Keep Your Eye on the Best: Contrastive Regression Transformer for Skill Assessment in Robotic Surgery

Dimitrios Anastasiou, Yueming Jin, Danail Stoyanov, and Evangelos Mazomenos

***Abstract*—This letter proposes a novel video-based, contrastive regression architecture, Contra-Sformer, for automated surgical skill assessment in robot-assisted surgery. The proposed framework is structured to capture the differences in the surgical performance, between a test video and a reference video which represents optimal surgical execution. A feature extractor combining a spatial component (ResNet-18), supervised on frame-level with gesture labels, and a temporal component (TCN), generates spatio-temporal feature matrices of the test and reference videos. These are then fed into an action-aware Transformer with multi-head attention that produces inter-video contrastive features at frame level, representative of the skill similarity/deviation between the two videos. Moments of sub-optimal performance can be identified and temporally localized in the obtained feature vectors, which are ultimately used to regress the manually assigned skill scores. Validated on the JIGSAWS dataset, Contra-Sformer achieves competitive performance (Spearman 0.65 - 0.89), with a normalized mean absolute error between 5.8% - 13.4% on all tasks and across validation setups. Source code and models are available at *https://github.com/anastadimi/Contra-Sformer.git*.**

***Index Terms*—Computer Vision for Medical Robotics, Deep Learning Methods, Surgical Skill Assessment, Contrastive Regression.**

I. INTRODUCTION

ROBOT-ASSISTED minimally invasive surgery (RMIS) is firmly established in clinical practice, offering enhanced visualization and manipulability compared to standard laparoscopy [1]. Operative performance assessment is a fundamental element of surgical education and practice, and similar to other surgical specialties, significant efforts have been devoted towards standardized objective skill assessment systems for RMIS [2]. Global rating scales (GRS) such as the Objective Structured Assessment of Technical Skills (OSATS) are established assessment tools. The OSATS comprise a list of core procedural components (*e.g.*, handling of instruments, and respect for tissue), assessed and scored on a Likert-style (typically 5-point) scale. Summing the individual GRS components produces an overall performance score (7-35 for OSATS [3]). Each component is assigned a score based on performance characteristics. For example, "time and motion" is scored with 5 when there is economy of movement, maximum efficiency and optimal outcome [3]. Nevertheless, RMIS evaluation with GRS is time-consuming, laborious and inherently subjective as different evaluators may assess GRS items differently. To address these limitations, several works have developed computational methods that evaluate surgical execution by processing intraoperative information (*e.g.*, surgical video and robot kinematics) [4].

Automated surgical skill assessment in RMIS can have a profound impact, streamlining the evaluation process, overcoming the need for manual assessment, and eliminating subjectivity [5]. Modeling optimal surgical execution can also introduce performance awareness in the design of actuation and control policies towards automation of surgical tasks where robotic systems mimic the performance of expert surgeons [1], [4]. The release of the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [6] containing synchronized kinematics and video captured using the da Vinci Surgical System, alongside atomic gestures and Global Rating Scale (GRS) score annotations (range of 6-30), provided the first structured benchmark to support activity in this space. Several methods for GRS estimation have been developed and validated on JIGSAWS.

Initial works utilized kinematic data to regress the GRS score by exploring different types of holistic features [5], and temporal convolutional neural networks (TCNs) [7]. Reported outcomes indicate that modeling surgical skills to regress GRS scores only using kinematic cues is challenging (Spearman's coefficient: 0.38-0.73). Furthermore, kinematic data are rarely available in real-world RMIS practice. More recent works propose video- or hybrid-based methods, leveraging spatial and temporal feature encoders to extract discriminative features or by using multi-task architectures [8]–[18]. Tang et al. propose an uncertainty-aware score distribution method based on 3D Convolutional Neural Networks (CNNs), where a distribution of different scores, instead of a single one, acts as the supervisory signal [17]. This method achieves

Manuscript received: September, 19, 2022; Revised December, 28, 2022; Accepted January, 19, 2023.

This paper was recommended for publication by Editor Eric Marchand upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in whole, or in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z, NS/A000050/]; Horizon 2020 FET (863146); and the EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) [EP/S021930/1]. Dimitrios Anastasiou is supported by an EPSRC DTP [EP/R513143/1]; and an ISAD award [EP/T517793/1]. Danail Stoyanov is supported by a RAE Chair in Emerging Technologies [CiET1819/2/36] and an EPSRC Early Career Research Fellowship [EP/P012841/1].

The authors are with the Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK {dimitrios.anastasiou.21, yueming.jin, danail.stoyanov, e.mazomenos}@ucl.ac.uk

Digital Object Identifier (DOI): see top of this page.

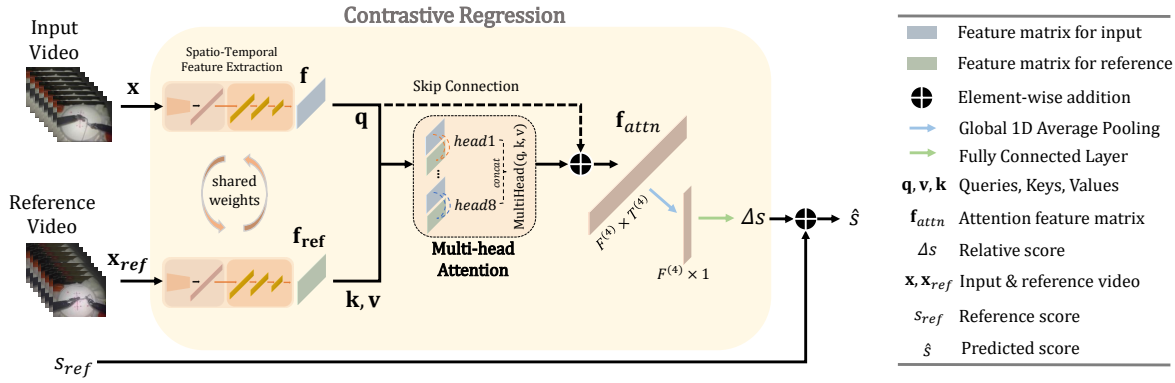

Fig. 1. Overview of our proposed framework Contra-Sformer. Input: RGB video frames for videos $\mathbf{x}$, $\mathbf{x}_{ref}$. Output: GRS score $\hat{s}$ of the video $\mathbf{x}$. Input and reference videos are fed into a stack of a ResNet-18 fined-tuned on surgical gestures and our TCN. Then, the relation between the two signals is modeled and computed by a multi-head attention block. The output of the multi-head attention block is then pooled across the temporal dimension followed by a fully connected layer to regress the relative score Δs. This is then added to the reference score s_{ref}, to predict the final score $\hat{s}$ of the input video $\mathbf{x}$.

good performance on JIGSAWS and other datasets (AQA-7 and MTL-AQA). Employing surgical gesture information (*e.g.,* pushing needle through the tissue, orienting needle) can be beneficial for capturing the subtle differences across time, during execution by different surgeons [8]. Wang et al. propose a multi-task learning framework, with primary task the GRS score prediction, and auxiliary tasks of gesture recognition and expertise classification [8]. Li et al. developed ViSA, reporting state-of-the-art performance on JIGSAWS. To model tool-tissue interaction, ViSA clusters local semantic features, produced by a 3D CNN, to generate abstract features for each group (*e.g.,* tools, tissue and background). These are fed to Bidirectional LSTMs and regress the GRS score [14]. Multimodal methods combining video and kinematics have also been proposed [2], [19], [20]. Liu et al. propose a unified multi-path framework for surgical skill assessment, with each path focusing on modeling a different aspect of skill (*e.g.*, semantic visual features, tools, events) [2]. Video features (extracted by ResNet-101), kinematic data, and gesture probabilities (obtained from an MS-TCN) are used. The model is supervised with regression loss and a self-supervised contrastive loss, achieving promising performance (see Table I).

Previous works attempt to directly regress the skill score for the target surgical video from its extracted feature vectors [8]–[10], [14], [16]–[18]. However, the manual GRS scores in JIGSAWS have considerable variation (6-30), while the accompanying videos have mostly similar appearance and context. Thus, it is challenging for deep models to robustly learn to distinguish between these scores, using only the total score value for supervision. We argue that the inherent dimensionality reduction of the regression task makes it very challenging to structure a model to learn representations that capture the subtle differences observed across the videos. Recently, the contrastive regression mechanism showed promise in a similar task of action quality assessment (AQA) [15]. Instead of learning representations that describe the skill score of a specific video, contrastive regression encourages the features to encode the difference among different videos.

In this paper, we propose a novel Contrastive Regression Transformer model, *Contra-Sformer*, for surgical skill assessment (formulated as a GRS regression task). Contra-Sformer focuses specifically on structuring features to express the level of similarity that each surgical execution (*i.e.,* test video) exhibits when compared to a reference one. Unlike [15], where the reference video is randomly selected according to the coarse category of the input test video, we set the reference video as the one with the highest assessment score for this particular surgical task. This is motivated by the current surgical training paradigm: an expert surgeon assesses a surgical execution by comparing it to an ideal execution, and deducing points when perceiving deviations. Therefore, *keeping the eye on the best* and using it as reference, our model learns how the performance in the input video spatio-temporally deviates from the reference. In Contra-Sformer, the regression model is optimized for estimating the difference in the GRS score between the two executions. We argue that by following this contrastive approach we obtain more discriminative and robust features, that are able to better generalize to the varied GRS scores among the different operators. Spatio-temporal feature extraction is implemented with a ResNet-18 and an enhanced TCN architecture. To capture similarities/deviations, we take advantage of the self-attention mechanism and use a multi-head attention block. With this modeling, we aim to encourage the generation of rich intra-video, action and skill-related features, as well as multi-aspect (tool usage, respect for tissue), inter-video features modeled by multi-head attention. Different to [15] where the two signals (i.e. reference, test) are combined with simple concatenation, we introduce multi-head attention to model similarity/deviation between videos. Also, in our work the action knowledge is implicitly embedded in the model by fine-tuning the feature extractor on atomic gestures. That allows the generation of gesture-related features, which help assess skill. This approach is different than others that use multi-task [2], [8] or segment-aware architectures [18] to encode action knowledge.

We evaluate the Contra-Sformer with Spearman's correlation coefficient (SCC) and Mean Absolute Error (MAE) on three tasks and three cross-validation schemes on JIGSAWS. We also validate the learned features for their ability to

represent skill similarity/deviation with ground truth error labels as defined in [21]. Our method achieves competitive performance compared to the state-of-the-art with a 5.8% - 13.4% normalized MAE, outperforming current methods on the knot-tying and suturing tasks. Our main contributions are summarized as follows:

1) Propose a novel contrastive regression framework for surgical skill assessment, integrating surgical domain knowledge by contrasting test inputs with a reference, selected as the optimal execution (highest GRS score).
2) Derive frame-level spatio-temporal features embedding action/gesture information, combining ResNet-18 outputs with a new temporal convolution network (TCN).
3) Propose multi-head attention to model the similarity/deviation between the input test and reference video. We show that moments of skill deviation/similarity can be identified from the derived spatio-temporal features.
4) Perform detailed analysis on the prediction error and introduce the MAE to complement SCC for evaluating regression performance of GRS score prediction in JIGSAWS.

II. METHODS

The overview of the proposed framework is illustrated in Fig. 1. Contra-Sformer takes a pair of video frames $\mathbf{x}$, $\mathbf{x}_{ref}$ as input, and outputs the GRS score $\hat{s}$ that corresponds to the video $\mathbf{x}$. Input and reference videos are encoded into compact high-level spatio-temporal feature matrices through a stack of a ResNet-18 trained on surgical gestures (*e.g.,* orienting needle, see [6]) and our enhanced TCN. To capture the difference between input and reference signals, we employ a multi-head attention block to take advantage of the self-attention mechanism and model the contrastive relation. The output of the multi-head attention block is then average pooled across the temporal dimension followed by a fully connected layer to regress the relative score. Finally, the reference score is added to the relative score to predict the final score of the input video.

A. Contrastive Regression from the Best

Typically, the task of estimating the GRS score (a real number) from an input video is formulated as a regression problem. Let $\mathbf{x}$ be the input video, then the corresponding regressed GRS score is expressed as $\hat{s} = h_w(\mathbf{x})$, where h_w is the mapping between $\mathbf{x}$ and $\hat{s}$, parameterized by w.

In this work, we model the regression task based on the similarity/deviation between test inputs and the reference. We hypothesize that this allows us to structure a model h_w that is able to capture and express the subtle differences across the videos and lead to an accurate estimation. Following [15], we develop a contrastive regression framework that estimates the relative score between an input test video and a reference video instead of regressing to the absolute score of the input. The problem is now expressed as:

$$\hat{s} = h_w(\mathbf{x}, \mathbf{x}_{ref}) + s_{ref} \tag{1}$$

where $\mathbf{x}_{ref}$ is a reference video, and s_{ref} is its corresponding GRS score label. Equation (1) can also be viewed as $\hat{s} = \Delta s + s_{ref}$, where $\Delta s \equiv h_w(\mathbf{x}, \mathbf{x}_{ref})$.

We set the reference video as the execution with the highest available GRS score in the dataset (for each task). The same reference video (per task) and its GRS score is used for training and validation, and is excluded from both training and validation sets. Unlike the suturing task, in both the knot tying and the needle passing tasks in JIGSAWS, there are no executions with perfect GRS scores. We therefore use the highest available ones as the reference. Upon visual inspection, all reference videos are free of major mistakes (*e.g.*, dropping the needle), that may affect the modeling of the similarity/deviation between input and reference.

B. Spatio-Temporal Feature Extraction

As shown in Fig. 1, Contra-Sformer first extracts spatio-temporal features of the input, $\mathbf{x}$, and the reference video, $\mathbf{x}_{ref}$. Let $\mathbf{x}_t \in \mathbb{R}^{C\times H\times W}$ denote the *t-th* frame for a video with T frames in total, and $\mathbf{y}_t \in \mathbb{R}^N$ denotes the corresponding one-hot encoded surgical gesture label of that frame. We employ the ResNet-18 [22] model to extract high-level spatial features from every frame $\mathbf{x}_t$ and optimize it as a frame-wise classifier using cross-entropy loss. We then take the outputs of the last average pooling layer as our spatial feature vector ($\mathbf{f}_t \in \mathbb{R}^{512}$). For every input video, we obtain the feature matrix $\mathbf{f} \in \mathbb{R}^{512\times T}$, representing spatial information in each frame. The spatial feature encoder is trained with per-frame gesture labels, leading to the generation of features that capture *local* deviations at frame level, with respect to the sequence and duration of the gestures that constitute each task. On the other hand, training under GRS label supervision will provide features expressing the similarity/deviation between the two videos at a *global* level.

To encode time cues within the videos, a TCN is employed. We differentiate from relevant works [2], [8] that use the multi-stage TCN (MS-TCN) [23], by introducing a customized TCN architecture, to improve temporal information modeling specifically for our contrastive regression task. We use standard 1D convolutions (four layers with output dimensions of $\{64, 32, 16, 16\}$; kernel size = 25; and stride = 1), instead of 1D dilated convolutions, followed by 1D max pooling (kernel size = 3, stride = 3) and batch normalization layers to compress the temporal information (see Fig. 2). Pooling layers are important for regressing the skill score, as they aggregate temporal information to represent groups of actions that constitute different aspects of skills. Formally, we input the feature matrix $\mathbf{f}^{(0)} \in \mathbb{R}^{F^{(0)}\times T^{(0)}}$ to the TCN, where $F^{(0)} = 512$ and $T^{(0)}$ is the initial temporal resolution. At each layer *l*, the following operations take place,

$$\mathbf{f}^{(l)} = \mathrm{BN}\left\{\mathrm{MaxPool}\{\mathrm{ReLU}(\mathbf{W}^{(l)} * \mathbf{f}^{(l-1)} + \mathbf{b}^{(l)})\}\right\} \tag{2}$$

where $*$ denotes the convolution operator, and $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ denote parameters and the bias term. The final output is a feature matrix $\mathbf{f}^{(4)} \in \mathbb{R}^{16\times T^{(4)}}$.

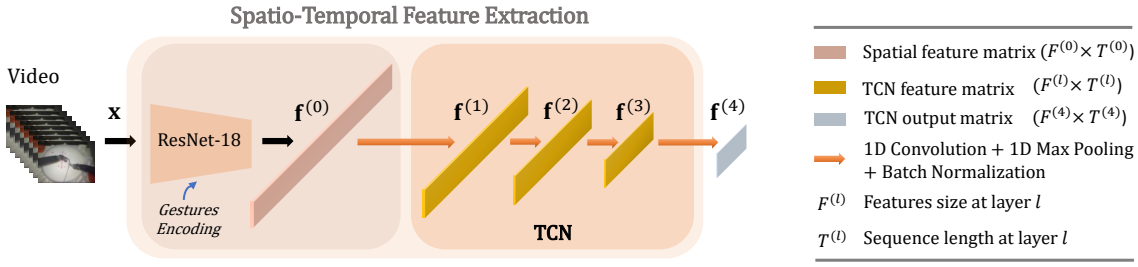

Fig. 2. Spatio-temporal feature extraction from the video frames is done by a ResNet-18 fined-tuned on surgical gesture labels, followed by our TCN for encoding time cues within the videos. The output is a high-level feature matrix $\mathbf{f}^{(4)} \in \mathbb{R}^{F^{(4)}\times T^{(4)}}$.

C. Relation Modeling via Action-Aware Transformer

Regressing the relative score between the current input video and the reference requires capturing the latent relations between them. One straightforward way is to concatenate their feature vectors and use it for regression [15]. However, simple concatenation does not encourage the generation of features representing the differences between the two videos. Instead, we propose to explicitly model the relation between the two executions by computing the similarity function (*i.e.,* dot product) between their corresponding high-level feature vectors.

GRS evaluates various aspects (*i.e.,* tissue handling, instrument movement) of surgical execution. Therefore, we implement the relation modeling block with a multi-head attention (8 heads, hidden dimensionality of 16 and no dropout) Transformer [24] block (see Fig. 1), to evaluate the similarity/deviation of the two feature sequences at different scales. Practically, this module compares the similarity/deviation of patterns indicative of high GRS score, present in the reference, with the patterns observed in the test video. Given the feature matrices $\mathbf{f} = \mathbf{f}^{(4)}$ and $\mathbf{f}_{ref} = \mathbf{f}^{(4)}_{ref}$, from the input and reference branches respectively, $\mathbf{f}$ are set as the queries $\mathbf{q}$ and $\mathbf{f}_{ref}$ as the keys $\mathbf{k}$ and values $\mathbf{v}$. Their relation is modeled as:

$$\mathbf{f}_{attn} = \text{Concat}\left(\mathbf{h}_1, \ldots, \mathbf{h}_n\right)\mathbf{W}^O + \mathbf{f} \tag{3}$$

$$\mathbf{h}_i = \text{softmax}\left(\frac{(\mathbf{f}\mathbf{W}_i^q)\cdot(\mathbf{f}_{ref}\mathbf{W}_i^k)^T}{\sqrt{d_k}}\right)\cdot(\mathbf{f}_{ref}\mathbf{W}_i^v) \tag{4}$$

where, $\mathbf{h}_i$ denotes the head i for $i = 1, ..., n$, d_k is the hidden dimensionality for queries and keys, and $\mathbf{W}_i^q, \mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^{T^{(4)}\times T^{(4)}}$, and $\mathbf{W}^O$ are learnable parameters. A skip connection, adding the output of that block to the queries, is introduced to maintain as much information as possible from the current input execution.

The feature matrix $\mathbf{f}_{attn} \in \mathbb{R}^{F^{(4)}\times T^{(4)}}$ is then averaged across the temporal dimension followed by a fully connected layer to compute the relative score $\Delta s \in \mathbb{R}$. The relative score is then added to the reference score, as shown in Eq. (1), to predict the final GRS score $\hat{s}$ for the input video.

D. Training Procedure

Contra-Sformer is trained end-to-end using a combination of the Mean Squared Error (MSE) and Mean Absolute Error (MAE) as the loss function, with proven effectiveness in similar quality assessment tasks [25]. It is defined as:

$$\mathcal{L} = \alpha\text{MSE}(\hat{s}, s) + (1-\alpha)\text{MAE}(\hat{s}, s), \tag{5}$$

where α is the hyper-parameter to balance the loss.

To optimize memory and computational time, training takes place in a two-step process. First, the feature encoder (ResNet-18) is trained using gesture labels and fixed using the weights corresponding to the minimum cross-entropy loss. Aggregating the individual feature vectors $\mathbf{f}_t \in \mathbb{R}^{512}$ from every frame t, a high-level representation of the video $\mathbf{f} \in \mathbb{R}^{512\times T}$ is formed. These feature matrices for both input and reference video are used as inputs to the TCN, which is trained jointly with the rest of the network in an end-to-end manner. Also, the whole video sequences of the executions, instead of frame clips, are used as inputs to maximize the information we provide the network with.

III. EXPERIMENTAL VALIDATION

A. Dataset and Evaluation Protocol

We develop and evaluate Contra-Sformer on the JIGSAWS dataset [6]. JIGSAWS is a benchmark dataset for surgical skill assessment and consists of three fundamental surgical tasks: suturing (SU), needle passing (NP), and knot tying (KT), all performed in a simulated environment using dry lab benchtop models and the da Vinci API. Each task is executed five times (trials) by the eight participating surgeons (2 experts, 2 intermediate, and 4 novices) while synchronized video and kinematic data are recorded. There are 39 videos for suturing, 28 for needle passing, and 36 for knot tying. Ground truth GRS scores (ranging from 6 to 30) provide the skill annotations and frame-level surgical gesture labels are provided for all tasks.

Cross-validation schemes for evaluating our method include the Leave-One-Supertrial-Out (LOSO) and Leave-One-User-Out (LOUO) setups as set forth in the JIGSAWS literature [6], as well as a random 4-Fold scheme used by [2], [15], [17]. In LOUO, all trials by a single surgeon are left out as the test set and the remaining are for training. In LOSO, the i-th trial of each surgeon is left out as the test set. LOUO evaluates model generalization to different surgeons, while LOSO to different trials performed by the same surgeon. The SCC is

TABLE I
COMPARISON TO THE STATE-OF-THE-ART METHODS ON JIGSAWS UNDER SPEARMAN'S CORRELATION COEFFICIENT AND MEAN ABSOLUTE ERROR ON THREE CROSS-VALIDATION SCHEMES AND ALL TASKS. **K**: KINEMATIC DATA, **V**: VIDEO DATA; *UTILIZATION OF SURGICAL GESTURE LABELS. CONTRA-SFORMER IS TRAINED AND EVALUATED IN ONE VIDEO LESS (REFERENCE) THAN THE METHODS PRESENTED.

Input	Method	Tasks & Scheme											
		KT			NP			SU			Across Tasks		
		LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold	LOSO	LOUO	4-Fold
		Spearmans's Correlation Coefficient (SSC)											
K	SMT-DCT-DFT	0.70	0.73	-	0.38	0.23	-	0.64	0.10	-	0.59	0.40	-
	DCT-DFT-ApEn	0.63	0.60	-	0.46	0.25	-	0.75	0.37	-	0.63	0.42	-
V	ResNet-LSTM	0.52	0.36	-	0.84	0.33	-	0.73	0.67	-	0.72	0.47	-
	C3D-LSTM	0.81	0.60	-	0.84	0.78	-	0.69	0.59	-	0.79	0.67	-
	C3D-SVR	0.71	0.33	-	0.75	-0.17	-	0.42	0.37	-	0.65	0.18	-
	USDL	-	-	0.61	-	-	0.63	-	-	0.64	-	-	0.63
	MUSDL	-	-	0.71	-	-	0.69	-	-	0.71	-	-	0.70
	CoRe + GART	-	-	0.86	-	-	0.86	-	-	**0.84**	-	-	**0.85**
	*S3D	0.64	0.14	-	0.57	0.35	-	0.68	0.03	-	0.63	0.18	-
	*ResNet-MTL-VF	0.63	0.72	-	0.73	0.48	-	0.79	0.68	-	0.72	0.64	-
	*C3D-MTL-VF	0.89	**0.83**	-	0.75	0.86	-	0.77	0.69	-	0.81	0.80	-
	ViSA	**0.92**	0.76	0.84	**0.93**	**0.90**	**0.86**	0.84	**0.72**	0.79	**0.90**	**0.81**	0.83
V + K	JR-GCN	-	0.19	0.75	-	0.67	0.51	-	0.35	0.36	-	0.43	0.56
	AIM	-	0.61	0.82	-	0.34	0.65	-	0.45	0.63	-	0.47	0.71
	MultiPath-VTP	-	0.58	0.78	-	0.62	0.76	-	0.45	0.79	-	0.55	0.78
	MultiPath-VTPE	-	0.59	0.82	-	0.65	0.76	-	0.45	0.83	-	0.57	0.81
V	***Contra-Sformer (ours)**	0.89	0.69	**0.87**	0.71	0.71	0.81	**0.86**	0.65	0.69	0.83	0.68	0.81
	SD of SCC	(±0.067)	(±0.190)	(±0.043)	(±0.104)	(±0.301)	(±0.112)	(±0.056)	(±0.138)	(±0.148)	(±0.079)	(±0.220)	(±0.110)
		Mean Absolute Error (MAE)											
V	ViSA	2.16	2.01	2.60	**1.66**	**2.16**	**2.05**	**2.58**	2.82	**2.70**	**2.13**	**2.33**	**2.45**
	***Contra-Sformer (ours)**	**1.75**	**1.39**	**2.10**	3.15	3.17	3.21	2.74	**2.58**	2.99	2.55	2.38	2.77
	SD of MAE	(±0.68)	(±0.73)	(±0.47)	(±0.89)	(±1.25)	(±1.12)	(±0.87)	(±1.06)	(±0.75)	(±0.82)	(±1.04)	(±0.82)

adopted as the main evaluation metric, comparing the rank of the predicted scores with the ground truth. SCC is defined as:

$$\rho = \frac{\sum_i (p_i - p)(q_i - q)}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \tag{6}$$

where p and q represent the ranking for each sample of two series respectively. We report SCC values averaged on all folds for each cross-validation scheme, and compute the average SCC across tasks using Fisher's z-value [2]. Since not all methods report average results using Fisher's z-value, we calculate these (Table I) based on the SCC values on individual tasks. Although useful as a metric, SCC alone cannot fully evaluate regression accuracy. Thus, we also report the Mean Absolute Error (MAE). Results in Table I are averaged on multiple runs with 5 different random seeds.

B. Implementation Details

We initialize the ResNet-18 on Image-Net and train it with $lr = 5e{-}6$ for 60 epochs. Video frames are resized to 240 $\times$ 240 and center cropped at 224 $\times$ 224. Random Horizontal Flip with $p = 0.1$ and Random Rotation of $\pm 5deg$ are used for KT and NP. To reduce redundancy and computational load, videos are downsampled to 5Hz. The rest of the framework is trained with the Adam optimizer for 300 epochs with $lr = 5e{-}4$ for SU and KT and, 600 epochs with $lr = 5e{-}5$ for NP. For KT we use batch size $=$ 16, and for SU and NP, we use full-batch training (*i.e.,* batch size $=$ # training samples). After experimentation, two values are considered for the hyperparameter α in Eq. (5), 0.5 for KT and NP, and 0.65 for SU. The videos were zero-padded to have equal length, according to the longest video in the dataset. Videos with the highest GRS scores in each task are used as reference: C004 for SU (GRS:30), E003 for KT (GRS:22) and F004 for NP (GRS:24). Following the literature, design choices and hyperparameter tuning is done experimentally on the averaged cross-validation test data. Contra-Sformer is implemented with PyTorch and trained on an NVIDIA RTX A6000 GPU.

C. Comparison with the State-Of-The-Art

SCC results, alongside a comparison with state-of-the-art methods (for GRS score regression in JIGSAWS), for each task and all three cross-validation schemes are listed in Table I. Overall, Contra-Sformer achieves competitive performance, outperforming the state-of-the-art on KT 4-Fold and SU LOSO, achieving SCC of 0.87 and 0.86 respectively. Contra-Sformer, along with the highest-performing video-based methods ViSA and C3D-MTL-VF, surpass MultiPath-VTPE [2] that utilizes both video and kinematic information in almost all tasks and schemes. Core+Gart does not report LOSO and LOUO results and was evaluated only on a random 4-Fold

scheme. Since a full comparison is not possible, comparing on the random 4-Fold setup, Contra-Sformer performs better on KT, while CoRe+GART performs better on the other tasks.

From the SCC values in Table I Contra-Sformer has comparable performance with the state-of-the-art methods, ViSA and C3D-MTL-VF, with ViSA showing better performance in LOUO, considered the most challenging setup [14]. However, SCC has limitations as a metric for regression accuracy (see Section III.D), especially in small evaluation sets and does not provide any information on the prediction error. Therefore, we evaluate our method using the MAE metric in addition to SCC and compare it with ViSA in Table I (see details in Section III.D).

D. Performance Metrics Analysis

SCC is the widely used evaluation metric in the JIGSAWS GRS score regression task. Although its efficiency is unquestioned, it only evaluates the rank of the predicted scores, without any information on the actual prediction error. Subsequently, a high SCC is not necessarily accompanied by a small prediction error, particularly in small test sets. For example, in a validation set with 2 videos with ground truth of 22 and 15 and predictions of 16 and 8 respectively, SCC would result in "1", while the prediction is actually poor.

To this end, we report MAE per task and cross-validation scheme (averaged on all folds and runs) together with the standard deviation (SD) of MAE values computed on the different folds for 5 different runs. We believe that such error analysis (on the prediction error) should complement the SCC in the evaluation of methods for GRS score estimation in the JIGSAWS dataset.

Results are presented in Table I. The MAE range is 1.39 - 3.21 and considering the range (6-30) of the GRS scores, corresponds to 5.8% - 13.4% Normalized MAE (NMAE), calculated by $NMAE = \frac{MAE}{range(GRS)} \cdot 100\%$. The slightly increased SD on LOUO is attributed to the unbalanced dataset (see III.A), where in some folds the training set contains only one user of the same expertise level. Also, LOUO has the smallest validation set (2-5 videos) and thus is more prone to show higher SD. Also, observing MAE and the SCC values from Table I, it is clear that a high SCC does not always correspond to a low prediction error and vice versa. Overall Contra-Sformer has small prediction errors (MAE: 1.39-2.10) in all GRS score ranges in KT. The error slightly increases in high GRS score ranges in SU (MAE: 2.39-3.85), and NP (MAE: 3.15-3.21). This is attributed to the small number of samples with high GRS scores in SU, and the small NP dataset.

Contra-Sformer outperforms ViSA, on all evaluation schemes of the KT task, by 18.98%, 30.85%, and 19.23% on LOSO, LOUO, and 4-Fold respectively. Contra-Sformer also achieves better LOUO performance in the SU, which is promising and indicates good generalization to unseen surgeons. ViSA has better performance in NP. Contra-Sformer performance in NP is mostly attributed to the smaller size of the NP dataset, compared to SU and KT (28, 39, and 36 respectively).

E. Ablation Studies

We validate: **1)** the key design components of Contra-Sformer, and **2)** the selection of the reference video, with ablation experiments on the most challenging (SU) of the three JIGSAWS tasks. We use the same set of hyperparameters as the ones used for our main experiment, since the performance overall remained consistent with different hyperparameters.

1) Contra-Sformer Design Components: The effectiveness of key design components in Contra-Sformer is studied with the following configurations: (i) We replace our TCN with the MS-TCN [23], using 10 layers and 64 feature maps, following [8]. As the original MS-TCN has been implemented for gesture recognition, we adapt it for the skill assessment task. Thus, we use two stages, as we observed degradation in the performance with more. The number of feature maps in the last convolutional layer is set to 16 in order to match the hidden dimensionality of the multi-head attention block; (ii) To test the benefit of contrastive regression, we regress the final GRS score directly from the input video without contrasting it with the reference. After the TCN stage, we use temporal average pooling followed by a fully connected layer to predict the final score; (iii) To validate the similarity/deviation modeling with the multi-head attention block, we combine the feature matrices of the input and the reference with simple concatenation (as in [15]); (iv) To further test the ability of our TCN to model temporal cues, we replace it with a 1D point-wise convolution layer and a max pooling layer to reduce feature and temporal dimensions. Results from the ablation studies are presented in Table II.

The performance in LOSO and 4-Fold is comparable for the MS-TCN and our TCN. However, in LOUO, the accuracy doubled when using our TCN. The results from ablation (ii) also show the benefit of contrastive regression, as regressing the score directly from an input video significantly reduces performance, with the decrease being more prominent in the LOUO setup. That is expected as no inter-video features are generated to model the subtle variations across different executions. Simple concatenation (ablation (iii)) does not allow the generation of rich inter-video features leading to reduced performance in all setups. This highlights the success of the multi-head attention block in modeling the similarity between input and reference execution, which contributes to the score prediction. Ablation (iv) shows lower performance when our TCN is replaced by a simpler 1D convolution module, proving the ability of our TCN to model temporal cues effectively. Also, the comparison between ablation (i) and (iv), suggests that temporal pooling layers are important for performance improvement in skill score estimation tasks, as they help in summarizing temporal information that could potentially represent groups of actions that constitute different aspects of skills. Comparing ablation (ii) and (iii) suggests that contrastive regression is beneficial only when the two signals (input test and reference) are fused in a more sophisticated method (*e.g.,* multi-head attention) instead of concatenation. Overall, the ablation studies justify the Contra-Sformer architecture, as it improves performance in LOUO, the most challenging evaluation scheme [14].

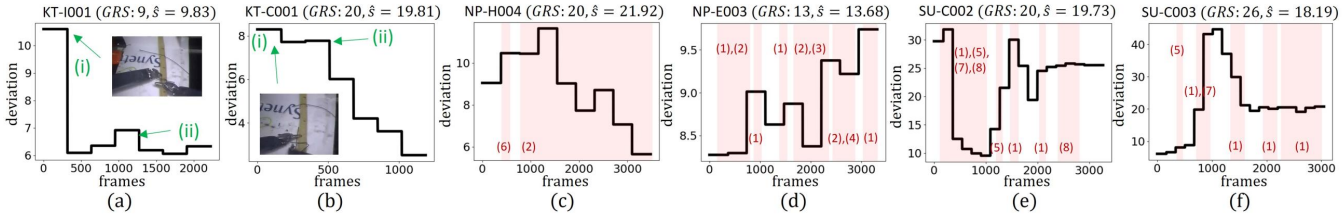

Fig. 3. Deviation of a test video with respect to the gold-standard reference video. High deviation corresponds to low similarity. Error annotations: (1) Out of view, (2) Erroneous gesture G2 (positioning the tip of the needle), (3) Erroneous gesture G3 (pushing needle through the tissue), (4) Erroneous gesture G4 (transferring needle from left to right), (5) Multiple attempts, (6) Erroneous gesture G6 (pulling suture with left hand), (7) Wrong positioning, (8) Not moving along the curve; (a) Annotations (i) and (ii) show instances of poor suture thread handling and multiple attempts to tie the knot (00min16s) and grasp the thread (00min38s). (b) Example where the test video's GRS score is very close to the highest. Multiple attempts to grab the suture (00min08s) and difficulty in securing the knot (00min15s). (c) The generation of this plot is shown in the supplementary material video. (d) NP execution. (e) SU execution. (f) Example with high prediction error.

2) Selection of Reference Video: To highlight the importance of selecting the trial with the highest GRS score as reference, we compare performance using different references. Table II lists results from two experiments: one with an intermediate score (I003, $GRS = 17$), and one with the lowest score (D004, $GRS = 8$) in SU as references.

Performance under LOUO decreases significantly (almost by half) with the intermediate and lowest scores as reference, while a smaller drop also occurs in LOSO. Evidently, Contra-Sformer can regress to values higher/lower than the reference video (Δs can be positive or negative depending on the reference score), with close performance in LOSO and 4-Fold, but robust GRS estimation across all schemes is only achieved when the highest GRS score is used as the reference. The underlying reason is that in JIGSAWS tasks, the sequence and execution of the surgical gestures comprising the task and the overall result should look fairly similar (high similarity/small deviation) across executions with high scores. Lower GRS scores are attributed to sub-optimal performance of surgical gestures. Our results also show that it is more efficient for the network to learn similarities/deviations of an input test video with the highest/lowest GRS score rather than a case with an intermediate score. Finally, although the best performance is achieved with the highest reference score, GRS scores close to the reference, tend to be underestimated in the SU task.

TABLE II
SCC RESULTS ON SU TASK IN ABLATION EXPERIMENTS.

Contra-Sformer Design Components					
Contr. Regres.	**Multi-head Att.**	**Temp. Enc.**	**LOSO**	**LOUO**	**4-Fold**
✓	✓	MS-TCN	0.80	0.33	0.67
✗	✗	our TCN	0.79	0.43	0.60
✓	✗	our TCN	0.67	0.43	0.58
✓	✓	1x1conv.	0.82	0.41	0.64
✓	✓	our TCN	**0.86**	**0.65**	**0.69**
Selection of Reference Video					
Reference Video		**GRS**	**LOSO**	**LOUO**	**4-Fold**
I003 (interm.)		17	0.78	0.32	0.66
D004 (lowest)		8	0.84	0.37	0.55
Ours (highest)		30	**0.86**	**0.65**	**0.69**

F. Visualizing the Deviation

To provide further insight into the contrastive regression mechanism and quantitatively validate the feature-generation approach, we extract and visualize as timeseries the attention weights of the contrastive inter-video features (main features contributing to the regression task) generated from the multi-head attention block. Weights are averaged across all heads resulting in a $T^{(4)} \times T^{(4)}$ attention weight matrix that corresponds to the temporal dimensions of the current and reference videos. We average the weights across the second dimension (which corresponds to the reference video), and obtain a $T^{(4)} \times 1$ array, which is then re-scaled (with nearest-neighbor interpolation) to match the initial temporal resolution of the execution. From this, we obtain averaged attention weights for each time point (*i.e.,* frame) in the current video. The weights represent skill-similarity between the two videos, thus inverse weights represent skill-deviation (high value corresponds to high deviation). In Fig. 3 we plot the deviation of the input with respect to the reference for two cases from each task.

In [21], procedural and executional error annotations are provided at gesture level, for SU and NP in JIGSAWS. Procedural errors are defined as any deviation from an ideal sequence of surgical gestures, and executional errors are defined as poor/failed manipulation of gestures within the task (*e.g.,* needle drop, multiple attempts). Procedural errors are annotated as unnecessary/problematic gestures and executional errors with the description of the error characterising the gesture. In Fig. 3 (c) - (f), red areas represent gesture errors. Almost all high deviation moments (peaks) in Fig. 3 (c) - (e), are located inside error areas, indicating that features learned by Contra-Sformer can successfully capture sub-optimal execution, contributing to distinguishing the relative performance of the test video. Fig. 3 (f) shows a less accurate example (prediction error of 7.81). For KT, we annotated the instances in the input video where sub-optimal actions clearly occur (these can be visually verified).

IV. CONCLUSIONS

A novel contrastive regression Transformer framework, Contra-Sformer, for automated skill assessment in RMIS is proposed. Instead of directly regressing the GRS score,

we formulate the regression task based on the skill-similarity/deviation of the test videos compared to a gold-standard reference, selected as the execution with the highest GRS score per task. A feature extractor block integrating ResNet-18 and an enhanced TCN, generates spatio-temporal features, encompassing fine-grained action/gesture information. The similarity between the two feature matrices is modeled with self-attention using a multi-head attention block.

We validate Contra-Sformer on JIGSAWS and report competitive results on all three surgical tasks and experimental setups. Our method outperforms the state-of-the-art on LOSO and 4-Fold experiments on the SCC metric and achieves normalized mean absolute error between 5.8% - 13.4% across tasks and experimental setups. Contra-Sformer can also capture the differences between test and reference video in delicate and complex surgical gestures (*i.e.*, SU task) and thanks to the contrastive regression mechanism estimates the GRS score with small prediction error. Ablation studies highlight the benefits and justify our contrastive regression approach and Contra-Sformer architecture (TCN, multi-head attention block). When optimized, Contra-Sformer generates features that faithfully represent the similarity/deviation between the two executions and encode information indicative of sub-optimal execution/errors, without requiring explicit error annotations. This is validated against manual error annotations from [21], and can be exploited for providing targeted feedback and real-time assessment to trainees.

Conceptually, Contra-Sformer is directly applicable to real surgical data (after re-training and hyperparameter search), with skill score and surgical gesture annotations available (similar to JIGSAWS). Since real surgical data consist of scenes with blood and tissue that can be deforming in time, different anatomies and light conditions across different executions, as well as non-constant field of view (as the surgical camera moves), appropriate data augmentations (e.g. cropping, scaling, flipping and color/contrast transformations), should be considered. Additionally, surgical procedures consist of sequences of different phases (e.g., dissection phase, suturing phase, anastomosis), further divided into individual tasks. It is expected that focused Contra-Sformer models would be developed for each phase/task separately. Contra-Sformer represents a novel approach for automated surgical skill assessment by constructing a score regression task on the basis of the comparison (similarity/deviation) between a test execution against a gold-standard one. This formulation offers the advantage of generating both intra-video features, that can vary across different surgical datasets, and inter-video features (input test and reference video) to support the regression task.

V. ACKNOWLEDGEMENT

We are grateful to Zhenqiang Li for sharing the MAE results of the ViSA paper [14].

REFERENCES

[1] C. D'Ettorre, A. Mariani, A. Stilli *et al.*, "Accelerating surgical robotics research: A review of 10 years with the da vinci research kit," *IEEE Robot. Autom. Mag.*, vol. 28, no. 4, pp. 56–78, 2021.

[2] D. Liu, Q. Li, T. Jiang *et al.*, "Towards unified surgical skill assessment," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 9517–9526, 2021.

[3] J. A. Martin, G. Regehr, R. Reznick *et al.*, "Objective structured assessment of technical skill (osats) for surgical residents," *BJS*, vol. 84, 1997.

[4] Y. Kassahun, B. Yu, A. T. Tibebu *et al.*, "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 4, p. 553–568, 2016.

[5] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 5, p. 731–739, 2018.

[6] N. Ahmidi, L. Tao, S. Sefati *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2025–2041, 2017.

[7] H. I. Fawaz, G. Forestier, J. Weber *et al.*, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 9, p. 1611–1617, 2019.

[8] T. Wang, Y. Wang, and M. Li, "Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels," *Med. Image Comput. Comput. Assist. Interv.*, vol. 12263 LNCS, p. 668–678, 2020.

[9] I. Funke, S. T. Mees, J. Weitz *et al.*, "Video-based surgical skill assessment using 3d convolutional neural networks," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 7, no. 14, p. 1217–1225, 2019.

[10] A. Soleymani, A. Akbar, S. Asl *et al.*, "Surgical skill evaluation from robot-assisted surgery recordings," *in Proc. Int. Symp. Med. Robot*, pp. 1–6, 2021.

[11] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6057–6066, 2018.

[12] D. Liu, T. Jiang, Y. Wang *et al.*, "Surgical skill assessment on in-vivo clinical data via the clearness of operating field," *Med. Image Comput. Comput. Assist. Interv.*, p. 476–484, 2019.

[13] J. Zhang, Y. Nie, Y. Lyu *et al.*, "Sd-net: joint surgical gesture recognition and skill assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 10, p. 1675–1682, 2021.

[14] Z. Li, L. Gu, W. Wang *et al.*, "Surgical skill assessment via video semantic aggregation," *Med. Image Comput. Comput. Assist. Interv.*, vol. 13437, p. 410–420, 2022.

[15] X. Yu, Y. Rao, W. Zhao *et al.*, "Group-aware contrastive regression for action quality assessment," *in Proc. IEEE Int. Conf. Comput. Vis.*, pp. 7919–7928, 2021.

[16] P. Parmar and B. T. Morris, "Learning to score olympic events," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 76–84, 2017.

[17] Y. Tang, Z. Ni, J. Zhou *et al.*, "Uncertainty-aware score distribution learning for action quality assessment," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 9836–9845, 2020.

[18] X. Xiang, Y. Tian, A. Reiter *et al.*, "S3d: Stacking segmental p3d for action quality assessment," *in Proc. Int. Conf. on Image Proces.*, pp. 2381–8549, 2018.

[19] J. Gao, W. S. Zheng, J. H. Pan *et al.*, "An asymmetric modeling for action assessment," *in Proc. Europ. Conf. Comput. Vis.*, p. 222–238, 2020.

[20] J. H. Pan, J. Gao, and W. S. Zheng, "Action assessment by joint relation graphs," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6330–6339, 2019.

[21] K. Hutchinson, Z. Li, L. A. Cantrell *et al.*, "Analysis of executional and procedural errors in dry-lab robotic surgery experiments," *Int J Med Robot.*, vol. 18, no. 3, 2022.

[22] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[23] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3575–3584, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *in Adv. Neural Inf. Process. Syst.*, vol. 30, p. 6000–6010, 2017.

[25] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," *in Proc. Conf. Comput. Vis. Pattern Recognit.*, pp. 304–313, 2019.