This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

CasCVS-Net: A Staged Multi-Task Cascade for Critical View of Safety Assessment

Bock-Zien Toh[1†], Yuanchuan Ren[1†], Tay Aw Yu[1], Ng Khee Ong[1], Zhehua Mao[1,2]✉, and Sophia Bano[1,2]

[1] Department of Computer Science, University College London, London, UK
[2] UCL Hawkes Institute, University College London, UK
[†] These authors contributed equally to this work.
✉ Corresponding author: z.mao@ucl.ac.uk

Abstract. Automated assessment of the Critical View of Safety (CVS) in laparoscopic cholecystectomy requires both recognition of the three CVS criteria and anatomical grounding in small, rare, and often occluded hepatocystic structures. Reliable grounding of this safety-critical anatomy remains the main bottleneck. We propose CasCVS-Net, a staged multi-task cascade that jointly performs object detection, semantic segmentation, and CVS assessment, trained on the Endoscapes dataset. The model couples the tasks through predicted anatomy: predicted boxes guide segmentation, and predicted masks provide region-level features for CVS classification. To reduce optimisation instability in this coupled setting, training progresses from detection through detection–segmentation to the full three-task cascade, followed by task-wise fine-tuning. Evaluation on the public unseen test set shows that CasCVS-Net improves over matched single-task baselines on all three tasks, achieving 32.0 detection mAP, 46.8 semantic mIoU, 15.3 rare-anatomy mIoU, and 67.2 CVS mAP. For CVS, our method achieves the best overall balance between mAP and bAcc, improving mAP over LG-CVS and SV2LSTG by 4.0 and 2.9 points, respectively, while achieving comparable mAP to SwinCVS with substantially higher bAcc (74.7 vs. 70.3). These results show that staged task coupling through predicted boxes and masks improves anatomical grounding for CVS assessment, particularly for rare hepatocystic structures.

Keywords: Critical View of Safety · Staged multi-task learning · Multi-task learning · Surgical scene understanding

1 Introduction

Bile duct injury is among the most serious complications of laparoscopic cholecystectomy, occurring in roughly 0.3–0.7% of cases and carrying high morbidity and cost [4]. Many injuries arise from misidentification of biliary anatomy rather than technical execution alone [5]. The Critical View of Safety (CVS), introduced by Strasberg et al. [3], aims to prevent such errors by requiring three criteria to be met before division of the cystic duct or artery: only the cystic

duct and cystic artery enter the gallbladder (C1), the hepatocystic triangle is cleared (C2), and the lower gallbladder is separated from the cystic plate (C3). Although reliable CVS achievement is associated with safer surgery [6], manual CVS documentation remains inconsistent [7], motivating automated assessment that is both accurate and anatomically grounded.

CVS assessment in deep learning is commonly formulated as three-criterion multi-label prediction, with methods differing in anatomical supervision and representation. Segmentation-based pipelines segment the hepatocystic anatomy before classification [8]; detection- and graph-based methods localise anatomy with bounding boxes and reason over the detected structures [9,10]; other methods detect landmarks or safety zones for intraoperative guidance [11,12]; and more recent transformer- or representation-learning classifiers operate directly at the image level [13,14]. Despite this progress, reliable anatomical grounding remains challenging, especially for the rare and small safety-critical structures such as the cystic plate and hepatocystic triangle, which are difficult to localise and segment. This motivates a detection-guided cascade, in which detection provides coarse localisation cues for segmentation, and segmentation in turn provides an anatomy-aware representation for CVS assessment.

We propose *CasCVS-Net*, a staged multi-task cascade linking object detection, semantic segmentation, and CVS prediction in a shared ResNet-50–FPN network with task-specific residual adapters. The framework is trained progressively: a detection head is trained first, and its predicted boxes are converted into class-aware spatial priors that guide training of the segmentation head; the predicted semantic masks then define the anatomical regions utilized by a mask-pooled CVS head. Because each head is trained on the same predicted boxes and masks it will receive at test time, the cascade needs no ground-truth boxes or masks at inference.

Our key contributions are: (i) a detection-guided segmentation cascade that uses predicted boxes as coarse priors for rare safety-critical anatomy; (ii) a mask-pooled CVS head that extracts region-level features from predicted semantic masks, keeping the training and inference pathways consistent; and (iii) an Endoscapes evaluation that reports detection and segmentation for the rare safety-critical anatomy separately, with ablations isolating staged training, box-to-mask guidance, and mask-to-CVS coupling.

2 Method

Given a laparoscopic frame, the model predicts object detections, a semantic segmentation map, and CVS criterion scores. Detection uses six foreground classes: cystic plate, hepatocystic triangle, cystic artery, cystic duct, gallbladder, and surgical instrument. Segmentation predicts the same six classes plus background. The CVS branch predicts the three criteria C1–C3.

The model is a cascaded multi-task predictor with a shared ResNet-50–FPN encoder [16,17] and task-specific heads (Fig. 1). The task-specific heads consist of Faster R-CNN [18] for detecting five anatomical structures and one surgical

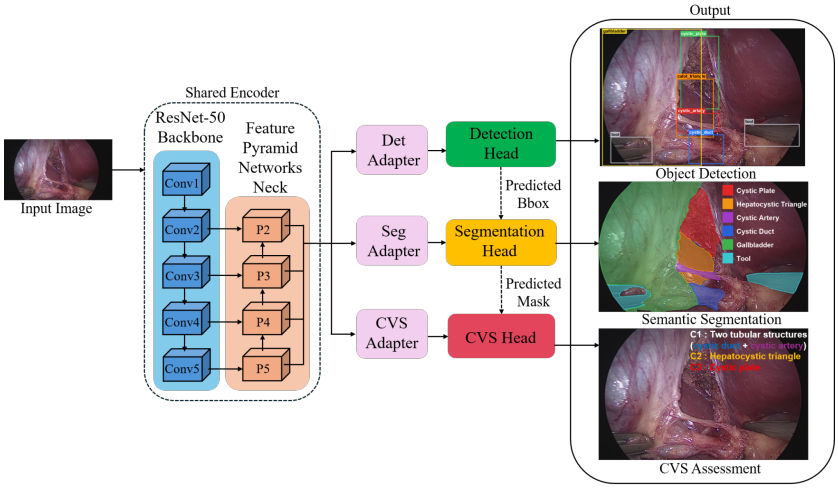


Fig. 1. Overview of the proposed staged cascade. A shared ResNet-50–FPN encoder produces multi-scale features, which are refined by task-specific adapters before the detection, segmentation, and CVS heads. At inference, predicted boxes guide segmentation and predicted masks define regions for CVS feature pooling.

instrument class, DeepLabV3+ [20] for segmentation, and a mask-pooled classifier for CVS assessment. Before each head, a lightweight residual adapter [19] refines every FPN level: a 1×1 then 3×3 convolution with a ReLU between, the final layer zero-initialised and added back to the level's features, so each adapter starts as the identity and learns only a task-specific residual. The three task adapters share this design but keep independent weights. At inference these heads form a cascade: the model maps a frame to boxes, then masks, then CVS scores using only its own predictions.

2.1 Box-to-Mask Coupling

The first cascade coupling converts detector outputs into a spatial prior for segmentation. From the detections $\mathcal{D} = \{(b_i, c_i, s_i)\}_{i=1}^{N}$, with box b_i, predicted class c_i, and confidence s_i, we retain those with $s_i \geq 0.1$, enlarge each box by 10%, and clip it to the image boundaries. The low threshold retains rare structures, while confidence weighting and the zero-initialised projector discount false positives. The enlarged boxes are rasterised into a seven-channel prior B whose background channel stays zero and whose six foreground channels are confidence-weighted:

$$B_c(p) = \max\Big(\{s_i \mid c_i = c,\ p \in b_i'\} \cup \{0\}\Big), \quad c \in \{1, \ldots, 6\}, \tag{1}$$

where b_i' is the enlarged box and p indexes pixels, so same-class overlaps take the maximum confidence while different classes occupy separate channels. In the

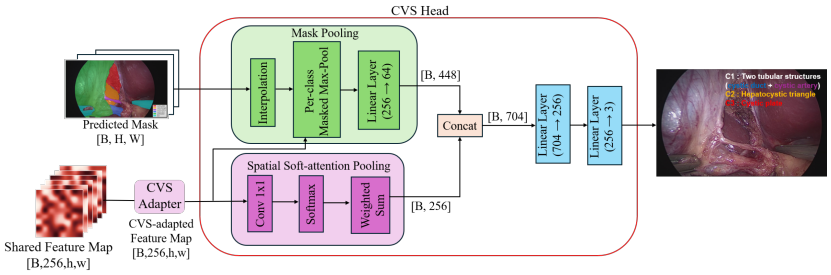


Fig. 2. Mask-pooled CVS head. The predicted semantic map is converted into hard class regions and downsampled to the CVS feature grid. Class-wise mask pooling extracts anatomical embeddings, while a spatial attention branch summarises global context. The concatenated representation is passed through fully connected layers to predict the three CVS criteria. Here B denotes the batch size; the mask branch stacks one 64-dimensional embedding per class ($7 \times 64 = 448$).

joint pretraining stage the values are binary, so this maximum reduces to a union of the overlapping boxes. The prior is resized, passed through a convolutional projector initialised to output zero, and added as a residual to the decoder feature of the DeepLabV3+ segmentation head, which produces per-pixel logits over the seven classes; the prior begins as a no-op and is learned as a correction. Predicted detections are detached before prior construction, so segmentation loss does not backpropagate through the detector.

2.2 Mask-Pooled CVS Head

The second coupling lets the CVS head utilize the predicted anatomy rather than only global image features. The segmentation logits are reduced to a hard label map by a per-pixel arg-max over the seven classes and downsampled to the CVS feature grid. Each class c then defines a region $\mathcal{R}_c$ of grid positions q labelled c over the CVS-adapted feature map F_{cvs}, and its embedding is the per-channel maximum $e_c = \max_{q \in \mathcal{R}_c} F_{\text{cvs}}(q)$ when $\mathcal{R}_c \neq \emptyset$ and $e_c = \mathbf{0}$ otherwise. Stacking the per-class embeddings, one slot per semantic class including background, yields the mask feature.

In parallel, a soft spatial-attention branch pools the same features into a global summary: a 1×1 convolution reduces the feature map to a single-channel score, and a softmax over the grid gives the attention weights that sum the features. The per-class mask embeddings and this summary are concatenated and passed through fully connected layers to produce the three CVS logits.

Because the regions come from a hard arg max semantic map, CVS loss does not backpropagate through mask membership into the segmentation decoder. The coupling therefore uses segmentation as a structured anatomical representation, while CVS supervision updates the CVS branch and any shared features trainable in the current stage.

2.3 Staged Training Strategy

Training the three tasks jointly from scratch is unstable: detection, dense segmentation, and image-level CVS impose conflicting demands on the shared encoder, and a weak detector provides unreliable inputs to the downstream heads. We therefore build the cascade in four stages. **(i) Detection:** we train the Faster R-CNN branch alone to establish the localisation and multi-scale features the cascade depends on. **(ii) Detection–segmentation:** we add the segmentation branch and train it together with detection, warm-starting its head from a pre-trained single-task segmentation model; the FPN, both heads, the adapters, and the late ResNet stages update while the early stages stay frozen. **(iii) Full cascade:** we add the CVS head and train all three tasks together with both couplings active, so each head learns from the predicted boxes and masks it will see at inference (this stage uses the binary box prior of Section 2.1); PCGrad [15] removes the conflicting component of each task gradient on the shared backbone and FPN before the update. **(iv) Fine-tuning:** finally, we fine-tune the heads one at a time in cascade order (detection, segmentation, CVS). Each pass trains only that head and its adapter with the rest frozen, recalibrating it on the now-fixed upstream predictions; the segmentation pass switches the box prior from its binary form to the confidence-weighted form of Eq. 1. This preserves the shared representation.

2.4 Training Objective

Each task head is trained with its own loss; because annotation availability differs by task, each sample contributes only to the losses for which its label is present. Detection uses the Faster R-CNN multi-task loss [18], the sum of the region-proposal-network and detection-head classification (cross-entropy) and box-regression (smooth-L_1) terms, $\mathcal{L}_{\text{det}} = \mathcal{L}^{\text{cls}}_{\text{rpn}} + \mathcal{L}^{\text{reg}}_{\text{rpn}} + \mathcal{L}^{\text{cls}}_{\text{roi}} + \mathcal{L}^{\text{reg}}_{\text{roi}}$. Segmentation uses a pixel-wise cross-entropy over the $K = 7$ classes, averaged over the valid pixels Ω, $\mathcal{L}_{\text{seg}} = -|\Omega|^{-1} \sum_{p\in\Omega} \log \text{softmax}_{y_p}(s_p)$, where $s_p \in \mathbb{R}^K$ are the class logits at pixel p and y_p its ground-truth label. CVS uses a class-balanced binary cross-entropy over the three criteria, $\mathcal{L}_{\text{cvs}} = -\frac{1}{3}\sum_{c=1}^{3}[w_c t_c \log \sigma(z_c) + (1-t_c)\log(1-\sigma(z_c))]$, with criterion logits z_c, soft consensus targets $t_c \in [0,1]$, the sigmoid σ, and inverse-frequency positive weights $w_1 = 5.40$, $w_2 = 7.92$, and $w_3 = 4.59$. The joint cascade stage minimises the fixed weighted sum $\mathcal{L} = 2.0\mathcal{L}_{\text{det}} + 1.5\mathcal{L}_{\text{seg}} + 0.5\mathcal{L}_{\text{cvs}}$, whose weights bring the three raw-loss magnitudes onto a comparable scale (the CVS term is largest, segmentation smallest); each fine-tuning stage instead optimises only its active head's loss.

3 Experiments

Dataset. We evaluate on Endoscapes2023, a publicly available laparoscopic cholecystectomy dataset with aligned annotations for object detection, semantic segmentation, and CVS assessment [1,2]. The dataset contains 201 videos with

an official video-level split of 120 training, 41 validation, and 40 test videos. CVS labels, taken as the consensus of three expert annotators, are available for 11,090 frames, bounding boxes for 1,933 frames, and public semantic masks for the Endoscapes-Seg50 subset. Detection and segmentation use the same six foreground classes; segmentation additionally includes background. The safety-critical anatomy is both rare and hard to localise: on the test split, the cystic plate and hepatocystic triangle appear in only 41.7% and 39.7% of frames, and their detection mAP is 8–19, compared with over 60 for the gallbladder and surgical instrument [1].

Metrics. Detection is evaluated with COCO-style bounding-box mAP@[0.5:0.95]. Segmentation is evaluated with semantic mean intersection-over-union (mIoU) over the seven classes. CVS is evaluated by average precision per criterion, mean AP over C1–C3, and balanced accuracy. Because these structures are the safety-critical bottleneck, we additionally report rare detection mAP and rare segmentation mIoU over the cystic plate and hepatocystic triangle. Since our segmentation metric is semantic mIoU rather than the official Endoscapes instance-segmentation AP, segmentation comparisons are restricted to our own framework-matched baselines.

Baselines. The primary controls are framework-matched single-task learning (STL) baselines for detection, segmentation, and CVS, each using the same ResNet-50–FPN encoder and corresponding task head as the proposed model. For CVS assessment, we compare our method with the best results reported for LG-CVS [9], SV2LSTG [10], and SwinCVS [13] on the public Endoscapes dataset using the official data split. While these results are taken from the respective original papers, the shared dataset and split provide a consistent basis for cross-paper comparison, with the best reported results selected to ensure a conservative evaluation against existing methods.

Implementation details. We implement all models in PyTorch. A shared preprocessing pipeline applies the same geometric transforms to images, boxes, and masks. During training, the input scale is sampled from 854×480, 1024×576, and 1280×720 at a fixed aspect ratio, and evaluation is bounded by 1280×720. Semantic masks are resized by nearest neighbour, images are normalised with ImageNet statistics, and augmentation is horizontal flip, translation, colour jitter, and cutout.

The full cascade trains with SGD (learning rate 0.0015, momentum 0.9, weight decay 5×10^{-4}, batch size 2) for up to 16 epochs, freezing all but the last two ResNet stages and disabling mixed precision; the detection, segmentation, and CVS fine-tuning stages use learning rate 0.0008 for up to 12, 12, and 6 epochs. The validation split is used for checkpoint and CVS-threshold selection, with the held-out test split reserved for final reporting. At inference, Faster R-CNN applies class-wise NMS at 0.5, score threshold 0.05, and at most 100 detections per image; the segmentation prior follows Section 2.1. Runtime and memory profiling are left for future deployment-oriented evaluation.

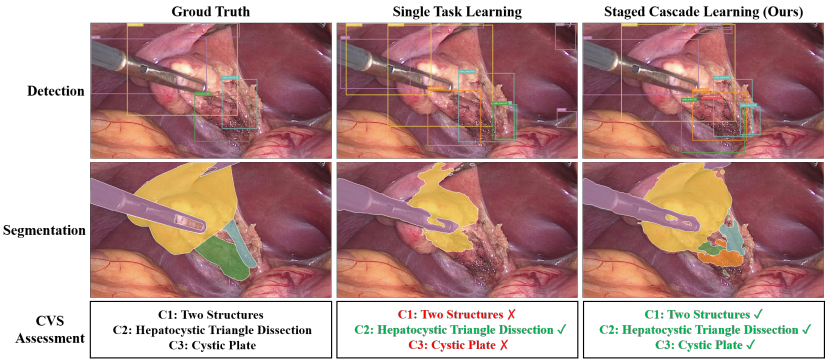


Fig. 3. Qualitative comparison on a test frame. Columns show ground truth, single-task learning (STL), and staged cascade learning (Ours); rows show detection, segmentation, and CVS assessment. Green ✓and red ✗ indicate correct and incorrect CVS predictions, respectively. Ours recovers rare safety-critical anatomy missed by STL and correctly predicts all three CVS criteria.

Table 1. Endoscapes test results (%). DET/SEG: detection mAP and semantic mIoU; rDET/rSEG: rare-class values for cystic plate and hepatocystic triangle; bAcc: balanced accuracy. Bold marks the better framework-matched DET/SEG result and the best CVS result; underlining marks the second-best CVS result.

					CVS	
Method	DET(↑)	rDET(↑)	SEG(↑)	rSEG(↑)	mAP(↑)	bAcc(↑)
LG-CVS [9]	–	–	–	–	63.2	**74.8**
SV2LSTG [10]	–	–	–	–	64.3	73.4
SwinCVS-Frozen [13]	–	–	–	–	**67.5**	70.3
Single-task (STL)	29.6	13.4	38.4	4.6	47.9	66.9
Ours	**32.0**	**16.4**	**46.8**	**15.3**	<u>67.2</u>	<u>74.7</u>

4 Results and Discussion

Comparison with single-task and published methods. Table 1 reports test results. Against single-task learning, the cascade raises detection mAP from 29.6 to 32.0, rare-anatomy mIoU from 4.6 to 15.3, and CVS mAP from 47.9 to 67.2. Its detection result lies between the published Deformable-DETR (32.7) and Faster R-CNN (30.0) results [1]. For CVS, CasCVS-Net improves mAP over LG-CVS and SV2LSTG by 4.0 and 2.9 points, respectively, while remaining within 0.3 points of SwinCVS-Frozen. Meanwhile, its bAcc of 74.7 is within 0.1 points of LG-CVS and exceeds SV2LSTG and SwinCVS-Frozen by 1.3 and 4.4 points, respectively, demonstrating the best overall balance between the two metrics. Fig. 3 qualitatively compares CasCVS-Net with the matched single-task baseline on a test frame.

Staging ablation. Table 2 compares the staged schedule against single-stage joint training of the same network, with identical couplings and PCGrad, from scratch. Without staging, detection fails to train (3.9 mAP), leaving the down-

Table 2. Staging ablation on the test set (%).

Training	DET(↑)	SEG(↑)	rSEG(↑)	CVS(↑)
Staged cascade (Ours)	**32.0**	**46.8**	**15.3**	**67.2**
Single-stage joint	3.9	27.2	0.1	28.0

Table 3. Cascade ablation on the test set (%); best result per column in bold.

Configuration	DET(↑)	SEG(↑)	rSEG(↑)	CVS(↑)
Full (both couplings)	32.0	46.8	15.3	**67.2**
without box-to-mask	**32.2**	48.0	14.4	62.1
without mask-to-CVS	29.6	**48.4**	**15.7**	55.8

stream tasks with unreliable localisation cues; rare-anatomy segmentation falls to 0.1 mIoU and CVS to 28.0 mAP. Training detection first is therefore essential for stable cascade training.

Cascade ablation. Table 3 ablates each coupling with the remaining training recipe fixed. Removing box-to-mask reduces rare-anatomy segmentation from 15.3 to 14.4 mIoU and CVS from 67.2 to 62.1 mAP, despite slightly higher aggregate segmentation (48.0 vs 46.8 mIoU). Removing mask-to-CVS lowers CVS further to 55.8 mAP despite the highest rare-anatomy mIoU (15.7), showing that segmentation alone is insufficient without explicit anatomy coupling. Mask-to-CVS is therefore the main contributor to CVS performance, while both couplings remain complementary.

5 Conclusion

CasCVS-Net couples detection, segmentation, and CVS assessment through predicted boxes and masks. The staged cascade achieves the best overall performance across the three tasks, including an improvement in rare-anatomy segmentation from 4.6 to 15.3 mIoU. Its CVS mAP of 67.2 exceeds LG-CVS and SV2LSTG by 4.0 and 2.9 points, respectively, and is within 0.3 points of SwinCVS-Frozen, while achieving substantially higher bAcc than SwinCVS-Frozen (74.7 vs. 70.3). The ablations identify mask-to-CVS coupling as the main contributor to CVS improvement.

The staged training strategy combines PCGrad, temporary freezing, and task-wise fine-tuning to mitigate task conflicts, preserve learned representations, and adapt individual branches, without increasing inference complexity. Although the overall results support the effectiveness of this strategy, finer-grained ablations could further clarify the contribution of each component. As the end-to-end evaluation uses predicted boxes and masks, it reflects the effects of upstream errors under normal inference, while future work will further investigate the individual training components and quantify the model's sensitivity to specific upstream error patterns.

Acknowledgments

This work was supported in whole, or in part, by the EPSRC Optical and Acoustic imaging for Surgical and Interventional Sciences (OASIS) [UKRI145]; the Department of Science, Innovation and Technology (DSIT) and the Royal Academy of Engineering under the Chair in Emerging Technologies programme; NIHR UCLH Biomedical Research Centre [NIHR203328]; EPSRC Robotic Assisted Imaging [EP/P012841/1]; ESRC Microrobotics [ES/T011866/1]; Horizon 2020 FET Open (EndoMapper) [863146].

Disclosure of Interests

The authors have no competing interests to declare.

References

1. 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 preprint arXiv:2312.12429 (2023)
2. Mascagni, P., Alapatt, D., Murali, A., et al.: Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. Scientific Data **12**(1), 331 (2025)
3. 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**(1), 101–125 (1995)
4. Brunt, L.M., Deziel, D.J., Telem, D.A., et al.: Safe cholecystectomy multi-society practice guideline and state-of-the-art consensus conference on prevention of bile duct injury during cholecystectomy. Surgical Endoscopy **34**(7), 2827–2855 (2020)
5. Way, L.W., Stewart, L., Gantert, W., et al.: 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), 460–469 (2003)
6. Manatakis, D.K., Antonopoulou, M.I., 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**(3), 640–648 (2023)
7. Mascagni, P., Fiorillo, C., Urade, T., et al.: Formalizing video documentation of the critical view of safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety. Surgical Endoscopy **34**(6), 2709–2714 (2020)
8. Mascagni, P., Vardazaryan, A., Alapatt, D., et al.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery **275**(5), 955–961 (2022)
9. Murali, A., Alapatt, D., Mascagni, P., et al.: Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging **43**(3), 1247–1258 (2024)
10. Murali, A., Alapatt, D., Mascagni, P., et al.: Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning. In: MICCAI 2023. LNCS, vol. 14228, pp. 647–657. Springer (2023)

11. Madani, A., Namazi, B., Altieri, M.S., et al.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of Surgery **276**(2), 363–369 (2022)
12. Tokuyasu, T., Iwashita, Y., Matsunobu, Y., et al.: Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. Surgical Endoscopy **35**(4), 1651–1658 (2021)
13. Nowak, F.M., Mazomenos, E.B., Davidson, B., Clarkson, M.J.: SwinCVS: a unified approach to classifying critical view of safety structures in laparoscopic cholecystectomy. International Journal of Computer Assisted Radiology and Surgery **20**(6), 1145–1152 (2025)
14. Baby, B., Srivastav, V., Jain, P.P., et al.: Multi-modal representations for fine-grained multi-label critical view of safety recognition. In: MICCAI 2025. LNCS, vol. 15970, pp. 423–432. Springer (2025)
15. Yu, T., Kumar, S., Gupta, A., et al.: Gradient surgery for multi-task learning. In: NeurIPS. vol. 33, pp. 5824–5836 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NeurIPS. pp. 91–99 (2015)
19. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NeurIPS. pp. 506–516 (2017)
20. Chen, L.C., Zhu, Y., Papandreou, G., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 833–851 (2018)