# SIFTA: Surgical Instrument Trajectory Forecasting with Anatomy

**Gary Sarwin**[1] gary.sarwin@vision.ee.ethz.ch
**Alessandro Carretta**[2,3] alessandro.carretta1@gmail.com
**Sara Sangalli**[1] sara.sangalli@vision.ee.ethz.ch
**Victor Staartjes**[2] VictorEgon.Staartjes@usz.ch
**Matteo Zoli**[3] matteo.zoli4@unibo.it
**Diego Mazzatenta**[3] diego.mazzatenta@isnb.it
**Luca Regli**[2] luca.regli@usz.ch
**Carlo Serra**[2] carlo.serra@usz.ch
**Ender Konukoglu**[1] ender.konukoglu@vision.ee.ethz.ch
[1] *Computer Vision Lab, ETH Zurich, Switzerland*
[2] *Department of Neurosurgery, University Hospital of Zurich, Zurich, Switzerland*
[3] *Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy*

**Editors:** Under Review for MIDL 2026

## Abstract

Surgical guidance can be delivered in various ways. In neurosurgery, spatial guidance and orientation are predominantly achieved through neuronavigation systems that reference pre-operative MRI scans. Recently, there has been growing interest in providing *live* guidance by analyzing video feeds from tools such as endoscopes. Existing approaches, including anatomy detection, orientation feedback, phase recognition, and visual question-answering, primarily focus on aiding surgeons in assessing the current surgical scene. This work explores finer scale guidance, in particular providing guidance by forecasting the trajectory of the surgical instrument, essentially trying to answer the question of *what to do next?*. To the best of our knowledge, this work is the first attempt to address this task for manually operated surgeries. Our approach leverages past locations of the surgical instrument as well as the relevant anatomical features. We further account for variations in inferred camera angles. For all these, we use a specialized detection model's outputs from the same video without relying on additional input or interaction. The underlying intuition is that the anatomical context informs the surgeon's next actions. The experiments on a comprehensive dataset containing pituitary surgery support this notion: we report an accuracy of 62.41% by classifying the predicted direction into the four principal directions of movement, compared to 53.25% using only the surgical instrument's historical locations. This demonstrates that anatomical features are a valuable asset in addressing this challenging task. Our findings indicate that trajectory prediction is very limited when relying only on the instrument's past locations, but becomes noticeably more feasible when anatomical context is incorporated. This represents an important early step toward more accurate predictive models. The code will be released upon acceptance.

**Keywords:** Surgical Guidance, Surgical Data Science, Anatomical Detection

## 1. Introduction

**Surgical guidance** is one of the hallmark problems in computed-aided interventions. The main goal is simply to build algorithms to help the surgeon during the surgery. This guidance can come in multiple forms, in neurosurgery the major ones being neuronavigation and surgical action or phase recognition (Härtl et al., 2013; Orringer et al., 2012; Garrow et al., 2021). While the former is crucial for helping surgeons orient themselves in the anatomy, the latter can help the organization around the surgery as well as the training of new surgeons. Both of these guidance forms have seen tremendous advances over the last years. In this work, we explore and assess the possibility of a new form of guidance, *forecasting surgical tool movement*. Related ideas have been explored in robotic surgery, where kinematics are available and incorporated (Weede et al., 2011; Qin et al., 2020; Kim et al., 2024; Saeidi et al., 2022). For manually operated surgeries, such forecasting could be particularly valuable; however, kinematic data are not accessible as conventional surgical instruments are rarely tracked and lack the sensors required to measure their motion. As a result, video-based methods remain the only practically obtainable source of motion cues in manually operated procedures. Accordingly, our approach operates exclusively on information extracted directly from the surgical video.

The forecasting problem can be coarsely defined as predicting the next movement of the surgical tool. The first step towards this direction is simply predicting the next location of the surgical tool. In the setting of surgical videos or video streams, this corresponds to predicting the location of the surgical tool in the next video frames given the previous ones. This is a narrower subproblem of the broader problem of predicting the surgeon's next action but it forms the first step. It can also be viewed as a narrower version of another broader problem, forecasting frames in a video, which may not be necessary for navigation purposes.

The applications of forecasting surgical tool movement can be multiple. First, it is a natural next step for surgical guidance. It goes one step further than providing information for self-orientation; it provides guidance towards how to move the surgical tool in a given scenario. Moreover, one can automatically derive surgical patterns and best practices from expert surgeons, leveraging vast video databases maintained by multiple institutions. These algorithms could contribute to the training and guidance of less experienced surgeons, democratizing access to expert knowledge and potentially supporting developments in autonomous surgical systems.

Recognition of surgical tools in surgical videos as well as recognition of surgical actions being performed in a set of observed frames have received ample attention from the research community. Advanced tools of today can make accurate predictions for both problems (Das et al., 2024; Padoy, 2019). Forecasting, however, is arguably a much more challenging problem than these two. In particular, anticipating the next motion of a surgical instrument requires modeling the behavior of an expert operator acting within a highly complex environment. A surgeon's movements reflect the rapid integration of anatomical context, procedural intent, and ergonomic constraints, all refined through extensive experience and often manifested as implicit, non-verbalizable motor strategies. As a result, trajectory forecasting demands capturing both fine-grained spatial structure and latent aspects of expert decision-making, making it substantially more difficult than standard recognition tasks.

Advances in this problem, especially through the recent works in structure recognition and detection (Staartjes et al., 2021; Sarwin et al., 2023, 2024), have opened up new opportunities towards tackling this challenging forecasting problem. Here, we build upon these recent advances and, to the best of our knowledge, propose one of the first attempts to solving the forecasting problem.

In this work, we propose a deep learning model that forecasts the surgical tool's position in subsequent frames from prior video frames. Motivated by the intuition that the visible anatomy largely guides the surgeon's next decision, the model integrates anatomical cues as contextual features for trajectory forecasting. The model uses a very simple neural network architecture to predict the location of a surgical tool in the next 8 or 16 frames given the previous 64 frames. The model input can vary, the simplest being the location of the surgical tool in the given frames. Here, we evaluate the forecasting accuracy using different levels of information on the scene as input, including the location of the surgical tool, and locations of anatomical structures as predicted by a detector (YOLO (Wang et al., 2023)), which is trained on surgical videos. Our experimental results on pituitary surgery videos suggest that the value of the anatomical features may be a key component towards solving the forecasting problem.

## 2. Methods

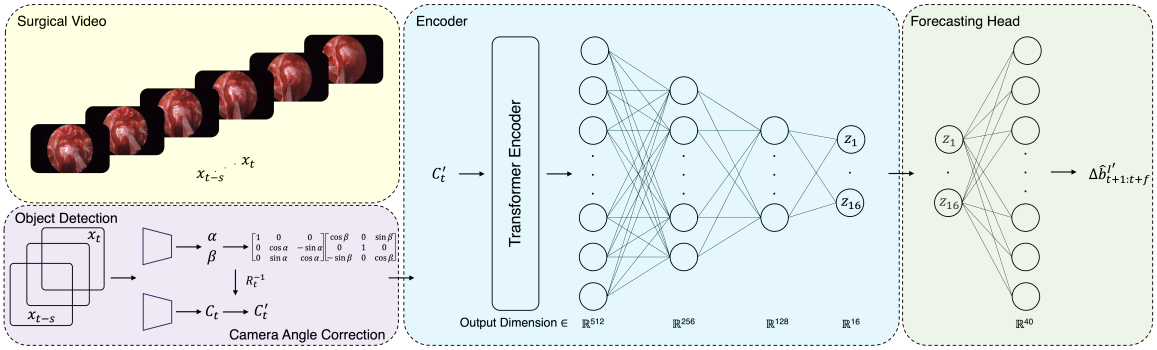


Figure 1: Overview of the proposed model pipeline. Frames extracted from a surgical video are processed through an object detection network, which identifies and localizes anatomical structures and surgical instruments across a sequence of frames. The detections withing these sequences are then rotated to a centered reference view using an additional encoder which approximates the orientation of the camera to compensate for variations in camera angle. These modified detections are then passed into an encoder to extract temporal features. Finally, the forecasting head predicts the bounding box changes of the instrument for the next $f$ frames.

### 2.1. Problem Formulation and Approach

Let $\mathbf{S}_t$ denote a sequence of endoscopic video frames $\mathbf{x}_{t-s:t}$, where $s$ is the sequence length and each frame $\mathbf{x}_t \in \mathbb{R}^{w \times h \times k}$ has width $w$, height $h$, and $k$ channels. Our goal is to forecast the change in the surgical instrument's bounding box over the next $f$ frames. Let the future frame-to-frame changes be

$$\Delta \mathbf{b}^I_{t+1:t+f} = \left[\Delta \mathbf{b}^I_{t+1}, \ldots, \Delta \mathbf{b}^I_{t+f}\right]^T \in \mathbb{R}^{f \times 4},$$

where each bounding box is parameterized by its center $(x, y)$ and width and height $(w, h)$, and $I$ indexes the surgical instrument class. Each $\Delta \mathbf{b}^I_{t+i}$ denotes the change between frame $t+i-1$ and $t+i$ in the surgical instrument's bounding box parameters.

The forecasting model operates on *centered detections*, an idealized representation in which the anatomical structures and instrument are expressed under a canonical reference camera pose. We use a prime $(')$ to denote quantities represented in this centered reference view. We first describe how centered detections are obtained and then how they are used for prediction.

### 2.2. Object Detection and Camera-Angle Normalization

To identify the surgical instrument and anatomical structures, we apply YOLOv7 (Wang et al., 2023) independently to each frame $\mathbf{x}_t$ in the sequence. For each frame, the detector produces raw detections

$$\mathbf{c}_t = \{\mathbf{y}_t, \mathbf{b}_t\},$$

where $\mathbf{y}_t \in \{0, 1\}^n$ indicates the presence of $n-1$ anatomical structures and the instrument, and $\mathbf{b}_t \in \mathbb{R}^{n \times 4}$ contains the corresponding bounding boxes, normalized with respect to the image size. Applying the detector across the sequence yields the detection set

$$\mathbf{C}_t = \mathbf{c}_{t-s:t}.$$

Since the detections in $\mathbf{C}_t$ are obtained under unconstrained endoscope orientations, we normalize them to a canonical centered view. To estimate the orientation of each frame, we employ the unsupervised framework of (Sarwin et al., 2024), which predicts pitch and yaw deviations from the centered view. These predictions define a rotation matrix $\mathbf{R}_t \in \mathbb{R}^{3 \times 3}$ mapping reference-view detections to the current orientation.

Centered detections are obtained by applying the inverse rotation:

$$\mathbf{C}'_t = \mathbf{R}_t^{-1} \mathbf{C}_t,$$

yielding detections expressed in the canonical coordinate frame. The centered sequence $\mathbf{C}'_t$ serves as the input to the forecasting network. An overview of the full pipeline is shown in Fig. 1.

### 2.3. Future Instrument Location Prediction

Given the centered detections $\mathbf{C}'_t$, the forecasting model predicts the future bounding-box changes

$$\Delta \hat{\mathbf{b}}'^I_{t+1:t+f} = \left[\Delta \hat{\mathbf{b}}'^I_{t+1}, \ldots, \Delta \hat{\mathbf{b}}'^I_{t+f}\right]^T.$$

The model parameters are optimized to ensure that the predicted changes match the ground-truth changes over the prediction horizon. For the $t$-th frame in training video $m$, the training loss is defined as

$$\mathcal{L}_{m,t} = \sum_{r=t+1}^{t+f} \left| \Delta \mathbf{b}'^{I}_{m,r} - \Delta \hat{\mathbf{b}}'^{I}_{m,r} \right| + \lambda \cdot \left( 1 - \frac{\langle \bar{\mathbf{v}}_{p,m,t:t+f}, \bar{\mathbf{v}}_{g,m,t:t+f} \rangle}{\|\bar{\mathbf{v}}_{p,m,t:t+f}\| \|\bar{\mathbf{v}}_{g,m,t:t+f}\|} \right), \tag{1}$$

where:

- $\Delta \mathbf{b}'^{I}_{m,r} = \mathbf{b}'^{I}_{m,r} - \mathbf{b}'^{I}_{m,r-1}$ is the ground-truth change in centered bounding-box values between frames $r-1$ and $r$,
- $\Delta \hat{\mathbf{b}}'^{I}_{m,r} = \hat{\mathbf{b}}'^{I}_{m,r} - \hat{\mathbf{b}}'^{I}_{m,r-1}$ is the predicted change over the same interval,
- $\bar{\mathbf{v}}_{g,m,t:t+f}$ and $\bar{\mathbf{v}}_{p,m,t:t+f}$ denote the cumulative displacement vectors of the ground-truth and predicted instrument motion across the prediction window. They are computed as

  $$\bar{\mathbf{v}}_{g,m,t:t+f} = \sum_{r=t+1}^{t+f} \Delta \mathbf{b}'^{I}_{m,r}, \qquad \bar{\mathbf{v}}_{p,m,t:t+f} = \sum_{r=t+1}^{t+f} \Delta \hat{\mathbf{b}}'^{I}_{m,r},$$

  where each vector represents the total movement of the bounding-box center over frames $t{+}1{:}t{+}f$, rather than instantaneous velocity, the subscripts $g$ and $p$ refer to ground truth and predicted quantities, respectively,
- $\langle \cdot, \cdot \rangle$ denotes the dot product,
- $\| \cdot \|$ denotes the Euclidean vector norm,
- $\lambda$ is a weighting factor that balances the directional-consistency term,
- $| \cdot |$ denotes the element-wise $L_1$ loss.

The first term encourages accurate frame-to-frame changes of the bounding box parameters, while the second penalizes deviations in the overall movement direction of the instrument across the prediction horizon. This prevents the model from producing locally correct movements that diverge from the correct global trajectory. The total training loss is computed by summing $\mathcal{L}_{m,t}$ over all frames and videos.

### 2.4. Training and Label Generation

The detection and forecasting networks are trained on disjoint sets of videos. For training the forecasting model, each sequence $\mathbf{S}_{m,t}$ is first passed through the detection network to obtain the raw detections $\mathbf{C}_t$. These detections are then mapped to the canonical centered view using the camera-angle normalization module, yielding the centered detections $\mathbf{C}'_t$, which serve as inputs to the forecasting network.

Ground-truth changes $\Delta \mathbf{b}'^{I}_{m,t+1:t+f}$ are obtained automatically by applying the detector to the future frames $\mathbf{x}_{t+1:t+f}$ and transforming the resulting bounding boxes to the centered view before computing frame-to-frame differences. These detections serve only to generate labels; at inference time, the forecasting network does not access future frames.

Finally, predicted bounding-box changes in the canonical view are rotated back to the original camera orientation:

$$\Delta\hat{\mathbf{b}}^{I}_{m,r} = \mathbf{R}_t\,\Delta\hat{\mathbf{b}}'^{I}_{m,r}.$$

## 3. Experiments and Results

### 3.1. Dataset

The medical dataset consists of 142 videos, each recorded during a unique pituitary surgery (transsphenoidal adenomectomy) performed on a different patient. These videos, collected over a 10-year period using various endoscopes and sourced from multiple centers, were made available under general research consent. Expert neurosurgeons provided annotations, identifying 16 classes in total: 15 representing distinct anatomical structures and one representing surgical instruments. Generally, there is one instance of each anatomical class per video, while multiple different instruments are grouped into a single instrument class. Of the 142 videos, 77 were allocated for training and validation of the object detection model, totalling approximately 10000 labeled images. The trained object detection model was then used to detect anatomy and instruments in each frame of 57 videos, which were used for training and validation of the model for surgical action forecasting. Finally, the remaining 8 were set aside for testing, with all frames in these videos processed by the object detection model to create the final test dataset. Although data originates from multiple sites, potential biases may still arise due to the proximity of the sites.

### 3.2. Implementation Details

The YOLO network and the model which approximates the camera angles were trained with identical parameters and implementations as reported in (Wang et al., 2023; Sarwin et al., 2023) and (Sarwin et al., 2024).

The forecasting model integrates a transformer encoder comprising six transformer encoder layers, each with five attention heads. The input to the transformer encoder has a size of $s \times n \times 5$, where $s$ represents the sequence length, set to 64 frames, and $n$ represents the 15 anatomical classes plus the additional surgical instrument class. Sinusoidal positional encodings are applied to these inputs to retain the temporal dimension of the sequence, and the transformer encoder output is fed through a series of three fully connected layers with output dimensions of 512, 256, and 128, respectively, using ReLU activation functions between the layers. The final fully connected layer outputs a latent representation $\mathbf{z}_t \in \mathbb{R}^{16}$ with 16 dimensions.

The decoder is implemented as a single linear layer, designed to take the latent vector $\mathbf{z}_t$ as input. This vector is processed through the layer to reach the final output of predicted sequence length × 4, where each predicted vector describes the bounding box changes.

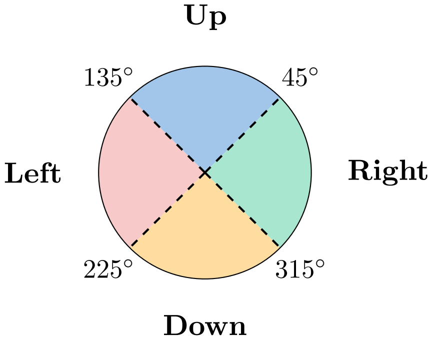
The output consists of frame-to-frame changes in bounding box coordinates, rather than absolute bounding box positions, to enable the model to effectively predict motion dynamics over the sequence. To enforce consistent cumulative movement, the model incorporates the direction-consistency term from the main loss (Eq. 1), which compares the cumulative displacement vectors of the predicted and ground-truth trajectories using cosine similarity.

For training, the AdamW optimizer (Loshchilov, 2017) is employed alongside a warm-up scheduler, which linearly increases the learning rate from 0 to 1e-4 over 60 warm-up epochs, and $\lambda$ was set to 0.5. The model is trained for a total of 75 epochs, and 150 for the task of predicting the instrument trajectory in the next 8 and 16 frames, respectively.

### 3.3. Results

**Object detection:** The trained object detection network achieves a mean average precision $mAP_{50}$ of 41.1%. More specifically, the average precision $AP_{50}$ for the instrument class equals 79.8%.

**Evaluation metric:** Due to the inherent uncertainty and variability in frame-to-frame instrument motion, accurately predicting exact displacements remains highly challenging. We therefore evaluate performance through a direction-classification metric, reflecting the idea that reliably capturing the coarse movement direction is a critical precursor to precise trajectory estimation. We classify the predicted direction of the instrument movement into four principal directions: *up*, *down*, *left*, and *right*. The classification is based on the angular range $\theta$ of the movement, measured in degrees, and is defined as follows:



Here, $\theta$ is computed as the angle of the cumulative displacement vector of the predicted bounding-box center over the forecast horizon.

**Accounting for noise in ground truth movement:** The ground truth motion of the surgical tools was estimated using their bounding box locations predicted with the detection network using the frames as input. These predictions contained jitter noise and this induces noise in the ground truth movements for the surgical tools. While this noise is desirable for training, since it adds to the robustness of the tool, it is detrimental to an accurate evaluation of the performance. We therefore additionally evaluated the performance of the tool for movements larger than a certain threshold, which are unlikely due to the jitter of the predictions of the detection network. To this end, in order to filter out noisy movements, we test movements of which the magnitude is greater than 0.1 and 0.05 of the image size, which when considering a horizontal movement and an image size of 1920 × 1280, corresponds to 192 and 96 pixels respectively. The filtering resulted in around 27000 test samples for the case of predicting 8 frames. These samples were then also used to generate the results of the models that predict 16 frames.

| **Model** | **# Frames $f$** | | | | | |
|---|---|---|---|---|---|---|
| | 8 | | | 16 | | |
| | 0.1 | 0.05 | 0 | 0.1 | 0.05 | 0 |
| Centered Anatomy + Instrument | **62.41** | **53.00** | **38.71** | 56.62 | 50.26 | 40.68 |
| Anatomy + Instrument | 61.39 | 51.97 | 38.30 | **57.20** | **50.50** | **40.90** |
| Instrument | 53.25 | 43.82 | 34.08 | 53.69 | 47.08 | 38.58 |
| Random | 24.70 | 24.91 | 24.97 | 25.31 | 25.26 | 25.15 |

Table 1: Direction classification accuracy (in %) for forecasting the next 8 and 16 frames under different model variants and movement thresholds. Random predictions are shown as a baseline.

**Forecasting results:** The results in Table 1 provide a quantitative evaluation of the forecasting model under different configurations and prediction horizons. Across all thresholds and both forecast lengths, incorporating anatomical detections improves direction classification accuracy compared to using instrument motion alone, indicating that anatomical context contributes additional useful information for anticipating tool movement.

The effect of compensating for camera angles shows a mixed trend. For short-term predictions ($f = 8$), centering the detections yields a modest improvement, suggesting that reducing variability caused by camera motion can stabilize the forecasts over shorter temporal windows. For longer predictions ($f = 16$), this benefit is not observed, which may be due to accumulated drift in the estimated camera pose, reducing the effectiveness of the compensation over extended sequences. As expected, accuracy decreases when evaluating all movements (threshold 0), reflecting the increased ambiguity associated with small displacements. Nevertheless, the anatomy-aware models consistently outperform the instrument-only model across all threshold levels, demonstrating the robustness of the contribution of anatomical cues. The random baseline performs near the expected accuracy of 25%, confirming that all learned models are capturing meaningful structure in the data. Qualitative results are visualized in Fig. 2.

Fig. 3 shows the directional confusion matrices for two forecasting settings (Ct. Anatomy + Instrument and Instrument, both with threshold 0.1). Incorporating anatomical context leads to an increased accuracy compared to when using instrument motion alone, and yields improvements in several movement classes, particularly for left and right directions. The instrument only model seems to predominantly predict up and down movements, which arguably correspond to the insertion and removal of the surgical tool. On the contrary, the model that utilizes anatomy features shows improved performance for left and right movements, reflecting an enhanced ability to capture more nuanced directional patterns. This supports the notion that anatomical context offers complementary spatial information that strengthens the coherence of the resulting forecasts.

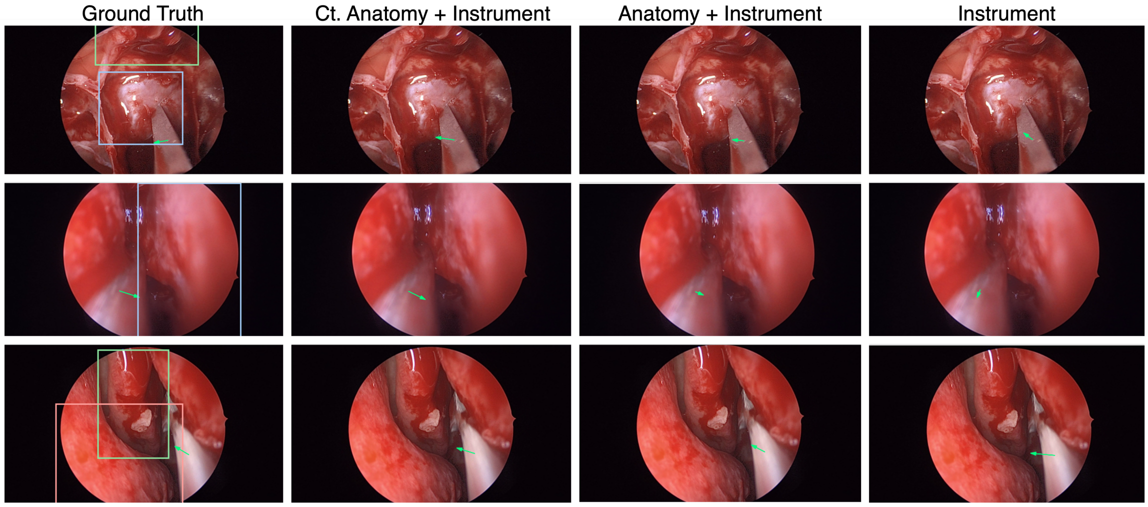


Figure 2: Qualitative comparison of the forecasting models for an 8-frame prediction horizon. Each predicted vector represents the overall displacement of the instrument's bounding-box center. Anatomy detections are shown in the ground-truth images. The centered variant (second column) additionally compensates for camera orientation by transforming detections to a learned reference view.

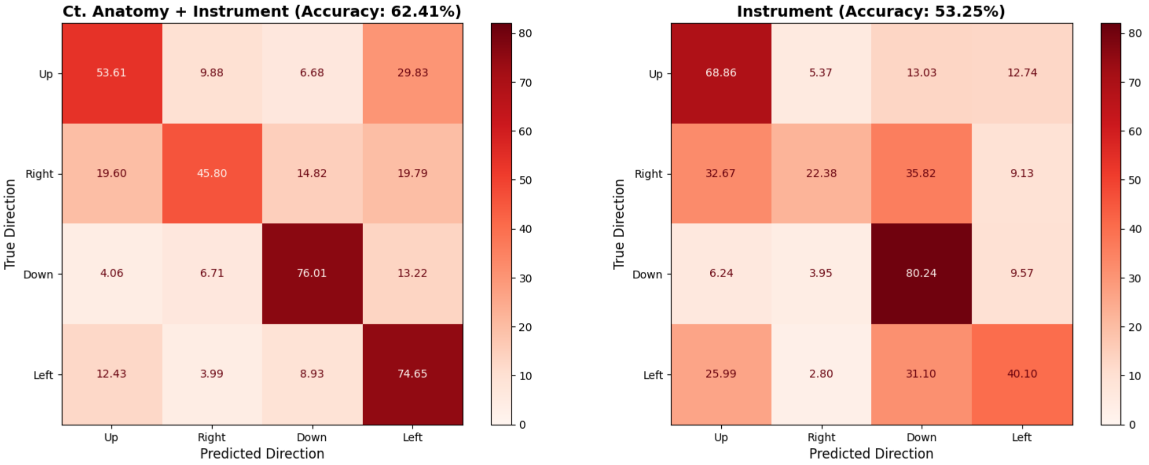


Figure 3: Confusion matrices for directional forecasting when using anatomical context (left) and instrument history alone (right). Incorporating anatomical information increases overall accuracy (62.41% vs. 53.25%) and reduces the frequency of large directional errors, with most misclassifications occurring toward neighboring directions.

## 4. Conclusion and Discussion

We proposed a novel framework for forecasting surgical tool trajectories from endoscopic videos by integrating anatomical structure detection with a transformer-based model. Our results demonstrate that using only the instrument's past locations provides insufficient information for reliable trajectory prediction, whereas incorporating anatomical context elevates the task into a feasible predictive setting.

Crucially, our approach enables unsupervised learning from expert demonstrations and thereby could support democratizing critical surgical knowledge. With improved accuracy, the model's predictions could offer real-time suggestions that reflect expert best practices, effectively guiding surgeons toward actions an expert would take.

However, several limitations must be acknowledged. The current framework does not differentiate between different surgical instruments and only considers one tool even when multiple are present. This is a notable restriction, as different instruments likely exhibit distinct movement patterns and are used for different surgical intents. Furthermore, variations in camera angles and inherent noise in detection outputs introduce uncertainties that impact the reliability of the predictions. In addition, the model predicts motion only in the image plane, meaning that depth changes or out-of-plane movements cannot be captured in its current form.

Our future work will focus on incorporating instrument classification, improving camera compensation strategies, extending the approach to multi-tool situations, implementing an autoregressive architecture, and incorporating visual features.

This work is a first step and we hope it will pave the way towards accurate forecasting models and ultimately accurate guidance on the next surgical action.

## Acknowledgments

This work was funded by the SNSF (Project $IZKSZ3\_218786$). We would like to thank Prof. Ron Kikinis for the input on surgical guidance.

## References

Adrito Das, Danyal Z Khan, Dimitrios Psychogyios, Yitong Zhang, John G Hanrahan, Francisco Vasconcelos, You Pang, Zhen Chen, Jinlin Wu, Xiaoyang Zou, et al. Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. *arXiv preprint arXiv:2409.01184*, 2024.

Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021.

Roger Härtl, Khai Sing Lam, Jeffrey Wang, Andreas Korge, Frank Kandziora, and Laurent Audigé. Worldwide survey on the use of navigation in spine surgery. *World neurosurgery*, 79(1):162–172, 2013.

Ji Woong Kim, Tony Z Zhao, Samuel Schmidgall, Anton Deguet, Marin Kobilarov, Chelsea Finn, and Axel Krieger. Surgical robot transformer (srt): Imitation learning for surgical tasks. *arXiv preprint arXiv:2407.12998*, 2024.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Daniel A Orringer, Alexandra Golby, and Ferenc Jolesz. Neuronavigation in the surgical management of brain tumors: current and future trends. *Expert review of medical devices*, 9(5):491–500, 2012.

Nicolas Padoy. Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies*, 28(2):82–90, 2019.

Yidan Qin, Seyedshams Feyzabadi, Max Allan, Joel W Burdick, and Mahdi Azizian. davincinet: Joint prediction of motion and surgical state in robot-assisted surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2921–2928. IEEE, 2020.

Hamed Saeidi, Justin D Opfermann, Michael Kam, Shuwen Wei, Simon Léonard, Michael H Hsieh, Jin U Kang, and Axel Krieger. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science robotics*, 7(62):eabj2908, 2022.

Gary Sarwin, Alessandro Carretta, Victor Staartjes, Matteo Zoli, Diego Mazzatenta, Luca Regli, Carlo Serra, and Ender Konukoglu. Live image-based neurosurgical guidance and roadmap generation using unsupervised embedding. In *International Conference on Information Processing in Medical Imaging*, pages 107–118. Springer, 2023.

Gary Sarwin, Alessandro Carretta, Victor Staartjes, Matteo Zoli, Diego Mazzatenta, Luca Regli, Carlo Serra, and Ender Konukoglu. Vision-based neurosurgical guidance: Unsupervised localization and camera-pose prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 736–746. Springer, 2024.

Victor E Staartjes, Anna Volokitin, Luca Regli, Ender Konukoglu, and Carlo Serra. Machine vision for real-time intraoperative anatomic guidance: a proof-of-concept study in endoscopic pituitary surgery. *Operative Neurosurgery*, 21(4):242–247, 2021.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

Oliver Weede, Holger Mönnich, B Müller, and Heinz Wörn. An intelligent and autonomous endoscopic guidance system for minimally invasive surgery. In *2011 IEEE international conference on robotics and automation*, pages 5762–5768. IEEE, 2011.