Original Article | Open Access

Artificial Intelligence Surgery

Tozzi *et al. Art Int Surg.* 2026;6:300-19 **DOI:**10.20517/ais.2025.122

Toward a standardized methodological framework for developing computer vision models in staging laparoscopy

Francesca Tozzi[1,#], Seyed Amir Mousavi[2,3,#], Robbe De Muynck[1,#], Dario Quintini[4,5], Adris Molnar[1], Femke Van Vaerenbergh[1], Xander De Lille[1], Matthias Van Liefferinge[1], Wim Ceelen[1,6], Wouter Willaert[1,6], Wesley De Neve[2,3], Niki Rashidian[1,7]

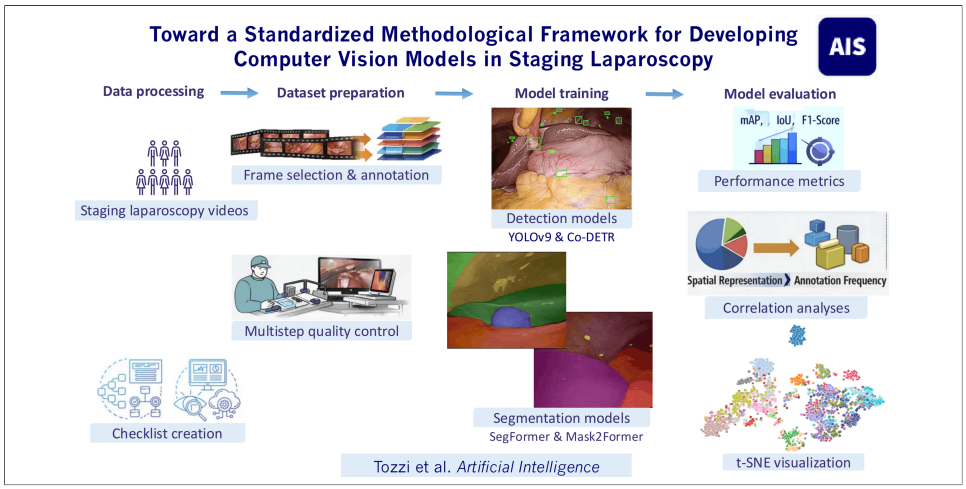
Keywords:
Anatomical detection, anatomical segmentation, annotation strategy, computer vision, deep learning, laparoscopic staging, peritoneal metastases

Citation: Tozzi F, Mousavi SA, De Muynck R, Quintini D, Molnar A, Van Vaerenbergh F, De Lille X, Van Liefferinge M, Ceelen W, Willaert W, De Neve W, Rashidian N. Toward a standardized methodological framework for developing computer vision models in staging laparoscopy. *Art Int Surg.* 2026;6:300-19. https://dx.doi.org/10.20517/ais.2025.122

Received: 31 Dec 2025
First Decision: 10 Mar 2026
Revised: 8 Apr 2026
Accepted: 22 Apr 2026
Published: 5 Jun 2026

Academic Editor:
Takeaki Ishizawa
Copy Editor:
Xing-Yue Zhang
Production Editor:
Xing-Yue Zhang

Check for updates



Abstract

Aim: To evaluate deep learning models for anatomical structure and peritoneal metastasis (PM) detection and segmentation during staging laparoscopy (SL) using a phase-independent dataset, and to quantify how annotation strategy and spatial representation relate to predictive performance.

Methods: A checklist covering 25 anatomical structures, one surgical instrument, and PM was defined. Detection models (YOLOv9, Co-DETR) and segmentation models (SegFormer, Mask2Former) were trained under two label configurations. Videos were split at the video level (60/20/20). To quantify annotation distribution and spatial representation, two class-level descriptors were derived from the training set: object count and area fraction (percentage of image area occupied by each class). Class-level associations between these descriptors and test-set performance [F1-score, Intersection over Union (IoU)] were evaluated using Spearman correlation.

Results: Thirty SL videos yielded 2,309 annotated frames (1,304/433/572 for training/

[1]Department of Human Structure and Repair, Ghent University, Ghent 9000, Belgium.
[2]Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon 21985, Republic of Korea.
[3]IDLab, Ghent University, Ghent 9052, Belgium.
[4]Gastroenterology and Endoscopy Unit, IRCCS ISMETT, Palermo 90127, Italy.
[5]Department of Precision Medicine in Medical, Surgical and Critical Care (MEPRECC), University of Palermo, Palermo 90127, Italy.
[6]Department of Gastrointestinal Surgery, Ghent University Hospital, Ghent 9000, Belgium.

[7]Department of General, Hepatobiliary Surgery and Liver Transplantation, Ghent University Hospital, Ghent 9000, Belgium.
[#]Authors contributed equally.

Correspondence to: Prof. Niki Rashidian, Department of General, Hepatobiliary Surgery and Liver Transplantation, Ghent University Hospital, Ghent 9000, Belgium. E-mail: nikdokht.rashidian@ugent.be

validation/testing). YOLOv9 reached mean mAP@50 of 0.52 and 0.61; Mask2Former achieved mean IoU of 0.51 and 0.61 and F1-scores of 0.65 and 0.73 for Sets A and B, respectively. Despite 4,094 annotations, PM remained difficult to segment (IoU 0.29-0.30; F1-score 0.45-0.46), due to low area fraction and high heterogeneity. For IoU, area fraction showed stronger correlations with performance than object count (ρ up to 0.66 *vs.* 0.48). Similar differences were observed for F1-score.

Conclusions: Anatomical detection and segmentation during SL are feasible but limited by small-target representation and heterogeneous intra-abdominal context. Spatial representation is more closely associated with segmentation performance than annotation frequency, supporting annotation strategies that address sparse pixel coverage in phase-independent intra-abdominal models.

INTRODUCTION

Laparoscopic videos contain a considerable amount of visual data. When correctly processed, these data have the potential to enhance clinical decision-making and surgical performance[1]. Machine learning (ML), a subfield of artificial intelligence (AI), employs algorithms that learn from patterns to support decision-making[2]. Specifically, modern-day computer vision (CV) utilizes ML techniques to locate and identify objects in digital images and videos[3]. CV has found numerous successful applications in medicine, particularly in oncology[4]. Models for skin lesion recognition and melanoma classification have demonstrated predictive performance exceeding that of expert dermatologists[5-7]. Similarly, CV systems for endoscopic polyp detection and classification have achieved high sensitivity and competitive specificity[8]. The number of studies applying CV to minimally invasive surgery has grown rapidly in the past decade. Applications range from surgical phase recognition and workflow automation to intraoperative guidance and assessment of errors and effectiveness[9,10]. Additional CV use cases include anatomy detection and segmentation, instrument tracking, and operative time prediction[1]. Despite its potential, the generalization of intraoperative anatomy detection and segmentation faces methodological challenges, largely driven by variations in anatomical configurations, patient characteristics, and pathology-related changes[1]. Deep learning (DL) models have been proposed to detect and segment intra-abdominal organs or instruments, but they are often restricted to specific procedures or steps and rarely attempt to recognize the entire abdominal cavity via broad intra-abdominal detection and segmentation[11-15]. Specifically, most existing models are trained on procedure-specific datasets with a limited number of labeled structures[1,12,14,16-18]. For instance, the Dresden Surgical Anatomy Dataset, one of the most comprehensive to date, was developed for robotic anterior rectal resections and includes annotations of 11 abdominal structures[19]. While these models for organ detection and segmentation perform well within their defined scope, a more generalized model capable of detecting and segmenting a broader range of structures across various abdominal surgeries, or in a general view of the abdominal cavity, such as during staging laparoscopy (SL), is still lacking[13,15,20]. This highlights the ongoing reliance on a limited set of anatomical structures and procedure-specific datasets, with most models being trained and validated within narrowly defined surgical contexts. Consequently, their generalizability to broader intra-abdominal views, such as those encountered during SL, remains limited. More recent work has also moved toward broader surgical scene understanding, including comprehensive analyses of laparoscopic anatomical-structure vision, clinically oriented systems for laparoscopic exploration, and endoscopic frameworks integrating 3D reconstruction, novel view synthesis, and deformable tissue modeling[21-25]. These developments reflect a broader shift from isolated visual subtasks toward richer and

more clinically relevant representations of the intraoperative scene. However, these approaches remain challenged by heterogeneous anatomy, non-rigid tissues, illumination changes, occlusions, and limited ground-truth data.

The Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) AI Taskforce has proposed standardized guidelines advocating for scalable video annotation frameworks adaptable to general surgery. In terms of spatial annotations, a hierarchical approach is recommended, starting from broad anatomical regions and progressing to finer structures, such as liver lobes, abdominal walls, or fascial layers[26]. Recently, a scoping review by Kamtam *et al.* analyzing 58 studies on surgical video segmentation has highlighted the ongoing lack of large and diverse datasets, underscoring the need for generalizable and cross-procedure models capable of consistent anatomical detection and segmentation across variable surgical contexts[20].

In this study, we assess the feasibility of developing CV models using a surgical phase-independent dataset of SL annotated with an expanded range of anatomical structures, including the peritoneal surfaces of the abdominal wall. We introduce a standardized annotation checklist and evaluate DL models for the detection and segmentation of a broad range of intra-abdominal organs and peritoneal metastasis (PM), highlighting the challenges in identifying small or infrequently labeled structures and anatomically variable regions.

METHODS

Data collection and video characteristics

This study was approved by the Institutional Review Board of Ghent University Hospital, Belgium (approval reference: BC09947) and in accordance with the Declaration of Helsinki. Written informed consent for data acquisition, analysis, and publication was obtained from all participants. From June 2020 to March 2023, a total of 30 videos were selected from the existing video databases of the departments of surgery at Ghent University Hospital. Eligibility criteria included SL performed for gastrointestinal malignancies with suspected or confirmed PM. Patients included in this study had pathology-confirmed PM at the patient level. Operative videos and lesion appearances were reviewed lesion by lesion by two surgeons, in conjunction with the corresponding clinical and intraoperative findings. Standard laparoscopic exploration was performed for each abdominal quadrant, including the central abdominal region, the upper and lower quadrants, the small bowel and the abdominal walls. No fixed imaging standard or lesion-camera distance was imposed, as the analyzed frames were derived from routine SL videos and reflected the natural progression of the surgical act, including variable zoom and viewing distance.

Videos were captured using 5- and 10-mm, 30° scopes (Olympus, Hamburg, Germany). Videos were stored utilizing a digital capture system (Stryker SDC3).

Video processing and data management

Patient data, including any patient information and metadata such as the date of recording and camera type contained in the videos, were removed. All data were pseudonymized according to the General Data Protection Regulation (GDPR) of the European Union. The videos were converted into a standardized MPEG-4 format. No changes were made to image intensity, contrast, resolution, or frame rate. Any surgical actions not relevant to abdominal exploration or any procedures performed following the exploration, if present, were edited out of the video. For instance, any surgical resections performed for therapeutic or diagnostic purposes, recorded after the exploration, were removed. However, manipulation of organs such as the small bowel to accurately assess the presence of PM was retained. Subsequently, the videos were uploaded to the web-based CV platform Encord, developed by Cord Technologies Limited in London, United Kingdom. This platform facilitates direct annotation of still frames, ensuring that the context from the source video is retained. Although SAM 2-based automated labeling was available within the platform

Table 1. Annotated classes

Original label set (Set A)		
SI	Bladder	Uterus
Omentum	Spleen	Ovary
Falciform ligament	Small bowel	Left paracolic gutter
Left liver lobe	Appendix	Right paracolic gutter
Right liver lobe	Cecum	Left pelvic sidewall
Gallbladder	Ascending colon	Right pelvic sidewall
Left hemidiaphragm	Transverse colon	Suprapubic region
Right hemidiaphragm	Descending colon	Anterior abdominal wall
Stomach	Sigmoid colon	PM
Merged label set (Set B)		
SI	Small bowel	Abdominal walls (merged):
Omentum	Appendix	*Left hemidiaphragm*
Falciform ligament	Colon (merged):	*Right hemidiaphragm*
Liver (merged):	*Cecum*	*Left paracolic gutter*
Left liver lobe	*Ascending*	*Right paracolic gutter*
Right liver lobe	*Transverse*	*Left pelvic sidewall*
Gallbladder	*Descending*	*Right pelvic sidewall*
Stomach	*Sigmoid*	*Suprapubic region*
Bladder	Uterus	*Anterior abdominal wall*
Spleen	Ovary	PM

SI: Surgical instrument; PM: peritoneal metastasis.

and was explored as an annotation-support option, it was not used to generate the final reference annotations because it did not provide sufficiently reliable results for dense multi-class labeling of phase-independent SL scenes, particularly for PM and fine-grained intra-abdominal structures.

Annotation checklist development

A structured checklist was developed by a multidisciplinary team through an iterative process to define a standardized annotation protocol, including explicit decision rules and quality-control steps to ensure consistent implementation. The team comprised an anatomist, two surgical oncologists, two CV engineers, and two surgeons specialized in CV and surgical AI. A review of anatomical guidelines and prior segmentation literature informed the checklist's development[26,27]. The protocol was pilot tested by two trained annotators and subsequently revised based on their feedback. The goal was to ensure anatomical accuracy, reduce inter-annotator variability, and improve consistency in frame labeling. The checklist included: (1) annotation instructions and frame selection guidelines; (2) anatomical definitions, landmarks, and labeling criteria; (3) example images; and (4) a comprehensive list of structures to annotate. Pilot testing on six videos with iterative feedback yielded a finalized checklist of 25 anatomical structures, one surgical instrument (SI), namely the laparoscopic fenestrated grasper, and one tumor class for PM, for a total of 27 labeled categories. These categories are listed in Table 1. The full checklist with annotation guidelines is available in Supplementary Data 1.

Annotation methodology

Following the checklist's instructions, free-hand spatial annotations were performed. Overall, annotations were generated within a multistep quality-control framework based on the standardized checklist. Each annotation was conducted by an annotating team consisting of seven trained medical annotators, including

an abdominal surgeon, surgical residents, medical students, and research fellows. One team leader oversaw data collection, annotations, team training, guidance, supervision, and coordination with the engineering team. Once the annotation process was completed, each annotated frame was carefully reviewed by the team leader and subsequently reassessed by two oncologic surgeons. Any discrepancies were resolved through group consensus. Finally, quality assessment was jointly performed by the team leader and one engineer. No automated near-complete reference masks were generated and subsequently corrected manually; Core-Set active learning was used exclusively for diversity-based frame selection, not for automatic annotation generation. Frames containing partially covered organs, as well as frames with suboptimal image quality (e.g. blur or low illumination), were retained to reflect the full spectrum of intraoperative visibility. All structures were segmented at the pixel level. Frame selection followed a content-aware random sampling strategy across videos, while ensuring representation of the anatomical structures defined in the annotation checklist, yielding 1,989 frames. This initial sampling was complemented by a diversity-based selection of 320 additional frames using a Core-Set active learning strategy[28], aimed at maximizing visual and anatomical diversity in the annotated dataset, for a total of 2,309 annotated frames.

Deep learning model development, metrics and label configurations

YOLOv9[29] and Co-DETR[30] were used for object detection, while SegFormer[31] and Mask2Former[32] were used for semantic segmentation. Specifically, object detection models (YOLOv9 and Co-DETR) were trained to localize anatomical structures and PM with bounding boxes, while semantic segmentation models (SegFormer and Mask2Former) were used to generate pixel-wise delineations of each class. YOLOv9[29] was chosen for its balance between speed and accuracy, making it suitable for real-time use in variable laparoscopic environments. Co-DETR[30] improves representation learning in DETR-style end-to-end detectors by training several parallel auxiliary prediction heads under one-to-many label assignments, which enriches supervision during training without increasing inference cost. For segmentation, SegFormer[31] employed a transformer-based hierarchical encoder to capture multi-scale contextual features, and Mask2Former[32] extended this approach with attention-based mask queries to better handle small or overlapping regions. Model selection was predefined to compare strong contemporary architectures from complementary design families within a coherent experimental framework, rather than to provide an exhaustive benchmark of all subsequently emerging model variants. For segmentation, SegFormer and Mask2Former were selected because they converged reliably and achieved competitive validation performance on our dataset, while offering strong multi-scale feature representations considered relevant for the variable viewing distances, scale variation, and optical conditions of phase-independent laparoscopic video, as well as for the segmentation of small PM. SegFormer uses hierarchical Transformer features without an explicit feature pyramid network, whereas Mask2Former combines deformable multi-scale attention with a pixel decoder tailored to fine-grained segmentation. These properties made them suitable for a focused feasibility assessment within a coherent and modern segmentation design space.

The dataset was randomly split at the video level using a 60/20/20 ratio for training, validation, and testing, respectively. SegFormer used a MiT-B5 backbone pretrained on ImageNet-1k, whereas Mask2Former used a Swin-Large backbone pretrained on ImageNet-22k. All segmentation models were trained with a fixed iteration-based schedule of 160,000 iterations, and the best checkpoint was selected as the iteration at which validation mIoU reached its maximum. SegFormer was optimized with AdamW using linear warmup followed by PolyLR, whereas Mask2Former was optimized with AdamW using PolyLR without warmup. Standard augmentation pipelines were used for all segmentation models and included multi-scale resizing, random cropping, random horizontal flipping ($P = 0.5$), and photometric distortion. SegFormer used RandomResize anchored at 896 px with ratio range 0.5-2.0 and 1,024 × 1,024 crops, whereas Mask2Former used RandomChoiceResize across 16 scales from 320 to 1,280 px with 640 × 640 crops. In addition, class-weighted loss terms were used in both model families to mitigate class imbalance. No explicit

small-object-specific methods such as copy-paste augmentation, focal loss, or lesion oversampling were included in the final training pipelines. The hardware/software environment included NVIDIA RTX 3090 GPUs, CUDA 11.2/11.3, PyTorch 1.12.1, and MMEngine 0.10.x. Full training configurations of the retained segmentation models are provided in Supplementary Table 1. Detection performance was evaluated using precision, recall, F1-score, and mean average precision at an Intersection over Union (IoU) threshold of 0.50 (mAP@50). Segmentation performance was evaluated using precision, recall, F1-score, and IoU. Precision and recall assessed the accuracy and completeness of predictions. F1-score was calculated as the harmonic mean of precision and recall. IoU measured spatial agreement between predicted and reference masks. mAP@50 was computed as the average area under the precision-recall curve across all classes at the specified threshold.

To evaluate the effect of class granularity on model performance, two label configurations were generated [Table 1]:

• Original Label Set (Set A): Each of the 25 anatomical structures. PM and SI annotated as separate categories.

• Merged Label Set (Set B): Thirteen functionally and topographically related structures grouped to reduce annotation sparsity. Right and left liver lobes; all colonic segments (cecum, ascending, transverse, descending, sigmoid); and abdominal wall regions (hemidiaphragms, paracolic gutters, pelvic sidewalls, suprapubic region, anterior abdominal wall) were merged. Additionally, PM and SI were evaluated as separate categories.

Statistical analysis and embedding visualization

Statistical analyses were descriptive and exploratory. Relationships between annotation features and model performance metrics, including F1-score and IoU, were assessed at the class level. Annotation features included the number of annotated objects and the area fraction, defined as the percentage of image pixels occupied by each class in each frame. Annotation features were derived from the training set, reflecting the data available to the model during learning, whereas performance metrics were computed on the test set to provide an unbiased estimate of model performance. Because annotation features were non-normally distributed, and because performance was not expected to scale linearly with increasing annotation count or annotated area, associations were primarily assessed using Spearman's rank correlation coefficients to evaluate monotonic relationships. Each variable was averaged per class, and pairwise univariate correlations were calculated to explore possible associations between annotation density, object size, and model performance. Logarithmic transformations (log10) of annotation features were additionally explored to account for potential plateau behavior. Correlation analyses were performed at the class level, with each annotated category treated as a single observation. Scatter plots were generated to visualize these relationships and to qualitatively assess the effect of annotation sparsity and class imbalance. In addition, complementary ordinary least squares (OLS) regression models using raw and log-transformed annotation features were fitted to further characterize the relationship between annotation characteristics and class-level segmentation performance, and residuals were examined to identify classes that overperformed or underperformed relative to what would be expected based on their spatial representation. A two-sided P-value < 0.05 was considered statistically significant and was interpreted cautiously given the exploratory nature of the analysis. Formal pairwise statistical comparison between model architectures was not performed, as the study was designed as an exploratory class-level feasibility analysis rather than a confirmatory benchmark. To further evaluate the internal feature representations learned by the segmentation network, dimensionality reduction was performed using t-distributed stochastic neighbor embedding (t-SNE) on the output of the penultimate layer[33]. The resulting two-dimensional representations

were visualized to assess class clustering and separability, providing insight into the discriminative structure of the learned embeddings across anatomical categories. Statistical analyses were performed using Python (version 3.10) with the SciPy and NumPy packages.

RESULTS

The dataset contained a total of 2,309 hand-annotated images from 30 videos of SL (females: n= 13, 43.3%; males: n= 17, 56.7%). Indications for SL were primarily gastric (n = 13, 43.3%), colorectal (n = 6, 20.0%) and pancreatic cancers (n = 6, 20.0%). Patients' demographics are presented in Table 2. The training, validation, and test sets were split randomly into 18/6/6 videos and 1304/433/572 frames, respectively. This corresponded to a total of 16,642 annotated objects, divided into 8,853, 3,039 and 4,750 objects in the training, validation and test sets, respectively. PM accounted for 4,094 annotations in total, comprising 1,861, 638, 1,595 in the training, validation and test sets. Table 3 summarizes the number of annotated objects per class. Figure 1 illustrates representative examples of the annotation framework and segmentation task, showing original laparoscopic frames, ground truth annotations, and model predictions. Representative visual examples of the detection models are provided in Supplementary Figure 1.

Detection models

Overall, mAP@50 for detection was 0.52 and 0.48 in Set A and 0.61 and 0.57 in Set B for YOLOv9 and Co-DETR, respectively. The best-performing anatomical structures with YOLOv9 were the left liver lobe (0.84), falciform ligament (0.81), and omentum (0.75) in Set A, and the abdominal wall (0.82), small bowel (0.77), and omentum (0.76) in Set B. With Co-DETR, the highest scores in Set A were for the omentum (0.77), left liver lobe (0.79), and falciform ligament (0.72), while in Set B the abdominal wall (0.92), liver (0.81), and omentum (0.75) performed best. SI were consistently well recognized, with mAP@50 0.95 in Set A and 0.96 in Set B with YOLOv9, and 0.97 in Set A and 0.89 in Set B with Co-DETR. PM detection achieved modest performance, with YOLOv9 reaching 0.39 in Set A and 0.25 in Set B, while Co-DETR achieved 0.44 in Set A and 0.40 in Set B. Performance of the detection models for all anatomical structures, SI, and PM is presented in Table 4.

Segmentation models

For semantic segmentation, Mask2Former outperformed SegFormer at the aggregate level, with overall F1-score improving from 0.58 to 0.67 and IoU from 0.43 to 0.54 when moving from Set A to the merged Set B. Classes with distinctive appearance achieved the best results, for example liver in Set B (IoU 0.83, F1-score 0.91 with SegFormer *vs.* IoU 0.88 and F1-score 0.94 for Mask2Former) and abdominal wall (IoU 0.87-0.88, F1-score 0.93 across the two models). SI was consistently strong (F1-score 0.92 in both sets with Mask2Former and 0.89-0.90 with SegFormer). Performance declined for small or visually ambiguous targets, with PM remaining challenging despite modest gains using SegFormer (IoU 0.14-0.13, F1-score 0.23-0.25), and Mask2Former (IoU 0.29-0.30, F1-score 0.45-0.46), which was consistent with limited pixel coverage and heterogeneous tumors' presentation. Performance of the segmentation models for all anatomical structures, SI, and PM is presented in Table 5.

Annotation analysis

Annotation analysis demonstrated substantial variation in labeling density and spatial representation across classes. A total of 4,094 PM instances were annotated, making PM the most frequently labeled class, yet PM occupied a comparatively small area fraction relative to larger organs such as the liver, omentum, and small bowel. Figure 2 shows the relationships between object count and area fraction and class-level F1-score across Sets A and B. Supplementary Figure 2 shows the corresponding relationships for IoU. Across both label configurations, the plots show that classes with higher annotation counts did not consistently achieve higher performance. In fact, PM showed the highest annotation counts but comparatively low IoU and

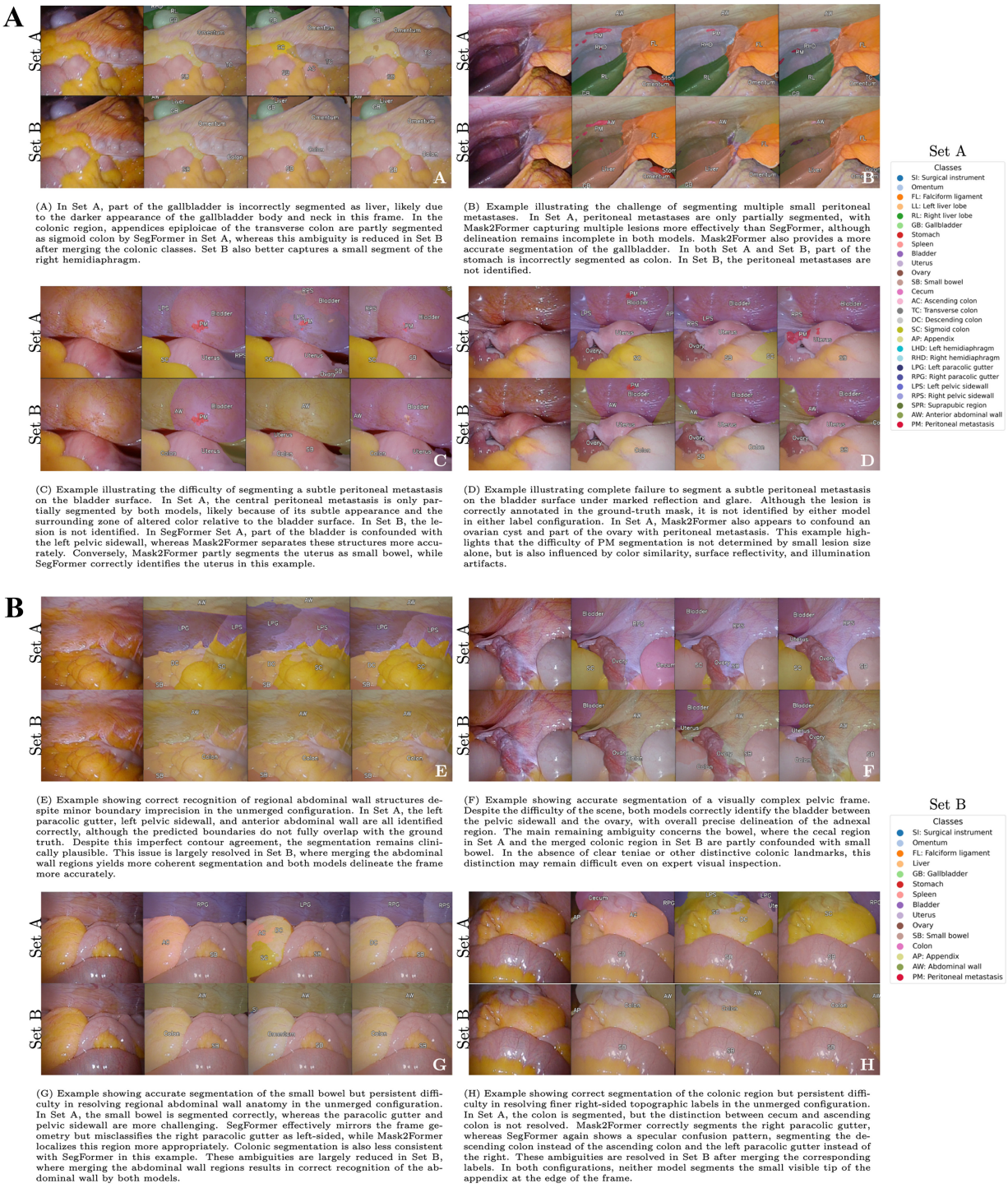


Figure 1. Semantic segmentation of intraoperative laparoscopic scenes. From left to right, each example shows the input frame, ground-truth annotation, SegFormer prediction, and Mask2Former prediction. The upper row corresponds to Set A and the lower row to Set B.

F1-score, whereas several classes with fewer annotations, including the gallbladder, omentum, and falciform ligament, achieved higher values for both metrics.

For the Mask2Former segmentation model, associations between annotation features (object count and area fraction) and performance differed between the two predictors. In the unmerged configuration (Set A), object count showed a statistically significant monotonic association with IoU (Spearman $\rho = 0.441$, $P =$

Table 2. Baseline patient characteristics

Patients' characteristics	N (%), Median (range)
Age	62 (45-78)
Sex (M)	17 (56.7%)
BMI (kg/m^2)	24.8 (21.1-35.3)
ASA score	
I	4 (13.3%)
II	16 (53.3%)
III	9 (30.0%)
IV	1 (3.3%)
Indication, n (%)	
Gastric cancer	13 (43.3%)
Colorectal cancer	6 (20.0%)
Pancreatic cancer	6 (20.0%)
Esophageal cancer	2 (6.7%)
Hepatocellular carcinoma	1 (3.3%)
Cholangiocarcinoma	1 (3.3%)
Appendiceal cancer	1 (3.3%)

BMI: Body mass index; ASA: American Society of Anesthesiologists score.

0.0212), while area fraction showed a stronger association (ρ = 0.482, P = 0.0109). In the merged configuration (Set B), object count showed a monotonic but non-significant association with IoU (ρ = 0.486, P = 0.0664), whereas area fraction showed a strong and statistically significant association (ρ = 0.668, P = 0.00651). After logarithmic transformation of object count and area fraction, Spearman coefficients remained unchanged, reflecting the invariance of rank-based correlation under monotonic transformations.

Comparable results were obtained when performance was assessed using F1-score. In Set A, object count was significantly associated with F1-score (Spearman ρ = 0.443, P = 0.0208), but less strongly than area fraction (ρ = 0.484, P = 0.0106). In Set B, object count was not significantly associated with F1-score (ρ = 0.486, P = 0.0664), whereas area fraction again showed a strong and statistically significant association (ρ = 0.668, P = 0.00651).

Complementary regression analyses supported these findings and further indicated that area-based predictors were more informative than annotation count. In Set A, the best-performing regression model used raw area fraction as predictor, with adjusted R^2 values of 0.194 for F1-score and 0.200 for IoU. For Set A, raw area fraction was positively associated with F1-score (β = 3.517, P = 0.0123) and IoU (β = 3.975, P = 0.0112). In Set B, the best-performing regression model used log-transformed area fraction as predictor, with adjusted R^2 values of 0.182 for F1-score and 0.224 for IoU. For F1-score, log-transformed area fraction showed a positive association that did not reach statistical significance (β = 0.111, P = 0.0635), whereas for IoU it showed a statistically significant positive association (β = 0.141, P = 0.0428). For Set B F1-score, regression inference was interpreted cautiously because residual normality was not satisfied, and the model was therefore used primarily to support residual-gap identification rather than formal hypothesis testing.

Residual analysis further showed class-specific deviations from the regression baseline. In Set A, left pelvic sidewall, ovary, suprapubic region, and ascending colon showed the largest negative performance gaps, approximately -0.20 to -0.30, whereas gallbladder, SI, and the liver lobes showed positive gaps of

Table 3. Distribution of annotated objects per class across the training, validation, and testing sets for sets A and B

Class	Set A						Set B					
	Training		Validation		Test		Training		Validation		Test	
SI	552	(6.2%)	169	(5.6%)	211	(4.4%)	552	(7.7%)	169	(7.0%)	211	(5.4%)
Omentum	633	(7.2%)	227	(7.5%)	267	(5.6%)	633	(8.8%)	227	(9.4%)	267	(6.9%)
Falciform ligament	239	(2.7%)	79	(2.6%)	112	(2.4%)	239	(3.3%)	79	(3.3%)	112	(2.9%)
Liver							451	(6.3%)	140	(5.8%)	167	(4.3%)
Left liver lobe	337	(3.8%)	97	(3.2%)	119	(2.5%)						
Right liver lobe	259	(2.9%)	98	(3.2%)	120	(2.5%)						
Gallbladder	81	(0.9%)	30	(1.0%)	63	(1.3%)	81	(1.1%)	30	(1.2%)	63	(1.6%)
Stomach	207	(2.3%)	73	(2.4%)	79	(1.7%)	207	(2.9%)	73	(3.0%)	79	(2.0%)
Spleen	19	(0.2%)	1	(0.0%)	24	(0.5%)	19	(0.3%)	1	(0.0%)	24	(0.6%)
Bladder	324	(3.7%)	133	(4.4%)	116	(2.4%)	324	(4.5%)	133	(5.5%)	116	(3.0%)
Uterus	78	(0.9%)	24	(0.8%)	13	(0.3%)	78	(1.1%)	24	(1.0%)	13	(0.3%)
Ovary	45	(0.5%)	18	(0.6%)	14	(0.3%)	45	(0.6%)	18	(0.7%)	14	(0.4%)
Small bowel	734	(8.3%)	248	(3.2%)	293	(6.2%)	734	(10.2%)	248	(10.3%)	293	(7.5%)
Colon							740	(10.3%)	225	(9.3%)	383	(9.8%)
Cecum	113	(1.3%)	8	(0.3%)	98	(2.1%)						
Ascending	107	(1.2%)	21	(0.7%)	74	(1.6%)						
Transverse	235	(2.7%)	104	(3.4%)	103	(2.2%)						
Descending	211	(2.4%)	79	(2.6%)	139	(2.9%)						
Sigmoid	360	(4.1%)	120	(3.9%)	159	(3.3%)						
Appendix	18	(0.2%)	7	(0.2%)	11	(0.2%)	18	(0.3%)	7	(0.3%)	11	(0.3%)
Abdominal wall							1,198	(16.7%)	402	(16.7%)	547	(14.0%)
Left hemidiaphragm	447	(5.0%)	127	(4.2%)	194	(4.1%)						
Right hemidiaphragm	247	(2.8%)	94	(3.1%)	123	(2.6%)						
Left paracolic gutter	372	(4.2%)	144	(4.7%)	139	(2.9%)						
Right paracolic gutter	190	(2.1%)	51	(1.7%)	122	(2.6%)						
Left pelvic sidewall	291	(3.3%)	124	(4.1%)	135	(2.8%)						
Right pelvic sidewall	301	(3.4%)	108	(3.6%)	190	(4.0%)						
Suprapubic region	224	(2.5%)	74	(2.4%)	104	(2.2%)						
Anterior abdominal wall	368	(4.2%)	143	(4.7%)	133	(2.8%)						
PM	1,861	(21.0%)	638	(21.0%)	1,595	(33.6%)	1,861	(25.9%)	638	(26.4%)	1,595	(40.9%)
Overall (anatomical structures only)	6,440		2,232		2,944		4,767		1,607		2,089	
Total	8,853		3,039		4,750		7,180		2,414		3,895	

SI: Surgical instrument; PM: peritoneal metastasis. Overall (anatomical structures only) excludes PM and SI; Overall (all classes) includes all categories.

approximately +0.31 for F1-score and +0.37 for IoU. In Set B, bladder, uterus, PM, and ovary showed the largest negative performance gaps, ranging approximately from -0.25 to -0.32, whereas gallbladder, SI, liver, spleen, and small bowel showed positive gaps of approximately +0.23 for F1-score and +0.27 for IoU.

Embedding visualization

t-SNE plots for Mask2Former for Set A and Set B are shown in Figure 3. In the unmerged configuration (Set A), class embeddings appeared more dispersed with partial overlap between categories. In contrast, the merged configuration (Set B) showed more compact and visually distinct clusters, particularly for larger

Table 4. Performance (mAP@50) of DL detection models in Set A and Set B

Class	YOLOv9		Co-DETR	
	Set A	Set B	Set A	Set B
SI	0.95	0.96	0.97	0.89
Omentum	0.75	0.76	0.77	0.75
Falciform ligament	0.81	0.77	0.72	0.59
Liver		0.88		0.81
Left liver lobe	0.84		0.79	
Right liver lobe	0.71		0.71	
Gallbladder	0.68	0.66	0.59	0.64
Stomach	0.71	0.63	0.47	0.45
Spleen	0.56	0.52	0.59	0.74
Bladder	0.39	0.40	0.38	0.24
Uterus	0.42	0.56	0.54	0.63
Ovary	0.25	0.20	0.32	0.15
Small bowel	0.68	0.77	0.62	0.65
Colon		0.57		0.49
Cecum	0.33		0.32	
Ascending	0.24		0.15	
Transverse	0.20		0.26	
Descending	0.36		0.26	
Sigmoid	0.65		0.54	
Appendix	0.55	0.33	0.22	0.21
Abdominal wall		0.82		0.92
Left hemidiaphragm	0.79		0.64	
Right hemidiaphragm	0.70		0.65	
Left paracolic gutter	0.66		0.52	
Right paracolic gutter	0.34		0.41	
Left pelvic sidewall	0.19		0.23	
Right pelvic sidewall	0.30		0.30	
Suprapubic region	0.23		0.18	
Anterior abdominal wall	0.34		0.32	
PM	0.39	0.25	0.44	0.40
Overall (anatomical structures only)	0.52	0.63	0.48	0.58
Overall (all classes)	0.52	0.61	0.48	0.57

SI: Surgical instrument; PM: peritoneal metastasis. Overall (anatomical structures only) excludes PM and SI; Overall (all classes) includes all categories.

organs and grouped regions, suggesting increased separation of the learned feature representations. t-SNE plots for SegFormer are provided in the Supplementary Figure 3.

DISCUSSION

This exploratory study investigated the feasibility of developing CV models for organs and PM detection and segmentation during SL using a phase-independent and anatomically comprehensive annotation framework. Four DL architectures were evaluated, two for detection and two for segmentation across two label-set

Table 5. Performance of DL segmentation models in Set A and Set B

Class	SegFormer								Mask2Former							
	P		R		IoU		F1-score		P		R		IoU		F1-score	
	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
SI	0.87	0.89	0.90	0.91	0.79	0.81	0.89	0.90	0.92	0.91	0.92	0.93	0.85	0.85	0.92	0.92
Omentum	0.70	0.74	0.95	0.94	0.67	0.70	0.80	0.83	0.87	0.85	0.91	0.94	0.80	0.80	0.89	0.89
Falciform ligament	0.77	0.92	0.68	0.65	0.56	0.61	0.72	0.76	0.92	0.91	0.67	0.73	0.63	0.68	0.77	0.81
Liver		0.91		0.96		0.83		0.91		0.92		0.96		0.88		0.94
Left liver lobe	0.72		0.81		0.61		0.76		0.81		0.89		0.73		0.85	
Right liver lobe	0.73		0.83		0.64		0.78		0.88		0.82		0.74		0.85	
Gallbladder	0.84	0.88	0.74	0.75	0.65	0.68	0.78	0.81	0.90	0.91	0.85	0.84	0.78	0.78	0.88	0.88
Stomach	0.84	0.81	0.68	0.61	0.60	0.53	0.75	0.70	0.94	0.93	0.64	0.66	0.61	0.63	0.76	0.77
Spleen	0.93	0.84	0.40	0.43	0.39	0.39	0.56	0.57	0.85	0.88	0.58	0.59	0.53	0.55	0.69	0.71
Bladder	0.47	0.70	0.48	0.41	0.31	0.35	0.48	0.52	0.58	0.63	0.54	0.44	0.39	0.35	0.56	0.52
Uterus	0.55	0.73	0.71	0.59	0.45	0.48	0.62	0.65	0.61	0.61	0.57	0.57	0.42	0.28	0.59	0.44
Ovary	0.73	0.76	0.59	0.43	0.48	0.38	0.65	0.55	0.43	0.65	0.29	0.27	0.20	0.23	0.34	0.38
Small bowel	0.78	0.83	0.94	0.92	0.74	0.77	0.85	0.87	0.86	0.86	0.94	0.93	0.82	0.81	0.90	0.90
Colon		0.83		0.65		0.58		0.73		0.87		0.78		0.70		0.82
Cecum	0.50		0.43		0.30		0.46		0.47		0.57		0.35		0.51	
Ascending	0.50		0.21		0.18		0.30		0.35		0.44		0.24		0.39	
Transverse	0.34		0.17		0.13		0.23		0.60		0.51		0.38		0.55	
Descending	0.76		0.33		0.30		0.46		0.66		0.58		0.45		0.62	
Sigmoid	0.76		0.49		0.42		0.59		0.85		0.81		0.70		0.83	
Appendix	0.42	0.43	0.08	0.03	0.07	0.03	0.13	0.05	0.57	0.65	0.37	0.64	0.29	0.48	0.45	0.65
Abdominal wall		0.91		0.96		0.88		0.93		0.92		0.95		0.87		0.93
Left hemidiaphragm	0.76		0.68		0.56		0.71		0.80		0.87		0.72		0.84	
Right hemidiaphragm	0.62		0.76		0.52		0.69		0.78		0.68		0.57		0.73	
Left paracolic gutter	0.81		0.63		0.54		0.71		0.78		0.77		0.63		0.77	
Right paracolic gutter	0.60		0.35		0.29		0.45		0.75		0.48		0.41		0.58	
Left pelvic sidewall	0.40		0.45		0.27		0.42		0.41		0.32		0.22		0.36	
Right pelvic sidewall	0.53		0.61		0.39		0.57		0.45		0.76		0.40		0.57	

Suprapubic region	0.36		0.36		0.22		0.36		0.50		0.33		0.25		0.40	
Anterior abdominal wall	0.50		0.67		0.40		0.57		0.56		0.57		0.40		0.57	
PM	0.28	0.35	0.22	0.18	0.14	0.13	0.25	0.23	0.68	0.71	0.34	0.38	0.29	0.30	0.45	0.46
Overall (anatomical structures only)	0.65	0.80	0.57	0.66	0.44	0.57	0.59	0.70	0.69	0.81	0.63	0.71	0.51	0.62	0.65	0.74
Overall (all classes)	0.63	0.77	0.56	0.63	0.43	0.54	0.58	0.67	0.70	0.81	0.63	0.71	0.51	0.61	0.65	0.73

P: Precision; R: recall, IoU: Intersection over Union; F1-score: harmonic mean of precision and recall; PM: peritoneal metastasis; SI: surgical instrument. Overall (anatomical structures only) excludes PM and SI; Overall (all classes) includes all categories.

configurations. The effect of class granularity on predictive performance was examined through correlation analyses. The best-performing detection model (YOLOv9) reached a mAP@50 of 0.52 and 0.61 in Set A and Set B, respectively, while the best-performing segmentation model (Mask2Former) achieved an overall IoU of 0.51 and 0.61 and an F1-score of 0.65 and 0.73 in Set A and Set B, respectively. Although PM segmentation yielded modest results, this study demonstrates both the feasibility and the challenges of incorporating a broad range of intra-abdominal structures (25 anatomical structures/regions), including peritoneal surfaces spanning different regions of the abdominal cavity. The complementary analysis of annotation features further identified limited area fraction (percentage of image pixels occupied by each class in each frame) as a critical determinant of PM segmentation performance, confirming that sparse spatial representation rather than annotation frequency alone was the primary constraint. Increasing annotation granularity by introducing detailed anatomical regions, even when defined according to a standardized checklist and annotated based on visible anatomical landmarks, introduced additional challenges for model discrimination. When structures were highly subdivided, feature similarity and contextual overlap reduced separability and degraded predictive performance. Conversely, regrouping related regions into broader topographic classes improved segmentation consistency and annotation efficiency, indicating that balanced label design is essential to achieve robust and interpretable models.

Developing a broadly generalizable anatomy detection and segmentation model remains challenging due to the considerable variability of anatomy and intraoperative conditions[15]. Previous datasets, such as the Dresden Surgical Anatomy Dataset, provided valuable annotations of eleven major abdominal structures (eight abdominal organs, abdominal wall, and two vascular structures) extracted from 32 robot-assisted procedures, yet were limited to rectal resections[19]. In contrast, the present dataset extends the anatomical spectrum to encompass the entire abdominal cavity, including regionally defined abdominal walls and peritoneal surfaces.

Unlike most existing CV models, which are trained on procedure-specific datasets where camera position, organ exposure, appearance and instrument presence follow predictable patterns, this work relies exclusively on SL recordings. In this context, the camera moves continuously without a defined operative phase or consistent field of view, producing substantial variability in illumination, depth, and visibility. This configuration represents a unique and underexplored scenario in the literature, which poses additional challenges for generalization but reflects more closely real intraoperative exploration. The present framework was intentionally designed to be phase-independent and not conditioned on camera direction or predefined anatomic priors. Although context-aware models incorporating spatial or procedural information may improve performance, the aim of this study was to evaluate feasibility under real-life conditions encountered during routine SL. Beyond detection and

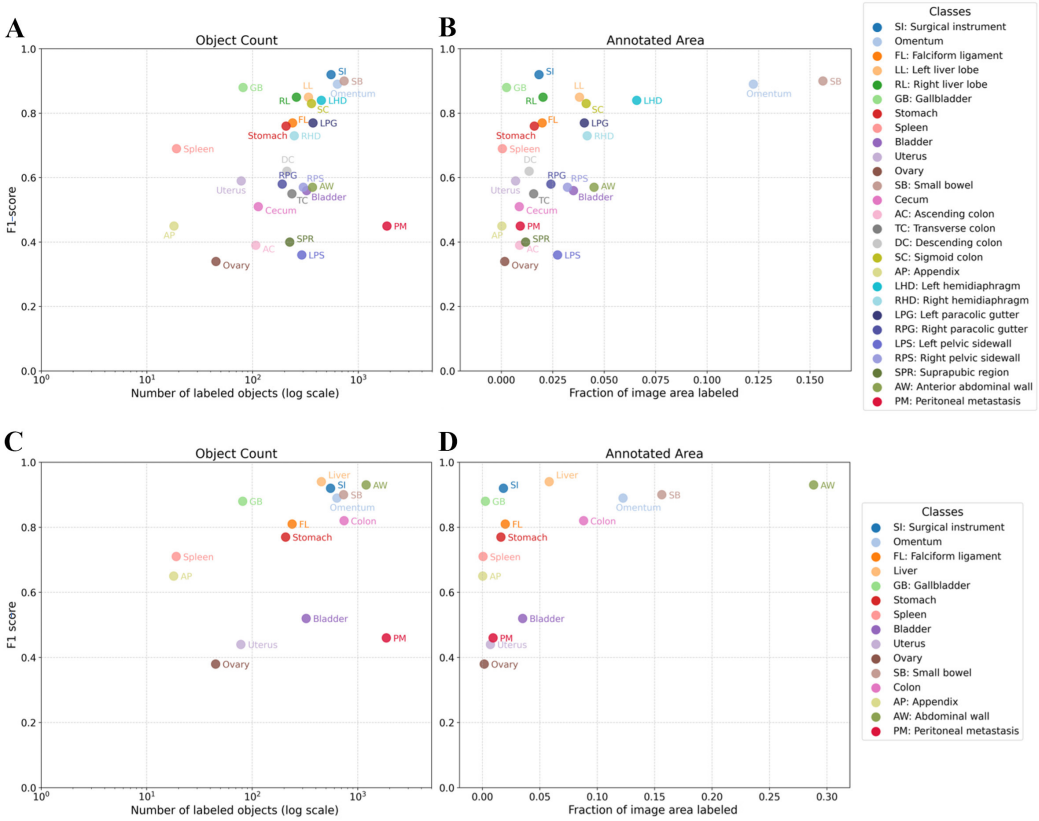


Figure 2. Class-level F1-score in relation to annotation count and annotated area fraction. (A) Set A: F1-score versus number of labeled objects; (B) Set A: F1-score versus fraction of image area labeled; (C) Set B: F1-score versus number of labeled objects; (D) Set B: F1-score versus fraction of image area labeled. The x-axis for object count is displayed on a logarithmic scale. Each point represents one annotated class.

segmentation, related areas of surgical CV such as feature matching, depth estimation, and endoscopic reconstruction also continue to face major challenges in minimally invasive surgery, particularly due to homogeneous tissue texture, variable illumination, frequent occlusions, soft-tissue deformability, and the limited availability of reliable ground-truth data. These broader methodological constraints further support the need for structured and generalizable annotation frameworks when developing phase-independent intra-abdominal CV models[34,35].

Across both label sets, over 4,000 PM annotations were performed, making them the most frequently labeled objects. However, when comparing the area fraction, PM occupied a markedly smaller area than larger organs such as the liver, omentum, and small bowel. The limited pixel coverage of PM likely contributed to lower detection and segmentation performance[36]. Annotation feature analyses supported this finding, showing that class-level segmentation performance had stronger monotonic associations with area fraction than with object count, and that logarithmic transformation of annotation features did not alter Spearman rank coefficients. Of note, despite being among the least frequently annotated organs and occupying a small area fraction, the gallbladder achieved some of the highest IoU and F1-scores, illustrating that distinctive, high-contrast structures can perform well even with limited annotation count and spatial coverage. This observation demonstrates that annotation frequency alone does not necessarily enhance model performance when the annotated surface area is minimal. Importantly, increasing the number of annotations alone does not necessarily improve model quality or generalization. Prior work has shown that annotation quality, labeling protocol, and the representativeness of selected samples can be as important as annotation volume, particularly in biomedical image analysis and annotation-efficient learning settings. Our sampling strategy

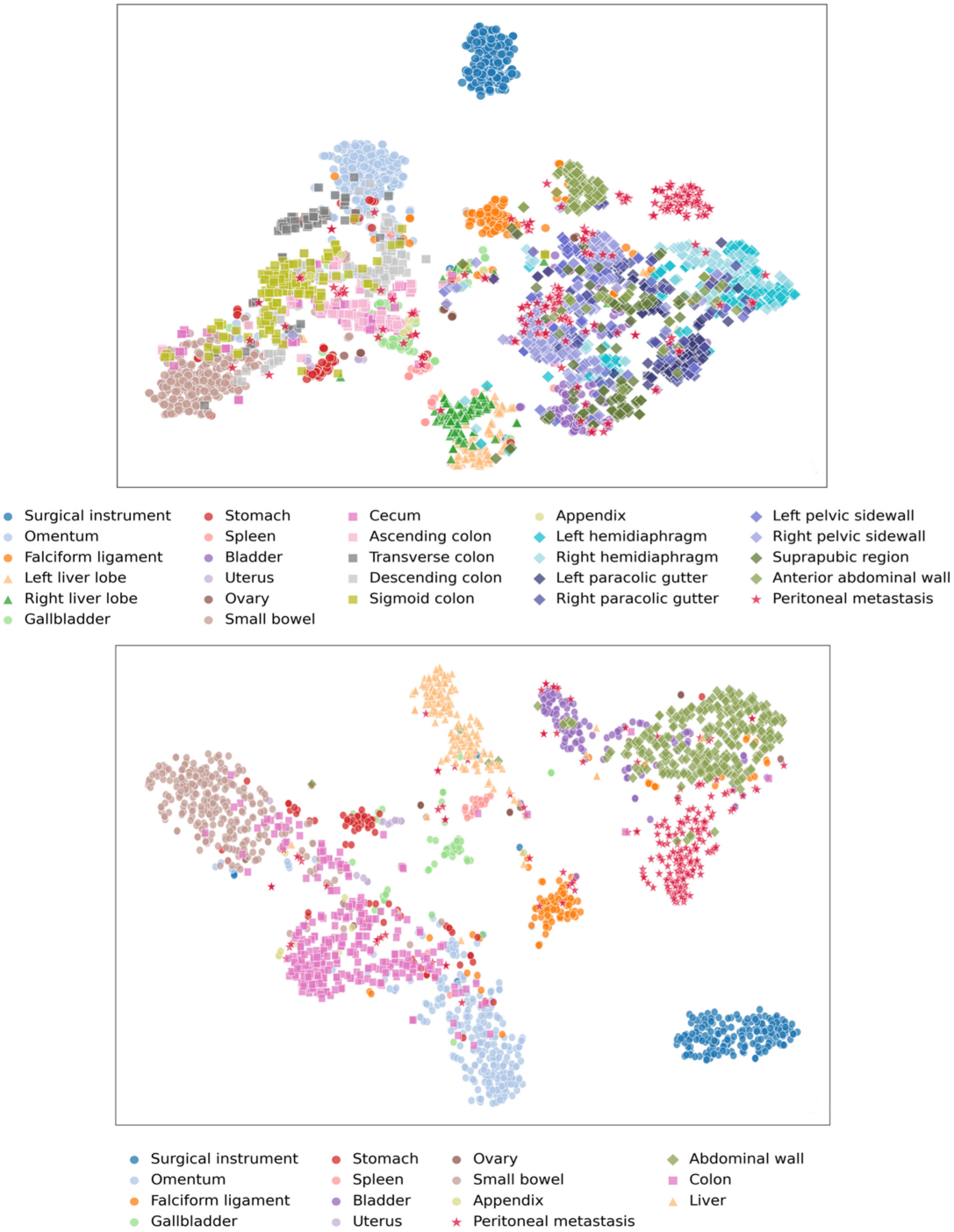


Figure 3. t-SNE visualization of feature embeddings for Mask2Former: (A) Set A, (B) Set B. t-SNE: t-distributed stochastic neighbor embedding.

therefore prioritized informative and diverse frames rather than raw frame count alone, and this was complemented by Core-Set active learning to improve coverage of visual and anatomical variability[23,27]. Earlier work by Yengera *et al.* showed that increasing the number of annotations does not necessarily improve model performance in surgical scene understanding, as excessive labeling may introduce redundancy and label noise, particularly when small or visually ambiguous regions occupy limited pixel areas[37]. This pattern mirrors what was observed in the present dataset, emphasizing that annotation quality

and spatial relevance are more critical than annotation quantity for effective intra-abdominal model development.

In the case of PM, this limitation is further compounded by intrinsic heterogeneity and by the fact that lesions arise on highly variable backgrounds across different organs and peritoneal regions. Unlike homogeneous organs, PM often appear as irregular deposits of varying size, color, and texture overlying a wide range of surrounding structures and peritoneal surfaces which alter local contrast and complicate discrimination from the background tissue. This visual overlap can confound pixel-based segmentation and reduce feature separability. In the t-SNE analysis [Figure 3], PM clusters show substantial overlap with adjacent anatomical structures, particularly in Set A within the abdominal wall region where PM are most commonly present. These observations indicate that small-object segmentation in intra-abdominal scenes is influenced not only by object size but also by contextual ambiguity and the inherent heterogeneity of oncologic disease.

Recent developments in small-target segmentation, such as the STS-Net framework proposed by Zhao *et al.*, highlight the potential of targeted data-augmentation and attention-based mechanisms to improve detection of small lesions by enhancing feature representation and balancing sample sizes. However, these methods remain computationally demanding and may underperform when applied to heterogeneous or large-area lesions, where amplification blocks and small-target-specific attention lose efficiency[36]. Adapting such techniques to intra-abdominal surgical scenes could therefore offer localized benefits for PM detection but would still require optimization for complex and variable visual contexts. The final training pipelines already incorporated multi-scale augmentation, cropping, photometric distortion, and class-weighted losses, but no explicit small-object-specific techniques such as copy-paste augmentation, focal loss, or lesion oversampling were used, and these remain relevant avenues for improving PM segmentation in future work.

Anatomical structures with well-defined contours and consistent visual appearance, such as the liver, spleen, gallbladder, and stomach, achieved higher accuracy across models[15]. SI were also reliably detected and segmented, reflecting their distinctive geometry and high contrast against soft tissue. Nevertheless, SI was among the most frequently annotated structures while occupying a relatively small area fraction. Comparable trends have been reported in prior surgical video datasets, where easily distinguishable organs with stable morphology and consistent exposure, such as the liver, consistently reached the highest detection and segmentation scores[1,15]. This supports the internal consistency of the current dataset and confirms its ability to reproduce established performance hierarchies. In contrast, when the annotated class set expands to include structures with less defined boundaries, variable texture, or partial exposure, such as peritoneal gutters or pelvic sidewalls, detection and segmentation performance declines, reflecting the visual heterogeneity and anatomical variability encountered in SL scenes.

These findings align with recent scoping evidence from Kamtam *et al.* who highlighted persistent limitations in dataset diversity and generalizability across 58 studies on surgical video segmentation. Despite major advances in DL-based scene understanding, most available datasets remain small, procedure-specific, and non-public, limiting reproducibility and cross-institutional development. Consistent with the present observations, segmentation performance depends strongly on the visual prominence and pixel coverage of each structure, with smaller or thin targets such as vessels, nerves, or peritoneal lesions showing lower predictive performances regardless of the amount of annotation[20].

Variability in surgical views, illumination, and annotation consistency continues to challenge model robustness, emphasizing the importance of standardized annotation protocols. The standardized checklist developed in this study addresses this limitation by providing a replicable framework suitable for multicenter

use and, to our knowledge, represents the first comprehensive annotation checklist specifically designed for SL. As suggested by Kamtam *et al.*, the creation of large open datasets, the implementation of generalized foundational models such as SAM, and the exploration of unsupervised training strategies are directly applicable to intra-abdominal scene understanding. Incorporating such approaches could enhance the generalizability and clinical maturity of SL-based CV models, ultimately supporting real-time intraoperative decision assistance[20]. Automated labeling tools such as SAM 2 may help reduce manual annotation burden for selected high-contrast structures, but in the present setting SAM 2 was not sufficiently reliable for dense annotation of PM and complex abdominal topography, underscoring the need for domain-adapted approaches in phase-independent SL.

While recent work has explored increasingly sophisticated architectures for organ detection and segmentation and intra-abdominal landmark detection in laparoscopic surgery, architectural advances alone have not resolved key limitations related to dataset scarcity, procedural specificity, and inconsistent annotation methodologies[15,38,39]. As highlighted by Zhou *et al.*, the lack of standardized taxonomies and evaluation protocols continues to hinder reproducibility and cross-dataset comparison in surgical CV[15]. These challenges are directly reflected in the present study, where predictive performance remained constrained by class imbalance and variability in visual representation despite extended anatomical coverage. In this context, the use of a standardized annotation checklist and a multi-expert review process represents a practical strategy to improve consistency, reproducibility, and generalizability in intra-abdominal CV research. Given the rapid evolution of the field, newer architectures continue to emerge. However, the aim of the present study was to evaluate the feasibility of a standardized methodological and annotation framework under a consistent experimental design, rather than to continuously rebuild the benchmark in response to newly released models. Although the present segmentation experiments focused on Transformer-based architectures, CNN-based approaches such as U-Net or DeepLabV3+ remain valid alternatives, may perform competitively in smaller datasets, and should be considered in future comparative studies to better define architecture-specific strengths in data-limited settings.

Of note, Peng *et al.* demonstrated an AI system for real-time quality control in laparoscopic liver surgery, underscoring clinical applicability when models are integrated into workflow-aware systems[40]. Although developed for a procedure-specific context, this illustrates how robust anatomical recognition frameworks can be translated into intraoperative feedback mechanisms once performance and interpretability are validated. Extending similar principles to phase-independent SL could enable real-time mapping of peritoneal disease, bridging algorithmic feasibility with intraoperative clinical utility. Together, these studies reinforce that the development of clinically applicable intra-abdominal CV systems will depend on larger, publicly accessible datasets and improved standardization of annotation strategies.

Among the limitations of this study, it should be noted that it was conducted retrospectively and within a single center. Certain anatomical structures were underrepresented, and not all types of intra-abdominal malignancies were included, which influence the generalizability of model performance. Notably, the present dataset was not designed as a balanced benign-versus-malignant lesion classification cohort. Rather, the primary aim was to evaluate the feasibility of PM detection and segmentation in real-life settings. The related question of discriminating malignant from benign peritoneal lesions was addressed separately in a previous pathology-grounded study[41]. In addition, the relatively limited sample size and absence of external validation restrict the extrapolation of these findings to other institutions and surgical settings.

Despite these limitations, a CV model capable of detecting and segmenting PM and anatomical regions during SL could serve as an objective tool to estimate tumor burden and lesion distribution in a manner analogous to the Peritoneal Cancer Index (PCI), while reducing dependence on operator experience and

subjective assessment. Future research should aim to overcome the current limitations of PM segmentation by developing robust, generalizable models capable of accurately identifying and quantifying tumor presence and location across diverse intra-abdominal contexts.

In conclusion, this study evaluated the feasibility of CV-based detection and segmentation of intra-abdominal anatomical structures and PM during SL performed in a phase-independent context. While detection and segmentation of major anatomical structures were achievable, PM and topography-based structure segmentation remained limited by visual heterogeneity, lack of clear landmarks or organ-specific features, and reduced pixel representation. Topographically informed label grouping improved segmentation consistency and annotation efficiency, underscoring that balanced annotation design is essential to improve robustness under sparse and heterogeneous labeling conditions. These findings provide a methodological foundation for future efforts to develop robust DL models for phase-independent intraoperative assessment in oncologic laparoscopy.

DECLARATIONS

Authors' contributions

Conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing - original draft, writing - review and editing, visualization, project administration: Tozzi F
Methodology, software, validation, formal analysis, resources, data curation, writing - review and editing, visualization: Mousavi SA, De Muynck R
Validation, data curation, writing - review and editing, visualization: Quintini D
Data curation, investigation: Molnar A, Van Vaerenbergh F, De Lille X, Van Liefferinge M
Resources, writing - review and editing: Ceelen W
Conceptualization, resources, methodology, writing - review and editing, supervision: Willaert W, De Neve W, Rashidian N

Availability of data and materials

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

AI and AI-assisted tools statement

During the preparation of this work, the authors used AI-assisted tools for image creation in the graphical abstract. The authors reviewed and edited all AI-assisted content and accept full responsibility for the published material.

Financial support and sponsorship

This research has received the Special Research Fund (BOF) of Ghent University for Interdisciplinary Research Projects (BOF/IOP/2022/049). Francesca Tozzi and Seyed Amir Mousavi are supported by this grant.

Conflicts of interest

Rashidian R (Niki Rashidian, or Nikdokht Rashidian) is a Junior Editorial Board Member of *Artificial Intelligence Surgery*. Rashidian R was not involved in any part of the editorial process for this manuscript, including reviewer selection, manuscript handling, or decision-making. The other authors declare that they have no conflicts of interest.

Ethical approval and consent to participate

This study was approved by the Institutional Review Board of Ghent University Hospital, Belgium (approval reference: BC09947). Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

Supplementary Materials

Supplementary Materials

REFERENCES

1. Anteby R, Horesh N, Soffer S, et al. Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. *Surg Endosc.* 2021;35:1521-33. DOI PubMed
2. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg.* 2019;270:414-21. DOI PubMed PMC
3. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg.* 2018;268:70-6. DOI PubMed PMC
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44-56. DOI PubMed
5. Haenssle H, Fink C, Toberer F, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol.* 2020;31:137-43. DOI PubMed
6. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* 2018;29:1836-42. DOI PubMed
7. Haenssle HA, Winkler JK, Fink C, et al. Skin lesions of face and scalp - classification by a market-approved convolutional neural network in comparison with 64 dermatologists. *Eur J Cancer.* 2021;144:192-9. DOI PubMed
8. Li M, Huang Z, Shan Q, et al. Performance and comparison of artificial intelligence and human experts in the detection and classification of colonic polyps. *BMC Gastroenterol.* 2022;22:517. DOI PubMed PMC
9. Loftus TJ, Vlaar AP, Hung AJ, et al. Executive summary of the artificial intelligence in surgery series. *Surgery.* 2022;171:1435-9. DOI PubMed PMC
10. De Backer P, Eckhoff JA, Simoens J, et al. Multicentric exploration of tool annotation in robotic surgery: lessons learned when starting a surgical artificial intelligence project. *Surg Endosc.* 2022;36:8533-48. DOI PubMed
11. Madad Zadeh S, Francois T, Calvet L, et al. SurgAI: deep learning for computerized laparoscopic image understanding in gynaecology. *Surg Endosc.* 2020;34:5377-83. DOI PubMed
12. Ryu K, Kitaguchi D, Nakajima K, et al. Deep learning-based vessel automatic recognition for laparoscopic right hemicolectomy. *Surg Endosc.* 2023;38:171-8. DOI PubMed
13. Den Boer RB, De Jongh C, Huijbers WTE, et al. Computer-aided anatomy recognition in intrathoracic and -abdominal surgery: a systematic review. *Surg Endosc.* 2022;36:8737-52. DOI PubMed PMC
14. Zygomalas A, Kalles D, Katsiakis N, Anastasopoulos A, Skroubis G. Artificial intelligence assisted recognition of anatomical landmarks and laparoscopic instruments in transabdominal preperitoneal inguinal hernia repair. *Surg Innov.* 2024;31:178-84. DOI PubMed
15. Zhou R, Wang D, Zhang H, et al. Vision techniques for anatomical structures in laparoscopic surgery: a comprehensive review. *Front. Surg.* 2025;12:1557153. DOI PubMed PMC
16. Igaki T, Kitaguchi D, Kojima S, et al. Artificial intelligence-based total mesorectal excision plane navigation in laparoscopic colorectal surgery. *Dis Colon Rectum*. 2022;65:e329-e33. DOI PubMed
17. Wang Z, Lu B, Long Y, et al. Autolaparo: a new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: Wang L, Dou Q, Fletcher PT, Speidel S, Shuo Li S, Editors. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2022 Sep 18-22; Singapore. Cham: Springer; 2022. DOI
18. Narihiro S, Kitaguchi D, Hasegawa H, Takeshita N, Ito M. Deep learning-based real-time ureter identification in laparoscopic colorectal surgery. *Dis Colon Rectum.* 2024;67:e1596-9. DOI PubMed
19. Carstens M, Rinner FM, Bodenstedt S, et al. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Sci Data.* 2023;10:3. DOI PubMed PMC
20. Kamtam DN, Shrager JB, Malla SD, et al. Deep learning approaches to surgical video segmentation and object detection: a scoping review. *Comput Biol Med.* 2025;194:110482. DOI PubMed
21. Salort-Benejam L, Agudo A. NeRFscopy: neural radiance fields for in-vivo time-varying tissues from endoscopy. *arXiv* 2026;arXiv:2602.15775. Available from https://arxiv.org/abs/2401.00044 [accessed 27 May 2026].
22. Chen H, Gou L, Fang Z, et al. Artificial intelligence assisted real-time recognition of intra-abdominal metastasis during laparoscopic gastric cancer surgery. *npj Digit. Med.* 2025;8:9. DOI PubMed PMC

23. Wang S, Li C, Wang R, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nat Commun*. 2021;12:5915. DOI PubMed PMC
24. Mascagni P, Alapatt D, Sestini L, et al. Computer vision in surgery: from potential to clinical value. *npj Digit. Med.* 2022;5:163. DOI PubMed PMC
25. Carstens M, Vasisht S, Zhang Z, et al. Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. *NPJ Digit Med*. 2025;9:59. DOI PubMed PMC
26. Meireles OR, Rosman G, Altieri MS, et al. ; SAGES Video Annotation for AI Working Groups. SAGES consensus recommendations on an annotation framework for surgical video. *Surg Endosc.* 2021;35:4918-29. DOI PubMed
27. Rädsch T, Reinke A, Weru V, et al. Labelling instructions matter in biomedical image analysis. *Nat Mach Intell.* 2023;5:273-83. DOI
28. Sener O, Savarese S. Active learning for convolutional neural networks: a core-set approach. *arXiv* 2017;arXiv:1708.00489. Available from https://doi.org/10.48550/arXiv.1708.00489 [accessed 27 May 2026].
29. Wang CY, Yeh IH, Mark Liao HY. Yolov9: learning what you want to learn using programmable gradient information. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, Editors. European Conference on Computer Vision 2024; 2024 Sep 29-Oct 4; Milan, Italy. Cham: Springer, 2024. pp. 1-21. DOI
30. Zong Z, Song G, Liu Y. DETRs with collaborative hybrid assignments training. *arXiv* 2022;arXiv:2211.12860. Available from https://doi.org/10.48550/arXiv.2211.12860 [accessed 27 May 2026].
31. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. *arXiv* 2021;arXiv:2105.15203. Available from https://doi.org/10.48550/arXiv.2105.15203 [accessed 27 May 2026].
32. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18-24; New Orleans, LA, USA. IEEE; 2022. pp. 1280-9. DOI
33. van der Maaten L, Hinton G. Visualizing data using t-SNE. *JMLR*. 2008;9:2579-605. Available from: https://jmlr.org/papers/v9/vandermaaten08a.html. [Last accessed on 27 May 2026].
34. Tan YZ, Cheng H, Ng KW, Ngiam KY, Gao Y, Khoo ET. Evaluating image matching with robust estimators: Bridging natural and surgical domains to enhance scene understanding. *IEEE J Biomed Health Inform*. 2025;29:8847-54. DOI PubMed
35. Yang Z, Dai J, Pan J. 3D reconstruction from endoscopy images: a survey. *Comput Biol Med*. 2024;175:108546. DOI PubMed
36. Zhao L, Wang T, Chen Y, et al. A novel framework for segmentation of small targets in medical images. *Sci Rep.* 2025;15:9924. DOI PubMed PMC
37. Yengera G, Mutter D, Marescaux J, Padoy N. Less is more: surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks. *arXiv* 2018;arXiv:180508569. Available from https://doi.org/10.48550/arXiv.1805.08569 [accessed 27 May 2026].
38. Raja MA, Loughran R, Mc Caffery F. NeuroEvolution of capsule networks for computer-aided laparoscopy. *TechRxiv* 2023. Available from https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.24648231.v1 [accessed 27 May 2026].
39. Cui R, Zhang J, Pei J, Wang K, Heng PA, Qin J. Topology-constrained learning for efficient laparoscopic liver landmark detection. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2025; 2025 Sep 23-25; Daejeon, Republic of Korea. Cham: Springer; 2026. pp 585–94. DOI
40. Peng Z, Wang Z, Yan Y, et al. Development of an AI-driven digital assistance system for real-time safety evaluation and quality control in laparoscopic liver surgery. *Front. Oncol.* 2025;15:1678525. DOI PubMed PMC
41. Tozzi F, Park HM, Mousavi SA, et al. Multimodal machine learning for staging laparoscopy: a combined image analysis and morphologic tool for the discrimination of peritoneal metastasis. *Int J Surg*. 2026;112:373-83. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.