This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Unified Surgical Skill Assessment

Daochang Liu[1,3,5], Qiyue Li [1], Tingting Jiang[1], Yizhou Wang[1,4], Rulin Miao[2], Fei Shan[2], Ziyu Li[2]
[1]NELVT, Department of Computer Science, Peking University
[2]Peking University Cancer Hospital, [3]Deepwise AI Lab
[4]Center on Frontiers of Computing Studies, Peking University
[5]Advanced Institute of Information Technology, Peking University
{daochang, liqiyue, ttjiang}@pku.edu.cn

Abstract

Surgical skills have a great influence on surgical safety and patients' well-being. Traditional assessment of surgical skills involves strenuous manual efforts, which lacks efficiency and repeatability. Therefore, we attempt to automatically predict how well the surgery is performed using the surgical video. In this paper, a unified multi-path framework for automatic surgical skill assessment is proposed, which takes care of multiple composing aspects of surgical skills, including surgical tool usage, intraoperative event pattern, and other skill proxies. The dependency relationships among these different aspects are specially modeled by a path dependency module in the framework. We conduct extensive experiments on the JIGSAWS dataset of simulated surgical tasks, and a new clinical dataset of real laparoscopic surgeries. The proposed framework achieves promising results on both datasets, with the state-of-the-art on the simulated dataset advanced from 0.71 Spearman's correlation to 0.80. It is also shown that combining multiple skill aspects yields better performance than relying on a single aspect.

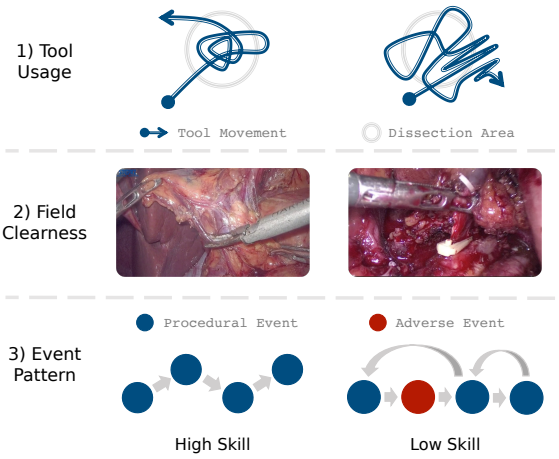

Figure 1. **Different aspects of surgical skills.** Surgical skills can be assessed from many aspects, *e.g.*, 1) the usage of surgical tools 2) the clearness of the operating field 3) the distribution of surgical events. This paper proposes a unified framework for surgical skill assessment, which exploits these different aspects and the interaction among them. Best viewed in color.

1. Introduction

Hundreds of millions of surgeries are performed worldwide annually [57]. The proficiency of the operating surgeon is a key factor affecting outcomes after surgery [8]. To ensure patient safety and reduce clinical errors, surgical skill assessment has become an indispensable part of surgical training [47] and credentialing [56].

Conventional surgical skill assessment is undertaken manually by experts with direct observation [47] or structured rating protocols [35]. Such human assessment is slow and hardly reproducible. Meanwhile, the prevalence of laparoscopic and robot-assisted surgeries nowadays brings a large volume of surgery videos captured by the built-in cameras in surgical devices, which lay the foundation for automatic learning-based approaches to provide efficient and repeatable skill assessment [52]. This paper works on automatic surgical skill assessment using surgical videos.

Surgical skills are complex with many facets. No universally accepted skill assessment criterion exists in the medical field currently [10]. By discussion with clinicians, we identify three important aspects from the medical literature that are likely to characterize surgical skills and also suitable for automatic assessment, *i.e.*, surgical tool usage, surgical field clearness, and surgical event pattern. *The first aspect* is the movement of surgical tools [19, 36], which could reflect the instrument handling proficiency and motion efficiency of the surgeon. As in Fig. 1, a high-skill surgeon will have a short and smooth tool trajectory con-

centrating on the dissection area, while a low-skill surgeon will have a lengthy and jerky trajectory dispersed in a large spatial range. *The second aspect* is the clearness of the operating field as a skill proxy [31]. Skill proxy means an indirect indicator that is statistically correlated to surgical skills. Concretely, a clear operating field ensures high visibility of anatomy structures, which is critical for the surgeon's performance. And it is more likely to link a poorly-performed surgery to an obscure field with excessive bleeding, burnt tissue, smoke, and thus limited anatomy visibility. *The third aspect* is the workflow of surgical events or actions [50]. For example in Fig. 1, a well-performed surgery tends to have a linear pattern of events following the optimal procedure. On the contrary, a loopy pattern with more jumps across events is more common in a poorly-performed surgery, because the surgeon could find some previous surgical step unsatisfactory and go back to fix it. Besides, skill-related factors are also major causes of adverse events in surgeries, such as bleeding and injury [60].

Prior works on automated surgical skill assessment, *e.g.* [46, 24, 31], mostly rely on one of these aspects. However, we believe the great complexity of surgical skills necessitates a combination of multiple aspects for an accurate assessment. Besides, the dependency relationships among different aspects also play important roles in skill assessment. For instance, tool usage needs to be more careful when the field clearness is impaired. This paper thus conceptualizes a unified framework to leverage the complementary information in different skill aspects and also capture dependencies among aspects. The proposed framework comprises multiple paths in parallel, with each corresponding to a skill aspect. Aspect-specific feature sequences extracted from surgical videos are forwarded along each path, subsequently transformed into skill score sequences for each aspect. We integrate a path dependency module into the framework to capture inter-path dependencies. In this module, feature sequences are aggregated from all the skill aspects to provide temporal importance weights for the score sequences. Lastly, the weighted score sequences are pooled over time and fused across paths as the final assessment prediction. One practical problem of surgical skill assessment is the limited amount of annotations for training. Therefore, apart from a classic supervised regression loss, the framework is additionally equipped with a self-supervised contrastive loss to learn without annotations. Specifically, we employ a predictive coding mechanism on the latent embedding of feature sequences.

On the other hand, existing approaches are usually validated on simulated surgical tasks, such as knot-tying in the JIGSAWS benchmark [18, 2]. However, a clinical dataset of real surgeries is more desirable. And it is better to also have event and tool annotations to support multi-aspect assessment. The clinical dataset in EndoVis Challenge 2019 [1] satisfies these requirements, which is unfortunately not publicly usable yet. Therefore, we collect a new clinical dataset consisting of twenty *in vivo* laparoscopic gastrectomies with comprehensive skill and event annotations. The proposed framework is validated by extensive experiments on our new clinical dataset and the simulated JIGSAWS dataset. On both datasets, instantiations of the proposed framework obtain state-of-the-art results. Experimental results are higher when multiple skill aspects are combined, validating the effectiveness of our unified approach. We also correlate the predicted skill scores with the input features on the temporal dimension to get insights on how the model understands surgical skills.

To summarize, our contributions are three-fold: 1) A unified framework assessing surgical skills from multiple aspects 2) A new clinical surgery dataset 3) Promising results on both simulated and clinical surgery datasets.

2. Related Works

2.1. Automatic Surgical Skill Assessment

Prior works roughly fall into three categories according to which skill aspect they are related to. *The first category* is tool-related and makes up the majority of the literature. Methods in this category rely on tool motion data from various sources, including video object tracking or detection [24, 5, 44], video spatiotemporal descriptors [65, 64, 62, 7], robotic kinematics [63, 55, 15, 9], external sensors [13, 4, 23], and virtual reality interfaces [25]. *The second category* is proxy-related. The clearness of operating field is identified as a skill proxy on clinical data [31]. In their method [31], indirect assessment via proxy performs better than direct skill prediction. *The third category* is event-related. They usually break down surgical trials into fine-grained events [46, 49, 34, 3, 51, 33]. These methods use surgical events differently from our framework. Our framework learns from the occurrence pattern of events to determine skills, while they mainly evaluate surgical skills in the individual event to provide detailed feedback. A recent method [54], which also belongs to the event-related category, tackles surgical skill assessment by multi-task learning with surgical gestures. In addition, some researchers adopt a purely learning-based approach [16] that is not related to any of these three categories.

Most previous studies take advantage of only one skill aspect. Our framework, instead, tries to unify multiple aspects for surgical skill assessment.

2.2. Action Quality Assessment

Action quality assessment is a field relevant to surgical skill assessment. Methods in this field aim at assessing the quality of actions in sports such as diving and gymnastics [6, 40, 17, 48, 39, 37, 61, 45, 42, 41, 29, 20, 26, 58],

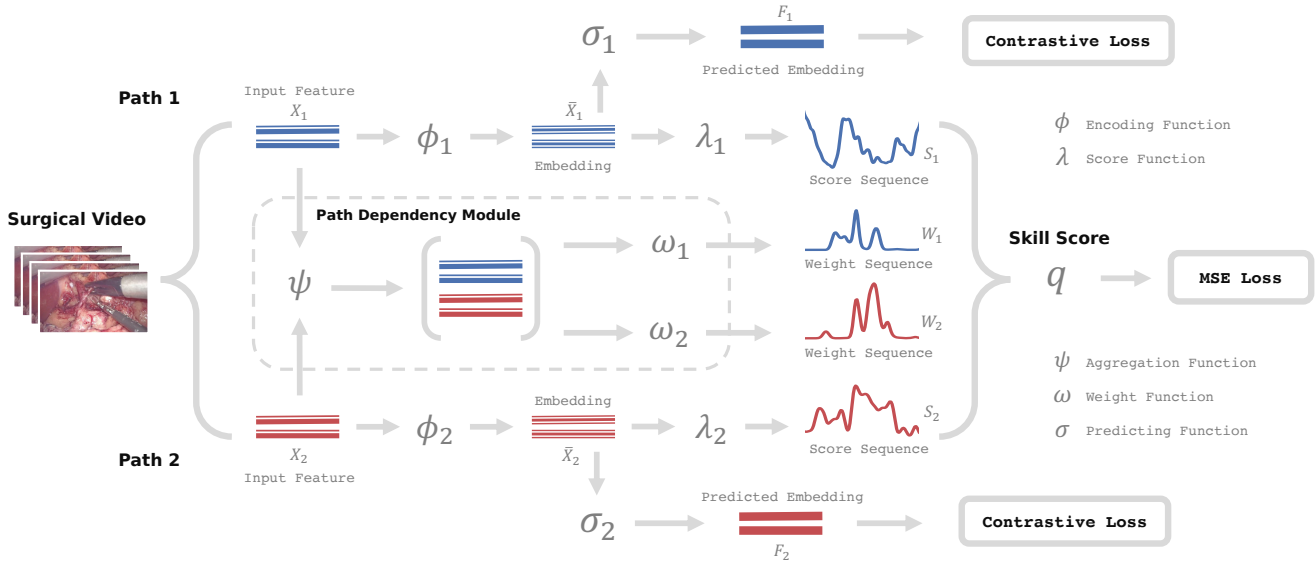

Figure 2. Multi-path framework for unified surgical skill assessment. Two paths are visualized for clarity and four paths are used in practice.

or actions in daily life such as drawing and going upstairs [30, 38, 12, 11]. The quality assessment of simulated surgical tasks is involved in the experiments of several methods above [11, 17, 48, 39, 12], but not as their central focuses. More importantly, medical domain knowledge is not sufficiently incorporated in these general-purpose methods. However, such domain knowledge is crucial for surgical skill assessment.

3. Method

In this section, we present a unified framework for surgical skill assessment, which takes in a surgery video and outputs a skill score. As shown in Fig. 2, our framework comprises multiple assessment paths, an inter-path dependency module, and a contrastive learning mechanism. This section introduces the framework in its general form and detailed instantiations are left to the experiment section 4.3.

3.1. Multi-Path Assessment

To characterize surgical skills from multiple aspects, our framework adopts a multi-path design, in which multiple paths with similar architectures are organized in parallel such that each path concentrates on one skill aspect. Concretely, four paths are included, with three of them corresponding to the previously mentioned skill aspects, *i.e.*, tool, proxy, and event. The rest one is a baseline visual path that evaluates surgical skills directly from semantic visual features, such as the features from pre-trained deep neural networks. The input to each path is a feature sequence extracted from the surgical video, which is intended to supply distinct information specific to each skill aspect. Feature sequences for different paths are of similar shapes:

$$X_m \in \mathbb{R}^{L \times D_m}, m \in \{\mathtt{V}, \mathtt{T}, \mathtt{P}, \mathtt{E}\} \tag{1}$$

where $\mathtt{V}, \mathtt{T}, \mathtt{P}, \mathtt{E}$ denote the visual, tool, proxy, event path respectively, X_m denotes the feature sequence input to the path m, L is the video length and D_m is the feature dimension. The extraction of these features is flexible and can adapt to the dataset, task, application and so on, as long as each focuses on its skill aspect. For example, the input to the event path can be a sequence of occurrence probabilities of surgical events, and the input to the tool path can be a sequence of spatial coordinates of surgical tools.

Along each path m, the feature sequence is first encoded into a high-level embedding sequence:

$$\bar{X}_m = \phi_m(X_m) \tag{2}$$

where ϕ_m represents an encoding function in the path m, and $\bar{X}_m \in \mathbb{R}^{L \times \bar{D}_m}$ is the resultant embedding with size $\bar{D}_m$. Afterward, the embedding sequence is converted into a score sequence $S_m \in \mathbb{R}^{L \times 1}$ indicating predicted surgical skill at each time step:

$$S_m = \lambda_m(\bar{X}_m). \tag{3}$$

The λ_m denotes a score function in the path m. In this way, each path gives an aspect-specific rating of surgical skills.

3.2. Path Dependency Module

In the assessment of surgical skills, the relation among skill aspects matters. Skill predictions in one path contribute to the overall assessment unequally at different time

steps, often depending on the information in other paths. For example, the surgical tool usage could become more important for skill assessment when the field clearness is reduced, and the event occurrence could become less important when no tool appears in the scene. To model such inter-path dependencies, we design a path dependency module in our framework, which mimics the phenomena in these examples by gathering information from all the paths to provide different temporal importance weights for each path. In detail, the importance weight $W_m \in \mathbb{R}^{L\times 1}$ for the path m is also a temporal sequence. To compute the weight W_m, feature sequences are collected from all the paths by an aggregation function ψ and then sent to a weight function ω_m:

$$W_m = \text{softmax}(\omega_m(\psi(X_{\mathtt{V}}, X_{\mathtt{T}}, X_{\mathtt{P}}, X_{\mathtt{E}}))). \tag{4}$$

The aggregation function is shared by paths while the weight functions are not. And a softmax function is imposed to normalize the weight sequence on the temporal dimension. Subsequently, the score sequences, weighted by the temporal importance, are averaged over time and paths to obtain an overall video-level skill score:

$$q = \frac{1}{4}\sum_{m}\sum_{i=0}^{L} S_{m,i}W_{m,i}, m \in \{\mathtt{V},\mathtt{T},\mathtt{P},\mathtt{E}\} \tag{5}$$

where $S_{m,i} \in \mathbb{R}$, $W_{m,i} \in \mathbb{R}$ denote the score and weight in the path m at time i, and $q \in \mathbb{R}$ is the video-level skill prediction. Lastly, a mean squared error (MSE) loss is utilized to supervise the skill prediction:

$$\mathcal{L}_{mse} = (y - q)^2. \tag{6}$$

The y is the ground truth skill annotation, which is within a predefined range of scores.

3.3. Self-Supervised Contrastive Loss

The scarcity of annotated data is a common concern for medical tasks, and surgical skill assessment is no exception. To alleviate this issue, we resort to a contrastive learning strategy. Inspired by the video predictive coding [21], we take future prediction as an auxiliary task to help the model learn temporal dynamics in a self-supervised manner. Concretely, in each path m, a predicting function σ_m is used to forecast the embedding in the future based on the recent past:

$$F_{m,i} = \sigma_m(\bar{X}_{m,i-1}) \tag{7}$$

where $\bar{X}_{m,i-1} \in \mathbb{R}^{\bar{D}_m}$ is the embedding in Eqn. 2 at time $i-1$, and $F_{m,i} \in \mathbb{R}^{\bar{D}_m}$ is the predicted embedding for time i. Then a contrastive loss is designed to encourage the similarity of the predicted embedding $F_{m,i}$ with the real embedding $\bar{X}_{m,i}$ at time i, and discourage its similarity with the embeddings in other time steps:

$$\mathcal{L}_{con} = -\sum_{m}\sum_{i=1}^{L} \log \frac{\exp(F_{m,i}\cdot \bar{X}_{m,i})}{\sum_{j\in\mathcal{N}_i}\exp(F_{m,i}\cdot \bar{X}_{m,j})} \tag{8}$$

where $\cdot$ represents the dot product, and $\mathcal{N}_i$ is a temporal neighborhood around time i including time i. This contrastive loss could assist the encoding function ϕ_m in Eqn. 2 in better capturing temporal dynamics in the surgical video.

Finally, we combine the self-supervised contrastive loss with the supervised MSE loss to train the framework:

$$\mathcal{L}_{full} = \mathcal{L}_{mse} + \mathcal{L}_{con}. \tag{9}$$

4. Experiments

This section first introduces the experimental setup and the implementation of our framework. We then present ablation studies and comparisons to state-of-the-art methods. At last, the correlation between model prediction and input features is examined.

4.1. Datasets

Simulated dataset. Experiments are first performed on the public JIGSAWS dataset [18, 2], which contains three simulated tasks for robotic-assisted surgery, *i.e.*, suturing (SU), needle-passing (NP), and knot-tying (KT). There are 78 egocentric videos for the suturing task, 56 for needle passing, and 72 for knot tying, with 206 videos in total. The duration of the video is 88 seconds on average. Each video is annotated with a skill-level global rating score (GRS) with a range from 6 to 30. We use the GRS as the ground truth of the surgical skill. The JIGSAWS dataset also provides annotations of fine-grained surgical gestures and kinematic data of robotic manipulators.

Clinical dataset. Experiments are also performed on a newly built clinical dataset. This dataset has 20 laparoscopic videos of *in vivo* surgeries for gastric cancer, including partial or total gastrectomy and related lymph node (LN) dissection. The videos have 960×540 resolution and 25 FPS. Different from the simulated JIGSAWS dataset, our dataset is collected from real operating rooms. This new dataset is very challenging, due to its extremely long duration (199 min. per video on average), frequent camera movement, changing illumination, and varying patient conditions. Example frames are given in Fig. 3.

For each video in our dataset, surgical skills are annotated by an expert surgeon on 7 metrics based on a modified OSATS protocol [35]. The global rating score (GRS) is defined as the sum of the 7 metrics. The GRS has a range from 7 to 35 and is used as the ground truth. In Fig. 4, we

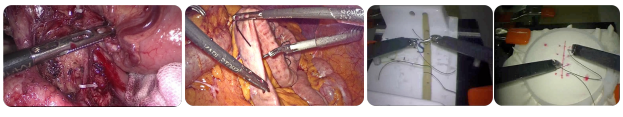

Figure 3. Example video frames.

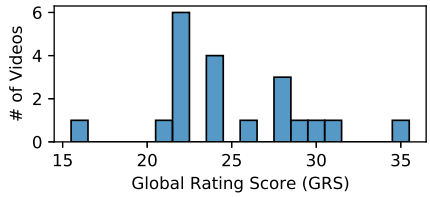

Figure 4. Skill distribution of the new clinical dataset.

# Videos	20
# Average frames per video	299K
Average duration per video	199 min.
# Surgical event classes	41
# Surgical event instances	1565
# Skill metrics	7
# Skill proxy	1

Table 1. Statistics of the new clinical dataset.

Surgical Event	#Ins.
Abdominal cavity exploration	31
Dissection of fusion tissue	19
Dissection of the greater omentum	24
LN dissection of subpyloric region (SR)	22
LN dissection of hepatoduodenal ligament region (HLR)	41
LN dissection of the superior pancreas (SP)	27
LN dissection of lesser curvature (LC)	21
LN dissection of the left gastroepiploic region (LGR)	22
Resection of the distal stomach	20
Specimen removal	20
Gastro-jejunal anastomosis	21
Jejuno-jejunal anastomosis	21
Irrigation and placement of the drains	17
Bleeding	279
Camera out	352

Table 2. Clinical surgical events used in this study.

plot the distribution of the GRS over the 20 videos. This dataset is also annotated with the proxy score of field clearness [31], as well as comprehensive surgical events. There are 41 classes of surgical events in total, including procedural events, adverse events, video events, *etc*. The statistics of our dataset are summarized in Table 1 and more details are given in the supplementary material. We use the 15 event classes listed in Table 2 in this study.

4.2. Experiment Setup

To keep consistent with existing literature, we adopt both four-fold cross-validation (4-Fold) and leave-one-user-out cross-validation (LOUO) when evaluating on JIGSAWS. The four-fold splits provided by [48] and the LOUO splits associated with the JIGSAWS dataset are used. On our clinical dataset, three-fold cross-validation is adopted.

Following prior works, we choose Spearman's rank correlation (SROCC) as the evaluation metric. For the JIGSAWS dataset, the average correlation across the three surgical tasks is computed by Fisher's z-value [40]. Besides, experiment results are averaged over multiple runs.

4.3. Framework Instantiation

4.3.1 Instantiation of input features

Input features to the paths in our framework are instantiated differently to carry aspect-specific information as follows.

Visual path input $X_{\mathtt{V}}$. We leverage the semantic features extracted from ResNet-101 [22] pre-trained on ImageNet as $X_{\mathtt{V}}$. The feature dimension $D_{\mathtt{V}}$ is 2048.

Tool path input $X_{\mathtt{T}}$. On the clinical dataset, to capture the surgical tool movement, we first apply an unsupervised tool segmentation method [32]. Then $X_{\mathtt{T}}$ is spatial histograms of the segmentation masks. Specifically, the mask in each frame is divided into 3×3, 4×4, and 5×5 spatial grids. The percentage of pixels belonging to surgical tools in each grid cell is taken as the feature. The feature dimension $D_{\mathtt{T}}$ thus equals $9 + 16 + 25 = 50$. Since the segmentation method is unsupervised, $X_{\mathtt{T}}$ does not involve extra data or annotations. On the simulated dataset, the robotic kinematic data associated with the dataset is used as $X_{\mathtt{T}}$. We use $D_{\mathtt{T}} = 14$ dimensions, including the 3D positions, 3D velocities, and gripper angles of the two patient-side manipulators.

Proxy path input $X_{\mathtt{P}}$. On the clinical dataset, the field clearness is used as a skill proxy. The frame-level scores of the proxy are extracted from a re-implementation of [31] as $X_{\mathtt{P}}$. The training and testing of method [31] follow the same cross-validation settings as in Section 4.2. Since the field clearness only works for clinical data, we employ another simple skill proxy for the simulated dataset, *i.e*., task completion time [11]. The $X_{\mathtt{P}}$ is set as a sequence with a constant value inversely proportional to the video length. On the simulated dataset, $X_{\mathtt{P}}$ does not involve extra data or annotations. $D_{\mathtt{P}}$ is 1 on both datasets.

Event path input $X_{\mathtt{E}}$. For the event path, we train Multi-Stage Temporal Convolutional Networks (MS-TCN) [14] to detect surgical events on the clinical dataset or surgical gestures on the simulated dataset. The event detection models are trained under the same cross-validation settings as in Section 4.2. The $X_{\mathtt{E}}$ is then set as the frame-level probabilities of events or gestures. Its dimension $D_{\mathtt{E}}$, which equals the number of classes, is 10 for suturing, 8 for needle-passing, 6 for knot-tying on the simulated dataset, or 15 on the clinical dataset.

4.3.2 Instantiation of functions

The aggregation function ψ is instantiated with the concatenation of the feature dimension. For all paths except the

Method ↓	Clinical 3-Fold	Simulated 4-Fold SU	NP	KT	Avg.
Ours (V)	0.201	0.642	0.666	0.729	0.681
Ours (T)	0.250	0.765	0.566	0.662	0.673
Ours (P)	0.469	0.396	0.333	0.803	0.554
Ours (E)	0.241	0.603	0.200	0.762	0.560
Ours (VT)	0.268	0.735	0.737	0.706	0.726
Ours (VTP)	0.525	0.791	**0.761**	0.784	0.779
Ours (VTPE)	**0.565**	**0.834**	0.756	**0.819**	**0.805**

Table 3. Ablation study on framework paths.

Method ↓	Clinical 3-Fold	Simulated 4-Fold SU	NP	KT	Avg.
Full Model	**0.565**	0.834	**0.756**	**0.819**	**0.805**
Without ω	0.509	0.808	0.681	0.754	0.752
Without ψ	0.522	0.766	0.555	0.777	0.712
Without σ	0.414	**0.853**	0.676	0.797	0.786

Table 4. Ablation study on framework components.

Clinical 3-Fold (mAP %)	Simulated 4-Fold (Acc. %) SU	NP	KT	Avg.
75.4	88.8	78.3	85.4	84.2

Table 5. Mean average precision (mAP) of surgical event detection and accuracy of surgical gesture detection.

proxy path, the encoding functions ϕ are instantiated with Temporal Convolutional Networks (TCN) [28]. The score functions λ, weight functions ω, predicting functions σ are chosen as frame-wise multilayer perceptrons (MLP). For the proxy path, since the input feature already represents the skill, the encoding function $\phi_{\mathtt{P}}$ and the score function $\lambda_{\mathtt{P}}$ are set as identity functions, and the weight function $\omega_{\mathtt{P}}$ is a constant function giving uniform weights at all time steps. The contrastive loss and the predicting function are removed from the proxy path.

4.3.3 Other implementation details

The proposed framework is implemented using PyTorch [43]. Model parameters are trained using mini-batch stochastic gradient descent with the Adam optimizer [27]. The embedding sizes $\bar{D}_{\mathtt{V}}, \bar{D}_{\mathtt{T}}, \bar{D}_{\mathtt{P}}, \bar{D}_{\mathtt{E}}$ are set to $20, 4, 1, 4$ respectively. We freeze the input features after extraction. This allows the model to have a larger temporal receptive field covering more video frames, which is necessary when handling the extremely long clinical videos and when learning long-term event patterns. The simulated videos are sampled at 5 FPS and the clinical videos are sampled at 0.5 FPS. The GRS is normalized within 0 and 1 during training. Our codes will be released to offer other details.

Method	SROCC
USDL [48]	0.161
Ours (VT)	**0.268**
MICCAI 2019 [31]	0.469
Ours (VTP)	**0.525**
Ours (VTPE)	**0.565**

Table 6. Comparison to the state-of-the-art on our clinical dataset. Methods in the same vertical slot can be directly compared.

Method	Input	SU	NP	KT	Avg.
USDL [48]	$\mathbb{V}$	0.64	0.63	0.61	0.63
Ours (VP)	$\mathbb{V}$	**0.68**	**0.71**	**0.80**	**0.73**
MUSDL [48] $*$	$\mathbb{V}$	0.71	0.69	0.71	0.70
ST-GCN [39, 59]	$\mathbb{VK}$	0.31	0.39	0.58	0.43
TSN [39, 11, 53]	$\mathbb{VK}$	0.34	0.23	0.72	0.46
JRG [39]	$\mathbb{VK}$	0.36	0.54	0.75	0.57
AIM [17]	$\mathbb{VK}$	0.63	0.65	**0.82**	0.71
Ours (VTP)	$\mathbb{VK}$	**0.79**	**0.76**	0.78	**0.78**
Ours (VTPE)	$\mathbb{VK}$	**0.83**	**0.76**	**0.82**	**0.80**

Table 7. Comparison to the state-of-the-art methods on the simulated dataset under the 4-Fold setting. Methods in the same vertical slot can be directly compared. $\mathbb{V}$: Surgical videos. $\mathbb{K}$: Robotic kinematics. $*$: Extra fine-grained skill annotations are used.

4.4. Ablation Studies

Effects of framework paths. A set of comparative experiments are performed on both datasets to inspect the effect of each skill aspect. The results of using every single path and combinations of multiple paths are reported in Table 3. When using a single path, the proxy path achieves higher results than other paths on the clinical data. In general, combining multiple paths improves performance. On both two datasets, the best average performance is obtained when all the paths are included. It is noticed that the event path performs badly on the needle-passing task, probably due to the inferior gesture detection accuracy on this task. The accuracy of gesture detection and the mean average precision (mAP) of event detection are reported in Table 5.

Effects of framework components. In Table 4, the results of our framework with one of the following components removed are presented: 1) weight functions ω, 2) aggregation function ψ, 3) predicting function σ. The contrastive loss is also discarded when the σ is removed. In general, all components contribute positively to the best performance. Besides, the predicting function σ for contrastive learning is especially important when the data is highly scarce, *i.e.*, on the clinical data. On the simulated suturing task, the contrastive learning yields no improvement possibly because sufficient training data is available. The encoding and score functions are not ablated over since they lie in the backbone of our framework.

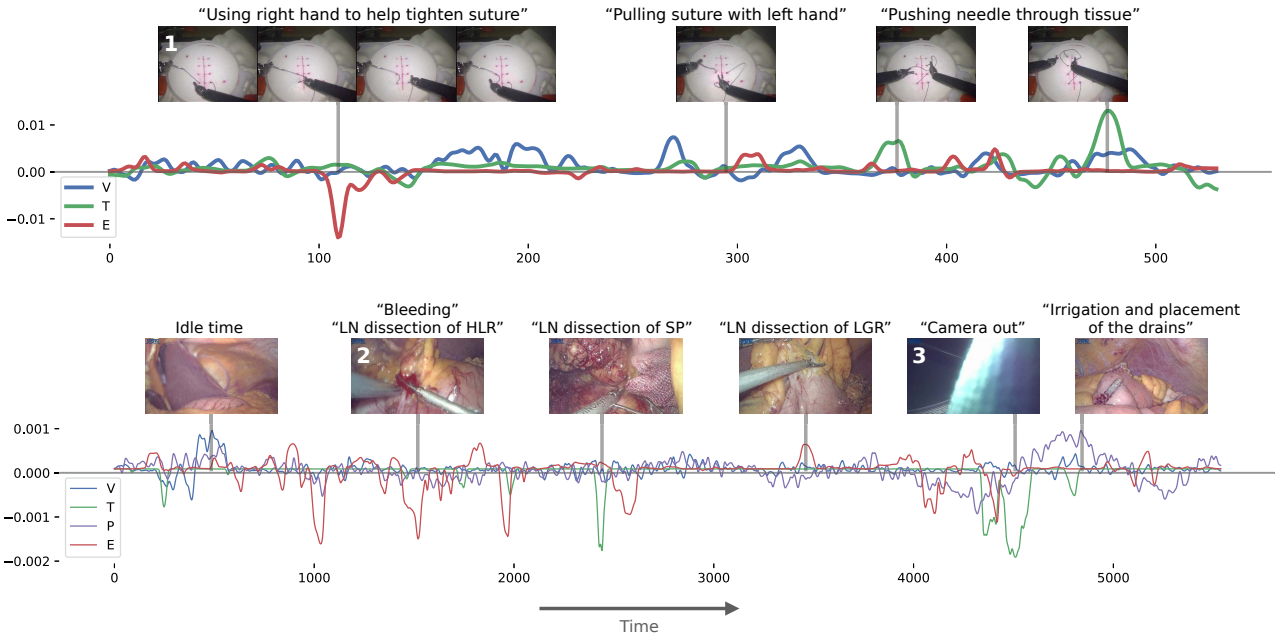

Figure 5. Result visualization. The upper part shows the result on a simulated surgery and the lower part on a clinical surgery. The weighted score sequences $S_m W_m$ from each path $m \in \{\mathtt{V}, \mathtt{T}, \mathtt{P}, \mathtt{E}\}$ are visualized. The higher score reflects the better surgical skill and vice versa. Corresponding surgical gestures or events are marked on the selected frames. $S_\mathtt{P} W_\mathtt{P}$ is not plotted for the simulated surgery because it is a constant sequence. Note that for each sequence the integral over time equals a video-level skill prediction.

Method	Input	SU	NP	KT	Avg.
DTC+DFT+ApEn [63]	$\mathbb{K}$	0.37	0.25	**0.60**	0.41
Ours (TP)	$\mathbb{K}$	**0.40**	**0.63**	0.55	**0.53**
JRG [39]	$\mathbb{VK}$	0.35	**0.67**	0.19	0.40
AIM [17]	$\mathbb{VK}$	**0.45**	0.34	**0.61**	0.47
Ours (VTP)	$\mathbb{VK}$	**0.45**	0.62	0.58	**0.56**
MTL-VF (ResNet) [54] $*$	$\mathbb{V}$	0.68	0.48	0.72	0.64
MTL-VF (C3D) [54] $*$	$\mathbb{V}$	0.69	0.86	0.83	0.80
Ours (VTPE)	$\mathbb{VK}$	**0.45**	0.65	0.59	**0.57**

Table 8. Comparison to the state-of-the-art methods on the simulated dataset under the LOUO setting. Methods in the same vertical slot can be directly compared. $\mathbb{V}$: Surgical videos. $\mathbb{K}$: Robotic kinematics. $*$: Extra surgical experience annotations are used.

4.5. Comparisons to State-of-the-Art

Clinical dataset. Due to the lack of existing results on our clinical dataset, two state-of-the-art methods are implemented to compare with, *i.e.*, USDL [48] and MICCAI 2019 [31]. Specifically, we run the public code of USDL and re-implement the method [31]. Moreover, for fairness, paths involving extra annotations are removed from our framework respectively in each comparison. We remove the proxy and event paths while comparing with USDL, and remove the event path while comparing with MICCAI 2019. Our method achieves promising performance as shown in Table 6.

Simulated dataset. Table 7 and Table 8 show the comparisons between the experimental results of our method and other approaches on the simulated dataset. Note that previous methods could adopt different modalities as input, some of which use extra annotations. Therefore, methods using the same data and annotation are grouped together for comparison. When comparing our framework with others, we remove the paths involving extra annotations. In each comparison, our method outperforms other counterparts respectively.

4.6. Visualization

We choose two videos from the simulated dataset and our clinical dataset to visualize the weighted score sequences $S_m W_m \in \mathbb{R}^{L\times 1}$ for each path $m \in \{\mathtt{V}, \mathtt{T}, \mathtt{P}, \mathtt{E}\}$ in Figure 5. For example, in the video from the simulated dataset, frames with number *1* record a miss of right robot hand while the surgeon is tightening suture, which leads to a simultaneous fall of event score. This error causes event repetition and interrupts the normal workflow of suturing. On the other hand, in the video from our clinical dataset, the frame with number *2* presents a detected bleeding event and a concurrent fall of event score. In addition, the frame with number *3* shows a camera out event with low tool scores. The camera out event likely corrupts the surgical tool segmentation algorithm, which can be regarded as a failure

case. A video demo is attached in the supplementary.

5. How the Model Understands Surgical Skills

To investigate what the model learns about surgical skills, the model outputs are analyzed quantitatively. Note that some input feature sequences are of medical or physical meanings, such as the $X_{\mathtt{E}}$ on the simulated dataset indicating the probabilities of surgical gestures. Therefore, we can get some insights by temporally correlating the weighted score sequences $S_m W_m \in \mathbb{R}^{L\times 1}$ to these meaningful input feature sequences. If take $X_{\mathtt{E}}$ on the simulated dataset as an example, the correlation between model outputs and a channel c in $X_{\mathtt{E}}$ is defined as $R_{\mathtt{E}}^{(c)}$ by the following equation:

$$R_{\mathtt{E}}^{(c)} = \frac{1}{3}\sum_{m} |\mathrm{srocc}(S_m W_m, X_{\mathtt{E}}^{(c)})|, m \in \{\mathtt{V}, \mathtt{T}, \mathtt{E}\} \quad (10)$$

where $X_{\mathtt{E}}^{(c)} \in \mathbb{R}^{L\times 1}$ is the selected channel of the input. SROCCs between this channel and weighted score sequences from different paths are computed and averaged. The resultant $R_{\mathtt{E}}^{(c)} \in [0, 1]$ indicates to what extent the skill predictions are correlated to the surgical gesture c. In the equation above, we take the absolute value of the SROCC since we care about the correlation regardless of whether it is positive or negative. The $S_{\mathtt{P}} W_{\mathtt{P}}$ is excluded from the computation since it is a constant sequence. Apart from $X_{\mathtt{E}}$, the tool feature sequence $X_{\mathtt{T}}$ also has physical meanings on the simulated data and $R_{\mathtt{T}}$ can be similarly computed.

The bar plots of $R_{\mathtt{E}}$ and $R_{\mathtt{T}}$ for the simulated suturing task are given in Fig. 6. For $R_{\mathtt{E}}$, it is interesting that the model outputs are most correlated to the surgical gesture *"Pushing needle through tissue"*, which is also an intuitively critical gesture for suturing. Model outputs are less correlated to transitional gestures such as *"Moving to center with needle in grip"* and *"Moving to end points"*. For $R_{\mathtt{T}}$, it is noticed that the model outputs are most correlated to the gripper angles and the position-z of the left manipulator, which are also highly active factors during suturing in practice. These findings are consistent with the human understanding of suturing.

Similarly, the $R_{\mathtt{E}}$ on the clinical data is also computed and visualized in Fig. 7. The $S_{\mathtt{P}} W_{\mathtt{P}}$ is included in the computation now. Currently, no remarkable correlation between the surgical event and the model output is observed, with all correlations are lower than 0.1. This demonstrates the simulated-clinical gap and thus the importance of clinical data. A larger number of clinical surgeries may lead to more evident findings in the future. More results are in the supplementary.

6. Conclusion and Future Work

This paper proposes a flexible and general framework to automatically assess surgical skills from multiple aspects.

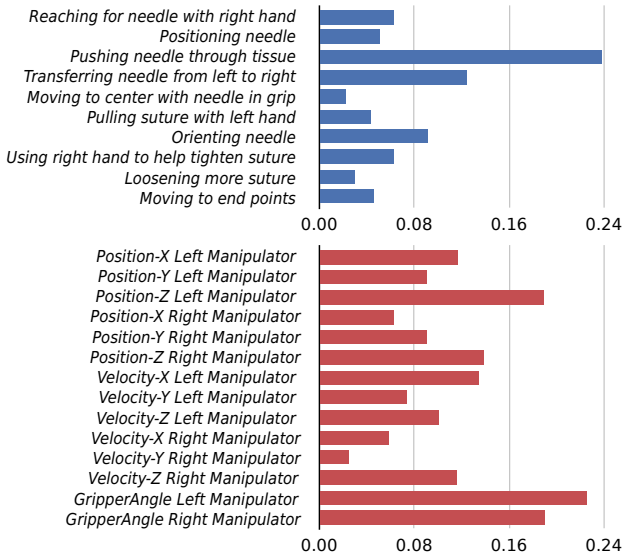

Figure 6. Blue: Correlations between model outputs and surgical gestures on the simulated suturing ($R_{\mathtt{E}}$). Red: Correlations between model outputs and tool features on the simulated suturing ($R_{\mathtt{T}}$).

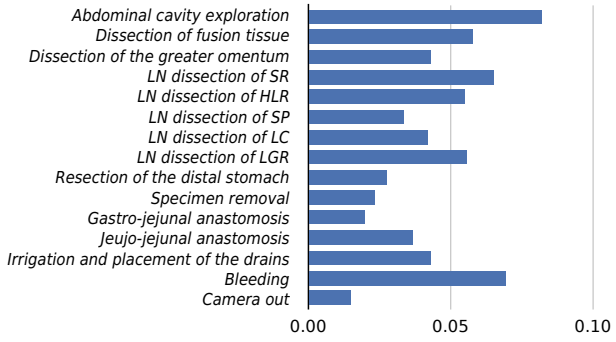

Figure 7. Correlations between model outputs and surgical events on the clinical data ($R_{\mathtt{E}}$).

The effectiveness of the proposed framework is validated by the experiments on both simulated and clinical surgery datasets. Within this framework, future works could focus on more advanced input features and composing functions. Our framework is also extendable to include more skill aspects beyond those used in this study. Future works could also research the flexible choice of paths and the adaptive fusion of paths. Besides, it is also desirable to have more data from the clinical environment for surgical skill assessment in the future.

Acknowledgments. This work was partially supported by NSFC-61625201 and NSFC-62061136001. We also acknowledge the Clinical Medicine Plus X-Young Scholars Project, and High-Performance Computing Platform of Peking University for providing computational resources.

References

[1] Endoscopic vision challenge 2019: Surgical workflow and skill analysis. 2

[2] Narges Ahmidi et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE TBE*, 2017. 2, 4

[3] Narges Ahmidi, Yixin Gao, Benjamín Béjar, S Swaroop Vedula, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In *MICCAI*, 2013. 2

[4] Narges Ahmidi, Piyush Poddar, Jonathan D Jones, S Swaroop Vedula, Lisa Ishii, Gregory D Hager, and Masaru Ishii. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *IJCARS*, 10(6):981–991, 2015. 2

[5] David P Azari, Lane L Frasier, Sudha R Pavuluri Quamme, Caprice C Greenberg, Carla M Pugh, Jacob A Greenberg, and Robert G Radwin. Modeling surgical technical skill using expert assessment for automated computer rating. *Annals of surgery*, 269(3):574–581, 2019. 2

[6] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am I a baller? basketball performance assessment from first-person videos. In *CVPR*, 2017. 2

[7] Vinay Bettadapura, Grant Schindler, Thomas Plötz, and Irfan Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *CVPR*, 2013. 2

[8] John D Birkmeyer et al. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 2013. 1

[9] Dayvid Castro, Danilo Pereira, Cleber Zanchettin, David Macêdo, and Byron LD Bezerra. Towards optimizing convolutional neural networks for robotic surgery skill evaluation. In *IJCNN*, 2019. 2

[10] Jian Chen et al. Objective assessment of robotic surgical technical skill: a systematic review. *The Journal of urology*, 201(3):461–469, 2019. 1

[11] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *CVPR*, June 2018. 3, 5, 6

[12] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR*, June 2019. 3

[13] Marzieh Ershad, Zachary Koesters, Robert Rege, and Ann Majewicz. Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In *MICCAI*, 2016. 2

[14] Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, June 2019. 5

[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *MICCAI*, 2018. 2

[16] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3D convolutional neural networks. *IJCARS*, 14(7):1217–1225, 2019. 2

[17] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. An asymmetric modeling for action assessment. In *ECCV*, 2020. 2, 3, 6, 7

[18] Yixin Gao et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2CAI*, 2014. 2, 4

[19] Ahmad Ghasemloonia, Yaser Maddahi, Kourosh Zareinia, Sanju Lama, Joseph C Dort, and Garnette R Sutherland. Surgical skill assessment using motion quality and smoothness. *Journal of surgical education*, 74(2):295–305, 2017. 1

[20] Andrew S Gordon. Automated video assessment of human performance. In *Proceedings of AI-ED*, 1995. 2

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 4

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[23] Matthew S Holden et al. Machine learning methods for automated technical skills assessment with instructional feedback in ultrasound-guided interventions. *IJCARS*, 14(11):1993–2003, 2019. 2

[24] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Fei-Fei Li. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *WACV*, 2018. 2

[25] Amod Jog, Brandon Itkowitz, May Liu, Simon DiMaio, Greg Hager, Myriam Curet, and Rajesh Kumar. Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. In *ICRA*, 2011. 2

[26] Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič. Trajectory based assessment of coordinated human activity. In *International Conference on Computer Vision Systems*, 2003. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[28] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, July 2017. 6

[29] Yongjun Li, Xiujuan Chai, and Xilin Chen. ScoringNet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In *ACCV*, 2018. 2

[30] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. Manipulation-skill assessment from videos with spatial attention network. In *ICCV Workshops*, 2019. 3

[31] Daochang Liu, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In *MICCAI*, 2019. 2, 5, 6, 7

[32] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In *MICCAI*, 2020. 5

[33] Francisco Luongo, Ryan Hakim, Jessica H Nguyen, Animashree Anandkumar, and Andrew J Hung. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery*, 2020. 2

[34] Anand Malpani, S Swaroop Vedula, Chi Chiung Grace Chen, and Gregory D Hager. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In *International Conference on Information Processing in Computer-Assisted Interventions*, 2014. 2

[35] JA Martin, Glenn Regehr, Richard Reznick, Helen Macrae, John Murnaghan, Carol Hutchison, and M Brown. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, 1997. 1, 4

[36] John D Mason, James Ansell, Neil Warren, and Jared Torkington. Is motion analysis a valid tool for assessing laparoscopic skill? *Surgical endoscopy*, 27(5):1468–1477, 2013. 1

[37] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. Falcons: Fast learner-grader for contorted poses in sports. In *CVPR Workshops*, 2020. 2

[38] Adeline Paiement, Lili Tao, Sion Hannuna, Massimo Camplani, Dima Damen, and Majid Mirmehdi. Online quality assessment of human movement from skeleton data. In *BMVC*, 2014. 3

[39] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, October 2019. 2, 3, 6, 7

[40] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, 2019. 2, 5

[41] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *CVPR*, 2019. 2

[42] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPR Workshops*, 2017. 2

[43] Adam Paszke et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[44] Fernando Pérez-Escamirosa et al. Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. *IJCARS*, 15(1):27–40, 2020. 2

[45] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, 2014. 2

[46] Carol E Reiley and Gregory D Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In *MICCAI*, 2009. 2

[47] Richard K Reznick. Teaching and testing technical skills. *The American journal of surgery*, 165(3):358–361, 1993. 1

[48] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, June 2020. 2, 3, 5, 6, 7

[49] Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D Hager, and René Vidal. Sparse hidden markov models for surgical gesture classification and skill evaluation. In *International Conference on Information Processing in Computer-Assisted Interventions*, 2012. 2

[50] Munenori Uemura et al. Procedural surgical skill assessment in laparoscopic training environments. *IJCARS*, 11(4):543–552, 2016. 2

[51] Balakrishnan Varadarajan, Carol Reiley, Henry Lin, Sanjeev Khudanpur, and Gregory Hager. Data-derived models for segmentation with application to surgical assessment and training. In *MICCAI*, 2009. 2

[52] S Swaroop Vedula, Masaru Ishii, and Gregory D Hager. Objective assessment of surgical technical skill and competency in the operating room. *Annual Review of Biomedical Engineering*, 2017. 1

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 6

[54] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *MICCAI*, 2020. 2, 7

[55] Ziheng Wang and Ann Majewicz Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *IJCARS*, 13(12):1959–1970, 2018. 2

[56] Kyle R Wanzel, Myléne Ward, and Richard K Reznick. Teaching the surgical craft: from selection to certification. *Current problems in surgery*, 39(6):583–659, 2002. 1

[57] Thomas G Weiser et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *The Lancet*, 385:S11, 2015. 1

[58] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE TCSVT*, 2019. 2

[59] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6

[60] Marieke Zegers, Martine C de Bruijne, Bertus de Keizer, Hanneke Merten, Peter P Groenewegen, Gerrit van der Wal, and Cordula Wagner. The incidence, root-causes, and outcomes of adverse events in surgical units: implication for potential prevention strategies. *Patient safety in surgery*, 5(1):13, 2011. 2

[61] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM MM*, 2020. 2

[62] Qiang Zhang and Baoxin Li. Relative Hidden Markov Models for video-based evaluation of motion skills in surgical training. *IEEE TPAMI*, 37(6):1206–1218, 2014. 2

[63] Aneeq Zia and Irfan Essa. Automated surgical skill assessment in RMIS training. *IJCARS*, 13(5):731–739, 2018. 2, 7

[64] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan Essa. Automated assessment of surgical skills using frequency analysis. In *MICCAI*, 2015. 2

[65] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, and Irfan Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *IJCARS*, 13(3):443–455, 2018. 2