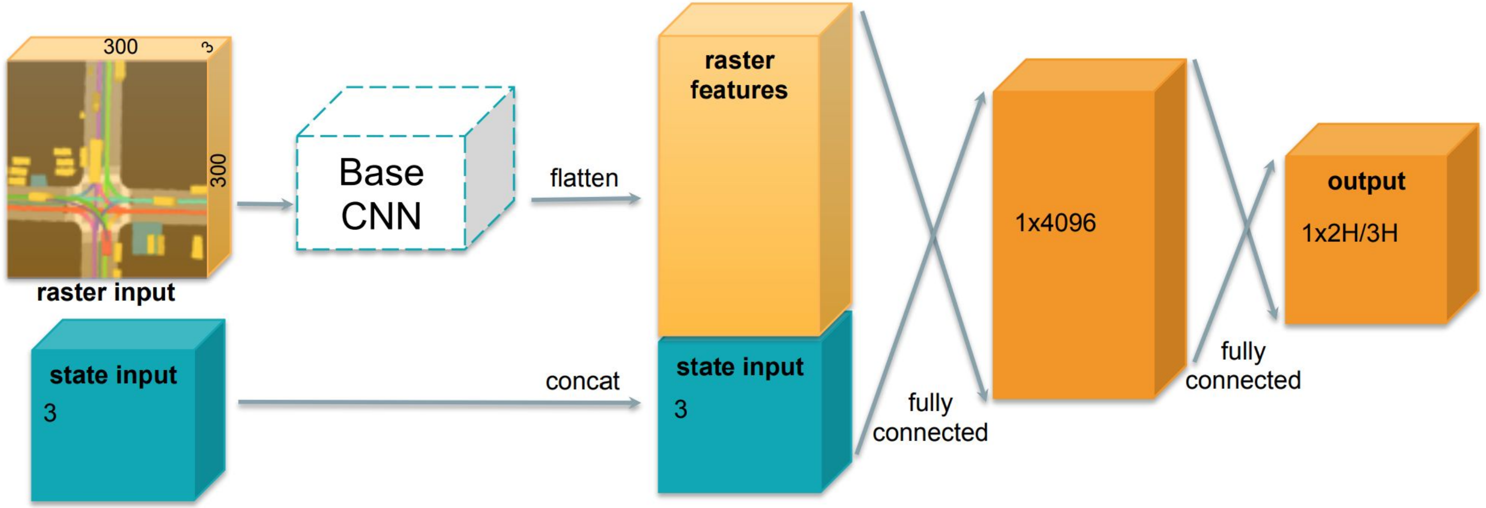
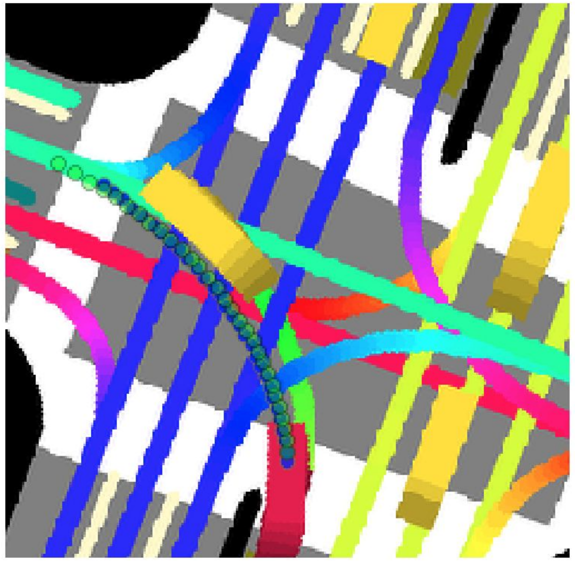
Predicting Motion of Vulnerable Road Users using High-Definition Maps and Efficient ConvNets

Fang-Chieh Chou, Tsung-Han Lin, Henggang Cui, Vladan Radosavljevic,
Thi Nguyen, Tzu-Kuo Huang, Matthew Niedoba, Jeff Schneider, Nemanja Djuric

Trajectory prediction in autonomous driving

- Goal: predicting future trajectory (x,y positions) of traffic actors around the ego vehicle.
- Inputs:
  - Reliable detection and tracking (current and past states of actors).
  - Scene context: HD map with lane graph, traffic light.
- Previous work (RasterNet) rasterizes the scene context and fuses the state input in one convolutional network for accurate vehicle trajectory predictions.





UBER ATG

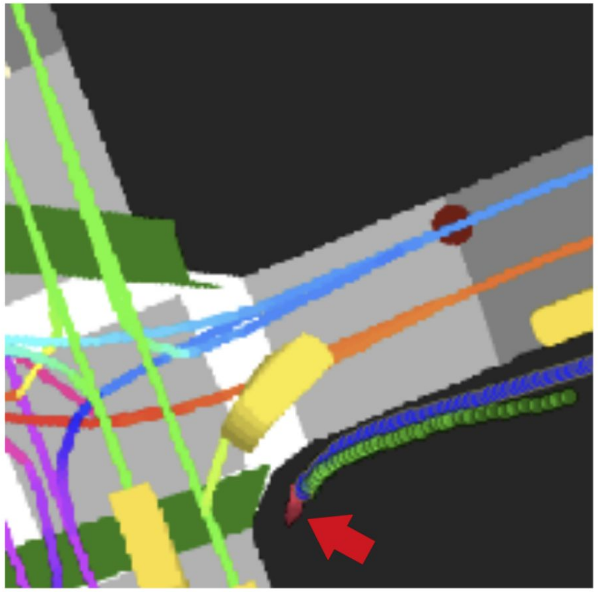
Djuric, N., et al. "Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving." *WACV* 2020.

Key Contributions

- Extended RasterNet for VRU (vulnerable road users, i.e. pedestrians and bicyclists in this work) trajectory prediction.
- Improved network architecture for backbone layers and raster-state fusion, for faster inference and more accurate predictions.
- Performed detailed ablation study on different rasterization settings to identify the optimal setting, and to provide insights into which parts of the system contribute the most to the accuracy.

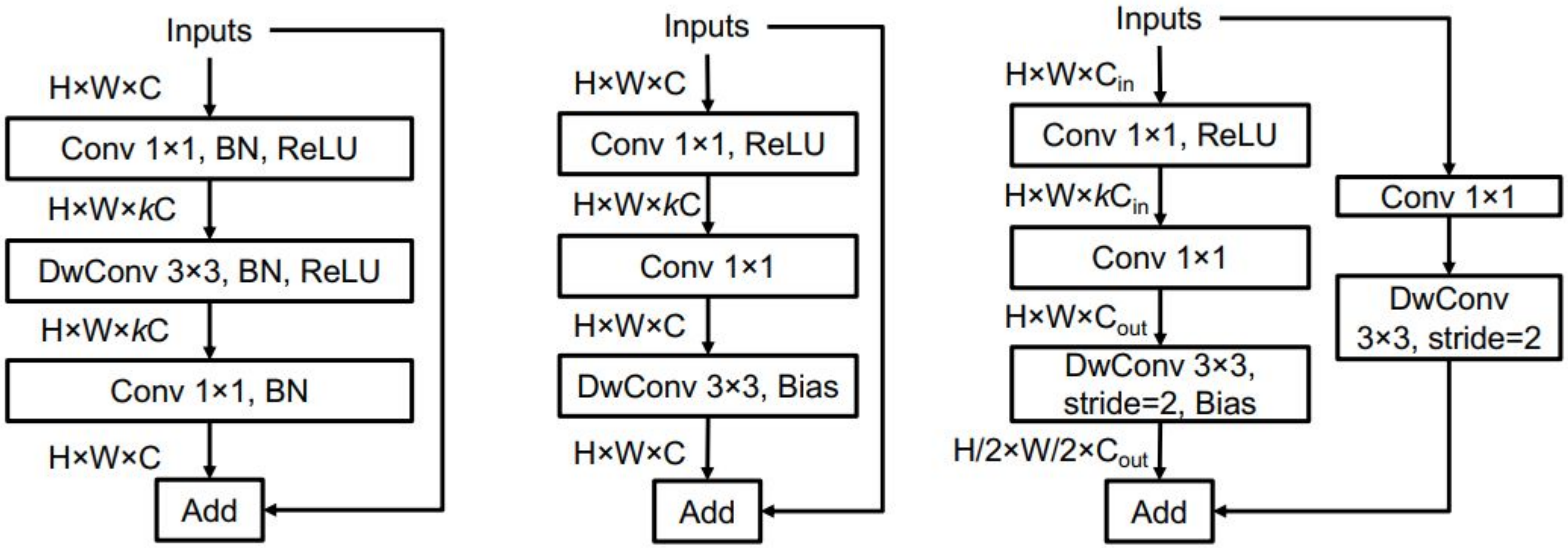
Trajectory prediction for VRUs

- We rasterize the target actor and surrounding scene context into an RGB image.
- By default, target actor is placed at bottom-center.
- Raster is rotated such that the target actor’s orientation points to north.
- Traffic light information:
  - Colored circles represents the traffic light color of the lane.
  - Green crosswalks means the vehicles have the right of way.



Improved backbone networks (FastMobileNet, FMNet)

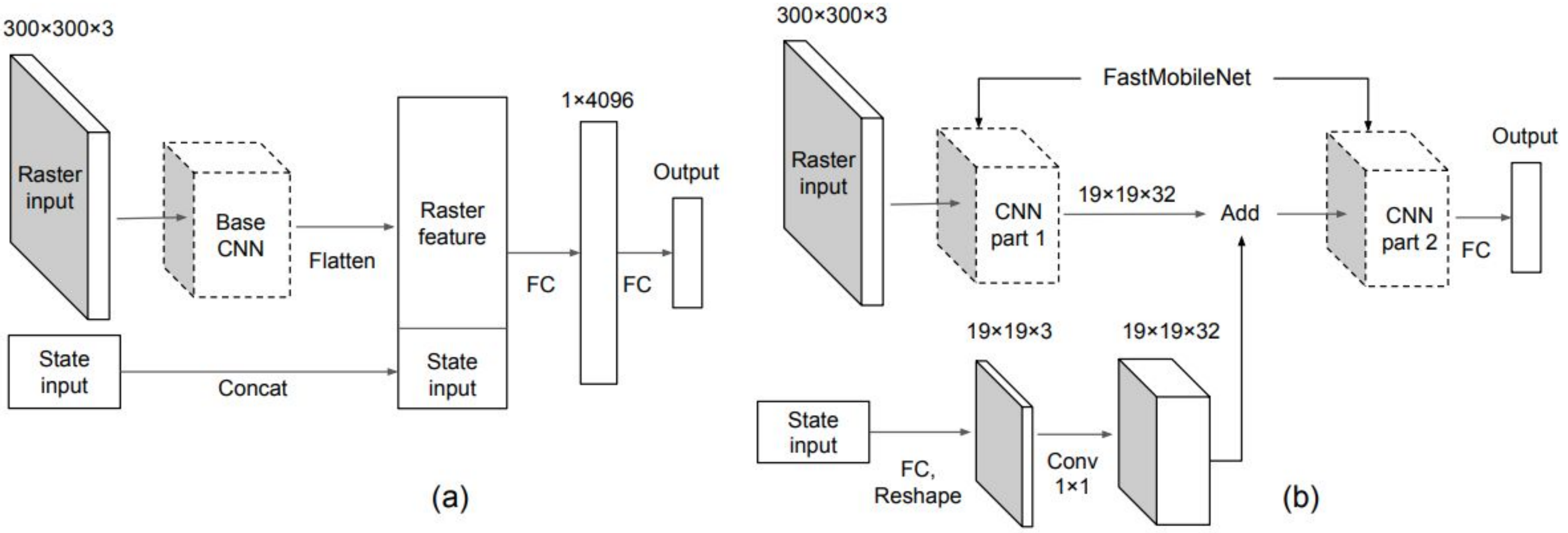
- Modified MobileNet-V2 (MNv2) architecture with faster inference.
- Moved the depthwise convolution to operate on less channels .
- Removed BatchNorm for faster inference.



(a) Regular block of MNv2 (b) Regular block of FMNet (c) Stride=2 block of FMNet

Improved state-raster spatial fusion

- Learned projection of state inputs to the 2D CNN feature map.
- Benefits: better metrics due to spatial fusion. Faster inference without expensive final FC fusion layer.



UBER ATG

Network architecture improvement results

- Latency is measured at batch-size=32 on a GTX 1080Ti.
- FMNet has much improved latency due to less number of tensor operations and memory access operations (MAC).
- Spatial fusion further improves latency and average displacement error (ADE).

Architecture	ADE [m]	Latency [ms]	FLOPS	Num. parameters	MAC	Num. ops
AlexNet	1.36	15.8	2.63G	70.3M	364 MB	**131**
ResNet18	1.29	36.2	6.26G	11.7M	163 MB	641
MNv2-0.5	1.27	21.3	308M	598K	146 MB	1542
MnasNet-0.5	1.28	18.3	323M	844K	113 MB	1490
FMNet	1.28	12.1	340M	565K	55 MB	336
FMNet with spatial fusion	**1.24**	**10.4**	**285M**	**558K**	**47 MB**	370

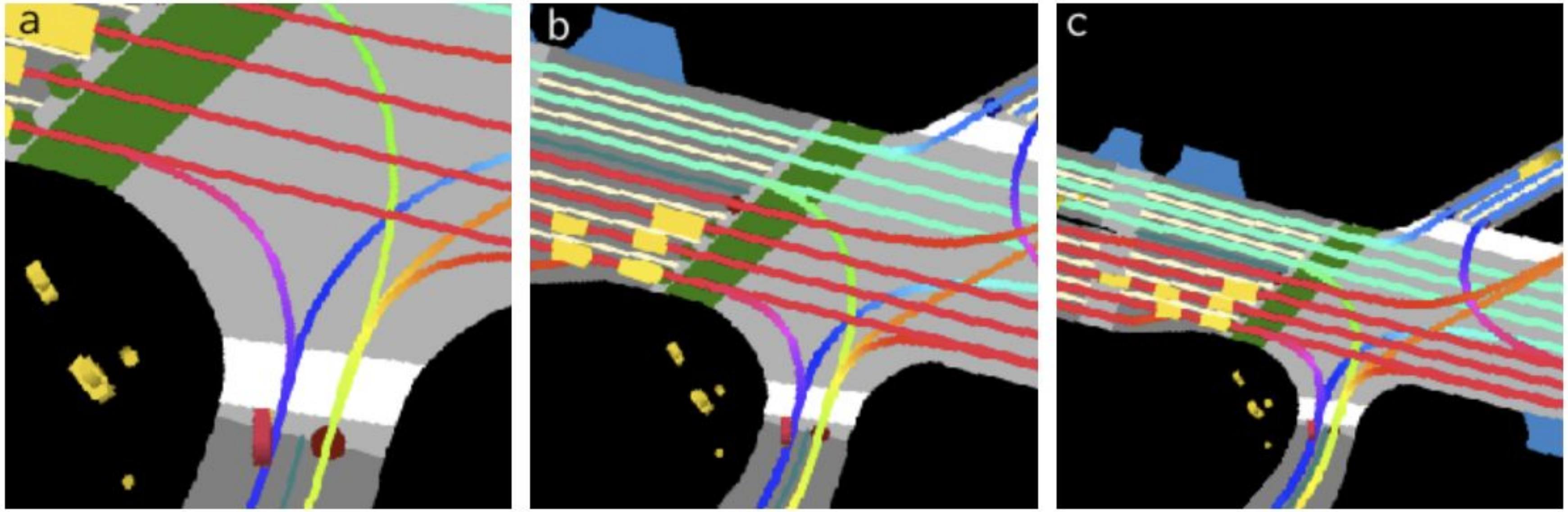
Network architecture improvement results

- Latency is measured at batch-size=32 on a GTX 1080Ti.
- FMNet has much improved latency due to less number of tensor operations and memory access operations (MAC).
- Spatial fusion further improves latency and average displacement error (ADE).

Architecture	ADE [m]	Latency [ms]	FLOPS	Num. parameters	MAC	Num. ops
AlexNet	1.36	15.8	2.63G	70.3M	364 MB	**131**
ResNet18	1.29	36.2	6.26G	11.7M	163 MB	641
MNv2-0.5	1.27	21.3	308M	598K	146 MB	1542
MnasNet-0.5	1.28	18.3	323M	844K	113 MB	1490
FMNet	1.28	12.1	340M	565K	55 MB	336
FMNet with spatial fusion	**1.24**	**10.4**	**285M**	**558K**	**47 MB**	370

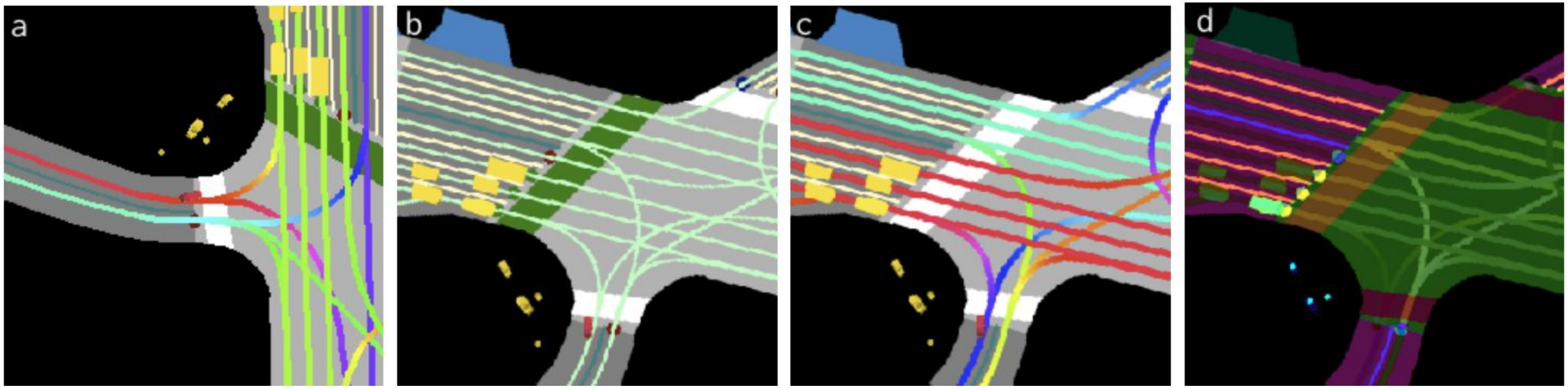
Rasterization setting ablation

Resolution=0.1m Resolution=0.2m Resolution=0.3m



Rasterization setting ablation

No raster rotation | No lane heading | No traffic light | Learned rasterization



Rasterization setting ablation

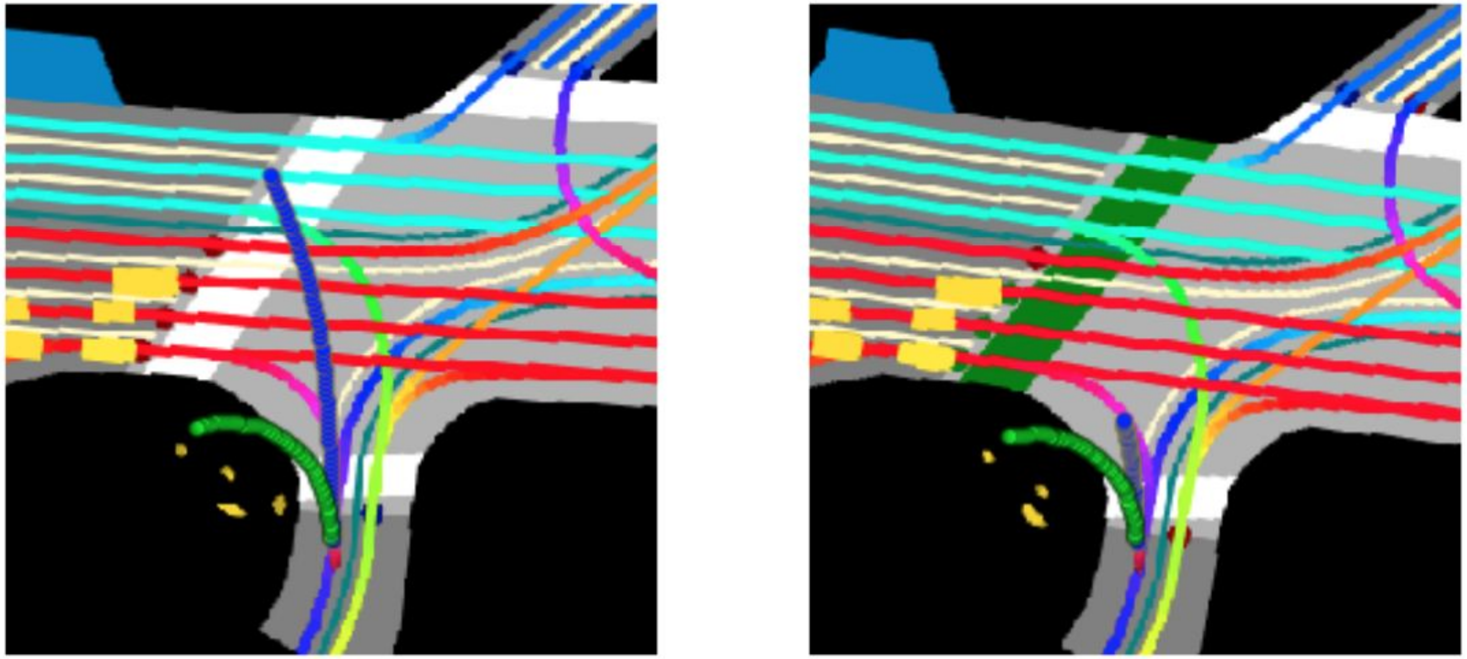
Approach	Resolution	Bicyclists Average	Bicyclists @1s	Bicyclists @5s	Pedestrians Average	Pedestrians @1s	Pedestrians @5s
UKF	–	2.89	0.80	6.60	0.67	0.22	1.22
Social-LSTM	–	3.79	1.85	6.61	0.53	0.29	0.95
RasterNet	$0.1m$	1.07	0.43	2.73	**0.51**	**0.17**	**0.90**
RasterNet	$0.2m$	1.07	0.44	2.72	0.52	0.18	0.93
RasterNet	$0.3m$	1.09	0.45	2.80	0.53	0.18	0.95
RasterNet w/o rotation	$0.2m$	1.29	0.49	3.30	0.58	0.20	1.02
RasterNet w/o traffic lights	$0.2m$	1.11	0.44	2.86	0.55	0.20	0.96
RasterNet w/o lane headings	$0.2m$	1.07	0.43	2.72	0.52	0.18	0.93
RasterNet with learned colors	$0.2m$	**1.05**	**0.42**	**2.70**	0.53	0.18	0.93

Rasterization setting ablation

Approach	Resolution	Bicyclists Average	Bicyclists @1s	Bicyclists @5s	Pedestrians Average	Pedestrians @1s	Pedestrians @5s
UKF	–	2.89	0.80	6.60	0.67	0.22	1.22
Social-LSTM	–	3.79	1.85	6.61	0.53	0.29	0.95
RasterNet	0.1m	1.07	0.43	2.73	**0.51**	**0.17**	**0.90**
RasterNet	0.2m	1.07	0.44	2.72	0.52	0.18	0.93
RasterNet	0.3m	1.09	0.45	2.80	0.53	0.18	0.95
RasterNet w/o rotation	0.2m	1.29	0.49	3.30	0.58	0.20	1.02
RasterNet w/o traffic lights	0.2m	1.11	0.44	2.86	0.55	0.20	0.96
RasterNet w/o lane headings	0.2m	1.07	0.43	2.72	0.52	0.18	0.93
RasterNet with learned colors	0.2m	**1.05**	**0.42**	**2.70**	0.53	0.18	0.93

Qualitative examples

- Model prediction reacts the traffic light state changes.



UBER ATG

Conclusions

- Successfully applied prior vehicle trajectory prediction method (RasterNet) to VRUs.
- Proposed architecture improvements, both in model backbone and raster-state input fusion, lead to better inference latency and prediction.
- Detailed rasterization ablation analysis reveals the factors that are important to accurate VRU trajectory prediction.
- Following completion of offline tests the system was successfully tested onboard SDVs.