THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

ProstaTD: Bridging Surgical Triplet from Classification to Fully Supervised Detection

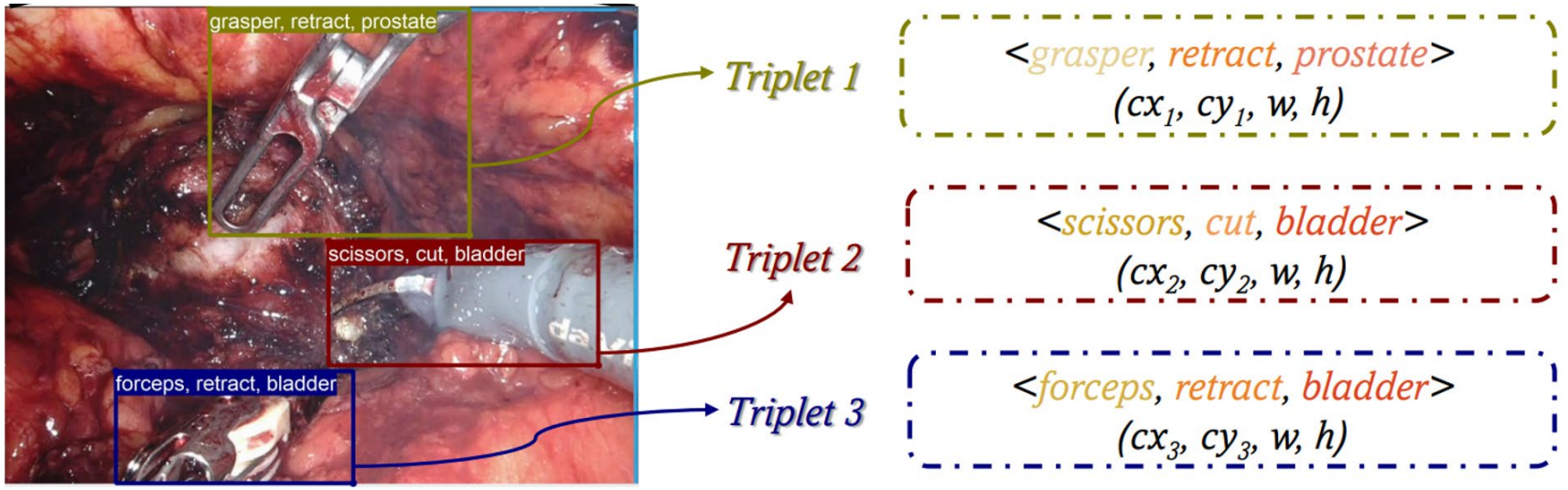
Paper ID: 1839

Presenter: Yiliang Chen

Backgroud

- **Definition:** A surgical triplet is made up of three components which is the **instrument**, the **action**, and the **target**, which is defined by medical experts.

- **Objective**: The goal is to automatically **recognize** these **triplets [tool, action, target]** from each frame of surgical videos.



Research Gap

Table 1: Comparison of existing surgical datasets with ProstaTD, highlighting the unique attributes required for surgical triplet detection. LC and RARP denote Laparoscopic Cholecystectomy and Robot-Assisted Radical Prostatectomy. Full BBox indicates that every frame is fully annotated with bounding boxes. Complete Surgery denotes that the dataset as a whole provides coverage for the entire surgical procedure. Triplet Boundary refers to the temporal start and end range of a complete surgical triplet. Multiple Sources indicates that the data is collected from various hospital.

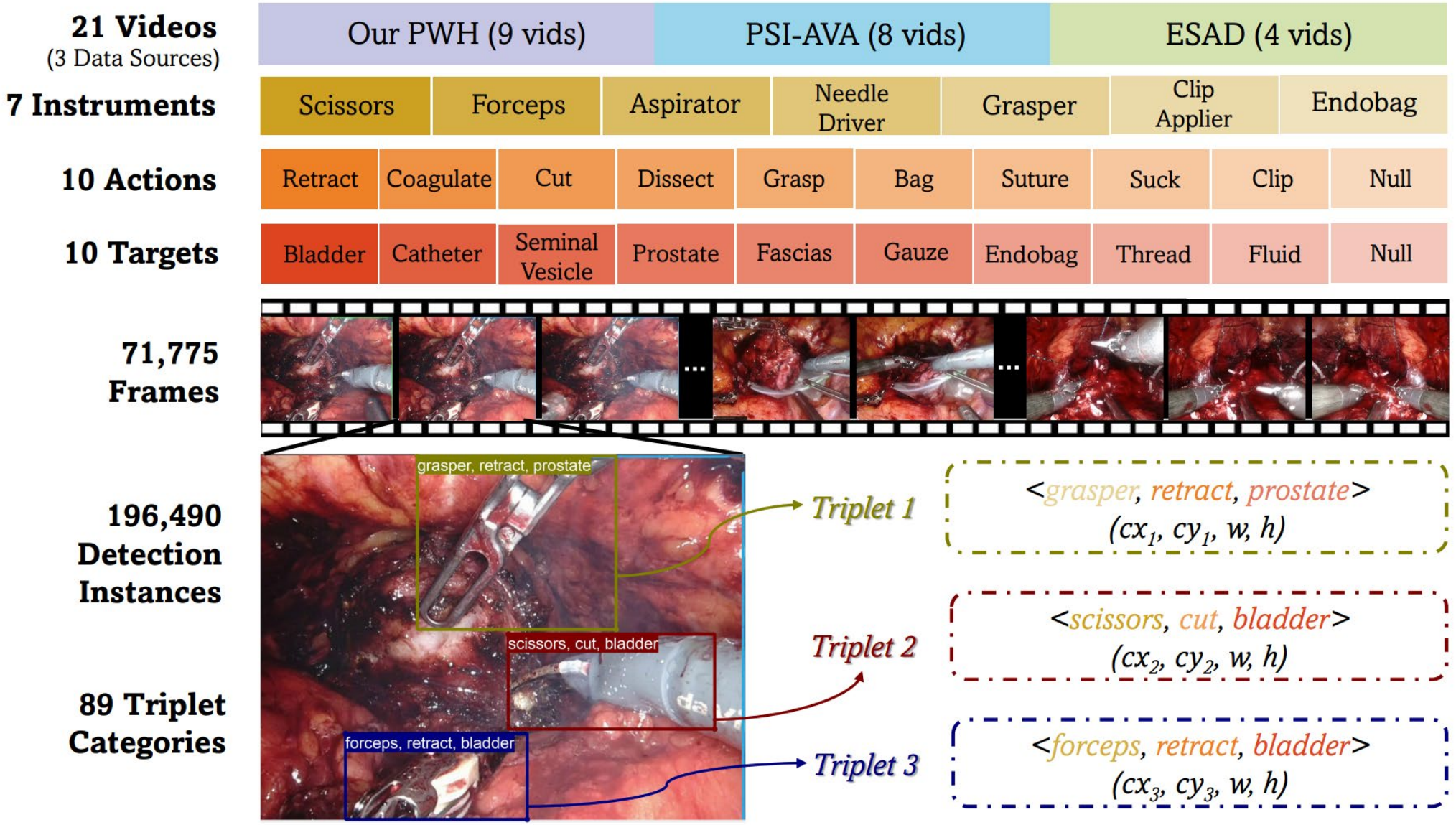
Dataset	Task	Attributes						Statistics	
		Supervised Detection	Full BBox	Triplet-like Structure	Triplet Boundary	Multiple Sources	Complete Surgery	No. Instances	No. Triplets
Cholec80-locations	LC	✓	✓				✓	6,471	–
CholecTrack20	LC	✓	✓				✓	65,200	–
ESAD	RARP	✓	✓				✓	46,753	–
PSI-AVA	RARP	✓					✓	5,804	–
CholecQ	LC	✓	✓	✓				14,480	17
CholecT45	LC			✓			✓	146,394	100
CholecT50	LC			✓			✓	161,988	100
ProstaTD (Ours)	RARP	✓	✓	✓	✓	✓	✓	196,490	89

Contributions

In summary, our contributions are fourfold:

- We introduce a new task with a new dataset for fully supervised surgical triplet detection at the procedure level. To the best of our knowledge, our ProstaTD is the *largest* surgical dataset with instance-level annotations.
- We construct a multi-institutional surgical triplet dataset featuring detailed labels (precise bounding boxes and standardized triplet boundaries) in more complex surgical scenarios.
- We release two open-source annotation tools, which are the first specifically tailored for surgical triplet annotation, along with an open-source evaluation toolkit for benchmarking surgical triplet detection, providing a foundation for surgical triplet analysis across diverse surgical procedures.
- We introduce the first benchmark for fully supervised surgical triplet detection, providing our tailored method as a baseline for comparison.

Dataset Structures



Results

Table 4: Detection performance on the ProstaTD dataset for I, V, T, and IVT components. We report mAP at IoU thresholds of 50% ("50") and 50:95 ("95"), together with inference speed (FPS). All results are reported as mean$_{\pm\text{std}}$ (%) over 5-fold cross-validation. Experiments are conducted with input size 640×640 on a single NVIDIA RTX 4090 GPU. Bold values with light green background indicate the best results, and underlined values with light purple background indicate the second-best.

Method	$mAP_I(\%)\uparrow$		$mAP_V(\%)\uparrow$		$mAP_T(\%)\uparrow$		$mAP_{IVT}(\%)\uparrow$		FPS ↑
	50	95	50	95	50	95	50	95	
Tripnet-Det*	$1.6_{0.4}$	–	$0.6_{0.3}$	–	$0.4_{0.1}$	–	$0.1_{0.0}$	–	**331.8**
RDV-Det*	$1.8_{0.5}$	–	$0.6_{0.4}$	–	$0.3_{0.1}$	–	$0.1_{0.0}$	–	146.6
Faster R-CNN	$73.3_{4.9}$	$63.2_{5.4}$	$48.4_{5.8}$	$42.1_{5.3}$	$43.5_{6.3}$	$37.6_{5.4}$	$25.9_{4.4}$	$22.6_{3.9}$	23.4
Cascade R-CNN	$69.5_{5.1}$	$59.6_{5.6}$	$44.6_{6.1}$	$38.5_{5.6}$	$39.5_{6.6}$	$33.6_{5.6}$	$21.6_{4.6}$	$18.7_{4.1}$	20.6
SSD	$74.6_{4.8}$	$64.5_{5.3}$	$50.2_{5.7}$	$43.7_{5.2}$	$45.4_{6.1}$	$39.4_{5.2}$	$27.1_{4.3}$	$23.8_{3.9}$	82.4
Vit-Det	$86.5_{2.8}$	$73.6_{2.8}$	$52.2_{4.6}$	$45.4_{3.8}$	$48.1_{4.8}$	$41.2_{3.8}$	$30.2_{3.9}$	$26.8_{3.5}$	16.8
Deformable-DETR	$75.4_{4.7}$	$65.0_{5.2}$	$51.1_{5.7}$	$44.5_{5.1}$	$46.3_{6.2}$	$40.1_{5.2}$	$27.5_{4.6}$	$24.0_{4.1}$	24.5
RT-DETR	**91.6**$_{0.9}$	**81.0**$_{1.6}$	$58.9_{4.7}$	$52.8_{3.3}$	**56.8**$_{2.4}$	$50.6_{2.2}$	$33.0_{3.8}$	$29.6_{3.3}$	66.3
YOLOv10	$88.4_{1.3}$	$80.7_{2.2}$	$59.4_{3.4}$	$\underline{54.9}_{2.5}$	$54.6_{3.2}$	$50.2_{3.1}$	$34.3_{4.1}$	$\underline{31.8}_{3.5}$	200.3
YOLOv11	$88.2_{1.4}$	$80.0_{2.5}$	$59.1_{3.2}$	$54.4_{2.1}$	$55.6_{3.6}$	**51.0**$_{3.5}$	$34.1_{3.7}$	$31.5_{3.3}$	185.2
YOLOv12	$88.8_{1.1}$	$80.4_{1.9}$	$\underline{59.9}_{3.1}$	$54.8_{2.2}$	$54.5_{2.1}$	$49.9_{1.9}$	$\underline{34.3}_{3.8}$	$31.5_{3.2}$	$\underline{204.1}$
TAPIR	$76.1_{4.5}$	$65.8_{4.8}$	$52.3_{5.1}$	$45.6_{4.6}$	$47.1_{5.4}$	$40.5_{4.7}$	$28.4_{4.7}$	$24.6_{4.2}$	10.6
MCIT-IG	$77.4_{4.4}$	$67.2_{4.6}$	$53.6_{4.9}$	$46.9_{4.3}$	$48.4_{5.1}$	$41.8_{4.5}$	$29.6_{4.5}$	$26.0_{4.0}$	16.0
TDnet (Ours)	$\underline{89.9}_{1.3}$	$\underline{81.0}_{2.0}$	**61.7**$_{2.9}$	**56.3**$_{2.1}$	$\underline{55.7}_{2.4}$	$\underline{50.8}_{2.7}$	**36.1**$_{3.4}$	**33.1**$_{3.1}$	126.6

* Weakly-supervised methods.

Thanks for your attention

:)

Yiliang Chen