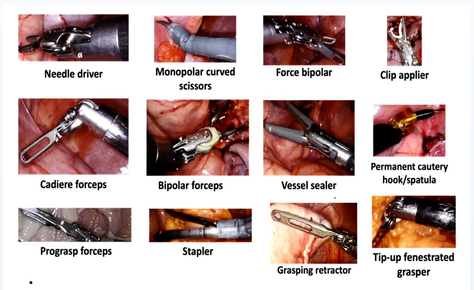
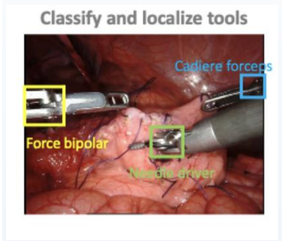
Semi-Supervised Surgical Tool Detection

DeAndre Bailey |Mentor: *Darinka Townsend Hinostroza* | Mentor: *Yuanhao Zou* | Advisor: *Dr. Chen Chen*
University of Central Florida, Orlando, FL

Introduction

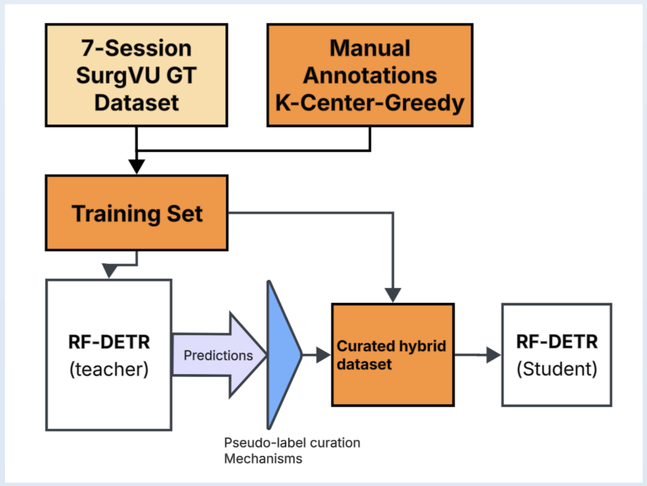
- Robotic-assisted surgery (da Vinci Systems) is now routine in surgery and annotations are necessary for workflow analysis, skill assessment, and OR safety monitoring. However, annotating every frame is very time consuming.
- The SurgVU26 MICCAI EndoVis Challenge (Category 1) poses this problem directly: detect and localize 12 surgical tool classes within video sessions with weak/noisy labels by developing a weakly/semi-supervised framework [1].





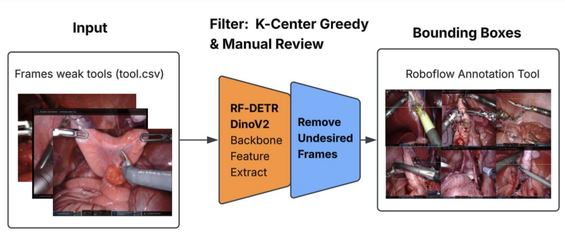
Objective

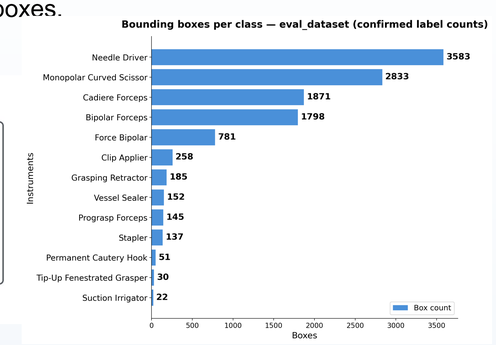
- Our objective is to build a complete detection pipeline that closes this annotation gap without relying on large-scale manual labeling.
- Gather diversity-sampled set of frames for the classes with the least training data, then used that teacher to generate pseudo-labels across all 155 sessions
- To build a strong first teacher model, we prioritized manual annotation on the tool classes with the largest data gaps.



Methodology

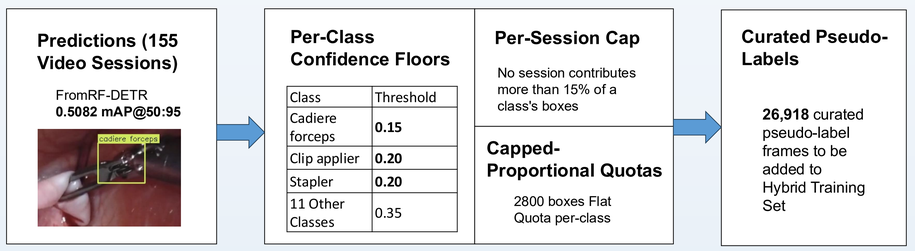
Annotations totaled to be **532 Annotated frames** and 537 bounding boxes added to the given validation set to make 11,668 bounding boxes.





Splitting this into a train and validation set to train RF-DETR-L (chosen object detection model) on it to achieve **0.5082 mAP@50:95** teacher model used to pseudo-label.

With this fine-tuned RF-DETR-L Model, the next step was to produce pseudo-labels from this model's predictions.



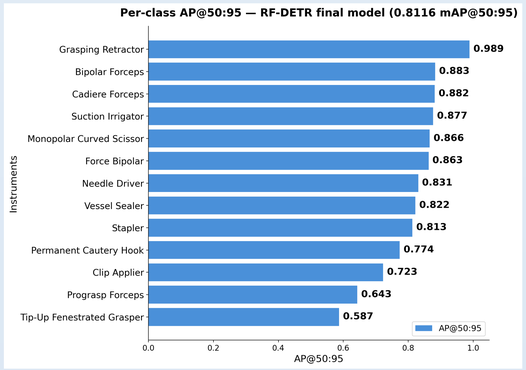
3 Mechanisms for noise reduction, class balance, session diversity bring together **26,918** curated pseudo-label frames to be added to Hybrid Training Set.

Curated Dataset

Combined curated pseudo-labels (PL) with ground-truths (GT) from the Train Split of the SurgVU Validation Set make a Curated Hybrid Training Set with **31,041** frames labeled, **42,605** Train Boxes and **3,459** Val Boxes

The Validation Set contains Ground Truths from SurgVU Validation Set and our manually annotated frames

class	train boxes	val boxes
monopolar curved scissors	8,851	760
needle driver	6,188	1,033
cadiere forceps	5,330	591
bipolar forceps	4,243	511
force bipolar	3,351	257
prograsp forceps	3,292	48
vessel sealer	3,006	33
stapler	3,004	28
permanent cautery hook/spatula	2,847	14
grasping retractor	859	70
tip-up fenestrated grasper	684	9
clip applier	654	94
suction irrigator	296	11
Total	**42,605**	**3,459**



Results

Model	mAP@50:95	Phase / dataset
Yolov8l	0.272	SurgVU Validation Set Spit
RF-DETR-L	0.352	SurgVU Validation Set Spit
RF-DETR-L (teacher)	0.5082	13-class Manual Annotations + SurgVU Validation Set Spit
RF-DETR-L	0.7979	Curated hybrid dataset
RF-DETR-L (best)	**0.8116**	Curated hybrid dataset + 3:1 Ratio + Weight Decay
D-FINE-L	0.7805	Curated hybrid dataset

After training the RF-DETR-L on Curated Hybrid Dataset we achieved a **0.7979** mAP@50:95.

To prevent the potential issue of Overfitting the ratio of PL to GT from ~6.5:1 to ~3:1 and introducing Weight decay to achieve **0.8116** mAP@50:95.

RF-DETR-L can be shown to outperform YOLOv8l in this domain [3].

Key Takeaways

- Targeted manual annotation of underrepresented tool classes, selected via diversity sampling (an active-learning-style frame selection **K-Center-Greedy** strategy)
- Manually Annotated **532 Fram**es on Weak Tool Classes
- Curated hybrid dataset combining ground-truth anchors with **26,918** curated pseudo-label frames (**31,041** total training frames)
- Three-strategy pseudo-label curation — noise reduction, class balance, session diversity — drove the performance gain, not just added data volume
- RF-DETR improved from **0.5082 → 0.8116 mAP@50:95** after retraining on the curated dataset

Future Work

In the Future, AP Scores for poor performing tool classes (Tip-Up Fenestrated Grasper, Prograsp Forceps, etc.) can be improved to match the other high performing classes. In addition, D-FINE, an Object Dection Model, could potentially be using for better localization and Decoupled Architecture for better Classification

References

[1] Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., & Jarc, A. (2025). Surgical Visual Understanding (SurgVU) Dataset. *arXiv:2501.09209*.
[2] Zia, A., Bhattacharyya, K., Liu, X., Berniker, M., et al. (2026). Intuitive Surgical SurgToolLoc and SurgVU Challenges Results: 2022–2025. *arXiv:2305.07152*.
[3] Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., & Peri, N. (2026). RF-DETR: Neural Architecture Search for Real-Time Detection Transformers. *ICLR 2026*. arXiv:2511.09554.
[4] Sener, O., & Savarese, S. (2018). Active Learning for Convolutional Neural Networks: A Core-Set Approach. *ICLR 2018*. arXiv:1708.00489.