Vision-Centric 3D Perception for Scalable Autonomous Driving

Hang Zhao

Assistant Professor
IIIS, Tsinghua University

CVPR22 Tutorial
2022.06.20

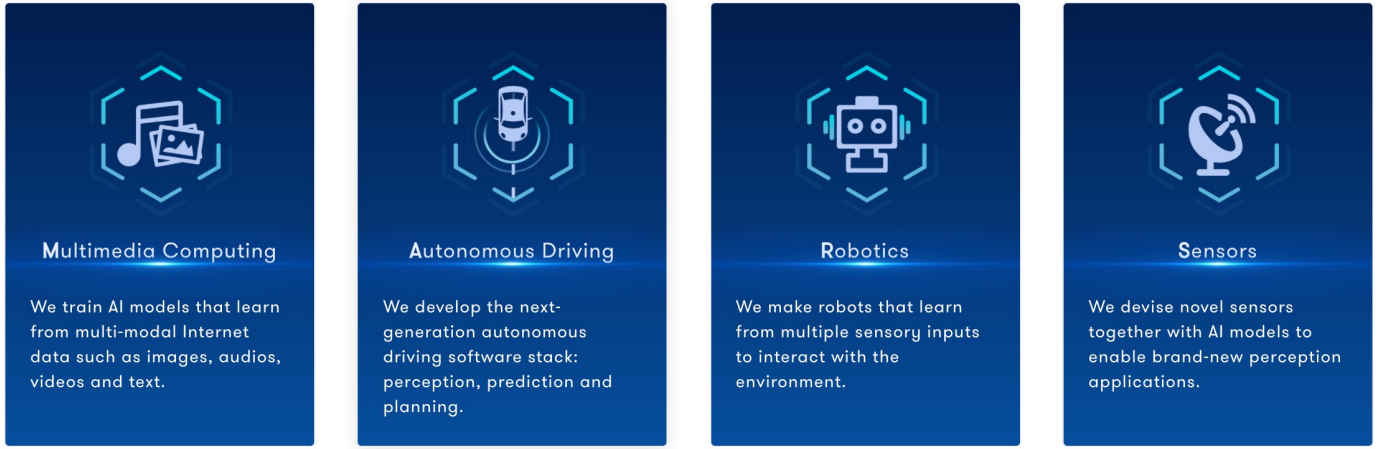
About

Assistant Professor @ IIIS, Tsinghua University

- PI, Leading MARS Lab
- Autonomous Driving
- Multimodal Learning

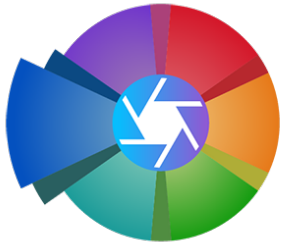
https://hangzhaomit.github.io

MARS Lab



http://group.iiis.tsinghua.edu.cn/~marslab/

Vision-Centric Autonomous Driving



VCAD

VISION-CENTRIC AUTONOMOUS DRIVING

DETR3D

Detection Transformer 3D

FUTR3D

Sensor Fusion Transformer 3D

MUTR3D

Multi-object Tracking Transformer 3D

HDMAPNET

High-Definition Map Learning

VECTOR MAPNET

https://vcad-ai.github.io

Why Vision-Centric?

- Under-explored
  - Rich attributes, beyond geometry
  - Holistic scene understanding

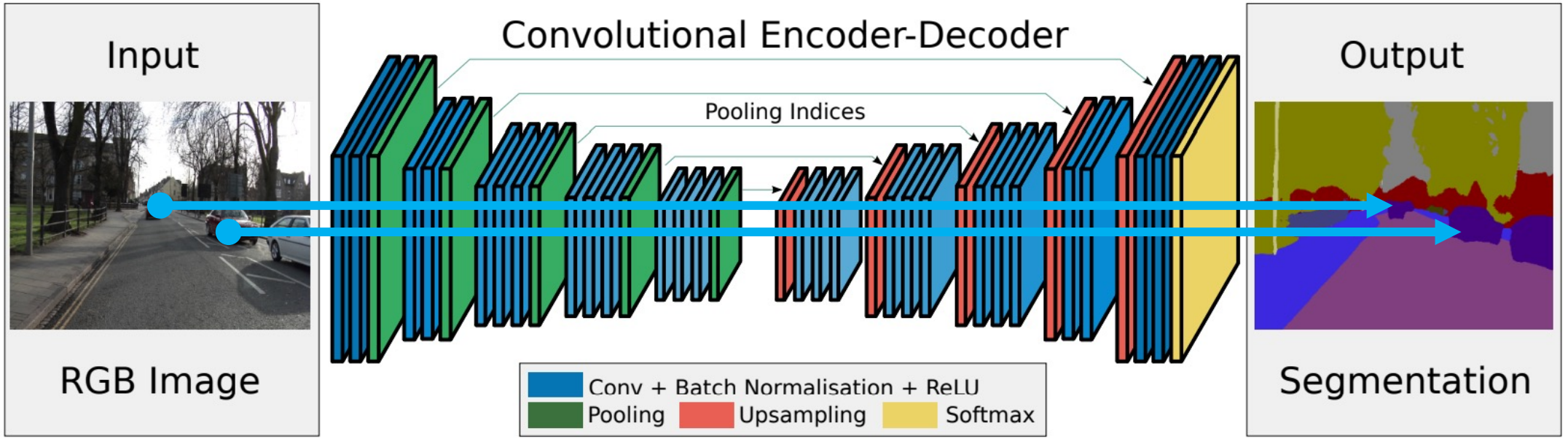
Why Vision-Centric?

- Scalable
  - Affordable for all vehicles
  - Data collection, compression, transmission, etc.

- **NOT** Vision-only
  - Sensor fusion is critical for safety

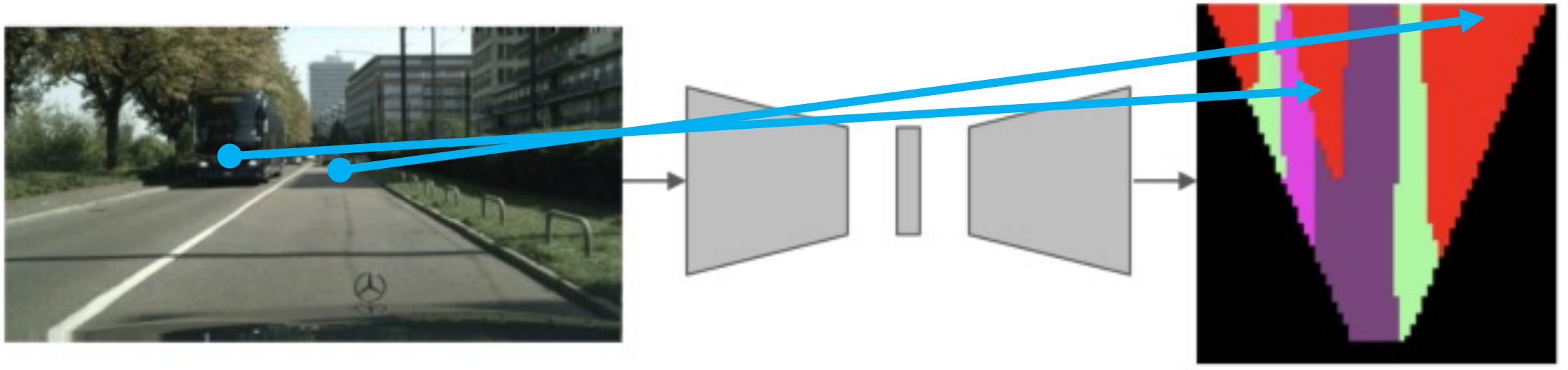
Perspective Perception

- Pixel-level alignment between input/output



BEV/3D Perception

- View Transformation



Our Work on VCAD

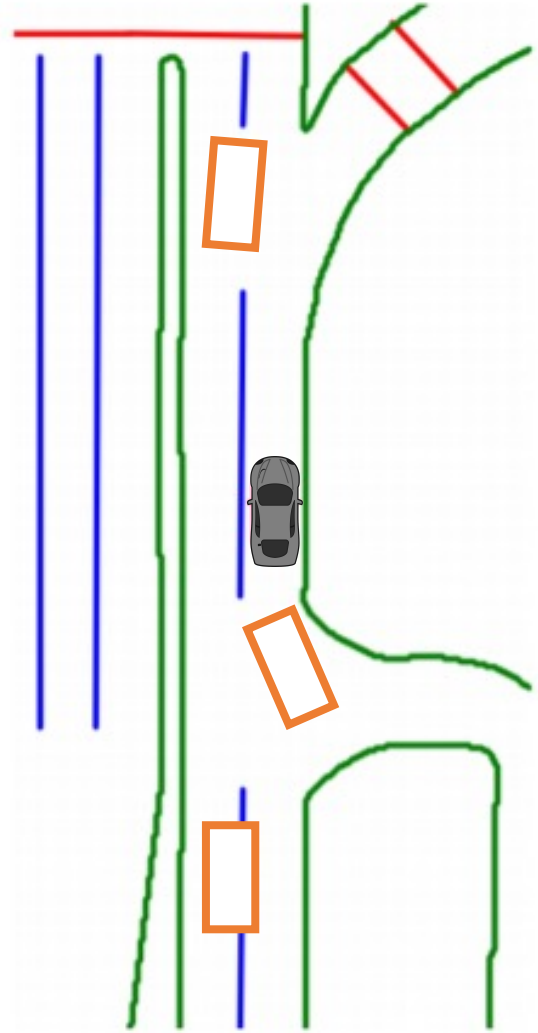
Surround View Vision

HDMapNet
VectorMapNet

Static Map Elements

DETR3D
FUTR3D
MUTR3D

Dynamic Objects



DETR3D
Detection Transformer 3D

DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries

CoRL 2021

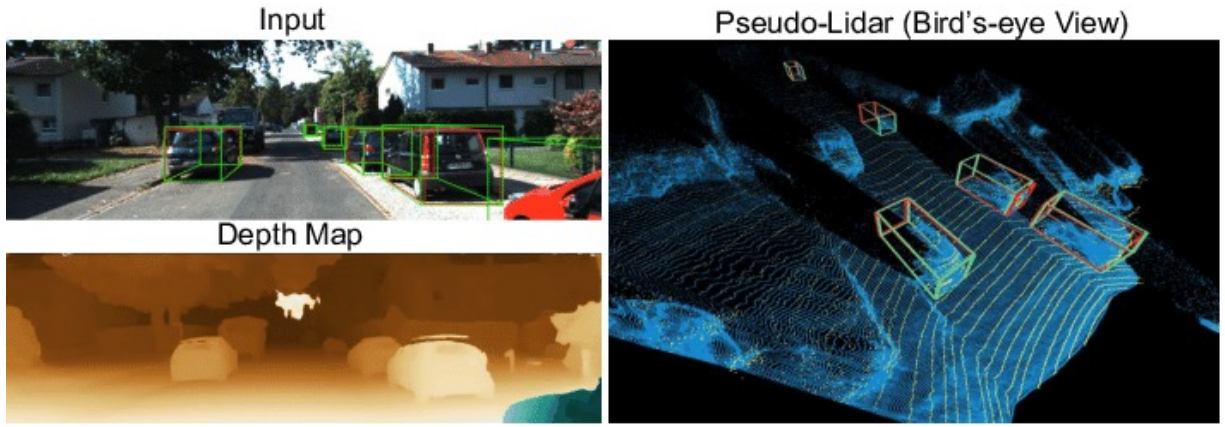
https://tsinghua-mars-lab.github.io/detr3d/

https://arxiv.org/abs/2110.06922

Yue Wang, Vitor Guizilini, Tianyuan Zhang
Yilun Wang, Hang Zhao, Justin Solomon

Related Work on Monocular 3D Detection

- Pseudo-LiDAR
- First predict per-pixel depth, then perform 3D detection on point clouds



Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving, Wang et al., CVPR 2019

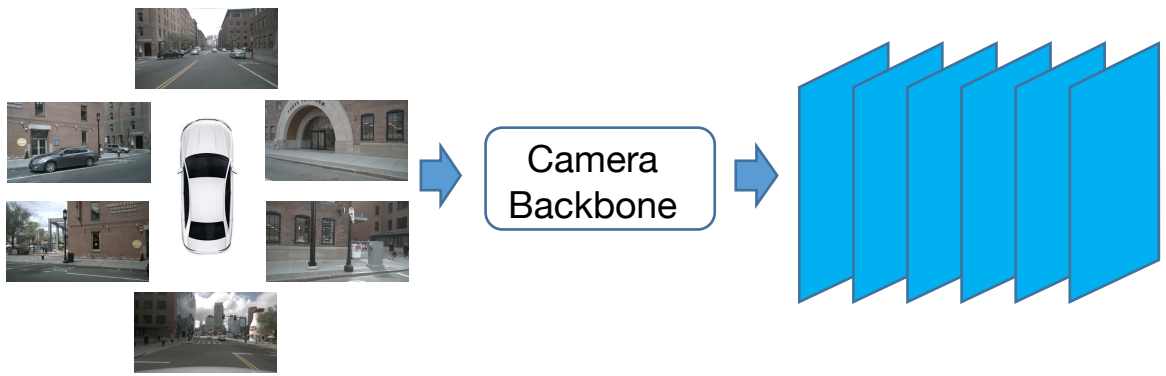
- Requires additional supervision
- Two stages, compounding error

Advantages of DETR3D

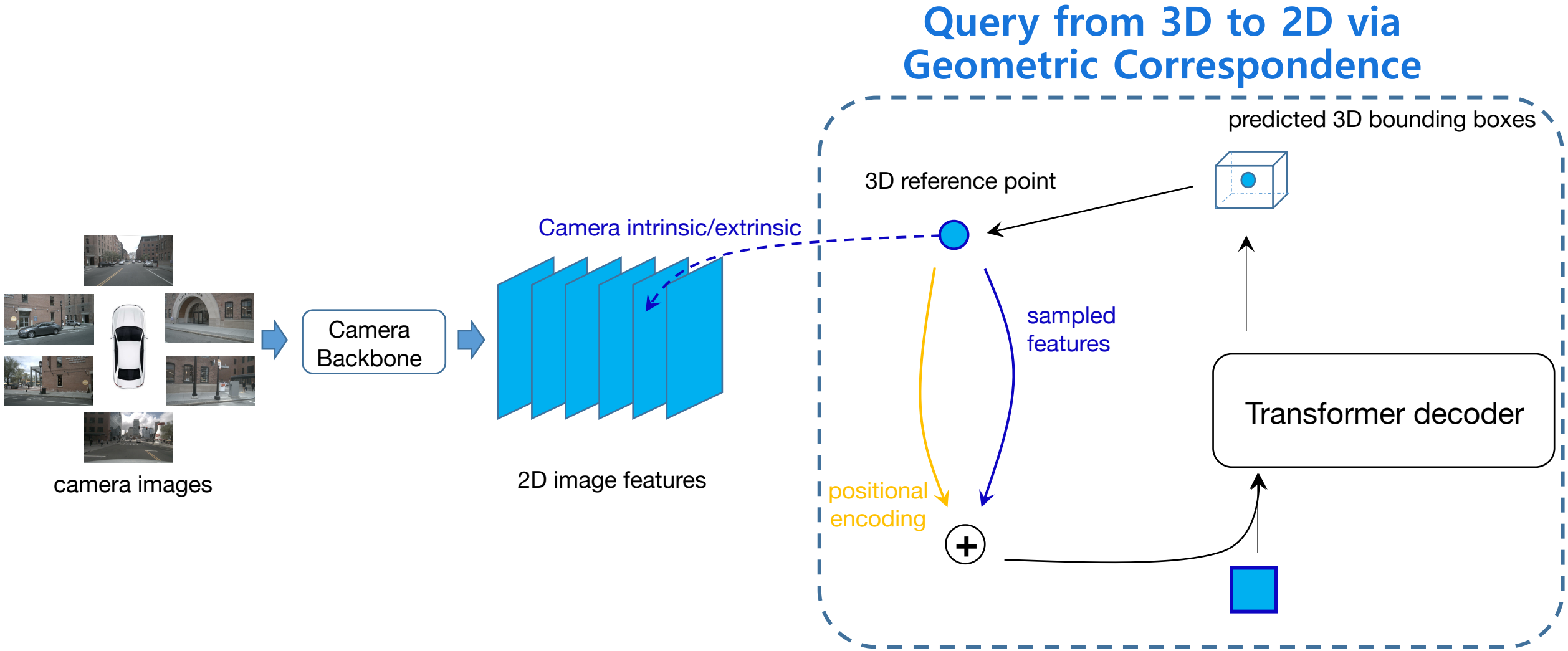
- ✓ Detection in the 3D space, even though the observations are 2D.
- ✓ Does not reconstruct 3D space explicitly.
- ✓ Avoids post-processing like NMS.

Model Architecture

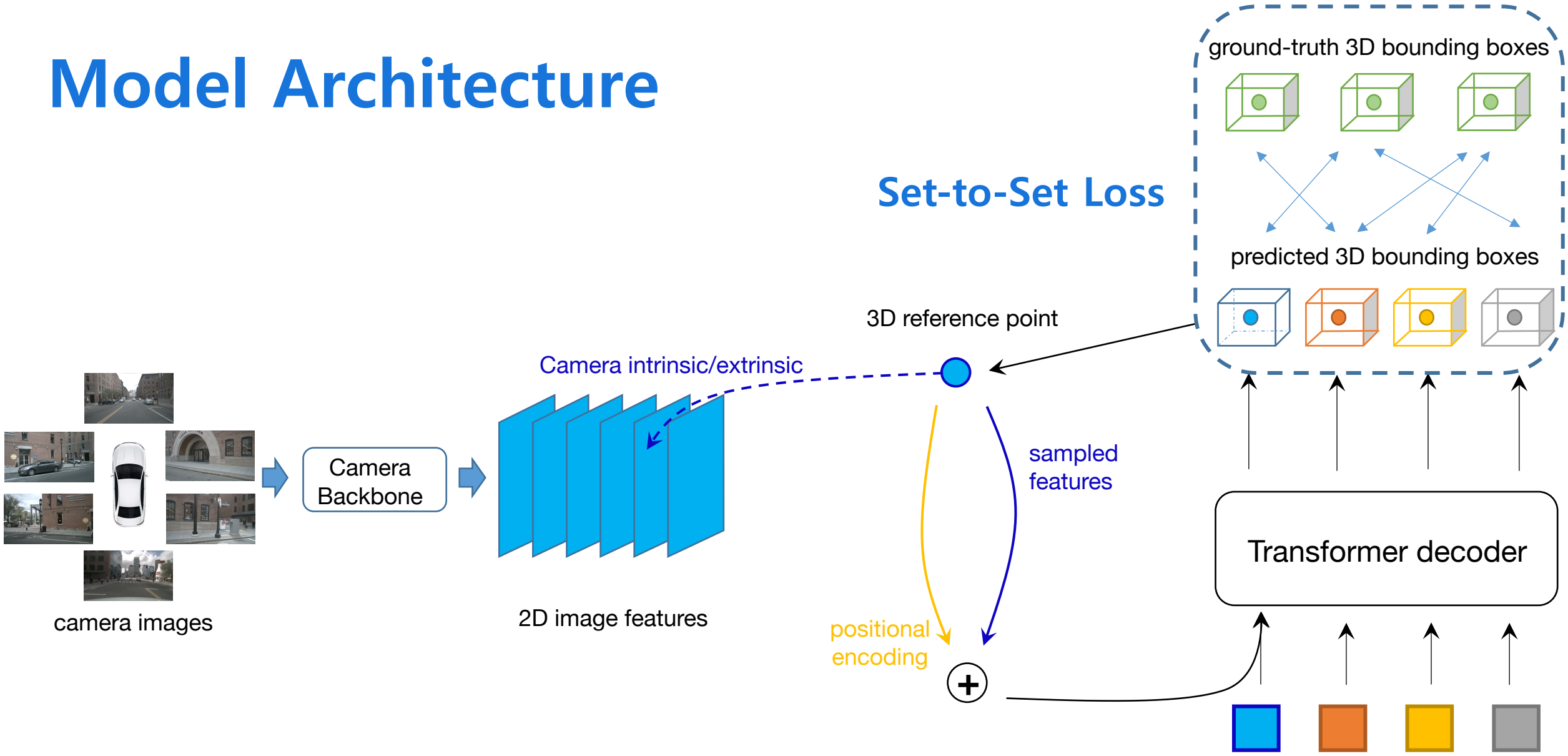
Feature Extraction Backbone



Model Architecture



Model Architecture



FUTR3D
Sensor Fusion Transformer 3D

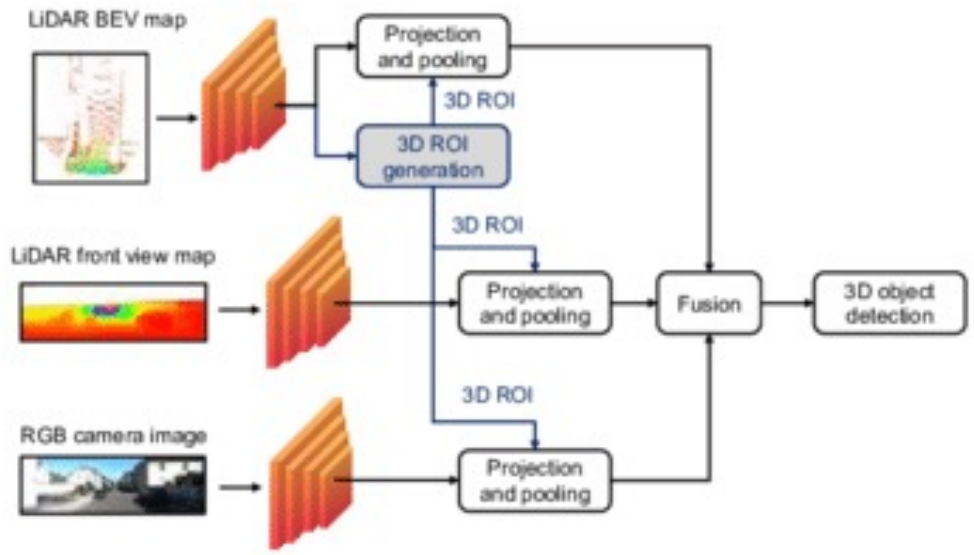
FUTR3D: A Unified Sensor Fusion Framework for 3D Detection

https://tsinghua-mars-lab.github.io/futr3d/
https://arxiv.org/abs/2203.10642

Xuanyao Chen, Tianyuan Zhang,
Yue Wang, Yilun Wang, Hang Zhao

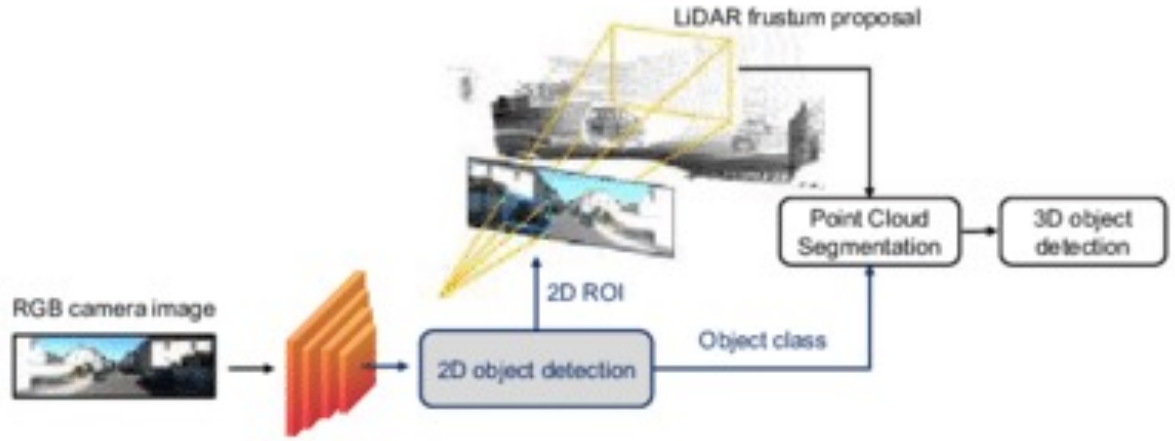
Related Work on LiDAR + Camera Fusion

- Object proposal from LiDAR
- Refinement with Camera



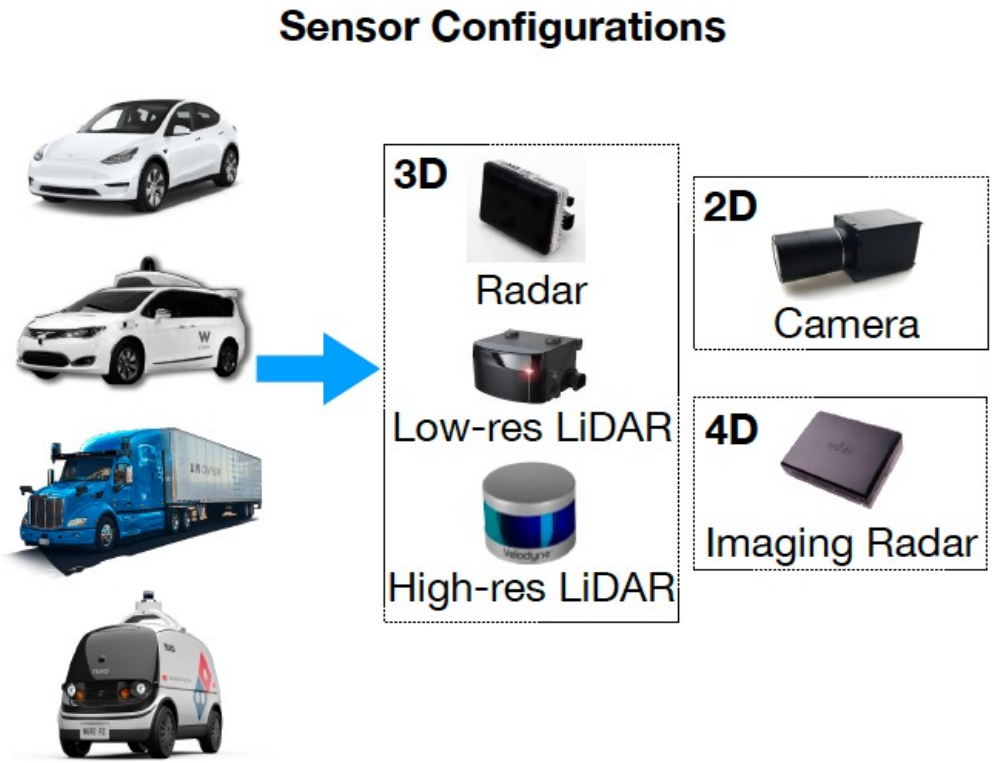
Multi-View 3D Object Detection Network for Autonomous Driving, Chen et al., CVPR 2017

- Object proposal from Camera
- Refinement with LiDAR



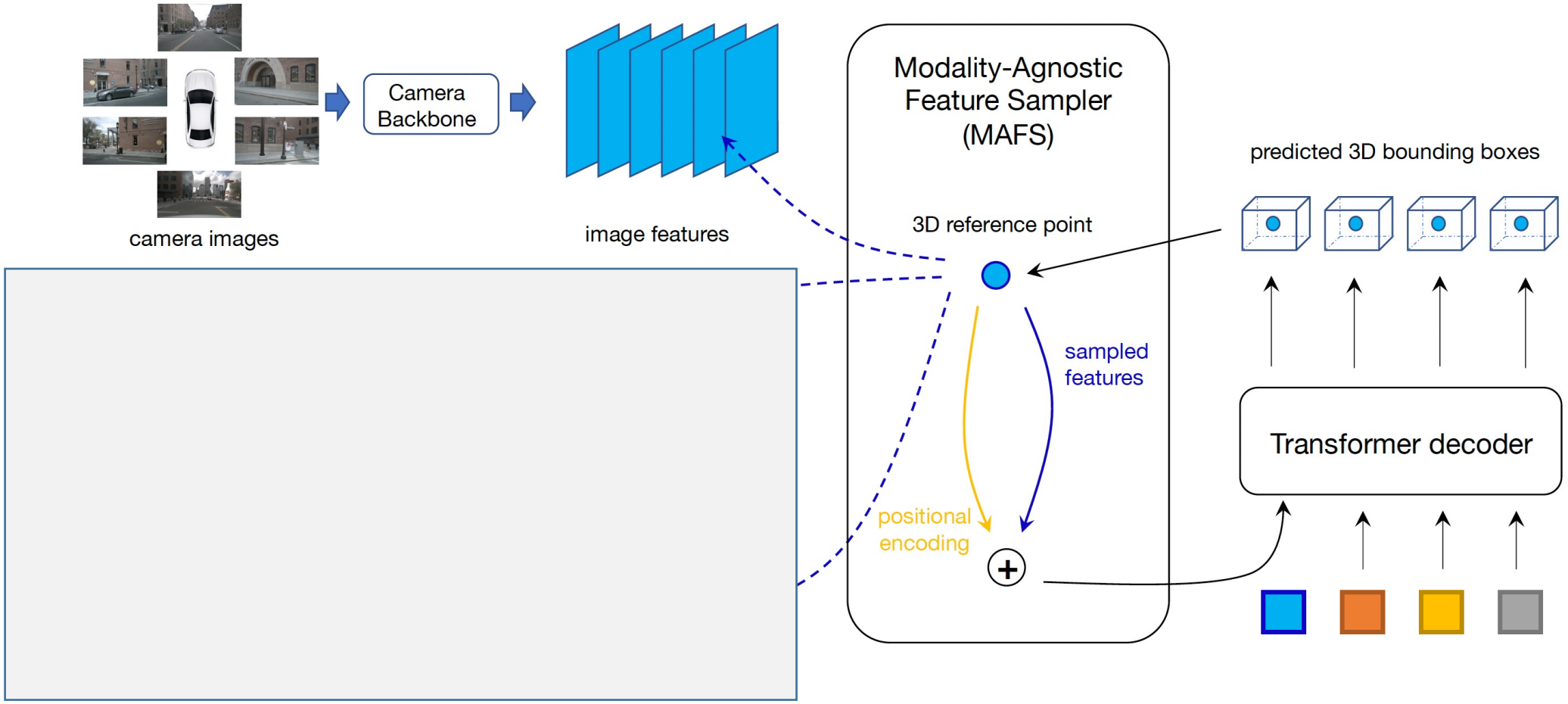
Frustum PointNets for 3D Object Detection from RGB-D Data, Qi et al., CVPR 2018

FUTR3D

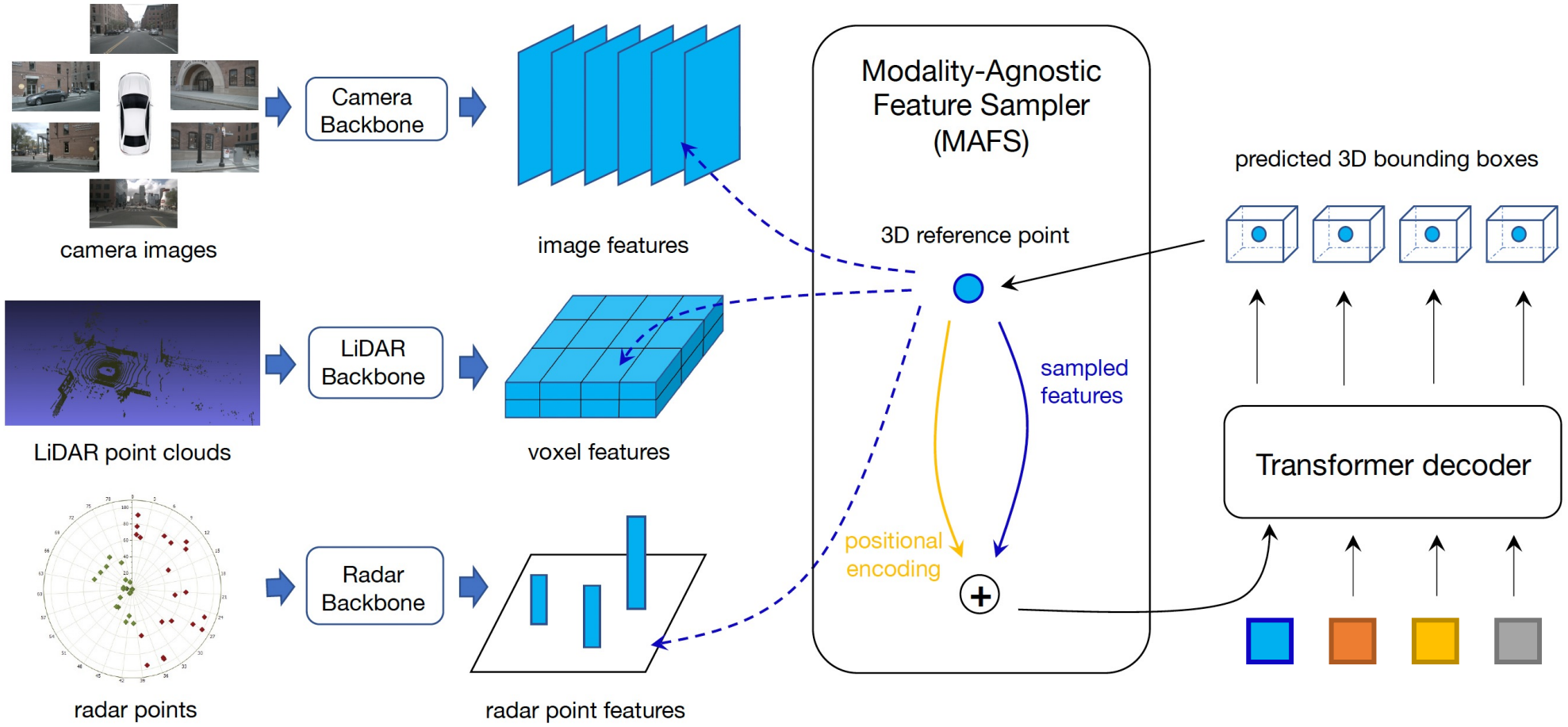


- Is there a simple, unified framework agnostic to sensor types?
- 3D Object Query

Model Architecture – DETR3D

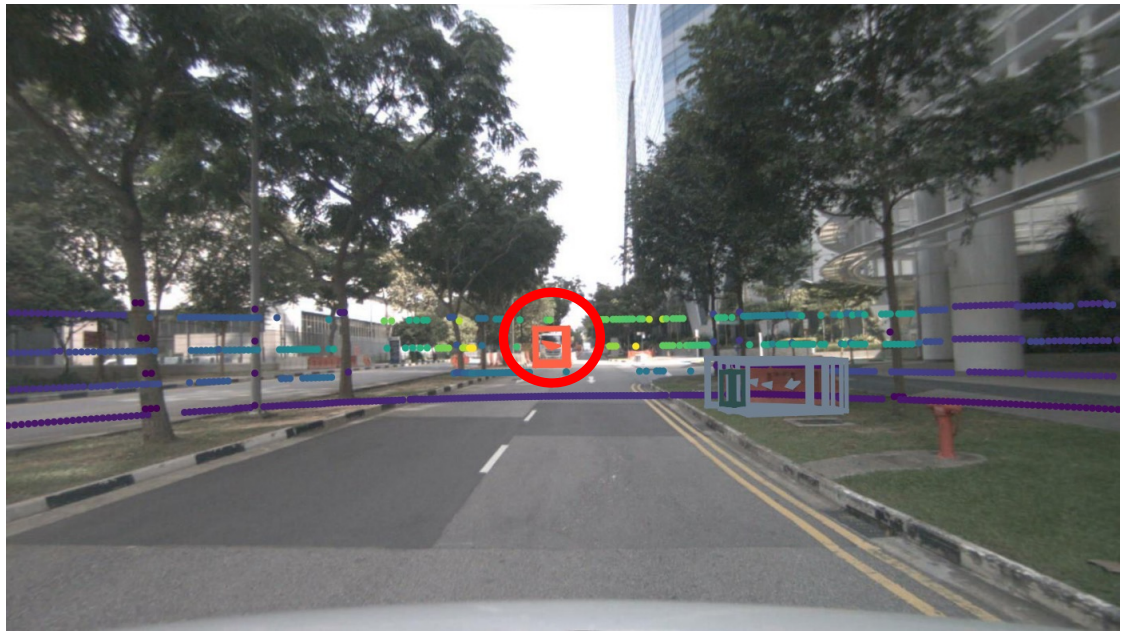


Model Architecture – FUTR3D

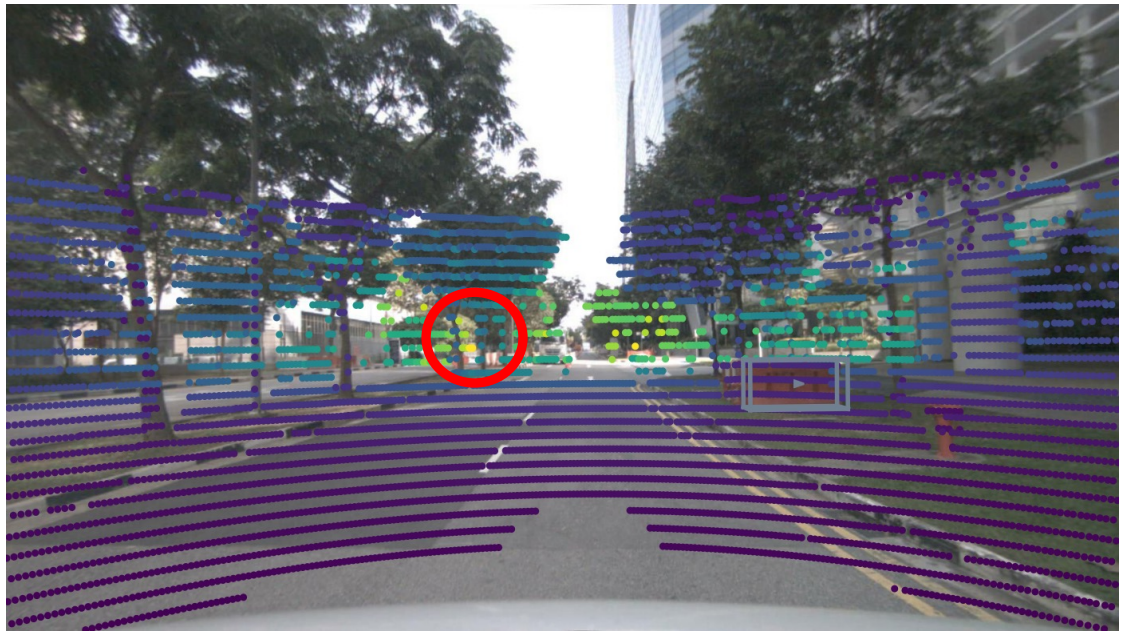


Qualitative Results

Camera helps in detecting distant objects



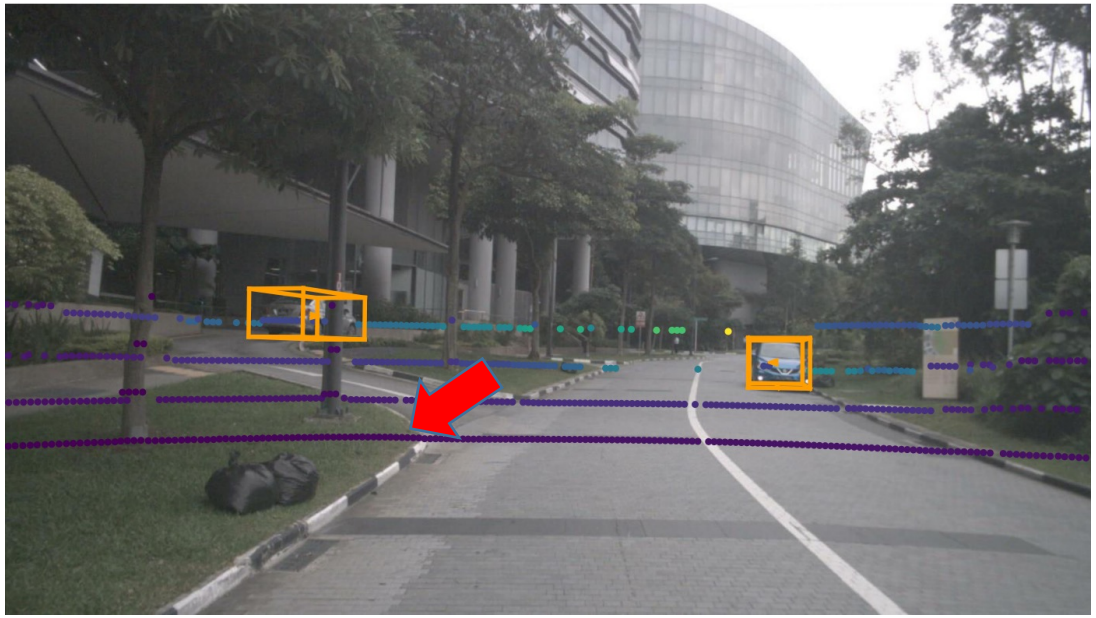
4-beam LiDAR + Cameras



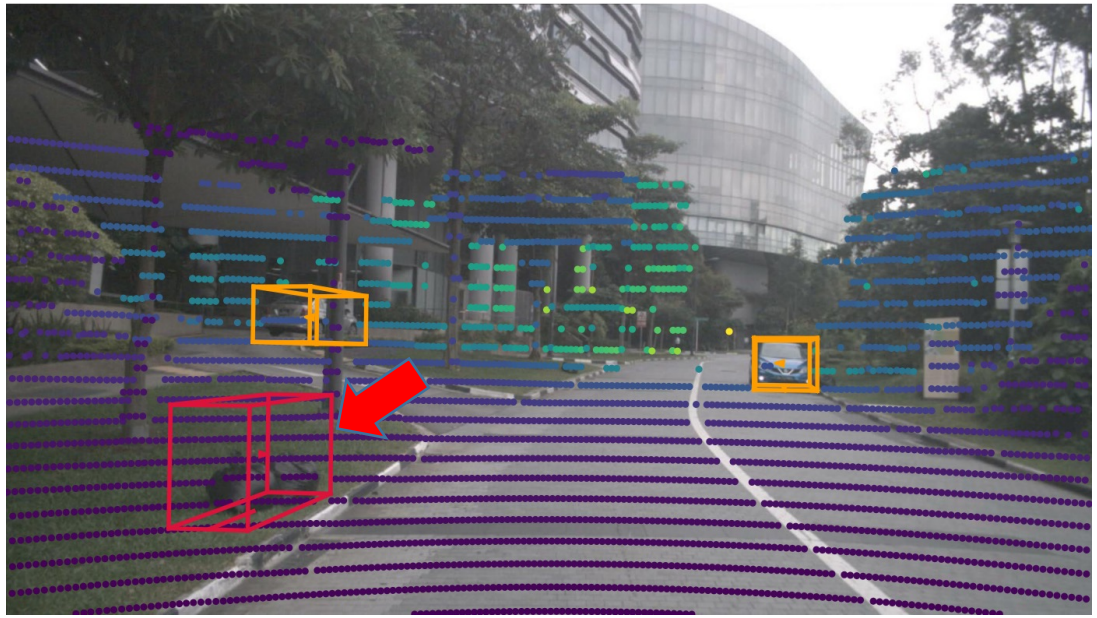
32-beam LiDAR

Qualitative Results

Without visual input, garbage bags detected as bicycles



4-beam LiDAR + Cameras



32-beam LiDAR

FUTR3D Enables Low-cost Solutions

4-beam LiDAR + Camera ~= 32-beam LiDAR

(a) **Camera**

methods	NDS ↑	mAP ↑
CenterNet* [40]	32.8	30.6
FCOS3D* [28]	41.5	34.3
PGD* [29]	42.8	36.9
FUTR3D	42.5	34.6

(b) **32-beam LiDAR**

methods	NDS ↑	mAP ↑
CenterPoint-Pillar ‡	59.7	49.1
CenterPoint-Voxel ‡	64.5	56.6
FUTR3D-Pillar	60.4	51.3
FUTR3D-Voxel	65.5	59.3

(c) **Low-resolution LiDAR**

methods	NDS ↑	mAP ↑
4-beam CenterPoint	53.6	38.5
4-beam FUTR3D	54.8	42.1
1-beam CenterPoint	36.9	14.5
1-beam FUTR3D	37.9	16.4

	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
32-beam LiDAR + Camera							
PointPainting [27]	67.8	62.8	29.6	25.5	32.5	29.3	19.0
FUTR3D	**68.0**	**64.2**	35.0	25.9	30.4	30.5	19.3
4-beam LiDAR + Camera							
PointPainting [27]	59.4	50.0	34.4	26.6	44.3	32.6	18.4
FUTR3D	61.5	54.9	43.4	26.5	35.2	36.8	18.0
FUTR3D + DDAD pretrain [20]	**62.9**	**56.8**	41.1	26.3	33.6	35.0	18.6
1-beam LiDAR + Camera							
PointPainting [27]	41.0	22.0	50.3	28.0	62.3	40.0	19.5
FUTR3D	**50.0**	**41.3**	61.0	26.9	40.8	59.0	18.7

MUTR3D
Multi-object Tracking Transformer 3D

MUTR3D: A Multi-camera Tracking Framework via 3D-to-2D Queries

CVPR WAD 2022

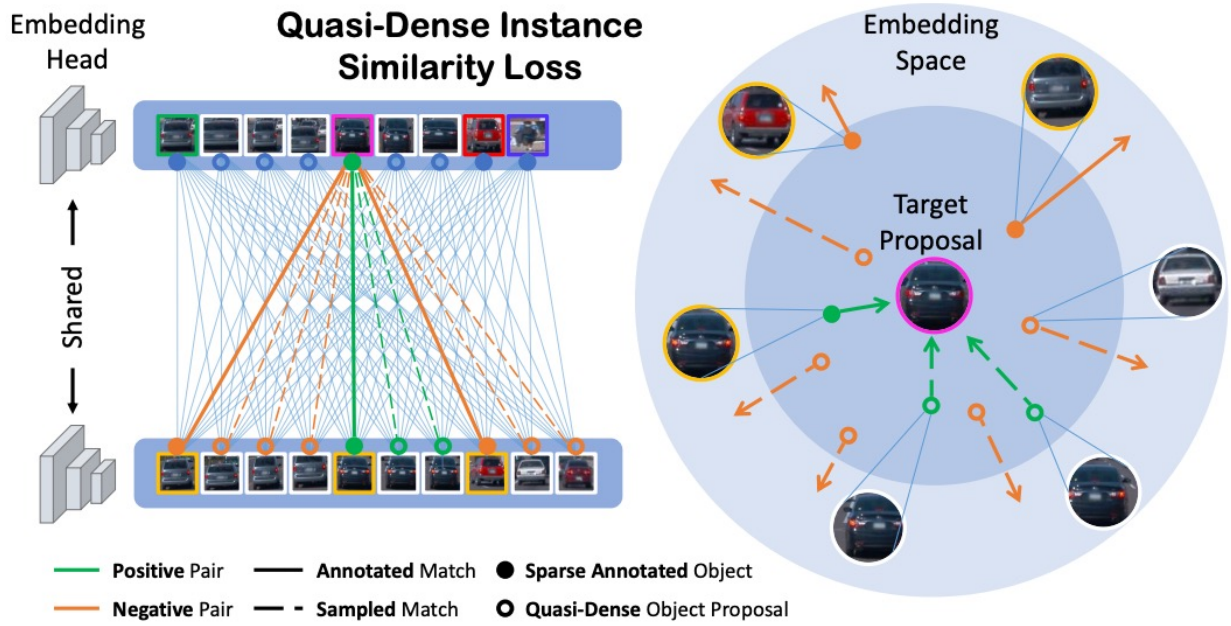
https://tsinghua-mars-lab.github.io/mutr3d/

https://arxiv.org/abs/2205.00613

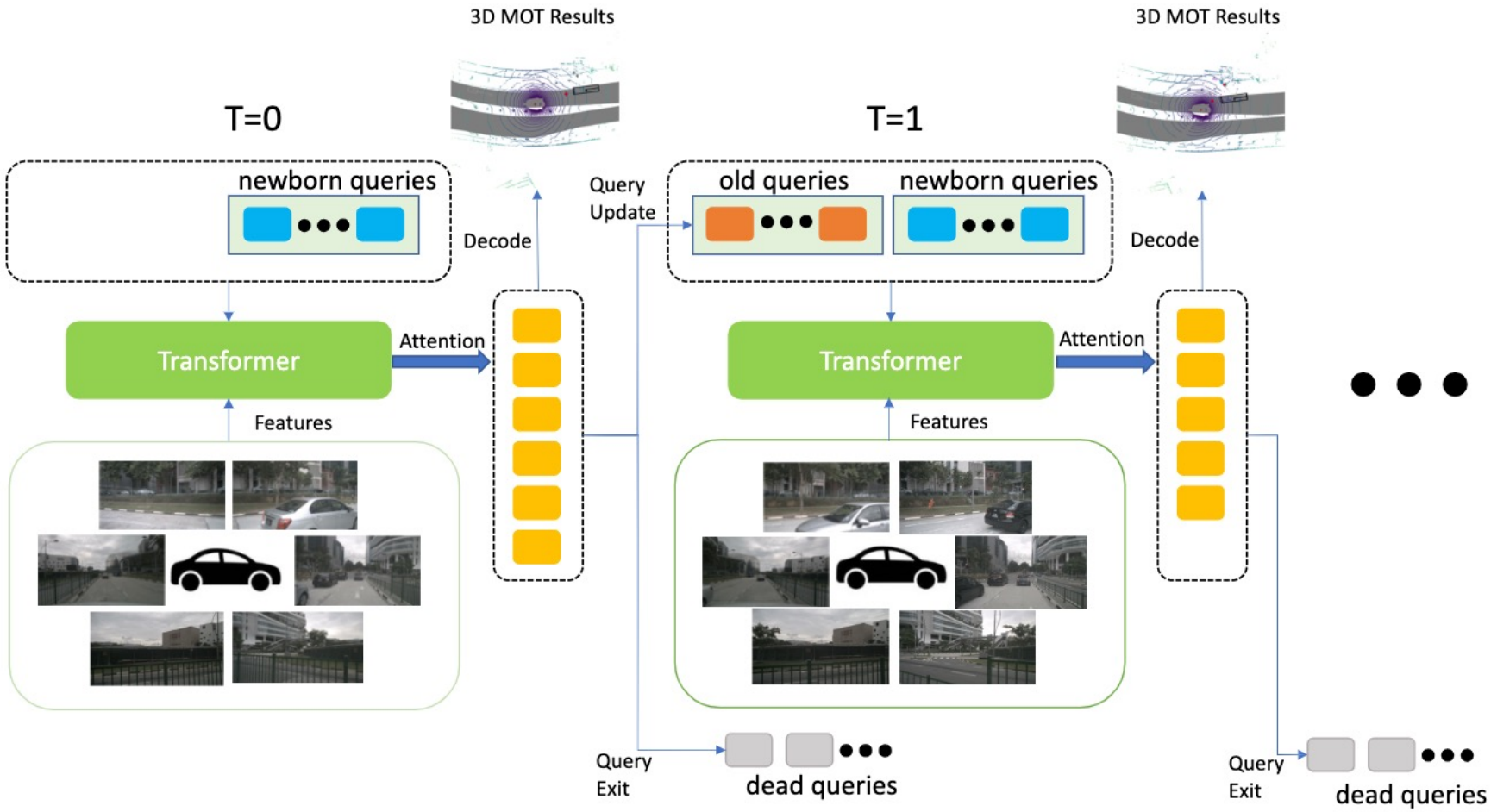
Tianyuan Zhang, Xuanyao Chen,
Yue Wang, Yilun Wang, Hang Zhao

Related Work on Detect-and-Track

- Traditional: Detection + Kalman Filter
- Learning to associate: Monocular Quasi-Dense 3D Object Tracking, Hu et al.



MUTR3D: Implicit Association with Queries



Quantitative Results

Table 1. **Comparison with state-of-the-art methods on nuScenes dataset.** For public camera-based 3D tracking, our algorithm achieves state-of-the-art results, outperforming QD3DT [11] by 0.052 in AMOTA on validation set and 0.053 on test split.

	Modality	AMOTA ↑	AMOTP ↓	RECALL ↑	MOTA ↑	IDS ↓	#params
Validation Split							
CenterPoint [43]	LiDAR	0.665	0.567	69.9%	0.562	562	9M
SimpleTrack [21]	LiDAR	0.687	0.573	72.5%	0.592	519	9M
DEFT [8]	Camera	0.201	N/A	N/A	0.171	N/A	22M
QD3DT [11]	Camera	0.242	1.518	39.9%	0.218	5646	91M
Ours	Camera	**0.294**	**1.498**	**42.7%**	**0.267**	**3822**	56M
Test Split							
CenterTrack [47]	Camera	0.046	1.543	23.3%	0.043	**3807**	20M
DEFT [8]	Camera	0.177	1.564	33.8%	0.156	6901	22M
QD3DT [11]	Camera	0.217	1.550	37.5%	0.198	6856	91M
Ours	Camera	**0.270**	1.494	**41.1%**	**0.245**	6018	56M

HDMapNet: An Online HD Map Construction and Evaluation Framework

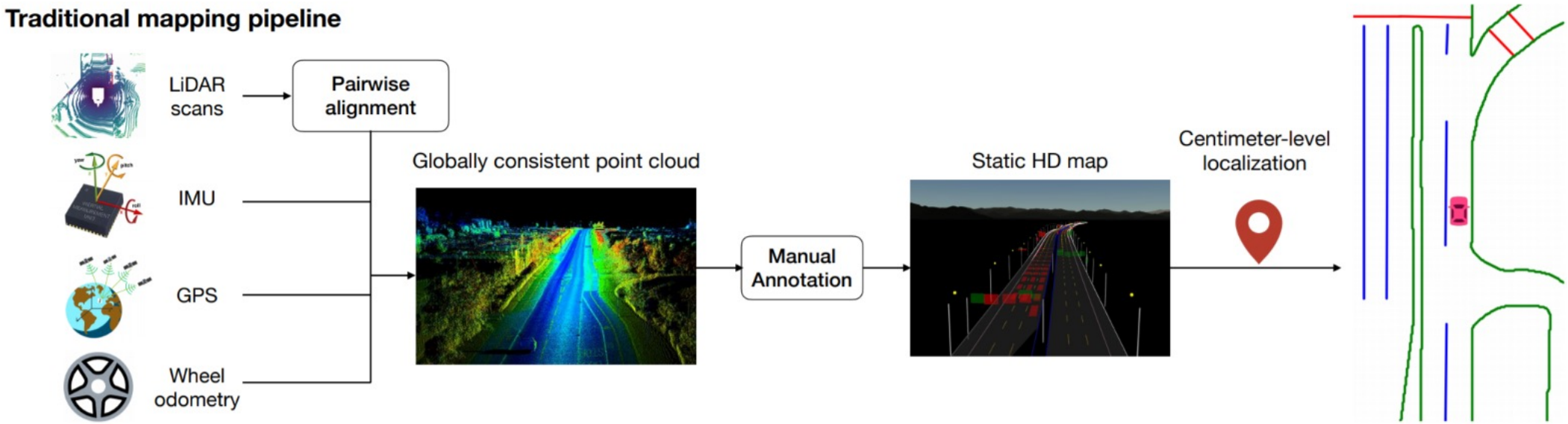
ICRA 2022, CVPRW 2021

https://tsinghua-mars-lab.github.io/HDMapNet/

https://arxiv.org/abs/2107.06307

Qi Li, Yue Wang, Yilun Wang, Hang Zhao

Traditional Mapping and Localization



☹ Involves a complex pipeline

☹ Costly and labor intensive

☹ Needs timely update

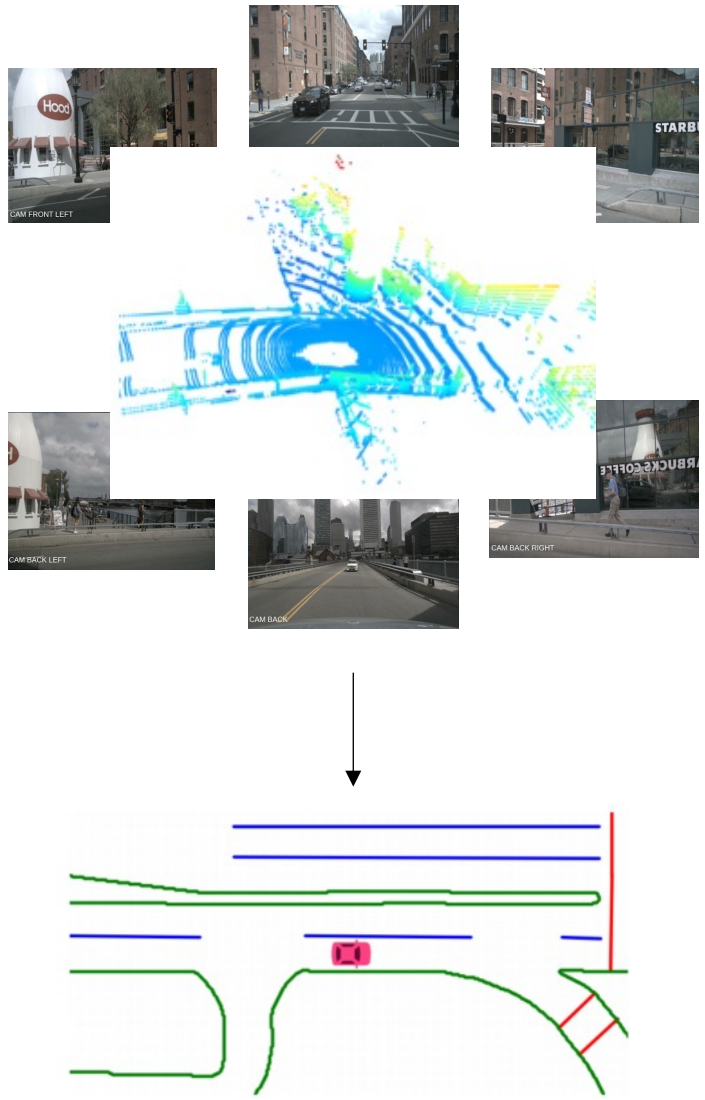
Inputs and Outputs

INPUT

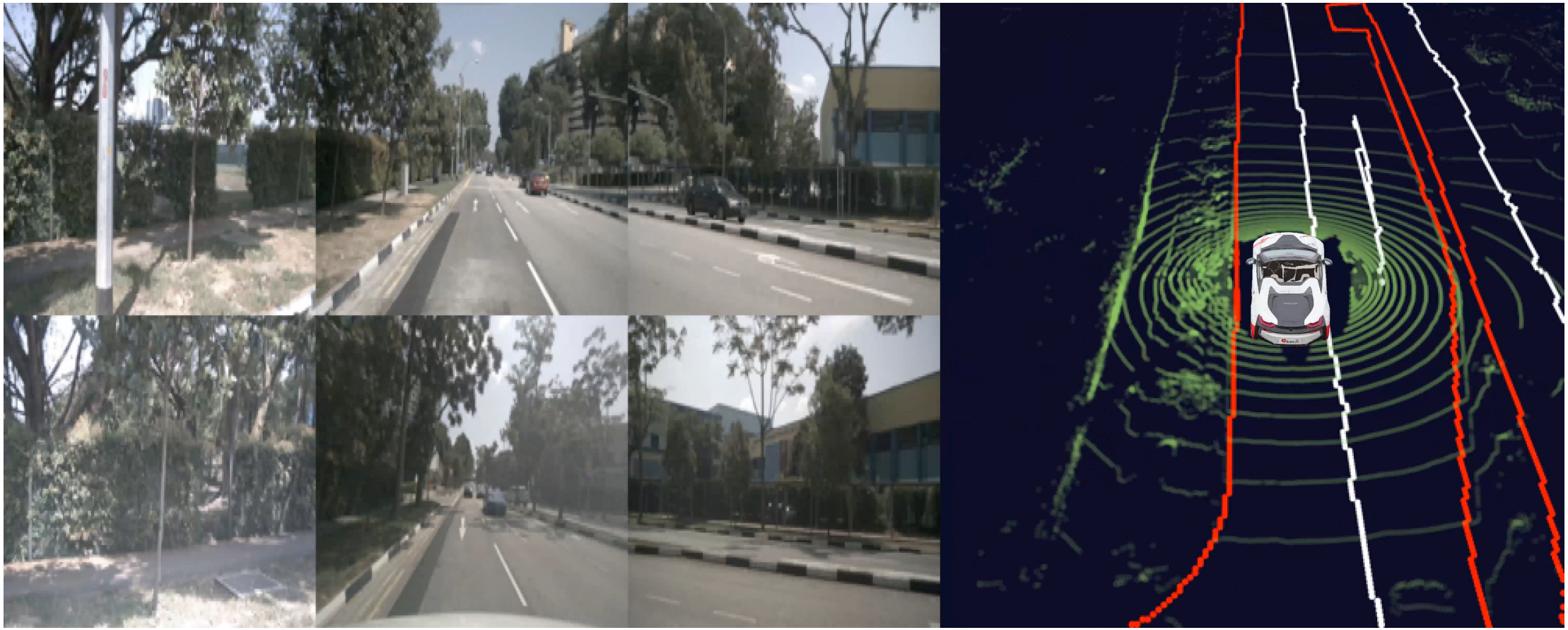
images	LiDAR	both
HDMapNet(Surr)	HDMapNet(LiDAR)	HDMapNet(Fusion)

OUTPUT

Vectorized BEV
HD Semantic Map

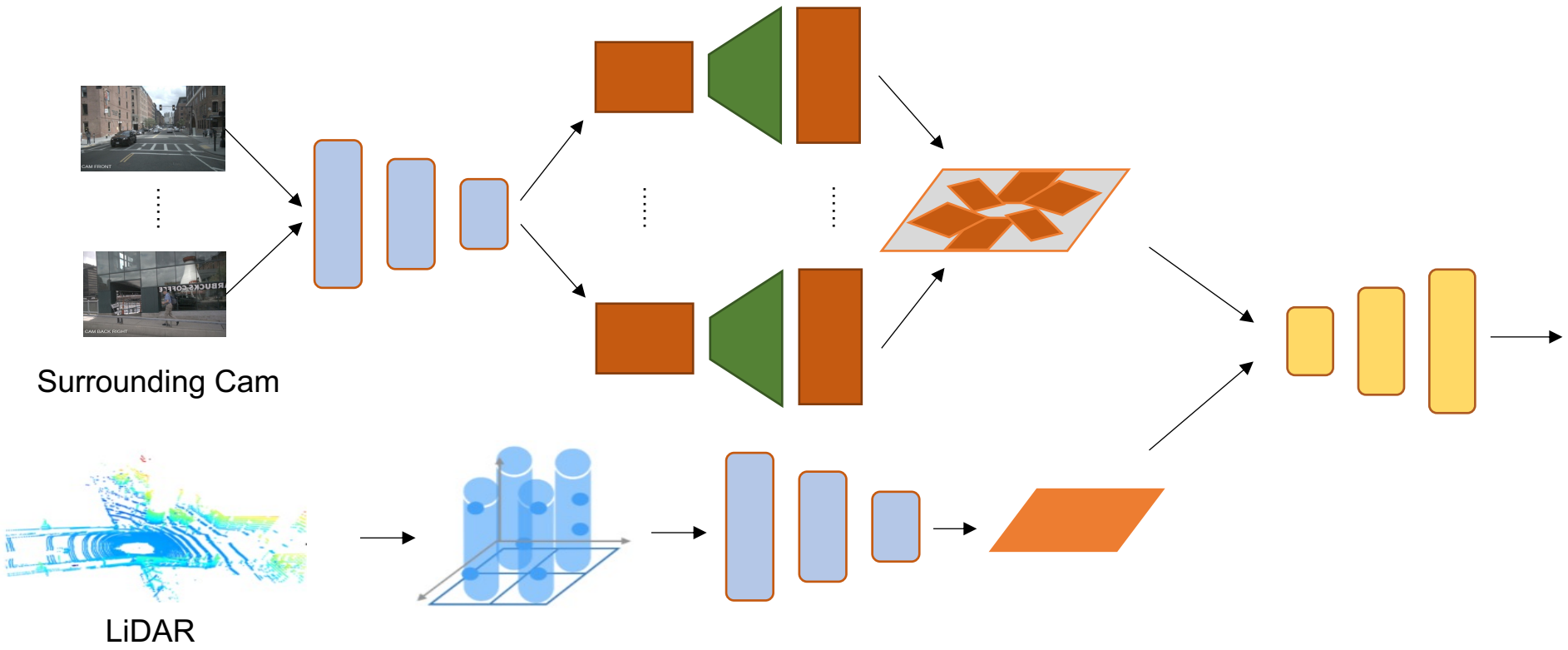


Results

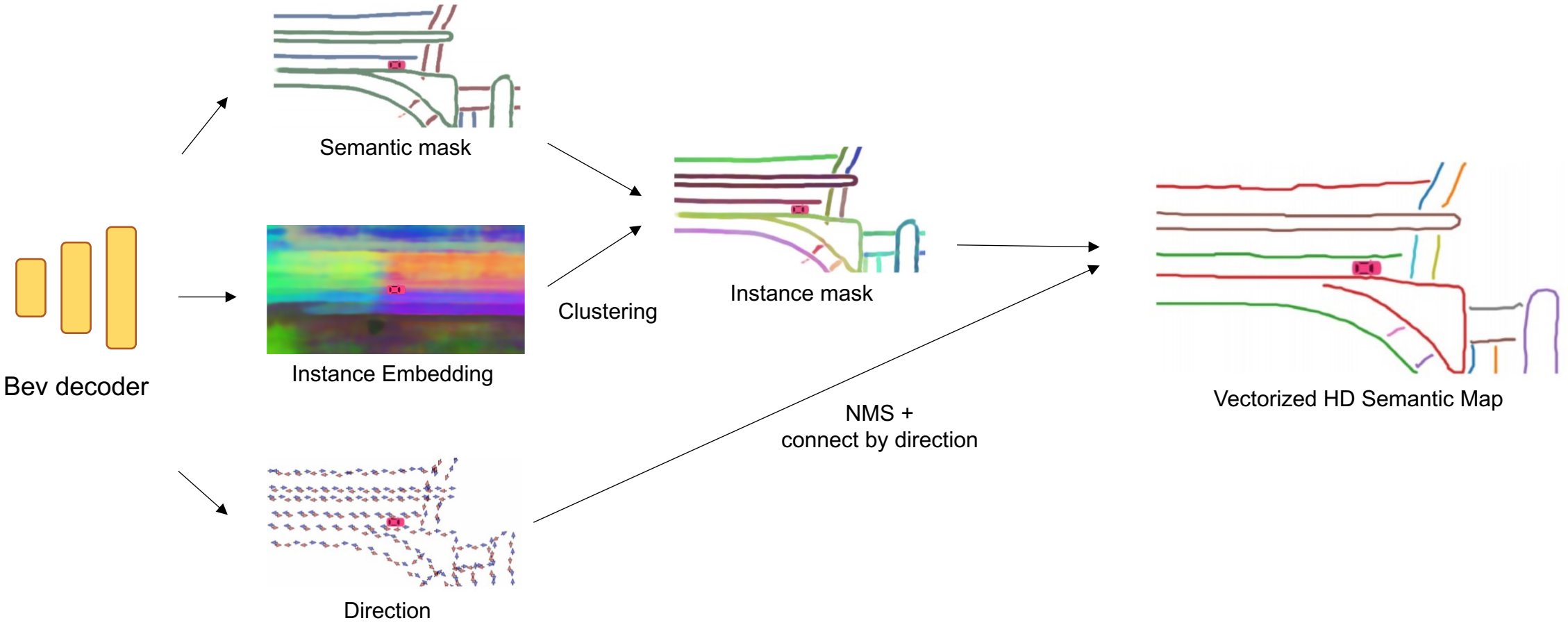


Model Architecture

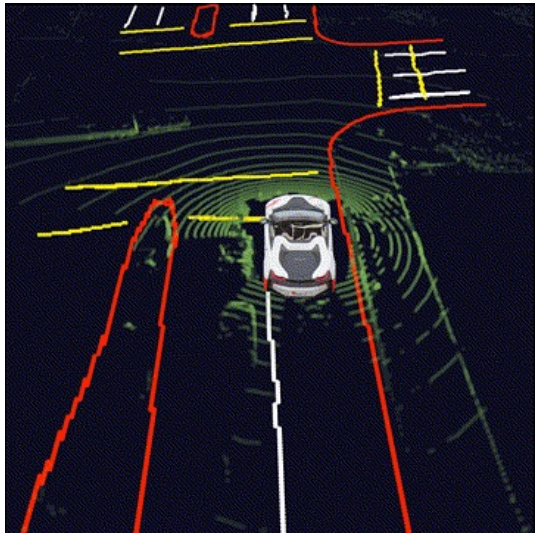
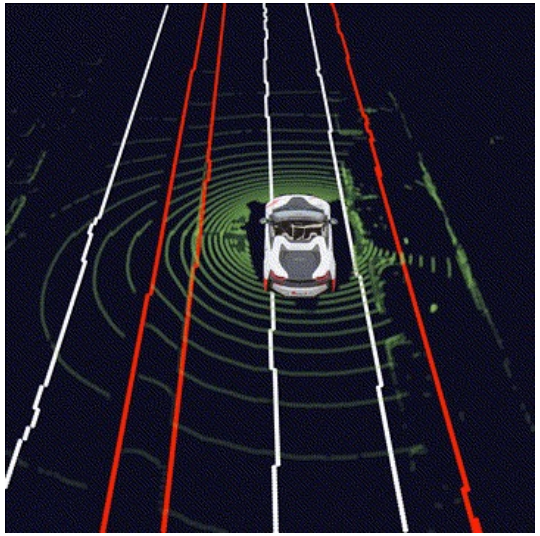
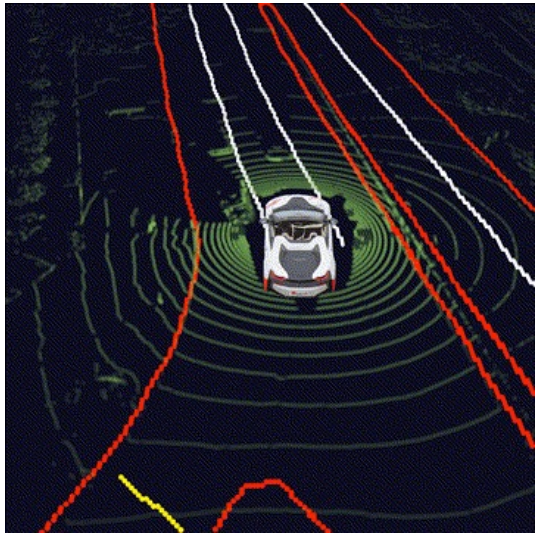
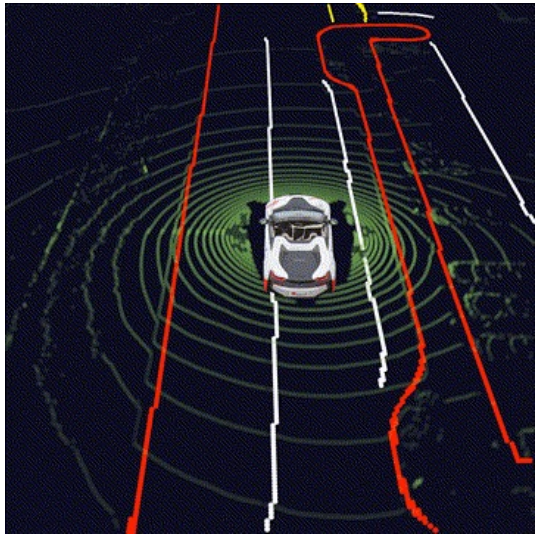
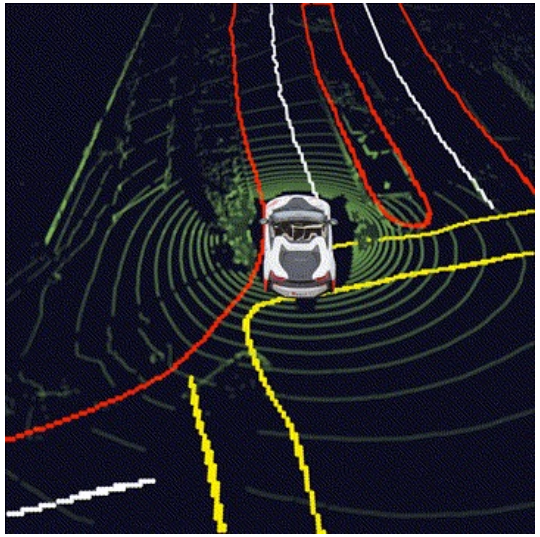
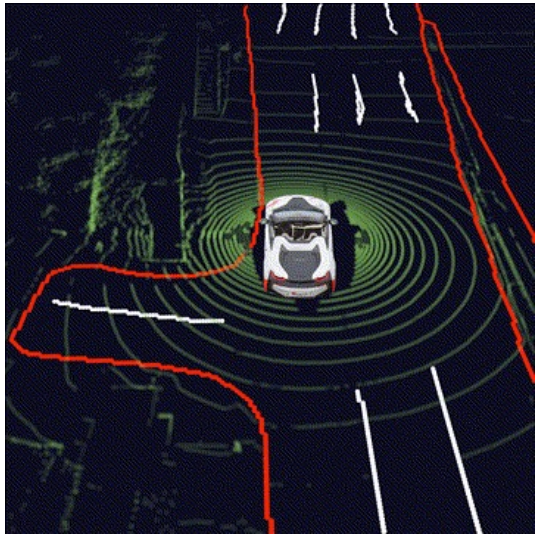
Projecting both cameras and LiDAR features to BEV space, and then decode.



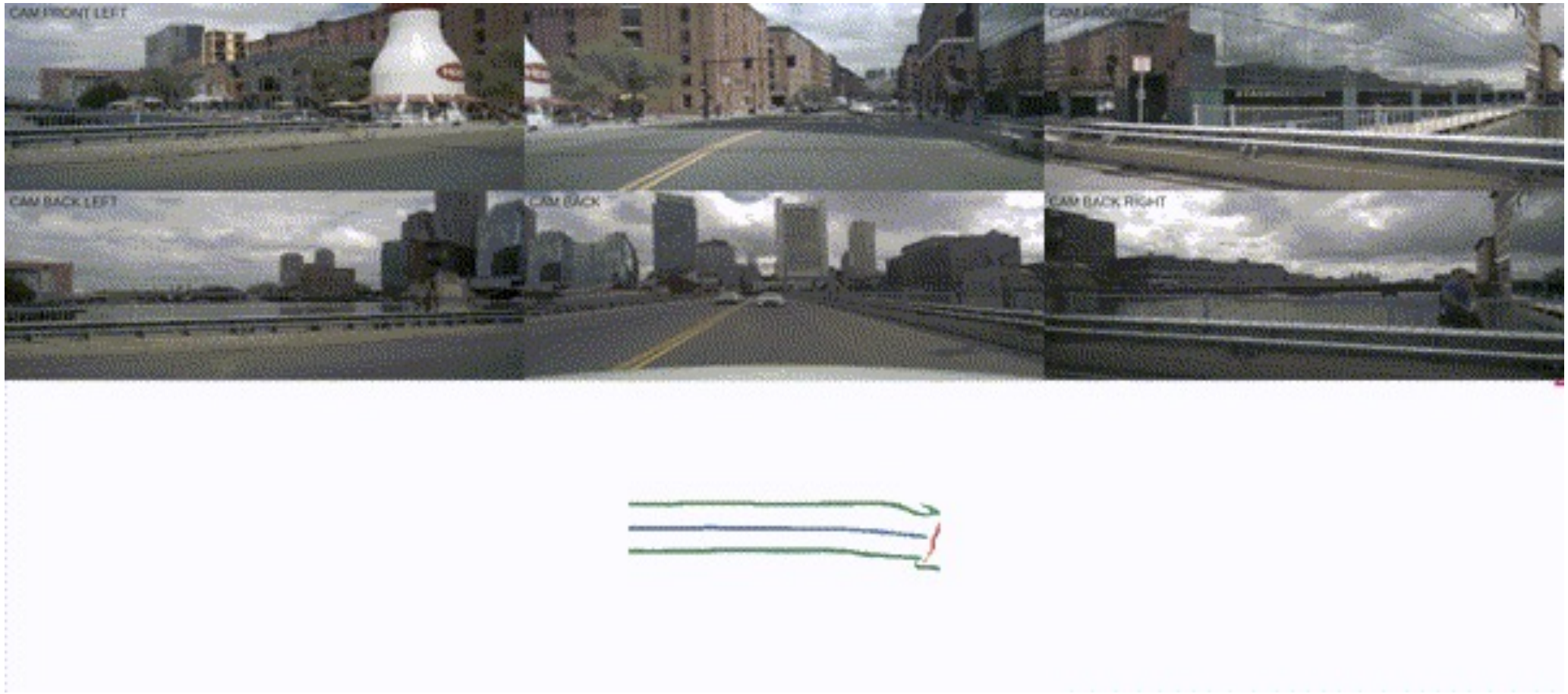
Vectorization



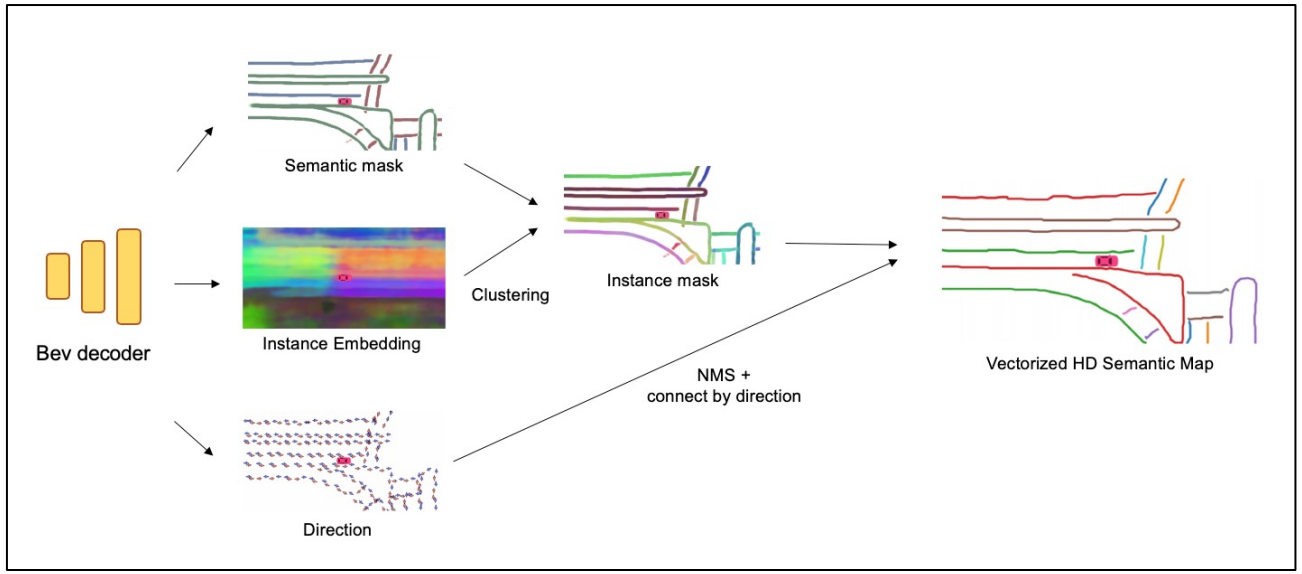
Qualitative Results

Temporal Map Aggregation



HDMapNet



- BEV segmentation + post-processing
- Do we have an end-to-end solution?
- **HDMapNet 2.0**

VectorMapNet: End-to-end Vectorized HD Map Learning

https://tsinghua-mars-lab.github.io/vectormapnet/

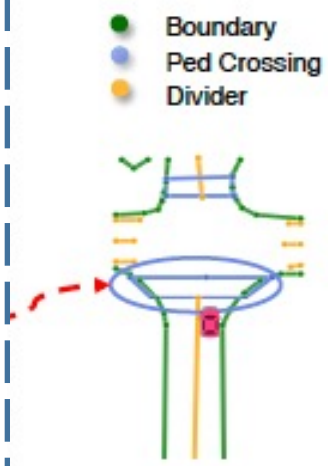
https://arxiv.org/abs/2206.08920

Yicheng Liu, Yue Wang, Yilun Wang, Hang Zhao

Model Architecture

Onboard Sensor Data

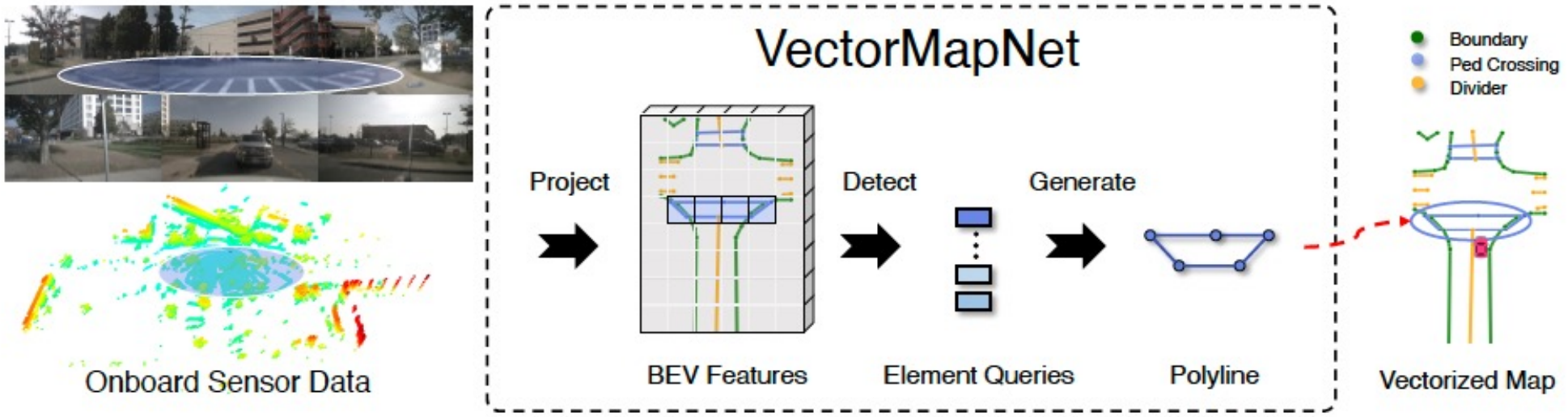
1. BEV Projection
2. Element Detection/Regression



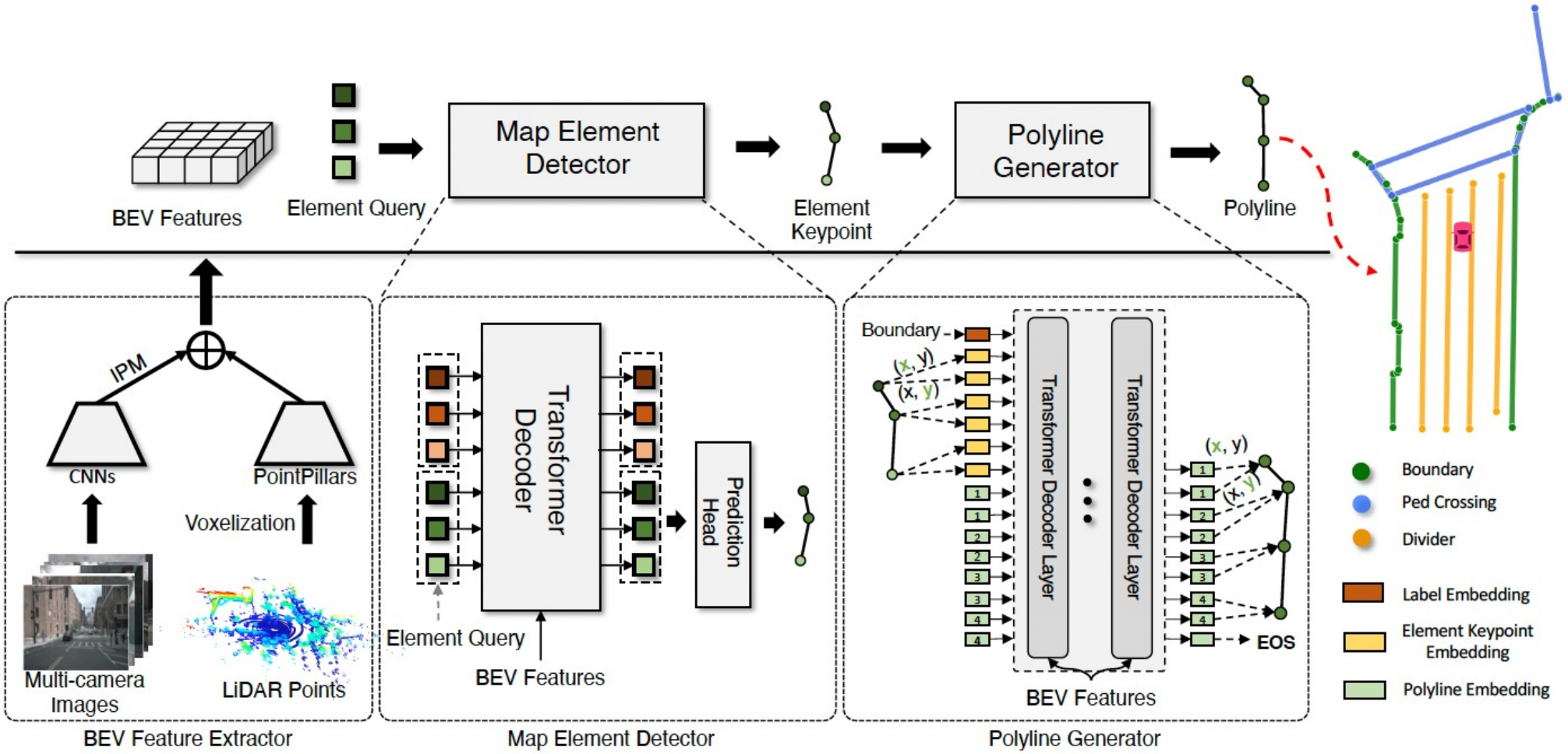
Vectorized Map

Model Architecture

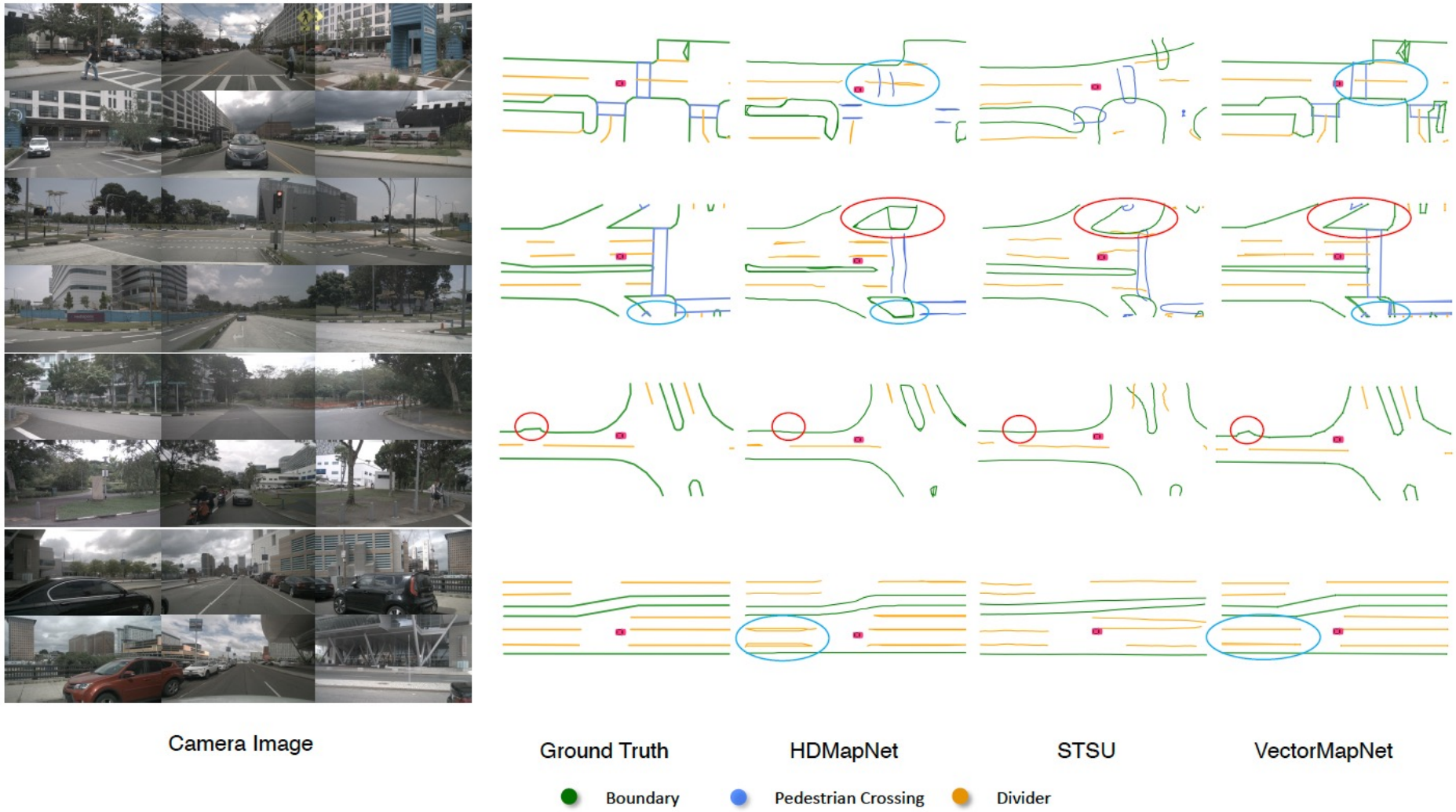
- Pose it as a detection problem
- Use *polyline* as the primitive for traffic elements
- Solve it in a DETR-way



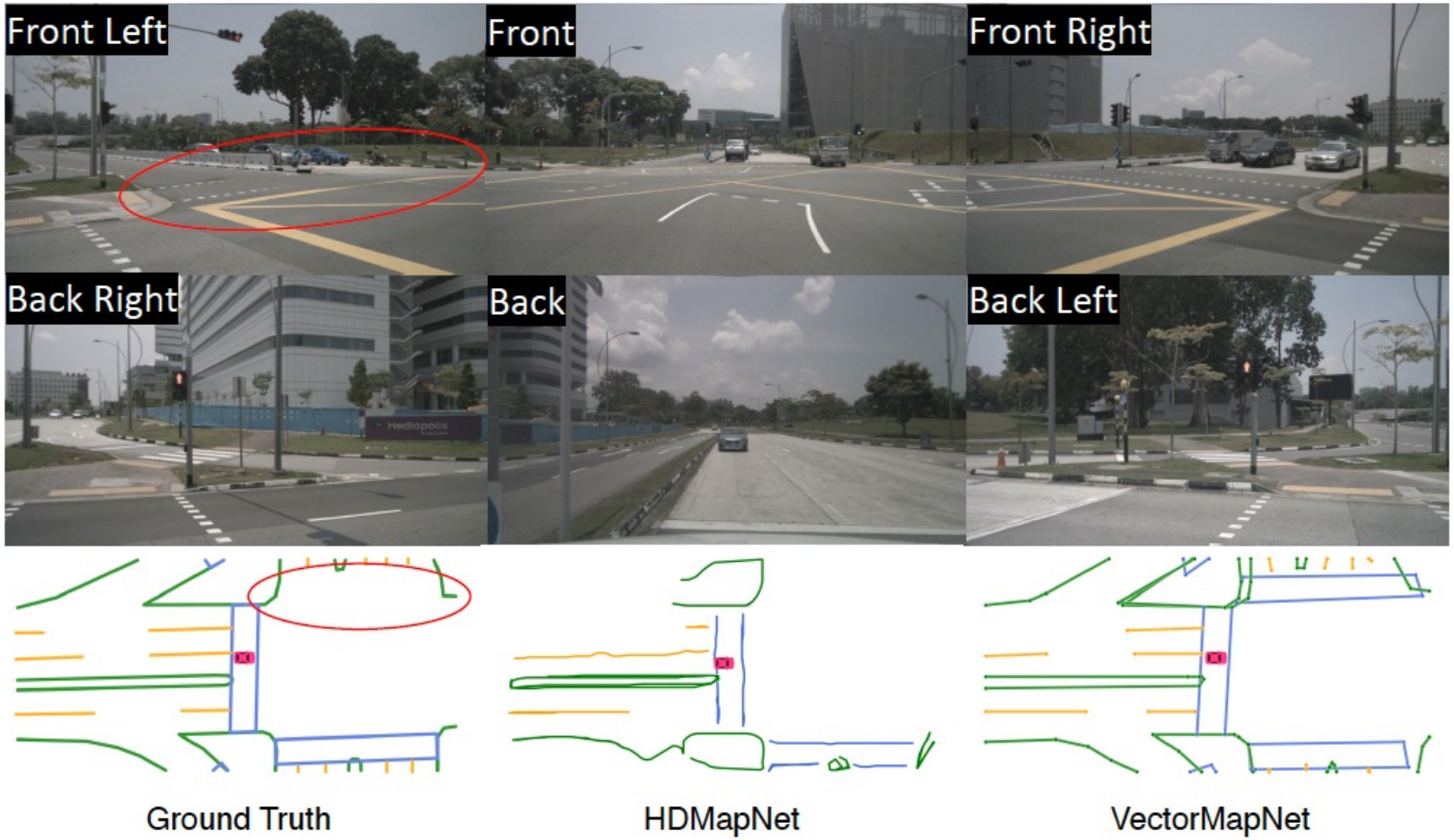
Model Architecture



Qualitative Results

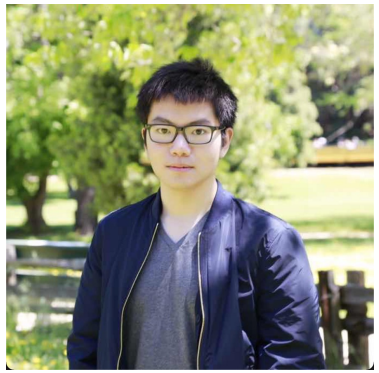


Qualitative Results



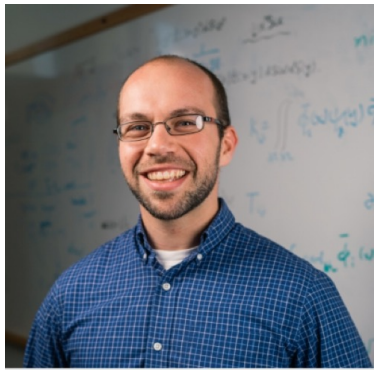
Acknowledgement

Yue Wang
MIT



Yilun Wang
Li Auto

Vitor Guizilini
Toyota Research



Justin Solomon
MIT

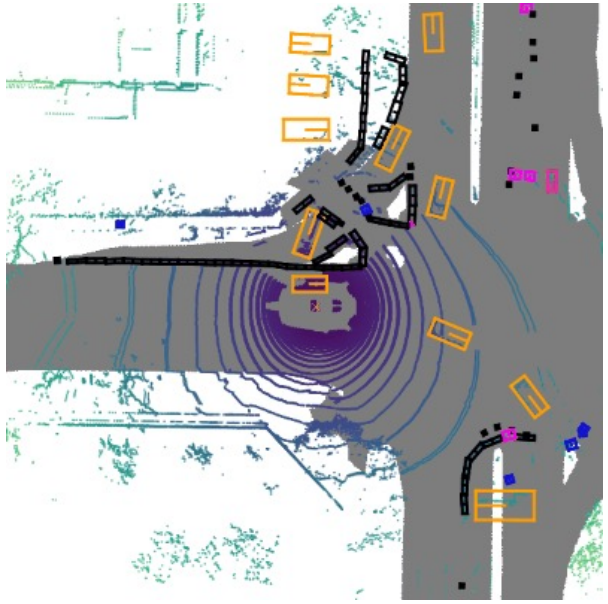
Qi Li
Tsinghua University

Tianyuan Zhang
CMU

Xuanyao Chen
Fudan University

Yicheng Liu

Acknowledgement



MMDetection3D

OpenMMLab's next-generation platform for general 3D object detection

- Different modalities/scenarios/tasks
- Natural integration with 2D detection
- High efficiency

Popular projects build upon MMDet3D

- FCOS3D, DETR3D, BEVFormer…
- TransFusion, BEVFusion…
- LiDAR R-CNN, SST, FCOS-LiDAR…

MARS Lab

THE END THANKS