SEDSkill: Surgical Events Driven Method for Skill Assessment from Thoracoscopic Surgical Videos

Xinpeng Ding[1], Xiaowei Xu[2(✉)], and Xiaomeng Li[1(✉)]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR, China
eexmli@ust.hk

[2] Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
xiao.wei.xu@foxmail.com

Abstract. Thoracoscopy-assisted mitral valve replacement (MVR) is a crucial treatment for patients with mitral regurgitation and demands exceptional surgical skills to prevent complications and enhance patient outcomes. Consequently, surgical skill assessment (SKA) for MVR is essential for certifying novice surgeons and training purposes. However, current automatic SKA approaches have inherent limitations that include the absence of public thoracoscopy-assisted surgery datasets, exclusion of inter-video relationships, and limited to SKA of a single short surgical action. This paper introduces a novel clinical dataset for MVR, which is the first thoracoscopy-assisted long-form surgery dataset to the best of our knowledge. Our dataset, unlike existing short video clips that contain single surgical action, includes videos of the whole MVR procedure that capture multiple complex skill-related surgical events. To tackle the challenges posed by MVR, we propose a novel method called **S**urgical **E**vents **D**riven **S**kill assessment (SEDSkill). Our key idea is to develop a long-form surgical events-driven method for skill assessment, which is based on the insight that the skill level of a surgeon is closely tied to the occurrence of inappropriate operations such as excessively long suture repairing times. SEDSkill incorporates an event-aware module that automatically localizes skill-related events, thus extracting local semantics from long-form videos. Additionally, we introduce a difference regression block to learn imperceptible discrepancies, which enables precise and accurate surgical skills assessment. Extensive experiments demonstrate that our proposed method outperforms state-of-the-art approaches. Our code is available at https://github.com/xmed-lab/SEDSkill.

Keywords: Surgical skill assessment · Long-form video · Thoracoscopy-assisted surgery

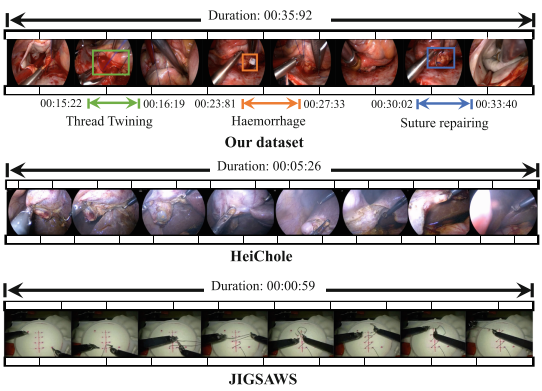
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 35–45, 2023.
https://doi.org/10.1007/978-3-031-43996-4_4

1 Introduction

Thoracoscopy-assisted mitral valve replacement (MVR) has become routine for the treatment of mitral valve regurgitation [4]. Compared to other surgeries such as laparoscopic operations, thoracoscopy-assisted MVR requires higher surgical skills due to the intricate structure of the heart, the mitral valve's proximity to other vital cardiac structures, and the geometric limitations of the surgical field [11]. Improving surgical skills can prevent avoidable complications [9], leading to better patient outcomes, such as improved long-term survival and reduced postoperative complications [2,3]. Therefore, surgical skill assessment (SKA), *i.e.*, evaluating the skill level of surgeons, is essential in the training and certification of novice surgeons [19,20,26].

Traditionally, SKA has been reliant on manual observation by experienced surgeons either in the operating room or via recorded videos, as described by Reznick in his work on teaching [19]. However, this method is subjective, time-consuming, and not very efficient for use in surgical education. To address these limitations, researchers have increasingly focused on developing automatic SKA tools. While current automatic SKA approaches [10,12,16,17,22] have demonstrated success on simulated and laparoscopic datasets, their application to thoracoscopy-assisted MVR poses several challenges. First, to the best of our knowledge, there are no publicly available clinical datasets for thoracoscopy-assisted surgery. Second, most existing methods [5–7,10,12,16,17] focus solely on the global information within a single video to perform SKA, such as regressing a singular skill score from the video. However, these methods disregard the inter-video information, such as subtle differences between various videos, that could be critical in predicting surgical skill scores [13]. For instance, differences in haemorrhage loss, suture repairing times, and thread twining times among videos can have a significant impact on the final scores. Generally, more thread winding, haemorrhage, and suture repairing can indicate a lower skill level; see Fig. 1(a).

To address the above challenges, we collect a new dataset for SKA, which is the first-ever long-form thoracoscopy-assisted MVR video dataset. Our dataset offers longer video duration and more surgical events with corresponding labels in comparison to the currently available public datasets such as JIGSAWS [8] or HeiChole [24]; see Fig. 1(b). Then, we present a novel **S**urgical **E**vents **D**riven **S**kill assessment (**SEDSkill**) method to address the limitations of current automatic methods for MVR assessment. Unlike prior work [10,12,17], our key idea is to develop a long-form surgical events-driven method for skill assessment, which is based on the crucial insight that the skill level of a surgeon is closely tied to the occurrence of inappropriate operations such as excessively long suture repairing times. To achieve it, we propose a **novel local-global difference method** that can *learn inter-video relations between both the global long-form and local surgical events correlated semantics*. The method includes an event-aware module and a difference regression module. The event-aware module can automatically localize skill-related surgical events and extract their corresponding features to represent the local event semantics. As surgical skill is highly correlated with



(a)

	Ours	HeiChole	JIGSAWS
Type	Clinic	Clinic	Simulated
Number of videos	50	48	206
Duration per video	35 min	5 min	88 sec
Number of skill events per video	Multiple	Single	Single
Annotation	(1) Skill score; (2) Category of skill events; (3) Start and end time for skill events;	Skill score	Skill score

(b)

Fig. 1. (a) Visualization of different surgical skill datasets: Our collected dataset, HeiChole [24] and JIGSAWS [8]. (b) Comparison of different datasets. Unlike other datasets where each video typically contains only one surgical action (e.g., dissection or suturing), our MVR dataset provides long-form videos with multiple skill-related events and their corresponding labels.

the occurrence of inappropriate events, this module is crucial for precise SKA. To enable the accurate detection of slight differences between videos, our difference regression module captures the relationships among videos and enhances the model's ability to detect subtle variations. By incorporating video-wise and event-wise difference learning, our framework can capture both local and global inter-video relations, thereby enabling precise SKA.

In summary, our contributions are three-fold: (1) We introduce a novel SEDSkill method that aims to design a long-form, surgical events-driven approach for SKA, and it is the first method designed specifically for SKA in thoracoscopic surgical videos. (2) We propose a local-global difference framework that can learn inter-video relations between both the global long-form and local surgical events correlated semantics, thereby enabling enhanced SKA performance. (3) Experimental results demonstrate that our method outperforms existing SKA methods, as well as methods designed for video quality assessment in computer vision. This indicates the great potential of our method for use in clinical practice. Our code will be publicly released upon paper acceptance.

2 Method

Figure 2 illustrates our SEDSkill framework for SKA, which takes a surgical video as input and regresses a surgical skill score. Our proposed framework consists of two main modules: (a) a basic regression module to output the skill score for each input, (b) a local-global difference module to learn both video-level and event-level inter-video differences for precise assessment.

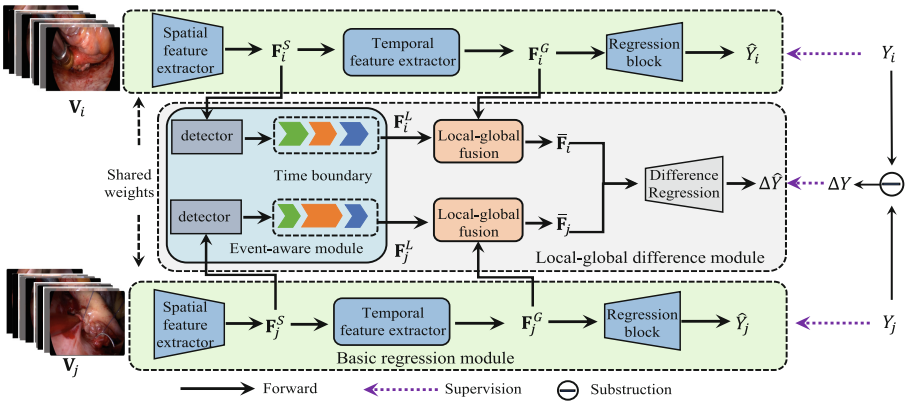


Fig. 2. Illustration of our proposed SEDSkill. SEDSkill consists of two main components: a basic regression module and a local-global difference module.

2.1 Basic Regression Module

The basic regression module aims to regress a surgical skill score, *i.e.*, Y_i, for a raw surgical video input $\mathbf{V}_i \in \mathbb{R}^{T_i \times H \times W}$, where T_i is the duration, H and W are the height and width of each frame. As shown in Fig. 2, the basic regression module consists of three components: a spatial feature extractor, a temporal feature extractor, and a regression block. Specifically, the video $\mathbf{V}_i$ is first fed into the spatial feature extractor to obtain the spatial feature, denoted as $\mathbf{F}_i^S \in \mathbb{R}^{T_i \times D_s}$, where D_s is the dimension of features, followed by a temporal feature extractor to model the intra-video relations to generate the global video feature $\mathbf{F}_i^G \in \mathbf{R}^{T \times D_t}$. Finally, a regression block consisting of several convolutional, max-pooling layers and a fully connected layer is applied to map $\mathbf{F}_i^G$ to the skill score $\hat{Y}_i$. Then, the loss function is to minimize the differences between predicted $\hat{Y}_i$ and the ground-truth Y_i as follows:

$$\mathcal{L}_{reg} = 1/N \sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2, \tag{1}$$

where N is the number of videos.

2.2 Local-Global Difference Module

Surgical Event-Aware Module for Local Information. Unlike prior datasets used for skill assessment, our MVR video dataset is much longer, ranging from 30 min to 1 h, and consists of multiple skill-related events such as thread twining, haemorrhage, and suture repairing; see Fig. 1. As surgical skill is highly correlated to the qualities of surgical events, directly using the basic regression module to predict a score from the long video would consider too many irrelevant parts, thus degrading the regression performance. To encourage the model

to focus on the skill-related parts and remove the less informative ones, we devise an event-aware module to localize the skill-related events, *i.e.*, haemorrhage, surgical thread twining and surgical suture repair, from the long videos and extract the local event-level features, as shown in Fig. 2(b).

Specifically, given the spatial feature, *e.g.*, $\mathbf{F}_i^S$, we introduce an event detector, which is a transformer-like network that maps the video features to the classified logits and regressed start/end time. The detailed architecture can refer to [28]. Formally, the prediction of the detector is a set for each time t which can be formulated as:

$$\mathcal{G}_t = \{\mathbf{p}_t, d_t^s, d_t^e\}, \tag{2}$$

where $\mathbf{p}_t \in \mathbb{R}^4$ consists of 4 values (including the background), which indicates the probability of event category, $d_t^s > 0$ and $d_t^e > 0$ denote the distance between time t to start and end time of events. Note that if p_t equals to zero, $d_t^s > 0$ and $d_t^e > 0$ are not defined. Then, following [28] the loss function for the detector is defined as:

$$\mathcal{L}_{det} = \sum_t \left(\mathcal{L}_{cls} + \lambda_{loc}\mathbf{1}_{c_t}\mathcal{L}_{loc}\right)/N_+, \tag{3}$$

where N_+ is the number of positive frames, $\mathcal{L}_{cls}$ is a focal loss [15] and $\mathcal{L}_{loc}$ is a DIoU loss [29]. $\mathbf{1}_{c_t}$ is the indicator function to identify where the time t is within a event. λ_{loc} is set to 1 following [28]. Note that the detector is pre-trained and fixed during the training of the basic regression module and the local-global difference module. After obtaining the pre-trained detector, we generate the event confidence map for each video denoted by $\mathbf{A}_i = [a_t]_{t=1}^{T_i}$, where $\mathbf{A}_i \in \mathbb{R}^{T_i \times 1}$, $a_t = \max \mathbf{p}_t$ is the confidence for each time t and T_i is the duration for the $\mathbf{V}_i$. Then, the local event-level feature is obtained by the multiplication of the confidence values and the global video feature, *i.e.*, $\mathbf{F}_i^L = \mathbf{A}_i \circ \mathbf{F}_i^S$, where $\mathbf{F}_i^L \in \mathbb{R}^{T_i \times D_s}$ and $\circ$ is the element-wise multiplication.

Local-Global Fusion. We introduce the local-global fusion module to aggregate the local (*i.e.*event-level) and global (long-form video) semantics. Formally, we can define the local-global fusion as $\overline{\mathbf{F}}_i = \text{Fusion}(\mathbf{F}_i^G, \mathbf{F}_i^L)$, where $\overline{\mathbf{F}}_i \in \mathbb{R}^{T_i \times (D_t + D_s)}$. This module can be implemented by different types and we will conduct an ablation study to analyze the effect of this module in Table 3.

Difference Regression Block. Most surgeries of the same type are performed in similar scenes, leading to subtle differences among surgical videos. For example, in MVR, the surgeon first stitches two lines on one side using a needle, and then passes one of the lines through to the other side, connecting it to the extracorporeal circulation tube. Although these procedures are performed in a similar way, the imperceptible discrepancies are very important for accurately assessing surgical skills. Hence, we first leverage the relation block to capture the inter-video semantics. We use the features of the pairwise videos, *i.e.*, $\overline{\mathbf{F}}_i$ and $\overline{\mathbf{F}}_j$, for clarity. Since attention [5,23] is widely used for capturing relations, we formulate the detailed relation block in the attention manner as follows:

$$\overline{\mathbf{F}}_{j\to i} = \text{Attention}\left(\mathbf{Q}_j; \mathbf{K}_i; \mathbf{V}_i\right) = \text{softmax}\left(\frac{\mathbf{Q}_j {\mathbf{K}_i}^\top}{\sqrt{D}}\right)\mathbf{V}_i, \tag{4}$$

where $\mathbf{Q}_j = \overline{\mathbf{F}}_j \mathbf{W}^q$, $\mathbf{K}_i = \overline{\mathbf{F}}_i \mathbf{W}^k$ and $\mathbf{V}_i = \overline{\mathbf{F}}_i \mathbf{W}^v$ are linear layers, $\sqrt{D}$ controls the effect of growing magnitude of dot-product with larger D [23]. Since $\overline{\mathbf{F}}_{j \to i}$ only learn the attentive relation from $\mathbf{F}_j$ to $\mathbf{F}_i$. We then learn the bi-direction attentive relation by $\overline{\mathbf{F}}_{i-j} = \text{Relation}(\overline{\mathbf{F}}_i, \overline{\mathbf{F}}_j) = \overline{\mathbf{F}}_{j \to i} + \overline{\mathbf{F}}_{i \to j}$, where $\overline{\mathbf{F}}_{i \to j} = \text{Attention}\,(\mathbf{Q}_i; \mathbf{K}_j; \mathbf{V}_j)$.

After that, we use the difference regression block to map $\overline{\mathbf{F}}_{i-j}$ to the difference scores $\hat{\Delta Y}$. Then, we minimize the error as follows:

$$\mathcal{L}_{diff} = (\hat{\Delta Y} - \Delta Y)^2, \tag{5}$$

where ΔY is the ground-truth of the difference scores between the pair videos, which can be computed by $|Y_i - Y_j|$. By optimizing $\mathcal{L}_{diff}$, the model would be able to distinguish differences between videos for precise SKA.

Finally, the overall loss function of our proposed method is as follows:

$$\mathcal{L} = \mathcal{L}_{reg} + \lambda_{diff} \mathcal{L}_{diff}, \tag{6}$$

where λ_{diff} is the hyper-parameter to control the weight between two loss functions (set to 1 empirically).

3 Experiments

Datasets. We collect the data from our collaborating hospitals. The data collection process follows the same protocol in a well-established study [2]. The whole procedure of the surgery is recorded by a surgeon's view camera. Each surgeon will submit videotapes when performing thoracoscopy-assisted MVR in the operating rooms. We have collected 50 high-resolution videos of thoracoscopy-assisted MVR from surgeons and patients, with a resolution of 1920×1080 and 25 frames per second. Each collected video lasts 30 min - 1 h. 50 videos are randomly divided into training and testing subsets containing 38, and 12 videos, respectively. To evaluate skill level, each video will be rated along various dimensions of technical skill on a scale of 1 to 9 (with higher scores indicating more advanced skill) by at least ten authoritative surgeons who are unaware of the identity of the operating surgeon. Furthermore, we also provide the annotations (including the category and corresponding start and end time) for three skill-related events, *i.e.*, haemorrhage, surgical thread twining and surgical suture repair times. The detailed annotation examples are illustrated in Fig. 1(a).

Implementation Details. Our model is implemented on an NVIDIA GeForce RTX 3090 GPU. We use a pre-trained inception-v3 [21] and MS-TCN [5] as the spatial and temporal feature extractors, respectively. For each video, we sample one frame per second. As the durations of different videos vary, we resample all videos to 1000 frames. We trained our model using an Adam optimizer with learning rates initialized at $1e-3$. The total number of epochs is 200, and the batch size is 4.

Table 1. Results on our MVR dataset. †: we implement existing action quality assessment methods on our dataset. ⋆: we run existing surgical skill assessment methods on our dataset.

Method	MAE	Corr
CoRe† [27]	2.89	0.40
TPT† [1]	2.68	0.42
C3D-LSTM⋆ [18]	3.01	0.20
C3D-SVR⋆ [18]	2.92	0.14
MTL-VF [25]	2.35	0.31
ViSA⋆ [14]	2.31	0.33
Ours	**1.83**	**0.54**

Table 2. Ablation study of the local-global difference module. "Base" refers to using the basic regression module containing only global information. "EAM" refers to the event-aware module capturing local information. "DRB" indicates the difference regression block extracting inter-video information.

Method	Base	EAM	DRB	MAE	Corr
Global	✓			2.82	0.25
Local		✓		2.49	0.33
Local-global	✓	✓		2.45	0.37
Global-difference	✓		✓	2.15	0.39
Local-difference		✓	✓	2.11	0.45
Full (ours)	✓	✓	✓	**1.83**	**0.54**

Evaluation Metrics. Following previous works [12,14], we measured the performance of our model using Spearman's Rank Correlation (Corr) and Mean Absolute Error (MAE). Lower values of Corr and MAE indicate better results.

3.1 Comparison with the State-of-the-Art Methods

We compare our method with existing state-of-the-art methods in action quality assessment (AQA) [1,27] and surgical skill assessment (SKA) [14,18]. Note that the spatial and temporal feature extractors for ViSA, CoRe and TPT as the same as our method. As shown in Table 1, our method achieved the best performance with an MAE score of 1.83 and a Corr score of 0.54. The comparison demonstrates that our method not only outperformed existing SKA methods but also outshined existing AQA methods by a clear margin in surgical skill assessment.

3.2 Ablation Study

Effectiveness of Proposed Modules. Table 2 shows the effectiveness of our proposed local-global difference module. We can see that using the local features from the event-aware module (EAM) can outperform the global ones, which indicates the importance of skill-related events. Furthermore, incorporating the difference regression module (DRB) can benefit both local and global features, *e.g.*, improving the MAE of Local from 2.49 to 2.11. Finally, the combination of all proposed modules can achieve the best performance, *i.e.*, the MAE of 1.83.

Analysis of Local-Global Fusion. We explore the effect of different types of local-global fusion in Table 3. "Concatenation" indicates concatenating the two features in the feature dimsension. "Multiplication" indicates the element-wise multiplication of the two features. The results show that the different fusion

Table 3. Analysis of local-global fusion.

	MAE	Corr
Concatenation	**1.83**	**0.54**
Multiplication	1.98	0.45
Attention	1.85	0.51

Table 4. Anlysis of the attention in the difference block. "Unidirectional" and "Bidirectional" indicate the unidirectional and bidirectional attentive relations (See Eq. 4).

Attention	MAE	Corr
Unidirectional ($\overline{\mathbf{F}}_{i\to j}$)	1.91	0.47
Bidirectional ($\overline{\mathbf{F}}_{i-j}$)	**1.83**	**0.54**

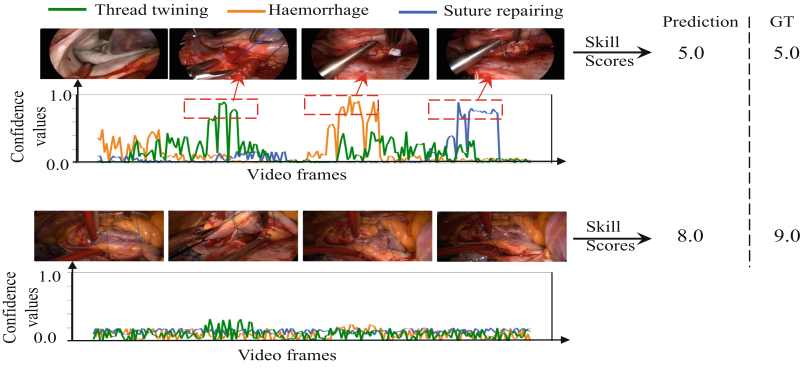


Fig. 3. Qualitative results of sampled pairwise videos. For each video, we visualize its confidence values along video frames for skill-related events. Specifically, the green, orange, and blue lines indicate the confidences scores of thread twining ($\mathbf{A}_0$), haemorrhage ($\mathbf{A}_1$) and suture repairing ($\mathbf{A}_2$) along video frames. It is worth noting that the occurrence of inappropriate surgical events such as haemorrhage and more thread twining times is highly correlated with the surgical skill level. Therefore, a lower confidence value indicates a lower probability of an event occurring, leading to a higher skill score.

methods can achieve comparable performance, indicating that different fusion methods can effectively aggregate local and global information. In this paper, we select concatenation as our default fusion method.

Effect of the Attention in the Difference Block. In Sect. 2.2, we implement the difference block by the attention, shown in Eq. 4. Here, we conduct the ablation study on the effect of different attentions in Table 4. The results indicate that using bidirectional attention, *i.e.*, $\overline{\mathbf{F}}_{i-j}$, can achieve better performance, compared with the uidirectional one.

Qualitative Results. Figure 3 shows the qualitative results of sampled videos to analyze the effectiveness of our method. The upper video presents a low surgical skill score, *i.e.*, 5.0, while the score for the lower video is higher, *i.e.*, 9.0. By comparing the two videos, the confidence lines generated by our model can find several factors that lower the skill score for the upper video, such as haemorrhage, multiple rewinds, and needle threading. Hence the upper video only obtains a skill score of 5.0, while the lower one achieves the better score, *i.e.*, 8.0.

4 Conclusion

This paper introduces a new surgical video dataset for evaluating thoracoscopy-assisted surgical skills. This dataset constitutes the first-ever collection of long surgical videos used for skill assessment from real operating rooms. To address the challenges posed by long-range videos and multiple complex surgical actions in videos, we propose a novel SEDSkill method that incorporates a local-global difference framework. In contrast to current methods that solely rely on intra-video information, our proposed framework leverages local and global difference learning to enhance the model's ability to use inter-video relations for accurate SKA in the MVR scenario.

Acknowledgement. This work was supported in part by a research grant from HKUST-BICI Exploratory Fund (HCIC-004) and in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: T45-401/22-N).

References

1. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part IV. LNCS, vol. 13664, pp. 422–438. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19772-7_25
2. Birkmeyer, J.D., et al.: Surgical skill and complication rates after bariatric surgery. N. Engl. J. Med. **369**(15), 1434–1442 (2013)
3. Brajcich, B.C., et al.: Association between surgical technical skill and long-term survival for colon cancer. JAMA Oncol. **7**(1), 127–129 (2021)
4. Carbello, B.: Mitral valve disease. Curr. Probl. Cardiol. **18**(7), 425–478 (1993)
5. Ding, X., Li, X.: Exploiting segment-level semantics for online phase recognition from surgical videos. arXiv preprint arXiv:2111.11044 (2021)
6. Ding, X., Wang, N., Gao, X., Li, J., Wang, X., Liu, T.: KFC: an efficient framework for semi-supervised temporal action localization. IEEE Trans. Image Process. **30**, 6869–6878 (2021)
7. Ding, X., et al.: Support-set based cross-supervision for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11573–11582 (2021)
8. Gao, Y., et al.: JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2cai, vol. 3 (2014)
9. Healey, M.A., Shackford, S.R., Osler, T.M., Rogers, F.B., Burns, E.: Complications in surgical patients. Arch. Surg. **137**(5), 611–618 (2002)
10. Jin, A., et al.: Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 691–699. IEEE (2018)

11. Kunisaki, C., et al.: Significance of thoracoscopy-assisted surgery with a minithoracotomy and hand-assisted laparoscopic surgery for esophageal cancer: the experience of a single surgeon. J. Gastrointest. Surg. **15**, 1939–1951 (2011)
12. Lavanchy, J., et al.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci. Rep. **11**(1), 5197 (2021)
13. Li, M., Zhang, H.B., Lei, Q., Fan, Z., Liu, J., Du, J.X.: Pairwise contrastive learning network for action quality assessment. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13664, pp. 457–473. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19772-7_27
14. Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 410–420. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_39
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
16. Liu, D., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: Shen, D., et al. (eds.) MICCAI 2019, Part V. LNCS, vol. 11768, pp. 476–484. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_53
17. Mason, J.D., Ansell, J., Warren, N., Torkington, J.: Is motion analysis a valid tool for assessing laparoscopic skill? Surg. Endosc. **27**, 1468–1477 (2013)
18. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28 (2017)
19. Reznick, R.K.: Teaching and testing technical skills. Am. J. Surg. **165**(3), 358–361 (1993)
20. Strasberg, S.M., Linehan, D.C., Hawkins, W.G.: The accordion severity grading system of surgical complications. Ann. Surg. **250**(2), 177–186 (2009)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
22. Uemura, M., et al.: Procedural surgical skill assessment in laparoscopic training environments. Int. J. Comput. Assist. Radiol. Surg. **11**, 543–552 (2016)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
24. Wagner, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the Heichole benchmark. arXiv preprint arXiv:2109.14956 (2021)
25. Wang, Tianyu, Wang, Yijie, Li, Mian: Towards accurate and interpretable surgical skill assessment: a video-based method incorporating recognized surgical gestures and skill levels. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part III. LNCS, vol. 12263, pp. 668–678. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_64
26. Wanzel, K.R., Ward, M., Reznick, R.K.: Teaching the surgical craft: from selection to certification. Curr. Probl. Surg. **39**(6), 583–659 (2002)
27. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7919–7928 (2021)

28. Zhang, C.L., Wu, J., Li, Y.: ActionFormer: localizing moments of actions with transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part IV. LNCS, vol. 13664, pp. 492–510. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19772-7_29
29. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12993–13000 (2020)