Medical Image Analysis 106 (2025) 103716

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

PitVis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery

Adrito Das [a,*], Danyal Z. Khan [a,b], Dimitrios Psychogyios [a], Yitong Zhang [a], John G. Hanrahan [a,b], Francisco Vasconcelos [a], You Pang [c], Zhen Chen [c], Jinlin Wu [c], Xiaoyang Zou [d], Guoyan Zheng [d], Abdul Qayyum [e], Moona Mazher [f], Imran Razzak [g], Tianbin Li [h], Jin Ye [h], Junjun He [h], Szymon Płotka [i,j,k,u], Joanna Kaleta [k], Amine Yamlahi [l], Antoine Jund [l], Patrick Godau [l,m,n], Satoshi Kondo [o], Satoshi Kasai [p], Kousuke Hirasawa [q], Dominik Rivoir [r,s], Stefanie Speidel [r,s], Alejandra Pérez [t], Santiago Rodriguez [t], Pablo Arbeláez [t], Danail Stoyanov [a,1], Hani J. Marcus [a,b,1], Sophia Bano [a,1]

[a] UCL Hawkes Institute, University College London, London, UK
[b] Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK
[c] Centre for AI and Robotics (CAIR) HKISI, CAS, Hong Kong, China
[d] Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[e] National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, UK
[f] Centre for Medical Image Computing, University College London, London, UK
[g] University of New South Wales, Sydney, Australia
[h] Shanghai AI Lab, Shanghai, China
[i] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[j] Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands
[k] Sano Center for Computational Medicine, Krakow, Poland
[l] German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Germany
[m] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Hospital Heidelberg, Heidelberg, Germany
[n] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[o] Muroran Institute of Technology, Hokkaido, Japan
[p] Niigata University of Health and Welfare, Niigata, Japan
[q] Konica Minolta Inc., Osaka, Japan
[r] National Center for Tumor Diseases, DKFZ, UKDD, TUD, HZDR, Dresden, Germany
[s] Centre for Tactile Internet, TUD, Dresden, Germany
[t] Universidad de los Andes, Bogota, Colombia
[u] Faculty of Mathematics and Computer Science, Jagiellonian University, Krakow, Poland

ARTICLE INFO

Dataset link: www.doi.org/10.5522/04/26531686, https://github.com/dreets/pitvis

Keywords:
Endoscopic vision
Instrument recognition
Step recognition
Surgical AI
Surgical vision
Workflow analysis

ABSTRACT

The field of computer vision applied to videos of minimally invasive surgery is ever-growing. Workflow recognition pertains to the automated recognition of various aspects of a surgery, including: which surgical steps are performed; and which surgical instruments are used. This information can later be used to assist clinicians when learning the surgery or during live surgery. The Pituitary Vision (PitVis) 2023 Challenge tasks the community to step and instrument recognition in videos of endoscopic pituitary surgery. This is a particularly challenging task when compared to other minimally invasive surgeries due to: the smaller working space, which limits and distorts vision; and higher frequency of instrument and step switching, which requires more precise model predictions. Participants were provided with 25-videos, with results presented at the MICCAI-2023 conference as part of the Endoscopic Vision 2023 Challenge in Vancouver, Canada, on 08-Oct-2023. There were 18-submissions from 9-teams across 6-countries, using a variety of deep learning models. The top performing model for step recognition utilised a transformer based architecture, uniquely using an autoregressive decoder with a positional encoding input. The top performing model for instrument recognition

* Corresponding author.
E-mail address: adrito.das.20@ucl.ac.uk (A. Das).
[1] These authors contributed equally as senior authors.

https://doi.org/10.1016/j.media.2025.103716
Received 22 September 2024; Received in revised form 5 May 2025; Accepted 30 June 2025
Available online 23 July 2025

utilised a spatial encoder followed by a temporal encoder, which uniquely used a 2-layer temporal architecture. In both cases, these models outperformed purely spatial based models, illustrating the importance of sequential and temporal information. This PitVis-2023 therefore demonstrates state-of-the-art computer vision models in minimally invasive surgery are transferable to a new dataset. Benchmark results are provided in the paper, and the dataset is publicly available at: https://doi.org/10.5522/04/26531686.

1. Introduction

The Endoscopic Vision (EndoVis) challenge[2] has existed since 2015, hosted by the Medical Image Computing and Computer Assisted Interventions (MICCAI) Society (Maier-Hein et al., 2022). Included are a wide range of challenges related to computer vision in minimally invasive surgeries: from polyp detection in colonoscopy videos in 2015 to action recognition on radical prostatectomy videos in 2022 (Maier-Hein et al., 2022). To the minimally invasive surgical computer vision community, the benefits of an EndoVis challenge are two-fold. Firstly, it pushes the boundaries of existing models (Maier-Hein et al., 2022). Secondly, it provides a curated public dataset (Maier-Hein et al., 2022). Building on this, the Pituitary Vision (PitVis) 2023 challenge was created as sub-challenge of EndoVis-2023 (Speidel et al., 2023). The PitVis-2023 challenge pertains to step and instrument recognition in the endoscopic transsphenoidal approach (eTSA) for pituitary adenoma resection.

The pituitary gland is found at the base of the brain (Ganapathy and Tadi, 2022). Tumours of the anterior pituitary gland, pituitary adenomas, have an estimated prevalence of 1 in 1000 of the general population (Russ et al., 2022; Agustsson et al., 2015). Symptoms typically include visual impairment (Russ et al., 2022; Ogra et al., 2014) and hormone imbalances (Ganapathy and Tadi, 2022; Russ et al., 2022). Left untreated, these symptomatic adenomas can cause blindness (Russ et al., 2022; Ogra et al., 2014) or, in cases such as Cushing's disease, be life limiting (Russ et al., 2022; Tritos and Biller, 2019). The gold standard treatment for most patients with a symptomatic pituitary adenoma is surgery, commonly via the eTSA (Ganapathy and Tadi, 2022; Wang et al., 2014).

The eTSA, also called endoscopic pituitary surgery, is a minimally invasive surgery where the tumour is removed by entering through a nostril, as displayed in Fig. 1(a) (Wang et al., 2014; Marcus et al., 2021). The endoscope allows the surgeon to see inside the patient, with the camera feed projected onto a monitor, and is used in conjunction with surgical instruments, as displayed in Fig. 1(b) (Wang et al., 2014; Marcus et al., 2021). The eTSA is performed heterogeneously (Consortium, 2023), and so there is variability in outcomes (Wang et al., 2014). Furthermore, it is a difficult procedure to master, requiring dedicated sub-speciality training (Frara et al., 2020).

Surgical workflow analysis breaks down a surgical procedure into its constituent components (Lalys and Jannin, 2013). At the highest level the surgery is broken down into multiple surgical phases, with each phase corresponding to a major surgical event (Lalys and Jannin, 2013). In turn, each phase is composed of multiple more granular steps (Lalys and Jannin, 2013). Here, each step achieves a certain surgical objective via the use of surgical instruments (Marcus et al., 2021). By splitting a surgical procedure in this way, surgeons can isolate where in a surgery technical errors can occur and how these errors can lead to adverse events (Marcus et al., 2021).

A major limiting factor in reaping the rewards of surgical workflow analysis is the time and labour costs of requiring expert surgeons to analyse each procedure, putting undue burden on an already overworked healthcare system (Wang et al., 2022). For minimally invasive surgeries such as the eTSA this comes down to surgeons manually annotating each surgical video (Marcus et al., 2021). Therefore, automated surgical workflow analysis, commonly referred to as surgical workflow recognition, can alleviate this cost, allowing for scalable analysis (Garrow et al., 2020).

The first clinical benefit of workflow recognition is the coaching of junior surgeons via interactive videos and automated performance metrics, which has shown to improve surgical performance (Khan et al., 2024b,c). A second benefit is found after a surgery, by automating the reporting of steps performed and instruments used (Das et al., 2023a; He et al., 2024). This can reduce the time spent on the writing of operation notes and hence improve surgical throughput (Khan et al., 2023; Das et al., 2023a). A third clinical benefit is during live surgery by automatically informing the wider operating room team (e.g. anaesthetists and theatre nurses) when a new step is to begin or when a new instrument is required (Das et al., 2024; Khan et al., 2024a). This has the potential to improve operating room efficiency (Khan et al., 2023; Garrow et al., 2020).

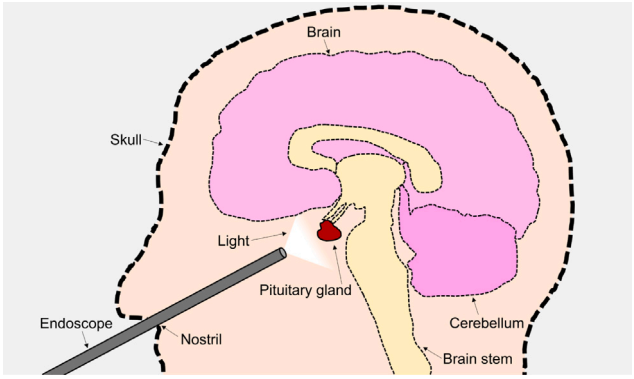
Motivated by these clinical benefits, the PitVis-2023 challenge was created. The challenge consisted of three tasks: (1) step recognition; (2) instrument recognition; and (3) step and instrument recognition. Participants were provided with 25-training-videos (public), along with per-second annotations of the current step and present instrument. Submitted models were evaluated on 8-testing-videos (private). For the public 25-training-videos, a 20-training to 5-validation split was suggested, and benchmark results are provided against these 5-validation-videos as a comparison for models created in the future. Monetary prizes totalling £3000 were awarded to the top two performing models for each of the 3-tasks, based on their performance on the 8-training-videos. Evaluation of performance included both spatial (F_1-score) and temporal (Edit-score) metrics.

18-submission were received from 9-teams across 6-countries. Models ranged from purely spatial encoding with just a Convolution Neural Network (CNN), to spatio-temporal methods which additionally utilised a Long Short Term Memory Network (LSTM). Transformers were also used and found to be effective in both step and instrument recognition. A commonality between the best performing models across all 3-tasks was the use of spatio-temporal representation learning. In CNN + LSTM based models this is through the use of hierarchical temporal modelling, such as with the use of 2-layered LSTMs. In transformer based models this is through the use of frame ordering, such as with the use of positional encoding.

The main contributions of the PitVis-2023 challenge are:

1. A thorough analysis of the state-of-the-art surgical workflow recognition models applied to endoscopic pituitary surgery: more granular than previous step recognition work and the first for instrument recognition in this surgery.
2. Providing benchmark results of surgical workflow recognition in endoscopic pituitary surgery: highlighting the challenges on a unique surgery not previously explored by the community.
3. The first curated public dataset of endoscopic pituitary surgery: 25-videos with each second annotated with its respective step and instrument.
4. A well-attended computer vision challenge associated with endoscopic pituitary surgery: with 18-submissions from 9-teams across 6-countries.

[2] https://opencas.dkfz.de/endovis/.



(a) Endoscopic pituitary surgery diagram.

(b) Endoscopic pituitary surgery operation.

Fig. 1. Visual representations of endoscopic pituitary surgery. Notice the entry of the endoscope and instruments through the nostril, and the use of a light source and monitor to visualise the endoscopic video feed.

The PitVis-2023 challenge followed the BIAS guidelines for transparent reporting of biomedical challenges (Speidel et al., 2023; Maier-Hein et al., 2020).[3] This paper also followed these guidelines, the details of which can be found in the supplementary material.

2. Related works

2.1. Difficulties

In minimally invasive surgery, workflow recognition is a difficult computer vision task for several reasons, including: (i) A variety in surgical practice across different hospitals throughout the globe, resulting in a lack of consensus of which steps are to be performed and instruments to be used (Garrow et al., 2020; Rueckert et al., 2024). (ii) A limited supply of well-curated large annotated public datasets, resulting in models focusing on some surgeries (e.g. laparoscopic cholecystectomy) and so their generalisability has not been well studied (Wang et al., 2022; Demir et al., 2023). (iii) Poor metric selection, often not representative of the underlying clinical motivation (Wang et al., 2022; Maier-Hein et al., 2024).

Additionally, there are several eTSA specific difficulties, including: (iv) Multiple steps and instruments with a high frequency of switching in an undetermined order, more so than in other surgeries (Marcus et al., 2021; Garrow et al., 2020; Das et al., 2022). This increases classification difficulty as the model predictions need to be more precise. (v) The small working space, leading to a thinner endoscope, and hence lense distortion (Das et al., 2022). This means features at the centre of the image appear smaller than features towards the edge of an image. This leads to instrument shafts, which are generally uninformative of the instrument class, to take up a large section of the image; whereas instrument tips, which are more informative of the instrument class, take up a small section of the image (Fig. 4). (vi) Occlusions due to bodily fluids, necessitating the need for the frequent withdrawal of the endoscope outside of the patient's body for cleaning, resulting in temporally inconsistent images (Das et al., 2023a, 2022). (vii) Many of the steps and instruments look similar. For example, instrument-9 (micro doppler probe) and instrument-18 (tissue glue applicator) look identical from a static image, and can only be distinguished by the action performed and the wider surgical context (Fig. 4).

2.2. Step recognition

Historically, a variety of machine learning models were used for step recognition across minimally invasive surgeries, but since 2016, deep learning models have dominated (Garrow et al., 2020; Demir et al., 2023). Typically, step recognition models consist of a 3-stage architecture: stage-1, a per-frame spatial encoder; followed by stage-2, where the per-frame spatial features are consecutively combined and sent to a temporal encoder; and finally stage-3, where the predicted spatial–temporal classifications are turned into a sequence and undergo processing (Garrow et al., 2020; Demir et al., 2023). For stage-1, CNNs are frequently used, although more recently Spatial Transformers (S-TFs) transformers or Spatio-Temporal Transformers (ST-TFs) have been found to be effective (Demir et al., 2023). For stage-2, Temporal Convolution Neural Networks (TCNs); Temporal Transformers (T-TFs); and Recurrent Neural Networks (RNNs) often used, particularly LSTMs and Gated Recurrent Units (GRUs) (Garrow et al., 2020; Demir et al., 2023). For stage-3, Hidden Markov Models (HMMs) were typically used (Garrow et al., 2020; Demir et al., 2023; Twinanda et al., 2017), but other methods, such as Temporal Smoothing Functions (TSFs), are also common (Das et al., 2022).

Step recognition has previously been achieved in eTSA. In this initial work, a 3-stage architecture was used: a CNN + LSTM + TSF (Das et al., 2022). In stage-1, the CNN ResNet50 was used as a spatial encoder, as it was shown to be the best spatial feature extractor through ablation studies (Das et al., 2022). In stage-2, the spatial features of 10-consecutive-frames were fed into an LSTM for temporal encoding (Das et al., 2022). Finally, in stage-3, a threshold smoothing function was used (Das et al., 2022). The smoothing function ensured the step predictions were consistent for a certain period of time before switching to another step, to reduce the number of the frequent yet short periods of incorrect predictions (Das et al., 2022). The model was trained on 40-videos and validated on 10-videos, achieving a 0.74 weighted-F_1-score in 7-step frame-level classification (5-fold-cross-validation) (Das et al., 2022).

2.3. Instrument recognition

The majority of computer vision models created for minimally invasive surgeries regarding instruments is to accomplish instrument segmentation, rather than instrument recognition (Wang et al., 2022; Rueckert et al., 2024). Instrument segmentation is an extension of instrument recognition, where the type of instrument needs to not only be classified (instrument recognition) but the boundaries of the instrument also needs to be predicted. Due to this more difficult task, more sophisticated models, utilising an encoder–decoder architecture are used. However, similar to step recognition models, the most common encoders are CNNs for spatial feature extraction and RNNs for temporal feature extraction (Wang et al., 2022; Rueckert et al., 2024). No work has been published for instrument recognition for the eTSA.

[3] https://zenodo.org/records/8315050.

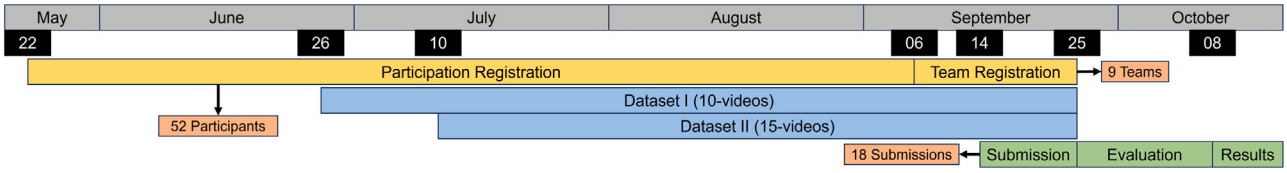


Fig. 2. A timeline of the challenge including dates of registration and dataset release. All dates are in 2023.

2.4. Multi-task recognition

Multi-task step and instrument recognition models connect single-task models at various stages in the neural network architecture (Twinanda et al., 2017; Psychogyios et al., 2024; Alabi et al., 2024). In doing so, they outperform single-task models in both tasks by sharing information (Das et al., 2023b; Mao et al., 2024). For example, in (Jin et al., 2020), a joint spatial–temporal (CNN + RNN) backbone is used for feature extraction in combination with a correlation loss function, so information gained from one task is shared with the other. However, multi-task models are not commonly used due to a lack of data (Wang et al., 2022; Alabi et al., 2024). No work has been published for multi-task step and instrument recognition for the eTSA.

3. Challenge description

The aim of the PitVis-2023 challenge was to develop Machine Learning (ML) models capable of step and instrument recognition in the eTSA. In doing so, these models provide surgical context that can be used as an assistive tool for clinicians.

3.1. Tasks

The challenge consisted of 3-tasks:

1. Surgical step recognition.
2. Surgical instrument recognition.
3. Multi-task steps and instrument recognition.

Representative images of the 12-steps and 19-instruments are displayed in Figs. 3 and 4 respectively. These steps and instruments are defined in (Marcus et al., 2021), and confirmed by two neurosurgical trainees (DZK and JGH) and one consultant neurosurgeon (HJM), based on the training dataset. For task-1; exactly one step is present at a given time, hence this is a multi-class problem. For task-2; zero, one, or two instruments may be present at a given time, hence this is a multi-label problem. Task-3 is a combination of task-1 and task-2, hence a multi-task problem.

3.2. Organisation

The PitVis-2023 challenge was a one-time event as part of EndoVis-2023 (Speidel et al., 2023), with all results presented publicly at the MICCAI-2023 conference in Vancouver, Canada. A timeline of the challenge organisation is displayed in Fig. 2. Organisation, communication, data sharing, and submissions were all done via the Synapse challenge website,[4] and no private communication with the organisers was permitted.

The organisation committee consisted of a collaboration between computer scientists and neurosurgeons from the UCL Hawkes Institute at University College London (UCL), London, United Kingdom (UK) and the Department of Neurosurgery at the National Hospital for Neurology and Neurosurgery (NHNN), London, UK respectively.

Advertisement was predominately done via social media.[5] 52-participants registered to download the data, with 9-teams across 6-countries successfully submitting 18-submissions. Prizes totalling £1000 per task were available to the top-2 teams of each task. Teams from the UCL Hawkes Institute were allowed to submit models, but illegible to win prizes.

25-annotated-videos were provided. A 20-training to 5-validation (01, 12, 21, 24, 25) split was suggested but not enforced. This split was based on step and instrument distributions (Section 4.2), such that the number of annotations for a class remained at an approximate 4:1 ratio, as is common in workflow recognition (Rueckert et al., 2024; Demir et al., 2023). The 8-testing-videos were not provided to the participants. The training and testing videos are similar to those of the intended use cases.

3.3. Model requirements

Only fully-automatic methods were permitted: the model must have taken an image input and output the predicted classification(s) as appropriate for the given task. For task-3, a multi-task model is defined as a single model that takes an image input and outputs both a predicted step classification and a predicted instrument classification congruently. Additionally, only online models were permitted: only information from frames up to and including the current frame can be used to classify the current frame. The inclusion of future frames' information or manual intervention to classify the current frame would generally improve model performance. However, these rules were implemented to mimic the real-time intra-operative use case.

Using instrument annotations for step recognition training, or using step annotations for instrument recognition training was permissible. Training on publicly available data was also permissible. Ideally, training would have been limited to just the dataset and annotations available for a specific task (i.e. step annotations only for step recognition training). This would have most likely resulted in weaker performing models, but provide better understanding of the methods. However, it would be impossible to monitor what data was used for training, and so these described measures were taken for practicality purposes.

Models were submitted as docker containers via Synapse on the challenge website, after detailed submission instructions were given. This included an example docker submission with the associated evaluation scripts, downloadable from GitHub.[6] The status of whether a submission was successfully submitted could also be found on the challenge website, but not the final evaluation scores. Participants were not required to publish their code, but were required to give detailed descriptions and diagrams of their model.

Finalised dockers were run on a single core of an NVIDIA-Tesla-V100-Tensor-Core-32-GB-GPU, and had to run in a reasonable time (less than 1 min of runtime for every 10 min of video). This runtime was implemented for practical purposes to allow for a quick evaluation, but also to mimic real-time performance such that recognition would occur at a reasonable frame rate (i.e. classification occurs on every other frame in the video).

[4] www.synapse.org/#!Synapse:syn51232283/wiki/621581.

[5] www.x.com/AdritoDas/status/1660677465956548609.

[6] www.github.com/dreets/pitvis/.

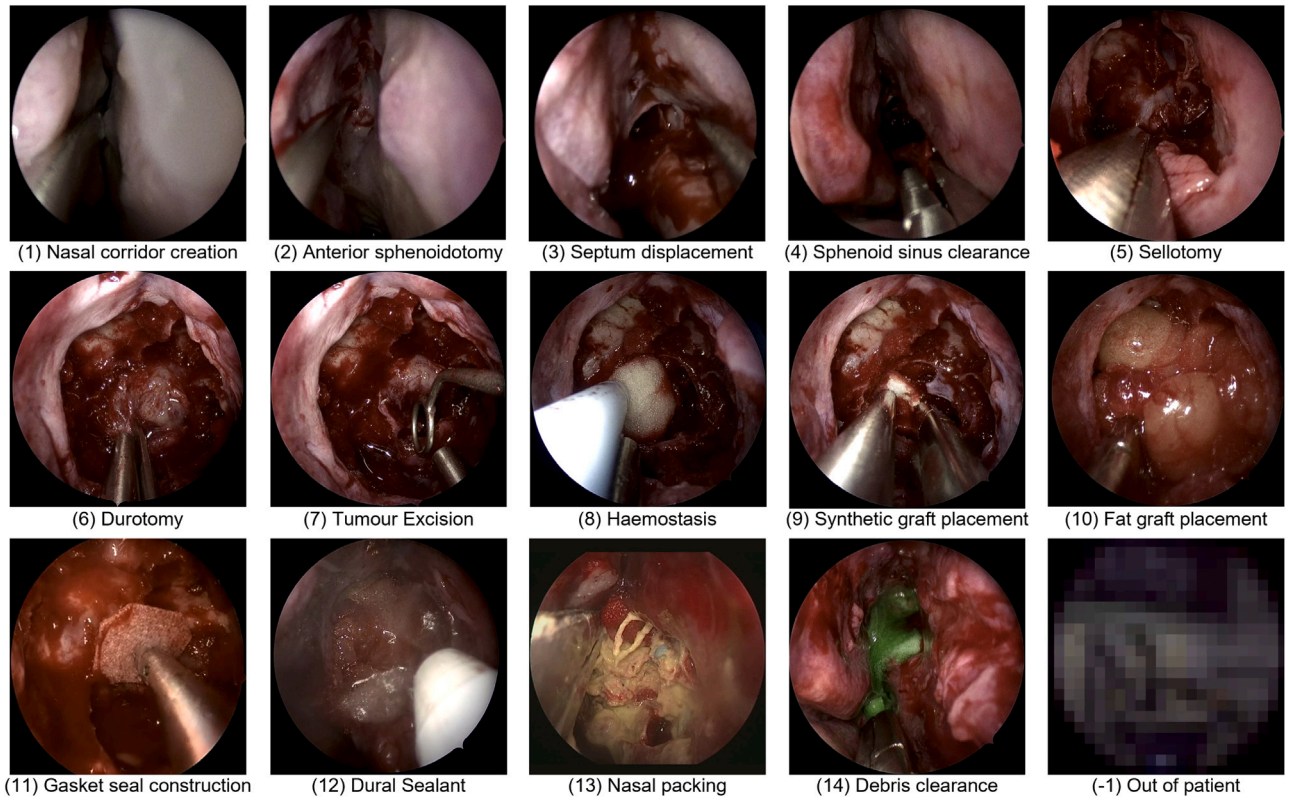


Fig. 3. Representative images of each of the 14-steps. Note step-11 and step-13 were not evaluated due to having insufficient volume to train on (Fig. 6), and 'out of patient' is not considered a class.

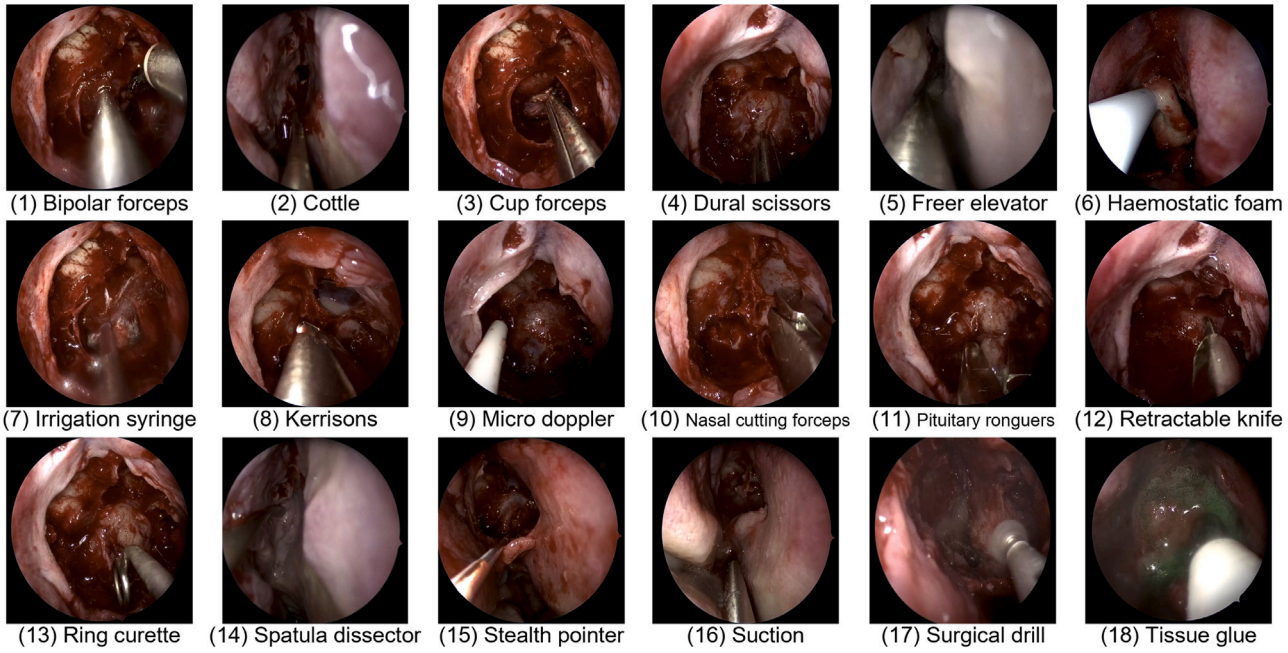


Fig. 4. Representative images of each of the 18-instruments, excluding the 'no instrument' class.

If any models were found to not satisfy these requirements they would have been disqualified. Fortunately, all models did satisfy the described requirements, resulting in 18 successful submissions.

3.4. Evaluation metrics

3.4.1. Spatial metric

Macro-F_1-score (Eq. (1)) was the chosen spatial metric. This is because F_1-score (Eq. (2)) ensures a high per-frame accuracy while also safeguarding against small precision or recall. Taking a macro-mean across classes ensures each class is treated equally so major classes do not dominate.

$$\text{Macro}\langle F_1 \rangle = \frac{1}{N} \sum_{i=1}^{N} (F_1)_i \ , \tag{1}$$

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \ , \tag{2}$$

where $\sum$ is the summation operator; N = total number of classes; TP = true positive; FP = false positive; FN = false negative. Note macro$\langle F_1 \rangle$ is used instead of macro-F_1-score in the above equations to prevent confusion with the subtraction (−) symbol.

3.4.2. Temporal metric

Edit-score (Eq. (3)) was chosen as the temporal metric (Lea et al., 2016). It is the reciprocal of the Levenshtein distance (Eq. (4)), which measures the number of edits (insertions, deletions, substitutions) required to change one temporal series into the other, and by doing so, penalises temporally inconsistent predictions (Lea et al., 2016). A series is defined as classifications without repeats. For example, classifications $[0, 0, 0, 1, 1, 0, 1, 1]$ are compressed to a $[0, 1, 0, 1]$ series.

$$\text{Edit} = \frac{1}{\text{Lev}} \ , \tag{3}$$

$$\text{Lev}(s,t) = \begin{cases} |s| & \text{if } |t| = 0, \\ |t| & \text{if } |s| = 0, \\ \text{Lev}\big(\text{tail}(s), \text{tail}(t)\big) & \text{if head}(s) \\ & = \text{head}(t), \\ 1 + \min \begin{cases} \text{Lev}\big(\text{tail}(s), t\big) \\ \text{Lev}\big(s, \text{tail}(t)\big) \\ \text{Lev}\big(\text{tail}(s), \text{tail}(t)\big) \end{cases} & \text{otherwise.} \end{cases} \tag{4}$$

where Lev is the Levenshtein distance; s, t are the ground-truth and predicted series of steps respectively; head(s) is the first value of a given series s; and tail(s) is all but the first value of a given series s. Note Edit is used instead of Edit-score in the above equations to prevent confusion with the subtraction (−) symbol.

3.4.3. Task specific metrics

The mean of Macro-F_1-score and Edit-score was chosen as the step recognition metric (Eq. (5)). This is so models are optimised for both frame-level accuracy and temporal consistency. Previous work has shown using purely spatial metrics leads to a high F_1-score but frequent inaccurate changes of steps for short periods of time (Das et al., 2022).

$$\text{Step\_metric} = \frac{\text{Macro}\langle F_1 \rangle + \text{Edit}}{2} \tag{5}$$

where Macro$\langle F_1 \rangle$ is defined in Eq. (2) with $N = 12$ for the 12 step classes, and Edit is defined in Eq. (3).

Macro-F_1-score was the chosen metric for instrument recognition with no Edit-score (Eq. (6)). This was because the usage of instruments is much more volatile and heavily dominated by the instrument-0 (no instrument) and instrument-16 (suction) class (Fig. 8). For example, a typical snippet of a ground-truth sequence is $[0, 11, 0, 0, 11, 16, 16, 11, 16]$, where an instrument such as instrument-11 (pituitary ronguers) will be briefly used between the dominating instrument-0 and instrument-16 classes. This means an incorrect prediction will be strongly penalised by temporal metrics. Moreover, as instrument recognition is a multi-label problem, a single sequence does not encapsulate all of the data, and so more sophisticated temporal metrics beyond Edit-score are required. After the results of this challenge, and the models are analysed, an appropriate temporal metric will be used for future work in an attempt to improve temporal consistency.

$$\text{Instrument\_metric} = \text{Macro}\langle F_1 \rangle \tag{6}$$

where Macro$\langle F_1 \rangle$ is defined in Eq. (2) with $N = 19$ for the 19 instrument classes.

The mean-average of the respective step and instrument recognition metric was chosen as the multi-task metric (Eq. (7)). This was done to treat both step and instrument recognition equally.

$$\text{Multitask\_metric} = \frac{\text{Step\_metric} + \text{Instrument\_metric}}{2} \tag{7}$$

where Step_metric and Instrument_metric are defined in Eqs. (5) and (6) respectively.

4. Dataset

The challenge dataset is the first publicly available annotated dataset of the eTSA. This section describes the dataset acquisition and analyses its properties.

4.1. Data acquisition

4.1.1. Videos

The NHNN (Queens Square, London, UK) provided all videos used in the PitVis challenge. This hospital is an academic tertiary neurosurgical centre, performing 150–200 pituitary operations each year (Khan et al., 2022). Videos of the eTSA were excluded if: the operation was a revision surgery within 6-months of the primary surgery; if large sections of the surgery were missing; or if the surgery was significantly different from a usual surgery. This curation was performed by two trainee neurosurgeons (DZK and JGH) and verified by a consultant neurosurgeon (HJM). The dataset size was determined by what was feasible to annotate in the challenge timeline.

The 25-training-videos were recorded between 02-Jul-2021 and 28-Dec-2022, and have written consent for public research use. The 8-testing-videos were recorded between 06-Dec-2018 to 07-Jan-2021, and have consent for research use within the organisers' institute (UCL). The study was registered with the UCL Institutional Review Board (IRB) (17819/011).

The surgeries were recorded using a high-definition endoscope (Hopkins Telescope with AIDA storage system, Karl Storz Endoscopy,[7] UK). The original videos have a variable Frames Per Second (FPS), with resolutions varying from 720p–2160p. These videos were uploaded from the hospital servers to the commercially available Touch Surgery™ Ecosystem,[8] an AI-powered surgical video management and analytics platform provided by Medtronic. Here, the videos were de-identified by blurring all images outside of the patient. The videos were then converted to a constant 24-FPS with 720p resolution using the publicly available Handbrake,[9] and stored as .mp4 files.

Additionally, a script to sample the videos at 1 FPS, and store them as .png images was provided on the GitHub. This sampling script was used by the organisers on the 8-testing-videos, and the .png images were fed into the submitted models for evaluation.

Table 1
An example of the .csv annotations given to participants. 'int_video' represents which video number the annotations are referring to. 'int_time' represents the frame number that row's annotations is representing. The numbers in the 'int_step', 'int_instrument1', and 'int_instrument2' columns represent the respective step and instrument class, as defined in Figs. 3 and 4. A '−1' in the 'int_step' and 'int_instrument1' column is indicative of 'out_of_frame', and removed for evaluation purposes. A '−2' in the 'int_instrument2' column is indicative of 'no annotation', and present as to not have an empty value. Note '...' indicates a break in the annotations for demonstration purposes.

int_video	int_time	int_step	int_instrument1	int_instrument2
25	0	−1	−1	−2
25	1	−1	−1	−2
...	...	...	...	...
25	2011	5	8	16
25	2012	5	16	−2
25	2013	5	16	−2
25	2014	5	0	−2

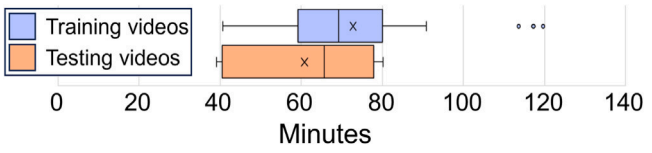


Fig. 5. Length distribution of the 25-training and 8-testing videos without the 'out of patient' class displayed as a box-whisker diagram. The middle horizontal line displays the median and the 'X' the mean. The left and right whiskers display the first (Q1) and fourth (Q4) quartile range respectively. The left and right boxes display the second (Q2) and third (Q3) quartile range respectively.

4.1.2. Annotations

For steps, each video was annotated by two trainee neurosurgeons (DZK and JGH) with any discrepancies solved via discussion and mutual agreement. For instruments, a third-party company Anolytics[10] was used. These annotations were not performed by clinical specialists, but verified by one neurosurgical trainee (DZK) and one research scientist (AD). All annotations were then verified by a consultant neurosurgeon (HJM) before being released.

Annotations were released as .csv files along with their associated videos, an example of which is displayed in Table 1. The map of the step or instrument to the corresponding integer was also provided.

As with all annotations, there can be errors, and in this challenge the most likely source is human error in misidentifying a step or instrument. These were reduced by the aforementioned multiple rounds of annotating and verification, and if any were found after release, they were immediately corrected and participants were informed.

4.2. Data analysis

4.2.1. Videos

The distribution of video lengths across all videos is displayed in Fig. 5. The mean and median of the 25-training-videos was 72.8 + 7.2 and 69.2 + 6.4 min respectively, where $+t$ indicates time, t, outside of the patient. This was slightly longer than the mean and median of the 8-testing-videos, which were 60.9 + 5.6 and 65.7 + 5.3 min respectively. The 'out of patient' frames, indicated by the '-1' class in annotation files were excluded during evaluation.

4.2.2. Steps

Step-11 (gasket seal construct) and step-13 (nasal packing) were only present in 2 and 1 training-videos respectively, and so were removed due to having insufficient volume to train on (Fig. 6), and any such frames were excluded during evaluation. A hypothetical step-0 (no step) class does not exist as every part of a video belongs to a step.

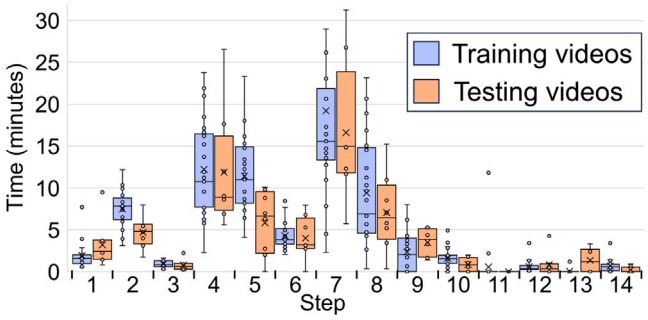


Fig. 6. Length distribution of steps across the 25-training and 8-testing videos displayed as a box-whisker diagram. The middle horizontal line displays the median and the 'X' the mean. The left and right whiskers display the first (Q1) and fourth (Q4) quartile range respectively. The left and right boxes display the second (Q2) and third (Q3) quartile range respectively. The '•' display each individual step length in the dataset.

Steps 1–8 are present in all 25-training-videos, with the remaining steps found in at least 18-training-videos. As displayed in Fig. 6, the length of steps are similar across the training and testing videos, but the step lengths themselves are varied. For example, step-7 (tumour excision) is the longest and step-14 (debris debulking) is the shortest with a with mean lengths of 19.2 and 0.7 min respectively. Moreover, as displayed in Fig. 7, the transition probabilities from one step to the next are not consistent. For example, step-8 (haemostasis) is often transitioned to and from out of sequence due to its short but frequent occurrences during surgery. This lack of consistency highlights the difficulty of step recognition in this dataset and the eTSA in general.

4.2.3. Instruments

A '-1' annotation indicates the 'out of patient' class and '-2' indicates a 'no secondary instrument' as to not have an empty entry in this column, and these frames were excluded during evaluation.

The majority of instruments are found in 20 or more training-videos. Exceptions to this are instrument-1 (bipolar forceps), found in 12-videos; and instrument-17 (surgical drill), found in 6-videos. As displayed in Fig. 8, the length distribution for instruments is dominated by instrument-0 (no instrument) and instrument-16 (suction) with mean lengths of 25.2 and 28.7 min respectively. The remaining instrument lengths are more clustered, although there is still some variance. There are also quite drastic differences between the training and testing dataset. For example, instrument-3 (cup forceps) and instrument-7 (irrigation syringe) have a relatively high usage in the training-videos, but very low usage in the testing-videos. This is likely due to time difference between when the training and testing surgeries were performed: leading to different availability of instruments, and variance in surgical technique. Similar to the steps, this highlights the difficulty of instrument recognition for the eTSA.

5. Methods

Table 2 displays a summary of the 9-teams from 6-countries, and the corresponding 18-submissions: 7 for Task-1; 6 for Task-2; and 5 for Task-3. All models use either a Spatial Encoder (S-E) (CNN; S-TF) or Spatio-Temporal Encoder (ST-E) (ST-TF), with the majority using a temporal encoder (LSTM; TCN; T-TF), and a few perform online post-processing (TSF). There are some which use multiple neural networks and combine them via an Ensemble. Detailed architectural diagrams of all models are displayed in Fig. 9.

Tables 3 and 4 display a summary of the training parameters and image augmentations. Although there are a few commonalities between

[7] www.karlstorz.com/.
[8] www.touchsurgery.com/.
[9] www.handbrake.fr/.
[10] www.anolytics.ai/.

To Step

From Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	End
Start	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1		100	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0		64	22	0	0	0	14	0	0	0	0	0	0	0
3	0	5		95	0	0	0	0	0	0	0	0	0	0	0
4	0	7	20		54	1	0	18	0	0	0	0	0	0	0
5	0	0	0	30		39	2	30	0	0	0	0	0	0	0
6	0	0	0	0	23		52	25	0	0	0	0	0	0	0
7	0	0	0	0	2	2		84	7	2	0	0	0	0	2
8	0	4	1	10	13	20	17		17	12	2	1	0	1	1
9	0	0	0	0	0	0	9	17		61	4	4	0	0	4
10	0	0	0	0	0	0	0	11	7		0	70	0	11	0
11	0	0	0	0	0	0	0	0	67	33		0	0	0	0
12	0	0	0	0	0	0	4	0	0	0	0		4	67	26
13	0	0	0	0	0	0	0	0	0	0	0	0		0	100
14	0	0	0	0	0	0	0	9	0	0	0	27	0		64

Fig. 7. Transition probabilities across the 25-training-videos. Each value represents the probability of going from one step to another (e.g. step-4 goes to step-5 with 54% probability). The 'out of patient' class was removed for these calculations. The greyed out 0 values are true zero and not rounded. The darker the blue colour the larger the transition probability, up to 100%.

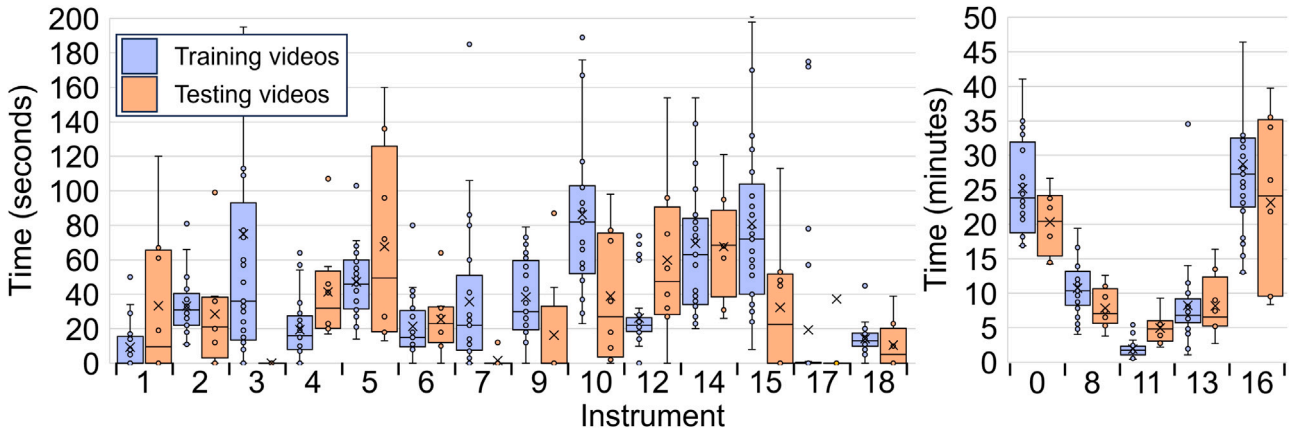


Fig. 8. Length distribution of instruments across the 25-training and 8-testing videos displayed as a box-whisker diagram. The time axis is presented as seconds in the left diagram and minutes in the right diagram — this is for improved visibility, as otherwise the minor class instrument length distributions would not be visible. The middle horizontal line displays the median and the 'X' the mean. The left and right whiskers display the first (Q1) and fourth (Q4) quartile range respectively. The left and right boxes display the second (Q2) and third (Q3) quartile range respectively. The '•' display each individual instrument length in the dataset.

the methods (Cross-Entropy Loss Function (CE) loss function; resizing input images), there are vast differences. The majority do not implement strong image augmentations; or any data balancing, whereas a majority do use the suggested validation split; pre-train on ImageNet; or use Adaptive Moment Estimation (Adam) for backpropagation. The remaining parameters are even: some use Rectified Linear Unit (ReLU); some remove the black borders of an image; and some use the task evaluation metric.

The remainder of this section provides an overview of each model (in alphabetical order by team name):

5.1. CAIR-POLYU-HK

CAIR-POLYU-HK consisted of You Pang; Zhen Chen; Xiaobo Qiu; and Zhen Sun, from the Hong Kong Institute of Science and Innovation, China. This team submitted only to task-1.

Their model consisted of 2-stages: CNN + TCN. The CNN was CSPDarknet53 (Bochkovskiy et al., 2020). The TCN was a 2-layered 10-window TeCNO (Czempiel et al., 2020).

Interestingly, CAIR-POLYU-HK had the largest batch size of 200, utilising an 80-GB NVIDIA-A100.

5.2. CITI

CITI consisted of Xiaoyang Zou; and Guoyan Zheng, from Shanghai Jiao Tong University, China. This team submitted to all 3-tasks: task-1 and task-3 used the same model architecture, and task-2 a simplified version.

Task-2's model consisted of 1-stage: a ST-E. This ST-E comprised of a ST-TF (Swin (Liu et al., 2021)) followed by a 2-layer Multi-Head Self-Attention (MHSA) (Zou et al., 2024). The ST-E took a 20-window sequential video frame input, outputting both step (just for training) and instrument (task-2&3) classifications.

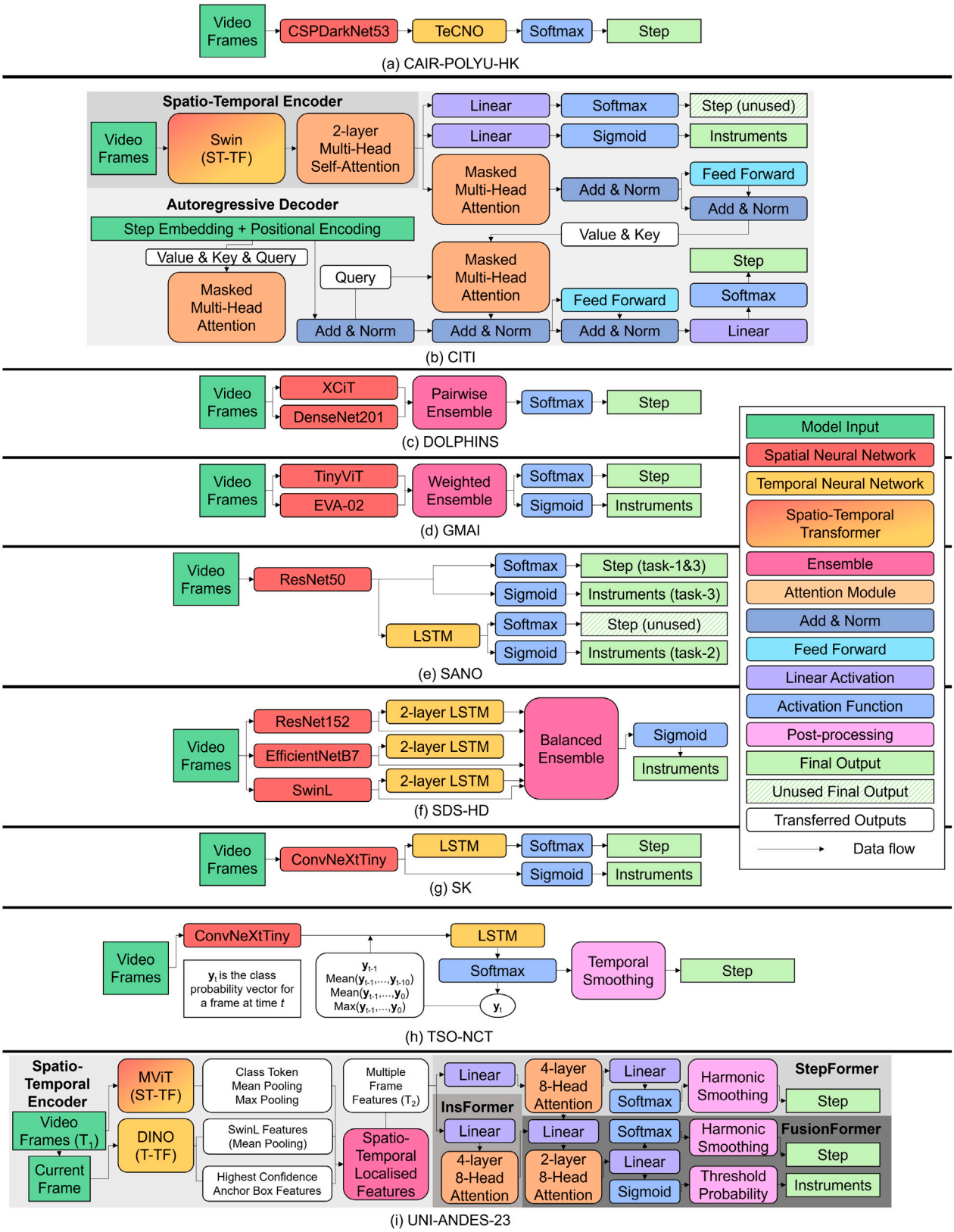


Fig. 9. Detailed model architectures for the 9-models. Each model architecture is given as a series of model modules, with the legend on the right identifying the colour used to represent each module, and the arrows' direction illustrating the flow of features extracted from the data (generally left to right). The various grey background for models (b) and (i) are there to visually indicate the different transformer components (e.g. the spatio-temporal encoder component). Simplified model architectures are displayed in Table 2.

Table 2
Team details (9-teams) and simplified model architectures for the successful 18-submissions. For the model columns, each row represents a different training component, and if a horizontal line is removed at a later stage it means the model features have been combined (e.g. in an Ensemble). () are given to indicate the type of model used for that stage. {} are given to indicate the window size of a temporal neural network (e.g. {24} represents 24-images have been turned into a sequence as an input). ⟨⟩ are given to indicate the task (step or instrument) for multi-task recognition if the same architecture is not used for both tasks. Citations: ARST (Zou et al., 2022); CSPDarknet53 (Bochkovskiy et al., 2020); ConvNeXtTiny (Liu et al., 2022); DenseNet201 (Huang et al., 2016), DINO (Zhang et al., 2022); EfficientNetB7 (Tan and Le, 2019); EVA-02 (Fang et al., 2024); MViT (Fan et al., 2021); ResNet152, ResNet50 (He et al., 2016); Swin, SwinL (Liu et al., 2021); TeCNO (Czempiel et al., 2020), TinyViT (Wu et al., 2022), Threshold Smoothing (Das et al., 2022), XCiT (El-Nouby et al., 2021). Detailed model architectures are displayed in Fig. 9.

Team	Institute	Task	Simplified model architecture		
			Stage-1	Stage-2	Stage-3
CAIR-POLYU-HK	Hong Kong Institute of Science and Innovation Hong Kong, China	1	CSPDarknet53(CNN)	TeCNO{10}(TCN)	–
CITI	Shanghai Jiao Tong University Shanghai, China	1,3	Swin{20}(ST-TF)	ARST{80}(ST-TF)⟨step⟩	–
		2		–	–
DOLPHINS	Imperial College London London, UK	1	XCiT(S-TF)	Pairwise ensemble	–
			DenseNet201(CNN)		
GMAI	Shanghai AI Lab Shanghai, China	1,2,3	TinyViT(S-TF)	Weighted ensemble	–
			EVA-02(S-TF)		
SANO	Sano Center for Computational Medicine Krakow, Poland	1,3	ResNet50(CNN)	–	–
		2		LSTM{5}	–
SDS-HD	German Cancer Research Center Heidelberg, Germany	2	ResNet152(CNN)	LSTM{15}	Balanced ensemble
			EfficientNetB7(CNN)	LSTM{15}	
			SwinL{1}(S-TF)	LSTM{12}	
SK	Muroran Institute of Technology Hokkaido, Japan	2	ConvNeXtTiny(CNN)	–	–
		3		LSTM{128}⟨step⟩	–
TSO-NCT	National Center for Tumor Diseases Dresden, Germany	1	ConvNeXtTiny(CNN)	LSTM{512}	Threshold smoothing(TSF)
UNI-ANDES-23	Universidad de los Andes Bogota, Colombia	1	MViT{24}(ST-TF)	StepFormer{24 × 8}(ST-TF)	Harmonic smoothing(TSF)
			DINO{24}(S-TF)		
		2,3	MViT{24}(ST-TF)	FusionFormer {24 × 10 × 2}(ST-TF)	Harmonic smoothing(TSF)⟨step⟩ Threshold probability⟨instrument⟩
			DINO{24}(S-TF)		

Task-1 and task-3's model consisted of 2-stages: ST-E + an autoregressive decoder. The ST-E is identical to the one described above. The autoregressive decoder was ARST (Zou et al., 2022). ARST took an 80-window input comprising of frame-wise visual features extracted by ST-E and shifted step outputs, outputting step classifications. It comprised of an initial Masked Multi-Head Attention (MMHA), followed by a mutual MMHA taking the Value and Key output of the ST-E after it has passed through a MMHA and the Query output from the initial MMHA (also passed to the normalisation layer). Positional encoding is added to embed the frame position for each step (as defined in Wu et al., 2022).

Table 3 CITI task-2&3 and task-1&3 represent the ST-E and ARST training parameters respectively.

5.3. DOLPHINS

DOLPHINS consisted of Abdul Qayyum; Moona Mazher; Imran Razzak; and Steven Niederer, from Imperial College London, United Kingdom. This team submitted only to task-1.

Their model consisted of 2-stages: a S-TF and CNN fused via pairwise ensemble. The S-TF was XCiT (El-Nouby et al., 2021). The CNN was DenseNet201 (Huang et al., 2016).

5.4. GMAI

GMAI consisted of Tianbin Li; Jin Ye; Junjun He; Yanzhou Su; Pengcheng Chen; and Junlong Cheng, from the Shanghai Artificial Intelligence Lab, China. This team submitted to all 3-tasks with the same model architecture.

Their model consisted of 2-stages: two S-TFs fused via weighted ensemble. The first S-TF is TinyViT (Wu et al., 2022), which utilities fast knowledge distillation. The second S-TF is EVA-02 (Fang et al., 2024), which utilities masked image modelling.

5.5. SANO

SANO consisted of Szymon Płotka; and Joanna Kaleta, from the Sano Center for Computational Medicine, Poland. This team submitted to all 3-tasks: task-1 and task-3 used the same model architecture, and task-2 adds to this.

For tasks-1 and task-3 their model consisted of 1-stage: a CNN. This CNN was ResNet50 (He et al., 2016), and used for both step (task-1 and task-3) and instrument (task-3) classification.

For task-2 their model consisted of 2-stages: CNN + LSTM. The CNN was the trained ResNet50 described above, with frozen weights. The LSTM had a 5-widow size, and used for both instrument (task-2) and step (just for training) classification. The details in Table 3 SANO task-2 represent the LSTM training parameters.

5.6. SDS-HD

SDS-HD consisted of Amine Yamlahi; Antoine Jund; Finn-Henri Smidt; Patrick Godau; and Lena Maier-Hein, from the German Cancer Research Center, Germany. This team only submitted to task-2.

Their model consisted of 3-stages: 3×{S-Es + LSTMs} fused together via balanced ensemble. The 3 S-Es are the CNN ResNet152 (He et al., 2016), the CNN EfficientNetB7 (Tan and Le, 2019), and the S-TF SwinL

Table 3
Training parameters and augmentations utilised by the models excluding UNI-ANDES-23. '/' implies implementation details for steps or instruments (e.g. CE/BCE means CE used for steps and BCE used for instruments). '|' implies implementation details from stage-1 to stage-2 (e.g. GeLU|Sigmoid means GeLU used for stage-1 and Sigmoid used for stage-2). Abbreviations: Adam (Adaptive Moment Estimation), AdamW (Adam with weight decay), BCE (Binary Cross-Entropy Loss Function), CE (Cross-Entropy Loss Function), ETE (End To End Temporal Training), GeLU (Gaussian error Linear Unit), HSV (Hue Saturation Value), mAP (mean Average Precision), RBG (Red Blue Green), ReLU (Rectified Linear Unit), SGD (Stochastic Gradient Descent), TS (Temporal Smoothing Loss Function), Sep (Separate Temporal Training), Val (Validation Dataset).

Team	CAIR-POLYU-HK	CITI		DOLPHINS	GMAI			SANO		SDS-HD	SK		TSO-NCT
Task	1	1&3	2&3	1	1	2	3	1&3	2	2	2	3	1
Loss	CE	CE		CE	CE			CE/BCE	CE/BCE	BCE	CE		CE|TS
Activation	ReLU	ReLU		ReLU	ReLU			ReLU	Softmax	ReLU	GeLU		GeLU|Sigmoid
Final activation	Softmax	Softmax	Sigmoid	Softmax	Softmax			Softmax	Sigmoid	Sigmoid	Sigmoid	Softmax	Softmax
Pre-trained	ImageNet	–		ImageNet	ImageNet			ImageNet		ImageNet	ImageNet		ImageNet
Multitask training	–	Yes		–	Yes			Yes		–	–	Yes	–
Temporal training	ETE	Sep	ETE	–	–			–	Sep	Sep	–	Sep	ETE
Removed borders	Yes	Yes		–	–			Yes		Yes	–		–
Augmentation probability	1.0	1.0		1.0	1.0			1.0		0.5	1.0		0.5
Resizing (pixels)	256 × 448	192 × 192		224 × 224	224 × 224			224 × 224		384 × 384	224 × 224		216 × 384
Rotation (degrees)	–	–		–	–			–		±45	±5		±15
Reflection	–	Horizontal		Horizontal	Horizontal			–		Horizontal&Vertical	–		–
Trans ation (x&y)	–	–		–	–			–		–	±5%		±5%
Scaling	–	–		–	–			–		±10%	±5%		±5%
Colour	–	ImageNet Normal-isation		ImageNet Normal-isation	–			ImageNet Normal-isation		Colour jitter Contrast equalisation	Blur HSV augmentations		RBG±15 Contrast ±0.2
Data balancing	–	–		–	–			–		Instrument upsampling	–		–
Validation	Suggested	Suggested		Suggested	–			Suggested		5-fold	12,15,17,20,22		Suggested
Training shuffling	Yes	Yes		Yes	Yes			Yes		Yes	Yes		No
Val shuffling	No	No		No	No			Yes		No	No		No
Trained epochs	30	10	8	50	20			40		10	50		200
Evaluation metric	Task	Task		Task	–			F_1-score	Task	F_1-score+mAP	Minimal loss		F_1-score
Best model choice	Val	Val		Val	Last epoch			Val		Val	Val		Val
Batch size	200	Video	4	25	16			128		64	128	32	512
Training hours	40	4	24	12	10			2		88	3	64	48
Backpropogation	SGD	Adam		Adam	AdamW			SGD		Adam	Adam		AdamW
Learning rate	1E−3	1E−4		1E−3	1E−3			1E−3	5E−3	2E−4 (>2E−5)	1E−4	1E−5	5E−4
Momentum	9E−2	–		–	–			9E−2		–	–		–
Decay	–	1E−3		–	–			–		1E−6	–		1E−2
GPU (NVIDIA)	A100	TITAN RTX		RTX A6000	V100			A100		V100	RTX4090		RTX A5000
GPU (GB)	80	24		48	32			2 × 80		32	24		24

(Liu et al., 2021). The 3 LSTMs are identical: 2-layers with 0.2-dropout (15-window, 15-window, 12-window). The balance ensemble has 6 prediction inputs: the 3 S-Es and the 3 LSTMs.

SDS-HD used a variety of alternative training techniques when compared to the other participants. Firstly, they balanced the data: 5-instrument classes (07; 10; 11; 12; 15) were upsampled and the remaining classes were downsampled. Secondly, they introduced both horizontal and vertical reflections, along with colour augmentations: colour jitter by modifying hue; saturation; and brightness, in addition to Contrast Limited Adaptive Histogram Equalization (CLAHE) augmentation. Thirdly, they utilised mean Average Precision (mAP) as an alternative evaluation metric along with the task specific macro-F_1-score. Finally, Adam backpropagation was enhanced via cosine annealing with a learning rate of 2E−4, with a minimum of 2E−5 and a 1E−6 decay rate.

5.7. SK

SK consisted of Satoshi Kondo; Satoshi Kasai; and Kousuke Hirasawa, from Muroran Institute of Technology, Niigata University of Health and Welfare, and Konica Minolta, Inc., Japan, respectively. This team submitted to task-2 and task-3, each with different model architectures.

For task-2, their model consisted of 1-stage: the CNN ConvNeXtTiny (Liu et al., 2022) for instrument classification. For task-3, their model consisted of 2-stages: CNN + LSTM. The trained CNN from task-2 was frozen for instrument classification and a 128-window LSTM was added for step classification. The details in Table 3 SK task-3 represent the LSTM training parameters.

5.8. TSO-NCT

TSO-NCT consisted of Dominik Rivoir; and Stefanie Speidel, from the National Center for Tumor Diseases, Germany. This team only submitted to task-1.

Their model consisted of 3-stages: CNN + LSTM + TSF. The CNN was ConvNeXtTiny (Liu et al., 2022). The LSTM had a 512-window size. The TSF was Threshold Smoothing (Das et al., 2022) with a 7-window size.

Inspired by Sufficient Statistics Model (SSM) (Ban et al., 2021), to propagate temporal features, for each frame, the softmax class scores of: the previous frame; the mean of the previous 10-frames, the mean and maximum of all previous frames, were fed into the LSTM in addition to the CNN spatial features. Per video, all temporal features (softmax scores and LSTM hidden state) are propagated across the unshuffled batches.

Threshold smoothing ensures a class transition only takes place after it has been predicted for a sufficient number of frames (in this case 7), otherwise it is left unchanged. In doing so, prediction consistency is improved in aims to increase Edit-score. Any steps not considered for evaluation (i.e. steps -1; 11; 13) were replaced with the most recent permitted step.

5.9. UNI-ANDES-23

UNI-ANDES-23 consisted of Alejandra Pérez; Santiago Rodriguez; Pablo Arbeláez; Nicolás Ayobi; and Nicolás Aparicio from Universidad de los Andes, Colombia. This team submitted to all 3-tasks: task-1 had its own model architecture; and task-2 and task-3 had a modified version of this architecture.

For all 3-tasks, their model consisted of 3-stages: a ST-E; a Spatio-Temporal Decoder (ST-D); and Harmonic Smoothing or Threshold Probability for step or instrument classification respectively.

In stage-1 for all 3-tasks, the ST-E is composed of two concatenated transformers. The first is the ST-TF MViT (Fan et al., 2021) with a 24-window size (6-s × 4-FPS), concatenating the class token; mean pooled features; and max pooled features. The second is the S-TF DINO (Zhang et al., 2022) acting on the final frame using SwinL (Liu et al., 2021), concatenating global max pooled features; and localised instrument features via anchor boxes.

For task-1, the ST-D (StepFormer) consists of an 8-window 4-layer 8-head attention transformer. For task-2, the ST-D (FusionFormer) consists of an identical transformer (InsFormer) combined with StepFormer (frozen weights) via a 2-layer 8-head attention transformer. For task-3, both StepFormer and InsFormer have frozen weights.

Harmonic Smoothing is an online post-processing TSF defined as follows: given the class probability vector of the current ($\mathbf{y}_t$) and previous frame ($\mathbf{y}_{t-1}$), if $\max\{\mathbf{y}_t\} < \max\{\mathbf{y}_{t-1}\}$, then $\hat{\mathbf{y}}_t = 2\left(\mathbf{y}_t^{-1} + \mathbf{y}_{t-1}^{-1}\right)^{-1}$ where $\hat{\mathbf{y}}_t$ is the updated class probability vector. This function is repeated for 750-iterations for improved temporal consistency, before the usual argmax function is applied for a final classification. Any steps not considered for evaluation were removed at this stage.

Threshold Probability is an online post-processing function defined as follows: if the second highest value in the class probability vector is less than 0.4, then only predict the first highest value's corresponding class; if at least two of the highest values in this vector are greater than or equal to 0.4 and this includes the value corresponding to the background class, then predict the two highest values' classes excluding the background class; in all other cases predict the two highest values' corresponding classes.

6. Results & discussion

6.1. Ranking method

Each video is considered one case of equal value, hence the rankings are determined by the tasks' evaluation metric mean-averaged across the 8-testing-videos (no missing results).

A Kruskal–Wallis test is run on the top-3 methods of each task (across the 8-testing-videos) to test the significance of the performances, with $p < 0.05$ implying significance.

6.2. Task-1

Results for the 7-submissions to 12-steps multi-class online recognition are displayed in Table 5, with £700 and £300 awarded to 1st and 2nd places respectively.

There is a strong performance, with the best models achieving 63% (CITI) and 54% (TSO-NCT) on the task metric. Macro-F_1-score is high, with the top 3-models achieving >50%, although there is a slow decline with the bottom 2-models achieving <7%. There is large variance in Edit-score, with the top 3-models achieving >46%, and the remaining <2%.

Although the best models use different architectures, a commonality between them is the use of propagating temporal features. For CITI and UNI-ANDES-23 via positional encoding, and for TSO-NCT via feeding classification vectors of previous frames back into the LSTM hidden state. It is clear models with temporal encoders and TSFs outperform those that are purely spatial, both in frame-level classification and significantly in temporal consistency.

For the top models Standard Deviation (std) is ≈10%, as can be more clearly seen in Fig. 10(d). Although there is some variance between videos, the performance is generally similar. In videos 26; 29; 33 CITI significantly outperforms the other models, whereas TSO-NCT outperforms CITI in videos 28; 31; 32. A Kruskal–Wallis test for comparing the step metric performance of CITI against TSO-NCT and UNI-ANDES-23 gives $p = 0.0587$ and 0.0233 respectively. For just macro-F_1-score $p = 0.9097$ and 0.0494, and for just Edit-score $p = 0.0356$ and 0.00325. This shows CITI is statistically significantly better than UNI-ANDES-23 in all aspects, but only temporally better than TSO-NCT. The differences between the several models' performances across the testing videos highlights the difficulty of creating a generalised model.

Figs. 10(a) and 10(b) displays the step confusion matrix for CITI and TSO-NCT respectively. Steps are often predicted as a neighbouring step, which is expected (Fig. 7). Step-8 (haemostasis) is special as it is used sporadically for short periods during a surgery, and therefore other steps are often predicted as it. The biggest difference between the models is overpredicting the dominant class step-7 (tumour excision) in TSO-NCT. Across both models there is poor performance for steps 3; 6; 9, suggesting these are inherently difficult steps to classify.

6.3. Task-2

Results for the 6-submissions to 19-instruments multi-label online recognition are displayed in Table 6, with £500 awarded to joint 1st (1st & 2nd).

There is a good performance, with the best models (SDS-HD and SANO) both achieving 42% on the task metric. The next top 2-models are not far behind, achieving >34% with the remaining bottom 2-models also not far behind, achieving >27%.

The top two models use the well-known architecture of CNN + LSTM (+ Ensemble for SDS-HD). They are able to outperform purely spatial models (SK and GMAI) as well as more sophisticated models that utilise temporal encoders; positional encoding; and multi-task training (CITI and UNI-ANDES-23).

There is varied std in the top models as displayed in Fig. 12. SDS-HD outperforms the other models in the majority of videos. However, it is outperformed significantly by SANO in video-31 and by CITI in video-27. A Kruskal–Wallis test for comparing the macro-F_1-score of SDS-HD against SANO and CITI gives $p = 0.8336$ and 0.5286 respectively. This shows the differences in the performances of these models are not statistically significant, and with more testing videos the rankings could change. As with step recognition, differences between the models' performance across the videos show the difficulty of a creating a generalised model.

Figs. 11(a) and 11(b) displays the instrument confusion matrix for SDS-HD and SANO respectively. Instruments are frequently misclassified as instrument-0 (no instrument) and instrument-16 (suction). This is to be expected as they are the dominant classes, suggesting one way to overcome these incorrect predictions is through data balancing. Across both models, instruments 4; 12; 13 are predicted poorly with 2; 6; 10 also poorly predicted by SANO. This disparity is likely due to the number of instrument classes and the visual similarity between them, as well as insufficient training data. Interestingly, instruments 16 and 17, the only two secondary instruments in the testing dataset, are predicted well as secondary instruments.

6.4. Task-3

Results for the 5-submissions to 12-steps and 19-instruments multi-task online recognition are displayed in Table 7, with £700 and £300 awarded to 1st and 2nd places respectively.

The performance is good, with the best models achieving 49% (CITI) and 41% (UNI-ANDES-23) on the task metric. The next top 2-models drop performance with <30%, and the worst model only achieves 16%. The std is <10% across all models.

CITI's model is identical to its previous task models, which already utilised multi-task learning: the strong step recognition (1st) compensates for the poorer instrument recognition (3rd). On the other hand, UNI-ANDES-23's model improves in both step (+0.4%) and instrument

Table 4
Training parameters and augmentations utilised by UNI-ANDES-23. Abbreviations: Adam (Adaptive Moment Estimation), AdamW (Adam with weight decay), BCE (Binary Cross-Entropy Loss Function), CE (Cross-Entropy Loss Function), GeLU (Gaussian error Linear Unit), Lion (Lightweight Interpolated Optimiser), ReLU (Rectified Linear Unit), SGD (Stochastic Gradient Descent), Val (Validation Dataset).

Network	MViT	DINO	StepFormer	InsFormer	FusionFormer
Loss	CE	CE	CE	BCE	BCE
Activation	ReLU	ReLU	GeLU	GeLU	GeLU
Final activation	–	–	Softmax	Sigmoid	Softmax/Sigmoid
Pre-trained	Kinetics400 + PSI-AVA	COCO	–	–	–
Temporal training	Yes				
Multitask training	Yes				
Removed borders	Yes		–		
Augmentation probability	1.0	1.0	–		
Resizing (pixels)	224 × 224	894 × 800	805 × 720		
Rotation (degrees)	–				
Reflection	–				
Translation (x&y)	–	Yes	–		
Scaling	–	Yes	–		
Colour	Jitter (0.4)	–	–		
Data balancing	Weighted sampling	Weights inverse of sample size		Weighted loss 2 × (step1,step14)	
Validation					
Training shuffling	No				
Val shuffling	No				
Trained epochs	16	12	50		
Evaluation metric	Task				
Best model choice	Val				
Batch size	12	4	3000		
Training hours	64	12	8		
Backpropogation	SGD	AdamW	Adam	Lion	Adam
Learning rate	1.25E−2	1E−4	1E−4	1E−5 (Adam 1E−4)	1E−4
Momentum	0.9	–	–	–	–
Decay	–	1E−4	–	1E−2	–
GPU (NVIDIA)	Quadro RTX8000				
GPU (GB)	48 GB				

Table 5
12-steps multi-class online recognition (task-1) rankings. Metrics are calculated across the 8-testing-videos, and given as percentages to one decimal place (mean ± std).

	Team	(Macro⟨F_1⟩ + Edit)/2	Macro⟨F_1⟩	Edit
1	CITI	62.9 ± 09.7	61.1 ± 10.6	64.7 ± 10.1
2	TSO-NCT	53.7 ± 11.2	58.2 ± 10.9	49.2 ± 13.0
3	UNI-ANDES-23	48.3 ± 07.3	50.1 ± 09.3	46.5 ± 08.2
4	SANO	20.5 ± 03.2	39.6 ± 06.5	01.4 ± 00.4
5	DOLPHINS	15.2 ± 04.0	28.9 ± 08.2	01.6 ± 00.7
6	GMAI	03.7 ± 00.2	06.8 ± 00.3	00.5 ± 00.1
7	CAIR-POLYU-HK	03.5 ± 00.8	05.8 ± 01.5	01.1 ± 00.3

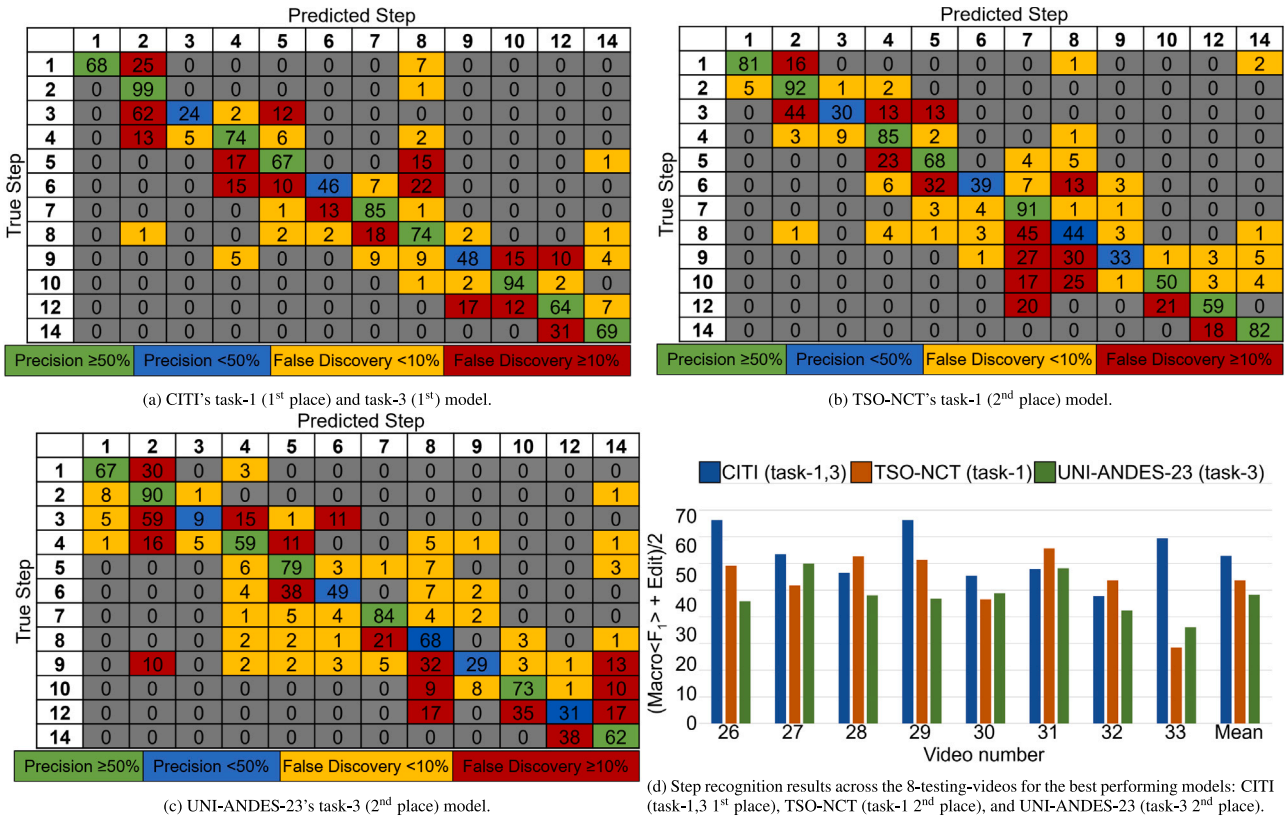
Table 6
19-instruments multi-label online recognition (task-2) rankings. Metrics are calculated across the 8-testing-videos, and given as percentages to one decimal place (mean ± std).

	Team	Macro⟨F_1⟩
1	SDS-HD	41.7 ± 15.4
2	SANO	41.6 ± 06.3
3	CITI	35.1 ± 18.5
4	SK	34.0 ± 17.0
5	GMAI	27.8 ± 08.7
6	UNI-ANDES-23	27.5 ± 13.5

(+4.9%) recognition due to the multi-task learning from the Fusion-Transformer. SK's instrument recognition model (4th) now incorporates step recognition via an LSTM achieving 25% on task-1's metric, which would have given them 4th place had they entered. SANO's model has decreased performance in both step (−1.4%) and instrument (−4.5%) recognition, this is due to their task-3 model not utilising the LSTM trained for instrument recognition in task-2. GMAI's model performs similarly poorly in both step (−0.2%) and instrument (−0.6%) recognition. It is likely a multi-task form of TSO-NCT's model, which came 2nd in task-1, would have performed well, given its similarity to the best models for instrument recognition. However, it is unlikely a multi-task form of DOLPHIN's and CAIR-POLYU-HK's task-1 models would have performed well given their poor performance in task-1.

The comparison of UNI-ANDES-23 task-3 model for each testing video is found in Figs. 10(d) (steps) and 12 (instruments). For steps, is able to outperform TSO-NCT in videos 27; 30; 33, but is always outperformed by CITI. For instruments, it performs similarly to the other models, significantly outperforming CITI in video 26, although it is never the best performing model. A Kruskal–Wallis test for comparing the multi-task metric of CITI against UNI-ANDES-23 and SK gives $p = 0.8335$ and 0.0087 respectively. This shows that although CITI is statistically significantly better than SK, the same is not true for UNI-ANDES-23. Breaking this down further, $p = 0.0742, 0.0063, 0.9164$ for step-macro-F_1-score, step-Edit-score, and instrument-macro-F_1-score respectively. As there is only statistical significance in the temporal metric, it can be inferred that CITI's main contribution is in temporal accuracy of step recognition.

This temporal accuracy can be more clearly seen in Fig. 13. This figure displays the step predictions of the top performing models against the ground-truth. From here the importance of temporal consistency is



(a) CITI's task-1 (1st place) and task-3 (1st) model.

(b) TSO-NCT's task-1 (2nd place) model.

(c) UNI-ANDES-23's task-3 (2nd place) model.

(d) Step recognition results across the 8-testing-videos for the best performing models: CITI (task-1,3 1st place), TSO-NCT (task-1 2nd place), and UNI-ANDES-23 (task-3 2nd place).

Fig. 10. In-depth details of the top models in step recognition across the 8-testing-videos: (a–c) Confusion matrices, mean-averaged and rounded to 0 decimal places. Each cell (i, j) represents the number of instances from true class i that were predicted as class j. The diagonal entries indicate correct classifications, while off-diagonal entries show misclassifications. (d) Per-video performance on the step recognition metric (Eq. (5)).

Table 7
12-steps and 19-instruments multi-task online recognition (task-3) rankings. Metrics are calculated across the 8-testing-videos, and given as percentages to one decimal place (mean ± std). The full definition of the multi-task metric is displayed in Eq. (7).

	Team	(Step_metric + Instrument_metric)/2	Step Macro⟨F_1⟩	Step Edit	Instrument Macro⟨F_1⟩
1	CITI	49.0 ± 09.4	61.1 ± 10.6	64.7 ± 10.1	35.1 ± 18.5
2	UNI-ANDES-23	40.5 ± 07.7	51.0 ± 08.8	46.3 ± 10.4	32.4 ± 11.7
3	SK	29.6 ± 09.1	41.2 ± 05.9	09.1 ± 02.0	34.0 ± 17.1
4	SANO	28.3 ± 06.4	39.6 ± 06.5	01.4 ± 00.4	36.2 ± 14.8
5	GMAI	15.5 ± 03.6	07.2 ± 00.7	00.5 ± 00.1	27.2 ± 06.9

more readily seen. For example, SANO's task-3 model has no temporal encoder, meaning step predictions are much more volatile, often changing from one frame to the next. Continuing, UNI-ANDES-23 implements harmonic smoothing for smoother predictions and TSO-NCT takes this further with threshold smoothing for even smoother predictions: These TSFs were originally created for the purpose of reducing this prediction volatility. CITI has stronger predictions overall, able to distinguish between steps 9, 10, and 12 where the other models are unable to do so.

Figs. 11(c) and 11(d) displays the instrument confusion matrix for CITI (1st) and UNI-ANDES-23 (2nd) respectively. When this is compared with the previously displayed confusion matrices, almost identical inferences can be made. One major difference is CITI overpredicts instrument-0 (no instrument) far less than other models, although it does overpredict instrument-0* (no secondary instrument) much more, reducing the precision of instrument-16 (suction). Similarly, Fig. 10(c) displays the step confusion matrix for the UNI-ANDES-23. This is again similar to the previous matrices. Two minor differences are a poorer step-12 performance and a greater overprediction of step-14.

6.5. Benchmarks

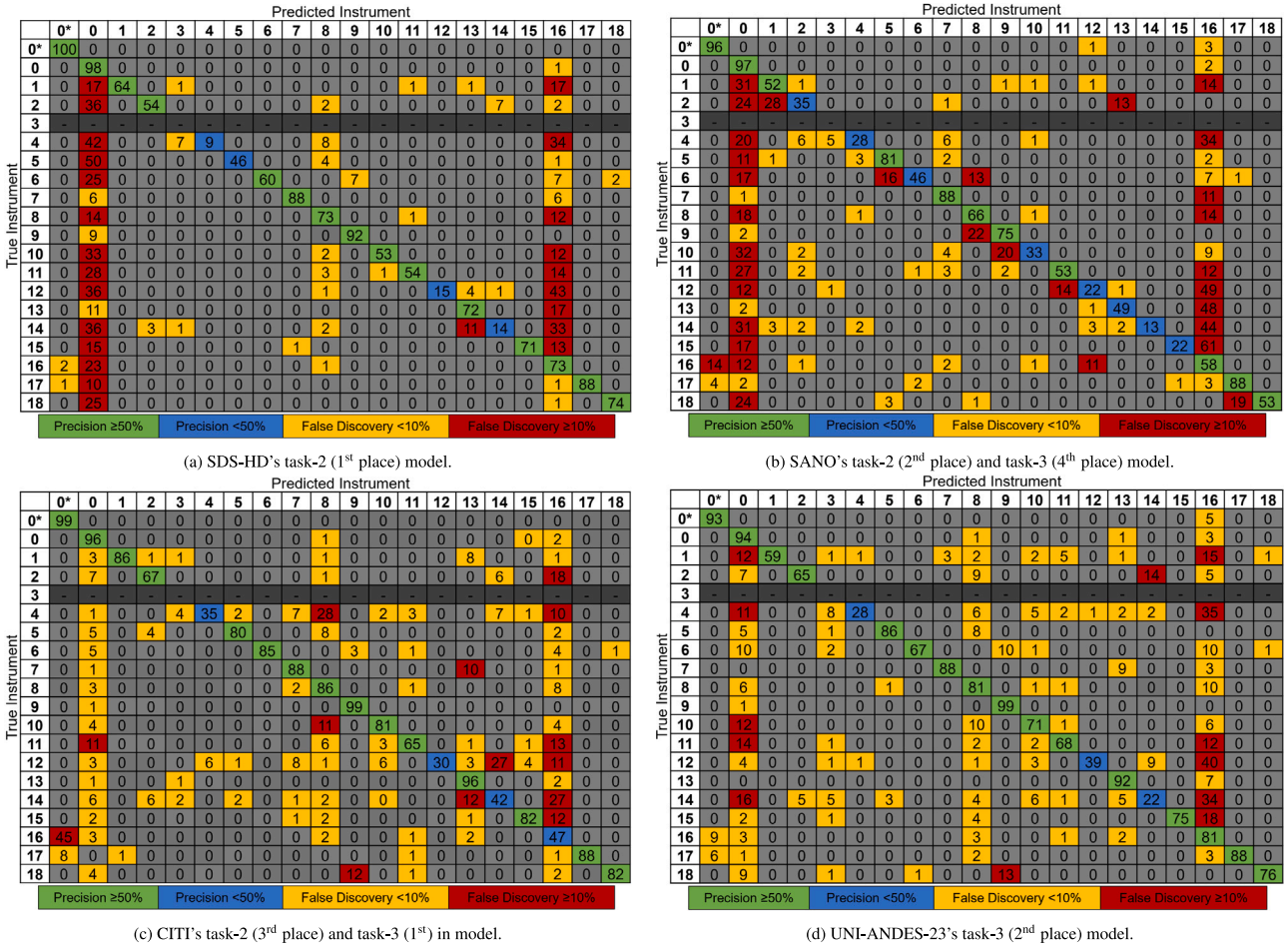
The 8-testing-videos were not released to the participants during the challenge. Instead, top results of the suggested validation split are provided in Table 8 to act as a benchmark for the community. The teams whose results are presented in this table all adhered to the suggested training–validation split, and so their models are optimised for this validation dataset.

The best performing models on the suggested validation dataset for each metric are identical to the testing dataset, implying these models have good generalisation. This is more strongly true for step recognition, where their performance drops lower (−7%) than instrument recognition (−47%). This is likely due to overfitting to the small number of images of each minor instrument class.

7. Conclusion

7.1. Principal findings

The PitVis-2023 challenge pertains to developing deep learning models for workflow recognition for the eTSA, with 3-tasks: (1) 12-step multi-class recognition; (2) 18-instrument multi-label recognition;



(a) SDS-HD's task-2 (1st place) model.

(b) SANO's task-2 (2nd place) and task-3 (4th place) model.

(c) CITI's task-2 (3rd place) and task-3 (1st) in model.

(d) UNI-ANDES-23's task-3 (2nd place) model.

Fig. 11. Instrument confusion matrices for the top models mean-averaged across the 8-testing-videos and rounded to 0 decimal places. Each cell (i, j) represents the number of instances from true class i that were predicted as class j. The diagonal entries indicate correct classifications, while off-diagonal entries show misclassifications. 0* indicates 'no secondary instrument'. Instrument-3 (cup forceps) is not present in the testing dataset and so greyed out.

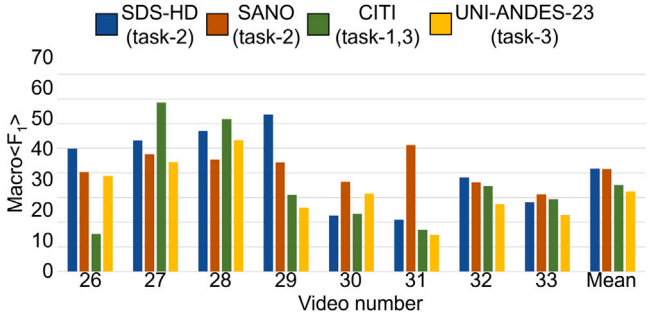


Fig. 12. Instrument recognition results across the 8-testing-videos for the best performing models: SDS-HD (task-2 1st place), SANO (task-2 2nd place), CITI (task-3 1st place), UNI-ANDES-23 (task-3 2nd place) . Per-video performance is given on the instrument recognition metric (Eq. (6)).

and (3) 12-step and 18-instrument multi-task recognition. It was run across 5-months as a sub-challenge of the EndoVis-2023 challenge, with results and awards presented at the MICCAI-2023 conference hosted in Vancouver, Canada on 08-Oct-2023. Participants were given access to the first curated public dataset of eTSA: comprising 25-videos, with annotations for each second indicating the corresponding surgical step and instrument used. Across the 3-tasks there were 18-submissions from 9-teams across 6-countries.

Table 8
Benchmark metric results for the suggested validation dataset, videos: 01, 12, 21, 24, 25. Bold indicates the best result for that column's task. Results are mean-averaged across the 5-videos and given to 0 decimal places. All teams used the suggested validation dataset split, and so models are optimised for these 5-videos. The respective metrics for task-1, task-2, and task-3 are defined in Eqs. (5), (6), and (7) respectively.

Team	Task-1	Task-2	Task-3
CITI	**70**	88	**79**
SANO	60	81	61
SDS-HD	–	**89**	–
TSO-NCT	67	–	–
UNI-ANDES-23	69	79	71

The 9-models utilise a variety of state-of-the-art computer vision and workflow recognition techniques and architectures. Training techniques include random augmentations; end-to-end training; multi-task training; and data balancing. Architectures are generally split into 2-stages. Stage-1 consists of an encoder: either purely spatial via a CNN or S-TF; or spatial–temporal via a ST-TF. Stage-2 consists of a ST-D: either a LSTM or ST-TF. An optional third stage is often used to improve performance, consisting of an online post-processing technique, usually a TSF. Some models also utilise ensembles. Performance was found to be strong for both established architectures (e.g. CNN + LSTM + TSF) as well as less established custom architectures utilising temporal propagation. A commonality between the best architectures was the use of a ST-D and TSF.

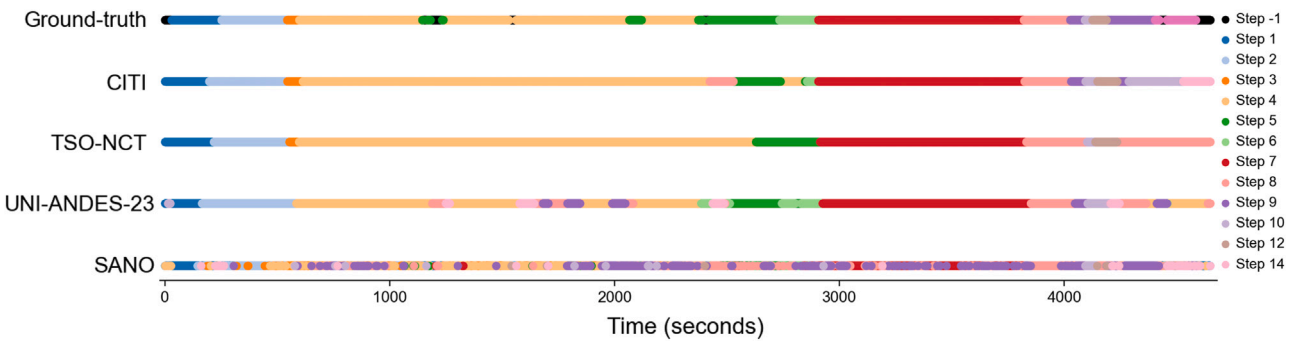


Fig. 13. Timeline of the ground-truth step classifications against the model predictions for video-26 in the testing dataset. The models are: CITI (task-1 1st place & task-3 1st place), TSO-NCT (task-1 2nd place), UNI-ANDES-23 (task-3 2nd place), SANO (task-3 4th place). Note the step '-1' annotations in the ground truth representing 'out of frame' were removed during evaluation.

7.2. Limitations

This challenge was limited primarily by the difficulty of data acquisition: obtaining consent; recording videos; and annotating videos. Firstly, all surgical videos come from a single hospital, and therefore lack diversity in: patient characteristics; surgical technique; and recording variations. Multi-centred data will improve model generalisability, although this is challenging due to patient confidentially; necessary ethics; and surgeon unease of sharing data. However, some of this can be circumvented via federated learning. Future work will address this issue through collaboration with other centres, within the UK and beyond. Moreover, releasing public datasets, as with this challenge, insensitivities other research groups to do the same.

Secondly, like with many medical datasets, the size of the dataset is small and restricted to just 25-training-videos and 8-testing-videos. This has meant minor classes either have much worse performance or are entirely removed. Although this is somewhat mitigated with data augmentation and utilising training techniques that compensate for imbalanced datasets, more data will certainly help. Potential alternative solutions to this are the use of more sophisticated computer vision techniques. For example, in the past year, pre-trained surgery specific transformers have emerged as foundational models (Schmidgall et al., 2024). Similarly, releasing and leveraging recorded videos without annotations via semi-supervised learning expands the training dataset without the labour costs of surgical annotations (Wijekoon et al., 2024; Ramesh et al., 2023).

Furthermore, although this challenge provides benchmark performances for workflow recognition in eTSA, the models' performances are not strong enough to be used in clinical practice. Steps are primarily misclassified as neighbouring steps, and so extra emphasis should be made to help clarify these boundaries. This may be achieved with upsampling these frames on step boundaries or loss functions that penalise neighbouring misclassifications. Although the challenge has ended, the website will remain, and the data is publicly available, along with the benchmark results. This means the community can improve on the models to better overcome the eTSA specific difficulties.

Other important factors to consider are: explainability of models, which is essential for a clinical setting; environmental impacts of model training, as some models were trained for long periods of time; and real-time implementation, which was enforced as models had to run at 10× speed on the 32-GB GPU.

7.3. Concluding statement

The Pituitary Vision 2023 Challenge showcases the efforts of the international minimally invasive surgical computer vision community on endoscopic pituitary surgery. The models created not only verify their generalisability on a new dataset, but advance the field, pushing it closer to usable clinical assistance.

CRediT authorship contribution statement

Adrito Das: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Danyal Z. Khan:** Writing – review & editing, Visualization, Methodology, Data curation, Conceptualization. **Dimitrios Psychogyios:** Project administration, Conceptualization. **Yitong Zhang:** Project administration, Conceptualization. **John G. Hanrahan:** Writing – review & editing, Data curation. **Francisco Vasconcelos:** Supervision, Resources, Conceptualization. **You Pang:** Methodology. **Zhen Chen:** Methodology. **Jinlin Wu:** Methodology. **Xiaoyang Zou:** Methodology. **Guoyan Zheng:** Methodology. **Abdul Qayyum:** Methodology. **Moona Mazher:** Methodology. **Imran Razzak:** Methodology. **Tianbin Li:** Methodology. **Jin Ye:** Methodology. **Junjun He:** Methodology. **Szymon Płotka:** Methodology. **Joanna Kaleta:** Methodology. **Amine Yamlahi:** Methodology. **Antoine Jund:** Methodology. **Patrick Godau:** Methodology. **Satoshi Kondo:** Methodology. **Satoshi Kasai:** Methodology. **Kousuke Hirasawa:** Methodology. **Dominik Rivoir:** Methodology. **Stefanie Speidel:** Methodology. **Alejandra Pérez:** Methodology. **Santiago Rodriguez:** Methodology. **Pablo Arbeláez:** Methodology. **Danail Stoyanov:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Hani J. Marcus:** Writing – review & editing, Supervision, Project administration, Data curation, Conceptualization. **Sophia Bano:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Ethics

The study was registered with UCL IRB (17819/011).

Data and publishing

The data for this challenge cannot be distributed but is available under a CC BY-NC-SA 4.0 licence: www.doi.org/10.5522/04/26531686. Data used in the challenge can be used for publication purposes only after the joint publication summarising the challenge results is published. For the purpose of open access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission.

Funding

This work was supported in whole, or in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z], the Engineering and Physical Sciences

Research Council (EPSRC), Canada [EP/W00805X/1, EP/Y01958X/1, EP/P012841/1], the Horizon 2020 FET [GA863146], the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies programme. Adrito Das is supported by the EPSRC, Canada [EP/S021612/1]. Danyal Z. Khan is supported by a National Institute for Health and Care Research (NIHR) Academic Clinical Fellowship and the Cancer Research UK (CRUK) Pre-doctoral Fellowship. John G. Hanrahan is supported by a NIHR Academic Clinical Fellowship. Hani J. Marcus is supported by WEISS [NS/A000050/1] and by the NIHR Biomedical Research Centre at UCL.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Adrito Das reports financial support was provided by Digital Surgery Limited. Danail Stoyanov reports a relationship with Digital Surgery Limited that includes: employment. Hani J. Marcus reports a relationship with Pandas Surgical that includes: equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the EndoVis-2023 organisation committee for running the grand challenge and the MICCAI-2023 committee for hosting the conference. With thanks to Digital Surgery Ltd, a Medtronic company, for access to Touch Surgery Ecosystem for video recording, annotation, and storage.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2025.103716.

Data availability

The training dataset has been released to the public (25-videos available at: www.doi.org/10.5522/04/26531686), along with the challenge code (available at: https://github.com/dreets/pitvis).

References

Agustsson, T., Baldvinsdottir, T., Jonasson, J., Olafsdottir, E., Steinthorsdottir, V., Sigurdsson, G., Thorsson, A., Carroll, P., Korbonits, M., Benediktsson, R., 2015. The epidemiology of pituitary adenomas in Iceland, 1955–2012: a nationwide population-based study. Eur. J. Endocrinol. 173, 655–664. http://dx.doi.org/10.1530/eje-15-0189.

Alabi, O., Vercauteren, T., Shi, M., 2024. Multitask learning in minimally invasive surgical vision: A review. arXiv, https://arxiv.org/abs/2401.08256.

Ban, Y., Rosman, G., Ward, T., Hashimoto, D., Kondo, T., Iwaki, H., Meireles, O., Rus, D., 2021. Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows. In: 2021 IEEE International Conference on Robotics and Automation. ICRA, http://dx.doi.org/10.1109/ICRA48506.2021.9561770.

Bochkovskiy, A., Wang, C., Liao, H., 2020. YOLOv4: Optimal speed and accuracy of object detection. arXiv, https://arxiv.org/abs/2004.10934.

Consortium, C., 2023. Machine learning driven prediction of cerebrospinal fluid rhinorrhoea following endonasal skull base surgery: A multicentre prospective observational study. Front. Oncol. 13, http://dx.doi.org/10.3389/fonc.2023.1046519.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S., Navab, N., 2020. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In: Lecture Notes In Computer Science, pp. 343–352. http://dx.doi.org/10.1007/978-3-030-59716-0_33.

Das, A., Bano, S., Vasconcelos, F., Khan, D., Marcus, H., Stoyanov, D., 2022. Reducing prediction volatility in the surgical workflow recognition of endoscopic pituitary surgery. Int. J. Comput. Assist. Radiol. Surg. 17, 1445–1452. http://dx.doi.org/10.1007/s11548-022-02599-y.

Das, A., Khan, D., Hanrahan, J., Marcus, H., Stoyanov, D., 2023a. Automatic generation of operation notes in endoscopic pituitary surgery videos using workflow recognition. Intell.-Based Med. 8, 100107. http://dx.doi.org/10.1016/j.ibmed.2023.100107.

Das, A., Khan, D., Williams, S., Hanrahan, J., Borg, A., Dorward, N., Bano, S., Marcus, H., Stoyanov, D., 2023b. A multi-task network for anatomy identification in endoscopic pituitary surgery. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 472–482. http://dx.doi.org/10.1007/978-3-031-43996-4_45.

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., 2024. Automated surgical skill assessment in endoscopic pituitary surgery using real-time instrument tracking on a high-fidelity bench-top phantom. Heal. Technol. Lett. 11 (6), 336–344. http://dx.doi.org/10.1049/htl2.12101.

Demir, K., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S., 2023. Deep learning in surgical workflow analysis: A review of phase and step recognition. IEEE J. Biomed. Heal. Inform. 27, 5405–5417. http://dx.doi.org/10.1109/JBHI.2023.3311628.

El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., Jegou, H., 2021. XCiT: Cross-covariance image transformers. arXiv, https://arxiv.org/abs/2106.09681.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C., 2021. Multiscale vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, http://dx.doi.org/10.1109/ICCV48922.2021.00675.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2024. EVA-02: A visual representation for neon genesis. Image Vis. Comput. 149, 105171. http://dx.doi.org/10.1016/j.imavis.2024.105171.

Frara, S., Rodriguez-Carnero, G., Formenti, A., Martinez-Olmos, M., Giustina, A., Casanueva, F., 2020. Pituitary tumors centers of excellence. Endocrinol. Metab. Clin. North Am. 49, 553–564. http://dx.doi.org/10.1016/j.ecl.2020.05.010.

Ganapathy, M., Tadi, P., 2022. Anatomy, head and neck, pituitary gland. In: StatPearls [Internet]. http://www.ncbi.nlm.nih.gov/books/NBK551529/ (Accessed August 2024).

Garrow, C., Kowalewski, K., Li, L., Wagner, M., Schmidt, M., Engelhardt, S., Hashimoto, D., Kenngott, H., Bodenstedt, S., Speidel, S., Müller-Stich, B., Nickel, F., 2020. Machine learning for surgical phase recognition: A systematic review. Ann. Surg. 273, 684–693. http://dx.doi.org/10.1097/sla.0000000000004425.

He, R., Xu, M., Das, A., Khan, D., Bano, S., Marcus, H., Stoyanov, D., Clarkson, M., Islam, M., 2024. PitVQA: Image-grounded text embedding LLM for visual question answering in pituitary surgery. arXiv, https://arxiv.org/abs/2405.13949.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Maaten, L., Weinberger, K., 2016. Densely connected convolutional networks. arXiv, https://arxiv.org/abs/1608.06993.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C., Heng, P., 2020. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. 59, 101572. http://dx.doi.org/10.1016/j.media.2019.101572.

Khan, D., Hanrahan, J., Baldeweg, S., Dorward, N., Stoyanov, D., Marcus, H., 2023. Current and future advances in surgical therapy for pituitary adenoma. Endocr. Rev. 44, 947–959. http://dx.doi.org/10.1210/endrev/bnad014.

Khan, D., Koh, C., Das, A., Valetopolou, A., Hanrahan, J., Horsfall, H., Baldeweg, S., Bano, S., Borg, A., Dorward, N., Olukoya, O., Stoyanov, D., Marcus, H., 2024a. Video-based performance analysis in pituitary surgery - Part 1: Surgical outcomes. World Neurosurg. http://dx.doi.org/10.1016/j.wneu.2024.07.218.

Khan, D., Luengo, I., Barbarisi, S., Addis, C., Culshaw, L., Dorward, N., Haikka, P., Jain, A., Kerr, K., Koh, C., Layard Horsfall, H., Muirhead, W., Palmisciano, P., Vasey, B., Stoyanov, D., Marcus, H., 2022. Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0). J. Neurosurg. 137, 51–58. http://dx.doi.org/10.3171/2021.6.jns21923.

Khan, D., Newall, N., Koh, C., Das, A., Aapan, S., Horsfall, H., Baldeweg, S., Bano, S., Borg, A., Chari, A., Dorward, N., Elserius, A., Giannis, T., Jain, A., Stoyanov, D., Marcus, H., 2024b. Video-based performance analysis in pituitary surgery - Part 2: Artificial intelligence assisted surgical coaching. World Neurosurg. http://dx.doi.org/10.1016/j.wneu.2024.07.219.

Khan, D.Z., Valetopoulou, A., Das, A., Hanrahan, J.G., Williams, S.C., Bano, S., Borg, A., Dorward, N.L., Barbarisi, S., Culshaw, L., Kerr, K., Luengo, I., Stoyanov, D., Marcus, H.J., 2024c. Artificial intelligence assisted operative anatomy recognition in endoscopic pituitary surgery. NPJ Digit. Med. 7 (1), http://dx.doi.org/10.1038/s41746-024-01273-8.

Lalys, F., Jannin, P., 2013. Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. (IJCARS) 9 (3), 495–511. http://dx.doi.org/10.1007/s11548-013-0940-5.

Lea, C., Reiter, A., Vidal, R., Hager, G., 2016. Segmental spatiotemporal CNNs for fine-grained action segmentation. In: Computer Vision – ECCV 2016. pp. 36–52. http://dx.doi.org/10.1007/978-3-319-46487-9_3.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. arXiv, https://arxiv.org/abs/2201.03545.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C., Stoyanov, D., Vedula, S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T., Hashizume, M., Heckmann-Nötzel, D., Kenngott, H., Kikinis, R., Mündermann, L., Navab, N., Onogur, S., Roß, T., Sznitman, R., Taylor, R., Tizabi, M., Wagner, M., Hager, G., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D., Joyeux, L., Lam, K., Leff, D., Madani, A., Marcus, H., Meireles, O., Seitel, A., Teber, D., Ückert, F., Müller-Stich, B., Jannin, P., Speidel, S., 2022. Surgical data science – from concepts toward clinical translation. Med. Image Anal. 76, 102306. http://dx.doi.org/10.1016/j.media.2021.102306.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M., Wiesenfarth, M., Kavur, A., Sudre, C., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M., Cheplygina, V., Cimini, B., Collins, G., Farahani, K., Ferrer, L., Galdran, A., Ginneken, B., Haase, R., Hashimoto, D., Hoffman, M., Huisman, M., Jannin, P., Kahn, C., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B., Litjens, G., Madani, A., Maier-Hein, K., Martel, A., Mattson, P., Meijering, E., Menze, B., Moons, K., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C., Shetty, S., Smeden, M., Summers, R., Taha, A., Tiulpin, A., Tsaftaris, S., Van Calster, B., Varoquaux, G., Jäger, P., 2024. Metrics reloaded: recommendations for image analysis validation. Nat. Methods 21, 195–212. http://dx.doi.org/10.1038/s41592-023-02151-z.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., Ginneken, B., Kopp-Schneider, A., Landman, B., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796. http://dx.doi.org/10.1016/j.media.2020.101796.

Mao, Z., Das, A., Islam, M., Khan, D., Williams, S., Hanrahan, J., Borg, A., Dorward, N., Clarkson, M., Stoyanov, D., Marcus, H., Bano, S., 2024. PitSurgRT: real-time localization of critical anatomical structures in endoscopic pituitary surgery. Int. J. Comput. Assist. Radiol. Surg. 19, 1053–1060. http://dx.doi.org/10.1007/s11548-024-03094-2.

Marcus, H., Khan, D., Borg, A., Buchfelder, M., Cetas, J., Collins, J., Dorward, N., Fleseriu, M., Gurnell, M., Javadpour, M., Jones, P., Koh, C., Layard Horsfall, H., Mamelak, A., Mortini, P., Muirhead, W., Oyesiku, N., Schwartz, T., Sinha, S., Stoyanov, D., Syro, L., Tsermoulas, G., Williams, A., Winder, M., Zada, G., Laws, E., 2021. Pituitary society expert Delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituit. 24, 839–853. http://dx.doi.org/10.1007/s11102-021-01162-3.

Ogra, S., Nichols, A., Stylli, S., Kaye, A., Savino, P., Danesh-Meyer, H., 2014. Visual acuity and pattern of visual field loss at presentation in pituitary adenoma. J. Clin. Neurosci. 21, 735–740. http://dx.doi.org/10.1016/j.jocn.2014.01.005.

Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C., Huang, S., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., Boels, M., Huo, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S., Xu, M., Wang, A., Wu, Y., Bai, L., Ren, H., Yamada, A., Harai, Y., Ishikawa, Y., Hayashi, K., Simoens, J., DeBacker, P., Cisternino, F., Furnari, G., Mottrie, A., Ferraguti, F., Kondo, S., Kasai, S., Hirasawa, K., Kim, S., Lee, S., Lee, K., Kong, H., Fu, K., Li, C., An, S., Krell, S., Bodenstedt, S., Ayobi, N., Perez, A., Rodriguez, S., Puentes, J., Arbelaez, P., Mohareri, O., Stoyanov, D., 2024. SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv, https://arxiv.org/abs/2401.00496.

Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., Exarchakis, G., Karargyris, A., Padoy, N., 2023. Dissecting self-supervised learning methods for surgical computer vision. Med. Image Anal. 88, 102844. http://dx.doi.org/10.1016/j.media.2023.102844.

Rueckert, T., Rueckert, D., Palm, C., 2024. Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. Comput. Biology Med. 169, 107929. http://dx.doi.org/10.1016/j.compbiomed.2024.107929.

Russ, S., Anastasopoulou, C., Shafiq, I., 2022. Pituitary adenoma. In: StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK554451/ (Accessed August 2024).

Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A., 2024. General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv, https://arxiv.org/abs/2403.05949.

Speidel, S., Maier-Hein, L., Stoyanov, D., Bodenstedt, S., Reinke, A., Bano, S., Jenke, A., Wagner, M., Daum, M., Tabibian, A., Das, Adrito, Zhang, Yitong, Vasconcelos, F., Psychogyios, D., Khan, Danyal Z., Marcus, H., Zia, Aneeq, Liu, X., Bhattacharyya, K., Wang, Ziheng, Berniker, M., Perreault, C., Jarc, A., Malpani, A., Glock, K., Xu, Haozheng, Xu, C., Huang, Baoru, Giannarou, S., 2023. Endoscopic Vision Challenge 2023. Zenodo, https://zenodo.org/record/8315050.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv, https://arxiv.org/abs/1905.11946.

Tritos, N., Biller, B., 2019. Medical management of cushing disease. Neurosurg. Clin. North Am. 30, 499–508. http://dx.doi.org/10.1016/j.nec.2019.05.007.

Twinanda, A., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M., Padoy, N., 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36, 86–97. http://dx.doi.org/10.1109/TMI.2016.2593957.

Wang, Y., Sun, Q., Liu, Z., Gu, L., 2022. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. Robot. Auton. Syst. 149, 103945. http://dx.doi.org/10.1016/j.robot.2021.103945.

Wang, F., Zhou, T., Wei, S., Meng, X., Zhang, J., Hou, Y., Sun, G., 2014. Endoscopic endonasal transsphenoidal surgery of 1 166pituitary adenomas. Surg. Endosc. 29, 1270–1280. http://dx.doi.org/10.1007/s00464-014-3815-0.

Wijekoon, A., Das, A., Herrera, R.R., Khan, D.Z., Hanrahan, J., Carter, E., Luoma, V., Stoyanov, D., Marcus, H.J., Bano, S., 2024. PitRSDNet: Predicting intra-operative remaining surgery duration in endoscopic pituitary surgery. Heal. Technol. Lett. 11 (6), 318–326. http://dx.doi.org/10.1049/htl2.12099.

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. TinyViT: Fast pretraining distillation for small vision transformers. arXiv, https://arxiv.org/abs/2207.10666.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H., 2022. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. arXiv, https://arxiv.org/abs/2203.03605.

Zou, X., Liu, W., Wang, J., Tao, R., Zheng, G., 2022. ARST: auto-regressive surgical transformer for phase recognition from laparoscopic videos. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 11, 1012–1018. http://dx.doi.org/10.1080/21681163.2022.2145238.

Zou, X., Yu, D., Tao, R., Zheng, G., 2024. An end-to-end spatial-temporal transformer model for surgical action triplet recognition. In: 12th Asian-Pacific Conference on Medical and Biological Engineering. pp. 114–120. http://dx.doi.org/10.1007/978-3-031-51485-2_14.