B^2 Loss for Evidential Classification of Critical View of Safety in Laparoscopic Cholecystectomy

Franciszek M. Nowak[1](✉), Evangelos B. Mazomenos[1], Brian Davidson[1,2], and Matthew J. Clarkson[1]

[1] UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, UCL, London, UK
[2] Division of Surgery and Interventional Science, UCL, London, UK
franciszek.nowak.23@ucl.ac.uk

Abstract. Automated recognition of the Critical View of Safety (CVS) in Laparoscopic Cholecystectomy is considered a deterministic classification task using majority vote labels. However, expert annotators' disagreement does not equate annotation error, rather reflecting informative ambiguity in judgement. Reducing such labels to hard targets ignores the underlying sampling process and discards information about observer's variability. Our work models annotator votes as Binomial observations of a latent visibility probability and learns its predictive distribution via a conjugate Beta-Binomial (B^2) loss with evidential parametrisation. The network outputs a Beta distribution over the latent visibility probability, enabling direct estimation of the probability that a majority of experts would judge a CVS criterion as visible. On Endoscapes2023, the proposed approach significantly improves average balanced accuracy by +2.1% compared to the state-of-the-art. We also introduce a precision-critical video-level evaluation protocol reflecting surgical decision-making, reporting recall under a strict precision constraint (precision = 1.0). Under this, the model maintains meaningful recall while avoiding false positives. Modelling annotator disagreement probabilistically provides a principled and clinically aligned alternative to deterministic CVS classification and supports precision-focused evaluation for surgical deployment. Code available at: *github.com/franeknowak/BetaBinomialLoss*.

Keywords: Laparoscopic Cholecystectomy · Critical View of Safety · Beta-Binomial Loss · Evidential Classification · Machine Learning

1 Introduction

Laparoscopic cholecystectomy (LC) is the minimally invasive removal of the gallbladder, one of the most frequently performed surgical procedures[1]. The most widely adopted intraoperative safety protocol is the Critical View of Safety (CVS) [4,5]. It requires the surgeon to achieve three criteria, C1-3, before committing to cutting vital structures. C1 requires that only the cystic duct and cystic artery enter the gallbladder; C2 requires clear dissection of the hepatocystic triangle from fat and fibrous tissue; and C3 requires visual separation of

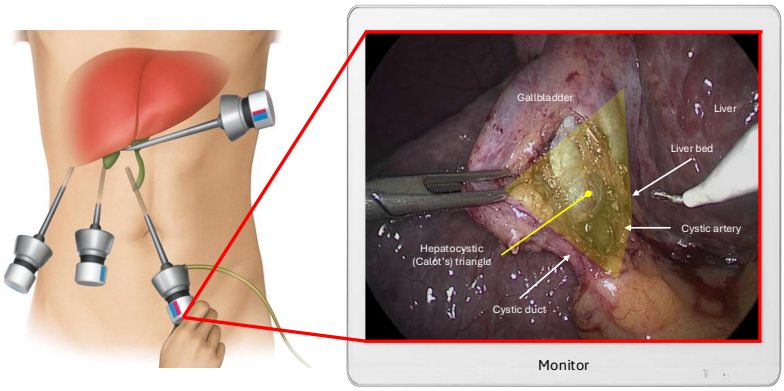

Fig. 1. Example frame from the laparoscopic procedure, annotated with anatomies relevant to the Critical View of Safety.

the gallbladder neck from the liver bed[19,13]. Achieving these before clipping the cystic structures makes it anatomically unlikely to misidentify the common bile duct for the cystic duct, thus avoiding severe vasculobiliary injury. CVS structures are visualised in Fig. 1.

Although CVS has shown benefits [17], the incidence (0.15-0.6%)[11,21,9,20] of bile duct injury (BDI) has remained largely unchanged [2,3]. This may reflect variable familiarity with the technique among clinicians, resulting in inconsistent achievement of the CVS (10.8-92.5%)[14,10]. Studies also report discrepancies between surgeons' perceived and actual attainment of the CVS [18,6]. This motivates data-driven, automated methods for validating CVS assessment.

State-of-the-art video analysis employs spatio-temporal encoding of short video clips to predict the presence of each CVS criterion. SV2LSTG[13] integrates graph-based reasoning and semantic segmentation masks, whereas SwinCVS[15] achieves comparable performance without dense annotations. Typically, CVS detection is treated as standard multi-label classification using majority-vote ground truth, optimised with conventional classification losses.

Two fundamental limitations remain. First, anatomical representation of CVS is subjective. The most comprehensive LC dataset to date, Endoscapes2023[12], is labelled by independent expert clinicians, with labels encoded as proportions $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$, reflecting annotator agreement. In that dataset the frames with at least one positive criterion are unanimously agreed only in 5.25% of the cases (Cohen's kappa between 0.33-0.55[12]). All previously published CVS detection systems collapse the disagreeing annotations to hard, majority-vote labels. Such conversion discards information about ambiguity in the images and implicitly assumes that an expert's minority vote is incorrect. Given that annotators are expert surgeons, this assumption is inappropriate. Instead, we propose that the provided soft-labels be interpreted as stochastic observations of a latent perceptual variable.

Second limitation comes from the current established standard of presenting results specifically at the frame-level: accuracy, balanced accuracy, and mean average precision. These metrics treat false-negatives and false-positives equally, ignoring the severe clinical consequences of the latter. As criteria can be achieved at any point in the procedure, frame-level metrics do not reflect this.

We propose methods to address both. We consider annotator responses as stochastic samples from a Binomial process, modelling a latent visibility probability θ; representing the probability that a randomly selected annotator would judge the criterion as visible. The objective is to estimate this latent probability and quantify its uncertainty, rather than predict a noisy hard label. Our contributions are:

- **Probabilistic modelling of annotator disagreement**: We model the three annotations as Binomial observations of a latent visibility probability, avoiding information loss in majority vote labels.
- **Beta-Binomial (B^2) loss**: We propose a statistically grounded loss derived from the marginal Beta-Binomial likelihood, enabling learning directly from annotator vote distributions.
- **Evidential uncertainty modelling**: We employ evidential learning to estimate both the latent visibility probability and the strength of evidence supporting it via a Beta-Binomial evidential framework.
- **Precision-critical video-level evaluation**: We introduce an evaluation protocol aligned with clinical requirements, prioritising precision and assessing criterion achievement at video level.

2 Methods

2.1 Dataset

We use the weakly labelled variant of the Endoscapes2023 dataset[13], without bounding boxes or segmentation masks. The dataset contains frames from 201 procedures, annotated with per-frame and per-video visibility of the three CVS criteria (C1-C3). Annotations are provided for every fifth frame, represented as the average from annotators who marked a criterion as visible, resulting in values in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Contents of the dataset are displayed in Table 1.

Table 1. Total number of annotated frames per class in the training, validation and testing subsets in Endoscapes2023. C1, C2, and C3 correspond to respective CVS criteria. 'No CVS' refers to the lack of any visible CVS criteria.

	Total	**No CVS**	**C1**	**C2**	**C3**
Training (120 videos)	6960	5023	1088	780	1245
Validation (41 videos)	2331	1757	381	291	389
Testing (40 videos)	1799	1031	423	302	501

2.2 Probabilistic Modelling of Annotator Votes

Probabilistic Formulation of Annotator Disagreement: Since annotations are provided as proportions in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ from three experts, we convert them to integer vote counts $k \in \{0, 1, 2, 3\}$. We model this vote count as a Binomial random variable with three trials and success probability θ, representing the latent probability that a randomly selected annotator would judge the criterion as visible. We assume annotators are exchangeable and independent expert observers, such that disagreement reflects perceptual ambiguity rather than a systematic error, as is implicit in any majority-vote labelling scheme. Therefore,

$$k \mid \theta \sim \text{Binomial}(3, \theta) \tag{1}$$

The probability of k annotators saying "visible" out of n annotators given θ is therefore:

$$p(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}. \tag{2}$$

Evidential Beta Parametrisation: To estimate the latent visibility probability, we adopt an evidential formulation. Instead of predicting a single scalar probability, the network outputs parameters of a Beta distribution:

$$\theta \sim \text{Beta}(\alpha_1, \alpha_0) \sim \frac{\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}}{B(\alpha_1, \alpha_0)}, \tag{3}$$

where α_1 and α_0 represent evidence supporting visibility and absence respectively, and $B(\cdot, \cdot)$ is the Beta function. The evidential formulation follows the framework of evidential deep learning proposed by Sensoy et al.[16], adapted here within a fully likelihood-based conjugate Beta-Binomial model.

Beta-Binomial Likelihood ($\mathbf{B^2}$ Loss): Marginalising over θ (Eq. 2, 3) yields the Beta-Binomial distribution:

$$k \sim \text{Beta-Binomial}(n = 3, \alpha_1, \alpha_0). \tag{4}$$

The per-frame log-likelihood is:

$$\log P(k \mid \alpha_1, \alpha_0) = \log \binom{3}{k} + \log B(k + \alpha_1, 3 - k + \alpha_0) - \log B(\alpha_1, \alpha_0). \tag{5}$$

During training we minimise the negative log-likelihood. This objective directly models annotator vote counts rather than majority labels, preserving information about inter-observer variability. This corresponds to the likelihood under the assumed Binomial-Beta generative model. Unlike cross-entropy, it does not assume deterministic ground truth.

Given the strong frame-level class imbalance in CVS visibility (Table 1), we also introduce a prevalence-aware prior. To discourage unjustified overconfidence, we penalise deviation from this prior via a KL divergence term:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\Big(\text{Beta}(\alpha_1, \alpha_0) \parallel \text{Beta}(\alpha_1^{\text{prior}}, \alpha_0^{\text{prior}})\Big), \tag{6}$$

where priors are $\alpha_1^{\text{prior}} = \pi\nu$ and $\alpha_0^{\text{prior}} = (1-\pi)\nu$ with π denoting empirical prevalence and ν a concentration parameter. $\mathcal{L}_{\text{KL}}$ constrains the concentration parameter S by pulling it toward a prevalence-aware prior, thus preventing pathological evidence growth in data-sparse regions.

Decision Quantity, Majority Probability: Evaluation in prior work is performed against majority-vote labels. Accordingly, the appropriate Bayesian decision quantity is not $\mathbb{E}[\theta]$, but the probability that a majority of annotators would vote positive. This formulation integrates over uncertainty in θ, rather than thresholding a point estimate:

$$P_{\text{maj}} = P(k \geq 2 \mid \alpha_1, \alpha_0). \tag{7}$$

For $n = 3$, this has a closed form:

$$P_{\text{maj}} = 3\mathbb{E}[\theta^2] - 2\mathbb{E}[\theta^3] = 3\frac{\alpha_1(\alpha_1+1)}{S(S+1)} - 2\frac{\alpha_1(\alpha_1+1)(\alpha_1+2)}{S(S+1)(S+2)}, \tag{8}$$

where $S = \alpha_1 + \alpha_0$ is the total concentration (evidence strength).

Interpretation of Uncertainty: The concentration parameter $S = \alpha_1 + \alpha_0$ controls the dispersion of the Beta distribution. Following the subjective logic decomposition of Audun Jøsang[8], total uncertainty mass for a Beta distribution (binary Dirichlet) can be expressed as $u = \frac{2}{S}$. Low S corresponds to diffuse distributions over θ, while high S indicates concentrated belief. Crucially, uncertainty is not an external auxiliary output; it directly influences the decision probability P_{maj} through the Beta moments. Thus, predictive uncertainty is already integrated into the classification probability.

2.3 Implementation details

We use a sequence model that predicts evidential outputs for the three CVS criteria from short, 5-frame video clips sampled at 1 FPS. Each frame is encoded using a DINOv3 ViT-B/16 backbone initialised from pretrained weights. The final six transformer blocks are first adapted on majority-vote labels, after which the encoder is frozen during evidential training. Each frame embedding in a sequence, $z_t \in \mathbb{R}^{768}$, is aggregated over time using a gated attention pooling module (LayerNorm–Linear–GELU–Linear), which outputs a softmax weight per-frame and forms a weighted sum $z_{\text{seq}} = \sum_t w_t\, z_t$. Finally, three independent evidential heads (one per C criterion; MLP with GELU) map z_{seq} to Beta parameters (α_0, α_1) by predicting non-negative evidence via softplus$(\cdot)$ and converting it to α, yielding a Beta distribution over latent visibility for each criterion. Our proposed "B^2 Model" is visualised on Fig. 2. Models' performance was compared across five seeds to reject initialisation bias. Results were statistically evaluated; tested for normality via Shapiro-Wilk and improvement significance with one-sided Wilcoxon signed-rank tests.

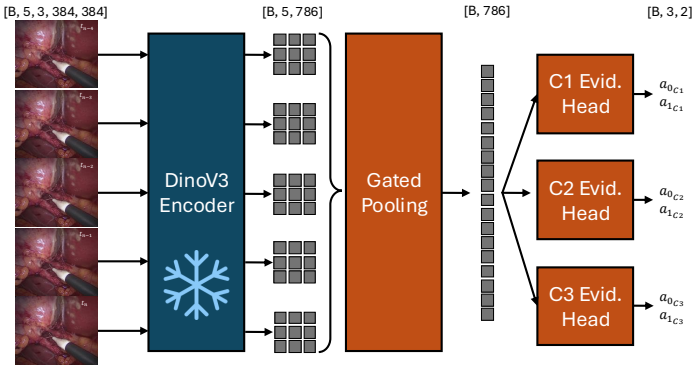

Fig. 2. Visualisation of B^2 model architecture. A sequence of five consecutive frames is passed through DINOv3 encoder. The image embeddings are aggregated using gated pooling module and the output encoding is processed by three independent evidential classifier heads, resulting in unique αs for each criterion.

3 Results

3.1 Per-Sequence

We compare the B^2 model's per-sequence results against the SwinCVS (E2E variant). We chose this comparison as it is 1) the best-performing model trained on Endoscapes2023 using exclusively weakly annotated images; 2) both models have similar architecture design based on five frames→encoder→temporal aggregation→classifier. We used code provided by the authors to run their model on five separate seeds to obtain the variance in the metrics (Table 2). Our proposed solution significantly outperforms SwinCVS across all measured metrics for averaged criteria scores. SwinCVS achieves improved performance only on C2 criterion in ACC, a metric which is skewed by data imbalance, and MAP which is less interpretable under soft, disagreement-driven labels, as it evaluates binary relevance while our model estimates majority-vote probability.

Table 2. Per-sequence results comparing B^2 and SwinCVS models. Each result provided with mean and standard deviation from five runs initialised on different seeds. ACC - Accuracy; BACC - Balanced accuracy; MAP - Mean average precision; **Bold** - statistically significant improvement (one-sided Wilcoxon signed-rank test, $p < 0.05$).

	Average C1-3			C1		
	ACC[%]	BACC[%]	MAP[%]	ACC[%]	BACC[%]	MAP[%]
SwinCVS	82.7±0.52	70.3±1.57	65.4±0.56	82.2±1.06	74.0±2.17	66.4±0.55
B^2 Model	**83.7±0.17**	**72.4±0.64**	**66.5±0.49**	82.5±0.36	73.9±0.86	65.9±0.88
	C2			C3		
	ACC[%]	BACC[%]	MAP[%]	ACC[%]	BACC[%]	MAP[%]
SwinCVS	**87.2±0.54**	69.9±2.27	**63.0±1.32**	78.7±0.59	67.0±3.02	66.9±1.27
B^2 Model	85.6±0.31	70.1±0.94	59.2±0.77	**82.8±0.20**	**73.3±0.63**	**74.5±0.29**

Table 3. Per-video means with 95% bootstrap CIs for B^2 Model. Format: mean [lower : upper]. † – threshold 0.5; * – threshold calibrated for Prec = 1.0 on validation.

		Precision [%]	Recall [%]	Specificity [%]
C1	Val†	60.0 [53.8 : 66.7]	95.5 [86.4 : 100]	26.3 [10.5 : 47.4]
	Test†	68.6 [62.9 : 75.8]	96.0 [88.0 : 100]	26.7 [6.7 : 46.7]
	Val*	100.0 [100 : 100]	27.3 [9.1 : 45.5]	100.0 [100 : 100]
	Test*	88.9 [66.7 : 100]	32.0 [16.0 : 52.0]	93.3 [80.0 : 100]
C2	Val†	59.3 [48.0 : 70.8]	72.7 [54.5 : 90.9]	42.1 [21.1 : 63.2]
	Test†	90.5 [79.2 : 100]	95.0 [85.0 : 100]	90.0 [75.0 : 100]
	Val*	100.0 [100 : 100]	13.6 [5.2 : 27.3]	100.0 [100 : 100]
	Test*	100.0 [100 : 100]	40.0 [20.0 : 60.1]	100.0 [100 : 100]
C3	Val†	62.1 [52.8 : 74.1]	90.0 [75.0 : 100]	47.6 [23.8 : 71.4]
	Test†	86.4 [74.1 : 100]	82.6 [65.2 : 95.7]	82.4 [64.7 : 100]
	Val*	100.0 [100 : 100]	30.0 [10.0 : 50.0]	100.0 [100 : 100]
	Test*	100.0 [100 : 100]	43.5 [26.0 : 65.2]	100.0 [100 : 100]

3.2 Per-Video

Per-sequence evaluation is useful for assessing class separability, but it does not reflect clinically realistic deployment. CVS confirmation is a binary surgical decision made once per procedure: a positive prediction grants permission to transect critical structures, making a single false positive potentially catastrophic. We therefore introduce a video-level evaluation protocol where the classification threshold is calibrated on the validation set to achieve precision = 1.0, and recall is subsequently reported on the test set. For each video, the B^2 model is applied to all frames and predictions are aggregated by taking the maximum across frames. This is consistent with clinical practice, where a criterion needs to be confirmed only once during the procedure. As shown in Table 3 using the default threshold of 0.5 results in substantial overprediction of criterion presence. Enforcing a precision-focused threshold ($C1 = 0.799; C2 = 0.929; C3 = 0.879$) reduces false positives but correspondingly lowers recall, indicating that stricter confirmation criteria improve reliability at the cost of missed detections.

4 Discussion

This work reframes CVS recognition from deterministic multi-label classification to probabilistic modelling of annotator agreement. Instead of reducing soft labels to majority vote targets, we treat annotator responses as stochastic observations of a latent visibility probability and learn it via a conjugate Beta-Binomial formulation. This preserves information about disagreement and avoids assuming that minority expert votes are incorrect.

The uncertainty provided by our model represents distributional uncertainty over the latent annotator visibility probability, where low concentration reflects weak evidential support in representation space. It does not correspond to posterior uncertainty over network parameters. Fig. 3 illustrates predicted majority probability against evidence on the test subset, showing structured separation

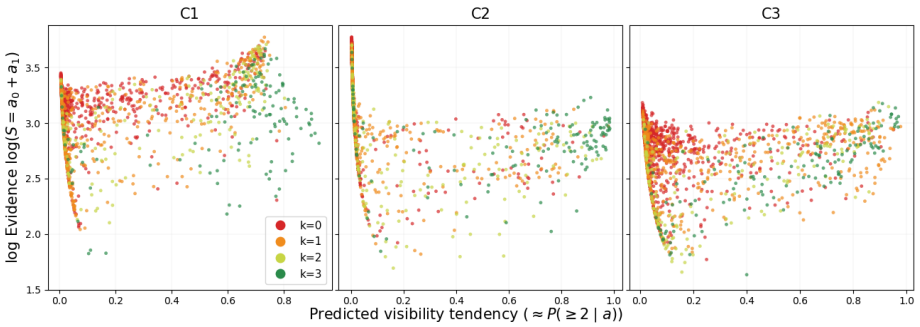

Fig. 3. Distribution of B^2 Model predicted binomial probability for $k \geq 2$ against logarithmically scaled evidence S. Colours indicate labels, how many annotators (k) considered the frame to have a criterion visible. Note: Best viewed in colour.

between ($k \geq 2$) and ($k \leq 1$) samples. Evidence is broadly distributed across vote counts, with a tendency toward higher concentration for positive samples. As uncertainty reflects evidence strength rather than error probability, no strong correlation with accuracy is observed.

Although the B^2 model demonstrates improved average per-sequence performance relative to SwinCVS (Table 2), frame-level metrics alone are insufficient to assess clinical relevance. CVS confirmation is a video-level decision. A criterion must be identified at least once during the procedure, and false positives carry greater clinical risk than false negatives. As shown in Table 3, directly transferring per-sequence evaluation methodology to the video level yields suboptimal precision. We therefore advocate evaluating CVS recognition at the video level by reporting recall under a strict precision constraint (precision = 1.0). This protocol better reflects clinical deployment, where automated confirmation must avoid false positives while identifying as many correctly achieved criteria as possible.

In terms of limitations, the evidential formulation does not capture epistemic uncertainty over model parameters; this could instead be approximated through ensembles of independently initialised models to obtain complementary variance-based estimates[7]. Furthermore, uncertainty quality is highly correlated with representation quality. Some CVS approaches incorporating semantic segmentation masks (e.g., SV2LSTG[13]) achieve stronger performance, suggesting that integrating this supervision could improve both predictive accuracy and evidential reliability.

In summary, we introduce a statistically grounded Beta-Binomial evidential framework for CVS recognition that models annotator disagreement explicitly and integrates uncertainty directly into predictive probabilities. The primary contribution lies in aligning probabilistic modelling with perceptual annotation structure and clinical decision-making, treating expert disagreement as a signal rather than noise, as well as providing a principled foundation for precision-critical surgical video analysis with the potential to extend our method to other multiple-expert annotation-driven recognition tasks.

Acknowledgments. Researcher's time was funded in whole, or in part, by the EPSRC-funded Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) [EP/S021930/1], the NIHR Central London Patient Safety Research Collaboration (CL PSRC) [NIHR204297] and Human-centred Machine Intelligence to optimise Robotic Surgical Training [EP/Z534754/1]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Abbott, T.E.F., Fowler, A.J., Dobbs, T.D., et al.: Frequency of surgical treatment and related hospital procedures in the UK: a national ecological study using hospital episode statistics. BJA: British Journal of Anaesthesia **119**(2), 249–257 (07 2017)
2. Agresta, F., Campanile, F.C., Vettoretto, N., et al.: Laparoscopic cholecystectomy: consensus conference-based guidelines. Langenbeck's Archives of Surgery **400**, 429–453 (5 2015)
3. Alius, C., Serban, D., Bratu, D.G., et al.: When critical view of safety fails: A practical perspective on difficult laparoscopic cholecystectomy. Medicina **59**, 1491 (8 2023)
4. de'Angelis, N., Catena, F., Memeo, R., et al.: 2020 WSES guidelines for the detection and management of bile duct injury during cholecystectomy. World Journal of Emergency Surgery **16**, 30 (12 2021)
5. Eikermann, M., Siegel, R., Broeders, I., et al.: Prevention and treatment of bile duct injuries during laparoscopic cholecystectomy: the clinical practice guidelines of the european association for endoscopic surgery (EAES). Surgical Endoscopy **26**, 3003–3039 (11 2012)
6. van de Graaf, F.W., van den Bos, J., et al.: Lacunar implementation of the critical view of safety technique for laparoscopic cholecystectomy: Results of a nationwide survey. Surgery **164**, 31–39 (7 2018)
7. Jain, A., Bates, S.: Deep ensembles for epistemic uncertainty: A frequentist perspective (2026), `https://arxiv.org/abs/2510.22063`
8. Jøsang, A.: Subjective Logic: A Formalism for Reasoning Under Uncertainty. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing (2016)
9. Kaman, L., Sanyal, S., Behera, A., et al.: Comparison of major bile duct injuries following laparoscopic cholecystectomy and open cholecystectomy. ANZ Journal of Surgery **76**, 788–791 (9 2006)
10. Manatakis, D.K., Antonopoulou, M., Tasis, N., et al.: Critical view of safety in laparoscopic cholecystectomy: A systematic review of current evidence and future perspectives. World Journal of Surgery **47**, 640–648 (3 2023)
11. Martin, D., Uldry, E., et al.: Bile duct injuries after laparoscopic cholecystectomy: 11-year experience in a tertiary center. BioScience Trends **10**, 197–201 (2016)
12. Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., Padoy, N.: Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. Scientific Data **12**(1), 331 (2025)

13. Murali, A., Alapatt, D., Mascagni, P., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv (10 2024)
14. Nijssen, M.A.J., Schreinemakers, J.M.J., Meyer, Z., et al.: Complications after laparoscopic cholecystectomy: A video evaluation study of whether the critical view of safety was reached. World Journal of Surgery **39**, 1798–1803 (7 2015)
15. Nowak, F.M., Mazomenos, E.B., et al.: SwinCVS: a unified approach to classifying critical view of safety structures in laparoscopic cholecystectomy. International Journal of Computer Assisted Radiology and Surgery **20**, 1145–1152 (4 2025)
16. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty (2018), `https://arxiv.org/abs/1806.01768`
17. Sgaramella, L.I., Gurrado, A., Pasculli, A., et al.: The critical view of safety during laparoscopic cholecystectomy: Strasberg yes or no? An Italian multicentre study. Surgical Endoscopy **35**, 3698–3708 (7 2021)
18. Stefanidis, D., Chintalapudi, N., Anderson-Montoya, B., et al.: How often do surgeons obtain the critical view of safety during laparoscopic cholecystectomy? Surgical Endoscopy **31**, 142–146 (1 2017)
19. Strasberg, S.M., Hertl, M., Soper, N.J.: An analysis of the problem of biliary injury during laparoscopic cholecystectomy. Journal of the American College of Surgeons **180**, 101–25 (1 1995)
20. Söderlund, C., Frozanpor, F., Linder, S.: Bile duct injuries at laparoscopic cholecystectomy: A single-institution prospective study. acute cholecystitis indicates an increased risk. World Journal of Surgery **29**, 987–993 (8 2005)
21. Wherry, D., Marohn, M., Malanoski, M., et al.: An external audit of laparoscopic cholecystectomy in the steady state performed in medical treatment facilities of the department of defense. Annals of Surgery **224**, 145–154 (8 1996)