Semi-Supervised Surgical Tool Detection via Curated Pseudo-Labeling

DeAndre Bailey
Center for Research in Computer Vision
University of Central Florida
Orlando, FL, USA

Darinka Townsend Hinostroza, Yuanhao Zou, Chen Chen
Center for Research in Computer Vision
University of Central Florida
Orlando, FL, USA

***Abstract*—Automated surgical tool detection is a prerequisite for workflow analysis, skill assessment, and operating-room safety monitoring in robotic-assisted surgery. However, exhaustively annotating bounding boxes for every tool in every frame is prohibitively time-consuming, and available data is weakly labeled, noisy, and severely class-imbalanced. We address the SurgVU26 MICCAI EndoVis Challenge (Category 1), which requires detecting and localizing surgical tool classes across 155 largely unlabeled video sessions. We present a semi-supervised pipeline that bootstraps a strong teacher detector from a small, diversity-sampled set of manual annotations, generates pseudo-labels across all sessions, and curates them through three complementary mechanisms—per-class confidence floors, per-session caps, and capped-proportional class quotas—before combining them with ground-truth anchors into a hybrid training set. Retraining RF-DETR on this curated dataset improves performance from 0.5082 to 0.7979 mAP@50:95, and a subsequent pseudo-label-ratio reduction combined with increased weight decay raises it further to 0.8116. We additionally show, via a localization-versus-classification analysis, that the model's residual weaknesses are concentrated in a small number of classes and are dominated by localization rather than classification failures. Our results indicate that *how* pseudo-labels are curated—not merely how many are added—drives the performance gain.**

***Index Terms*—surgical tool detection, semi-supervised learning, pseudo-labeling, object detection, robotic surgery**

I. Introduction

Robotic-assisted surgery using systems such as the da Vinci platform is now routine, and reliable perception of surgical tools underpins downstream applications including workflow analysis, skill assessment, and safety monitoring. Detecting and localizing tools automatically requires large quantities of annotated video, yet manual bounding-box annotation of every frame does not scale to the volume of data these systems generate.

The SurgVU26 MICCAI EndoVis Challenge (Category 1) poses this problem directly: detect and localize surgical tool classes within video sessions labeled only weakly and noisily [1]. The challenge provides 155 video sessions, the vast majority of which lack frame-level bounding-box annotations, alongside severe class imbalance in what labels do exist.

Our objective is to build a complete detection pipeline that closes this annotation gap without relying on large-scale manual labeling. We bootstrap a strong teacher model by manually annotating a small, diversity-sampled set of frames targeting the classes with the least data, use that teacher to generate pseudo-labels across all sessions, and—critically—curate those pseudo-labels rather than using them wholesale. Our contributions are:

- A semi-supervised, teacher–student detection pipeline for surgical tool detection that requires only a small manual-annotation budget.
- A three-mechanism pseudo-label curation strategy (noise reduction, class balance, session diversity) that we show drives the performance gain beyond simply adding data.
- An empirical analysis isolating localization from classification error, showing residual weaknesses are localization-dominated and concentrated in specific classes.

II. Related Work

Semi-supervised object detection and pseudo-labeling. Teacher–student pseudo-labeling generates supervision from a model's own confident predictions on unlabeled data. Our approach follows this paradigm but emphasizes curation of the resulting labels.

DETR-family detectors. We build primarily on RF-DETR, a neural-architecture-search-derived real-time detection transformer [3], and compare against D-FINE, which reformulates bounding-box regression as fine-grained distribution refinement [5], and a YOLOv8 baseline.

Active learning for annotation efficiency. To allocate a limited manual-annotation budget effectively, we select frames using K-Center-Greedy diversity sampling, a core-set approach to active learning [4]. An alternative feature-based selection strategy, used by team PDMYR in the 2025 SurgToolLoc/SurgVU challenge, extracts per-frame features with a pre-trained backbone (ImageNet or CLIP), runs K-Means clustering (1,000 clusters) over the full 1 fps-sampled dataset, and selects the frame nearest each cluster centroid for labeling [2]. After this initial round, PDMYR identified class imbalance and visually challenging conditions (fog, bleeding, lens contamination, exposure errors) as remaining gaps, and addressed them via human-in-the-loop mining of the sequences with the poorest model predictions and augmentations targeting the observed defect types [2]. Notably, PDMYR also reported that the challenge's official instrument-presence labels were unreliable, and instead extracted ground-truth instrument identity via OCR on the da Vinci system's on-

screen GUI—an approach that parallels our own OCR-based recovery of tool-name mappings from the robot UI.

Surgical tool detection benchmarks. The SurgVU dataset [1] and prior SurgToolLoc/SurgVU challenge results [2] establish the task and evaluation setting we address.

III. Problem Setup

The task presents three intertwined difficulties. First, labels are *weak and noisy*: across 155 sessions, most frames are unlabeled, and the available supervision is imperfect. Second, the data is *severely class-imbalanced*—needle driver appears in 3,583 confirmed boxes while suction irrigator appears in only 22, a roughly 160× gap—so rare tools are nearly absent from the raw data. Third, *manual annotation of every frame is infeasible* at this scale. These conditions motivate a semi-supervised approach in which a small, carefully chosen manual budget seeds an automated labeling process, with curation used to counteract the noise and imbalance that naive pseudo-labeling would otherwise amplify.

IV. Method

A. Teacher Bootstrapping

To build a strong first teacher, we prioritized manual annotation of the tool classes with the largest data gaps. We selected frames for annotation using K-Center-Greedy diversity sampling [4], an active-learning strategy that chooses a representative, non-redundant subset rather than annotating arbitrary or consecutive frames. This yielded a compact manual-annotation set concentrated on underrepresented classes. Combined with the ground-truth split of the provided validation data, these annotations trained an RF-DETR teacher model achieving 0.5082 mAP@50:95.

B. Pseudo-Label Generation

The fine-tuned teacher was applied across all 155 video sessions to produce candidate pseudo-labels. Used directly, these candidates would inherit the teacher's systematic errors and reproduce the dataset's extreme class imbalance—so we curate them before use.

C. Pseudo-Label Curation

Three mechanisms transform raw teacher predictions into trustworthy training data:

- **Noise reduction (per-class confidence floors).** Each pseudo-label must exceed a class-specific confidence threshold. Most classes use a default floor of 0.35; underrepresented classes use lower floors (e.g., 0.15–0.20) to preserve scarce-class signal that a uniform threshold would discard.
- **Session diversity (per-session cap).** No single recording session may contribute more than 15% of a class's selected boxes, preventing overfitting to the visual characteristics of any one procedure.
- **Class balance (capped-proportional quotas).** A per-class quota (2,800 boxes) prevents abundant classes from swamping the training set relative to naturally rarer tools.

These filters yield 26,918 curated pseudo-label frames.

TABLE I
Main Results on the SurgVU Validation Set

Model	Phase / Dataset	mAP@50:95
YOLOv8l	SurgVU validation split	0.272
RF-DETR-L	SurgVU validation split	0.352
RF-DETR-L (teacher)	Manual annot. + val split	0.5082
RF-DETR-L	Curated hybrid dataset	0.7979
RF-DETR-L (best)	Hybrid + 3:1 ratio + WD	**0.8116**

D. Hybrid Dataset and Student Training

The curated pseudo-labels are combined with ground-truth anchor frames to form a hybrid training set of 31,041 frames (42,605 training boxes; 3,459 validation boxes). We retrain RF-DETR (the student) on this hybrid set. To address overfitting observed during training (Section V-C), we additionally reduce the pseudo-label-to-ground-truth ratio from approximately 6.5:1 to 3:1 and increase weight decay from 1×10^{-4} to 5×10^{-4}.

V. Experiments and Results

A. Main Results

Retraining on the curated hybrid dataset improves RF-DETR substantially over the teacher, and the ratio/weight-decay refinements yield a further gain (Table I).

Training on the curated hybrid dataset alone raises RF-DETR to 0.7979 mAP@50:95; the ratio reduction and weight-decay increase together bring the best model to 0.8116.

B. Per-Class Performance

The best model's per-class AP@50:95 ranges from 0.989 (grasping retractor) down to 0.587 (tip-up fenestrated grasper), with prograsp forceps (0.643) as the other principal weak class. The majority of classes exceed 0.80.

C. Overfitting Diagnosis

Examining the per-epoch training curve on the curated dataset revealed that training loss continued to fall by roughly 35% after the validation-mAP peak at epoch 3, while validation mAP oscillated within a narrow band and never surpassed that peak. This divergence—improving training loss with static validation performance—is characteristic of overfitting to pseudo-label-specific patterns rather than a learning-rate or diminishing-returns effect. This observation motivated the ratio reduction and increased regularization described above, which raised the ceiling from 0.7979 to 0.8116.

D. Localization vs. Classification Analysis

To understand residual errors, we decomposed detection failures into localization (was the tool found?) and classification (once found, was it labeled correctly?). Across all classes, classification accuracy given correct localization never fell below 0.976, whereas localization recall was the limiting factor for the weakest classes—tip-up fenestrated grasper (0.778) and prograsp forceps (0.854). The model reliably recognizes tools it localizes; its weakness is failing to localize certain instruments, not confusing them with others.

E. Ablations

We isolated individual curation mechanisms at the selection stage. Removing per-class confidence floors (using a uniform 0.35 threshold) materially reduces usable data only for the two thinnest affected classes—clip applier (-62%) and grasping retractor (-47%)—while leaving abundant classes unchanged, since they remain quota-bound regardless. Removing the per-session cap lets three of four scarce classes reach 100% of their available candidates, showing the session cap—not the class quota—was the true volume bottleneck for scarce classes. These results indicate that different mechanisms matter for different, specific reasons, and that curation quality, not raw volume, drives the improvement.

VI. Discussion and Limitations

Our strongest results rest on single training runs without repeated seeds, so reported gaps—while large enough to be unlikely to arise from noise—should be interpreted with that caveat. Residual weakness remains concentrated in a small number of classes, and our localization analysis suggests future gains for these classes should target localization rather than classification. A portion of apparent misses for prograsp forceps were near-misses—correct class and location falling just short of the 0.5 IoU threshold—suggesting annotation-style differences also contribute to measured error for this class.

VII. Future Work

Several directions remain. First, we plan to extend the ratio/weight-decay refinement to D-FINE, whose fine-grained distribution-refinement design targets the localization precision our analysis identifies as the key bottleneck. Second, we will complete the remaining curation ablations (isolating the class-quota mechanism) with full training runs to quantify each mechanism's effect on final mAP rather than selection-stage volume alone. Third, a decoupled architecture—generating boxes with one model and classifying them with a separate stage—may improve the localization-limited classes. Finally, we will format predictions to the official scored class set for challenge submission.

VIII. Conclusion

We presented a semi-supervised surgical tool detection pipeline that closes a large annotation gap through curated pseudo-labeling. By bootstrapping a teacher from a small, diversity-sampled manual set and curating its pseudo-labels for noise, class balance, and session diversity, we improved RF-DETR from 0.5082 to 0.8116 mAP@50:95. Our ablations show the curation strategy—not merely added data volume—drives the gain, and our error analysis localizes residual weakness to a small, specific, and actionable set of classes.

Acknowledgment

This work was conducted through the CRCV Research Experiences for Undergraduates (REU) program at the University of Central Florida, supported by the National Science Foundation.

References

[1] A. Zia, M. Berniker, R. Nespolo, C. Perreault, Z. Wang, B. Mueller, R. Schmidt, K. Bhattacharyya, X. Liu, and A. Jarc, "Surgical Visual Understanding (SurgVU) Dataset," *arXiv:2501.09209*, 2025.

[2] A. Zia, K. Bhattacharyya, X. Liu, M. Berniker, *et al.*, "Intuitive Surgical SurgToolLoc and SurgVU Challenges Results: 2022–2025," *arXiv:2305.07152*, 2026.

[3] I. Robinson, I. Robicheaux, P. Popov, M. Ramanan, and D. Peri, "RF-DETR: Neural Architecture Search for Real-Time Detection Transformers," *arXiv:2511.09554*, 2026.

[4] O. Sener and S. Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," in *Proc. ICLR*, 2018, *arXiv:1708.00489*.

[5] Y. Peng *et al.*, "D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement," in *Proc. ICLR*, 2025.