Computers in Biology and Medicine 191 (2025) 110222

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Attention in surgical phase recognition for endoscopic pituitary surgery: Insights from real-world data

Ángela González-Cebrián [a,*], Sara Bordonaba [b], Javier Pascau [c], Igor Paredes [b,d,e], Alfonso Lagares [b,d,e], Paula de Toledo [a]

[a] Computer Science and Engineering Department, Universidad Carlos III de Madrid, Madrid, Spain
[b] Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Madrid, Spain
[c] Bioengineering Department, Universidad Carlos III de Madrid, Madrid, Spain
[d] Department of Surgery, Faculty of Medicine, Universidad Complutense de Madrid, Madrid, Spain
[e] NeuroSurgery Department, Hospital Universitario 12 de Octubre, Madrid, Spain

ARTICLE INFO

Keywords:
Endoscopic pituitary surgery
Surgical phase recognition
Real-world data
ResNet-50
Spatial features
Attention
SAHC
Temporal features

ABSTRACT

Background and objective: Surgical Phase Recognition systems are used to support the automated documentation of a procedure and to provide the surgical team with real-time feedback, potentially improving surgical outcome and reducing adverse events. The objective of this work is to develop a model for endoscopic pituitary surgery, a challenging procedure for phase recognition due to the high variability in the order of surgical phases.
Methods: A dataset of 69 pituitary endoscopic videos was collected and labelled by two surgeons in seven different phases. The architecture proposed comprises a Convolutional Neural Network to identify spatial features in individual frames, and a Segment Attentive Hierarchical Consistency Network (which combines Temporal Convolutional Networks with attention mechanisms) to learn temporal relationship information between frames and segments at different temporal scales. Finally, predictions are refined with an adaptative mode window.
Results: We have built and made publicly available the largest pituitary endoscopic surgery database to date, named PituPhase. We have built a model with a 73 % accuracy (75 % using a 10 s relaxed boundary). This result is comparable to other state-of-the-art methods in this surgical domain despite the challenges of the dataset (only 10 % of the videos are complete and only 3 % present all phases in the same order, versus 90 % and 50 % respectively in other studies).
Conclusions: Attention mechanisms in combination with Temporal Convolutional Networks and adaptive mode windows improve the performance of Surgical Phase Recognition systems and are robust to missing video sections and high variability in phase order.

1. Introduction

The pituitary gland, or hypophysis, is an endocrine gland located at the skull's base responsible for internal homeostasis [1] which can be affected by intracranial tumours. Specifically, pituitary neuroendocrine tumours account for 15 % of benign intracranial tumours [2]. In the 1990s [3] a new technique emerged for the treatment of these tumours using the endoscope as a source of vision and removing them in a minimally invasive way, so it was no longer necessary to retract the cerebral lobes to reach deep intracranial locations [4]. This surgery has advantages over previous techniques, improving the visualization of deep structures and reducing patients' intraoperative and hospitalisation time. Therefore, endoscopic pituitary surgery is a safe and effective approach widely applied nowadays. Nevertheless, it is a procedure with intricate anatomical aspects where precise visual oversight of every surgical action is essential for achieving favourable outcomes [5]. This underscores the significance of continuously refining the technique [4]. Acquiring a comprehensive understanding of the surgical

Abbreviations: AI, Artificial Intelligence; SPR, Surgical Phase Recognition; CNN, Convolutional Neural Network; TCN, Temporal Convolutional Network; SAHC, Segment Attentive Hierarchical Consistency; ReLU, Rectified Linear Unit; RCDL, Residual Causal Dilated Layer.

* Corresponding author.
E-mail addresses: angegonz@pa.uc3m.es (Á. González-Cebrián), sara.bordonaba.imas12@h12o.es (S. Bordonaba), jpascau@ing.uc3m.es (J. Pascau), igorparedes@gmail.com (I. Paredes), alfonsolagares@pdi.ucm.es (A. Lagares), mtoledo@inf.uc3m.es (P. de Toledo).

https://doi.org/10.1016/j.compbiomed.2025.110222
Received 2 October 2024; Received in revised form 11 April 2025; Accepted 14 April 2025
Available online 17 April 2025

procedure, along with its various phases, is crucial for gaining expertise in technique development.

Since the early 2000s, enhancing efficiency in the healthcare sector has been related to monitoring and preventing adverse events within operating rooms [6]. Applying decision support systems based on Artificial Intelligence (AI) can be a significant aid in improving patient safety. Specifically, the automatic identification of surgical phases, referred to as Surgical Phase Recognition (SPR), has the potential to serve as a critical component of intraoperative decision support systems by enhancing communication among operating room professionals. The aim is to identify the action being performed by the surgeon at any given moment during the operation.

These systems could aid in the automatic indexing of surgical videos providing a valuable educational resource. Likewise, SPR-based systems contribute to establish a standardised surgical protocol, essential for educational purposes and post-operative patient monitoring [7]. Other potential applications include training or evaluating new surgeons, who currently require a long learning curve [8], and selecting relevant frames and video segments to incorporate them in an automatically generated surgery report. Furthermore, optimising operating room utilisation and efficient scheduling of medical teams could be achieved by predicting the time required to complete an intervention [9]. In 2018, a study was published including the design of regressors to predict surgery times while surgery was being performed, which is also closely related to the different phases of the surgical procedure [10].

In the future, implementing SPR in clinical decision support systems will allow to determine which patients can undergo surgery with guaranteed success and will help to mitigate surgical errors by providing early warnings upon detecting anomalies in the technique, thereby lessening the negative impacts of the surgery [11].

In 2012, the first SPR algorithms emerged, aimed at recognising the phases of laparoscopic cholecystectomy and were based on machine learning techniques [12]. These early systems used biomedical data, such as signals from the anaesthesia system or signals of a surgical tool tracking system, which yielded a general statistical description of the surgical process. They were also aided by annotations taken during the operation by a technician, such as the action being performed by the nurse or surgeon at any given moment [13]. Yet, they did not use images and were based on the complete temporal sequence, preventing their application to an online surgical scenario.

Later on, the models were trained with features manually extracted from the video frames. The images were processed, and values such as the gradient in a specific direction, the histogram, or the maximum and minimum intensity were extracted [14]. The classification models were trained through vectors containing those features; however, details that could be important for classification were lost during the extraction process. Furthermore, specifically in videos of endoscopic surgeries, the design and choice of features for classification is a challenge of great difficulty due to the visual complexities of this type of video, as the camera is not static, so there is motion blur and some artefacts (see Fig. 1 (a) and (b)). In addition, the lenses are smeared with blood, blurring the image and opaquing the vision.

More recent works designed models such as EndoNet in 2017 [15] which applied tool recognition for phase detection in laparoscopic cholecystectomy surgeries. EndoNet already used images as a direct source of information for feature extraction through Convolutional Neural Networks (CNN) [15]. Thanks to these, inherent characteristics of the images were learned to perform the classification, avoiding the undesired loss of information. These updated models, besides, incorporate temporal information for classification to obtain the temporal dependencies between frames. The oldest studies used Hidden Markov Models (HMM) for that purpose. However, these are statistical models with pre-defined dependencies that cannot accurately represent the subtle movements of surgical videos [16]. For this reason, other alternatives, like EndoLSTM or MTRCNet-CL [17], which used a Long Short-Term Memory (LSTM), obtained better results [18]. Nevertheless,

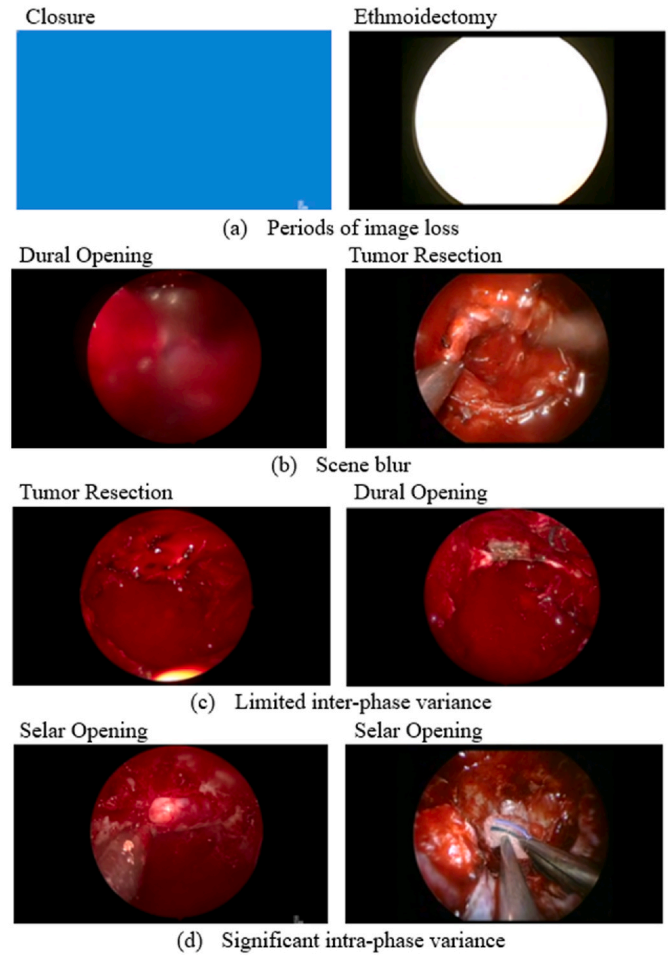

Fig. 1. Challenges in Surgical Phase Recognition in endoscopic pituitary surgery. The text in the upper left corner of each image indicates the phase to which it belongs. In each row, a different difficulty is represented (a) artefacts, (b) blurred images caused by the presence of smoke or fluids on the lens, as well as by the movement of the cameras during operation, (c) low variability between phases, (d) high variability within the same phase.

these models lacked parallelisation capabilities, so they process data sequentially, which makes harder and slows down the training of these networks. In addition, LSTM lacked ability to capture long-term video context, making them suboptimal for SPR, as surgery videos usually last about an hour.

Another approach that has been tested relied on Temporal Convolutional Networks (TCN) [19] with dilated convolutions, an encoder-decoder architecture characterized by its ability to capture both high- and low-level features as well as by its larger receptive field in comparison with LSTM achieving a better performance [7]. This research line led to the emergence of Multi-Stage TCNs [20] which consist of various blocks of stacked TCN in such a way that each of them refines the prediction made by the previous one. This method was used for the first time in TeCNO [7] outperforming various state-of-the-art LSTM approaches. In addition to these neural networks, some studies, like SV-RCNet [16], applied certain constraints based on domain knowledge at the end of the classification called Prior Knowledge Inference (PKI), refining model results. Utilizing some filters, called Temporal Smoothing Functions (TSF), also help to improve model performance in recent works [11], reducing volatility in the predictions.

Nowadays, one of the most advanced fields of artificial intelligence in video understanding are attention networks [21]. These are characterised by the ability to establish long-range temporal relationships, whereas previous methods had more limited memory. Endoscopic pituitary surgery takes more than an hour, so these new methods seem promising. In addition, transformers try to learn, not only frame-level features, but also high-level segment semantics. They have already shown to be effective in laparoscopic cholecystectomy with works such as MS-ASCT [22], SAHC [23] and OperA [24], all of which achieved state of the art performance and even obtained better results than the recurrent neural network-based methods used to date. The same is true for endoscopic pituitary surgery, as seen in the PitVis-2023 Challenge [25], where the top-performing architecture incorporated attention.

Endoscopic pituitary surgery has been the object of only three studies [11,25,26], as opposed to laparoscopic cholecystectomy that has been investigated thoroughly. This is mainly due to the great complexity involved, as it is a longer procedure (Cholec80 videos averaged 34 min compared to more than 1 h in PituPhase dataset) and there is great intra-phase variability accompanied by minute inter-phase variability (see Fig. 1(c) and (d)). Moreover, cholecystectomies are more standardised, so the phases follow a sequential order making it easier to model temporal information while in the case of endoscopic pituitary surgeries, there are different workflows with large variations in the step order complicating the classification [26].

Additionally, endoscopic pituitary surgery presents a unique challenge due to its smaller surgical working field, often resulting in obscured images by instruments or bleeding. Consequently, there's a frequent need to remove the endoscope for cleaning or readjusting the field of vision, leading to brief periods of image loss, as illustrated in Fig. 1. This results in rapid fluctuations in predictions, diminishing clinical confidence in these systems [11]. Usually, frames where the endoscope is placed outside the body are not taken into account during the development of SPR systems, either because these moments are removed from the videos [23] or because they are included as part of other phases [27]. Currently, there is only one public database of endoscopic pituitary surgery published for the PitVis-2023 Challenge [25].

On the other hand, virtually all articles published to date have been conducted using publicly available databases, and most of them with Cholec80, where videos are standardized, presenting much shorter durations than real surgeries and omitting unusual phases that are present in real-world data, such as the included in our work. The study conducted by Kirtac, K et al. [27] demonstrated strong differences between a real-world database and Cholec80, wherein even with 384 videos, compared to the 80 in the public database, similar results could not be achieved with differences of up to 20 % in metrics such as precision and recall. Therefore, we consider studies like ours to be essential for advancing the field of SPR, as working with standardized datasets generate models that are unable to detect minority phases present in real surgeries, thus rendering the application of these systems in a real clinical setting less reliable.

This paper presents the construction and publication of the largest real-world labelled dataset of endoscopic pituitary surgery including a phase to identify when the endoscope is outside the body. Also, a CNN is trained for spatial frame-level feature extraction. Then, for the first-time, attention is combined with TCN's to model temporal information in endoscopic pituitary surgery with a Segment-Attentive Hierarchical Consistency (SAHC) network. Finally, adaptative mode windows were used to refine the classification. Relaxed boundary evaluation metrics were used for this problem achieving the state-of-the-art results and even improving them demonstrating the importance of introducing temporal information into the model. Additionally, this approach will be tested using the PitVis Challenge database, evaluating the generalisation of the network in an external database. To our knowledge, no other research group has trained a deep learning-based model for SPR in neurosurgery including a phase out of the patient, nor has applied attention combined with TCN methods, one of the newest approaches of current research.

2. Methods

2.1. Data

PituPhase-SurgeryAI dataset comprised 69 anonymized surgical procedures, with an average duration of 82 min. Both, complete and incomplete surgeries, were included. Specifically, 21 % of surgeries were incomplete. Pituitary endoscopies were conducted on patients with an average age of 60.24 ± 14.18 years by neurosurgeons of *Hospital 12 de Octubre*. All the patients were diagnosed with pituitary adenoma except two of them, one with meningioma and another one with craniopharyngioma. Every video was acquired via the endoscope at a frequency of 25 frames per second, with each frame maintaining a resolution of 1920×1080 pixels. This study was conducted with the formal approval of the Ethics Committee of *Hospital 12 de Octubre*.

Data were partitioned at patients' level, allocating 80 % for training, i.e., 55 patients and 20 % for testing, i.e., 14 patients. Within the training set of 55 patients, five additional 80-20 partitions were created to generate a fivefold cross-validation dataset, which was utilized to fine-tune the network hyperparameters during the training process.

In order to test our model on an external database, we used the data presented at the PitVis Challenge [25], a set of 25 videos of endoscopic pituitary surgeries with an average duration of 72.8 min labelled in 12 different steps. For more details see https://doi.org/10.5522/04/26531686.v2.

2.2. Surgical phases definition

A phase is a major event occurring during a surgical procedure, composed of several steps [28]. There is a previous work from a consensus group using the Delphi methodologies to identify the main surgical phases in endoscopic pituitary surgery that we have used as a reference [28]. In our PituPhase dataset the decision on the name and number of phases was made by consensus between two neurosurgeons before the labelling task. Initially, they described nine different phases (see Table I) largely corresponding to those described in the referred consensus document [28]. The first two phases (Oto1 and Oto2) were performed by the otolaryngologist, so they were not present in most of the videos and we do not consider them. Nonetheless, they will be taken into account for the recording of future interventions. Therefore, the videos were finally labelled in seven different phases. This includes an "Outside the body" phase (P0), which was defined to group all those frames in which the endoscope was placed outside the patient. Fig. 2 shows an example frame depicting each phase.

Table 1
Phases of transnasal endoscopic surgery.

ID	Name	Start	End
Oto1	Optics Preparation	Recording connected	Enter the nose
Oto2	Middle turbinate resection	Optic insertion, dissection and cutting of the turbinate	Hemostasis of the turbinate
P0	Outside the body	Endoscope exits the body	Endoscope enters the body
P1	Nasoseptal flap preparation	Creation of the flap from the nasal septum	Placement of the flap in the maxillary sinus, choana
P2	Ethmoidectomy and sphenoidal sinus opening	Ethmoid resection; resection of the posterior septum	Sphenoid sinus drilling and removal of its mucosa
P3	Selar opening	Drilling of rostrum and sella turcica	Exposed sella dura
P4	Dural opening	Dura coagulation	End and beginning of resection
P5	Tumour resection	Tumor resection	Hemostasis
P6	Dural closure	Placement of fat/ Duragen	Placement of the nasoseptal flap/ Tisseel

Fig. 2. Representative frame of each of the phases. The text in the top left corner of each image specifies its corresponding phase.

Doctors encountered challenges in pinpointing the precise moment of transition between two distinct phases during the labelling process. This difficulty may have led to transitions between phases being less accurate than desired.

2.3. Labelling strategy

We used Label-studio [29] for labelling the surgeries in a fast and easy way based on the methodology described in Ref. [30]. To reduce labelling costs, an inter-annotator reliability test was conducted to evaluate the concordance between annotators. Both surgeons annotated the same 20 videos and a kappa coefficient was calculated with them [31]. All calculations were performed by treating the annotation of each second of video as an independent observation. After confirming annotation similarity, as will be explained in the results section, each surgeon annotated a subset of the remaining videos.

2.4. Overall Aproach

Our SPR system began by extracting spatial features from individual frames using a ResNet-50. Next, temporal information was extracted by applying the SAHC network, where TCNs were used to extract temporal features at both the frame level and the segment level, and an attention module was employed to establish relationships between frames and segments at different temporal scales. This helped to correct misclassifications caused by ambiguous frames. Finally, temporal filters were used to reduce the volatility of predictions, achieving a smoother result. The entire workflow is summarized in Fig. 3 and it can process 182 frames per second. The following sections explain in detail what each of these blocks consists of.

To compare the results obtained in PituPhase database with SAHC, we also trained a LSTM as a temporal feature extractor. Although this model is simpler, as mentioned in the introduction, it initially yielded good results in SPR for other surgeries, such as cholecystectomy [17], so it is a known baseline against which to compare.

2.5. Spatial Feature Extractor

CNNs are neural networks comprising several structured layers that identify the most representative visual features in their input. The first layers detect the simplest patterns, such as straight lines, curves or corners of the image and, as the network deepens, the layers become more specialized in recognising more complex patterns [32]. In this study, a Residual Network (ResNet) [33] was used to extract spatial features from surgical frames, aligning with the consistency observed in other state-of-the-art SPR studies that also rely on this network architecture as a feature extractor [23]. Moreover, after conducting several experiments, this network yielded the most promising results, therefore this choice not only maintains coherence with prevailing methodologies in the SPR domain but also reflects our empirical findings regarding its efficacy. Different network depths were explored and we finally settled on a ResNet-50, as it gave the best results with the lowest possible computational cost. Number of classes in the last fully-connected layer was reduced to 7. Pre-trained networks were used so that the initialisation of the weights was not random but according to the results obtained on the ImageNet database.

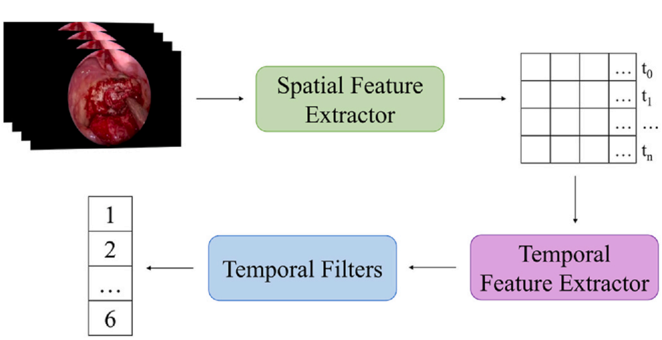

Fig. 3. Diagram of the overall approach. First, spatial features are extracted at the frame level (Spatial Feature Extractor). Each frame corresponds to a specific time point (t_0, t_1, …, t_n). Once the matrix of features is built by stacking the vector from each frame, temporal features are extracted (Temporal Feature Extractor). Finally, temporal filters are applied to smooth the results. This process results in a vector indicating the class to which each frame belongs.

The frames used to train the networks were extracted by sampling the videos at one frame per second with the FFmpeg software [34]. However, a clear class imbalance was observed due to the difference in the duration of the phases (see Fig. 5). To address this problem, we used weighted cross-entropy as the loss function, giving more importance to minority classes. For this purpose, the weights for each class were calculated as the total number of images divided by the number of images in each class. Despite this, the resection phase had a large majority of frames, outnumbering the other classes by more than twice and even four times for the shorter phases. In order to balance the number of frames with the other classes and to reduce the training time of the network, the number of frames in P5 was reduced to a total of 50,000.

To save memory and reduce the number of network parameters the frames were pre-processed. A centre crop and a rescaling were applied, reducing the initial resolution from 1920 × 1080 to 224 × 224, matching the resolution of ImageNet images. Afterwards, the frames were normalized by subtracting the mean pixel value in each channel across the entire database and dividing by the std.

All the Python code was implemented using the PyTorch library [35]. An NVIDIA GeForce RTX 3090 graphic card was used to train the networks. We used Adam optimiser with a learning rate of 5e-7, a weight decay of 0.01 and a step size of 10, lowering the learning rate by a factor of 0.5 every 10 epochs. The batch size was 128, the maximum possible with the available computational resources. The training images within a batch were shuffled, ensuring the presence of images of all phases in each batch. Also, an early stopping criterion was included so the model stopped training when there were three consecutive epochs in which validation losses increase. The model was trained for 40 epochs, taking 20 h to train.

2.6. Segment-Attentive Hierarchical Consistency network

An attentional network combined with TCN was used to capture temporal information directly from spatial features extracted by a backbone CNN. This strategy, presented by Xinpeng Ding in the context of Cholec80 [23], arose from the idea that surgical phases should be classified taking into account information at high-segment level and not at low-frame level. The idea of SAHC was to extract high-level information through video segments and low-level information through frames and use high-level information to refine frame-level classifications solving the ambiguity generated by very similar frames belonging to two different phases.

The combination of TCNs with attention particularly improves the codification of temporal information. TCNs extract temporal features, not only at low-level, but also at high-level, with a larger receptive field than a LSTM. Attention establishes relationships between frame and segment characteristics. Therefore, its combination can correct CNN misclassifications caused by ambiguous frames. This architecture has never been tried before in neurosurgery.

As shown in Fig. 4, originally SAHC comprised four main parts. First, the images were used to train a spatial and temporal feature extractor at frame-level. The spatial features were obtained through a ResNet-50 previously trained in the same way as explained in section 2.5 and resulted in a vector of dimensions N x 2048 where N is the number of frames for each video. In order to obtain 2048 features per frame, the last classification layer in ResNet-50 was removed. These feature vectors were then used to extract temporal features through 11 Stacked Residual Causal Dilated Temporal Convolution Layers (RCDL). As a result, the dimensions of frame-level feature vectors were reduced to N x 64.

Segment-level features were extracted through the Hierarchical Segment-level Feature Extractor. First, the segment was created merging the features at frame level with a Temporal Fusion Layer (TFL). Then, a set of 10 RCDL obtained the temporal characteristics of that segment. This same procedure was repeated two more times to obtain three different temporal scales. At this point, the Attention module learned the relationships between segments and frames in order to refine the consistency of the predictions at frame-level with the segment-frame hierarchical consistency loss. All the code was extracted from Ref. [23]. In the same way that the spatial encoder, the SAHC network was trained using an NVIDIA GeForce RTX 3090 graphic card for 100 epochs over 40 min.

We made some changes to the proposed SAHC [23] diagram (Fig. 4), adapting the network to PituPhase database through a hyperparameter

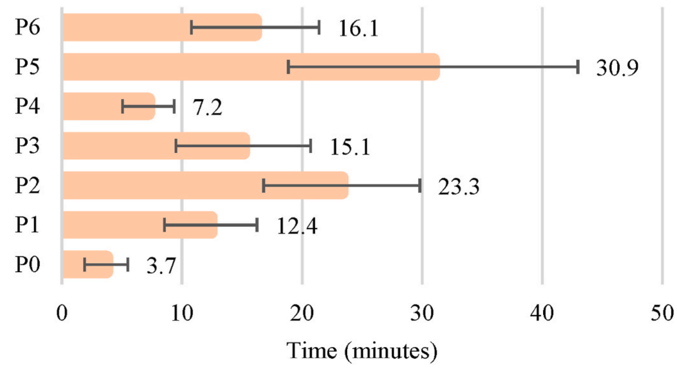

Fig. 5. Duration of the phases. The orange bar indicates the average duration of each phase, which is also specified numerically to the right of each bar. The standard deviation is represented by the error bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

selection performing a grid search using a five-fold cross-validation dataset. First of all, as mentioned before, in our study existed a significant disparity in the number of frames per class as we dealt with real-world data. To address this issue, we opted for weighted cross entropy as the loss function. This choice allowed us to mitigate the imbalance effectively. Notably, we conducted our analysis on the entire dataset without restricting the number of frames for any phase, otherwise such limitations could potentially distort the temporal statistics of the surgical procedure.

Next, we increased the number of RCDL layers in the frame-level temporal feature extractor from 11 to 15, which helped the model to have a wider field of view in terms of frame-to-frame relationship

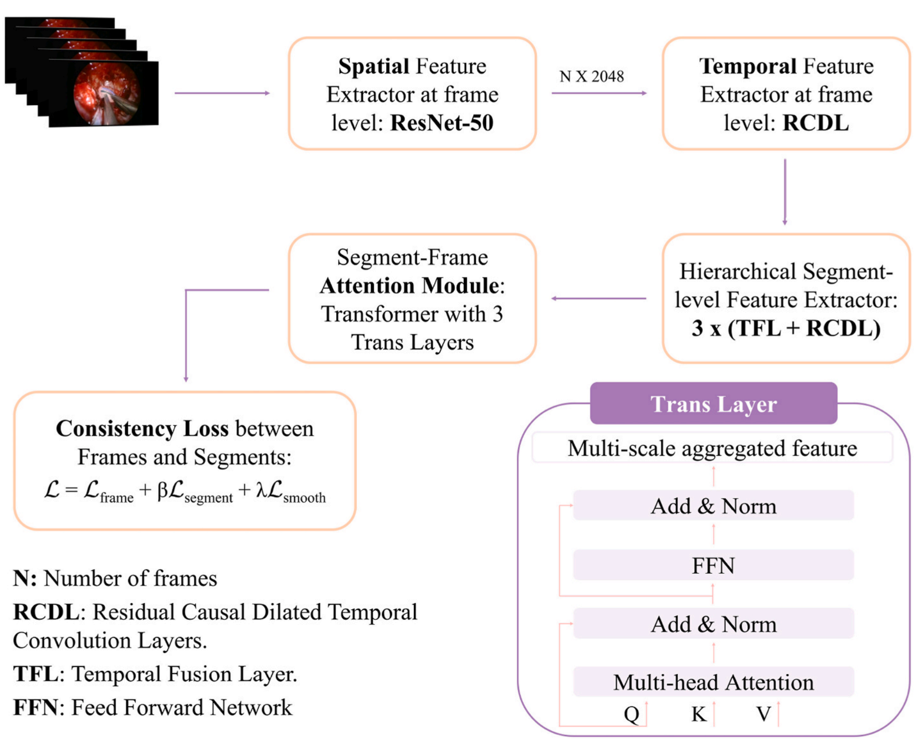

Fig. 4. Segment-Attentive Hierarchical Consistency (SAHC) network implementation. First, spatial features are extracted at the frame level using ResNet-50. Then, RCDL are used to extract temporal features at the frame level. The TFL aggregates features from the frames (with a kernel size of 11), yielding temporal features at the segment level. This process is repeated three times, generating three different temporal scales. Finally, the attention module establishes relationships between the frames (Queries) and the segments (Keys and Values). There is a Trans Layer for each of the different temporal scales. The loss function used for network training combines three terms: cross-entropy loss at the frame level, cross-entropy loss at the segment level and a smoothing term based on the mean squared error of classification probabilities between adjacent frames. β and λ are hyperparameters that determine the weight of each term in the function.

search. The disparity can be explained by the extended duration of PituPhase surgeries compared to the Cholec80 cholecystectomies, which typically endured around 30 min, contrasting with PituPhase surgeries lasting 82 min. Consequently, in PituPhase database frames with greater temporal separation exhibited a heightened correlation, potentially holding significance for classification purposes. In addition, the surgeons' movements must be very precise and were therefore slower, which made them last longer. We decided to set the limit at 15 layers, as there comes a point where considering additional seconds no longer proves beneficial due to the significant amount of noise generated by lens cleaning, blood, and smoke. Including extra frames would inevitably introduce such noise, thereby adversely affecting the model's performance.

Furthermore, reducing the number of heads in the attention block by half resulted in improved performance of the model. This suggests that with 8 heads, there was noise in the data, as having more heads than necessary can lead the model to focus on irrelevant details for classification, which might adversely affect its performance. By reducing the number of heads to 4, the model focused its attention on the most important features of our data.

Moreover, in the decoding layer of the transformer, the dimension of the hidden layer was increased from 512 to 2048, allowing the network to have greater representation capacity to capture more complex and abstract relationships between frames and segments. This was explained by the higher duration of PituPhase surgeries and the increased heterogeneity of workflow in endoscopic pituitary surgery, especially when working with real-world data. Also, this transformer layer transformed and combined information from attention and previous network information. With a larger dimension, there was a richer and more complex flow of information in the network, enhancing the model's learning capacity.

Finally, the decoder layers of the transformer were increased to two, resulting in enhanced representational capacity and improved ability to capture complex patterns in the input data.

2.7. Temporal filters

After obtaining a frame classification based in both spatial and temporal features extracted by Neural Networks (NN). We used a Temporal Filter (TF) as a modal window with an adaptative length extracting the most frequent phase within n images. We decided to implement an adaptive mode window due to the significant variation in phase durations. If we were to use a general mode window, for instance, if the window size is 60 frames, the shorter phases of 60 frames would be adversely affected. The first n images in each surgery were unchanged from the SAHC prediction. In particular, the temporal filter was applied in frames belonging to P1, P2 and P4 using a window of length 31, 96 and 172 respectively reducing the volatility of the predictions. The window utilizes exclusively previous frames to ensure that the system can be employed online, thereby enabling its use during the surgery in real-time.

2.8. Activation maps

Activation maps are commonly employed in current studies to elucidate the decision-making process of CNNs during image classification. Therefore, in this work we used these maps in order to detect which patterns characterise the surgical phases. Grad-CAM, or Gradient-weighted Class Activation Mapping [36], represents a technique for visualizing class activation maps. In this approach, the final convolutional layer of the CNN is examined, together with an analysis of the gradient information flowing into this layer. The gradients represent how important the activation of a pixel is for a particular class. The objective is to identify significant areas crucial for the classification of each class. This is achieved by multiplying the gradients with the activations, revealing the important class pixels. Subsequently, a heatmap is generated to highlight pixels exhibiting high activation, which is overlaid onto the input image for visualization. This process was applied to a randomly selected subset of frames that were accurately classified by the CNN.

2.9. Evaluation metrics

Four metrics were employed to assess the performance of the models: accuracy, recall, precision, and F1-score. Accuracy denotes the percentage of correct detections along the entire surgery. Conversely, recall and precision were computed separately for each phase and subsequently macro-averaged. This evaluation criteria was followed due to the significant variance in the duration of phases, which can conceal misclassifications in shorter phases within accuracy but are exposed by precision and recall. These metrics are commonly used in related work [11,26,27] so we can make a fair comparison.

Besides this, we present metrics with relaxed boundaries, a methodology initially introduced in the m2cai16-workflow challenge and implemented in Ref. [23]. The relaxed boundary metric delineates a 10-s 'relaxed' interval surrounding each ground truth phase transition. Within these intervals, consecutive phases are considered correctly classified. For instance, both phase 4 and phase 5 are deemed accurately classified during the 10 s preceding and succeeding the transition from phase 4 to 5. This approach was deemed pertinent due to its utility in mitigating errors stemming from the challenges articulated by surgeons in accurately delineating transitions between two phases, as mentioned upon in Section 2.2. Consequently, the application of the relaxed boundary yields higher scores across evaluated methodologies [37].

3. Results

3.1. Inter-annotator reliability and agreement

According to reference values of Cohen's kappa published in the study [38], there was a near-perfect level of agreement between both annotators as the kappa coefficient was 0.83. Therefore, there was sufficient reliability to combine their annotations in the same dataset. These results confirmed that a third annotator's opinion was not necessary to obtain a reliable labelling.

Furthermore, the difficulties in the annotation of the transitions between phases were particularly evident in the 'Outside the body' and 'Dural opening' phases, which coincide with the shortest durations, amplifying the effect on the kappa coefficient. This challenge faced by annotators is reflected in Table II, where an almost perfect level of agreement is observed for all phases, except for the aforementioned.

3.2. PituPhase dataset

The PituPhase dataset is available at https://doi.org/10.21950/YDGPZM and is the largest database of endoscopic pituitary surgery for SPR published to date. The surgeries were labelled in seven different phases, whose durations are shown in Fig. 5.

Endoscopic pituitary surgery is a procedure highly contingent upon

Table 2
Inter-annotator agreement, per-phase and overall, across 20 videos assessed by Cohen's kappa.

Phase	κ
Outside the body	0.71
Nasoseptal flap preparation	0.82
Ethmoidectomy and sphenoidal sinus opening	0.89
Selar opening	0.85
Dural opening	0.77
Tumour resection	0.88
Dural closure	0.83
Overall	0.83

the characteristics of the patient's tumour, leading to notable divergence in the sequencing and duration of its phases. For example, the nasoseptal flap preparation (P1) exhibits variability, with instances where it is omitted entirely or executed either at the outset or the conclusion of the surgical process. Moreover, in cases involving notably large tumours that have eroded the Sella turcica and invaded the sphenoid sinus, resection may precede the opening phase, with extreme scenarios even obviating the need for such an opening. As depicted in Fig. 6 and Table III, PituPhase showcases a significant number of potential variations in procedural workflows. It is worth noting that only 3 % of the surgeries followed a consecutive order of phases and that 84.5 % of surgeries were missing some of the phases, either because they were not necessary in the procedure or because the video was incomplete.

3.3. Spatial feature exraction

Two approaches were employed to address the data set's imbalance.

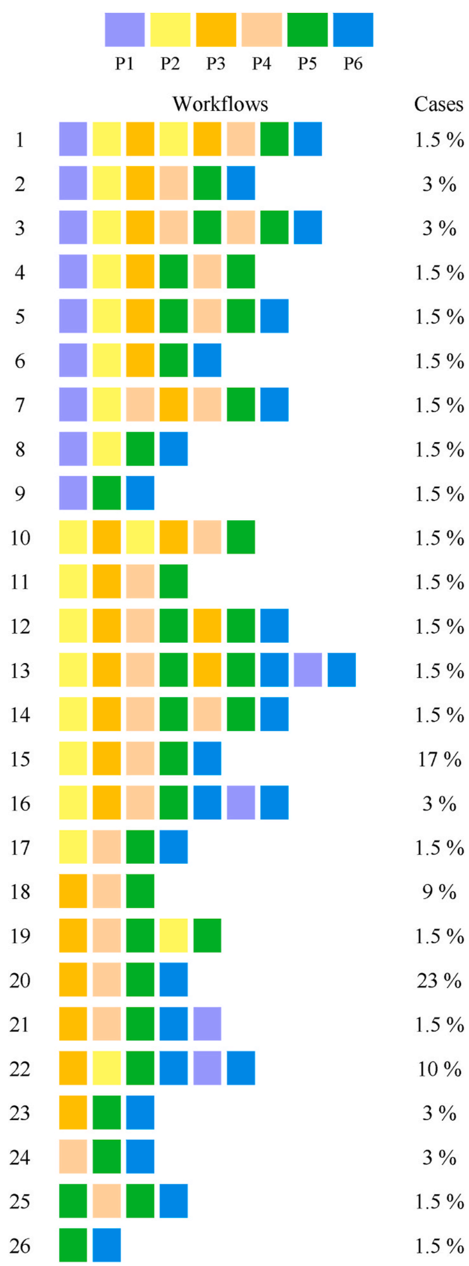

Fig. 6. Phase variations in PituPhase. Cases indicate the percentage of surgeries following that specific workflow.

Table 3
Phase transitions in PituPhase. The numbers indicate the relative (in percent) transitions from one phase (row-axis) to another (column-axis).

Next Phase (%)						
Phase	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**
P1	–	43.75	0	0	0	56.25
P2	7.14	–	82.14	3.57	3.57	3.57
P3	0	3.08	–	86.15	10.77	0
P4	0	0	5.80	–	92.75	1.45
P5	0	0	4.26	14.89	–	80.85
P6	100	0	0	0	0	–

Initially, we implemented weighted cross-entropy as loss function, which enabled enhancement of the model's performance across the underrepresented classes. Notably, the most significant improvement was observed in 'Dural opening', with its recall escalating from 19 % to 30 % in validation data. Furthermore, the impact of reducing the volume of images in the resection phase resulted in enhanced outcomes, thereby diminishing the model's training duration and elevating the recall rates for the nasoseptal flap preparation and ethmoidectomy phases by up to 20 % in validation data. This improvement was achieved while upholding the performance levels of the remaining phases.

Attempts were made to freeze certain network layers. Nevertheless, superior outcomes were attained through the whole network training, prompting us to pursue this approach. We attribute this to the substantial disparity between PituPhase and ImageNet, suggesting that the specific patterns extracted from ImageNet may not be conducive to effective classification in our context.

In summary, ResNet-50 yielded noteworthy results, with an accuracy and a recall rate of 52 %, a precision rate of 55 %, and a F1-score of 53 %.

3.4. Activation maps

Fig. 7 reveals the intricate dynamics of our network, showcasing its strategic emphasis on specific image regions during frame classification. Notably, the classification process predominantly centres around the areas housing the surgical tools, just as humans would do, a trend particularly pronounced in phases characterized by higher recall. Conversely, phases with lower recall, specifically P1, deviate from this conventional pattern. For instance, during the selar opening, the motor tool appearance assumes paramount significance, whereas in the context of nasoseptal flap preparation, the emphasis shifts towards anatomical considerations. Comparable findings were replicated in the case of cholecystectomy surgeries as reported in Ref. [39]. Furthermore, it is noteworthy that the network focused its attention in the spatial positioning of the tool's tip, a pivotal discriminator crucial for discerning among various instruments.

3.5. Temporal semantic extractor

Table IV illustrates that the inclusion of temporal data in the classification resulted in a notable enhancement of performance, with accuracy raising by 71 %. LSTM metrics are much worse than those obtained with SAHC. In addition, through SAHC, all phases were better identified, leading to an increase in the individual recall of each phase as well as in its average, from 52 % to 66 %. This improvement was particularly evident in closure phase, with an increase of 23 % as shown in Fig. 8.

Classification excluding temporal information exhibited substantial noise due to their disordered sequence wherein consecutive frames often corresponded to different phases, as depicted in Fig. 9. However, given that each phase had defined durations, multiple rapid transitions between phases were implausible. Integration of temporal information into the classification model yielded significantly smoother outcomes with reduced noise, thereby mitigating volatility. These results were in line with surgeons' assertions, who mentioned that it is not possible for them to classify individual frames into phases, and they need the

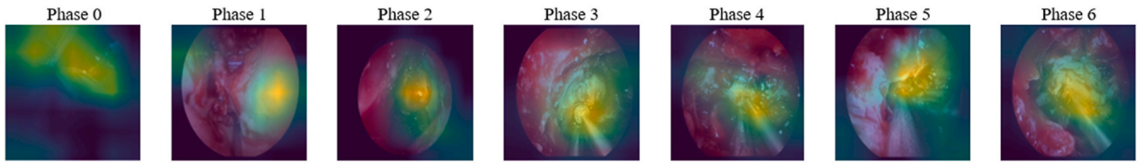

Fig. 7. Gradient activation maps os the ResNet-50. Each frame illustrates which pixels are the most importants to predict that surgical phase.

Table 4
Strict boundary evaluation metrics for comparison of model performance after addition of temporal information in PituPhase. CNN refers to the results obtained after ResNet-50, TF to the results obtained after passing Temporal Filters (adaptative online windows). Each metric is displayed as the mean ± std, between patients for the accuracy and between phases for recall, precision and F1-score.

Model	Accuracy	Recall	Precision	F1-score
CNN	0.52 ± 0.11	0.52 ± 0.22	0.55 ± 0.16	0.53 ± 0.18
LSTM	0.43 ± 0.16	0.20 ± 0.32	0.63 ± 0.27	0.34 ± 0.29
SAHC	0.71 ± 0.11	0.66 ± 0.27	0.71 ± 0.10	0.68 ± 0.15
SAHC + TF	0.73 ± 0.11	0.66 ± 0.28	0.74 ± 0.15	0.70 ± 0.19

temporal contextual information to perform the classification. Model outcomes were further refined through the application of the adaptative mode window described in the methodology, which enabled the correction of subtle network errors, resulting in a 2 % increase in accuracy.

Table V showed the relaxed boundary metrics. Both surgeons differed in annotations for phase transitions in the same videos, with an average discrepancy of ±62 s. For that reason, we proposed the 60 s relaxed boundary as the most appropriate metric to measure performance in our problem dealing with real-world surgeries. Thus, an 81 % of accuracy demonstrated the effective performance of the network to our specific problem, surpassing the results obtained by other studies in endoscopic pituitary surgery.

3.6. SAHC in PitVis challenge dataset

Table VI summarizes the performance of the SAHC method on the dataset provided for the PitVis-2023 Challenge. Similar to what we

Table 5
Relaxed boundary evaluation metrics of 10 and 60 s in PituPhase. Each metric is displayed as the mean ± std, between patients for the accuracy and between phases for recall, precision and F1-score.

Seconds	Accuracy	Recall	Precision	F1-score
10	0.75 ± 0.10	0.70 ± 0.24	0.79 ± 0.11	0.75 ± 0.16
60	0.81 ± 0.08	0.81 ± 0.16	0.88 ± 0.08	0.84 ± 0.10

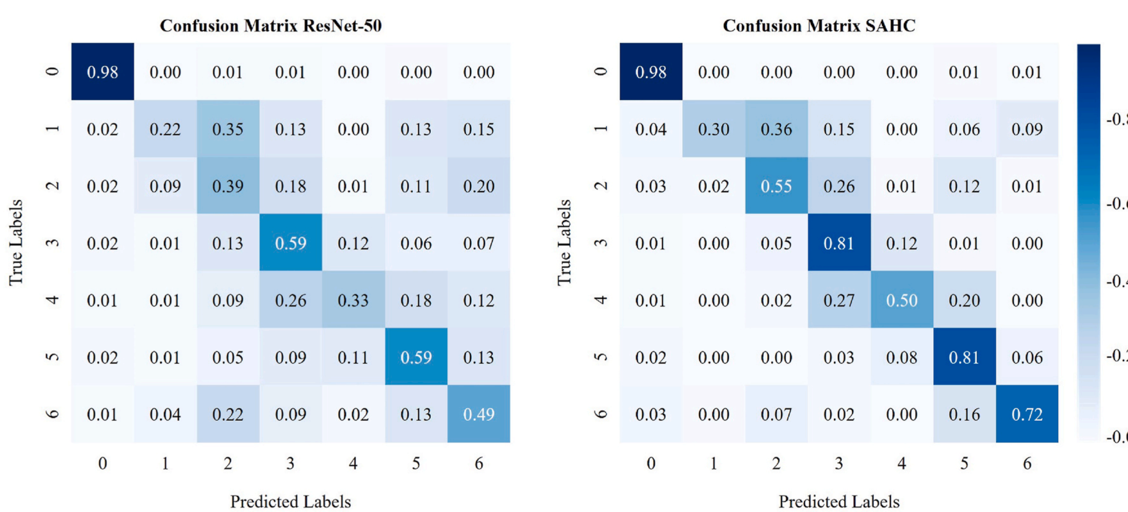

Fig. 8. Normalized confusion matrix for ResNet-50 (left) and SAHC (right) in PituPhase. The X and Y axes represent the predicted labels and ground truth respectively. The element (a,b) of the matrix represents the probability of predicting class a when the ground truth is class b. The probabilities on the diagonal indicate the recall of each phase.

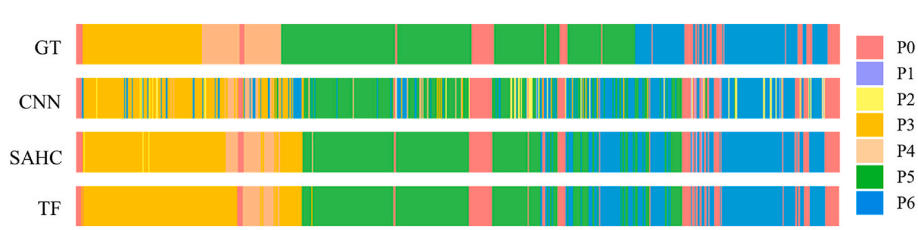

Fig. 9. Color-coded ribbon of phase recognition results vs. ground truth in one patient of the PituPhase test set. The horizontal axis of the ribbon represents the time progression. The top ribbon is the ground truth (GT), the second one corresponds with the CNN results, the third one takes into account the temporal information after SAHC network and the bottom ribbon was obtained after the mode window. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

observed in PituPhase, the incorporation of temporal information proves to be crucial for achieving robust results in PitVis. When comparing our approach to CITI [25], the top-performing method in the Challenge with an F1-score of 61 %, we obtain a comparable performance, highlighting the effectiveness of our strategy on PitVis dataset.

4. Discussion

The ability to automatically analyse and identify the phases of surgical workflows is key for the development of future smart operating rooms. The importance of including spatial and temporal information for effective classification is undisputed. However, the techniques used for achieving a robust classification are still being explored in current research. In this paper we introduce an attentional strategy and show its value for workflow modelling in real-world endoscopic pituitary surgery with improved performance.

In particular, a database is constructed from videos of endoscopic pituitary surgery labelled by two surgeons at the *Hospital 12 de Octubre* standardising the surgery into seven different phases. A near-perfect level of agreement between both annotators is achieved. However, in those phases where there is a major disagreement between both annotators, the model could be less accurate with the predictions. The same happens with phase transitions, where sometimes the discrepancy is longer than a minute. “Outside the body” phase (P0) has the lower kappa coefficient due to it is a short and frequent phase which requires a lot of time to be carefully annotated. However, is the easiest to identify by the model as its spatial features are very different from the rest (all other phases are characterized by red tissue and blood). Consequently, utilizing solely the ResNet-50 architecture, the “outside the body” class is very well defined, with a recall of 98 %, the highest for all phases. This phase exhibits a very variable temporal pattern, as it can appear at many different moments during the surgery, so temporal information is not expected to improve its classification accuracy. It is worth noting, that this high accuracy is preserved when the SAHC attention network is incorporated, with an identical recall rate of 98 %.

Regarding the rest of the phases, the classification of individual frames is much more difficult because, as illustrated in Fig. 1, there are similar frames belonging to different phases, and very different frames belonging to the same phase, presenting an important challenge for model learning. Furthermore, the fact that some tools, e.g., the hoover or the motor, are used in several different phases, and the presence of the same artefacts in images of different phases make the recognition task more difficult.

The attention maps shown in Fig. 7 help us to understand the network behaviour, which relies heavily on tools for the classification of individual frames. This can result in reduced network’s efficacy when tools from different commercial brands (that look different) are used, and can impact outcomes when two instruments are very similar or when the same tool is used in different phases.

The introduction of temporal information in the classification is key to obtain coherent results. Nevertheless, the metrics shown in Table 4 demonstrate that the decision on the type of network to be used to consider such information is extremely important to obtain good results. CNNs, by relying solely on visual features extracted from an individual frame, result in a very noisy outcome with unrealistic phase sequences (see Fig. 9). With LSTM we obtain very poor results because, as previously shown in other studies [7], the field of view of the network is much shorter than in the case of a TCN. Therefore, it is not able to correctly learn the relationships between phases and the results in PituPhase, a highly variable database, are very poor. This demonstrates that we are dealing with an extremely difficult database, as is often characteristic of real-world data. The inclusion of temporal information with SAHC, combining attention with TCNs, increases the CNN model accuracy by 21 %.

Fig. 8, show that, upon temporal information consideration, each class is predominantly confused with its adjacent ones. This observation aligns logically with the notion that the model’s primary challenge lies in precisely discerning transitions between phases. Such difficulty is justified, given that medical professionals themselves often struggle to pinpoint these transitions accurately. Discrepancies in annotations made by different surgeons persist even over extended durations, illustrating the difficulty of the delineation of these boundaries. P6, which has a 23 % increase in recall when SAHC is applied, is the phase with the greatest improvement. This can be attributed to the fact that P6 is characterized by highly marked temporal features, as it always occurs at the end of surgery after resection and is consistently present, unless the video is incomplete and lack that particular segment of the procedure, what happens in 11 surgeries of PituPhase. With a more identifiable pattern, SAHC is capable of correcting the errors made by ResNet-50 in this phase with greater ease. On the other hand, the most challenging phases for the model to identify are P1 and P4, which are the shortest phases (see Fig. 5) and among those that are also more difficult to identify for the surgeons (lower inter-annotator agreement (see Table 2)).

The surgeries included in PituPhase do not always contain a recording of the procedure from beginning to end. Furthermore, pituitary endoscopic surgery is a procedure highly dependent on the characteristics of the patient’s tumour. Thus, there are steps, such as the preparation of the nasoseptal flap, which are not always necessary or might even be performed at different times during the operation. The same applies to openings. It is possible that the patient has a very large tumour which has eroded the Sella turcica and invaded the sphenoid sinus, so that when opening, the surgeon has to remove part of the tumour tissue, which correspond to the resection phase, and then need to open again to properly remove the remaining tumour mass. Therefore, there are many variations of the workflow in this surgery and modelling this information robustly would require many more time series than the 55 recordings available for training and validation in PituPhase. Compared to other real world data state-of-the-art work [27], which used more than 300 surgeries to train and validate the network, PituPhase is small. Although future work involves increasing the number of recordings available, it is important to understand that data scarcity is an inherent feature of research in this field and models that can be trained on small datasets are very valuable, for example to use them in specific surgeries rarely performed.

Our findings support those previously observed in endoscopic pituitary surgical context. However, PituPhase exhibited a significantly higher degree of heterogeneity, with only 10 % of the surgeries being complete, compared to other studies where at least 50 % of surgeries were complete [11,26]. Consequently, we must face real-world database challenges, underscoring the robust performance of the SAHC network in identifying surgical phases in practical settings. Indeed, we achieve same accuracy levels, starting from a ResNet with a 10 % lower accuracy, further evidencing the SAHC’s effectiveness in capturing temporal information.

Results are comparable to those reported in real-world databases for laparoscopic cholecystectomy [27]. Our method achieves a precision of 74 %, whereas a precision of 69 % was reported for cholecystectomy. Despite this, the accuracy of our model is 9 % lower. However, this disparity must be contextualized within the fact that our study comprises only 69 surgeries, whereas cholecystectomy dataset includes over 300 cases. Moreover, pituitary endoscopy typically involves a notably longer duration than laparoscopic cholecystectomy. The average length of our videos is 82 min, whereas the real-world cholecystectomy

Table 6
Evaluation of SAHC on the PitVis dataset (strict boundary metrics).

Model	Accuracy	Recall	Precision	F1-score
CNN	0.54 ± 0.10	0.46 ± 0.21	0.42 ± 0.20	0.44 ± 0.20
SAHC	0.66 ± 0.11	0.60 ± 0.20	0.58 ± 0.20	0.59 ± 0.20
SAHC + TF	0.68 ± 0.09	0.64 ± 0.22	0.63 ± 0.16	0.63 ± 0.18
CITI [25]	–	–	–	0.61 ± 0.11

procedures typically last 54 min [27]. Additionally, laparoscopic cholecystectomy tends to follow more linear flow, rendering the modelling of such information comparatively easier and necessitating fewer videos to yield better results.

Concerning our results on the PitVis dataset with respect to the models trained for the challenge, a direct comparison is not feasible as the test set is not publicly available. Therefore, we have used the validation set as a proxy for evaluation. The PitVis Challenge evaluates performance using both the F1-score and the Edit-score, penalizing noise in the temporal sequence but not significant shifts in phase timing. In PituPhase, the inclusion of the "*outside the body*" phase introduces inherent noise, and transitions between phases may naturally contain fluctuations without disrupting the surgical workflow. Given these characteristics, we consider that the Edit-score metrics does not align with our objectives sufficiently well. Therefore, the other metrics considered in the PitVis Challenge, the F1-score, is used for comparison. Our results improve those of the best performing method, highlighting the robustness of the SAHC approach.

Comparing the results of the proposed method (SAHC + TF) in both the PituPhase dataset and the PitVis dataset, the F1-score is higher in PituPhase. This difference can be attributed to the size and characteristics of each dataset. While PitVis consists of 25 surgeries, PituPhase includes 69, providing a larger and more diverse set of samples for the model to learn from. Additionally, the PitVis dataset has a higher annotation granularity, making phase recognition more challenging for the model.

The use of the 60 s relaxed boundary metric is very well adapted to our objectives, where surgeons differ in their annotations regarding transitions between phases by an average time of ±62 s. This metric accounts for a more realistic definition of the objectives, as the model cannot achieve results that the surgeons themselves are not able to produce. Using this metric, the accuracy of the model is 81 % (see Table 5). With a relaxed boundary of 10 s before and after each transition, our accuracy is 75 %. Other authors [25] suggest using the Edit-score as an evaluation metric. However, we advocate for relaxed boundary metrics, as we believe they better reflect real clinical conditions. Transitions between surgical phases are inherently uncertain, and some degree of variability is expected. The Edit-score excessively penalizes this transition noise, whereas our approach acknowledges it as an intrinsic characteristic rather than a flaw. Furthermore, the Edit-score would not penalize a situation that has an important impact on the use case when a phase transition occurs significantly earlier or later than expected. In contrast, our method imposes a time constraint, limiting the tolerated uncertainty to 10 s in one case and 60 s in another case. This approach was chosen in consensus with surgeons as it provides a more realistic representation of surgical workflow. Therefore, we argue that relaxed boundary metrics are better suited for practical applications.

This study has a number of limitations. First, the small number of surgeries used, not all of which are complete. Thus, we consider that the current data is sufficient for modelling spatial features, but limited for temporal information. On the other hand, the videos used for training have situations (scene blur, …) which may have had a negative impact on the results of the model. Future work to try to better detect these adverse effects or shorter phases developing an ensemble with specialized sub-models will help to improve performance. Another limitation to mention is that the model can be over-fitted to local surgical technique, as both the data and the labels were extracted from the consensus of surgeons at the same institution. We are already working in a multicentre analysis and the introduction of prospective data would help to improve the generalisability of our model, facilitating its integration into the clinical setting.

In summary, this paper presents a surgical phase recognition model in endoscopic pituitary surgery that can classify surgical phases with 73 % accuracy. This result is comparable to other work in less challenging surgical phase recognition tasks. Our model uses attention mechanisms in combination with TCNs for the first time in this specific surgery. Potential applications of the model are plentiful, ranging from automatically labelling existing surgical videos for their use in training, to structured, objective and personalised assessment of specific competencies of novice surgeons. Also, combining the model with Natural Language Processing and generation, surgery reports could be prepared automatically, improving current annotations that are often template-based, losing 50 % of information [40]. Correlation of intraoperative video data with postoperative outcomes could be analysed for complication prediction and early detection of errors or adverse events [41]. Ultimately, an automatic indexing of surgeries could be done to build a databank for educational purposes. As a consequence, safer, more effective and individualised surgery would be achieved, which is why these systems are in high demand and are the future of operating rooms [16].

Our future efforts will focus on increasing and release the available database, as well as testing other possible attention architectures such as OperA [24] which also seem promising for endoscopic pituitary surgery. Furthermore, we aim to conduct a multi-centre study involving surgeries performed at various hospitals, allowing for a more comprehensive analysis of the phases involved creating a generalisable and standardised workflow for endoscopic pituitary surgery. Given the significant disagreement among surgeons in identifying the exact point of transition between surgical phases, and the resulting difficulty in defining a ground truth, it is worthwhile to research different success metrics for both algorithm training and results assessment. This issue is also identified in related work [23]. For the multicentric study we will examine the intra-observer agreement among surgeons and between surgeons and the current algorithm, and we will collaborate with end users to determine the most appropriate definition of ground truth and the best quality metric. Likewise, training a network capable of classifying actions would be ideal for enhancing our understanding of these procedures and providing more opportunities to detect errors. Lastly, integrating multi-label learning into this context holds the potential to provide significant assistance in our objectives.

4.1. Conclusion

This research presents a new database of endoscopic pituitary surgeries for surgical phase recognition and automatic analysis of the surgical workflow named PituPhase. For the first time in this surgery, an attention model together with TCNs was trained on this dataset obtaining a very good performance and facing the real-world data challenges with 74 % precision, even if the recording doesn't include the complete surgery. Furthermore, with metrics adapted to the specifications of our problem, an accuracy of 81 % was achieved. As a novelty, the model detected when the endoscope was placed outside the body with a 98 % recall. Future work involves adding new recordings to obtain a more robust classification. This work has a wide range of applications, such as automatically documenting the procedure and updating the surgical team in real on the stage of the surgery.

CRediT authorship contribution statement

Ángela González-Cebrián: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Sara Bordonaba:** Software, Investigation. **Javier Pascau:** Writing – review & editing, Methodology, Conceptualization. **Igor Paredes:** Writing – review & editing, Validation, Methodology, Investigation, Data curation, Conceptualization. **Alfonso Lagares:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Paula de Toledo:** Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Ethics approval

Comité de Ética de la Investigación del HOSPITAL UNIVERSITARIO 12 DE OCTUBRE (N. ° 23/037) signed on date 2023/03/16. Besides, inform consent was obtained from all patients to include their surgical videos in the publicly available dataset.

Funding

This work is partially funded by Grant TED2021-130944B-C21 by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR".

Funding for APC

Universidad Carlos III de Madrid (Agreement CRUE-Madroño 2025).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.C. Pérez-López, M.J. Abenza-Abildúa, Historia de la cirugía de la hipófisis, Neurosci. History 8 (1) (2020) 29–38 [Online]. Available: https://nah.sen.es/es/178-revistas/volumen-8/numero-1/452-historia-de-la-cirugia-de-la-hipofisis. (Accessed 2 March 2023).

[2] J. Torales, et al., Cirugía endoscópica endonasal en tumores de hipófisis. Resultados en una serie de 121 casos operados en un mismo centro y por un mismo neurocirujano, Endocrinol. Nutr. 61 (8) (Oct. 2014) 410–416, https://doi.org/10.1016/J.ENDONU.2014.03.011.

[3] M. Henríquez, E. Monnier B, E. Ortiz P, L. Nicklas, S. Henríquez, V. Resumen, Cirugía hipofisiaria endoscópica transesfenoidal, con realización de colgajo nasoseptal: Evaluación del impacto de la técnica en la olfación. Serie de casos, Revista de otorrinolaringología y cirugía de cabeza y cuello 77 (1) (Mar. 2017) 27–34, https://doi.org/10.4067/S0718-48162017000100004.

[4] H. Pablo Rojas, P. Horacio José, R. Rafael Herrera, J. Luis Ledesma, E. Rubín, L. Agustín Ramírez Stieben, Transnasal endoscopic skull base surgery: analysis of complications in the first 120 procedures, Surg. Neurol. Int. 2022 (523) (2022), https://doi.org/10.25259/SNI_980_2022.

[5] R. Torino, Surgical Neurology International SNI: Revista Argentina de Neurocirugia, a supplement to SNI OPEN ACCESS, 2012, https://doi.org/10.4103/2152-7806.104403.

[6] F. Miyawaki, K. Masamune, S. Suzuki, K. Yoshimitsu, J. Vain, Scrub nurse robot system - intraoperative motion analysis of a scrub nurse and timed-automata-based model for surgery, IEEE Trans. Ind. Electron. 52 (5) (Oct. 2005) 1227–1235, https://doi.org/10.1109/TIE.2005.855692.

[7] T. Czempiel, et al., TeCNO: surgical phase recognition with multi-stage temporal convolutional networks, Lect. Notes Comput. Sci. 12263 (LNCS) (Mar. 2020) 343–352, https://doi.org/10.1007/978-3-030-59716-0_33.

[8] M. Takeuchi, et al., Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence, Ann. Surg Oncol. 29 (1996), https://doi.org/10.1245/s10434-022-11996-1.

[9] T. Golany, et al., Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy, Surg. Endosc. 36 (12) (Dec. 2022) 9215–9223, https://doi.org/10.1007/S00464-022-09405-5/FIGURES/4.

[10] G. Yengera, J. Marescaux, and N. Padoy, "Less is more: surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks". arXiv preprint arXiv:1805.08569.

[11] A. Das, S. Bano, F. Vasconcelos, D.Z. Khan, H.J. Marcus, D. Stoyanov, Reducing prediction volatility in the surgical workflow recognition of endoscopic pituitary surgery, Int. J. Comput. Assist. Radiol. Surg. 17 (8) (Aug. 2022) 1445–1452, https://doi.org/10.1007/S11548-022-02599-Y.

[12] N. Padoy, T. Blum, S.A. Ahmadi, H. Feussner, M.O. Berger, N. Navab, Statistical modeling and recognition of surgical workflow, Med. Image Anal. 16 (3) (Apr. 2012) 632–641, https://doi.org/10.1016/J.MEDIA.2010.10.001.

[13] G. Forestier, et al., Multi-site study of surgical practice in neurosurgery based on surgical process models, J. Biomed. Inf. 46 (5) (Oct. 2013) 822–829, https://doi.org/10.1016/J.JBI.2013.06.006.

[14] U. Klank, et al., Automatic feature generation in endoscopic images, Int. J. CARS 3 (2008) 331–339, https://doi.org/10.1007/s11548-008-0223-8.

[15] A.P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, N. Padoy, EndoNet: a deep architecture for recognition tasks on laparoscopic videos, IEEE Trans. Med. Imag. 36 (1) (Jan. 2017) 86–97, https://doi.org/10.1109/TMI.2016.2593957.

[16] Y. Jin, et al., SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network, IEEE Trans. Med. Imag. 37 (5) (May 2018) 1114–1126, https://doi.org/10.1109/TMI.2017.2787657.

[17] Y. Jin, et al., Multi-task recurrent convolutional network with correlation loss for surgical video analysis, Med. Image Anal. 59 (Jan. 2020) 101572, https://doi.org/10.1016/J.MEDIA.2019.101572.

[18] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (Nov. 1997) 1735–1780, https://doi.org/10.1162/NECO.1997.9.8.1735.

[19] C. Lea, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks: a unified approach to action segmentation, Lect. Notes Comput. Sci. 9915 (LNCS) (2016) 47–54, https://doi.org/10.1007/978-3-319-49409-8_7/TABLES/1.

[20] Y.A. Farha, J. Gall, MS-TCN: multi-stage temporal convolutional network for action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Mar. 2019, pp. 3575–3584. Available: https://arxiv.org/abs/1903.01945v2.

[21] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: Proceedings of the 38 Th International Conference on Machine Learning, vol. 139 PMLR, 2021. Available: https://arxiv.org/abs/2102.05095v4.

[22] B. Zhang, J. Meng, B. Cheng, D. Biskup, S. Petculescu, and A. Chapman, "Friends Across Time: Multi-Scale Action Segmentation Transformer for Surgical Phase Recognition," In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-6) IEEE. Available: https://arxiv.org/abs/2401.11644v1.

[23] X. Ding, X. Li, Exploring segment-level semantics for online phase recognition from surgical videos, IEEE Trans. Med. Imag. 41 (11) (Nov. 2021) 3309–3319, https://doi.org/10.1109/TMI.2022.3182995.

[24] T. Czempiel, M. Paschali, D. Ostler, S.T. Kim, B. Busam, N. Navab, OperA: attention-regularized transformers for surgical phase recognition, Lect. Notes Comput. Sci. 12904 (LNCS) (Mar. 2021) 604–614, https://doi.org/10.1007/978-3-030-87202-1_58.

[25] A. Das, et al., PitVis-2023 Challenge: Workflow Recognition in Videos of Endoscopic Pituitary Surgery, Sep. 2024, https://doi.org/10.5522/04/26531686 arXiv preprint arXiv:2409.01184.

[26] D.Z. Khan, et al., Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0), J. Neurosurg. 137 (1) (Nov. 2021) 51–58, https://doi.org/10.3171/2021.6.JNS21923.

[27] K. Kirtac, et al., Surgical phase recognition: from public datasets to real-world data, Appl. Sci. 12 (17) (Sep. 2022) 8746, https://doi.org/10.3390/APP12178746/S1.

[28] H.J. Marcus, et al., Pituitary society expert Delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection, Pituitary 24 (6) (Dec. 2021) 839–853, https://doi.org/10.1007/S11102-021-01162-3.

[29] Open source data labeling | label studio. https://labelstud.io/. (Accessed 14 August 2023).

[30] E. Fischer, et al., A methodology for the annotation of surgical videos for supervised machine learning applications, Int. J. Comput. Assist. Radiol. Surg. 18 (2023) 1673–1678, https://doi.org/10.1007/s11548-023-02923-0.

[31] J.L. Cerda, L.P. Villarroel Del, Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa, Rev. Chil. Pediatr. 79 (1) (2008) 54–58, https://doi.org/10.4067/S0370-41062008000100008.

[32] K. O'shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. 2016 *IEEE Conference On Computer Vision And Pattern Recognition* (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[34] FFmpeg. https://ffmpeg.org/. (Accessed 30 August 2023).

[35] J. Ansel, et al., PyTorch 2: faster machine learning through dynamic Python bytecode transformation and graph compilation [conference paper], in: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, vol. 2, 2024, https://doi.org/10.1145/3620665.3640366 (ASPLOS '24).

[36] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359, https://doi.org/10.48550/arXiv.1610.02391.

[37] S. Ramesh, et al., Dissecting Self-Supervised Learning Methods for Surgical Computer Vision, Jul. 2022, p. 1, https://doi.org/10.48550/arxiv.2207.00449.

[38] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1) (Mar. 1977) 159, https://doi.org/10.2307/2529310.

[39] H.H. Yeh, A.M. Jain, O. Fox, S.Y. Wang, PhacoTrainer: a multicenter study of deep learning for activity recognition in cataract surgical videos, Transl. Vis Sci. Technol. 10 (13) (Nov. 2021), https://doi.org/10.1167/TVST.10.13.23.

[40] S. Wu, et al., Deep learning in clinical natural language processing: a methodical review, J. Am. Med. Inf. Assoc. 27 (3) (Mar. 2020) 457–470, https://doi.org/10.1093/JAMIA/OCZ200.

[41] E.M. Bonrath, L.E. Gordon, T.P. Grantcharov, Characterising 'near miss' events in complex laparoscopic surgery through video analysis, BMJ Qual. Saf. 24 (8) (Aug. 2015) 516–521, https://doi.org/10.1136/BMJQS-2014-003816.