Linus L. Kienle*, Anna Hilsmann, Peter Eisert, Michael Knoke, and Eric L. Wisotzky

Surgical Instrument Detection on the Instrument Stand using Neural Networks

https://doi.org/10.1515/cdbme-2025-0235

Abstract: Efficient detection of surgical instruments in operating room (OR) settings is critical for advancing surgical workflow management systems. While deep learning-based object detection has achieved promising results in minimally invasive surgery, the detection of instruments on the instrument stand during open procedures remains underexplored. This study introduces two annotated datasets simulating real-world OR scenes, featuring instruments from different manufacturers. We evaluate two detection approaches: a standalone YOLOv8x model and a custom two-stage pipeline that combines a YOLOv8x with a ResNet-34 model. Both models were trained and evaluated on the first dataset and subsequently tested on a second dataset containing instruments from different manufacturers to assess cross-manufacturer generalizability. On the first dataset, the two-stage pipeline achieved a mAP50 of 99.2%, outperforming the standalone YOLOv8x model (98.4%) by improving detection accuracy of stacked instruments. However, both models exhibited notable performance drops (YOLOv8x: 75.9%, pipeline: 78.4% mAP50) when applied to instruments from different manufacturers, highlighting the impact of inter-manufacturer variability in instrument morphology. Our findings emphasize the need for more comprehensive, multi-source datasets to enable robust and generalizable instrument detection solutions in diverse surgical environments.

Keywords: deep learning, convolutional neural network, surgical instrument detection, surgical workflow analysis.

1 Introduction

Surgical workflow and operating room (OR) management systems are integrated systems that support surgeons and OR staff with time-consuming manual and information-related tasks. These systems recognize the current surgical context and collect relevant information that can be displayed, processed, or used for automatic assistance functions. Potential capabilities of such systems include efficient documentation and automatic report generation, surgical phase detection, and automated surgical count [1]. Therefore the integration of such systems in surgical practice has the potential to improve OR efficiency, reduce documentation burdens, and prevent instrument retention events [2].

An important informational aspect for the development of these functionalities is the detection of surgical instruments in the operating environment [3]. Camera-based detection systems integrated with deep learning offer a practical and cost-effective solution and can be easily incorporated into existing OR infrastructure. Research in surgical instrument detection has focused primarily on identifying instruments from laparoscopic images during minimally invasive procedures [4]. Despite the rise in minimally invasive surgeries, most procedures are still performed using open surgical techniques [5]. Accurate detection and counting of surgical instruments on the instrument stand is essential for preventing the misplacement or retention of instruments, and enabling efficient and reliable documentation. These capabilities are particularly important in open surgeries, where a large number of instruments are used and handled outside the endoscopic view. Despite this relevance, publicly available datasets for instrument detection on the instrument stand in open procedures remain extremely limited. Existing datasets rarely depict realistic OR settings and often lack images with occlusions, stacked instruments [6, 7], or sufficient variety in instrument types and manufactures [8], which significantly affects model robustness.

In this work, we propose a two-stage instrument detection pipeline that addresses the introduced real-world challenges and present a dataset that captures realistic OR scenes of the instrument stand.

2 Materials and Methods

We evaluated two architectures for surgical instrument detection on the acquired datasets. Initially a standalone YOLOv8x network [9], which is a state-of-the-art object detection model was employed. In early experiments we observed that the standard Non-Maximum Suppression (NMS) post-processing step impeded reliable detection of stacked instruments.

***Corresponding author: Linus L. Kienle,** Fraunhofer HHI & Klinikum Fulda, Berlin, Germany, e-mail: linus.kienle@hhi.fraunhofer.de
Anna Hilsmann, Fraunhofer HHI, Berlin, Germany
Peter Eisert, Eric L. Wisotzky, Fraunhofer HHI & Humboldt Universität zu Berlin, Berlin, Germany
Michael Knoke, Klinik für Hals-, Nasen-, Ohrenheilkunde, Charité – Universitätsmedizin Berlin, Berlin, Germany

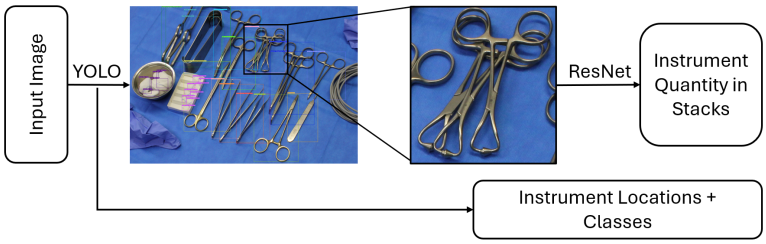

Fig. 1: The proposed instrument detection pipeline. The YOLO network identifies instruments positioned on the instrument table, detecting, in this example, a stack of towel clamps. The section of the image containing the detected stack is then given to the ResNet, which determines the number of instruments inside the stack.

To address this limitation, we developed a two-stage pipeline combining a YOLOv8x detector with a ResNet-34 model, see Figure 1. This architecture leverages domain-specific knowledge that certain instrument classes are typically arranged in stacks and that stacking generally occurs only between instruments of the same class. In the first stage, the YOLOv8x model was employed to detect instrument location within the images, classify instrument type, and determine whether the detection represented a single instrument or a stack. In the second stage, the ResNet quantified the number of instruments in each detected stack.

2.1 Dataset

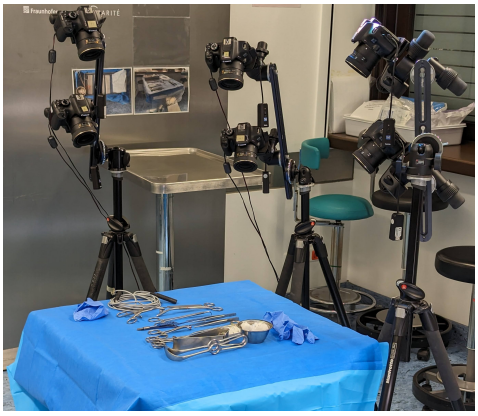

Fig. 2: Data acquisition setup. Data Acquisition at the Charite Hospital Center using six Canon 550D cameras capturing the scene on the operating stand from various perspectives.

To authentically recreate OR scenes of the instrument stand we mimic realistic instrument stand configurations in our OR lab. To achieve that, publicly available images of instrument stands were analyzed and experienced scrub nurses were consulted regarding best practices for instrument placement. For the instrument selection, we focused on ENT surgery and selected a variety of instruments commonly used across multiple surgical disciplines, drawing from the vast range of existing surgical tools. To enhance scene authenticity, we also included frequently encountered non-instrument items such as gauze balls, dissecting swabs, and suture material.

Capturing of high-resolution still images has been performed using six Canon EOS 550D cameras equipped with 35mm lenses, positioned at varying heights in a circular arrangement around the instrument stand to ensure diverse perspectives, as shown in Figure 2. We captured two datasets of the same instruments from different manufacturers (Table 1). The first dataset, encompasses 26 instruments from 16 distinct classes, manufactured by Aesculap, Germany. The second dataset features instruments manufactured by KLS Martin, Germany and Karl Storz, Germany. This second set represents a subset of the first, including 19 instruments across 12 different classes. All surgical instruments as well as gauze balls and dissecting swabs were annotated with precisely fitted bounding boxes.

2.2 Experiment

The experiment was designed to evaluate the generalization capabilities of both models with respect to surgical instrument sets from different manufacturers. Initially, both models were trained and tested on the first dataset with an 80%/ 20% split ratio. Subsequently, to assess their generalization ability, the models were applied to the secondary dataset, which featured instruments from different manufacturers. This experimental design enabled quantitative analysis of cross-domain performance and robustness to variations in instrument appearance.

Tab. 1: Number of images and annotated objects in our datasets.

Dataset	No. Images	No. Objects	Manufacturer
No. 1	846	8720	Aesculap
No. 2	106	2257	KLS Martin/Karl Storz

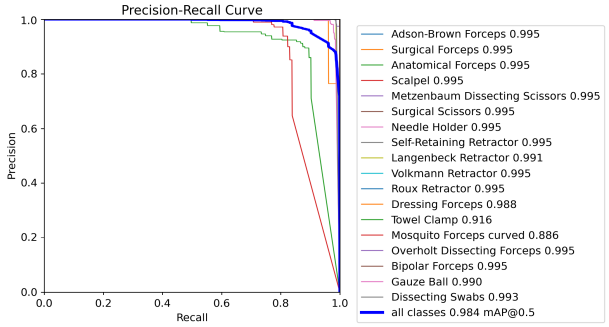

Fig. 3: PR-Curve of the standalone YOLOv8x model on the first dataset. The classes with the lowest AP are the towel clamps and the mosquito forceps.

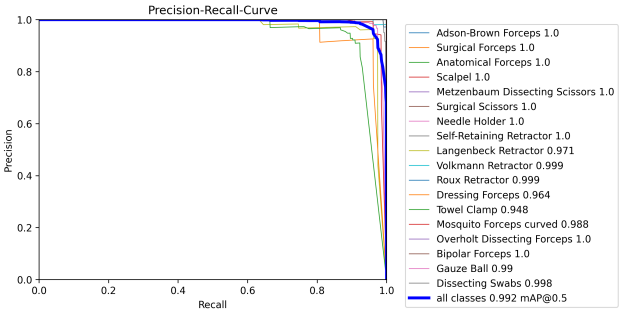

Fig. 4: PR-Curve of the detection pipeline on the first dataset. There is an increase in detection accuracy compare to the standalone YOLO model, particularly for typically stacked instruments like the towel clamps and mosquito forceps.

The standalone YOLOv8x network was trained at 960 pixels input image resolution, utilizing MSCOCO pre-trained weights. Training was conducted over 300 epochs with a batch size of 16 and an early stopping mechanism to prevent overfitting. We increased the Intersection over Union (IoU) threshold parameter of the NMS algorithm to 0.9, achieving an favorable balance between detection of instruments in stacks and suppression of duplicate detections.

To train the two-stage pipeline, the original dataset was modified by merging bounding boxes of adjacent instruments of the same class when they exceeded a specified IoU threshold of 0.3. Subsequently, they were assigned a new class label representing a stacked instrument configuration. First, the YOLOv8x model was trained on this modified data. For the second stage, the ResNet training set was generated by extracting images of instrument stacks from the same modified dataset. The ResNet-34, was trained using a cross-entropy loss function to quantify instrument count. The classifier head was trained for 10 epochs before fine-tuning the full network for 10 additional epochs.

3 Results and Discussion

The performance of the stand-alone YOLOv8x model, trained and evaluated on the first dataset, is shown in Figure 3 illustrating the precision-recall curve (PR-Curve). The model demonstrated robust performance, achieving a mAP50 (mean average precision) of 98.4%, with the AP (average precision) for most instrument classes exceeding 95%. These results are similar to literature [8].

However, it was observed that two specific classes, towel clamps and mosquito forceps, exhibited the lowest AP values among all classes, at 91.6% and 88.6% respectively. These lower precision rates are notable as these instruments are typically found in stacks on the instrument table (cf. Figure 1), a factor that complicates their individual detection.

Our proposed instrument detection pipeline can effectively address this shortcoming. Trained and evaluated on the first dataset, the pipeline achieved a mAP50 of 99.2%, outperforming the standalone YOLOv8x model as shown in Figure 4. While YOLOv8x showed limitations in reliably detecting stacked instruments, the integrated two-stage pipeline mitigated this issue. The increase in AP for stacked instruments, 3% for towel clamps and 10% for mosquito forceps, shows that the detection pipeline successfully overcomes this challenge, while still maintaining high AP for other instrument classes.

To evaluate detection transferability over different instrument manufacturers, both models were trained on the first dataset and tested on the second one, showing a significant decline in performance as presented in Figures 5 and 6. The mAP50 of the standalone YOLOv8x model dropped to 75.9%, a considerable decrease from the mAP50 of 98.4% achieved on the first dataset. Similarly, the performance of the two-stage pipeline dropped to a mAP of 78.4%.

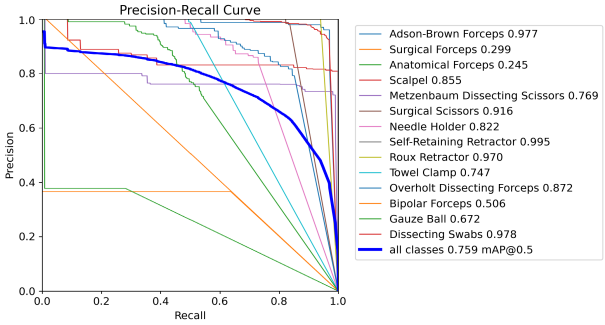

Fig. 5: Precision-Recall curve demonstrating the performance of our detection pipeline when tested on the second dataset, which includes medical instruments from manufacturers KLS Martin and Karls Storz.

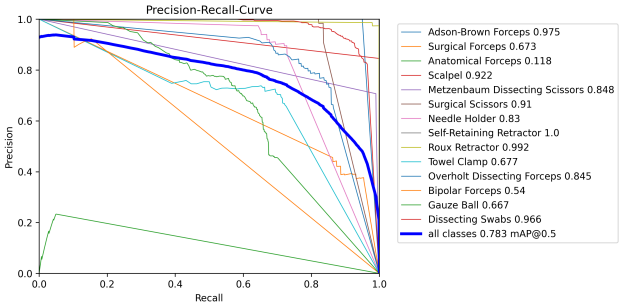

Fig. 6: Precision-Recall curve demonstrating the performance of our detection pipeline when tested on the second dataset, which includes medical instruments from manufacturers KLS Martin and Karls Storz.

For both models, certain instrument classes such as the Self-Retaining Retractor, Roux Retractors, Adson-Brown Forceps, and Scalpel maintained detection performance with an AP exceeding 90%. In contrast, instruments that exhibited significant variations in shape or color between manufacturers posed greater challenges for the detection models. The anatomical and surgical forceps were particularly affected. As depicted in Figure 7(A), the anatomical forceps from manufacturer Aesculap features a metallic pin at the middle of the forceps. This pin served as a robust feature for the models to distinguish between anatomical and surgical forceps. However, the anatomical forceps from manufacturer KLS Martin lacks this pin, leading to poor detection accuracy. Similarly, the morphology and color of the bipolar forceps differed significantly, as shown in Figure 7(B), resulting in the models inability to recognize them accurately.

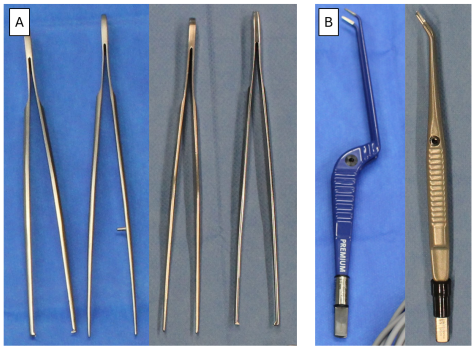

Fig. 7: Instruments across manufacturers. (A) Comparison of the Anatomical and Surgical Forceps from manufacturer Aesculap (left) and KLS Martin (right). (B) Bipolar forceps from Aesculap (left) and Karl Storz (right).

4 Conclusion

In this study, we introduce a novel dataset for surgical instrument detection on instrument stands, comprising images from realistic OR environments with instruments from multiple manufacturers. We evaluated both a state-of-the-art YOLO object detection model and a novel two-stage detection pipeline designed specifically to improve the detection of stacked surgical instruments. An experimental analysis was conducted to assess the generalization capabilities of both models when detecting instruments from different manufacturers. Results demonstrated that even minor variations in morphology or color between manufacturers significantly impacted detection accuracy. These findings underscore the necessity of a comprehensive dataset encompassing instruments from diverse manufacturers to develop robust surgical instrument detection systems.

Author Statement

The authors state no funding involved, no conflict of interest.

References

[1] Franke S, Meixensberger J, Neumuth T. Multi-perspective workflow modeling for online surgical situation models. J Biomed Inform. 54:158–66, 2015.

[2] Wisotzky EL, Rosenthal JC, Meij S, et al. Telepresence for surgical assistance and training using eXtended reality (during and after pandemic periods). J Telemed Telecare. 31(1):14-28, 2025.

[3] Dergachyova O, Bouget D, Huaulmé A, et al. Automatic data-driven real-time segmentation and recognition of surgical workflow. Int J CARS. 11(6):1081–9, 2016.

[4] Fujii R, Hachiuma R, Kajita H, Saito H. Surgical Tool Detection in Open Surgery Videos. Appl Sci. 12(20):10473, 2022.

[5] Mattingly AS, Chen MM, Divi V, et al. Minimally Invasive Surgery in the United States, 2022: Understanding Its Value Using New Datasets. J Surg Research. 281:33–6, 2023.

[6] Lee JD, Chien JC, Hsu YT, Wu CT. Automatic Surgical Instrument Recognition—A Case of Comparison Study between the Faster R-CNN, Mask R-CNN, and Single-Shot Multi-Box Detectors. Appl Sci. 11(17):8097, 2021.

[7] Badilla-Solórzano J, Spindeldreier S, Ihler S, Gellrich NC, Spalthoff S. Deep-learning-based instrument detection for intra-operative robotic assistance. Int J CARS. 17(9):1685–95, 2022.

[8] Bajraktari F, Fleissner K, Pott P. A comparison of two CNN-based instrument detection approaches for automated surgical assistance systems. Curr Dir Biomed Eng. 9(1):599-602, 2023.

[9] Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics. 2023