Improving Autonomous Driving Pipeline with Graph Neural Networks

Xinshuo Weng, PhD candidate
Robotics Institute, Carnegie Mellon University

December 26, 2020
Overview

• Standard perception and prediction pipeline for autonomous driving

• Application of Graph Neural Networks for interaction modeling in tracking

• Extension of interaction modeling in perception and prediction pipeline

• Moving forward: the inverted perception and prediction pipeline
Perception and prediction are important components in the autonomous driving stack.
Standard Perception and Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Perception

Prediction
Standard Perception and Prediction Pipeline

- Sensor Data
  - 3D Object Detection
  - 3D Multi-Object Tracking
  - Trajectory Forecasting

LiDAR

RGB
Standard Perception and Prediction Pipeline

Sensor Data → 3D Object Detection → 3D Multi-Object Tracking → Trajectory Forecasting

Detection results
Standard Perception and Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Tracking results
Standard Perception and Prediction Pipeline

Sensor Data → 3D Object Detection → 3D Multi-Object Tracking → Trajectory Forecasting

Forecasting results
What is the state of the art?
Better models from better (bigger) data!

<table>
<thead>
<tr>
<th></th>
<th>KITTI</th>
<th>NuScenes</th>
<th>Argo*</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenes</td>
<td>22</td>
<td>1000</td>
<td>113</td>
<td>1150</td>
</tr>
<tr>
<td>Ann. Lidar Fr.</td>
<td>15K</td>
<td>40K</td>
<td>22K</td>
<td>230K</td>
</tr>
<tr>
<td>Hours</td>
<td>1.5</td>
<td>5.5</td>
<td>1</td>
<td>6.4</td>
</tr>
<tr>
<td>3D Boxes</td>
<td>80K</td>
<td>1.4M</td>
<td>993k</td>
<td>12M</td>
</tr>
<tr>
<td>2D Boxes</td>
<td>80K</td>
<td>–</td>
<td>–</td>
<td>9.9M</td>
</tr>
<tr>
<td>Lidars</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Cameras</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Avg Points/Frame</td>
<td>120K</td>
<td>34K</td>
<td>107K</td>
<td>177K</td>
</tr>
<tr>
<td>LiDAR Features</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Maps</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Visited Area (km²)</td>
<td>–</td>
<td>5</td>
<td>1.6</td>
<td>76</td>
</tr>
</tbody>
</table>

* Mined trajectory data not counted for the Argo dataset

150x increase!
State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Dataset with multi-modal ground truth

In contrast to prior dataset with single future ground truth

What are the right metrics for evaluation?

State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Monocular 3D Detection (KITTI)

Publication trend on Mono3DOD in Autonomous Driving

15x increase (3 years)

Image credit to Patrick Langechuan Liu, https://towardsdatascience.com/monocular-3d-object-detection-in-autonomous-driving-2476a3c7f57e
State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Lidar-based 3D Detection (KITTI)

27% increase (2 years)
State of the Art

Sensor Data → 3D Object Detection → 3D Multi-Object Tracking → Trajectory Forecasting

2D MOT (KITTI)

Accuracy vs Speed

X. Weng, J. Wang, D. Held, K. Kitani.
3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020
State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Feature Extraction

Data Association

Trajectory Forecasting

Recent trend:

Jointly optimized


State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Goal-conditioned forecasting

Different goals could lead to different forecasts

State of the Art

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Road context / physical constraint helps

Using road structure semantics as inputs eliminates physically impossible trajectories

CoverNet: Multimodal Behavior Prediction using Trajectory Sets. CVPR 2020
State of the Art

End-to-end perception and prediction pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Memory operator: Read from memory: Write to memory:

Memory of BEV Feature Maps

Memory of Object Trajectories

HD map at frame t

3D Object Detection

Track-Detection Association

Trajectory Estimation

LIDAR data at frame t

Detects at frame t

Occluded object

Newborn object

Separately optimized

1. Suboptimal performance
2. Slow inference speed

All modules are optimized for the end goal: trajectory prediction

Jointly optimized

State of the Art

Lots of progress on each individual module

Feature representation learning for each object/agent is independent of others!

how does one object/agent affect the other agent in the pipeline?
Using GNNs for Interaction Modeling in Multi-Object Tracking
What are the Issues of Feature Learning?

• Goal: learn discriminative features

• Issues in feature representation learning?
  • Independent feature extraction
    • No communication, ignoring the context information
  • Feature from one modality
    • E.g., 2D appearance, or 2D motion, or 3D motion, or 3D appearance
    • Not utilize all information that is complementary

Prior work

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Improve Feature Learning for 3D MOT

- How can we address these issues?
- Shouldn’t feature depending on other objects?
  - A novel interaction modeling mechanism
- Can we utilize all modalities?
  - Extract multi-modal features
  - i.e., 2D motion + 2D appearance + 3D motion + 3D appearance

Our work

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Multi-Modal Feature Extraction

• How do we do?
  
  • (a) Obtain the appearance / motion features from the 3D point cloud
    • LSTM for 3D motion from 3D box trajectories
    • PointNet for 3D appearance from point cloud

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Multi-Modal Feature Extraction

- How do we do?
  - (b) Obtain the appearance / motion features from the 2D image
    - LSTM for 2D motion from 2D box trajectories
    - CNN for 2D appearance from 2D image patches

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Interaction Modeling with GNNs

- (c) Learn discriminative object features through interaction with GNNs

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Interaction Modeling with GNNs

• How do we do?
  • Construct the graph with nodes for objects in frame $t$ and frame $t+1$
  • Define connectivity between nodes (edges)
  • Perform graph convolution for each node
    $$n_{i}^{t'} = \sigma_4(n_{i}^{t}) + \sum_{j \in \mathcal{N}(i)} \sigma_3(A_{ij}(n_{j}^{t+1} - n_{i}^{t}))$$
  • Edge regression
    $$A_{ij} = \text{Sigmoid}((\text{ReLU}(\sigma_1(n_{i}^{t} - n_{j}^{t+1}))))$$

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020.
Interaction Modeling with GNNs

• Training
  • Batch triplet loss
    \[
    L_{\text{tri}} = \max\left( \| n_i^t - n_j^{t+1} \|, \min_{d_x \in D, id_x \neq id_i} \| n_i^t - n_s^{t+1} \|, \min_{o_r \in O, id_r \neq id_j} \| n_r^t - n_j^{t+1} \| + \alpha, 0 \right)
    \]
  • Affinity loss
    • Binary cross-entropy loss
      \[
      L_{\text{bce}} = -\frac{1}{MN} \sum_i \sum_j A_{ij}^g \log A_{ij} + (1 - A_{ij}^g) \log(1 - A_{ij})
      \]
    • Cross-entropy loss
      \[
      L_{\text{ce}} = -\frac{1}{M} \sum_i A_{ij}^g \log\left(\frac{\exp A_{ij}}{\sum_i^{M} \exp A_{ij}}\right)
      \]
Multi-Modal Feature Extraction

• Is encoding the multi-modal features really useful?
  • Answer: Yes
  • We should encode different features so that they can compliment each other

<table>
<thead>
<tr>
<th>Feature Extractor</th>
<th>sAMOTA (%) ↑</th>
<th>AMOTA (%) ↑</th>
<th>AMOTP (%) ↑</th>
<th>MOTA (%) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D A</td>
<td>88.31</td>
<td>41.62</td>
<td>76.22</td>
<td>79.42</td>
</tr>
<tr>
<td>2D M</td>
<td>64.24</td>
<td>23.95</td>
<td>61.13</td>
<td>54.88</td>
</tr>
<tr>
<td>3D A</td>
<td>88.27</td>
<td>41.55</td>
<td>76.29</td>
<td>77.38</td>
</tr>
<tr>
<td>3D M</td>
<td>88.57</td>
<td>41.62</td>
<td>76.22</td>
<td>81.84</td>
</tr>
</tbody>
</table>

A: appearance feature, M: motion feature

Use feature from
single modality

Use feature from
multiple modalities:
Performance increased!
Interaction Modeling with GNNs

- Is feature interaction using GNN useful to 3D MOT?
  - Answer: Yes
  - We should let objects communicate and encode the context information

![Accuracy vs. Number of Layers](image)

Performance largely increased with GNN layers = 3 v.s. 0!
Interaction Modeling with GNNs

• Initially, there are confusions in the affinity matrix

• After a few layers, the confusion is suppressed, and a more discriminative matrix is obtained

\[
\begin{pmatrix}
0.899 & 0.007 & 0.002 \\
0.001 & 0.982 & 0.993 \\
0.003 & 0.992 & 0.957
\end{pmatrix}
\]

\[
\begin{pmatrix}
1.0 & 0.0 & 0.0 \\
0.0 & 1.0 & 0.0 \\
0.0 & 0.0 & 1.0
\end{pmatrix}
\]

A good estimation

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020
Qualitative Results

X. Weng, Y. Wang, Y. Man, K. Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020
References

• Presented work
  • Weng et al. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020

• Independent works
  • Guillem Braso and Laura Leal-Taixe. Learning a Neural Solver for Multiple Object Tracking. CVPR 2020
  • Li et al. Graph Networks for Multiple Object Tracking. WACV 2020
Introducing GNNs for interaction modeling
Extending GNNs for Interaction Modeling to the Entire Pipeline
Extending GNNs for interaction modeling in joint detection and tracking
Prior Work: Joint MOT without Relation Modeling

Input images → Past tracklets → Jointly trained

Object Detection → Data Association

Prior Work: Relation Modeling for Data Association only

Input images → Past tracklets

Off-the-shelf detections → Relation Modeling Via GNNs → Jointly trained

Data Association

Ours: Relation Modeling for Joint MOT

Input images → Past tracklets → Jointly trained

Relation Modeling Via GNNs → Object Detection → Data Association
GNNs for Joint Object Detection and MOT

- Now we have a method for MOT with interaction modeling
- Can we do joint detection and MOT with GNNs?
  - No detection but anchors
  - Add a detection head (MLP) to classify/regress objects

Y. Wang, X. Weng, K. Kitani. Joint Detection and Multi-Object Tracking with Graph Neural Networks and Complete Feature Learning. arXiv 2020
## Quantitative Evaluation on the MOT Challenges

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA(%)↑</th>
<th>IDF1(%)↑</th>
<th>MT(%)↑</th>
<th>ML(%)↓</th>
<th>IDS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMT [27]</td>
<td>44.5</td>
<td>49.2</td>
<td>34.7</td>
<td>22.1</td>
<td>684</td>
</tr>
<tr>
<td>Lif.TsimmInt [81]</td>
<td>47.2</td>
<td>57.6</td>
<td>27.0</td>
<td>29.8</td>
<td>554</td>
</tr>
<tr>
<td>MDP_SubCNN [11]</td>
<td>47.5</td>
<td>55.7</td>
<td>30.0</td>
<td>18.6</td>
<td>628</td>
</tr>
<tr>
<td>CDA_DDAL [82]</td>
<td>51.3</td>
<td>54.1</td>
<td>36.3</td>
<td>22.2</td>
<td>544</td>
</tr>
<tr>
<td>MPNTrack [40]</td>
<td>51.5</td>
<td>58.6</td>
<td>31.2</td>
<td>25.9</td>
<td>375</td>
</tr>
<tr>
<td>EAMTT [83]</td>
<td>53.0</td>
<td>54.0</td>
<td>35.9</td>
<td>19.6</td>
<td>776</td>
</tr>
<tr>
<td>AP_HWDPL [84]</td>
<td>53.0</td>
<td>52.0</td>
<td>29.1</td>
<td>20.2</td>
<td>708</td>
</tr>
<tr>
<td>NOMTwSDP [85]</td>
<td>55.5</td>
<td>59.1</td>
<td>39.0</td>
<td>25.8</td>
<td>427</td>
</tr>
<tr>
<td>RAR15 [86]</td>
<td>56.5</td>
<td>61.3</td>
<td>45.1</td>
<td>14.6</td>
<td>428</td>
</tr>
<tr>
<td>Tube.TK [35]</td>
<td>58.4</td>
<td>53.1</td>
<td>39.3</td>
<td>18.0</td>
<td>854</td>
</tr>
<tr>
<td><strong>GSDT (Ours)</strong></td>
<td><strong>60.7</strong></td>
<td><strong>64.6</strong></td>
<td><strong>47.0</strong></td>
<td><strong>10.5</strong></td>
<td><strong>477</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA(%)↑</th>
<th>IDF1(%)↑</th>
<th>MT(%)↑</th>
<th>ML(%)↓</th>
<th>IDS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSORT.2 [13]</td>
<td>61.4</td>
<td>62.2</td>
<td>32.8</td>
<td>18.2</td>
<td>1,423</td>
</tr>
<tr>
<td>NOMTwSDP16 [85]</td>
<td>62.2</td>
<td>62.6</td>
<td>32.5</td>
<td>31.1</td>
<td>406</td>
</tr>
<tr>
<td>VMaxx [87]</td>
<td>62.6</td>
<td>49.2</td>
<td>32.7</td>
<td>21.1</td>
<td>1,389</td>
</tr>
<tr>
<td>RAR16wVGG [86]</td>
<td>63.0</td>
<td>63.8</td>
<td>39.9</td>
<td>22.1</td>
<td>482</td>
</tr>
<tr>
<td>TAP [88]</td>
<td>64.8</td>
<td>73.5</td>
<td>38.5</td>
<td>21.6</td>
<td>571</td>
</tr>
<tr>
<td>CNNMTT [89]</td>
<td>65.2</td>
<td>62.2</td>
<td>32.5</td>
<td>21.3</td>
<td>946</td>
</tr>
<tr>
<td>POI [90]</td>
<td>66.1</td>
<td>65.1</td>
<td>34.0</td>
<td>20.8</td>
<td>3,093</td>
</tr>
<tr>
<td>Tube.TK.POI [35]</td>
<td>66.9</td>
<td>62.2</td>
<td>39.0</td>
<td>16.1</td>
<td>1,236</td>
</tr>
<tr>
<td>CTracker.V1 [34]</td>
<td>67.6</td>
<td>57.2</td>
<td>32.9</td>
<td>23.1</td>
<td>1,897</td>
</tr>
<tr>
<td><strong>GSDT (Ours)</strong></td>
<td><strong>66.7</strong></td>
<td><strong>69.2</strong></td>
<td><strong>38.6</strong></td>
<td><strong>19.0</strong></td>
<td><strong>959</strong></td>
</tr>
</tbody>
</table>

### MOT17

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA(%)↑</th>
<th>IDF1(%)↑</th>
<th>MT(%)↑</th>
<th>ML(%)↓</th>
<th>IDS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPNTrack [40]</td>
<td>58.8</td>
<td>61.7</td>
<td>28.8</td>
<td>33.5</td>
<td>1,185</td>
</tr>
<tr>
<td>Lif.T [81]</td>
<td>60.5</td>
<td>65.6</td>
<td>27.0</td>
<td>33.6</td>
<td>1,189</td>
</tr>
<tr>
<td>Tube.TK [35]</td>
<td>63.0</td>
<td>58.6</td>
<td>31.2</td>
<td>19.9</td>
<td>4,137</td>
</tr>
<tr>
<td>CTrackerV1 [34]</td>
<td>66.6</td>
<td>57.4</td>
<td>32.2</td>
<td>24.2</td>
<td>5,529</td>
</tr>
<tr>
<td>CTracker17 [30]</td>
<td>67.8</td>
<td>64.7</td>
<td>34.6</td>
<td>24.6</td>
<td>3,039</td>
</tr>
<tr>
<td><strong>GSDT (Ours)</strong></td>
<td><strong>66.2</strong></td>
<td><strong>68.7</strong></td>
<td><strong>40.8</strong></td>
<td><strong>18.3</strong></td>
<td><strong>3,318</strong></td>
</tr>
</tbody>
</table>

### MOT20

<table>
<thead>
<tr>
<th>Method</th>
<th>MOTA(%)↑</th>
<th>IDF1(%)↑</th>
<th>MT(%)↑</th>
<th>ML(%)↓</th>
<th>IDS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SORT20 [12]</td>
<td>42.7</td>
<td>45.1</td>
<td>16.7</td>
<td>26.2</td>
<td>4,334</td>
</tr>
<tr>
<td><strong>GSDT (Ours)</strong></td>
<td><strong>67.1</strong></td>
<td><strong>67.5</strong></td>
<td><strong>53.1</strong></td>
<td><strong>13.2</strong></td>
<td><strong>3,133</strong></td>
</tr>
</tbody>
</table>
Qualitative Analysis
Visualization on MOT20 test sequences

MOT20-04

MOT20-07
Visualization on MOT17 test sequences

MOT17-03

MOT17-07
Visualization on 2DMOT2015 test sequences

2DMOT2015 – AVG-TownCentre

2DMOT2015 – PETS09-S2L2
Our Method with/without GNNs
Our method with GNNs

Our method without GNNs

Objects that are detected in one video but are missing in the other are marked with **bold** bounding boxes
Objects that are detected in one video but are missing in the other are marked with **bold** bounding boxes.
Our method with GNNs

Our method without GNNs

Objects that are detected in one video but are missing in the other are marked with bold bounding boxes
Our Joint MOT Method with GNNs vs. Existing Joint MOT Methods without GNNs
Ours – Joint detection and data association with GNNs

FairMOT\textsuperscript{[1]} – Joint detection and data association without GNNs

Objects that are detected in one video but are missing in the other are marked with \textbf{bold} bounding boxes

\textsuperscript{[1]} Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, A Simple Baseline for Multi-Object Tracking, ArXiv, 2020
Our Joint MOT Method with GNNs vs. Existing Methods using GNNs for Data Association only
Ours – **Joint detection and data association** with GNNs

MPNTrack$^[2]$ – **Data association only** with GNNs

Objects that are detected in one video but are missing in the other are marked with **bold bounding boxes**

Ours – **Joint detection and data association** with GNNs

MPNTrack\(^{[2]}\) – **Data association only** with GNNs

Objects that are detected in one video but are missing in the other are marked with **bold** bounding boxes

Extending GNNs for interaction modeling in joint tracking and forecasting
GNNs for Joint MOT and Forecasting

- Most prior works separate MOT and forecasting (sequential order)
- Why not good?
  - No interaction modeling
  - Optimization of entire pipeline is impossible
  - Slow inference
  - Error propagation
- What can we do?

Pipeline from Prior Work

3D Multi-Object Tracking

X. Weng, Y. Ye, K. Kitani. Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv 2020
GNNs for Joint MOT and Forecasting

- Parallelize MOT and forecasting
  - Share the feature learning process -> faster inference
  - Use GNNs -> interaction modeling
  - Add a trajectory forecasting head in parallel -> alleviate error propagation
  - End-to-end training

X. Weng, Y. Ye, K. Kitani. Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv 2020
GNNs for Joint MOT and Forecasting

- Is it useful to do joint optimization?
  - Add forecasting is useful to tracking

- How does adding forecasting affect 3D MOT?
  - Add joint optimization with forecasting improves performance on tracking

<table>
<thead>
<tr>
<th>Metrics</th>
<th>w/o forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>sAMOTA(%)↑</td>
<td>90.17</td>
</tr>
<tr>
<td>AMOTA(%)↑</td>
<td>42.81</td>
</tr>
<tr>
<td>AMOTP(%)↑</td>
<td>76.94</td>
</tr>
<tr>
<td>MOTA(%)↑</td>
<td>82.91</td>
</tr>
<tr>
<td>MOTP(%)↑</td>
<td>78.11</td>
</tr>
<tr>
<td>IDS↓</td>
<td>5</td>
</tr>
</tbody>
</table>

Improvement on 5 out of 6 entries!

3D MOT evaluation without forecasting module

X. Weng, Y. Ye, K. Kitani. Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv 2020
GNNs for Joint MOT and Forecasting

- Is it useful to do joint optimization?
  - Add forecasting is useful to tracking
  - Add MOT is useful to forecasting

- How does adding 3D MOT affect trajectory forecasting?
  - Add joint optimization with 3D MOT improves performance on forecasting

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>w/o MOT+DSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-1.0s</td>
<td>ADE↓</td>
<td>0.663</td>
</tr>
<tr>
<td></td>
<td>FDE↓</td>
<td>1.121</td>
</tr>
<tr>
<td></td>
<td>ASD↑</td>
<td>1.796</td>
</tr>
<tr>
<td></td>
<td>FSD↑</td>
<td>3.168</td>
</tr>
<tr>
<td></td>
<td>ADE↓</td>
<td>1.729</td>
</tr>
<tr>
<td>KITTI-3.0s</td>
<td>FDE↓</td>
<td>3.086</td>
</tr>
<tr>
<td></td>
<td>ASD↑</td>
<td>3.196</td>
</tr>
<tr>
<td></td>
<td>FSD↑</td>
<td>5.776</td>
</tr>
</tbody>
</table>

Performance improved after adding MOT!

Forecasting evaluation without MOT

X. Weng, Y. Ye, K. Kitani. Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv 2020
GNNs for Joint MOT and Forecasting

- Is interaction modeling useful?
  - Yes, performance improved for both MOT and forecasting

---

X. Weng, Y. Ye, K. Kitani. Joint 3D Tracking and Forecasting with Graph Neural Network and Diversity Sampling. arXiv 2020
Qualitative Results
Moving Forward:
The Inverted Perception and Prediction Pipeline
Standard Perception and Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Is this really the best place to perform prediction?
Standard Perception and Prediction Pipeline

Can we do prediction here?

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting
Limitation of the Standard Pipeline

- Expensive to scale as it requires labeled object poses
- Hard to annotate in 3D space
Can we scale forecasting performance without requiring human labels?
We hypothesize yes!

A forecast-then-detect pipeline that is less expensive to scale
SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting

- **Standard pipeline:**
  - Detection $\rightarrow$ MOT $\rightarrow$ Trajectory Forecasting

- **Our new pipeline**
  - Sequential Pointcloud Forecasting $\rightarrow$ Detection $\rightarrow$ MOT

- **Differences**
  - Invert the order of forecasting
  - Forecast at the sensor level, instead of at the object level

---

**Conventional pipeline**
- Sensor data in *past* M frames
- Object trajectories in *past* M frames
- Object trajectories in *future* N frames

**Proposed new pipeline**
- Sensor data in *future* N frames
- Object trajectories in *future* N frames

Weng et al. Inverting the Pose Forecasting Pipeline with SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
SPF$^2$: Sequential Pointcloud Forecasting for Sequential Pose Forecasting

- Any advantage?
  - Forecasting module does not require annotation for training
  - Forecasting performance can scale with unlabeled data
- Inherited from the Sequential Pointcloud Forecasting
The challenging first step -- Sequential Pointcloud Forecasting (SPF)
SPF: Sequential Pointcloud Forecasting

- Advantages:
  - Remove the need of labels for training the forecasting module
  - Prediction represents the entire scene, including background
    - Easier to incorporate scene context information

Weng et al. Inverting the Pose Forecasting Pipeline with SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
An effective approach for SPF: SPFNet
SPFNet
• Four modules
  • (a) Shared point cloud encoder
  • (b) LSTM for temporal modeling
  • (c) Shared point cloud decoder
  • (d) Losses

Weng et al. Inverting the Pose Forecasting Pipeline with SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting, CoRL 2020
Results
Is SPFNet Effective to the Proposed SPF Task?

- Outperform baselines devised using existing techniques

**Table 1:** Quantitative evaluation for the proposed SPF task on the KITTI and nuScenes datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>Ours+Point</th>
<th>Ours+RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-1.0s</td>
<td>CD↓</td>
<td>2.37</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>EMD↓</td>
<td>211.47</td>
<td>128.81</td>
</tr>
<tr>
<td>KITTI-3.0s</td>
<td>CD↓</td>
<td>3.91</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>EMD↓</td>
<td>267.42</td>
<td>175.54</td>
</tr>
<tr>
<td>nuScenes-1.0s</td>
<td>CD↓</td>
<td>1.25</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>EMD↓</td>
<td>135.94</td>
<td>78.37</td>
</tr>
<tr>
<td>nuScenes-3.0s</td>
<td>CD↓</td>
<td>2.97</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>EMD↓</td>
<td>128.26</td>
<td>91.83</td>
</tr>
</tbody>
</table>

Weng et al. Inverting the Pose Forecasting Pipeline with SPF$^2$: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Is Our SPF$^2$ Pipeline Competitive?

- Yes, also requires less human annotations for training

### Table 3: Evaluation for the perception and trajectory forecasting pipeline on the KITTI and nuScenes datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI-1.0s</td>
<td>AADE↓</td>
<td>1</td>
<td>0.792</td>
<td>0.524</td>
<td>1.099</td>
<td>0.470</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>0.623</td>
<td>0.340</td>
<td>0.443</td>
<td>0.382</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AFDE↓</td>
<td>1</td>
<td>1.285</td>
<td>0.886</td>
<td>1.708</td>
<td>0.889</td>
<td>0.405</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>1.152</td>
<td>0.511</td>
<td>0.546</td>
<td>0.613</td>
<td>–</td>
</tr>
<tr>
<td>KITTI-3.0s</td>
<td>AADE↓</td>
<td>1</td>
<td>1.692</td>
<td>1.362</td>
<td>2.720</td>
<td>1.432</td>
<td>0.408</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>1.593</td>
<td>0.984</td>
<td>1.231</td>
<td>0.725</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AFDE↓</td>
<td>1</td>
<td>2.670</td>
<td>2.267</td>
<td>3.938</td>
<td>2.536</td>
<td>0.504</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.385</td>
<td>1.512</td>
<td>1.405</td>
<td>1.118</td>
<td>–</td>
</tr>
<tr>
<td>nuScenes-1.0s</td>
<td>AADE↓</td>
<td>1</td>
<td>1.186</td>
<td>1.117</td>
<td>2.030</td>
<td>1.214</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>0.907</td>
<td>0.762</td>
<td>0.826</td>
<td>0.881</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AFDE↓</td>
<td>1</td>
<td>1.490</td>
<td>1.310</td>
<td>2.337</td>
<td>1.563</td>
<td>0.825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>1.231</td>
<td>0.763</td>
<td>0.849</td>
<td>1.197</td>
<td>–</td>
</tr>
<tr>
<td>nuScenes-3.0s</td>
<td>AADE↓</td>
<td>1</td>
<td>1.794</td>
<td>2.224</td>
<td>4.954</td>
<td>2.417</td>
<td>1.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>1.658</td>
<td>1.426</td>
<td>1.760</td>
<td>1.938</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AFDE↓</td>
<td>1</td>
<td>2.850</td>
<td>3.224</td>
<td>6.765</td>
<td>3.479</td>
<td>1.043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.538</td>
<td>1.652</td>
<td>1.845</td>
<td>2.766</td>
<td>–</td>
</tr>
</tbody>
</table>

Weng et al. Inverting the Pose Forecasting Pipeline with SPF$^2$: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Can We Scale Forecasting Performance with More Data?

- Yes, adding more unlabelled point clouds improves performance

Weng et al. Inverting the Pose Forecasting Pipeline with SPF\(^2\): Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Visualization

- Reasonably reconstruct the object locations and scene background
  - Aligns well with GT
- Limitation: object shape not sharp enough
  - Hidden representation is compressed too much
Summary

• Standard perception and prediction pipeline for autonomous driving

• Application of Graph Neural Networks for interaction modeling in tracking

• Extension of interaction modeling in perception and prediction pipeline

• Moving forward: the inverted perception and prediction pipeline
Improving Autonomous Driving Pipeline with Graph Neural Networks

Xinshuo Weng, PhD candidate
Robotics Institute, Carnegie Mellon University

December 26, 2020