Qualitative Results

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

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[1])
- Tracking-by-attention: End-to-end tracking (with transformers)
- Feature aggregation in video object detection (e.g., SELSA[2])

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
MOTA = 1 - \[
\frac{[FP] - [TP] + [IDS]}{[gDet]}
\]
HOTA: HOTA, DetA, AssA
HOTA = \sqrt{DetA \cdot AssA}
HOTA = \frac{TP}{TP + FP + FN} \cdot AssA

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

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


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
MOTA = 1 - \[
\frac{[FP] - [TP] + [IDS]}{[gDet]}
\]
HOTA: HOTA, DetA, AssA
HOTA = \sqrt{DetA \cdot AssA}
HOTA = \frac{TP}{TP + FP + FN} \cdot AssA

Object Detection Metrics:

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

Tracking Metrics:

<table>
<thead>
<tr>
<th>Tracking Metrics</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APs</th>
<th>APm</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>52.617</td>
<td>51.111</td>
<td>54.874</td>
<td>50.169</td>
<td>79.415</td>
<td>405</td>
</tr>
<tr>
<td>SELSAMOT (n=2.5, k=5)</td>
<td>51.389</td>
<td>48.791</td>
<td>54.675</td>
<td>47.045</td>
<td>79.045</td>
<td>383**</td>
</tr>
<tr>
<td>SELSAMOT (n=2.5, k=10)</td>
<td>50.968</td>
<td>48.477</td>
<td>54.134</td>
<td>47.269</td>
<td>78.987</td>
<td>366**</td>
</tr>
<tr>
<td>SELSAMOT (n=5, k=5)</td>
<td>51.458</td>
<td>48.836</td>
<td>54.782</td>
<td>48.601</td>
<td>79.062</td>
<td>363**</td>
</tr>
</tbody>
</table>

Object Detection Metrics:

<table>
<thead>
<tr>
<th>Object Detection Metrics</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>APs</th>
<th>APm</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>64.294</td>
<td>71.284</td>
<td>45.574</td>
<td>11.271</td>
<td>31.587</td>
<td>55.248</td>
</tr>
<tr>
<td>SELSAMOT (n=2.5, k=5)</td>
<td>61.866</td>
<td>71.062</td>
<td>44.805</td>
<td>10.522</td>
<td>31.603</td>
<td>54.457</td>
</tr>
<tr>
<td>SELSAMOT (n=2.5, k=10)</td>
<td>62.315</td>
<td>71.187</td>
<td>46.058</td>
<td>10.585</td>
<td>31.931</td>
<td>54.935</td>
</tr>
<tr>
<td>SELSAMOT (n=5, k=5)</td>
<td>61.221</td>
<td>69.243</td>
<td>44.419</td>
<td>8.454</td>
<td>30.726</td>
<td>53.989</td>
</tr>
</tbody>
</table>

*In part, due to high IDFs. **In part, due to high IDFN.