MultiXNet: Multiclass Multistage Multimodal Motion Prediction

Djuric, N., Cui, H., Su, Z., Wu, S., Wang, H., Chou, F.-C., San Martin, L., Feng, S., Hu, R., Xu, Y., Dayan, A., Zhang, S., Becker, B. C., Meyer, G. P., Vallespi-Gonzalez, C., Wellington, C. K.

UBER

ATG

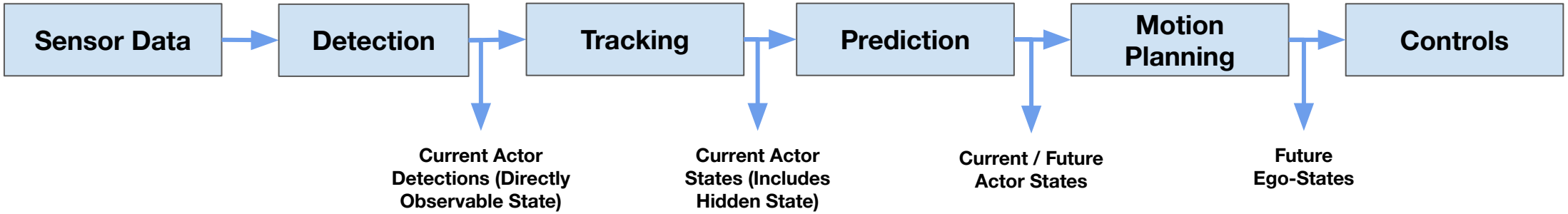
IV21 32nd IEEE INTELLIGENT VEHICLES SYMPOSIUM

Introduction

- We address two critical tasks in self-driving technology
  - Understanding surroundings of an autonomous vehicle (AV)
  - Predicting how these surroundings will change in the near future
- We propose a novel end-to-end model that performs both simultaneously
  - Historical lidar sweep data and map info used as inputs
  - Run detection and prediction for three actor types (vehicles, pedestrians, and bicyclists), with multimodal trajectory predictions for vehicles
- The method achieves state-of-the-art performance over the baselines

Autonomy System

- Typical AV system processes sensor data in a sequence of modules
  - Detection takes in sensor data and outputs detected objects and their states
  - Prediction takes in current states of traffic actors, and outputs their trajectories

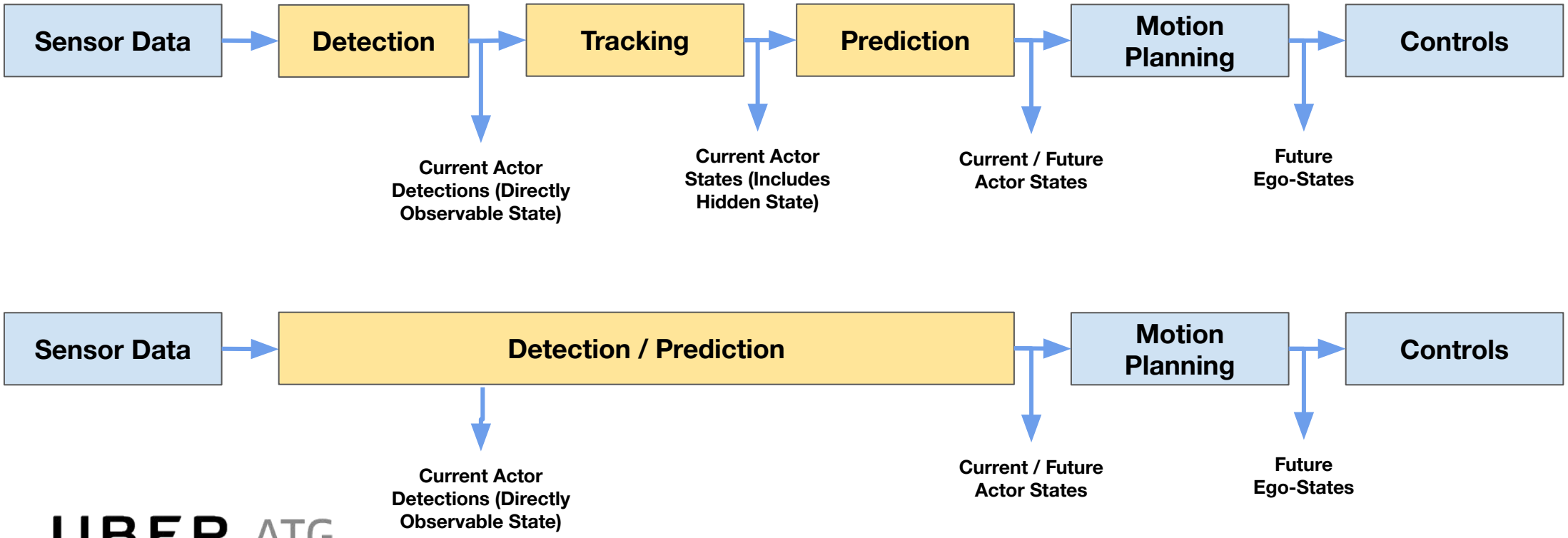


UBER ATG

IV21 32nd IEEE INTELLIGENT VEHICLES SYMPOSIUM

Autonomy System

- In previous work Detection and Prediction modules running in a sequence
  - Using an end-to-end Detection/Prediction module instead?



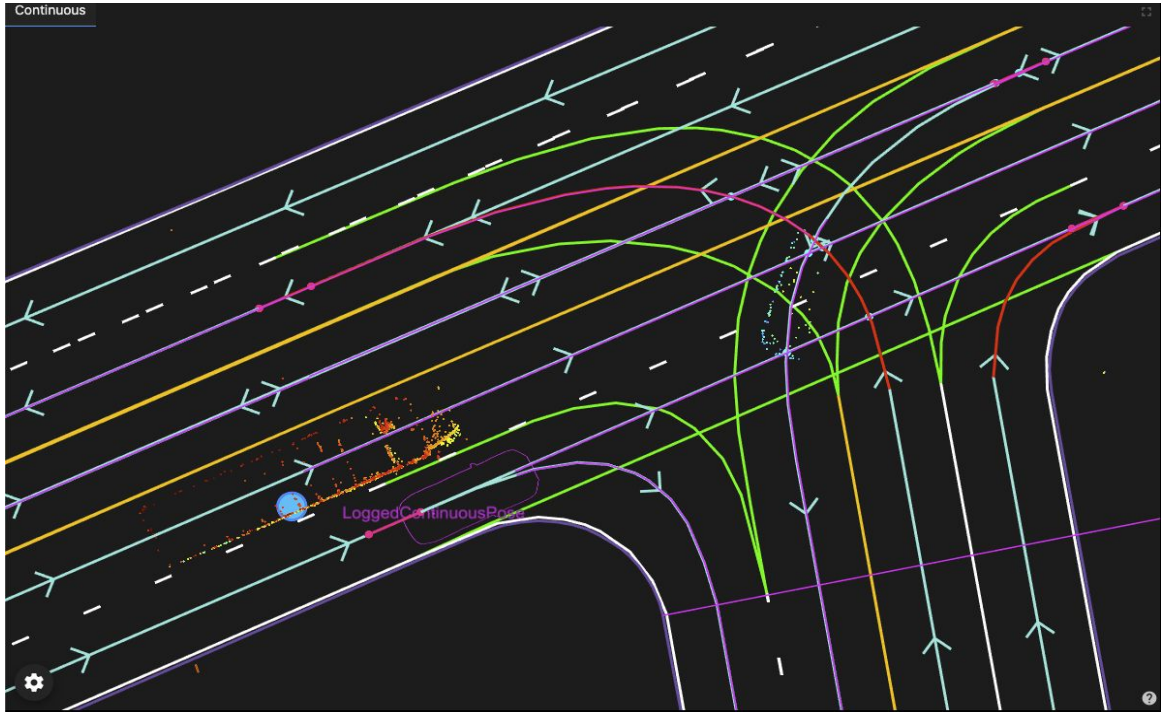
UBER ATG

Joint detection/prediction

- Existing work mostly focuses on one of the tasks, without considering joint models
- **Pros of joint training**
  - Unified model with end-to-end training
  - Simpler online system
- **Cons of joint training**
  - More complex deep model

Example output

- Goal: Produce object detections, their states, and future trajectories



Example output

- Goal: Produce **object detections, their states**, and future trajectories

Example output

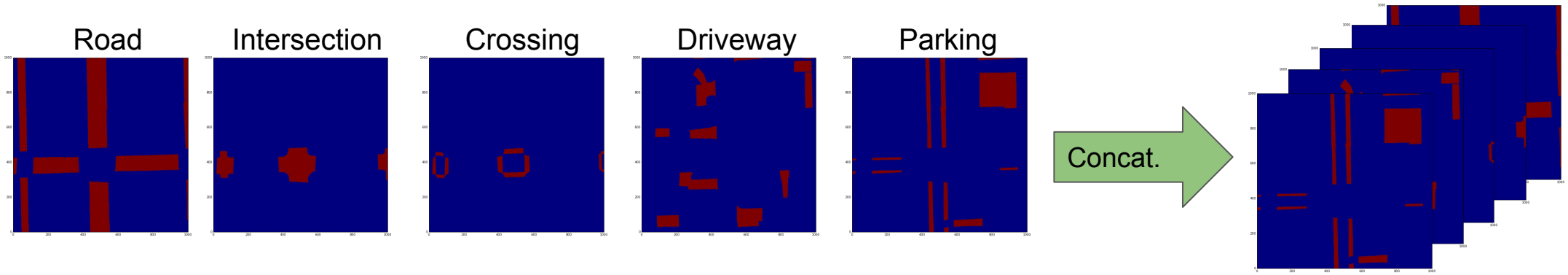
- Goal: Produce object detections, their states, and **future trajectories**

Example output

- Goal: Produce object detections, their states, and **future trajectories**

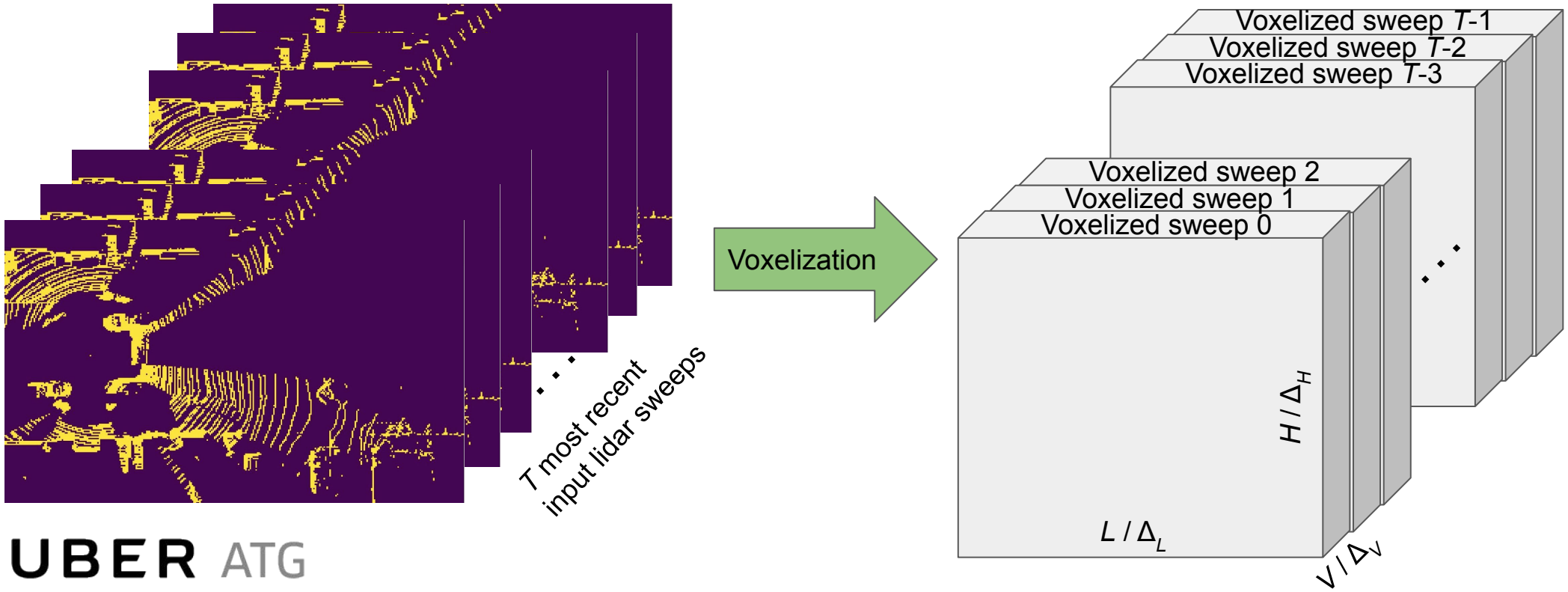
Input representation

- The model takes historical lidar sweeps and current map as input
- Generate rasterized image of the world with pre-specified resolution centered on the AV
  - Different map elements are represented as a binary image and concatenated



Lidar sweep voxelization

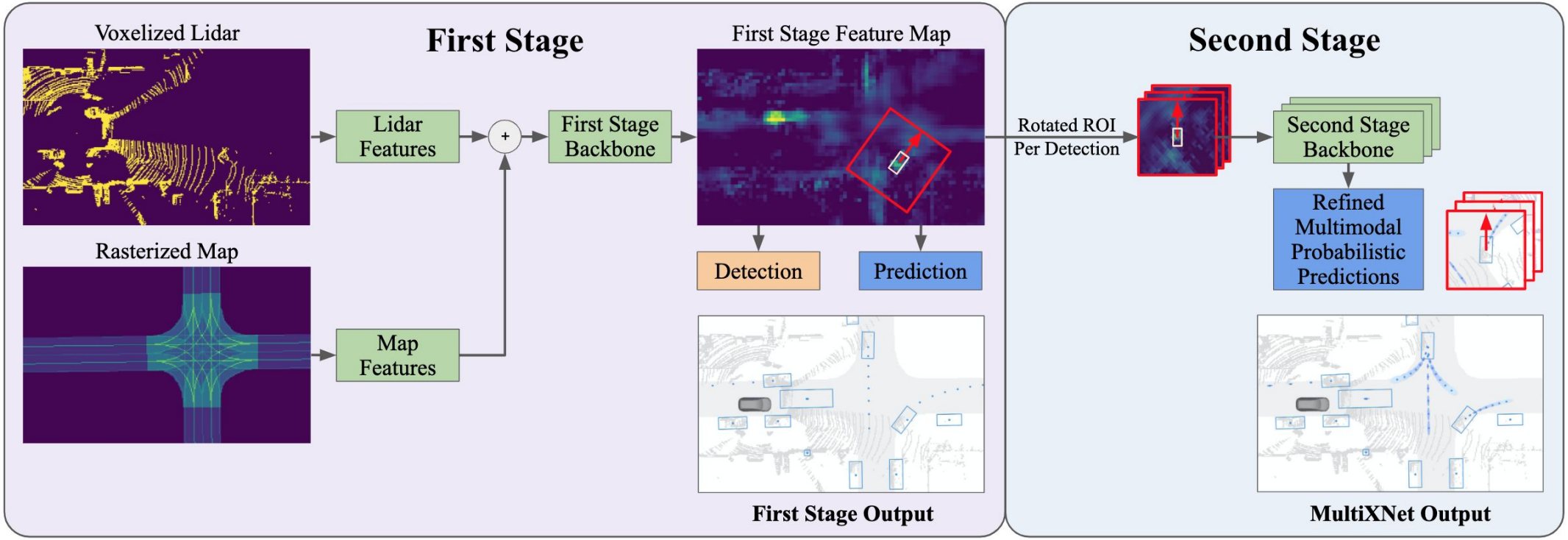
- Given a lidar sweep, its returns are encoded in a binary 3D occupancy image centered on the AV, with length L (150m), width W (100m), height V (3.2m)
  - Resolution Δ_L and Δ_W are set to 0.16m, Δ_V to 0.2m
  - Voxel value is set to 1 if at least one lidar point falls within it, and 0 otherwise



UBER ATG

MultiXNet

- Raster-based end-to-end model architecture



UBER ATG

Detection and prediction loss

- Detection/prediction losses computed for background (*bg*) and foreground (*fg*) cells
  - Existence probability, length, width, *x*/*y* position, and actor heading

$$\mathcal{L}_{bg} = \ell_{focal}(1 - \hat{p})$$

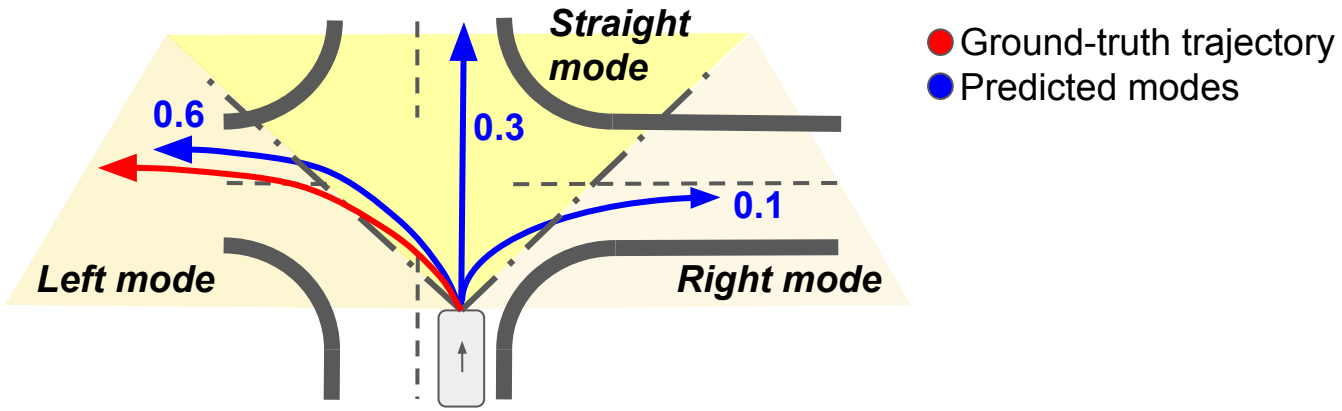
$$\mathcal{L}_{fg(h)} = 1_{h=0}\Big(\ell_{focal}(\hat{p}) + \ell_1(\hat{l} - l) + \ell_1(\hat{w} - w)\Big) + \ell_1(\hat{c}_{xh} - c_{xh}) + \ell_1(\hat{c}_{yh} - c_{yh}) + \ell_1(sin\hat{\theta}_h - sin\theta_h) + \ell_1(cos\hat{\theta}_h - cos\theta_h)$$

- Overall loss is summed over all horizons, over all grid cells

$$\mathcal{L} = 1_{\text{bg cell}}\mathcal{L}_{bg} + 1_{\text{fg cell}}\sum_{h=0}^{H}\lambda^h\mathcal{L}_{fg(h)}$$

Multi-modal prediction loss

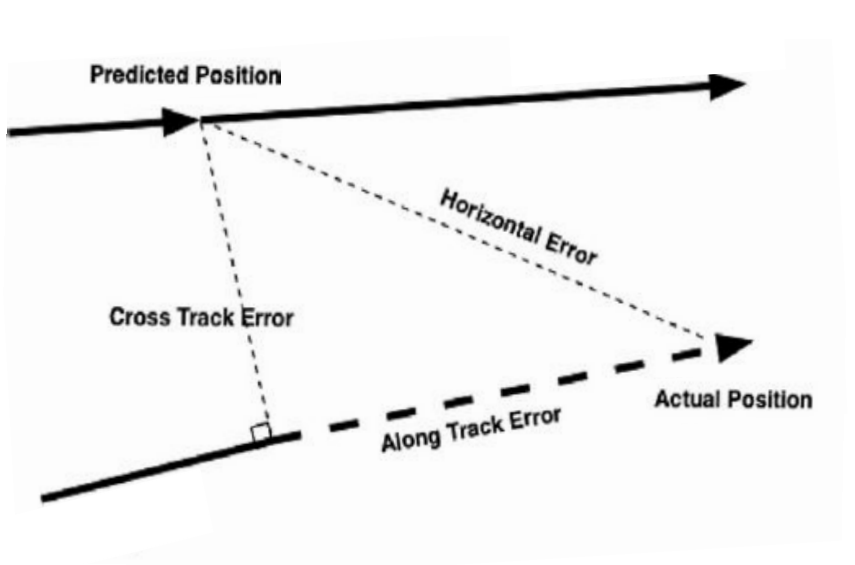
- Multi-modal prediction loss is used in the second stage, where we assume left-turning, right-turning, and going-straight modes, and predict each trajectory and their probability



- During training only the mode assigned to the region where the ground-truth is located is updated (in this case the left-turning mode), and its probability pushed towards 1

Uncertainty-aware loss

- Uncertainty-aware loss
  - Along-/cross-track decomposition
  - Gaussian vs. Laplace distribution
  - Negative log-likelihood vs. KL-divergence
- KL-divergence
  - *Ground-truth* and *predicted* AT/CT uncertainty



$$Laplace(0, b_{AT}),\ b_{AT}(t) = \alpha_{AT} + \beta_{AT} t$$

$$Laplace(\hat{e}_{AT}, \hat{b}_{AT})$$

$$KL_{AT} = \log \frac{\hat{b}_{AT}}{b_{AT}} + \frac{b_{AT} \exp\left(-\frac{|\hat{e}_{AT}|}{b_{AT}}\right) + |\hat{e}_{AT}|}{\hat{b}_{AT}} - 1$$

Results

- Comparison to the state-of-the-art baselines on nuScenes data

	Vehicles			Pedestrians			Bicyclists		
Method	**AP**	**DE**	**CT**	**AP**	**DE**	**CT**	**AP**	**DE**	**CT**
SpAGNN	-	145.0	-	-	-	-	-	-	-
IntentNet (DE)	61.0	117.4	37.7	64.4	83.5	47.3	30.9	184.6	73.7
IntentNet (AT/CT)	60.3	118.3	37.8	63.4	83.6	46.8	31.8	**173.0**	70.2
MultiXNet	60.6	**105.0** (**104.2**)	**29.7** (**29.1**)	66.1	**80.1**	**43.8**	32.6	203.1	**54.8**

Results

- Ablation study on ATG4D data

			Vehicles			Pedestrians			Bicyclists		
Unc.	2nd	Mm.	AP	DE	CT	AP	DE	CT	AP	DE	CT
			83.9	90.4	26.0	88.4	61.8	32.9	83.2	51.7	23.5
KL-L			84.1	91.9	22.8	88.2	**57.1**	**30.4**	84.6	49.9	21.1
	✓		84.6	**82.2**	22.2	88.7	63.2	33.2	84.3	51.6	23.8
KL-L	✓		84.4	83.3	20.4	88.4	57.6	30.6	83.9	52.0	21.7
	✓	✓	84.0	82.4 (**81.4**)	22.4 (21.8)	88.5	62.6	33.0	84.2	51.2	23.7
KL-L	✓	✓	84.2	83.1 (82.1)	**20.2** (19.8)	88.4	57.2	30.5	84.6	**48.5**	20.7
KL-G	✓	✓	85.0	84.5 (83.4)	20.6 (**19.7**)	88.7	58.1	31.3	84.7	50.4	**20.6**
NL-L	✓	✓	84.6	85.3 (84.0)	**20.2** (19.8)	88.4	58.4	31.0	83.9	50.4	20.9
NL-G	✓	✓	84.5	88.5 (87.7)	21.1 (20.7)	88.6	59.3	32.0	83.5	51.8	21.0
KL-L	no rot.	✓	84.4	86.9 (86.0)	21.9 (21.2)	88.6	57.3	30.7	84.0	50.4	21.5

Results

- Ablation study on ATG4D data

			Vehicles			Pedestrians			Bicyclists		
Unc.	2nd	Mm.	AP	DE	CT	AP	DE	CT	AP	DE	CT
			83.9	90.4	26.0	88.4	61.8	32.9	83.2	51.7	23.5
KL-L			84.1	91.9	22.8	88.2	**57.1**	**30.4**	84.6	49.9	21.1
	✓		84.6	**82.2**	22.2	88.7	63.2	33.2	84.3	51.6	23.8
KL-L	✓		84.4	83.3	20.4	88.4	57.6	30.6	83.9	52.0	21.7
	✓	✓	84.0	82.4 (**81.4**)	22.4 (21.8)	88.5	62.6	33.0	84.2	51.2	23.7
KL-L	✓	✓	84.2	83.1 (82.1)	**20.2** (19.8)	88.4	57.2	30.5	84.6	**48.5**	20.7
KL-G	✓	✓	85.0	84.5 (83.4)	20.6 (**19.7**)	88.7	58.1	31.3	84.7	50.4	**20.6**
NL-L	✓	✓	84.6	85.3 (84.0)	**20.2** (19.8)	88.4	58.4	31.0	83.9	50.4	20.9
NL-G	✓	✓	84.5	88.5 (87.7)	21.1 (20.7)	88.6	59.3	32.0	83.5	51.8	21.0
KL-L	no rot.	✓	84.4	86.9 (86.0)	21.9 (21.2)	88.6	57.3	30.7	84.0	50.4	21.5

Results

- Ablation study on ATG4D data

			Vehicles			Pedestrians			Bicyclists		
Unc.	2nd	Mm.	AP	DE	CT	AP	DE	CT	AP	DE	CT
			83.9	90.4	26.0	88.4	61.8	32.9	83.2	51.7	23.5
KL-L			84.1	91.9	22.8	88.2	**57.1**	**30.4**	84.6	49.9	21.1
	✓		84.6	**82.2**	22.2	88.7	63.2	33.2	84.3	51.6	23.8
KL-L	✓		84.4	83.3	20.4	88.4	57.6	30.6	83.9	52.0	21.7
	✓	✓	84.0	82.4 (**81.4**)	22.4 (21.8)	88.5	62.6	33.0	84.2	51.2	23.7
KL-L	✓	✓	84.2	83.1 (82.1)	**20.2** (19.8)	88.4	57.2	30.5	84.6	**48.5**	20.7
KL-G	✓	✓	85.0	84.5 (83.4)	20.6 (**19.7**)	88.7	58.1	31.3	84.7	50.4	**20.6**
NL-L	✓	✓	84.6	85.3 (84.0)	**20.2** (19.8)	88.4	58.4	31.0	83.9	50.4	20.9
NL-G	✓	✓	84.5	88.5 (87.7)	21.1 (20.7)	88.6	59.3	32.0	83.5	51.8	21.0
KL-L	no rot.	✓	84.4	86.9 (86.0)	21.9 (21.2)	88.6	57.3	30.7	84.0	50.4	21.5

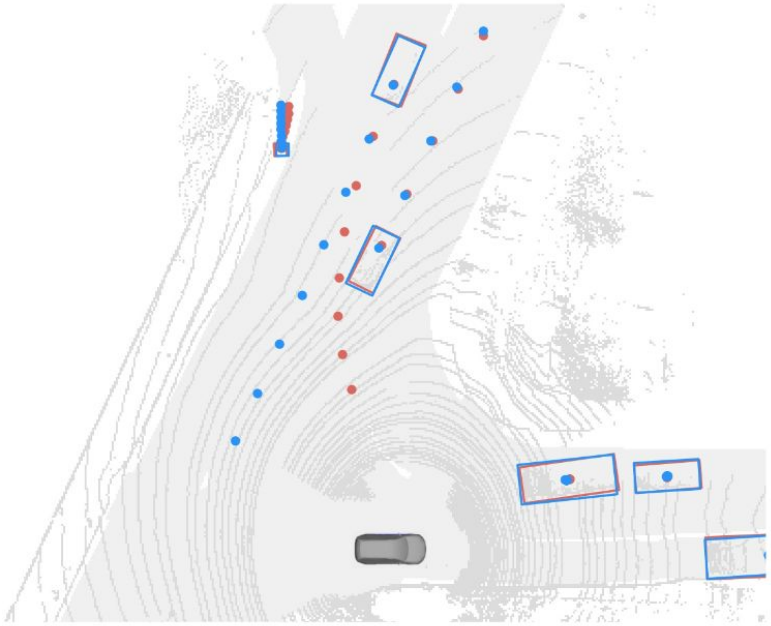
Results

- Ablation study on ATG4D data

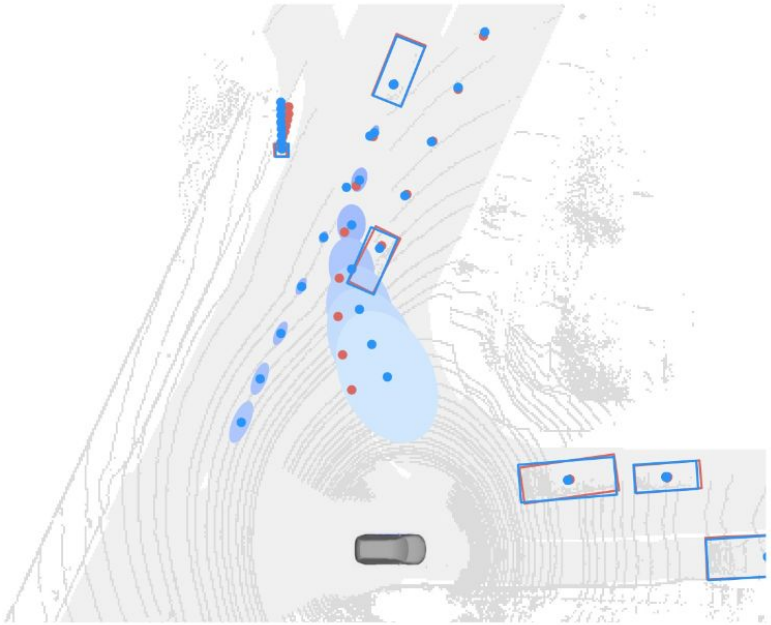
			Vehicles			Pedestrians			Bicyclists		
Unc.	2nd	Mm.	AP	DE	CT	AP	DE	CT	AP	DE	CT
			83.9	90.4	26.0	88.4	61.8	32.9	83.2	51.7	23.5
KL-L			84.1	91.9	22.8	88.2	**57.1**	**30.4**	84.6	49.9	21.1
	✓		84.6	**82.2**	22.2	88.7	63.2	33.2	84.3	51.6	23.8
KL-L	✓		84.4	83.3	20.4	88.4	57.6	30.6	83.9	52.0	21.7
	✓	✓	84.0	82.4 (**81.4**)	22.4 (21.8)	88.5	62.6	33.0	84.2	51.2	23.7
KL-L	✓	✓	84.2	83.1 (82.1)	**20.2** (19.8)	88.4	57.2	30.5	84.6	**48.5**	20.7
KL-G	✓	✓	85.0	84.5 (83.4)	20.6 (**19.7**)	88.7	58.1	31.3	84.7	50.4	**20.6**
NL-L	✓	✓	84.6	85.3 (84.0)	**20.2** (19.8)	88.4	58.4	31.0	83.9	50.4	20.9
NL-G	✓	✓	84.5	88.5 (87.7)	21.1 (20.7)	88.6	59.3	32.0	83.5	51.8	21.0
KL-L	no rot.	✓	84.4	86.9 (86.0)	21.9 (21.2)	88.6	57.3	30.7	84.0	50.4	21.5

Results

- Case study analysis



IntentNet



MultiXNet

Ground-truth trajectory

Predicted trajectory

UBER ATG

IV21 32nd IEEE INTELLIGENT VEHICLES SYMPOSIUM

Conclusion

- Joint models show great promise
- State-of-the-art performance with reduced overall system complexity
- Avenues to improve the performance
  - New sensor inputs (camera, radar)
  - Physical feasibility of trajectories
  - Stronger map constraints
- Questions?