Towards Modular and Differentiable Autonomous Driving

Xinshuo Weng
Robotics Institute, Carnegie Mellon University

RI PhD Thesis Defense
June 7, 2022

Committee Members:
Kris Kitani (Chair), Matthew P. O'Toole, Deva Ramanan, Marco Pavone
What does it mean by a modular and differentiable autonomy stack?

Modular design for interpretability?

End-to-end training?
Traditional Autonomy Stack - Sequential Cascading

- LiDAR
- RGB

1. Sensors & Map
2. Object Detection
3. Multi-Object Tracking
4. Motion Prediction
5. Planning & Control

Map
Traditional Autonomy Stack - Sequential Cascading

1. Sensors & Map
2. Object Detection
3. Multi-Object Tracking
4. Motion Prediction
5. Planning & Control

Detection results
Traditional Autonomy Stack - Sequential Cascading

- Sensors & Map
- Object Detection
- Multi-Object Tracking
- Motion Prediction
- Planning & Control

Tracking results
Traditional Autonomy Stack - Sequential Cascading

- Sensors & Map
- Object Detection
- Multi-Object Tracking
- Motion Prediction
- Planning & Control

Forecasting results
Traditional Autonomy Stack - Sequential Cascading

- Sensors & Map
- Object Detection
- Multi-Object Tracking
- Motion Prediction
- Planning & Control
Traditional Autonomy Stack - Sequential Cascading

Advantages?
- Interpretability
- Easy to develop with separate teams
- etc.

Disadvantages
- Compounding errors
- Slower inference speed
- Not optimized for the end task
- etc.
Direct Planning / End-to-End Approach

Advantages?
- No explicit compounding errors
- Simplicity & Faster speed
- Optimized for the end task
- etc.

Disadvantages
- Lack of interpretability
- Data hungry
- etc.
Can We Combine the Best of Two Worlds?

- End-to-end optimization
- Retain modular design for interpretability

Interpretable Neural Motion Planning

Image cropped from Raquel Urtasun’s talk at CVPR ’21 WAD
Other alternatives or directions for autonomy stack integration?
Uncertainty Propagation through the Stack

Forward propagation with *uncertainties*

- Which uncertainty should we use?
- How can we incorporate it in downstream modules?

Back propagation

1. Sensors & Map
2. Object Detection
3. Multi-Object Tracking
4. Motion Prediction
5. Planning & Control
What is the Best Order/Structure?

For example,

- Perception-then-Prediction
- Perception-Prediction Parallelization
- Prediction-then-Perception
My research aims to build an integrated and modular autonomous system that

• Can propagate different kinds of useful uncertainties

• Has been optimized with the best ordering/structure

More robust and more effective
High performance of individual Modules is also important!
Modular Improvements for Perception and Prediction

Detection

Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW '19

When We First Met: Visual-Inertial Person Localization for Co-Robot Rendezvous. IROS '20

3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS '20

GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR '20

Re-ID and Tracking

Multi-Modality Task Cascade for 3D Object Detection. BMVC '21

Multi-Echo LiDAR for 3D Object Detection. ICCV '21

AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. ICCV '21

Prediction

Multi-Echo LiDAR for 3D Object Detection. ICCV '21

Learning Shape Representations for Person Re-Identification under Clothing Change. WACV '21

Visio-Temporal Attention for Multi-Camera Multi-Target Association. ICCV '21

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
Towards modular and differentiable autonomous driving

- Uncertainty propagation through the stack
- Optimization of the stack ordering

**AutoSelect**
- Automatic & Dynamic Detection Selection for Tracking (AutoSelect)
  - Weng, Kitani. arXiv ‘21

**MTP**
- Multi-hypothesis Tracking and Prediction (MTP)
  - Weng, Ivanovic, Pavone. IV ‘22

**Affinipred**
- Affinity-based Prediction (Affinipred)
  - Weng, Ivanovic, Kitani, Pavone. CVPR ‘22
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. IV 2022
Object Detection and Tracking Is Not Perfect

- Real-world results on KITTI (10 FPS)
- Can be even worse on nuScenes with a lower FPS
Prediction errors increase by \textbf{10-30x} in the presence of tracking errors.
Multi-hypothesis Tracking and Prediction

Prediction for each hypothesis

Multiple matching hypotheses

Explore different tracking results

Past tracklets + detections

Increase the likelihood of including accurate tracking results for prediction

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. IV 2022
Are Different Hypotheses Complimentary?

- Underlying assumption:
  - Using more hypotheses may include more accurate tracking results as inputs

\[ K = 1 \quad \text{vs} \quad K = 20 \]

The No. of tracking errors persistent in all hypotheses is reduced

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. IV 2022
MTP Increases Prediction Robustness

**TABLE II**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Targets</th>
<th>Methods</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>IDS</td>
<td>STP, H=1, k=20</td>
<td>2.820</td>
<td>4.514</td>
</tr>
<tr>
<td></td>
<td></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</td>
<td>FRAG</td>
<td>STP, H=1, k=20</td>
<td>1.621</td>
<td>2.155</td>
</tr>
<tr>
<td></td>
<td></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>

Prediction errors decrease by up to 70% by using our MTP

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. IV 2022
MTP Increases Prediction Robustness

Standard perception-prediction
Single hypothesis

Unstable

MTP (Ours)
Multiple hypotheses

More stable and robust

Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. IV 2022
Generating multiple hypotheses from the affinity matrix is useful!

Is that the best we can do?
Rethinking the Perception-then-Prediction Pipeline

- Is the “hard” matching step really needed?
- Also, loss of information for MTP

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022
Affinity-based Prediction (Affinipred)

Directly using affinity matrices as inputs

Propagate matching uncertainty to trajectory prediction

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone.
Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022
Separated Temporal/Social Modeling Does Not Apply

Multi-Agent Trajectories

Standard Multi-Agent Trajectory Models

Our inputs: raw detections, no identity & trajectory

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022
Affinipred with Joint Social-Temporal Transformer

Tokens are extracted from detections, no trajectory anymore

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone.
Inject Matching Uncertainty: Affinity-based Attention

Learn two sets of keys and queries: self-attention and cross-attention
Inject Matching Uncertainty: Affinity-based Attention

Affinity matrices between consecutive frames

Idealized setting with object identities available

Realistic setting using affinity representing the probability of having similar identities

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022
MTP extracts “more” information from affinity matrices.

Affinipred “fully” exploits the uncertainties in affinity matrices.

Error decrease by up to 88%!
The more tracking errors, the larger gain.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>minADE_S</th>
<th>minFDE_S</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITTI</td>
<td>PointRCNN+AB3DMOT+PTP, S=20</td>
<td>0.185</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>MTP [40], S=20</td>
<td>0.162</td>
<td>0.238</td>
</tr>
<tr>
<td></td>
<td>MTP [40], S=400</td>
<td>0.146</td>
<td>0.203</td>
</tr>
<tr>
<td></td>
<td>Affinipred (Ours), S=20</td>
<td>0.129</td>
<td>0.194</td>
</tr>
<tr>
<td>nuScenes</td>
<td>Megvii+AB3DMOT+PTP, S=10</td>
<td>2.320</td>
<td>3.819</td>
</tr>
<tr>
<td></td>
<td>MTP [40], S=10</td>
<td>1.585</td>
<td>2.512</td>
</tr>
<tr>
<td></td>
<td>MTP [40], S=200</td>
<td>1.325</td>
<td>1.979</td>
</tr>
<tr>
<td></td>
<td>Affinipred (Ours), S=10</td>
<td>0.977</td>
<td>1.628</td>
</tr>
</tbody>
</table>

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone.
Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022.
Affinipred Is More Robust to Perception Noises

Standard tracking-prediction

Error due to tracklet orientation change

Affinipred

Prediction robust to tracking errors

Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone. Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022
Towards modular and differentiable autonomous driving

- Uncertainty propagation through the stack
- Optimization of the stack ordering

Perception-Prediction Parallelization

Wang, Kitani, Weng. ICRA ’21
Weng*, Yuan*, Kitani. ICRA ‘21

Prediction-first approach e.g., Prediction-then-Perception

Weng, et al. CoRL ’20
Weng, et al. in submission ‘22
Weng, et al. in submission ‘22
• Differences
  • Predicting sensor data rather than object trajectories
  • Prediction before perception
• Major advantage: self-supervised training for prediction

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
Detections on predicted point clouds match with GT at a close distance
Tracking on Predicted Point Clouds

Stable tracking of detected objects on predicted point clouds

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
Simple Baseline: SPFNet
LSTM Auto-encoder

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
Limitations: Deterministic & Blurry Predictions

- The use of a deterministic model (LSTM autoencoder)
- High compression of information (feature vectorization)
S2Net: Stochastic Sequential Pointcloud Forecasting

\[ z_0 \rightarrow z_1 \rightarrow z_2 \]

Temporally-Dependent Latent Variables

Consistent future prediction

LSTM

Frame 0 Past Future Frame 1

Sample K times

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
Uncertainty Is Mostly Around the Boundary

GT Point Cloud

Prediction with Uncertainty (K=5)

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
Uncertainty Can Potentially Help Downstream Prediction with Uncertainty (K=5)

3D detection on predicted point clouds

<table>
<thead>
<tr>
<th>Inputs</th>
<th>mAP</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x, y, z)</td>
<td>69.96</td>
<td>76.44</td>
<td>72.81</td>
</tr>
<tr>
<td>(x, y, z) + uncertainty</td>
<td>70.41</td>
<td>78.48</td>
<td>72.76</td>
</tr>
</tbody>
</table>

Incorporating uncertainty -> higher precision

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani.
S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
Limitations: Deterministic & Blurry Predictions

- The use of a deterministic model (LSTM autoencoder)
- High compression of information (feature vectorization)
Simple Feature Pyramid Fails to Converge

Feature Pyramid

Reduced to learn residual in the hidden space

LSTM

Frame Pyramid

Frame 0

Past

Future

Frame 1

Frame 1 (predicted)

Frame 2 (predicted)

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
S2Net with Time-Synchronized Pyramid LSTMs

Time-Synchronized Pyramid LSTMs

Resolve training instability

Frame 0
Past
Frame 1
Future
Frame 1 (predicted)
Frame 2 (predicted)

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani.
S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
More Fine-Grained Predictions with S2Net

Red points: high CD error
Cyan points: low CD error
Better structure!
Sharper boundary!
More accurate object positions!
More Fine-Grained Predictions with S2Net

Zoom-in visualization shows better shape details

Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022
Sequential Pointcloud Forecasting is applicable to detection, tracking and trajectory prediction

How about planning?
Most top methods are based on Imitation Learning

Scalability & efficiency

Only one method (No. 8) is based on Reinforcement Learning

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani. Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transformer. In Submission 2022
Combining Imitation with Auxiliary Tasks

Sensor inputs

**Supervised Semantic Task**

Driving Policy

Pixel/object labels

Although helpful:

1. Require labels
2. Non-predictive representations

- **Segmentation, CVPR ’17; IROS ’20**
- **Implicit Affordances, CVPR ’20**
- **3D Detection, CVPR ’22**

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Imitation with Predictive Representations

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
State-of-the-art Driving Performance on Carla

Closed-loop roll-out on unseen towns/routes

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani. Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transform. In Submission 2022
Learning Informative Representations via Joint Multi-Modal, Multi-View and Multi-Frame Attention

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFo 

SPIFF (Joint Attention)

Conv Blocks (RGB)
Conv Blocks (LiDAR)

Split
Flatten
Un-flatten

Self-supervised Predictive Task
Unlabeled Sensor data

Supervised Semantic Task
Pixel/object labels

Driving Policy

Single modality (RGB only)
Single view (front only)

Multi-modal + Multi-view + Multi-frame

Multi-modal
Multi-view
Multi-frame

Conv Blocks
Add & Layer norm
Feed forward
Add & Layer norm
Multi-head attention

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFo

In Submission 2022
Using Partial Observations Leads to Training Confusion

Large peak of errors during training caused by the partial observability

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Using Partial Observations Leads to Inertial Problem at Testing

Learning informative representations via joint attention over all observations is critical

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani. Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFormer. In Submission 2022
State-of-the-art Driving Performance on Carla

Informative + Predictive Representations

- Multi-modal
- Multi-view
- Multi-frame

Unlabeled sensor data

SPIFF (Joint attention)

Self-supervised Predictive Task

Methods | Average DS ↑
---|---
CILRS [32], ICCV ’19 | 8.22 ± 1.50
LBC [33], CoRL ’19 | 14.26 ± 1.90
Transfuser [2], CVPR ’21 | 25.02 ± 0.81
Rails [5], ICCV ’21 | 34.45 ± 1.39
LAV [1], CVPR ’22 (Concurrent) | 48.75 ± 2.16
SPIFF (ours) | 50.97 ± 1.84

Closed-loop evaluation on 26 Carla official testing routes

Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani. Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFo rm. In Submission 2022
Towards robust autonomy by integrating and propagating uncertainty

Exploring effective and differentiable autonomy stack by re-ordering
Lessons Learned and Open Problems

• Uncertainty matters
  • Keep “soft” information in hard decision making
  • Even uncalibrated uncertainty may help
  • The combinatorial effect of uncertainty is still unclear

• Point cloud forecasting may be useful but challenging
  • Improvements seen in the image domain typically apply
  • Architecture-level improvements needed for realistic and shape-aware predictions
  • Unclear the best way to integrate with the downstream tasks

• Cross-domain exploration may open doors for new research problems
Towards Modular and Differentiable Autonomous Driving

Xinshuo Weng

www.xinshuoweng.com

Robotics Institute, Carnegie Mellon University

Thank you!