Journal of Robotic Surgery
https://doi.org/10.1007/s11701-021-01316-2

ORIGINAL ARTICLE

# Evaluating robotic-assisted surgery training videos with multi-task convolutional neural networks

**Yihao Wang[1] · Jessica Dai[2] · Tara N. Morgan[2] · Mohamed Elsaied[1] · Alaina Garbens[2] · Xingming Qu[1] · Ryan Steinberg[2] · Jeffrey Gahan[2] · Eric C. Larson[1]**

Received: 2 July 2021 / Accepted: 3 October 2021


## Abstract

We seek to understand if an automated algorithm can replace human scoring of surgical trainees performing the urethrovesical anastomosis in radical prostatectomy with synthetic tissue. Specifically, we investigate neural networks for predicting the surgical proficiency score (GEARS score) from video clips. We evaluate videos of surgeons performing the urethral anastomosis using synthetic tissue. The algorithm tracks surgical instrument locations from video, saving the positions of key points on the instruments over time. These positional features are used to train a multi-task convolutional network to infer each sub-category of the GEARS score to determine the proficiency level of trainees. Experimental results demonstrate that the proposed method achieves good performance with scores matching manual inspection in 86.1% of all GEARS sub-categories. Furthermore, the model can detect the difference between proficiency (novice to expert) in 83.3% of videos. Evaluation of GEARS sub-categories with artificial neural networks is possible for novice and intermediate surgeons, but additional research is needed to understand if expert surgeons can be evaluated with a similar automated system.

**Keywords** Surgical training · Robotic-assisted surgery · Deep learning · Skill evaluation · Keypoint detection

## Introduction

Robotic-assisted surgery is a type of minimally invasive surgery that allows for increased precision, flexibility, and control, resulting in quicker patient convalescence [1]. While robotic surgery has achieved great success and widespread adoption in recent years, this success remains inseparable from the skill of the surgeon operating the robot. Therefore, efficient and reliable training of residents remains of paramount importance. The current model for training residents (surgical trainees) is to always have an experienced surgeon overseeing the resident throughout their training period, which is both labor intensive and costly, and therefore difficult to scale to meet current needs [2, 3]. To help mitigate this, Johnson et al. [4] designed a training scenarios for robotic prostatectomy procedures with synthetic tissue. From video of trainees practicing on this synthetic tissue, a validated assessment of robotic skill was assigned to each video from four experienced surgeons. For scoring, they employed the "Global Evaluative Assessment of Robotic Skills" (GEARS) score which contains various elements for an evaluator to consider including: Depth Perception (DP), Bi-manual Dexterity (BD), Efficiency (E), Force Sensitivity (FS), Autonomy (A), Robotic Control (RC). In their work, they establish that ratings of expert, intermediate, and novice (based on surgeons' real-world experiences) did translate to significant differences in GEARS scores from the synthetic tissue model. However, the downside of this learning procedure is that is still requires experienced surgeon's to review trainee videos—ideally with multiple reviewers to increase the accuracy and reproducibility of the ratings. In this work, we seek to automate the process of scoring trainees on the synthetic tissue model. Using computer vision, we develop an automate the surgical evaluation process that closely correlates with expert review, thus potentially expediting the learning process and reducing the need for oversight from an experienced surgeon.

✉ Eric C. Larson
eclarson@smu.edu

Yihao Wang
yihaow@smu.edu

[1] Department of Computer Science, Southern Methodist University, Dallas, USA

[2] Department of Urology, University of Texas Southwestern Medical Center, Dallas, USA

Published online: 28 October 2021

We leverage the dataset of Johnson et al. [4] at the University of Texas, Southwestern Medical Center (with permission) to inform the design and evaluate our model. Therefore, in this work, our goals were set as follows:

1. Design and evaluate algorithms for tracking the location of various surgical instruments, from standard video, in real time.
2. Design and evaluate novel deep learning architectures capable of predicting surgical proficiency using validated scoring metrics from training videos and investigate the time segments from the video that influence the model prediction.

An overview of our processing pipeline is shown in Fig. 1. First, each video is divided into numerous video clips. For each video clip, we perform object detection for tracking specific parts of surgical instruments, saving the position of each object over time. From these time series positions, a sequence scoring model is designed to predict the GEARS score (a regression). We conduct extensive experimental validation using leave-one-subject-out cross-validation. We evaluate on the total GEARS score (a possible score of 30) and on each sub-category (or sub-domain) of GEARS (a possible score of 5 for each category). Of the 18 videos, 15 were classified into the correct category (83.3%) using the conventional GEARS rating system (expert >25, intermediate 20–25, novice <20). The model accurately predicted novice surgeons in 11/11 cases (100%) and intermediate surgeons correctly in 4/5 cases (80%). The model failed to correctly identify an expert surgeon in any case (0/2). For GEARS sub-categories, 108 GEARS predictions were made by the reviewers (18 video segments × 6 GEARS domains). Compared to human scoring, the model accurately predicted the individual sub-category score within 1 point in 86.1% of predictions and within 2 points in 100% of predictions.

## Related work

Because of the varied works that our methods build on, we divide our related work discussion into several categories comprising object detection, sequence modeling, multi-task learning, and robotic surgery assessment.

*Object detection and location tracking* Object detection is a challenging computer vision task that aims to identify predefined objects in a image or a video [5]. It has been a problem historically investigated for more than 40 years. However, due to the emergence of convolution neural networks (CNNs) [6], the vision community has achieved state-of-the-art results in many recognition tasks. We use a key point detector based on the U-Net architecture, which systematically encodes images and decodes into a given label space [7]. For example, Hasan et al. used the U-Net architecture to segment surgical instruments in a video frame [8]. Differently than our approach, they do not segment key points of the instrument; rather, they try to find a mask around each instrument. Perhaps most similar to our approach is that of Islam et al. [9] and Shvets et al. [10], that used a masking U-Net for identifying individual portions of robotic instruments [9, 10]. While these works did not use key points,

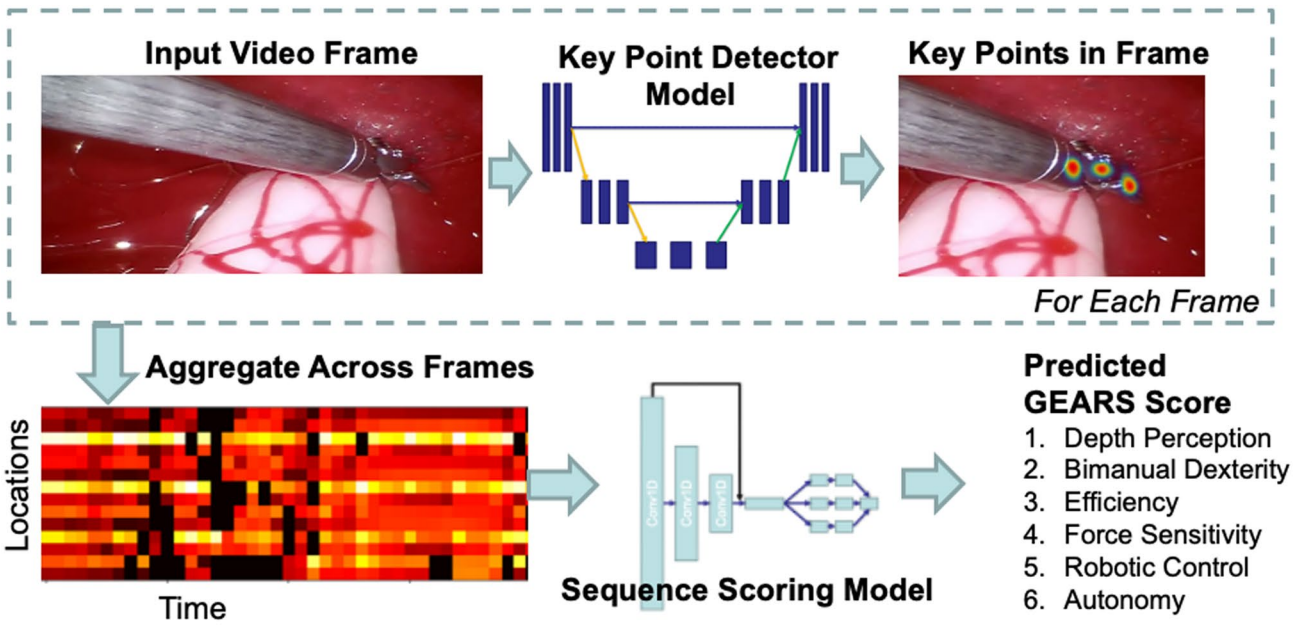


**Fig. 1** An overview of the processing pipeline employed in our analysis

their approach could be used to find angles of instruments and relative location of arm joints.

*Sequential data modeling* In addition to key point detection, our model makes use of sequence analysis to score each time series of object positions. Sequential series analysis aims to predict the future data based on current and historical data. temporal convolutional neural networks (CNN) [11] are a popular modeling procedure for sequential analysis that can be trained with parallelism, which is more complex for recurrent neural networks (RNN). For many applications, they have been shown to take less training time with prediction performance similar to RNNs [12]. Moreover, a number of more recent models make use of a hybrid approach, whereby filter outputs are multiplied by weighted vectors, called attention [13]. Similarly, we also use weighted multipliers in our model over times when video segments are given greater importance in the final calculation.

*Multi-task learning* (MTL) is a popular method for exploiting information from multiple classification tasks in a single modeling framework. In neural networks, this information sharing is achieved using shared weight representations between tasks—it has been shown to increase accuracy and generalization in a number of applications [14]. In the context of robotic surgery proficiency, each sub-scale of the GEARS score can be treated as a separate classification task from a shared set of features (the input video), which improves model reliability.

*Robotic surgery assessment using deep learning* Many works employ convolutional networks for robotic surgery assessment. Zhao et al. used traditional 2-D convolutional networks for tracking surgical instruments in video [15]. The main aim of this work was not assessment, but to track the instruments for verification. Another related work to ours is from Law et al. [16]. They employ as hourglass network to locate specific parts of a robotic instrument (tips, wrist, arm) form the raw video and compute motion features. From these features, they employ machine learning to make a binary prediction related to surgical proficiency. For labeling, they use a binary version of the GEARS score as evaluation—the authors state that they would prefer to use a non-binary GEARS scoring, but found the results unreliable. Similarly, Lee et al. used instrument tracking to analyze surgical proficiency using convolutional tracking networks. Based on the output of the tracking network, they used a mix of traditional machine learning modeling to predict three levels of surgical skill [17]. Their models' accuracy ranged from 57% up to 83% accuracy for classifying the three levels. In our own previous work [18], we used a detector with a simple convolutional sequence model to assess GEARS scores with up to 78% accuracy among five different levels for each GEARS sub-category. While our work clearly takes inspiration from previous methods, we make a number of improvements and extensions to the system that increases throughput, requires less data, and can be used to assess sub-categories of the GEARS score using its entire dynamic range.

## Methods

In this work, we employ a dataset collected in our previous work [4]. The aim of this dataset was to evaluate how the GEARS surgical scoring system performed for a prostatectomy using artificial tissue. The dataset includes 18 robotic urological videos that recorded how trainees and experts sutured artificial urinary tissue in the final step of a prostatectomy procedure. These 18 videos are segmented into 74 non-overlapping segments and each segment is rated according to the GEARS rating system. All of the video segments are evaluated by more than one physician, so they have multiple GEARS ratings. Most discrepancies are small between raters.

In addition, the videos were mostly visually consistent—each of the 18 videos was a different surgeon performing with the same surgical instruments, in the same environment, and completing the same task. The surgeon was operating two robotic arms to suture artificial urinary tissue. This visual consistency is advantageous for creating an automated system because it reduces the need for large-scale data collection. That is, the model can train effectively with fewer examples because we only seek to analyze videos of synthetic tissue with consistent visual characteristics. Therefore, while the relatively few videos in the dataset might be considered a limitation of our study, the visual consistency greatly mitigates this limitation. Table 1 shows other general information about the dataset.

The "Global Evaluative Assessment of Robotic Skills" (GEARS) score were used to evaluate the skill of trainees [4, 19]. This scoring contains various elements for an evaluator to consider including: Depth Perception (DP), Bi-manual Dexterity (BD), Efficiency (E), Force Sensitivity (FS), Autonomy (A), Robotic Control (RC). We use the same scoring sheet and instructions available from [19]. In this dataset, the videos are sampled at rate of thirty frames per second (30 FPS). For each frame, specific parts of surgical instruments can be tracked using our key point detector. We believe that the movement of

**Table 1** General dataset information

| | |
|---|---|
| Total videos | 18 |
| Total segments | 74 (56 train, 18 eval.) |
| GEARS per video | 24 |
| Video length | 8–12 min, Avg.: 10.4 min |
| Segment length | 30s–5 min, Avg.: 2.2 min |
| Experience | 2 Fac., 5 Fel., 11 Res. |

surgical instrument can give us clues regarding the skill of the trainee. In conversations with our surgical team, we identified seven unique objects in the video that play important roles during operation. These seven objects are the robotic arm Rear Joint, Front Joint, and Claw tip (for both the left and right arm) and the Needle. For each object, we locate the coordinates of it in the image $(x, y)$ and save these coordinates for further processing. For the Needle, we identify two points corresponding to the start and end points. These points are marked manually for a subset of training frames and used to train our "key point detector model." After training, these points are extracted automatically by the model.

## Model architectures

Figure 1 presents the overview of our method. First, the key point object detector is used to identify surgical instruments in each video frame and the coordinates of surgical instruments are saved. Second, the positions of surgical instruments are concatenated together as features and used to train a sequence scoring model.

*Key point detector* The overview of the network is shown in Fig. 2. The key points detector is a modified U-Net architecture [7]. It receives a video frame and generates probabilistic heatmaps that match the resolution to the input image, 640 by 360 pixels. One heatmap is generated for each key point and one heatmap for the background. Each heatmap

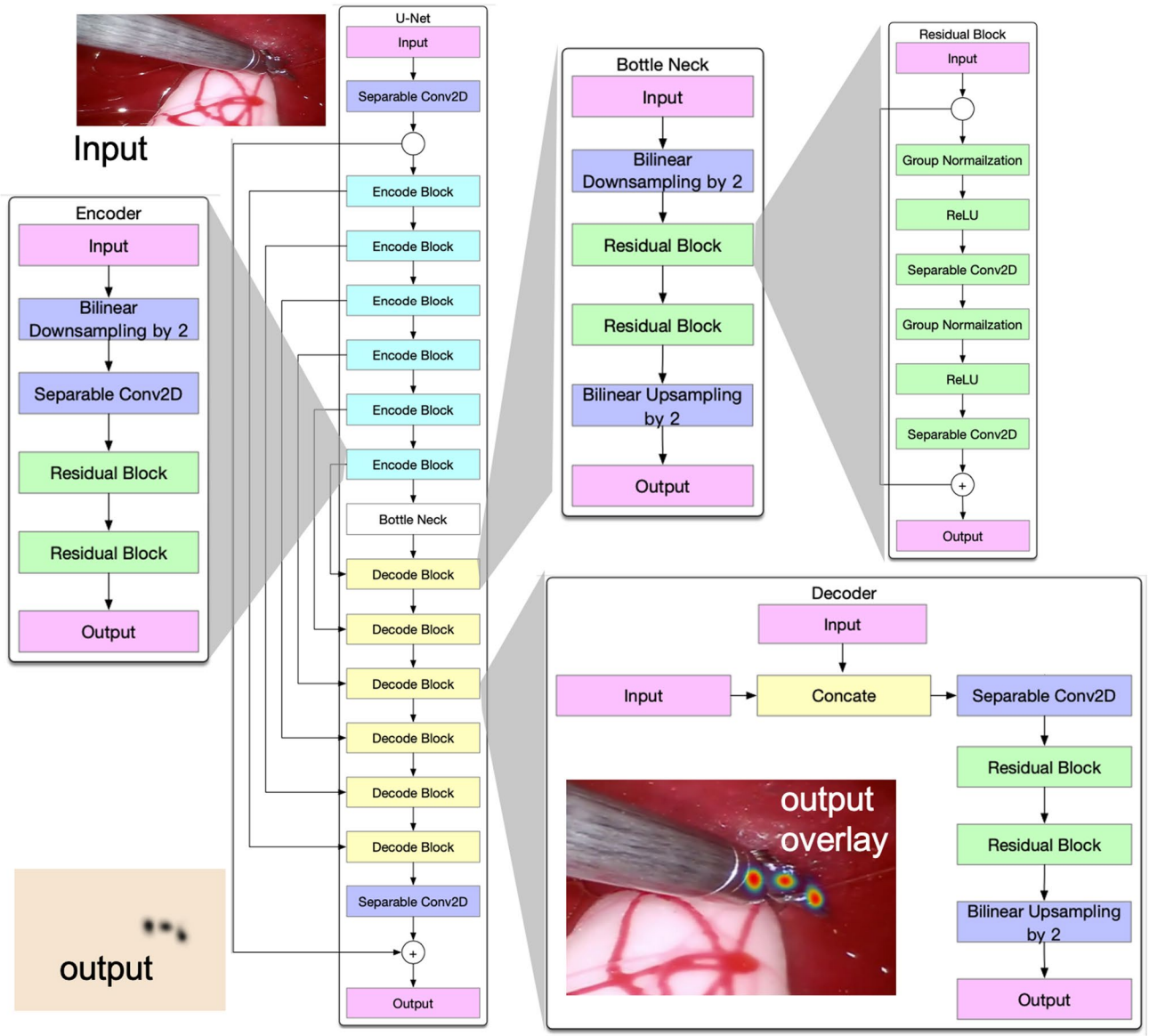


**Fig. 2** Our key point detector model, which estimates locations of joints and instruments in the video. Various breakout overlays are also shown

generated corresponds to the probability that the surgical instrument is located at each pixel location. That is, eight of nine outputs predict the probability of joints and the tip of the needle. The ninth output is the "background," which is used for training the detector, but is discarded before providing the predicted key points to the next phase. Finally, the model uses a Hough vote to calculate the coordinate of the joint [20]. The Hough vote helps to assign probability to groups of pixels by looking at all heatmaps together. Once the voting is complete, the detector gives an $x$ and $y$ coordinate for any detected key points. We save this centroid to represent the location of the object, resulting in a 16-dimensional vector (six joints with $x$ and $y$ pixel locations in the frame of the video and the two endpoints of the needle).

*Training details* The U-net architecture used is a convolutional neural network that contains six encoding blocks, six decoding blocks, and a bottleneck block. The encoding blocks downsample the input image and the decoding block combines the previous block output and the encoding block output from early layers. "Downsampling" and "Upsampling" are image resizing methods that change the size of the feature maps. Feature maps are sets of images that show how the neural network responds at each layer. The bottleneck is the smallest feature map in the network. By hierarchically using downsampling and upsampling, the U-Net architecture can effectively reduce the dimensionality of the input data. The model also consists of a series of residual blocks that employ "identity mapping" [21]. Batch Normalization is a widely used technology in neural networks that allows the network to learn how to scale different feature maps. Layer Normalization has the same purpose as Batch Normalization, but has better performance if the batch size is small. Therefore, the batch normalization typically used in U-net is replaced with group normalization due to the limited size of the batch used in our model [22]. The network's final layer is a pixel-wise Softmax across the nine heatmaps (i.e., a probability is given for each pixel). Cross-entropy to calculate the prediction's loss during training. In total, the key point detector is trained on 560 labeled video frames from the dataset.

*Multi-task sequence scoring model* In previous work comparing single task modeling to a multi-task modeling for robotic surgery scoring, we found that the multi-task model consistently out performed the single task model [18]. Therefore, we only consider the multi-task GEARS score in this work. By training our model in this way, we incentivize the model to extract patterns from the sequences that are important for for all GEARS sub-categories. Then, a separate regression network is trained more specifically to each GEARS sub-category. As shown in Fig. 3, the scoring architecture processes the sequence of key point locations across all the frames of the video. Two pathways are traversed in the model, a time convolution and gate processing pathway. The time convolution extracts meaningful patterns from the sequence over time. The gate processing path informs which time segments are most meaningful. This pathway uses the outputs from the time convolutions and previous gate processing blocks. The two pathways are merged using the "weighted sum" block, which uses the gated blocks to aggregate the time convolution patterns into a single vector representation. This weighted sum is how our network "focuses" on certain time segments in the video. Segments that are judged to be important for assessing GEARS scores are given greater weight. Also, note that these weights are learned by the network so each video is given a dynamically calculated weight. After aggregation, the network splits into separate branches for each GEARS sub-category (i.e., multi-task branches). These branches collapse the features from the convolution into a single score for each category. Finally, the six categories are concatenated into a single vector that denotes the predicted GEARS score for each subcategory. The total GEARS score is calculated as the sum of all sub-categories.

*Model training* We separate the video segments into training and testing data as a validation procedure to allow better generalization of the model and prevent memorization. That is, we divide up the dataset into 74 segments, reserving 56 for training and 18 videos for evaluation (assigned randomly, but ensuring that each surgeon has one video segment for testing). To make our prediction as close as possible to human rated GEARS score, *mean squared error* is used as a loss function for each GEARS score.

## Results

We separate our analyses into four sections that build from one another. First, we describe the key point detector accuracy, followed by two analyses of the scoring network. We analyze the GEARS sub-category accuracy and the "novice, intermediate, expert" distinguishing ability of the network. Finally, we discuss the attention mechanism of the scoring network, investigating its ability to discover relevant portions of the video that are influential to GEARS scoring. We conclude that the model is accurate, but requires further data collection and investigation to understand how it assigns importance to time segments of video.

### Key point detector evaluation

We first evaluate the key point detector qualitatively with visual examples, shown in Fig. 4. These four examples showcase different instrument positions and crossovers. In each case, the detector identifies reasonable locations for the key points. To evaluate the key point detector more quantitatively, we randomly choose 80 frames from videos

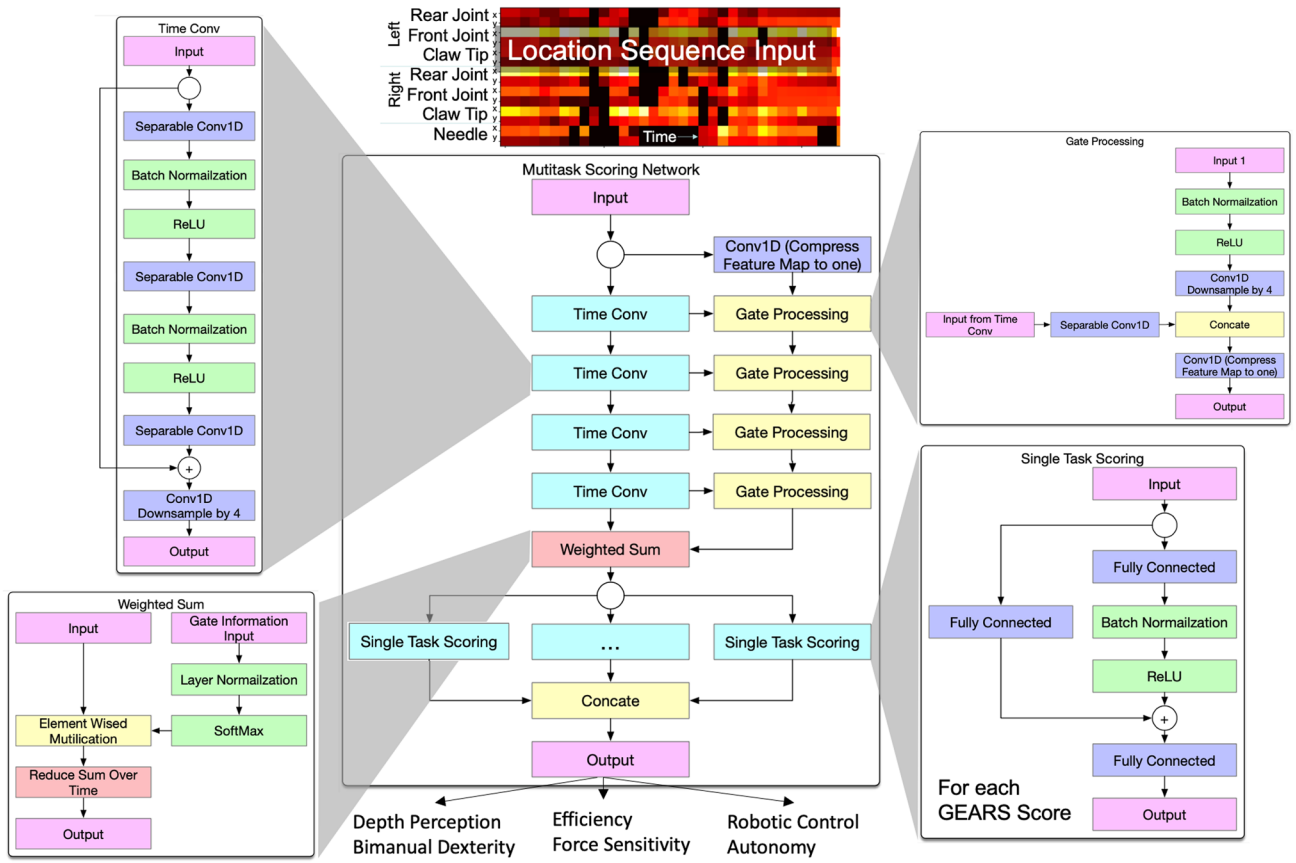


**Fig. 3** Our Scoring network, which processes estimated locations of joints and instruments in the video. Top heatmap: Visualization of object position from video

**Fig. 4** Results from our key point detector. Specific parts of surgical instruments were clearly identified with high confidence. Coordinates of surgical instruments for each frame are saved and sent to our scoring model

throughout our dataset for evaluation. We selected frames that contained most of the key points in them (that is, both left and right instruments, without many occluded points). We did not include the key points associated with the needle because many frames do not have a clear view of the needle (or the needle is placed out of the frame). This results in 420 key points among the 80 frames. We labeled each of these key points in each frame manually. We then compared the key point centroid locations from the model with our manual labels. Two evaluation criteria are employed: (1) if a key point is found in the image (detection) and (2) if detected, how far away are the points, normalized by the image width. Table 2 summarizes these results. Of the 420 key points, 297 are detected properly (leaving 123 missed key points). Upon further investigation of these misses, we found that many of the missed point were momentary "drops" in the instrument tracking. That is, the instrument was detected

**Table 2** Evaluation of key point detector

| Evaluation metric | Value | Evaluation metric | Value |
|---|---|---|---|
| Key points detected | 297/420 (70.7%) | Average difference | 16.9px (2.6%) |
| Median difference | 11.4px (1.8%) | STD of difference | 14.7px (2.3%) |

soon after losing tracking (typically within 5 frames). On rare occasions, some "drops" lasted as long 1.5 s. These were typically the result of blurred video segments. Moreover, it was very rare that a single frame missed more than 1 or 2 key points. Only one of the 80 frames had 3 misses, for example.

When the instrument was detected, the found location was within 16.9 pixels of the actual location, on average. This translates to about 2.6% of the image width. Thus, we conclude that the key point detector is reliable for tracking most frames and when the instrument is found, the location of the centroid is accurate. Finally, we note that perfect tracking accuracy is not necessarily required because the scoring network can process the sequence in a way that helps to correct for these imperfections in tracking.

*GEARS sub-category evaluation* In this analysis, we investigate the sequential scoring model's ability to identify each GEARS score on a scale of 1–5. We define a "matching" prediction to our human ratings for each GEARS sub-category based on the multiple human ratings that video has received. A prediction for the GEARS sub-category is considered "valid" if the prediction is within one point of the human rating. Of the 108 GEARS predictions (18 video segments × 6 GEARS domains), 86.1% of all GEARS sub-categories are valid. If we expand the definition to define a valid entry as when the predicted score is within two points, then 100% of all GEARS sub-categories are valid. To further elucidate this performance for each GEARS score, Table 3 shows the results per sub-category. In this table, a match is considered valid when the prediction is within one point of the human reviewers. Table 3 reveals an important limitation of our modeling: the autonomy score only matches for 72.5% of the results, the lowest of all sub-categories. This is likely because the model does not have direct information to understand the autonomy of a given trainee. For instance, verbal commands that would indicate the trainee has decreased autonomy, are not captured by the model. However, when a surgeon receives verbal instructions there may be some characteristics (such as pausing in tool motions) that indicate feedback is occurring. We hypothesize that this is why the model does not fail entirely on the autonomy sub-category but struggles most with this prediction.

*Novice-expert assessment* In this analysis, we summarize the results of the models by aggregating to a single score based upon the conventional GEARS rating system. Specifically, for each video, we sum all GEARS scores for each sub-category to achieve a single score of proficiency. We define levels of proficiency as is typically applied to sums of GEARS scores: expert >25, intermediate 20–25, and novice <20. Of the 18 videos, 15 were classified into the correct category (83.3%). The model accurately predicted novice surgeons in 11/11 cases (100%). The model classified intermediate surgeons correctly in 4/5 cases (80%). However, the model failed to correctly identify an expert surgeon in any case (0/2). There are several possible reasons for why the model fails to identify videos that are performed by experts. First, fewer examples of experts were available for training, which is crucially important for a machine learning algorithm such as this. We hypothesize this is the main reason for decreased accuracy. Even so, while the training data do not include many expert examples, experts are not the primary use case for our model. That is, we expect trainees to be the main users, where novices and intermediate surgeons are far more common. It is also possible that the model performs well at finding errors in the videos (which are more numerous for novices), but struggles when few errors are seen (experts). To elucidate this further, we also investigate what parts of the videos the model selected as "most influential" in its weighted sum merging block (which aggregates frames over time).

## Important video segments analysis

To analyze the importance of segments within each video, we saved the "weighted sum" merge block output from the scoring model for each video. We refer to these as "importance weights" or IWs. These IWs give a weight for each frame in the video. Larger IWs result in that segment of the video having more influence on the assigned GEARS score. Smaller IWs indicate that the time segment is largely ignored. When analyzing the distribution of IWs, we see that the model clearly assigned certain portions with higher weights and many portions were assigned weights near zero, indicating that the model had clear preferences for certain time segments. Given this observation, we grouped IWs in each video based on percentile of all IWs across the dataset. Anything above the 30th percentile was deemed "high impact" and anything in the range of 15th–30th percentile was deemed "medium impact." Using this definition, 9 of the 18 videos had segments deemed high impact and 6 videos with "medium impact", each lasting about 10 s in duration. Three videos did not have any IWs above the 15th percentile. Interestingly, these three videos

**Table 3** Evaluation of model performance for each GEARS sub-category

| Sub-category | Mean Diff. | Matching | Sub-category | Mean diff. | Matching |
|---|---|---|---|---|---|
| Depth perception | 0.69 | 94.4% | Force Sens. | 0.72 | 83.3% |
| Bi-manual Dext. | 1.04 | 77.8% | Autonomy | 0.85 | 72.2% |
| Efficiency | 0.64 | 94.4% | Robotic control | 0.74 | 88.9% |

included the two expert surgeons from our dataset. Of the 15 videos with high- and medium-impact segments, we recruited a surgeon to review the videos and observe the time segments for any maneuvers or techniques that might be informative for the GEARS scores. For all time segments, the reviewer responded that a technique was employed at that time that could influence their rating of the GEARS score. While the reviewer assessed that the time segment could be influential, they did not always agree that it was *most* influential for determining the score. We did observe that about half the time segments coincided with poor technique, and about half with good technique. Further research is required to understand the role of IWs. Even so, it is clear that the time segments chosen by the model are influential, but the direction of influence (negative or positive) does not appear to be easily explained.

## Conclusion

In this paper, we developed a deep learning method to evaluate robotic-assisted surgery video. Our proposed method detects the positions of surgical instruments with a key point detector model. The recognizer tracks surgical instrument positions over time and uses these time series features to infer the proficiency (GEARS score). Specifically, we present a convolutional multi-task model capable of using the instrument positions to predict six categories of GEARS scores reliably. Experimental results demonstrate that the proposed method achieves good performance on a surgery video dataset of trainees performing a prostatectomy with scores matching human raters in 86.1% of all GEARS sub-categories. Furthermore, the model can detect the difference between proficiency (novice to expert) in 83.3% of videos.

**Funding** Internal research and development funds were employed.

**Availability of data and material** Videos available upon request, with proper data management plan and IRB reciprocity approval.

**Code availability** All source code is custom developed and available upon request (https://github.smu.edu/48066464/UTSW-Surgery-Project)

## Declarations

**Conflict of interest** Authors Wang, Dai, Morgan, Elsaied, Garbens, Qu, Steinberg, Gahan, and Larson declare that they have no conflict of interest.

**IRB approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** was obtained from all individual participants included in the study. This article does not contain any studies with animals performed by any of the authors.

## References

1. Davies B (2015) Robotic surgery—a personal view of the past, present and future. Int J Adv Robot Syst 12(5):54. https://doi.org/10.5772/60118
2. Fard MJ, Pandya AK, Chinnam RB, Klein M, Ellis R (2016) Distance-based time series classification approach for task recognition with application in surgical robot autonomy: task and gesture recognition in robotic minimally invasive surgery. Int J Med Robot Comput Assist Surg 1:3. https://doi.org/10.1002/rcs.1766
3. Ghani KR, Miller DC, Linsell S, Brachulis A, Lane B, Sarle R, Dalela D, Menon M, Comstock B, Lendvay TS, Montie J, Peabody JO (2016) Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. Eur Urol 69(4):547–550
4. Johnson BA, Timberlake M, Steinberg RL, Kosemund M, Mueller B, Gahan JC (2019) Design and validation of a low-cost, high-fidelity model for the urethrovesical anastomosis in radical prostatectomy. J Endourol Ja 33:331–336
5. Han J, Zhang D, Cheng G, Liu N, Dong X (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100
6. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. IEEE, vol 86, issue 11, pp 2278–2324. https://doi.org/10.1109/5.726791
7. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the international conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
8. Hasan SMK, Linte CA (2019) U-NetPlus: a modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp 7205–7211
9. Islam M, Atputharuban DA, Ramesh R, Ren H (2019) Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. IEEE Robot Autom Lett 4(2):2188–2195
10. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI (2018) Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE International Conference on machine learning and applications (ICMLA). IEEE, pp 624–628
11. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882
12. Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of CNN and RNN for natural language processing. CoRR https://arxiv.org/abs/1702.01923
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762

14. Ramachandram D, Taylor GW (2017) Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process Mag 34(6):96–108
15. Zhao Z, Voros S, Weng Y, Chang F, Li R (2017) Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method. Comput Assist Surg 22(sup1):26–35
16. Law H, Ghani K, Deng J (2017) Surgeon technical skill assessment using computer vision based analysis. In: Proceedings of the machine learning for healthcare conference, pp 88–99
17. Lee D, Yu HW, Kwon H, Kong H-J, Lee KE, Kim HC (2020) Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. J Clin Med 9(6):1964
18. Gahan J, Steinberg R, Garbens A, Qu X, Larson E (2020) Machine learning using a multi-task convolutional neural networks to accurately assess robotic skills. J Urol 203(Supplement 4):e505–e505
19. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252
20. Tombari F, Di Stefano L (2010) Object recognition in 3d scenes with occlusions and clutter by hough voting. In: 2010 Fourth Pacific-Rim Symposium on image and video technology. IEEE, pp 349–355
21. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: European Conference on computer vision. Springer, pp 630–645
22. Wu Y, He K (2018) Group normalization. In: Proceedings of the European Conference on computer vision (ECCV), pp 3–19