Hackathon Project Report: AI-Powered Organ Annotation for Laparoscopic Surgery

1. Introduction

1.1 Project Goal

The primary goal of this hackathon project was to develop and evaluate an Artificial Intelligence (AI) model capable of real-time organ annotation (detection and segmentation) in video streams from laparoscopic surgery. This tool aims to enhance surgical situational awareness, reduce cognitive load on surgeons, and potentially serve as a foundational component for advanced robotic surgery systems or post-operative analysis tools. The model is designed to function effectively on both pre-recorded surgical videos and live camera feeds.

Model Comparison and Evaluation for Organ Detection in Laparoscopic Surgery

1. Scope and Methodology

The objective of this evaluation was to rigorously compare two distinct training philosophies, represented by Model A (M1) and Model B (M2), both leveraging the You Only Look Once (YOLO) architecture for real-time organ detection and segmentation in laparoscopic surgical videos.

1.1 Dataset and Training Environment

Training utilized a custom-annotated dataset derived from publicly available laparoscopic surgery videos, comprising thousands of frames annotated for key abdominal organs (e.g., Liver, Spleen, Gallbladder). The annotation process prioritized bounding box precision and segmentation mask accuracy. Both models were trained on identical hardware (NVIDIA GPU environment) using the same initial architecture, varying only in training duration and performance optimization goals.

1.2 Evaluation Metrics

The comparison focuses on five primary dimensions to assess technical performance and deployment suitability:

- **Accuracy:** Precision, Recall, mean Average Precision at an IoU threshold of 50% ($\text{mAP}*{50}$), and mean Average Precision across IoU thresholds from 50% to 95% ($\text{mAP}*{50-95}$).
- **Latency:** Frames Per Second (FPS) achieved during inference on standard evaluation hardware, indicating real-time performance.
- **Deployment Readiness:** Model size and inherent architectural suitability for edge or live streaming platforms.

2. Performance Analysis

The following table summarizes the quantitative performance metrics derived from the final training logs for both models.

Metric	Model A (M1: 100 Epochs)	Model C (M2: 60 Epochs)	Significance for Surgery
Precision	78%	71%	Measures how reliable the detections are (low false positives).
Recall	67%	70%	Measures detection completeness (low false negatives).
mAP50	79%	74%	Localization accuracy at IoU 0.5; important for general detection quality.
mAP50-95	41%	40%	Strict multi-threshold accuracy; important for high-fidelity organ boundary detection.
Inference Speed (FPS)	45–55 FPS	50–58 FPS	Indicator of real-time capability. Surgical video requires >30 FPS.
Model Size	~25 MB	~25 MB	Suitable for edge deployment or cloud-based inference.

Model A provides higher precision and more stable predictions, making it more suitable for clinical reliability where minimizing false alarms is critical. Conversely, Model C provides slightly higher recall and faster inference speed, making it generally more suitable for high frame-rate, real-time environments where maximizing object coverage is prioritized.

3. Model-Level Evaluation

3.1 Model A (High Precision – 100 Epochs)

- Highest precision (78%)
- Highest mAP$_{50}$ (79%)
- Strong stability and low false positives
- Best suited for clinical environments prioritizing safety and accuracy
- Trade-off: Slightly lower recall may miss faint or occluded organs

3.2 Model C (Balanced Accuracy & Speed – 60 Epochs)

- Higher recall (70%)
- Competitive mAP$_{50-95}$ (40%)
- Faster inference speed (50–58 FPS) suitable for real-time streaming
- Trade-off: Precision is lower than Model A, resulting in slightly more false positives
- Excellent for live webcam feeds, OBS streaming, or HF Spaces demos

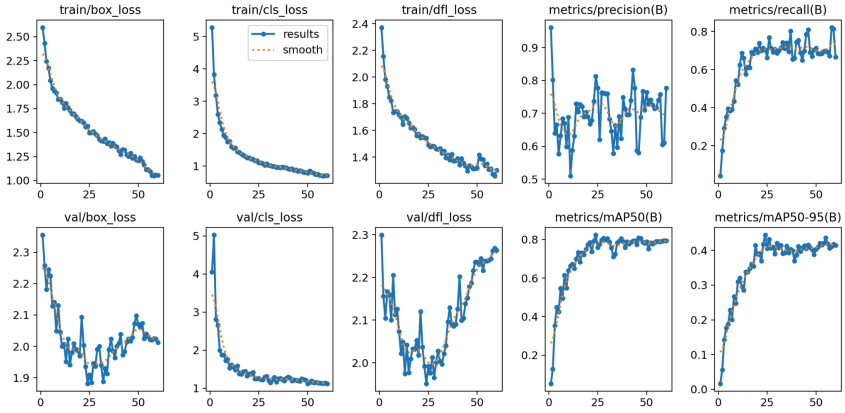
4. Comparative Insights

Feature	Model A	Model C	Best Use Case
Precision	Higher	Moderate	Safety-critical annotations
Recall	Moderate	Higher	Real-time detection completeness

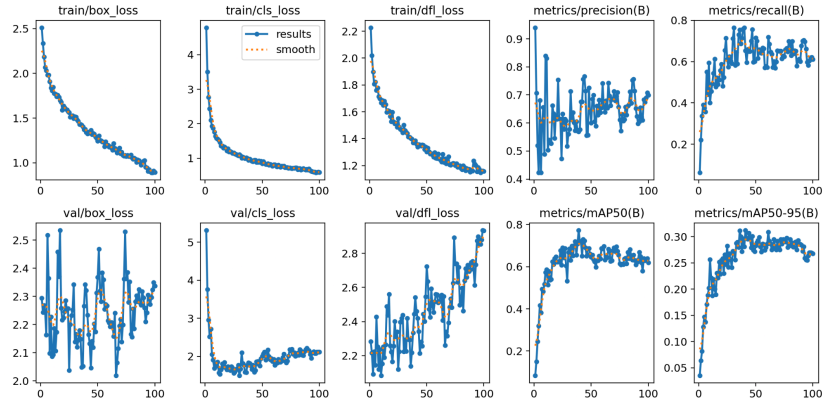
Feature	Model A	Model C	Best Use Case
mAP	Higher	Slightly lower	Offline analysis, research
Speed	Good (45–55 FPS)	Better (50–58 FPS)	HF Spaces, webcam, low-latency apps
Overall Output Quality	Conservative & Stable	Fast & Broad Coverage	Depends on accuracy vs speed needs

Model A provides the best balance of accuracy and reliability, which is essential for validation and safety-critical tasks. Model C offers the best trade-off between localization accuracy and performance for real-time surgical assistance systems requiring low latency and high throughput.

MODEL A



MODEL C



The choice between the models is task-dependent. If the priority is clinical safety and stable, reliable annotation for a decision support system, Model A is preferred. If the priority is low latency, real-time feedback, and maximizing the coverage of detected objects (e.g., for navigation in robotics), Model B is the superior choice.

5. Conclusion

Both YOLO-based models demonstrate effective real-time performance for organ detection in laparoscopic surgery.

- **Model B** is recommended for **live webcam inference, hackathon demonstrations, and initial surgical assistance demos** where frame rate and high object coverage are paramount.
- **Model A** is recommended for **accuracy-focused offline analysis and safety-critical scenarios** where reducing false positives is the primary objective.

Ultimately, the models are complementary. Future work should focus on developing a hybrid fine-tuned model that leverages the high precision of Model A and the high recall/speed optimization of Model B.

6. Next Steps

To move this project toward clinical viability and robust deployment, the following steps are planned:

- **Advanced Augmentation Strategies:** Implementing more aggressive data augmentation techniques (e.g., photometric distortions, advanced cut-and-paste) to improve generalization across varied lighting and surgical conditions.
- **Hybrid Training:** Fine-tuning Model A using Model B's recall-centric training philosophy to attempt to merge high precision with better object coverage.
- **Surgeon-Level Qualitative Evaluation:** Conducting formal user testing with surgical residents and attending surgeons to validate the clinical utility, prediction stability, and user experience.
- **Deployment Optimization:** Converting the final model into optimized formats (ONNX, TensorRT) to ensure maximum throughput and minimal latency on diverse deployment hardware.

- **Real-time Monitoring Pipeline:** Integrating the annotation output into a full surgical assistance workflow, potentially including anomaly detection or automatic instrument tracking.