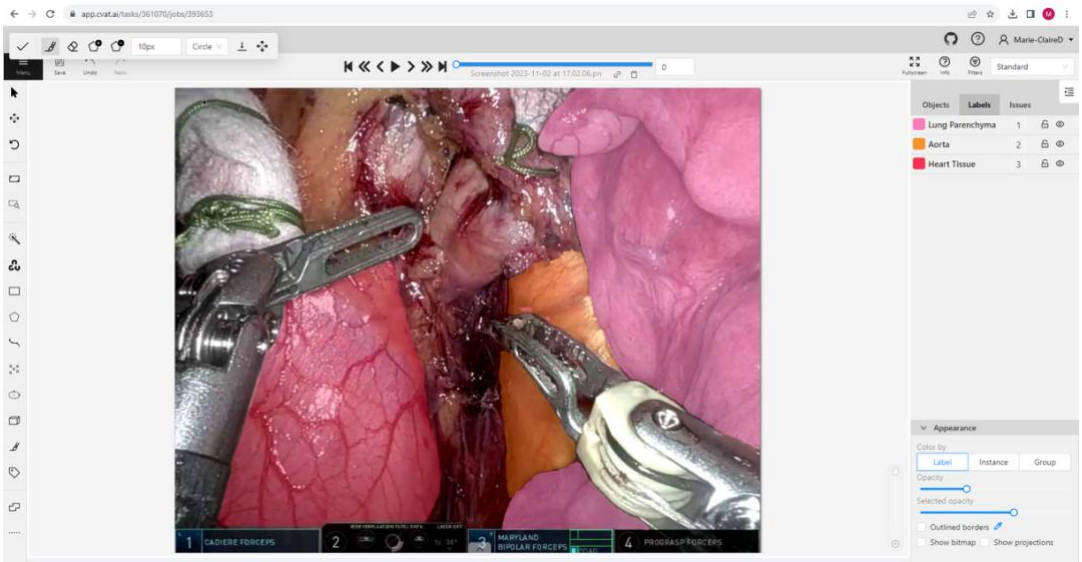


Figure 9. Exploration of annotation platforms. a. LabelMe annotation platform interface. b. CVAT annotation platform interface (local version).

ANVIL annotation platform

ANVIL is an annotation platform specifically designed for the analysis of multimedia content. Due to the flexibility of the feature set it can be adapted for a wide range of annotation tasks, including sequence annotation of surgical videos since it allows users to mark and label segments over time. In addition, ANVIL is highly customizable, enabling users to define their own coding schemes. Figure 10 provides an example overview of ANVIL's annotation platform interface, consisting of four separate windows. A project specific coding scheme was prepared for the purpose of orientation detection, as presented in figure 11.

Installation of ANVIL

To download ANVIL, the right version of Java and the ANVIL software are required.

- Download Java: Index of java-local/jdk/9.0.1+11 (huaweicloud.com)
- Download Anvil: http://www.anvil-software.de/download/index.html

For further instruction, please be referred to ANVIL's own website: http://www.anvil-software.de/download/index.html#

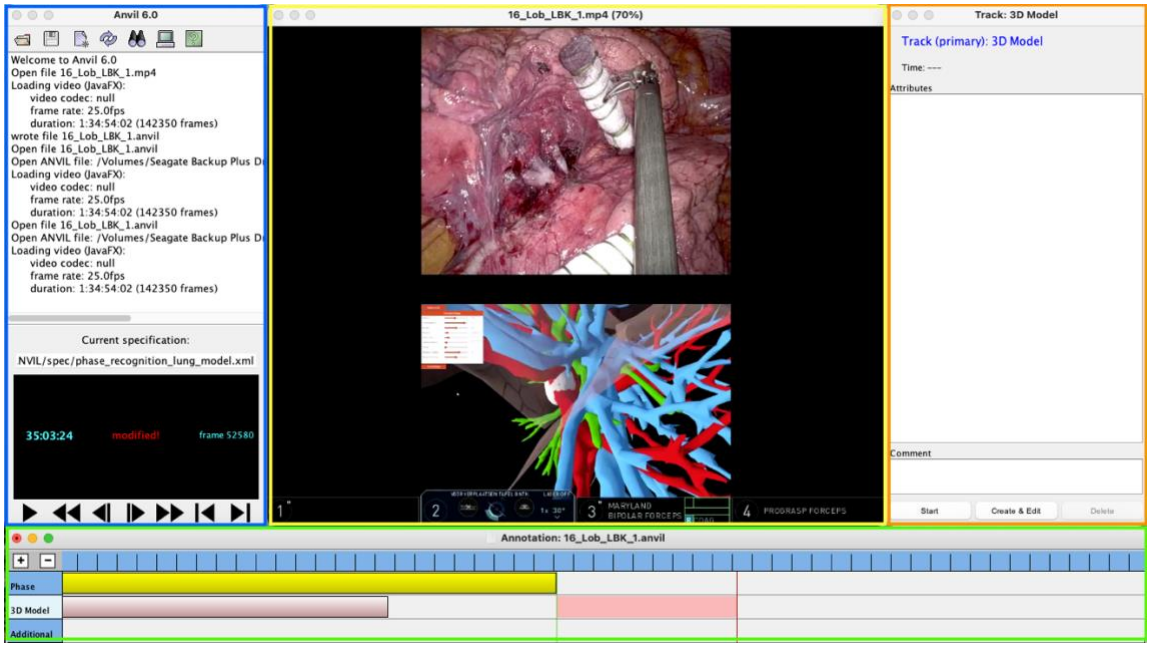


Figure 10. ANVIL annotation platform interface. Blue: Main window to open and save files, display video Information including frame rate, video duration, current frame and timestamp & navigate through the video. Yellow: Video footage. Orange: Tracking window to assign attributes/comments to a selected video sequence. Green: Video timeline with specified annotation categories, to assign color-coded elements on multiple tracks in time-alignment.

Open documents

- To open a document, click the 'folder' icon in the upper left corner (blue section figure 10) or select 'file' from the menu bar.
- When opening a new video, choose the appropriate specification file.
- For existing ANVIL files, the specification file loads automatically. Ensure the video and ANVIL file are in the same folder, or manually select the video file if they are located separately.

Labelling

1. Navigate through the video using either the cursor on the video timeline (green section figure 10) or the navigation arrows (blue section figure 10).
2. Initiate a video sequence by clicking 'Start' in the tracking section (orange section figure 10), which will show a green line marking the starting point.
3. Progress the video by clicking 'Play' (▷) or manually moving the red line on the timeline.
4. The selected video sequence is indicated in red, between the green and red line.
5. To label a section, click 'Create & Edit' in the tracking section (orange section figure 10) or right-click on the selection and choose 'Create & Edit' (figure 12a).
6. Select a label from the predefined options and add comments if necessary (figure 12b).

7. For adjustments to existing labels, use the ‘edit’ button (figure 12a).
8. Ensure continuity in labeling, particularly for orientation tracking elements, by placing labels sequentially without gaps.

Export annotations

- Once all necessary frames are labeled, export the annotations on a frame-by-frame basis.
- Go to ‘file’ in the menu, click ‘export’, then choose ‘Annotation Frame-By-Frame’.
- In the export menu (figure 12c), make sure ‘Exclude end frame’ is unchecked.
- Click ‘OK’ to save the annotations and a summary of the labels used (figure 13).

Figure 11. Specification file for Anvil annotation

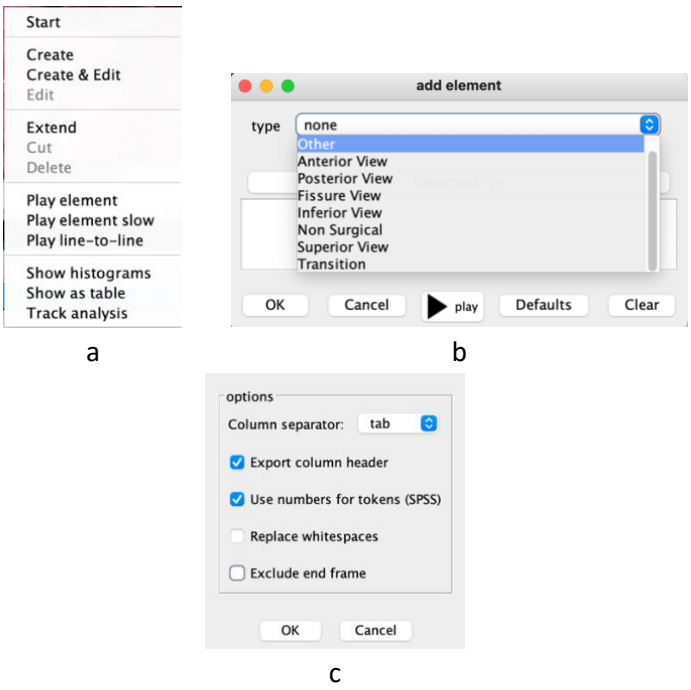


Figure 12. a. Selection menu right-mouse click. b. Selection menu to add element/label to the selected section. c. Export menu

Frame	Time	Phase	Phase:type	3D Model		3D Model:type	Additional	Additional:type
0	0.0	1	1	-1000	-1000	1	1	
1	0.04	1	1	-1000	-1000	1	1	
2	0.08	1	1	-1000	-1000	1	1	
3	0.12000000000000001	1	1	-1000	-1000	1	1	
4	0.16	1	1	-1000	-1000	1	1	
5	0.2	1	1	-1000	-1000	1	1	
6	0.24000000000000002	1	1	-1000	-1000	1	1	
7	0.28	1	1	-1000	-1000	1	1	
8	0.32	1	1	-1000	-1000	1	1	
9	0.36000000000000004	1	1	-1000	-1000	1	1	
10	0.4	1	1	-1000	-1000	1	1	
11	0.44	1	1	-1000	-1000	1	1	
12	0.48000000000000004	1	1	-1000	-1000	1	1	
13	0.52	1	1	-1000	-1000	1	1	
14	0.56	1	1	-1000	-1000	1	1	
15	0.6000000000000001	1	1	-1000	-1000	1	1	

Figure 13. Left: Example of frame-by-frame annotation file. Right: Example of annotation labels file.

Appendix C. Revision Results

	Video	Total # of transitions	Correct # of Transitions	Transition Score	Cohen's Kappa	Cohen's Kappa Interpretation
1st revision						
	05	56	35	62.50%	0.96	Almost Perfect
	13	14	11	78.57%	0.97	Almost Perfect
2nd revision						
	04	25	18	72.00%	0.98	Almost Perfect
	08	48	41	85.42%	0.83	Strong
3rd revision						
	16	19	18	94.74%	1	Almost Perfect

Figure 14. Results inter-rater agreement revision sessions 1, 2 and 3, including inter-rater transition agreement scores, Cohen's Kappa scores and Cohen's Kappa score interpretation [21].

	Video	Cohen's Kappa	Cohen's Kappa Interpretation
1st revision			
	02	0.75	Moderate
2nd revision			
	20	0.96	Almost Perfect
3rd revision			
	25	0.95	Almost Perfect

Figure 15. Results intra-rater agreement revision sessions 1, 2 and 3, including Cohen's Kappa scores and Cohen's Kappa score interpretation [11].

Appendix D. Dataset Distribution, Dataset Split & Feature Maps

Video ID	Lung Lobe	Type of Resection
1	RLL	Lobectomy
2	RUL	Lobectomy
3	LLL	Segmentectomy
4	RLL	Segmentectomy
5	RUL	Segmentectomy
6	LUL	Lobectomy
7	RLL	Lobectomy
8	RUL	Lobectomy
9	LLL	Lobectomy
10	RUL	Lobectomy
11	RLL	Lobectomy
12	LUL	Lobectomy
13	LUL	Segmentectomy
14	RLL	Lobectomy
15	RUL	Segmentectomy
16	LUL	Lobectomy
17	RUL	Lobectomy
18	LLL	Lobectomy
19	LLL	Lobectomy
20	LLL	Lobectomy
21	RLL	Lobectomy
22	RLL	Segmentectomy
23	RLL	Lobectomy
24	LUL	Segmentectomy
25	RLL	Segmentectomy
26	RML	Lobectomy
27	RLL	Lobectomy

	Lobectomy	Segmentectomy	Total
Left	7	3	10
Left Upper Lobe (LUL)	3	3	6
Left Lower Lobe (LLL)	4	0	4
Right	12	5	17
Right Upper Lobe (RUL)	4	2	6
Right Middle Lobe (RML)	1	0	1
Right Lower Lobe (RLL)	7	3	10

Figure 16. Distribution of the type of lung resections across our dataset

Fold 1	
Training set	3, 6, 9, 18, 12, 1, 2, 4, 5, 7, 23, 10, 25, 26
Validation set	24, 20, 15, 27, 21

Fold 2	
Training set	15, 12, 24, 4, 23, 21, 26, 10, 6, 2, 20, 3, 9, 27
Validation set	18, 7, 5, 1, 25

Fold 3	
Training set	27, 6, 9, 12, 26, 18, 4, 5, 23, 21, 2, 15, 24, 1
Validation set	20, 10, 7, 25, 3

Fold 4 (Final dataset)	
Training set	12, 7, 2, 9, 23, 25, 24, 3, 26, 5, 15, 27, 20, 6
Validation set	21, 10, 1, 4, 18

Fold 5	
Training set	5, 7, 26, 24, 21, 27, 12, 18, 2, 4, 25, 10, 9, 1
Validation set	6, 20, 3, 23, 15

Figure 17. Randomized dataset splits for cross-validation

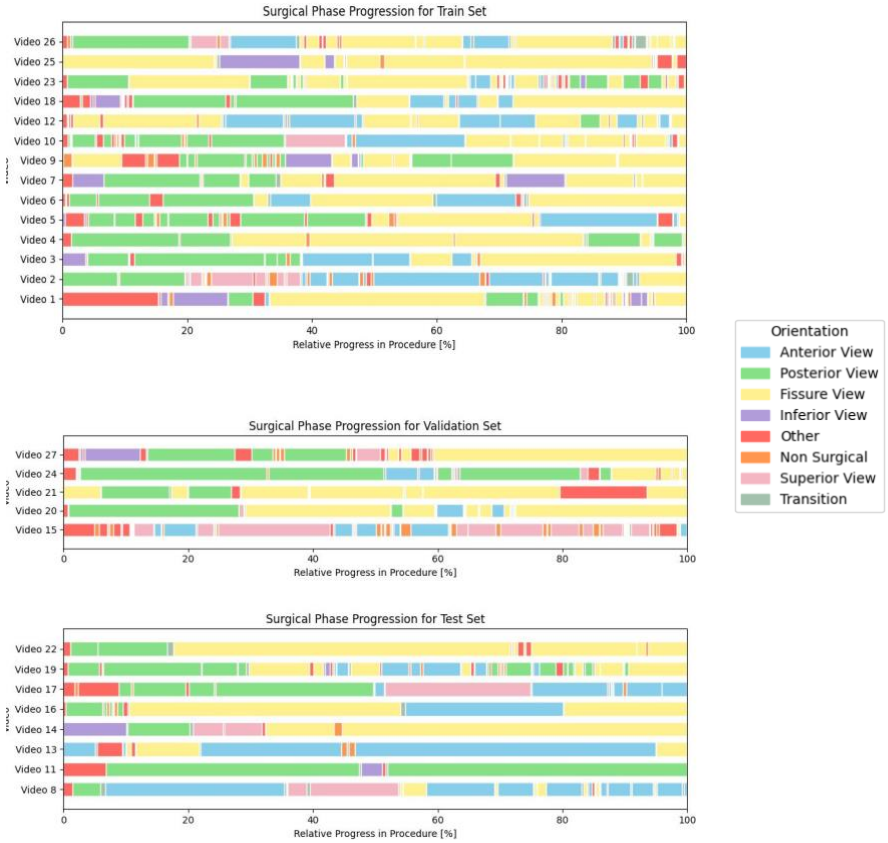


Figure 18. Orientation distribution across videos of the training, validation and test set of our Final Dataset

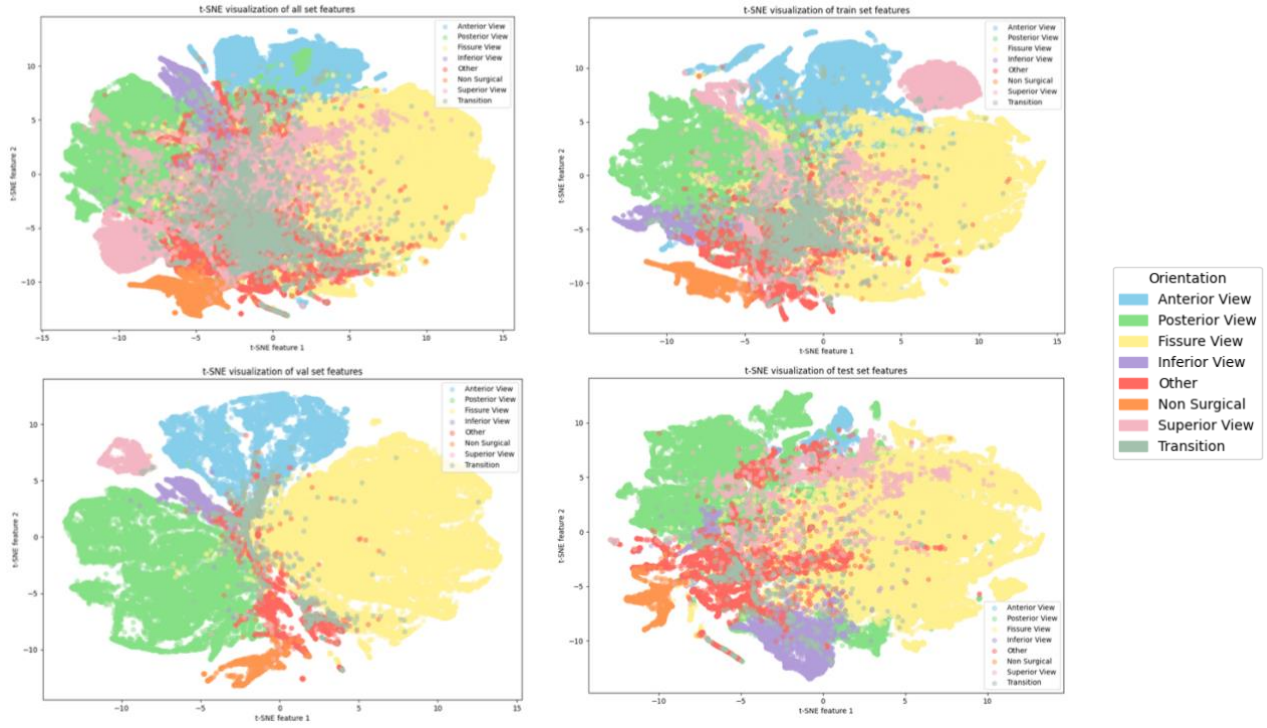


Figure 19. t-Distributed Stochastic Neighbor Embedding (t-SNE) Feature Maps illustrating the diverse distribution of lung orientation labels across the complete final dataset, the training set, the validation set and the test set. The color-codes differentiate among various labels, offering a visual representation of the dataset's complexity. Each point of the maps corresponds to a specific feature, plotted according to the feature similarities. It provides information about the clustering tendency and the ability to distinguish different lung orientations.

Appendix E. Orientation Distribution, Sequences & 3D Model Usage

Label	Percentage
Anterior	15.4%
Posterior	25.4%
Fissure	41.8%
Inferior	3.2%
Other	12.0%
Non-Surgical	2.2%
Total	100%
3D Model	4.2%

Figure 20. Distribution of orientations & 3D model usage across all video's

		Min. Duration			**Max Duration**			**Median Duration**			**Average Duration**		
Label	Sequences	Frames	Minutes	Seconds	Frames	Minutes	Seconds	Frames	Minutes	Seconds	Frames	Minutes	Seconds
Other	460	1	0.00	0.04	57646	38.43	2305.84	376	0.25	15.02	1359	0.91	54.37
Anterior	115	22	0.01	0.88	54814	36.54	2192.56	2282	1.52	91.28	7007	4.67	280.28
Posterior	145	28	0.02	1.12	64895	43.26	2595.80	3297	2.20	131.88	9162	6.11	366.49
Fissure	219	29	0.02	1.16	145947	97.30	5837.88	2036	1.36	81.44	9979	6.65	399.15
Inferior	32	88	0.06	3.52	32489	21.66	1299.56	2165	1.44	86.58	5204	3.47	208.16
Non-Surgical	122	144	0.10	5.76	4215	2.81	168.60	781	0.52	31.22	948	0.63	37.93
3D Model	102	18	0.01	0.72	40032	26.69	1601.28	225.5	0.15	9.02	2208	1.47	88.32

Figure 21. Sequences & sequence duration of orientations & 3D model usage across all videos

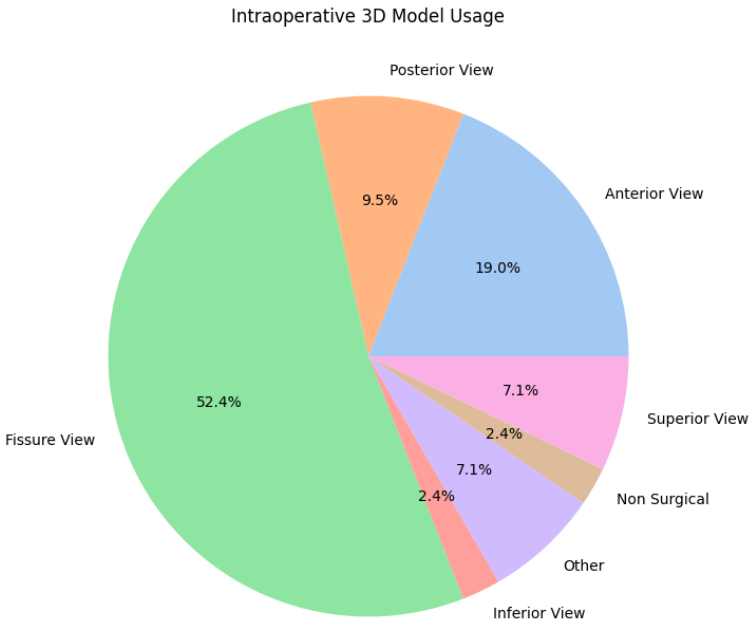


Figure 22. Distribution of 3D model usage across all orientations

Appendix F. Hyperparameter Tuning

Hyperparameter	Definition	Experiments
Data augmentation	To increase the diversity of your training dataset by applying various transformations (like rotation, scaling, cropping) to the original data. This helps in reducing overfitting and improves model generalization.	**Rotation, scaling and shifting**
Weights	Refers to the importance assigned to each label in the dataset, The Median Frequency Method extracts the median frequency of the occurrence of each label in a specified dataset and assigns weights accordingly. This approach can help to address class imbalance by assigning higher weights to less frequent labels, ensuring that the model pays more attention to these during training.	**Without weights** Median Frequency Weighting - Training Set
Input Height x Width	The dimensions of the input images fed into the model. (resolution), which affects the amount of detail the model can extract.	**224 x 224** 900 x 900: model becomes too slow and GPU cannot handle a batchsize > 50. 450 x 450:
Batch Size	The number of training samples processed before the model's internal parameters are updated. A larger batch size provides a more accurate estimate of the gradient, but requires more memory and computational power.	20, 40, 50, **80**, 450
Learning Rate	The step size at each iteration of the training process, controlling how much the model's weights are adjusted during training and is crucial for convergence and performance.	CNN: **0.0005** MS-TCN: **0.0007**
Early Stopping Metric	To stop training process if the model stops improving on a designated validation metric. This prevents overfitting and ensures that the model stops training when it achieves optimal performance.	Validation accuracy
Min Epoch	The minimum number of complete passes through the entire training dataset. It sets a lower bound to ensure that the model is exposed to the data sufficiently.	CNN: 1, 5-16, 20 MS-TCN:
Max Epoch	The maximum number of complete passes through the entire training dataset allowed for the Convolutional Neural Network. It prevents excessive training time and potential overfitting	CNN: 2, 7-20, 30 MS-TCN:
MSTCN Layers	The number of layers in the Multi-Stage Temporal Convolutional Network. Each layer contributes to the model's ability to learn and represent temporal features in the data.	8, 15
MSTCN Feature Maps	The number of distinct feature maps generated by each layer in the Multi-Stage Temporal Convolutional Network. Feature maps are the outputs of the convolutional layers and represent learned features.	32 **64**
MSTCN Feature Dimensions	The size of the feature representations in each layer of the Multi-Stage Temporal Convolutional Network. This affects the model's capacity and computational requirements.	2048
MSTCN Stages	The number of sequential stages in the Multi-Stage Temporal Convolutional Network. Each stage is designed to capture temporal relationships at different scales or complexities.	0, 1, 2, **3**
Dataset	The collection of data samples used for training, validating, and testing the model. The quality, size, and diversity of the dataset significantly influence the model's performance.	Initial dataset **5-fold Cross Validation** **Best dataset Cross Validation**

Table 3. List of Hyperparameters

For model training, all videos were processed at 1 fps. The ResNet50 model was pretrained on ImageNet and finetuned on our dataset. For actual training, videos were downsized to 224x224 pixels and data augmentation was performed including shifting, scaling and rotation. The learning rate was set to 0.0005 and the batch size was set to 80 frames. An early stopping metric was

determined as an unchanging validation accuracy for 3 epochs. The training process of the feature extraction stage was extended over a minimum of 15 and a maximum of 16 epochs, for all 5 folds of cross-validation. Given the imbalanced multi-class problem that orientation recognition involves, softmax activations were employed and a weighted cross-entropy loss was applied. The class weights were initially determined through median frequency balancing, to alleviate the imbalance between orientations [22]. However, this proved to give inferior results due to the difference in dataset composition of the train and test set and was therefore not continued with. Resnet50 requires a squared input., for which enlarging the input height and width to increase resolution was considered but didn't show improved results while prolonging the training process. The best combination of minimum and maximum epochs, was based on several experimental runs, mainly observing whether the loss function was still decreasing.

For the TeCNO model, the number of MS-TCN stages was set to 3, each stage including 8 layers of TCN's and each layer resulting in an output of 64 feature maps. The MS-TCN model is trained with a maximum number of 16 epochs and a learning rate of 0.0007. The early stopping metric was based on an unchanging validation accuracy, with a persistence of 3 epochs. Several experiments were performed with the number of layers, the feature maps and the stages of the MSTCN stage. The best combination was based on the performance on the test dataset. Most important hyperparameters were chosen among the options provided in the table, with our final selection in bold.

After cross validation results were interpreted both as an average over all folds as for the model showing best performance. Selection of the final model is based on the best performance in test results observed during the testing phase.

Appendix G. 3D model orientation – Clinical workflow

The intraoperative use of the dynamic 3D model by PulmoSR (Medical VR, Amsterdam, the Netherlands) offers surgeons intraoperative patient-specific visualizations. These 3D models are increasingly being utilized for both preparation and intraoperative guidance in lobectomy or segmentectomy lung surgeries at Cardiothoracic Surgery Department of the EMC (Rotterdam, the Netherlands).

During surgery, the 3D model guides surgeons in gaining a better understanding of intraoperative anatomy, especially of the initially invisible bronchovascular anatomy which divide the lung into its various lobes and segments. Hence, the 3D model can aid in anatomy recognition, assist in the selection of surgical approaches, and serve as a verification step before dissection and stapling. Currently, transformation of the 3D model's orientation to match the intraoperative orientation of the lung is performed manually. This often required an additional medical specialist with experience in this subject, since these adjustments require clinical insight to recognize the current orientation and can be time-consuming. Extensice transformation of the 3D model, to e.g., the interlobal fissure view, is iterative process consisting of multiple steps.

Therefore, the automatic recognition of intraoperative orientation and the subsequent automatic adjustment of the 3D model could significantly enhance surgical guidance. However, the actual implementation of this requires extensive research in both medical and technical fields. An interesting area for further exploration is whether actual real-time automatic orientation detection is necessary for continuous model adjustment. Our study results show that the 3D model is used in only 4.2% across all videos, suggesting that real-time application may not be essential. Instead, rapid rapid orientation detection and subsequent model adjustment upon request could be sufficient for improved intraoperative guidance. The implementation of such an application could be in two development steps. We conceptualize both steps below.

NEXT STEP - Automatic Intraoperative Orienatation Detection & Preset 3D Model Views

1. Automatic intraoperative detection of lung orientation.
2. Based on a large dataset of intraoperative images in five different lung orientations, for each type of pulmonary lobectomy/segmentectomy resection, an average transformation of the 3D model for each orientation is determined serving as the starting orientation.
3. The 3D model is automatically transformed to this standard orientation.
4. Manual adjustments of the 3D model orientation can be performed to fully correspond to intraopertive anatomy.

An example of this development phase is provided in figure.

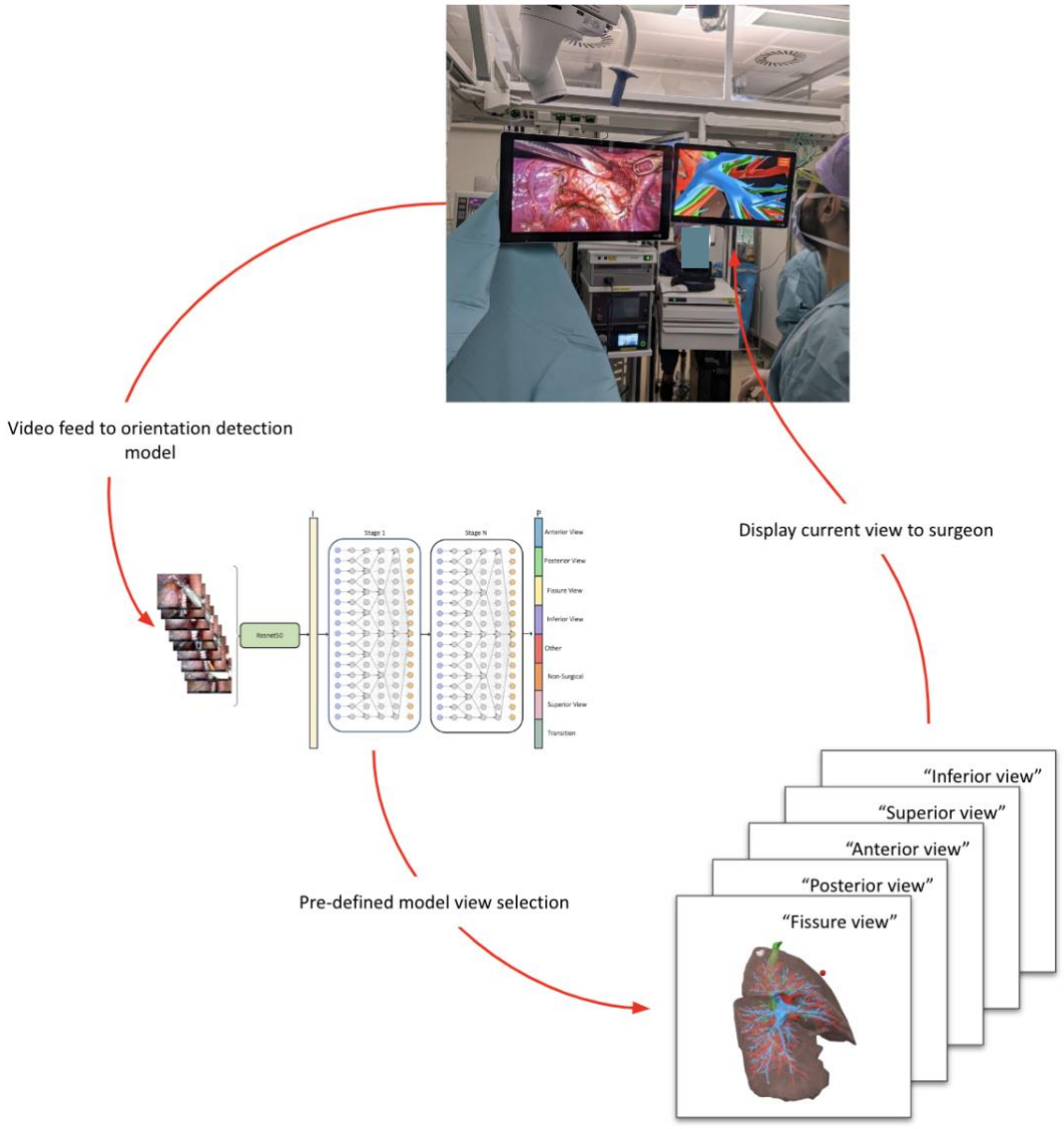


Figure 23. Future implementation of automatic intraoperative orientation detection for dynamic 3D model adjustment

Future - Automatic Intraoperative Orienatation Detection & Automatic 3D Model Registration

1. Automatic intraoperative detection of lung orientation, e.g., through generation of a point cloud or detection of anatomical landmarks such as the lung parenchyma, pericardial tissue, thoracic aorta and exposed bronchovasculare.
2. Automatic registration of the intraoperative orientation to the 3D model.
3. Automatic dynamic adjustment of the 3D model, for real-time intraoperative registration.

An example of this development phase is provided in figure

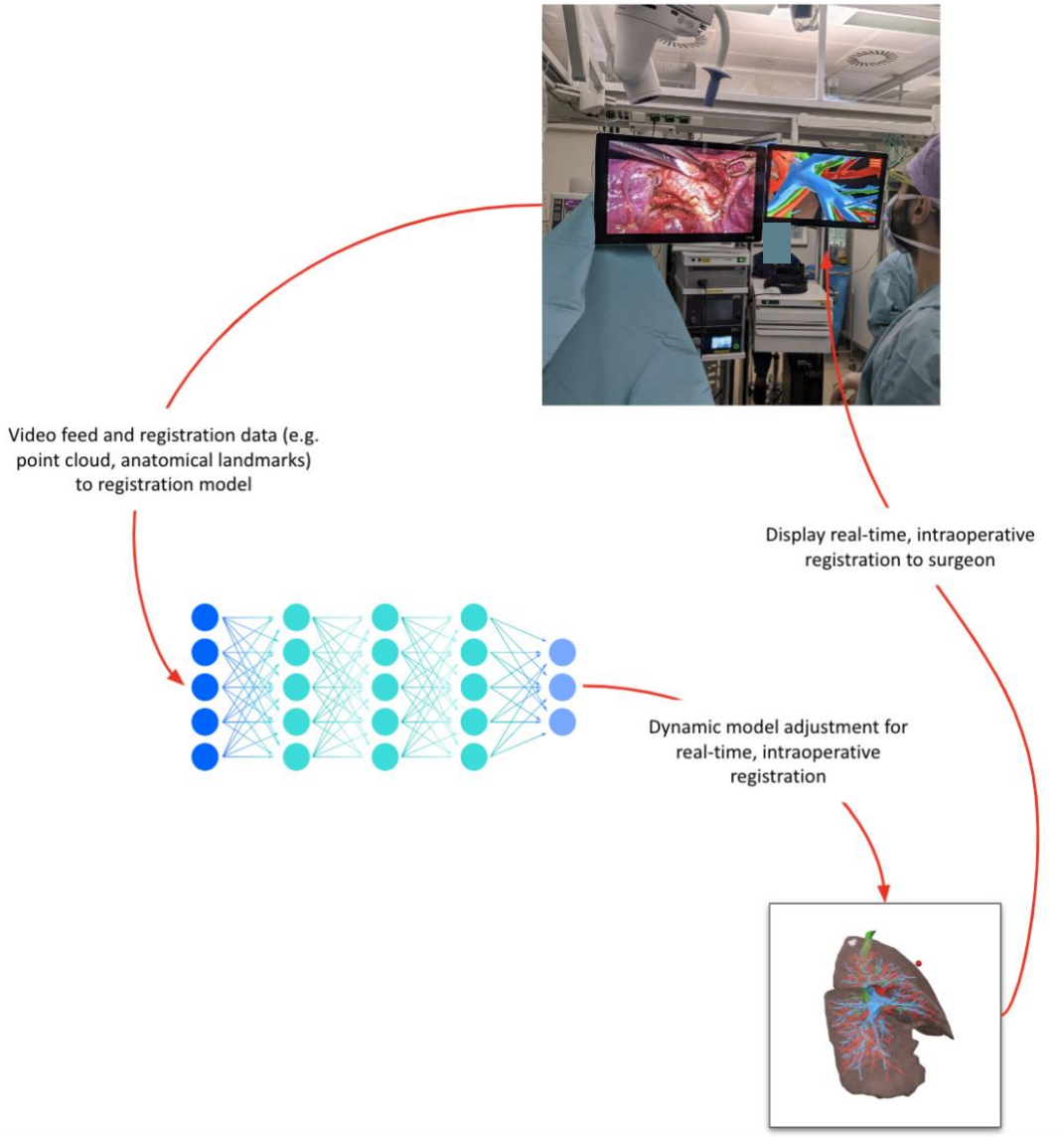


Figure 24. Implementations of automatic intraoperative orientation detection & subsequent automatic dynamic 3D model registration