A Multi Instance Learning Approach for Critical View of Safety Detection in Laparoscopic Cholecystectomy

Yariv Colbeci YARIV@THEATOR.IO
Maya Zohar MAYA@THEATOR.IO
Daniel Neimark DANIELN@THEATOR.IO
Dotan Asselmann DOTAN@THEATOR.IO
Omri Bar OMRI@THEATOR.IO
Theator Inc., Palo Alto, CA, USA.

Abstract

Surgical procedures have a clear designated goal, which makes the art of performing surgery a task-oriented action. The performing surgeon follows specific workflow steps that describe the actions needed to reach the surgery goal. In *ectomy* procedures, such as Cholecystectomy and Appendectomy, the goal is to dissect and remove a specific organ. Safety measures are set to prevent injuries, and the surgeon needs to follow protective methods to avoid misidentification. In Laparoscopic Cholecystectomy (LC), this method is known as *Critical View of Safety* (CVS). This work illustrates that machine learning can detect CVS accurately enough to be used routinely in the clinical setting, both for educational purposes and in other assessment scenarios. We formulate CVS detection as a supervised Multi Instance Learning (MIL) problem and propose an attention-based MIL model that is trained and evaluated on more than 2,000 surgical videos. It achieves 82.6% mean unweighted accuracy in detecting LC CVS criteria and 84.2% accuracy in the final task of CVS detection.

1. Introduction

Laparoscopic surgeries are minimally invasive surgeries performed in the abdominal or pelvic cavities with the aid of a video feed coming from an endoscope. Laparoscopic Cholecystectomy (LC) is a common procedure performed in this approach, with nearly one million procedures performed annually in the United States alone (Pucher et al., 2018).

Despite the substantial benefits of laparoscopy, such as reduced pain and a shorter hospitalization period (Fuchs, 2002), one drawback of this technique is that since a live video feed is used to provide the surgeon with the field of view, it also limits the visual depth and affects the visual perception.

A severe intraoperative complication of LC, mostly happens due to visual perceptual illusion or misidentification, is Bile Duct Injury (BDI), occurring in about 0.4% of the cases and increasing mortality rate by about 2% (Way et al., 2003; Pucher et al., 2018).

Critical View of Safety (CVS) is a method first introduced in 1995 (Strasberg et al., 1995) and adopted by surgeons to safely perform LC procedures and avoid misidentification of the biliary tracts (Nassar et al., 2021). In order to correctly identify the bile duct structures before division, the surgeon must obtain a clear view of the area and make sure three mandatory criteria are met (Fig. 1). First, the hepatocystic triangle must be cleared

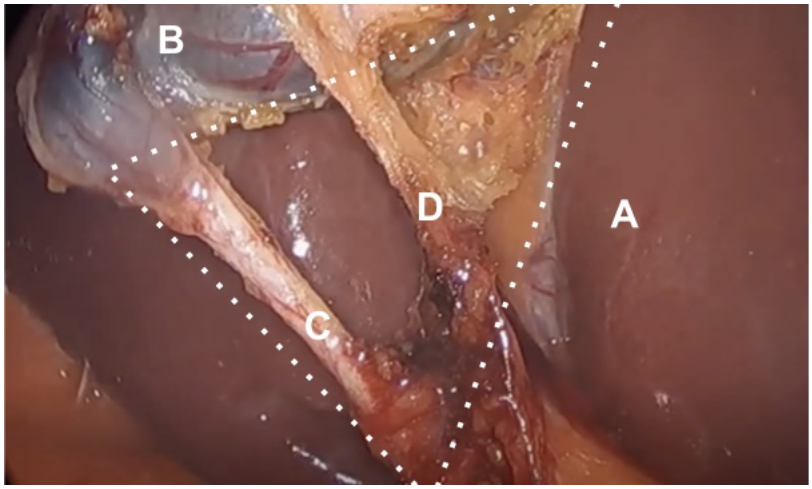

Figure 1: An example of a complete Critical View of Safety when all criteria were achieved. **A.** Liver. **B.** Gallbladder. **C.** Cystic Duct - the left structure. **D.** Cystic Artery - the right structure. The white dashed line represents the hepatocystic triangle that should be cleared of fat and tissues (first criterion). The lower part of the gallbladder is separated from the liver (second criterion), and only two structures enter the gallbladder (third criterion).

of fat and fibrous tissue (clean triangle). Second, the lowest part of the gallbladder should be separated from the liver (gallbladder separation). And third, only two structures should be seen entering the gallbladder (two structures). Correctly performed, CVS execution significantly lowers the incidence of BDI and intraoperative complications (Sgaramella et al., 2021; Strasberg and Brunt, 2010).

Novel AI techniques, such as Computer Vision algorithms, can improve surgical practice and patient outcomes (Maier-Hein et al., 2017). Integrating automatic video analysis methods into the routine work of surgeons can contribute significantly to educational purposes, such as preoperative preparation and postoperative insights and debriefing.

Furthermore, real-time detection of intraoperative events can support surgeons' decision-making during critical phases of the operation and provide better anatomical understanding and orientation. In addition, reliable agnostic models trained on a variety of procedures are unbiased to the different perspectives of surgeons. Thus, using these models for real-time decision support can result in better outcomes and reduce the variability between surgeons.

CVS is an important milestone in LC, making it a valuable task for automation using machine learning. Supporting the surgeon workflow with automated detection of CVS in the future would improve surgical practice and help assess CVS achievement, potentially resulting in fewer mistakes or misidentifications that might lead to bile duct injuries.

In this work, we suggest a new method for CVS detection. We formulate the problem as an explicit criteria classification task to infer CVS achievement. We set heuristic rules based on surgical steps annotations for time frame localization and use an attention-based Multi Instance Learning (MIL) model with a multi-label classifier for criteria detection.

We train and test our method on a large-scale database of more than 2,000 annotated LC videos. This study aimed to develop a method that can provide accurate detection of the CVS time frame in different scenarios and reliable predictions of each CVS criterion.

Generalizable Insights about Machine Learning in the Context of Healthcare

This study focuses on the challenging task of identifying the Critical View of Safety in Laparoscopic Cholecystectomy with machine learning. Although CVS has a significant clinical value for surgery and is done routinely in the operating room (OR), it still lacks data and studies exploring how to identify its criteria automatically. Using machine learning for CVS detection is an example of an important key moment or milestone that has significant value to the clinical community but is limited by data and has a complex labeling process. Thus, limiting the ability to develop and explore ML-based methods. In this work, we propose a new method for CVS detection. Our approach builds on top of the latest developments in computer vision and applies attention-based networks that significantly improve previous studies' results. We train and evaluate our approach on a large-scale dataset and propose a relatively simple method that does not require a complex labeling process. We evaluate our method on surgical videos curated from several medical centers, which supports our approach's generalization and robustness. Finally, ML-based applications in the medical field are often referred to as black boxes, making it difficult for physicians to understand and interpret the results. In contrast, our method uses an attention-based MIL, which allows us to illustrate the specific frames the model focuses on during the prediction process (Fig. 4 and Appendix). Therefore significantly add to the interpretability of the model.

2. Related Work

Extensive research is being conducted in the field of surgical video analysis (Bar et al., 2020; Jin et al., 2018; Yu et al., 2019; Zohar et al., 2020). Previous studies demonstrated the use of computer vision techniques for the foundational task of surgical step recognition, i.e., parsing a surgical video into key segments that represent the surgeon's workflow (Neimark et al., 2021b; Hashimoto et al., 2019; Twinanda et al., 2016). These studies lay the foundations for more complex tasks like CVS detection, which relies on pre-detecting the specific step the surgeon is focusing on.

A recent work by Mascagni et al. (2021) introduced a multi-label classification method for LC CVS criteria detection in static images captured from videos when CVS was achieved. Their approach requires accurate anatomical segmentation of the hepatocystic anatomy as input and manually selected frames that include CVS criteria. The results reported set a baseline for CVS detection, demonstrating the ability to accurately detect CVS in images with mean unweighted accuracy of 71.4%. However, this approach will be hard to scale and use in production scenarios where segmentation and specific frames are absent.

Our method relies on transfer learning which is widely used in image and video recognition tasks (Girshick et al., 2014; Carreira and Zisserman, 2017; Yosinski et al., 2014). Recent studies also applied transfer learning in the surgical domain. They demonstrated the effectiveness of taking pre-trained models, trained on one procedure type, and using them to fine-tune on different procedure types (Neimark et al., 2021b). Here, we suggest

that feature representations extracted by a model pre-trained on surgical step recognition can be successfully applied to other surgical-related tasks, such as CVS detection.

A multi instance learning approach has been used before for several medical-related problems. In Sadafi et al. (2020), MIL is used to classify blood cell disorders and helps with the complicated task of manually labeling all cells in an image. The work of Pal et al. (2021) developed a new sparse attention-based MIL algorithm for abnormal cell detection in cervical histopathology images. A recent work introduced DEEMD, a computational pipeline using deep neural networks within a multi instance learning framework. DEEMD uses MIL to identify putative treatments effective against SARS-CoV-2 based on morphological analysis in images (Saberian et al., 2021). Another example of using MIL in the medical field is DeepLION, which shows accurate cancer detection by modeling the T cell receptors correlation and using MIL to assign an adjusted weight for each receptor during the prediction process (Xu et al., 2022).

3. Cohort

3.1. Dataset

Our dataset contains 2,223 videos of LC, curated from 12 different medical centers (see Table 6 in the Appendix for demographics information). Each video undergoes a meticulous annotation and validation process by a team of specialists trained to annotate the surgical workflow. The annotation process was validated by a group of surgeons to ensure high agreement with our annotation (Korndorffer Jr et al., 2020). The dataset was manually annotated with surgical steps and CVS, including time frame and CVS criteria (Fig. 1).

The dataset was randomly divided into three subsets (Table 3.1); 25% (556) of the videos for the test set, 20% (333) of the remaining for the validation set, and the rest for the training set (1334).

Table 1: CVS dataset. Each column presents the number of positive and negative samples of each criterion and its related percentage. The right column shows the number of total CVS samples.

	Clean Triangle		Gallbladder Separation		Two Structures		CVS	
	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos
Training	1122 (67%)	561 (33%)	1009 (60%)	674 (40%)	1377 (82%)	306 (18%)	1458 (87%)	225 (13%)
Validation	305 (70%)	136 (30%)	281 (64%)	160 (36%)	366 (83%)	75 (17%)	384 (87%)	57 (13%)
Test	365 (57%)	279 (43%)	340 (53%)	304 (47%)	467 (73%)	177 (27%)	505 (78%)	139 (22%)

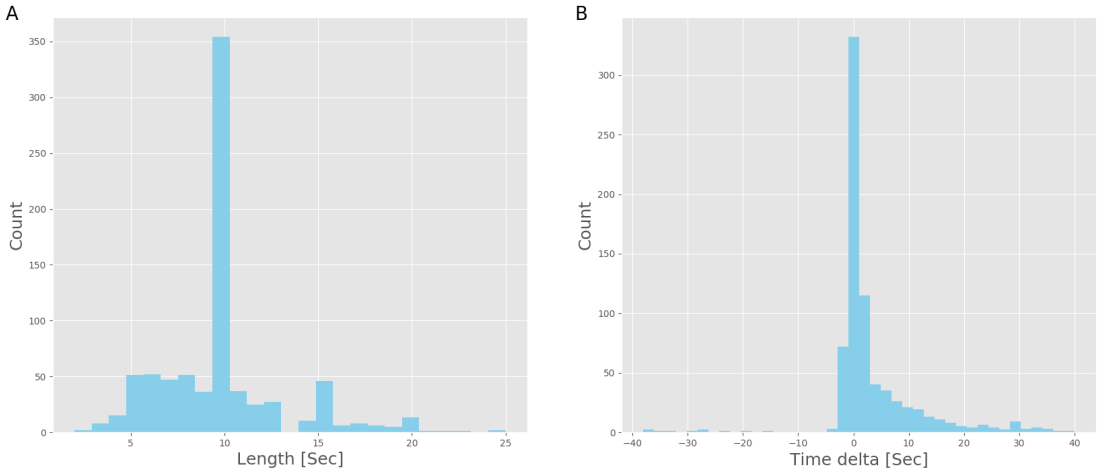

Figure 2: **A.** Distribution of CVS segment length. **B.** Distribution of the time delta between cystic structures division and CVS end time.

4. Methods

4.1. Problem formulation

We mapped out CVS detection to three separate sub-tasks: (1) Time frame localization, (2) Binary classification of CVS (i.e., achieved or not achieved), and (3) Multi-label classification of the three criteria.

4.2. Implementation details

We define heuristic rules to tackle the localization problem, derived from the structured procedure workflow and based on statistics drawn from our dataset. A correct CVS appears right before the division of the cystic structures (Strasberg and Brunt, 2010). We used surgical steps annotations to locate the division step start time and then extracted a clip that ends when the division starts. As CVS mostly appears in a scope of 30 seconds before the division (Fig. 2B), with an average duration of 10 seconds (Fig. 2A), we chose these values as our region of interest.

For the criteria classification task, we used a similar architecture to the one suggested in VTN (Neimark et al., 2021a). With ViT (Dosovitskiy et al., 2020) as a feature extractor and replacing the Longformer with a multi-label classifier that detects all three criteria simultaneously. We initialized the network with pre-trained model weights, trained on the LC steps recognition task (Bar et al., 2020). To discern whether CVS was achieved or not, we first need to determine which criteria were achieved. Thus, we focused on exploring different methods for CVS criteria detection.

The interesting work of Mascagni et al. (2021) shows promising results for CVS detection. However, it requires that the dataset be annotated with anatomical segmentation and relevant frames for CVS detection. This is not feasible for a large-scale dataset like the one

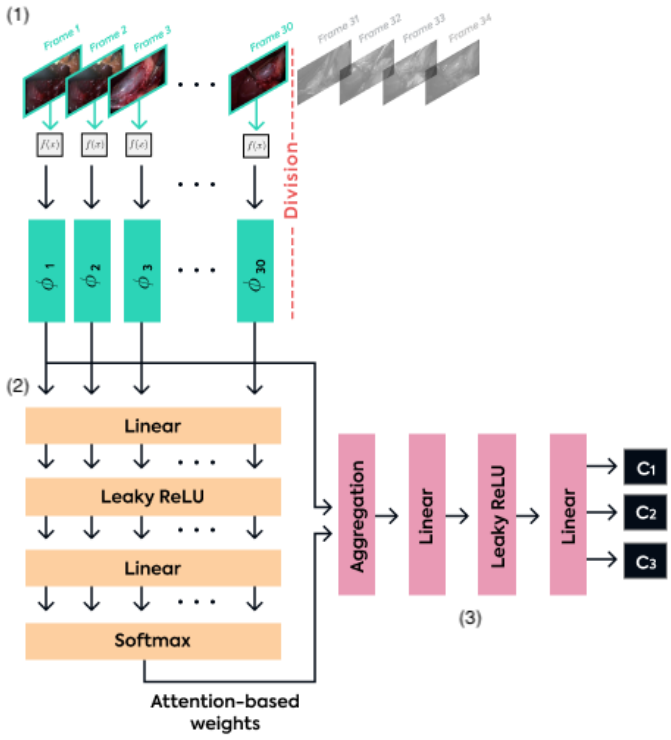

Figure 3: High-level description of the attention-based MIL architecture for CVS criteria detection. The network includes three modules: (1) ViT backbone, (2) An attention block, and (3) a multi-label classification head.

described in our work. In the absence of other CVS detection methods to compare to, and in order to allow a fair comparison between our suggested approach and other ML-based methods, we chose to include different types of models, such as MLP and LSTM, as a baseline.

Multilayer Perceptron (MLP). We used a small MLP network consisting of two linear layers as a baseline. The first layer reduces the dimension of each feature vector to 256, followed by flattening the temporal dimension into a single dimension. The second layer is a 3 class classification layer. The learning rate was set to 0.0005.

Long Short-Term Memory (LSTM). We used an LSTM network (Goodfellow et al., 2016) with a hidden dimension of size 256. We tried two aggregation functions applied to the output; *mean*, and *take last*, with a learning rate of 0.2 and 0.002, respectively. The aggregation function reduces the temporal dimension to one vector of size 256, which is the input to the classification head. *mean* refers to the average vector across the temporal axis, and *take last* is the vector representing the last second of the sequence.

Multi Instance Learning (MIL). This method refers to the case in which a single label is assigned to a bag of instances. Meaning, if there is at least one positive instance in the bag,

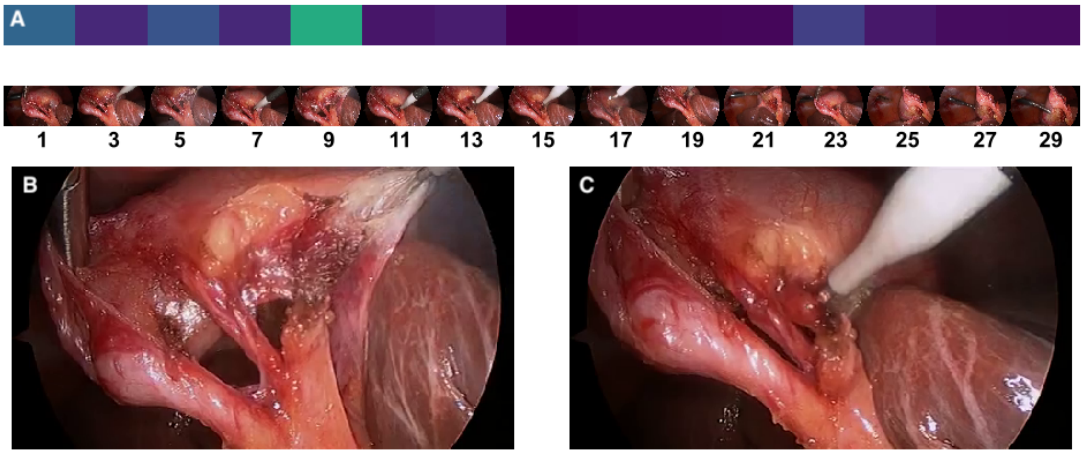

Figure 4: Demonstrating the frames the MIL model focuses on for classification. **A.** Attention heatmap aligned with the corresponding video frames. A warmer color represents a higher attention value. To create this figure, we subsampled the frames, taking 15 frames out of 30. **B.** The frame with the highest attention value (frame 9). **C.** The frame with the lowest attention value (frame 15). Additional examples are provided in the appendix.

it will be labeled as positive, and if all the instances in the bag are negative, it will be labeled as negative. This definition makes MIL ideal for the task of CVS criteria detection, as the input is a sequence of 30 frames, labeled as a single input sample, and each criterion might only be identified in one or a few specific frames of the entire input sequence. In contrast to our method's fixed window, the annotators assign a single label to the whole sequence based on specific frames they see during the annotation process. Thus, many frames in our input will be redundant for criteria detection, while some hold relevant information for successful identification. The challenge is to automatically focus on those frames during classifications, similarly to what the annotators undergo when making a decision. Based on the success of recent studies that apply attention modules to sequential signals , we use an attention-based MIL model (Ilse et al., 2018; Sheng, 2020). Figure 3 illustrates the network architecture. The input frames are processed using a ViT network to produce a feature representation for each frame. The resulting features are then applied to an attention block that learns to select the frames to focus on. Finally, the weights from the attention block are aggregated along with the input features, and processed using a small MLP network that produces the final predictions for each criterion. Figure 4 further emphasizes the importance of attention in our approach (additional examples of the importance of attention are provided in the appendix.). When focusing on the frames with the highest and lowest attention weights, it is clear that the first provides the information needed to identify CVS criteria, while in the latter, it is not possible to discern successfully which criteria were achieved.

All features-based models were trained using a single GPU for 100 epochs, with a batch size of 16 samples and a learning rate of 0.01. We use Stochastic Gradient Descent, with a

momentum of 0.9. A Sigmoid function was applied to each model output, and we used a binary cross-entropy loss function.

Fixed features or end-to-end training. Using a features-based setup has many benefits. It allows shorter training runtime, requires less computing power, and utilizes other pre-trained models trained on related tasks. However, previous studies showed that fine-tuning the backbone instead of using it only for feature extraction improves results significantly (Neimark et al., 2021a). We explored this approach and trained the entire network in Figure 3 in an end-to-end manner. Now, the inputs are the actual frames (images) instead of feature representations. We kept the input window size the same and used one frame for each second. Since end-to-end training also allows using image augmentations, we randomly resize the shorter side of all frames to a $[256, 320]$ scale and randomly crop all frames to 224×224. We initiate the model weights to a pre-trained model and finetune the entire backbone. The end-to-end experiments were done using 4 GPUs, with a batch size of 4 and a learning rate of 0.0012.

5. Results

To evaluate our approach, we use a mean unweighted accuracy, by calculating the unweighted accuracy for each criterion on the same test set and averaging them. All ablation experiments, including architecture selection and hyperparameter tuning, were done using the training and validation sets. The results reported in this section are all done on the same test set, which was kept aside during the exploration process.

5.1. Model architecture experiments

We compared several architectures to find the optimal architecture for CVS criteria detection (Table 2). To allow a fair comparison between the different architectures, we trained all models using the same set of fixed features. The attention MIL model outperformed the others, with more than 1% increase in each criterion accuracy.

Table 2: A comparison between different architectures for CVS criteria detection. We compared the unweighted accuracy of each criterion and the mean value of all criteria.

	Clean Triangle	Gallbladder Separation	Two Structures	Mean
MLP	83.28%	80.19%	75.78%	79.75%
LSTM (*mean* agg.)	81.07%	81.02%	78.20%	80.10%
LSTM (*take last* agg.)	83.06%	80.80%	78.39%	80.75%
Attention MIL	**84.64%**	**82.23%**	**79.47%**	**82.11%**

5.2. Multi-label *vs.* single-label classification

We examine two approaches for CVS criteria detection. First, a multi-label classification setup that combines all three criteria into a single model (i.e., a single 30-frame sequence is labeled with all three criteria labels). Second, a single-label classification setup for each criterion separately (i.e., three different models are trained, and the 30-frame sequence is

labeled only with one criterion value). We trained the MIL model in both ways to assess the contributions of multi-label training (Table 3). When trained separately, we replaced the Sigmoid function with a Softmax one. The results implies that the whole is greater than the sum of its parts. Multi-label classification outperforms separate models results and produces better accuracy.

Table 3: Multi-label classification *vs.* single-label comparison. Models were trained similarly, other than replacing the end classification function from Sigmoid to Softmax. We report the unweighted accuracy of each criterion and the mean value of three criteria together.

	Clean Triangle	Gallbladder Separation	Two Structures	Mean
Single-label	82.45%	81.83%	76.95%	80.41%
Multi-label	**84.64%**	**82.23%**	**79.47%**	**82.11%**

5.3. Fixed features *vs.* end-to-end training

Here, we compare between training from fixed features and training the entire network end-to-end. In Table 4, we show that although there is a slight improvement (0.5%) in the average performance, the criteria results were inconclusive. The results imply that end-to-end training has good potential, but the gain was minor in this framework. We argue that the relatively small dataset is insufficient to train a large network in an end-to-end manner as transformer-based networks require large-scale datasets to generalize (Devlin et al., 2018). More data will reveal the potential of such a setup even further.

Table 4: A comparison between fixed features and end-to-end training. We report the unweighted accuracy of each criterion and the mean value of all three. The mean unweighted accuracy improves when trained end-to-end.

	Clean Triangle	Gallbladder Separation	Two Structures	Mean
Fixed features	**84.64%**	82.23%	79.47%	82.11%
End-to-end	82.26%	**83.41%**	**82.14%**	**82.60%**

5.4. CVS detection

The end result we are interested in is evaluating the ability to detect whether CVS was achieved or not, based on the different criteria predictions. Here, we assess the performance of both fixed features and end-to-end training on detecting CVS by using only criteria identification. We consider a CVS to be positive when all three criteria are positive. The end-to-end architecture achieves better results in the final task of CVS detection compare to the fixed features alternative (Table 5).

Table 5: CVS detection results, using the criteria detection model to identify if CVS was achieved or not. The end-to-end architecture performed better in every metric, with a 4% gain in accuracy compared to the fixed features architecture.

	Accuracy	Unweighted Accuracy	Specificity	Sensitivity	F1-score
Fixed features	80.1%	77.2%	82.4%	71.9%	61%
End-to-end	**84.2%**	**80%**	**87.3%**	**72.7%**	**66.4%**

5.5. Partial CVS detection

Although the common practice treats CVS as a binary decision, it seems counterintuitive to consider partial CVS as if CVS wasn't achieved at all. We believe that surgeons can still gain valuable information even when not all criteria were achieved. For example, in high complexity procedures, when it is sometimes hard to obtain a complete CVS, partial CVS detection can still have a clinical value. We define partial CVS if at least two criteria exist but not necessarily all three. The features-based model achieved an accuracy of 84.9%, with unweighted accuracy of 88.1% and 82.3% for positive and negative classes, respectively.

6. Discussion

Previous studies demonstrated the advantages of implementing AI techniques in surgical-related tasks. Building on top of their findings, we explore the challenging downstream task of CVS detection. Our goal was to develop a robust method for CVS detection in a production clinical setting - namely, fitted to operate in the OR tomorrow.

Our proposed approach applies attention-based multi instance learning and produces high accuracy of 84.2% in detecting whether CVS was achieved or not. This result represents a 10% increase in performance compared to other methods suggested before, by only using the raw video as input.

Comparing the results of our approach with the one suggested by Mascagni et al. (2021) and with the alternative baseline models explored in this work indicates that CVS detection fits and can benefit from the formulation of a MIL problem. The MIL approach achieves better performance and also simplifies the labeling process. Moreover, the attention-based method enables illustrating the frames the model focused on during inference. This significantly adds to our method's interpretability, which is an important property for ML-based applications in the healthcare domain.

Although the vast majority of surgeons agree on the importance of CVS in procedures such as Laparoscopic Cholecystectomy, getting a group of experts to agree on CVS achievement and reach consensus is not trivial and leads to many biases in what is supposed to be a methodic milestone. Future work will explore the usability of our method in real-time scenarios and study its ability to provide immediate support for better decision-making.

An AI-based system for CVS detection, which learns from numerous surgical procedures done by various surgeons, can set a solid and robust standardization for CVS achievement and allow better supervision as part of the procedure routine, ultimately leading to better surgical best practices and improved patients outcomes.

Limitations The results presented in this work are limited to a single surgical procedure, Laparoscopic Cholecystectomy. We choose to focus on this procedure since CVS is a well-known method routinely performed in the OR. However, further research is needed in order to validate our approach in other *ectomy* procedures, such as Laparoscopic Appendectomy. The dataset described in this study was randomly divided into three subsets: training, validation, and test. Other types of sets, leaving out all videos from a specific medical center, surgeon, or technique, could shed more light on the method's robustness. In addition, our approach is based on first identifying the CVS location by using the segment before the division of the cystic structures. Our analysis was done by using the manual annotations of the surgical steps. Interesting future research should evaluate our approach in an end-to-end manner, in which both the division step and the CVS are automatically detected.

Acknowledgments

We thank Ross Girshick for providing valuable feedback on this manuscript and helpful suggestions on several experiments.

References

Omri Bar, Daniel Neimark, Maya Zohar, Gregory D Hager, Ross Girshick, Gerald M Fried, Tamir Wolf, and Dotan Asselmann. Impact of data on generalization of ai for surgical intelligence applications. *Scientific reports*, 10(1):1–12, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

KH Fuchs. Minimally invasive surgery. *Endoscopy*, 34(02):154–159, 2002.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Daniel A Hashimoto, Guy Rosman, Elan R Witkowski, Caitlin Stafford, Allison J Navarrete-Welton, David W Rattner, Keith D Lillemoe, Daniela L Rus, and Ozanan R Meireles. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Annals of surgery*, 270(3):414, 2019.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 691–699. IEEE, 2018.

James R Korndorffer Jr, Mary T Hawn, David A Spain, Lisa M Knowlton, Dan E Azagury, Aussama K Nassar, James N Lau, Katherine D Arnow, Amber W Trickey, and Carla M Pugh. Situating artificial intelligence in surgery: a focus on disease severity. *Annals of Surgery*, 272(3):523–528, 2020.

Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

Pietro Mascagni, Armine Vardazaryan, Deepak Alapatt, Takeshi Urade, Taha Emre, Claudio Fiorillo, Patrick Pessaux, Didier Mutter, Jacques Marescaux, Guido Costamagna, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of Surgery*, 2021.

Ahmad HM Nassar, Hwei J Ng, Arkadiusz Peter Wysocki, Khurram Shahzad Khan, and Ines C Gil. Achieving the critical view of safety in the difficult laparoscopic cholecystectomy: a prospective study of predictors of failure. *Surgical Endoscopy*, 35(11):6039–6047, 2021.

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021a.

Daniel Neimark, Omri Bar, Maya Zohar, Gregory D. Hager, and Dotan Asselmann. "train one, classify one, teach one" - cross-surgery transfer learning for surgical step recognition. In *Medical Imaging with Deep Learning*, 2021b. URL https://openreview.net/forum?id=cTB4Qz3RzCl.

Anabik Pal, Zhiyun Xue, Kanan Desai, Adekunbiola Aina F Banjo, Clement Akinfolarin Adepiti, L Rodney Long, Mark Schiffman, and Sameer Antani. Deep multiple-instance learning for abnormal cell detection in cervical histopathology images. *Computers in Biology and Medicine*, 138:104890, 2021.

Philip H Pucher, L Michael Brunt, Neil Davies, Ali Linsk, Amani Munshi, H Alejandro Rodriguez, Abe Fingerhut, Robert D Fanelli, Horacio Asbun, and Rajesh Aggarwal. Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis. *Surgical endoscopy*, 32(5):2175–2183, 2018.

M Sadegh Saberian, Kathleen P Moriarty, Andrea D Olmstead, Ivan R Nabi, François Jean, Maxwell W Libbrecht, and Ghassan Hamarneh. Deemd: Drug efficacy estimation against

sars-cov-2 based on cell morphology with deep multiple instance learning. *arXiv preprint arXiv:2105.05758*, 2021.

Ario Sadafi, Asya Makhro, Anna Bogdanova, Nassir Navab, Tingying Peng, Shadi Albarqouni, and Carsten Marr. Attention based multiple instance learning for classification of blood cell disorders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–256. Springer, 2020.

Lucia Ilaria Sgaramella, Angela Gurrado, Alessandro Pasculli, Nicola de Angelis, Riccardo Memeo, Francesco Paolo Prete, Stefano Berti, Graziano Ceccarelli, Marco Rigamonti, Francesco Giuseppe Aldo Badessi, et al. The critical view of safety during laparoscopic cholecystectomy: Strasberg yes or no? an italian multicentre study. *Surgical endoscopy*, 35(7):3698–3708, 2021.

Lori Sheng. Multiple instance learning with mnist dataset using pytorch, 2020. URL https://medium.com/swlh/multiple-instance-learning-c49bd21f5620.

Steven M Strasberg and L Michael Brunt. Rationale and use of the critical view of safety in laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 211(1): 132–138, 2010.

Steven M Strasberg, Martin Hertl, and Nathaniel J Soper. An analysis of the problem of biliary injury during laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 180(1):101–125, 1995.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Lawrence W Way, Lygia Stewart, Walter Gantert, Kingsway Liu, Crystine M Lee, Karen Whang, and John G Hunter. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Annals of surgery*, 237(4):460, 2003.

Ying Xu, Xinyang Qian, Xuanping Zhang, Xin Lai, Yuqian Liu, and Jiayin Wang. Deeplion: Deep multi-instance learning improves the prediction of cancer-associated t cell receptors for accurate cancer detection. *Frontiers in genetics*, 13:860510, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

Felix Yu, Gianluca Silva Croso, Tae Soo Kim, Ziang Song, Felix Parker, Gregory D Hager, Austin Reiter, S Swaroop Vedula, Haider Ali, and Shameema Sikder. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA network open*, 2(4):e191860–e191860, 2019.

Maya Zohar, Omri Bar, Daniel Neimark, Gregory D Hager, and Dotan Asselmann. Accurate detection of out of body segments in surgical video using semi-supervised learning. In *Medical Imaging with Deep Learning*, pages 923–936. PMLR, 2020.

Appendix A. Demographics information

Table 6: Age group and sex distribution of the videos in our dataset.

Age Group	#videos	%
0-18	20	1
19-33	293	13
34-48	415	19
49-64	475	21
65-78	345	16
79-	144	6
untagged	531	24

Sex	#videos	%
Male	587	27
Female	1116	50
untagged	520	23

Appendix B. Additional examples of the importance of attention

Demonstrating the frames the MIL model focuses on for classification. A warmer color represents a higher attention value. To create this figure, we subsampled the frames, taking 15 frames out of 30.

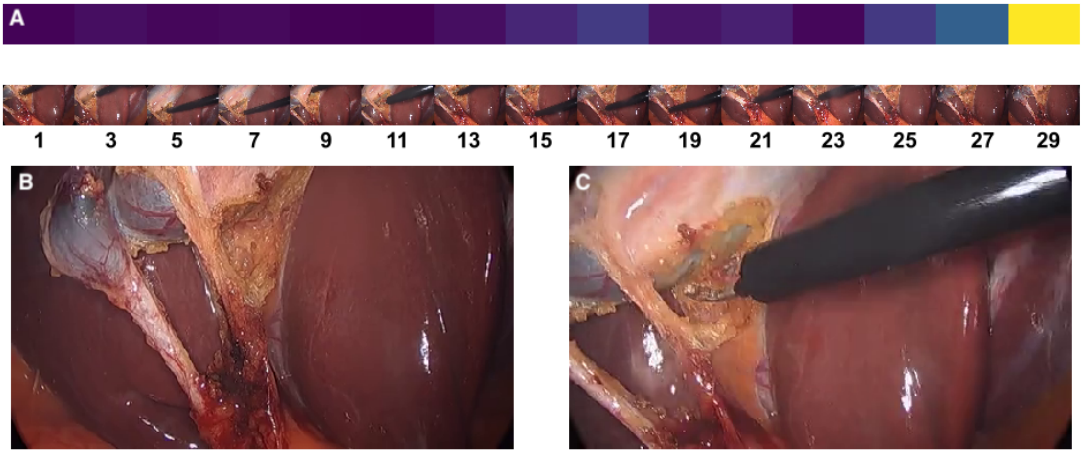

Figure 5: A positive CVS example **A.** Attention heatmap aligned with the corresponding video frames. **B.** The frame with the highest attention value (frame 29). **C.** The frame with the lowest attention value (frame 11).

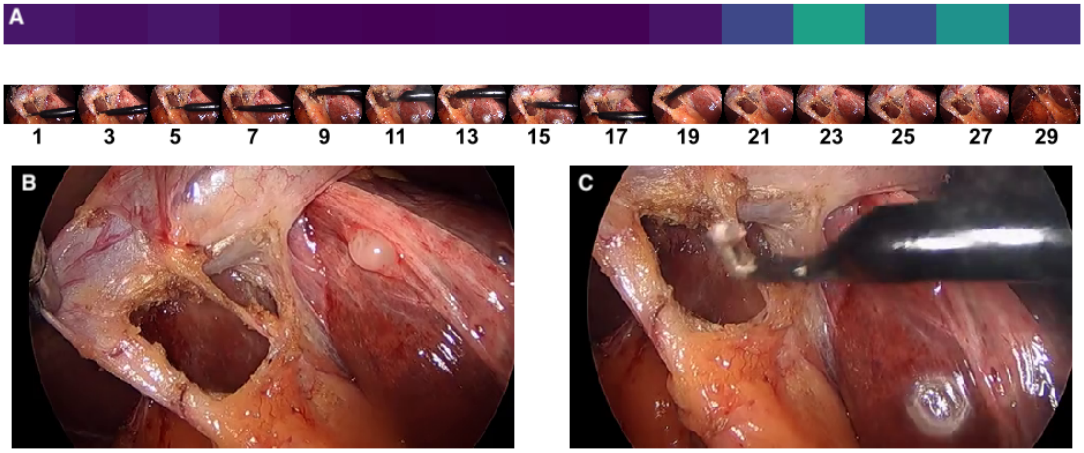

Figure 6: A positive CVS example **A.** Attention heatmap aligned with the corresponding video frames. **B.** The frame with the highest attention value (frame 23). **C.** The frame with the lowest attention value (frame 11).

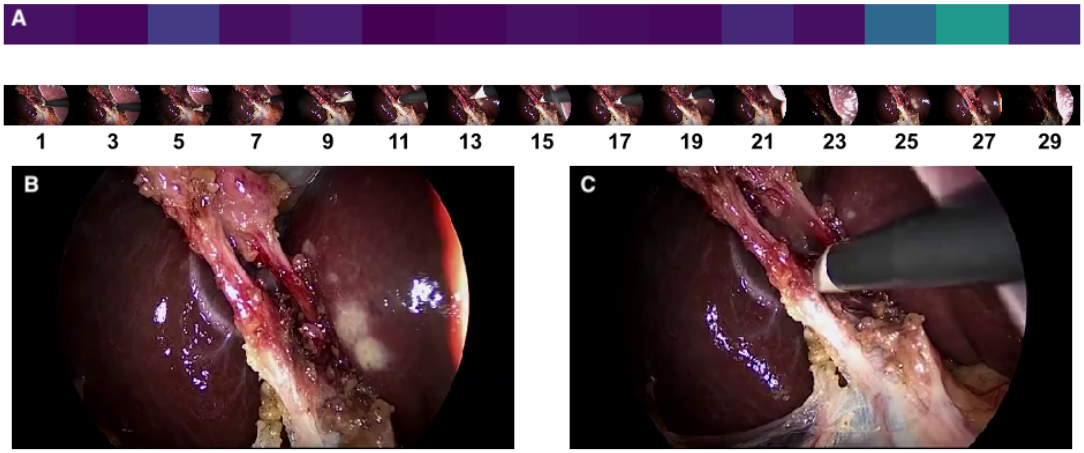

Figure 7: A negative CVS example that shows uniform attention weights across all frames. **A.** Attention heatmap aligned with the corresponding video frames. **B.** The frame with the highest attention value (frame 27). **C.** The frame with the lowest attention value (frame 11).

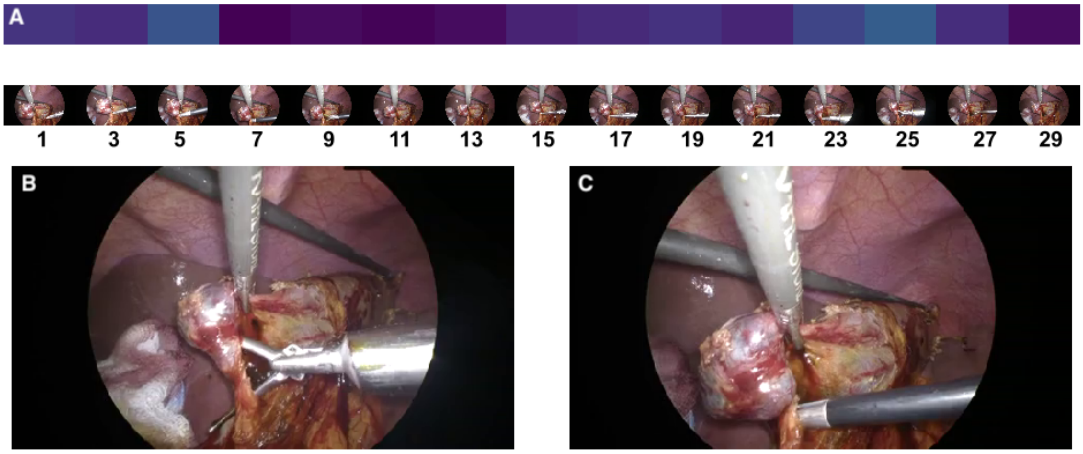

Figure 8: A negative CVS example that shows uniform attention weights across all frames. **A.** Attention heatmap aligned with the corresponding video frames. **B.** The frame with the highest attention value (frame 25). **C.** The frame with the lowest attention value (frame 7).

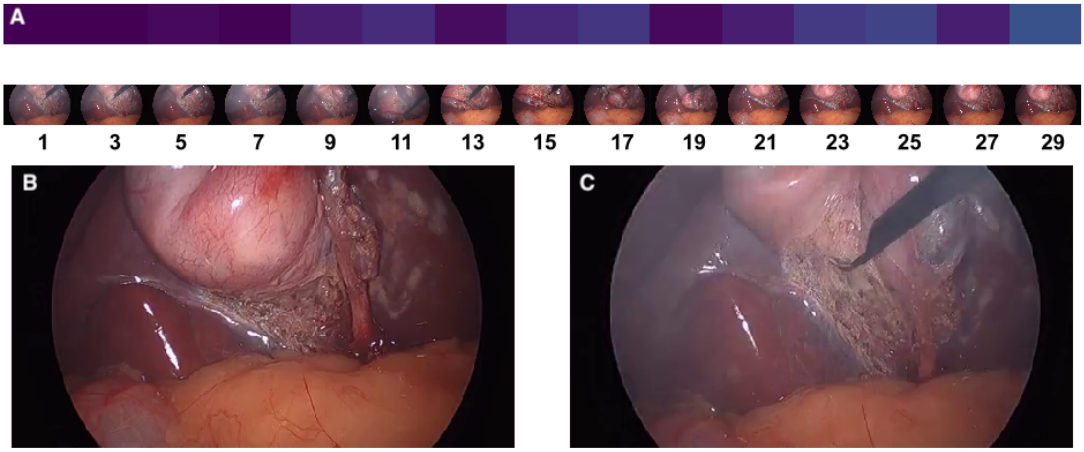

Figure 9: A negative CVS example **A.** Attention heatmap aligned with the corresponding video frames. **B.** The frame with the highest attention value (frame 29). **C.** The frame with the lowest attention value (frame 1).