A Modular Yet Highly-Integrated Pipeline for Perception and Prediction

Xinshuo Weng
Robotics Institute, Carnegie Mellon University

RI PhD Thesis Proposal
December 2, 2021

Committee Members:
Kris Kitani (Chair), Matthew P. O'Toole, Deva Ramanan, Marco Pavone
Perception and prediction are important components in the autonomous driving stack
Standard Perception-then-Prediction Pipeline

- LiDAR
- Sensor Data
- 3D Object Detection
- 3D Multi-Object Tracking
- Trajectory Forecasting
- RGB
Standard Perception-then-Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Detection results
Standard Perception-then-Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Tracking results
Standard Perception-then-Prediction Pipeline

1. Sensor Data
2. 3D Object Detection
3. 3D Multi-Object Tracking
4. Trajectory Forecasting

Forecasting results
Standard Perception-then-Prediction Pipeline

Sensor Data

3D Object Detection

3D Multi-Object Tracking

Trajectory Forecasting

Perception

Prediction
Standard Perception-then-Prediction Pipeline

Advantages?
- Interpretability
- Fast development cycle
- etc.

Disadvantages
- Cascading errors
- Slower inference speed
- etc.
Modular Improvement
Improving Each Individual Module in Isolation

Tracking

Detection

Prediction

Park, Weng, Man, Kitani. BMVC ‘21
Man, Weng, Sivakumar, O’Toole, Kitani. ICCV ‘21
Weng, Kitani. ICCVW ‘19
Weng, Wu, Beainy, Kitani. WACV ’18

Li, Weng, Xu, Kitani. ICCV ‘21
Weng, Wang, Held, Kitani. IROS ‘20
Weng, Wang, Man, Kitani. CVPR ‘20

Yuan, Weng, Ou, Kitani. ICCV ‘21
Manglik, Weng, Ohn-Bar, Kitani. IROS ‘19
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

Tracking

Detection

Prediction

Affinity-based Prediction
Weng, Ivanovic, Kitani, Pavone. In submission '21
Weng*, Yuan*, Kitani. ICRA/RA-L '21

Weng, Wang, Levine, Kitani, Rhinehart. CoRL '20
Weng, Nan, Lee, McAllister, Gaidon, Rhinehart, Kitani. In submission '21

Weng, Kitani. arXiv '21
Wang, Kitani, Weng. ICRA '21

Weng, Ivanovic, Kitani, Pavone. arXiv '21
Project 1: Enable Agent-Controllability and Diverse Prediction for the Prediction-then-Perception Pipeline


Diagram:
- Tracking
  - Detection
  - Prediction
- Diversity & Agent-Controllability
Project 2: End-to-End Prediction-then-Perception Framework for Driving in Simulation


End-to-End Prediction-then-Perception for Planning

Control

Closed-Loop Evaluation in Simulation
Modular Improvement
Improving Each Individual Module in Isolation

Tracking

Detection

Prediction

Park, Weng, Man, Kitani. BMVC ‘21
Man, Weng, Sivakumar, O’Toole, Kitani. ICCV ‘21
Weng, Kitani. ICCVW ‘19
Weng, Wu, Beainy, Kitani. WACV ‘18

Li, Weng, Xu, Kitani. ICCV ‘21
Weng, Wang, Held, Kitani. IROS ‘20
Weng, Wang, Man, Kitani. CVPR ‘20

Yuan, Weng, Ou, Kitani. ICCV ‘21
Manglik, Weng, Ohn-Bar, Kitani. IROS ‘19
Modular Improvement for Perception and Prediction

Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

Efficient 3D Multi-Object Tracking with 3D Kalman Filter

Social-Aware Representation Learning for 3D Multi-Object Tracking

Agent-Aware Transformer
(Joint Social & Temporal Modeling + Preserve Time & Agent Information)

Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting

Xinshuo Weng, Kris Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019

Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
Performance Gap between LiDAR-based and Monocular 3D Detection

**LiDAR-based 3D detection methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Easy</th>
<th>Hard</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMLab-PartA^2</td>
<td>Moderate</td>
<td></td>
<td>77.86</td>
<td>85.94</td>
<td>72.00</td>
<td>GPU @ 2.5 Ghz (Python + C/C++)</td>
</tr>
<tr>
<td>HRI-FusionRCNN</td>
<td></td>
<td></td>
<td>77.84</td>
<td>86.60</td>
<td>69.15</td>
<td>1 core @ 2.5 Ghz (C/C++)</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td></td>
<td>77.63</td>
<td>86.61</td>
<td>76.06</td>
<td>GPU @ 2.5 Ghz (Python + C/C++)</td>
</tr>
</tbody>
</table>


**Top monocular 3D detection method in 2019**

<table>
<thead>
<tr>
<th>Method</th>
<th>code</th>
<th>Easy</th>
<th>Hard</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>MonoGRNet</td>
<td></td>
<td>12.90</td>
<td>11.29</td>
<td>11.34</td>
<td>0.04s</td>
</tr>
</tbody>
</table>


- KITTI 3D detection leaderboard (car category only, snapshot in 2019)
- Why such a gap?
Why Such a Big Performance Gap?

Direct regression (Chen et al. CVPR ’16)
- Dominant monocular approach in 2018

Frustum PointNet (Qi et al. CVPR ’18)
- Two-stage LiDAR-based method

• Similarity: 2D box proposals + ROI features + box regression
• Difference: ROI features are in 2D or 3D spaces
• Combine the best of two worlds?
  • Borrow successful LiDAR-based architecture
  • Estimate the depth or point cloud
Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

- Depth estimation -> pseudo-LiDAR

- Instance masks -> point cloud frustums

- Bounding box consistency (BBC) loss
  - 2D proposals vs. projected 2D boxes

- Up to 4x improvement over prior art

<table>
<thead>
<tr>
<th>Method</th>
<th>AP\textsubscript{BEV} / AP\textsubscript{3D} (in %), IoU = 0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>Moderate</td>
</tr>
<tr>
<td>Mono3D [6]</td>
<td>5.2 / 2.5</td>
</tr>
<tr>
<td>Deep3DBox [30]</td>
<td>10.0 / 5.6</td>
</tr>
<tr>
<td>MLF-MONO [58]</td>
<td>22.0 / 10.5</td>
</tr>
<tr>
<td>Ours</td>
<td>41.9 / 31.5</td>
</tr>
</tbody>
</table>

Xinshuo Weng, Kris Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019
Detection Are Accurate for Near and Mid Range Objects
Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

- 1st place on KITTI monocular 3D detection for a while in 2019
- Impact: opened the field of pseudo-LiDAR based 3D detection
  - Alone with another pseudo-LiDAR approach (Wang et al, CVPR ‘19)

Most of them are pseudo-LiDAR based methods

Image credit to Patrick Langechuan Liu, [https://towardsdatascience.com/monocular-3d-object-detection-in-autonomous-driving-2476a3c7f57e](https://towardsdatascience.com/monocular-3d-object-detection-in-autonomous-driving-2476a3c7f57e)
Modular Improvement for Perception and Prediction

Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

Efficient 3D Multi-Object Tracking with 3D Kalman Filter

Social-Aware Representation Learning for 3D Multi-Object Tracking

Agent-Aware Transformer (Joint Social & Temporal Modeling + Preserve Time & Agent Information)

Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting

Xinshuo Weng, Kris Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019

Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
3D Multi-Object Tracking is Under-Explored

• Observations
  • Only a few publications per year before 2019
  • No commonly used public code base
  • Increased complexity and giving less attention to practical considerations
  • Lack of 3D MOT evaluation

• Motivation
  • Provide a public code base with a strong baseline to standardize 3D MOT evaluation

• KITTI 2D MOT evaluation
• Only 5 of methods included are for 3D MOT
AB3DMOT: A Baseline for 3D Multi-Object Tracking

- The state of Kalman filter is extended to 3D
  - 3D location, 3D velocity, 3D size, heading orientation
- Runtime speed on KITTI: 207 FPS (Car), 470 FPS (Pedestrians), 1242 FPS (Cyclists)
AB3DMOT is Accurate for Short-Term Tracking but Not Very Robust to Identity Switches
Missing Standardized 3D MOT Evaluation

- Commonly used benchmark: KITTI
  - nuScenes, Waymo, Argoverse datasets not released before 2019
- However, KITTI only provides 2D MOT evaluation
- The common practice for evaluating 3D MOT methods is
  - Project 3D boxes to the image plane
  - Run KITTI 2D evaluation code with a 2D IoU matching criteria

Our solution: replace the matching criteria with 3D IoU or distance
AB3DMOT Has Been Widely Used in the Community

• Baseline and evaluation code have received 1,200 stars up to date
  • [https://github.com/xinshuoweng/AB3DMOT](https://github.com/xinshuoweng/AB3DMOT)

• 3D MOT evaluation used in nuScenes tracking challenge

• Baseline adopted by others to win various challenges or serve as a baseline
  • nuScenes, Argoverse, Waymo, JRDB, etc

Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020
Modular Improvement for Perception and Prediction

Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

Efficient 3D Multi-Object Tracking with 3D Kalman Filter

Social-Aware Representation Learning for 3D Multi-Object Tracking

Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting

Xinshuo Weng, Kris Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019

Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
Accurate Data Association Requires Discriminative Features

- Limitations in prior work
  - Features are not discriminative

- Why?
  - Independent feature extraction
  - Single-modality feature extraction

- Our solutions
  - Feature interaction via GNNs
  - Multi-modal feature extraction

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020
Discriminative Feature Learning
Message Passing and Triplet Loss

- Affinity-aware message passing
  \[ n_i' = \sigma_1(n_i) + \sum_{j \in \mathcal{N}(i)} \sigma_2(\mathcal{A}_{ij}(n_j - n_i)) \]
  - MLPs
  - Affinity value
  - Neighborhood of node \( i \)
  - Node feature \( i \)
  - Node \( i \)
  - Node \( j \)
  - Node \( r \)
  - Feature space

- Push object features closer if their affinity value is large, \textit{i.e.}, likely have the same ID

- Batch triplet loss
  \[ \mathcal{L} = \max \left( ||n_i - n_j|| - \min_{s \in \mathcal{N}(i)} ||n_i - n_s|| - \min_{r \in \mathcal{N}(j)} ||n_r - n_j|| + \alpha, 0 \right) \]
  - Push object features closer if have the same ID,
  - Push object features further away if they have different IDs
More Identity-Consistent Multi-Object Tracking

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020
Modular Improvement for Perception and Prediction

Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud

Efficient 3D Multi-Object Tracking with 3D Kalman Filter

Social-Aware Representation Learning for 3D Multi-Object Tracking

Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting

Xinshuo Weng, Kris Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019

Xinshuo Weng, Jianren Wang, David Held, Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani. GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR 2020

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
Joint Social-Temporal Representation Learning for Multi-Agent Trajectory Prediction

Multi-Agent Trajectories

Agent 1

Agent 2

Agent 3

Multi-Agent Trajectories

Standard Multi-Agent Trajectory Models

Agent-Aware Transformer

(Joint Social & Temporal Modeling + Preserve Time & Agent Information)

Our Multi-Agent Trajectory Model

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
Loss of Information by Flattening

Agent-Aware Transformer
(Joint Social & Temporal Modeling + Preserve Time & Agent Information)

- Tokens are permutation-invariant in self-attention (no order of information)
- Loss of time  ➔  Temporal positional encoding
- Loss of agent identity  ➔  Agent-Aware Attention

Ye Yuan, Xinshuo Weng, Yanglan Ou, Kris Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021
Agent-Aware Attention

Two sets of keys & queries:

\[ Q_{self} = QW_{self}^{Q}, \quad K_{self} = KW_{self}^{K} \]
\[ Q_{other} = QW_{other}^{Q}, \quad K_{other} = KW_{other}^{K} \]

Mask \( M \)

\( i \)

\( j \)

Attention Weight Matrix \( A \)
Scene-Consistent Prediction & Attention Visualization

(a) Sample 1

(b) Sample 1 (Attention)

(c) Sample 2

(d) Sample 3
A Modular Yet Highly-Integrated Pipeline from Partial Integration towards Full Integration

Tracking

Detection

Prediction

Affinity-based Prediction

Weng, Ivanovic, Kitani, Pavone. arXiv ’21
Weng*, Yuan*, Kitani. ICRA/RA-L ’21

Weng, Wang, Levine, Kitani, Rhinehart. CoRL ’20

Weng, Nan, Lee, McAllister, Gaidon, Rhinehart, Kitani. In submission ‘21

Weng, Kitani. arXiv ’21
Wang, Kitani, Weng. ICRA ’21
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

Automatic and Dynamic Detection Selection for Tracking

Perception-then-Prediction Integration

Joint Detection and Tracking with Graph Neural Networks

Parallelized Tracking and Prediction

Towards Full Integration with Prediction-then-Perception

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
Automatic and Dynamic Detection Selection for 3D Tracking

One of the most critical hyper-parameters: *detection threshold*

- Manual threshold search is cumbersome
- Using a fixed threshold is sub-optimal  
  
  Require category-specific ablation study
  Optimal threshold may vary across frames/objects

- Learn an adaptive threshold?

Optimal Detection Threshold Varies Across Frames

- Preserve true positive (TP) tracked objects (blue overlaid green) with low scores
- Filter false positive (FP) tracked objects (red) with high scores

Detection Threshold Is Indeed Dynamic

• Learned per-frame thresholds over frames
• Trends are smooth (optimal thresholds gradually change)
Sometimes Even Per-Frame Threshold is Not Sufficient

- We might not be able to separate TPs and FPs in a single frame
- Instance-level threshold might be necessary (a true positive classifier?)

Sometimes Even Per-Frame Threshold is Not Sufficient

Instance-level AutoSelect

Filter out high-score FPs and preserve low-score TPs in the same frame
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

**Perception-then-Prediction Integration**

- **Detection** → **Tracking**
- **Tracking** → **Prediction**

**Multi-hypothesis Tracking and Prediction**

- **Detection** → **Tracking** → **Prediction**
- **Tracking** → **Detection** → **Prediction**

**Affinity-based Prediction**

**Joint Detection and Tracking with Graph Neural Networks**

- **Detection** → **Feature Extraction** → **Tracking**
- **Tracking** → **Feature Extraction** → **Prediction**

**Parallelized Tracking and Prediction**

**Perception-Prediction Parallelization**

**Towards Full Integration with Prediction-then-Perception**

- **Detection** → **Prediction** → **Tracking**
- **Prediction** → **Detection** → **Tracking**

**References**

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
Trajectory Prediction is Challenging in the Wild

- Key difference: using tracking results as inputs (not GT past trajectories)
- Prediction errors can be caused due to upstream perception errors

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Using GT past trajectories as inputs

Using past tracklets from tracking as inputs

- Inconsistent predictions (sudden velocity & orientation change) caused by identity switches

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Prediction is not Robust to Tracking Errors

- If using past GT trajectories as inputs, prediction errors are low
- If using past tracklets from tracking as inputs, prediction errors are large (28x)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Eval. Targets (# of obj)</th>
<th>Inputs to Prediction</th>
<th>minADE&lt;sub&gt;k&lt;/sub&gt;</th>
<th>minFDE&lt;sub&gt;k&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI</td>
<td>Objects with IDS (33)</td>
<td>GT past trajectories</td>
<td>0.100</td>
<td>0.171</td>
</tr>
<tr>
<td>KITTI</td>
<td>Objects with IDS (33)</td>
<td>past tracklets</td>
<td>2.820</td>
<td>4.514</td>
</tr>
<tr>
<td>KITTI</td>
<td>Objects with FRAG (330)</td>
<td>GT past trajectories</td>
<td>0.177</td>
<td>0.306</td>
</tr>
<tr>
<td>KITTI</td>
<td>Objects with FRAG (330)</td>
<td>past tracklets</td>
<td>1.621</td>
<td>2.155</td>
</tr>
<tr>
<td>nuScenes</td>
<td>Objects with IDS (4160)</td>
<td>GT past trajectories</td>
<td>0.473</td>
<td>0.825</td>
</tr>
<tr>
<td>nuScenes</td>
<td>Objects with IDS (4160)</td>
<td>past tracklets</td>
<td>8.345</td>
<td>13.892</td>
</tr>
<tr>
<td>nuScenes</td>
<td>Objects with FRAG (3365)</td>
<td>GT past trajectories</td>
<td>0.621</td>
<td>1.108</td>
</tr>
<tr>
<td>nuScenes</td>
<td>Objects with FRAG (3365)</td>
<td>past tracklets</td>
<td>14.520</td>
<td>21.815</td>
</tr>
</tbody>
</table>

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Multi-hypothesis Tracking and Prediction

A list of data association hypotheses

- By reasoning about multiple possible hypotheses of tracking results
  - Increase the likelihood of including accurate tracking results for prediction
- Assumption: different hypotheses are complimentary
MTP: Multi-hypothesis Tracking and Prediction

- Different hypothesis is indeed complimentary
  - The No. of tracking errors persistent in all hypotheses are reduced with more hypotheses
- Significantly improve the accuracy metrics: minADE/FDE

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Targets</th>
<th>Methods</th>
<th>minADE$k$</th>
<th>minFDE$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI IDS</td>
<td>STP, H=1, k=20</td>
<td>MTP (Ours), H=5, k=100</td>
<td>1.099</td>
<td>1.768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=10, k=200</td>
<td>0.844</td>
<td>1.332</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=20, k=400</td>
<td>0.707</td>
<td>1.093</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=5, k=20, sampling</td>
<td>1.118</td>
<td>1.802</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=10, k=20, sampling</td>
<td>0.876</td>
<td>1.390</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=20, k=20, sampling</td>
<td>0.747</td>
<td>1.173</td>
</tr>
<tr>
<td>KITTI FRAG</td>
<td>STP, H=1, k=20</td>
<td>MTP (Ours), H=5, k=100</td>
<td>1.436</td>
<td>1.862</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=10, k=200</td>
<td>1.385</td>
<td>1.765</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=20, k=400</td>
<td>1.305</td>
<td>1.627</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=5, k=20, sampling</td>
<td>1.448</td>
<td>1.888</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=10, k=20, sampling</td>
<td>1.404</td>
<td>1.801</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MTP (Ours), H=20, k=20, sampling</td>
<td>1.335</td>
<td>1.688</td>
</tr>
</tbody>
</table>

4x of improvements with MTP
MTP Increases the Likelihood of Preserving Accurate Predictions

Single-hypothesis perception-then-prediction (Inaccurate)

MTP (Ours) More accurate predictions

Only the best prediction sample is visualized

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

Perception-then-Prediction Integration

Towards Full Integration with Prediction-then-Perception

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
The step of computing affinity matrices is necessary
- But is the “hard” matching step really needed?

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Affinity-based Prediction

Directly using affinity matrices as inputs

Affinity matrices

Data Association

Trajectory Prediction

Objects’ past tracklets

Removed, no past tracklets anymore

Predictions

Uncertainty propagation between tracking and prediction

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Proposed Approach

Builds on top of AgentFormer

- Tokens are now extracted from each individual detection
- Joint social & temporal modeling is preserved

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Agent-Aware Attention -> Affinity-based Attention

Two sets of keys & queries:

\[ Q_{\text{self}} = QW_{\text{self}}^Q, \quad K_{\text{self}} = KW_{\text{self}}^K \]
\[ Q_{\text{other}} = QW_{\text{other}}^Q, \quad K_{\text{other}} = KW_{\text{other}}^K \]

\[ M \odot Q_{\text{self}} K_{\text{self}}^T \]
\[ (1 - M) \odot Q_{\text{other}} K_{\text{other}}^T \]

Attention Weight Matrix \( A \)

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

Automatic and Dynamic Detection Selection for Tracking
Multi-hypothesis Tracking and Prediction
Affinity-based Prediction

Perception-then-Prediction Integration

Joint Detection and Tracking with Graph Neural Networks
Parallelized Tracking and Prediction

Towards Full Integration with Prediction-then-Perception

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
Limitation of the Cascaded Pipeline

• Slower inference speed (redundant modules)
• Cascading errors
• Can we go beyond the cascaded pipeline?

Inaccurate predictions when using erroneous tracklets as inputs
Task Parallelization & Merge Similar Modules

Sensor Data → 3D Object Detection

3D Multi-Object Tracking
- Feature Extraction
- Matching

Trajectory Forecasting
- Feature Extraction
- Trajectory Decoder

Sensor Data → 3D Object Detection

Shared Feature Learning

3D Multi-Object Tracking
- Matching

Trajectory Forecasting
- Trajectory Decoder

Similar components, which aims to encode object features from past information

Module-specific components

Cascaded Pipeline

Parallelized Tracking and Prediction Pipeline

Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and DiversitySampling. ICRA 2021
Parallelized Tracking and Forecasting

- Prediction does not depend on tracking results (two parallel heads)
- Improve computational efficiency by sharing the feature learning process
Posterior Collapse in CVAE-based Prediction

Learn a generative model $p_{\theta}(x|\psi)$

Data

Context feature $\psi$

Future trajectories $x$

Low sample efficiency!

Generator $G_{\theta}(x|z, \psi)$

Latent space

Trajectory Space

Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Diversity Sampling Function

$S_\gamma(\psi)$

Latent space

Diversity Sampling Function (DSF)

Latent codes $\{z_1, \ldots, z_N\}$

Generator $G_\theta(x|z, \psi)$

Optimizer

Data

Context feature $\psi$

Future trajectories $x$

Diversity loss on samples $\{x_1, \ldots, x_N\}$

Trajectory Space

Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
## Results on KITTI

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Metrics</th>
<th>w/o MOT+DSF</th>
<th>w/o DSF</th>
<th>+App</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADE↓</td>
<td>0.663</td>
<td>0.582</td>
<td>0.692</td>
<td><strong>0.471</strong></td>
</tr>
<tr>
<td>KITTI-1.0s</td>
<td>FDE↓</td>
<td>1.121</td>
<td>0.978</td>
<td>1.085</td>
<td><strong>0.763</strong></td>
</tr>
<tr>
<td></td>
<td>ASD↑</td>
<td>1.796</td>
<td>1.730</td>
<td>1.683</td>
<td><strong>2.351</strong></td>
</tr>
<tr>
<td></td>
<td>FSD↑</td>
<td>3.168</td>
<td>3.052</td>
<td>3.194</td>
<td><strong>4.071</strong></td>
</tr>
<tr>
<td></td>
<td>ADE↓</td>
<td>1.729</td>
<td>1.564</td>
<td>2.104</td>
<td><strong>1.319</strong></td>
</tr>
<tr>
<td>KITTI-3.0s</td>
<td>FDE↓</td>
<td>3.086</td>
<td>2.893</td>
<td>3.536</td>
<td><strong>2.299</strong></td>
</tr>
<tr>
<td></td>
<td>ASD↑</td>
<td>3.196</td>
<td>3.416</td>
<td>2.951</td>
<td><strong>5.843</strong></td>
</tr>
<tr>
<td></td>
<td>FSD↑</td>
<td>5.776</td>
<td>6.168</td>
<td>5.895</td>
<td><strong>10.123</strong></td>
</tr>
</tbody>
</table>

- Joint training with tracking improves prediction performance
- DSF increases the pairwise distance between prediction samples
- PTP can simultaneously track and predict trajectories
Joint Detection and Multi-Object Tracking with Graph Neural Networks

- Replace prediction with a detection head
- Replace detection inputs with anchors
- Cannot remove the detection-tracking dependency during testing
Graph Representation Learning can Improve Performance

FairMOT (IJCV ’21)
Joint detection and tracking without GNNs

Our GSDT
Joint detection and tracking with GNNs

Objects that are missing but detected in the other are marked with **bold**
*The more bold boxes, the worse performance it has*

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Joint Detection and Tracking Optimization Is Also Critical

MPNTrack (CVPR ‘20)
GNN-based data association

Our GSDT -- Joint detection and tracking with GNNs

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
A Modular Yet Highly-Integrated Pipeline
from Partial Integration towards Full Integration

Automatic and Dynamic Detection Selection for Tracking
Multi-hypothesis Tracking and Prediction
Affinity-based Prediction

Perception-then-Prediction Integration

Feature Extraction
Detection
Tracking
Prediction
Joint Detection and Tracking with Graph Neural Networks
Parallelized Tracking and Prediction
Perception-Prediction Parallelization

Towards Full Integration with Prediction-then-Perception

Yongxin Wang, Kris Kitani, Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021
Xinshuo Weng*, Ye Yuan*, Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting

- Perception-then-prediction pipeline:
  - Detection → MOT → Trajectory Forecasting

- Prediction-then-perception pipeline
  - Sequential Pointcloud Forecasting → Detection → MOT

- Differences
  - Invert the order of prediction
  - Prediction at the sensor data level (not object positions)

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF²: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Advantages?

• Does not require human annotation for prediction
• Point cloud data is prevalent nowadays
• The key is this new task -- Sequential Pointcloud Forecasting (SPF)
Performance Scaling with More Point Cloud Data

- CD: Chamfer distance (lower is better)
- EMD: Earth Mover’s Distance (lower is better)
SPFNet

- How to learn representations from large-scale point cloud sequences?
- Baseline: FC-LSTM autoencoder model
- Four modules
  - Frame-wise point cloud encoder
  - FC-LSTM for temporal modeling
  - Frame-wise point cloud decoder
  - Losses: chamfer distance

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Predicted Point Clouds Preserves Objects

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Detection on Predicted Point Clouds

- Green: detected 3D boxes
- Yellow: GT 3D boxes
- Detections mostly match with GT on predicted point clouds from SPFNet

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Tracking on Predicted Point Clouds

- Color represents the object identity
- Predicted point clouds from SPFNet preserves objects
SPF² Outperforms in-the-Wild Perception-then-Prediction

<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>

• Evaluation is performed for end-to-end prediction
• Prediction-then-perception has the potential for better full integration
Deterministic & Blurry Predictions

- The use of a deterministic model (LSTM autoencoder)
- High compression of information (feature vectorization)

Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020
Stochastic Predictions with Temporally-Dependent CVAE

Prior work

Time

Point cloud sequence

Deterministic model

Single prediction

A similar LSTM autoencoder but with latent space

Sample a sequence of latent code at every timestamp for decoding

Temporal dependency to ensure consistent prediction over time

Sample another sequence of latent code…

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2021
Proposed Works
Project 1: Enable Agent-Controllability and Diverse Prediction for the Prediction-then-Perception Pipeline

Limited Diversity in Point Cloud Prediction

Two limitations

- Using a single latent space makes it challenging to disentangle diverse behavior generation for each object from background
- Random sampling can easily lead to posterior collapse

4 future prediction samples in one frame

Hard to see visible differences across samples (red -> large error w.r.t GT)
Prediction with Context-Object Split Latent Spaces
Control the Future Behavior Generation for Each Object!

Scene Decomposition → Scene Composition

Encoder: Off-the-shelf panoptic seg. Mask
Decoder: Off-the-shelf Panoptic seg. Mask

Context latent space for background
Temporal dependency
Foreground object latent spaces

Can replace panoptic segmentation mask with unsupervised clustering techniques to preserve self-supervised training scheme.
How to Ensure Consistency and Diversity of Objects’ Behaviors?

Can we sample object behavior that is consistent with other objects and background?
Can we predict point clouds of objects and background that cover multiple modes of future?

Collision between sampled motion of different foreground objects
Inconsistency!

Similar sampled motions for the same object latent space
Posterior collapse!
Diverse Prediction with Scene Consistency

A hierarchical sampling technique

Scene-Consistent Sampling (SCS)

Global sampling to ensure diversity (similar to DSF proposed in PTP)

Local sampling to ensure consistency between all objects and background

Temporal dependency

\[ z_1, z_2, z_3, \ldots, \text{Encoder} \]

\[ z_1, z_2, z_3, \ldots, \text{Scene Decomposition} \]

\[ z_1, z_2, z_3, \ldots, \text{Scene Composition} \]

\[ z_1, z_2, z_3, \ldots, \text{Decoder} \]
Project 2: End-to-End Prediction-then-Perception Framework for Driving in Simulation

Is the Learned Latent Space Universal? Universal Latent Representation for End-to-End Prediction-then-Perception

End-to-end training for the entire pipeline
Ego’s waypoint prediction depends on both future trajectories of surrounding objects and the learned universal latent representation.
Reducing Prediction Uncertainty with Action-Conditioned Prediction

Scene Decomposition → Global latent → Scene Composition

local latent

Global latent

Point cloud prediction
Panoptic segmentation
Trajectories
Ego’s waypoint prediction

Steer
Gas
Brake
Control commands
Action state
Closed-Loop Driving in Simulation with Joint Prediction-then-Perception for Planning

End-to-End Prediction-then-Perception for Planning

Controller (e.g., autopilot in Carla)

Simulator (e.g., Carla)

Updated state and sensor data

Ego’s waypoint prediction

Action-conditioned prediction

Control commands

Steer

Gas

Brake
Summary

• Modular improvement
  • Improving each individual module in isolation

• A modular yet highly-integrated pipeline
  • From partial integration towards full integration

• Proposed works
  • (Dec. 21 – Mar. 22) Enable agent-controllability and diverse prediction for the prediction-then-perception pipeline
  • (Mar. 22 – Jun. 22) End-to-end prediction-then-perception framework for driving in simulation
References (Presented Works)

• **Weng**, Kitani. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW 2019

• **Weng**, Wang, Held, Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS 2020

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

• **Weng**, Wang, Levine, Kitani, Rhinehart. Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020

• Wang, Kitani, **Weng**. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. ICRA 2021

• **Weng***, Yuan*, Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. ICRA 2021

• Yuan, **Weng**, Ou, Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV 2021


• **Weng**, Ivanovic, Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. arXiv 2021

• **Weng**, Nan, Lee, McAllister, Gaidon, Rhinehart, Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In submission 2021

• **Weng**, Ivanovic, Kitani, Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-Based Prediction. In submission 2021
Open Questions

• Which integration is the best for the modularized pipeline?
  • Perception-then-prediction
  • Parallelized perception and prediction
  • Prediction-then-perception

• Integration with sensor measurement uncertainty

• Modularized pipeline vs. direct planning from sensor data
A Modular Yet Highly-Integrated Pipeline for Perception and Prediction

Xinshuo Weng

www.xinshuoweng.com

Robotics Institute, Carnegie Mellon University

Thank you!