GES-Net: A Gesture, Error, and Smoothness Aware Network for Surgical Skill Assessment

Runlong He[1,2(✉)], Dimitrios Anastasiou[1,2], Jialang Xu[1,2], Mobarak I. Hoque[1,3], Matthew W. E. Boal[4], Freweini M. Tesfai[4], Abdolrahim Kadkhodamohammadi[6], Matthew J. Clarkson[1,2], Danail Stoyanov[1,5], Nader Francis[4], and Evangelos B. Mazomenos[1,2(✉)]

[1] UCL Hawkes Institute, University College London, UK
[2] Dept of Medical Physics & Biomedical Engineering, University College London, UK
[3] Division of Informatics, Imaging and Data Science, University of Manchester, UK
[4] The Griffin Institute, Northwick Park and St Mark's Hospital, UK
[5] Dept of Computer Science, University College London, UK
[6] Medtronic Digital Surgery, UK
runlong.he.23@ucl.ac.uk, e.mazomenos@ucl.ac.uk

Abstract. With robotic-assisted surgery (RAS) increasingly adopted into clinical practice, automated skill assessment becomes essential for its transformative potential in surgical analytics and education. Existing RAS skill assessment methods are limited to black-box models that purely predict skill scores from input videos. This design offers no assessment transparency and only a crude, uninterpretable numerical score as feedback to users. Moreover, the majority of relevant studies are developed and validated in controlled simulation settings, with unproven generalization to real-world clinical contexts. This work introduces GES-Net, a multi-task learning framework integrating surgical gesture recognition, error detection, and skill assessment. GES-Net comprises two core components: semantic alignment and multi-dimensional evaluation. For semantic alignment, the model fuses visual features with gesture semantics through gated cross-attention and scheduled masked guidance, while a multi-stage temporal reasoning module captures complex gesture transitions. For multi-dimensional evaluation, GES-Net employs soft error weighting to dynamically emphasize error segments and uses jitter features to quantify operational smoothness. This design enables both accurate RAS skill assessment and transparent diagnostic feedback. Validated on two publicly available datasets JIGSAWS and RARP-skill, our approach outperforms recent baselines with SCC improvements of 0.05 in LOSO, 0.07 in LOUO on the JIGSAWS suturing task, and 0.05 on RARP-skill. Our code and data are available at https://github.com/HRL-Mike/GES-Net.

Keywords: Robotic Surgical Skill Assessment · Multi-task Learning · Multimodal Models · Surgical Data Science.

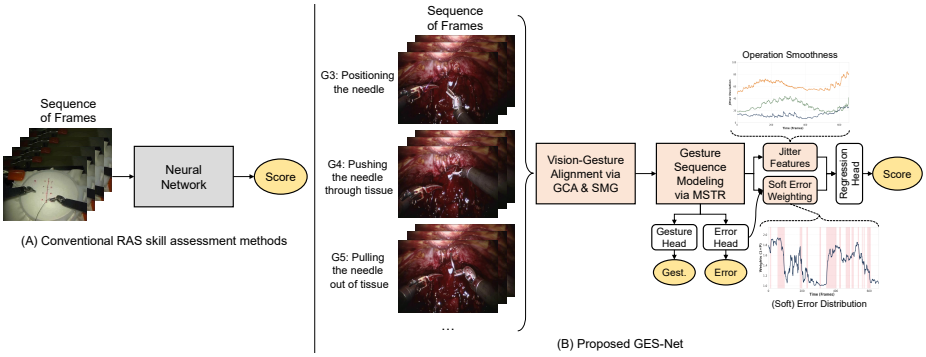

Fig. 1. Conventional RAS skill assessment methods use black-box models to predict scores and validate on simulation data. GES-Net establishes a multi-task learning framework that integrates gesture modeling and error detection to enhance skill assessment accuracy, while providing interpretable visualizations to support surgical training.

1 Introduction

Robot-assisted surgery (RAS) has transformed contemporary surgical practice through enhanced precision and safety, leading to improved patient outcomes [12]. Yet, the inherent complexity of RAS systems necessitates comprehensive training to ensure surgeons attain clinical proficiency. Traditionally, RAS evaluation involves experienced surgeons reviewing videos and completing structured scoring frameworks, such as the Global Evaluative Assessment of Robotic Skills (GEARS) [20]. This faces significant scalability challenges, due to substantial labor requirements for experts, while suffering from inter-rater inconsistencies and insufficient standardization [4]. As demand for RAS expertise and case volumes increase, automated, data-driven techniques for competency evaluation is a critical priority for advancing RAS training and maintaining optimal operational performance [23, 25].

Prior work in surgical skill assessment predominantly focuses on estimating an overall performance score from input videos, as shown in Fig 1. Since a single score provides limited supervision and interpretability [2], strategies such as contrastive learning [3, 28] and weakly-supervised learning [8, 17] have been explored to extract more discriminative representations under sparse label supervision. Other works decompose surgical videos into fine-grained actions through multi-modal approaches to assess surgical skills [10, 11, 26]. These approaches fail to explicitly model the relationship between errors and skill scores. Clinical evaluators directly consider operational errors as key references for scoring dimensions such as bimanual dexterity, tissue handling, and efficiency. This underscores an underexplored intrinsic relationship between errors and surgical skill.

Our work presents a novel approach to harness the underexplored relationship between surgical errors and skill scores. We propose GES-Net, an innovative multi-task learning framework that jointly models surgical gestures, errors, and

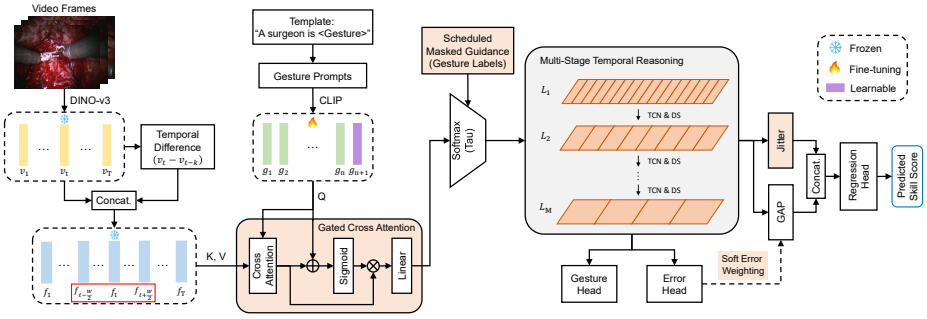

Fig. 2. GES-Net: The network forms of visual and textual encoders, gated cross attention block, scheduled masked guidance module, multi-stage temporal reasoning module, and multi-task heads.

motion smoothness for automated RAS skill assessment. Experimental validation on two public benchmarks demonstrates the effectiveness of GES-Net: compared to existing baseline methods, it achieves Spearman's rank correlation coefficient (SCC) improvements of 0.07 on the LOUO scheme for JIGSAWS suturing and 0.05 on the RARP-skill dataset. We illustrate that GES-Net effectively learns semantic features of different gestures and the temporal smoothness of operator performance, thereby enhancing the interpretability and reliability of robotic surgical skill assessment. Our key contributions include:

- We propose GES-Net, a multi-task learning framework that jointly models surgical gesture recognition, error detection, and operational smoothness to effectively assess RAS skill with limited score annotations.
- Specifically, we introduce scheduled masked guidance to align visual features with gesture semantics, soft error weighting to adaptively focus on error segments, and jitter features to quantify operational smoothness.
- Through extensive experiments on both simulation and clinical settings, we demonstrate the effectiveness of GES-Net for accurate and clinically-interpretable surgical skill assessment.

2 Method

We frame surgical skill assessment as a regression task and normalize the labels to $[0, 1]$ for stable training. Fig. 2 shows the architecture of GES-Net and the description of each module follows.

Visual and Textual Encoders: For each frame t, we extract visual features $v_t \in \mathbb{R}^{768}$ using a pre-trained DINO-v3 backbone. To capture motion dynamics, we compute temporal difference features as $\Delta v_t = v_t - v_{t-k}$ and form the final representation by concatenation: $f_t = [v_t; \Delta v_t] \in \mathbb{R}^{1536}$. A sliding window extracts local temporal context for subsequent cross-modal fusion. For semantic encoding, we use CLIP to generate embeddings for n surgical gestures. To

adapt these generic embeddings to the surgical domain, we introduce learnable offsets $\delta \in \mathbb{R}^{n\times 512}$ and add a global gesture token $g_{n+1} \in \mathbb{R}^{512}$ for holistic video representation. This yields $n+1$ semantic observers for guiding cross-attention.

Multimodal Fusion with Gated Cross Attention: Inspired by [16], we introduce a Gated Cross-Attention (GCA) module to transform dense visual features into sparse, gesture-specific semantic representations. Given textual query Q and visual key-value pairs $\{K, V\}$, we first compute the cross-attention context [24]: $C = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V$. To adaptively control information flow, a learnable gate $g \in [0, 1]$ is generated as: $g = \sigma(W_q Q + W_v C)$. The gated output is then computed as $out = \text{Linear}(g \odot C)$, where $\odot$ denotes element-wise multiplication. This mechanism enables each semantic observer to selectively attend to visual regions relevant to its corresponding surgical gesture. The GCA module employs two cascaded layers for progressive refinement. The first layer uses gesture prompt features as query Q to interact with the local visual window $\mathcal{W}_t$, while the second layer refines the output context using the same query, further enhancing semantic discriminability [21].

Scheduled Masked Guidance: To align semantic observers with surgical gestures while reducing label dependency, we incorporate a Scheduled Masked Guidance (SMG) module that encourages autonomous reasoning from visual cues. At each time step t, we compute the importance weights $W_t \in \mathbb{R}^{n+1}$ of the observer features $\mathcal{S}_t$ via temperature-scaled softmax: $W_t = \text{Softmax}(\text{Linear}(\mathcal{S}_t)/\tau)$, where τ controls the sharpness of the distribution. During training, labeled frames ($y_t \geq 0$) receive one-hot guidance to activate the corresponding gesture and global observer channels. To reduce label dependency, we randomly mask gesture labels in the middle 80% of surgical frames, where gesture transitions frequently occur. The masking rate follows a linear decay schedule: $M_r = \max(start\_pct - epoch \times decay, 0)$ with $start\_pct = 35\%$ and $decay = 0.5\%$ per epoch. Masked labels force the model to rely on W_t for inference, thereby encouraging the GCA module to extract discriminative features from visual signals. The SMG module is supervised by an alignment loss $\mathcal{L}_{align}$, computed as cross-entropy between predicted logits $\hat{y}_t$ and labels y_t over non-masked frames $\mathcal{T} = \{t : y_t \neq -100\}$: $\mathcal{L}_{align} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \ell_{ce}(\hat{y}_t, y_t)$. This maintains semantic alignment while progressively learning gesture inference from visual context.

Multi-Stage Temporal Reasoning: The refined semantic stream is processed by a Multi-Stage Temporal Reasoning (MSTR) module to capture long-range surgical dependencies. Unlike traditional methods that model temporal patterns on raw visual features, our model operates on a gesture semantic stream, where each frame t is represented by $n+1$ observer features weighted by W_t. The architecture consists of $M = 4$ sequential stages, each with 10 layers of 1D dilated residual convolutions. By exponentially increasing the dilation factor $d = 2^l$ at layer l, the MSTR module expands its receptive field to cover thousands of frames, enabling the model to reason across entire surgical procedures and recognize logical gesture transitions. Each stage produces auxiliary predictions for gestures and errors, providing multi-task supervision that stabilizes hidden representations for final skill regression.

Auxiliary Heads and Soft Error Weighting: To ground temporal representations in surgical domain knowledge, the MSTR module is equipped with two auxiliary prediction heads: a gesture recognition head and an error detection head. Both heads consist of a 1×1 convolution and Softmax activation, supervised by frame-wise cross-entropy losses to ensure the model captures surgical workflow (what is being done) and execution quality (how well it is done). For each stage $s \in \{1, \ldots, M\}$, the auxiliary loss averaged over T frames is: $\mathcal{L}_{aux}^{(s)} = \frac{1}{T} \sum_{t=1}^{T} \left(\ell_{ce}(\hat{y}_{g,t}^{(s)}, y_{g,t}) + \ell_{ce}(\hat{y}_{e,t}^{(s)}, y_{e,t}) \right)$ where ℓ_{ce} denotes the cross-entropy loss, $\hat{y}_g$ and $\hat{y}_e$ are predicted probabilities for gestures and errors, maintaining semantic signals throughout temporal reasoning. Since annotators weigh error-containing frames more heavily when scoring surgical skill, we propose a Soft Error Weighting (SEW) mechanism that replaces global average pooling. Using error probability $\hat{p}_{e,t}$ and hidden features out_t from MSTR's final stage, we compute the global representation as: $v_{global} = \frac{\sum_{t=1}^{T}(1+\hat{p}_{e,t}) \cdot out_t}{\sum_{t=1}^{T}(1+\hat{p}_{e,t})}$ This up-weights error-prone frames during aggregation, allowing the model to focus on informative evidence of surgical skill.

Smoothness Modeling and Regression Head: Motion smoothness of maneuvers represents another critical dimension of surgical proficiency. While the SEW mechanism provides a content-aware weighted summary emphasizing error-prone segments, the resulting v_{global} inherently smooths the temporal axis. To explicitly evaluate operation smoothness, we introduce a jitter feature v_{jitter} to capture transient instability of surgical gestures, calculated as the root-mean-square (RMS) deviation of frame-wise representations out_t from the global semantic mean: $v_{jitter} = \mathrm{RMS}_t\{\|out_t - v_{global}\|\}$, thus v_{jitter} quantifies stochastic fluctuations in the feature space. High jitter values often correlate with surgical hesitations, tremors, or redundant movements, subtle cues of lower skill that may not trigger explicit error labels but significantly impact procedural smoothness.

For the final assessment, we concatenate v_{global} and v_{jitter}, feeding them into a regression head consisting of an MLP with layer normalization, ReLU activation, and sigmoid output: $\hat{S} = \mathrm{Sigmoid}(\mathrm{MLP}([v_{global}; v_{jitter}]))$. The training objective is: $\mathcal{L}_{total} = \mathcal{L}_{skill} + \alpha \cdot (\mathcal{L}_{mstr} + \mathcal{L}_{align})$ where $\mathcal{L}_{skill} = \mathrm{MSE}(\hat{S}, S) + \mathrm{MAE}(\hat{S}, S)$ is the primary regression loss, and the multi-stage temporal reasoning loss is: $\mathcal{L}_{mstr} = \frac{1}{M} \sum_{s=1}^{M} \left(\mathcal{L}_{aux}^{(s)} + \mathcal{L}_{sm}^{(s)} \right)$ where $\mathcal{L}_{sm}^{(s)}$ is the temporal smoothing loss [21]. We set $\alpha = 0.1$ to prioritize skill prediction.

3 Experimental Setup

Datasets and Evaluation Metrics: We evaluate GES-Net on two publicly available datasets: JIGSAWS [1] and RARP-skill [7]. JIGSAWS comprises three surgical tasks: knot tying, needle passing (NP), and suturing (SU). We focus on NP and SU, as they exclusively contain binary error annotations [9]. These tasks include 28 and 39 videos respectively, captured at 640×480 resolution with 30 fps, annotated with 15 gestures and binary error labels. For JIGSAWS, we

use the standard Leave-One-Supertrial-Out (LOSO) and Leave-One-User-Out (LOUO) protocols. The RARP-skill dataset contains 33 videos from a suturing task in robot-assisted radical prostatectomy procedures, captured at 1920×1080 resolution with 60 fps. Each video is annotated with 8 gestures [15] and binary error labels [27]. We perform 4-fold cross-validation on RARP-skill. Performance is evaluated using SCC and mean absolute error (MAE) for skill assessment, and accuracy for auxiliary tasks, with mean and standard deviation reported across all folds.

Implementation: We employ DINO-v3 [22] with ViT-B/16, pre-trained on LVD-142M [13]. Videos are sampled at 5 Hz and resized to 224×224. The CLS token embedding from the final layer serves as the 768-dimensional visual representation per frame. We utilize pre-trained CLIP [18] (ViT-B/32) to obtain semantic representations of surgical gestures. Each gesture description is embedded via the prompt template "A surgeon is `<gesture>`." (e.g., "A surgeon is tying a knot" for G6), producing 512-dimensional text embeddings. We optimize the model using Adam optimizer with learning rates of 6e-4 (JIGSAWS) and 1e-3 (RARP-skill). A cosine annealing schedule decays the learning rate to 5e-6. Training uses batch size 8 for 200 epochs per fold. The model is implemented in PyTorch and trained on a RTX 6000 GPU.

4 Results and Discussions

Table 1. Performance comparison on JIGSAWS dataset.

Method	Needle Passing (NP)				Suturing (SU)			
	LOSO		LOUO		LOSO		LOUO	
	SCC↑	MAE↓	SCC↑	MAE↓	SCC↑	MAE↓	SCC↑	MAE↓
C3D-LSTM [14]	0.84	—	0.78	—	0.69	—	0.59	—
C3D-MTL-VF [26]	0.75	—	<u>0.86</u>	—	0.77	—	0.69	—
MultiPath-VTPE [11]	—	—	0.65	—	—	—	0.45	—
ViSA [10]	**0.93**	**1.66**	**0.90**	<u>2.16</u>	0.84	<u>2.58</u>	<u>0.72</u>	2.82
Contra-Sformer [3]	0.71	3.15	0.71	3.17	<u>0.86</u>	2.74	0.65	<u>2.58</u>
ReCAP [17]	0.85	—	—	—	0.83	—	—	—
GES-Net (Ours)	<u>0.91</u>	<u>1.92</u>	**0.90**	**2.10**	**0.91**	**2.21**	**0.79**	**2.25**
SD	±0.02	±0.96	±0.22	±1.40	±0.10	±0.68	±0.23	±1.01

Baseline Comparison: We compare GES-Net with existing surgical skill assessment baselines. Since the RARP-skill dataset lacks kinematic data, we reimplement several vision-based methods using their official code to ensure fair comparison. Table 1 shows that GES-Net outperforms previous SOTA methods on JIGSAWS. In the SU task, our method improves SCC by 0.05 and 0.07 over ViSA [10] on LOSO and LOUO, respectively. In the NP task, GES-Net surpasses ReCAP [17] by 0.06 in SCC on LOSO and C3D-MTL-VF [26] by 0.04 in SCC on LOUO, demonstrating strong generalization. GES-Net achieves 2.12 (±0.15)

MAE across the four cross-validation schemes, significantly lower than Contra-Sformer [3] (2.91±0.30) and ViSA (2.31±0.51), indicating superior accuracy and stability. For auxiliary tasks under LOUO settings, GES-Net achieves error detection accuracy of 0.653 (±0.036) on NP and 0.715 (±0.079) on SU, while attaining gesture recognition accuracy of 0.730 (±0.126) on NP and 0.750 (±0.094) on SU, comparable to recent baselines [19,21]. GES-Net performs comparably to ViSA on NP while outperforming it on SU. This reflects task characteristics: NP exhibits relatively uniform motion patterns across skill levels, whereas SU shows greater variability between novices and experts. The higher temporal variance in SU (std: 43.63s vs. 24.42s for NP) further supports this observation.

Table 2. Performance comparison on RARP-skill dataset.

Method	4-fold	
	SCC↑	MAE↓
CoRe+GART [28]	0.57	2.02
ViSA [10]	0.60	2.10
Contra-Sformer [3]	0.61	<u>1.63</u>
CoRe+PECoP [5]	<u>0.71</u>	2.28
GES-Net (Ours)	**0.76**	**1.38**
SD	±0.15	±0.11

As shown in Table 2, GES-Net achieves SOTA performance on the RARP-skill clinical suturing dataset, outperforming CoRe+PECoP by 0.05 in SCC and Contra-Sformer by 0.25 in MAE. Additionally, GES-Net achieves an error detection accuracy of 0.691 ± 0.049 and a gesture recognition accuracy of 0.682 ± 0.042. These results demonstrate GES-Net's robustness and strong generalization from simulation to real clinical data.

Table 3. Ablation study on JIGSAWS and RARP-skill datasets.

Model	JIGSAWS (SU)		RARP-skill	
	LOSO		4-fold	
	SCC↑	MAE↓	SCC↑	MAE↓
GES-Net (ResNet)	<u>0.89</u> (±0.11)	<u>3.01</u> (±0.60)	0.69 (±0.24)	1.86 (±0.22)
GES-Net (w/o GCA)	0.85 (±0.16)	3.49 (±0.79)	0.71 (±0.15)	1.67 (±0.27)
GES-Net (w/o SMG)	0.82 (±0.11)	3.81 (±0.84)	0.64 (±0.22)	1.94 (±0.51)
GES-Net (w/o Jitter)	0.84 (±0.13)	3.18 (±0.98)	0.64 (±0.30)	1.78 (±0.28)
GES-Net (w/o SEW)	0.87 (±0.17)	3.08 (±0.80)	<u>0.73</u> (±0.17)	<u>1.63</u> (±0.20)
GES-Net	**0.91** (±0.10)	**2.21** (±0.68)	**0.76** (±0.15)	**1.38** (±0.11)

Ablation Study: We perform an ablation study examining different visual encoders and architectural components. We replace DINO-v3 features with ResNet50 embeddings and systematically remove the GCA, SMG, Jitter, and SEW mod-

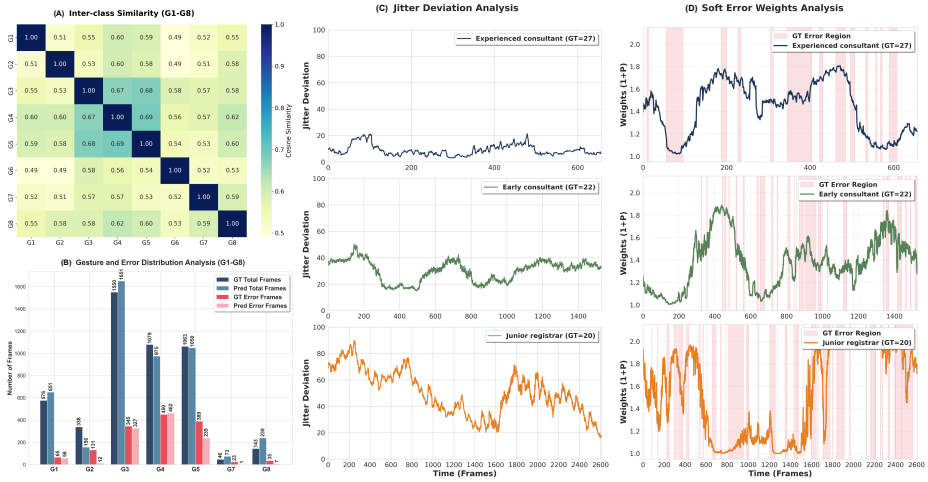

Fig. 3. Qualitative analysis on the RARP-skill dataset.

ules. Table 3 shows that DINO-v3 provides substantially larger SCC improvements on RARP-skill (0.07) than on JIGSAWS (0.02), as its register tokens [6] effectively filter irrelevant background noise in surgical scenes, such as vignetting at endoscope peripheries and noise from non-operative regions, enabling better focus on instrument-tissue interactions. The SMG module and Jitter features are the most influential components, with the former improving SCC by 0.09 on JIGSAWS suturing and the latter improving SCC by 0.12 on RARP-skill, indicating that scheduled masking helps establish robust gesture-visual associations while motion fluidity provides critical skill assessment cues.

Qualitative Analysis: We visualize gesture feature similarity, gesture-error distributions, temporal jitter, and SEW on the RARP-skill dataset. Fig. 3(A) shows high inter-class similarity among core suturing gestures G3 (positioning the needle tip), G4 (pushing the needle through tissue), and G5 (pulling the needle out of tissue), confirming that GES-Net effectively captures gesture semantics. Fig. 3(B) presents ground-truth and predicted gesture-error distributions, showing good performance on G1 and G3-G5 but struggles with terminal gestures G7 (cutting the suture) and G8 (dropping the needle). Fig. 3(C) shows that experienced consultants exhibit lower temporal jitter and minimal fluctuations compared to early consultants and junior registrars, reflecting smoother operations. Fig. 3(D) presents SEW analysis, where experienced consultants produce localized error clusters while junior registrars generate fragmented errors across multiple stages, and the model displays conservativeness toward certain ground-truth error annotations. Beyond predicting skill scores, GES-Net provides interpretable feedback through gesture-error analysis, smoothness metrics, and soft error distribution for comprehensive surgical skill assessment.

5 Conclusion

In this study, we introduced GES-Net, a multi-task learning framework designed for surgical skill assessment, integrating gesture recognition and error detection. GES-Net demonstrates superior performance on both JIGSAWS and RARP-skill datasets, surpassing existing baselines. This indicates robust generalization from simulation to clinical settings, underscoring its practical value for objective surgical skill evaluation in clinical practice. Ablation studies validate its design, highlighting improvements by masked guidance, soft error weighting and jitter features. Beyond accurate predictions, GES-Net provides interpretable visualizations, enhancing transparency in surgical skill assessment. Future work will extend binary error detection to multi-class, increasing interpretability of assessment outputs, and integrate kinematic features in smoothness modeling.

Acknowledgments. This study was funded in whole, or in part, by Medtronic Digital Surgery, the EPSRC [EP/Z534754/1, EP/W00805X/1, EP/T517793/1, EP/S021930/1, UKRI145], the DSIT and the RAEng under the Chair in Emerging Technologies.

Disclosure of Interests. DS reports employment with Medtronic Ltd; board membership, consulting or advisory, and equity or stocks with Odin Medical Ltd and Panda Surgical Ltd; and board membership and consulting or advisory with EnAcuity Ltd. The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., et al.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans. Biomed. Eng. **64**(9), 2025–2041 (2017)
2. Anastasiou, D., Caramalau, R., Sirajudeen, N., Boal, M., Edwards, P., Collins, J., et al.: Exploring pre-training across domains for few-shot surgical skill assessment. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 212–222 (2025)
3. Anastasiou, D., Jin, Y., Stoyanov, D., Mazomenos, E.: Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery. IEEE Robot. Autom. Lett. **8**(3), 1755–1762 (2023)
4. Boal, M.W., Anastasiou, D., Tesfai, F., Ghamrawi, W., Mazomenos, E., Curtis, N., et al.: Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review. Br. J. Surg. **111**(1), znad331 (2024)
5. Dadashzadeh, A., Duan, S., Whone, A., Mirmehdi, M.: Pecop: Parameter efficient continual pretraining for action quality assessment. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. pp. 42–52 (2024)
6. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International Conference on Learning Representations. vol. 2024, pp. 2632–2652 (2024)

7. He, R., Tesfai, F.M., Boal, M.W., Sirajudeen, N., Anastasiou, D., Xu, J., et al.: Surgfusion-net: Diversified adaptive multimodal fusion network for surgical skill assessment. arXiv preprint arXiv:2603.00108 (2026)
8. Hu, Z.J., Ranne, A., Abdelaal, A.E., Bhattacharyya, K., Burdet, E., Okamura, A.M., et al.: Model selection and real-time skill assessment for suturing in robotic surgery. arXiv preprint arXiv:2601.12012 (2026)
9. Hutchinson, K., Li, Z., Cantrell, L.A., Schenkman, N.S., Alemzadeh, H.: Analysis of executional and procedural errors in dry-lab robotic surgery experiments. Int. J. Med. Robot. Comput. Assist. Surg. **18**(3), e2375 (2022)
10. Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 410–420 (2022)
11. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., et al.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9522–9531 (2021)
12. Moglia, A., Georgiou, K., Georgiou, E., Satava, R.M., Cuschieri, A.: A systematic review on artificial intelligence in robot-assisted surgery. Int. J. Surg. **95**, 106151 (2021)
13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., et al.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
14. 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)
15. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
16. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., et al.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. Advances in Neural Information Processing Systems **38**, 100092–100118 (2026)
17. Quarez, J., Modat, M., Ourselin, S., Shapey, J., Granados, A.: Recap: Recursive cross attention network for pseudo-label generation in robotic surgical skill assessment. In: MICCAI Workshop on Agentic AI for Medicine. pp. 156–166 (2025)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
19. Rao, M., Qin, Y., Kolouri, S., Wu, J.Y., Moyer, D.: Zero-shot prompt-based video encoder for surgical gesture recognition. Int. J. Comput. Assist. Radiol. Surg. **20**(2), 311–321 (2025)
20. Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond, V., Rojas, A., et al.: Robotic surgery training: construct validity of global evaluative assessment of robotic skills (gears). J. Robot. Surg. **10**(3), 227–231 (2016)
21. Shao, Z., Xu, J., Stoyanov, D., Mazomenos, E.B., Jin, Y.: Think step by step: Chain-of-gesture prompting for error detection in robotic surgical videos. IEEE Robot. Autom. Lett. **9**(12), 11513–11520 (2024)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
23. Sirajudeen, N., Boal, M., Anastasiou, D., Xu, J., Stoyanov, D., Kelly, J., et al.: Deep learning prediction of error and skill in robotic prostatectomy suturing. Surg. Endosc. **38**(12), 7663–7671 (2024)

24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
25. Walshaw, J., Fadel, M.G., Boal, M., Yiasemidou, M., Elhadi, M., Pecchini, F., et al.: Essential components and validation of multi-specialty robotic surgical training curricula: a systematic review. Int. J. Surg. **111**(4), 2791–2809 (2025)
26. Wang, T., Wang, Y., Li, M.: Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 668–678 (2020)
27. Xu, J., Sirajudeen, N., Boal, M., Francis, N., Stoyanov, D., Mazomenos, E.B.: Sedmamba: Enhancing selective state space modelling with bottleneck mechanism and fine-to-coarse temporal fusion for efficient error detection in robot-assisted surgery. IEEE Robot. Autom. Lett. **10**(1), 232–239 (2024)
28. 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)