LETTER

The Institution of Engineering and Technology **WILEY**

Automated surgical skill assessment in endoscopic pituitary surgery using real-time instrument tracking on a high-fidelity bench-top phantom

Adrito Das[1] | Bilal Sidiqi[1] | Laurent Mennillo[1] | Zhehua Mao[1] | Mikael Brudfors[2] | Miguel Xochicale[1,3] | Danyal Z. Khan[1,4] | Nicola Newall[1,4] | John G. Hanrahan[1,4] | Matthew J. Clarkson[1,5] | Danail Stoyanov[1] | Hani J. Marcus[1,4] | Sophia Bano[1]

[1]UCL Hawkes Institute, University College London, London, UK

[2]NVIDIA, London, UK

[3]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

[4]Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK

[5]Department of Medical Physics and Biomedical Engineering, University College London, London, UK

Correspondence Adrito Das, UCL Hawkes Institute, University College London, London, UK. Email: adrito.das.20@ucl.ac.uk

Abstract

Improved surgical skill is generally associated with improved patient outcomes, although assessment is subjective, labour intensive, and requires domain-specific expertise. Automated data-driven metrics can alleviate these difficulties, as demonstrated by existing machine learning instrument tracking models. However, these models are tested on limited datasets of laparoscopic surgery, with a focus on isolated tasks and robotic surgery. Here, a new public dataset is introduced: the nasal phase of simulated endoscopic pituitary surgery. Simulated surgery allows for a realistic yet repeatable environment, meaning the insights gained from automated assessment can be used by novice surgeons to hone their skills on the simulator before moving to real surgery. Pituitary Real-time INstrument Tracking Network (PRINTNet) has been created as a baseline model for this automated assessment. Consisting of DeepLabV3 for classification and segmentation, StrongSORT for tracking, and the NVIDIA Holoscan for real-time performance, PRINTNet achieved 71.9% multiple object tracking precision running at 22 frames per second. Using this tracking output, a multilayer perceptron achieved 87% accuracy in predicting surgical skill level (novice or expert), with the 'ratio of total procedure time to instrument visible time' correlated with higher surgical skill. The new publicly available dataset can be found at https://doi.org/10.5522/04/26511049.

1 | INTRODUCTION

Benign tumours of the pituitary gland, pituitary adenomas, are common, associated with systemic morbidity and mortality, and the majority are curable with surgery [1–3]. The endoscopic transsphenoidal approach (eTSA), is a minimally invasive surgery where these tumours are removed by entering through a nostril [4]. However, this surgery has a steep learning curve, with superior surgical skill generally associated with superior patient outcomes [3, 5, 6].

Objective Structured Assessment of Technical Skills (OSATS) measures surgical skill by assessing how well aspects of a surgical task are performed on a scale of 1–5 [7]. For example, for the aspect of instrument handling, a value of 1 indicates 'Repeatedly makes tentative or awkward moves with instruments', and a value of 5 indicates 'Fluid moves with instruments and no awkwardness' [8]. However, it is not operation specific; liable to interpreter variability; and is a time-consuming manual process requiring surgical experts [9, 10]. Data-driven metrics
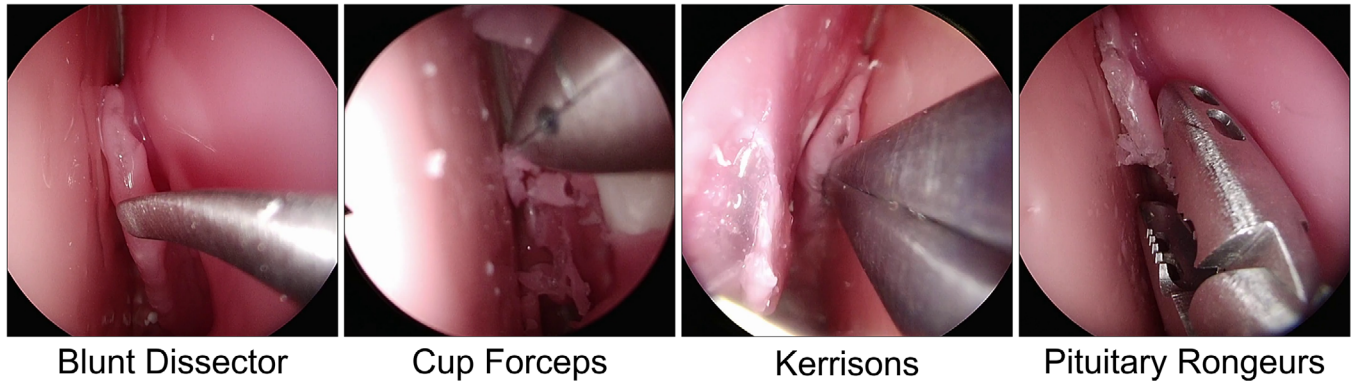
FIGURE 1   Representative images of the four instrument classes used in the nasal phase of endoscopic pituitary surgery.

may be more specific, objective, reproducible, and easier to automate.

Neural networks can automatically and accurately determine surgical skill [11, 12]. More specifically, instrument tracking has been shown to be associated with OSATS in minimally invasive surgeries [11]. However, the models have been tested on limited datasets with a focus on laparoscopic surgeries [11]. Pedrett et al. [11] provides a comprehensive list of these datasets, which are videos of: isolated tasks (e.g. peg transfers in JIGSAWS [13]), real surgery (e.g. Cholec-80 with no publicly available surgical skill assessment [14]), on robotic surgery (e.g. ROSMA [15]), or include instrument tracking data from built in methods (e.g. [16]) or wearable sensors (e.g. [17]).

Here, this previous work is extended to be tested on videos of a high-fidelity bench-top phantom of the full nasal phase of eTSA. These videos are therefore of a non-laparoscopic, non-private, non-robotic, and non-task-isolated surgery with no tracking data. This phantom is commonly used in neurosurgical training to simulate real surgery, and so surgical skill is an important measure to track a novice surgeon's progress until they are able to perform real surgery. Additionally, the insights gained from the automated assessment can be used to isolate specific areas of improvement for the novice surgeon. In real surgery, surgeons are already of sufficient skill, and surgical skill assessment has the alternative use of correlating certain practices with patient outcomes.

Moreover, instrument tracking in eTSA provides a unique computer vision challenge due to: (I) A non-fixed endoscope leading to large camera movements; (II) The frequent withdrawal of instruments leading to instruments having a range of sizes; (III) The use of niche instruments leading to heavy class imbalance; (IV) The smaller working space requiring the use of a wide lens, distorting images (see Figure 1). To overcome these challenges, Pituitary Real-time INstrument Tracking Network (PRINTNet) has been created, and the output is used to demonstrate correlations between instrument tracking and surgical skill. Therefore, this paper's contributions are:

1 The first public dataset containing both instrument and surgical skill assessment annotations in a high-fidelity bench-top phantom of eTSA.

2 A baseline network capable of automated classification, segmentation, and tracking of the instruments in the nasal phase of eTSA, integrated on a NVIDIA Clara AGX for real-time assistance in surgical training sessions.

3 Statistical analysis between instrument tracking and surgical skill assessment in eTSA.

2 | RELATED WORK

Instrument classification in eTSA has been attempted in the PitVis-EndoVis MICCAI-2023 sub-challenge [18], where 25 videos and 8 videos of real eTSA (complete videos) were used for training and testing, respectively.

Instrument segmentation and tracking is yet to be explored for eTSA, though it has been attempted in minimally invasive surgeries since 2016 [19, 20]. Modern models use encoder–decoder architectures, utilising U-Net [21] and its variants for segmentation [19], and early forms of SORT [22] for tracking [20].

The most similar study to this paper linking instrument tracking to surgical skill assessment is the one conducted on robotic thyroid surgery [23]. Twenty-three videos (simulation and real) were used for training the four-instrument-class tracking model, and 40 simulation videos were used for training the surgical assessment model, with 12 simulation videos used for testing [23]. Mask R-CNN and DeepSORT were used for segmentation and tracking, respectively, achieving 70.0% area under curve (AUC) for tracking a tool tip within 1 mm [23]. A random forest (RF) model was shown to be the best predictor of surgical skill, achieving 83% accuracy in distinguishing between novice, intermediate, and expert surgeons [23]. It was found that 'economy of motion' was the most important predictive factor where camera motion is minimal [23].

Other studies that use tool tracking for surgical skill assessment include one on real non-robotic laparoscopic cholecystectomy [24]. Here, instruments in 80 videos (15 tests) of the Calot triangle dissection phase were tracked [24]. The model consisted of YoloV5 for detection, followed by a Kalman filter and the Hungarian algorithm for tracking, achieving 83% multiple object tracking accuracy (MOTA) and 83% accuracy
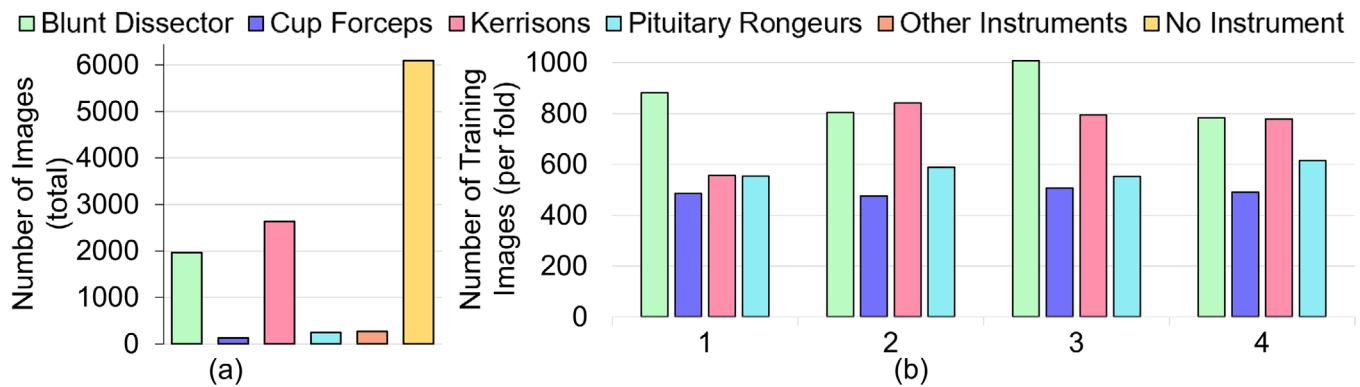
FIGURE 2 Distribution of instruments: (a) Total number of images before data balancing; (b) Number of images per fold after data balancing.

in binary skill assessment via RF [24]. Alternative models, such as those utilising aggregation of local features, have also been used [25]. This model consisted of stacked convolution neural networks (CNNs) followed by bidirectional long short-term memorys (LSTMs) and temporal pooling [25]. On 24 videos (four fold) of the Calot triangle and gallbladder dissection phases of real non-robotic laparoscopic cholecystectomy, the model achieved 46% Spearman's rank correlation on a 1–5 scale [25]. An identical model trained on 30 videos (four fold) of the three isolated robotic tasks found in the JIGSAWS dataset [13] achieved 83% Spearman's rank correlation on a 1–6 scale [25]. This paper extends these methods to a new and unique dataset, in order to test their capability.

3 | DATASET DESCRIPTION

3.1 | Videos

During a surgical training course at the National Hospital for Neurology and Neurosurgery, London, UK, 15 simulated surgeries videos (11426 images) were recorded, one per participating surgeon, using a commercially available high-fidelity bench-top phantom of the nasal phase of eTSA[1] [8]. The participants were recruited from multiple neurosurgical centres within the United Kingdom, with self-reported skill levels (10 novice, 5 expert), receiving tutorials and teaching beforehand. A high-definition endoscope (Olympus S200 visera elite endoscope) was used to record the surgeries at 25 frames per second (FPS) with 720×1080 pixels[2] resolution, and stored as .mp4 files in a surgical video management and analytics platform (Medtronic, Touch Surgery Ecosystem[2]). Ethical approval was granted by the Institutional Review Board (IRB) at University College London (UCL) (17819/011) with informed participation consent.

3.2 | Instrument annotations

Each video was sampled at 1-FPS with 720×1080 pixels[2] resolution, and stored as .png files. Third party annotators (Anolytics[3]) manually annotated each image for instrument boundary and class, which was then verified by two neuro-

surgical trainees and one consultant neurosurgeon. No image contained multiple instruments, and only visible parts of the instrument were annotated if obscured. Figure 2a displays the distribution of the instruments.

3.3 | Surgical skill assessments

Modified OSATS (mOSATS), OSATS curated for pituitary videos, was created, leading to 10 aspects each measured between 1 and 5 [8]. Each video was assessed by two neurosurgical trainees and verified by one consultant neurosurgeon. Inter-rater reliability was calculated using Cohen's Kappa, resulting in 0.949 (confidence interval [CI] 0.983–0.853) for the six general surgical aspects and 0.945 (CI 0.981–0.842) for the four eTSA specific aspects, as defined in the first and second column, respectively, under 'mOSATS Assessment' in Figure 3. Figure 4 displays the mOSATS distribution.

4 | METHODS

4.1 | Instrument segmentation and tracking

4.1.1 | PRINTNet

The simplified diagram of the created architecture is displayed in the dashed green box of Figure 3. The encoder is ResNet50 [26], with no pre-training: a well understood, strong performing, and lightweight CNN commonly used for medical imaging tasks [20], particularly for eTSA recognition [27, 28]. The decoder is DeepLabV3 [29], commonly used in eTSA segmentation [30], which utilises Atrous (also called dilation) convolutions, as opposed to skip connections found in other decoders. These convolutions skip a certain number of pixels (the dilation rate), which increases the receptive field without sacrificing spatial resolution or increasing the number of weights (and so computationally efficient), allowing object features to be captured on multiple spatial scales [29]. This is particularly important for instrument segmentation in eTSA, given the frequency in which instruments are entering and exiting the endoscopic view, and so the same instrument will be found in a variety of sizes.
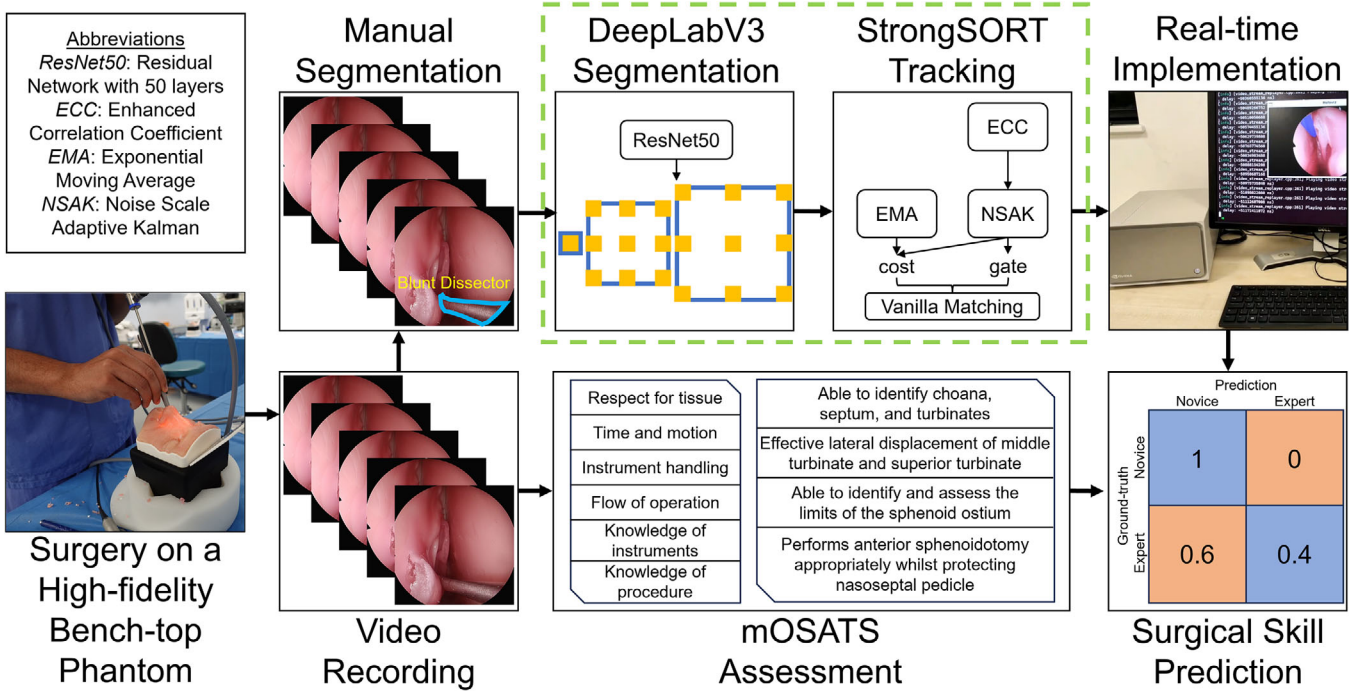
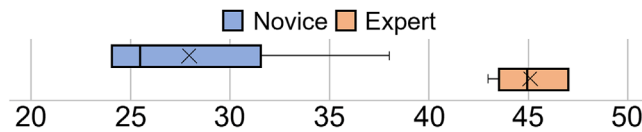**FIGURE 3** Complete workflow diagram of this study.

FIGURE 4 Distribution of mOSATS (10 aspects, max 50) across the 15 videos.

Simple Online and Realtime Tracking (SORT) begins with object detection using a CNN as a feature extractor, followed by object estimation via velocity predictions, and finally ensuring the new objects detected and predicted trajectories of the old objects match [22]. DeepSORT extends SORT through the use of a feature bank (storing features from previous frames), and matching these with the previous predictions [31]. StrongSORT extends DeepSORT through the use of an improved feature extractor, feature bank (now updater), velocity prediction algorithm, and matching algorithm [32]. Moreover, StrongSORT compensates for camera motion by estimating global rotation and translation between frames [32], which is of importance for instrument tracking in eTSA. PRINTNet utilises StrongSORT, replacing the object detection model with DeepLabV3.

4.1.2 | Real-time implementation

The implementation is done via the NVIDIA Holoscan SDK[4] and runs on a NVIDIA Clara AGX[5] [33]. The Holoscan SDK builds a TensorRT[6] engine, which optimises models through reductions in floating point precision, smaller model size, and dynamic memory allocation [33].

4.1.3 | Metrics

Mean Interval over Union (mIoU) was the evaluation metric for segmentation models. Multiple Object Tracking Precision (MOTP) was the evaluation metric for tracking models, and MOTA is given as a secondary metric. MOTA is calculated on every frame, and for frames where the ground-truth classification is unknown, it is assumed the ground-truth classification is unchanged since last known. MOTP is calculated only on frames where ground-truth segmentations, and hence bounding boxes, are known. For these segmentation and tracking metrics a 100% score indicates perfect overlap between the predicted and ground-truth annotation, with 0% indicating no overlap or a missclassification.

FPS was the metric used to compare the speeds of the models. A 25-FPS model would match the native video frame rate and allow for real-time tracking, whereas a lower frame rate model would mean some frames in the video will be skipped.

4.1.4 | Dataset split

Four-fold cross-validation was implemented, as 15 videos are not sufficiently large for a reliable training to testing split. The folds were chosen such that each fold contained approximately the same number of images of a given instrument, but images from one video were only present in one fold. Five instrument classes (Blakesly, Irrigation Syringe, Retractable Knife, Dual Scissors, and Surgical Drill) were removed from the analysis as they appeared in less than four videos, and so could not be present in each fold. This left four instrument classes (Blunt Dissector, Cup Forceps, Kerrisons, and Pituitary Ronguers) as

displayed in Figure 1. Figure 2a displays the stark data imbalance between the instrument classes. To mitigate the effect of over-training on dominant classes, images of the Blunt Dissector and Kerrisons were downsampled by 600 and 1200, respectively, and images of the Cup Forceps and Pituitary Ronguers were upsampled by 400. This was done per fold, and sampled images were chosen at random. Figure 2b displays the resampled dataset (per fold).

4.1.5 | Implementation details

To improve segmentation model training and generalisation, the following augmentation techniques were applied in sequence at random: horizontal flips, vertical flips, rotation, and colour jitters. As a compromise between having a sufficiently large batch size for finding optimal weights during gradient descent and a sufficiently high image resolution for meaningful feature extraction, models were training with a batch size of 16 with training images resized to 288×512 pixels^2, which was able to run on a single NVIDIA Tesla V100 Tensor Core 32-GB GPU.

Cross-entropy was the loss function and Adam with learning rate 0.00006 was the optimiser, as these choices resulted in improved convergence over focal loss; and Dice loss; and other optimiser variations. Each model was run for 50 epochs where the loss function was shown to be sufficiently small (< 0.04) across all folds with minimal changes in subsequent epochs (< 0.005 change after 100 epochs), and so restricting training to 50 epochs limits overfitting and reduces computational time. The model weights of the final (50th) epoch was evaluated on the testing dataset with no early stopping procedure as to be a consistent choice which would not bias the model on any given fold.

The code is written in Python 3.8 using PyTorch 1.8.1 using CUDA 11.2, and is available at https://github.com/dreets/printnet. All videos and annotations are available at https://doi.org/10.5522/04/26511049.

4.2 | Surgical skill assessment

In total, 34 metrics were extracted from the tracking data (see Figure 6). In summary, it consisted of time (e.g. instrument visible time), motion (e.g. acceleration), and usage metrics (e.g. number of instrument switches).

For each metric, a Pearson correlation coefficient (PCC) was calculated against each mOSATS aspect and summed mOSATS. A PCC of 1.0 or -1.0 indicates direct positive or negative correlation, respectively, with 0.0 indicating no correlation.

Then, two classification tasks were then performed: multi-class mOSATS (mean averaged and rounded) and binary-class skill level (novice/expert). For each task, a Linear, Support Vector Machine (SVM); RF; and multilayer perceptron (MLP) model were trained, and boosted via Analysis of Variance (ANOVA) feature selection. A naïve classifier that only predicts the dominant class would achieve 33.3% accuracy in multi-class by predicting '3' and 66.7% accuracy in binary-class by predicting 'novice'.

5 | RESULTS AND DISCUSSION

5.1 | Instrument tracking and segmentation

5.1.1 | Instrument segmentation

It is found that Blunt Dissector and Kerrisons are segmented well, with much worse performances for Cup Forceps and Pituitary Ronguers (see Table 1). This is due to the heavy data imbalance (mIoU = 0 for misclassifications), which is difficult to account for given the small number of images used for testing, even if balance sampling was implemented during training (see Figure 2).

This difficulty in classification is likely because instrument handles are very similar, and take up a large portion of an image due to the image distortion, and so instruments must be distinguished by their relatively small tips. This can be more clearly seen in Figure 5b where PRINTNet struggles to identify the boundary of the Pituitary Rongeur, but is able to identify the boundary of Kerrisons (Figure 5a), a dominant class. This again implies poor classification rather than poor segmentation, which is verified by ablation studies showing $82.2 \pm 0.2\%$ mIoU in binary segmentation.

When compared to other segmentation models, DeepLabV3 has the highest overall mIoU, although closely followed by SegFormer, which also has a significantly higher Cup Forceps mIoU. Given more data, it is likely SegFormer will outperform DeepLabV3, as the transformer encoder performs better with larger datasets [34], extracting both local and global spatial features [35]. U-Net performs worse, as the skip connections between the CNN encoder and upsampling decoder prevents derogation of local and not global spatial information [21].

5.1.2 | Instrument tracking

StrongSORT has the highest MOTP as it accounts for camera motion, although at a lower FPS when compared to SORT due to this extra computation (see Table 2). All models have an identical and high MOTA as classification is determined by the same DeepLabV3 backbone.

Moreover, occasionally, PRINTNet incorrectly predicts an instrument's classification, segmentation, and tracking, such as in Figure 5c, caused by overpredicting the Blunt Dissector tracking paths from previous frames. These incorrect predictions increase the difficulty of surgical skill analysis as some metrics, such as time of instrument usage, may not be reliable.

5.1.3 | Real-time implementation

The accelerated PRINTNet runs at 22 FPS with a 100-ms delay at FP16 precision on the NVIDIA Clara AGX. This is sufficient for real-time use, so PRINTNet can be used during surgical training courses. (See Supplementary Material for a live demonstration of this setup.)

TABLE 1 Segmentation models' mIoU for each of the four instrument classes across the four folds. The highest mIoU for a given instrument is displayed in bold.

Model	Blunt Dissector	Cup Forceps	Kerrisons	Pituitary Ronguers	All Instruments	No instrument
U-Net	63.8 ± 09.2	22.1 ± 18.3	62.1 ± 23.6	18.6 ± 14.7	41.6 ± 9.1	98.4 ± 0.9
SegFormer	63.4 ± 12.7	**24.4 ± 17.9**	60.2 ± 21.5	31.9 ± 24.7	45.0 ± 11.5	98.2 ± 0.6
DeepLabV3	**66.9 ± 15.3**	11.8 ± 10.6	**73.4 ± 28.0**	31.9 ± 22.9	**46.0 ± 09.1**	**98.7 ± 0.5**

(a) Kerrisons   (b) Pituitary Rongeurs   (c) Blunt Dissector
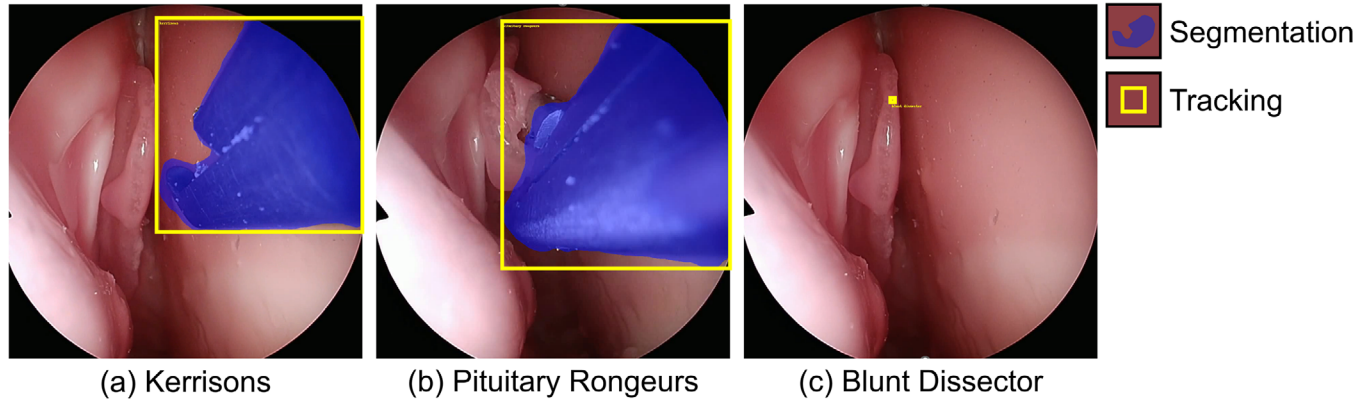
FIGURE 5 Qualitative results of PRINTNet: (a) a strong example where the classification, segmentation, and tracking are accurate; (b) a common example where the classification and tracking are accurate, but the segmentation could be improved at the instrument tip; (c) an uncommon example where classification, segmentation, and tracking are all inaccurate. (See the Supplementary Material for the full video.)

TABLE 2 Tracking models' performance across the four folds. The highest value for a given evaluation metric is displayed in bold. Note that the detection frequency was set to 5.

Model	MOTP (%)	MOTA (%)	FPS (mean)
SORT	59.1 ± 03.1	77.9 ± 07.1	**24.7 ± 00.8**
DeepSORT	62.9 ± 05.0	77.9 ± 07.1	12.8 ± 00.7
StrongSORT	**71.9 ± 05.5**	77.9 ± 07.1	10.6 ± 02.9

TABLE 3 Accuracy in surgical skill classification across the four folds. The highest value for a given metric is displayed in bold.

Model	Multi-class mean mOSATS (%)	Binary-class skill level (%)
Linear	**39.9 ± 24.9**	80.0 ± 16.3
Support Vector Machine	**46.7 ± 26.7**	80.0 ± 26.7
Random Forest	40.0 ± 38.9	**73.3 ± 24.9**
MultiLayer Perceptron	26.7 ± 24.9	**86.7 ± 16.3**

5.2 | Surgical skill assessment

Distinguishing between expert and novice skill level achieved a high 87% accuracy (see Table 3), in line with similar studies [23–25]. However, there was poor accuracy in multi-class mean mOSATS classification, although comparable to similar studies

[25]. This highlights the complexity of the problem, with the implication that more data is required.

Across the 10 aspects, time-based metrics were stronger predictors than motion-based metrics. This is seen in Figure 6 where PCC for summed mOSATS is shown. Specifically, 'ratio of total procedure time to instrument visible time' is found to be positively correlated with mOSATS, indicating instrument efficiency (i.e. a reduced idle time) is correlated with higher surgical skill. Interestingly, it is found that the use of a Blunt Dissector or Cup Forceps is negatively correlated with mOSATS whereas Kerrisons and Pituitary Rongeurs are positively correlated.

The limited correlation between motion-based metrics and mOSATS is an opposing result to that found in robotic thyroid surgery, where instrument motion in the absence of camera motion was a strong predictor [23]. Removing this camera motion is tricky, as large endoscope movements are required to navigate through the nasal phase of eTSA in order to get through the nostril (for both novice and expert surgeons), which outweighs the more subtle movements of the instruments. Although StrongSORT does compensate for this motion, more sophisticated models are needed.

6 | CONCLUSION

Rating surgical skill via instrument tracking during minimally invasive surgery in an objective and reproducible manor remains a difficult task. Existing models have focused on real and robotic laparoscopic surgery, and these models have now been
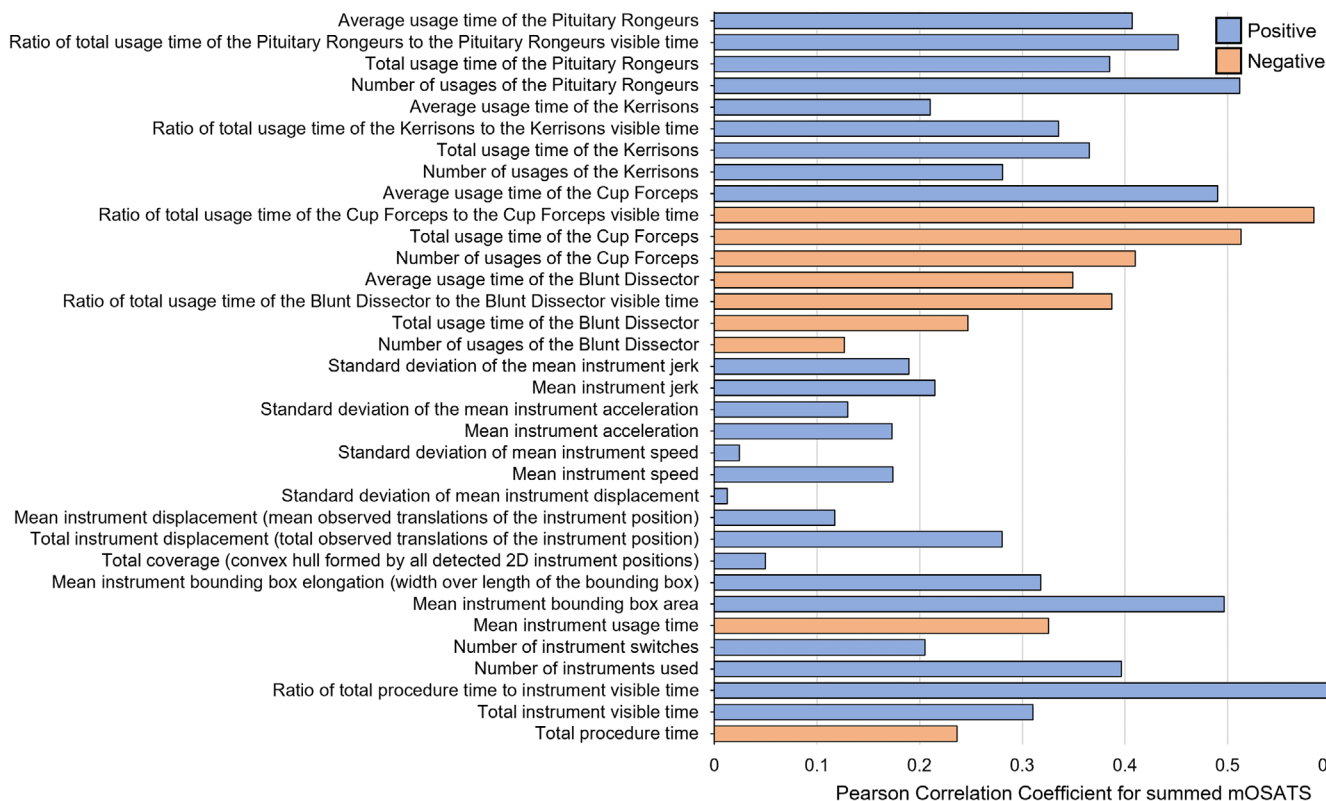
FIGURE 6 Pearson Correlation Coefficient of the 34 metrics for summed mOSATS.

extended to simulated endoscopic surgery. Fifteen videos of the nasal phase of eTSA were performed on a high-fidelity bench-top phantom during a training course and were recorded. They were later assessed for surgical skill by expert surgeons, and instruments were manually segmented. The created model, PRINTNet, designed to classify, segment, and track the instruments during the nasal phase of eTSA achieved 67% and 73% mIoU for the dominant Blunt Dissector and Kerrisons classes, with 72% MOTP. Eighty-seven percent accuracy was achieved with a MLP when using the PRINTNet tracking output to predict whether a surgeon was a novice or expert. Moreover, real-time speeds were achieved when run on a NVIDIA Clara AGX, allowing for real-time feedback for surgeons during training courses. This continuous monitoring of surgical skill allows novice surgeons to consistently improve their skill on simulated surgery before they are sufficiently skilled to perform real surgery. Future work will involve: modifying the model, such as with the use of temporal [24] or anchor free methods [36], collecting a larger dataset, and extending this work to real eTSA – linking instrument tracking to both surgical skill and real patient outcomes. For now, this paper provides a new and unique publicly available dataset and baseline network, which can be improved on by the community.

AUTHOR CONTRIBUTIONS

Adrito Das: Conceptualization; data curation; formal analysis; software; visualization; writing—original draft; writing—review and editing. **Bilal Sidiqi**: Formal analysis; methodology. **Laurent Mennillo**: Formal analysis; visualization. **Zhehua Mao**: Software. **Mikael Brudfors**: Software. **Miguel Xochicale**: Software. **Danyal Z Khan**: Conceptualization; data curation; writing—original draft; writing—review and editing. **Nicola Newall**: Data curation. **John G Hanrahan**: Data curation. **Matthew Clarkson**: Resources; writing—review and editing. **Danail Stoyanov**: Funding acquisition; resources; supervision. **Hani J Marcus**: Conceptualization; data curation; supervision; writing—original draft; writing—review and editing. **Sophia Bano**: Conceptualization; resources; supervision; writing—review and editing.

ACKNOWLEDGEMENTS

[NS/A000050/1] and by the NIHR Biomedical Research Centre at UCL.

CONFLICT OF INTEREST STATEMENT

DATA AVAILABILITY STATEMENT

ORCID

Adrito Das https://orcid.org/0000-0001-5682-9545

ENDNOTES

1 https://www.store.upsurgeon.com/products/tnsbox/
2 https://www.touchsurgery.com/
3 https://www.anolytics.ai/
4 https://github.com/nvidia-holoscan/holoscan-sdk
5 https://www.nvidia.com/en-gb/clara/intelligent-medical-instruments
6 https://developer.nvidia.com/tensorrt

REFERENCES

1. Asa, S.L.: Practical pituitary pathology: What does the pathologist need to know?. Arch. Path. Lab. Med. 132(8), 1231–1240 (2008)
2. Ezzat, S., Asa, S.L., Couldwell, W.T., Barr, C.E., Dodge, W.E., Vance, M.L., McCutcheon, I. E.: The prevalence of pituitary adenomas: A systematic review. Cancer 101(3), 613–619 (2004)
3. Khan, D.Z., Hanrahan, J.G., Baldeweg, S.E., Dorward, N.L., Stoyanov, D., Marcus, H.J.: Current and future advances in surgical therapy for pituitary adenoma. Endocr. Rev. 44(5), 947–959 (2023)
4. Marcus, H.J., Khan, D.Z., Borg, A., Buchfelder, M., Cetas, J.S., Collins, J.W., Dorward, N.L., Fleseriu, M., Gurnell, M., Javadpour, M., Jones, P.S., Koh, C.H., Layard Horsfall, H., Mamelak, A.N., Mortini, P., Muirhead, W., Oyesiku, N.M., Schwartz, T.H., Sinha, S., Stoyanov, D., Syro, L.V., Tsermoulas, G., Williams, A., Winder, M.J., Zada, G., Laws, E.R.: Pituitary society expert delphi consensus: Operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituitary 24(6), 839–853 (2021)
5. Leach, P., Abou Zeid, A.H., Kearney, T., Davis, J., Trainer, P.J., Gnanalingham, K.K.: Endoscopic transsphenoidal pituitary surgery: Evidence of an operative learning curve. Neurosurgery 67(5), 1205–1212 (2010)
6. McLaughlin, N., Laws, E.R., Oyesiku, N.M., Katznelson, L., Kelly, D.F.: Pituitary centers of excellence. Neurosurgery 71(5), 916–926 (2012)
7. Martin, J.A., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents: Objective structured assessment of technical skill. Br. J. Surg. 84(2), 273–278 (1997)
8. Newall, N., Khan, D.Z., Hanrahan, J.G., Booker, J., Borg, A., Davids, J., Nicolosi, F., Sinha, S., Dorward, N., Marcus, H.J.: High fidelity simulation of the endoscopic transsphenoidal approach: Validation of the upsurgeon tns box. Front. Surg. 9, 1049685 (2022)
9. Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., Murakami, S., Saeki, S., Mukaida, H., Takiyama, W.: Using the objective structured assessment of technical skills (OSAts) global rating scale to evaluate the skills of surgical trainees in the operating room. Surg. Today 43(3), 271–275 (2012)
10. Khan, D.Z., Koh, C.H., Das, A., Valetopoulou, A., Hanrahan, J.G., Horsfall, H.L., Baldeweg, S.E., Bano, S., Borg, A., Dorward, N.L., Olukoya, O., Stoyanov, D., Marcus, H.J.: Video-based performance analysis in pituitary surgery - part 1: Surgical outcomes. World Neurosurg. (2024). https://doi.org/10.1016/j.wneu.2024.07.218
11. Pedrett, R., Mascagni, P., Beldi, G., Padoy, N., Lavanchy, J.L.: Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review. Surg. Endosc. 37(10), 7412–7424 (2023)
12. Khan, D.Z., Newall, N., Koh, C.H., Das, A., Aapan, S., Horsfall, H.L., Baldeweg, S.E., Bano, S., Borg, A., Chari, A., Dorward, N.L., Elserius, A., Giannis, T., Jain, A., Stoyanov, D., Marcus, H.J.: Video-based performance analysis in pituitary surgery - part 2: Artificial intelligence assisted surgical coaching. World Neurosurg. (2024). https://doi.org/10.1016/j.wneu.2024.07.219
13. Gao, Y., Vedula, S.S., Reiley, C.E., Narges, A., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Bejar, B., Yuh, D.D., Chen, C.C.G., Vidal, R., Khudanpur, S., Hager, G.D.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. Modeling and monitoring of computer assisted interventions (M2CAI)–MICCAI Workshop (2014). https://doi.org/10.1109/TBME.2016.2647680
14. Lavanchy, J.L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., Beldi, G.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci. Rep. 11(1), 8933 (2021)
15. Rivas-Blanco, I., Pérez-del Pulgar, C.J., Mariani, A., Tortora, G., Reina, A.J.: A surgical dataset from the da vinci research kit for task automation and recognition. arXiv (2021). https://arxiv.org/abs/2102.03643
16. 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. Int. J. Comput. Assist. Radiol. Surg. 15(1), 27–40 (2019)
17. Soangra, R., Sivakumar, R., Anirudh, E.R., Reddy Y, S.V., John, E.B.: Evaluation of surgical skill using machine learning with optimal wearable sensor locations. PLoS ONE 17(6), e0267936 (2022)
18. Das, A., Khan, D.Z., Psychogyios, D., Zhang, Y., Hanrahan, J.G., Vasconcelos, F., Pang, Y., Chen, Z., Wu, J., Zou, X., Zheng, G., Qayyum, A., Mazher, M., Razzak, I., Li, T., Ye, J., He, J., Płotka, S., Kaleta, J., Yamlahi, A., Jund, A., Godau, P., Kondo, S., Kasai, S., Hirasawa, K., Rivoir, D., Pérez, A., Rodriguez, S., Arbeláez, P., Stoyanov, D., Marcus, H.J., Bano, S.: Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. arXiv (2024). https://arxiv.org/abs/2409.01184
19. Rueckert, T., Rueckert, D., Palm, C.: Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. Comput. Biol. Med. 169, 107929 (2024)
20. Wang, Y., Sun, Q., Liu, Z., Gu, L.: Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. Rob. Auton. Syst. 149, 103945 (2022)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. arXiv (2015). https://arxiv.org/abs/1505.04597
22. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE, Piscataway, NJ (2016)
23. Lee, D., Yu, H.W., Kwon, H., Kong, H.J., Lee, K.E., Kim, H.C.: 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 (2020)
24. Fathollahi, M., Sarhan, M.H., Pena, R., DiMonte, L., Gupta, A., Ataliwala, A., Barker, J.: Video-based surgical skills assessment using long term tool tracking, pp. 541–550. Springer Nature, Switzerland (2022)
25. Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation, pp. 410–420. Springer Nature, Switzerland (2022)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv (2015). https://arxiv.org/abs/1512.03385
27. Das, A., Bano, S., Vasconcelos, F., Khan, D.Z., Marcus, H.J., Stoyanov, D.: Reducing prediction volatility in the surgical workflow recognition of endoscopic pituitary surgery. Int. J. Comput. Assist. Radiol. Surg. 17(8), 1445–1452 (2022)

28. Das, A., Khan, D.Z., Hanrahan, J.G., Marcus, H.J., Stoyanov, D.: Automatic generation of operation notes in endoscopic pituitary surgery videos using workflow recognition. Intell.-Based Med. 8, 100107 (2023)

29. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv (2017). https://arxiv.org/abs/1706.05587

30. Das, A., Khan, D.Z., Williams, S.C., Hanrahan, J.G., Borg, A., Dorward, N.L., Bano, S., Marcus, H.J., Stoyanov, D.: A multi-task network for anatomy identification in endoscopic pituitary surgery, pp. 472–482. Springer Nature, Switzerland (2023)

31. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. arXiv (2017). https://arxiv.org/abs/1703.07402

32. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. arXiv (2022). Available from: https://arxiv.org/abs/2202.13514

33. Sinha, S., Dwivedi, S., Azizian, M.: Towards deterministic end-to-end latency for medical ai systems in nvidia holoscan. arXiv (2024). https://arxiv.org/abs/2402.04466

34. Sourget, T., Hasany, S.N., Mériaudeau, F., Petitjean, C.: Can segformer be a true competitor to u-net for medical image segmentation?, pp. 111–118. Springer Nature, Switzerland (2023)

35. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv (2021). https://arxiv.org/abs/2105.15203

36. Ding, G., Zhao, X., Peng, C., Li, L., Guo, J., Li, D., Jiang, X.: Anchor-free feature aggregation network for instrument detection in endoscopic surgery. IEEE Access 11, 29464–29473 (2023)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Das, A., Sidiqi, B., Mennillo, L., Mao, Z., Brudfors, M., Xochicale, M., Khan, D.Z., Newall, N., Hanrahan, J.G., Clarkson, M.J., Stoyanov, D., Marcus, H.J., Bano, S.: Automated surgical skill assessment in endoscopic pituitary surgery using real-time instrument tracking on a high-fidelity bench-top phantom. Healthc. Technol. Lett. 1–9 (2024). https://doi.org/10.1049/htl2.12101