

Surgical Instrument Tracking: a Lightweight YOLO Approach Across Diverse Domains

Nathalia Frederico[1(✉)], Rafael Miranda[1], Saulo Silva[2], and Jefferson Morais[1]

[1] Federal University of Pará (UFPA), Belém, Brazil
almeidadifre@gmail.com, rafaeldoriamm@gmail.com, jmorais@ufpa.br
[2] Instituto Federal de Goiás (IFG), Luziânia, Brazil
saulo.silva@ifg.edu.br

Abstract. Artificial intelligence has the potential to improve robotic and laparoscopic surgery through accurate surgical instrument tracking. However, widespread clinical implementation is currently hindered by complex visual interferences, such as bleeding and smoke, and the prohibitive computational costs of state-of-the-art models. Therefore, this paper presents a framework to unify three public surgical datasets (Cholec, Robust, and Badilla) by standardizing annotations (bounding boxes, keypoints, and semantic masks) to the YOLO format. Using the unified dataset, a lightweight YOLOv11-s architecture was trained and evaluated. This integration aims to enhance the generalization of the model in various surgical procedures and clinical domains, mitigating the limitations of single-dataset training. The model achieved a mean Average Precision (mAP@0.5) of 90.2%, and the inference speed was 5.1 ms per frame. The system demonstrated resilience to severe class imbalances that are natural in surgical scenarios, maintaining high accuracy even for underrepresented instruments. The experimental analysis indicates computational viability in low-resource environments, as validated by retrospective data. Further research can employ data augmentation in minority classes to achieve robustness on a larger dataset and deployment in real clinical scenes, focusing on interoperability under restricted operational settings.

Keywords: Surgical dataset integration · Yolov11 · Low-resource tracking

1 Introduction

Robotic surgery is a constantly growing trend in medicine. Encouraged by its potential to improve surgical results in terms of predictability, decision-making, precision, safety, and time optimization [14].

Solutions based on artificial intelligence (AI), which are still in the early stages of clinical integration, have a promising future in robotic surgery. However, concerns about reliability exist among the medical community [51].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 461–477, 2027.
https://doi.org/10.1007/978-3-032-30491-9_29

These concerns converge on one of the most demanding areas of surgical practice: the precise tracking of surgical instruments, given the dynamic and complex context that involves various types of instruments and the challenging visual conditions involved in their handling, such as bleeding, smoke, and variable lighting [44].

Patient interest in robotic surgery continues to increase, resulting in better results and a shorter hospital stay [46]. To achieve an interface between surgeons and robots, the current paradigm suggests the need for integrating AI and its new ways of interpreting modern surgery [41].

The history of robotic surgery dates back to the 1980s, with various models designed for this purpose, from the first, the Programmable Universal Machine for Assembly (PUMA), to the most established and widely used model worldwide, the Da Vinci Surgical System [28]. The implementation of robotic systems in healthcare is characterized by exponential growth [31].

Currently, AI is transforming robotic surgery by integrating machine learning (ML) and computer vision (CV) algorithms to refine accuracy, assist in preoperative planning, and, in the future, expand the level of medical and care autonomy in surgical procedures [1].

Controlled studies have demonstrated that the exhaustive training of surgeons through extensive simulator training shortens the learning curve for skill and precision, potentially reducing the frequency of accidents and complications [12]. The ethical priority in the surgical context should be considered: "never the first time on the patient" [7].

In this context, the challenge itself lies in the transition from computer lab models to simulation centers, considering that expensive computational costs and a lack of metrics for educational utility are the main barriers to implementation in educational institutions.

To address these challenges, this study evaluates the YOLOv11-s architecture to optimize the trade-off between latency and precision. By unifying specialized public datasets, we propose a robust methodology that covers real laparoscopic, robotic, and simulated scenarios.

The remainder of this paper is organized as follows: Sect. 2 reviews related works; Sect. 3 details the experimental methodology; Sect. 4 discusses the results; Sect. 5 addresses limitations; Sect. 6 concludes the paper; Sect. 7 presents an ethics statement; and finally, Sect. 8 displays the availability of data and code.

2 Related Works

2.1 Systematic Search

The exponential growth of the field motivated a literature review on the recent period (2021–2025). The strings that involved four dimensions (application, problem, technique and resource) were subsequently combined: ("Surgical Instrument*" OR "Surgical tool*" OR "Surgical Simulation*") AND ("Clutter*" OR "Bin-picking" OR "Occlusion*") AND ("Machine Learning" OR "Deep Learning") AND ("Dataset*" OR "Data collection" OR "Training set*").

Focusing on ML and CV models applied to healthcare issues, the literature searches deploy important databases in the fields of Engineering and Computing, as follows: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Web of Science, PubMed, and Google Scholar.

PRISMA (Preferred Reporting items for Systematic Reviews and Meta-Ana- lyses) guidelines constitute a standard protocol for identifying and screening studies [40]. Its selection justified the goal of presenting transparency and methodological rigor in the process of searching the bibliographic bases of this study.

Through the steps of Identify, Screen, Elect, and Synthesize, the protocol allows comparison with the following previously established criteria. Four inclusion criteria were used: published in English; recent (2021–2025); open access full-text; and consistent with the objectives of this work. Five exclusion criteria were used: secondary literature, such as literature reviews, systematic reviews, meta-analyses, editorials, or surveys; incomplete access; duplicates; non related studies; and gray literature.

From an initial pool of 219 screened records, 48 articles were ultimately included for qualitative synthesis after a strict application of predefined inclusion and exclusion criteria.

After the systematic screening of 219 studies, 48 were ultimately included after their eligibility was assessed on the basis of predefined inclusion and exclusion criteria. The following section begins the analysis of state-of-the-art (SOTA) methods.

2.2 SOTA Models for Surgical Instrument Tracking

Surgical instrument tracking has evolved from feature-based manual methods to deep learning architectures capable of navigating the complexities of the intraoperative environment. The following subsections provide a comparative synthesis of the most relevant studies from the past five years to provide a thorough overview of the current state of the art.

The technical literature was organized along three main lines: the Task performed (from basic detection to Surgical Workflow Recognition), Data Availability, and Model Architecture.

The Education purpose topic also looked at studies that were specifically designed for medical training, skill testing, or feedback from a simulator. We chose to contextualize different approaches, ranging from real-time assisted methodologies to designs with purely pedagogical purposes.

Considering the following performance metrics grouped by methodological function, studies were compared using this general framework:

- **Classification:** Accuracy (Acc), Precision (Prec), Recall (Rec), and F1-Score are calculated to measure the level of ability to find patterns (categories or identities).
- **Overlap:** Sørensen–Dice Coefficient and Intersection over Union (IoU) are often used to point the match of segmentation boundaries.

- **Error:** Statistical Errors demonstrate changes in position or orientation of the objects by the Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Pose Error.
- **Flow:** Detection and tracking tasks consist on verify spatial localization and temporal coherence, it may uses the mean Average Precision (mAP), Higher Order Tracking Accuracy (HOTA), and Multi-Object Tracking Accuracy (MOTA).
- **Speed:** At last, to indicate computational efficiency, its imperative the urge to the system perform in a short time given a clinical situation, therefore, we use the Frames Per Second (FPS) rate for this purpose.

One-stage detectors are becoming popular on real-time detection scene, as simulator's feedback requires low latency, these architectures in turn ensure a single end-to-end pipeline, by object localization and classification into a single pass through the network.

In this scenario, the YOLO (You Only Look Once) family has become the most prevalent technique [8,9,11,16,24,30,42–45,49,56,59,60]. Recent studies have seek for balancing elevated inference speed (FPS) and localization precision (Flow). Additionally, innovations such as Omni-Dimensional Dynamic Convolutions (ODConv) and Large Receptive Field blocks (RepLK), were design to capture both fine details and larger structures [49].

For tasks requiring precise pixel-level delimitation, architectures based on U-Net [3,23,25,52,58,60] and DeepLabV3+ [13,23] remain the primary benchmarks. The integration of robust backbones, such as ResNet-50/101 [2,5,13, 17,19,22,32,47,48,59] and ConvNeXt [20,59], aims to extract features that are resilient to visual interferences.

Specific mechanisms like SEBlock (Squeeze-and-Excitation) and CBAM (Convolutional Block Attention Module) were helpful to ignore background noise and to carefully observe reflective properties of metal surgical tools [2,13].

In order to handle natural switch between similar tools, it is important to keep identities consistent in video sequences. Recurrent neural networks as Long Short-Term Memory (LSTM) and its variants enables the understanding of temporal surgical workflow [10,15,29,50,54]. ByteTrack and BoT-SORT work as association algorithms for strong tracking [42,44].

A recent study calls attention to the need of datasets that shows clinically pertinent trajectories for training models, especially focusing to tracking continuity on instruments frequently switching [44].

The moment trend is the Foundation Models, like Segment Anything Model (SAM) [36,47,57], that act on zero-shot segmentation, in other words, presents an accurate segment of objects or classes unknown during training, relying on general visual features or external tips. A lightweight version of that architecture is the MobileSAM, that is making an effort to real-time apps on less processing settings [57].

On the other hand, there is also a combination of hybrid techniques (Transformers like Swin-UNet with Convolutional Neural Networks) to balance global attention with local efficiency [20,36,56,59]. Presenting a good performance on

hard tasks such 3D pose estimation, which requires spatial context and specific fine details.

2.3 Research Gap and Challenges

The SOTA analysis established that significant advances in raw accuracy occurred, however, there are still challenges to address, such as: High Computational Dependency, current SOTA models, especially Foundation Models, often need high power Graphic Process Units (GPUs) [35], making its widespread use in educational institutions unfeasible. Requiring architectures that provides energy efficiency and low latency on modest hardware.

As educational concerns stands out: Generalization power (Models simulated datasets struggle to generalize on real-world domain gap); and Pedagogical metrics (few studies integrate tracking with motion economy (Path Length) or smoothness (Jerk) [10,38]. Such gaps limits the capacity to provide actionable educational feedback, since there is a disconnection between CV metrics (mAP, Dice) and clinical proficiency indicators.

Due to these questions, the experimental method of this research shift the paradigm of "maximum precision at any cost" to "resilient efficiency".

3 Experimental Design, Materials and Methods

The methodological framework employed was made under the Cross-Industry Standard Process for Data Mining (CRISP-DM), an worldwide accepted guide for methodical and reproducible researches about machine learning models [21]. That chosen was driven to break the usual 'black-box' algorithms and so facilitate posterior audit and validation, considering the principles of Explainable Artificial Intelligence (XAI) [4], and always subordinate to clinical reason [37].

An adaptation of the CRISP-DM protocol follows these steps checklist: Problem in Business Terms; Data Understanding and Preparation; Model and Evalute; Deployment [18]. These steps were adapted to the surgical context discussed, so that the development of an object detection solution with few computational resources will facilitate future adoption in educational environments, consequently improving the surgical learning curve and, in turn, making robot-assisted procedures safer.

The experimental methodology includes two principle targets: Low-resource Optimization (by apllying Nano and Small versions of one-stage detector YOLO-v11 and examine its performance under hardware limitations, aiming to found settings of an optimal balance between FPS and efficient classification); and Unify Diverse Domains (considering the benchmarks identified in the Related Works section, a data-centric pipeline evaluate the model robustness across a range of scenarios).

The experiments were held on Google Colab, linked to Google Drive as a environment to manage data. The computer system available contains NVIDIA

T4 GPU with 16 GB of VRAM, capable of ran deep learning frameworks based on Python.

While the NVIDIA T4 is a high-performance accelerator utilized here for efficient training and validation, the selected YOLOv11-s architecture is inherently lightweight. This choice aims to balance high-speed inference with low computational overhead, making it significantly more efficient for real-time surgical applications compared to larger, more resource-intensive models.

All workflow and supplementary materials were organized to ensure methodological transparency and enable complete reproducibility [after the pair revision is finished, to avoid author identification, a link to the repository will be shared]. Primary datasets used in the development of this work are properly cited and referenced according to their original license agreement.

3.1 Business and Healthcare Understanding

Initial stage is to translate clinical needs into business terms. Track instruments is not an end in itself, it is a way to enable Computer-Assisted Interventions (CAI), including skill assessment, safety zone estimation, and create an efficient interface people-machine [44]. As to turn raw video from simulators into useful teaching feedback, to mitigate problems with cost and technical difficulty.

Accurate tracking grant an effective automated workflow identification [53]. Allows the system to discern between actions like suturing and dissection. In robotic surgery, instruments are categorized by their technical function, allying their visual design, examples include Graspers (change shape when pulled back [62]), Dissectors (like Hooks, that often produce surgical smoke [43]), and Needle Holders (which requires millimeter accuracy for Jerk and Path Length metrics [38]).

There is also difficult optical properties, stainless steel's specular reflections, tissues or fluids that frequently are blocking the view [55,58]. Finally, shift from laboratory models to real-time application call for minimal latency, as also to prevent the disruption in educational process [42].

On this occasion, "business terms" means a system that is technically accurate, useful for teaching, and financially possible.

3.2 Data Understanding

Google Colab was used to technical audit the data repositories that were found in the literature. This stage aimed to select benchmarks datasets that would help built a robust model capable of deal with a range of scenarios.

In Table 1 we present the study carried out in five public repositories, in order to posteriously choose for this work. These datasets demonstrated different surgical approaches, from controlled environments for multi-class complex clinical scenarios.

OSI26 is a dataset designed to be a balanced baseline, with 66 examples for each class. This consistency allows the model to learn rare instrument features

Table 1. Technical Review and Description of Surgical Datasets

Dataset	Vol.	Cls	Inst.	Type	Src	Res.	Challenges	Bal.
OSI26 [49]	0.23	26	1,716	Image	Sim.	1080p	Low: Occ., reflex	Uni.
CholecTrack20 [44]	32.9	7	65,247	Video	Real	480p	High: Blood, Smoke	Unb.
ROBUST-MIPS [22]	3.29	10	11,560	Seq.	Real	540p	High: Sev. Occ.	Unb.
m2cai16 [26]	0.13	7	3,929	Image	Real	334p	Mod: Blur, Overlap	Unb.
Badilla-S. [5]	0.98	18	1,019	Mask	R/S	480p	High: Intra-op noise	Unb.

Abbreviations: Vol: Volume (GB); **Cls**: Classes; **Inst**: Instances; **Src**: Source; **Res**: Resolution; **Occ**: Occlusion; **Bal**: Balance (Uni: Uniform/Unb: Unbalanced); **Sim.**: Simulated; **R/S**: Real/Synthetic.

giving the same importance as common ones. Originally, it had 1,716 instances, but after the authors add more data, the augmentation technique was to better separate overlapping tools in simulated environments [49].

The next evaluated was CholecTrack20, in contrast with the previous dataset, this presented a high class imbalance, natural in real world surgical routine. On this particular set, "Grasper" was the prevalent identity (52.62%), following by "Hook" (32.61%), "Bipolar" (3.61%) and "Specimen Bag" (5.37%), "Scissors", "Clipper" and "Irrigator" got less than 2%, along with visual challenges (such blur, blood, smoke, occlusion) [44].

ROBUST-MIPS focusing on pose estimation in complex structures, not just classify identities but also make effort to understand the global scene around among partial occlusions [22]. Unlike the other datasets, ROBUST-MIPS organizes annotations by surgical procedure type (Proctocolectomy, Rectal Resection, and Sigmoid), with instrument identities encoded at the keypoint level. This required an additional extraction step during Data Preparation to map skeletal nodes to the Universal Taxonomy.

Finally, the m2cai16-tool-locations provides a historical baseline for real-world bounding boxes, while the dataset by Badilla-Solórzano et al. utilizes Mask-Based Object Insertion (MBOI) to create ∼27,000 synthetic images, focusing on edge precision and semantic segmentation of specific robotic components [5,26].

3.3 Data Preparation

This stage focuses on the structural unification of the CholecTrack20 [44], ROBU- ST-MIPS [22], and Badilla-Solórzano et al. [5] datasets. To mitigate cloud-storage latency and Input/Output (I/O) overhead associated with large-scale surgical data, we implemented a partitioned storage architecture.

By segmenting the Unified Dataset into sub-directories (keeping each folder below the 10,000-item synchronization limit of Google Drive), we ensured high-throughput data streaming and avoided the common timeout errors of distributed file systems.

A core contribution of this preparation is the adoption of a Universal Taxonomy of 7 classes: Grasper, Hook, Scissors, Clipper, Irrigator, Bipolar, and Specimen Bag. This taxonomy considered the recent pioneer dataset for multi-perspective tracking by Nwoye et al. [44], following this protocol, additional categories from ROBUST-MIPS and Badilla-Solórzano datasets were semantically mapped and filtered to align with the selected Ground Truth.

That preparation was necessary to ensure that robotic, laparoscopic and simulated environments could all functioning together. First, each dataset had a type of annotation: sequential JSON (CholecTrack20), instance IDs (ROBUST-MIPS), and binary masks (Badilla-Solórzano). Making necessary a geometry extraction algorithm to change vector and matrix data into the YOLO PyTorch format (.txt), converging coordinates to normalized XYWH values based on the size of the frame [27].

Also, all the frames had to be resized to the standard 640×640 resolution, conforming YOLOv11 architecture demand [42]. As a low-resource environment, using less VRAM, that configuration preserve the tool-tip details.

Before the ultimate training with full data, we run a pilot phase with a stratified subset of 361 images (splitting in 70/20/10 for training, validation and testing, respectively) to validate hyperparameter configurations and taxonomic correct correspondence.

3.4 Modeling and Evaluation

The data-centric pipeline followed the modules of: Data Understanding (script made to study the architectural profile, verify spatial integrity and class imbalances of pre-selected datasets); Data Preparation (this module worked on math normalization, conversion of the different annotations into the standard YOLO format); Preliminary Analysis (the pilot phase of validation using a sample of the unified dataset); Model Training (full dataset split in 80% training and 20% validation to run the neural network training loop); Finally, the module Zero-Shot functioning as a test to validate the generalization of the model in an unannotated surgical footage.

To the pilot run, YOLOv11-n (Nano) architecture was execute, as the most recent version of a lightweight approach, with 2.59 million parameters and a lightweight computational footprint of 6.4 GFLOPs [27,61]. This first execution run for 10 epochs, consolidates the 7-class taxonomic and the model performance on a small scale, validating the pre-processing made.

As the pilot was completed, the ultimate detection task scaled to YOLOv11-s (Small) architecture, to better balance between mAp and Speed, essential in surgical procedures. The standard pre-trained weights of MS COCO dataset were used by Transfer Learning, that process allows fine-tuning focus on the specific surgical domain of the present research, considering that the weights ensure model already learned patterns (edges and textures) from millions of natural images before [33].

The ‘s’ (small) variant of the YOLOv11 family was specifically selected to ensure a lightweight framework. This version provides a strategic optimization,

delivering the precision required for delicate surgical tracking while maintaining a small parameter footprint and high frames-per-second (FPS) rates.

Following, YOLOv11 architecture have itself an automated augmentation pipeline, that improve the resistance of the system among expected visual challenges, by making a Mosaic Augmentation, where four aleatory training images are set together, as to avoid the model memorizing scenarios and focusing on the tools in different scenarios and spatial localizations, as so looking closer to their texture and shape by HSV color jittering [6].

At this point, the final training ran over 50 epochs with a batch size of 16, AdamW algorithm contributed to weight decay regularization, that mechanism optimize the learning rate by minimize losses on localization and classification [34], this stage started in a learning rate of 0.000909 and a momentum of 0.9.

According to the protocol for surgical detection settings evaluation [44], we employed the following multi-metric check to evaluate spatial accuracy and computational efficiency: mAP (mAP50 tested the capacity to identify tools that are 50% overlapping each other and mAP50-95 gave a complete evaluation across IoU thresholds from 50% to 95%, showing the level of precision for tool-tip track [33]; Prec and Rec (trade-off between false positives and false negatives, crucial for clinical safety [37]); and F1-score (confirm if the model is reliable across the 7-class taxonomy by calculate the harmonic mean between Prec and Rec).

A common issue in the process of recognize surgical instruments is the similarity inherent of them [62], for this matter, the Confusion Matrix checks the possible mistakes (like identify a "Irrigator" when it is a "Scissor").

Its also important to evaluate the practicability of low-resource solutions, outlined by [42]: Inference Speed (ms/frame) for measure the real-time feedback without any lag, and Computational Complexity (GFLOPs) for acknowledgment of computing environments required.

This work thoroughly addresses the modeling and evaluation stages of the computational pipeline; however, it is a fundamentally experimental analysis validated through retrospective datasets. The last step in the method, Deployment, demands further experimentation.

To turn this architecture from a computational proof-of-concept into a useful clinical tool, the trained YOLOv11-s model will need to be added directly to the video processing pipelines in operating rooms. Focusing on testing the model during live and planned intraoperative procedures. This will ensure that it works well with existing robotic and laparoscopic hardware.

4 Results and Discussion

Image pre-processing was done on a localized Solid State Drive (SSD) runtime to make sure that the computer worked quickly and that the data were accurate. This method got around the built-in latency and metadata synchronization problems of distributed cloud storage, which made it possible to resize and verify images quickly before putting them all together into a final dataset.

The unified YOLOv11-s model was able to detect things in the very diverse surgical dataset. The Precision-Recall (PR) Curve (Fig. 1) shows that the network got a mean Average Precision (mAP@0.5) of 0.902 for all instrument classes. The model also got the global F1-confidence score of 0.87 at a confidence threshold of 0.480. This means that it had a strong balance between precision (keeping false positives to a minimum) and recall (keeping false negatives to a minimum).

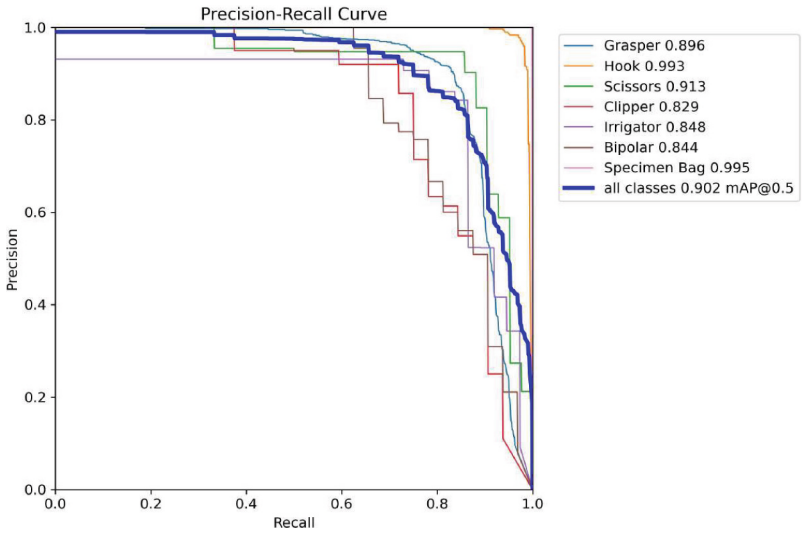

Fig. 1. Precision-Recall (PR) curve for the unified taxonomy (mAP@0.5: 90.1%). Detailed metrics in Table 2.

The model demonstrated significant resilience to the severe class imbalances inherent in the unified dataset. As detailed in Table 2, underrepresented classes such as 'Scissors', 'Hook', and 'Specimen Bag' achieved excellent mAP@0.5 scores of 91.2%, 99.3%, and 99.5%, respectively. This performance is visually corroborated by the PR curve, which maintains high stability and a large Area Under the Curve (AUC) across individual classes despite the extreme variation in sample representation.

We also qualitatively present the model's performance, as to enrich quantitative previous analysis, in Fig. 2, where its showed inference results of predicted bounding boxes on top of the real frames of the unified dataset used.

The network was able to frame several tools at once (like Grasper and Hook interactions) with tight bounding-box regression. The model consistently exhibited elevated classification confidence, irrespective of the background environment, effectively abstracting the mechanical essence of the instruments across diverse laparoscopic and robotic domains, undeterred by surgical smoke, tissue occlusion, or reflective glare.

Table 2. Per-class performance metrics for the unified surgical tracking model.

Class	Precision (P)	Recall (R)	mAP@0.5
Grasper	0.913	0.826	89.5%
Hook	0.976	0.974	99.3%
Scissors	0.946	0.838	91.2%
Clipper	0.846	0.750	82.5%
Irrigator	0.810	0.865	84.7%
Bipolar	0.843	0.688	84.3%
Specimen Bag	0.947	1.000	99.5%
All Classes	**0.897**	**0.849**	**90.1%**

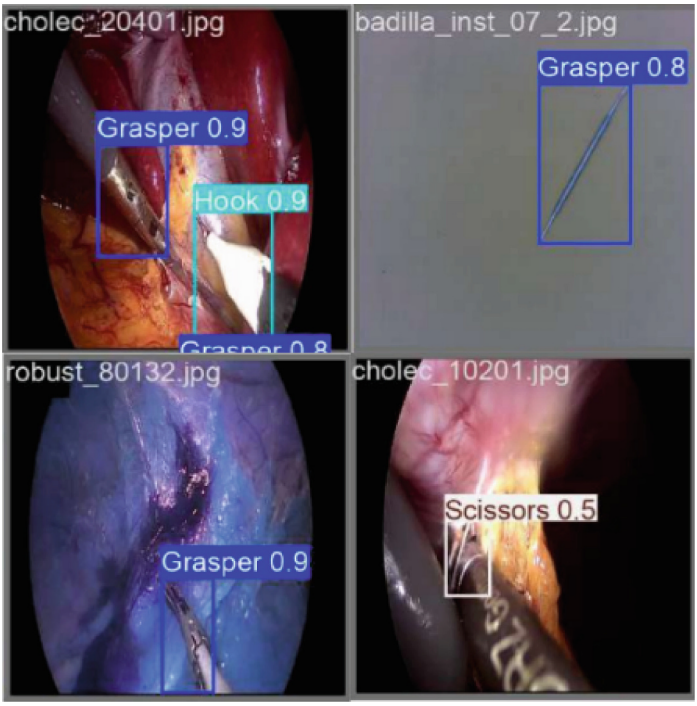

Fig. 2. Qualitative tracking results demonstrating cross-domain generalization. Top-left: CholecTrack20 (Grasper and Hook). Top-right: Badilla-Solórzano (synthetic background). Bottom-left: ROBUST-MIPS (colorectal view). Bottom-right: Accurate detection of a minority class (Scissors). The YOLOv11-s model maintains precise bounding box localization across distinct surgical environments.

An examination of the class distribution indicates a pronounced natural imbalance, with the 'Grasper' class prevailing in the dataset. This distribution

accurately reflects real-world surgical workflows, where graspers are used all the time to move tissue around, but scissors or clips are only used for a short time.

The model did not show any prediction bias toward the majority class, even though there was a big difference between the two groups. The Precision-Recall metrics show that the minority classes were very accurate. For example, "Scissors" had a mAP@0.5 of 91.3% and "Hook" had a mAP@0.5 of 99.3%.

The internal loss mechanisms of the YOLOv11 architecture (like Focal Loss) did a good job of punishing misclassifications of underrepresented classes. This made sure that the whole surgical toolkit could be used without overfitting to the most common instruments.

A zero-shot inference test was conducted using external, unannotated surgical footage to further validate the model's ability to generalize in real-world situations without any restrictions.

In accordance with the qualitative evaluation methodology established by Myo et al. [42], publicly accessible laparoscopic cholecystectomy videos from YouTube (Dr. R. K. Mishra's surgical channel [39]) were employed.

This evaluation substantiates the architecture's resilience, illustrating its capacity to precisely monitor instruments amidst novel illumination profiles and hardware discrepancies absent in the initial training distribution.

5 Limitations and Error Analysis OR Limitations and Gaps to Address

The OSI26 dataset was very useful for finding surgical instruments, but it had a lot of methodological problems during the integration phase. The main problem is that its metadata doesn't have standard class names; it only has numbers for its 26 categories.

Since there is no official documentation that links these IDs to specific surgical instruments, keeping this dataset would hurt the taxonomic consistency needed for a multi-dataset comparative analysis. Because of this, the OSI26 was left out to make sure that the evaluation of the proposed model was reliable and could be repeated.

Also, integrating large-scale surgical datasets was hard because of problems with the latency of cloud-based file systems. The distributed cloud environment often had I/O bottlenecks and synchronization latencies because there were so many small metadata files and the raw surgical frames were so high resolution.

To make sure the data was safe and to get around these problems, a localized caching strategy had to be put in place. This meant moving compressed archives to a SSD local runtime for atomic processing. These problems show how important it is to manage data pipelines well in computer vision tasks, where the physical limits of cloud storage can make things take a lot longer.

The Badilla-Solórzano dataset encodes class identity as raw pixel values in grayscale masks without a publicly available ID-to-instrument mapping. The correspondence adopted here was derived through structural analysis and cross-referencing with the original publication, mapping six out of eighteen detected pixel classes to the Universal Taxonomy.

The remaining twelve classes were excluded to preserve taxonomic consistency, resulting in 116 annotated frames from this source. The assumed mapping has not been independently verified with the original authors, representing a potential source of label noise in the Badilla-sourced training samples.

The high mAP@0.5 achieved across all classes suggests that the derived mapping is consistent with the instrument categories described in the original publication.

Qualitative evaluation also pointed out some specific cases where the visual model had trouble. For example, in the synthetic subset (which has isolated instruments on black backgrounds), a closed surgical scissor was sometimes wrongly labeled as a "Irrigator" on Badilla's dataset.

This mistake can be explained geometrically: when there is only a dark background and no anatomical tissue to give scale or functional cues, a closed scissor looks like the long, narrow metal shaft of an irrigator. Including these failure cases shows the real limits of the architecture and shows how important it is to have contextual anatomical cues when doing complex visual recognition tasks.

To mitigate this specific misclassification, future training protocols should incorporate targeted data augmentation—such as synthetic background randomization—and background-aware learning strategies. This would prevent the network from overfitting to homogeneous dark contexts and force it to rely strictly on the geometric features of the instrument tips.

6 Conclusions

This study introduced a data-driven approach for the identification of surgical instruments across diverse domains. A YOLOv11-s architecture was trained to identify tools in different surgical settings by geometrically combining different types of annotations—bounding boxes, keypoints, and semantic masks—from the Cholec, Robust, and Badilla datasets.

The unified model had a mAP@0.5 of 90.2% and an inference speed of 5.1 milliseconds per frame, which means it could be used for real-time intraoperative processing. Furthermore, the model maintained precision for minority classes such as 'Scissors' and 'Hook', demonstrating stability against the class imbalance inherent to surgical workflows.

Promising directions for future research include the implementation of targeted data augmentation techniques, such as generative synthetic oversampling, to balance minority classes during training; and the prospective deployment of the architecture in clinical and educational settings to evaluate interoperability under operational constraints.

Ethics Statement. The authors affirm that the present study complies with the ethical standards required for publication in the ICCSA. This research does not involve human participants, animal experimentation, or data derived from social media platforms. The datasets utilized in this study are publicly available and comply with the principles of Open Access scientific research.

References

1. Andrade, T.N., Morais, O.J., Ramirez, B.F.D.: Artificial intelligence in medicine: Surgical robots, digital neural networks, health big data, infodemiology, and bioprinting for tissue regeneration. In: Albino, J.P., Valente, V.C.P.N. (eds.) Artificial Intelligence and its Interdisciplinary Applications, chap. 11. e-Publicar, , Curitiba, Brazil (2023). https://doi.org/10.47402/ed.ep.c2023210411201
2. Ángeles Cerón, J.C., et al.: Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion. Med. Image Anal. **81**, 102569 (2022). https://doi.org/10.1016/j.media.2022.102569
3. Arevalo-Ancona, R.E., et al.: Advanced dual-branch U-Net decoder for precise and robust surgical instrument and organ segmentation. Biomed. Signal Process. Control (2025). https://doi.org/10.1016/j.bspc.2025.108348. in Press
4. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
5. Badilla-Solórzano, J., et al.: Deep-learning-based instrument detection for intraoperative robotic assistance. Int. J. Comput. Assisted Radiol. Surg. **17**(9), 1685–1695 (2022). https://doi.org/10.1007/s11548-022-02715-y
6. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934 (2020), https://arxiv.org/abs/2004.10934
7. Bresler, L., et al.: Residency training in robotic surgery: The role of simulation. J. Visceral Surgery **157**(3S2), S123–S129 (2020). https://doi.org/10.1016/j.jviscsurg.2020.03.006, supplement 2
8. Chellamani, G.K.: A robust dataset for surgical instrument detection to aid autonomous robotic surgery. Data Brief **61**, 111798 (2025). https://doi.org/10.1016/j.dib.2025.111798
9. Chong, N., Yang, F.: CMMCAN: lightweight feature extraction and matching network for endoscopic images based on adaptive attention. Comput. Mater. Continua **79**(2), 3145–3164 (2024). https://doi.org/10.32604/cmc.2024.052217
10. Das, A.: PitVis-2023 challenge: workflow recognition in videos of endoscopic pituitary surgery. Med. Image Anal. **99**, 103716 (2025). https://doi.org/10.1016/j.media.2025.103716
11. Deol, E.S., et al.: Artificial intelligence model for automated surgical instrument detection and counting: An experimental proof-of-concept study. Patient Safety Surgery **18**(1), 24 (2024). https://doi.org/10.1186/s13037-024-00406-y
12. Domene, C.E.: Robotic surgery – a step towards the future. ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo) **27**(4), 233–233 (2014). https://doi.org/10.1590/S0102-67202014000400001
13. Duan, X., Li, L., Li, X., Lyu, S., Ding, L., Li, C.: DINOv2-enhanced surgical instrument segmentation with lightweight CNN. Procedia Comput. Sci. **271**, 225–231 (2025). https://doi.org/10.1016/j.procs.2025.10.135
14. Duarte, L.D.C.: Artificial intelligence in surgery: the frontier between technological assistance and medical autonomy. Revista Caderno Pedagógico **22**(12), 1–17 (2025). https://doi.org/10.54033/cadpedv22n12-344
15. ElMoaqet, H., et al.: Deep learning architectures for single-label and multi-label surgical tool classification in minimally invasive surgeries. Appl. Sci. **15**(11), 6121 (2025). https://doi.org/10.3390/app15116121

16. Franco-González, I.T., et al.: Tracking 3D motion of instruments in microsurgery: a comparative study of stereoscopic marker-based vs. deep learning method for objective analysis of surgical skills. Inf. Med. Unlocked **51**, 101593 (2024). https://doi.org/10.1016/j.imu.2024.101593
17. Fujii, R.: Surgical tool detection in open surgery videos. Appl. Sci. **12**(20), 10473 (2022). https://doi.org/10.3390/app122010473
18. Géron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd edn. O'Reilly Media, Sebastopol (2019)
19. Graëff, C., et al.: Introducing surgical workflow recognition in orthopaedic surgery with timestamp supervision. Comput. Biol. Med. **197**, 110995 (2025). https://doi.org/10.1016/j.compbiomed.2025.110995
20. Guan, B., et al.: Enhancing spatial perception in robot-assisted minimally invasive surgery with edge-preserving depth estimation and pose tracking. BMC Surgery **25**, 455 (2025). https://doi.org/10.1186/s12893-025-03198-9
21. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 4th edn. Elsevier, Amsterdam (2020)
22. Han, Z., et al.: ROBUST-MIPS: a combined skeletal pose and instance segmentation dataset for laparoscopic surgical instruments. Scientific Data (2026). https://doi.org/10.1038/s41597-026-06938-5
23. Hu, Z.: Automatic segmentation of surgical instruments in endoscopic spine surgery: a deep learning-based analysis. Asian J. Surg. **48**(3), 1977–1979 (2025). https://doi.org/10.1016/j.asjsur.2024.09.128
24. Hua, S., et al.: Deep learning-based recurrent laryngeal nerve detection during endoscopic thyroidectomy. Intelligent Medicine (August 2025). https://doi.org/10.1016/j.imed.2025.08.005, in Press
25. Jha, D.: Validating polyp and instrument segmentation methods in colonoscopy through Medico 2020 and MedAI 2021 challenges. Med. Image Anal. **99**, 103307 (2025). https://doi.org/10.1016/j.media.2024.103307
26. Jin, A., et al.: Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: Proc. IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 691–699 (2018). https://doi.org/10.1109/WACV.2018.00081
27. Jocher, G., Qiu, J., Chaurasia, A.: YOLO11 by Ultralytics. Ultralytics (2023). https://github.com/ultralytics/ultralytics
28. Kezan, R.D., Chiavegatto, M.A., Marco, D.C.B.: Challenges and benefits of robotic surgery in the modern world. Revista Eletrônica Acervo Saúde **23**(11), e14031 (2023). https://doi.org/10.25248/reas.e14031.2023
29. Khalil, S., Shah, H.A., Bednarik, R.: Improving microsurgical suture training with automated phase recognition and skill assessment via deep learning. Comput. Biol. Med. **192**, 110238 (2025). https://doi.org/10.1016/j.compbiomed.2025.110238
30. Koskinen, J., et al.: Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery. Comput. Biol. Med. **141**, 105121 (2022). https://doi.org/10.1016/j.compbiomed.2021.105121
31. Lavado, D.M.: Sorting Surgical Tools from a Cluttered Tray: Object Detection and Occlusion Reasoning, Coimbra, Portugal (2018). Master's dissertation, Faculdade de Ciências e Tecnologia
32. Lavanchy, J.L.: Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci. Rep. **11**, 5197 (2021). https://doi.org/10.1038/s41598-021-84295-6

33. Lin, T.Y.: Microsoft COCO: Common Objects in Context. In: Proc. European Conference on Computer Vision (ECCV), pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. 7th International Conference on Learning Representations (ICLR). New Orleans, LA, USA (2019). https://openreview.net/forum?id=Bkg6RiCqY7
35. Mahmood, T., Cho, S.W., Park, K.R.: DSRD-Net: dual-stream residual dense network for semantic segmentation of instruments in robot-assisted surgery. Expert Syst. Appl. **202**, 117420 (202https://doi.org/10.1016/j.eswa.2022.117420
36. Matasyoh, N.M., Mathis-Ullrich, F., Zeineldin, R.A.: SAMSurg: surgical instrument segmentation in robotic surgeries using vision foundation model. IEEE Access **12**, 193950–193959 (2024). https://doi.org/10.1109/ACCESS.2024.3520386
37. Medina, F.M.: Desafios da Formação Médica na Modernidade: A Experiência da Sociedade Brasileira de Oftalmologia. Revista Brasileira de Oftalmologia **84**, e0076 (2025). https://doi.org/10.37039/1982.8551.20250076
38. Mikhail, D., et al.: Quantitative analysis of instrument motion paths in cataract surgery across a resident's training. Ophthalmology Sci. **6**(2), 101014 (2026). https://doi.org/10.1016/j.xops.2025.101014
39. Mishra, D.R.K.: D. R. K. Mishra - YouTube Channel. YouTube (2026) https://www.youtube.com/channel/UC5GRDQvh5Ze-agGcaMeveKw
40. Moher, D., et al.: Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Systematic Reviews **4**(1), 1 (jan 2015https://doi.org/10.1186/2046-4053-4-1
41. Morrell, A.L.G., et al.: The history of robotic surgery and its evolution: when illusion becomes reality. Revista do Colégio Brasileiro de Cirurgiões **48**, e20202798 (2021). https://doi.org/10.1590/0100-6991e-20202798
42. Myo, N.N., et al.: A two-point association tracking system incorporated with YOLOv11 for real-time visual tracking of laparoscopic surgical instruments. IEEE Access **13**, 12225–12238 (2025). https://doi.org/10.1109/ACCESS.2025.3529710
43. Nwoye, C.I., Padoy, N.: SurgiTrack: Fine-grained multi-class multi-tool tracking in surgical videos. Med. Image Anal. **101**, 103438 (2025). https://doi.org/10.1016/j.media.2024.103438
44. Nwoye, C.I.: Cholectrack20: A multi-perspective tracking dataset for surgical tools. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8942–8952 (2025)
45. Olei, S., et al.: The AENEAS project: Intraoperative anatomical guidance through real-time landmark detection using machine vision. Mayo Clinic Proceedings: Digital Health **4**(1), 100308 (2026). https://doi.org/10.1016/j.mcpdig.2025.100308
46. Pereira-Arias, J.G., et al.: How to build a robotic program. Archivos Españoles de Urología **72**(3), 227–238 (2019). pMID: 30945649
47. Rai, U., Xu, H., Giannarou, S.: SurgPose: Generalisable surgical instrument pose estimation using zero-shot learning and stereo vision. arXiv preprint arXiv:2505.11439, May 2025. https://doi.org/10.48550/arXiv.2505.11439, https://doi.org/10.48550/arXiv.2505.11439, preprint
48. Ramesh, S., et al.: Dissecting self-supervised learning methods for surgical computer vision. Med. Image Anal. **88**, 102844 (2023). https://doi.org/10.1016/j.media.2023.102844
49. Ran, B.: Surgical instrument detection algorithm based on improved yolov7x. Sensors **23**(11), 5037 (2023). https://doi.org/10.3390/s23115037

50. Roß, T., et al.: Beyond rankings: Learning (more) from algorithm validation. Medical Image Analysis **86**, 102765 (2023). https://doi.org/10.1016/j.media.2023.102765
51. Ruiz, N., et al.: Accuracy and reliability of artificial intelligence in surgical decision-making: a literature review. Cureus **17**(10), e95337 (2025). https://doi.org/10.7759/cureus.95337
52. Saadati, S., Hashemi, M., Nezhat, C.: Unsupervised deep learning for semantic segmentation using laparoscopic videos: A self-detection and self-learning approach. Mach. Learn. Appl. **22**, 100767 (2025). https://doi.org/10.1016/j.mlwa.2025.100767
53. Sasaki, K.: Automated Surgical Workflow Identification by Artificial Intelligence in Laparoscopic Hepatectomy: Experimental Research. Int. J. Surg. **105**, 106856 (2022). https://doi.org/10.1016/j.ijsu.2022.106856
54. Sugiyama, T., et al.: Artificial intelligence-integrated video analysis of vessel area changes and instrument motion for microsurgical skill assessment. Scientific Reports **15**, 27898 (2025). https://doi.org/10.1038/s41598-025-27898-z
55. Wang, Y.H., et al.: SMILEtrack: similarity learning for occlusion-aware multiple object tracking. In: Proc. AAAI Conference on Artificial Intelligence, vol. 38, pp. 5740–5748 (2024). https://doi.org/10.1609/aaai.v38i6.28386
56. Wiese, L., et al.: Detection of surgical instruments based on synthetic training data. Computers **14**(2), 69 (2025). https://doi.org/10.3390/computers14020069
57. Wu, Z., et al.: Augmenting efficient real-time surgical instrument segmentation in video with point tracking and Segment Anything. Healthcare Technol. Lett. (2024). https://doi.org/10.1049/htl2.12111
58. Xie, D., et al.: Surgical instruments hyalinization: Occlusion removal in minimally invasive endoscopic surgery. Biomimetic Intell. Robot. **3**(3), 100105 (2023). https://doi.org/10.1016/j.birob.2023.100105
59. Xu, H., et al.: Surgripe challenge: benchmark of surgical robot instrument pose estimation. Med. Image Anal. **105**, 103674 (2025). https://doi.org/10.1016/j.media.2025.103674
60. Xu, Z., et al.: Spine endoscopic atlas: an open-source dataset for surgical instrument segmentation. Scientific Data **12**, 1611 (2025). https://doi.org/10.1038/s41597-025-05897-7
61. Yang, X., et al.: CUE-YOLO: an efficient and robust detector for metal surface defect detection with adaptive attention and task alignment. IEEE Access **14**, 37020–37037 (2026). https://doi.org/10.1109/ACCESS.2026.3670661
62. Zhang, T., Yuan, X., Xu, H.: Surgical Instrument Segmentation via Segment-Then-Classify Framework with Instance-Level Spatiotemporal Consistency Modeling. J. Imaging **11**(10), 364 (2025). https://doi.org/10.3390/jimaging11100364