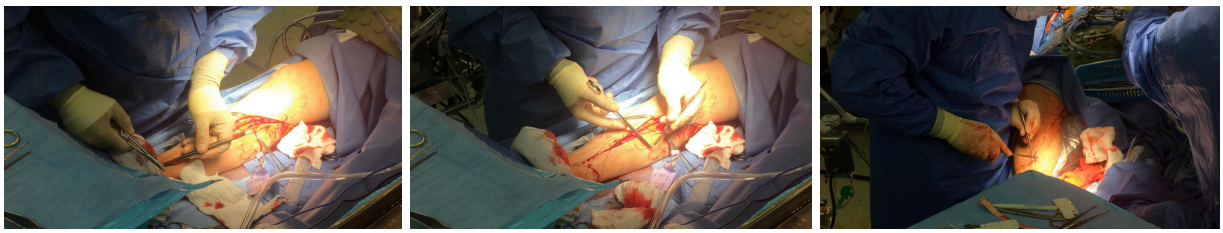
1. Exploratory Data Analysis

1.1 Dataset Overview and Visualization

The dataset includes 71 labeled pictures from leg surgery procedures, with 61 used for training and 10 for testing. Each image shows boxes drawn around hands and surgical tools, such as tweezers and needle drivers, including cases where the hands are empty. The dataset also contains three unlabeled surgery videos - two that are similar to the labeled pictures (ID), and one that is different because it was filmed on a different day with separate camera equipment (OOD). Visualization of images from the in-distribution data:

Visualization of images from the out-of-distribution video:



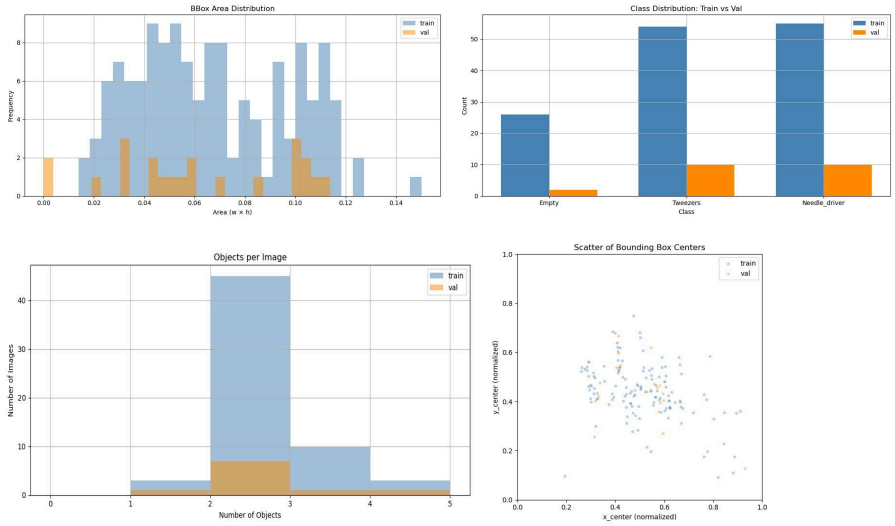
1.2 Insights from Visual Inspection

The surgical dataset captures collaborative operating room environments where multiple surgeons work together on leg procedures using needle drivers and tweezers. The imagery is characterized by focused lighting that illuminates the surgical field, enhancing tool and hand visibility for computer vision applications, though shadows can sometimes obscure hands outside the direct light. The consistent visual elements—including blue surgical uniforms, white gloves, sterile drapes, and controlled lighting—create a standardized aesthetic that reflects the precision of surgical procedures. However, this uniformity across both in-distribution and out-of-distribution data may lead to model overfitting on color cues, potentially limiting the system's ability to generalize to surgical environments with different attire or lighting conditions.

1.3 Data Distribution Analysis

The dataset demonstrates proportionally balanced class distributions between training and validation splits, though the 'Empty' category consistently appears with the lowest frequency across both subsets. The annotated portion of the dataset is notably constrained, comprising only 71 manually labeled images extracted from in-domain surgical video sequences—61 designated for training and 10 for validation, which inherently limits the diversity of surgical scenarios and camera perspectives available for model

development. Complementing this labeled data are three video sequences that provide valuable temporal context: two in-domain videos that share the same procedural and environmental characteristics as the annotated frames, and one out-of-domain video captured on a different day with an altered camera configuration. While these video sequences offer the advantage of temporal consistency and continuity for understanding surgical workflows, they lack the ground truth annotations necessary for supervised learning, creating a gap between available visual data and training-ready labeled examples. Plots for the distribution:



2. Experimental Methodology: Twelve-Stage Progressive Training Approach

The study employed YOLO11 Nano (yolo11n.pt) as the primary architecture, selected for its optimal balance between detection accuracy and computational efficiency in detecting small surgical tools and hands for real-time applications. The implementation relied on the Ultralytics library with standard loss components including box, classification, and distribution focal loss (DFL). The experimental pipeline was organized into twelve progressive training stages, systematically leveraging both in-distribution (ID) and out-of-distribution (OOD) video data through iterative semi-supervised learning with rigorous confidence-based filtering.

Stage 1: Baseline Model Establishment

The initial phase established a performance baseline using the original dataset of 61 training images, trained for 70 epochs with comprehensive data augmentation. The augmentation parameters were carefully selected to preserve surgical context integrity: minimal hue variation (hsv_h=0.015), moderate saturation changes (hsv_s=0.7), brightness adjustments (hsv_v=0.4), small rotation angles (degrees=5.0), minor position shifts (translate=0.05), scale variations (scale=0.5), shearing transformations (shear=2.0), no horizontal flipping (fliplr=0.0) to maintain anatomical correctness, full mosaic augmentation (mosaic=1.0), and minimal MixUp augmentation (mixup=0.1). This baseline model served as the foundation for all subsequent refinement stages.

Stages 2-6: Iterative In-Distribution Semi-Supervised Refinement

Five consecutive refinement iterations were performed using in-distribution video data through a systematic four-step process. First, pseudo-label generation extracted frames from ID videos using the current model with an initial confidence threshold of 0.5 during inference to capture comprehensive detections. Second, high-confidence filtering applied a stringent threshold of 0.85, retaining only frames with reliable detections while deleting low-quality predictions to maintain dataset integrity. Third, the dataset combination merged filtered pseudo-labeled frames with the original 61 training images, creating expanded datasets with corresponding YAML configuration files. Fourth, model refinement trained each iteration for 10 epochs using the same augmentation parameters, with the refined model serving as input for the next iteration. This process was repeated five times, progressively refining the training dataset with high-quality ID pseudo-labeled data while maintaining strict quality controls.

Stages 7-11: Iterative Out-of-Distribution Semi-Supervised Refinement

Following ID refinement, five additional iterations targeted out-of-distribution video data to enhance model robustness across diverse surgical environments. The process mirrored ID refinement while addressing challenges from different camera setups and imaging conditions. OOD pseudo-label generation applied to the ID-refined model to OOD video frames, testing initial generalization capabilities. High-confidence filtering maintained the same 0.85 threshold to ensure consistent quality standards across domains. Dataset combination merged filtered OOD pseudo-labels with original training data, with each iteration beginning fresh to prevent error accumulation while maintaining foundational knowledge. Model refinement used 10-epoch training sessions with consistent parameters, systematically exposing the model to varied OOD scenarios while preserving core surgical detection capabilities.

Stage 12: Final Model Evaluation

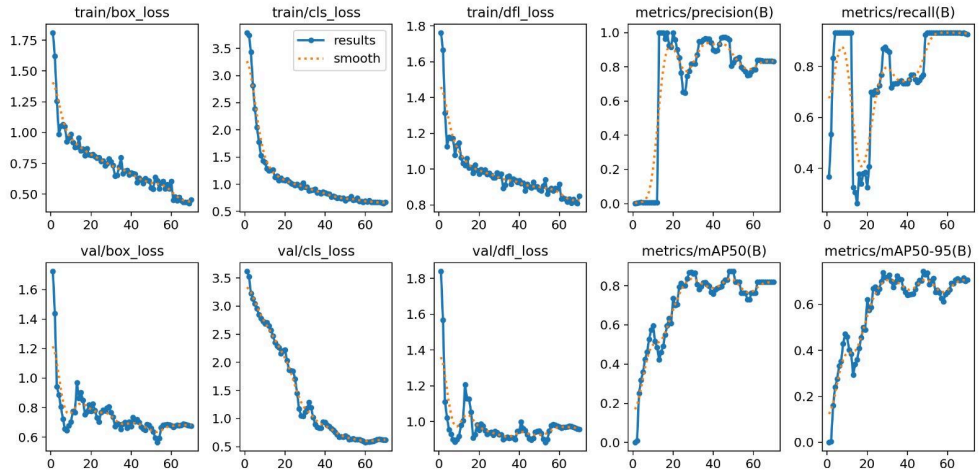
The final refined model (yolo11n_ood_refined.pt) underwent a comprehensive evaluation on both ID and OOD video sequences to assess whether the final model could generalize better across diverse surgical environments. While it helped suppress some false positives, the performance gains in terms of accuracy and mAP were modest

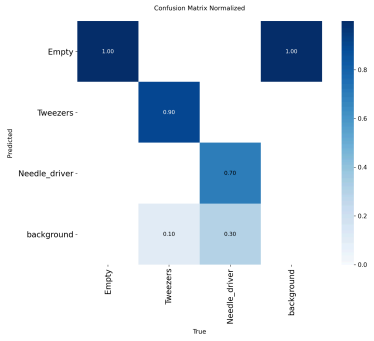
Training Configuration Summary and Progressive Dataset Expansion

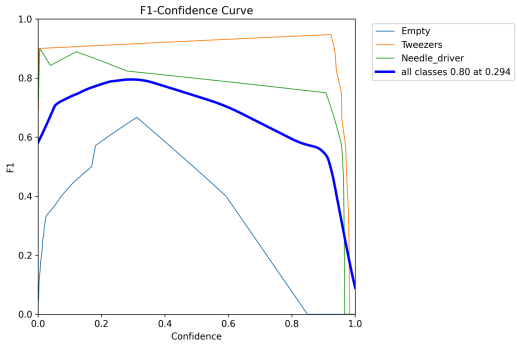
The methodology encompassed 170 total epochs: 70 for baseline establishment, 50 across five ID refinement iterations, and 50 across five OOD refinement iterations. The confidence threshold of 0.85 for pseudo-label filtering was maintained consistently throughout all stages. Training utilized AdamW optimizer with adaptive learning rate scheduling. The dataset expansion progressed through three phases: Stage 1 used 61 original labeled images, Stages 2-6 progressively accumulated ID pseudo-labeled frames with the original dataset, and Stages 7-11 independently combined original data with OOD pseudo-labeled frames for each iteration. This twelve-stage progressive approach enabled comprehensive utilization of labeled and unlabeled video data while maintaining training stability through conservative epoch allocation and rigorous filtering, producing a robust surgical tool detection system capable of performing effectively across diverse surgical environments.

Statistics

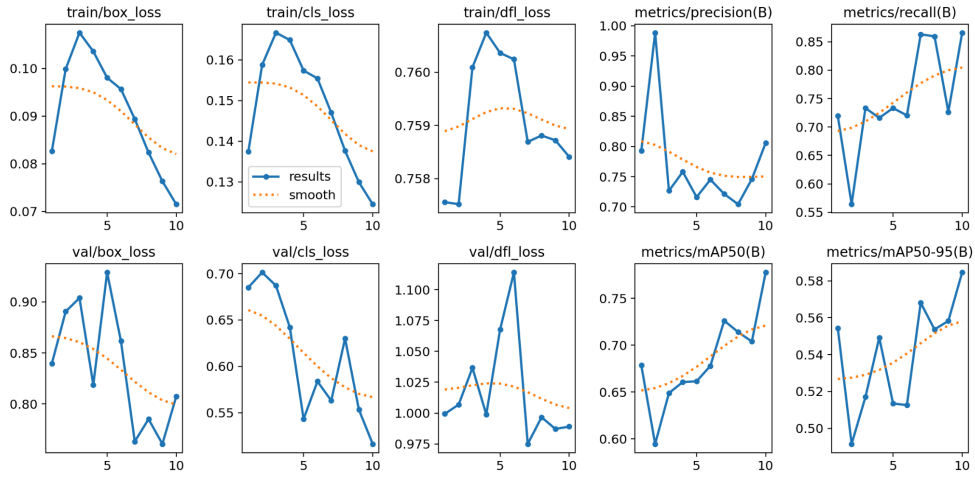
In-distribution(ID):

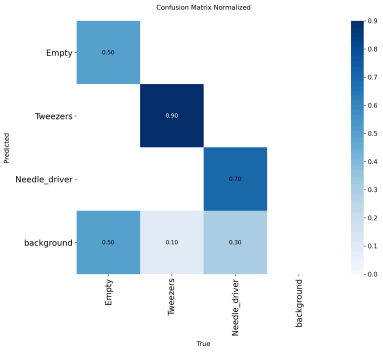


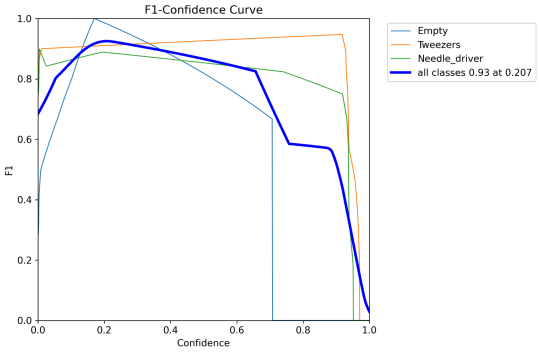




Out of distribution(OOD):







3. Conclusion

The progressive four-stage training methodology successfully demonstrated the effectiveness of semi-supervised learning for surgical instrument detection under severely constrained labeled data. Starting with only 60 annotated images, the systematic pipeline—combining data augmentation, pseudo-labeling, and iterative refinement—achieved strong performance, exceeding 90% mAP@50 for hands and surgical tools.

Experimental results show consistent performance improvements across stages, with the largest gains occurring during in-distribution pseudo-labeling. Incorporating out-of-distribution data helped reduce false positives, particularly on inactive instruments, but did not substantially improve overall detection accuracy. This outcome highlights the challenges of generalizing to new visual domains using noisy pseudo-labels alone.

The final model remains sensitive to camera setup variations and benefits primarily from increased label diversity rather than direct accuracy gains. Future work could explore manual curation of a small OOD-labeled validation set, enhanced pseudo-label filtering strategies, or the use of larger architectures to improve generalization under domain shift.

The submitted video was generated using the ID-refined model, which exhibited more stable and reliable predictions compared to the OOD-refined model.