Multimodal Transformers for Real-Time Surgical Activity Prediction

Keshara Weerasinghe**[1], Seyed Hamid Reza Roodabeh**[1], Kay Hutchinson[1], Homa Alemzadeh[1]

***Abstract*— Real-time recognition and prediction of surgical activities are fundamental to advancing safety and autonomy in robot-assisted surgery. This paper presents a multimodal transformer architecture for real-time recognition and prediction of surgical gestures and trajectories based on short segments of kinematic and video data. We conduct an ablation study to evaluate the impact of fusing different input modalities and their representations on gesture recognition and prediction performance. We perform an end-to-end assessment of the proposed architecture using the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset. Our model outperforms the state-of-the-art (SOTA) with 89.5% accuracy for gesture prediction through effective fusion of kinematic features with spatial and contextual video features. It achieves the real-time performance of 1.1-1.3ms for processing a 1-second input window by relying on a computationally efficient model.**

I. INTRODUCTION

Surgical robots translate the intricate movements of a surgeon's hands, wrists, and fingers into precise actions performed by miniature surgical instruments and offer many advantages, including improved visual perception, heightened surgical dexterity, reduced incision size [1], and shortened postoperative recovery periods [2]. Their adoption in surgical specialties such as urology, gynecology, and general surgery is not only enhancing surgical precision and quality, but is also opening avenues for the development of autonomous systems [3], [4], [5], [6], automated skill assessment [7], [8], [9], and error detection [10], [11], [12]. However, the development of these systems in robot-assisted minimally invasive surgery (RMIS) requires the understanding and perception of surgical activities carried out during surgical operations to support lower-level analysis [13], [14] of procedures.

Multiple modalities of data are available from surgical robots including video and kinematic data that can be used separately or in combination for the recognition and prediction of surgical gestures. Previous works have proposed methods for the recognition of gestures based on kinematic data from the surgical robot [15], [16], [17], [18], [19] or video data of the surgical scene [20], [21], [22], [23]. More recently, there have been several efforts that utilize the fusion of kinematic and video data for gesture recognition [24], [25]. However, most prior works have concentrated on the recognition of gestures based on the observation of a complete trial of a surgical task as opposed to short temporal segments observed at runtime. This limitation hinders the practical implementation of runtime gesture recognition in realistic environments for applications such as safety monitoring [10], [11], [12], training [26], autonomy [27], [28], and teleoperation [4], [3]. Recognizing a surgical gesture within a short temporal segment (e.g., a 1-second window) can help with timely intervention and feedback during simulated or real surgical tasks. However, window-based gesture recognition and prediction are more challenging compared to analyzing the entirety of a surgical trial due to the significantly reduced availability of contextual information. To our knowledge, fusion of multiple data modalities and the impact of different representations of modalities for window-based gesture recognition and prediction have not been studied before. Moreover, the *end-to-end* performance evaluation of gesture recognition and prediction, which is important for real-time interventions in real-world deployment, has not been explored in previous works.

To address these challenges, we propose methods that utilize the rich information embedded in different modalities by fusing kinematic and video data and exploring different representations of these modalities on the end-to-end performance of surgical activity (gesture and trajectory) prediction.

The main contributions of the paper are as follows:

- We propose a multimodal transformer model that utilizes the fusion of different modalities to recognize surgical gestures based on short temporal segments of surgical activity data (1-second), which is then used to predict surgical gestures and trajectories for a short temporal segment in the future.
- We conduct an ablation study on the impact of different modalities (including video and robot kinematics) and different representations of certain modalities (e.g., features extracted using ResNet50 [29], Spatial CNN [30], and contextual representations [31] of video data) on gesture recognition and prediction.
- We perform an end-to-end evaluation of our proposed model on the publicly available JIGSAWS dataset and show that our model can outperform a previous transformer model [15] in gesture prediction and trajectory prediction with a prediction accuracy of **89.5%** vs. 84.6%, while achieving a real-time performance of, respectively, 1.3ms and 1.1ms for a 1-second window in gesture recognition and prediction.

II. PRELIMINARIES

A. Surgical Gestures and Context

Previous works have modeled surgical procedures using a hierarchy [32] of steps, phases, tasks, gestures, and low-level

*This work was supported in part by the National Science Foundation under Grants CNS-2146295 and DGE-1842490.

** denotes equal contribution.

[1]Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA. {cjh9fw,ydq9ag,kch4fk,ha4d}@virginia.edu.

The code for this paper is available at https://github.com/UVA-DSA/MTRSAP

motion primitives [31]. Gestures are defined as purposeful actions imbued with semantic content that are specific and often involve particular instruments or objects. [33] proposed a framework that further decomposes surgical gestures into a sequence of elementary instrument movements referred to as "motion primitives", which encompass basic actions such as pushing, pulling, and grasping. [33] defined "context" as a set of states that describe the status and interactions among surgical tools, objects, and the physical environment which can be inferred from video data [31], [34]. A change in context happens as the result of the execution of a motion primitive within a gesture or task [33]. In this paper, we leverage context as an alternative representation of the information within the video to recognize surgical activities.

B. JIGSAWS Dataset

JIGSAWS [35] is a publicly available dataset of surgical tasks performed using the da Vinci robot [36]. It includes synchronized kinematic, video, and gesture transcripts collected from executions of three fundamental surgical tasks on a bench-top model by eight surgeons of three expertise levels. We evaluate our methods using 39 trials of the Suturing task.

The kinematic data in JIGSAWS captures the Cartesian positions ($\mathbf{p} \in \mathbb{R}^3$), rotation matrices ($\mathbf{R} \in \mathbb{R}^{3\times3}$), linear velocities ($\mathbf{v} \in \mathbb{R}^3$), rotational velocities ($\boldsymbol{\omega} \in \mathbb{R}^3$), and grasper angles ($\theta$) for left and right tools for both patient-side manipulators (PSM) and master-side manipulators (MTM), resulting in a total of 76 features sampled at 30Hz. The video data is collected at 30fps from an endoscopic camera. The dataset also contains manual annotations for gestures based on a predefined surgical activity vocabulary (see Table 2 in [35]), along with the skill levels of the subjects.

C. Gesture Recognition

Gesture recognition plays a vital role in identifying the present state of surgical procedures, enabling the detection of safety violations [10], [11], [12] and facilitating the prediction of future surgical actions [37], [15] with enhanced accuracy and confidence. Early research on surgical gesture recognition relied on probabilistic graphical models like Hidden Markov Models [38], [13], whereas contemporary studies predominantly focus on the utilization of deep learning (DL) techniques [20], [15] based on video [30],[23],[22],[21] and/or kinematic [39],[17],[15], [19] data. Specifically, Temporal Convolutional Networks (TCNs) [30], [20], [40] have been shown to efficiently capture temporal information for action segmentation based on video data from the JIGSAWS dataset. [25] employs the TCN in a parallel two-stream network with weighted fusion, and [24] utilizes TCN and LSTM to leverage multimodal data in improving surgical gesture recognition, but does not address gesture and trajectory prediction. Although both works also consider real-time implementations of their models, in general, limited attention has been directed toward real-time recognition based on short temporal segments through multimodal fusion. Furthermore, a gap remains in understanding the significance of different representations of certain modalities and their impact on gesture recognition accuracy and real-time performance as it is imperative for developing systems such as online safety monitoring [11], [12] for RMIS.

D. Gesture and Trajectory Prediction

Prediction of surgical activities including gestures and tool trajectories is an area of growing interest and applicability such as in visual window [41], [42] and surgical instrument [43] tracking. The precise and timely predictions of surgical states and trajectories can improve the success rates of teleoperated surgical procedures [44], [45] and guide the surgical tools in real-time autonomous operations [27], [28].

While some earlier studies relied on conventional methods, such as silhouette-based instrument tracking using Kalman filters [43], recent research is leveraging DL techniques. For instance, daVinciNet [37] adopts a fusion approach involving encoded multimodal features using LSTM [46] along with feature and temporal attention mechanisms. Similarly, [15] proposes a pipeline of three consecutive transformer models [47] for gesture recognition, gesture prediction and trajectory prediction based on kinematic features.

III. METHODS

In this section, we introduce our model for gesture recognition, prediction, and trajectory prediction within the context of short temporal segments. This model is built upon an adaptation of the original transformer model proposed by [47] for Natural Language Processing (NLP) and incorporates the fusion of multimodal data. The transformer model has proven its excellence in NLP tasks due to its capabilities of identifying long-term patterns from historical data [48], [49], [50], [51] as well as in the domain of sequence generation [52], [53]. We aim to utilize the strengths of the transformer architecture and introduce changes that can adapt to the context of RMIS for runtime surgical gesture recognition, prediction, and trajectory prediction. We also evaluate the impact of different modalities and their representations on the recognition and prediction performance.

Our proposed model is structured as a three-part pipeline, including the stages for Feature Extraction and Transformation, Gesture Recognition, and Gesture/Trajectory Prediction, as illustrated in Figure 1. More specifically, we process the input data features for an observation window, spanning from $t+1$ to $t+W_{obs}$, recognize the gesture(s) being performed in that window, and use this output along with other features to predict future gesture(s) and trajectory coordinates for a prediction window $t+W_{obs}+1$ to $t+W_{obs}+W_{pred}$.

The Gesture Recognition stage consists of a transformer encoder whose output is fed as input to another transformer model, separately trained for Gesture/Trajectory Prediction. We utilize multi-task learning to train a single transformer model to simultaneously predict both gestures and trajectories for a prediction window W_{pred}. This unified end-to-end model architecture is different from previous works [15] and [37], and particularly well-suited for real-time tasks, such as error detection and recovery [11], where timely response and reduced computational complexity are needed.

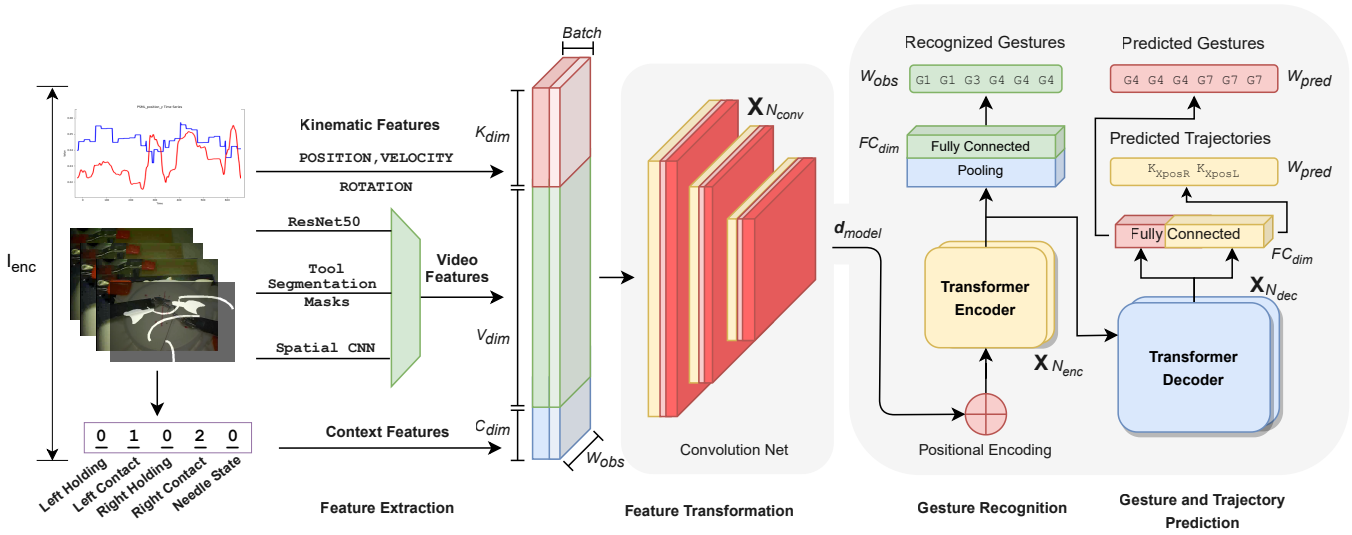


Fig. 1: Overall Architecture for End-to-End Real-Time Surgical Activity Recognition and Prediction

A. *Feature Extraction and Transformation*

The first stage of the pipeline transforms the multimodal input data into features representing the rich information embedded in the data. Of the 76 kinematic features in the JIGSAWS dataset, we explore using subsets of features including just the 38 kinematic features from the PSM side (K_{38}) and only the 14 kinematic features representing the position, velocity, and gripper angles from PSM (K_{14}).

In transfer learning, pre-trained CNNs are widely used to extract latent features from raw images as input for downstream tasks. We utilize different SOTA methods to extract feature representations from video data, including V_{Res} extracted using pre-trained ResNet50 [29], $V_{Spatial}$ extracted using a Spatial CNN proposed for action segmentation [30], and V_{Seg} surgical instrument and object segmentation masks extracted using memory networks, which have been shown to be effective in capturing interactions between instruments and objects [34]. Since the raw image features of the segmentation masks are high dimensional and the actual objects only occupy a relatively small area of the image, we first resize each segmentation frame by a factor of ten and then apply Principal Component Analysis (PCA) [54] to the resized images to extract a more compact representation of each frame. We also use the surgical context [33] defined as a state vector C representing the interactions between surgical instruments (e.g., graspers, scissors) and objects (e.g., needle, thread) in the surgical scene, which has been proposed as a fine-grained representation of surgical activity [33], [31]. and can be inferred from video data using a combination of knowledge and data-driven methods [34], [31].

We perform an ablation study on the above set of features to assess the effects of different modalities and their respective representations on gesture recognition and prediction. This study aims to determine the most suitable fusion of features for the subsequent feature transformation stage.

We draw upon the insights presented in [20] for the transformation of selected features from the kinematic and video data, which are then given as input to our model. We adopt the encoder component of the TCN model in [40] to efficiently capture features from the fused inputs as shown in Figure 1. The TCN encoder employs a stack of hierarchical $N_{conv} = 3$ temporal convolutional layers, pooling, and non-linear activations which effectively capture robust temporal relationships while enhancing computational efficiency, as shown in [40]. We modify the final convolutional filter to output a feature vector of dimension d_{model} that is subsequently fed into the encoder component of the transformer model for gesture recognition. During gesture and trajectory prediction, the feature transformation stage is skipped, and input features are directly fed into the encoder.

B. *Gesture Recognition*

For runtime gesture recognition, we leverage the encoder component derived from the transformer encoder-decoder [47] architecture. In NLP, the encoder module of the transformer architecture performs the comprehension and extraction of information embedded within the input text [51] and is mainly used for classification tasks. We regard the recognition of surgical gestures based on time series data as an adaptation of transformers for the time series classification, as evident in [50] and [49]. We also employ the decoder module of the transformer architecture to predict gestures and trajectories, as these tasks entail generative aspects that align well with the decoder's capabilities.

In the classical transformer implementation [47], an embedding layer transforms the input data into sequential token embeddings before sending it to the encoder. This step is not needed in surgical gesture recognition where numerical data is used instead of textual input. Our encoder uses a multi-headed attention mechanism to generate a feature vector of dimension d_{model}. A fully connected layer $FC_{dim} = 10$ is appended to the output of the encoder as depicted in Figure 1 to attain the desired dimension for the gesture output vector O_{enc}, representing each gesture class.

Real-time gesture recognition is achieved by using an observation window W_{obs} with a duration of 30 samples

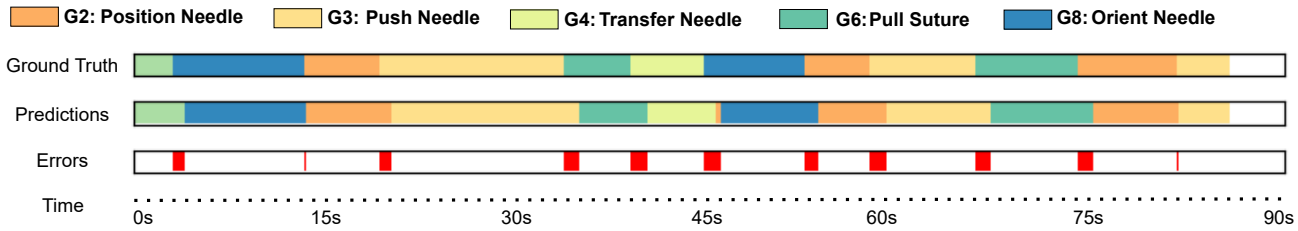


Fig. 2: A sample timeline of a Suturing trial, illustrating the gestures executed throughout the trial. Top Row: Actual gestures, Middle Row: Predicted gestures, Bottom Row: Error intervals (often occurring when transitioning to the next gestures).

(corresponding to $1s$) representing a short temporal segment as the input to the model. The output consists of gesture labels at runtime of length W_{obs} (see Figure 2). A tumbling window approach [55], instead of a sliding window, is used to decrease computational overhead.

C. Gesture and Trajectory Prediction

Our integrated framework for simultaneous gesture and trajectory prediction hinges on the recognition module's output to accurately generate gesture labels for the observation window W_{obs}. These labels, coupled with the encoder's output, and the original input features prior to the feature extraction stage, form the input to the transformer decoder. The decoder comprises N_{dec} layers, with each layer utilizing H_{dec} attention heads. This design allows the model to effectively consider various aspects of the input data and observed gestures when predicting the correct gesture at each time step. The hidden dimensions of both the transformer encoder and decoder are identical.

The decoder's output undergoes two linear transformations to produce the final outputs. One transformation maps the output to a set of probability weights for the gesture prediction task, while the other maps the output to the 3D Cartesian trajectory coordinates of the two robot end-effectors, represented by six real-valued numbers (X, Y, Z coordinates for each end-effector). We adopted a cumulative L_2 function over the prediction window to compute the regression loss for trajectory prediction. This loss is calculated based on the differences between the ground truth and predicted trajectory variables. We employed categorical cross-entropy over the prediction window for our classification loss which measures the disparity between the ground truth and predicted gesture labels. The ultimate loss function is a weighted combination of these two terms. The weights are hyper-parameters that are fine-tuned to achieve optimal model performance between trajectory prediction and gesture prediction.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

The experiments were done on a PC with an Intel Core i9-12900K 3.20GHz, 32GB RAM, and an NVIDIA GeForce RTX 3080 Ti 12GB GPU running Ubuntu 20.04.6 LTS.

We used the Leave-One-User-Out (LOUO) [56] cross-validation strategy to evaluate the model performance and to conduct the ablation study of the impact of different modalities and their representations. For surgical gesture recognition, we used the original sampling rate of data at $30Hz$ with $W_{obs} = 30$ samples which is equivalent to $1s$. For gesture prediction and trajectory prediction, we used the same temporal window length of 1 second, but downsampled the data from $30Hz$ to $10Hz$, thus having $W_{pred} = 10$.

Across all input configurations, for gesture recognition we maintained model hyper-parameters at $N_{enc} = 3$, $H_{enc} = 2$, and $d_{model} = 60$, whereas for gesture and trajectory prediction we maintained $N_{dec} = 2$ and $H_{dec} = 4$. Model training was done using the Adam optimizer [57] with a dynamic learning rate, as outlined in [47]. The training duration spanned 20 epochs with a batch size of 10.

B. Metrics

We evaluate the performance of each module and the overall end-to-end pipeline using the following metrics.

Gesture Recognition and Prediction: We use the standard accuracy and edit score metrics [58] to compare the recognized/predicted gestures to the ground truth labels. We evaluate the window-based classification using the F1@X metric [40], which defines recognized/predicted windows that overlap with actual windows by more than "X" percent as true positives and those with less overlap as false positives.

Trajectory Prediction: In order to assess the performance of the trajectory prediction module, we analyze the difference between the ground truth and predicted left and right end-effector trajectories within the Cartesian endoscopic reference frame using the standard metrics of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

Inference Time: We report the inference times for different stages of the pipeline to assess the impact of different feature representations on real-time performance. The inference times are measured for a 1-second window, averaged across trials and subjects. The inference time for gesture recognition encompasses both feature extraction and recognition.

C. Results

1) Gesture Recognition: Table I shows the results of the ablation study on the impact of different modalities and their representations on the performance of runtime gesture recognition for the **Suturing** task. We observe that utilizing a subset of kinematic features, K_{14}, yields superior results compared to utilizing all 38 kinematic features, K_{38}. Thus, K_{14} represents the essential characteristics [13] of a surgical gesture while omitting non-significant kinematic variables that are typically associated with the Suturing task. These results suggest that utilizing a refined subset of kinematic

TABLE I: Performance of gesture recognition in the ablation study of input features for Suturing task: K_{38} = 38 Kinematic features, K_{14} = 14 Kinematic features, C = Surgical Context features [33], $V_{Spatial}$ = Video features from Spatial CNN [30], V_{Res} = Video features from ResNet50 [29], V_{Seg} = Video features from Tool Segmentation Masks [34]

Input Features	Accuracy (%)	Edit Score (%)	F1@10 (%)	F1@25 (%)	F1@50 (%)	Inference Time (ms)
K_{38}	71.1	66.8	69.0	68.0	60.4	0.60
K_{14}	74.8	72.3	72.5	71.3	64.6	**0.55**
C	74.3	71.3	75.6	74.2	65.6	0.57
$V_{Spatial}$	**80.7**	**81.2**	**82.0**	**80.8**	**72.3**	1.52
V_{Res}	69.8	66.6	70.5	68.8	58.9	1.67
V_{Seg}	47.7	45.2	47.6	44.5	33.6	1.45
K_{14} + C	78.6	76.2	77.8	76.5	70.0	**0.58**
K_{14} +$V_{Spatial}$	83.5	**84.0**	86.3	85.8	79.0	1.14
K_{14} + V_{Res}	76.2	71.4	76.1	75.0	66.8	1.64
K_{14} + V_{Seg}	57.5	55.0	59.0	57.8	46.8	1.14
C + $V_{Spatial}$	**84.4**	83.4	**86.5**	**86.1**	**80.6**	1.24
K_{14}+ C + $V_{Spatial}$	**87.1**	**83.9**	**87.3**	**86.5**	**81.1**	1.32
K_{14} + V_{Seg} + C	71.3	69.3	73.4	72.3	65.4	1.33
K_{14} + V_{Seg} + C + $V_{Spatial}$	71.5	69.1	72.7	71.8	64.9	**1.31**

TABLE II: Performance of gesture prediction in the ablation study of input features for Suturing task: G = Surgical Gestures for the observation window. Other notations are the same as Table I.

Input Features	Ground truth Gestures			Recognized Gestures			Inference Time (ms)
	Accuracy (%)	Edit Score (%)	F1 @ 10,25,50 (%)	Accuracy (%)	Edit Score (%)	F1 @ 10,25,50 (%)	
K_{14} + G	85.4	87.0	81.3, 81.1, 78.5	80.1	82.8	78.1, 77.8, 77.1	**0.49**
K_{14} + G + C	88.8	90.4	85.0, 83.9, 77.8	85.5	87.2	84.7, **84.4**, **82.9**	0.89
K_{14} + G + $V_{Spatial}$	**89.5**	**91.3**	**87.8**, 84.4, 80.3	86.0	88.2	**86.2**, 84.2, 80.5	1.08
K_{14} + G + $V_{Spatial}$ + C	87.1	90.7	86.6, **85.3**, **82.6**	**86.5**	**89.8**	85.3, 82.3, 81.7	1.30
K_{14} + G + V_{Seg}	86.6	88.8	81.1, 78.3, 77.0	83.3	84.9	80.3, 78.2, 74.0	1.19
daVinciNet [37]	84.3	-	-	-	-	-	-
Transformer (MTMs) [15]	84.6	-	-	-	-	-	-
Transformer (PSMs) [15]	84.0	-	-	-	-	-	-

TABLE III: Performance of trajectory prediction in the end-to-end study of input features for Suturing task. RMSE and MAE are expressed in millimeters (mm):

Input Features	Metric	x1	y1	z1	x2	y2	z2
K_{14} + G	RMSE	5.3	4.5	5.9	6.16	6.37	6.74
	MAE	4.79	4.1	4.87	5.75	6.04	6.09
	MAPE	11	9.7	10.23	12.68	13.2	13.72
K_{14} + G + C	RMSE	5.11	4.32	5.8	5.34	5.56	5.65
	MAE	4.43	3.77	4.68	4.23	4.34	4.45
	MAPE	10.3	9.1	10.03	10.66	10.69	11.26
K_{14} + G + $V_{Spatial}$	RMSE	4.8	**4.09**	5.23	**4.57**	**4.44**	**4.37**
	MAE	4.13	**3.55**	4.19	**3.77**	**3.29**	**3.02**
	MAPE	7.6	**8.4**	9.45	**8.52**	**9.06**	9.74
K_{14} + G + $V_{Spatial}$ + C	RMSE	**4.75**	4.14	**5.17**	5.2	4.6	4.41
	MAE	**3.91**	3.6	**3.99**	3.8	3.73	4.1
	MAPE	**7.4**	8.7	**8.78**	9.35	9.2	**9.11**
K_{14} + G + V_{Seg}	RMSE	4.92	4.49	5.76	5.4	4.9	4.55
	MAE	4.3	4.02	4.61	4.16	3.96	4.12
	MAPE	8.84	9.5	9.7	9.45	9.79	9.36
daVinciNet [37]	RMSE	**2.53**	**1.89**	**2.96**	**3.15**	**3.5**	**3.91**
	MAE	**2.07**	**1.51**	**2.46**	**2.78**	**3.06**	3.50
	MAPE	**6.43**	**4.72**	**6.35**	**6.13**	**6.11**	**6.67**
Transformer [15]	RMSE	3.15	3.03	3.30	3.93	4.21	4.22
	MAE	2.86	2.85	3.00	3.60	3.88	3.84

TABLE IV: Gesture recognition accuracy compared with related work for Suturing task under LOUO cross-validation

	Data Sources	Trial (%)	1s Window (%)
Transformer [15]	PSM	-	89.2
Fusion-KVE [24]	PSM + Video	-	86.3
MA-TCN [25]	PSM + Video	83.4	-
TCN [20]	Video	81.4	-
Our model	PSM + Video + Context	**86.1**	87.3

features can lead to more precise and effective surgical gesture recognition with decreased computational overhead.

Notably, the method of feature extraction from video as well as the fusion of video and kinematic features affect gesture recognition performance as shown in Table I. $V_{Spatial}$ extracted using Spatial CNN [30] contributes to significant improvements compared to using features extracted using ResNet50 [29] due to the inherent capability of capturing spatio-temporal information from a video. Surgical context C improves performance by 5% due to the ability of capturing contextual features highly specific for robotic surgery [33] which signifies the temporal relationships of gesture sequences and transitions defined in a surgical process. We also observe that certain input feature combinations, such as K_{14} + V_{Seg} + C + $V_{Spatial}$, can result in overfitting and ultimately lead to decreased performance of the model. We also observe that the expertise level of the subjects and the different approaches taken by the subjects to perform the same task impacts the recognition accuracy. Moreover, certain expert subjects maintain similar movements over multiple trials and novice subjects exhibit inconsistent movements. This also impacts the model's gesture recognition performance.

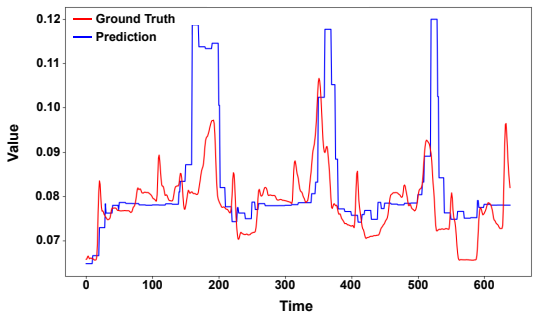
As shown in Table IV, our model achieves comparable performance to SOTA gesture recognition models by demonstrating an accuracy of **87.3%** for a 30-sample window (1s), while maintaining a low inference time of just **1.3ms**. This

enables real-time performance by meeting the constraint of performing inference faster than the input data acquisition rate of 30Hz (33.33ms) [59]. We note that the SOTA model presented by [15] attains a slightly higher accuracy of 89.3% for gesture recognition. The source code for [15] was not available, so we were unable to reproduce their results.

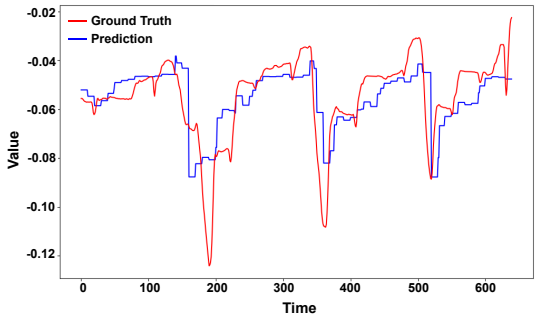
2) Gesture Prediction: Table II shows the results from our ablation and end-to-end experiments, which investigate the impact of different video feature representations on gesture prediction performance. When employing solely $K_{14} + G$ as features, our results show agreement with prior works [37], [15]. Notably, introducing context C as an additional feature led to major enhancements in performance, yielding an improvement of nearly 3% in accuracy and edit score. This underscores the significance of context features in enhancing the predictive capacity for future surgical activities. Furthermore, the inclusion of $V_{Spatial}$ as a video data representation results in substantial performance improvements, leading to the combination $K_{14}+G+V_{Spatial}$ outperforming the state-of-the-art with an accuracy of **89.5%** and an edit score of **91.3%**. Similar to gesture recognition, integrating both kinematic and video features yields superior prediction results. Although the utilization of V_{Seg} as a complement to kinematic features did yield improvements compared to baseline kinematic inputs, its impact was not as substantial as the spatio-temporal features or context variables. Additionally, the combination of three inputs $K_{14} + G + V_{Spatial} + C$ did not demonstrate significant performance enhancements and only increased the inference time of the prediction model. In summary, $K_{14} + G + V_{Spatial}$ is the most effective set of features, providing the best performance while maintaining reasonable inference times. Figure 2 shows the output of the gesture prediction over a sample suturing trial, using our top performing feature configuration.

End-to-end gesture prediction results are shown in Table II where the output of the recognition model is used instead of the ground truth gesture labels from the observation window. Our best model outperforms previous work by about 2.5% in accuracy, and using $K_{14}+G+V_{Spatial}+C$ seems to be more robust to inaccuracies in the gesture recognition outputs.

3) Trajectory Prediction and End-to-End Evaluation: Table III shows end-to-end results for our trajectory prediction module. We ran the full model without the use of ground truth observation gestures (similar to the second column of the Table II) and computed RMSE, MAE and MAPE metrics for Cartesian coordinates of each robot end-effector. In terms of the effectiveness of each input feature/video representation configuration in predicting accurate trajectories, the results are almost the same as gesture prediction. The $K_{14} + G + V_{Spatial} + C$ configuration still performs best on average, although $K_{14} + G + V_{Spatial} + C$ also appears to be comparable (e.g. for $x1$ and $z1$ coordinates). The inference times are also the same as gesture prediction since both predictions are generated simultaneously at the output of the transformer decoder. Figure 3 shows the predicted x and z trajectories for the right patient-side grasper during a trial using our top performing feature configuration. The Fundamentals of Laparoscopic Surgery guidelines [60] recommend a tool trajectory error of up to 1 mm in Suturing. However, this is not achieved by SOTA, including the daVinciNet model [37]. Our approach using $K_{14} + G + V_{Spatial} + C$ as input is not the most accurate, but it provides the best trade-off between accuracy and inference time. We observe performance fluctuations among different subjects, which appears to be a function of their level of expertise. The subjects with higher expertise have a better economy of motion and smoother trajectories than others. When learning on novices and intermediates and predicting for experts, we observe signs of over-compensation by the prediction model.



(a) PSMR Position X



(b) PSMR Position Z

Fig. 3: Trajectory Prediction results for X-axis and Z-axis position of the right instrument for a subject in the Suturing task

V. CONCLUSIONS

We presented a multimodal transformer architecture for real-time surgical gesture and trajectory prediction toward improving safety and autonomy in RMIS. This architecture outperforms the SOTA gesture prediction models by utilizing advanced video feature extraction techniques and achieves real-time performance by relying on a single transformer model. We evaluated the efficacy of multiple input feature configurations for both the recognition and prediction tasks and the end-to-end pipeline. Our results indicate that the fusion of kinematic data with spatial and contextual video features consistently yields the best performance. Future work will focus on validating our proposed method using data collected from a wider range of surgical tasks, participants with a variety of surgical skills, and actual surgical procedures and on applying it to real-time safety monitoring.

REFERENCES

[1] A. Kumar, N. Yadav, S. Singh, and N. Chauhan, "Minimally invasive (endoscopic-computer assisted) surgery: Technique and review," *Annals of maxillofacial surgery*, vol. 6, no. 2, p. 159, 2016.

[2] J. Finkelstein, E. Eckersberger, H. Sadri, S. S. Taneja, H. Lepor, and B. Djavan, "Open versus laparoscopic versus robot-assisted laparoscopic prostatectomy: the european and us experience," *Reviews in urology*, vol. 12, no. 1, p. 35, 2010.

[3] S. Bonne, W. Panitch, K. Dharmarajan, K. Srinivas, J.-L. Kincade, T. Low, B. Knoth, C. Cowan, D. Fer, B. Thananjeyan *et al.*, "A digital twin framework for telesurgery in the presence of varying network quality of service," in *2022 IEEE 18th international conference on automation science and engineering (CASE)*. IEEE, 2022, pp. 1325–1332.

[4] G. Gonzalez, M. Balakuntala, M. Agarwal, T. Low, B. Knoth, A. W. Kirkpatrick, J. McKee, G. Hager, V. Aggarwal, Y. Xue *et al.*, "Asap: A semi-autonomous precise system for telesurgery during communication delays," *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 1, pp. 66–78, 2023.

[5] J. Han, J. Davids, H. Ashrafian, A. Darzi, D. S. Elson, and M. Sodergren, "A systematic review of robotic surgery: From supervised paradigms to fully autonomous robotic approaches," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 18, no. 2, p. e2358, 2022.

[6] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G. D. Hager, and P. Jannin, "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, pp. 691–696, Sep. 2017. [Online]. Available: https://doi.org/10.1038/s41551-017-0132-7

[7] M. J. Fard, S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1850, 2018.

[8] C. E. Reiley, E. Plaku, and G. D. Hager, "Motion generation of robotic surgical tasks: Learning from expert demonstrations," in *2010 Annual international conference of the IEEE engineering in medicine and biology*. IEEE, 2010, pp. 967–970.

[9] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, vol. 13, pp. 731–739, 2018.

[10] M. S. Yasar, D. Evans, and H. Alemzadeh, "Context-aware monitoring in robotic surgery," in *2019 International symposium on medical robotics (ISMR)*. IEEE, 2019, pp. 1–7.

[11] M. S. Yasar and H. Alemzadeh, "Real-time context-aware detection of unsafe events in robot-assisted surgery," in *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2020, pp. 385–397.

[12] Z. Li, K. Hutchinson, and H. Alemzadeh, "Runtime detection of executional errors in robot-assisted surgery," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 3850–3856.

[13] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery: a review," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, 2021.

[14] K. Hutchinson, I. Reyes, Z. Li, and H. Alemzadeh, "Evaluating the task generalization of temporal convolutional networks for surgical gesture and motion recognition using kinematic data," *IEEE Robotics and Automation Letters*, 2023.

[15] C. Shi, Y. Zheng, and A. M. Fey, "Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8017–8024.

[16] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584.

[17] R. DiPietro, N. Ahmidi, A. Malpani, M. Waldram, G. I. Lee, M. R. Lee, S. S. Vedula, and G. D. Hager, "Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, no. 11, pp. 2005–2020, 2019.

[18] G. Menegozzo, D. Dall'Alba, C. Zandona, and P. Fiorini, "Surgical gesture recognition with time delay neural network based on kinematic data," in *2019 International symposium on medical robotics (ISMR)*. IEEE, 2019, pp. 1–7.

[19] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Multi-task recurrent neural network for surgical gesture recognition and progress prediction," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 1380–1386.

[20] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 47–54.

[21] J. Zhang, Y. Nie, Y. Lyu, H. Li, J. Chang, X. Yang, and J. J. Zhang, "Symmetric dilated convolution for surgical gesture recognition," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 409–418.

[22] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 467–475.

[23] D. Sarikaya and P. Jannin, "Surgical gesture recognition with optical flow only," *arXiv preprint arXiv:1904.01143*, 2019.

[24] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 371–377.

[25] B. Van Amsterdam, I. Funke, E. Edwards, S. Speidel, J. Collins, A. Sridhar, J. Kelly, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery with multimodal attention," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1677–1687, 2022.

[26] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, 2015.

[27] M. Hwang, J. Ichnowski, B. Thananjeyan, D. Seita, S. Paradis, D. Fer, T. Low, and K. Goldberg, "Automating surgical peg transfer: Calibration with deep learning can exceed speed, accuracy, and consistency of humans," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 2, pp. 909–922, Apr. 2023. [Online]. Available: https://doi.org/10.1109/tase.2022.3171795

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

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 36–52.

[31] K. Hutchinson, Z. Li, I. Reyes, and H. Alemzadeh, "Towards surgical context inference and translation to gestures," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6802–6809.

[32] D. Neumuth, F. Loebe, H. Herre, and T. Neumuth, "Modeling surgical processes: A four-level translational approach," *Artificial intelligence in medicine*, vol. 51, no. 3, pp. 147–161, 2011.

[33] K. Hutchinson, I. Reyes, Z. Li, and H. Alemzadeh, "Compass: a formal framework and aggregate dataset for generalized surgical procedure modeling," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–12, 2023.

[34] Z. Li, I. Reyes, and H. Alemzadeh, "Robotic scene segmentation with memory network for runtime surgical context inference," *arXiv preprint arXiv:2308.12789*, 2023.

[35] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, no. 3, 2014.

[36] S. DiMaio, M. Hanuschik, and U. Kreaden, "The da vinci surgical system," *Surgical robotics: systems applications and visions*, pp. 199–217, 2011.

[37] Y. Qin, S. F. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, "davincinet: Joint prediction of motion and surgical state in robot-assisted surgery," 2020.

[38] T. E. Murphy, "Towards objective surgical skill evaluation with hidden markov model-based motion recognition," 2004.

[39] I. Gurcan and H. Van Nguyen, "Surgical activities recognition using multi-scale recurrent networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2887–2891.

[40] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[41] Z. Wang, B. Zi, H. Ding, W. You, and L. Yu, "Hybrid grey prediction model-based autotracking algorithm for the laparoscopic visual window of surgical robot," *Mechanism and Machine Theory*, vol. 123, pp. 107–123, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094114X18300107

[42] Y. Sun, B. Pan, Y. Fu, and G. Niu, "Visual-based autonomous field of view control of laparoscope with safety-RCM constraints for semi-autonomous surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 2, Feb. 2020. [Online]. Available: https://doi.org/10.1002/rcs.2079

[43] C. Staub, C. Lenz, G. Panin, A. Knoll, and R. Bauernschmitt, "Contour-based surgical instrument tracking supported by kinematic prediction," in *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2010, pp. 746–752.

[44] M. M. Rahman, M. V. Balakuntala, G. Gonzalez, M. Agarwal, U. Kaur, V. L. N. Venkatesh, N. Sanchez-Tamayo, Y. Xue, R. M. Voyles, V. Aggarwal, and J. Wachs, "Sartres: a semi-autonomous robot teleoperation environment for surgery," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 4, pp. 376–383, 2021.

[45] S. Bonne, W. Panitch, K. Dharmarajan, K. Srinivas, J.-L. Kincade, T. Low, B. Knoth, C. Cowan, D. Fer, B. Thananjeyan, J. Kerr, J. Ichnowski, and K. Goldberg, "A digital twin framework for telesurgery in the presence of varying network quality of service," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 2022, pp. 1325–1332.

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

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[48] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," *arXiv preprint arXiv:2202.07125*, 2022.

[49] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 2114–2124.

[50] M. Liu, S. Ren, S. Ma, J. Jiao, Y. Chen, Z. Wang, and W. Song, "Gated transformer networks for multivariate time series classification," *arXiv preprint arXiv:2103.14438*, 2021.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[52] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[53] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[54] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[55] K. Patroumpas and T. Sellis, "Window specification over data streams," in *International Conference on Extending Database Technology*. Springer, 2006, pp. 445–464.

[56] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[58] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[59] G. Quellec, K. Charrière, M. Lamard, Z. Droueche, C. Roux, B. Cochener, and G. Cazuguel, "Real-time recognition of surgical tasks in eye surgery videos," *Medical image analysis*, vol. 18, no. 3, pp. 579–590, 2014.

[60] S. F. committee. (2019) Fundamentals of laparoscopic surgery. [Online]. Available: https://www.flsprogram.org/technical-skills-training-curriculum/