Artificial Intelligence In Medicine 172 (2026) 103309

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

# SurgflowNet: Leveraging unannotated video for consistent endoscopic pituitary surgery workflow recognition

Anjana Wijekoon [a,b,*], Adrito Das [a], Zhehua Mao [a,b], Danyal Z. Khan [a,c], John G. Hanrahan [a,c], Danail Stoyanov [a,b], Hani J. Marcus [a,c], Sophia Bano [a,b]

[a] *UCL Hawkes Institute, University College London, London, United Kingdom*
[b] *Department of Computer Science, University College London, London, United Kingdom*
[c] *Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, United Kingdom*

ARTICLE INFO

*Keywords:*
Endoscopic pituitary surgery
Surgical workflow recognition

ABSTRACT

Surgical workflow recognition has the potential to accelerate training initiatives through the analysis of surgical videos, improve intraoperative efficiency, and support preemptive postoperative care. Unlike well-explored minimally invasive surgeries, where surgical workflows are consistent across patients, automating endoscopic pituitary surgery workflow recognition is challenging. Pituitary surgery involves a large number of steps, diverse sequences, optional steps, and frequent transitions, making it challenging for current state-of-the-art (SOTA) methods, which struggle with transferability. Progress is largely limited by the lack of annotated data that captures the complexity of pituitary surgery, and obtaining such annotations is both time-consuming and resource-intensive. This paper presents SurgflowNet, a novel spatio-temporal model for consistent pituitary workflow recognition leveraging unannotated data. We utilise a limited yet fully annotated dataset to infer quasi-labels for unannotated videos and curate a balanced dataset to train a robust frame encoder using the student–teacher framework. A spatio-temporal network that combines the resulting frame encoder and an LSTM network is trained with a consistency loss to ensure stability in step predictions. With a 5% improvement in macro $F_1$-score and 13.4% in Edit Score over the SOTA, SurgflowNetdemonstrates a significant improvement in workflow recognition for endoscopic pituitary surgery.

## 1. Introduction

Pituitary adenomas are benign pituitary gland tumours that can cause vision loss due to mass effect or changes in appearance and bodily function from hormone imbalance, potentially leading to increased mortality [1]. Most of these tumours can be effectively treated using the eTSA, a minimally invasive surgery performed through the nostrils [2]. This skill is difficult to master, with many years of dedicated training required [1,3]. Breaking down surgery into phases and steps for analysis has been shown to provide a framework for surgical coaching, improving patient outcomes [3,4], and recent advancements in machine learning have automated this process through intra-operative workflow recognition [5–8]. Contrary to other minimally invasive surgeries such as cholecystectomy, eTSA surgeries feature a large number of surgical steps, defined as 'sequence of activities used to achieve a surgical objective [2], along with diverse step sequences and multiple optional steps, leading to highly variable workflows and surgery durations. Additionally, frequent interruptions due to the endoscope moving in and out result in numerous step transitions within a single surgery (see examples in Fig. 1). Many steps also look visually similar to each other, which further complicates accurate step recognition. For instance, in the publicly available PitVis Challenge dataset [9], each video features a unique step sequence, while across all videos there is an average of 71 ± 34 step transitions, and the tumour excision step duration varies between 3 and 68 min. Training a robust step recognition model in a fully supervised manner would therefore require an extensive volume of comprehensively annotated video data, which is resource-intensive. In the past, to overcome these challenges, eTSA workflow recognition has either merged steps or omitted steps, leading to incomplete solutions [6,8]. Instead, in this paper, we address the primary research question of *how to leverage unannotated video data to reduce reliance on full supervision when developing a step recognition model for eTSA*.

We propose a two-stage student–teacher framework for training an endoscopic pituitary surgery step recognition model that leverages both annotated and unannotated video data. Our key contributions are:

[*] Correspondence to: Charles Bell House, 43-45 Foley Street, London W1W 7TY, United Kingdom.
*E-mail address:* a.wijekoon@ucl.ac.uk (A. Wijekoon).

https://doi.org/10.1016/j.artmed.2025.103309
Received 15 July 2025; Received in revised form 30 October 2025; Accepted 13 November 2025
Available online 26 November 2025
0933-3657/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

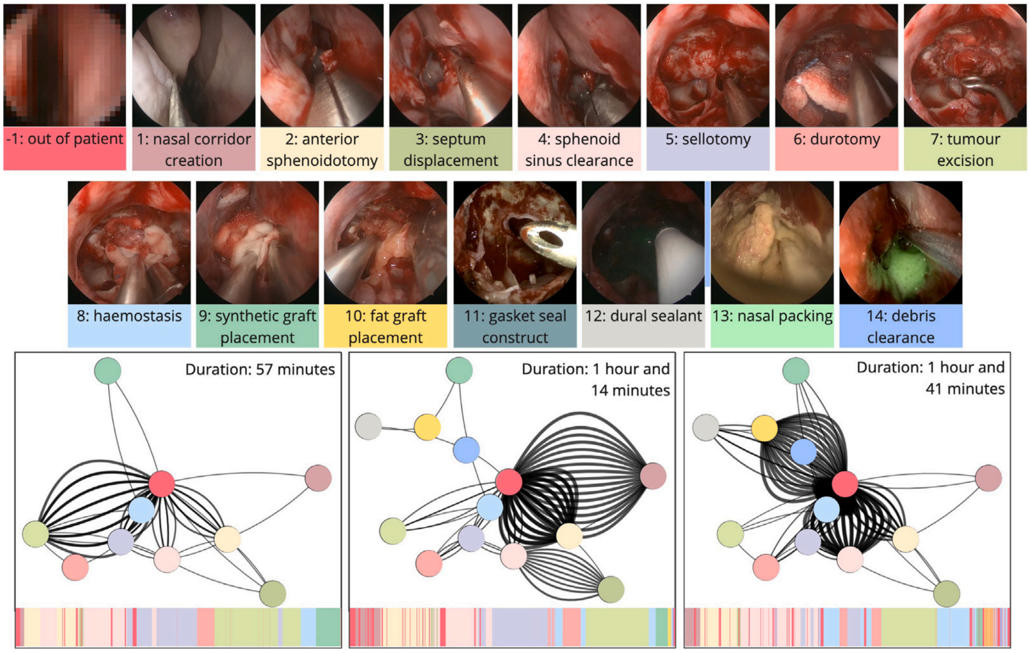

-1: out of patient
1: nasal corridor creation
2: anterior sphenoidotomy
3: septum displacement
4: sphenoid sinus clearance
5: sellotomy
6: durotomy
7: tumour excision
8: haemostasis
9: synthetic graft placement
10: fat graft placement
11: gasket seal construct
12: dural sealant
13: nasal packing
14: debris clearance
Duration: 57 minutes
Duration: 1 hour and 14 minutes
Duration: 1 hour and 41 minutes

**Fig. 1.** eTSA surgical workflow; top: surgical steps and out-of-patient example frames; bottom: three examples illustrate interlaced steps, diversity of step transitions and variability in step and surgery durations.

- a pipeline to curate a quasi-labelled dataset from annotated and unannotated surgical video, mitigating class imbalance;
- a two-step frame encoder training approach that applies semi-supervised learning in a student–teacher framework;
- a spatio-temporal model trained with a consistency loss to achieve step prediction stability; and
- a comprehensive evaluation in comparison to SOTA step recognition models, semi-supervised and self-supervised approaches, as well as ablated variants.

To the best of our knowledge, this is one of the most fine-grained step recognition frameworks to date for endoscopic pituitary surgery, leveraging unannotated data to achieve consistent workflow recognition. SurgflowNet contributes to improving an array of clinical applications, including intraoperative guidance, surgical coaching, and postoperative documentation. Intraoperatively, accurate step recognition can help surgeons identify critical points—such as the start of the sella phase, where anatomy segmentation models can be activated to provide decision support for safer opening [10]; and can also improve estimates of remaining surgery duration for the anaesthetics team and scheduling staff [11]. For surgical coaching, SurgflowNet will automate the extraction of step sequences and durations, streamlining a process that is currently performed manually, expanding the pool of videos available for surgical trainees. Postoperative surgical note generation will be more comprehensive, replacing our previous step recognition model [12], and automatic detection of out-of-patient frames will allow surgical videos to be anonymised before storage, protecting patient privacy.

The rest of the paper is organised as follows: Section 2 reviews related work, followed by the details of the proposed method in Section 3. Section 4 presents the evaluation and implementation details. Comparative results, ablation study and the qualitative evidence are presented in Section 5 followed by a discussion on limitations of this work. Finally, Section 6 concludes with a summary of findings and future directions.

## 2. Related work

Surgical workflow analysis seeks to systematically break down surgical procedures into discrete clinically meaningful units of procedures and errors [2]. This is a hierarchical process where procedures are broken down into phases, defined as 'major events occurring during a surgical procedure' each composed of a series of steps, generating unique workflows [13]. During each step (e.g. nasal corridor creation), surgical instruments (e.g. Freer elevator) are used to perform manoeuvres via a series of gestures [2].

Surgical phase recognition is one of the most explored workflow analysis tasks in surgical vision [14]. The most prevalent approaches employ a spatio-temporal neural network trained in a supervised manner, utilising labelled data. Earlier networks relied heavily on a fine-tuned ResNet50 encoder for feature extraction [15–18], and more recently, ConvNeXt and ViT encoders are used [19,20]. Temporal dependencies are learned with Long Short-Term Memory (LSTM) [15], Temporal Convolutional Network (TCN) [16] or Transformers [17,21,22]. One of the latest comparative studies by Rivoir et al. revealed that the state-of-the-art approach for laparoscopic and robotic surgeries is a ConvNeXt-LSTM architecture trained end-to-end with a partially frozen encoder [19]. It outperformed several other commonly used architectures and training strategies, including TeCNO [16] and Trans-SVNet [23].

In endoscopic pituitary surgery, workflow analysis focuses on recognising surgical steps, providing greater granularity needed in clinical applications. Step recognition in endoscopic pituitary surgery is a significantly challenging yet developing research area [5,6,24]. The PitVis Challenge made significant contributions in the step recognition task, achieving 0.611 ± 0.106 macro $F_1$-score and 0.647 ± 0.101 edit-score over 12 surgical steps (video-level) [8]. However, these networks are limited by the insufficient training data to robustly recognise rare surgical steps, including when the endoscope is outside the patient.

Self-supervised learning strategies have been employed in surgical vision to address the limitations of annotated data and to develop foundational models for various downstream tasks, including surgical phase recognition. Earlier works have utilised contrastive learning [18,25,26], knowledge distillation [20,27], mask reconstruction [28] and student–teacher framework [29,30] techniques for training networks used in phase recognition in cholecystectomy and cataract surgeries. Evidence from the computer vision research has shown that self-distillation methods, such as BYOL [31] and DINOv2 [32], are

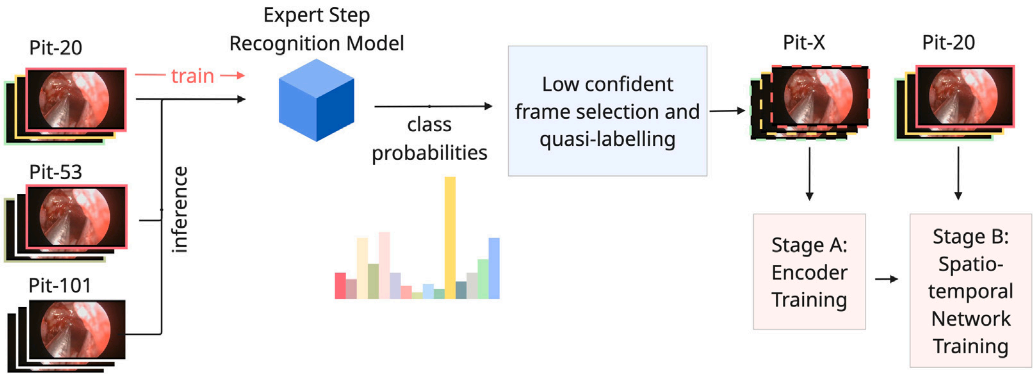


**Fig. 2.** Workflow of Pit-X curation and SurgflowNet training; Frames with coloured borders are annotated (Pit-20 and part of Pit-53); white borders are unannotated (Pit-101); dashed coloured borders indicate quasi-labelled frames (Pit-X).

computationally efficient compared to contrastive learning approaches like SimCLR [33] and MoCo [34]. Additionally, DINOv2 [32] and LEMON [27] have demonstrated the effectiveness of curated datasets for self-supervision, resulting in enhanced representation learning.

Building on these insights, our approach to distilling pituitary surgical workflow knowledge from raw video data begins with the curation of a diverse pre-training dataset through a tailored filtering pipeline. In contrast to the literature, our pipeline focuses on identifying cases where current SOTA methods fail, such as frames around a step transition. We adopt a student–teacher framework for training the frame encoder, as it enables effective utilisation of both unannotated and annotated data (both ground truth and quasi-labelled) by combining self-supervision with semi-supervision. Therefore, we contrast with prior works that rely solely on full supervision [6,19] and address unique challenges in eTSA by adapting recent advances in self-supervision to leverage unannotated video.

## 3. Methods

SurgflowNet is a spatio-temporal neural network trained in two stages for consistent step recognition. We first curate the Pit-X dataset by leveraging both annotated and unannotated data focused on identifying ambiguous or challenging cases (Section 3.1.2). Stage A trains a frame-level ConvNeXt encoder on the curated Pit-X dataset via a student–teacher framework applying semi-supervised learning over two steps (Section 3.2). In Stage B, the encoder-LSTM network is fine-tuned end-to-end for consistent step recognition (Section 3.3). Overall workflow is illustrated in Fig. 2.

### *3.1. Pit-X dataset curation*

#### *3.1.1. Source datasets*

We utilise three eTSA video datasets with varying levels of annotations, fully annotated, partially annotated and unannotated. They are referred to as Pit-20, Pit-53 and Pit-101 where the numbers indicate the number of videos. Pit-20 comprised 20 out of 25 videos publicly available from the PitVis challenge [9]. Pit-20 videos are annotated for 14 surgical steps (8 core and 6 optional), with the out-of-patient frames representing a 15th class. Pit-53 dataset is only partially annotated — each out-of-patient frame is labelled as the preceding surgical step, which exaggerates the frame count for each surgical step. The surgical videos were annotated by two trainee neurosurgeons, with discrepancies resolved through discussion and mutual agreement, and a consultant neurosurgeon subsequently verified all annotations. Videos in Pit-101 are not annotated. To the best of our knowledge, this represents the largest surgical video collection used to date in eTSA workflow analysis. Figs. 3 and 4 summarises the data distribution statistics across the three datasets.

The videos were collected from surgeries performed by three lead surgeons at a single centre, National Hospital for Neurology and Neurosurgery (NHNN), London, UK, between 2018 and 2024. A high-definition endoscope (Hopkins Telescope, Karl Storz Endoscopy) was used to record the videos. The data collection and subsequent use for research purposes received ethical approval from the Institutional Review Board at University College London (UCL) (17819/011), and informed consent was obtained. All videos were uploaded and analysed using Touch Surgery Ecosystem, an AI-powered surgical video management and analytics platform provided by Medtronic.[1] Using their internal software, all images outside of the patient were blurred to de-identify the patient and surgical team. The videos were then reduced to 720p (1280 × 720) resolution at 24-frames per second (FPS) using the publicly available software handbrake,[2] and stored as mp4 files. Images were sampled from the videos at 1-frames per second (FPS); centre cropped to 720 × 720 to remove the excessive black borders; resized to 256 × 256; and stored in PNG format.

#### *3.1.2. Pit-X dataset curation pipeline*

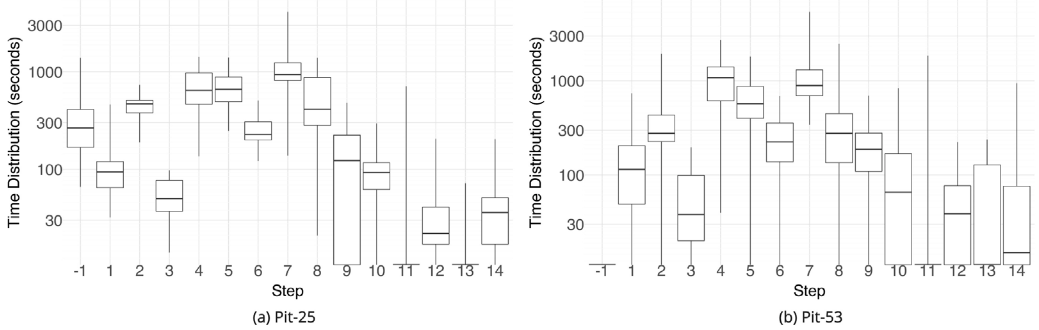
The curation pipeline serves two objectives: (1) mitigating severe class imbalance; and (2) identifying frames that current step recognition models find challenging or ambiguous to enhance the robustness of the frame encoder and consistency of step recognition. The limited but fully annotated dataset, Pit-20, is used to train a SOTA step recognition model (we selected the best-performing model from Rivoir et al. [19]), which is then used to obtain step predictions for all three datasets. To understand where SOTA models struggle with pituitary workflow recognition, we analyse the confidence of these predictions.

We identify frames with low step prediction confidence as those where the difference between the predicted probabilities (after softmax) of the most likely class and the second-most likely class falls below a predefined threshold, $\epsilon$. The threshold $\epsilon$, defines the confidence gap where it is considered uncertain. In this work, we have set $\epsilon = 0.9$, which has been selected considering the balance between the number of low confidence frames chosen for each step class (a higher $\epsilon$ returns more frames, and a lower $\epsilon$ returns fewer frames). When the preceding and succeeding frames have confident and matching step predictions, we use this consistency to infer a quasi-label for a low-confidence frame.

From all available frames, we select a stratified subset of 116,491 frames across the three datasets, capping 10,000 instances per step class, to form Pit-X. For each step, the low-confidence frames with their quasi-labels are prioritised over confident frames, encouraging the model to learn from uncertain examples during self-supervised and semi-supervised training. The final composition of Pit-X across source datasets and steps is presented in Fig. 5.

[1] https://www.medtronic.com/covidien/en-gb/products/touch-surgery.html

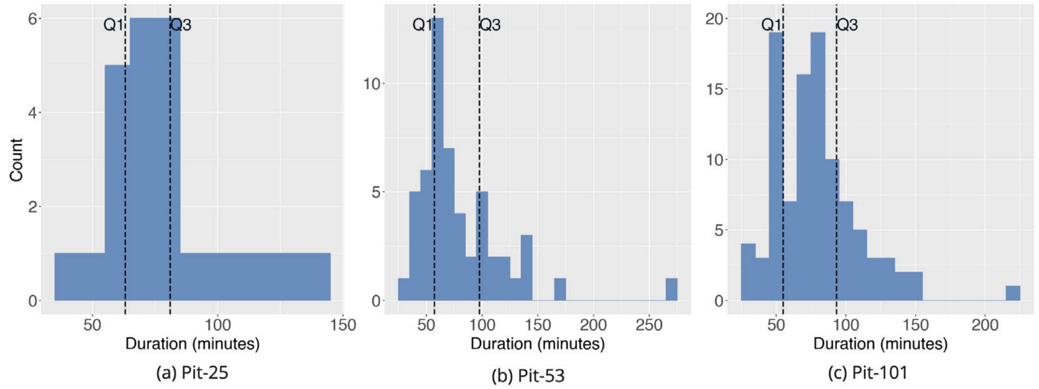
[2] https://handbrake.fr/



(a) Pit-25

(b) Pit-53

**Fig. 3.** Distribution of step durations shown as box plots, indicating the minimum, 25th percentile, median, 75th percentile, and maximum for: (a) the Pit-25 dataset – comprising the Pit-20 videos and the validation set; and (b) the Pit-53 dataset – where out-of-patient frames are annotated with the preceding surgical step. Distribution for Pit-101 is unavailable as it is fully unannotated.



(a) Pit-25

(b) Pit-53

(c) Pit-101

**Fig. 4.** Surgery duration statistics: (a) Pit-25; (b) Pit-53; and (c) Pit-101 annotated with 25th (Q1) and the 75th (Q3) percentiles. Median durations are 73 min, 67 min and 75 min; and inter-quartile ranges are 63–81 min, 57–97 min and 55–93 min for Pit-25, Pit-53 and Pit-101 respectively.

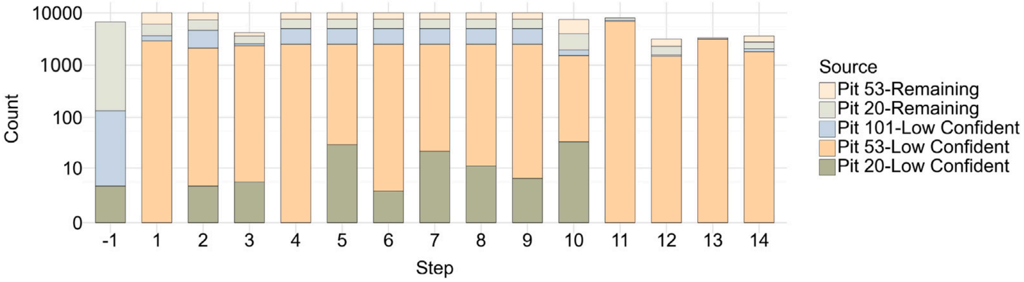


**Fig. 5.** Composition of the Pit-X dataset across source datasets and step classes. As the ‘expert’ model is trained on Pit-20, the number of ‘low-confidence’ instances from Pit-20 is limited, and for readability, the $y$-axis is presented on a log10 scale.

### 3.2. Stage A: Encoder training

Stage A training is performed in two steps, utilising the Pit-X dataset. We select Bootstrap Your Own Latent (BYOL) [31] student–teacher framework for Stage A training due to its effective adaptability for both self-supervised and semi-supervised learning (Fig. 6). The online network in Blue consists of a ConvNeXt encoder, initialised with ImageNet pre-trained weights, 2 fully connected layers as the projector and two prediction heads. The target network in Green consists of an encoder and a projector that are architecturally identical to those on the online network. ConvNeXt [35] was selected empirically as the encoder architecture over ResNet50 [16], ResNet50-GN [19] and Swin-ViT [36]. During training, the online network is updated via gradient descent and the target network is updated with the exponential moving average (ema) of the online network.

Step 1 utilises Pit-X in a self-supervised manner to progressively match the online latent predictions with the target projections using *cosine distance* loss, $\ell 1$. Minimising $\ell 1$ encourages similarity in feature space while remaining invariant to transformations. Step 2 introduces a step prediction head to the online network and utilises Pit-X in a semi-supervised manner. The combined losses from latent representation matching and *cross-entropy* loss, $\ell 2$ calculated between step predictions and quasi-labels further train the online network to discern between surgical steps. During step 2, $\ell 1$ and $\ell 2$ losses are given equal weights. At the end of Stage A, all network components are discarded except the ConvNeXt encoder of the online network, which is referred to as the pituitary encoder or $E^p$.

### 3.3. Stage B: Spatio-temporal network fine-tuning

The spatio-temporal network for step recognition consists of the pituitary encoder and an LSTM network. It is trained end-to-end in a supervised manner using the limited but fully annotated Pit-20 dataset, with the encoder partially frozen to retain the distilled knowledge from

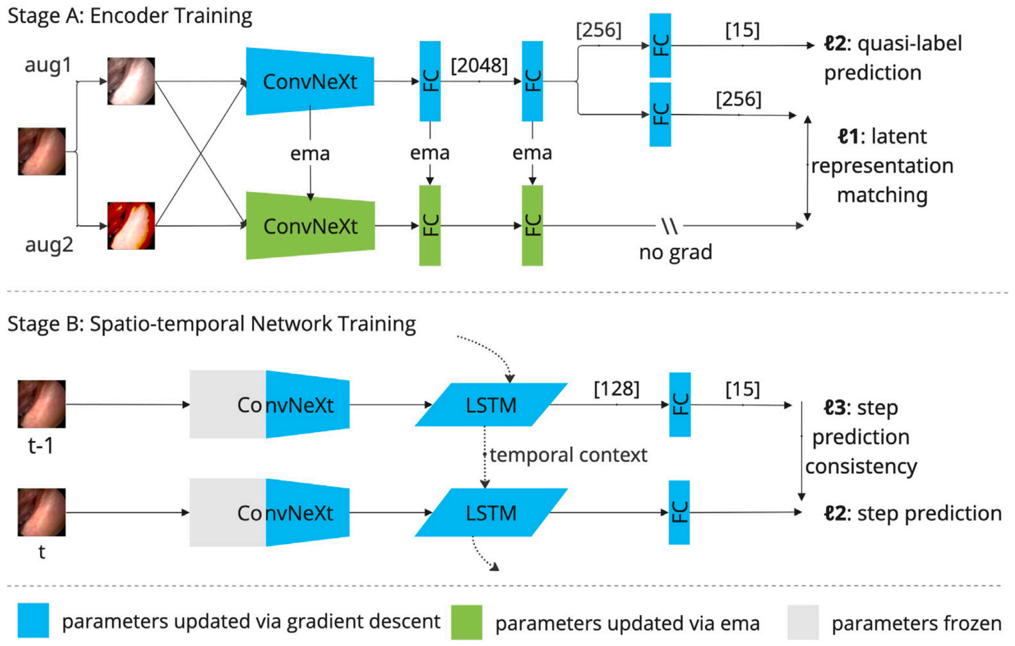

Stage A: Encoder Training
aug1
aug2
ConvNeXt
ConvNeXt
FC
[2048]
FC
[256]
FC
[15]
ℓ2: quasi-label prediction
FC
[256]
ℓ1: latent representation matching
ema
ema
ema
FC
FC
\\
no grad
Stage B: Spatio-temporal Network Training
ConvNeXt
LSTM
[128]
FC
[15]
ℓ3: step prediction consistency
t-1
temporal context
ConvNeXt
LSTM
FC
ℓ2: step prediction
t
parameters updated via gradient descent
parameters updated via ema
parameters frozen

**Fig. 6.** Two stage training of the SurgflowNet for consistent step recognition; Stage A trains a frame-level ConvNeXt encoder on the curated Pit-X dataset using self-supervision and semi-supervision; Stage B trains a spatio-temporal network using cross-entropy and consistency losses for accurate and stable step prediction.

Stage A while allowing task-specific fine-tuning. In addition to *cross-entropy loss* (**$\ell 2$**) for improving step prediction accuracy, we introduce a *Smooth L1 loss*, **$\ell 3$**, to enforce step prediction consistency. Given $p_i^t$ and $p_i^{t-1}$ are prediction probabilities for class $i$ at time $t$ and $t-1$, **$\ell 3$** for a sequence of length $S$ is calculated as follows, where $C$ is the number of classes.

$$\ell 3 = \frac{1}{(S-1)\times C}\sum_{t=2}^{S}\sum_{i=1}^{C} SmoothL1(p_i^t, p_i^{t-1}) \qquad (1)$$

**$\ell 3$** loss reduces prediction volatility by minimising the difference in step predictions between consecutive frames. Stage B is trained on a weighted combination of **$\ell 2$** and **$\ell 3$**, ($\ell 2 + \alpha \ell 3$). To find the balance between step prediction accuracy and consistency, the most optimal weight $\alpha \in \mathbb{R}^+$ is determined via a hyper-parameter search.

## 4. Evaluation

### *4.1. Evaluation methodology*

The proposed SurgflowNet is compared with the following existing SOTA models. Their implementations were adapted from respective publications and GitHub repositories.

- **Supervised step recognition models: ConvNeXt-LSTM** [19] and **EndoViT-TCN** [28] are the SOTA for cholecystectomy workflow recognition, outperforming previously well-known methods [19] such as TeCNO [16] and Trans-SVNet [23]. **ResNet50-LSTM-TSF** [6] from the pituitary workflow domain proposed a post-processing temporal smoothing function (TSF) to improve step prediction consistency.
- **Semi-supervised methods: SRC_MT** [37] and **ABCL** [38] are SOTA semi-supervised learning methods for medical image classification. While both follow the well-known mean-teacher framework [39], ABCL further considers the imbalance of training data via an adaptive consistency loss. Both methods were applied to ConvNeXt encoders. To ensure fairness with models that leverage temporal context, we extract frame-level features after semi-supervised learning and train an LSTM network following the 2-stage approach from Rivoir et al. [19].
- **Self-supervised encoders: DINOv2** [32] from the public domain and **EndoViT** [28] from the surgical vision domain are SOTA self-supervised image encoders. For fairness, we fine-tune the encoder, extract frame-level features and train an LSTM network following the 2-stage approach from Rivoir et al. [19]. Including these encoder networks enables us to compare our approach against SOTA feature representation learning methods.

We also conduct an ablation study with the following variants:

- **No Stage A:** Stage B uses a partially frozen ConvNeXt encoder initialised with ImageNet pre-trained weights;
- **Stage A self-supervised:** $E^p$ is trained only on Step 1 with **$\ell 1$** loss;
- **Stage A semi-supervised:** $E^p$ is trained in a semi-supervised manner – **$\ell 1$** loss is calculated for all instances and **$\ell 2$** loss is calculated only for instances with ground-truth;
- **No Pit-X curation** Stage A 2-step training utilises all available data with quasi-labels;
- **No partial freeze** Stage B trained end to end without partially freezing $E^p$;
- **No $\ell 3$ loss:** Stage B trained only on **$\ell 2$** loss; and
- **No $\ell 3$ loss+TSF** [6]**:** Stage B trained only on **$\ell 2$** loss and TSF [6] applied in post-processing.

### *4.2. Performance metrics*

Comparative analysis results are reported using the following metrics. All metrics are calculated per video and averaged to report mean±margin of error (*moe*) corresponding to a 95% confidence interval. Statistical significance is evaluated for the performance metrics of each method compared with those of the best-performing method. The *moe* is derived from the t-distribution ($t_{0.975,n-1} \times \frac{s}{\sqrt{n}}$; $s$: *sample standard deviation*; $n$: *number of videos*) and statistical significance is assessed using the Wilcoxon signed-rank test, both accounting for the limited number of test samples and potential non-normality of the metrics.

- **Macro $F_1$-score** is the unweighted mean of F1-scores across all step classes, without being influenced by class imbalance.

- **Balanced Accuracy** is the unweighted mean of recall across all step classes. Compared to accuracy, recall captures the performance across both common and rare step classes.
- **Edit Score** is the normalised Levenshtein distance between the sequences of ground truth and step predictions, measuring the minimal edits needed to match them. A higher score indicates better alignment, effectively penalising misclassifications and over-segmentation in step predictions [8].
- **Combined Score** calculates the arithmetic mean of macro $F_1$-score and Edit score [8].

### 4.3. Implementation details

At Stage A, the data loader randomly applied colour jitter, greyscale, horizontal flip, Gaussian blur, and resized cropping in addition to ImageNet normalisation. The optimal output dimension for the projector and latent predictor was empirically determined as 256 over 128 and 512. Each step in Stage A was trained for 100 epochs using Stochastic gradient descent (SGD) optimiser with a learning rate (lr)=1e−2, weight decay=1e−5 and batch size=64. The target decay rate, $\tau$ for *ema* was updated using cosine annealing, starting at 0.99 and reaching 1.00 over the 100 epochs.

At Stage B, the first 55% parameters of $E^p$ were frozen, and the network was trained in a supervised manner for a maximum of 200 epochs with SGD optimiser (lr=1e−4, weight decay=1e−2) with sequence length 128. For Stage B training, a hyper-parameter search was performed by fixing the $\boldsymbol{\ell 2}$ loss weight and varying the $\boldsymbol{\ell 3}$ loss weight $\alpha$ from 0.0 to 1.5 in steps of 0.1. The best performance (combined score) was achieved with $\alpha = 1.0$, corresponding to equal weights for $\boldsymbol{\ell 2}$ and $\boldsymbol{\ell 3}$. The details of the hyperparameter search are included in Appendix A.

The limited but fully annotated Pit-20 dataset was used to train Stage B. Five videos, held out from the PitVis public dataset, were used as a validation set for early stopping. Training was halted if the macro $F_1$-score score on the validation set did not improve for 5 consecutive epochs. For consistency and comparability with prior work, we intentionally used the same PitVis challenge test set, comprising 8 videos excluded from all source datasets (Pit-20, Pit-53, and Pit-101), as our test set. These test videos were used to report the final step recognition performance.

#### 4.3.1. Code and resource usage

The code was written in PyTorch, and the repository will be available on GitHub.[3] A 32 GB NVIDIA Tesla V100 Tensor Core GPU was used for training both stage A and B. In Stage A, one epoch took 2619.20 ± 26.42 s for Step 1 and 1802.46 ± 1.66 s for Step 2, requiring up to 2 GB of GPU memory. In Stage B, an epoch took on average 516.27 ± 1.39 s and utilised up to 8 GB of GPU memory. SurgflowNet PyTorch model achieved a 110.42 ± 3.24 FPS on GPU, indicating potential for real-time deployment, though further optimisation may be required for fully real-time performance in clinical settings.

## 5. Results and discussion

### 5.1. Comparative study

Table 1 summarises the performance of the proposed SurgflowNet against selected baselines. SurgflowNet outperformed all seven baselines, exceeding the second best ConvNeXt-LSTM [19] by 5.5% in macro $F_1$-score and ResNet-LSTM-TSF [6] by 13.4% in Edit Score. SurgflowNet achieved the best combined score of 0.545 and was statistically significant over the second-best method ConvNeXt-LSTM [19]. Additional significant testing results with Wilcoxon signed-rank test p-values are included in Appendix C. Semi-supervised methods that also utilised unannotated data in encoder training failed to surpass SurgflowNet or supervised SOTA baselines. However, ABCL outperforming SRC_MT suggests that mitigating class imbalance improves representation learning for surgical workflow analysis. Applying TSF in ResNet-LSTM-TSF [6] yielded the second-highest Edit Score, yet did not surpass SurgflowNet, while all other methods performed significantly poorly in step prediction consistency. A class-wise breakdown of SurgflowNet performance and comparison with the second-best performing ConvNeXt-LSTM [19] showed significant improvements on steps 1(+16%), 3(+20%), 5(+15%), and 12(+35%); minor gains on steps 8(+5%), 9(+4%), and 10(+1%); and slight decreases on steps −1(−4%), 2(−4%), 4(−4%), 6(−4%), and 7(−6%). The detailed confusion matrix is included in Appendix B.

### 5.2. Ablation study

In Table 2, SurgflowNet performance is compared against the ablated variants. The $\boldsymbol{\ell 3}$ loss–ablated variant outperformed SurgflowNet across all metrics except the Edit Score. SurgflowNet achieved the highest Edit Score, emphasising the role of $\boldsymbol{\ell 3}$ loss in reducing step prediction volatility. Both $\boldsymbol{\ell 3}$ loss and TSF affected step recognition performance, but TSF [6] had a more severe impact, unlike in the 7-step task, where TSF improved step recognition accuracy [6]. This makes $\boldsymbol{\ell 3}$ loss a more generalisable solution towards achieving step prediction consistency.

Fig. 7 presents the surgical workflow of two test videos, comparing ground truth with $\boldsymbol{\ell 3}$ loss ablated variant and SurgflowNet step predictions. When $\boldsymbol{\ell 3}$ loss is ablated, predictions are volatile, allowing quick step transitions, either correct or incorrect, as highlighted by Orange markers. In contrast, the lack of quick transitions in SurgflowNet has impacted the ability to swiftly correct step predictions as highlighted by Purple markers. These observations correlate with the Combined Scores in Table 2 highlighting that SurgflowNet finds the balance between consistency and accuracy.

### 5.3. Comparison with the PitVis challenge

The PitVis challenge focused on a 12-step recognition task. ‘Out of patient’ was not considered as a class and steps 11 and 13 were not evaluated due to insufficient representative data in the training set. To compare SurgflowNet with PitVis challenge submissions, we adopt a similar evaluation methodology and compare the results against those reported in Das et al. [8]. We updated the labels of Pit-20, the validation set and the test set to emulate the 12-step workflow and train SurgflowNet Stage B for the 12-step classification task while Stage A remained unchanged. SurgflowNet outperformed PitVis challenge best-performing model in macro F1-score achieving 0.682 ± 0.11, a 7.1% increase. SurgflowNet was the second best overall with an Edit Score of 0.555 ± 0.10 and a Combined Score of 0.619 ± 0.07.

### 5.4. Discussion

With these results, we have shown that SurgflowNet consistently outperformed all seven baselines on macro $F_1$-score and Edit Score, demonstrating strong overall performance and step prediction consistency. The ablation study highlighted the performance gain from each component in the two-stage approach and the importance of the $\boldsymbol{\ell 3}$ loss as a more generalisable approach to enhancing temporal consistency. When adapted to the 12-step classification task used in the PitVis challenge, SurgflowNet achieved a 7.1% improvement in macro $F_1$-score over the challenge’s top submission. In qualitative evidence, we have shown that, while SurgflowNet reduces prediction volatility, it sometimes delays the correction of misclassifications.

While SurgflowNet demonstrates strong performance, we note the following implications. First, as described in Section 3.1.2, our approach can select low-confidence frames that are misclassified. We

[3] https://github.com/anjanaw/pit-surgflownet.git

**Table 1**
Performance comparison with SOTA baselines; Methods where a temporal network was added for fair comparison are indicated by split columns Spatial and Temporal. Each value is mean±*moe*; the best performance is in bold; second-best is marked†; asterisk (*) indicates statistical significance over the second best with $p < 0.05$. B-Accuracy stands for Balanced Accuracy; SurgflowNet achieved the best performance across every metrics.

| | Model | | Metric | | | |
|---|---|---|---|---|---|---|
| | Spatial | Temporal | macro $F_1$-score↑ | B-Accuracy↑ | Edit Score↑ | Combined Score↑ |
| Supervised | ResNet-LSTM-TSF [6] | | 0.364 ± 0.04 | 0.421 ± 0.06 | 0.365 ± 0.06† | 0.365 ± 0.05 |
| | EndoViT-TCN [28] | | 0.343 ± 0.06 | 0.428 ± 0.08 | 0.154 ± 0.05 | 0.249 ± 0.04 |
| | ConvNeXt-LSTM [19] | | 0.536 ± 0.09† | 0.597 ± 0.07† | 0.253 ± 0.06 | 0.394 ± 0.06† |
| Semi-supervised | SRC-MT [37] | LSTM | 0.370 ± 0.05 | 0.467 ± 0.05 | 0.177 ± 0.04 | 0.273 ± 0.04 |
| | ABCL [38] | LSTM | 0.469 ± 0.08 | 0.542 ± 0.07 | 0.161 ± 0.04 | 0.315 ± 0.04 |
| Self-supervised | DINOv2 [32] | LSTM | 0.491 ± 0.12 | 0.559 ± 0.10 | 0.105 ± 0.04 | 0.298 ± 0.07 |
| | EndoViT [28] | LSTM | 0.380 ± 0.05 | 0.469 ± 0.07 | 0.219 ± 0.04 | 0.299 ± 0.04 |
| Proposed | **SurgflowNet** | | **0.591 ± 0.12** | **0.647 ± 0.12** | **0.499 ± 0.07*** | **0.545 ± 0.09*** |

**Table 2**
Performance comparison with ablated variants; SurgflowNet with the $\ell 3$ loss achieved the best Combined Score, balancing between accuracy and consistency.

| Ablation | macro $F_1$-score↑ | B-Accuracy↑ | Edit Score↑ | Combined Score↑ | Δ vs SurgflowNet (Combined Score %) |
|---|---|---|---|---|---|
| No Stage A | 0.527 ± 0.08 | 0.597 ± 0.07 | 0.252 ± 0.04 | 0.389 ± 0.05 | −15.6 |
| Stage A self-supervised | 0.457 ± 0.10 | 0.530 ± 0.08 | 0.491 ± 0.08 | 0.474 ± 0.06 | −7.1 |
| Stage A semi-supervised | 0.482 ± 0.08 | 0.555 ± 0.08 | 0.256 ± 0.05 | 0.369 ± 0.04 | −17.6 |
| No PitX curation | 0.505 ± 0.09 | 0.575 ± 0.09 | 0.390 ± 0.08 | 0.448 ± 0.06 | −9.7 |
| No partial freeze | 0.557 ± 0.11 | 0.614 ± 0.12 | 0.380 ± 0.10 | 0.469 ± 0.08 | −7.6 |
| No $\ell 3$ loss | **0.608 ± 0.10** | **0.656 ± 0.10** | 0.230 ± 0.04 | 0.419 ± 0.06 | −12.6 |
| No $\ell 3$ loss + TSF [6] | 0.581 ± 0.09 | 0.620 ± 0.09 | 0.496 ± 0.06† | 0.538 ± 0.06† | −0.7 |
| **SurgflowNet** | 0.591 ± 0.12† | 0.647 ± 0.12† | **0.499 ± 0.07** | **0.545 ± 0.09** | – |

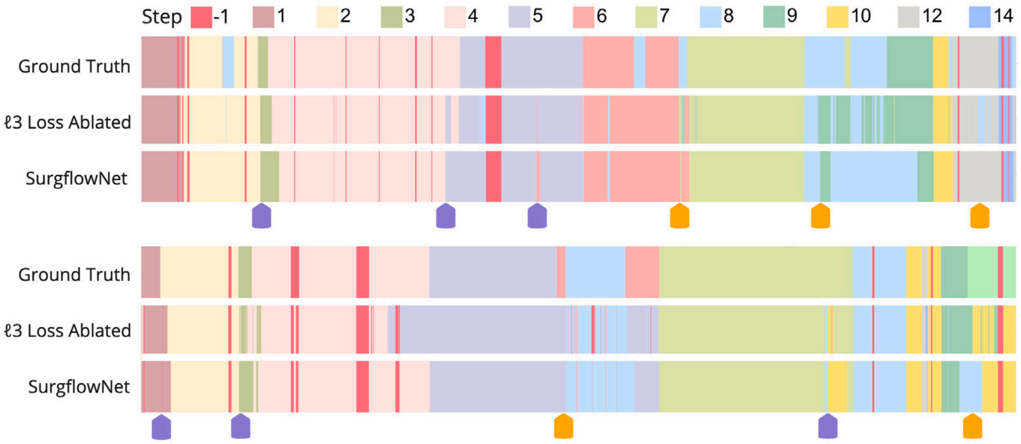


**Fig. 7.** Workflow prediction sequence comparison between SurgflowNet and $\ell 3$ loss ablated variant; Purple markers highlight instances where SurgflowNet fails to swiftly correct step predictions due to $\ell 3$ loss; Orange markers highlight volatile step transitions made by the $\ell 3$ loss ablated variant.

mitigate this by checking for consistency with preceding and succeeding frames; nonetheless, a few frames may still be mislabelled and affect Stage A Step 2 training. Second, model performance is sensitive to hyperparameter choices, particularly the weighting between loss terms (see Appendix A). We recommend exploring optimal $\alpha$ when applying SurgflowNet to other datasets or clinical domains to achieve the best balance between accuracy and consistency. Thirdly, the proposed data curation pipeline and SurgflowNet training depend on a small, fully annotated dataset, such as Pit-20.

The current experiments are based on data collected from three lead surgeons at a single centre (NHNN, London). Accordingly, the findings presented in this work are limited to a specific surgical protocol [40]. However, to the best of our knowledge, this is the first instance of a pituitary surgery dataset of this scale being used in any surgical AI task for pituitary surgery. Extending to multi-centre datasets is a natural next step, which we are actively working towards. However, these efforts face practical challenges, including ethics approvals, ensuring patient privacy, and complying with data-sharing restrictions, which have limited timely access to surgical video data.

As detailed in Section 1, SurgflowNet contributes to improving and automating several clinical applications. While some of these applications, such as intraoperative guidance [41] and surgical note generation [12], have demonstrated effectiveness in prior studies, others, such as the use of remaining surgery duration, are currently undergoing preclinical validation with target clinical user groups. Ongoing work will validate these applications and establish the utility of SurgflowNet to improve the clinical workflows.

## 6. Conclusion

This paper presented SurgflowNet– a novel spatio-temporal network for consistent step recognition in endoscopic pituitary surgery. The proposed combination of self-supervision and semi-supervision, leveraging

**Table A.3**
Hyper-parameter search for $\alpha$; each value is mean±*moe*. The best for each metric is highlighted in bold text. The best combined score is achieved at $\alpha = 1.0$.

| $\ell 3$ loss weight ($\alpha$) | macro $F_1$-score↑ | B-Accuracy↑ | Edit Score↑ | Combined Score↑ |
|---|---|---|---|---|
| 0.0 | 0.608 ± 0.10 | 0.656 ± 0.10 | 0.230 ± 0.04 | 0.419 ± 0.06 |
| 0.1 | 0.530 ± 0.13 | 0.594 ± 0.12 | 0.407 ± 0.14 | 0.469 ± 0.12 |
| 0.2 | 0.628 ± 0.11 | **0.679 ± 0.10** | 0.333 ± 0.07 | 0.481 ± 0.08 |
| 0.3 | 0.593 ± 0.13 | 0.653 ± 0.12 | 0.351 ± 0.13 | 0.472 ± 0.11 |
| 0.4 | **0.640 ± 0.09** | 0.668 ± 0.09 | 0.326 ± 0.07 | 0.483 ± 0.06 |
| 0.5 | 0.627 ± 0.11 | 0.679 ± 0.11 | 0.400 ± 0.11 | 0.514 ± 0.09 |
| 0.6 | 0.616 ± 0.09 | 0.671 ± 0.07 | 0.384 ± 0.07 | 0.500 ± 0.06 |
| 0.7 | 0.618 ± 0.13 | 0.673 ± 0.12 | 0.419 ± 0.11 | 0.519 ± 0.09 |
| 0.8 | 0.599 ± 0.12 | 0.653 ± 0.10 | 0.414 ± 0.06 | 0.507 ± 0.08 |
| 0.9 | 0.625 ± 0.12 | 0.664 ± 0.10 | 0.392 ± 0.12 | 0.508 ± 0.08 |
| 1.0 | 0.591 ± 0.13 | 0.647 ± 0.13 | **0.499 ± 0.08** | **0.545 ± 0.10** |
| 1.1 | 0.623 ± 0.13 | 0.666 ± 0.11 | 0.447 ± 0.10 | 0.535 ± 0.09 |
| 1.2 | 0.599 ± 0.09 | 0.661 ± 0.09 | 0.402 ± 0.10 | 0.501 ± 0.07 |
| 1.3 | 0.607 ± 0.09 | 0.665 ± 0.09 | 0.395 ± 0.11 | 0.501 ± 0.09 |
| 1.4 | 0.624 ± 0.08 | 0.674 ± 0.09 | 0.433 ± 0.08 | 0.528 ± 0.06 |
| 1.5 | 0.596 ± 0.13 | 0.670 ± 0.11 | 0.362 ± 0.14 | 0.479 ± 0.10 |

the quasi-labelled Pit-X dataset, achieved SOTA performance compared to existing step recognition methods, with a notable 5.5% improvement in macro $F_1$-score. To the best of our knowledge, this is the first instance leveraging unannotated data for granular step recognition in eTSA. The consistency loss introduced for spatio-temporal training improved the Edit Score by 13.4%, demonstrating greater step prediction consistency. SurgflowNet presents a significant advancement in eTSA workflow recognition to date, accelerating the development of tools for surgical skill development, improving intra-operative efficiency and supporting post-operative reporting and decision-making.

## CRediT authorship contribution statement

**Anjana Wijekoon:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Adrito Das:** Writing – review & editing, Data curation, Conceptualization. **Zhehua Mao:** Writing – review & editing, Conceptualization. **Danyal Z. Khan:** Writing – review & editing, Data curation. **John G. Hanrahan:** Writing – review & editing, Data curation. **Danail Stoyanov:** Writing – review & editing, Resources, Funding acquisition. **Hani J. Marcus:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Sophia Bano:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. Danail Stoyanov reports a relationship with Medtronic Ltd that includes employment. Danail Stoyanov reports a relationship with Odin Medical Limited that includes board membership, consulting or advisory, and equity or stocks. Danail Stoyanov reports a relationship with Panda Surgical Limited that includes equity or stocks, board membership, and consulting or advisory. Danail Stoyanov reports a relationship with EnAcuity Limited that includes board membership and consulting or advisory. Hani J. Marcus reports a relationship with Panda Surgical Limited that includes board membership, employment and 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 thank Digital Surgery Ltd, a Medtronic company, for access to Touch Surgery Ecosystem for video recording, annotation and storage. This work received funding from the UCL Hawkes Institute, United Kingdom - formerly the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) (203145/Z/16/Z); the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom grant AID-PitSurg: AI-enabled Decision support in Pituitary Surgery (EP/Y01958X/1); the EPSRC, United Kingdom grant CARES: Context aware Augmented Reality for Endoscopic transsphenoidal Surgery (EP/W00805X/1); Adrito Das is supported by EPSRC, United Kingdom [EP/S021612/1]; Danyal Z. Khan and John G. Hanrahan are supported by the National Institute for Health and Care Research (NIHR) Academic Clinical Fellowship; Hani J. Marcus is supported by WEISS, United Kingdom [NS/A000050/1] and by NIHR Biomedical Research Centre at UCL. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

## Appendix A. Hyper-parameter search for $\alpha$

Table A.3 presents the results of the hyper-parameter search for $\alpha$. With a lower $\alpha$ SurgflowNet achieve better precision indicated by higher macro $F_1$-score and balanced accuracy scores. However, increasing $\alpha$ does not improve consistency indicated by lower edit scores when $1.1 \leq \alpha \leq 1.5$. The balance between precision and consistency is achieved at $\alpha = 1$, indicated by the best combined score. Accordingly, we set $\alpha = 1$ for Stage B training of SurgflowNet.

## Appendix B. SurgflowNet performance by surgical step

Fig. B.8 presents the performance of SurgflowNet for each step class as a confusion matrix. Each cell shows the row-normalised proportion of frames with true class i predicted as class j. Values are reported as mean±margin of error (95% confidence interval), computed using the normal approximation to the binomial distribution for each class. Several surgical steps, such as 13 and 14, are often misclassified as step 8 (haemostasis). Row 11 is empty due to step 11 (gasket seal construct) not being present in the test set. Compared to the second-best performing model, ConvNeXt-LSTM [19], SurgflowNet shows significant improvements on steps 1(+16%), 3(+20%), 5(+15%), and 12(+35%); minor gains on steps 8(+5%), 9(+4%), and 10(+1%); and slight decreases on steps −1(−4%), 2(−4%), 4(−4%), 6(−4%), and 7(−6%).

## Appendix C. Comparative study statistical significance test results

Statistical significance testing results (p-values) for the comparative study are included in Table C.4.

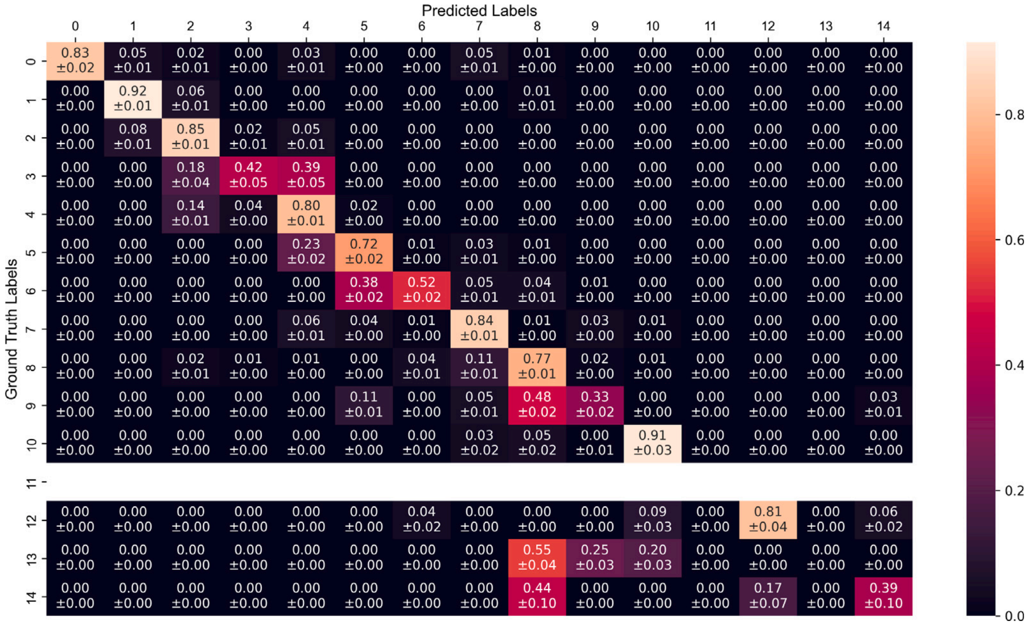


**Fig. B.8.** Normalised confusion matrix with margin of error.

**Table C.4**
Statistical significance testing results. Wilcoxon signed-rank test $p$-value of each method when compared with the best-performing SurgflowNet for statistically significant performance difference; $p < 0.05$ indicates statistical significance with 95% confidence interval; underlined instances are where statistical significance was not achieved.

| | Model | | Metric | | | |
|---|---|---|---|---|---|---|
| | Spatial | Temporal | macro $F_1$-score↑ | B-Accuracy↑ | Edit Score↑ | Combined Score↑ |
| Supervised | ResNet-LSTM-TSF [6] | | 0.0078 | 0.0078 | 0.0156 | 0.0078 |
| | EndoViT-TCN [28] | | 0.0078 | 0.0078 | 0.0078 | 0.0078 |
| | ConvNeXt-LSTM [19] | | 0.2500 | 0.1953 | 0.0078 | 0.0078 |
| Semi-supervised | SRC-MT [37] | LSTM | 0.0156 | 0.0156 | 0.0078 | 0.0078 |
| | ABCL [38] | LSTM | 0.0781 | 0.0547 | 0.0078 | 0.0078 |
| Self-supervised | DINOv2 [32] | LSTM | 0.0156 | 0.0234 | 0.0078 | 0.0078 |
| | EndoViT [28] | LSTM | 0.0078 | 0.0078 | 0.0078 | 0.0078 |

## References

[1] Khan DZ, Hanrahan JG, Baldeweg SE, Dorward NL, Stoyanov D, Marcus HJ. Current and future advances in surgical therapy for pituitary adenoma. Endocr Rev 2023;44(5):947–59.

[2] Marcus HJ, Khan DZ, Borg A, Buchfelder M, Cetas JS, Collins JW, Dorward NL, Fleseriu M, Gurnell M, Javadpour M, Jones PS, Koh CH, Layard Horsfall H, Mamelak AN, Mortini P, Muirhead W, Oyesiku NM, Schwartz TH, Sinha S, Stoyanov D, Syro LV, Tsermoulas G, Williams A, Winder MJ, Zada G, Laws ER. Pituitary society expert delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituitary 2021;24(6):839–53. http://dx.doi.org/10.1007/s11102-021-01162-3.

[3] Khan DZ, Newall N, Koh CH, Das A, Aapan S, Horsfall HL, Baldeweg SE, Bano S, Borg A, Chari A, Dorward NL, Elserius A, Giannis T, Jain A, Stoyanov D, Marcus HJ. Video-based performance analysis in pituitary surgery - part 2: Artificial intelligence assisted surgical coaching. World Neurosurg 2024. http://dx.doi.org/10.1016/j.wneu.2024.07.219.

[4] Khan DZ, Koh CH, Das A, Valetopolou A, Hanrahan JG, Horsfall HL, Baldeweg SE, Bano S, Borg A, Dorward NL, Olukoya O, Stoyanov D, Marcus HJ. Video-based performance analysis in pituitary surgery-part 1: Surgical outcomes. World Neurosurg 2024. http://dx.doi.org/10.1016/j.wneu.2024.07.218.

[5] Khan DZ, Luengo I, Barbarisi S, Addis C, Culshaw L, Dorward NL, Haikka P, Jain A, Kerr K, Koh CH, Layard Horsfall H, Muirhead W, Palmisciano P, Vasey B, Stoyanov D, Marcus HJ. Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0). J Neurosurg 2022;137(1):51–8. http://dx.doi.org/10.3171/2021.6.jns21923.

[6] Das A, Bano S, Vasconcelos F, Khan DZ, Marcus HJ, Stoyanov D. Reducing prediction volatility in the surgical workflow recognition of endoscopic pituitary surgery. Int J Comput Assist Radiol Surg 2022;17(8):1445–52.

[7] Das A, Khan DZ, Hanrahan JG, Marcus HJ, Stoyanov D. Automatic generation of operation notes in endoscopic pituitary surgery videos using workflow recognition. Intelligence-Based Med 2023;8:100107.

[8] Das A, Khan DZ, Psychogyios D, Zhang Y, Hanrahan JG, Vasconcelos F, Pang Y, Wu J, Zou X, Zheng G, Qayyum A, Mazher M, Razzak I, Li T, Ye J, He J, Pł otka S, Kaleta J, Yamlahi A, Jund A, Godau P, Kondo S, Kasai S, Hirasawa K, Rivoir D, Pérez A, Rodriguez S, Arbeláez P, Stoyanov D, Marcus HJ, Bano S. 2024,

[9] Das A, Khan DZ, Hanrahan JG, Bano S, Stoyanov D, Marcus HJ. PitVis-2023 challenge: Endoscopic pituitary surgery videos. 2024, http://dx.doi.org/10.5522/04/26531686.v2.

[10] Mao Z, Das A, Khan DZ, Williams SC, Hanrahan JG, Stoyanov D, Marcus HJ, Bano S. ConsisTNet: a spatio-temporal approach for consistent anatomical localization in endoscopic pituitary surgery. Int J Comput Assist Radiol Surg 2025;1–10.

[11] Wijekoon A, Das A, Herrera RR, Khan DZ, Hanrahan J, Carter E, Luoma V, Stoyanov D, Marcus HJ, Bano S. PitRSDNet: Predicting intra-operative remaining surgery duration in endoscopic pituitary surgery. Heal Technol Lett 2024;11(6):318–26.

[12] Das A, Khan DZ, Hanrahan JG, Marcus HJ, Stoyanov D. Automatic generation of operation notes in endoscopic pituitary surgery videos using workflow recognition. Intelligence-Based Med 2023;8:100107.

[13] Lalys F, Jannin P. Surgical process modelling: a review. Int J Comput Assist Radiol Surg 2014;9(3):495–511.

[14] Demir KC, Schieber H, Weise T, Roth D, May M, Maier A, Yang SH. Deep learning in surgical workflow analysis: a review of phase and step recognition. IEEE J Biomed Health Inform 2023;27(11):5405–17.

[15] Zisimopoulos O, Flouty E, Luengo I, Giataganas P, Nehme J, Chow A, Stoyanov D. Deepphase: surgical phase recognition in cataracts videos. In: Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part IV 11. Springer; 2018, p. 265–72.

[16] Czempiel T, Paschali M, Keicher M, Simson W, Feussner H, Kim ST, Navab N. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: 23rd international conference on medical image computing and computer assisted intervention. Springer; 2020, p. 343–52.

[17] Czempiel T, Paschali M, Ostler D, Kim ST, Busam B, Navab N. Opera: Attention-regularized transformers for surgical phase recognition. In: Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1, 2021, proceedings, part IV 24. Springer; 2021, p. 604–14.

[18] Ding X, Liu Z, Li X. Free lunch for surgical video understanding by distilling self-supervisions. In: International conference on medical image computing and computer-assisted intervention. Springer; 2022, p. 365–75.

[19] Rivoir D, Funke I, Speidel S. On the pitfalls of batch normalization for end-to-end video learning: a study on surgical workflow analysis. Med Image Anal 2024;94:103126.

[20] Yamlahi A, Tran TN, Godau P, Schellenberg M, Michael D, Smidt F-H, Nölke J-H, Adler TJ, Tizabi MD, Nwoye CI, et al. Self-distillation for surgical action recognition. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 637–46.

[21] Pérez A, Rodríguez S, Ayobi N, Aparicio N, Dessevres E, Arbeláez P. Must: Multi-scale t ransformers for surgical phase recognition. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024, p. 422–32.

[22] Funke I, Rivoir D, Krell S, Speidel S. Tunes: A temporal u-net with self-attention for video-based surgical phase recognition. IEEE Trans Biomed Eng 2025.

[23] Gao X, Jin Y, Long Y, Dou Q, Heng P-A. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: Medical image computing and computer assisted intervention–mICCAI 2021: 24th international conference, strasbourg, France, September 27–October 1, 2021, proceedings, part IV 24. Springer; 2021, p. 593–603.

[24] González-Cebrián Á, Bordonaba S, Pascau J, Paredes I, Lagares A, de Toledo P. Attention in surgical phase recognition for endoscopic pituitary surgery: Insights from real-world data. Comput Biol Med 2025;191:110222.

[25] Funke I, Jenke A, Mees ST, Weitz J, Speidel S, Bodenstedt S. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In: International workshop on computer-assisted and robotic endoscopy. Springer; 2018, p. 85–93.

[26] Xia T, Jia F. Against spatial–temporal discrepancy: Contrastive learning-based network for surgical workflow recognition. Int J Comput Assist Radiol Surg 2021;16(5):839–48.

[27] Che C, Wang C, Vercauteren T, Tsoka S, Garcia-Peraza-Herrera LC. Surg-3M: A dataset and foundation model for perception in surgical settings. 2025, arXiv preprint arXiv:2503.19740.

[28] Batić D, Holm F, Özsoy E, Czempiel T, Navab N. EndoViT: pretraining vision transformers on a large collection of endoscopic images. Int J Comput Assist Radiol Surg 2024;19(6):1085–91.

[29] Yu T, Mutter D, Marescaux J, Padoy N. Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. In: International conference on information processing in computer-assisted interventions (IPCAI), rennes, France, juin 2019. 2019.

[30] Zhang J, Barbarisi S, Kadkhodamohammadi A, Stoyanov D, Luengo I. Self-knowledge distillation for surgical phase recognition. Int J Comput Assist Radiol Surg 2024;19(1):61–8.

[31] Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, et al. Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 2020;33:21271–84.

[32] Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, Fernandez P, HAZIZA D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang P-Y, Li S-W, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P. DINOv2: Learning robust visual features without supervision. Trans Mach Learn Res 2024.

[33] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020, p. 1597–607.

[34] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 9729–38.

[35] Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11976–86.

[36] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 10012–22.

[37] Liu Q, Yu L, Luo L, Dou Q, Heng PA. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans Med Imaging 2020;39(11):3429–40.

[38] Huynh T, Nibali A, He Z. Semi-supervised learning for medical image classification using imbalanced training data. Comput Methods Programs Biomed 2022;216:106628.

[39] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv Neural Inf Process Syst 2017;30.

[40] Marcus HJ, Khan DZ, Borg A, Buchfelder M, Cetas JS, Collins JW, Dorward NL, Fleseriu M, Gurnell M, Javadpour M, et al. Pituitary society expert delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituitary 2021;24(6):839–53.

[41] Mao Z, Das A, Islam M, Khan DZ, Williams SC, Hanrahan JG, Borg A, Dorward NL, Clarkson MJ, Stoyanov D, et al. PitSurgRT: real-time localization of critical anatomical structures in endoscopic pituitary surgery. Int J Comput Assist Radiol Surg 2024;1–8.