Benchmarking variability in semantic segmentation in minimally invasive abdominal surgery

LT Castro[1*], C Barata[2], P Martins[1], F Afonso[1], M Pascoal[1], C Santiago[2], L Mennillo[3], P Mira[1], D Stoyanov[3], M Chand[4], S Bano[3], AS Soares[5, 6*]

[1]General Surgery Department, Hospital Prof. Dr. Fernando Fonseca, Lisbon, Portugal.

[2]Instituto de Sistemas e Robótica, LARSyS, Instituto Superior T´ecnico, Lisbon, Portugal.

[3]Computer Science Department, University College London, London, United Kingdom.

[4]Division of Surgery and Interventional Sciences, University College London, London, United Kingdom.

[5]Lisbon School of Medicine, University of Lisbon, Lisbon, Portugal.

[5]General Surgery Department, Instituto Portuguˆes de Oncologia Lisboa, Lisbon, Portugal.

*Corresponding author(s). E-mail(s): lauracastro@campus.ul.pt; antoniosoares@medicina.ulisboa.pt;

Abstract

Purpose – Anatomical identification during abdominal surgery is subjective given unclear boundaries of anatomical structures. Semantic segmentation of these structures relies on an accurate identification of the boundaries which carries an unknown uncertainty. Given its inherent subjectivity, it is important to assess annotation adequacy. This study aims to evaluate variability in anatomical structure identification and segmentation using MedSAM by surgical residents. **Methods –** Images from the Dresden Surgical Anatomy Dataset and the Endoscapes2023 Dataset were semantically annotated by a group of surgery residents using MedSAM in the following classes: abdominal wall, colon, liver, small bowel, spleen, stomach and gallbladder. Each class had 3 to 4 sets of annotations. Inter-annotator variability was assessed through DSC, ICC, BIoU and using the Simultaneous Truth and Performance Level Estimation algorithm to obtain a consensus mask and by calculating Fleiss' Kappa agreement between all annotations and reference. **Results –** The study showed strong inter-annotator agreement among surgical residents, with DSC values of 0.84–0.95 and Fleiss' Kappa between 0.85 and 0.91. Surface area reliability was good to excellent (ICC = 0.62–0.91), while boundary delineation showed lower reproducibility (BIoU = 0.092–0.157). STAPLE consensus masks confirmed consistent overall shape annotations despite variability in boundary precision. **Conclusion –** The study demonstrated low variability in the semantic segmentation of intraperitoneal organs in minimally invasive abdominal surgery, performed by surgical residents using MedSAM. While DSC and Fleiss' Kappa values confirm strong inter-annotator agreement, the relatively low BIoU values point to challenges in boundary precision, especially for anatomically complex or variable structures. These results establish a benchmark for expanding annotation efforts to larger datasets and more detailed anatomical features.

Keywords: semantic segmentation, surgery, abdominal anatomy, agreement

1 Background

1.1 Introduction

Learning to identify anatomical structures is a fundamental requirement to adequately perform surgical procedures. Surgical residents acquire this skill during their training primarily by observing or performing surgeries under the supervision of consultant surgeons. The guidance provided to residents during these procedures is predominantly verbal. Despite the critical nature of this training, there currently exists no method to process intraoperative video at scale, leading to potential discrepancies in the understanding of precise anatomical boundaries in the operative setting. This gap highlights the need for more systematic and scalable approaches to training and evaluating anatomical identification during surgery, particularly with the advent of supervised AI algorithms in medical imaging [1].

1.2 Related work

Supervised learning, a key component in developing AI algorithms for medical imaging, relies heavily on high-quality annotations. However, the identification of anatomical structures is inherently subjective, influenced by the experience level and familiarity of the annotator with the anatomical structures in question [2]. The gold standard for medical imaging segmentation has been manual annotation, which is a time-consuming process that often requires a high degree of expertise [3]. Annotators can achieve high agreement but will likely create different contours when asked to annotate the same structure, as shown by Yang *et al.* [4]. This is further exacerbated when evaluating multiple segmentation annotations. Consequently, there exists a degree of variability in how different annotators identify these structures. While there already exists experience in this assessment in the context of radiology [5] and dermoscopy [6] images, there is little knowledge regarding surgical images.

Furthermore, automation of imaging segmentation through computer vision algorithms such as the Medical Segment Anything Model (MedSAM)[7], which uses a bounding box mechanism, has the potential to help users efficiently define cohesive areas within medical images and potentially accelerating the annotation process[7].

In the intraoperative setting, understanding of the relevant anatomy is still mostly based on verbal discussion during an operation. The senior surgeon is not always able to outline every structure, as they are scrubbed and therefore unable to interact with non-sterile equipment. This variability is crucial to understand because it can affect the performance and reliability of AI algorithms trained on these annotations [4]. Despite its importance, the degree of variability in anatomical identification among annotators has not been adequately characterized in the context of intraoperative images, posing an unsolved challenge.

1.3 Objective

This study aims to address the identified knowledge gap by analyzing annotations performed by surgical residents for intra-abdominal organ segmentation within a large dataset, using MedSAM. The primary objective is to quantify inter-annotator variability in the segmentation of abdominal anatomy in minimally invasive surgery and to evaluate whether segmentation using MedSAM yields consistent results across annotators with different levels of surgical experience. The study also aims to validate the use of MedSAM for intraoperative organ segmentation and to identify areas of higher and lower certainty within each organ. Through this analysis, we seek to provide a clearer understanding of segmentation variability and offer recommendations to support surgical training and the development of reliable AI models for intraoperative use.

2 Methods

2.1 Overview

The proposed study assesses the annotation performance of surgical residents on intraoperative images from the open-source Dresden Surgical Anatomy Dataset [8] and from the Endoscapes2023 Dataset [9] using a semi-automated segmentation tool, Medical Segment Anything Model (MedSAM) [7]. 5 annotators were asked to segment the same intra-abdominal structure per image among the 7 classes selected for this study (abdominal wall, colon, liver, small bowel, spleen, stomach and gallbladder). The segmentation masks were then merged using the STAPLE algorithm [10] and compared to the expert-annotated reference labels provided in the dataset. Figure 1 shows an overview of the analysis made in this study.

2.2 Datasets

The Dresden Surgical Anatomy Dataset (DSAD) [8] provides expert semantic segmentation labels of 11 anatomical structures in laparoscopic videos of surgeries performed on 32 patients. The segmentation masks of intra-abdominal organs were manually generated by several resident surgeons using a polygon annotation tool, before independent review by a board-certified consultant surgeon in minimally invasive surgery. Out of the eleven anatomical classes provided in the dataset, six were selected for annotation in this study (abdominal wall, liver, spleen, colon, small bowel, and stomach), with respectively 1206, 1023, 1191, 1374, 1168, and 1430 labels per class. The other five intra-abdominal structures of the dataset were not selected because of their

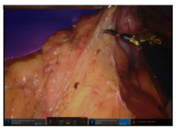
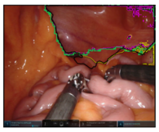
Population	Interface	Annotations	Comparison	Metrics
General Surgery Residents (annotators)	MedSAM 	3 sets: - Abdominal wall - Liver - Spleen - Colon - Small bowel - Stomach 4 sets: - Gallbladder	- Between Annotators - Annotators vs Reference 	- Variability - STAPLE - Fleiss Kappa - ICC - BIoU - DSC
DSAD (reference)	Polygon-based tool 	3 sets of all structures, revised by experts	- Annotators vs Reference	- DSC
Endoscapes (reference)	Instance + semantic segmentation 	5 trained annotators, 3 surgeons and 2 computer scientists with domain- specific expertise	- Annotators vs Reference	- DSC

Fig. 1: Visual representation of the analysis made in this study. Annotations done by general surgery residents using MedSAM were compared to references (DSAD and Endoscapes).

.

retroperitoneal position (ureter, pancreas), extraperitoneal position (vesicular glands), or relatively small size (inferior mesenteric artery, intestinal veins). Crucially, the DSAD was chosen for this study to avoid data leakage and bias during the annotation process.

The Endoscapes2023 [9] is a dataset of 201 laparoscopic cholecystectomy videos with annotations targeted at automated assessment of the Critical View of Safety by three clinical experts. It's contents are divided into 3 sub-datasets. For the purpose of this work we used the *Endoscapes-Seg50* sub-dataset, which contains 493 frames from 50 videos annotated with instance and semantic segmentation masks for 5 anatomical structures/regions (Gallbladder, Cystic Duct, Cystic Artery, Cystic Plate, Hepatocystic Triangle Dissection). Only the gallbladder annotations were used in this work.

These datasets were not used to train MedSAM. Masks from these datasets served as reference for this analysis.

2.3 Medical Segment Anything Model

The Medical Segment Anything Model (MedSAM) [7] is a deep-learning foundation model for promptable medical image segmentation. It shares the same architecture as the original Segment Anything Model (SAM) [11], which is designed to semiautomatically generate segmentation masks from prompt inputs (points, bounding boxes, binary masks, and text). The MedSAM model has been specifically trained on a curated corpus of publicly available medical image segmentation datasets to adapt the original SAM model to the medical domain. In the context of data annotation, the main benefit of using this model lies in its ability to generate accurate segmentation masks in a time-efficient way when compared with manual segmentation, which is time-consuming and has variable degrees of consistency, as well as enabling the analysis of large-scale datasets. In this study we are now aiming to acess inter-annotator variability using MedSAM (bounding box prompting).

2.4 Annotations

A total of 5 general surgery residents performed the annotations in this study(annotator 1,2,3,4 and 5). Not all annotators annotated all structures, specifically: annotators 1 and 2 annotated all structures, annotator 3 annotated 3 structures (abdominall wall, colon and stomach), annotator 4 annotated 3 structures (liver, gallbladder and spleen) and annotator 5 annotated 2 structures (galldladder and small intestine). Annotators 1 and 2 were 1st year residents. Annotator 3 was a 2nd year resident, annotator 4 was a 4th year resident and annotator 5 was a 6th year resident. Only the most senior of these residents (annotator 5) had annotation experience before this study.

Image annotations were performed using (MedSAM) [7] through bounding box prompting to generate segmentation masks. Among these annotations, 1206 images of the abdominal wall, 1023 images of the liver, and 1191 images of the spleen, 1374 images of the colon, 1168 images of the small bowel, 1430 images of the stomach classes were annotated by five annotators, generating three sets of annotation masks per image w.r.t their class. 493 images of the gallbladder were annotated by 4 annotators, generating 4 sets of annotation masks.

The MedSAM graphic user interface was used to annotate all images using bounding box prompts. The instructions for annotating each anatomical structure were provided in the DSAD [8], with additional guidelines put in place to minimize annotation variability during the annotation process, namely: each structure or separate portion of the structure had to be annotated using only one bounding box; when there was an object such as surgical instrument, gauze or other element occluding part of the structure, the annotation had to be performed using one bounding box on each side of the structure (the occluding object could not be encompassed within the bounding box).

3 Evaluation Metrics

In line with current recommendations [3], the Dice Similarity Coefficients (DSC) between each annotator's mask and the reference mask were calculated for each image, following:

$$\text{DSC} = \frac{2|\text{Mask}_A \cap \text{Mask}_B|}{|\text{Mask}_A| + |\text{Mask}_B|}$$ Where:

- $|\text{Mask}_A \cap \text{Mask}_B|$ is the intersection area of the two masks.
- $|\text{Mask}_A|$ is the area of Mask A. • $|\text{Mask}_B|$ is the area of Mask B.

However, the DSC is a pair-wise comparison, whereas the present study aims to assess variability between multiple annotators and the reference. To estimate the degree of variability between annotators, the intersection areas of all masks were assessed and presented as the intersection of 3, 2, and 1 masks over the union area of all masks, according to the formula below. These results were then mapped as heat maps to identify the annotated areas with greater and lesser intersections between annotators and reference. The general formula for *n* masks is:

$$\text{Variability} = \frac{{}_i\big|\big|\bigcap_{i=1}^{n} \text{Mask}\,{}_i\big|}{\big|\bigcup_{i=1}^{n} \text{Mask}}$$ Where:

- $\left|\bigcap_{i=1}^{n} \text{Mask}\,{}_i\right|$ is the intersection area of all *n* masks.

- $\left|\bigcup_{i=1}^{n} \text{Mask}_i\right|$ is the union area of all *n* masks.

A boundary intersection-over-union (BIoU) analysis was performed to better capture differences in boundary delineation, which are especially important in abdominal organ segmentation where even small contour discrepancies can significantly affect clinical interpretation and downstream shape-based analyses.

The general formula is:

$$\text{Boundary IoU} = \frac{|\partial A \cap \partial B|}{|\partial A \cup \partial B|}$$

Where:

- $\partial A \cap \partial B$ is the intersection of the boundaries of A and B.
- $\partial A \cup \partial B$ is the union of the boundaries of A and B.
- $|\cdot|$ denotes the cardinality (number of pixels) in the boundary sets.

To better assess interannotator variability, Intraclass Correlation Coefficient (ICC) was measured between each annotator and reference for each structure. It is a statistical measure used to assess reliability or agreement of measurements made by different observers. High ICC indicated high reliability.

ICC (2,1) was calculated following:

$$ICC(2,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$$

Where:

- n: number of subjects
- k: number of raters

ICC was calculated for surface area, which reflects size and boundary complexity of the masks.

Additionally, according to [4], the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [10] was used to obtain consensus annotation masks from the sets of individually generated annotations. STAPLE is an expectation maximization method designed to combine multiple segmentation masks and estimate a consolidated mask to assess the performance of each segmentation approach. The method considers a collection of segmentations and computes a probabilistic estimate of the true segmentation, as well as measures the performance level of each segmentation. The STAPLE algorithm was applied with a foreground prior probability of 1.0, assigning equal sensitivity and specificity priors to all annotators. The resulting probabilistic consensus map was binarized using a confidence threshold of 0.5 to generate the final consensus segmentation mask.

These consolidated masks were then compared with the external reference masks from the Dresden dataset, using the Dice Similarity Coefficient.

Finally, the Fleiss' kappa coefficient was used to assess the pixel-wise interannotator reliability of segmentation masks generated by multiple annotators [4]. Fleiss' kappa measures the level of agreement between annotators, correcting for the likelihood of chance agreement. The Fleiss' kappa coefficient was computed following:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where $\Pr(a)$ represents the actual observed agreement, and $\Pr(e)$ represents the chance agreement.

The script used to compute the statistical analysis presented in this study can be found in this shared GitHub repository.

4 Results

4.1 Generated annotations

A total of 5 annotators independently created 21423 new annotations using MedSAM, as can be seen in table 1. DSC was calculated between annotated masks and references. Figure 2 shows the usage of the MedSAM interface and example annotations between annotators and reference (DSAD).

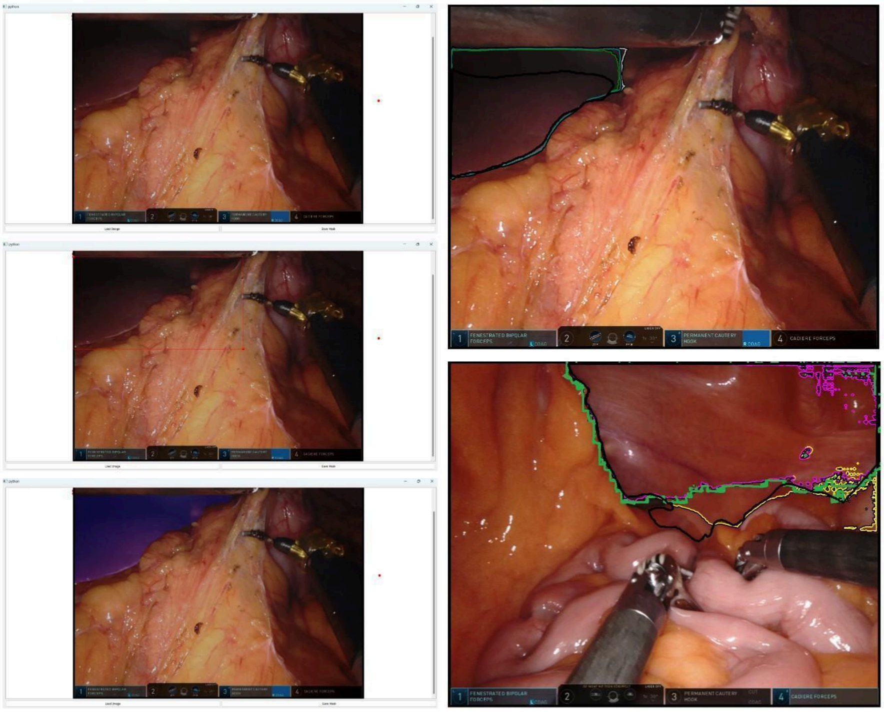


Fig. 2: Left column: MedSAM interface and annotation process. Right column: annotations from general surgery residents (colors) and reference(black).

Table 1: Generated annotations. Abdominal wall, liver, spleen, colon, small bowel and stomach had 3 sets of annotations. The gallbladder had 4 sets of annotations. Mean DSC between annotations and references ranged from 0.84 to 0.91.

	Abdom. wall	Liver	Spleen	Colon	Sm. bowel.	Stomach	Gallbladder
No. annotators	3	3	3	3	3	3	4
No. images	1206	1023	1191	1374	1168	1430	493
Tot. annotations	3614	3064	3566	4122	3499	4290	1968
Mean DSC	0.87	0.91	0.88	0.85	0.85	0.88	0.84
(95% CI)	(0.87-0.88)	(0.90-0.91)	(0.87-0.88)	(0.84-0.85)	(0.86-0.87)	(0.87-0.88)	(0.84-0.85)

4.2 Qualitative variability assessment

As can be seen in table 2, there is a significant intersection between the masks of all anatomical structures, indicating high agreement between annotators. The colon and stomach had the lowest average of 3 mask intersections, while the liver has the highest.

Table 2: Average percentage of mask intersections in different anatomical structures.

Intersection Type	Abdominal wall	Liver	Spleen	Colon	Small intestine	Stomach	Gallbladder
4-mask intersection	-	-	-	-	-	-	63.04
3-mask intersection	79.84	81.50	79.98	70.00	78.94	75.03	13.37
2-mask intersection	11.91	9.89	12.00	13.32	9.44	13.67	8.53
1-mask intersection	8.24	8.61	8.01	16.68	11.62	11.30	10.59

Figure 3 represents the areas of overlap as overlay, representative for high and low agreement.

The gallgladder had 4 annotations and had a 4-mask intersection of 63.04%. The other structures analysed were annotated by 3 annotators and had 3 mask intersection average between 70% (colon) and 81.5% (liver), indicating high agreement between annotators.

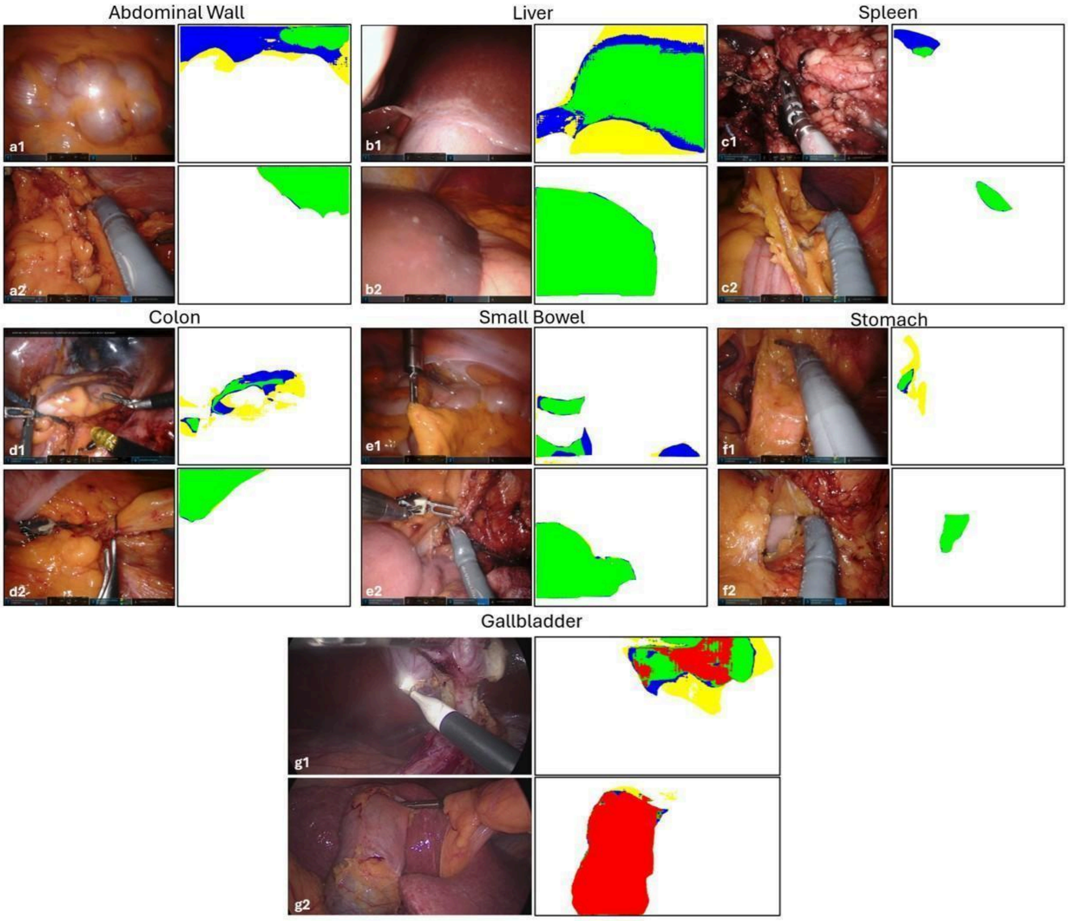


Fig. 3: Areas of annotation mask overlap per class. Red represents the areas of 4mask intersection, green represents areas of 3-mask intersection, blue represents areas of 2-mask intersection, and yellow represents areas of one-mask intersection. Figures suffixed by 1 show low agreement and figures suffixed by 2 show high agreement.

4.3 Quantitative variability assessment

4.3.1 ICC

ICC was calculated for surface area between each annotator's mask and the reference mask.

Table 3: Mean Surface Area ICC for Different Anatomical Structures. ICC values ranged from 0.63 (abdominal wall) and 0.91 (spleen).

Structure	Surface area	95% CI
Abdominal wall	0.63	0.60-0.66
Liver	0.86	0.84-0.87
Spleen	0.91	0.90-0.92
Colon	0.62	0.59-0.66
Small bowel	0.67	0.59-0.74
Stomach	0.83	0.82-0.85
Gallbladder	0.64	0.60-0.68

Mean surface area intraclass correlation coefficients (ICCs) and corresponding 95% confidence intervals (CIs) are presented in Table 3. Reliability was excellent for the spleen (ICC = 0.91) and good for the liver (ICC = 0.86) and stomach (ICC = 0.83). Moderate reliability was observed for the abdominal wall (ICC = 0.63), colon (ICC = 0.62), small bowel (ICC = 0.67), and gallbladder (ICC = 0.64). Overall, reproducibility was highest for solid abdominal organs, particularly the spleen and liver, whereas structures with greater shape variability and less well-defined boundaries, such as the bowel and gallbladder, demonstrated lower agreement.

4.3.2 Boundary Intersection Over Unit (BIoU)

Table 4 represents the average Boundary Intersection Over Unit (BIoU) between new annotations and reference.

Table 4: BIoU for Different Anatomical Structures

Structure	Average BIoU	95% Confidence Interval
Abdominal wall	0.124	0.122-0.126
Liver	0.146	0.144-0.149
Spleen	0.157	0.154-0.160
Colon	0.092	0.090-0.093
Small bowel	0.129	0.127-0.130
Stomach	0.122	0.120-0.124
Gallbladder	0.148	0.144-0.152

The highest agreement in boundary delineation was observed for the spleen (BIoU = 0.157, 95% CI: 0.154–0.160), followed by the gallbladder (BIoU = 0.148, 95% CI: 0.144–0.152) and liver (BIoU = 0.146, 95% CI: 0.144–0.149). Lower boundary agreement was noted for the abdominal wall (BIoU = 0.124, 95% CI: 0.122–0.126), stomach (BIoU = 0.122, 95% CI: 0.120–0.124), and small bowel (BIoU = 0.129, 95% CI: 0.127–0.130). The lowest BIoU was found for the colon (BIoU = 0.092, 95% CI: 0.090–0.093).

Boundary reproducibility was highest for solid organs (with clearly defined borders), while hollow organs and structures with greater shape variability demonstrated lower agreement.

4.3.3 Fleiss' Kappa agreement

Table 5 represents the average Fleiss' Kappa agreement between new annotations and reference. The agreement is extremely high across all classes, from 0.85 (gallbladder) to 0.91(liver), indicating high similarity between annotations and reference. The lowest agreement was found for the gallbladder, colon and abdominal wall classes.

Table 5: Fleiss' Kappa Agreement for Different Anatomical Structures

Structure	Average Fleiss' Kappa agreement	95% Confidence Interval
Abdominal wall	0.86	0.85-0.87
Liver	0.91	0.90-0.91
Spleen	0.89	0.88-0.90
Colon	0.86	0.85-0.86
Small bowel	0.89	0.89-0.90
Stomach	0.89	0.89-0.90
Gallbladder	0.85	0.84-0.86

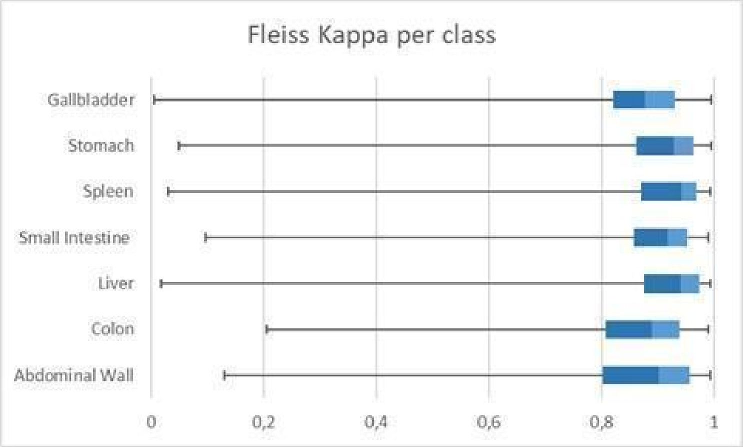


Fig. 4: Boxplot of Fleiss Kappa per class. The first quartile values range widely from 0 to 0.86. In the second and third quartiles variability is minimal (ranging from 0.86 to 0.91).

4.4 Consensus assessment using the STAPLE algorithm

The STAPLE algorithm was used to create a consensus mask between annotators and compare it with the reference mask.
As can be seen in table 6, there is an extremely high DSC between the STAPLE consensus masks and the reference in this dataset. The lowest agreement was found for the gallbladder class.

Table 6: Dice similarity coefficient average between dataset reference and STAPLE masks. Mean DSC was above 0.84 for all structures. The liver and small bowel achieved the highest mean DSC (0.95).

Structure	Mean DSC	95% Confidence Interval
Abdominal wall	0.93	0.93-0.94
Liver	0.95	0.95-0.96
Spleen	0.91	0.92-0.93
Colon	0.94	0.92-0.93
Small bowel	0.95	0.95-0.95
Stomach	0.93	0.92-0.93
Gallbladder	0.84	0.82-0.86

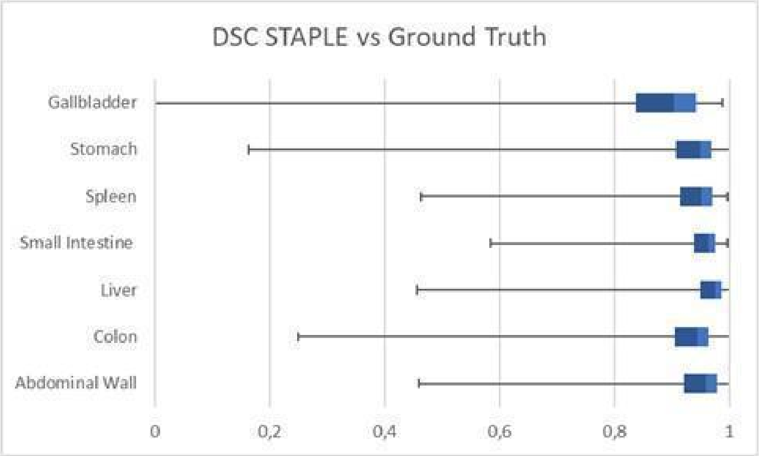


Fig. 5: Boxplot of DSC calculated between STAPLE masks and reference masks. The first quartile values range widely from 0 to 0.84. In the second and third quartiles variability is minimal (ranging from 0.84 to 0.95).

4.5 Annotator performance analysis

An annotator performance analysis was performed by calculating mean DSC between the annotated masks and references. Table 5 demonstrates average annotator DSC per organ.

Table 5: Mean DSC per annotator per structure

Structure	Annotator 1	Annotato r 2	Annotator 3	Annotator 4	Annotator 5
Abdominal wall	0.89	0.90	0.83	-	-
Liver	0.92	0.90	-	0.90	-
Spleen	0.90	0.90	-	0.84	-
Colon	0.85	0.84	0.86	-	-
Small bowel	0.90	0.89	-	-	0.80
Stomach	0.88	0.88	0.87	-	-
Gallbladder	0.85	0.83	-	0.88	0.83
Average DSC	0.88	0.89	0.85	0.87	0.82

Higher DSC were observed for annotators 1 and 2. These annotators annotated all 7 structures. The lowest DSC were observed for annotator 5. This annotator only annotated 2 structures (small bowel and gallbladder).

4.6 Qualitative analysis of factors influencing lower performance

We performed a qualitative analysis of masks with the lowest mask intersection percentages (3-mask intersection for the abdominal wall, liver, spleen, colon, small bowel and stomach, and 4-mask intersection for the gallbladder). Representative images of each class are represented in Figure 6.

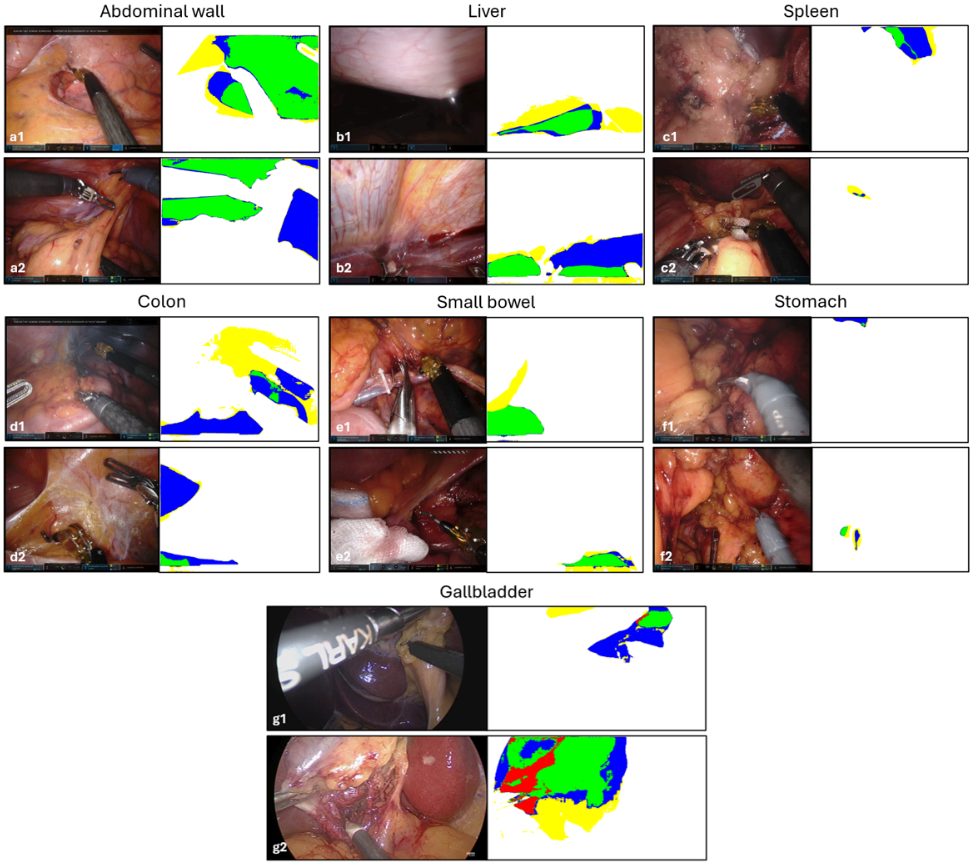


Fig. 6: Examples of masks with the lowest mask intersection percentages, per structure. Red represents the areas of 4-mask intersection, green represents areas of 3-mask intersection, blue represents areas of 2-mask intersection, and yellow represents areas of one-mask intersection.

Transversally to all structures, surgical conditions influenced accuracy mainly due to shadows and surgical smoke. Anatomical positioning difficultated interface identification as interference with other structures created areas of low contrast, particularly in the case of background structures, which generally achieved lower accuracy.

Besides these general limitations, accuracy of segmentation was mainly influenced by two categories of factors: organ-specific factors and MedSAM-related factors.

Organ-specific challenges varied by structure. For the liver and spleen, segmentation was hindered by frequent shadows and their appearance as background structures.

The presence of mesocolic fat in the colon and discontinuous visualization complicated boundary identification. The variable position and discontinuous visualization of the small bowel often created areas of low contrast, making segmentation difficult. For the stomach, the presence of gastroepiploic vessels and coverage by the greater omentum were the main challenges. Finally, the gallbladder segmentation accuracy was heavily influenced by the stage of dissection, with the cystic plate and surrounding peritoneum presented as confounding structures.

The MedSAM-related factors included a tendency to demarcate regions based on textural similarity rather than anatomical boundaries, particularly under poor lighting conditions, or interference from surgical instruments or gauze within the field.

5 Discussion

The findings of this study demonstrate that the overall variability in intra-abdominal organ segmentation in laparoscopic images, when combining human assessment by a surgery resident and the MedSAM algorithm, is minimal. This low variability was consistently observed across all classes of intraperitoneal organs studied, including the abdominal wall.

These results were obtained through three distinct analyses: the intersection of different annotations as a percentage of total annotation pixels, the Dice similarity coefficients (DSC) between STAPLE consensus masks and dataset reference ground truth, and Fleiss' kappa values. When examining the percentage of total annotation pixels (Table 2) and Fleiss' kappa agreement (Table 5 and Figure 4), the gallbladder and colon classes frequently showed the lowest values, while the liver class consistently exhibited some of the highest agreement. We hypothesize that these differences are attributable to organ-specific factors; for example, the presence of colonic haustrae and mesocolic fat appendages may complicate segmentation, whereas the liver's more compact structure and fewer interposed tissues likely facilitate more consistent delineation.

Comparing the DSC between the STAPLE consensus masks and the reference ground truth (Table 5 and Figure 5), the mean DSC exceeded 0.80, indicating a very high level of agreement. Notably, the boxplots reveal that while the first quartile values range widely from 0 to 0.8 (representing 25% of the data), this subset reflects annotations with comparatively poorer agreement or performance, highlighting an area where annotation consistency or model reliability still requires improvement. Future research should focus on identifying the factors contributing to this variability. However, in the second and third quartiles(covering 75% of the data)variability is minimal (ranging from 0.8 to 1), with mean DSC values above 0.8 across all structures. This indicates a strong correlation among new masks as assessed by STAPLE and

Fleiss' kappa, and a high degree of overlap with expert manual annotations, as reflected by elevated DSC scores.

An analysis of interobserver reliability for surface area annotations further supports the robustness of this metric across different abdominal structures. Excellent reliability was observed for solid, well-defined organs such as the spleen (ICC = 0.91) and liver (ICC = 0.86), with good agreement for the stomach (ICC = 0.83). In contrast, structures with greater shape variability and less distinct boundaries,including the abdominal wall (ICC = 0.63), colon (ICC = 0.62), small bowel (ICC = 0.67), and gallbladder (ICC = 0.64),showed moderate reproducibility. These results suggest that surface area measurements are highly reproducible for solid organs but more sensitive to segmentation variability in morphologically complex or flexible structures. Overall, MedSAM provides reliable performance in capturing surface area of anatomically consistent structures, while more variable shapes present greater challenges for consistent boundary delineation.

Boundary delineation results, as reflected by generally low Boundary Intersection over Union (BIoU) values, highlight a key limitation of the segmentation approach. The highest boundary agreement was observed for the spleen (BIoU = 0.157), gallbladder (BIoU = 0.148), and liver (BIoU = 0.146), yet even these values suggest modest reproducibility. Lower agreement was seen for the abdominal wall (BIoU = 0.124), stomach (BIoU = 0.122), and small bowel (BIoU = 0.129), with the colon exhibiting the lowest boundary reproducibility (BIoU = 0.092). These findings indicate once more that MedSAM is more reliable in capturing the overall shape and surface area of anatomical structures than in precisely defining their boundaries. Reduced BIoU scores for hollow and morphologically complex organs likely reflect the inherent challenges of consistent boundary delineation due to shape variability and indistinct anatomical landmarks. While the method provides robust global segmentation, further refinement is needed to enhance boundary accuracy, especially for complex or flexible structures.

The consistency of these findings across annotations created independently by surgery residents with varying levels of operative experience underscores the reliability of the results. Although boundary-specific metrics remain challenging, overall agreement metrics support the validity and reproducibility of the annotations used in this study.

Interestingly, annotators 1 and 2, both first-year residents, consistently achieved some of the highest DSC scores across all structures. These annotators had the least clinical experience, but the most experience with medical imaging segmentation using MedSAM, as they annotated the most images in the study. Since annotation was performed using MedSAM, a semi-automatic segmentation tool based on bounding box initialization, repeated use likely contributed to their improved performance. By annotating more images, these residents may have developed a better understanding of how to optimally position bounding boxes to guide MedSAM's segmentation, effectively compensating for some of the model's limitations. This increased familiarity probably translated into more consistent and accurate annotations, reflected in their higher agreement scores. Similarly, some of the lowest DSC scores were achieved by annotator 5. Although annotator 5 had the most clinical experience (6th year resident), they had the least experience using MedSAM, as they only annotated 2 structures. These findings suggest that, beyond clinical expertise, experience in using MedSAM is a critical factor for achieving higher agreement—a key consideration for future annotation workflows and training protocols.

Finally, we identified several factors affecting segmentation performance (Figure 6), which can be categorized as organ-specific—such as the presence of mesocolic fat, vessels, blood, shadows, or poor lighting conditions—and MedSAM-related factors, including a tendency to demarcate regions based on textural similarity rather than anatomical boundaries, especially in areas of textural heterogeneity, or interference from surgical instruments or gauze within the field. These factors appear to negatively impact interface identification between structures, thereby reducing segmentation accuracy.

5.1 Limitations

While the study presents robust evidence for low variability in segmentation, it is important to acknowledge its limitations. The dataset used for this study involved the segmentation of whole organs in their intraperitoneal position during elective surgeries. Consequently, the findings may not be directly applicable to the segmentation of extraperitoneal organs, different parts of the same anatomical structure, or cases involving more severe disease states.

There is significant concern regarding the low number of annotators. The low number was a convenience sample where the goal was to have a high number of annotations by annotator, as opposed to a lower number from a higher number of annotators. This decision stemmed from the will to address extensively what the difference in annotation that stemmed from differences in anatomical structures found intraoperatively, hence the emphasis on increasing the number of annotations per annotator. These limitations should be further addressed in future work.

5.2 Implications Moving Forward

The results of this study have significant implications for the future of surgical training and AI development in medical imaging. The evidence of low variability among surgery residents in annotating laparoscopic images of whole intraperitoneal anatomy using MedSAM suggests that it is feasible to scale annotations performed by residents to other datasets. This finding also sets a benchmark for evaluating the performance of annotations by medical students or non-clinical annotators. However, the observed challenges in precise boundary delineation indicate that additional training or refinement may be necessary when annotating structures with complex or variable borders and highlights possible limitations of AI algorithms trained based on these masks Furthermore, this study opens the door for surgery residents to annotate more complex structures, such as different parts of the same anatomical structures, more severe disease states, and dissection planes. To build on these findings, future research should focus on several key areas:

- **Segmentation of Extraperitoneal Organs and Complex Structures:** Extending the analysis to include extraperitoneal organs, different parts of the same anatomical structure, blood vessels and other structures, as well as more severe disease states to determine if the low variability observed in this study holds in these contexts.
- **Impact of Surgical Experience:** Investigating how different levels of surgical experience affect annotation quality and consistency to provide more tailored training and support.
- **Task Shifting and Non-Clinical Annotators:** Exploring the potential for task shifting, allowing medical students or non-clinical annotators to perform initial annotations, which can then be refined by more experienced annotators. This approach could significantly increase the scalability of annotation efforts.

In conclusion, this study provides strong evidence of low variability in intraabdominal organ segmentation among surgical residents, laying the groundwork for more efficient and scalable annotation practices in medical imaging. By addressing the identified limitations and exploring new avenues for research, the medical community can continue to enhance the accuracy and consistency of anatomical annotations, ultimately improving the quality of AI models and surgical training programs.

6 Conclusion

This study makes several key contributions to the field of surgical data annotation and algorithm validation for intraoperative image segmentation. First, it demonstrates that semantic segmentation of intraperitoneal organs using MedSAM, when guided by surgical residents, results in consistently high inter-annotator agreement, indicating low variability across users with differing levels of clinical experience. This confirms the reliability of MedSAM-assisted annotations for laparoscopic anatomy and intraoperative images. The study provides quantitative validation of MedSAM for intraoperative use, showing that surface area features are particularly robust, with high inter-class correlation (ICC > 0.87) across all structures. This positions MedSAM as a viable tool for generating high-quality ground truth segmentations in minimally invasive surgery datasets.

Furthermore, this work reveals that experience with the annotation tool, rather than clinical seniority, was the strongest predictor of segmentation quality. Annotators who labeled more images, even if less clinically experienced, produced more consistent results—highlighting the importance of tool-specific training in annotation workflows. However, a notable limitation of MedSAM was observed in boundary delineation accuracy, particularly for hollow or morphologically complex organs, where boundary agreement was modest. This highlights an area for future refinement to improve the precision of boundary segmentation.

Finally, by identifying specific factors contributing to segmentation uncertainty—including anatomical complexity, textural variability, and interference from surgical instruments—this study offers valuable insights into current model limitations and guidance for future model development.

Together, these findings establish a validated framework for scaling surgical annotation using semi-automated tools like MedSAM and offer a foundation for task-shifting annotation to medical students or non-clinical staff, thereby accelerating dataset creation and AI development in surgical imaging.

7 Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

[1] Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., Watanabe, Y., Alseidi, A., Redan, J.A., Alfieri, S., Costamagna, G., *et al.*: Computer vision in surgery: from potential to clinical value. npj Digital Medicine **5**(1), 163 (2022)

[2] Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. European radiology **29**, 1391–1399 (2019)

[3] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods, 1–18 (2024)

[4] Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., Antani, S.K.: Assessing inter-annotator agreement for medical image segmentation. IEEE Access **11**, 21300–21312 (2023)

[5] Moli`ere, S., Hamzaoui, D., Granger, B., Montagne, S., Allera, A., Ezziane, M., Luzurier, A., Quint, R., Kalai, M., Ayache, N., Delingette, H., Renard-Penna, R.: Reference standard for the evaluation of automatic segmentation algorithms: Quantification of inter observer variability of manual delineation of prostate

contour on mri. Diagnostic and Interventional Imaging **105**(2), 65–73 (2024) https://doi.org/10.1016/j.diii.2023.08.001

[6] Ribeiro, V., Avila, S., Valle, E.: Handling Inter-Annotator Agreement for Automated Skin Lesion Segmentation (2019). https://arxiv.org/abs/1906.02415

[7] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)

[8] Carstens, M., Rinner, F.M., Bodenstedt, S., Jenke, A.C., Weitz, J., Distler, M., Speidel, S., Kolbinger, F.R.: The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. Scientific Data **10**(1), 1–8 (2023)

[9] Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia Vazquez, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., & Padoy, N. (2024). Endoscapes2023, A Critical View of Safety and Surgical Scene Segmentation Dataset for Laparoscopic Cholecystectomy (version 1.0.0). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/czwq-jh81

[10] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., & Girshick, R.B. (2023). Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992-4003.