Spatial-Temporal Graph Transformer for Surgical Skill Assessment in Simulation Sessions

Kevin Feghoul, Deise Santana, Mehdi El Amrani, Mohamed Daoudi, Ali Amad

Spatial-Temporal Graph Transformer for Surgical Skill Assessment in Simulation Sessions

Kevin Feghoul[1,2], Deise Santana Maia[2], Mehdi El Amrani[3], Mohamed Daoudi[2,4], and Ali Amad[1]

[1] Univ. Lille , Inserm, CHU Lille, UMR-S1172 LilNCog, F-59000 Lille, France
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3] Department of Digestive Surgery and Transplantation, CHU Lille, PRESAGE, Univ. Lille, France
[4] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France

Abstract. Automatic surgical skill assessment has the capacity to bring a transformative shift in the assessment, development, and enhancement of surgical proficiency. It offers several advantages, including objectivity, precision, and real-time feedback. These benefits will greatly enhance the development of surgical skills for novice surgeons, enabling them to improve their abilities in a more effective and efficient manner. In this study, our primary objective was to explore the potential of hand skeleton dynamics as an effective means of evaluating surgical proficiency. Specifically, we aimed to discern between experienced surgeons and surgical residents by analyzing sequences of hand skeletons. To the best of our knowledge, this study represents a pioneering approach in using hand skeleton sequences for assessing surgical skills. To effectively capture the spatial-temporal correlations within sequences of hand skeletons for surgical skill assessment, we present STGFormer, a novel approach that combines the capabilities of Graph Convolutional Networks and Transformers. STGFormer is designed to learn advanced spatial-temporal representations and efficiently capture long-range dependencies. We evaluated our proposed approach on a dataset comprising experienced surgeons and surgical residents practicing surgical procedures in a simulated training environment. Our experimental results demonstrate that the proposed STGFormer outperforms all state-of-the-art models for the task of surgical skill assessment. More precisely, we achieve an accuracy of 83.29% and a weighted average F1-score of 81.41%. These results represent a significant improvement of 1.37% and 1.28% respectively when compared to the best state-of-the-art model.

Keywords: Graph Convolutional Networks · Transformer · Surgical Skill Assessment · Hand Skeleton · Simulation · Education

1 Introduction

Surgical skill assessment refers to the process of evaluating and measuring surgeon's technical proficiency and competence in executing surgical procedures. It

delivers targeted feedback that enables efficient skill development through the provision of guidance, ultimately resulting in better patient treatment. Traditionally, evaluation has been performed by senior surgeons using both global and task-specific checklists [5, 12]. However, classical surgical skill assessment checklists have several limitations, such as having a restricted scope, being prone to evaluator bias, lacking standardization, being a time-intensive and expensive process. Therefore, the development of automated tools to evaluate surgical skills is of significant interest. Collection and analysis of tool motion or video data can lead to an accurate assessment of the trainee's surgical proficiency. The proficiency can be quantified numerically through metrics such as the average OSATS score [12], or categorized into novice or expert levels, providing a clear and objective evaluation.

The conventional approach for automatically evaluating surgical proficiency relies on analyzing instrument motion, which can be obtained from various data sources such as video object tracking [14], video spatial-temporal features [24], and robotic kinematics [8, 20]. Other techniques focus solely on utilizing video data. For instance, Funke et al. [3] proposed to use a Temporal Segment Network [19] by fine-tuning a pre-trained 3D Convolutional Neural Network on a stack of video frames. In [10], the authors proposed a unified multi-path framework for automatic video-based surgical skill assessment, taking into account various aspects of surgical skills, such as surgical tool usage, intraoperative event patterns, and other skill proxies. To capture the relationships between these factors, a path dependency module has been specially designed.

In recent years, Graph Convolutional Networks (GCNs) have become the de facto choice for modeling relational data due to their ability to capture both the local and global structure of graphs. This has resulted in GCNs achieving state-of-the-art performance in various tasks related to spatial-temporal data [6,17,21]. Similarly, Transformers [18] have revolutionized the field of natural language processing and have become the go-to method for various natural language processing (NLP) tasks. In addition to language-related applications, the Transformer architecture has also been applied to tasks beyond NLP, such as skeleton-based action recognition, and has produced outstanding results, as demonstrated in studies such as [13, 16, 23].

In this study, we explored the potential of using hand skeleton sequences for surgical skill assessment. Our framework offers several advantages, including (1) being lighter and easier to train than models that process entire video sequences, and (2) providing an affordable alternative to expensive robotic surgical systems that can provide kinematics data, since the hand skeleton can be extracted from affordable mobile phones. Additionally, hand skeleton detection is performed in real-time, which ensures its practicality and suitability for use in real-world scenarios. As far as our knowledge extends, this is the first attempt to use hand skeleton sequences for evaluating surgical proficiency. Considering the graph structure of the hand skeleton and the dynamic spatial-temporal patterns in sequences of hand movement, we propose the STGFormer framework that combines the strengths of spectral GCNs for learning spatial-temporal represen-

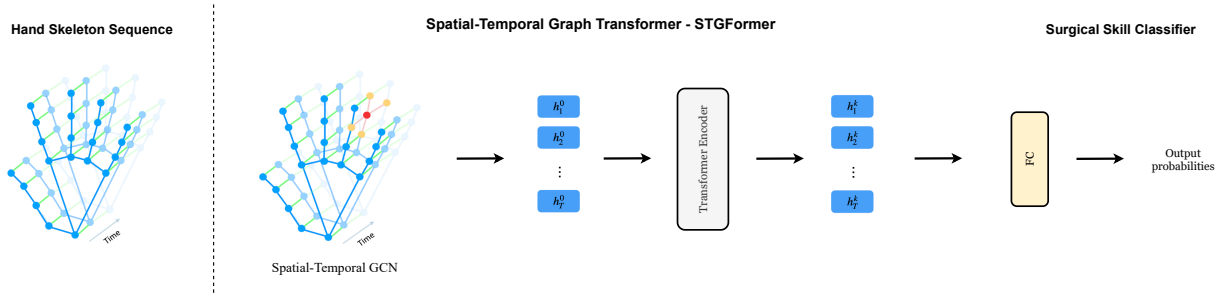


Fig. 1: Illustration of our STGFormer based surgical skill assessment framework, which is composed of two key components: Spatial-Temporal Graph Transformer and Surgical Skill Classifier.

tations and Transformers for capturing long-range dependencies. Our framework has shown to outperform all existing state-of-the-art spatial-temporal skeleton-based deep learning models for surgical skill evaluation.

The contributions of this work are twofold and can be summarized as follows: (1) we propose to use sequences of hand skeleton for the task of surgical skill assessment. This approach offers several advantages, such as being non-invasive, objective, and extensible to operating rooms. Moreover, hand skeletons can be extracted from inexpensive devices, such as a smartphone. By analyzing hand dynamics, practitioners can gain valuable insights into their performance, which can be used for improvement and ultimately lead to better patient treatment; (2) we developed a new spatial-temporal model that learns the dynamic spatial-temporal correlations of hand skeletons. It consists of a spectral GCN for spatial-temporal feature learning follows by a Transformer encoder for capturing global temporal dependencies. This combination of proven techniques leads to the best prediction performances compared to existing state-of-the-art models.

2 Proposed Approach

This section introduces our STGFormer framework, which is illustred in Figure 1. The framework consists of two essential components: (1) a spectral GCN responsible for learning spatial-temporal representation from hand skeleton sequences, and (2) a Transformer encoder designed to capture global temporal patterns.

2.1 Spectral Graph Convolutional Networks

In order to learn higher-level feature representations, we constructed a spatial-temporal graph and employed a spectral domain GCN.

Graph Construction In this study, we constructed an undirected spatial-temporal graph $\mathcal{G} = (V, E)$ to obtain high-level representations of a hand skeleton sequence consisting of N joints over T frames. The set of nodes in the graph is represented by V, while E denotes the set of edges. The construction process is outlined as follows:
Nodes: the nodes in the graph consist of all joints in the sequence, expressed as $V = \{v_{ti} \mid t = 1, .., T, i = 1, .., N\}$. Each node v_{ti} is initialized with its 3D coordinate information. In this study, N is equal to 21.
Edges: the set of edges E is defined as the union of intra-skeleton connections, E_{intra}, and inter-frame connections, E_{inter}, in the graph, defined as follows:

$$E_{intra} = \{v_{ti}v_{tj} \mid (i, j) \in H, t \in \{1, .., T\}\} \tag{1}$$

$$E_{inter} = \{v_{ti}v_{(t+1)i} \mid i \in \{1, .., N\}, t \in \{1, .., T-1\}\} \tag{2}$$

In Eq. 2., H represents the set of naturally connected hand joints.

Graph Learning We trained a spectral deep GCNs based on the previously constructed graph $\mathcal{G}$. We define the graph convolution operator as in [9]:

$$\widetilde{H}^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{3}$$

where $\widetilde{A} = A + I_n$ denotes the adjacency matrix of the undirected graph $\mathcal{G}$ with inserted self-connections, I_n represents the identity matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ is the diagonal degree matrix, $W^{(l)}$ is a learnable weight matrix, and $\sigma(.)$ an activation function. H^l represents the matrix of activations in the l^{th} layer; $H^0 = X$, where X is the matrix of input node feature.

2.2 Transformer Encoder

In order to capture complex temporal patterns in the hand skeleton sequences, we feed the final high-level representation, previously extracted from the GCN, to a Transformer encoder. To be more specific, we concatenate the newly obtained representations of every joint j_{ti} into a vector h_t^0 for each frame. This concatenation process is illustrated in Eq. 4. Next, we combine the initial representations h_t^0 from all frames into a vector h^0 as depicted in Eq. 5. This vector h^0 serves as the input for a Transformer encoder.

$$h_t^0 = [j_{t1}, j_{t2}, .., j_{tN}] \tag{4}$$

$$h^0 = [h_1^0, h_2^0, .., h_T^0] \tag{5}$$

The Transformer is an advanced neural network architecture that relies on the self-attention mechanism, enabling the model to effectively process input sequences and generate predictions. Unlike traditional recurrent neural networks,

which are limited by sequential processing, the Transformer can simultaneously attend to different parts of the input sequence, making it highly efficient at capturing long-term dependencies.

The self-attention mechanism in the Transformer calculates a weighted sum of the input sequence, with the weights being learned during the training process. This allows the model to assign importance to different positions in the sequence, focusing on the most relevant information for prediction. By considering the entire input sequence rather than just past representations, the Transformer can effectively capture contextual information and make accurate predictions.

A crucial component of the Transformer is the Multi-Head Attention (MHA) module. It enhances the model's ability to capture long-range dependencies and enables simultaneous attention across multiple representation subspaces at different positions. The MHA achieves this by utilizing Query-Key-Value (QKV) pairs. Each QKV triple is transformed into separate linear projections, and the scaled dot-product attention mechanism is applied. The scaled dot-product attention can be defined as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where $\frac{1}{\sqrt{d_k}}$ is used to counteract the vanishing gradient problem cause by the softmax function.

Each head of the MHA module is computed in parallel. The MHA module can be mathematically represented by the following equations:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{7}$$

$$\text{with } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

where $W_i^Q \in \mathbb{R}^{d_m \times d_k}, W_i^K \in \mathbb{R}^{d_m \times d_k}, W_i^V \in \mathbb{R}^{d_m \times d_v}$, $W_O \in \mathbb{R}^{hd_v \times d_m}$ represent the query, key, value, and output projection learnable weight matrices, respectively. h and d_m correspond to the number of heads and the output dimension of the encoder block. In this study, we choosed $d_k = d_v = d_m/h$.

2.3 Surgical Skill Classifier

After the forward pass through k-th Transformer encoder layer, the learned representation h^k, as shown in Eq. 9, is utilized as input to a fully connected neural network. This network is responsible for making predictions about whether the hand skeleton sequence is related to a senior surgeon or a surgical resident.

$$h^k = [h_1^k, h_2^k, .., h_T^k] \tag{9}$$

3 Experimental Results

This section presents the dataset collected for the surgical skill assessment task, as well as the results obtained using our proposed approach and several state-of-the-art deep learning-based models.

3.1 Dataset

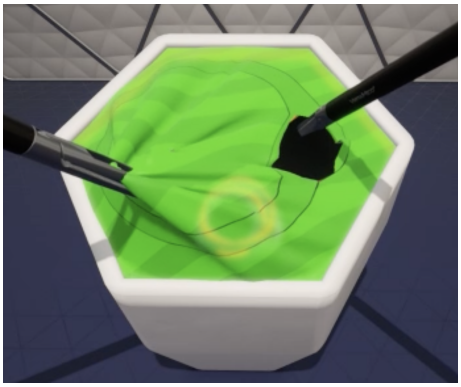
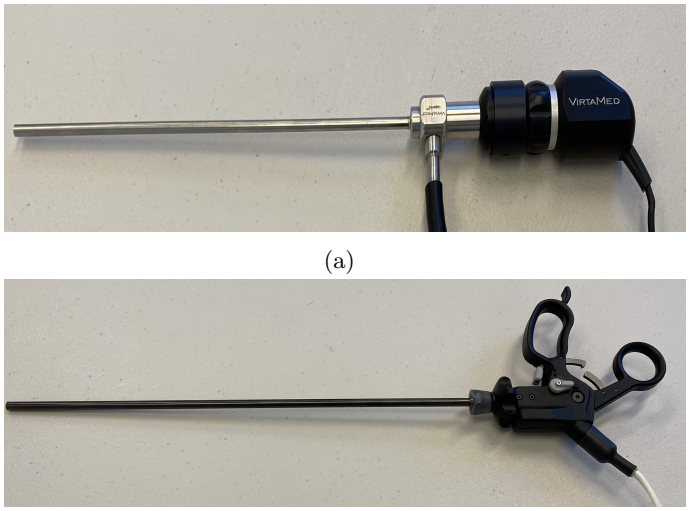


Fig. 2: Illustration of the circular cutting exercise using the VirtaMed simulator.

Data Collection We gathered data from a total of 16 participants, consisting of 4 experienced surgeons and 12 surgical residents. The participants executed a circular cutting exercise using the VirtaMed medical simulator, as depicted in Figure 2. The first step of the cutting exercise was to use a laparoscope, as illustrated in Figure 3a, to enter the virtual environment and position the view at the correct location. Following that, the participants utilized an atraumatic grasper tool (Figure 3b) to apply tension to the tissue and execute a precise cut along a circular incision between two lines using a pair of scissors (Figure 3b). For each participant, we recorded their hand movements while they performed the exercise using a smartphone equipped with 4K recording capability.

The circular cutting exercise was conducted in a simulated environment at the PRESAGE medical simulation center (Plateforme de Recherche et d'Enseignement par la Simulation pour l'Apprentissage des Attitudes et des GEstes), which is a department affiliated with the Faculty of Medicine at the University of Lille. This simulation center accurately replicates surgical training scenarios, making it an ideal setting for developing surgical skills. It is common for surgical novices to practice on medical simulator, where they perform tasks from curricula such as the Fundamentals of Laparoscopic Surgery (FLS) [15].



(a)

(b)

Fig. 3: (a) Laparoscope; (b) Atraumatic Grasper / Scissors.

FLS Program The circular cutting exercise is an important part of the training of residents and is included in the FLS program. Initiated in 2004, the FLS program was designed to deliver standardized training for laparoscopic procedures and encompasses theoretical knowledge as well as practical skills. Surgeons often need to complete the FLS program to obtain certification in laparoscopic surgery.

The circular cutting exercise is a crucial component of the FLS program, along with few other simulation exercises, and holds significant importance within the training curriculum. Despite seeming straightforward, it remains an important aspect of the training curriculum. This exercise helps residents develop precise control over laparoscopic instruments, particularly scissors, and improves their hand-eye coordination. It also enables them to understand the tactile feedback and resistance encountered when cutting tissue using these instruments, and enhances their depth perception skills by accurately assessing the distance and thickness of simulated tissue.

3.2 Data Preprocessing

We used the method from [22] to extract the hand skeleton of both hands from the recorded videos of each individual. The hand landmark model outputs a set of 21 3D coordinates for each frame, based on the hand intra connectivity structure, as illustrated in Figure 1. We opted to rely solely on right hand landmarks, as left hand detection was unreliable and played a minor role in this task. Indeed, the right hand was primarily responsible for the cutting, while the left hand primarily held the tissue with limited movement. Afterward, we normalized each

hand skeleton sequence by subtracting the coordinate of the first wrist joint (v_{00}) from each joint. Finally, we generated non-overlapping sliding windows of 20 seconds, which correspond approximately to 600 data points. As a result, we have a varying number of data sequences for each subject, which are directly dependent on the time taken to complete the exercise, with a duration of the recordings ranging from 1 minute and 33 seconds to 6 minutes and 17 seconds, with an average duration of 3 minutes and 6 seconds.

Table 1: Surgical skill assessment: comparison with state-of-the-art methods.

Method	Acc	F1-score
SoCJ [2]	80.39	77.55
TCN [1]	80.08	78.25
LSTM [7]	81.21	79.36
DeepGRU [11]	81.42	79.48
Transformer [18]	80.53	78.19
GCN [9]	81.92	80.13
ST-GCN [21]	79.14	79.54
ASTGCN [6]	79.30	79.49
STGFormer (ours)	**83.29**	**81.41**

3.3 Results

Evaluation framework In line with the JIGSAWS [4] dataset, which is widely used as a benchmark for evaluating surgical skill assessment, our study also takes into account the surgeon's experience as a valuable indicator of surgical proficiency. In our case, given the existence of two distinct groups of practitioners, namely senior surgeons and surgical residents, we formulate the surgical skill assessment as a binary classification task.

Our evaluation strategy involved utilizing a subject-independent 6-fold cross-validation to enhance the robustness of our evaluation. This approach was necessary because the data sequences of hand movements from the same subjects are likely to exhibit correlations. To ensure fairness in distributing the limited number of surgeons in our dataset across each fold, we generated all possible combinations of two surgeons, resulting in a total of six combinations. This ensured that each surgeon had an equal presence in both the training and test sets.

Additionally, we ensured that surgical residents were evenly distributed across the six folds to maintain homogeneity. To assess the performance of our

model, we employed accuracy as well as the weighted average F1-score. The inclusion of the F1-score allowed us to account for imbalanced class distributions within our dataset.

Surgical Skill Classification We compared our approach with eight state-of-the-art models that we re-implemented. Our approach was compared to classical deep learning-based methods such as TCN [1], LSTM [7], DeepGRU [11], and Transformer [18], trained directly on sequences of raw hand landmarks. In addition, we compared our approach with state-of-the-art spatial-temporal graph-based models, including GCN [9], ST-GCN [21], and ASTGCN [6]. The ST-GCN consists of multiple spatial-temporal convolutional blocks, each of which includes two temporal gated convolution layers and one spatial graph convolution layer in the center. The ASTGCN consists of multiple blocks, each composed of a spatial-temporal attention mechanism and a spatial-temporal convolution that utilizes graph convolutions to capture spatial patterns and standard convolutions to describe temporal features simultaneously. We also compared our framework with a model trained on handcrafted features, namely the SoCJ descriptor [2], which extracts a descriptor from the hand skeleton based on its geometric shape. These features are then input into a LSTM model.

In Table 1, we presented the results of our STGFormer model and above mentioned state-of-the-art baselines. Our STGFormer achieves the best performance in terms of both evaluation metrics, achieving 83.29%, and 81.41% in terms of accuracy, and F1-score respectively, as shown in Table 1, which represent an improvement of 1.37% and 1.28% when compared to the best state-of-the-art model.

The SoCJ approach, which involves extracting spatial descriptors, exhibits the lowest F1-score among the evaluated methods. In addition, even when compared to a LSTM model trained directly on raw data, the SoCJ descriptor proves to be inefficient, highlighting the limitations of the descriptor extraction process for our particular task.

As part of our ablation study, we observed that STGFormer outperformed both the GCN and Transformer models by a significant margin, achieving an accuracy and F1-score improvement of at least 1.37% and 1.28% respectively. This outcome clearly demonstrates the effectiveness of combining graph-based and transformer-based approaches in the context of learning surgical skill evaluation. These results highlight the importance of incorporating spatial and temporal information for accurate and robust assessment of surgical skills.

Therefore, based on these findings, we can draw several conclusions regarding the effectiveness of using temporal data either individually or in combination with spatial data. Firstly, the superiority of STGFormer over the GCN and Transformer models suggests that leveraging both spatial and temporal information provides a more comprehensive understanding of surgical skill performance. By capturing the interplay between spatial relationships and temporal dynamics, STGFormer can extract more informative features, leading to improved accuracy and F1-score.

Secondly, the performance gap between STGFormer and the other models implies that solely relying on either spatial or temporal data may not be sufficient for accurate surgical skill assessment. Spatial information alone might not capture the dynamic nature of the surgical procedure, while temporal information alone might lack the contextual understanding provided by spatial relationships. Therefore, combining both spatial and temporal data, as done in STGFormer, proves to be crucial for achieving superior performance for the particular task of surgical skill assessment.

4 Conclusion

This study demonstrates the feasibility of utilizing hand skeleton sequences for accurate surgical skill assessment. The successful development of automated surgical skill assessment holds significant importance in training aspiring surgeons and enhancing their proficiency in performing safe interventions. In order to achieve this goal, we proposed a novel approach called STGFormer, which effectively captures spatial-temporal correlations and long-range dependencies in the hand skeleton sequences of practitioners as they perform tasks within a simulated environment. Extensive experiments were conducted on a dataset comprising both senior surgeons and surgical residents, and our STGFormer framework achieved an accuracy of 83.29% and a weighted average F1-score of 81.41%. These results strongly support the efficiency of our approach to accurately distinguish between senior surgeons and surgical residents, highlighting its potential as a valuable tool for evaluating surgical skills.

In a future study, we plan to extend our research by investigating a multimodal approach that combines hand skeleton sequences with RGB data to improve the accuracy of surgical skill assessment. This integration aims to leverage the complementary information provided by both modalities, further enhancing the robustness and effectiveness of our assessment framework.

References

1. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
2. De Smedt, Q., Wannous, H., Vandeborre, J.P.: Skeleton-based dynamic hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–9 (2016)
3. Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3d convolutional neural networks. International journal of computer assisted radiology and surgery **14**, 1217–1225 (2019)
4. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., 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)

5. Goh, A.C., Goldfarb, D.W., Sander, J.C., Miles, B.J., Dunkin, B.J.: Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of urology **187**(1), 247–252 (2012)
6. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 922–929 (2019)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Evaluating surgical skills from kinematic data using convolutional neural networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 214–221. Springer (2018)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
10. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9522–9531 (2021)
11. Maghoumi, M., LaViola, J.J.: Deepgru: Deep gesture recognition utility. In: Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14. pp. 16–31. Springer (2019)
12. Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British journal of surgery **84**(2), 273–278 (1997)
13. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., Chiaberge, M.: Action transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recognition **124**, 108487 (2022)
14. Pérez-Escamirosa, F., Alarcón-Paredes, A., Alonso-Silverio, G.A., Oropesa, I., Camacho-Nieto, O., Lorias-Espinoza, D., Minor-Martínez, A.: Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. International Journal of Computer Assisted Radiology and Surgery **15**(1), 27–40 (2020)
15. Peters, J.H., Fried, G.M., Swanstrom, L.L., Soper, N.J., Sillin, L.F., Schirmer, B., Hoffman, K., Committee, S.F., et al.: Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery **135**(1), 21–27 (2004)
16. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III. pp. 694–701. Springer (2021)
17. Slama, R., Rabah, W., Wannous, H.: Str-gcn: Dual spatial graph convolutional network and transformer graph encoder for 3d hand gesture recognition. In: IEEE FG. pp. 1–6 (2023)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2740–2755 (2018)

20. Wang, Z., Majewicz Fey, A.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International journal of computer assisted radiology and surgery **13**, 1959–1970 (2018)
21. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)
22. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)
23. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3229–3237 (2021)
24. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. International journal of computer assisted radiology and surgery **13**, 443–455 (2018)