Automated Detection of Surgical Instruments

Presented to: Hazem Abbas

CISC/CMPE 452

Computer Engineering/Science

Queen's University

Prepared by: Team 61,

Iram Hasan, 20127302

Harminder-Singh Saini, 20207736

12/04/2024

Table of Contents

Table of Figures

Motivation

The motivation behind this research study on You Only Look Once (YOLO) models lies in evaluating and enhancing their effectiveness in automating the detection and tracking of surgical instruments—an essential aspect of modern surgical procedures. The YOLOv7 model, as highlighted in recent studies like Jiang et al. (2023), has demonstrated significant potential in instance segmentation and object detection for surgical tools. Additionally, its advancements in speed and accuracy have positioned it as an ideal candidate for automating surgical instrument detection.

However, existing literature highlights key challenges that hinder its validation in real-time surgical settings. These challenges include varying object perspectives, inconsistent lighting conditions within surgical environments, and the need for greater robustness in predictions. Building on the strengths of YOLOv7 and earlier versions, this study seeks to address these limitations through targeted improvements and modifications, aiming to develop a more robust and accurate YOLO model tailored for real-world surgical applications.

Problem Description

Accurate identification and tracking of surgical instruments are essential for ensuring patient safety and optimizing operational efficiency in medical procedures. Traditionally, this process is performed manually at least three times: before the procedure begins, prior to the closure of the initial wound layer, and at the procedure's conclusion. However, manual methods are prone to human error, which can compromise patient safety. Moreover, efficient inventory management of surgical instruments is crucial for minimizing delays, ensuring the availability of the correct tools, and streamlining hospital workflows.

This study aims to overcome the limitations of current YOLO algorithms, such as challenges with lighting conditions, instrument angles, and image quality, while expanding the scope of surgical instrument recognition to cover a broader range of tools and variations. The proposed approach involves developing a neural network model optimized for real-time performance in diverse surgical environments. This includes addressing challenges like variable lighting, different instrument orientations, and the need for fast, precise detection.

To achieve these goals, the study will enhance the dataset used for training by incorporating a wider range of instruments and their variants while improving data collection techniques to ensure robust detection under varied conditions. Furthermore, optimization strategies such as attention mechanisms and dynamic hyperparameter tuning will be explored to maximize the model's accuracy and efficiency, paving the way for improved automation in surgical settings.

Contribution

Both authors contributed to the research on YOLO models and their application in surgical tool detection, collaboratively drafting the initial proposal and providing the foundational background and motivation for this study. Iram played a key role in the initial stages of dataset research, proposing multiple datasets and identifying the one chosen for this paper. Harminder offered critical input and made the final dataset selection.

Iram initiated the coding process by developing a preliminary template for the implementation, which served as the foundation for further development. Building on this, Harminder created a comprehensive implementation of the YOLOv8 and YOLOv11 models. Prior to implementation, Iram evaluated the advantages of these models, particularly their relevance to the segmentation and detection of surgical tools. He also introduced modifications to the YOLO models to enhance their final performance, while Harminder conducted the model evaluation, generating a normalized confusion matrix for YOLOv11.

Iram performed the initial data analysis, identifying the study's strengths and limitations. Both authors contributed to interpreting the results, developing the discussion, and formulating the paper’s conclusions, ensuring a balanced and thorough collaboration throughout the project.

Related Work

Surgical Instrument Recognition Based on Improved YOLOv5

(K. Jiang, S. Pan, L. Yang, J. Yu, Y. Lin, and H. Wang, 2023)

This study focuses on an enhanced YOLOv5 model tailored for surgical instrument recognition, incorporating several modifications to improve its accuracy, efficiency, and resource utilization. The authors introduced the Squeeze-and-Excitation (SE) attention module into the model architecture, enabling dynamic reweighting of channel-wise feature maps to improve the extraction of key features. They also proposed a new loss function that leveraged global parameters for faster loss convergence, stabilizing the training process and accelerating model optimization.

Further, the traditional convolutions in the C3 module were replaced with a more efficient convolutional algorithm, which significantly reduced computational complexity and memory usage. This modification made the model faster and less resource-intensive without compromising accuracy. The dataset used for training was also augmented with images of eight representative surgical instruments, including scissors, haemostats, and speculums, to improve robustness and introduce variability into the training data.

The results demonstrated a precision of 88.7%, a marked improvement over the original YOLOv5, particularly in recognizing small and overlapping objects commonly encountered in surgical environments. The enhanced model also achieved faster convergence, reducing training time and computational costs, which is critical for scalability. However, limitations persisted, such as the lack of variability in imaging conditions (e.g., lighting and occlusions) and the absence of real-world testing. The model was not validated in practical surgical environments where instruments may be partially occluded, heavily contaminated, or arranged in complex configurations. Future work should focus on expanding the dataset to include more diverse imaging conditions and conducting real-world trials to evaluate its practical applicability

Surgical Instrument Detection Algorithm Based on Improved YOLOv7x

(B. Ran, B. Huang, S. Liang, and Y. Hou, 2023)

This study aimed to enhance the YOLOv7x model for surgical instrument detection by improving the backbone network and incorporating modules to expand its feature extraction capabilities. Building on similar objectives outlined by Jiang et al., the authors introduced the RepLK Block module into the YOLOv7x backbone. This addition expanded the effective receptive field, allowing the network to capture more comprehensive shape features critical for detecting complex surgical instruments.

In addition to backbone improvements, the authors integrated the ODConv structure into the neck module, alongside a spatial pyramid pooling (SPP) module and a path aggregation network (PAN). The PAN aggregated feature maps from multiple backbone layers, enriching the contextual information used by the detection head. The ODConv module, in particular, improved feature extraction efficiency, enabling the model to capture richer contextual details compared to standard convolution operations.

These enhancements led to significant performance improvements over the baseline YOLOv7 model. The modified YOLOv7x achieved an F1 Score of 94.7%, reflecting a 4.6% increase, and an average precision of 91.3%, marking a 3% improvement. The model demonstrated superior accuracy in detecting surgical instruments, even under challenging scenarios such as densely packed arrangements and occlusions.

Despite these advancements, the study faced limitations. The training dataset consisted of only 452 images, which significantly restricted the model’s generalizability and its ability to perform well under varied real-world conditions. Additionally, the model was not tested in diverse clinical environments with different instrument types, lighting conditions, or surgical scenarios. These shortcomings highlight the need for future work to address these gaps.

To further enhance the model’s robustness and practical applicability, the authors recommended expanding the dataset to include a broader variety of instruments and imaging conditions. Moreover, conducting trials in actual surgical environments would provide valuable insights into the model’s performance under real-world constraints, ultimately refining its design for clinical use. While the study presents promising improvements to YOLOv7x, the limited dataset and lack of real-world validation underscore the need for more comprehensive testing and optimization.

Dataset Comparisons and Limitations

Jiang et al. Study

The dataset in the Jiang et al. study consisted of 740 original images, specifically curated to create a comprehensive surgical instrument dataset. Among these, 540 images featured non-stacked surgical instruments, while 200 images showcased stacked instruments across eight instrument types. Each image was labeled, and the dataset was categorized into eight classes: haemostat, speculum, napkin tong, scissors, tweezers, colposcope, attractor, and stripper.

Despite its utility, this dataset faced a limitation in size, which increased the risk of overfitting when training deep learning models. To address this, the researchers used various image augmentation techniques to artificially expand the dataset and enhance generalizability. These techniques included panning, cropping, brightness adjustments, noise introduction, angle rotation, horizontal flipping, and the cutout method. These augmentations effectively improved the model's ability to generalize but did not completely mitigate the limitations posed by the dataset's relatively small size.

Ran et al. Study

The Ran et al. study utilized a smaller dataset of 452 images, featuring 26 representative surgical instruments from an orthopedic surgical instrument set. These instruments included surgical scissors, forceps, hooks, bone knives, bone forceps, periosteal elevators, nerve strippers, surgical clamps, surgical knife handles, aspirator tips, scrapers, and spreaders, among others. A unique aspect of this dataset was the focus on instruments with similar shapes to address challenges in distinguishing visually similar tools.

The dataset was organized into various arrangements:

- 156 images of single instruments.
- 72 images of multiple instruments without crossing occlusion.
- 224 images of multiple instruments with crossing occlusion.

To enhance variability, images were captured under different lighting conditions and standardized to a resolution of 1920 × 1080 pixels. Manual annotations ensured high-quality labels. Data augmentation techniques, including translation, scaling, flipping, brightness adjustment, and hue and saturation transformations, were employed. Advanced methods like mix-up and mosaic augmentation were also incorporated. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio, ensuring a balanced distribution across all categories.

Despite these efforts, the dataset's small size limited its generalizability, especially given the wide variety of tools it aimed to classify. While the study introduced a more comprehensive list of instruments and attempted to improve variability, the dataset's scale hindered its ability to robustly generalize across diverse surgical settings.

Our Dataset

In comparison, our dataset is significantly larger, consisting of 3,009 images across four classes: scalpel, straight dissection clamp, straight mayo scissors, and curved mayo scissors. Each image is carefully labeled to account for both object detection and occlusion status, ensuring precise annotations crucial for training YOLO models. The dataset's size provides a clear advantage over the smaller datasets used in the studies, offering more training data to enhance model performance and reduce the risk of overfitting.

However, the dataset also has limitations. While its focus on four surgical tools ensures high accuracy in detecting these specific objects, the limited diversity in object types may restrict the model's ability to generalize to unseen instruments.

The detailed annotations and clear imagery make this dataset ideal for training YOLO models, which excel in real-time object detection and segmentation tasks. The provision of precise bounding boxes and segmentation masks further supports effective training. However, the limited number of object classes and controlled imaging conditions highlight the need for future expansion. Introducing more diverse object types, imaging conditions, and real-world scenarios will be critical for improving the model's robustness and adaptability for practical deployment in surgical environments.

Implementation

For implementation, we first looked at the YOLOv8 model, weighing its advantages compared to the older models discussed in the research such as the YOLOv7x and the YOLOv5. Despite their modifications, we found the enhancements in YOLOv8's architecture to be promising both from a resource perspective and for improved model accuracy. The backbone and head modifications to the model to improve multi-scale detection, and the improved CSPNet structure offering better feature propagation and gradient flow showed great promise in our initial implementation, where the unmodified YOLOv8 model outperformed the modified YOLOv5 and YOLOv7 models in terms of accuracy using our dataset. Furthermore, we saw utility in the model scaling enhancements made in the YOLOv8 model, adjusting image size parameters to see how the model performed under different conditions. Finally, we implemented image augmentation to our training method to better augment our dataset and provide more validity to our results. The YOLOv8 showed significant improvements across training, test, and validation, even surpassing the YOLOv11 model in mAP. Despite these improved results, there is some skepticism towards the validity of the YOLOv8 and its generalizability to a real-world setting due to the small amount of data used to train this model.

Secondly, we noticed drastic improvement in training of the model and computational efficiency of YOLOv8 in contrast to YOLOv5 or YOLOv7. This can be attributed to the modifications made to how the model has an anchor-free mechanism to simplify the detection pipeline and reduce sensitivity to hyperparameter tuning for anchors.

While conducting this study we noticed the recent release of YOLOv11 which we felt showed more promise and would be interesting to investigate in our study as well. This model improves implementation of YOLO further by introducing C3K2 blocks and the SPFF module for efficient multi-scale feature extraction, alongside a C2PSA attention mechanism to enhance the detection of small and occluded objects. It outperforms YOLOv8 in speed and accuracy, with a notable 2% improvement in mean Average Precision (mAP) on the COCO dataset. YOLOv11 also extends its capabilities to pose estimation, oriented object detection, and other tasks, positioning itself as a versatile and robust solution for diverse computer vision applications.

After the initial implementation of our dataset, we looked through the research to try and make modifications to our model to obtain better performance. We saw a promising method was to perform image augmentation on top of our original dataset to better improve validity. We used geometric transformations such as random flipping, scaling, and rotation, as well as brightness, contrast, and saturation adjustments. Furthermore, we added gaussian blurs and noise to the data and introduced equalization techniques, specifically, contrast limited adaptive histogram (CLAHE). This led to improvements, particularly for the YOLOv8 model showing that the augmentation did have a signficant effect on model performance.

Results and Discussion

YOLOv8

The YOLOv8 model demonstrates strong overall performance in detecting surgical tools, achieving high true positive rates (TPRs) for most classes. Notably, the model achieves TPRs of 0.98 for the "Scalpel," 0.97 for the "Straight Dissection Clamp," and 0.96 for the "Curved Mayo Scissors," reflecting its ability to accurately classify these instruments. However, the "Straight Mayo Scissors" class shows a slightly lower TPR of 0.94, with 28% of its predictions misclassified as "Curved Mayo Scissors." Additionally, the "Straight Dissection Clamp" exhibits challenges, with 37% of predictions incorrectly classified as "Straight Mayo Scissors." These misclassifications highlight difficulties in distinguishing tools with similar shapes and features, especially under conditions of overlapping or limited dataset variability in lighting and angles.

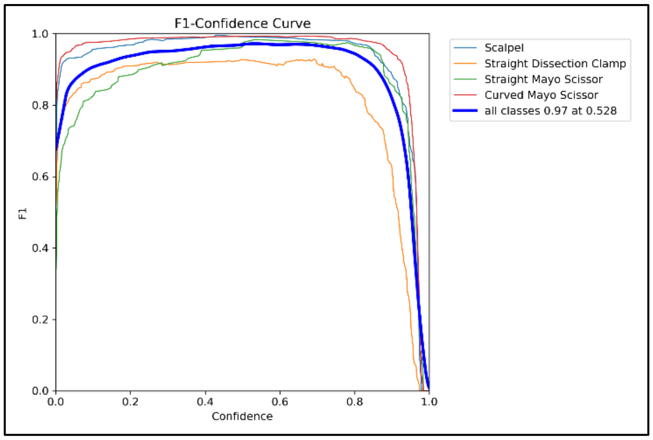


Figure 1: F1 Curve for testing YOLOv8

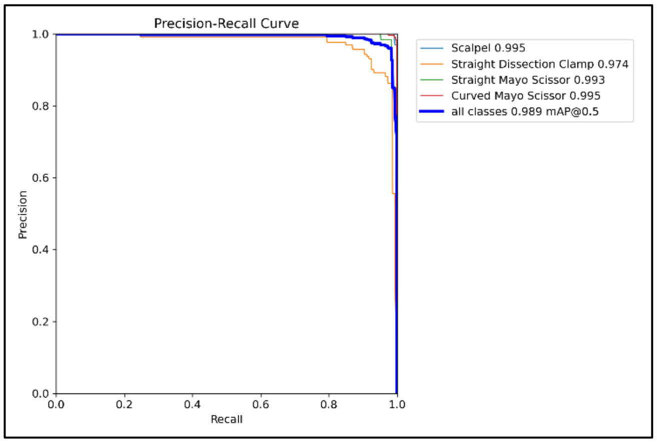


Figure 2: Precision-Recall Curve for testing YOLOv8

Error Analysis

Confusion between "Straight Mayo Scissors" and "Curved Mayo Scissors" arises primarily from their similar geometric profiles. Limited representation of diverse angles, lighting, and orientations in the dataset makes it challenging for the model to consistently recognize the curvature or straightness of these tools. Similarly, the overlap in features between "Straight Dissection Clamp" and "Straight Mayo Scissors" further exacerbates misclassification. On the other hand, the model shows excellent performance on the "background" class, with misclassification rates as low as 1-3%, demonstrating its reliability in distinguishing surgical tools from non-tool regions. However, the model's reliance on augmented patterns raises potential concerns about overfitting, as performance gains may stem from memorized augmented features rather than robust generalization.

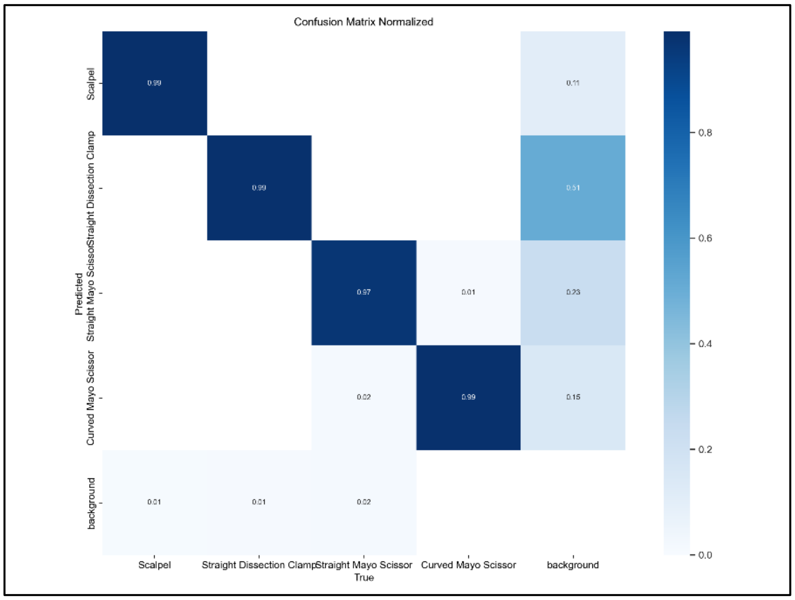


Figure 3: Normalized Confusion Matrix for testing YOLOv8

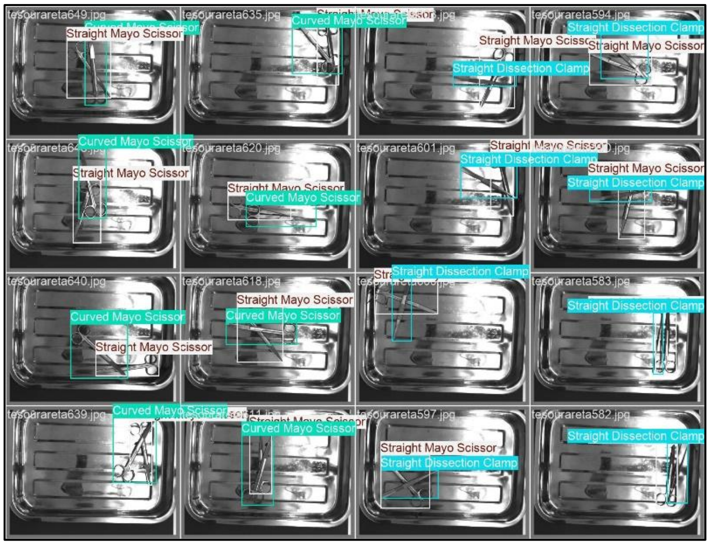


Figure 4: Model Labels for testing YOLOv8

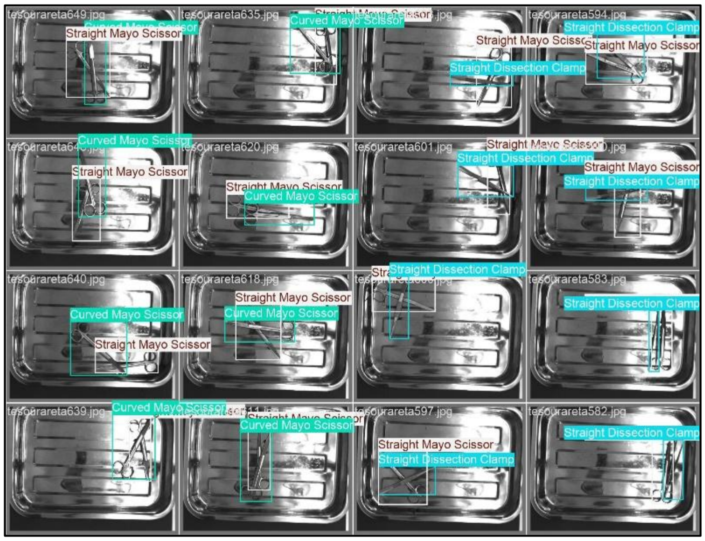


Figure 5: Model Predictions for testing YOLOv8

Recommendations

To address these challenges and improve performance:

1. **Feature Enhancement**: Emphasize distinctive geometric features like curvature, edges, and shape details to better differentiate similar tools.

2. **Data Augmentation**: Use rotations, varied angles, and lighting adjustments to improve the model's ability to generalize to real-world scenarios.

3. **Dataset Expansion**: Add diverse samples, including varied poses, occlusions, and challenging conditions, to enhance robustness.

The YOLOv8 model shows strong potential for surgical tool detection but requires targeted refinement to overcome challenges with inter-class confusion and overfitting. By emphasizing better feature representation, improving dataset diversity, and incorporating robust validation strategies, the model can achieve greater reliability and accuracy, making it more suitable for deployment in real-world surgical environments.

YOLOv11

The YOLOv11 model showed great performance in detecting surgical tools, with consistently strong metrics across training, validation, and testing phases. During training and validation, the model achieves a mean F1-score of 0.98, demonstrating a well-balanced relationship between precision and recall. In testing, the F1-score slightly decreases to 0.97, reflecting a minor gap in generalization. Among the tools, Scalpel and Curved Mayo Scissors show the highest performance, with F1-scores nearing 0.99 across all datasets. Straight Mayo Scissors follow closely, maintaining an average F1-score of 0.98. However, the Straight Dissection Clamp poses the greatest challenge, with its F1-score averaging 0.95 in testing, indicating consistent difficulties in accurate detection and classification.

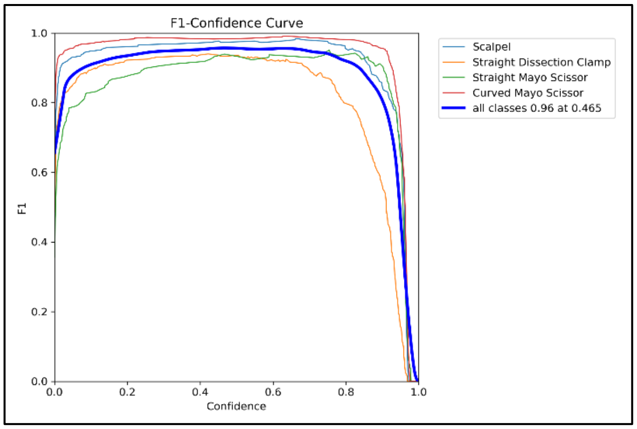


Figure 6: F1 Curve for testing YOLOv11

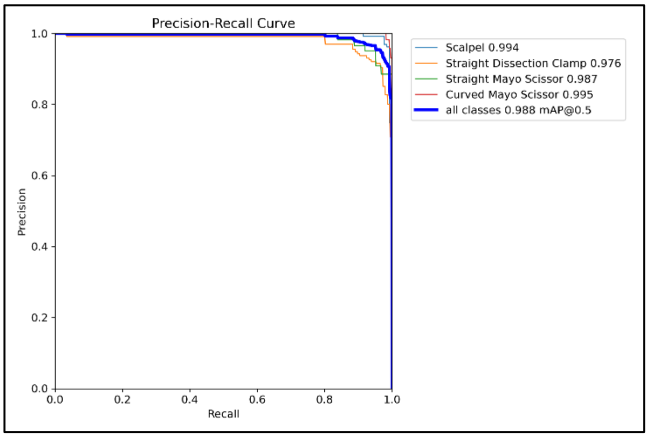


Figure 7: Precision-Recall Curve for testing YOLOv11

Error Analysis

Confusion matrices reveal class-specific trends in misclassification. Scalpel and Curved Mayo Scissors are rarely misclassified, benefiting from their distinct visual features such as unique shapes and sizes, which make them easily distinguishable from other classes. Conversely, the Straight Dissection Clamp is often confused with the Curved Mayo Scissors (~3-4% in testing) and the background (~5%), likely due to visual similarities, partial occlusions, or insufficient representation of edge cases in the training dataset. Misclassification of the background class as surgical tools is minimal, with error rates between 1-3% across datasets, underscoring the model's robustness in distinguishing tools from non-tool regions.

Precision and recall curves provide additional insight into the model's performance. Precision remains consistently high across all classes, approaching 1.0 even during testing, indicating very few false positives—a critical requirement for high-stakes applications such as surgical tool detection. However, recall is slightly lower for tools like the Straight Dissection Clamp, dropping to 0.95 in testing compared to 0.97 during validation. This suggests occasional false negatives, where some instances of the tool are not detected. In contrast, Scalpel and Curved Mayo Scissors maintain robust recall rates of approximately 0.99, indicating strong generalization for these classes.

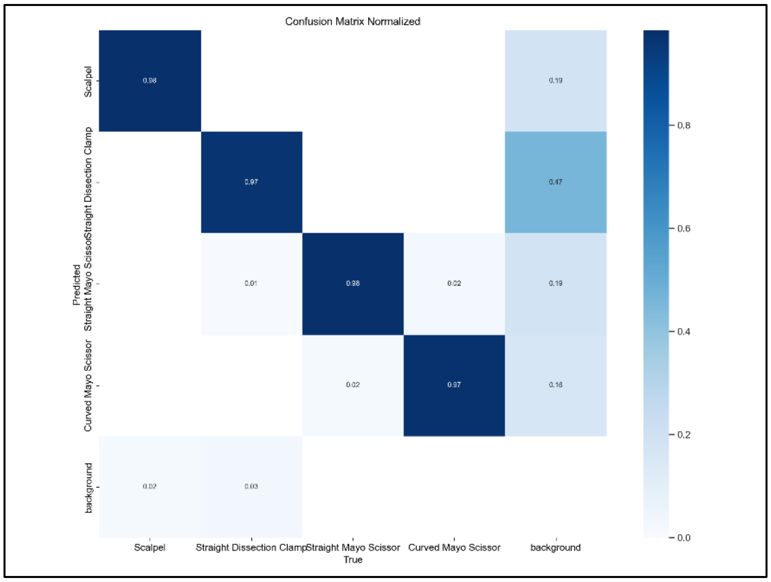
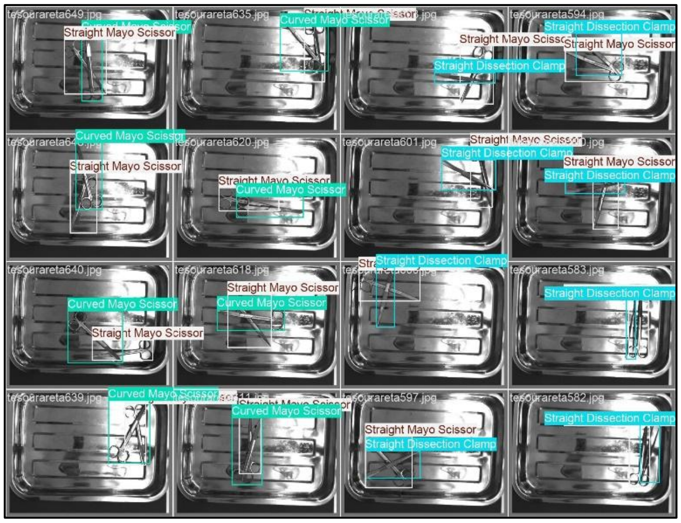


Figure 8: Normalized Confusion Matrix for testing YOLOv11

Challenges and Observations

Error trends reveal persistent challenges in distinguishing tools with overlapping shapes or similar visual characteristics. For instance, Straight Dissection Clamp and Curved Mayo Scissors are sometimes confused, especially in cases of occlusion or partial visibility of features. Tools occupying minimal space within the image frame, or those affected by poor lighting or shadows, are more susceptible to being misclassified as the background.



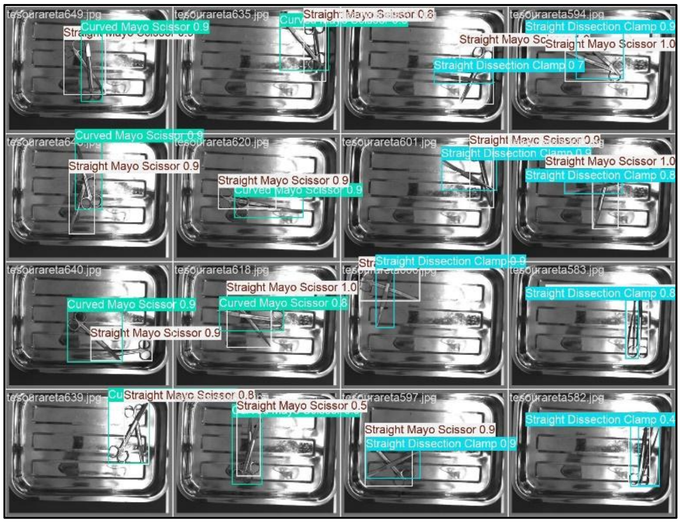

tesourareta649.jpg
Curved Mayo Scissor 0.9
tesourareta635.jpg
Straight Mayo Scissor 0.8
tesourareta594.jpg
Straight Dissection Clamp 0.9
Straight Mayo Scissor 1.0
Straight Dissection Clamp 0.7
Curved Mayo Scissor 0.9
tesourareta620.jpg
tesourareta601.jpg
Straight Mayo Scissor 0.9
Straight Mayo Scissor 0.9
Straight Mayo Scissor 1.0
Straight Dissection Clamp 0.8
Straight Mayo Scissor 0.9
Curved Mayo Scissor 0.9
tesourareta640.jpg
tesourareta618.jpg
Straight Dissection Clamp 0.9
tesourareta583.jpg
Straight Mayo Scissor 1.0
Curved Mayo Scissor 0.9
Curved Mayo Scissor 0.8
Straight Dissection Clamp 0.8
Straight Mayo Scissor 0.9
tesourareta639.jpg
Straight Mayo Scissor 0.8
Straight Mayo Scissor 0.5
tesourareta597.jpg
tesourareta582.jpg
Straight Mayo Scissor 0.9
Straight Dissection Clamp 0.9
Straight Dissection Clamp 0.4

Figure 9: Model Labels for testing YOLOv11

Figure 10: Model Predictions for testing YOLOv11

Recommendations for Improvement

To address these challenges and further enhance performance, several refinements are proposed:

1. **Feature Extraction Enhancements**:
   Fine-tuning feature extraction layers to emphasize subtle differences between visually similar tools could improve classification accuracy. Incorporating orientation-sensitive features and emphasizing edge details, such as curvature or shape variations, would help distinguish tools like the Straight Dissection Clamp from others.
2. **Confidence Threshold and NMS Refinements**:
   Adjusting the confidence thresholds and refining the Non-Maximum Suppression (NMS) algorithm could help improve recall for challenging tools like the Straight Dissection Clamp without compromising overall precision.
3. **Augmented Training Data**:
   Expanding the training dataset to include more diverse cases, such as occluded tools, varied lighting conditions, and additional angles, could improve the model's robustness. Synthetic data generation methods could also be used to supplement underrepresented scenarios.

These enhancements aim to address the limitations identified in the error analysis, enabling the model to better handle edge cases and improve performance in real-world surgical settings. By implementing these recommendations, the YOLOv11 model can be refined to deliver even greater reliability and accuracy, making it well-suited for deployment in critical medical applications.

Conclusion

In conclusion, the YOLOv11 model demonstrates exceptional performance in detecting surgical tools, achieving impressive mean Average Precision (mAP) scores of 0.991 on validation data and 0.988 on test data, alongside strong precision across all classes. The YOLOv8 model also performs competitively; however, its performance shows potential signs of overfitting due to limited training data and the image augmentation techniques applied. Among the instruments evaluated, the Scalpel and Curved Mayo Scissors are detected with near-perfect accuracy, while the Straight Mayo Scissors also deliver consistently strong results. The primary challenge lies with the Straight Dissection Clamp, which requires improved recall from similar tools.

To address these challenges and further improve the model, targeted refinements in data augmentation, feature engineering, and post-processing will be critical. Such enhancements could boost the model's performance and robustness, making it even more suitable for deployment in surgical settings, where accuracy and reliability are paramount.

However, this study does not yet address several key concerns related to real-world validation, largely due to resource constraints. Future work should focus on replicating the study with datasets containing contaminated surgical tools and diverse imaging conditions to evaluate model robustness. Additionally, real-time testing in live surgical environments will be essential to accurately assess the model's effectiveness in practical applications and ensure its readiness for clinical deployment.

References

[1] K. Jiang, S. Pan, L. Yang, J. Yu, Y. Lin, and H. Wang, "Surgical Instrument Recognition Based on Improved YOLOv5," *Applied Sciences*, vol. 13, no. 21, pp. 1–17, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/21/11709

[2] B. Ran, B. Huang, S. Liang, and Y. Hou, "Surgical Instrument Detection Algorithm Based on Improved YOLOv7x," *Sensors*, vol. 23, no. 11, pp. 1–18, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5037