PnPNet: End-to-End Perception and Prediction with Tracking in the Loop

Ming Liang*, Bin Yang*, Wenyuan Zeng,
Yun Chen, Rui Hu, Sergio Casas, Raquel Urtasun

Uber ATG | University of Toronto
PnPNet: Tracking in-the-loop

- End-to-End Perception & Prediction, *tracking in-the-loop*

- Performs **discrete-continuous tracking** between detection and prediction
- **Explicit memory** for past tracks and their features
- Exploits long history information with a new **trajectory representation**
- **End-to-end** optimization of multi-tasks

Model Architecture

- We start with a joint perception & prediction architecture
- We add a **discrete tracker** that links detections across time
- We **smooth** the updated **trajectories** in continuous space
- We perform motion forecasting from **trajectory-level object feature**
- The model runs in a recurrent fashion, and **memorizes past trajectories**

A New Object Trajectory Representation

- **History observation feature**: per-frame feature extraction given the trajectory
- **History motion feature**: location displacement over the trajectory
- Feature fusion and temporal modelling

Discrete Tracking

- For each *past track*:
  - It’s associated with a current detection → Multi-Object Tracking
  - It’s unassociated, therefore need to “hallucinate” its current state → Single-Object Tracking

- For each *current detection*:
  - It’s associated with a past track → Multi-Object Tracking
  - It’s unassociated, therefore need to “birth” a new track → Multi-Object Tracking

Multi-Object Tracking

- Handle “newborn” objects by adding \textit{null} nodes at past tracks side
- Learnable matching function
- Hungarian algorithm for optimal assignment

Single-Object Tracking

- Perform on unassociated tracks
- Inherit the spirit from Siamese tracker [2], but replaces correlation with a learnable match function
- Produce more accurate estimations by exploiting observations

Continuous Tracking

- **Classification**: re-estimate the object confidence
- **Regression**: Smooth the past trajectory
- **Post-Process**: NMS, keep top-50 confident objects

Motion Forecasting

- Prediction header: simple regression based prediction
- Input features: trajectory features after tracking

End-to-End Learning

- We adopt multi-task loss for detection, tracking and prediction
- Video-centric training, with online estimations from previous modules & time steps

3D Detection Results on nuScenes

- The detection module of PnPNet achieves state-of-the-art performance

<table>
<thead>
<tr>
<th>Methods</th>
<th>AP ↑</th>
<th><a href="mailto:AP@0.5m">AP@0.5m</a></th>
<th>@1m</th>
<th>@2m</th>
<th>@4m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapillary [40]</td>
<td>47.9</td>
<td>10.2</td>
<td>36.2</td>
<td>64.9</td>
<td>80.1</td>
</tr>
<tr>
<td>PointPillars [22]</td>
<td>70.5</td>
<td>55.5</td>
<td>71.8</td>
<td>76.1</td>
<td>78.6</td>
</tr>
<tr>
<td>Megvii [55]</td>
<td>82.3</td>
<td>72.9</td>
<td>82.5</td>
<td>85.9</td>
<td>87.7</td>
</tr>
<tr>
<td>PnPNet, det only</td>
<td><strong>82.7</strong></td>
<td><strong>73.7</strong></td>
<td><strong>83.3</strong></td>
<td><strong>86.2</strong></td>
<td><strong>87.5</strong></td>
</tr>
</tbody>
</table>

Multi-Object Tracking Results on nuScenes

- Compared with state-of-the-art, PnPNet achieves 8.0% gain in AMOTA
- Compared with a Kalman Filter based tracker, PnPNet achieves 4.6% gain in AMOTA
- PnPNet also produces more complete trajectories

<table>
<thead>
<tr>
<th>Methods</th>
<th>AMOTA↑</th>
<th>AMOTP↓</th>
<th>RECALL↑</th>
<th>MOTA↑</th>
<th>MOTP↓</th>
<th>MT↑</th>
<th>ML↓</th>
<th>FP↓</th>
<th>IDS↓</th>
<th>FRAG↓</th>
<th>TID↓</th>
<th>LGD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>StanfordIPRL-TRI [13]</td>
<td>73.5%</td>
<td>0.53</td>
<td>73.8%</td>
<td>62.3%</td>
<td>0.26</td>
<td>1978</td>
<td>1053</td>
<td>6340</td>
<td>367</td>
<td>341</td>
<td>0.79</td>
<td>1.08</td>
</tr>
<tr>
<td>PnPNet, KF tracker</td>
<td>76.1%</td>
<td>0.52</td>
<td>79.1%</td>
<td>64.8%</td>
<td>0.24</td>
<td>2351</td>
<td>745</td>
<td>7555</td>
<td>802</td>
<td>628</td>
<td>0.51</td>
<td>0.97</td>
</tr>
<tr>
<td>PnPNet</td>
<td>81.5%</td>
<td>0.44</td>
<td>81.6%</td>
<td>69.7%</td>
<td>0.26</td>
<td>2518</td>
<td>804</td>
<td>6771</td>
<td>152</td>
<td>310</td>
<td>0.30</td>
<td>0.57</td>
</tr>
</tbody>
</table>

AMOTA/AMOTP: MOTA/MOTP averaged over different recall thresholds;
TID: average track initialization duration in seconds;
LGD: average longest gap duration in seconds.

Joint P&P Results on nuScenes & ATG4D

Absolute gain in **perception** metrics

![Bar chart showing absolute gain in perception metrics for nuScenes and ATG4D.](chart1)

Relative error reduction in **prediction** metrics

![Bar chart showing relative error reduction in prediction metrics for nuScenes and ATG4D.](chart2)

Qualitative Results