# Automated Performance Metrics for Objective Surgical Skill Assessment in Laparoscopic Training

Asaf Arad[1], Julia Leyva I Torres[1], Kristian Nyborg Jespersen[1], Nicolaj Boelt Pedersen[1], Pablo Rey Valiente[1], Alaa El-Hussuna[2] and Andreas Møgelmose[1]

[1]*Department of Electronic Systems, Aalborg University, Denmark*

[2]*Open Source Research Collaboration, Aalborg, Denmark*

Keywords: Automated Performance Metrics, Laparoscopic Surgery, Multi Object Tracking.

Abstract: The assessment of surgical skill is critical in advancing surgical training and enhancing the performance of surgeons. Traditional evaluation methods relying on human observation and checklists are often biased and inefficient, prompting the need for automated and objective systems. This study explores the use of Automated Performance Metrics (APMs) in laparoscopic surgeries, using video-based data and advanced object tracking techniques. A pipeline was developed, combining a fine-tuned YOLO11 model for detection with state-of-the-art multi-object trackers (MOTs) for tracking surgical tools. Metrics such as path length, velocity, acceleration, jerk, and working area were calculated to assess technical performance. BoT-SORT emerged as the most effective tracker, achieving the highest HOTA and MOTA, enabling robust tool tracking. The system successfully extracted APMs to evaluate and compare surgical performance, demonstrating its potential for objective assessment. This work validates state-of-the-art algorithms for surgical video analysis, contributing to improved surgical training and performance evaluation. Future efforts should address limitations like pixel-based measurements and dataset variability to enhance the system's accuracy and applicability, ultimately advancing patient safety and reducing training costs.



## 1 INTRODUCTION

Accurate assessment of surgical performance is a cornerstone of surgical training, especially as the field advances toward proficiency-based methodologies (Jin et al., 2018; Ebina et al., 2022a; Ebina et al., 2022b; Guerin et al., 2022). Traditional methods of evaluating trainees, which rely on human observers and task-specific checklists, are accessible, but suffer from bias and time inefficiencies (D'Angelo et al., 2015). These limitations highlight the need for automated and objective assessment systems which provide consistent and detailed feedback to trainees.

APMs have emerged as a promising solution, offering objective and data-driven evaluations as suggested by Ebina et al. (Ebina et al., 2022b). By leveraging video-based data and computer vision techniques, APMs can be used to analyze surgical performance with higher precision and reproducibility. Unlike traditional observer-based methods, APMs eliminate bias and offer a standardized approach for evaluating surgical skills, allowing the surgeons to potentially improve upon surgical training programs

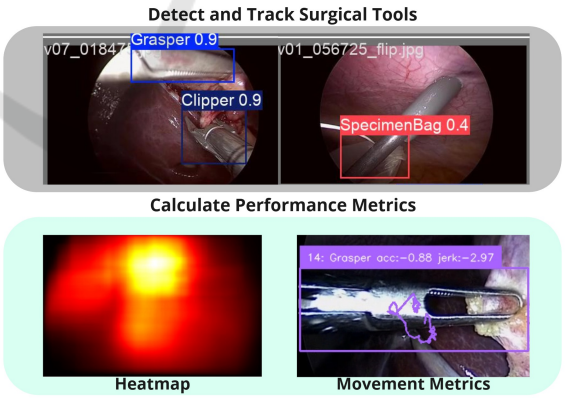



Figure 1: Simple illustration showcasing the detection of surgical instruments used to calculate popular performance metrics for surgical skill evaluation.

(Buckley et al., 2015; Peng et al., 2019; Sallaberry, Tori, and Nunes, 2022).

Recent advancements in artificial intelligence, particularly in object detection and object tracking, have enabled the development of robust APM systems. Technologies such as Convolutional Neural



Arad, A., Leyva I. Torres, J., Jespersen, K. N., Pedersen, N. B., Valiente, P. R., El-Hussuna, A. and Møgelmose, A.
Automated Performance Metrics for Objective Surgical Skill Assessment in Laparoscopic Training.
DOI: 10.5220/0013380300003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 824-830
ISBN: 978-989-758-728-3; ISSN: 2184-4321

Networks (CNNs) allow automatic tracking and analysis of surgical instruments, providing a detailed understanding of tool usage, motion efficiency, and task completion. These capabilities pave the way for generating a wide range of performance metrics, including time efficiency, path consistency, and tool utilization patterns, offering valuable insights for both trainees and trainers (Twinanda et al., 2016; Jin et al., 2018; Rivas-Blanco et al., 2021).

This work focuses on the identification and implementation of common APMs for surgical performance evaluation in laparoscopic surgeries; a minimally invasive surgical method using small incisions and a camera to guide the procedure (Rivas-Blanco et al., 2021). Specifically, the project aims to develop a pipeline of an image recognition system, which is capable of tracking surgical instruments and calculating performance metrics based on their motion and usage during procedures (Figure 1). The main contribution of this work is validation of state-of-the-art trackers' abilities to extract popular metrics used for evaluating surgical skills from 2D captured videos of laparoscopic surgeries. In future work, these metrics could serve as a basis for comparative studies with traditional surgical skill evaluation techniques, providing insights into their alignment with surgical expertise.

## 2 RELATED WORKS

When identifying the most commonly used APMs, the following was found: Metrics such as path length (the distance traveled by a tool), operative time (total and sub-task durations), velocity (tool movement speed), jerk (smoothness of motion), acceleration (rate of motion change), and tool angle (orientation of tools) are frequently implemented for their ability to evaluate motion efficiency and skill precision in both laparoscopic and robotic surgeries (Buckley et al., 2015; Ebina et al., 2022b; Guerin et al., 2022).

Certain metrics, such as force measurements, are specific to robotic surgeries due to the built-in sampling of instrument sensor data, which enables precise evaluation of applied forces (Sallaberry, Tori, and Nunes, 2022; Trejos et al., 2014). Other metrics can also be categorized into procedure-specific metrics, such as stitching techniques used in bladder suturing (Chen et al., 2018). Similarly, deep features extracted via deep learning approaches can be considered a separate category (Jin et al., 2018; Reiley et al., 2011; Rivas-Blanco et al., 2021; Moglia et al., 2021).

Figure 2 summarizes the 10 most commonly used APMs, according to an unstructured literature review conducted by us, based on 26 relevant papers found through Google Scholar and PubMED using the keywords *Automatic Performance Metrics* AND (*Surgery* OR *Surgery Training*) and *Objective Evaluation* AND (*Surgery* OR *Surgery Training*).

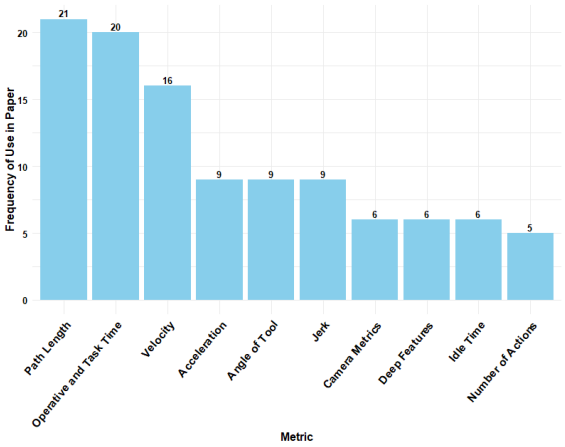



Figure 2: Table showcasing the 10 most popular performance metrics found, ordered by most frequently found (left) to least frequently found (right).

The majority of popular APMs rely on tracking surgical tools, often through video data, to provide the positional and motion information needed for performance evaluation. However, tool tracking in surgical environments is challenging due to factors such as low-texture organ surfaces, visual artifacts from fluids or reflections, and dynamic environmental conditions like blood or smoke (Schmidt et al., 2024).

## 3 METHODOLOGY

The methodology for this project was designed to develop a comprehensive pipeline for assessing surgical performance. The process involves the detection, tracking, and analysis of surgical tools within laparoscopic videos, as illustrated in Fig. 3.

The workflow begins with video input of surgical procedures, where the tools used need to be detected and localized. A YOLO11 object detection model, was fine-tuned on a surgical tool dataset to detect the tools' bounding boxes in each frame. The detected bounding boxes were then passed to state-of-the-art MOTs to establish temporal consistency between frames. The selection of the most suitable tracker was based on MOT metrics such as HOTA, MOTA, and IDF1. The tracked tools' motion data, including path length, velocity, acceleration, jerk, and working area, were then aggregated into APMs. These metrics were compiled into an output table, facilitating comparative analysis of surgical videos and providing possible

insights into the performance of these procedures.

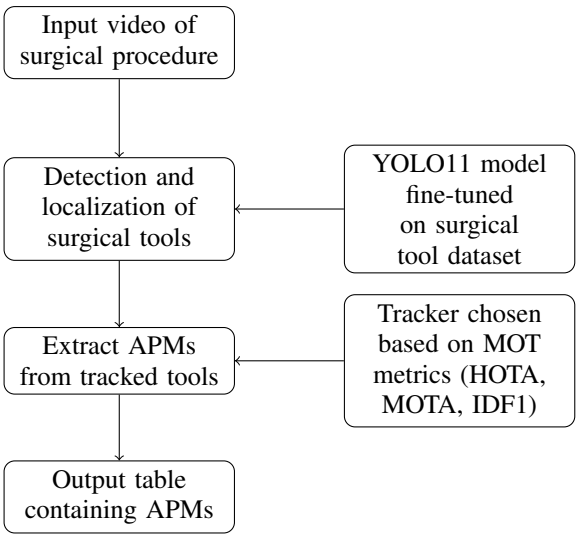



Figure 3: Flow Chart of the steps for extracting APMs from surgical videos.

## 3.1 Identifying State of the Art Trackers

Three state-of-the-art MOTs have been identified and implemented: BOT-SORT, ByteTrack, and UniTrack. These three trackers were selected based on their good performance on benchmarks from (Benchmark", n.d.) and (Code", n.d.).

- BOT-SORT enhances the Simple Online and Realtime Tracking (SORT) framework with improved Kalman filtering, camera motion compensation, and a combination of Intersection over Union (IoU) and Re-Identification (ReID) (Aharon, Orfaig, and Bobrovsky, 2022).
- ByteTrack leverages low-confidence detections to recover missed objects, using a two-step association process (Zhang et al., 2022).
- UniTrack employs a unified model for multiple tracking tasks and uses reconstruction-based similarity for better associations (Wang et al., 2021).

## 3.2 Tracker Evaluation

To evaluate the trackers, MOT metrics have been calculated, including Higher Order Tracking Accuracy (HOTA), Multi-Object Tracking Accuracy (MOTA), and Identity F1 (IDF1) for each tracker. These evaluation metrics quantify how well the trackers perform on a given data set. MOTA prioritizes detection (identification and localization) accuracy, IDF1 prioritizes association (maintaining the identity of objects between frames) accuracy, and HOTA was created to balance the accuracy of both detection and association to align with human perception (J Luiten., 2021).

## 3.3 Datasets

The dataset was employed by a combination of publicly available datasets from laparoscopic videos. The dataset used for model training is *m2cai16-tool-locations* (Jin et al., 2018), which includes 2,532 frames annotated with bounding box coordinates for seven surgical tools (Grasper, Bipolar, Hook, Scissors, Clipper, Irrigator and Specimen Bag). This dataset was chosen because of its detailed annotations and diversity of tool types, which facilitated robust model training. Examples are shown in figure 4.

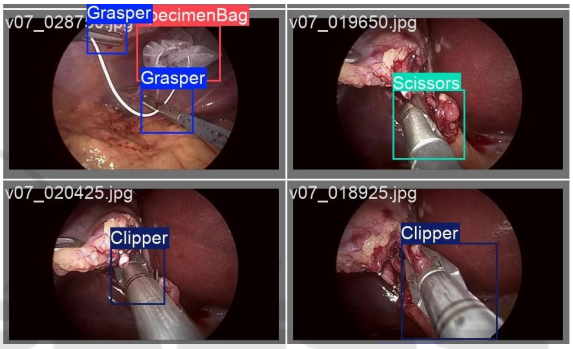



Figure 4: Frame examples of *m2cai16-tool-locations* dataset.

For testing the trackers' performance, we used the *Cholec80-Boxes* dataset (Abdulbaki Alshirbaji et al., 2024), which contains annotations at 1 fps for 5 videos from *Cholec80* (Twinanda et al., 2016). As opposed to *m2cai16-tool-locations*, this dataset contains sequential frames arranged in chronological order.

## 3.4 Fine Tuning Using YOLO11

As all the identified trackers use YOLO for object detection, a YOLO model was fine-tuned to adapt it in our domain. For this, YOLO11 was chosen as it is the latest, and best performing, version at the time of writing (Jocher and Qiu, 2024). The model was fine-tuned exclusively with the *m2cai16-tool-locations* dataset using a dataset split of 70%/15%/15% (Training, validation, test).

For fine-tuning, the medium model, YOLO11m, was chosen, as it has a good trade-off between complexity and performance. The model was trained for 500 epochs using the default hyperparameters provided by Ultralytics (Jocher and Qiu, 2024).

Figure 5 illustrates the training and validation loss across the epochs, stopping at epoch 215 due to early

stopping. The model achieved its best performance at epoch 115, with a box validation loss of 1.38 and a CLS validation loss of 1.01. The corresponding precision and recall scores were 0.48 and 0.91, respectively.

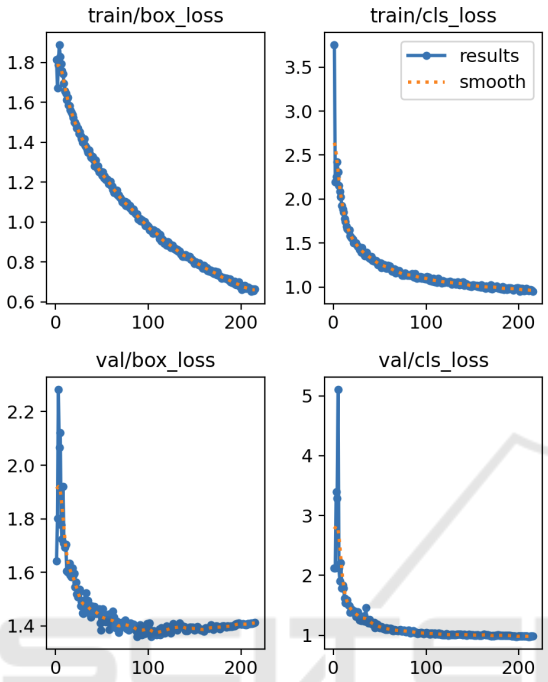



Figure 5: Training/Validation loss (Y-axis) of the fine-tuned YOLO model for each epoch (X-axis).

Testing the model on the remaining 15% of the data resulted in a mean average precision (mAP@0.5) of 0.957 across all classes, with the Hook class achieving the highest mAP@0.5 (0.992), and the Grasper class achieving the lowest mAP@0.5 (0.896). Table 1 shows the confusion matrix of our classification model where columns represent the true classes and rows the prediction. The lowest detection score was observed for the *Grasper*, which was misclassified as background in 0.16% of cases. However, it is worth noting that the *Grasper* appeared far more frequently in the dataset (223 instances) compared to the other classes *Bipolar* (72), *Hook* (49), *Scissors* (66), *Clipper* (48) *Irrigator* (57), and *Specimen Bag* (77). Generally, the classification errors primarily stem from missed detections as opposed to confusion between the classes.

## 3.5 Acquisition of APMs

In order to be able to quantitatively assess surgical performance, the following APMs were chosen for calculation: path length (Eq. 1), average velocity (Eq. 2), average acceleration (Eq. 3), and average jerk

Table 1: Confusion matrix for the YOLO11 test results, highlighting which classes the model confuses with each other.

| *Predicted / True* | Grasper | Bipolar | Hook | Scissors | Clipper | Irrigator | SpecimenBag | Background |
|---|---|---|---|---|---|---|---|---|
| Grasper | 0.84 | | | 0.03 | | | 0.03 | 0.64 |
| Bipolar | | 0.94 | | | | | | |
| Hook | | | 0.96 | | | | | |
| Scissors | | | | 0.92 | | | | 0.03 |
| Clipper | | | | 0.02 | 0.92 | | | 0.03 |
| Irrigator | | | | | | 0.93 | | 0.12 |
| SpecimenBag | | | | | | | 0.87 | 0.15 |
| Background | 0.16 | 0.06 | 0.04 | 0.03 | 0.08 | 0.07 | 0.10 | |

(Eq. 4). These were selected based on the most commonly used APMs derived from the literature search (see 2) that we also deemed possible to calculate using the datasets that were available. In addition, working area (Eq. 5) can be managed as it depends only on the object position appearance. The APMs are acquired using the center of the bounding box from the tracker results. The spatial metrics are obtained using Savitzky-Golay smoothing filter of the position of the tools suggested by Ebina et al. (Ebina et al., 2022b).

$$\mathcal{PL} = \sum_{i=1}^{n-1} \|p_{tip}(i+1) - p_{tip}(i)\| \quad (1)$$

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} \|\frac{d}{dt} p_{tip}(i)\| \quad (2)$$

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} \|\frac{d^2}{dt^2} p_{tip}(i)\| \quad (3)$$

$$\bar{\jmath} = \frac{1}{n}\sum_{i=1}^{n} \|\frac{d^3}{dt^3} p_{tip}(i)\| \quad (4)$$

$$\mathcal{WA} = (x_{(q=97.5\%)} - x_{(q=2.5\%)}) \cdot (y_{(q=97.5\%)} - y_{(q=2.5\%)}) \quad (5)$$

Where $p_{tip}(i)$ represents the tool tip in frame $i$, and $x_{(q=97.5\%)}$ and $x_{(q=2.5\%)}$ are the 97.5-percentile and 2.5-percentile, respectively, of the bounding box center position in the X-axis, and similarly for the Y-axis.

# 4 RESULTS

We evaluated state-of-the-art trackers using standard metrics (MOTA, HOTA, and IDF1). Among the evaluated trackers, BoT-SORT achieved the highest scores and proved to be the most effective for our dataset. As a result, it was selected for computing the APMs, which are used to compare the technical

skill and performance across different laparoscopic surgery videos.

## 4.1 Trackers

Table 2: MOT metrics from BoT-SORT, ByteTrack, and UniTrack of the videos from *Cholec80-Boxes* dataset.

| *Tracker* \ *Metric* | **MOTA** | **HOTA** | **IDF1** |
|---|---|---|---|
| **BoT-SORT** | 54.46 | 12.61 | 9.17 |
| **ByteTrack** | 51.17 | 11.44 | 8.48 |
| **UniTrack** | 48.77 | 11.71 | 9.37 |

Table 2 shows the results of MOTA, HOTA and IDF1 metrics obtained from each tracker. BoT-SORT was selected as the main tracker due to its superior HOTA score (12.61), reflecting its strong ability to balance detection accuracy and trajectory association. Although BoT-SORT only demonstrated marginally better performance on individual metrics, it demonstrated consistently solid performance overall. It also achieved the highest MOTA (54.46) and performed well in IDF1 (9.17) compared to the others. ByteTrack, although slightly behind BoT-SORT, showed strong performance in MOTA (51.17) but had the lowest HOTA score (11.44). UniTrack was the best in IDF1 (9.37) but lagged behind in MOTA and HOTA, indicating a less balanced performance overall.

## 4.2 APMs

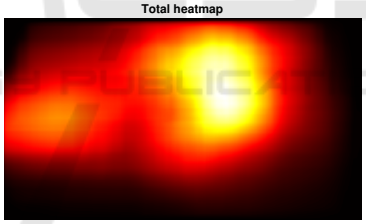
Table 3 shows the APM scores of the videos from *Cholec80-Boxes* dataset. The videos have approximately the same length and were captured using the same surgical instrument types as mentioned in section 3.3. The chosen APMs were acquired using the calculations in section 3.5. The scores for each video were acquired using BoT-SORT tracking outputs for each tool type, then the total scores were calculated as a weighted mean of all instruments. The average weighting was acquired using the appearance temporal ratio of each tool type in the entire video, i.e., the ratio between the number of frames each tool appeared and the total number of frames. In addition, the path length score is represented by the total path length per minute in each video, for scalability.

It can be observed that *video04* and *video02* have the longest path length and the largest working area. They also have the highest average velocity, acceleration, and jerk, i.e. the movements in those videos are incoherent and might often have jitter and shakiness. In contrast, *video05* has the shortest path length and the smallest working area, and a smaller average velocity, acceleration, and jerk. Thus, it might have smoother and less shaky movements and more coherent and accurate movements during the operation.

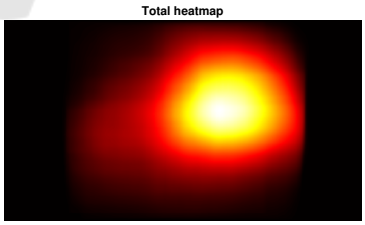
Table 3: Examples of APM calculations from different videos.

| | **PL** $[pixel/min]$ | $\bar{\mathbf{v}}$ $[pixel/s]$ | $\bar{\mathbf{a}}$ $[pixel/s^2]$ | **J** $[pixel/s^3]$ | **WA** $[pixel^2]$ |
|---|---|---|---|---|---|
| **video01** | 3213 | 57.67 | 30.16 | 32.24 | 149250 |
| **video02** | 3771 | 80.86 | 38.38 | 40.43 | 122654 |
| **video03** | 1715 | 63.15 | 30.81 | 32.84 | 110318 |
| **video04** | 3258 | 65.14 | 36.64 | 39.06 | 147245 |
| **video05** | 1888 | 56.94 | 29.49 | 31.17 | 81156 |

Furthermore, in order to analyze and comprehend the APMs acquired from tracking a video, we generate heat maps of the total number of appearances and the locations of the surgical tools from laparoscopic videos. Heat maps illustrate the working area metric, and may serve as an indicator of technical skill, as it shows how focused the surgeon is during the procedure. It may leverage comparison methods between different videos (Jin et al., 2018). Figure 6 shows examples of heat maps obtained from 2 aforementioned tracked videos: *video04* (fig. 6a) and *video05* (fig. 6b). It may be deemed from the visualization of the heat maps that *video05* has more focused and accurate movements, compared to *video04*, as matching the comparison of the aforementioned metrics in table 3.




(a) video04.




(b) video05.

Figure 6: Heat map of bounding box occurrences and locations of tools in two full laparoscopic videos.

# 5 DISCUSSION

In this study, we focused on automating the assessment of surgical performance in laparoscopic surgeries by tracking surgical tools and computing APMs.

Our findings show that by leveraging video-based analysis and state-of-the-art object tracking techniques, we were able to generate detailed performance metrics such as path length, velocity, acceleration, jerk, and working area. These metrics may offer an objective and data-driven approach to evaluate surgical performance.

While we can successfully extract the aforementioned APMs from input videos, we are currently not able to evaluate how these APMs translate to surgical performance. Making such an evaluation would require thorough analysis of the videos by one or more expertly trained surgeons. This analysis could potentially be correlated with the extracted APMs to determine what classifies as good surgical performance. Thus, given the APMs of the videos, we can only compare the scores between different videos. Future work is required to determine how well the extracted APMs can predict surgical performance.

Moreover, it is not currently possible for us to evaluate the accuracy of the APMs we have calculated. The datasets we have acquired do not include ground truth data regarding the positioning of the tools relative to the environment they are in. Because of this, we cannot make conclusions about how good our APM estimates are, but only conclude about comparison between one video's surgery performance over another. One factor that could also potentially introduce error is the spatial component in some APMs, such as velocity, acceleration, jerk, and path length, because the calculation requires depth information. The tools move around in all directions, making it hard to estimate depth from 2D videos. It might be possible to estimate this if the camera parameters are known, but these parameters are not available for the datasets that we have used.

Additionally, the nature of laparoscopic surgeries introduces further complexity. These surgeries often involve a person manually holding the camera, leading to slight movements that can add noise to the APM calculations. This camera movement is not taken into account when calculating the APMs.

Although APMs are still in their infancy, this emerging field holds significant promise for transforming surgical training and enhancing the performance of experienced surgeons. APMs can provide actionable feedback to surgeons, potentially reducing the reliance on high surgical volumes for skill acquisition. With the current constraints on surgical care limiting procedure volumes, the development and refinement of APMs offer a viable solution to optimize training and performance evaluation. Future research should focus on expanding the repertoire of APMs and advancing their accuracy and applicability in assessing surgical performance.

# 6 CONCLUSION

In this project, we developed an automated system to evaluate surgical performance, providing a foundation for improved assessment methods. The primary objective was to create a pipeline capable of processing laparoscopic surgery videos, detecting and tracking surgical instruments, and calculating APMs to support objective evaluation.

The fine-tuned YOLO11 model demonstrated strong performance on the *m2cai16-tool-locations* dataset, achieving a mean average precision (mAP@0.5) of 0.957. However, on the *Cholec80-Boxes* dataset, it scored a lower mAP@0.5 of 0.65, highlighting dataset-dependent variability. Despite this, the model provided a robust foundation for tracking.

For tracking, ByteTrack, BoT-SORT, and UniTrack were evaluated using established MOT metrics such as HOTA, MOTA, and IDF1. These MOT metrics helped identify BoT-SORT as the most effective tracker for this application, balancing detection accuracy and identity association across the evaluation metrics.

The system successfully extracted APMs, such as path length, velocity, acceleration, jerk, working area, and usage time distribution, enabling comparative analyses between surgical videos. Future work should address our limitations to improve APM accuracy.

In conclusion, this work validates state-of-the-art detection and tracking algorithms' ability to compute APMs from 2D laparoscopic surgery videos. This project lays a foundation to improve the quality and fidelity of surgical training, offering a potential in enhancing patient safety and reducing training costs.

# ACKNOWLEDGMENT

We would like to thank Aalborg University for providing the computational resources needed for this project.

# REFERENCES

Abdulbaki Alshirbaji, T. et al. (Aug. 2024). *Cholec80-Boxes: Bounding-Box Labels for Surgical Tools in Five Cholecystectomy Videos*. Zenodo. DOI: 10.5281/

zenodo . 13170928. URL: https : / / doi . org / 10 . 5281 / zenodo.13170928.

Aharon, N., R. Orfaig, and B.-Z. Bobrovsky (2022). "BoT-SORT: Robust Associations Multi-Pedestrian Tracking". In: *arXiv preprint arXiv:2206.14651*.

Benchmark", " O. T. (n.d.). Accessed on: Dec. 11, 2024. [Online]. Available: https://www.MOTChallenge.net.

Buckley, C. E. et al. (2015). "Zone calculation as a tool for assessing performance outcome in laparoscopic suturing". In: *Surgical Endoscopy* 29, pp. 1553–1559.

Chen, J. et al. (2018). "Use of automated performance metrics to measure surgeon performance during robotic vesicourethral anastomosis and methodical development of a training tutorial". In: *The Journal of urology* 200.4, pp. 895–902.

Code", " with (n.d.). Accessed on: Dec. 11, 2024 [Online]. Available: https://www.paperswithcode.com.

D'Angelo, A.-L. et al. (2015). "Idle time: an underdeveloped performance metric for assessing surgical skill." In: *American journal of surgery* 209, pp. 645–651.

Ebina, K. et al. (2022a). "Automatic assessment of laparoscopic surgical skill competence based on motion metrics". In: *PLoS one* 17.11, e0277105.

— (2022b). "Objective evaluation of laparoscopic surgical skills in wet lab training based on motion analysis and machine learning". In: *Langenbeck's archives of surgery* 407.5, pp. 2123–2132.

Guerin, S. et al. (2022). "Review of automated performance metrics to assess surgical technical skills in robot-assisted laparoscopy". In: *Surgical Endoscopy*, pp. 1–18.

J Luiten. A Ošep., P. D. e. a. (2021). "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking". In: *International Journal of Computer Vision* 129, pp. 548–578. URL: https://doi.org/10.1007/s11263-020-01375-2.

Jin, A. et al. (2018). *Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks*. arXiv: 1802 . 08774 [cs.CV]. URL: https://arxiv.org/abs/1802.08774.

Jocher, G. and J. Qiu (2024). *Ultralytics YOLO11*. Version 11.0.0. URL: https : / / github . com / ultralytics / ultralytics.

Moglia, A. et al. (2021). "A systematic review on artificial intelligence in robot-assisted surgery". In: *International Journal of Surgery* 95, p. 106151.

Peng, W. et al. (2019). "An automatic skill evaluation framework for robotic surgery training". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 15.1, e1964.

Reiley, C. E. et al. (2011). "Review of methods for objective surgical skill evaluation". In: *Surgical endoscopy* 25, pp. 356–366.

Rivas-Blanco, I. et al. (2021). "A review on deep learning in minimally invasive surgery". In: *IEEE Access* 9, pp. 48658–48678.

Sallaberry, L. H., R. Tori, and F. L. Nunes (2022). "Automatic performance assessment in three-dimensional interactive haptic medical simulators: A systematic review". In: *ACM Computing Surveys* 55.7, pp. 1–35.

Schmidt, A. et al. (2024). "Tracking and mapping in medical computer vision: A review". In: *Medical Image Analysis* 94, p. 103131. ISSN: 1361-8415. DOI: https: / / doi . org / 10 . 1016 / j . media . 2024 . 103131. URL: https : / / www . sciencedirect . com / science / article / pii / S1361841524000562.

Trejos, A. L. et al. (2014). "Development of force-based metrics for skills assessment in minimally invasive surgery". In: *Surgical endoscopy* 28, pp. 2106–2119.

Twinanda, A. et al. (Feb. 2016). "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos". In: *IEEE Transactions on Medical Imaging* 36. DOI: 10.1109/TMI.2016.2593957.

Wang, Z. et al. (2021). "Do different tracking tasks require different appearance models?" In: *Advances in Neural Information Processing Systems* 34, pp. 726–738.

Zhang, Y. et al. (2022). "ByteTrack: Multi-Object Tracking by Associating Every Detection Box". In.