Journal of Robotic Surgery (2025) 19:131
https://doi.org/10.1007/s11701-025-02284-7

RESEARCH

Optimizing intraoperative AI: evaluation of YOLOv8 for real-time recognition of robotic and laparoscopic instruments

Sébastien Frey[1,2,3,8] · Federica Facente[1,3,4] · Wen Wei[4] · Ezem Sura Ekmekci[3] · Eric Séjor[2,4] · Patrick Baqué[1,2] · Matthieu Durand[1,5,6] · Hervé Delingette[3] · François Bremond[7] · Pierre Berthet-Rayne[4] · Nicholas Ayache[3]

Received: 21 November 2024 / Accepted: 11 March 2025

Abstract

The accurate recognition of surgical instruments is essential for the advancement of intraoperative artificial intelligence (AI) systems. In this study, we assessed the YOLOv8 model's efficacy in identifying robotic and laparoscopic instruments in robot-assisted abdominal surgeries. Specifically, we evaluated its ability to detect, classify, and segment seven different types of surgical instruments. A diverse dataset was compiled from four public and private sources, encompassing over 7,400 frames and 17,175 annotations that represent a variety of surgical contexts and instruments. YOLOv8 was trained and tested on these datasets, achieving a mean average precision of 0.77 for binary detection and 0.72 for multi-instrument classification. Optimal performance was observed when the training set of a specific instrument reached 1300 instances. The model also demonstrated excellent segmentation accuracy, achieving a mean Dice score of 0.91 and a mean intersection over union of 0.86, with Monopolar Curved Scissors yielding the highest accuracy. Notably, YOLOv8 exhibited superior recognition performance for robotic instruments compared to laparoscopic tools, a difference likely attributed to the greater representation of robotic instruments in the training set. Furthermore, the model's rapid inference speed of 1.12 milliseconds per frame highlights its suitability for real-time clinical applications. These findings confirm YOLOv8's potential for precise and efficient recognition of surgical instruments using a comprehensive multi-source dataset.

Keywords Computer-assisted surgery · Computer vision · Surgical instrument detection · Instrument segmentation · Robotic surgery · YoloV8

Introduction

Minimally invasive surgery has transformed surgical care, with robotic platforms becoming increasingly prevalent in developed countries [1]. Every day, thousands of robotic-assisted procedures generate vast amounts of endoscopic video data. This data provides an opportunity for artificial intelligence (AI) applications, which could soon assist surgeons by enhancing intraoperative decision-making [2, 3]. Recent advancements in computer vision (CV) and deep learning (DL) have driven substantial research interest, paving the way for AI systems capable of recognizing surgical instruments to understand the surgical scene comprehensively [4–7]. For example, the delineation of surgical instruments can help with the overlay of 3D models in augmented reality (AR) projections [8]. However, tracking the surgeon's instruments and movements is not new. Before the widespread use of CV for instrument recognition, instrument tracking could be accomplished using external sensors

Sébastien Frey and Federica Facente have contributed equally to this work.

✉ Sébastien Frey
freysebastien6@gmail.com

1 Université Côte d'Azur, Nice, France
2 Department of General Surgery, Pasteur 2 Hospital, University Hospital of Nice, Nice, France
3 Epione Team, Université Côte d'Azur, Inria, Sophia-Antipolis, Nice, France
4 Caranx Medical, Nice, France
5 Urology, Andrology, Renal Transplant Unit, Pasteur 2 Hospital, University Hospital of Nice, Nice, France
6 INSERM U1081 – CNRS UMR 7284, Nice University Côte d'Azur, Nice, France
7 Stars Team, Université Côte d'Azur, Inria, Sophia-Antipolis, Nice, France
8 Hôpital L'Archet, University Hospital of Nice, 151, Route de Saint-Antoine, Nice, France

Published online: 31 March 2025

such as electromagnetic, optical markers, or more recently, Intuitive Surgery's dV-logger® using kinematic information[9]. Because of the inevitable intrusion into the surgical setup, these solutions were too restrictive to gain the interest of the surgeon community. Offering simplicity through vision-based techniques alone, CV is now the preferred field. Although it has been in development for several decades, it is only recently that advances in machine learning (ML) and computing power have allowed its generalizability. The increased improvement in computing power, visual data storage, and deep learning methods are offering new ways of analyzing the surgical field.

From a CV perspective, surgical instrument recognition can be categorized in three approaches: detection, segmentation, and classification. Initially, detection using a bounding box and binary segmentation have been the most studied methods, yielding excellent results. Current state-of-the-art methods have reached a precision of up to 90–95% [10, 11]. Initial efforts used classical ML algorithms such as support vector machines or naive Bayes approaches [12, 13]. Subsequently, CNN-based methods have been proposed in numerous works, showing promising results thanks to their strong ability to extract features from pixel-wise semantic segmentation. The Endoscopic Vision (EndoVis) 2017 robotic instrument segmentation challenge was one of the preliminary works, where the most successful methods used a U-net neural network based on a fully convolutional neural network (FCN) structure [14].

The discrimination of instrument types and instances remains, however, a challenging task with unsatisfying results. While the initial algorithm would efficiently detect instruments, type-based segmentation was left at 54.2% in the EndoVis 2017 Challenge [14]. To improve the per-pixel classification, some focused on adding features such as depth perception, saliency maps, pose estimation, or attention mechanism [15, 16]. Other techniques focused on the learning strategy through domain adaptation, data augmentation, or unsupervised methods [17–19]. More recently, transformer-based solutions have been proposed [16, 20]. Using attention mechanisms, these models can track and predict instance segmentation, which is an attractive strategy for dynamic objects.

Nevertheless, several problems remain. First, most algorithms have been only tested on the EndoVis 2017 dataset, which includes porcine surgical images. Second, most published studies have used either robotic or laparoscopic instruments, but never both simultaneously. Nonetheless, robotic-assisted procedures often, if not always, require an assistant using laparoscopic instruments. Instrument recognition should logically include both sets of instruments for optimal understanding of the surgical scene. Third, there is a paucity of data on type-based segmentation on corrupted images: blurred, bloody, covered by smoke, covered by tissue, lack of light, or poor quality. Fourth, transformer architectures have recently become popular in this field, but they are known to be large, computationally expensive, and require large datasets and large memory [20]. Most reported works lack information on their real-time performance in surgical video. Ideally, the deep learning algorithm should work in real-time, incorporate robotic and laparoscopic instruments, and be efficient in both clean and occluded environments.

At this moment, some of the original algorithms have become robust and have stood the test of time. Our research aims to address these gaps. We hypothesize that a deep learning model, trained on a diverse, multi-source dataset encompassing both robotic and laparoscopic instruments, can achieve efficient and accurate real-time recognition even in varied and challenging conditions. In this study, we evaluate the performance of such an approach, exploring its potential integration into robotic surgery to enhance intraoperative support. To date, this paper delves into the current potential of neural networks to analyze day-to-day surgical instrument activity, and study how they can integrate into daily robotic surgery practice.

Methods

Experiments

Model description

The You Only Look Once (YOLO) model was initially developed for detection tasks only but has evolved significantly over the years. Enhancements now enable it to classify, segment, estimate poses, and detect oriented objects. For this study, we employed the eighth version of this model [24]. YOLOv8 incorporates multiple detection heads and features a self-attention mechanism combined with a pyramid network for multi-scale object detection. This design allows the model to focus on various parts of an image and efficiently detect objects of different sizes and scales. Given the constant movement and varying scales of surgical instruments, YOLOv8's capabilities make it well-suited for this application. Moreover, its rapid run-time for object detection and segmentation is ideal for real-time surgical settings. The decision to use YOLOv8 was made after a a preliminary bench comparison with the Mask-R CNN model from Detectron 2 and the SAM on the initial two datasets described below.

Tasks

We aim to evaluate the performance of YOLOv8 model across three specific tasks related to robotic and laparoscopic

Table 1 Number of instruments instances across datasets

	Instrument	Total	EndoVis	PSI-AVA	RoboTool	URAS
R	Monopolar Curved Scissors	4174	791	2609	277	497
	Prograsp Forceps	3185	1698	1071	283	133
	Bipolar forceps	4322	1180	2503	87	552
	Large Needle Driver	3644	2051	1343	127	123
L	Laparoscopic Grasper	382		275	16	91
	Laparoscopic Clip Applier	104		102	2	
	Laparoscopic Vacuum	1364		1126	38	200
	Total	17175	5720	9029	830	1596

Empty cells correspond to instruments non-present in the corresponding dataset

R robotic instruments, *L* laparoscopic instruments

instruments used in robot-assisted abdominal surgeries. First, we focus on binary instrument detection, which involves detecting the presence of any instrument by drawing bounding boxes around them. This task helps us understand the model's ability to accurately locate surgical instruments within the images. Second, we tested the multi-instrument detection, meaning the model not only localizes instruments but also classifies each detected instrument into predefined categories. This task is crucial for determining the model's ability to differentiate between various types of instruments. Finally, we evaluate the binary and the instance segmentation performances, requiring the model to provide precise boundaries for each instrument, respective to its category.

Implementation details

The training for YOLO was performed in PyTorch using the default configuration. The model used was Yolov8n-seg. The training was done for 300 epochs with earlystop set to stop the training if there was no significant improvement during five epochs. The batch size used was 16, and the training had the usual mosaic augmentations [25]. The model was trained with a 11GB GeForce GTX 1080 Ti graphics card.

Datasets

EndoVis 2017 — robotic instrument segmentation [14]

EndoVis2017 is an open dataset from the MICCAI 2017 Endoscopic Vision SubChallenge. It contains 8 x 225-frame robot-assisted nephrectomy videos, performed on the porcine abdomen, captured at 2 Hz. Ground truth labels were provided for left-view frames only. Six distinct surgical instruments were labeled, including Monopolar Curved Scissors, bipolar forceps, large needle drive, Prograsp Forceps, vessel sealer, and grasping retractor. For the purpose of this study, only the vessel sealer was not considered in our analysis.

PSI-AVA dataset [22]

The PSI-AVA dataset contributed to the recognition of tasks for holistic surgical understanding. It contains 8 robotic surgical videos of robot-assisted radical prostatectomies, performed on a Da Vinci SI3000. In addition to phase and step annotation, seven distinct surgical instruments were labeled, including Monopolar Curved Scissors, bipolar forceps, Large Needle Driver, Prograsp Forceps, suction instrument, clip applier, and laparoscopic instruments. In total, 5804 instrument instances are available in the dataset.

Table 2 Dataset split and the corresponding number of frames

Set	Total	EndoVis	PSI-AVA	RoboTool	URAS
Training set	4459	1536	2091	383	449
Validation set	494	156	232	22	84
Test set	2449	1190	1125	46	88
Total	7402	2882	3448	451	621

Robotool dataset [23]

The RoboTool dataset was released in 2021 and contains 514 manually annotated images from 20 robot-assisted abdominal surgical procedures freely available online. This dataset is originally annotated for binary instrument and background segmentation. We manually updated this dataset by adding the type of instrument. This dataset contains a wide range of interference factors in the frames, including smoke, blur, or darkness (Supplementary File 1).

URAS dataset

We developed our dataset using urological robot-assisted surgeries (URAS). Surgeries were performed on Da Vinci X from Intuitive Surgery® and registered using the Intuitive Hub under an MP4 video format. Ten surgeries were chosen for this study, comprising radical prostatectomy, partial nephrectomy, and adenomyomectomy. Anonymization of videos was the initial step, using a local modification of the IODA algorithm [26]. The resolution was reduced to 678 x 541 pixels and the frame rate was lowered at 10 frames per second (fps). The following instruments were labeled: Monopolar Curved Scissors, bipolar forceps, large needle drive, Prograsp Forceps, suction instrument, and laparoscopic grasper. Only frames from the left eye were annotated. Annotations were performed by one surgical resident and reviewed by two urological specialists.

Training, validation, and test sets (Table 1 and 2)

The overall dataset contained 7402 frames. The training dataset consisted of a compilation of the training set of EndoVis 2017 and PSI-AVA datasets, 85% of RoboTool dataset, and 449 frames of URAS dataset.

The validation set comprised the corresponding EndoVis 2017 and PSI-AVA validation set, along with 5% of the RoboTool dataset, and 84 frames from the URAS dataset.

We used the independent test sets provided by the EndoVis 2017 and PSI-AVA datasets without modification to ensure consistency and comparability with other studies utilizing these datasets. For the RoboTool dataset, we selected 10% of the data to serve as an independent test set. For the URAS dataset, we chose 88 frames as the test set. To prevent any overlap and ensure data integrity, we divided the data into training, validation, and test sets on a patient-wise basis. This means that all video frames from a single patient were kept within the same set, ensuring no patient's video frames appeared in both the training and test sets. This method maintains the independence of the test sets, guaranteeing that our model's performance is evaluated on entirely unseen data and

Table 3 Multi-instrument detection results

Instrument	AP50	n° instances
Monopolar Curved Scissors	0.94	1360
Prograsp Forceps	0.69	989
Bipolar forceps	0.89	1450
Large Needle Driver	0.82	1403
Laparoscopic Grasper	0.45	84
Laparoscopic Clip Applier	0.55	38
Laparoscopic Vacuum	0.73	320
mAP50	0.72	–

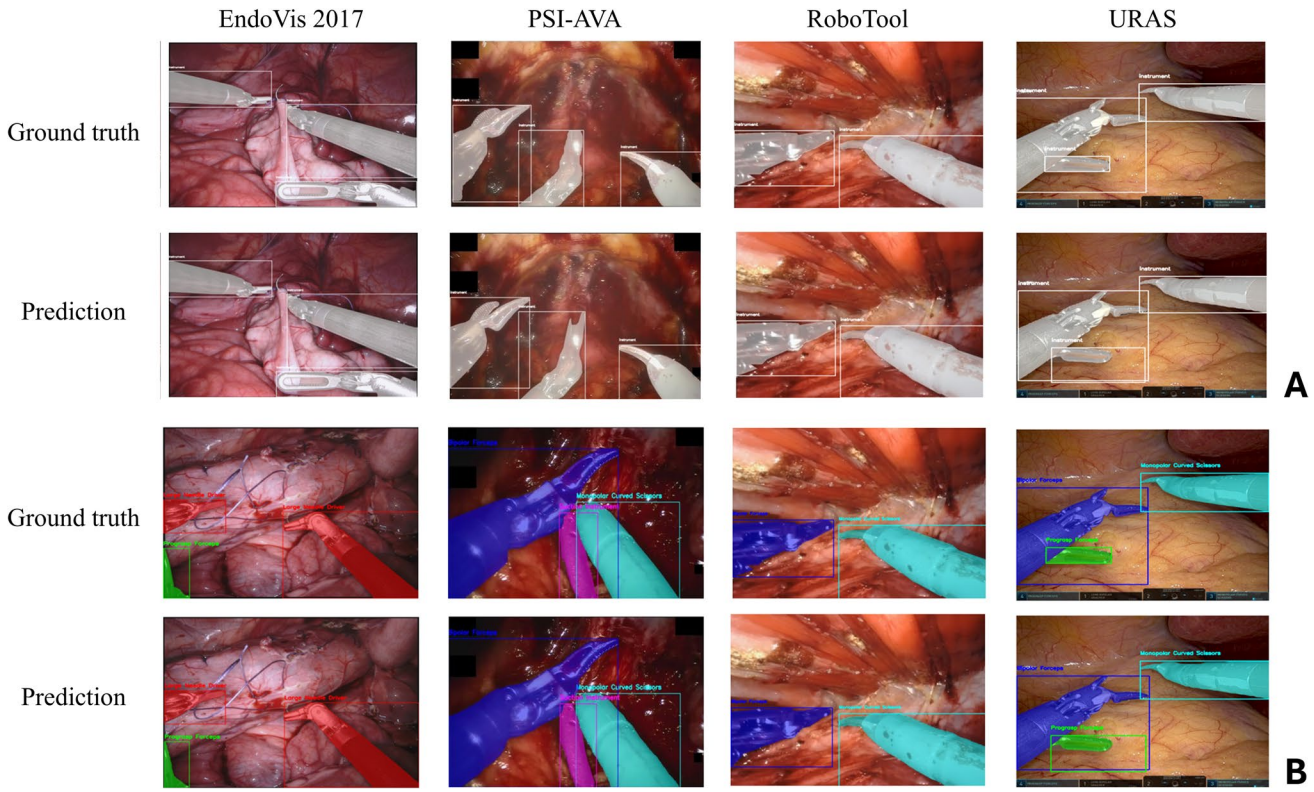

Fig. 1 Binary detection (**A**) and instance segmentation of instruments (**B**)

preventing any data leakage that could artificially inflate performance metrics.

Performance indicators

The quality metrics used to evaluate the performance of the model are described here. For the first and second tasks, we used Average Precision (AP) and the mean AP (mAP) as metrics. The AP and mAP at an Intersection over Union (IoU) threshold of 0.50 (AP50 and mAP50, respectively) were used to evaluate object detection models. A prediction is considered correct if the IoU between the predicted bounding box and the ground truth box exceeds 0.50, indicating at least a 50% overlap. Formally, mAP50 is given by:

$$\text{mAP50} = \frac{1}{N}\sum_{i=1}^{N} AP50_i$$

where N is the number of classes. The AP for each class is determined by:

$$\text{AP} = \int_{r=0}^{1} p(r)dr$$

Where p is the precision and r is the recall defined as:

Table 4 Dice score of instance segmentation

Instrument	Total	EndoVis	PSI-AVA	RoboTool	URAS
Monopolar Curved Scissors	0.93	0.91	0.94	0.91	0.94
Prograsp Forceps	0.86	0.89	0.80	0.84	0.78
Bipolar forceps	0.89	0.91	0.88	0.89	0.91
Large Needle Driver	0.81	0.80	0.83	0.92	–
Lap. Grasper	0.77		0.77	–	0.90
Lap. Clip Applier	0.89		0.89	–	
Lap. Vacuum	0.87		0.87	0.77	0.95
Average	0.86	0.87	0.85	0.87	0.90
Weighted Mean	0.87	0.88	0.87	0.89	0.89

$$p = \frac{TP}{TP + FP}$$

$$r = \frac{TP}{TP + FN}$$

where True Positives (TP) are the predicted bounding boxes that have an IoU ≥0.50 with a ground truth box and are correctly classified. False Negatives (FN) refers to ground truth bounding boxes that do not have any predicted bounding box with an IoU ≥0.50 or are misclassified.

This metric provides a comprehensive measure of a model's detection performance across different classes and varying levels of object overlap, using an IoU threshold to define true positive detections.

For the third task, we used the IoU, also known as the Jaccard Index, and the Dice score (DS). The respective formulas were as follows:

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}$$

$$Dice\ score_i = \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i}$$

TP_i, FP_i, and FN_i are the true positives, false positives, and false negatives for the *i-th* class, respectively. The true positives are foreground pixels of class *i* that are correctly identified, the false negatives are foreground pixels of class *i* that are not correctly identified, and the false positives are background pixels (class different from *i*) that have been annotated as class *i* by the model. The mean IoU and mean Dice score were computed for each dataset as well as for a separate total test set.

For instance segmentation, the performance metrics are calculated only when the corresponding instrument is correctly detected. An average Dice score and Intersection over Union (IoU) are computed across all classes and frames for each dataset, along with a Weighted Mean that accounts for the instrument's test-set instances.

Results

The overall use of YOLOv8 in analyzing surgical instruments intraoperatively can be seen in the multimedia file (Supplementary file 2). The online performance was performed with an inference speed of 1.12 ms/frame for all tasks.

Task 1: binary instrument detection

The binary detection of instruments across the independent test set, using the YOLOv8 model, reached a mAP50 of 0.77. High scores have already been reported previously for binary detection, yet, our dataset includes corrupted frames which might decrease the ability to detect, although it gives more of a "real-life" exposure [27].

Task 2: multi-instrument detection

When classification was incorporated along with detection, the model reached a mAP50 of 0.72 on the totally independent test set. The detailed performance of the model across different types of instruments is presented in Table 3. Notably, the model exhibited high precision in detecting Monopolar Curved Scissors (AP50: 0.94), Large Needle Driver (AP50: 0.82), and bipolar forceps (AP50: 0.89), while the performance was lower for Laparoscopic instruments. One major reason, among others, may be the high number of instances available for these three robotic surgical instruments, exceeding 1300 instances. This comes in line with common surgical procedures as robotic instruments are handled by the primary surgeon while laparoscopic ones are handled by the assistant.

Task 3: binary and instrument segmentation

The binary segmentation of surgical instruments achieved an IoU score of 0.93 and a Dice score of 0.89. This high accuracy

demonstrates the model's proficiency in precisely segmenting surgical instruments within the video frames and offers higher accuracy than current state-of-the-art methods (Fig. 1A).

The model demonstrated surprising multi-instrument segmentation capabilities, achieving an average Dice score of 91% and an average IoU of 86% across all instruments (Table 4 and 5). Notably, monopolar scissors achieved the highest segmentation accuracy with a Dice score of 96% and an IoU of 93%, indicating instrument delineation very close to the ground truth. The lowest performance was attained by the laparoscopic grasper with a Dice score of 77% and an IoU of 85%. Overall, these performances were stable among the different datasets.

Discussion

The analysis of surgical images and videos represents an emerging field with significant potential. Accurate AI recognition of surgical instruments within the operative field serves as a critical foundation for numerous advanced applications. Instrument segmentation, for instance, ensures that the precise location and orientation of each tool are identified relative to key anatomical structures. This real-time monitoring could help prevent surgical instruments from coming dangerously close to vital areas, such as arteries, thereby minimizing the risk of accidental punctures or tears. Moreover, sophisticated detection and segmentation techniques facilitate automatic data collection on instrument use throughout surgical procedures. Such data can be leveraged to optimize surgical methods, enhance training simulations, or develop AI algorithms that provide real-time assistance to surgeons. For example, analyzing patterns in instrument use may reveal opportunities to improve procedural efficiency or pinpoint common sources of errors, contributing to better surgical practices. Beyond these applications, detection and classification pave the way for more autonomous robotic functions. In repetitive tasks like suturing or tissue retraction, AI systems could assist automatically, reducing the surgeon's workload. For example, recognizing a needle holder could trigger the system to perform automated stitching, alleviating surgeon fatigue. Ultimately, these advancements hold the promise of enhancing the safety, speed, and consistency of robot-assisted surgeries, particularly for complex or minimally invasive procedures. In this investigation, the ability of YOLOv8 to detect and classify instruments on our global database was lower than expected, with an mAP50 of 72%. However, these performances can be explained by an evident dependency of the algorithm on the number of annotated images. Performances reached 82% or more when there were more than 1,300 annotations and stayed below 73% when there were fewer than 500. Laparoscopic instruments were largely less annotated than robotic instruments, due

Table 5 IoU results of instance segmentation

Instrument	Total	EndoVis	PSI-AVA	RoboTool	URAS
Monopolar Curved Scissors	0.96	0.95	0.97	0.95	0.97
Prograsp Forceps	0.91	0.93	0.87	0.90	0.84
Bipolar forceps	0.94	0.95	0.93	0.94	0.95
Large Needle Driver	0.87	0.87	0.86	0.96	–
Lap. Grasper	0.85		0.84	–	0.95
Lap. Clip Applier	0.94		0.94	–	
Lap. Vacuum	0.93		0.92	0.87	0.97
Average	0.91	0.92	0.90	0.92	0.94
Weighted Mean	0.92	0.92	0.91	0.94	0.93

to their often-futile presence in robot-assisted surgery. It is therefore reasonable that the algorithm should have difficulty adapting to the heterogeneity of the data. When only laparoscopic instruments are considered, YOLOv8 shows great performances, with a mAP50 exceeding 95.6%, as shown by Le et al. [28]. To alleviate this problem, we could add to our merged database, a database dedicated to laparoscopic instruments such as Endoscape and/or Cholec80 [29, 30]. In addition, our merged database includes several corrupted images, which may also be the cause of training difficulties for the algorithm. Even if the initial intention was to create a set more in line with the reality of our exercise, it would be appropriate to specifically study images with an oversight or detection error, in order to understand the impact of corrupted images on training.

In terms of segmentation, the model demonstrated outstanding multi-instrument segmentation capabilities, achieving an average Dice score of 0.91 and an average IoU of 0.86 across all instruments. Notably, the Monopolar Curved Scissors achieved the highest segmentation performance with a Dice score of 0.96 and an IoU of 0.93, indicating precise boundary delineation. This represents a significant improvement over previous methods, which often reported lower accuracy and struggled with the segmentation of various instrument types. The best current CNN-based and pixel-classification methods have reached performances of 65.18% and 66.3% for instance segmentation on EndoVis 2017, respectively for ISINet and SurgNet algorithms [16, 31]. Transformer-based and mask-classification methods, such as the MATIS algorithm from Ayobi et al, have achieved higher accuracy than the previously cited methods, with an IoU as high as 71.36% [20]. Yet, the highest score in instance segmentation was achieved by SAM with an IoU of 88.2% [21]. Nevertheless, this model needed to be prompted. Recently, Sheng et al proposed Surgical-DeSAM to avoid such constraints. Their algorithm uses DETR to obtain a bounding box, which prompts in turn SAM. They achieved an IoU of 82.41% on EndoVis 2017 [32]. Despite being initially developed for object detection, YOLOv8 proved to be highly efficient for segmentation tasks. This is largely due to its robust architecture, which integrates a feature pyramid network and self-attention mechanisms, enabling it to effectively detect and segment objects of different sizes and scales. In addition, the model can perform efficiently even when multiple instruments are present and overlapping (Fig. 1B). The model showed consistency in performance across the different datasets, despite some having more or less inference factors (Tables 4 and 5).

When our algorithm is tested in real-life conditions on a surgical video sequence, its performance for the various tasks studied became apparent (Video Appendix A-3 and A-4). Despite an execution speed theoretically sufficient for real-time analysis (1.12 ms/frame), we observed a lack of instrument detection. This also leads to a segmentation deficit, since this task depends on prior instrument detection with a bounding box. However, instrument classification and segmentation, when instruments are correctly detected, are visually correct. To optimize detection, beyond the approaches already mentioned, adding a multi-object tracking function to the algorithm could be an effective solution. Multi-object tracking is a computer vision task that involves tracking the movements of several objects over time in a video sequence. The aim is to determine not only the class and location of each object, but also its trajectory throughout the video, including in situations where the objects are partially or totally obscured by other elements in the scene. In general, multiple-object tracking is a two-stage process: object detection and association of detected objects across frames. This type of model can be easily integrated with a detection model such as YOLO. A popular model in this field is ByteTrack, introduced in 2022 [33]. ByteTrack is innovative in its ability to retain low confidence bounding boxes, which are usually discarded after initial detection filtering, for use in a second association step. Occulted detection boxes often have confidence scores below threshold, but still contain relevant object information, distinguishing them from pure background detections. By preserving them for association, ByteTrack improves tracking robustness. A very recent paper by Myo *et al.* tested YOLOv8 + ByteTrack under surgical conditions [34]. Testing the algorithm on the ROBUST-MIS database of the EndoVis 2019 challenge, they obtained better results, with a real-time segmentation speed of around 45 fps, sufficient for a real-time application. By categorizing images according to the positions of separate, crossed or overlapping instruments, they also demonstrated that ByteTrack was able to improve segmentation performance in all categories.

Overall, our results may have different clinical impact. First, YOLOv8 allows full instrument segmentation, which will be necessary to avoid overlay of AR images in the surgical fields. Second, this quick processing time is crucial for potential clinical real-time applications, ensuring timely decision-making and effective use in tasks for surgical guidance and real-time diagnostics. Last but not least, it proved its effectiveness in classifying and segmenting despite different types of instruments and environments. As minimally invasive surgery can be laparoscopic or robotic, having a generalized algorithm capable of performing on both types of instruments could be of interest, and comes in line with the current evolution toward computer-aided surgery [35, 36].

The current analyzed model reveals several limitations that impact its effectiveness in practical surgical scenarios. First, if an instrument is absent from the model's predefined toolset (i.e., not annotated in the GT reference frame), it risks misclassifying it as another instrument. For instance,

in Supplementary File 3a, non-fenestrated bipolar forceps were not included in our model, yet the model still classified them as bipolar forceps. Also, when the background does not consist of anatomic structures, such as the endobag, it leads to poor segmentation results. Second, when the number of instances of certain instruments is very low, such as with the clip applier and laparoscopic grasper (Table 1), the model struggles either to classify them incorrectly (Supplementary File 3c) or misses their detection entirely (Supplementary File 3b). This limitation underscores the model's dependency on adequate training data representation. Moreover, instruments that are too small within the image frame pose a challenge for accurate classification, often leading to misclassifications as other tools (Supplementary File 3d). In addition, images with significant blur (Supplementary File 3e) present another obstacle, as the model may not detect the instrument. These limitations underscore the need for further refinement and robustness in the model's training and inference processes. Addressing these challenges could significantly enhance its reliability and applicability in real-world surgical environments, ensuring more accurate instrument recognition and segmentation.

Conclusion

The YOLOv8 model exhibited strong performance in detecting and segmenting both robotic and laparoscopic surgical instruments, even in complex and varied surgical settings. Its capability to maintain high accuracy, especially in binary segmentation, and its adaptability across diverse datasets highlight the model's robustness. However, challenges remain, including the misclassification of instruments not included in the training data and difficulties with low-quality images, indicating the need for further enhancements. Rather than focusing solely on improving precision, this study highlights the model's potential to optimize surgical assistance by streamlining workflow and supporting surgeons more effectively during procedures.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11701-025-02284-7.

Acknowledgments None.

Author contributions S.F., F.F., W.W., E.S., H.D., P.B.R., F.B. and N.A. conceptualized the project. S.F., F.F., W.W., H.D., P.B.R., F.B. and N.A. designed the methodology. S.F., F.F., E.S.E., W.W., H.D., P.B.R., F.B. and N.A. performed the formal analysis and investigation. S.F., F.F. and W.W. wrote the main manuscript. S.F., F.F., W.W., E.S., H.D., P.B.R., F.B. and N.A. reviewed and edited the manuscript. S.F. and F.F. collected the resources. P.B., M.D., H.D., P.B.R., F.B. and N.A. supervised the project.

Funding This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA0002. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur. Agence Nationale de la Recherche, ANR-19-P3IA0002.

Data availability The URAS dataset remains private until further notice. No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Hospital of Nice.

Consent to participation Informed consent was obtained from all individual participants included in the study.

References

1. Fairag M, Almahdi RH, Siddiqi AA, Alharthi FK, Alqurashi BS, Alzahrani NG, Alsulami A, Alshehri R (2024) Robotic revolution in surgery: diverse applications across specialties and future prospects review article. Cureus 16(1):e52148. https://doi.org/10.7759/cureus.52148
2. Sejor E, Berthet-Rayne P, Frey S (2022) Calling on the next generation of surgeons. Surg Innov 30:15533506221124500. https://doi.org/10.1177/15533506221124501
3. Rodler S, Ganjavi C, De Backer P, Magoulianitis V, Ramacciotti LS, De Castro Abreu AL, Gill IS, Cacciamani GE (2024) Generative artificial intelligence in surgery. Surgery 175(6):1496–1502. https://doi.org/10.1016/j.surg.2024.02.019
4. Gumbs AA, Grasso V, Bourdel N, Croner R, Spolverato G, Frigerio I, Illanes A, Abu Hilal M, Park A, Elyan E (2022) The advances in computer vision that are enabling more autonomous actions in surgery: a systematic review of the literature. Sensors (Basel) 22(13):4918. https://doi.org/10.3390/s22134918
5. Mascagni P, Vardazaryan A, Alapatt D, Urade T, Emre T, Fiorillo C, Pessaux P, Mutter D, Marescaux J, Costamagna G, Dallemagne B, Padoy N (2022) Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Ann Surg 275(5):955–961. https://doi.org/10.1097/SLA.0000000000004351
6. Choksi S, Szot S, Zang C, Yarali K, Cao Y, Ahmad F, Xiang Z, Bitner DP, Kostic Z, Filicori F (2023) Bringing artificial Intelligence to the operating room: edge computing for real-time surgical phase recognition. Surg Endosc 37(11):8778–8784. https://doi.org/10.1007/s00464-023-10322-4
7. den Boer RB, Jaspers TJM, de Jongh C, Pluim JPW, van der Sommen F, Boers T, van Hillegersberg R, Van Eijnatten MAJM, Ruurda JP (2023) Deep learning-based recognition of key anatomical structures during robot-assisted minimally invasive esophagectomy. Surg Endosc 37(7):5164–5175. https://doi.org/10.1007/s00464-023-09990-z
8. Hofman J, De Backer P, Manghi I, Simoens J, De Groote R, Van Den Bossche H, D'Hondt M, Oosterlinck T, Lippens J, Van Praet C, Ferraguti F, Debbaut C, Li Z, Kutter O, Mottrie A, Decaestecker K (2023) First-in-human real-time AI-assisted instrument deocclusion during augmented reality robotic surgery. Healthc Technol Lett 11(2–3):33–39. https://doi.org/10.1049/htl2.12056

9. Sorriento A, Porfido MB, Mazzoleni S, Calvosa G, Tenucci M, Ciuti G, Dario P (2020) Optical and electromagnetic tracking systems for biomedical applications: a critical review on potentialities and limitations. IEEE Rev Biomed Eng 13:212–232. https://doi.org/10.1109/RBME.2019.2939091
10. Zia A, Bhattacharyya K, Liu X, Berniker M, Wang Z, Nespolo R, Kondo S, Kasai S, Hirasawa K, Liu B, Austin D 2023. Surgical tool classification and localization: results and methods from the MICCAI 2022 SurgToolLoc challenge. arXiv preprint arXiv:2305.07152.
11. Yu L, Wang P, Yu X, Yan Y, Xia Y (2020) A holistically-nested U-net: surgical instrument segmentation based on convolutional neural network. J Digit Imaging 33(2):341–347. https://doi.org/10.1007/s10278-019-00277-1
12. Lo BPL, Darzi A, Yang G-Z (2003) Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. Med Image Comput Comput-Assist Interv - MICCAI 2003(2878):230–237
13. Bouget D, Benenson R, Omran M, Riffaud L, Schiele B, Jannin P (2015) Detecting surgical tools by modelling local appearance and global shape. TMI 34:2603–2617
14. Allan M, Shvets A, Kurmann T, Zhang Z, Duggal R, Su YH, Rieke N, Laina I, Kalavakonda N, Bodenstedt S, Herrera L (2019). 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426.
15. Sestini L, Rosa B, De Momi E, Ferrigno G, Padoy N (2021) A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. IEEE Robot Autom Lett 6(2):2938–2945
16. Ni ZL, Zhou XH, Wang GA, Yue WQ, Li Z, Bian GB, Hou ZG (2022) SurgiNet: pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. Med Image Anal 76:102310. https://doi.org/10.1016/j.media.2021.102310
17. Liu J, Guo X, Yuan Y (2022) Graph-based surgical instrument adaptive segmentation via domain-common knowledge. IEEE Trans Med Imaging 41(3):715–726. https://doi.org/10.1109/TMI.2021.3121138
18. Liu J, Guo X, Yuan Y (2021) Prototypical interaction graph for unsupervised domain adaptation in surgical instrument segmentation. In: de Bruijne M et al (eds) Medical image computing and computer assisted intervention – MICCAI 2021. MICCAI 2021. lecture notes in computer science, vol 12903. Springer, Cham. https://doi.org/10.1007/978-3-030-87199-4_26
19. Wei M, Budd C, Garcia-Peraza-Herrera LC, Dorent R, Shi M, Vercauteren T (2023). SegMatch: A semi-supervised learning method for surgical instrument segmentation. arXiv preprint arXiv:2308.05232.
20. Ayobi N, Pérez-Rondón A, Rodríguez S, Arbeláez P (2023). Matis: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.
21. Wang A, Islam M, Xu M, Zhang Y, Ren H (2023). Sam meets robotic surgery: an empirical study on generalization, robustness and adaptation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland (pp. 234-244).
22. Valderrama N et al (2022) Towards Holistic Surgical Scene Understanding. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S (eds) Medical image computing and computer assisted intervention– MICCAI 2022. MICCAI 2022. lecture notes in computer science, vol 13437. Springer, Cham. https://doi.org/10.1007/978-3-031-16449-1_42
23. Garcia-Peraza-Herrera LC, Fidon L, D'Ettorre C, Stoyanov D, Vercauteren T, Ourselin S (2021) Image compositing for segmentation of surgical tools without manual annotations. IEEE Trans Med Imaging 40(5):1450–1460
24. Ultralytics, YOLOv8 Mosaic, https://yolov8.org/yolov8-mosaic
25. Glenn Jocher, Ayush Chaurasia and Jing Qiu. Ultralytics YOLOv8, version 8.0.0. 2023. https://github.com/ultralytics/ultralytics; Accessed: 2024-June-18.
26. Schulze A, Tran D, Daum MTJ et al (2023) Ensuring privacy protection in the era of big laparoscopic video data: development and validation of an inside outside discrimination algorithm (IODA). Surg Endosc 37(8):6153–6162. https://doi.org/10.1007/s00464-023-10078-x
27. De Backer P, Van Praet C, Simoens J, Peraire Lores M, Creemers H, Mestdagh K, Allaeys C, Vermijs S, Piazza P, Mottaran A, Bravi CA, Paciotti M, Sarchi L, Farinha R, Puliatti S, Cisternino F, Ferraguti F, Debbaut C, De Naeyer G, Decaestecker K, Mottrie A (2023) Improving augmented reality through deep learning: real-time instrument delineation in robotic renal surgery. Eur Urol 84(1):86–91. https://doi.org/10.1016/j.eururo.2023.02.024
28. Le HB, Kim T, Ha M, Tran A, Nguyen DT, Dinh XM (2023) Robust Surgical Tool Detection in Laparoscopic Surgery using YOLOv8 Model. Proc ICSSE. https://doi.org/10.1109/ICSSE58758.2023.10227217
29. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36(1):86–97. https://doi.org/10.1109/TMI.2016.2593957
30. Murali A, Alapatt D, Mascagni P, et al (2024). The Endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: official splits and benchmark. Accessed August 29, 2024. http://arxiv.org/abs/2312.12429
31. González C, Bravo-Sánchez L, Arbelaez P (2020) Isinet: an instance-based approach for surgical instrument segmentation. International conference on medical image computing and computer-assisted intervention. Springer International Publishing, Cham, pp 595–605
32. Sheng Y, Bano S, Clarkson MJ et al (2024) Surgical-DeSAM: decoupling SAM for instrument segmentation in robotic surgery. Int J CARS. https://doi.org/10.1007/s11548-024-03163-6
33. Zhang Y, Sun P, Jiang Y, et al (2022). Bytetrack: multi-object tracking by associating every detection box. Accessed August 29, 2024. http://arxiv.org/abs/2110.06864
34. Myo N, Boonkong A, Khampitak K, Hormdee A (2024) Real-time surgical instrument segmentation analysis using YOLOv8 with bytetrack for laparoscopic surgery. IEEE. https://doi.org/10.1109/ACCESS.2024.3412780
35. Makary J, van Diepen DC, Arianayagam R, McClintock G, Fallot J, Leslie S, Thanigasalam R (2022) The evolution of image guidance in robotic-assisted laparoscopic prostatectomy (RALP): a glimpse into the future. J Robot Surg 16(4):765–774. https://doi.org/10.1007/s11701-021-01305-5
36. Schmidt A, Mohareri O, DiMaio S, Yip MC, Salcudean SE (2024) Tracking and mapping in medical computer vision: a review. Med Image Anal 94:103131. https://doi.org/10.1016/j.media.2024.103131