



# SELSAMOT: Sequence-Level Semantic Aggregation for Multi-Object Tracking

Elaine Sui\*, Serena Yeung  
{esui, syyeung}@stanford.edu

Stanford  
Computer Science

## Motivation & Background

- Multi-object tracking is crucial for **scene understanding** from video
- End Goal: Apply tracking to videos of surgeons performing open surgery to assess surgical skill and help facilitate training
- Challenges: Motion blur, object deformation, occlusion, re-identification
- Current tracking algorithms do not leverage the dependency between frames in a video

**We propose to use contextual information in a video to enhance per-frame object detections and subsequently improve tracking**

## Related Work

- Tracking-by-detection: Object detector + deterministic tracking algorithm (e.g. ByteTrack<sup>[1]</sup>)
- Tracking-by-attention: End-to-end tracking (with transformers)
- Feature aggregation in video object detection (e.g. SELSA<sup>[2]</sup>)

## Problem Statement

**Input:** Video frames (current and context)

**Output:** Bounding boxes and track IDs for each instance

**Dataset:** MOT17 & CrowdHuman (> 25k frames total)

**Metrics:**

CLEARMOT: MOTA, MOTP, IDS

$$\text{MOTA} = 1 - \frac{|\text{FN}| - |\text{FP}| + |\text{IDS}|}{|\text{gtDet}|}$$

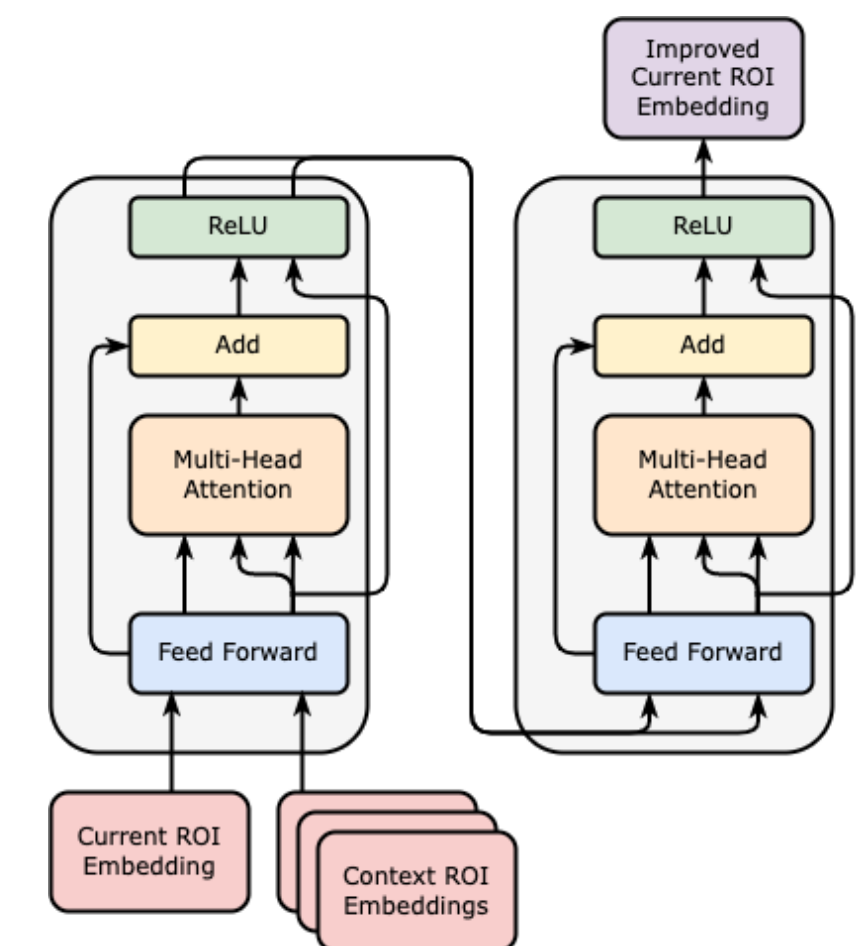
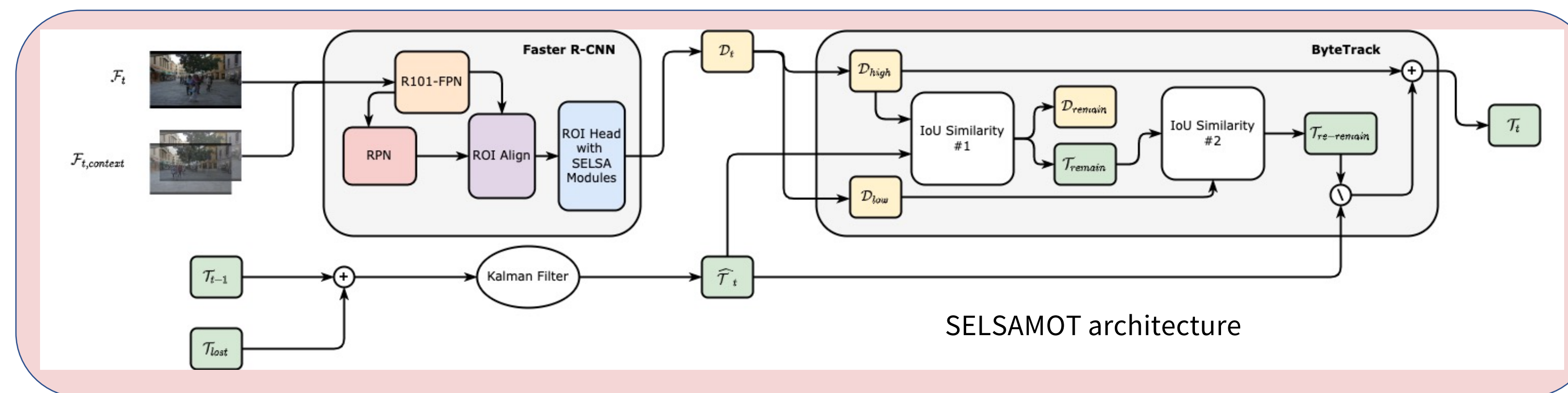
HOTA: HOTA, DetA, AssA

$$\text{HOTA} = \sqrt{\text{DetA} \cdot \text{AssA}}$$

$$\text{DetA} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}| + |\text{FP}|} \quad \text{AssA} = \frac{1}{|\text{TP}|} \sum_{c \in \{\text{TP}\}} \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|}$$

\*n=2.5, k=5 at training time. Denoted are settings at test time.  
\*\*In part, due to high IDFN.

## Method



## Qualitative Results

**Baseline (Faster R-CNN with ResNet101-FPN)**



**SELSAMOT**



## Quantitative Results

**Tracking Metrics:**

|                        | HOTA          | DetA          | AssA          | MOTA          | MOTP          | IDS          |
|------------------------|---------------|---------------|---------------|---------------|---------------|--------------|
| Baseline               | <b>52.617</b> | <b>51.111</b> | <b>54.874</b> | <b>50.169</b> | <b>79.415</b> | 405          |
| SELSAMOT (n=2.5, k=5)  | 51.389        | 48.791        | 54.675        | 47.645        | 79.045        | 383**        |
| SELSAMOT (n=2.5, k=10) | 50.968        | 48.477        | 54.134        | 47.269        | 79.087        | 366**        |
| SELSAMOT (n=5, k=5)    | 51.458        | 48.836        | 54.782        | 48.601        | 79.062        | <b>363**</b> |

**Object Detection Metrics:**

|                        | AP            | AP50          | AP75          | APs           | APm           | API           |
|------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Baseline               | 42.294        | <b>71.284</b> | 45.574        | <b>11.271</b> | 31.587        | <b>55.248</b> |
| SELSAMOT (n=2.5, k=5)  | 41.866        | 71.062        | 44.805        | 10.522        | 31.603        | 54.457        |
| SELSAMOT (n=2.5, k=10) | <b>42.315</b> | 71.187        | <b>46.058</b> | 10.585        | <b>31.931</b> | 54.935        |
| SELSAMOT (n=5, k=5)    | 41.221        | 69.243        | 44.419        | 8.454         | 30.726        | 53.989        |

## Discussion

**Conclusion:** SELSAMOT achieves performance that is on-par with the baseline across all object detection and tracking metrics

**Limitation:** Training speed due to bottleneck in ROI head

**Future Work:**

- Evaluating the robustness of these ROI features
- Adding the ROI features as additional input to the tracking algorithm (similar to DeepSORT)
- Extending MOT to a weakly supervised setting

## Acknowledgements

This work was done for the CS231N course final project. We would like to thank Professor Serena Yeung and the rest of the MARVL lab for their guidance.

## References

- [1] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2021.  
[2] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. ICCV 2019, 2019.