EdgeCVS: Democratization of surgical AI with a Distilled Edge-Deployable Critical View of Safety (CVS) model

Amine Yamlahi[1,2,5], Jakob Hennighausen[1], Pascal Hansen[1,3,4], David Leeb[1], and Lena Maier-Hein[1,2,3,4,5,6]

[1] Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany m.elyamlahi@dkfz-heidelberg.de

[2] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and university medical center Heidelberg

[3] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany

[4] Faculty of Mathematics and Computer Science, Heidelberg University, Germany

[5] Medical Faculty, Heidelberg University, Germany

[6] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

Abstract. Critical View of Safety (CVS) assessment has become a central safeguard against bile duct injury during laparoscopic cholecystectomy (LC). Current deep learning-based solutions achieve strong performance but rely on massive backbone architectures (up to >1B parameters) that hinder real-world deployment, particularly in resource-constrained operating rooms. We challenge the assumption that such scale is necessary. Because CVS assessment is confined to a specific surgical phase and anatomical region, we hypothesize that its visual representation occupies a more compact region of feature space. Guided by the hypothesis that this reduced semantic dispersion enables substantial model compression without sacrificing accuracy, we introduce **EdgeCVS**, a phase-gated distillation framework that compresses a high-capacity teacher into **EdgeCVS-5M**, a 5M-parameter edge-efficient student. By incorporating anatomical spatial priors and phase-gated pseudo-labels, the compact model recovers complex decision boundaries. EdgeCVS-5M achieves mAP close to the most recent state-of-the-art models while reducing the parameter count to 1.6%, enabling real-time inference on CPUs. These results demonstrate that accurate and deployable edge-based surgical AI solutions are feasible, advancing the democratization of surgical AI. Code available at https://github.com/IMSY-DKFZ/edgecvs.

Keywords: Knowledge distillation · Critical View of Safety · Edge Computing

1 Introduction

With more than 750,000 procedures performed annually in the United States alone [19,20], laparoscopic cholecystectomy ranks among the most commonly performed surgical interventions worldwide. Despite its routine nature, it remains associated with a significant risk of bile duct injury (BDI) [3]. To mitigate BDIs, surgeons visually assess the Critical View of Safety (CVS) criteria prior to dissection. However, this real-time visual assessment is highly subjective and ambiguous, necessitating automated AI solutions for consistent evaluation.

Recent benchmarks like Endoscapes [11] and the MICCAI Surgical Lighthouse SAGES-CVS challenge [1] have introduced CVS as a classification task based on anatomical visibility to support surgeons in decision-making. This has spurred the exploration of diverse architectures, including spatio-temporal models [13], vision-language models [2], and Multi-Task Learning (MTL) approaches [10] or Graph Neural Network methods such as LG-CVS [12] and LG-DG [18] that leverage the spatial annotations available in a subset of the Endoscapes dataset. Among these, using large backbones such as DINOv2 ViT-Giant [14] and EVA02-Large [5] in a multi-task fashion, as seen in the winning solutions of the top 3 teams in the CVS challenge [1], emerge as highly performant methods for CVS detection. However, these current state-of-the-art (SOTA) methods present a significant deployment barrier: their massive backbones require high-end hardware and technical infrastructure not currently available in standard hospitals. To truly democratize surgical AI, solutions must target the edge-device "sweet spot" of 2.5 to 10 million parameters [6], ensuring real-time inference on standard, accessible hardware. We address this challenge with the following contributions. First, guided by the hypothesis that low diversity enables substantial model compression, we demonstrate that the visual space underlying CVS exhibits significantly lower semantic dispersion compared to general surgical tasks, owing to its strict phase restriction and anatomical focus. Second, by strategically enriching supervision with task-aligned pseudo-labels and spatial priors (segmentation masks), compact models can approximate—and effectively recover—the decision boundaries of much larger networks. These findings culminate in **EdgeCVS**, a phase-gated distillation framework designed to compress massive, hardware-intensive teacher models into a deployable 5M-parameter edge-efficient student that runs in real time on standard CPU hardware (see Figure 1).

2 Methods

2.1 EdgeCVS

Our approach was guided by the following key design decisions:

MTL: Because CVS assessment is inherently defined by anatomical visibility, we incorporated segmentation as an auxiliary task to enforce spatial priors and embed the geometry of the hepatocystic triangle into the model's latent space.

Distillation at the core: Rather than training a compact model from scratch, we centered the framework around transferring structured knowledge from a high-capacity teacher. We adopted a dual-head distillation strategy, transferring both classification logits and segmentation outputs to preserve geometric

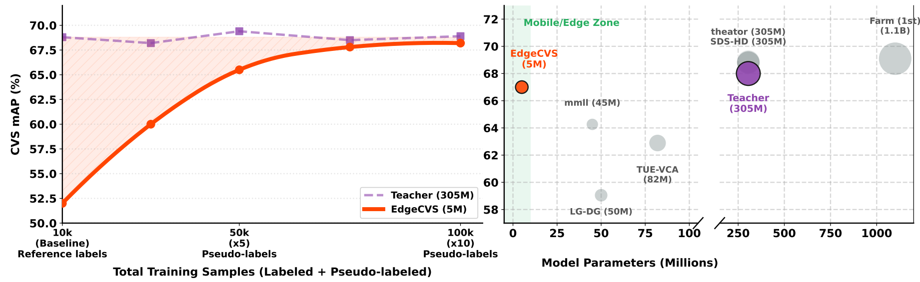

Fig. 1. Targeted signal enrichment enables a compact edge model to approximate a massive teacher. (a) While the 305M-parameter teacher quickly reaches a performance ceiling, the 5M edge model continues to improve as task-aligned pseudo-labels are introduced. (b) Comparison with top-performing teams from the SAGES-CVS challenge 2024: By trading model capacity for structured supervision and manifold densification, the compact model approaches SOTA mAP while using 98.4% fewer parameters, establishing an efficiency frontier for CPU-deployable CVS estimation. The methods are plotted at the parameter count of their single largest ensemble member.

and semantic consistency. This allows the edge model to inherit refined decision boundaries learned in a richer representational space.

Pseudo-labeling of public data: We leveraged unlabeled frames from publicly available surgical datasets to generate pseudo-labels for both the classification and segmentation tasks. By restricting pseudo-label generation to the Calot's triangle dissection (CTD) phase, we ensured semantic alignment and maximized the usefulness of additional supervision for compact model training. The full training pipeline is depicted in Figure 2.

Implementation Details: Our implementation is based on our SAGES-CVS Challenge 2024 submission [1], where our team (SDS-HD) ranked 1st in the CVS Robustness category with a model demonstrating strong generalization across domain shifts between data subsets, including country and device type. We adopted this model, based on an EVA02-Large backbone (305M parameters), as our high-capacity teacher. The MTL networks use a linear layer as the primary classification head for CVS criteria prediction and an auxiliary Atrous Spatial Pyramid Pooling (ASPP) [4] decoder for anatomical segmentation. Unlike methods where spatial features feed back into the encoder (*e.g.*, DeepCVS, LG-DG), our decoupled segmentation head contributes only the auxiliary segmentation loss (Dice plus cross-entropy) to the weighted training objective, which also includes binary cross-entropy for classification, and is discarded during inference to maximize speed. For the distillation process, pseudo-labels were generated as soft labels to capture the teacher's uncertainty. To reduce prediction variability, we averaged the teacher's logits across multiple intermediate training epochs following the approach detailed in [23].

To achieve edge deployment, the student model is instantiated using the EdgeNeXt-Small [8] backbone (5M parameters). All models were trained using the AdamW [7] optimizer with a base learning rate of 10^{-5} and a batch size of 16

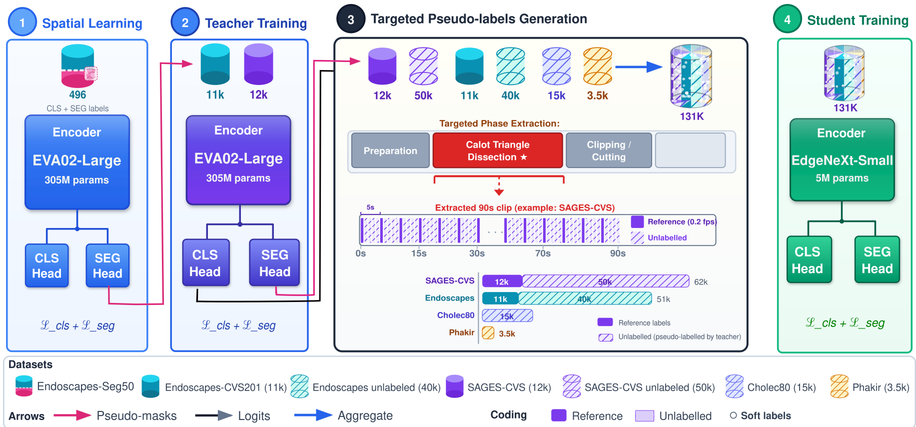

Fig. 2. The EdgeCVS pipeline compresses a massive teacher into an edge-efficient student in four stages. (1) Spatial Learning: A segmentation model is first trained in a MTL setup on the Endoscapes-Seg50 subset, which contains human reference annotations for both anatomical segmentation (SEG) and classification (CLS), enabling the model to learn task-specific geometry. **(2) Teacher Training:** The model from Stage 1 is then used to generate SEG pseudo-labels for Endoscapes-CVS201 and SAGES-CVS, allowing full multi-task training of the EVA02-Large teacher using both classification labels and segmentation supervision. **(3) Targeted Pseudo-Labeling:** The trained teacher is applied to strictly phase-gated unlabeled frames extracted from the CTD phase of publicly available datasets to generate dense, task-aligned soft classification and segmentation targets. **(4) Student Distillation:** The dual-head knowledge is transferred to a 5M-parameter student, enabling recovery of the teacher's decision boundaries within a constrained parameter budget.

and 64 for the teacher and student, respectively. Teacher pre-training and student distillation were conducted for 20 epochs, utilizing cosine annealing learning rate schedules. All models process inputs at a resolution of 448×448. Training and GPU inference evaluations were performed on an NVIDIA RTX 4090 (24 GB), while edge-simulated CPU inference latency was benchmarked on an AMD Ryzen 9 7950X 16-Core processor.

2.2 Validation Methodology

The purpose of the experiments was to investigate the following research questions (RQs). **RQ1:** Does CVS estimation occupy a more compact feature space due to its phase-gated and anatomy-specific nature? **RQ2:** Can reduced model size be traded for targeted signal enrichment in this low-dimensional surgical task? **RQ3:** How much task-relevant data is required for a compact edge model to match the performance of a large teacher network?

Datasets: In total, we utilize 1,288 videos across four cholecystectomy datasets: SAGES-CVS [1] (1,000 videos), Endoscapes-CVS201 [11,9] (201 videos), Cholec80-

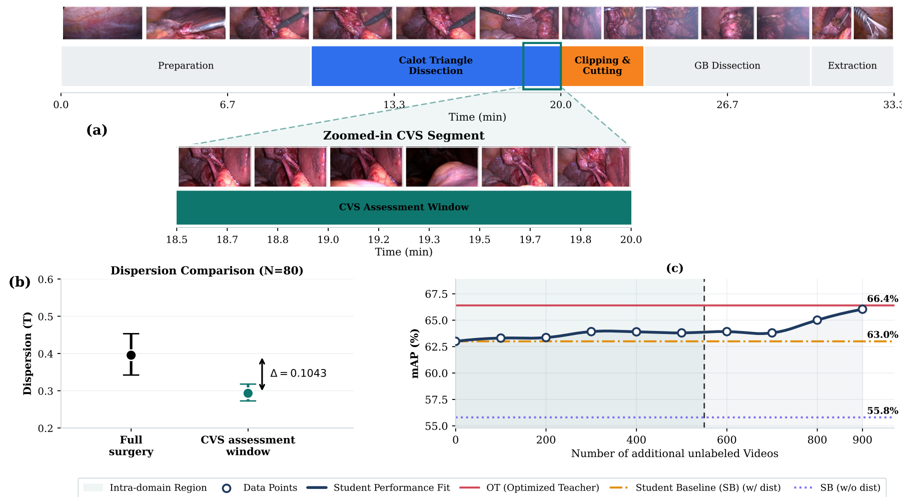

Fig. 3. CVS semantic dispersion and student data scaling. CVS assessment is strictly confined to a short transitional window during laparoscopic cholecystectomy. **(a) Evaluation Segment:** This strict phase constraint isolates the visual field to the hepatocystic triangle, drastically limiting scene diversity. **(b) Semantic Dispersion:** By sampling one frame per video ($N = 80$) to eliminate temporal leakage, the Mean Pairwise Distance (MPD) of high-level semantic embeddings demonstrates that phase-gated CVS frames (SAGES-CVS) are tightly clustered and significantly less visually diverse than full-procedure videos (Cholec80). **(c) Parameter-Data Exchange Rate:** We trace the mAP of the 5M-parameter student on the SAGES-CVS benchmark as the training pool is incrementally densified with phase-gated pseudo-labels.

CVS [16] (80 videos), and Phakir [17] (7 videos). While SAGES-CVS and Endoscapes-CVS201 provide task-related CVS labels, these are sparsely annotated (0.2 FPS). To bridge this gap, we adopt a 1 FPS sampling rate, where labeled frames constitute only 20% of the total sequence. The remaining 80% of unlabeled frames—along with the CTD phases from Cholec80-CVS and Phakir—are leveraged for dense unsupervised training via pseudo-labeling. To ensure reproducibility, we follow the official dataset splits and adopt the official SAGES-CVS challenge metrics mean Average Precision (mAP) and Brier score for evaluation.

Models were trained on the combined SAGES-CVS and Endoscapes datasets, with all final results reported on the official SAGES-CVS test set. To evaluate data scaling (RQ3) without cross-domain confounding, teacher baselines were trained strictly on the SAGES-CVS target dataset.

Experimental design: RQ1. While prior work [15] has used intrinsic image dimension to quantify visual diversity in static image datasets, this concept cannot be straightforwardly extended to videos in a statistically sound manner due to the strong temporal correlation between adjacent frames. To quantify the visual diversity of the CVS task, we instead analyzed the semantic dispersion of CVS frames by repeating the following process for 1,000 bootstrap iterations on the Cholec80 dataset. First, we extracted exactly one random frame per unique video ($N = 80$ in total). For each frame, we collected embeddings from the penultimate linear layer on an EVA02-Large model initialized with ImageNet pretrained weights, effectively mapping each image into the teacher's expressive latent feature space. To measure semantic dispersion, we computed the Mean Pairwise Cosine Distance (MPD) across the $N = 80$ embeddings for both the phase-gated CVS frames and the full-procedure Cholec80 frames. Finally, to assess statistical robustness, we estimated 95% confidence intervals for the MPD via bootstrap resampling (1,000 iterations) and evaluated the significance of the observed dispersion difference (Δ) using a permutation test with 1,000 shuffles.

RQ2. To investigate RQ2, we leveraged models of identical backbones with different sizes from the EVA02 family (Tiny, Small, Base, Large), spanning approximately 5M to 305M parameters. We progressively injected four complementary information streams to analyze capacity-dependent learning dynamics across model scales.

Baseline: We first trained the model using only the reference CVS classification labels to establish a baseline.

Spatial prior (+SP): Because CVS assessment depends on anatomical visibility, we incorporated segmentation pseudo-masks within an MTL framework. This encourages the model to encode the geometry of the hepatocystic triangle, introducing a structured spatial inductive bias. Specifically, we leveraged our model trained in step 1 (Figure 2) to generate segmentation masks on both Endoscapes and SAGES-CVS datasets.

Teacher Supervision (+TS, Same Data): To isolate the impact of signal quality from data volume, we distilled the teacher's soft predictions—both classification logits and segmentation outputs—into the compact model using only the original labeled dataset. This serves as a standard distillation control group, testing whether transferring the teacher's nuanced confidence scores on already-seen data is sufficient to close the performance gap.

Manifold Densification (+MD, New Data, Increased Density): To test the effect of data scaling, we expanded beyond the labeled set by leveraging a dense pool of unlabeled frames extracted specifically from the CTD phase, corresponding to 900 cholecystectomy training videos. By having the teacher generate targeted soft labels on this massive pool—across SAGES-CVS, Endoscapes, Cholec80, and Phakir—we increased task-aligned sample density. This distinguishes +MD from +TS by forcing the compact model to explicitly trade this new data volume for its reduced parameter capacity.

RQ3. To investigate RQ3, we determined performance as a function of training data volume, shown in Figure 3 (c). First, to densify the manifold without external bias, we progressively added phase-gated, unlabeled videos strictly from the target dataset (intra-domain). Once this intra-domain pool was exhausted, we incrementally padded the training set with pseudo-labeled external videos from

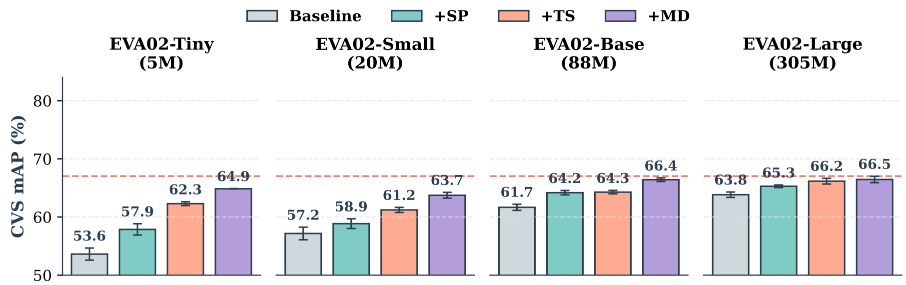

Fig. 4. The compact edge model progressively recovers teacher-level performance through targeted signal enrichment. Mean Average Precision (mAP) is evaluated across transformer backbones ranging from 5M to 305M parameters. While large models saturate early and show minimal benefit from additional supervision, the 5M edge model exhibits pronounced data plasticity. Sequentially adding spatial priors (+SP), teacher supervision (+TS), and phase-gated manifold densification (+MD) enables the compact model to systematically close the capacity gap and ultimately approaches the teacher's performance ceiling. Error bars denote the standard deviation over three random seeds.

the remaining datasets (Endoscapes, Cholec80, and Phakir), with Endoscapes treated purely as a pseudo-labeled source.

3 Results

RQ1. Frames spanning the full surgical workflow in Cholec80 exhibit high semantic dispersion (see Figure 3), with an MPD of 0.39. In contrast, frames strictly confined to the phase-gated CVS assessment window in SAGES-CVS show substantially lower visual diversity, with an MPD of 0.29. This marked reduction in dispersion ($\Delta = 0.1043$) provides empirical evidence consistent with the hypothesis that the CVS task occupies a significantly more concentrated region of the visual manifold compared to general surgical video data.
RQ2. Figure 4 shows that under baseline supervision, performance scales with parameter count: EVA02-Large achieves 63.8% mAP, EVA02-Base 61.7%, EVA02-Small 57.2%, and EVA02-Tiny 53.6%. Incorporating spatial priors via MTL (+SP) improves all models; yet the large models plateau at this stage, showing lower gains with teacher supervision (+TS) or manifold densification (+MD). In contrast, the compact models continue to benefit from additional signals, ultimately approaching the peak performance of the larger backbones.
RQ3. As shown in Figure 3 (c), the student improves steadily as the target-domain pool is densified, with a further gain once out-of-domain videos are added, reaching a peak of 66.3%. This suggests that both intra-domain densification and cross-source diversity contribute to closing the capacity gap.

EdgeCVS benchmarking: Table 1 benchmarks EdgeCVS against the top-performing teams from the SAGES-CVS Challenge and previously published

Table 1. EdgeCVS yields competitive performance on the test set while reducing the number of model parameters by two orders of magnitude. **Seg.**: auxiliary segmentation supervision. **Ens.**: number of models ensembled. Latency (FPS) measured on NVIDIA RTX 4090 (GPU) and AMD Ryzen 9 7950X 16-Core (CPU). †Latency reported for a single model checkpoint; ensemble inference scales linearly. Parameter counts for external teams are approximate and estimated from their reported architectures; ranges reflect uncertainty in component-level configurations.

Method / Team	**Params**	**Seg.**	**Ens.**	**CVS (mAP %) SAGES**	**Brier Score (Avg) ↓**	**Latency (FPS)† GPU**	**CPU**
SAGES-CVS Challenge (Top Teams)							
1. Farm (DINOv2-G)	∼1.1B+		7x	69.1	0.058	-	-
2. theator (EVA02-L)	∼305M		6x	68.9	0.022	-	-
3. SDS-HD (EVA02-L) (Ours)	∼305M	✓	5x	68.8	0.024	-	-
4. mmll (RN50+GNN)	∼50–60M	✓	3x	64.3	0.044	-	-
5. TUE-VCA (PVT-v2)	∼82M		5x	62.9	0.052	-	-
Baseline							
LG-DG [18,1]	∼30–44M	✓	-	59.1	0.124	-	-
Ours							
EVA-CVS (Teacher, Single)	*305M*	✓	-	*66.9*	*0.039*	*32*	*1.4*
EVA-CVS-Ens (Teacher, 5× Ensemble)	*305M*	✓	*5x*	*68.2*	*0.038*	-	-
EdgeCVS-5M (Single)	**5M**	✓	-	**66.5**	**0.029**	**230**	**40**
EdgeCVS-25M (5× Ensemble)	**5M**	✓	**5x**	**67.6**	**0.030**	-	-

Endoscapes baselines. Our single 5M-parameter model achieves 66.5 mAP on SAGES-CVS, nearly matching the single 305M-parameter teacher (66.9 mAP) while using 98.4% fewer parameters. Since challenge teams report only ensemble results, we also report a 5-model EdgeCVS ensemble (25M total) for direct comparison, reaching 67.6 mAP — within 1.5 points of the winning entry (>1B total parameters) and competitive with the 305M teacher ensemble (68.2 mAP). Notably, EdgeCVS runs at 40 FPS on CPU, while all challenge-winning solutions operate below 1.4 FPS.

4 Discussion

In this work, we demonstrate that real-time CPU-based surgical AI for automated CVS assessment is feasible through the introduction of EdgeCVS, a highly efficient dual-modality distillation framework. By demonstrating that the phase-gated CVS task occupies a compact visual manifold with low semantic dispersion, we challenge the prevailing reliance on massive, GPU-dependent architectures. Through the targeted integration of spatial priors and phase-gated manifold densification via pseudo-labeling, we effectively replaced raw parameter capacity with task-aligned data density. Our lightweight 5M-parameter model recovers the complex decision boundaries of a 305M-parameter teacher, achieving close to SOTA-competitive performance while uniquely enabling real-time inference (40 FPS) on standard CPU hardware.

Our analysis reveals a fundamental paradox in CVS assessment: although the task is semantically challenging, it occupies a visual manifold with low semantic dispersion. This property can be leveraged for efficient compression. In large-scale tasks such as ImageNet, compressing a teacher model with more than 100 million parameters into a student with fewer than 10 million typically leads to a persistent 5–10% performance gap [22,21]. In contrast, EdgeCVS reduces parameter counts by two orders of magnitude, demonstrating that extreme compression is feasible in phase-gated medical tasks when the low-dimensional manifold is sufficiently densified. Semantic dispersion is quantified in the teacher's pretrained feature space, consistent with the representation used for knowledge distillation. EdgeCVS is intentionally designed for phase-gated tasks with restricted visual variability. Future work will, however, extend this phase-gated distillation framework to other surgical and general computer vision tasks that share similar temporal and spatial constraints.

In conclusion, our findings demonstrate that model scaling requirements are fundamentally task-dependent. By linking visual manifold structure to compression potential, we show that accurate, real-time, CPU-based surgical AI is achievable without reliance on massive architectures. This work lays the foundation for a broader paradigm of geometry-aware, edge-deployable medical AI systems, and its low hardware barrier opens the door to prospective clinical studies in real-world operating rooms.

Acknowledgments. Funding was provided by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health" and at the Surgical AI Hub Germany and the National Center for Tumor Diseases (NCT), Heidelberg's Surgical Oncology Program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alapatt, D., Eckhoff, J., Lyu, Z., Ban, Y., Mazellier, J.P., Choksi, S., Yang, K., Chiang, P.H., Zorzetti, N., Cannas, S., et al.: The sages critical view of safety challenge: A global benchmark for ai-assisted surgical quality assessment. arXiv preprint arXiv:2509.17100 (2025)
2. Baby, B., Srivastav, V., Jain, P.P., Yuan, K., Mascagni, P., Padoy, N.: Multi-modal representations for fine-grained multi-label critical view of safety recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–432. Springer (2025)
3. Burns, R., Connor, K.L., Guest, R.V., Johnston, C.C., Harrison, E.M., Wigmore, S.J., Sherif, A.E.: Risk factors and mitigating measures associated with bile duct injury during cholecystectomy: meta-analysis. BJS open **9**(4), zraf076 (2025)
4. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017), `http://arxiv.org/abs/1706.05587`
5. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024)
6. Koonce, B.: Mobilenetv3. In: Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, pp. 125–144. Springer (2021)

7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
8. Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Shahbaz Khan, F.: Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In: European conference on computer vision. pp. 3–20. Springer (2022)
9. Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. Scientific Data **12**(1), 331 (2025)
10. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N.: 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). https://doi.org/10.1097/SLA.0000000000004351, epub 2020 Nov 16
11. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., 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)
12. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N.: Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging **43**(3), 1247–1258 (2023)
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. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: arXiv preprint arXiv:2304.07193 (2023)
15. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The intrinsic dimension of images and its impact on learning. arXiv preprint arXiv:2104.08894 (2021)
16. Ríos, M.S., Molina-Rodriguez, M.A., Londoño, D., Guillén, C.A., Sierra, S., Zapata, F., Giraldo, L.F.: Cholec80-cvs: An open dataset with an evaluation of strasberg's critical view of safety for ai. Scientific Data **10**(1), 194 (2023)
17. Rueckert, T., Rauber, D., Maerkl, R., Klausmann, L., Yildiran, S.R., Gutbrod, M., Nunes, D.W., Moreno, A.F., Luengo, I., Stoyanov, D., et al.: Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the phakir 2024 challenge. Medical Image Analysis p. 103945 (2026)
18. Satyanaik, S., Murali, A., Alapatt, D., Wang, X., Mascagni, P., Padoy, N.: Optimizing latent graph representations of surgical scenes for unseen domain generalization. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1243–1250 (2024)
19. Shaffer, E.A.: Epidemiology of gallbladder stone disease. Best practice & research Clinical gastroenterology **20**(6), 981–996 (2006)
20. Stinton, L.M., Shaffer, E.A.: Epidemiology of gallbladder disease: cholelithiasis and cancer. Gut and Liver **6**(2), 172–180 (2012)

21. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
22. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: European conference on computer vision. pp. 68–85. Springer (2022)
23. Yamlahi, A., Kalinowski, P., Godau, P., Younis, R., Wagner, M., Müller, B., Maier-Hein, L.: Smarter self-distillation: Optimizing the teacher for surgical video applications. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 522–531. Springer (2025)