2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

MGRFormer: A Multimodal Transformer Approach for Surgical Gesture Recognition

Kevin Feghoul[1,2] Deise Santana Maia[2] Mehdi El Amrani[3] Mohamed Daoudi[2,4] Ali Amad[1]
[1] Univ. Lille , Inserm, CHU Lille, UMR-S1172 LilNCog, F-59000 Lille, France
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3] Department of Digestive Surgery and Transplantation, CHU Lille, PRESAGE, Univ. Lille, France
[4] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France

***Abstract*— Automatic surgical gesture recognition has the potential to revolutionize the field of surgery by enhancing patient care, surgical training, and our understanding of surgical skills. By integrating kinematic data, which precisely captures hand movements, with video data for contextual understanding, multimodal machine learning can greatly enhance the accuracy of surgical gesture recognition systems by capturing complementary knowledge. Recent research has highlighted the capabilities of Transformer-based models for temporal action segmentation. A key component of these models is the iterative refinement module, which enhances predictions using contextual data. In this study, we propose MGRFormer, a novel multimodal framework that leverages the interaction between kinematics and visual data at the refinement stage for the task of surgical gesture recognition. We evaluated our MGRFormer on the VTS dataset, and the results demonstrated that our approach outperformed unimodal and multimodal state-of-the-art methods by a large margin.**

I. Introduction

The field of surgery has undergone remarkable advancements in recent decades, driven by cutting-edge technologies and innovative techniques that have revolutionized patient care [7], [39]. As surgical procedures become increasingly intricate and precise, the demand for highly skilled surgeons is higher than ever. An essential component of surgical training involves the mastery of surgical gestures. Surgical residents are required to become proficient in these complex techniques to ensure the safety of their patients, achieve surgical accuracy and efficiency, and build their professional confidence.

The emerging field of automatic surgical gesture recognition holds significant promise for advancing surgical education. This technology aims to accurately classify and segment fine-grained surgical gestures automatically, thereby providing real-time feedback and objective assessments. To enhance the robustness of surgical gesture recognition methods, it is essential to consider multiple modalities, such as motion sensor and video data [30], [34], [40], which can capture distinct and complementary patterns. This integration of different data sources can provide a more comprehensive understanding of a surgeon's actions, allowing for the identification of correlations between hand movements, instrument motions, and visual cues in video data. Additionally, this multimodal approach can improve fault tolerance. For instance, unpredictable events such as occlusions can occur in the video, motion sensor data can act as a backup, ensuring the model's accurate performance even under challenging conditions.

However, accurately recognizing surgical gestures from multimodal data presents several challenges, including differences in representation and scale, synchronization issues, high dimensionality, and potential data loss due to sensor failures or occlusions. A key challenge is the effective integration of data from different modalities through multimodal machine learning. This involves deciding on the best stage for data fusion (early, late, or intermediate level) and developing efficient methods for this integration.

The Transformer architecture [41] has become the predominant choice for a wide range of tasks, including multimodal learning [1], [12], [18], as well as temporal action segmentation. Inspired by the success for the Transformer for temporal action segmentation and multimodal learning, we introduce MGRFormer, a novel attention-based multimodal framework for the task of surgical gesture recognition. This framework is designed to leverage the complementary information from kinematic and video modalities during the refinement stage. To the best of our knowledge, no prior work has explored multimodal fusion at this stage. To validate the effectiveness of our proposed approach, we conducted extensive experiments using the VTS dataset [15].

The contributions of this paper are three-fold and can be summarized as follows: (1) we propose a new multimodal fusion framework that leverages the joint relationship between kinematic and video modalities at the refinement stage; (2) to validate the proposed method and to further demonstrate the complementarity between the modalities, we provided a unimodal and multimodal benchmark; (3) our MGRFormer outperformed other state-of-the-art methods by a significant margin on the VTS dataset.

II. Related Work

A. Temporal Action Segmentation

Temporal action segmentation refers to the localization of individual actions within a video sequence. Traditional methods for identifying actions within video sequences typically involved using a sliding window approach, followed

The proposed work was supported by the French State, managed by the National Agency for Research (ANR) under the Investments for the Future program with reference ANR-16-IDEX-0004 ULNE, and with the support of PRESAGE Lille.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

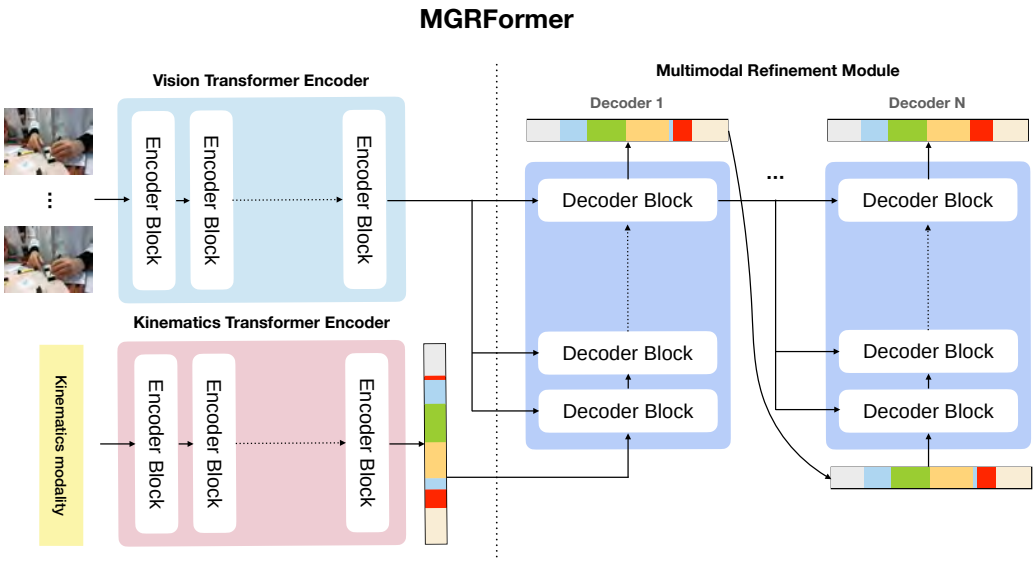


Fig. 1: Illustration of the MGRFormer framework, consisting of two Unimodal Encoders and a Multimodal Refinement Module for iterative cross-refinement using the output predictions of one modality and the Encoder features of the other modality.

by non-maximum suppression to select the most relevant candidates [21], [35]. Alternative approaches explore the use of Bayesian model [4], Conditional Random Fields [32], [38], and Markov models [23]. Modern approaches for modeling long-range dependencies among actions involve the use of deep neural networks, which encompass a variety of architectures such as Recurrent Neural Networks [8], [37], Temporal Convolutional Networks [11], [24], [27], [28], Graph Neural Networks [19], [45], and recent Transformers [9], [10], [42], [43]. Particularly, the ASFormer [43] has established itself as a state-of-the-art solution for temporal action segmentation. It employs a multi-stage process in which an initial stage generates the initial prediction, followed by subsequent refinement stages responsible for refining and fine-tuning the initial prediction. Despite the vast success of the Transformer for temporal segmentation, there have been relatively few attempts to extend its use for multimodal applications. In this study, we propose to exploit the complementary information during the refinement stage by leveraging cross-attention mechanisms between two modalities through the use of multiple Transformer decoders.

B. Surgical Gesture Recognition

Numerous studies proposed the used of kinematics and video data, either independently or in combination, for the task of surgical gesture recognition.

A rich variety of deep temporal models have been employed using kinematics data, including Convolutional Neural Networks [20], Temporal Convolutional Networks [15], [26], Recurrent Neural Networks [5], [6], [15], and Transformer [36]. On the other hand, additional methods have been developed solely using video data. These solutions involve employing 3D Convolutional Neural Networks [13], Symmetric Dilated Convolution [44], and Deep Reinforcement Learning [29].

Recent studies have explored multimodal learning strategies to enhance surgical gesture recognition systems. Some studies [25], [31], [34], [38] have reported consistent improvements when combining kinematics and video data compared to the two individual modalities. The integration of these two modalities has been investigated at the input level [25], [31], intermediate level [33], [40], and prediction level [34]. However, a very limited number of studies have explored more complex multimodal approaches. In their paper [34], the authors introduced Fusion-KVE, a novel approach that integrates visual features, kinematics data, and system events. This method employs individual networks for each input modality and then combines their predictions using a weighted voting scheme. Long et al. [30] proposed MRG-Net, an approach that leverages the complementary information between kinematics and visual features using a graph convolutional network. Van Amsterdam et al. [40] introduced MA-TCN, which utilizes multimodal attention mechanisms to weight kinematic and visual features. Unlike the aforementioned methods, we propose an iterative refinement module that leverages kinematics and video data in conjunction with a Transformer encoder. This enhances the contextual understanding of gesture predictions and substantially reduces over-segmentation errors.

C. RGB-D based Gesture Recognition

The field of multimodal gesture recognition is constantly evolving, with numerous studies exploring various modalities to enhance performance. The integration of RGB and depth data has been extensively investigated for its potential to significantly improve gesture recognition systems. This integration provides crucial spatial context by adding depth information to the RGB data, offering a more comprehen-

sive understanding of gestures. In their study, Hu et al. [17] introduced a novel deep bilinear framework designed to learn time-varying information from multimodal data. Furthermore, for capturing rich modality-temporal patterns, they proposed a novel action feature representation, which encodes the context of RGB-D actions into a tensor structure. Zhou et al. [47] introduced a novel spatial-temporal representation learning framework consisting of decoupled spatial and temporal representation learning networks, denoted as DSN and DTN, respectively, and a recoupling representation learning network denoted as RCM. To effectively exploit multimodal interactions between unimodal branches, they proposed a cross-modal adaptive posterior fusion module, termed CAPF. Furthermore, building upon the previously mentioned work, Zhou et al. [46] introduced a new video data augmentation technique, ShuffleMix, which mask randomly two video pairs along the temporal dimension and then mixes them. They also enhanced the RCM module with a multi-head mechanism that independently generates an attention map for each frame. Furthermore, they introduced a novel cross-modal Complement Feature Catcher (CFCer) for multimodal fusion, aimed at improving the results of late fusion.

III. Proposed Approach

In this study, we present the MGRFormer architecture to tackle the task of surgical gesture recognition. Fig. 1 shows an overview of the proposed framework, which is composed of three key components: (1) a Kinematics Transformer Encoder; (2) a Vision Transformer Encoder; and (3) a Multimodal Refinement Module. We first extract kinematic and visual features using the Kinematics Transformer Encoder and Vision Transformer Encoder, respectively. Both encoders are designed based on the ASFormer Encoder [43]. Next, the initial predictions from one modality and the extracted features from the other modality are passed through a series of successive decoders to perform an incremental cross-refinement. Our MGRFormer is designed to predict the probability distributions of surgical gestures for each time step.

A. Unimodal Feature Encoder

The first part of our framework consists of the Kinematics Transformer Encoder and the Vision Transformer Encoder, which extract kinematic and visual features. The Kinematics Transformer Encoder processes input kinematics data, denoted as x_{kin}, with dimensions $T \times d_{kin}$, while the Vision Transformer Encoder processes visual features, denoted as x_{vis}, with dimensions $T \times d_{vis}$. T represents the sequence length, and d_{kin} and d_{vis} represent the dimensions of kinematics and visual features, respectively. Regarding the visual features, we used either image features extracted from a pre-trained ResNet-18 [16] or frame-wise feature sequences extracted from pre-trained I3D [3]. The dimension of the ResNet-18 feature is 512-d, while the dimension of the I3D features is 1024-d. For the I3D feature, we added the RGB and flow predictions.

The initial stage of our framework involves linearly projecting each input feature, x_{kin} and x_{vis}, onto embedding vectors z_{kin} and z_{vis}, respectively. This projection is performed to adjust the dimensionality of the input features. Next, each embedding vector is fed to a series of encoder blocks. Finally, a fully-connected layer is used to generate initial predictions for either the kinematics or visual modality, which are denoted as $\hat{y}_{kin}$ or $\hat{y}_{vis}$. In our framework, it is important to note that only one modality is selected to generate initial predictions.

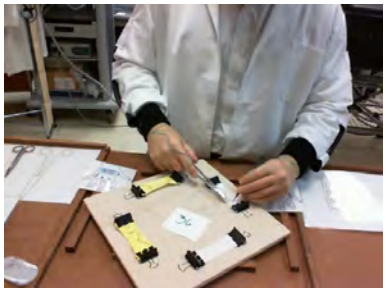
The encoder consists of a sequence of encoder blocks, with each encoder block comprising a temporal convolution layer followed by a single-head self-attention layer. A residual connection is applied around each of these two sublayers, followed by a ReLU activation function and instance normalization. The readers can refer to [43] for more details on the encoder part of the ASFormer.

Regarding the computation of the self-attention mechanism, we used a local window of size w, as proposed in [2]. This choice was motivated by the consideration that videos can be very long and can demand substantial computational resources for self-attention calculations. The size of the local window increases exponentially with the number of layers ($w = 2^i, i = 1, 2, \ldots$). This design allows for a transition from a local to a global focus, expanding the receptive field to effectively encompass the entire video sequence. Additionally, we double the dilation rate of the temporal convolution layer as the encoder depth increases, maintaining consistency with the self-attention layer.

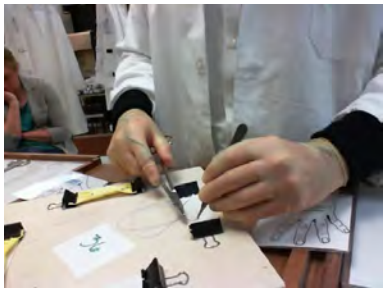
B. Multimodal Refinement Module

Iterative refinement is a crucial component of modern state-of-the-art methods for the task of temporal action segmentation. It enhances the initial predictions generated by the backbone encoder by incorporating higher-level contextual information that captures temporal relationships between surgical gestures. This contextual information aids in refining predictions, making them more coherent and consistent.

As previously mentioned in Sec. II, multimodal learning has the potential to enhance surgical gesture recognition by incorporating multiple modalities. However, traditional fusion techniques may not be suitable and could lead to suboptimal performance. For instance, early fusion, which combines modalities at the input level, may not effectively capture modality-specific patterns and can result in information loss due to incompatibilities in data scales, dynamics, or representations. As for late fusion, as the classical ASFormer generates multiple prediction outputs, there is no straightforward solution for effectively aggregating the outputs from different modalities. A simple solution will be to calculate either the mean or maximum between the different outputs of the different modalities, but this can lead to segmentation errors. To the best of our knowledge, no studies have explored multimodal fusion learning at the refinement stage. In this section, we present our proposed Multimodal Refinement Module, which incorporates Transformer decoders to exploit the complementary information between the kinematics and



(a) Frontal view



(b) Side view

Fig. 2: Suturing exercise on tissue extracted from the VTS dataset, seeing from two different views.

visual modalities during the refinement stage. We will first introduce the design of a single decoder and then extend it to multiple decoders for further iterative refinement.

The first decoder takes as input either the initial predictions from the kinematics or from the visual modality. Subsequently, a fully-connected layer is utilized to adjust the dimensions of these predictions. The decoder comprises a sequence of decoder blocks, each consisting of a temporal convolution layer and a cross-attention layer. Our approach involves computing cross-attention between the encoder features from the visual modality and the preceding decoder block's output responsible for refining kinematic initial predictions, and likewise, between the encoder features from the kinematics modality and the preceding decoder block's output for refining videos initial predictions.

Specifically, drawing inspiration from the decoder design in [43], we form the query Q and key K by concatenating the output from the encoder with the output from the previous decoder block. The value V, on the other hand, is solely derived from the output of the preceding decoder block. This cross-attention mechanism enables each position in one modality's encoder to attend to all positions in the refinement process of the other modality.

We are expanding from the use of a single decoder to the use of multiple decoders to perform further iterative refinement. Regarding each intermediate decoder, we compute cross-attention between the decoder features and the output predictions from the previous decoder.

We proposed different combinations for performing multimodal refinement using the kinematics (k) and video (v) modalities. We denote the process of refining our initial predictions derived from the kinematics modality using the encoder features from the video modality as MGRFormer$_{k\rightarrow v}$. Conversely, MGRFormer$_{v\rightarrow k}$ refers to the situation where we refine our initial predictions from the video modality with the encoder features from the kinematics modality. As we will demonstrate later, a double refinement process can further enhance predictions. For instance, we can refine the predictions from MGRFormer$_{k\rightarrow v}$ using the kinematics encoder features, resulting in MGRFormer$_{k\rightarrow v+k}$. In Sec. IV, we will report the performance for all possible combinations of double refinement between the kinematics and video modalities.

C. Loss Function

The loss function $\mathcal{L}$ is composed of two parts: a frame-wise classification loss and a smooth loss. The frame-wise classification loss is calculated as the negative log-likelihood of the correct class, and the smooth loss computes the squared error between the probabilities of successive frames. The loss function is defined as:

$$\mathcal{L} = \frac{1}{T}\sum_{t} -log(y_{t,\hat{c}}) + \lambda \frac{1}{TC}\sum_{t}\sum_{c}(y_{t-1,c} - y_{t,c})^2$$

Here, $y_{t,\hat{c}}$ denotes the predicted probability for the ground truth label $\hat{c}$ at time t. T represents the total number of points and C the number of distinct gestures. The term λ is fix at 0.60 in our experiments, balancing the classification loss and the smooth loss. The smooth loss aims to encourage consistency in the prediction probabilities between successive frames, which is particularly important for gesture recognition tasks. To train the model, we sum the losses associated with the predictions from both the encoder and decoders.

D. Implementation details

Both Transformer encoders and decoders are composed of 10 blocks each. Each input modality is processed by one Transformer encoder. For single refinement, we employed three Transformer decoders. In the case of double refinement, we further enhance the predictions using an additional decoder.

As mentioned in Section III-A, the input features x_{kin} and x_{vis} are projected onto the embedding vectors z_{kin} and z_{vis}, whose dimension was set to 128. Following the approach in [43], we applied dropout to the input features of the encoder with a rate of either 0.2 or 0.3, which was chosen through empirical experimentation. In all experiments, we trained our models using the Adam optimizer [22] with a learning rate of 0.0005.

IV. Experimental Results

A. Dataset

We conducted all our experiments using the Variable Tissue Simulation (VTS) dataset [14]. The dataset included twenty-four participants performing a suturing exercise on

Method	Modality	Features	Acc	F1-Macro	Edit	F1@10	F1@25	F1@50
LSTM [15]	kin	✗	81.26 ± 7.46	77.05 ± 9.26	84.69 ± 8.20	88.07 ± 7.33	83.69 ± 10.37	68.13 ± 18.74
GRU [15]	kin	✗	82.23 ± 7.25	78.20 ± 8.76	84.94 ± 7.70	88.01 ± 6.98	83.82 ± 10.17	68.86 ± 18.31
MS-TCN++ [15]	kin	✗	82.40 ± 6.97	78.92 ± 8.50	86.30 ± 8.42	89.30 ± 7.01	85.79 ± 9.82	71.12 ± 17.94
ASFormer [43]	kin	✗	82.66 ± 6.08	79.46 ± 7.18	88.65 ± 7.34	91.36 ± 5.73	87.68 ± 8.63	72.55 ± 16.51
MS-TCN++ [28]	frontal	ResNet-18	77.84 ± 8.31	73.34 ± 10.70	77.80 ± 10.88	81.36 ± 11.83	78.21 ± 13.93	63.87 ± 17.06
ASFormer [43]	frontal	ResNet-18	79.25 ± 8.21	75.20 ± 9.95	84.17 ± 8.27	87.16 ± 8.95	83.86 ± 11.40	69.00 ± 16.64
MS-TCN++ [28]	frontal	I3D	82.85 ± 5.86	78.85 ± 7.72	86.33 ± 7.24	89.98 ± 5.98	87.39 ± 7.96	74.33 ± 15.56
ASFormer [43]	frontal	I3D	82.72 ± 6.74	78.90 ± 8.25	88.28 ± 6.91	91.28 ± 5.09	88.35 ± 8.21	73.80 ± 17.53
MS-TCN++ [28]	side	ResNet-18	84.45 ± 5.97	81.71 ± 6.95	82.35 ± 10.19	87.01 ± 8.50	85.32 ± 9.51	77.01 ± 12.54
ASFormer [43]	side	ResNet-18	85.44 ± 4.92	82.87 ± 5.74	86.26 ± 8.53	90.44 ± 6.19	88.98 ± 7.00	80.41 ± 11.24
MS-TCN++ [28]	side	I3D	86.83 ± 4.81	84.14 ± 6.10	86.68 ± 8.62	90.85 ± 6.56	89.83 ± 7.82	82.48 ± 11.78
ASFormer [43]	side	I3D	87.43 ± 4.59	85.29 ± 4.91	89.24 ± 7.32	92.89 ± 5.05	91.61 ± 6.33	85.05 ± 10.43

TABLE I: Unimodal surgical gesture recognition

two distinct tissue simulators, each representing different materials: tissue paper to simulate friable tissue, and rubber balloons to mimic arterial conditions. Each participant completed the task twice with both materials. The study involved eleven medical students, one resident, and thirteen attending surgeons. One surgeon, who happened to be left-handed, was excluded from the study. A total of 96 procedures were carried out, each with a duration ranging from 2 to 6 minutes.

Kinematics data for both hands were captured using electromagnetic motion sensors, while video data were concurrently recorded by two cameras: one frontal camera focused on the simulation material, and another wide-angle camera encompassing the surrounding area, as we can see in Fig. 2. The sensors and both cameras were synchronized to ensure simultaneous recording.

The suturing exercise were segmented into six gestures: ”pass the needle through the material”, ”pull the suture”, ”perform an instrumental tie”, ”lay the knot”, ”cut the suture”, and ”the background gesture”.

B. Evaluation metrics

We evaluated our approach for surgical gesture recognition using two types of metrics: frame-wise and segmentation. Frame-wise metrics include accuracy (ratio of correctly classified gestures) and macro F1-score (average F1-scores for each gesture class, treating all classes equally). Segmentation metrics used are the segmental edit score, and segmental F1 score (F1@k) with thresholds at 10%, 25%, and 50%, assessing the overlap between predicted and actual gesture segments.

C. Evaluation framework

Following prior works [15], we employed a subject-independent 5-fold cross-validation strategy to train all our models. In each iteration, we divided our dataset into training, validation, and test sets, following the procedure outlined in [15]. For each evaluation metric, we reported the mean and standard deviation across all the folds.

D. Results

We conducted unimodal and multimodal surgical gesture recognition using the VTS dataset, employing kinematic and video modalities. For the frontal and side view videos, we employed ResNet-18 and I3D extracted features in order to evaluate the effectiveness of our proposed method in handling both image and video features.

1) Unimodal: In Table I, we present the performance of the ASFormer model, alongside results from several state-of-the-art methods, across three modalities: kinematics, frontal-view, and side-view video. The ASFormer consistently outperforms other methods across all input modalities. Regarding the kinematics modality, we observed significant improvements of at least 0.54%, 2.35%, and 2.06% in terms of macro F1 score, Edit score, and F1@10, respectively. For both frontal and side view modalities with ResNet-18 features, the ASFormer consistently outperformed the MS-TCN++ across all types of extracted features. Specifically, for the side view modality, the ASFormer surpassed the MS-TCN++ by 1.16%, 3.91%, and 3.43% in terms of the macro F1 score, Edit score, and F1@10, respectively.

The ASFormer exhibits the best performances for all evaluation metrics by using the side view modality combined with I3D features. Conversely, the ASFormer shows the poorest performance when employing the frontal view modality with ResNet-18 features. Furthermore, for both the frontal and side view modalities, we observe that using I3D features yields superior performance compared to ResNet-18 features. This enhancement can be attributed to the fact that I3D features are better suited for capturing temporal correlations among adjacent frames. In contrast, ResNet-18 features are extracted from individual images, neglecting the contextual information provided by neighboring frames.

2) Multimodal: We present the results regarding the fusion of the kinematics and video modalities in Tables II, III, V, and VI. We benchmarked our method against several state-of-the-art multimodal methods that integrate kinematics

Method	Acc	F1-Macro	Edit	F1@10	F1@25	F1@50
Fusion-KV [34]	81.94 ± 6.95	77.28 ± 9.14	83.33 ± 9.49	87.21 ± 8.09	83.18 ± 11.15	68.25 ± 18.54
MGR-Net [30]	77.70 ± 5.92	73.87 ± 7.15	81.49 ± 9.41	85.08 ± 7.70	80.64 ± 9.83	62.17 ± 15.16
MA-TCN [40]	79.91 ± 6.23	75.64 ± 8.17	82.02 ± 9.38	86.21 ± 8.03	82.32 ± 10.57	66.38 ± 16.78
MS-TCN++ (early)	82.01 ± 6.83	79.07 ± 7.96	82.54 ± 10.29	86.65 ± 9.01	83.97 ± 10.84	71.26 ± 17.24
MS-TCN++ (late)	82.77 ± 6.89	79.56 ± 8.38	86.69 ± 9.23	89.32 ± 7.39	85.52 ± 10.71	70.84 ± 17.97
ASFormer (early)	81.15 ± 6.68	77.35 ± 7.89	85.66 ± 7.68	88.59 ± 7.03	86.01 ± 8.97	72.25 ± 15.05
ASFormer (late)	81.85 ± 5.59	77.82 ± 7.13	84.04 ± 8.43	88.12 ± 7.10	85.02 ± 9.38	71.58 ± 14.62
MGRFormer $v \rightarrow k$	82.80 ± 6.15	79.29 ± 7.36	88.06 ± 7.91	91.55 ± 6.02	88.50 ± 8.52	73.61 ± 16.15
MGRFormer $k \rightarrow v$	83.85 ± 5.83	80.35 ± 7.43	88.34 ± 7.79	91.28 ± 6.01	88.12 ± 8.59	74.71 ± 15.49
MGRFormer $v \rightarrow v+k$	80.66 ± 6.44	76.69 ± 8.16	84.67 ± 8.57	87.93 ± 7.79	85.31 ± 9.09	70.39 ± 15.44
MGRFormer $k \rightarrow k+v$	83.81 ± 5.36	80.47 ± 6.39	88.22 ± 7.38	91.81 ± 5.51	89.14 ± 7.95	76.28 ± 15.41
MGRFormer $v \rightarrow k+v$	82.07 ± 6.32	78.46 ± 7.82	87.22 ± 7.35	90.40 ± 6.34	87.43 ± 9.41	73.40 ± 16.76
MGRFormer $k \rightarrow v+k$	**84.05 ± 5.56**	**80.66 ± 6.53**	**89.14 ± 7.69**	**92.30 ± 5.50**	**89.80 ± 7.95**	**76.40 ± 16.70**

TABLE II: Multimodal surgical gesture recognition: kinematics + frontal view (ResNet-18 features). Regarding the notation for MGRFormer, the prediction derived from the modality on the left side of the arrow is refined using the modalities on the right side. For instance, MGRFormer$_{k \rightarrow v+k}$ denotes the process where the kinematics prediction is first refined with video features, followed by a subsequent refinement using kinematics features.

Method	Acc	F1-Macro	Edit	F1@10	F1@25	F1@50
Fusion-KV [34]	81.82 ± 6.69	77.70 ± 8.30	84.42 ± 9.87	87.62 ± 8.18	83.32 ± 10.56	68.69 ± 17.77
MGR-Net [30]	78.88 ± 4.91	75.56 ± 5.72	81.63 ± 8.22	85.93 ± 6.85	82.62 ± 8.65	64.79 ± 13.78
MA-TCN [40]	83.15 ± 5.50	80.04 ± 6.52	84.50 ± 9.05	88.38 ± 7.24	85.98 ± 8.42	73.33 ± 14.94
MS-TCN++ (early)	85.17 ± 5.63	83.21 ± 6.64	84.77 ± 9.60	89.22 ± 7.51	88.01 ± 8.53	80.05 ± 12.35
MS-TCN++ (late)	86.81 ± 5.07	83.90 ± 6.91	82.83 ± 10.28	88.00 ± 7.92	86.20 ± 9.21	78.35 ± 12.85
ASFormer (early)	85.76 ± 4.68	83.42 ± 5.42	86.93 ± 6.72	90.76 ± 5.76	89.26 ± 6.78	80.80 ± 10.67
ASFormer (late)	85.53 ± 4.43	83.02 ± 5.12	85.69 ± 8.01	89.67 ± 6.58	88.10 ± 7.45	80.11 ± 11.20
MGRFormer $v \rightarrow k$	85.95 ± 4.32	83.47 ± 4.84	89.24 ± 6.52	92.78 ± 4.81	91.16 ± 6.06	81.58 ± 11.38
MGRFormer $k \rightarrow v$	87.40 ± 4.03	85.17 ± 4.42	89.53 ± 6.35	93.08 ± 4.28	**91.78 ± 5.16**	84.02 ± 9.51
MGRFormer $v \rightarrow v+k$	84.97 ± 4.63	82.16 ± 5.36	86.51 ± 7.94	90.19 ± 6.22	88.81 ± 7.14	79.62 ± 11.32
MGRFormer $k \rightarrow k+v$	86.75 ± 4.17	84.34 ± 5.01	88.58 ± 6.37	91.91 ± 4.47	90.37 ± 5.67	82.68 ± 9.29
MGRFormer $v \rightarrow k+v$	84.81 ± 4.62	82.16 ± 5.15	86.70 ± 6.92	90.43 ± 5.57	88.98 ± 6.54	80.17 ± 10.48
MGRFormer $k \rightarrow v+k$	**87.61 ± 3.75**	**85.47 ± 4.06**	**89.74 ± 6.23**	**93.40 ± 4.22**	**91.77 ± 5.01**	**85.12 ± 8.73**

TABLE III: Multimodal surgical gesture recognition: kinematics + side view (ResNet-18 features)

Method	Acc	F1-Macro	Edit	F1@{10,25,50}		
Vision Encoder	85.49	83.02	81.48	86.73	85.00	76.34
Kinematics Encoder	83.64	80.36	83.09	87.41	83.62	69.12
One Decoder	86.09	83.47	86.54	90.09	88.88	80.36
Two Decoders	86.15	83.75	87.91	91.51	89.97	82.19
Three Decoders (ours)	**87.40**	**85.17**	**89.53**	**93.08**	**91.78**	**84.02**
Four Decoders	85.26	82.71	87.52	91.14	89.79	81.19

TABLE IV: Comparative results from varying the number of decoders in MGRFormer$_{k \rightarrow v}$, using kinematics data and side view video with ResNet-18 features. The performance when using only vision and kinematics encoders is also included.

with frontal and side view videos, using ResNet-18 features. These techniques include Fusion-KV [34], MGR-Net [30], and MA-TCN [40]. Specifically for MGR-Net, we re-implemented the entire framework excluding the LSTM module, as its inclusion leads to lower performance. Our comparison also featured MS-TCN++ [28], a state-of-the-art approach in action segmentation, which employs an iterative refinement. It should be noted that this particular refinement is different from the one presented in our work. Furthermore, we tested MS-TCN++ under two classical multimodal fusion settings: early and late fusion. The results of these comparisons are detailed in Tables II and III. Our MGRFormer outperformed all the aforementioned state-of-the-art methods by a large margin in merging kinematics

Method	Acc	F1-Macro	Edit	F1@10	F1@25	F1@50
ASFormer (early)	83.62 ± 6.24	**80.15 ± 7.91**	88.09 ± 7.23	91.66 ± 5.32	89.32 ± 7.28	**76.86 ± 15.46**
ASFormer (late)	**83.73 ± 5.98**	80.13 ± 7.48	86.73 ± 7.50	90.20 ± 5.92	87.61 ± 8.69	74.92 ± 16.09
MGRFormer $v \rightarrow k$	82.74 ± 6.41	79.21 ± 8.28	88.00 ± 8.04	91.30 ± 6.19	88.69 ± 8.54	74.92 ± 16.10
MGRFormer $k \rightarrow v$	83.12 ± 5.69	79.77 ± 6.88	**89.53 ± 7.41**	**92.44 ± 5.47**	**89.48 ± 8.07**	74.79 ± 15.40
MGRFormer $v \rightarrow v + k$	82.60 ± 6.22	79.07 ± 7.87	87.85 ± 7.64	91.20 ± 6.24	88.41 ± 9.15	75.45 ± 15.11
MGRFormer $k \rightarrow k + v$	83.21 ± 5.89	79.29 ± 6.98	87.47 ± 8.25	90.90 ± 6.41	87.86 ± 9.37	74.09 ± 15.88
MGRFormer $v \rightarrow k + v$	83.12 ± 6.01	79.57 ± 7.38	88.60 ± 6.67	91.79 ± 5.36	89.05 ± 8.21	75.81 ± 14.95
MGRFormer $k \rightarrow v + k$	82.95 ± 5.58	79.44 ± 6.73	88.36 ± 7.80	92.01 ± 5.67	89.30 ± 8.44	74.14 ± 15.79

TABLE V: Multimodal surgical gesture recognition: kinematics + frontal view (I3D features)

Method	Acc	F1-Macro	Edit	F1@10	F1@25	F1@50
ASFormer (early)	87.62 ± 4.44	85.20 ± 4.85	88.55 ± 7.73	92.23 ± 5.61	91.16 ± 6.79	84.09 ± 10.57
ASFormer (late)	87.68 ± 4.06	85.13 ± 4.61	86.62 ± 7.89	90.83 ± 6.17	89.69 ± 6.85	82.89 ± 10.40
MGRFormer $v \rightarrow k$	86.90 ± 5.38	84.57 ± 5.76	88.26 ± 7.98	91.91 ± 6.15	90.82 ± 7.16	83.86 ± 10.02
MGRFormer $k \rightarrow v$	**88.39 ± 3.94**	**86.03 ± 4.59**	89.55 ± 7.68	93.46 ± 4.97	92.38 ± 5.95	**86.29 ± 9.92**
MGRFormer $v \rightarrow v + k$	87.24 ± 4.70	84.47 ± 5.20	89.11 ± 8.11	92.36 ± 5.94	91.23 ± 7.34	84.47 ± 10.07
MGRFormer $k \rightarrow k + v$	87.47 ± 4.25	85.31 ± 4.72	87.81 ± 7.81	91.85 ± 5.68	90.32 ± 6.86	83.46 ± 10.72
MGRFormer $v \rightarrow k + v$	87.44 ± 4.73	85.09 ± 5.30	89.54 ± 6.68	92.61 ± 5.54	91.51 ± 6.72	84.93 ± 9.89
MGRFormer $k \rightarrow v + k$	88.10 ± 3.80	85.89 ± 4.27	**89.91 ± 7.42**	**93.51 ± 5.00**	**92.40 ± 6.00**	85.66 ± 9.63

TABLE VI: Multimodal surgical gesture recognition: kinematics + side view (I3D features)

with both video perspectives. Specifically, for the side view modality, MGRFormer$_{k \rightarrow v+k}$ exceeded the performance of Fusion-KV, MGR-Net, and MA-TCN by minimum margins of 5.43%, 5.24%, and 5.02%, respectively, in terms of macro F1-score, Edit score, and F1@10. It also surpassed both the early and late fusion variants of MS-TCN++, but to a lesser extent. We observed that both multimodal versions of MS-TCN++ outperformed the other three baseline models. This enhancement is likely due to MS-TCN++'s iterative refinement module, which boosts the network's accuracy by repeatedly refining gesture segment predictions.

To demonstrate the effectiveness of the multimodal refinement module, we conducted an ablation study on the number of decoders in MGRFormer$_{k \rightarrow v}$, where we fused kinematics and side view video with ResNet-18 features. As shown in Table IV, it was found that selecting three decoders for iterative cross-refinement yielded the best performance across all metrics. It was observed that adding another decoder beyond three did not lead to further improvement, while it did add more complexity to the overall model. Furthermore, we can observe that using at least one decoder significantly improves performance compared to both the vision and kinematics encoders, which demonstrates the utility of the cross-refinement module.

The MGRFormer consistently outperformed each modality when used individually. When integrating kinematics data and frontal view video features extracted using ResNet-18, the MGRFormer$_{k \rightarrow v+k}$ model significantly outperformed each input modality when used separately, as demonstrated in Table II. We observed enhancements of 1.20%, 0.49%, and 0.94% in terms of macro F1 score, Edit score, and F1@10, compared to the unimodal ASFormer trained on the kinematics data. Similarly, improvements of 5.46%, 4.97%, and 5.14% were noted in comparison to the ASFormer trained the ResNet-18 features extracted from the frontal view modality. As for the fusion of kinematics data and the side view video with ResNet-18 extracted features, we observed significant improvements of at least 2.60% in macro F1 score, 1.09% in Edit score, and 2.04% in F1@10, compared to the best results obtained from each of the two individual modalities (see Table III). Similar improvements were observed with I3D extracted features, as shown in Tables V and VI.

When comparing the results of the ResNet-18 and I3D features in combination with the kinematics modality, the MGRFormer exhibits slightly superior performance when utilizing the I3D features in regard of the side view modality (see Table. III and VI). However, the opposite effect can be observed when employing the frontal view modality (see Table. II and V).

To demonstrate the relevance of our proposed MGRFormer architecture in comparison to conventional fusion techniques, we conducted a comparative analysis with traditional multimodal fusion methods, specifically early fusion and late fusion. More precisely, ASFormer (early) concatenates the

kinematics and video modalities at the input level, while ASFormer (late) adds the predictions from both the encoders and decoders of the different modalities. For this particular case, it is worth noting that both ASFormer instances for each input modality must have the same number of encoders and decoders to add the predictions from both modalities of the same stage. Combining kinematics and side view modalities with ResNet-18 features resulted in significant improvements compared to ASFormer (late), we achieved a 2.45% increase in F1 macro score, a 4.05% improvement in the Edit score, and a 3.73% enhancement in F1@10 (see Table III) with MGRFormer$_{k\rightarrow v+k}$. Similarly, when compared to ASFormer (early), we observed an increase of 2.05%, 2.81%, 2.64%, respectively in terms of F1 macro score, Edit score, and F1@10. As for the frontal view modality, as we can see in Table. II, we can observe an improvement of 2.84%, 5.10%, and 4.18% in terms of F1 macro score, Edit score, and F1@10, respectively, in favor of MGRFormer$_{k\rightarrow v+k}$, compared to the ASFormer (late). Additionally, we observed improvements of 3.31%, 3.48%, and 3.71%, respectively, compared to the ASFormer (early). Regarding the use of I3D features with the side view modality, we noticed an improvement compared to both baselines, but to a lesser extent, as depicted in Table VI. However, when considering the frontal view modality with I3D features, Table V shows that MGRFormer$_{k\rightarrow v+k}$ exhibits only marginal improvements in terms of Edit score, F1@10, and F1@25 when compared to ASFormer (early) and ASFormer (late). These results suggest the superiority of our proposed method compared to traditional multimodal approaches. Despite the significant performance gain we achieved, in terms of complexity, our MGRFormer, with a single cross-refinement stage, is comparable to the ASFormer (early) since both models have the same number of decoders. MGRFormer with the double cross-refinement adds only one additional decoder, which is a reasonable increase in complexity and is only slightly more complex than the single cross-refinement stage. In contrast, when comparing our proposed approach to ASFormer (late), our method involves half the number of encoders and decoders. ASFormer (late) requires training two separate models, each with a single encoder and three decoders.

Finally, regarding the different settings associated to our MGRFormer, we can observe that the one-stage refinement MGRFormer$_{k\rightarrow v}$ outperforms MGRFormer$_{v\rightarrow k}$ for each combination of kinematics data and both video views with respect to the ResNet-18 and I3D features, as we can see in Table II, III, V, and VI. For instance, we can see in Table VI that MGRFormer$_{k\rightarrow v}$ outperforms MGRFormer$_{v\rightarrow k}$ in terms of F1 macro score, Edit score, and F1@10 by 1.46%, 1.29%, and 1.55%, respectively. These findings underscore the superior performance of our framework in leveraging video encoder features for refining initial kinematics predictions over the inverse approach. The advantage of MGRFormer$_{k\rightarrow v}$ can be attributed to the richer spatiotemporal context provided by video data, which is critical for iterative refinement. Surgical gestures, characterized by intricate, fine-grained movements and interactions with various tools and tissues, are more discernible in video data. This modality not only captures the detailed visual context of the surgical site, including tool-tissue interactions and surgeon hand movements, but also the subtleties necessary for accurately identifying gestures. Conversely, while kinematic data is valuable, it lacks the visual nuances essential for distinguishing between closely related gestures, focusing instead on motion trajectories. This is further supported by the findings in Table IV, where training the Transformer encoder with side view video data outperformed kinematic data across several metrics, including accuracy and macro F1 scores, as well as F1@25 and F1@50, demonstrating the video data's superior contextual robustness for gesture segmentation. On the other hand, kinematics data, which surpassed the side view video in terms of Edit score and F1@10 metrics, can also enhance video predictions through our proposed multimodal refinement module, albeit to a lesser extent than by fusing kinematic predictions with video features. These improvements can be attributed to the fact that kinematic data offers precise temporal information about tool movements, leading to enhanced edit scores and F1@10 metrics.

Furthermore, we reported results for all possible combinations of double refinements involving kinematics and video modalities. As shown in Tables II, III, and VI, MGRFormer$_{k\rightarrow v+k}$ outperformed the other combinations in each evaluation metric. This outcome is not surprising, as MGRFormer$_{k\rightarrow v}$ had previously demonstrated the best results for one-stage refinement. In comparing single and double refinement, we observed that the double refinement with the MGRFormer$_{k\rightarrow v+k}$ setting works better than MGRFormer$_{k\rightarrow v}$ when fusing kinematics and video data from both views with ResNet-18 features, as shown in Tables II and III. Specifically, for the I3D features, MGRFormer$_{k\rightarrow v+k}$ yields better results than MGRFormer$_{k\rightarrow v}$ in terms of Edit score, F1@10, and F1@50 when fusing kinematics and side view video, as depicted in Table VI. In contrast, when fusing with frontal view video, MGRFormer$_{k\rightarrow v}$ surpasses MGRFormer$_{k\rightarrow v+k}$ across all six evaluation metrics (see Table V).

V. Conclusion

This paper presents a new multimodal fusion framework called MGRFormer, which involves iterative cross-refinement between the output predictions of one modality and the encoder features of the other modality. This approach permits leveraging the complementary information between the kinematics and video modalities at the refinement stage. The effectiveness of our approach has been validated on the VTS dataset, where our MGRFormer outperformed traditional multimodal fusion techniques by a large margin. Additionally, combining kinematics and video data with our approach consistently led to performance improvements compared to the two modalities individually. In future work, we plan to extend our framework to more than two modalities.

REFERENCES

[1] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

[2] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Y. Cheng, Q. Fan, S. Pankanti, and A. Choudhary. Temporal sequence modeling for video event detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2227–2234, 2014.

[5] 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*, 14(11):2005–2020, 2019.

[6] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*, pages 551–558. Springer, 2016.

[7] R. W. Dobbs, W. R. Halgrimson, S. Talamini, H. T. Vigneswaran, J. O. Wilson, and S. Crivellaro. Single-port robotic surgery: the next generation of minimally invasive urology. *World journal of urology*, 38:897–905, 2020.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[9] D. Du, B. Su, Y. Li, Z. Qi, L. Si, and Y. Shan. Do we really need temporal convolutions in action segmentation? In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1014–1019. IEEE, 2023.

[10] Z. Du and Q. Wang. Dilated transformer with feature aggregation module for action segmentation. *Neural Processing Letters*, pages 1–17, 2022.

[11] 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*, pages 3575–3584, 2019.

[12] K. Feghoul, D. S. Maia, M. Daoudi, and A. Amad. Mmgt: Multimodal graph-based transformer for pain detection. In *31st European Signal Processing Conference (EUSIPCO 2023)*, 2023.

[13] 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*, pages 467–475. Springer, 2019.

[14] A. Goldbraikh, A.-L. D'Angelo, C. M. Pugh, and S. Laufer. Video-based fully automatic assessment of open surgery suturing skills. *International Journal of Computer Assisted Radiology and Surgery*, 17(3):437–448, 2022.

[15] A. Goldbraikh, T. Volk, C. M. Pugh, and S. Laufer. Using open surgery simulation kinematic data for tool and gesture recognition. *International Journal of Computer Assisted Radiology and Surgery*, 17(6):965–979, 2022.

[16] 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*, pages 770–778, 2016.

[17] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 335–351, 2018.

[18] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020.

[19] Y. Huang, Y. Sugano, and Y. Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14024–14034, 2020.

[20] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14:1611–1617, 2019.

[21] S. Karaman, L. Seidenari, and A. Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, volume 1, page 5, 2014.

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

[23] H. Kuehne, J. Gall, and T. Serre. An end-to-end generative framework for video segmentation and recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[24] 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*, pages 156–165, 2017.

[25] C. Lea, G. D. Hager, and R. Vidal. An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In *2015 IEEE winter conference on applications of computer vision*, pages 1123–1129. IEEE, 2015.

[26] 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*, pages 47–54. Springer, 2016.

[27] P. Lei and S. Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6742–6751, 2018.

[28] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, and J. Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[29] D. Liu and T. Jiang. Deep reinforcement learning for surgical gesture segmentation and classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 247–255. Springer, 2018.

[30] Y. Long, J. Y. Wu, B. Lu, Y. Jin, M. Unberath, Y.-H. Liu, P. A. Heng, and Q. Dou. Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13346–13353. IEEE, 2021.

[31] A. Murali, A. Garg, S. Krishnan, F. T. Pokorny, P. Abbeel, T. Darrell, and K. Goldberg. Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 4150–4157. IEEE, 2016.

[32] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 612–619, 2014.

[33] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. 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.

[34] 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)*, pages 371–377. IEEE, 2020.

[35] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE, 2012.

[36] 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)*, pages 8017–8024. IEEE, 2022.

[37] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016.

[38] L. Tao, L. Zappella, G. D. Hager, and R. Vidal. Surgical gesture segmentation and recognition. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part III 16*, pages 339–346. Springer, 2013.

[39] M. Tonutti, D. S. Elson, G.-Z. Yang, A. W. Darzi, and M. H. Sodergren. The role of technology in minimally invasive surgery: state

of the art, recent developments and future directions. *Postgraduate medical journal*, 93(1097):159–167, 2017.

[40] 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*, 41(7):1677–1687, 2022.

[41] 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*, 30, 2017.

[42] J. Wang, Z. Wang, S. Zhuang, Y. Hao, and H. Wang. Cross-enhancement transformer for action segmentation. *Multimedia Tools and Applications*, pages 1–14, 2023.

[43] F. Yi, H. Wen, and T. Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

[44] 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*, pages 409–418. Springer, 2020.

[45] J. Zhang, P.-H. Tsai, and M.-H. Tsai. Semantic2graph: Graph-based multi-modal feature for action segmentation in videos. *arXiv preprint arXiv:2209.05653*, 2022.

[46] B. Zhou, P. Wang, J. Wan, Y. Liang, and F. Wang. A unified multi-modal de-and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[47] B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, D. Zhang, Z. Lei, H. Li, and R. Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022.