Carnegie Mellon University
The Robotics Institute

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

THE ROBOTICS INSTITUTE

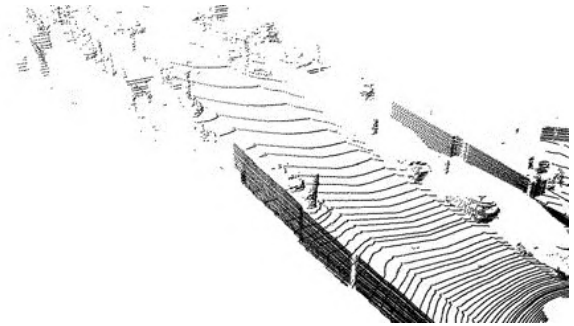
What does it mean by a modular and differentiable autonomy stack?

Modular design for interpretability ?

End-to-end training?

THE ROBOTICS INSTITUTE

Traditional Autonomy Stack - Sequential Cascading



LiDAR

RGB

Sensors & Map

↓

Object Detection

↓

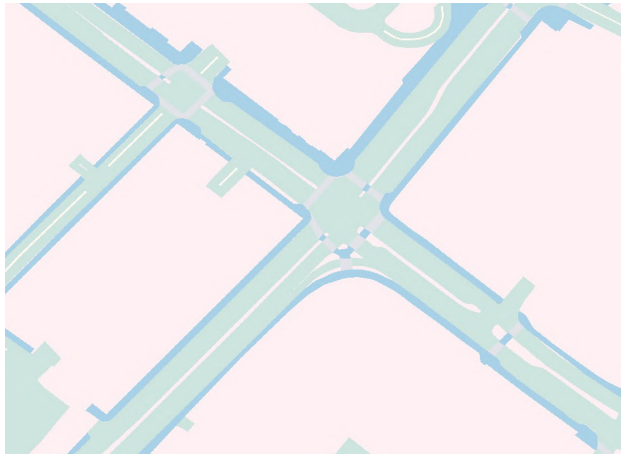
Multi-Object Tracking

↓

Motion Prediction

↓

Planning & Control



Map

THE ROBOTICS INSTITUTE

Traditional Autonomy Stack - Sequential Cascading

Sensors & Map

↓

Object Detection

↓

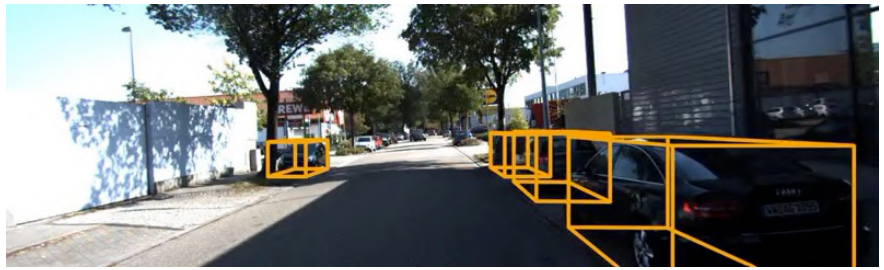
Multi-Object Tracking

↓

Motion Prediction

↓

Planning & Control



Detection results

Traditional Autonomy Stack - Sequential Cascading

Sensors & Map

↓

Object Detection

↓

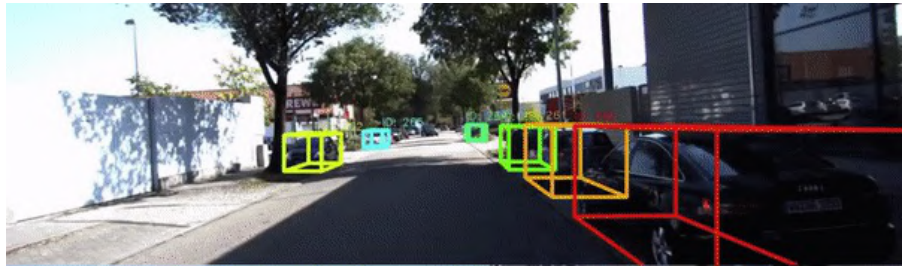
Multi-Object Tracking

↓

Motion Prediction

↓

Planning & Control



Tracking results

Traditional Autonomy Stack - Sequential Cascading

Sensors & Map

↓

Object Detection

↓

Multi-Object Tracking

↓

Motion Prediction

↓

Planning & Control

Forecasting results

THE ROBOTICS INSTITUTE

Traditional Autonomy Stack - Sequential Cascading

Sensors & Map

↓

Object Detection

↓

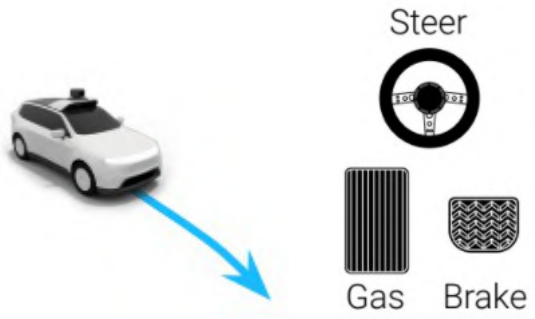
Multi-Object Tracking

↓

Motion Prediction

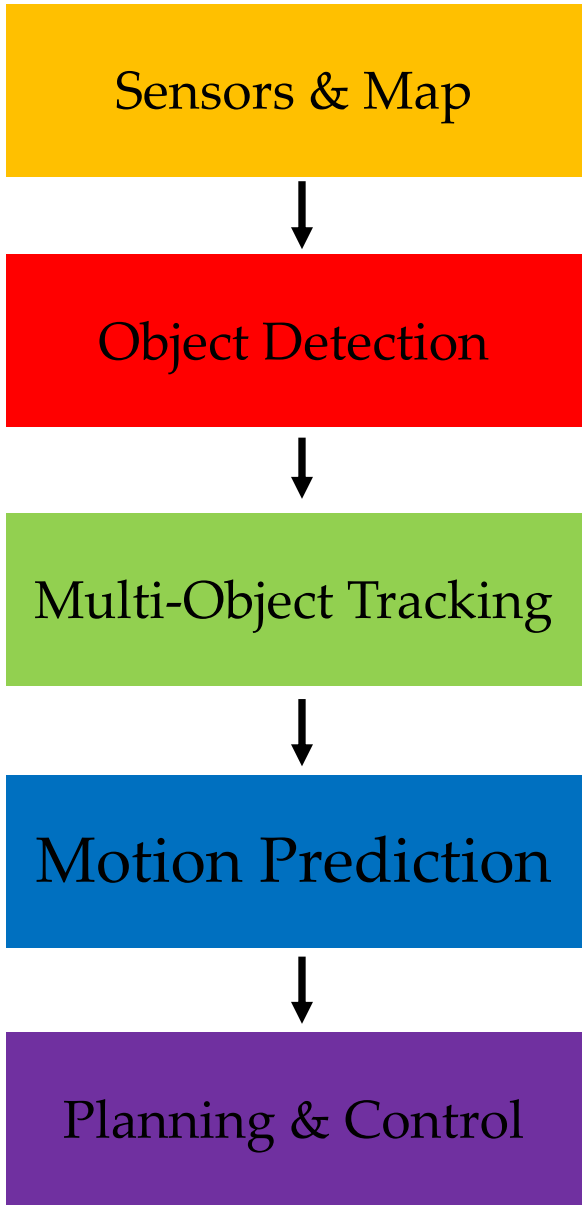
↓

Planning & Control



THE ROBOTICS INSTITUTE

Traditional Autonomy Stack - Sequential Cascading



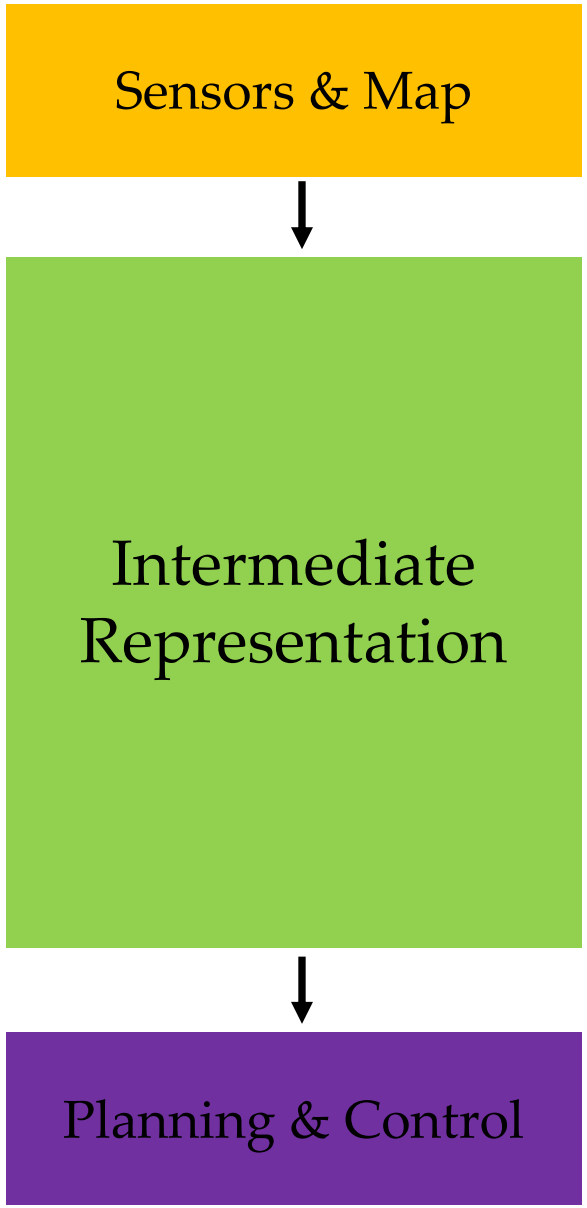
Advantages?

- 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



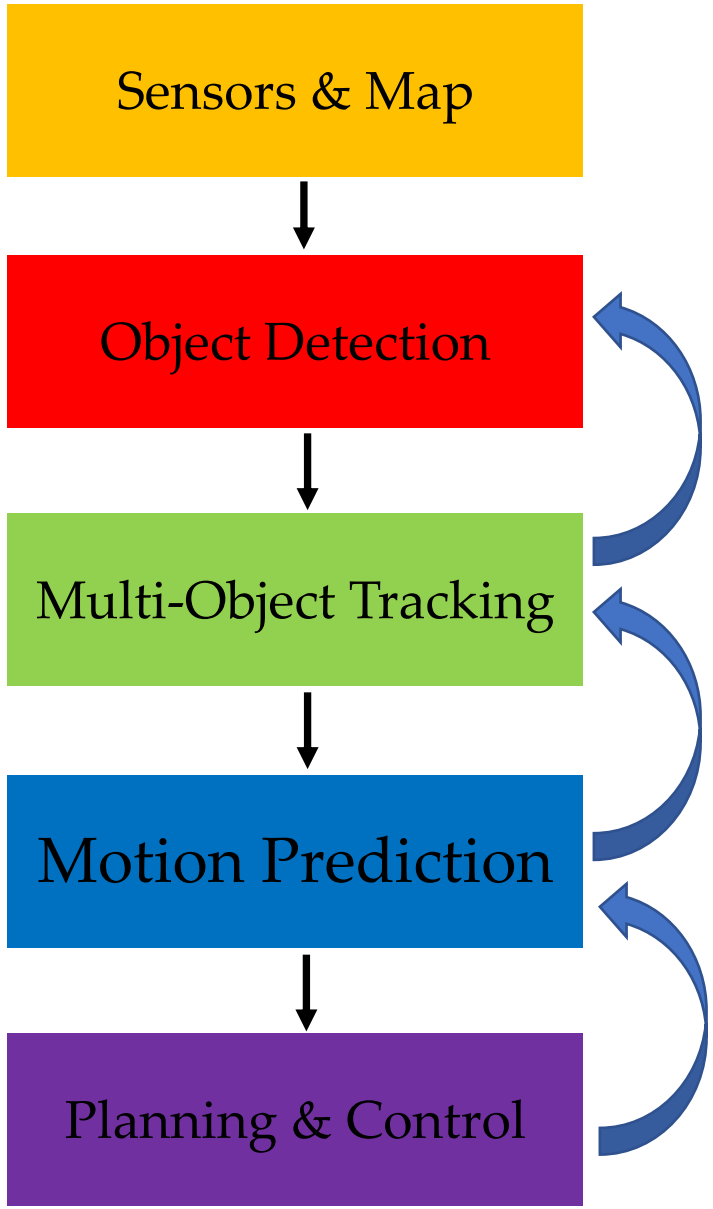
Advantages?

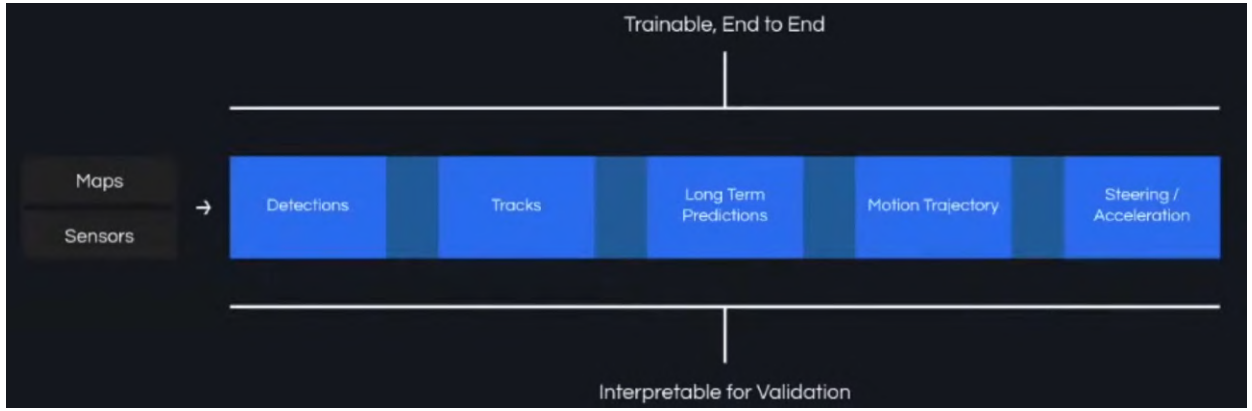
- No explicit compounding errors
- Simplicity & Faster speed
- Optimized for the end task
- *etc.*

Disadvantages

- Lack of interpretability
- Data hungry
- *etc.*

Can We Combine the Best of Two Worlds?





Interpretable Neural Motion Planning

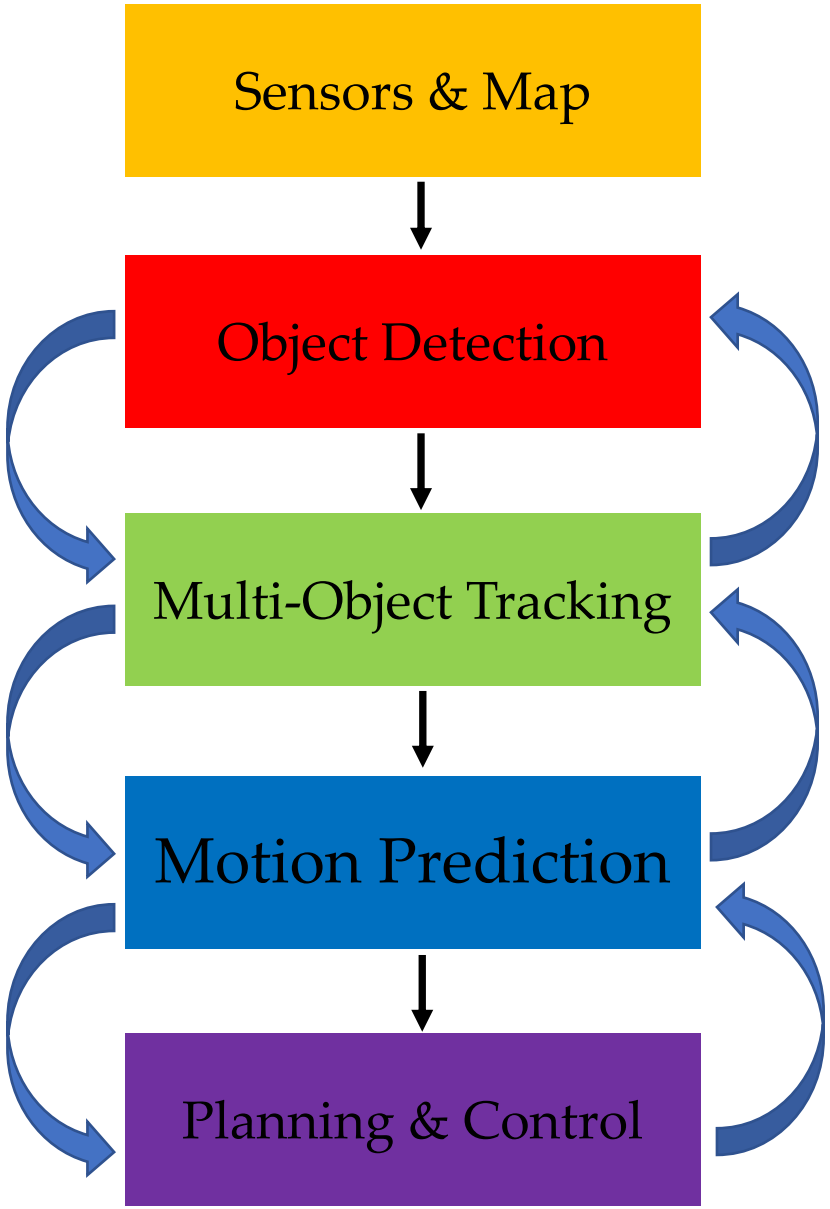
- End-to-end optimization
- Retain modular design for interpretability

Image cropped from Raquel Urtasun's talk at CVPR '21 WAD

Other alternatives or directions for autonomy stack integration?

Uncertainty Propagation through the Stack

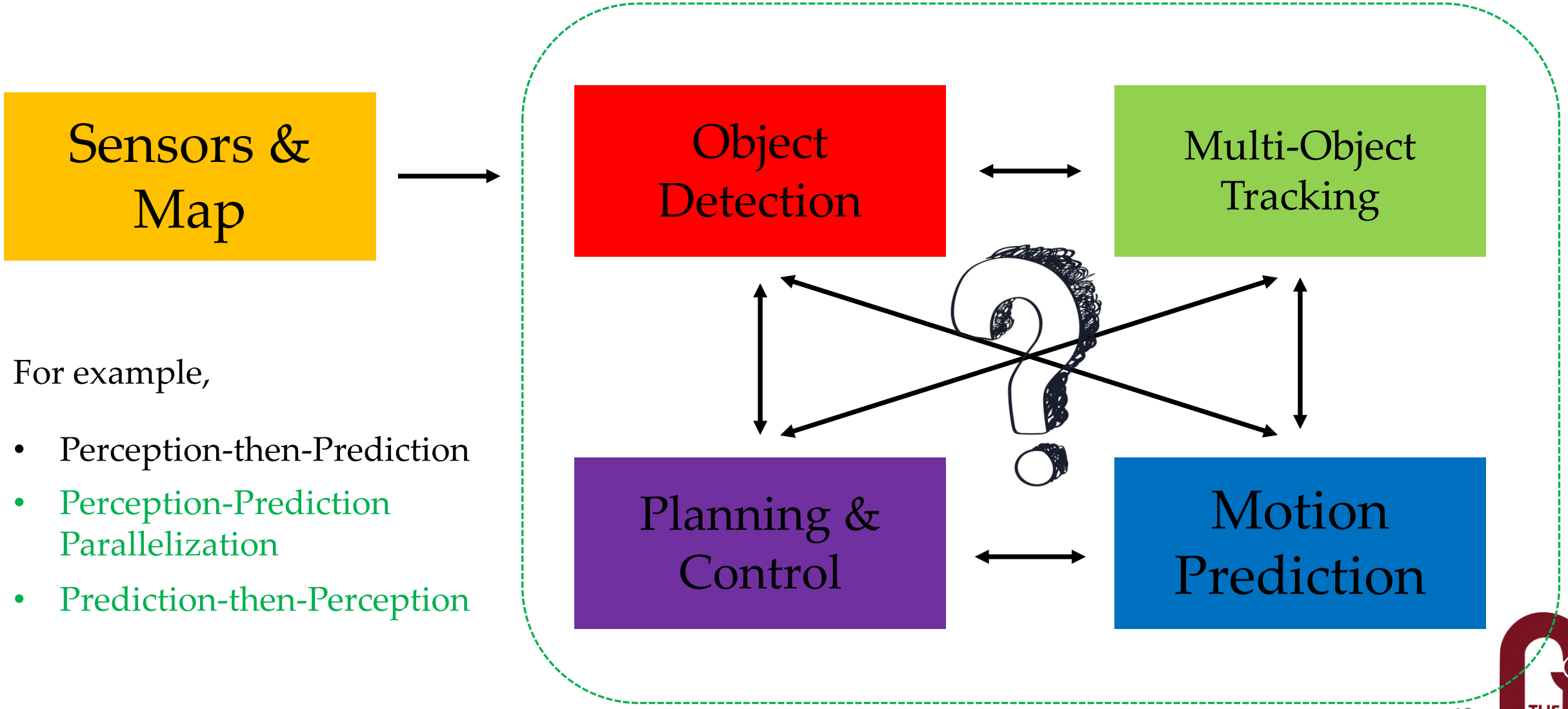
Forward propagation with *uncertainties*



Back propagation

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

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

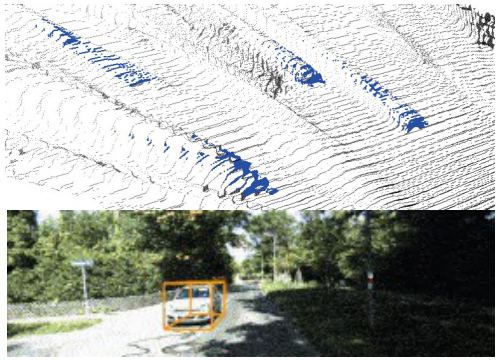
- Can propagate different kinds of useful uncertainties
- Has been optimized with the best ordering/structure

More robust and more effective

THE ROBOTICS INSTITUTE

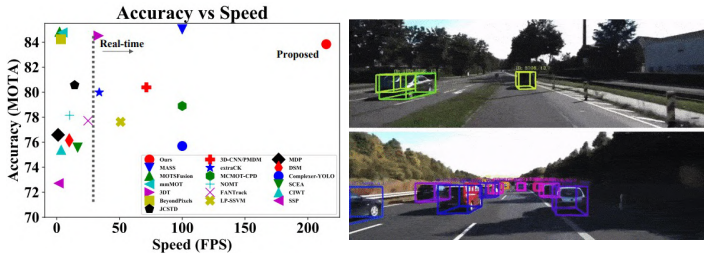
High performance of individual Modules is also important!

Modular Improvements for Perception and Prediction

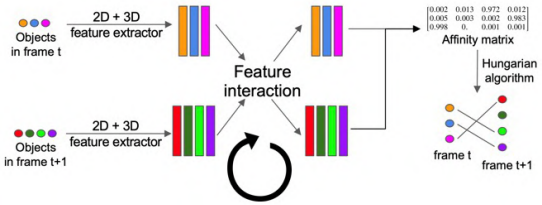


Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW '19

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

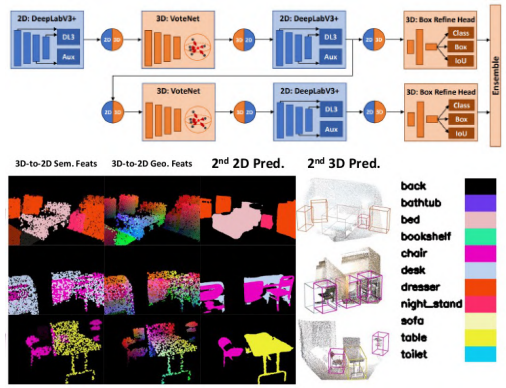


3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS '20

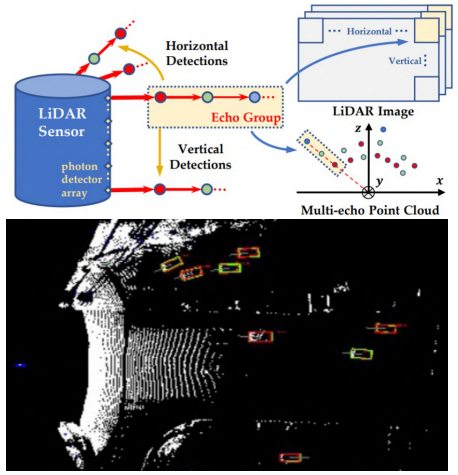


GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning. CVPR '20

Re-ID and Tracking



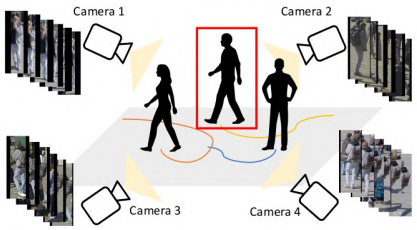
Multi-Modality Task Cascade for 3D Object Detection. BMVC '21



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

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

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



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

Detection

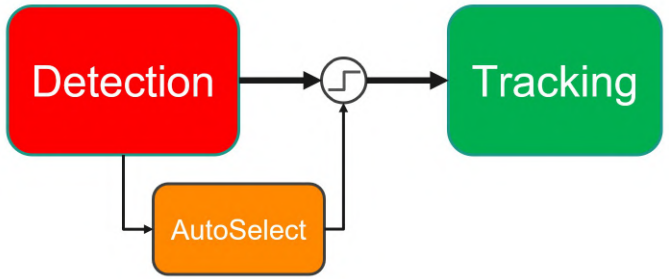
Prediction

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
...

THE ROBOTICS INSTITUTE

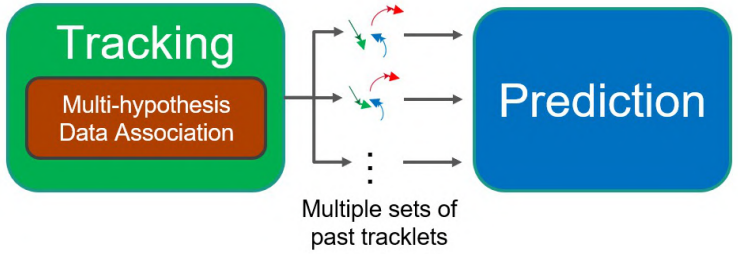
Towards modular and differentiable autonomous driving

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



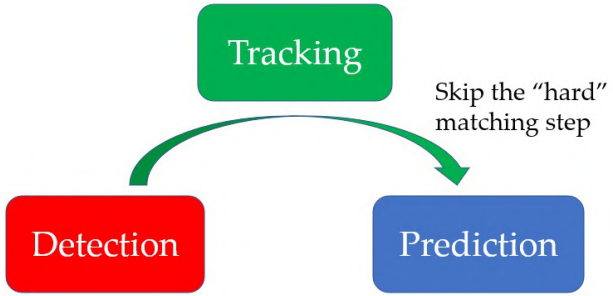
Automatic & Dynamic Detection Selection for Tracking (AutoSelect)

Weng, Kitani. arXiv '21



Multi-hypothesis Tracking and Prediction (MTP)

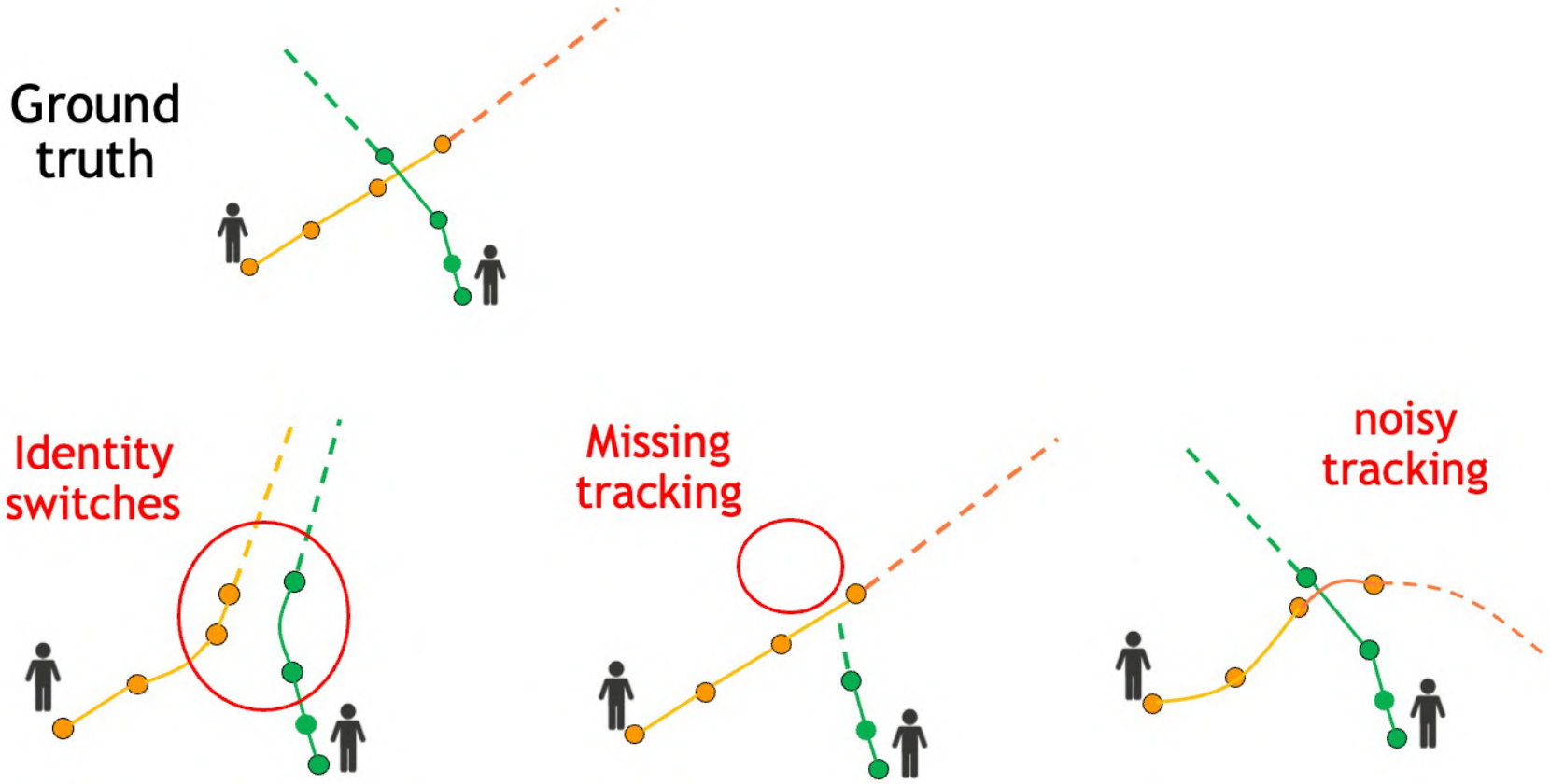
Weng, Ivanovic, Pavone. IV '22



Affinity-based Prediction (Affinipred)

Weng, Ivanovic, Kitani, Pavone. CVPR '22

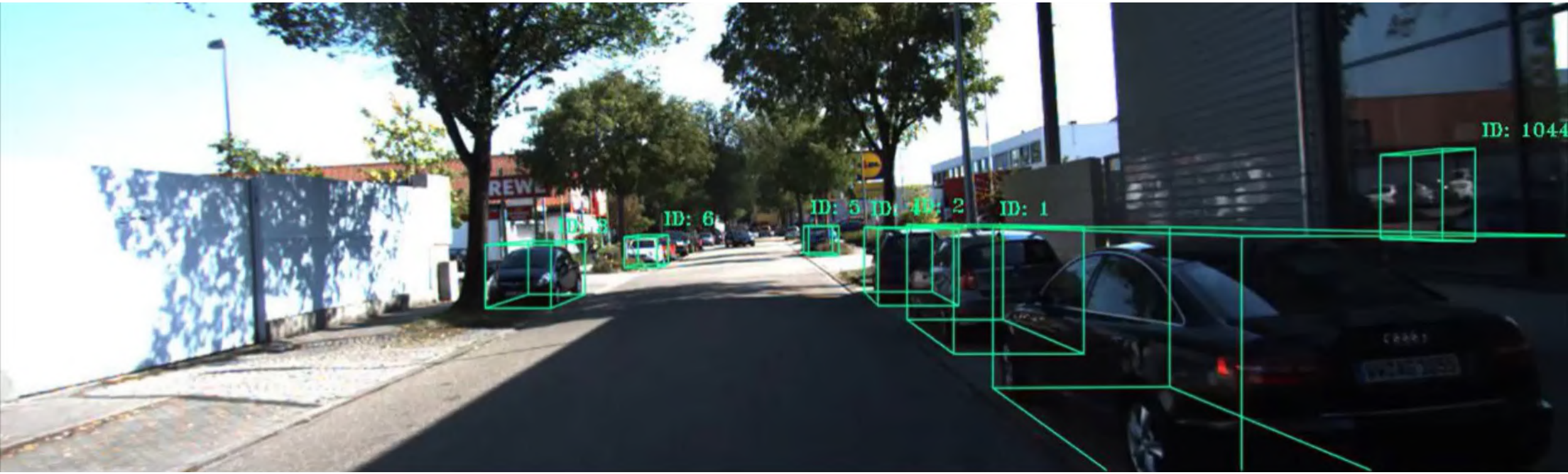
Trajectory Prediction is Challenging in the Wild



- 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

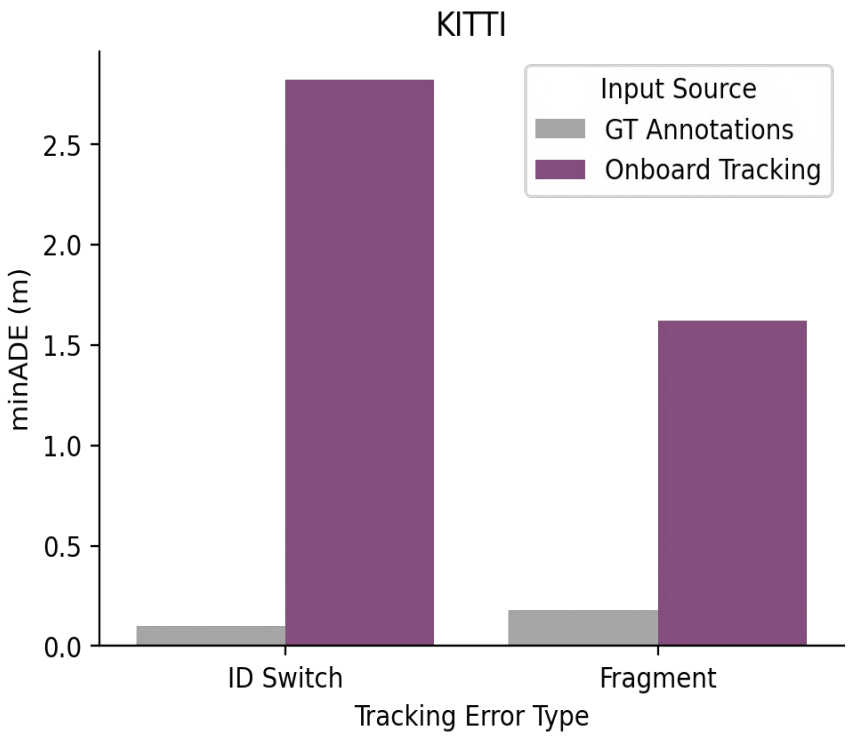
Object Detection and Tracking Is Not Perfect

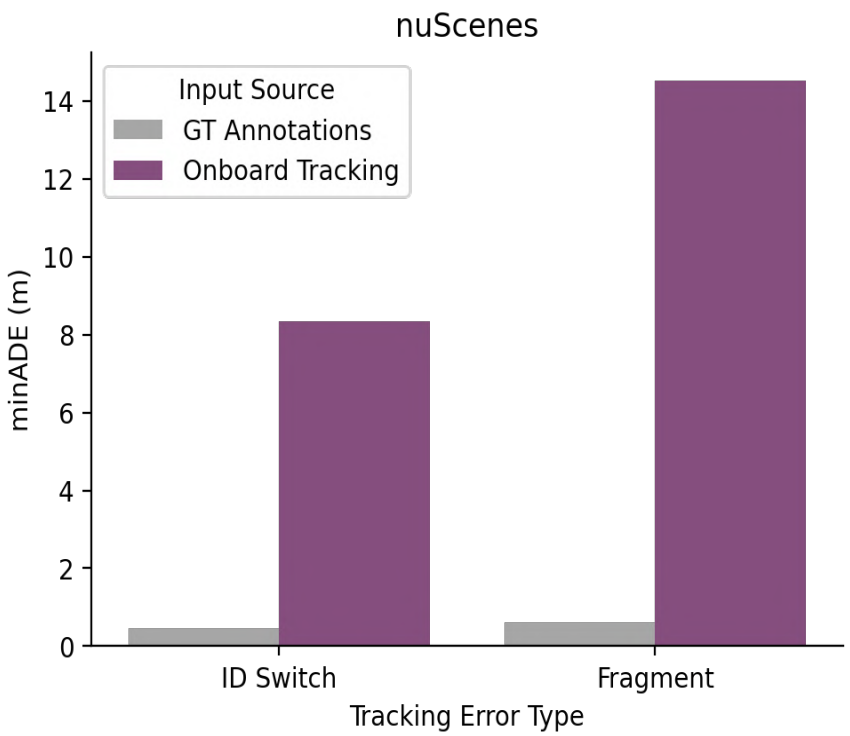


- Real-world results on KITTI (10 FPS)
- Can be even worse on nuScenes with a lower FPS

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

Tracking Errors Detrimentally Affect Prediction





Prediction errors increase by ***10-30x*** in the presence of tracking errors

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

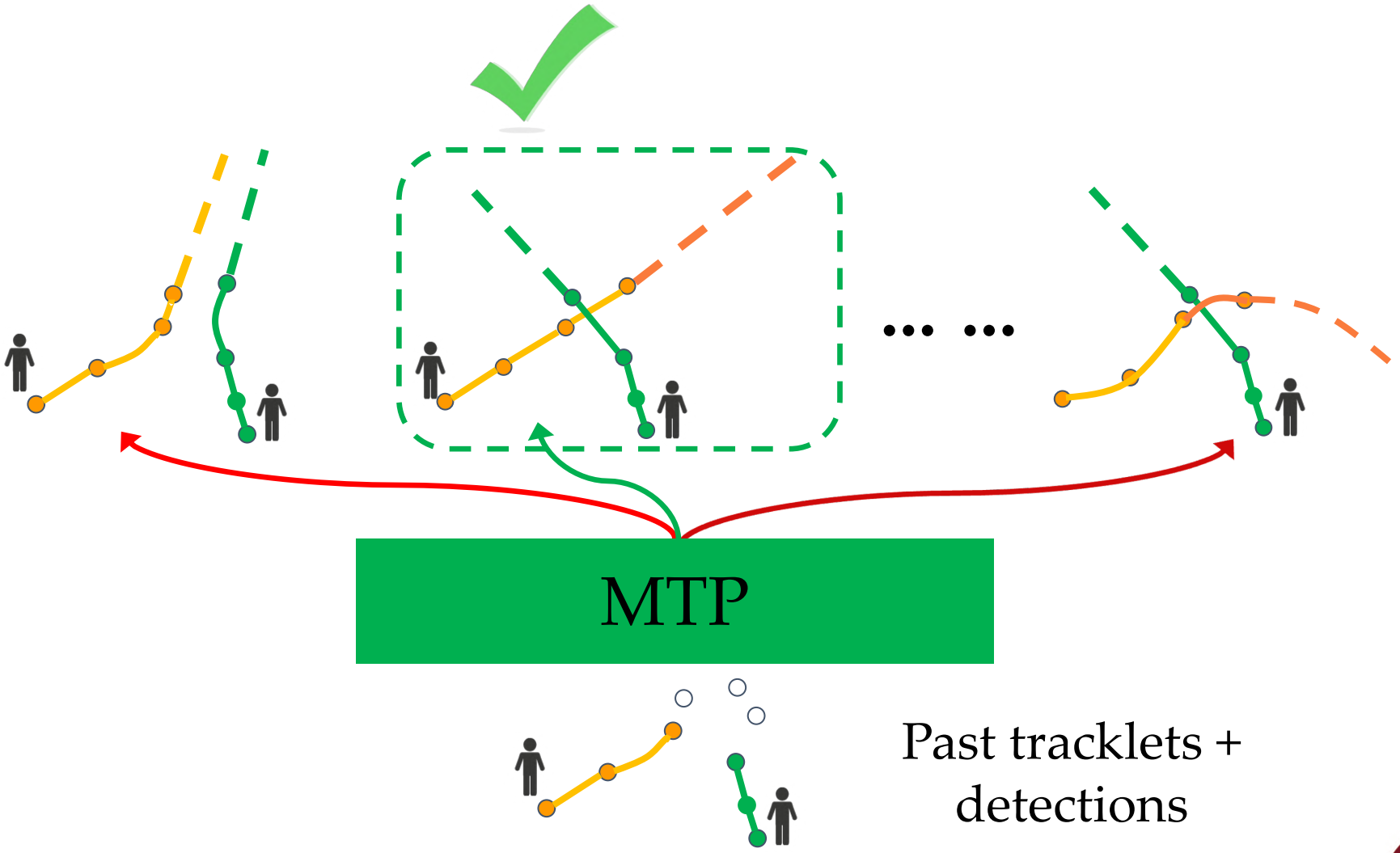
Multi-hypothesis Tracking and Prediction

Prediction for each hypothesis

Multiple matching hypotheses

↓

Explore different tracking results

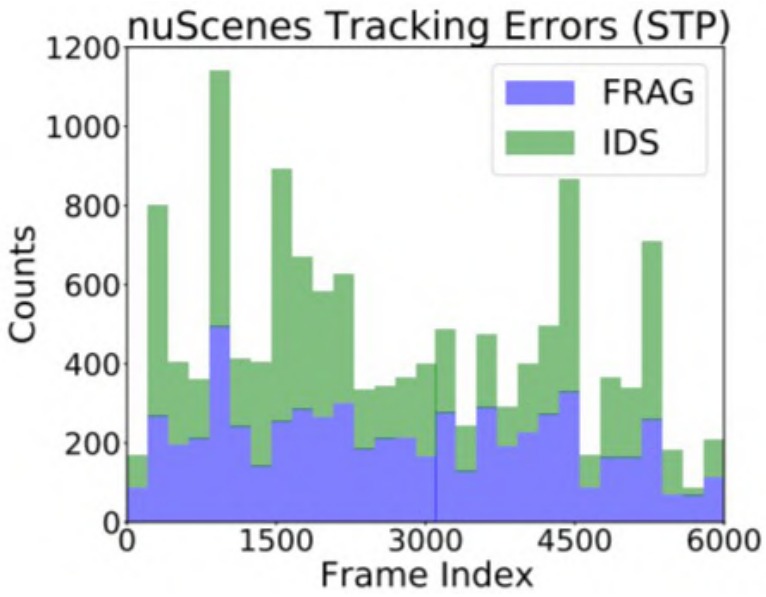


Increase the likelihood of including accurate tracking results for prediction

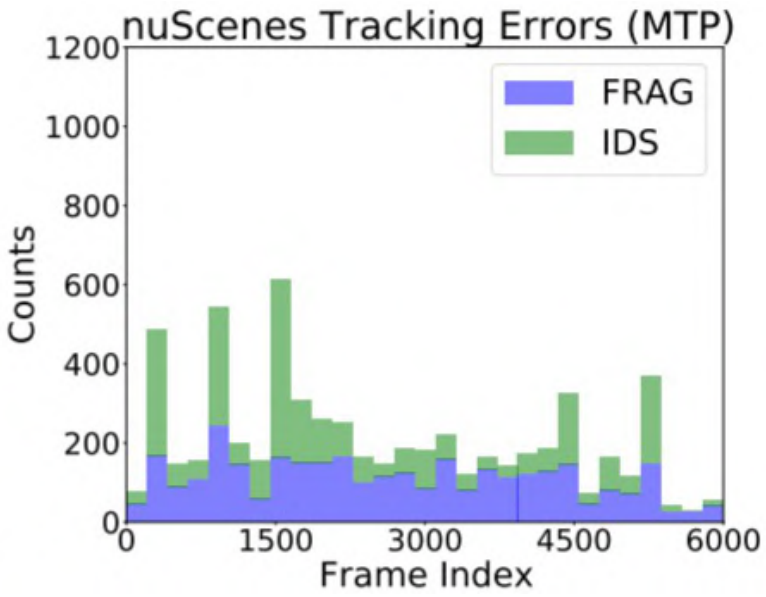
Xinshuo Weng, Boris Ivanovic, Marco Pavone. MTP: Multi-hypothesis Tracking and Prediction for Reduced Error Propagation. IV 2022

Are Different Hypotheses Complimentary?

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



K = 1



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

PREDICTION PERFORMANCE FOR OBJECTS WITH IDS/FRAG.

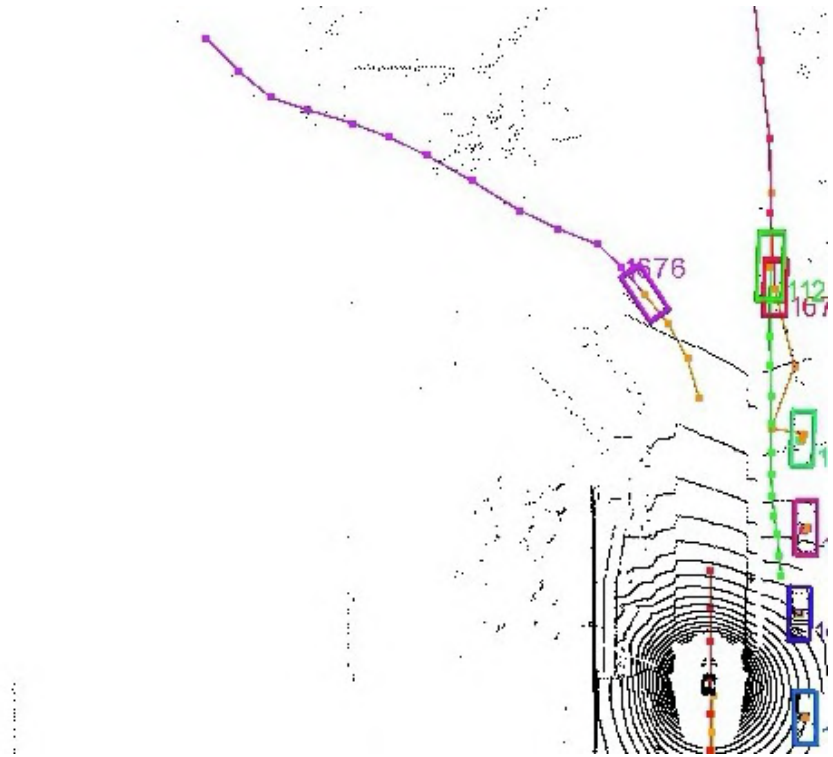
Datasets	Targets	Methods	minADE_k	minFDE_k
KITTI	IDS	STP, H=1, k=20	2.820	4.514
		MTP (Ours), H=5, k=100	1.099	1.768
		MTP (Ours), H=10, k=200	0.844	1.332
		MTP (Ours), H=20, k=400	0.707	1.093
		MTP (Ours), H=5, k=20, sampling	1.118	1.802
		MTP (Ours), H=10, k=20, sampling	0.876	1.390
		MTP (Ours), H=20, k=20, sampling	0.747	1.173
KITTI	FRAG	STP, H=1, k=20	1.621	2.155
		MTP (Ours), H=5, k=100	1.436	1.862
		MTP (Ours), H=10, k=200	1.385	1.765
		MTP (Ours), H=20, k=400	1.305	1.627
		MTP (Ours), H=5, k=20, sampling	1.448	1.888
		MTP (Ours), H=10, k=20, sampling	1.404	1.801
		MTP (Ours), H=20, k=20, sampling	1.335	1.688

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

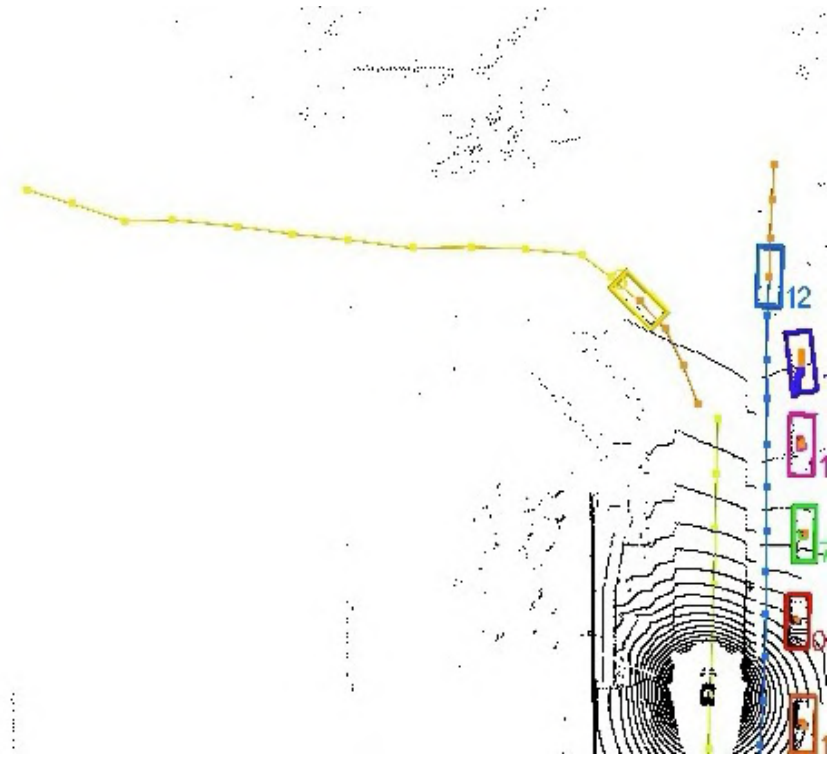
THE ROBOTICS INSTITUTE

MTP Increases Prediction Robustness



Standard perception-prediction
Single hypothesis

Unstable



MTP (Ours)
Multiple hypotheses

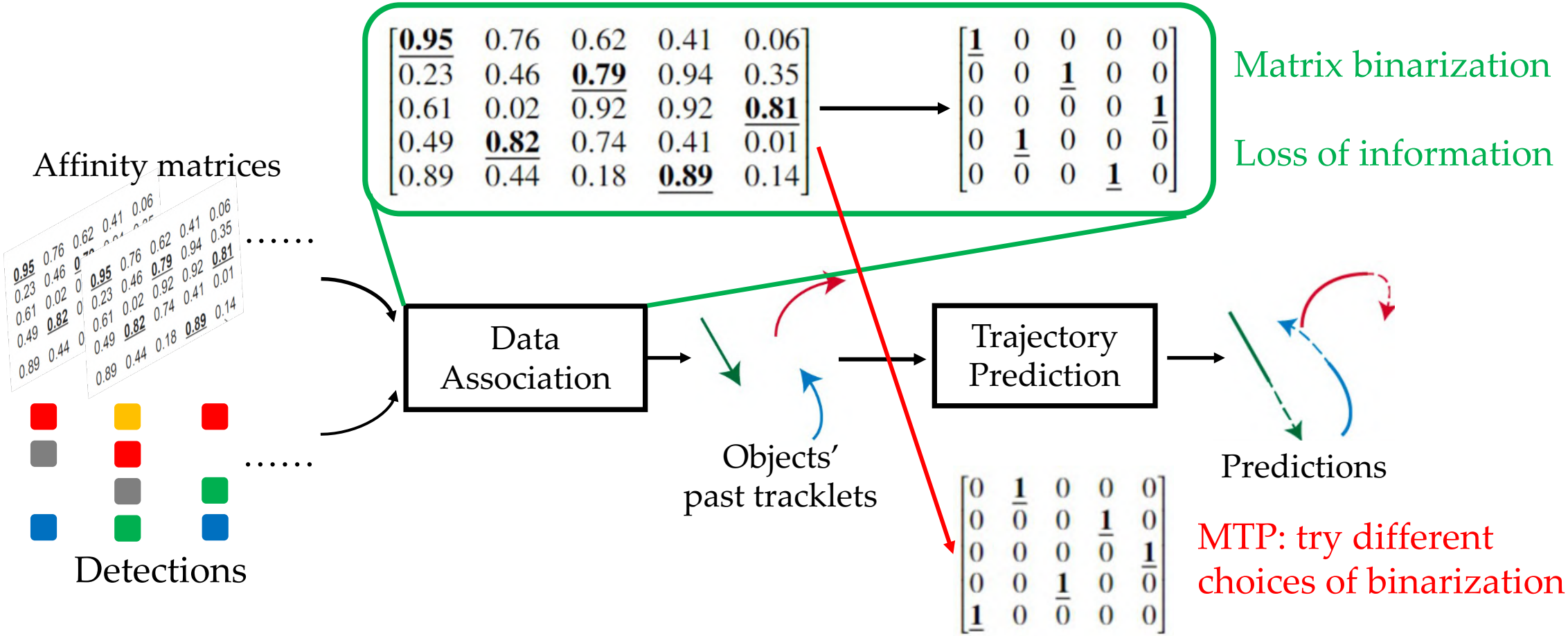
More stable and robust

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?

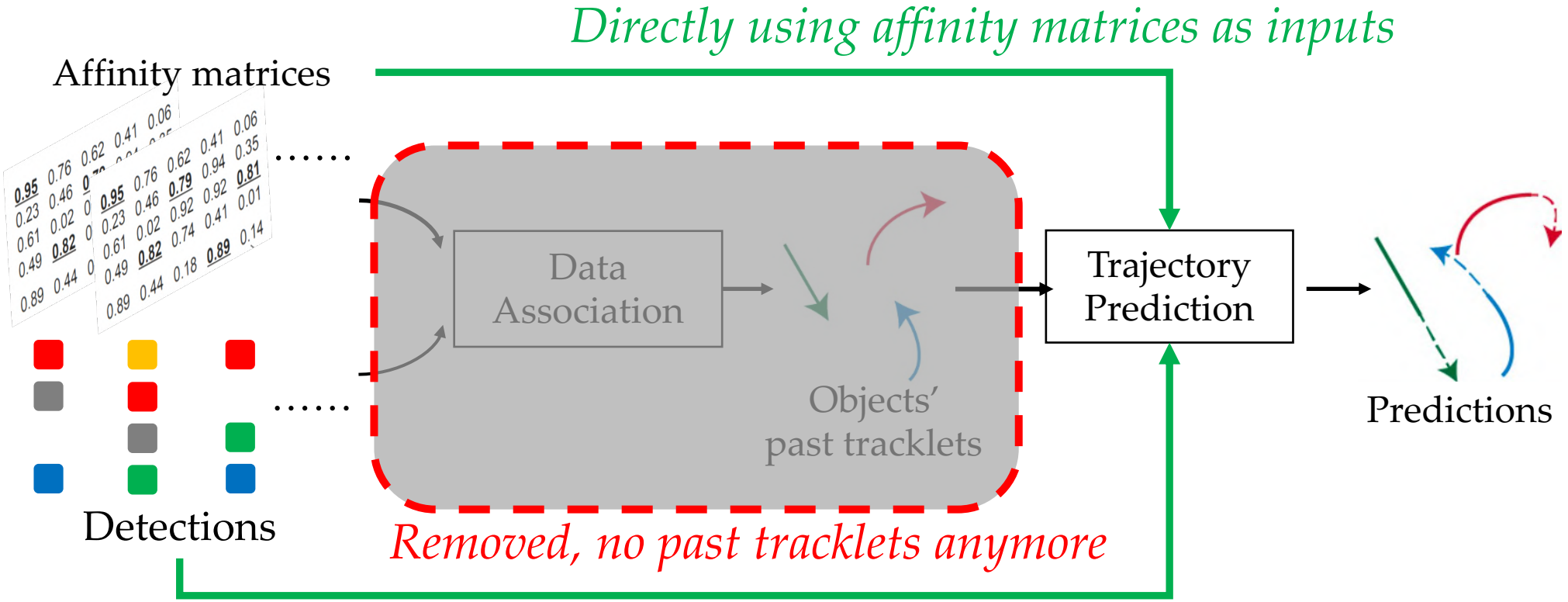
Rethinking the Perception-then-Prediction Pipeline



- 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

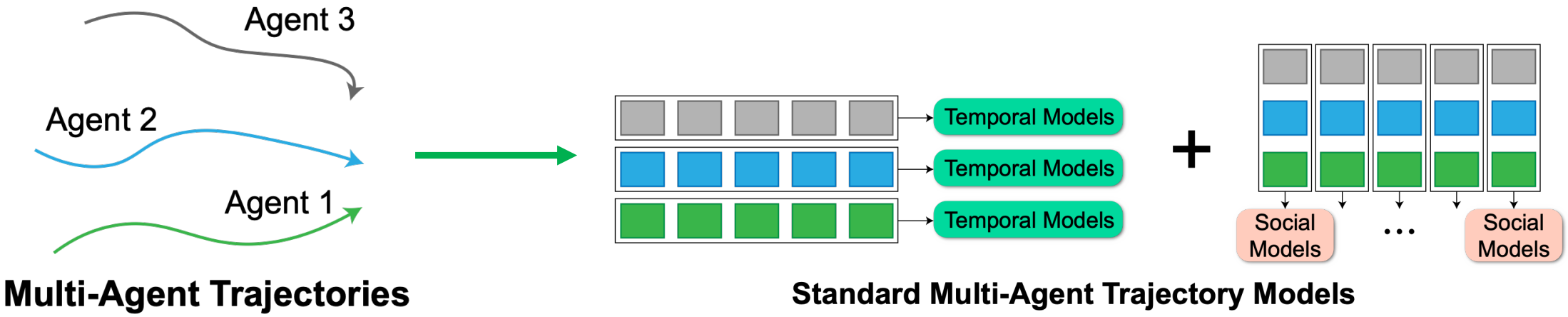
Affinity-based Prediction (Affinipred)

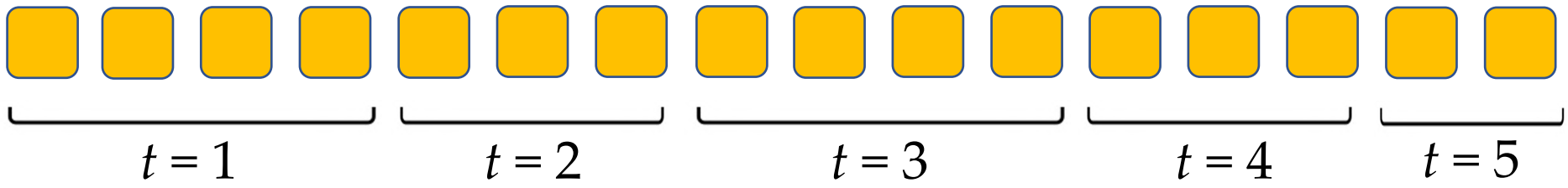


Propagate matching uncertainty to trajectory prediction

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



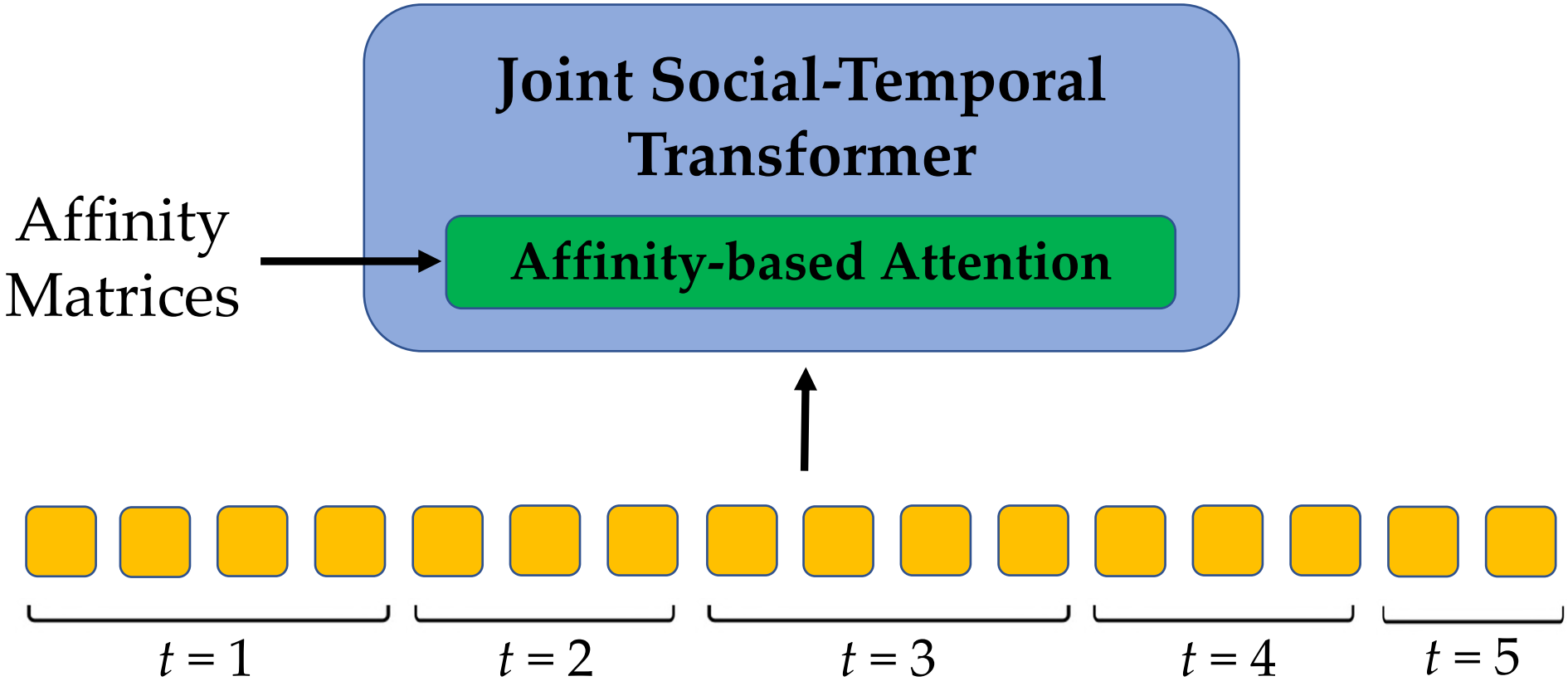


Our inputs: raw detections, no identity & trajectory

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

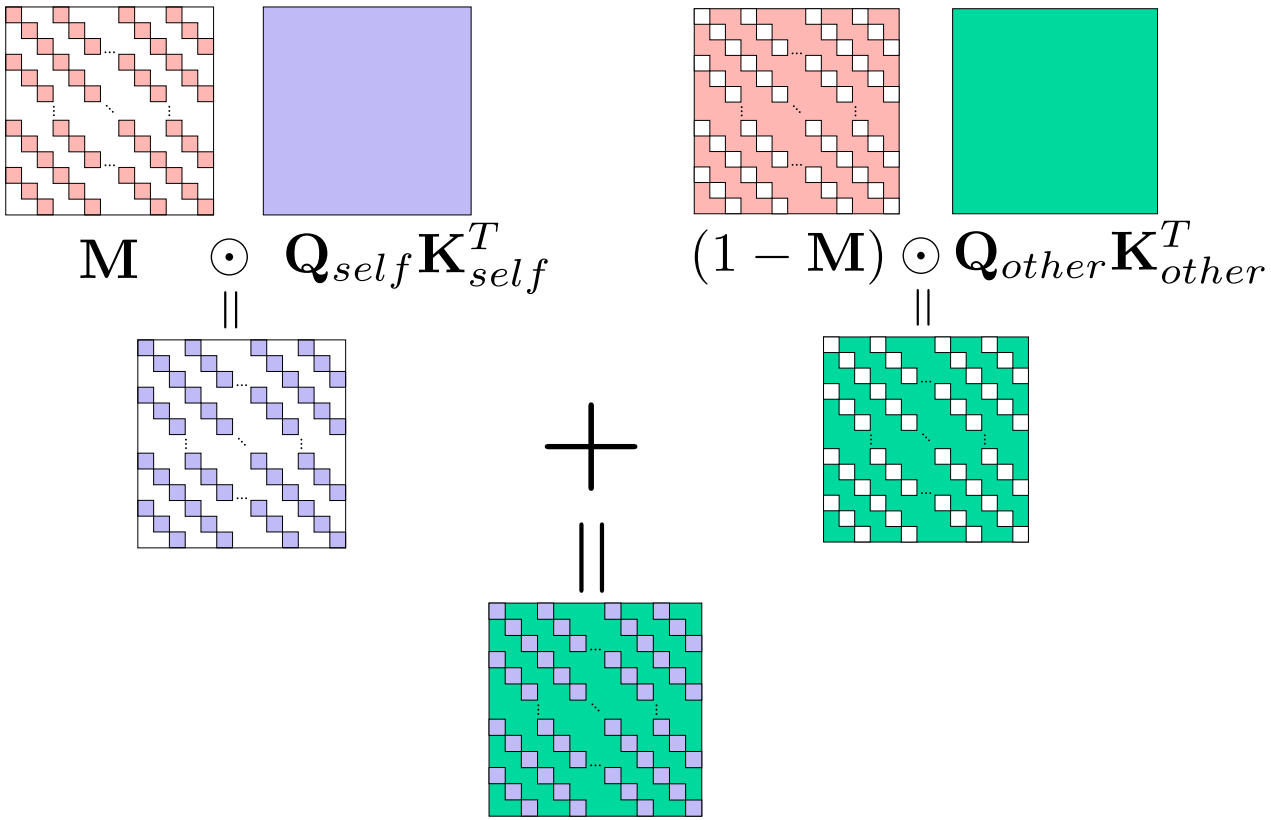
THE ROBOTICS INSTITUTE

Affinipred with Joint Social-Temporal Transformer



Tokens are extracted from detections, no trajectory anymore

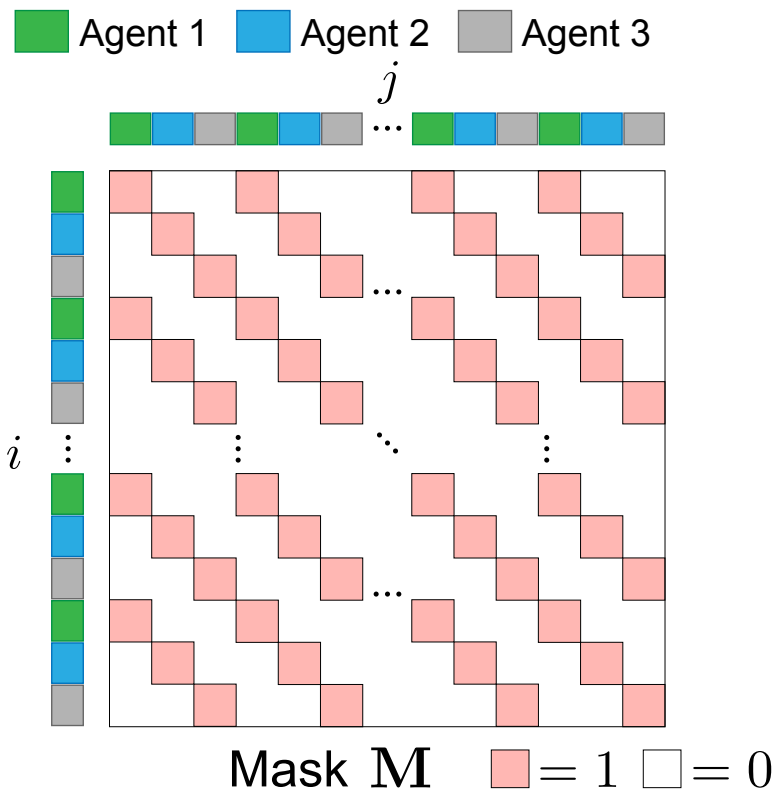
Inject Matching Uncertainty: Affinity-based Attention



Learn two sets of keys and queries: self-attention and cross-attention

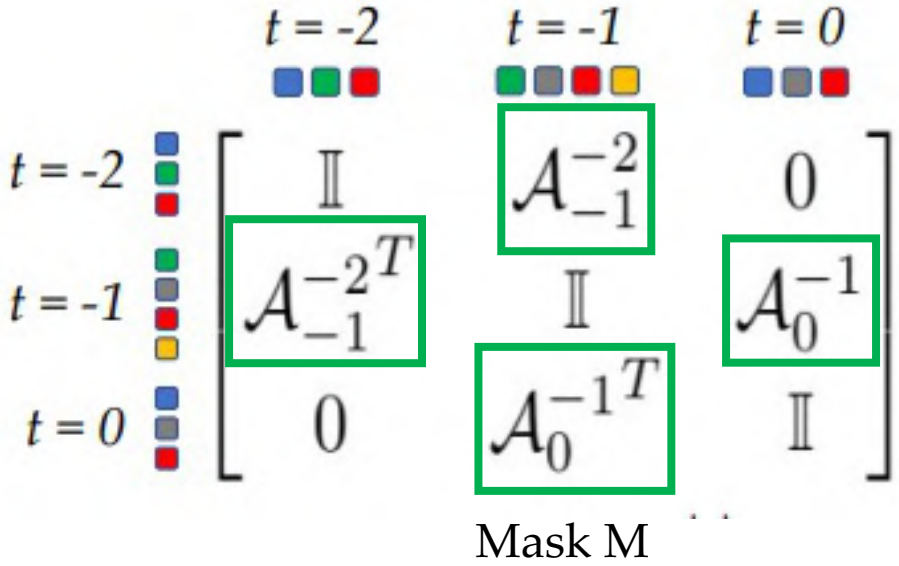
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone.
Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022

Inject Matching Uncertainty: Affinity-based Attention



Idealized setting with object identities available

Affinity matrices between consecutive frames



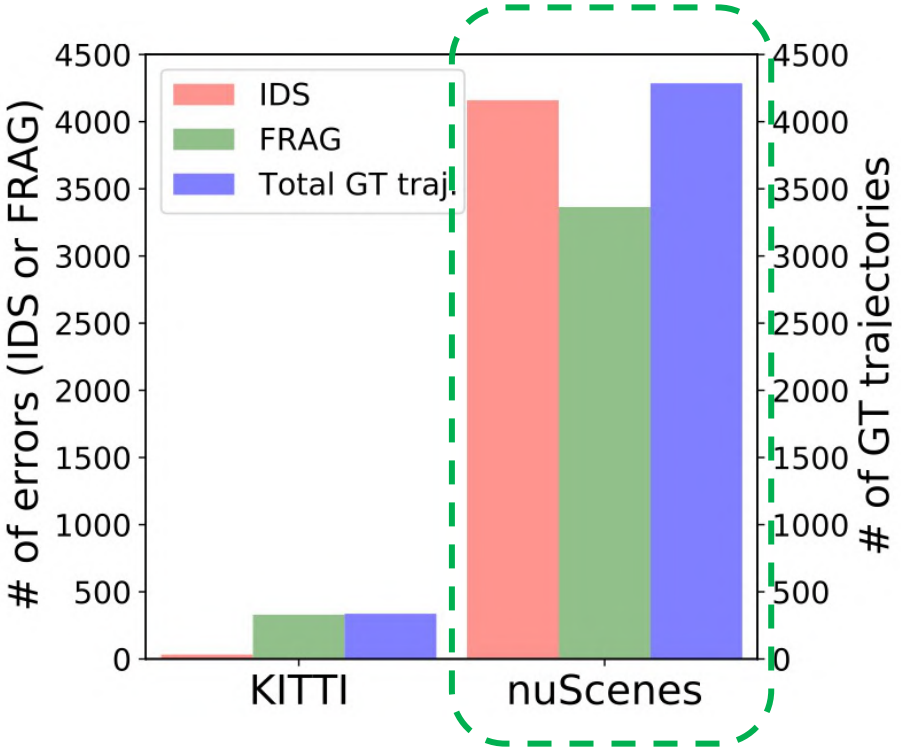
Realistic setting using affinity representing the probability of having similar identities

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

THE ROBOTICS INSTITUTE

Less Cascading Errors

Between Tracking and Prediction



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

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

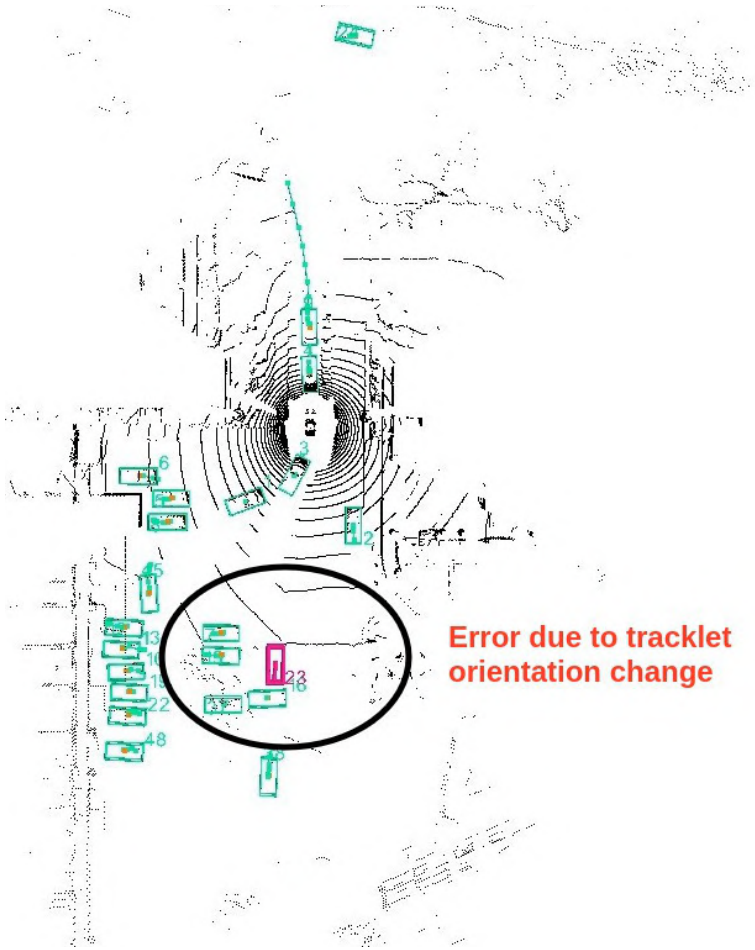
MTP extracts "more" information from affinity matrices

Affinipred "fully" exploits the uncertainties in affinity matrices

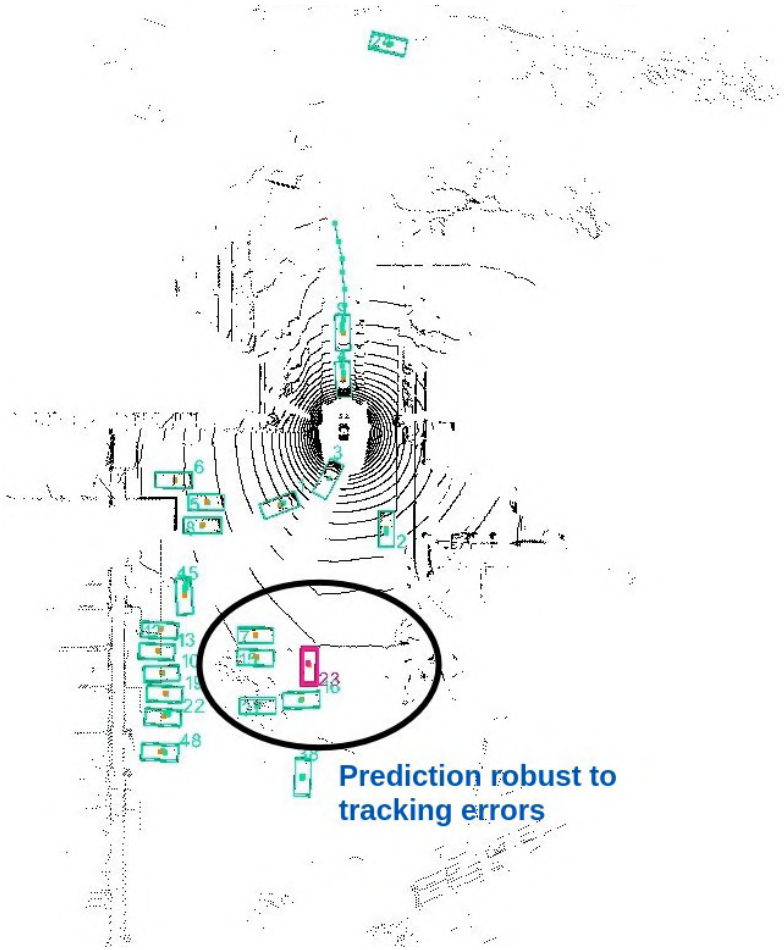
Xinshuo Weng, Boris Ivanovic, Kris Kitani, Marco Pavone.
Whose Track Is It Anyway? Improving Robustness to Tracking Errors with Affinity-based Trajectory Prediction. CVPR 2022

THE ROBOTICS INSTITUTE

Affinipred Is More Robust to Perception Noises



Standard tracking-prediction

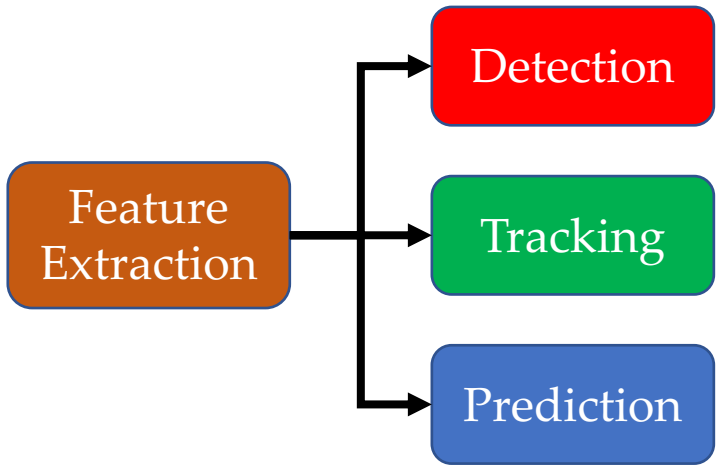


Affinipred

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

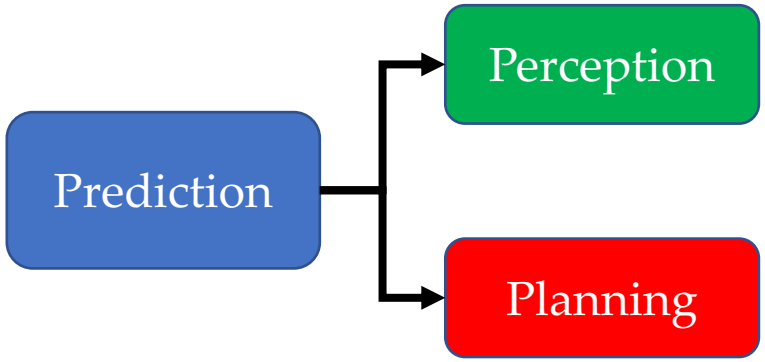
Towards modular and differentiable autonomous driving

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



Perception-Prediction Parallelization

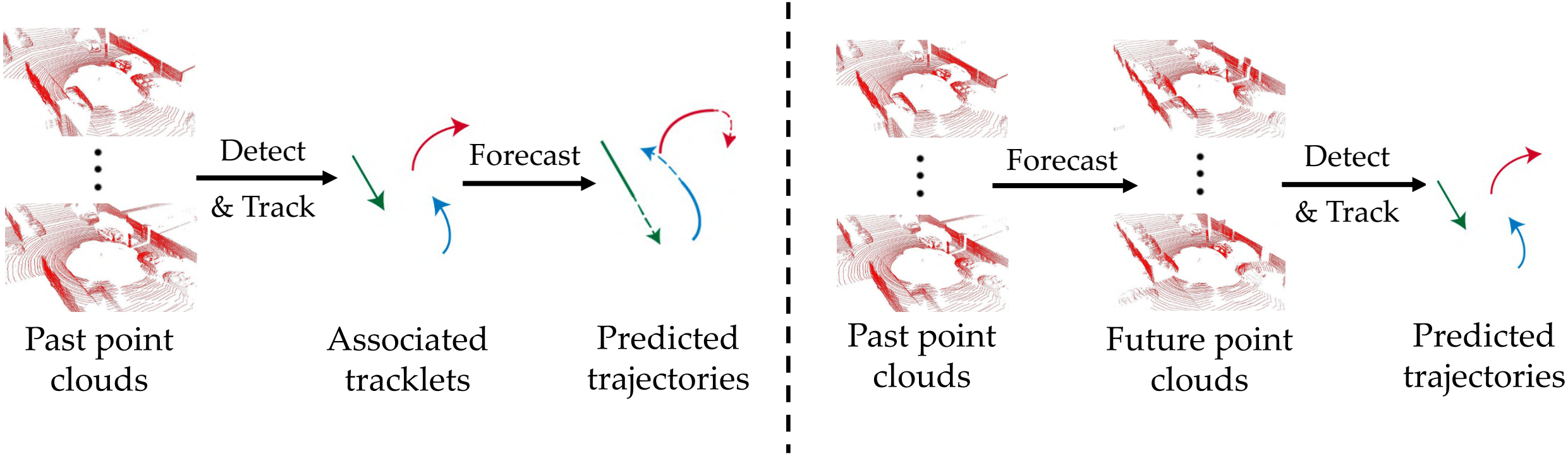
Wang, Kitani, **Weng**. ICRA '21
Weng*, Yuan*, Kitani. ICRA '21



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

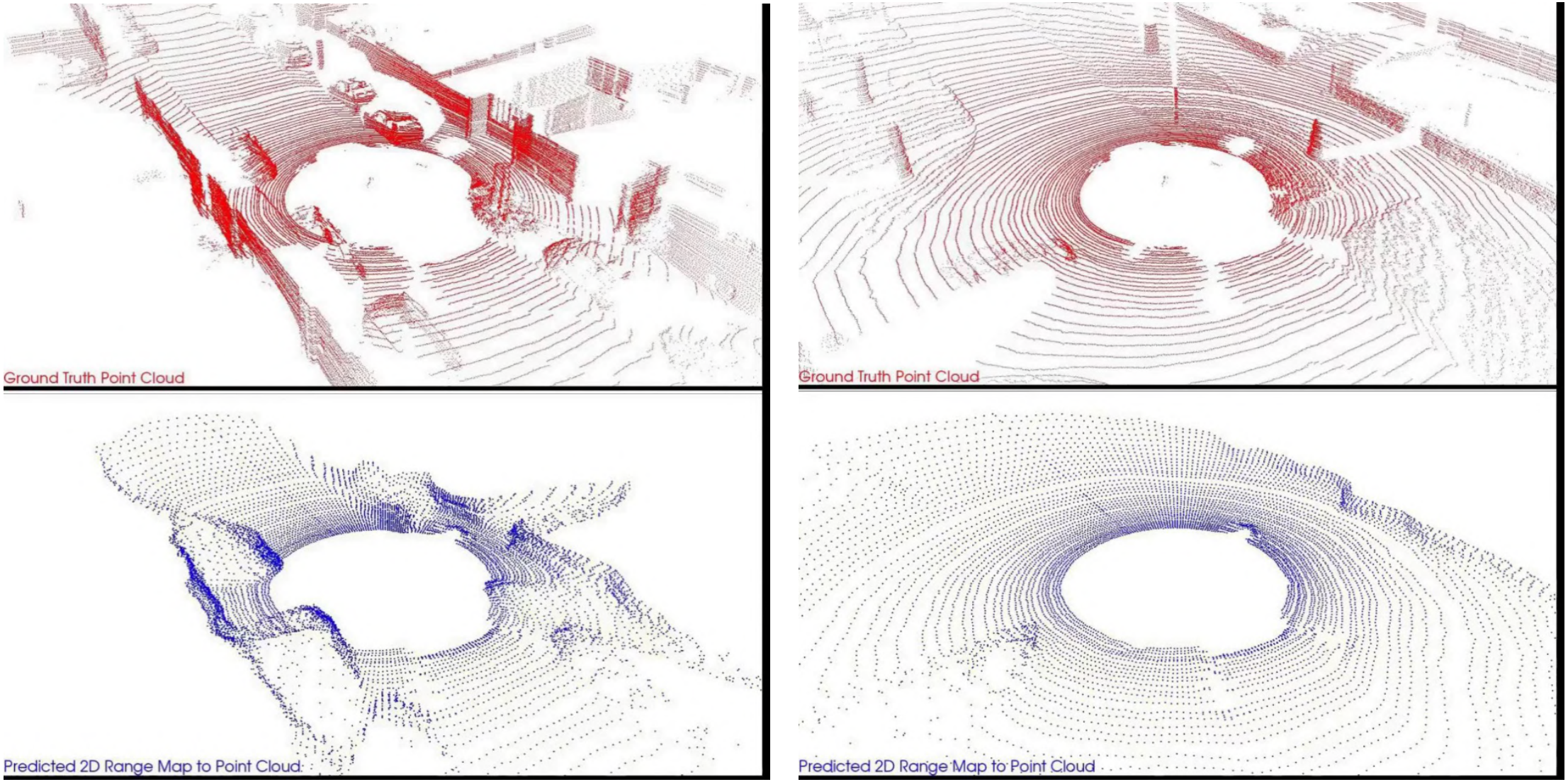
Prediction-then-Perception



- 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

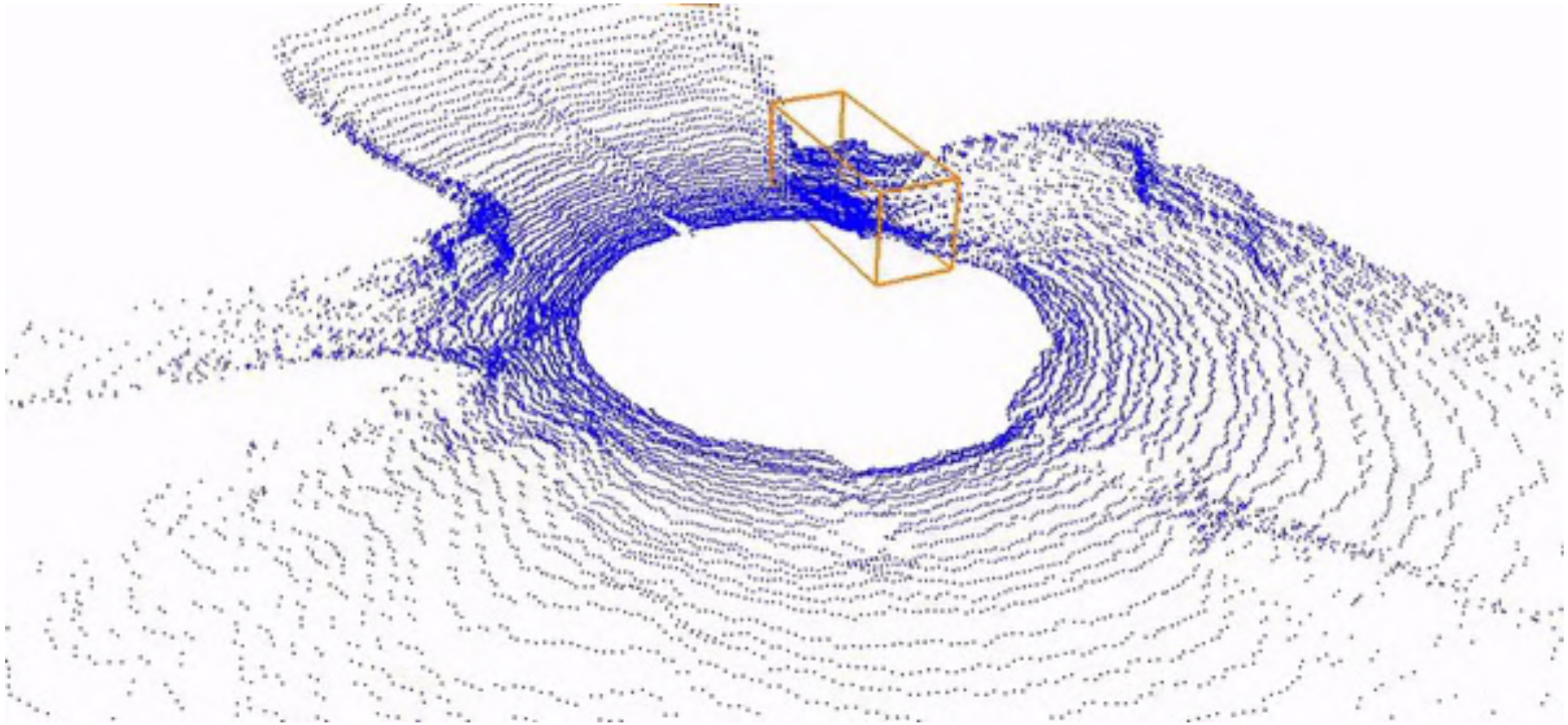
Predicted Point Clouds Preserves Objects



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

THE ROBOTICS INSTITUTE

Detection on Predicted Point Clouds



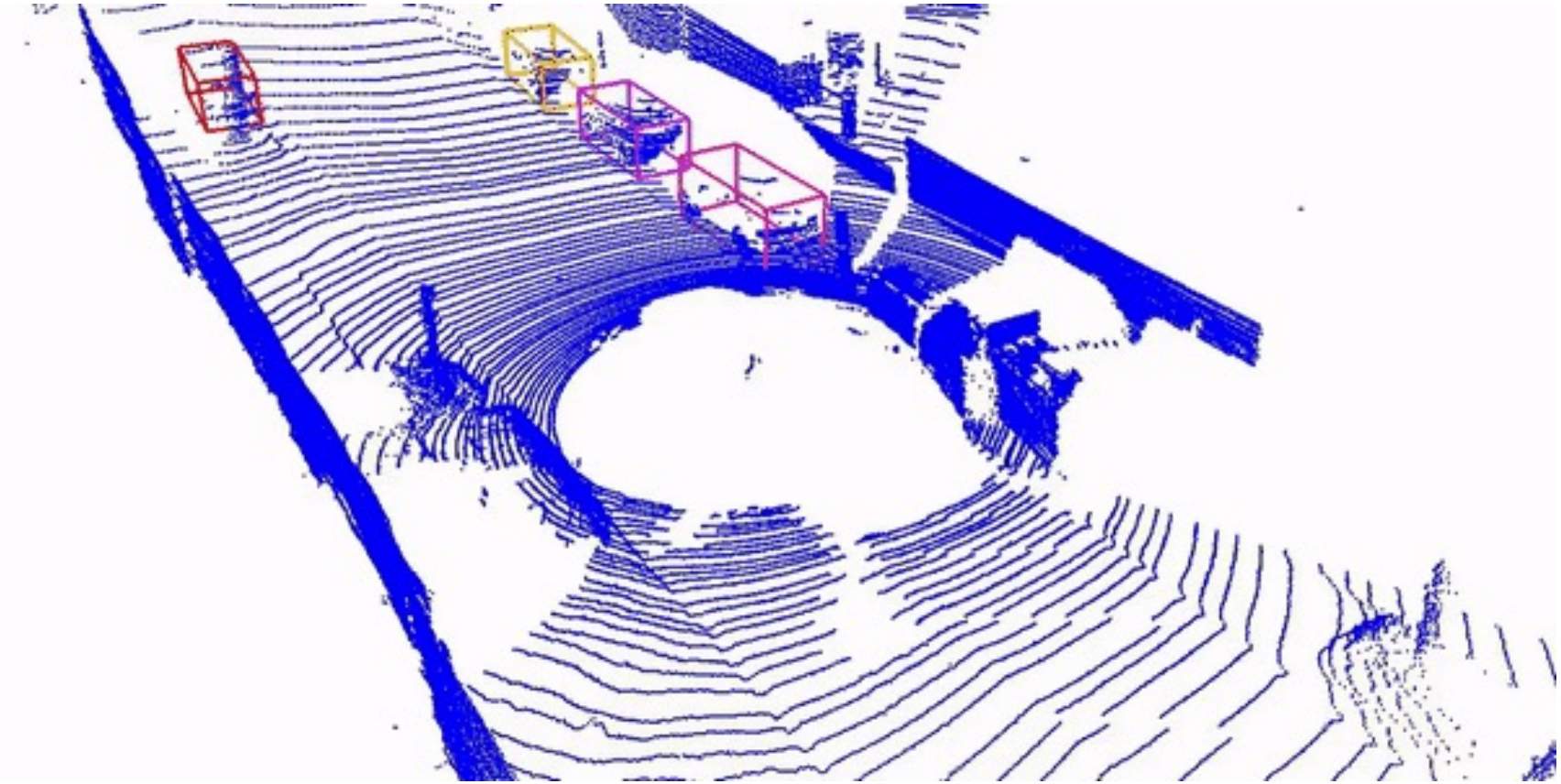
Green: detections

Yellow: GT

Detections on predicted point clouds match with GT at a close 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

Tracking on Predicted Point Clouds



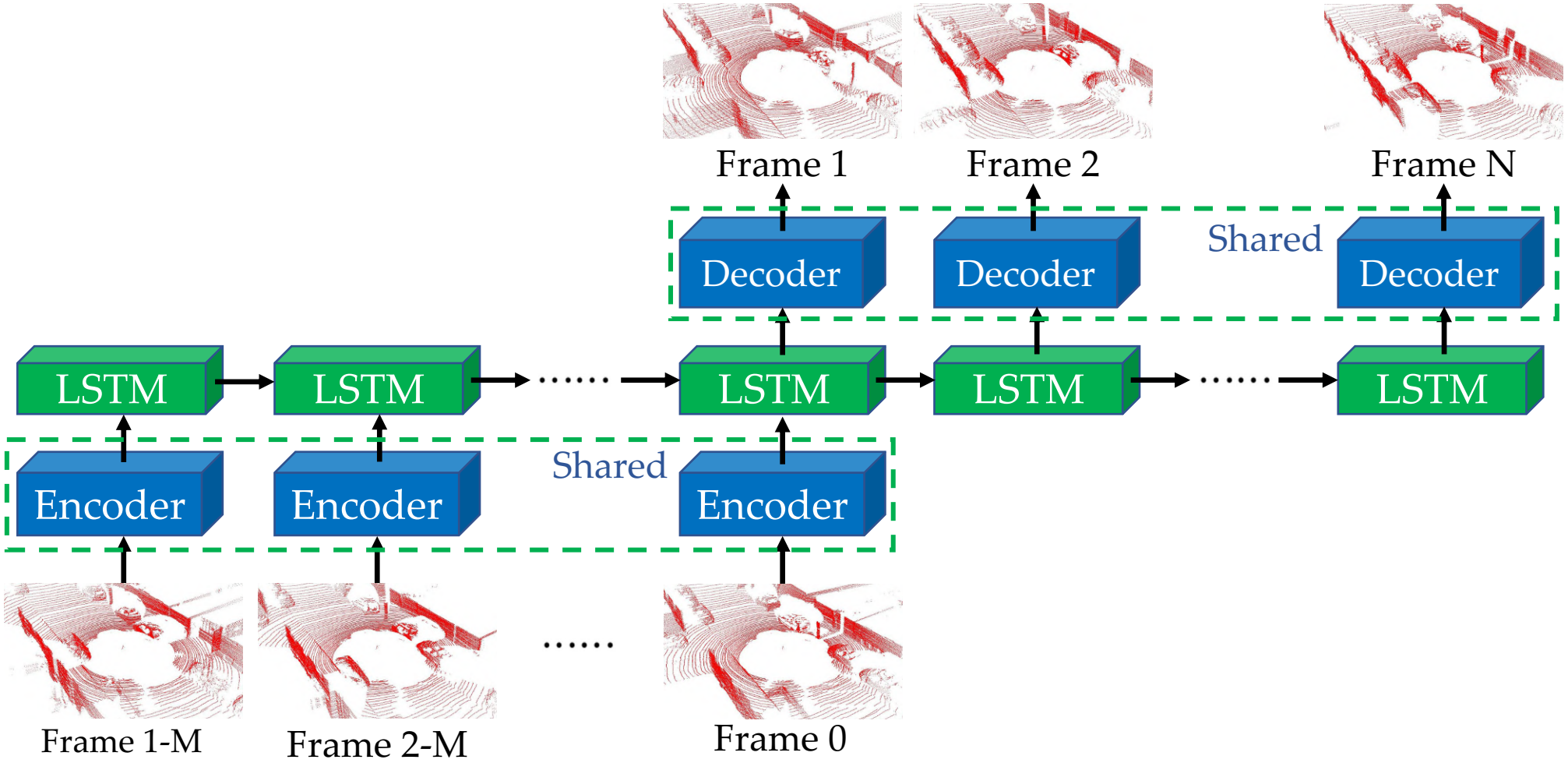
Stable tracking of detected objects on predicted point clouds

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

THE ROBOTICS INSTITUTE

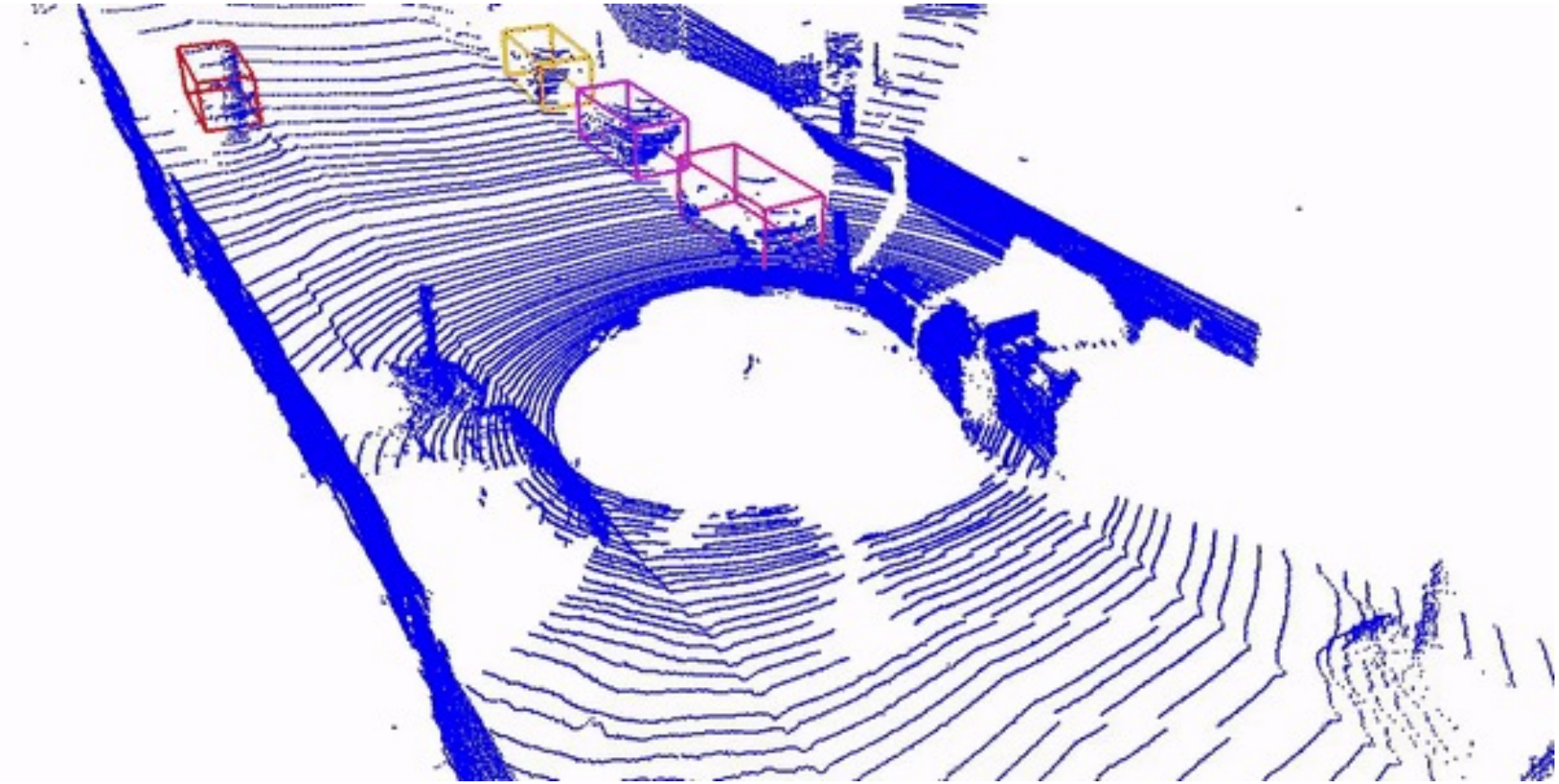
Simple Baseline: SPFNet

LSTM Auto-encoder



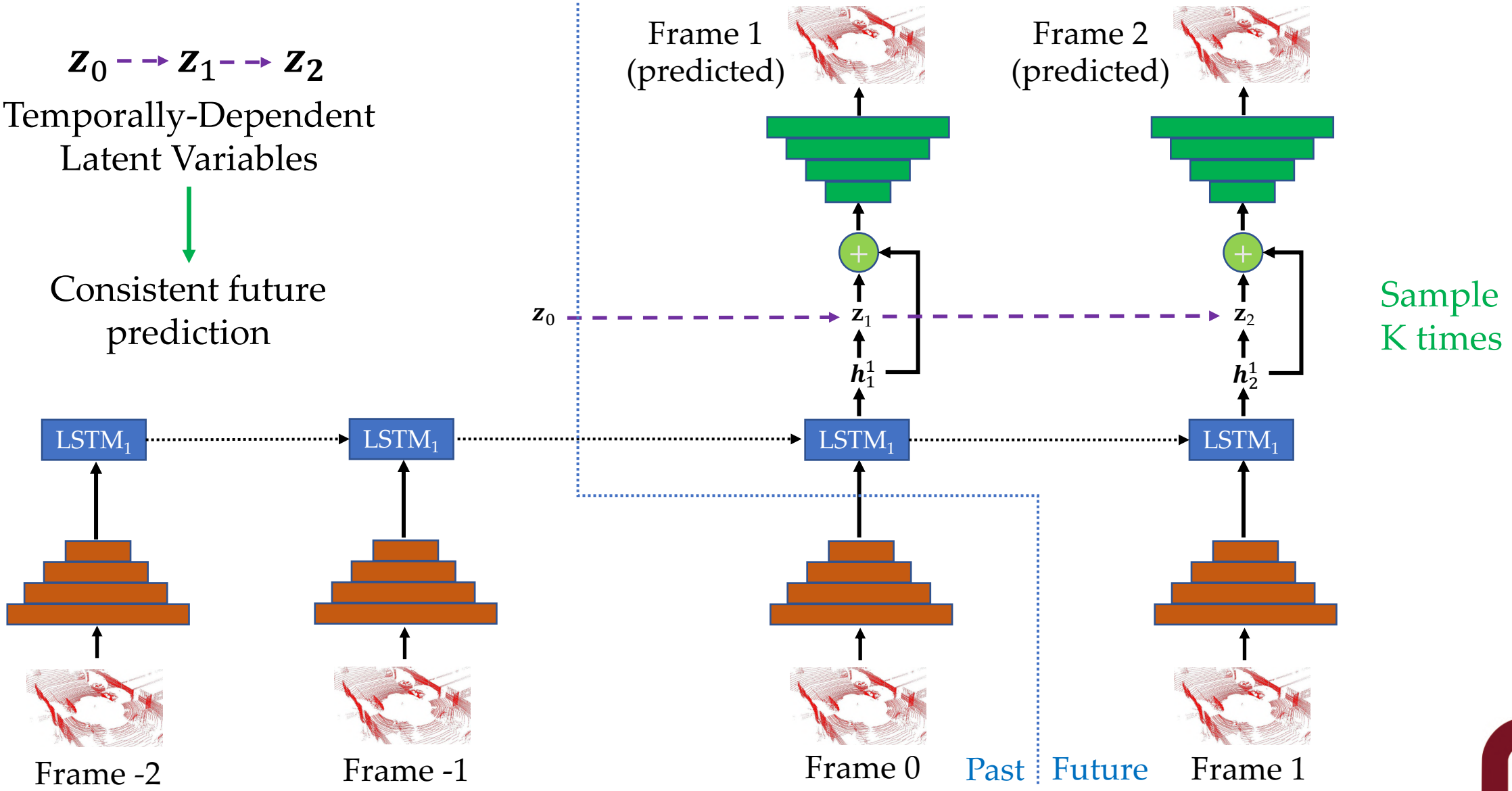
Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, Nick Rhinehart.
Inverting the Pose Forecasting Pipeline with SPF2: Sequential Pointcloud Forecasting for Sequential Pose Forecasting. CoRL 2020

Limitations: Deterministic & Blurry Predictions



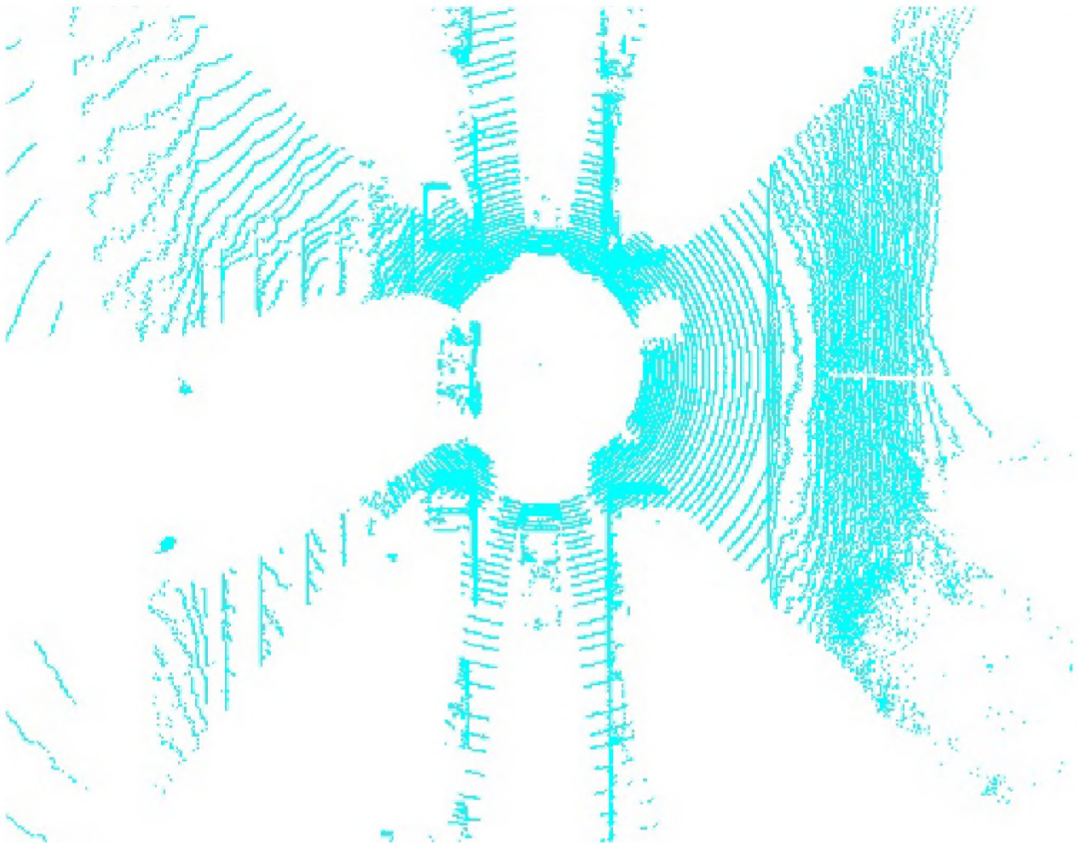
- The use of a deterministic model (LSTM autoencoder)
- High compression of information (feature vectorization)

S2Net: Stochastic Sequential Pointcloud Forecasting

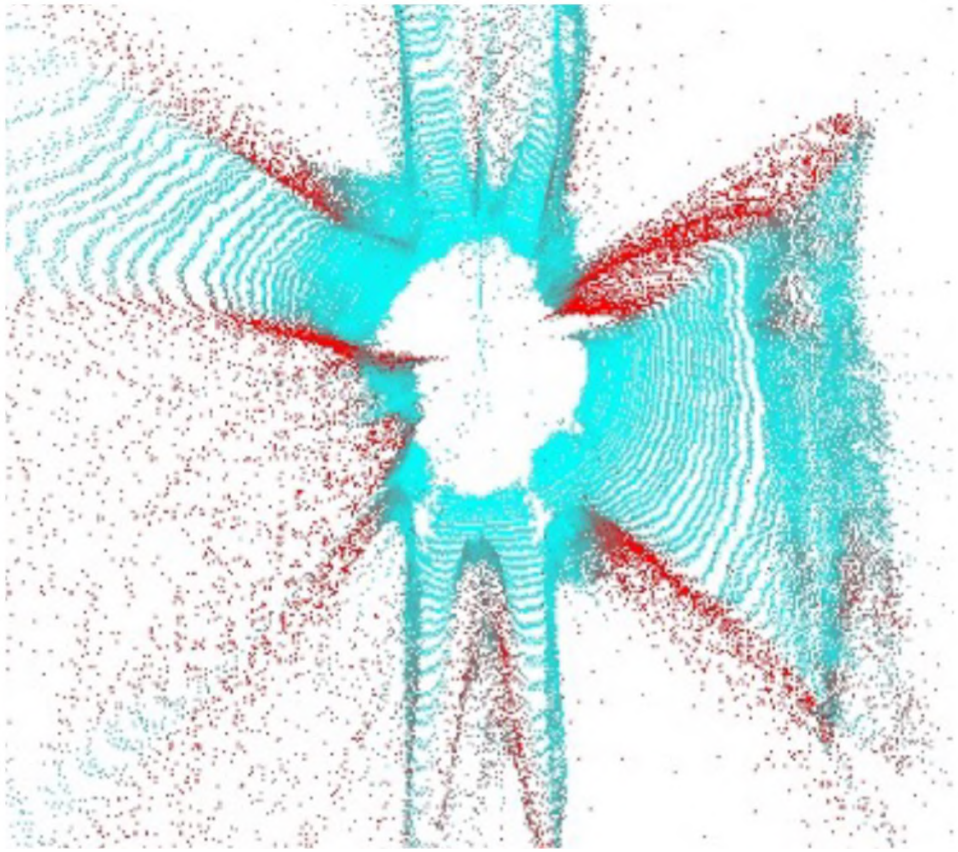


Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022

Uncertainty Is Mostly Around the Boundary



GT Point Cloud

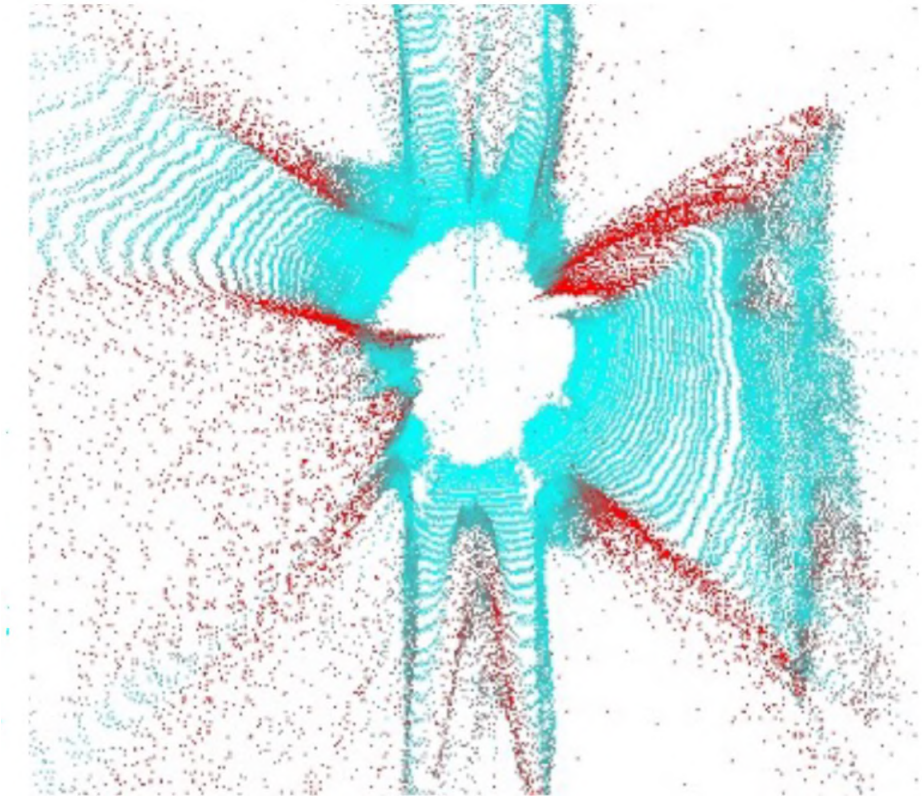


Prediction with Uncertainty (K=5)

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

THE ROBOTICS INSTITUTE

Uncertainty Can Potentially Help Downstream



Prediction with Uncertainty (K=5)

3D detection on predicted point clouds

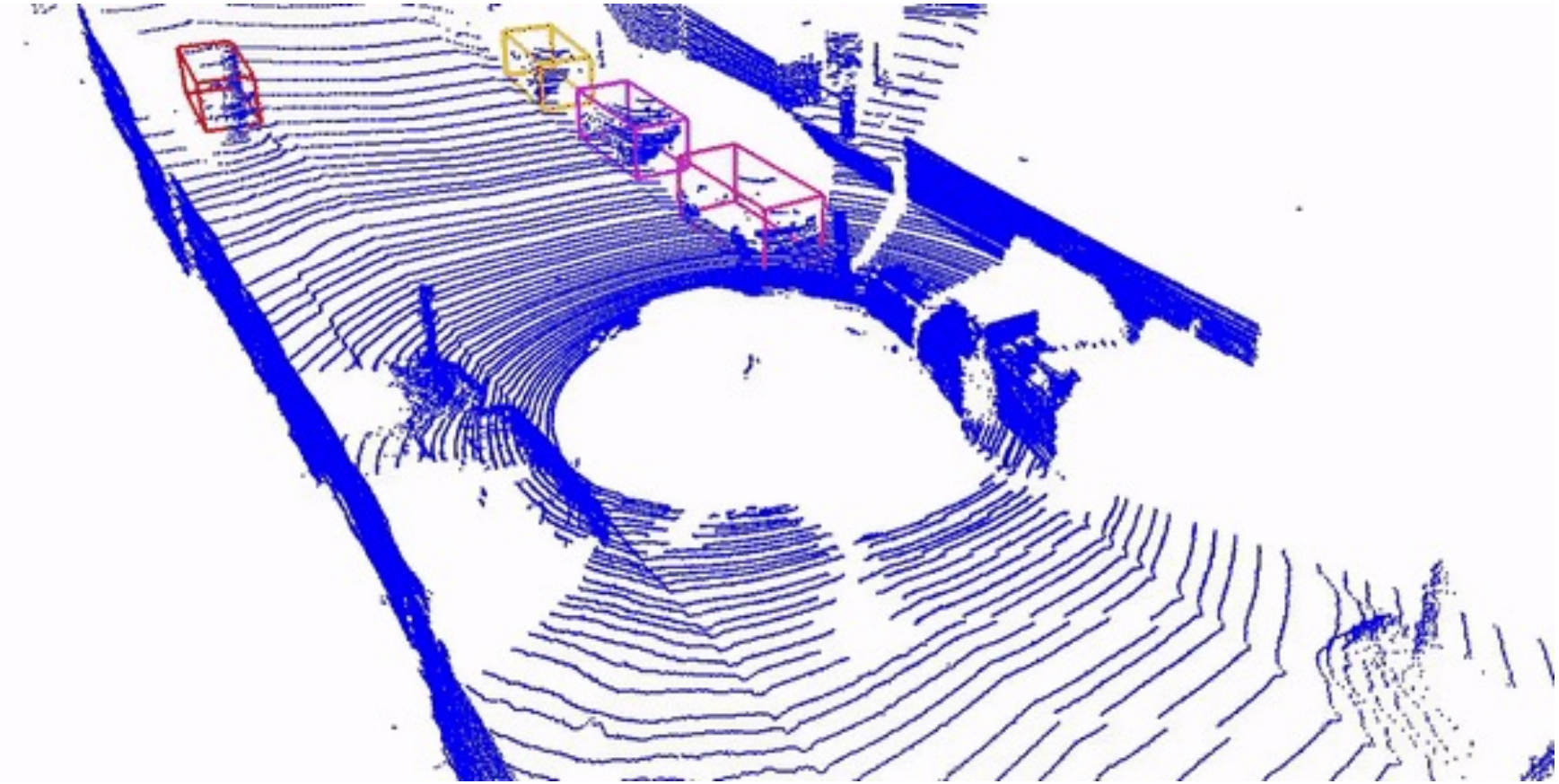
Inputs	mAP	Precision	Recall
(x, y, z)	69.96	76.44	72.81
(x, y, z) + uncertainty	70.41	78.48	72.76

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

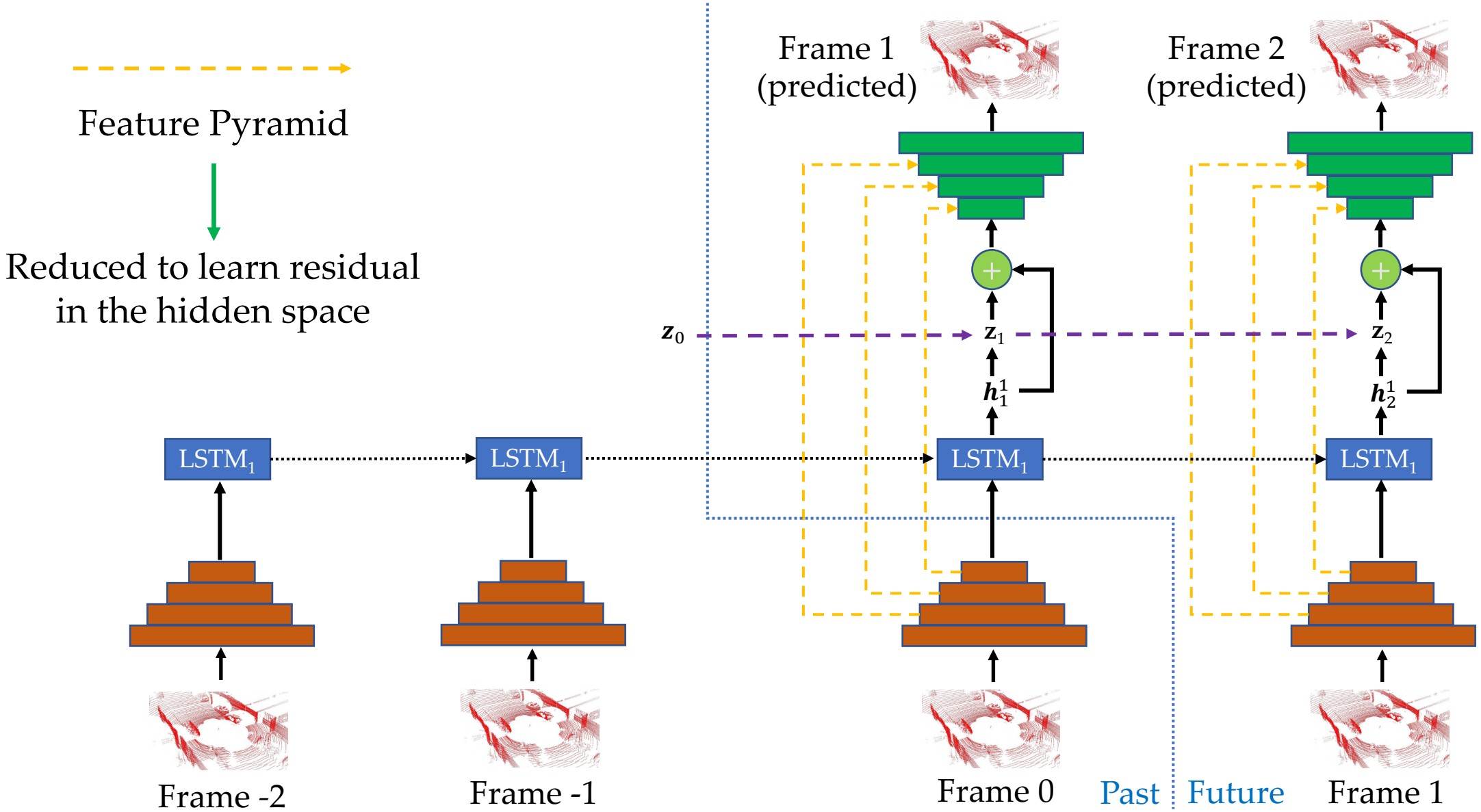
THE ROBOTICS INSTITUTE

Limitations: Deterministic & Blurry Predictions



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

Simple Feature Pyramid Fails to Converge

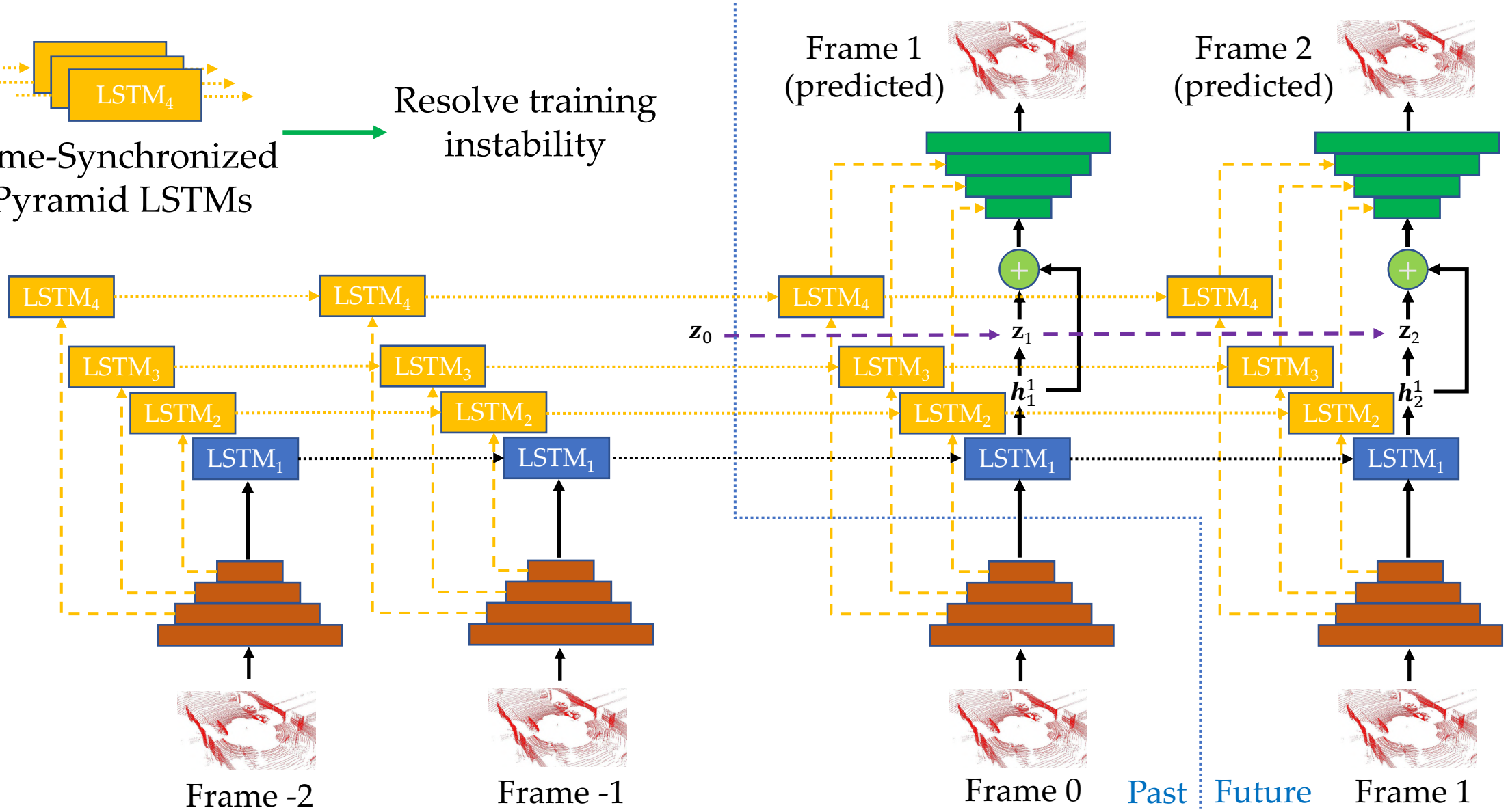


Xinshuo Weng, Junyu Nan, Kuan-Hui Lee, Rowan McAllister, Adrien Gaidon, Nicholas Rhinehart, Kris Kitani. S2Net: Stochastic Sequential Pointcloud Forecasting. In Submission 2022

S2Net with Time-Synchronized Pyramid LSTMs

Time-Synchronized Pyramid LSTMs

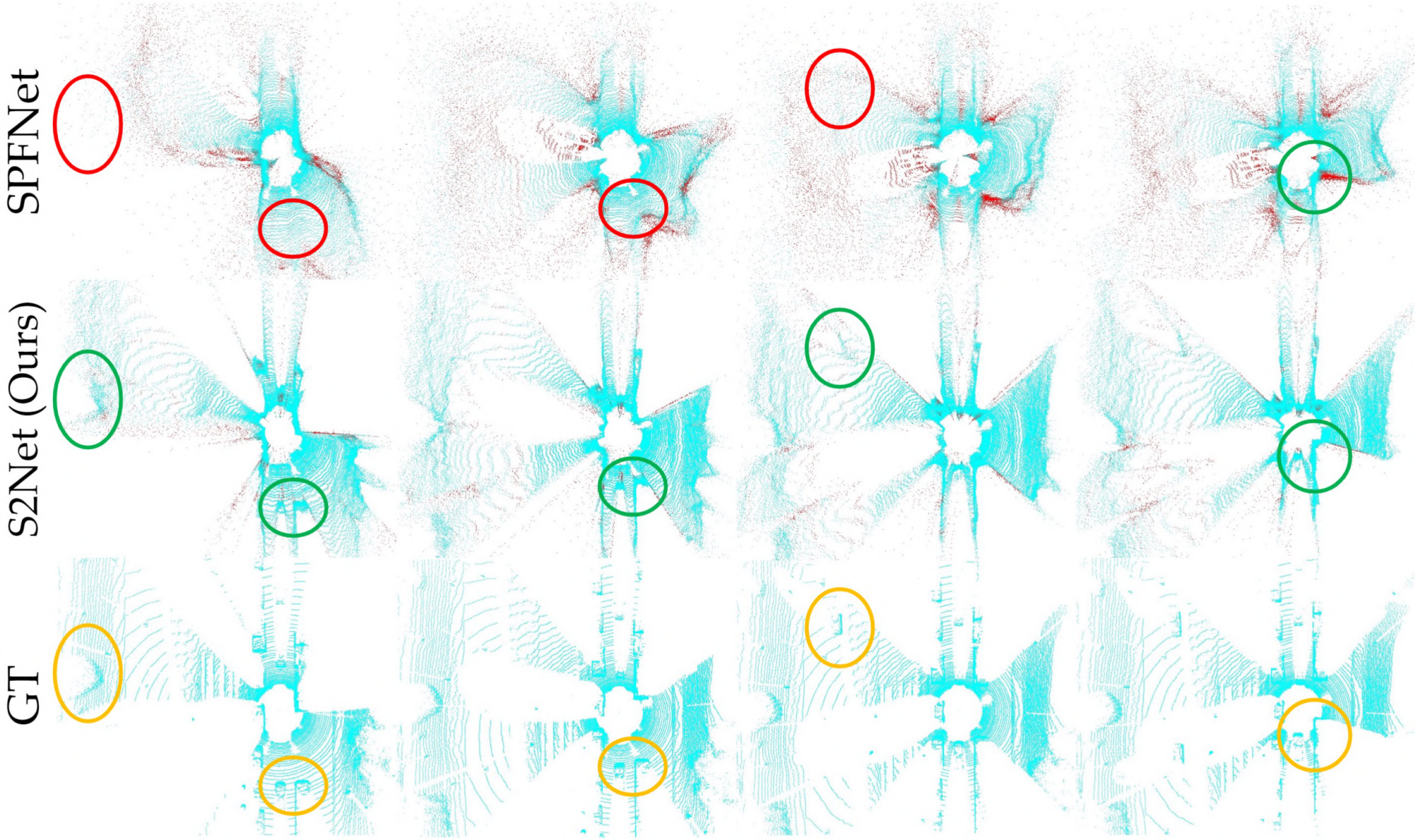
Resolve training instability



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

THE ROBOTICS INSTITUTE

More Fine-Grained Predictions with S2Net



Red points: high CD error

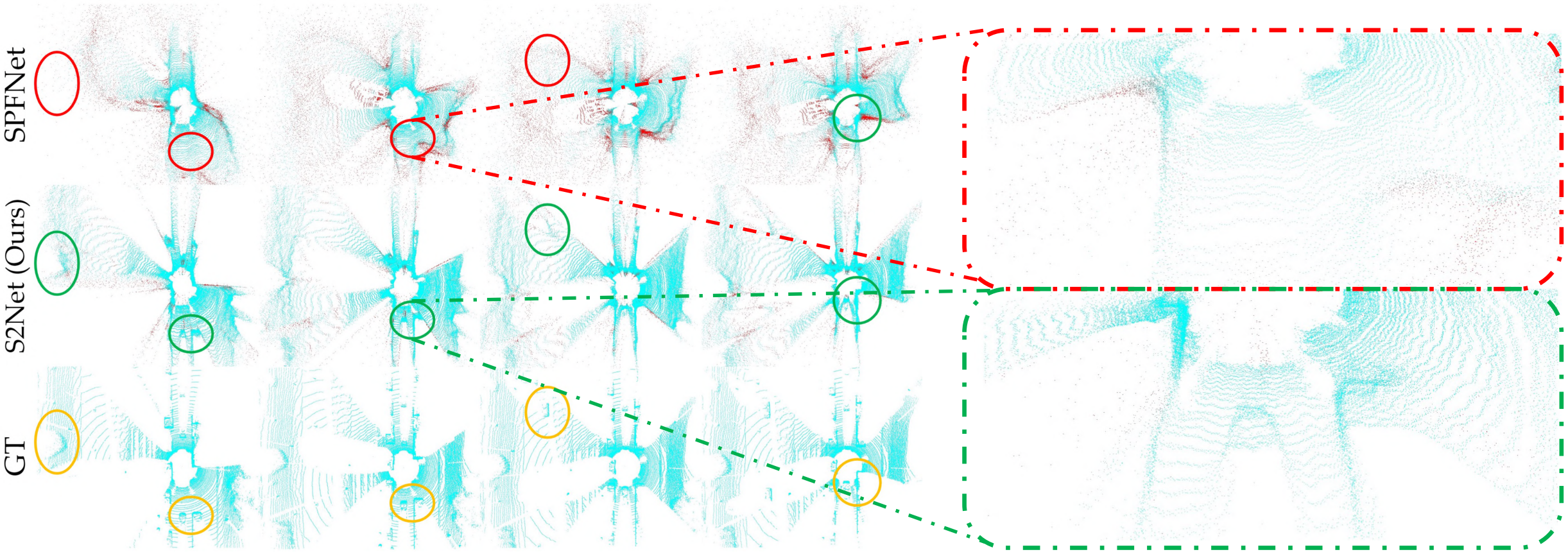
Cyan points: low CD error

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?

Revisit Popular End-to-End Driving Frameworks

Most top methods are based on
Imitation Learning

Scalability & efficiency

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

CARLA Autonomous Driving Leaderboard Home Get started Submit

Leaderboard

Sensors Track – CARLA 0.9.10.1

Copy CSV

Team	Submission	Driving score	Route completion	Infraction penalty	Collisions pedestrians	Collisions vehicles
	Units	%	%	[0, 1]	Infractions/Km	Infractions/Km
Anonymous	TCP	75.14	85.63	0.87	0.00	0.32
WOR	Learning from All Vehicles (LAV)	61.85	94.46	0.64	0.04	0.70
DP	TransFuser	61.18	86.69	0.71	0.04	0.81
Attention Fields	Latent TransFuser	45.20	66.31	0.72	0.02	1.11
RaphaeL	General Reinforced Imitation for Autonomous Driving (GRIAD)	36.79	61.85	0.60	0.00	2.77
DP	TransFuser+	34.58	69.84	0.56	0.04	0.70
WOR	World on Rails	31.37	57.65	0.56	0.61	1.35
MaRLn	MaRLn	24.98	46.97	0.52	0.00	2.33
Attention Fields	Neural Attention Fields (NEAT)	21.83	41.71	0.65	0.04	0.74

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

THE ROBOTICS INSTITUTE

Combining Imitation with Auxiliary Tasks

Sensor inputs

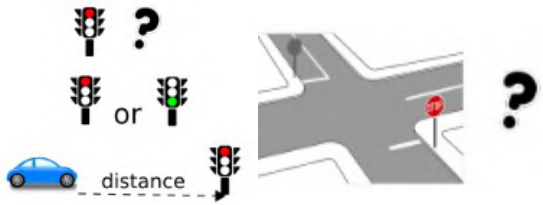
Or

Supervised Semantic Task

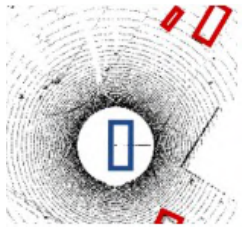
Driving Policy

Pixel/object labels

Segmentation, CVPR '17; IROS '20



Implicit Affordances, CVPR '20



3D Detection, CVPR '22

Although helpful:

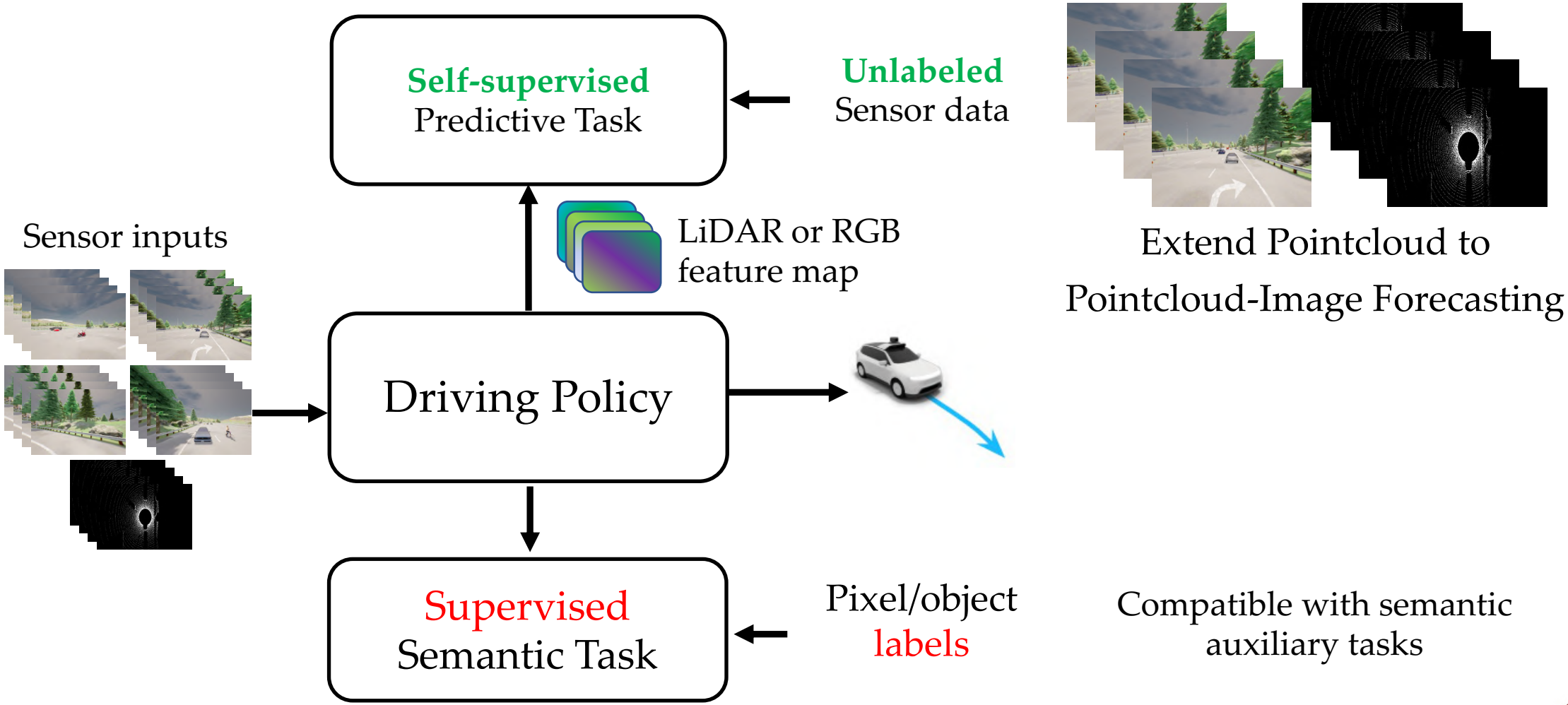
(1) Require labels

(2) Non-predictive representations

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

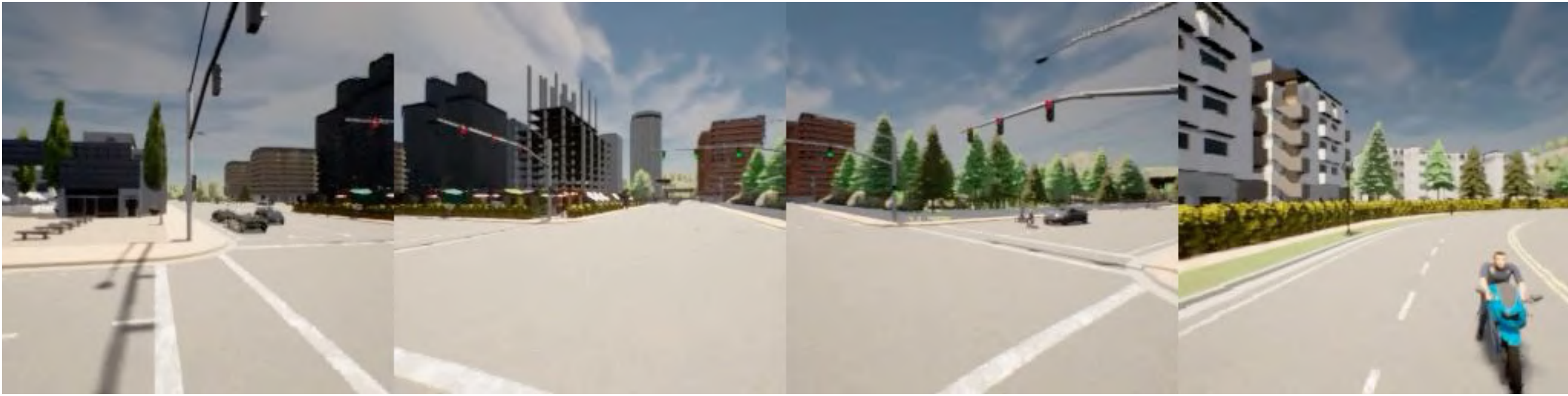
THE ROBOTICS INSTITUTE

Imitation with Predictive Representations



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



Left camera | Front camera | Right camera | Back camera

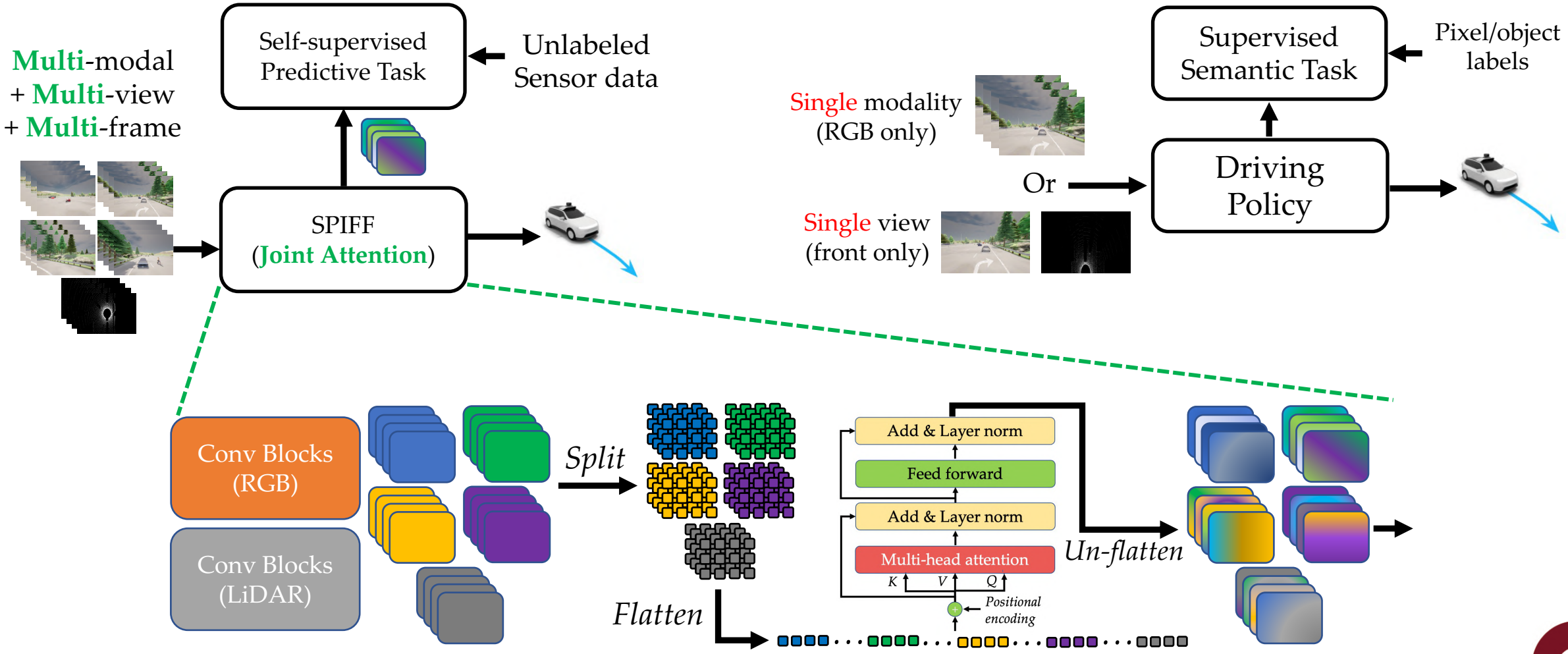
Closed-loop roll-out on unseen towns/routes

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 transFormer. In Submission 2022

THE ROBOTICS INSTITUTE

Learning Informative Representations

via Joint Multi-Modal, Multi-View and Multi-Frame Attention



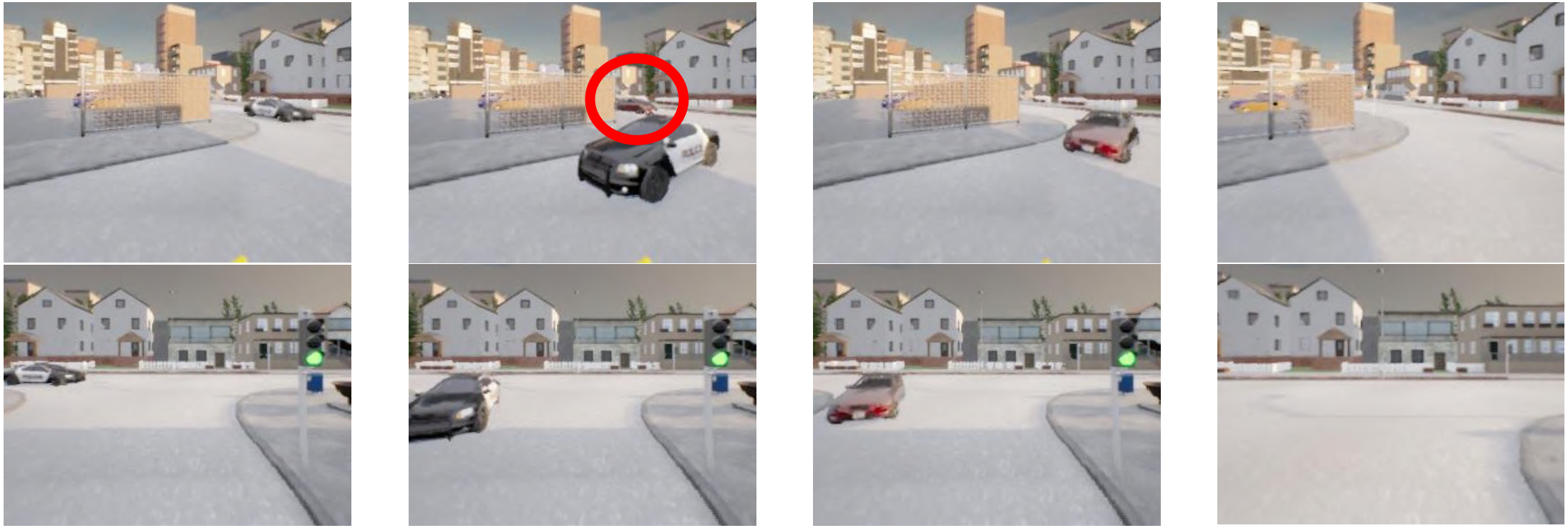
Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFormer. In Submission 2022

THE ROBOTICS INSTITUTE

Using Partial Observations Leads to Training Confusion

Left camera

Front camera



Frame 400 *Frame 470* *Frame 540* *Frame 610*

Still stops, why?

Large peak of errors during training caused by the partial observability

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

THE ROBOTICS INSTITUTE

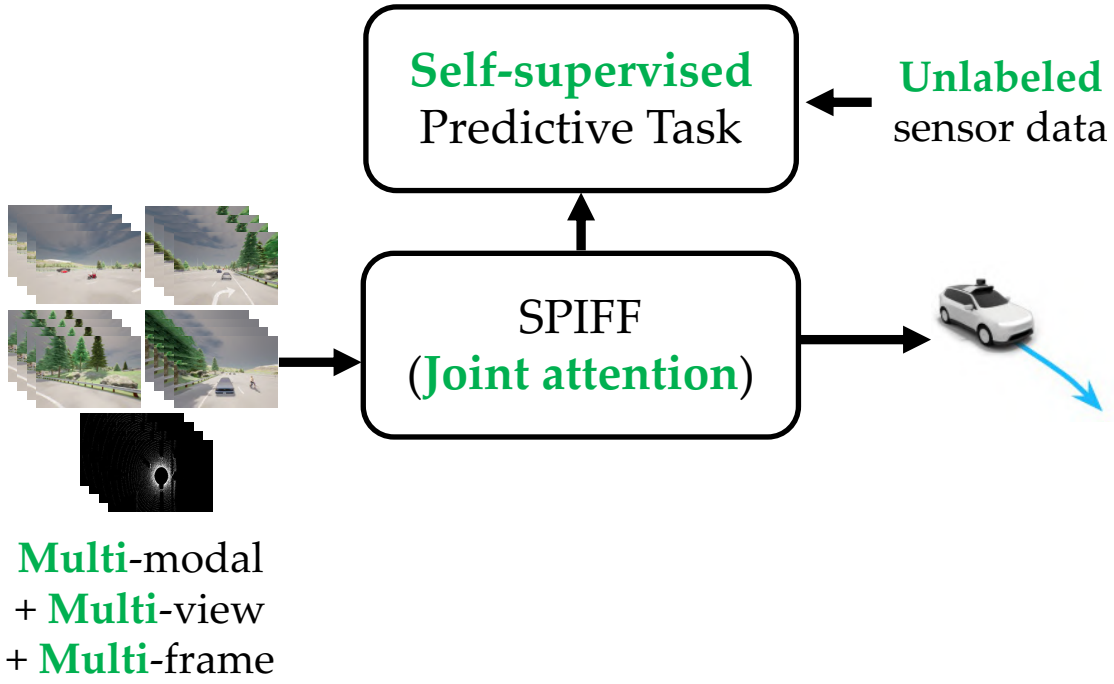
Using Partial Observations Leads to Inertial Problem at Testing

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

THE ROBOTICS INSTITUTE

State-of-the-art Driving Performance on Carla



Informative + Predictive Representations

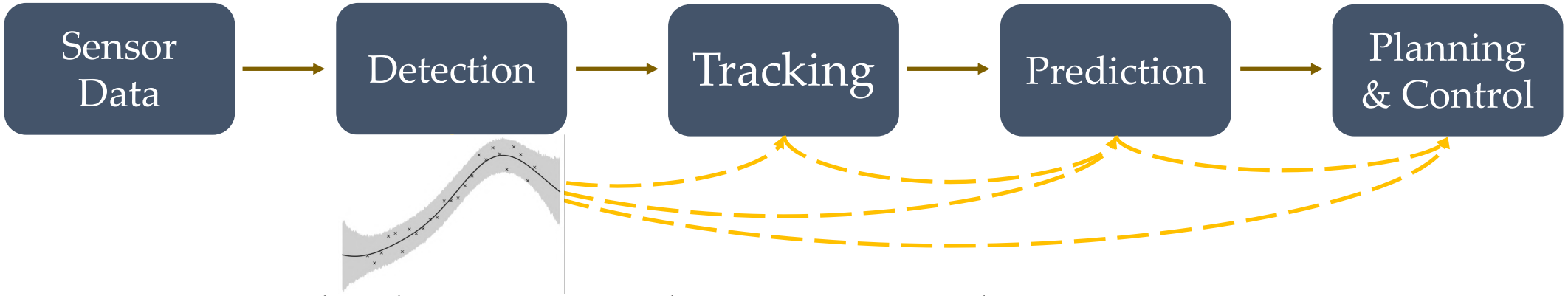
Methods	Average DS ↑
CILRS [32], ICCV '19	8.22 ± 1.50
LBC [33], CoRL '19	14.26 ± 1.90
Transfuser [2], CVPR '21	25.02 ± 0.81
Rails [5], ICCV '21	34.45 ± 1.39
LAV [1], CVPR '22 (Concurrent)	48.75 ± 2.16
SPIFF (ours)	50.97 ± 1.84

Closed-loop evaluation
on 26 Carla official testing routes

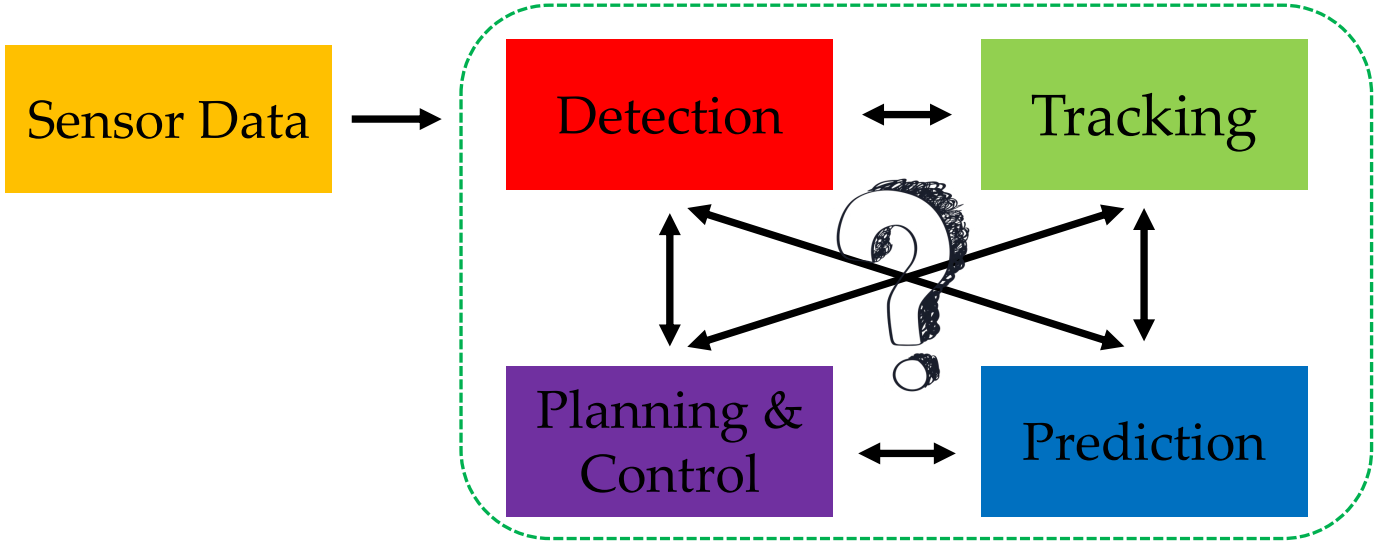
Xinshuo Weng, Junyu Nan, Rowan McAllister, Nicholas Rhinehart, Adrien Gaidon, Kris Kitani.
Learning Predictive Scene Representations for End-to-End Driving with SPIFF: Sequential Pointcloud-Image Forecasting and transFormer. In Submission 2022

THE ROBOTICS INSTITUTE

Summary



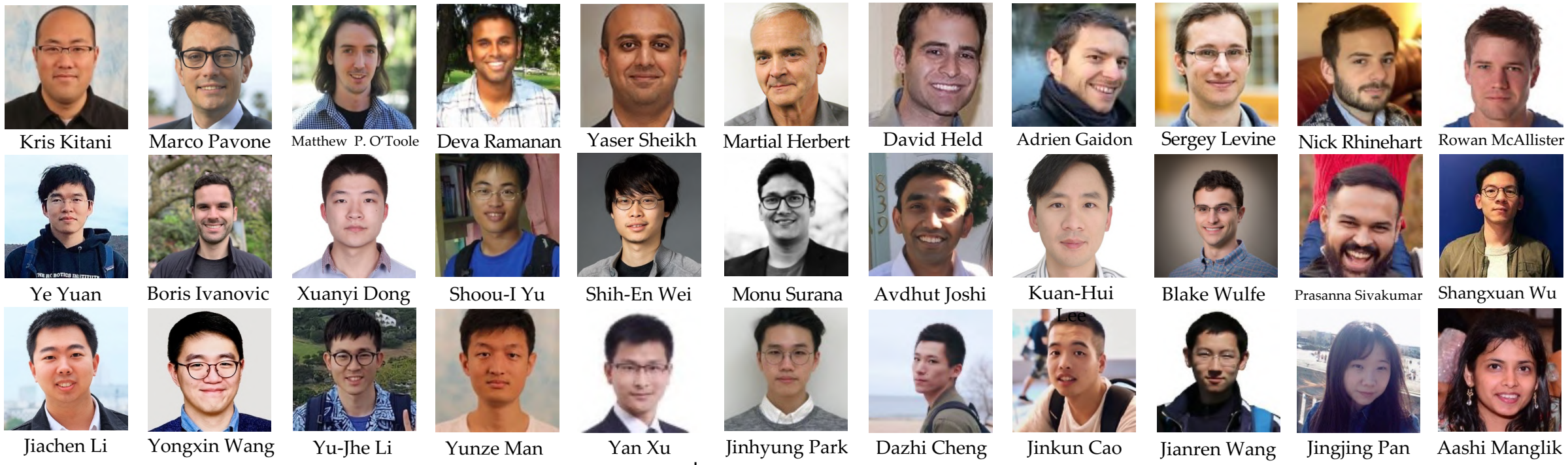
Towards robust autonomy by integrating and propagating uncertainty



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



and many more……

Thank you!

Towards Modular and Differentiable Autonomous Driving

Xinshuo Weng

www.xinshuoweng.com

Robotics Institute, Carnegie Mellon University