Enhanced 3D Perception and 3D Reasoning for end-to-end Autonomous Driving (AD)

Zhen Li
Assistant Professor
The Chinese University of Hong Kong, Shenzhen (CUHKSZ)

May 7, 2024 at Chong Qing

End-to-end Autonomous Driving (L2)

Tesla FSD 12.3.6

Huawei Wenjie Zhijia

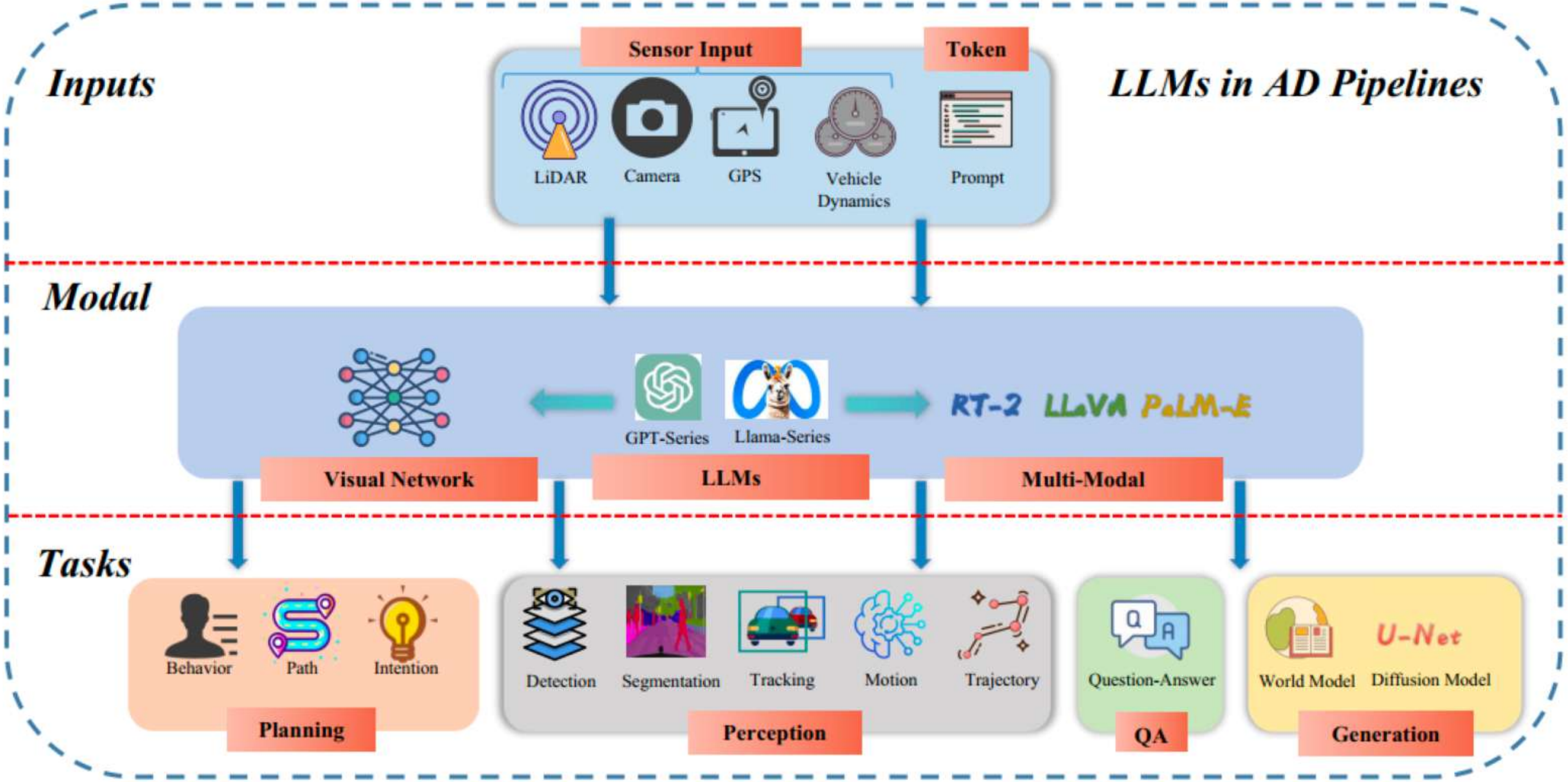
Hardware for Telsa and Wenjie

Tesla- Vision Centric

Wenjie with LiDAR

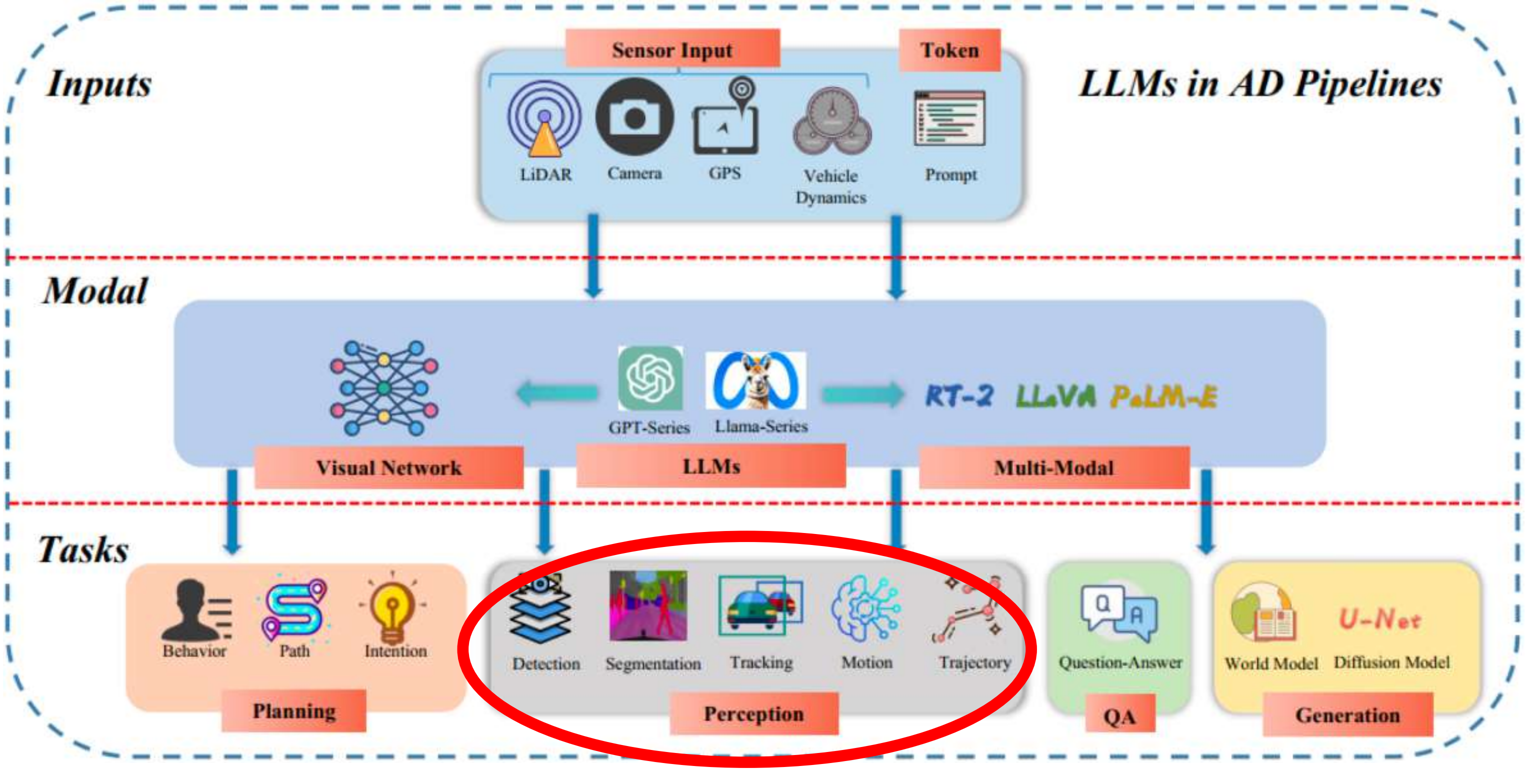
https://zhuanlan.zhihu.com/p/646724439

Large Models for Autonomous Driving



Awesome-LLM-for-Autonomous-Driving-Resources
https://github.com/Thinklab-SJTU/Awesome-LLM4AD/tree/main

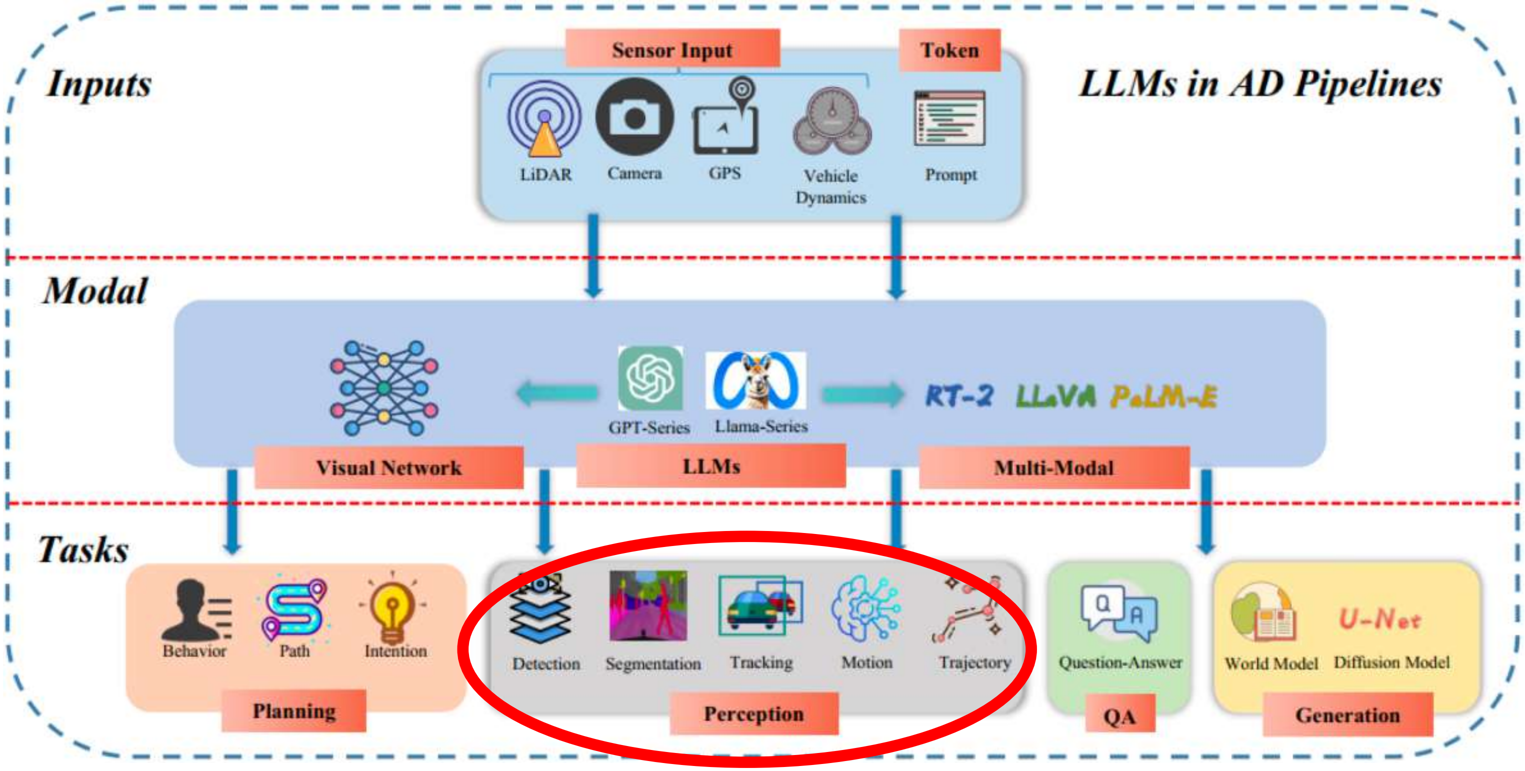
Outline



Enhanced 3D Perception for AD

- Monocular (front/ego view)/multimodality 3D Lane Detection
- 3D/4D Occupancy (world model)
- Semantic Segmentation and Semantic Completion

Outline



Enhanced 3D Perception for AD

- ➢ **Monocular (front/ego view)**/Multimodality 3D Lane Detection
- ➢ 3D/4D Occupancy (world model)
- ➢ Semantic Segmentation and Semantic Completion

LATR: 3D Lane Detection from Monocular Images with Transformer

Yueru Luo[1,2] Chaoda Zheng[1,2] Xu Yan[1,2] Tang Kun[3]
Chao Zheng[3] Shuguang Cui[2,1] **Zhen Li**[2,1,✉]

[1]FNii, CUHK-Shenzhen [2]SSE, CUHK-Shenzhen
[3]Tencent Map, T Lab

ICCV 2023 Oral

ICCV23
PARIS

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

FNii
未来智联网络研究院

腾讯地图

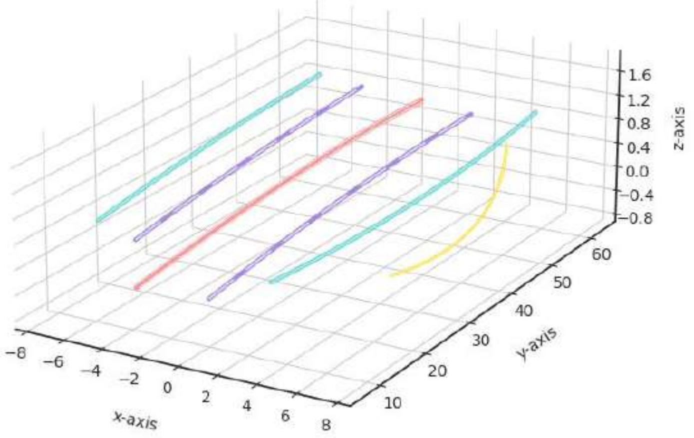
Monocular 3D Lane Detection

Goal: localize lane boundaries in 3D space using a single monocular image.

Challenges:

- slenderness and elongation of lanes.
- depth absence in monocular images.

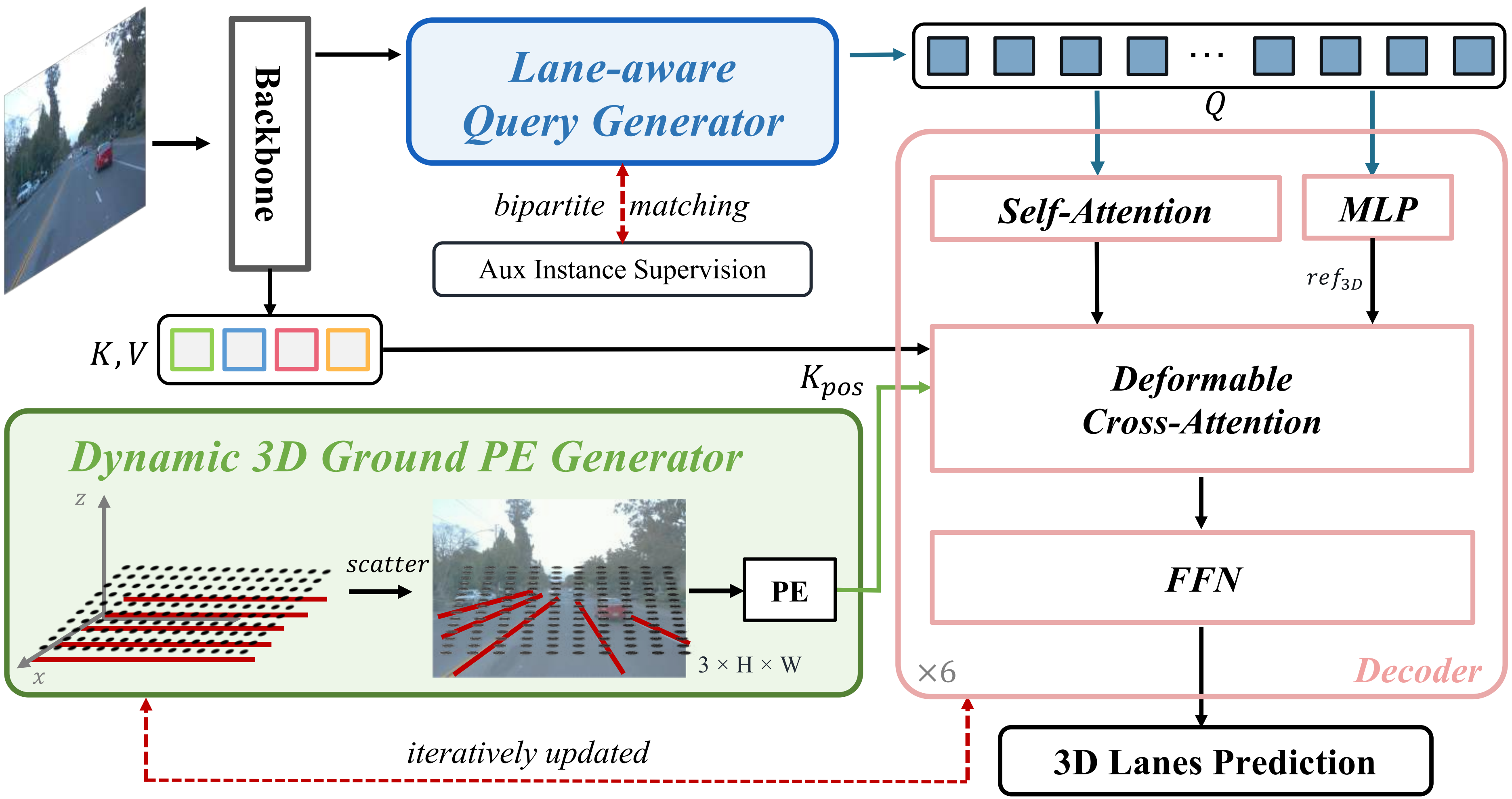
Predict



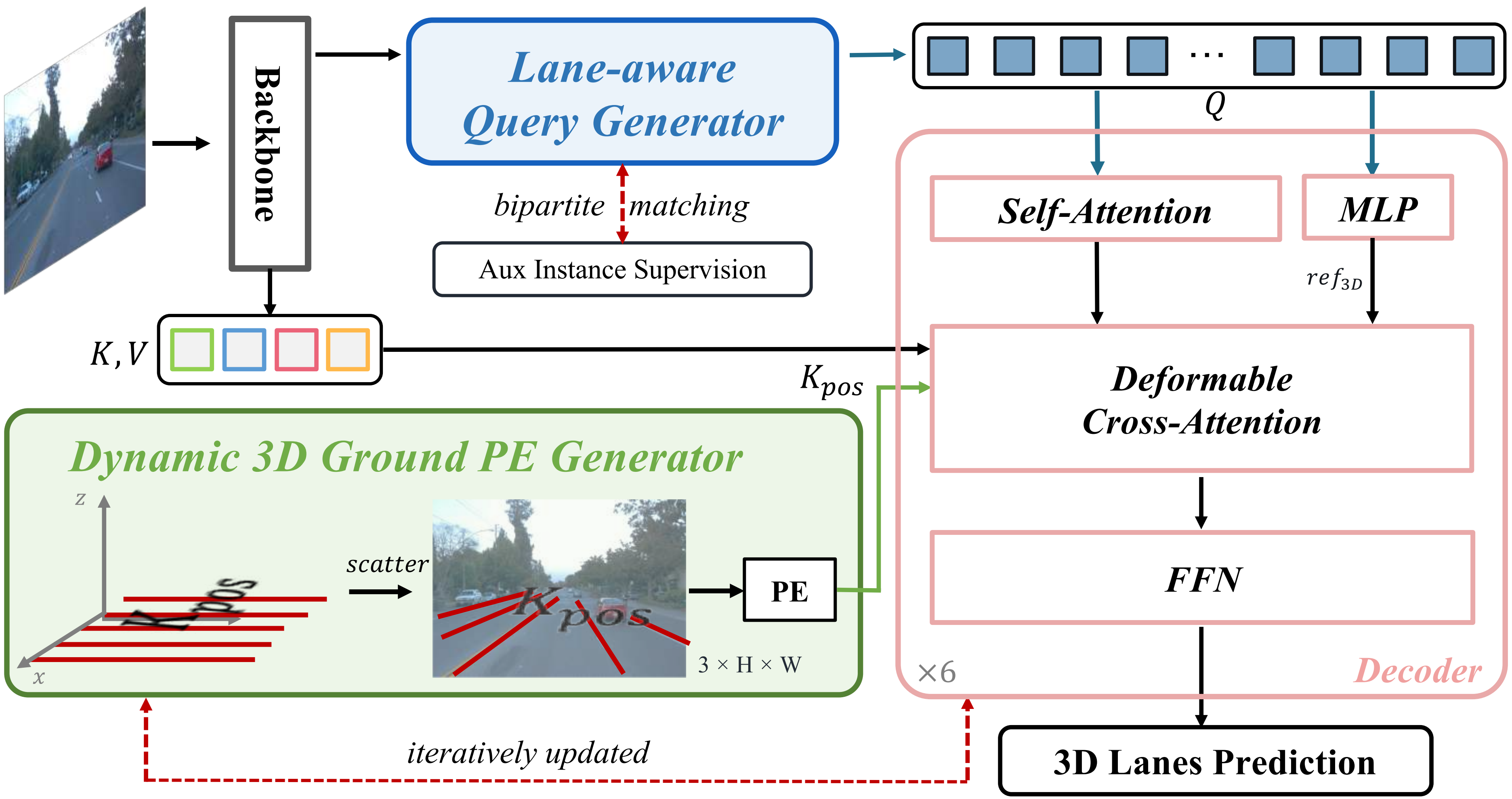
How to achieve end-to-end monocular 3D lane detection w/o surrogate representation?

☹ design anchors ☹ post-processing ☹ distortion caused by IPM

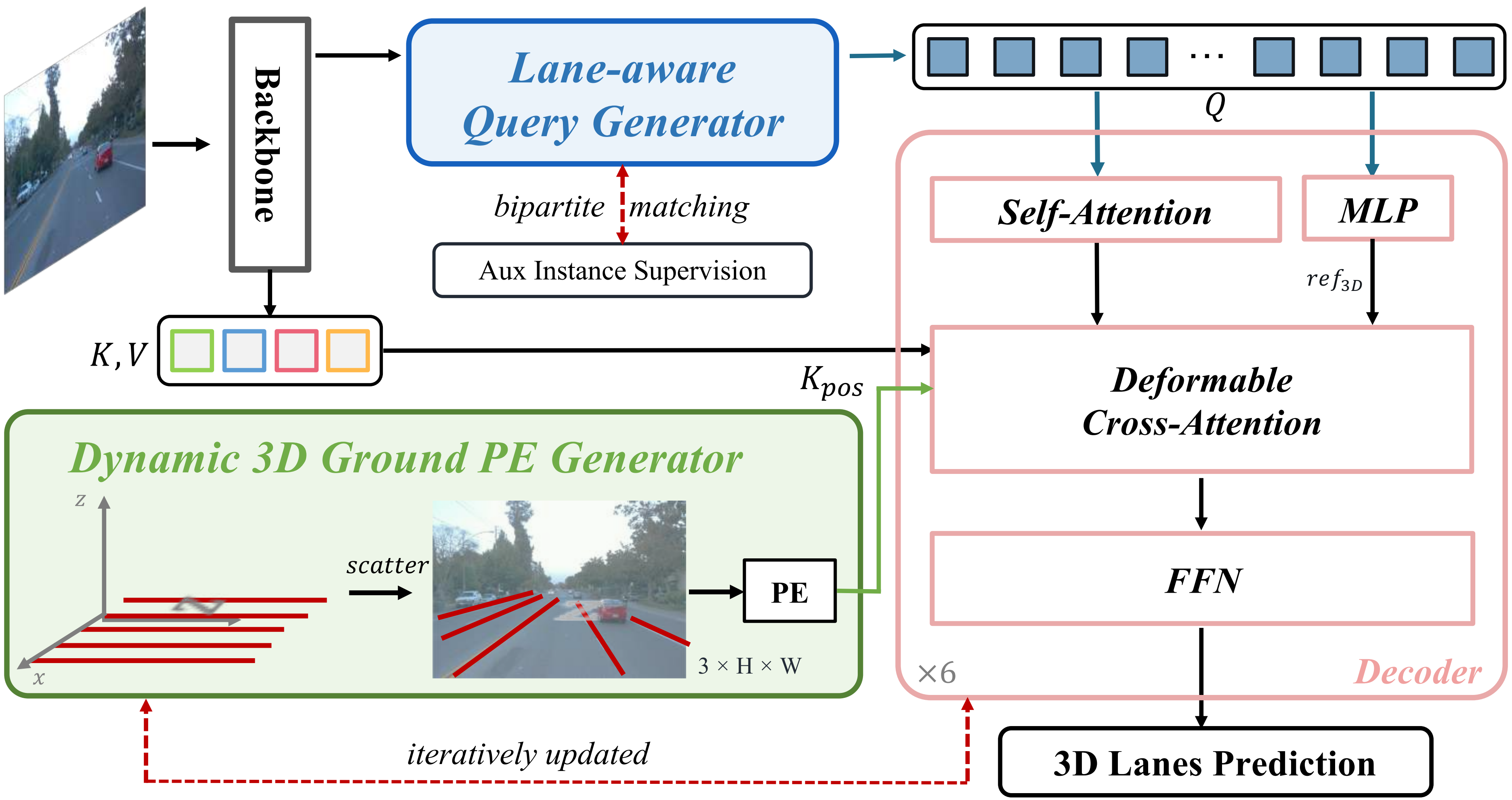
Our Framework



Our Framework

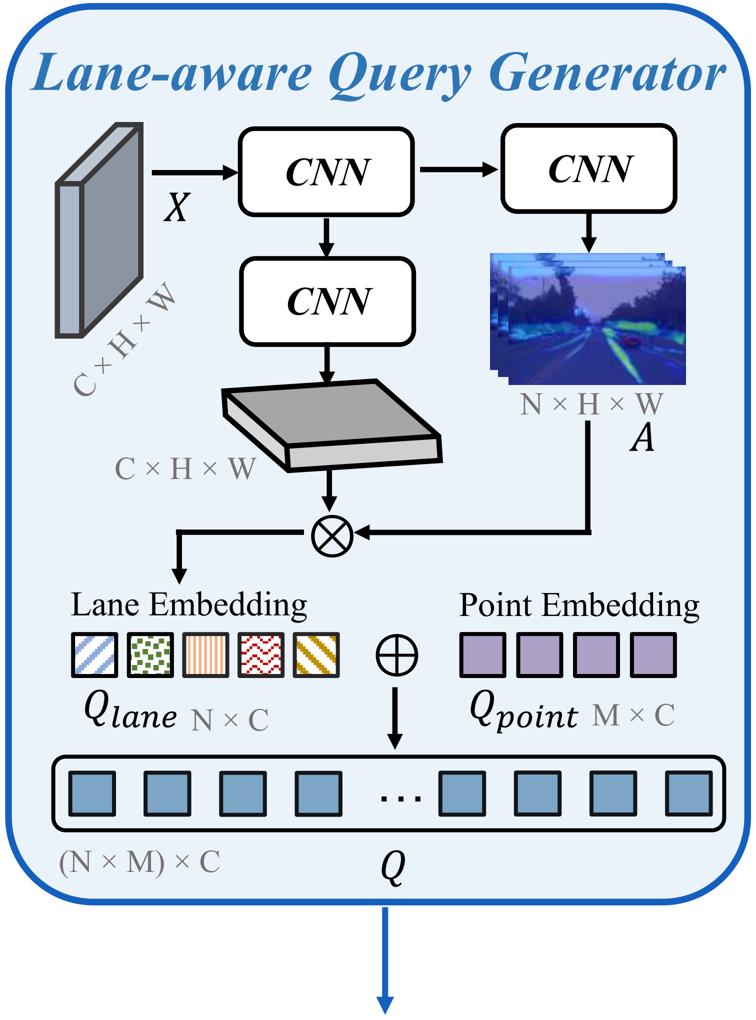


Our Framework

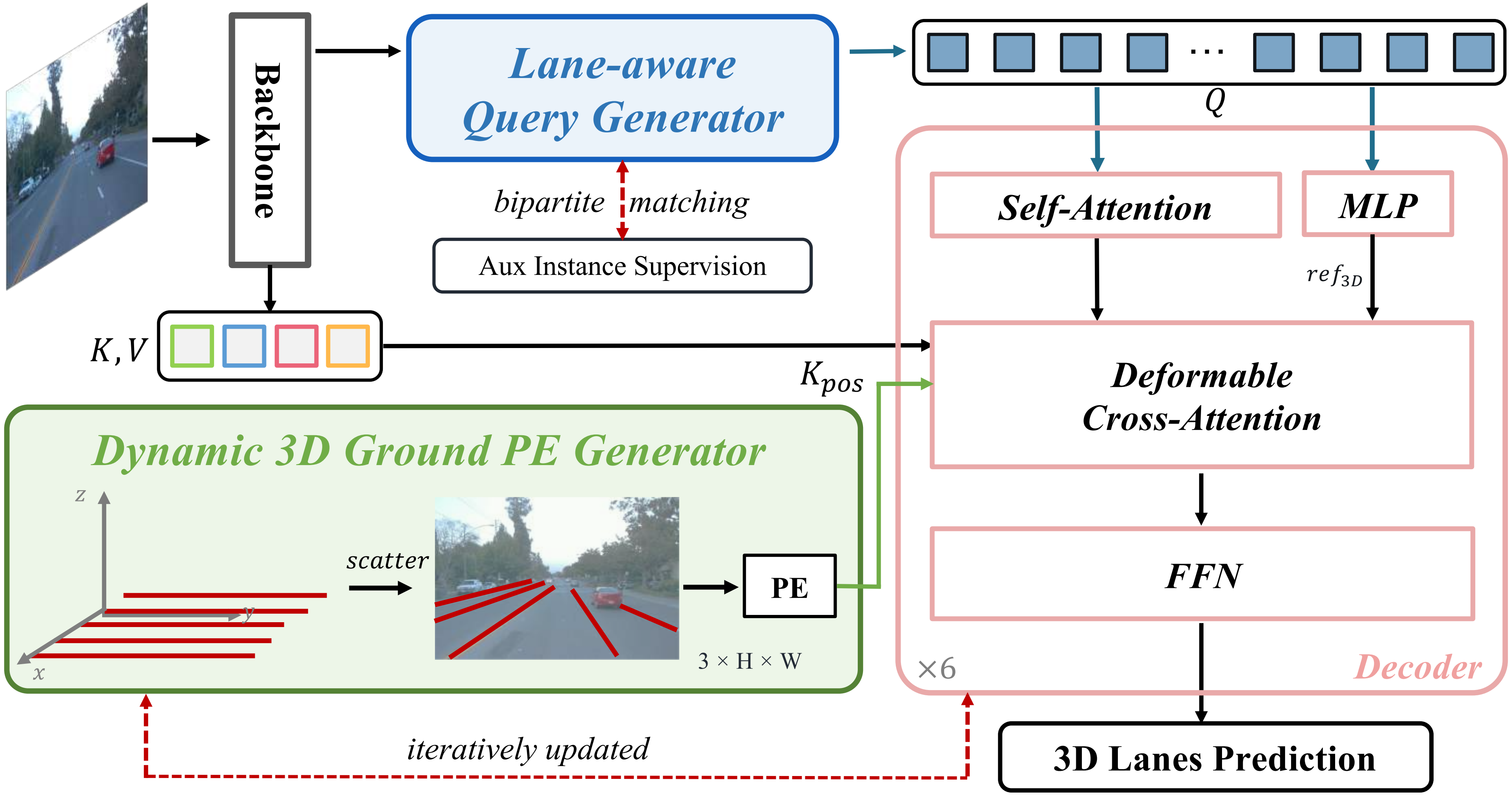


Lane-Aware Query Generator

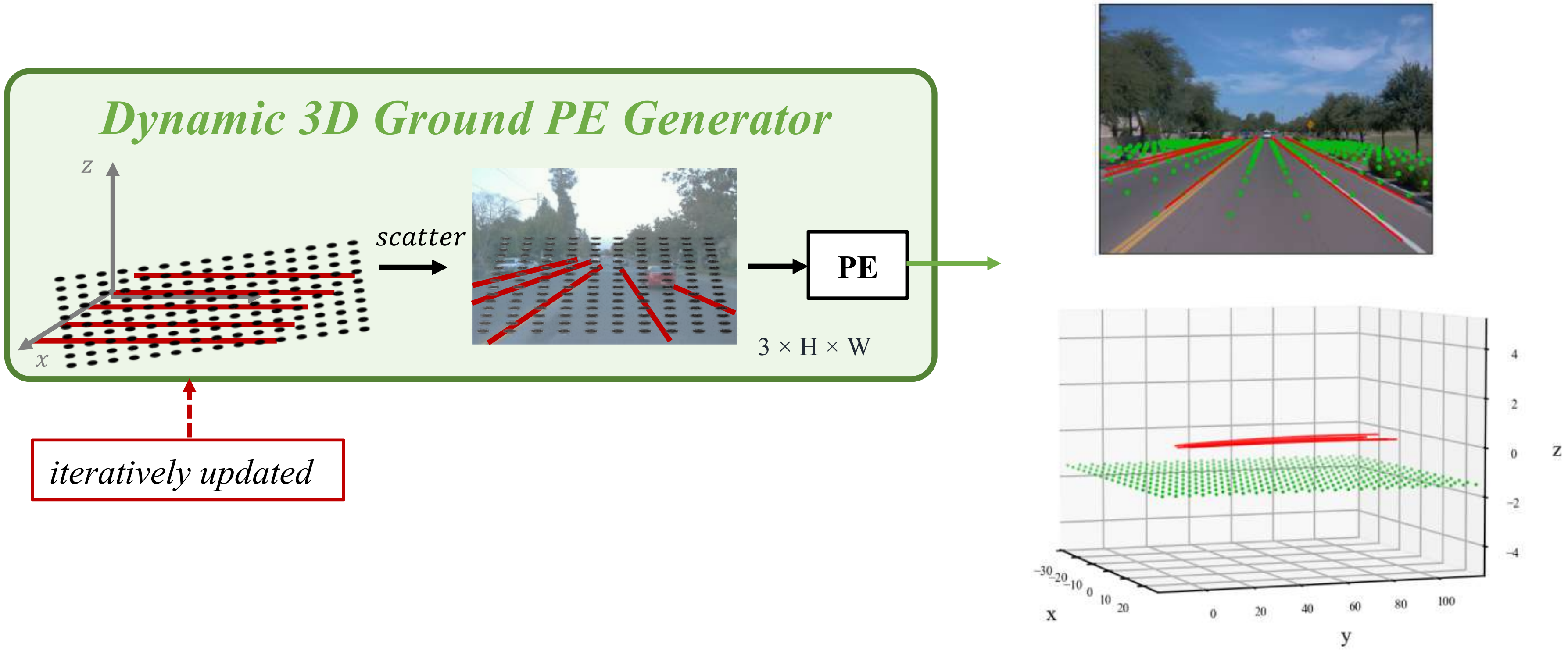
- Lane-aware:
  Instance Activation Map-based features
- Holistically:
  lane-level embedding
- Locally:
  point-level embedding



Dynamic 3D Ground Positional Embedding

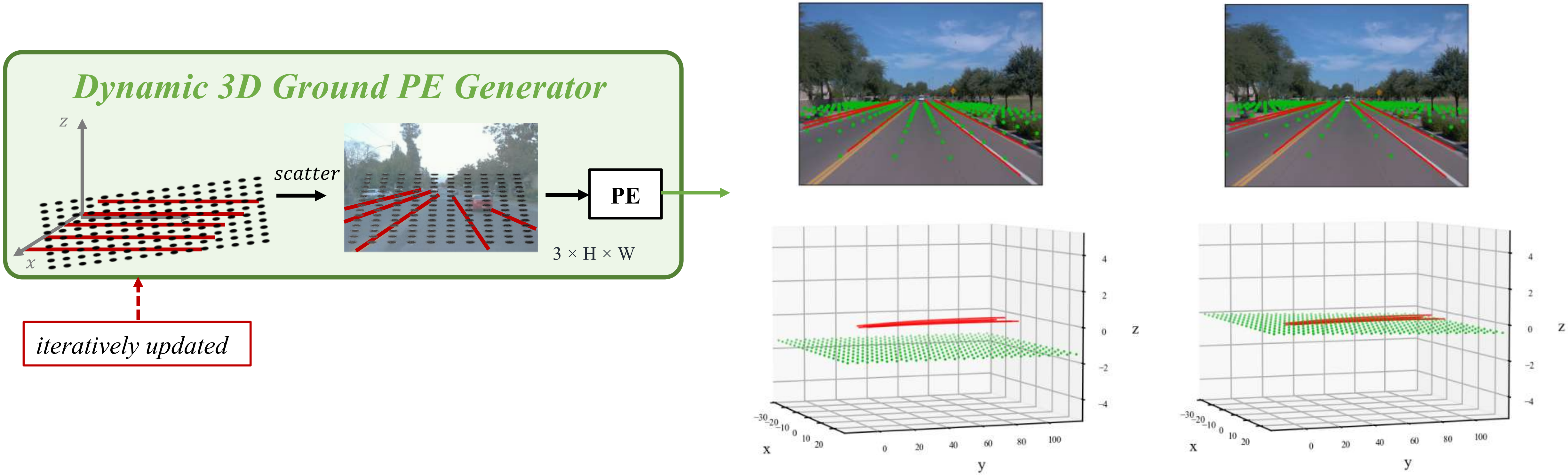


Dynamic 3D Ground Positional Embedding



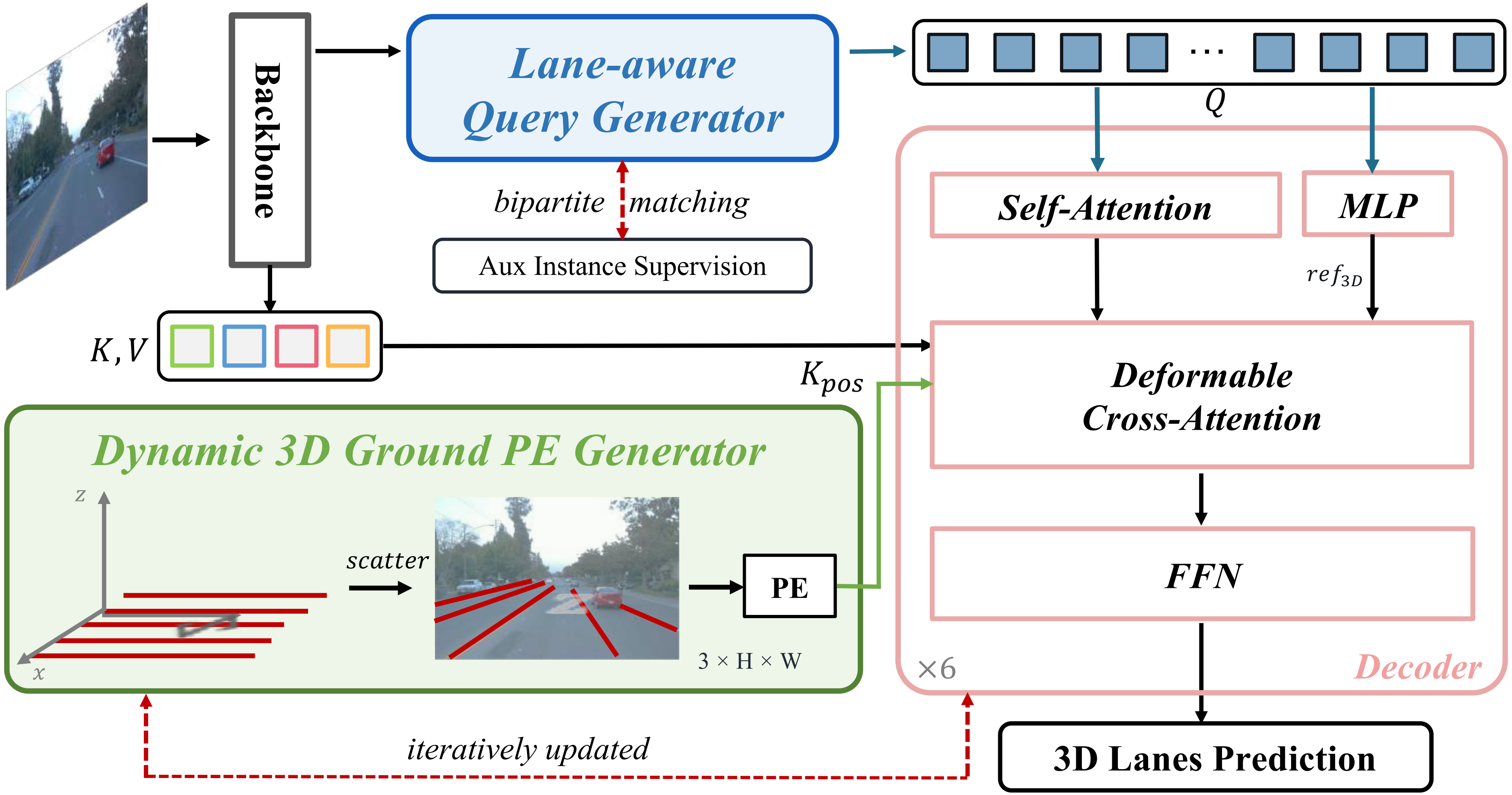
encode the 3D plane as the positional embedding for image features.

Dynamic 3D Ground Positional Embedding



encode the 3D plane as the positional embedding for image features.

Decoder


Backbone
Lane-aware Query Generator
bipartite matching
Aux Instance Supervision
Q
Self-Attention
MLP
ref_{3D}
K, V
K_{pos}
Deformable Cross-Attention
FFN
Dynamic 3D Ground PE Generator
z
x
scatter
PE
3 × H × W
×6
Decoder
iteratively updated
3D Lanes Prediction

Experimental Results

- Results on OpenLane

Methods	F1 ↑	Category Accuracy ↑	X error (m) ↓		Z error (m) ↓	
			near	*far*	*near*	*far*
3DLaneNet [7]	44.1	-	0.479	0.572	0.367	0.443
GenLaneNet [8]	32.3	-	0.593	0.494	0.140	0.195
Cond-IPM	36.6	-	0.563	1.080	0.421	0.892
Persformer* [3]	50.5	89.5	0.319	0.325	0.112	0.141
CurveFormer [1]	50.5	-	0.340	0.772	0.207	0.651
Persformer-Res50†	53.0	89.2	0.321	0.303	0.085	0.118
LATR-Lite	61.5	91.9	0.225	0.249	0.073	0.106
LATR	**61.9** ↑11.4	**92.0** ↑2.5	**0.219** ↓0.100	**0.259** ↓0.066	**0.075** ↓0.037	**0.104** ↓0.037

Experimental Results

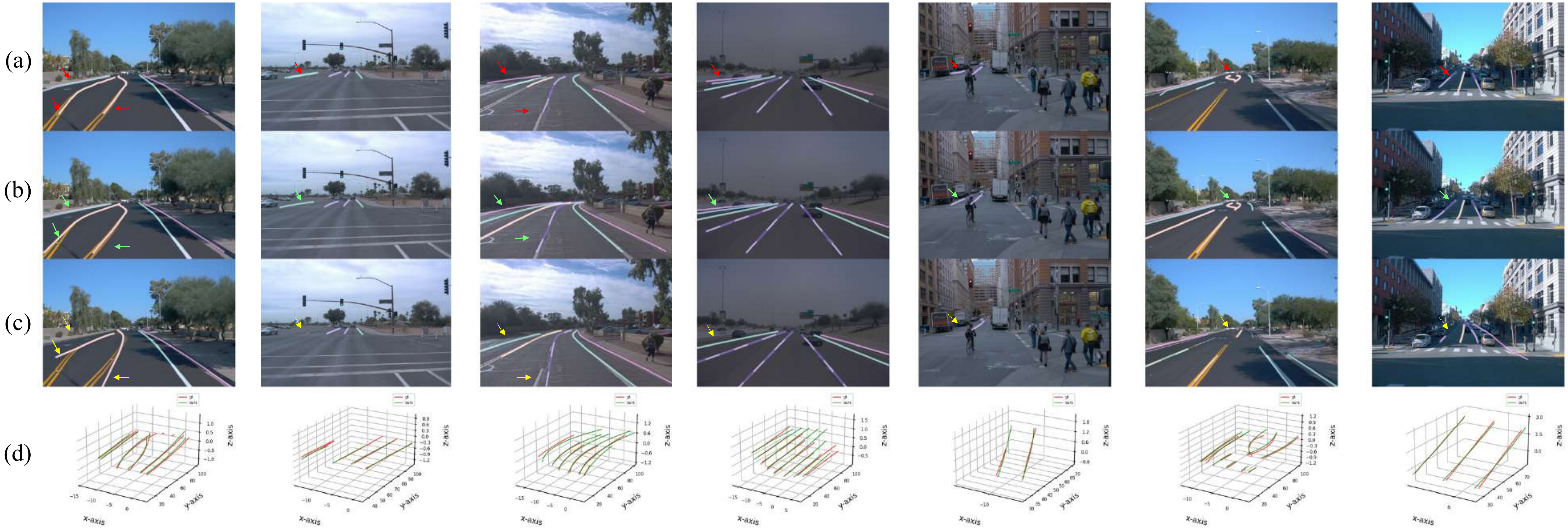
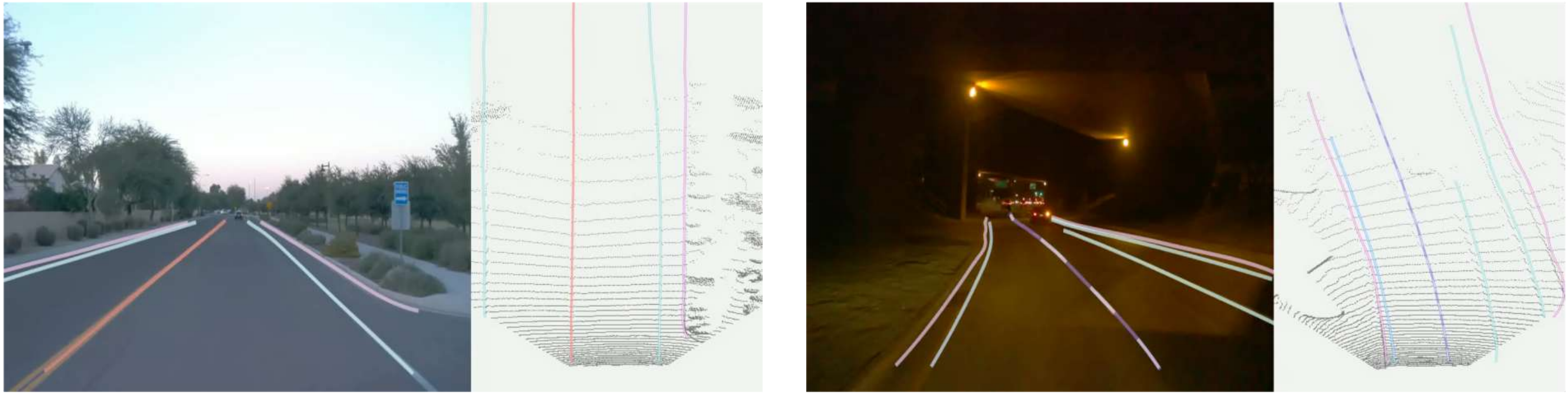
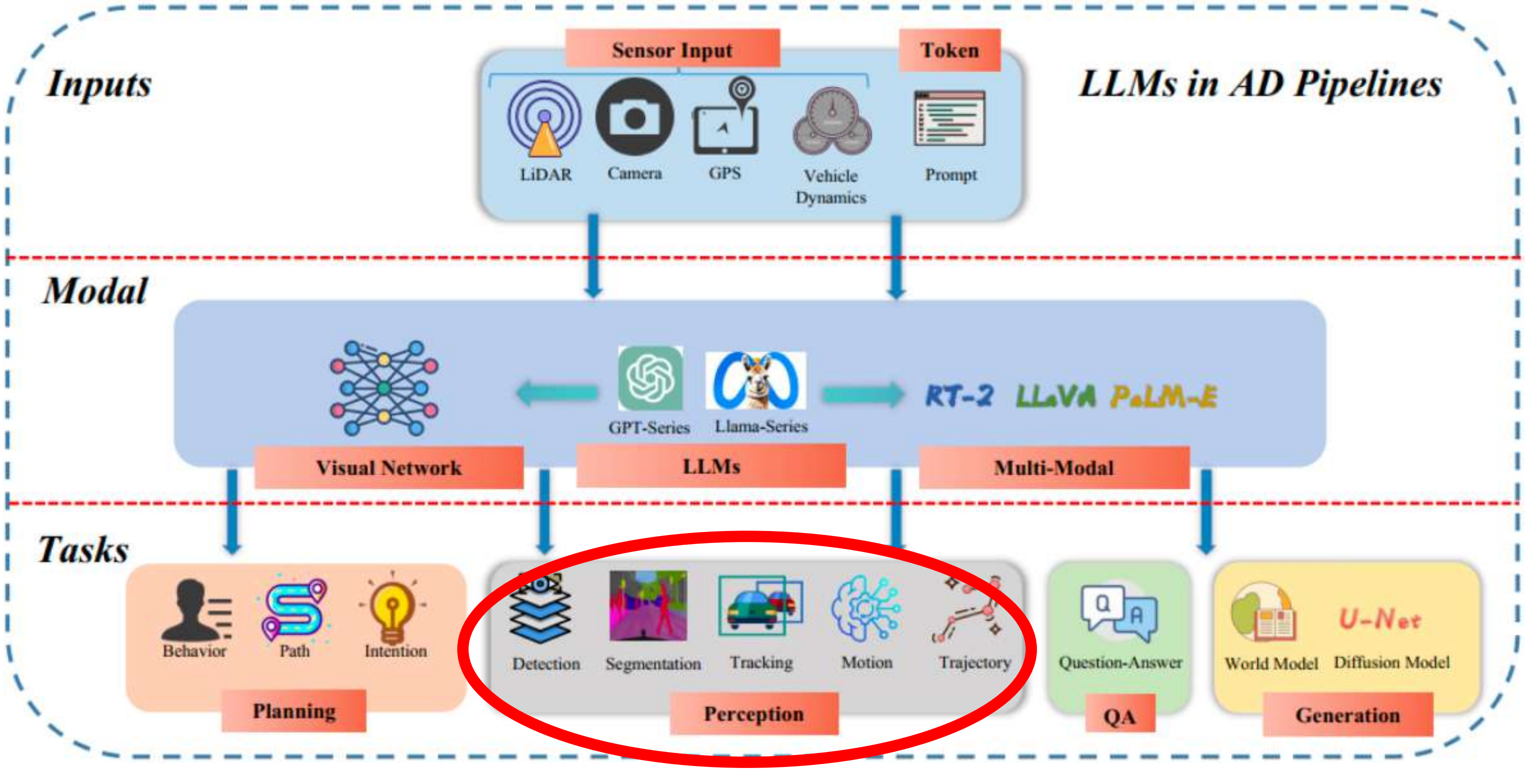


Figure 4. **Qualitative evaluation on OpenLane *val* set.** The rows (a), (b), (c) illustrate ground truth 3D lanes, prediction from LATR and Persformer [3] with 2D projection, respectively. Here, different colors indicate specific categories. Row (d) demonstrates the ground truth (red) and prediction of LATR (green) in 3D space. Best viewed in color (zoom in for details).

Results Video



Outline



Enhanced 3D Perception for AD

- Monocular (front/ego view)/**Multimodality** 3D Lane Detection
- 3D/4D Occupancy (world model)
- Semantic Segmentation and Semantic Completion

DV-3DLane: End-to-End Multi-Modal 3D Lane Detection with Dual-View Representation

Yueru Luo[1,2] Shuguang Cui[2,1] Zhen Li[2,1,✉]

[1]FNii, CUHK-Shenzhen , [2]School of Science and Engineering, CUHK-Shenzhen

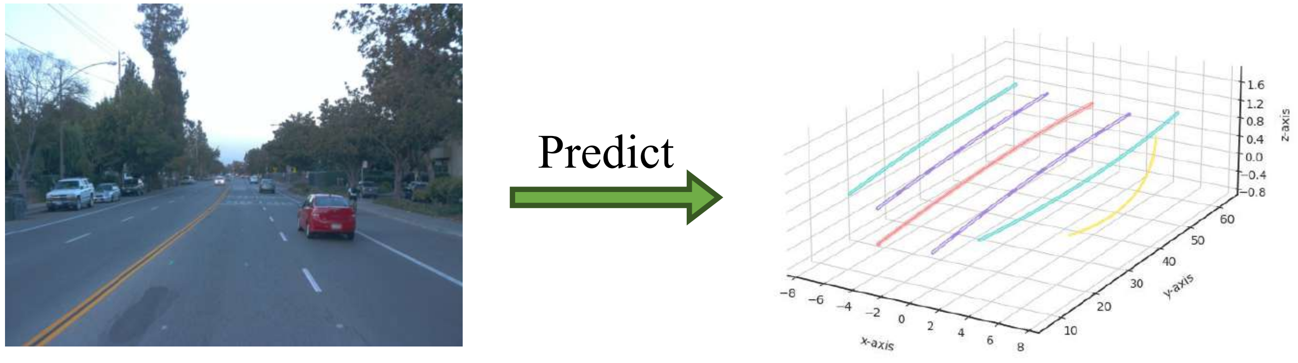
(ICLR 2024)

3D Lane Detection

Goal: localize lane boundaries in 3D space based on inputs.

Inherent Challenges: slenderness and elongation of lanes.

Existing Solutions: mainly formulate based on **monocular** images.



Monocular Solutions

depth loss

illumunation sensitivity

Challenges +

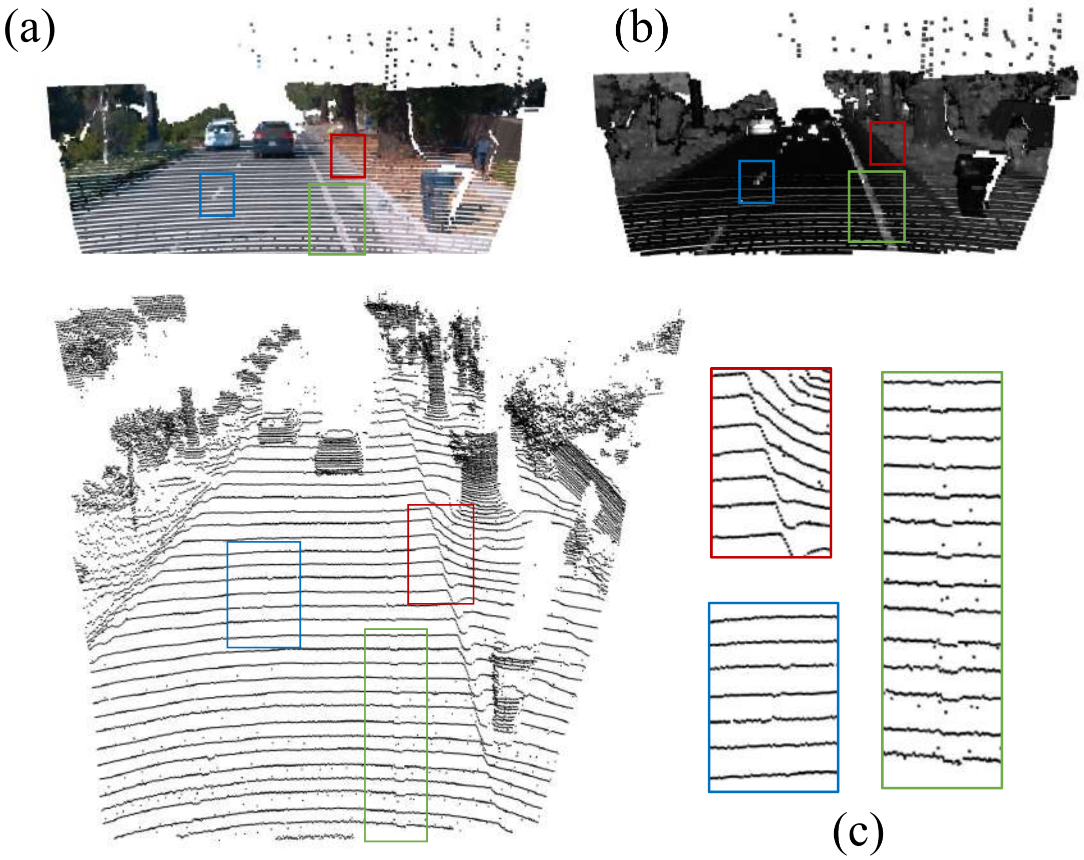
Can We Utilize **Multi-modal** Data to Facilitate 3D Lane Detection?

Multi-modal 3D Lane Detection

Reflection

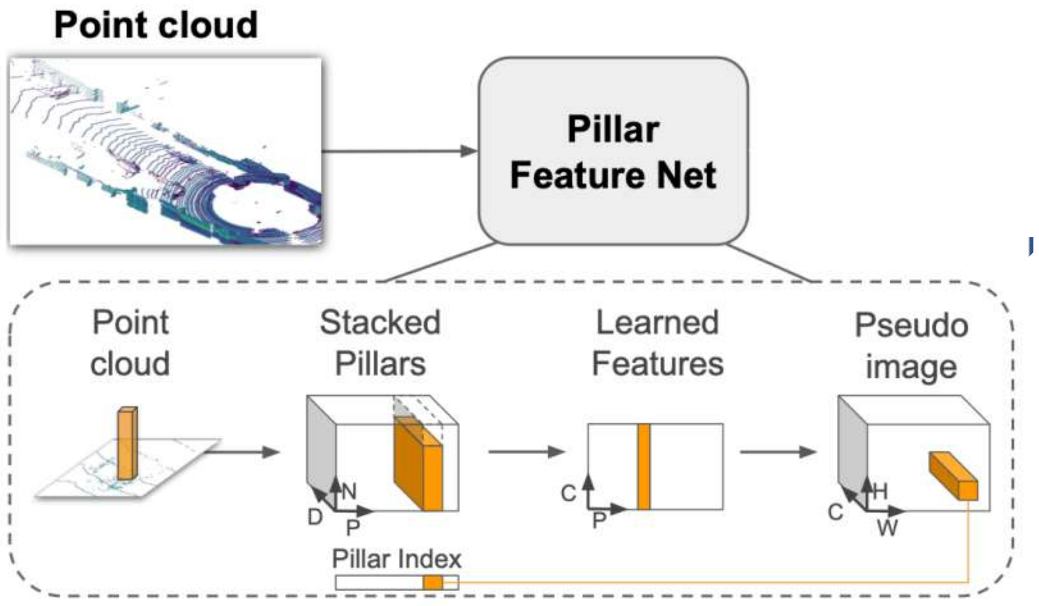
Geometry

LiDAR **CAN HELP** detect 3D lanes!



Luo, Y., et al. (2022). M ²-3DLaneNet: Exploring Multi-Modal 3D Lane Detection. Arxiv.

End-to-End Multi-modal 3D Lane Detection with Dual-view Representation

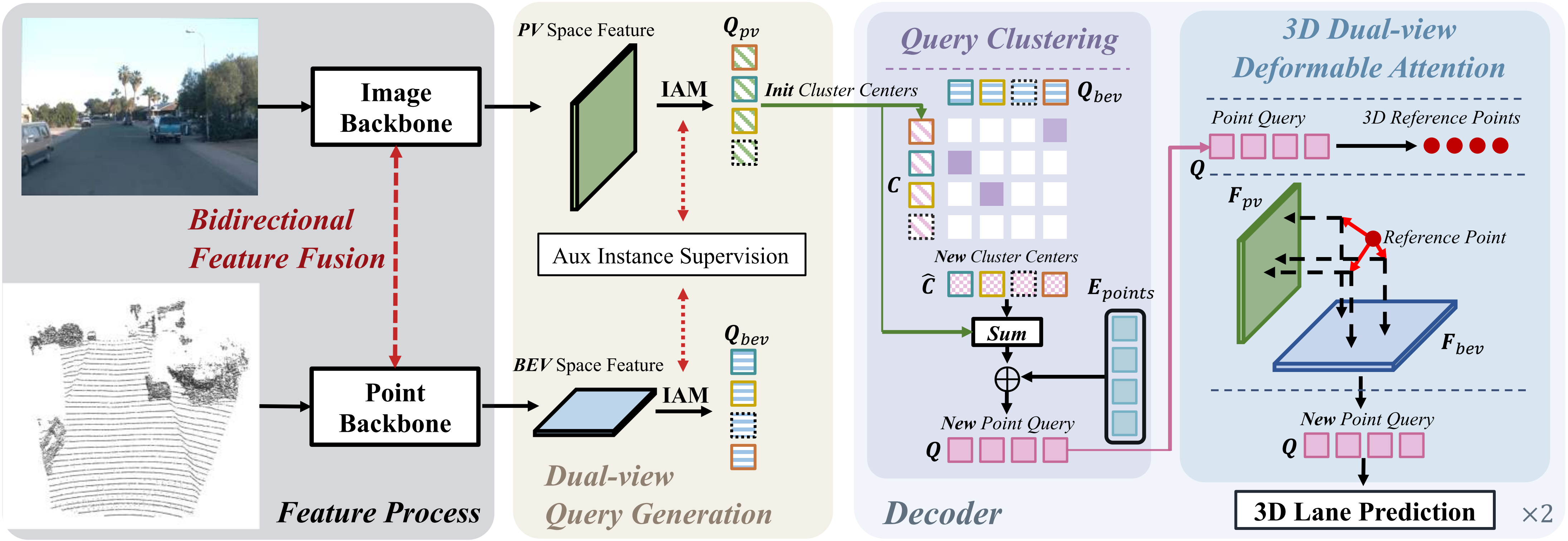


DV-3DLane

Luo, Y., et al. (2023). LATR: 3D Lane Detection from Monocular Images with Transformer. ICCV.
Lang, Alex H., et al. (2019). Pointpillars: Fast encoders for object detection from point clouds. CVPR.

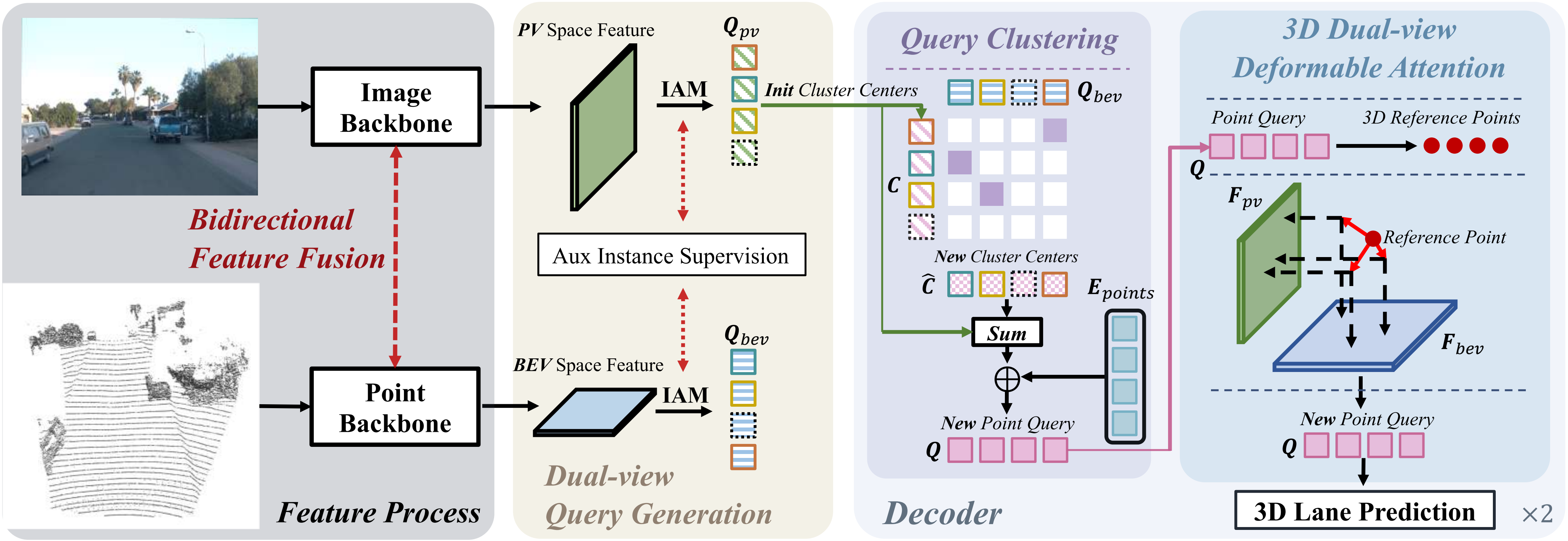
DV-3DLane

- Bidirectional feature fusion
- Unified query generator
- 3D dual-view deformable attention



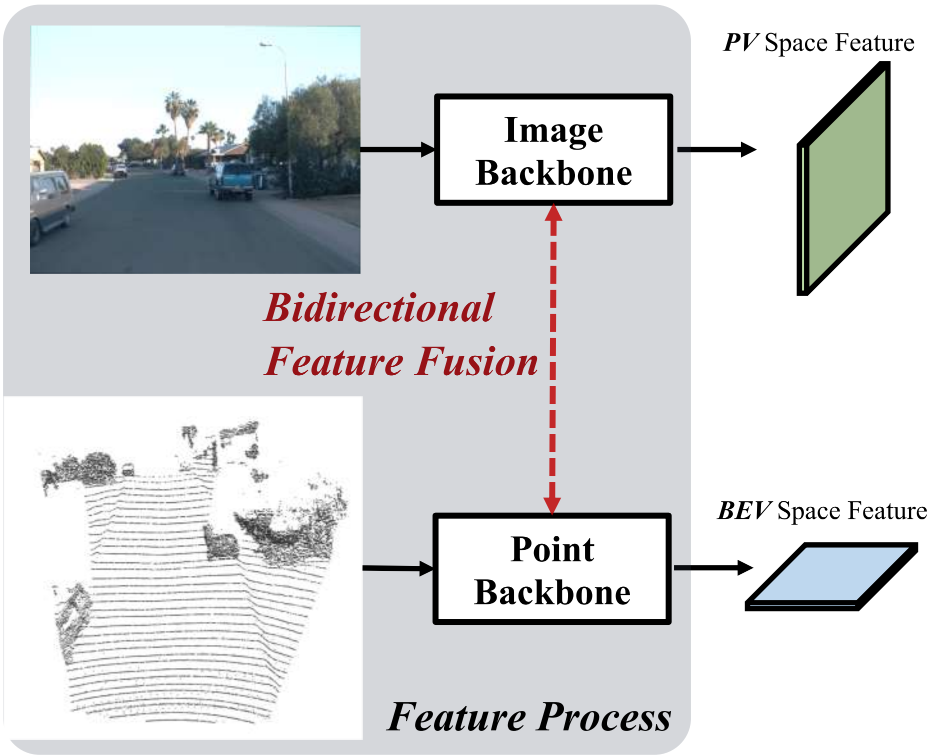
DV-3DLane

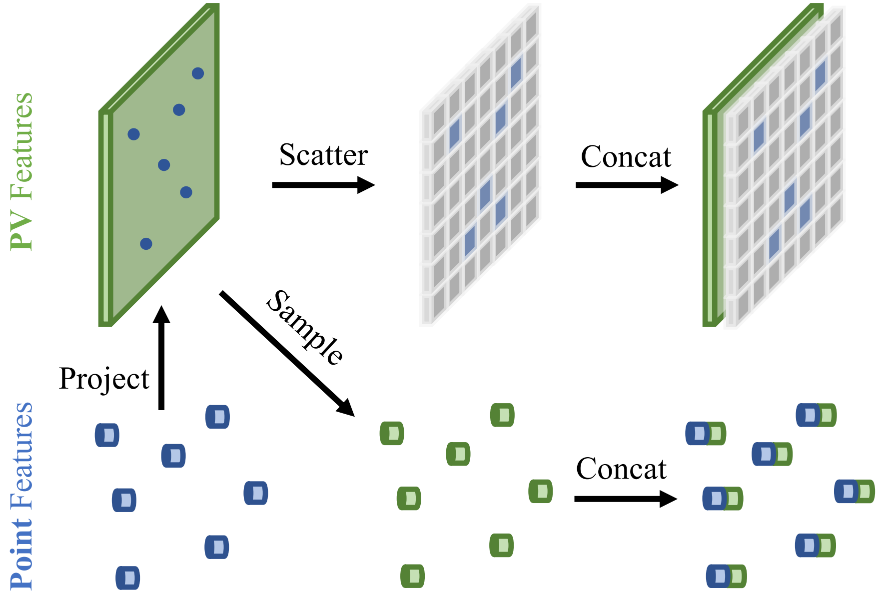
- Bidirectional feature fusion
- Unified query generator
- 3D dual-view deformable attention



DV-3DLane

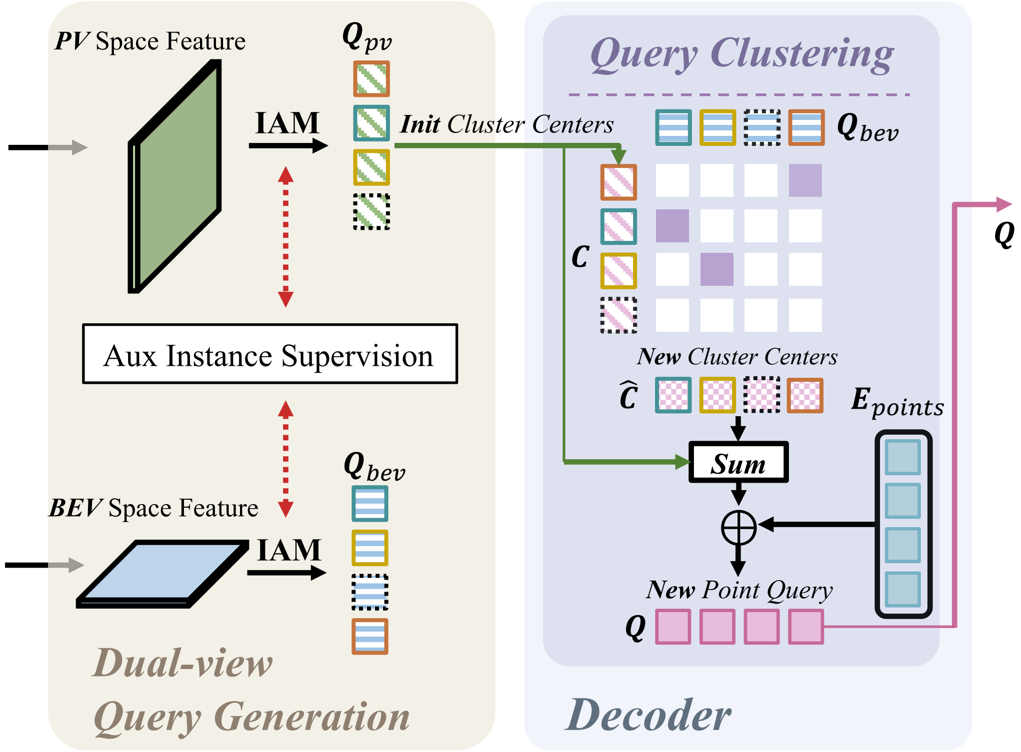
- **Bidirectional feature fusion**

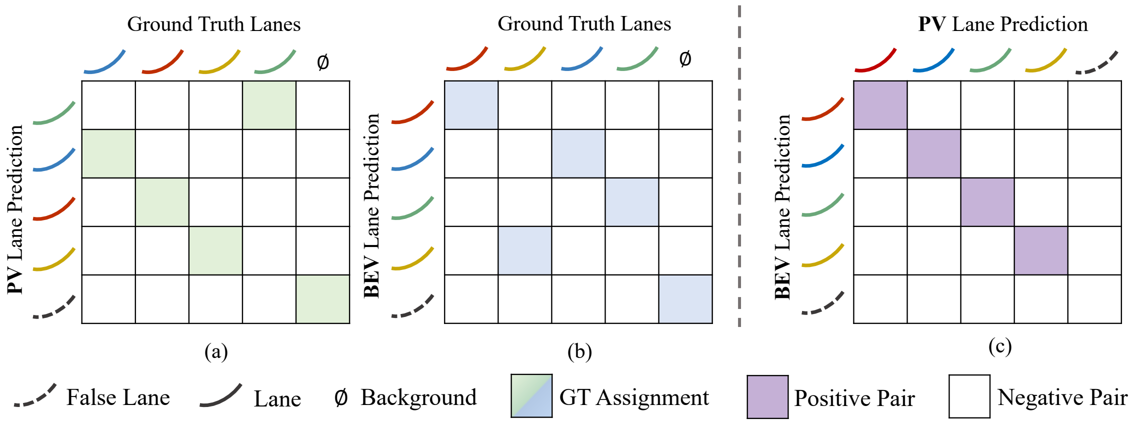




DV-3DLane

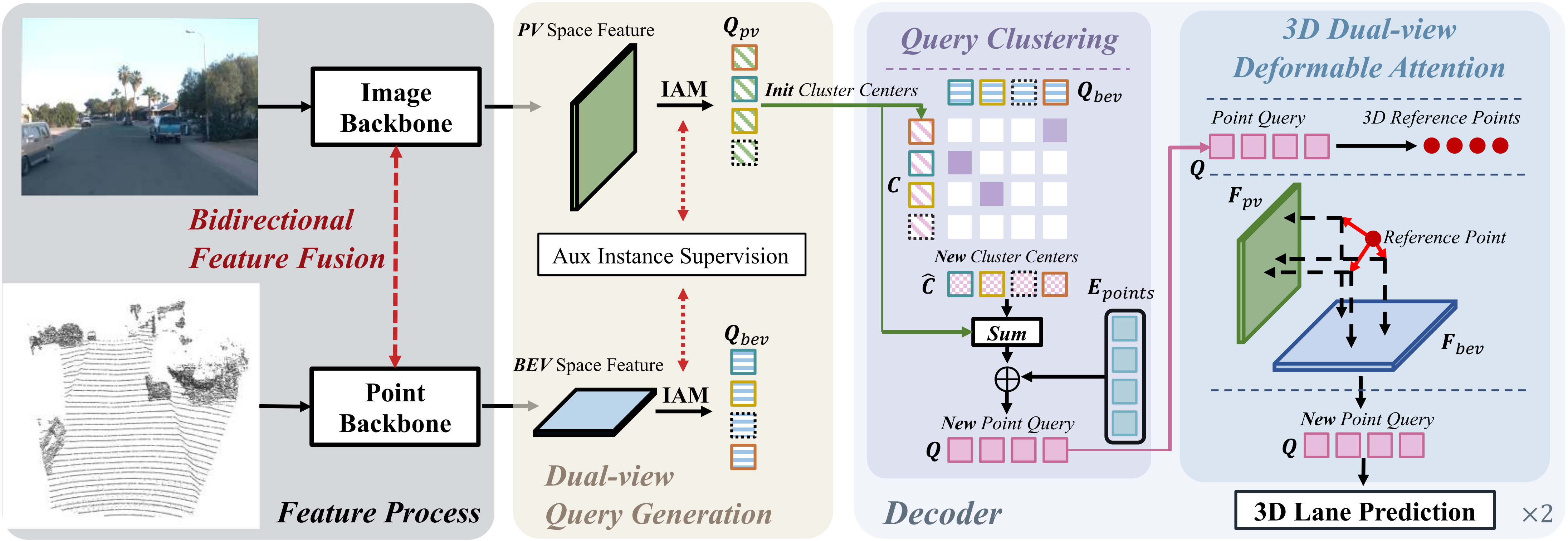
- **Unified query generator**
  1. Dual-view query generation
  2. Dual-view query clustering





DV-3DLane

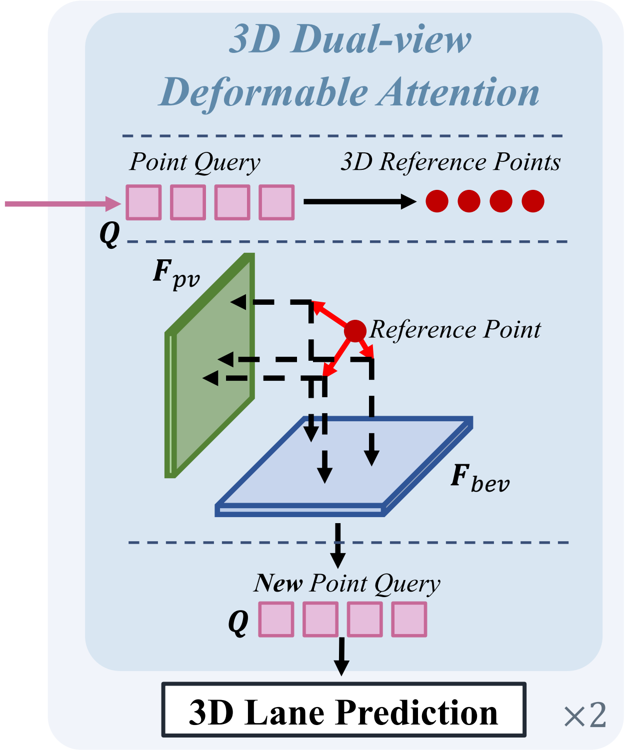
- Bidirectional feature fusion
- Unified query generator
- 3D dual-view deformable attention

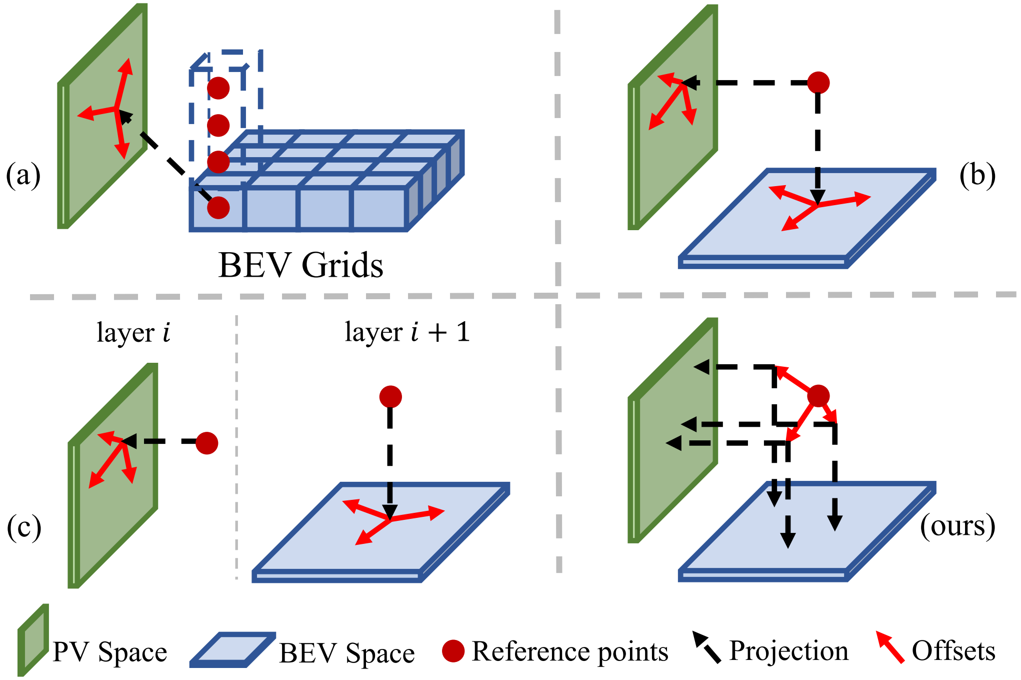


DV-3DLane

➢ 3D dual-view deformable attention

Consistantly sample features from dual-view spaces.

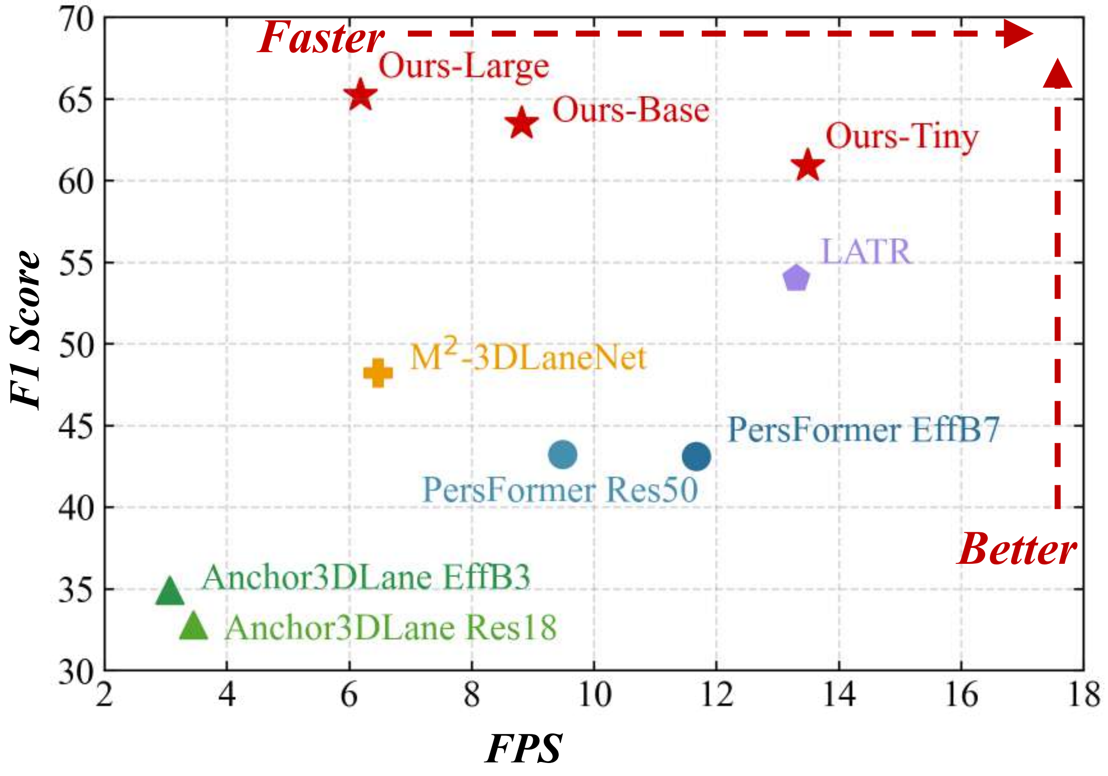


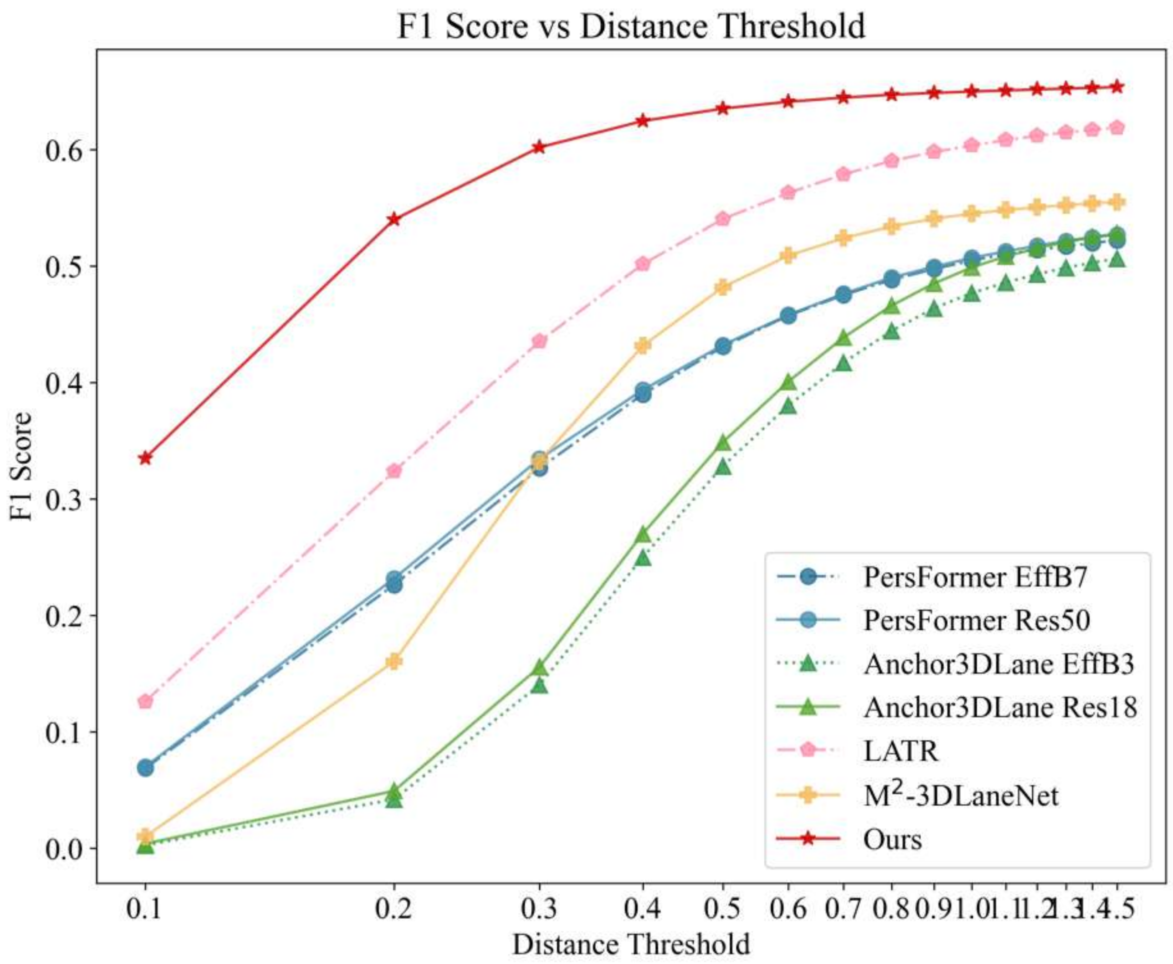


Comparison with other methods.

Li, Z., et al. (2022). Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. ECCV.
Yang, Z., et al. (2022). Deepinteraction: 3d object detection via modality interaction. NeurIPS.
Chen, X., et al. (2023). Futr3d: A unified sensor fusion framework for 3d detection. ICCV.

Results on OpenLane





Results on OpenLane

Dist.	Methods	Backbone	Modality	F1 ↑	*Acc.* ↑	X error (m) ↓		Z error (m) ↓	
						near	*far*	*near*	*far*
1.5 m	PersFormer	EffNet-B7	C	50.5	89.5	0.319	0.325	0.112	0.141
	Anchor3DLane†	EffNet-B3	C	52.8	89.6	0.408	0.349	0.186	0.143
	M^2-3DLaneNet	EffNet-B7	C+L	55.5	88.2	0.283	<u>0.256</u>	0.078	0.106
	Anchor3DLane†	ResNet-18	C	50.7	89.3	0.422	0.349	0.188	0.146
	PersFormer	ResNet-50	C	52.7	88.4	0.307	0.319	0.083	0.117
	LATR	ResNet-50	C	<u>61.9</u>	<u>92.0</u>	<u>0.219</u>	0.259	<u>0.075</u>	<u>0.104</u>
	DV-3DLane-Tiny (Ours)	ResNet-18	C+L	63.4	91.6	0.137	0.159	0.034	0.063
	DV-3DLane-Base (Ours)	ResNet-34	C+L	65.4	92.4	0.118	0.131	0.032	0.053
	DV-3DLane-Large (Ours)	ResNet-50	C+L	**66.8**	**93.3**	**0.115**	**0.134**	**0.029**	**0.049**
	Improvement	-	-	*↑4.9*	*↑1.3*	*↓0.104*	*↓0.122*	*↓0.046*	*↓0.055*
0.5 m	PersFormer	EffNet-B7	C	36.5	87.8	0.343	0.263	0.161	0.115
	Anchor3DLane†	EffNet-B3	C	34.9	88.5	0.344	0.264	0.181	0.134
	M^2-3DLaneNet	EffNet-B7	C+L	48.2	88.1	0.217	0.203	0.076	0.103
	Anchor3DLane†	ResNet-18	C	32.8	87.9	0.350	0.266	0.183	0.137
	PersFormer	ResNet-50	C	43.2	87.8	0.229	0.245	0.078	0.106
	LATR	ResNet-50	C	<u>54.0</u>	<u>91.7</u>	<u>0.171</u>	<u>0.201</u>	<u>0.072</u>	<u>0.099</u>
	DV-3DLane-Tiny (Ours)	ResNet-18	C+L	60.9	91.8	0.097	0.124	0.033	0.062
	DV-3DLane-Base (Ours)	ResNet-34	C+L	63.5	92.4	0.090	0.102	0.031	0.053
	DV-3DLane-Large (Ours)	ResNet-50	C+L	**65.2**	**93.4**	**0.082**	**0.101**	**0.028**	**0.048**
	Improvement	-	-	*↑11.2*	*↑1.7*	*↓0.089*	*↓0.100*	*↓0.044*	*↓0.051*

Results on Various Scenarios:

Dist.	Methods	Backbone	Modality	All	Up & Down	Curve	Extreme Weather	Night	Intersection	Merge & Split
1.5 m	PersFormer	EffNet-B7	C	50.5	42.4	55.6	48.6	46.6	40.0	50.7
	Anchor3DLane†	EffNet-B3	C	52.8	48.5	50.7	56.9	43.6	48.5	50.7
	M^2-3DLaneNet	EffNet-B7	C+L	55.5	53.4	60.7	56.2	51.6	43.8	51.4
	PersFormer	ResNet-50	C	52.7	46.4	57.9	52.9	47.2	41.6	51.4
	LATR	ResNet-50	C	<u>61.9</u>	<u>55.2</u>	<u>68.2</u>	<u>57.1</u>	<u>55.4</u>	<u>52.3</u>	<u>61.5</u>
	Anchor3DLane†	ResNet-18	C	50.7	45.3	53.7	48.5	51.6	45.3	48.5
	DV-3DLane-Tiny	ResNet-18	C+L	63.4	59.9	69.8	62.2	58.8	53.5	60.6
	DV-3DLane-Base	ResNet-34	C+L	65.4	60.9	**72.1**	64.5	61.3	55.5	61.6
	DV-3DLane-Large	ResNet-50	C+L	**66.8**	**61.1**	71.5	**64.9**	**63.2**	**58.6**	**62.8**
	Improvement	-	-	*↑4.9*	*↑5.9*	*↑3.9*	*↑7.8*	*↑7.8*	*↑6.3*	*↑1.3*
0.5 m	PersFormer	EffNet-B7	C	36.5	26.8	36.9	33.9	34.0	28.5	37.4
	Anchor3DLane†	EffNet-B3	C	34.9	28.3	31.8	30.7	32.2	29.9	33.9
	M^2-3DLaneNet	EffNet-B7	C+L	48.2	40.7	48.2	<u>49.8</u>	<u>46.2</u>	38.7	44.2
	PersFormer	ResNet-50	C	43.2	36.3	42.4	45.4	39.3	32.9	41.7
	LATR	ResNet-50	C	<u>54.0</u>	<u>44.9</u>	<u>56.2</u>	47.6	<u>46.2</u>	<u>45.5</u>	<u>55.6</u>
	Anchor3DLane†	ResNet-18	C	32.8	26.5	27.6	31.2	30.0	28.1	31.7
	DV-3DLane-Tiny	ResNet-18	C+L	60.9	56.9	65.9	60.0	56.8	50.7	57.6
	DV-3DLane-Base	ResNet-34	C+L	63.5	58.6	**69.3**	62.4	59.9	53.9	59.3
	DV-3DLane-Large	ResNet-50	C+L	**65.2**	**59.1**	69.2	**63.0**	**62.0**	**56.9**	**60.5**
	Improvement	-	-	*↑11.2*	*↑14.2*	*↑13.1*	*↑13.2*	*↑15.8*	*↑11.4*	*↑4.9*

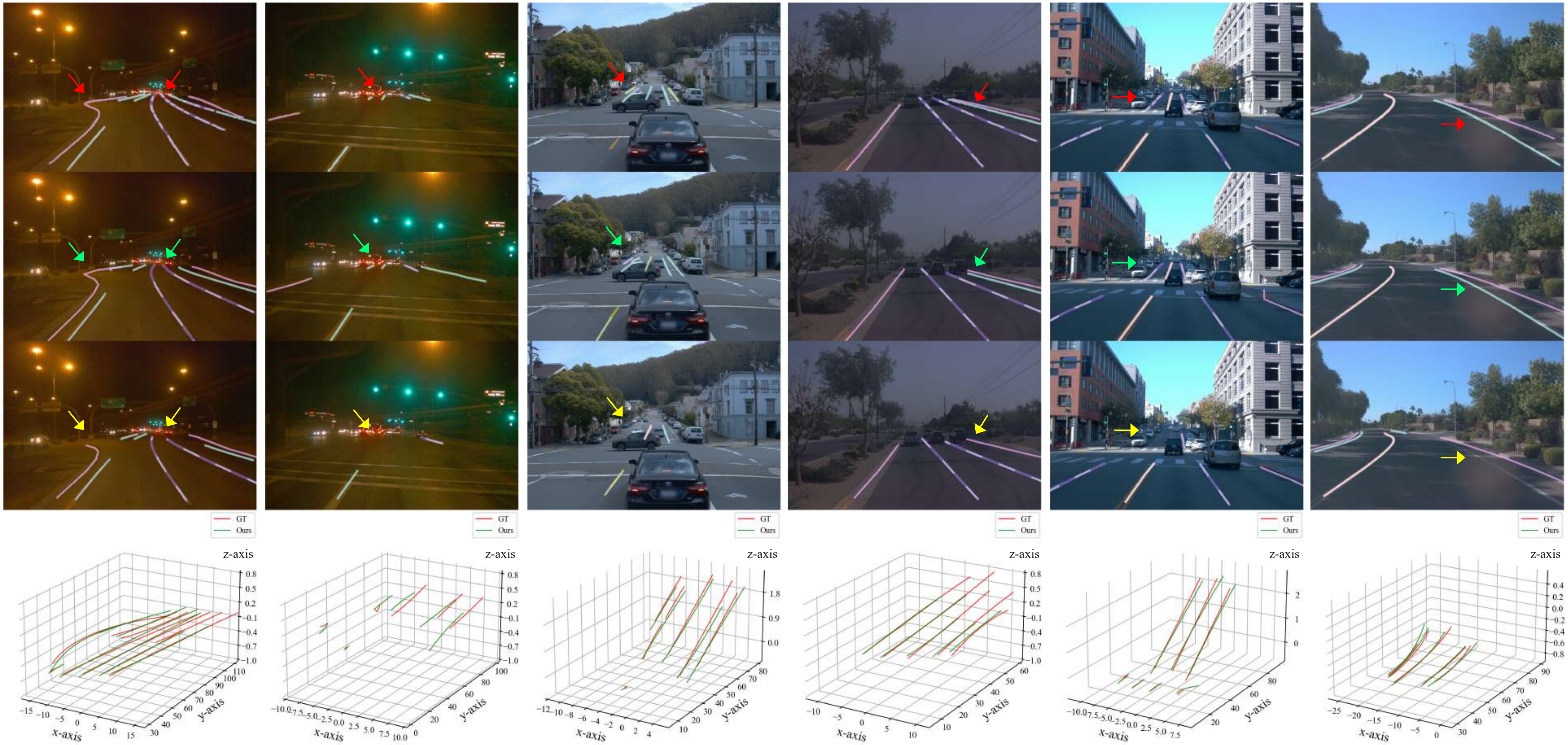
Visualization

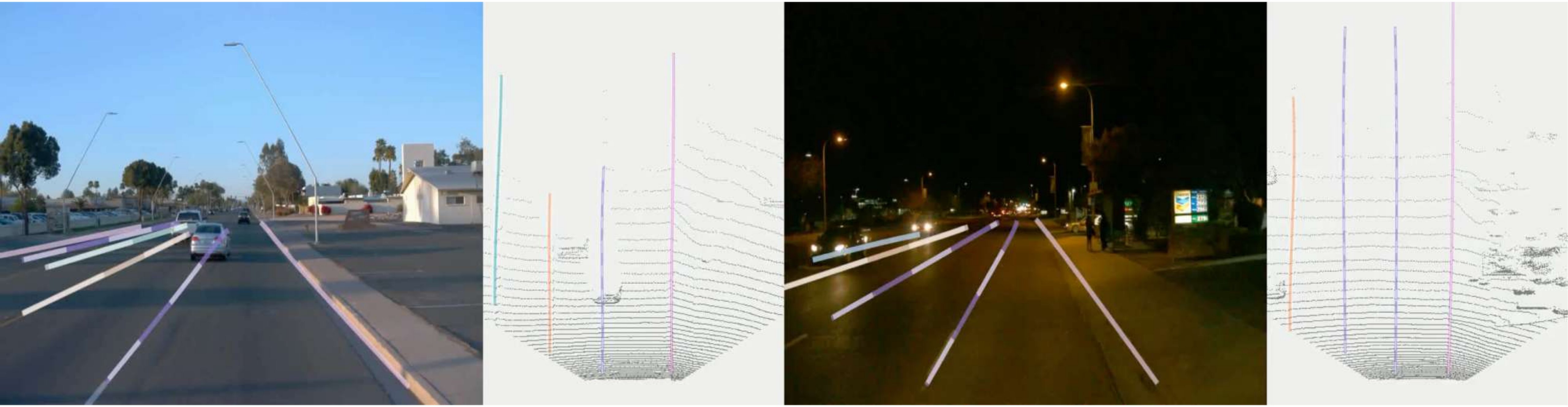
(a)

(b)

(c)

(d)





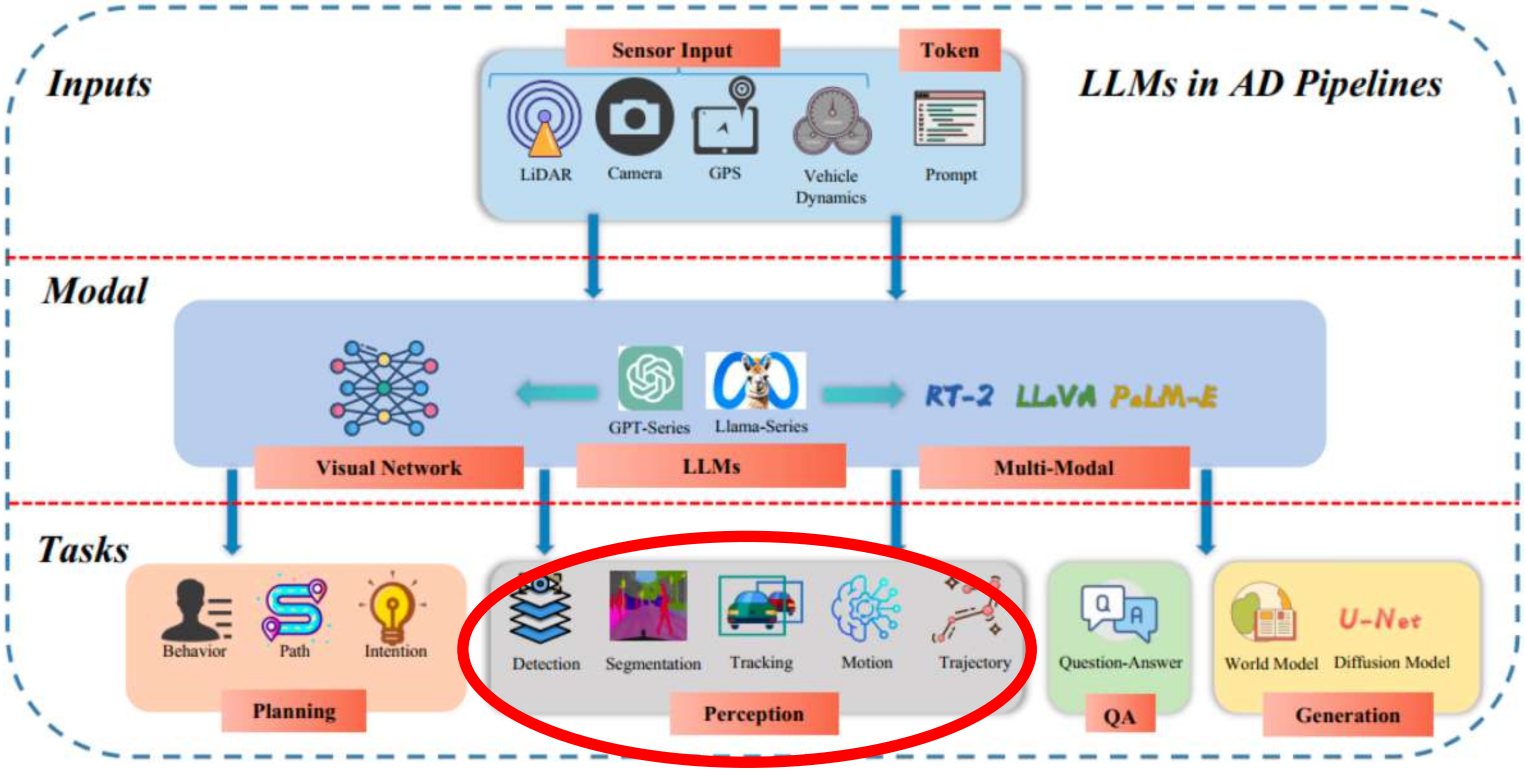
Thanks!

Deep Bit Lab

Paper

Code

Outline



Enhanced 3D Perception for AD

- ➢ Monocular (front/ego view)/Multimodality 3D Lane Detection
- ➢ **3D/4D Occupancy (world model)**
- ➢ Semantic Segmentation and Semantic Completion

RadOcc: Learning Cross-Modality Occupancy Knowledge through Rendering Assisted Distillation

Haiming Zhang [1,2], Xu Yan [3], Dongfeng Bai [3], Jiantao Gao [3],
Pan Wang [3], Bingbing Liu [3], Shuguang Cui [2,1,†] and ***Zhen Li*** *[2,1†]*

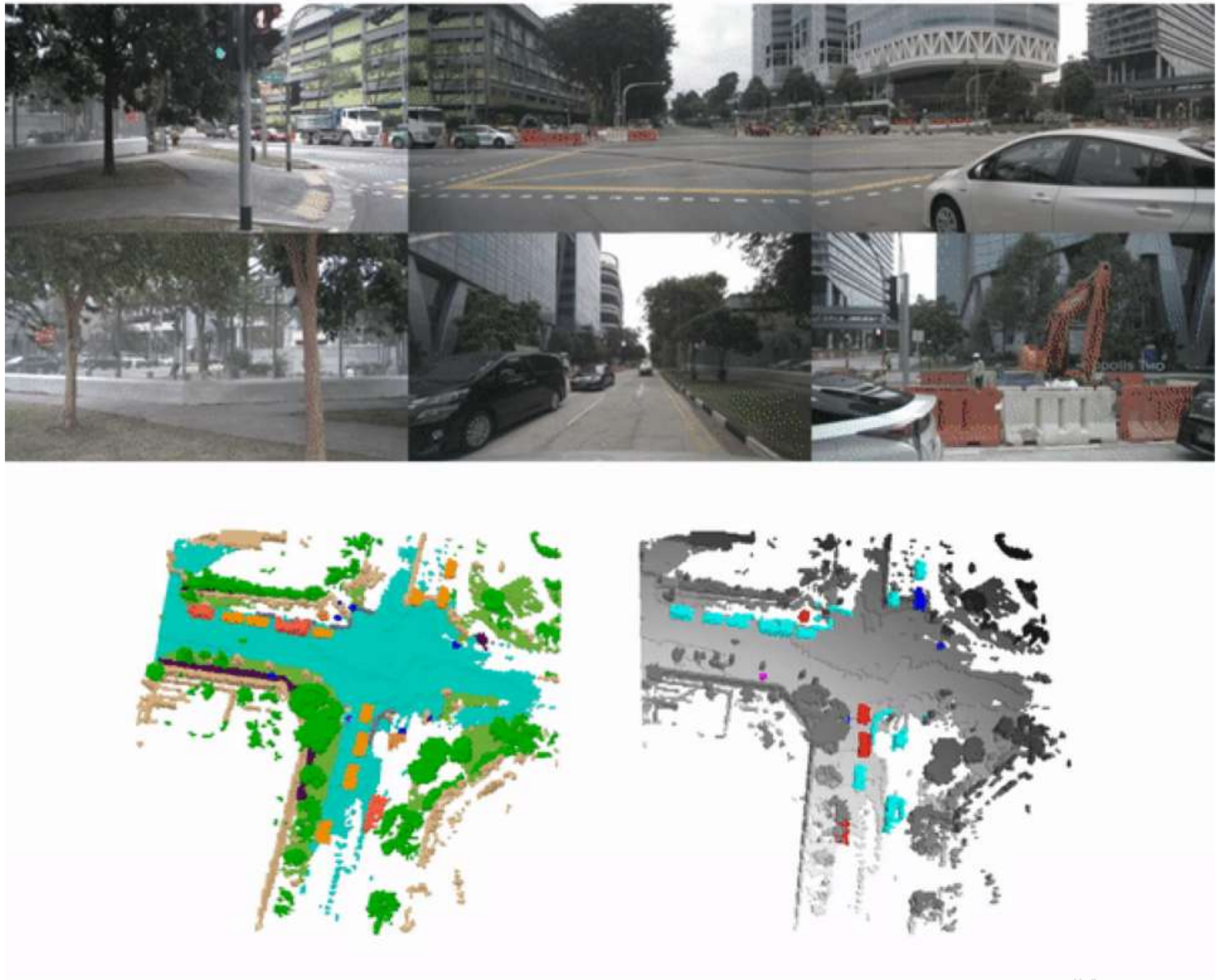
[1] The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen),
[2] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen),
[3] Huawei Noah's Ark Lab

(AAAI 2024)

Background

Vision-based 3D Occupancy Prediction:

- Inputs: Multi-view camera images
- Outputs: 3D semantic occupancy
- Advantages: cost-effectiveness, general object representation, suitable for unified models
- ...



Challenges

Challenges of Vision-based 3D Occupancy Prediction:

- Lack of geometric priors;
- 2D to 3D transformations;
- Semantic complete 3D scene details perception;
- ...

Related Work

Three Typical Solutions:

- Forward projection methods;
- Backward projection methods;
- Forward-Backward projection methods;

Related Work

Three Typical Solutions:

- Forward projection methods;

  LSS-based

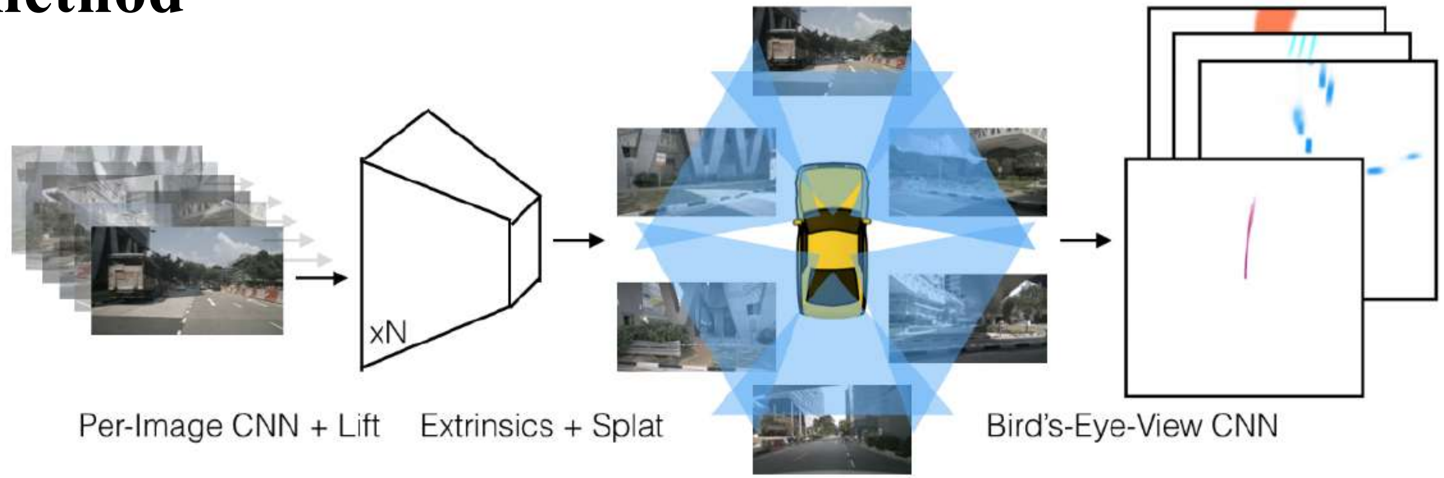
- Backward projection methods;

- Forward-Backward projection methods;

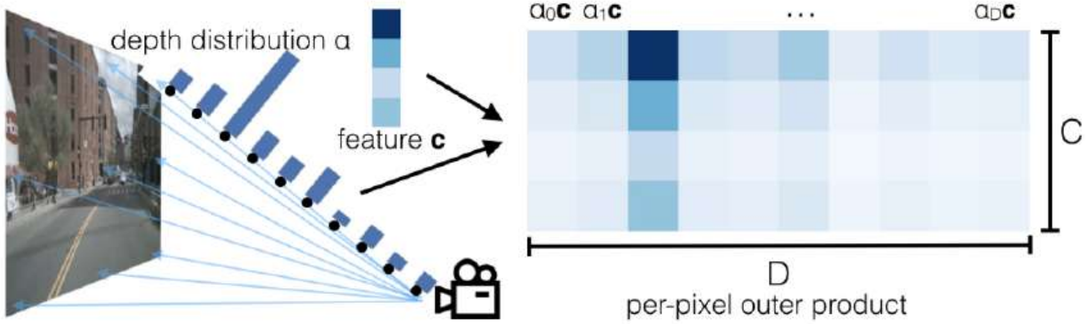
Related Work

LSS-based method

Framework



Lifting



Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d.

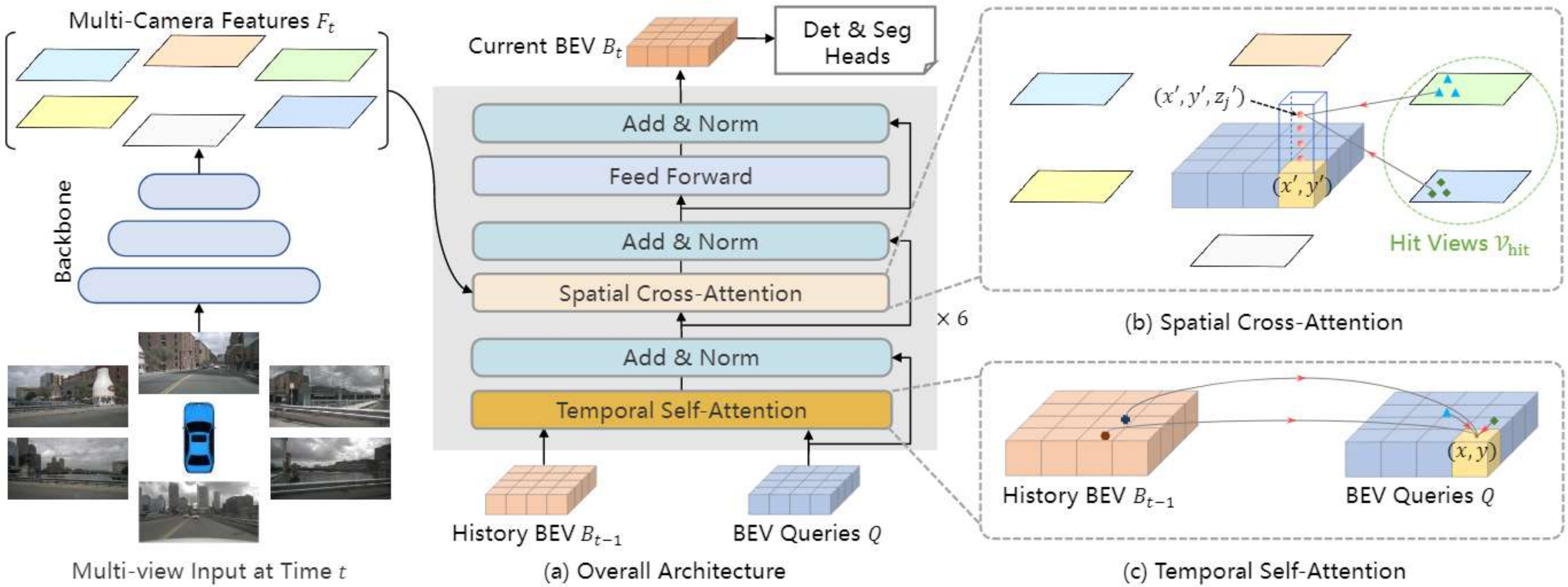
Related Work

Three Typical Solutions:

- Forward projection methods;
  - *LSS-based*
- Backward projection methods;
  - *BEVFormer*
- Forward-Backward projection methods;

Related Work

BEVFormer



Bevformer: Learning bird's-eye-view representation from multi-camera images via spatio-temporal transformers.

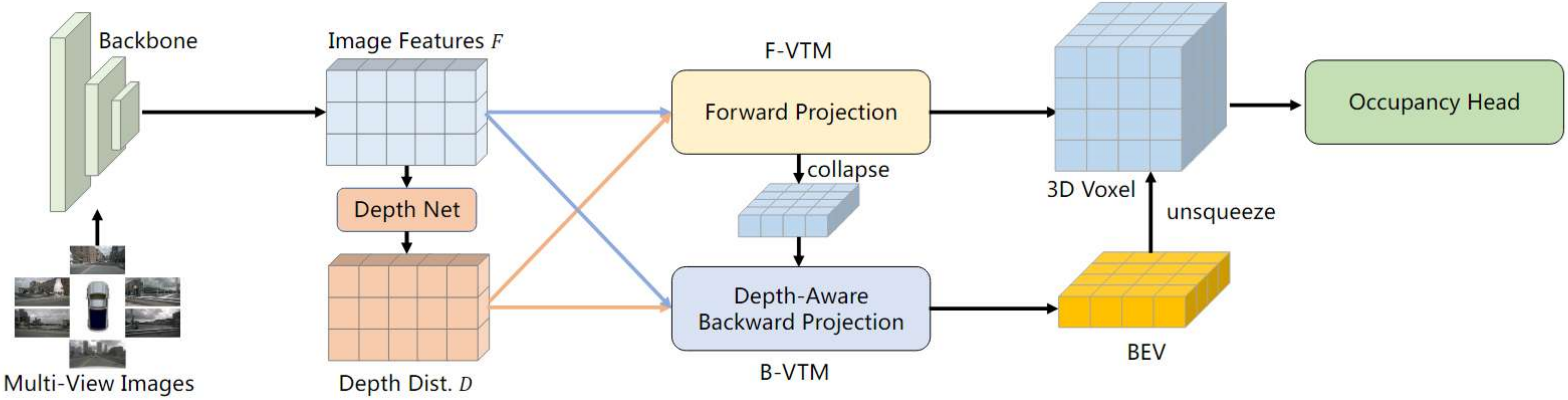
Related Work

Three Typical Solutions:

- Forward projection methods;
  - *LSS-based*
- Backward projection methods;
  - *BEVFormer*
- Forward-Backward projection methods;
  - *FB-OCC*

Related Work

FB-OCC



FB-OCC: 3D Occupancy Prediction based on Forward-Backward View Transformation.

Related Work

Issues:

- More and more complex model structure;
- Heavy model parameters;
- Long-term training;
- ... ...

Motivation

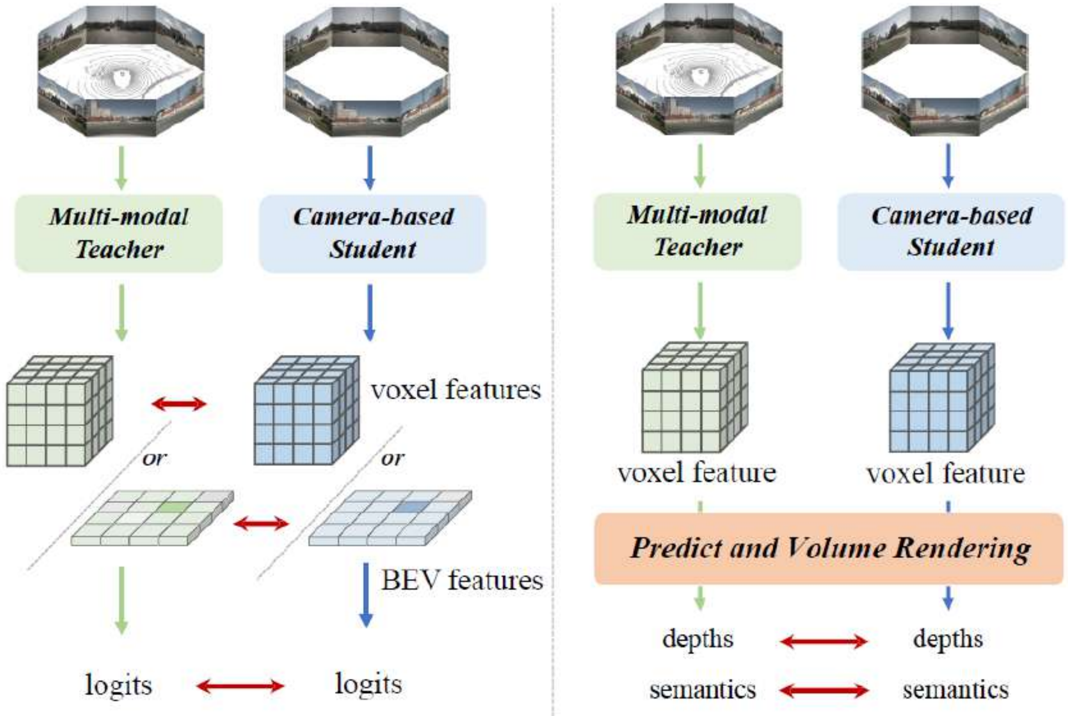
Can we enhance existing models benefiting from knowledge distillation?

Strengths:

- Knowledge distillation could improve student model while do not introduce burden during inference;
- Multi-modality model tend to achieve high performance more easily;

Motivation

Can we enhance existing models benefiting from knowledge distillation?



(a) Voxel / BEV / logits Distillation (b) Rendering Assisted Distillation

- Simply align the features or logits do not obtain satisfied results;
- We need to explore more effective knowledge distillation paradigm for 3D occupancy representation;

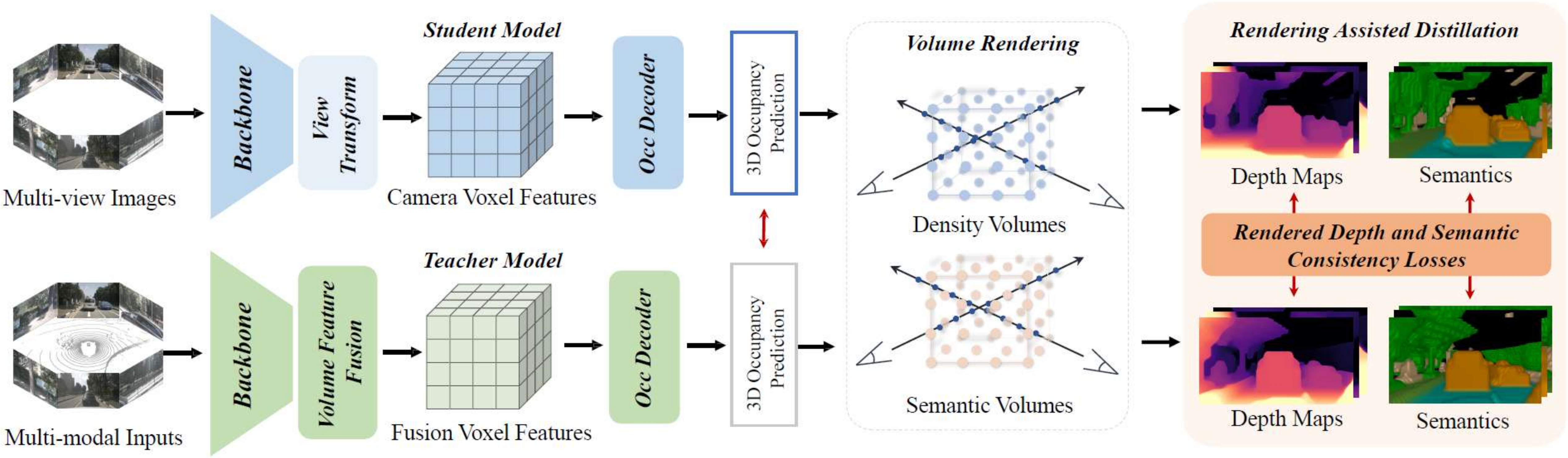
RadOcc

Proposed Method

RadOcc, Rendering **a**ssisted **d**istillation paradigm for 3D **Occ**upancy prediction.

RadOcc

Framework



- Two models: teacher model and student model;
- Teacher model takes multi-modality data (images and LiDAR point cloud) as inputs;
- Student model can be any of a existing vision-based occupancy prediction model;

RadOcc

Volume Rendering

Voxel-based volume rendering

Algorithm 1: The pseudocode of volume rendering.

Input: V^D, V^S, K, T, H, W, step_size
Output: D, S
 # get ray origin and direction of each pixel
$\text{rays}_o, \text{rays}_d \leftarrow$ **get_rays**$(K,\ T,\ H,\ W)$
 # get sampled points on each ray
$\mathcal{P} \leftarrow$ **get_points**$(\text{rays}_o,\ \text{rays}_d,\ \text{step\_size})$
 # get the distance between the sampled point and ray origin
dist $\leftarrow$ **get_distance**$(\text{rays}_o,\ \text{rays}_p)$
 # inject density and semantic on each sample point
$\mathcal{P}^D \leftarrow$ **grid_sample**$(V^D,\ \mathcal{P})$
$\mathcal{P}^S \leftarrow$ **grid_sample**$(V^S,\ \mathcal{P})$
 # calculate interval of each sampled point pair
delta $\leftarrow$ dist[..., 1 :] − dist[..., : −1]
 # Eqn. (1)-(3) in manuscript
$D, S \leftarrow$ **Render**$(\mathcal{P}^D,\ \mathcal{P}^S,\ \text{delta},\ \text{dist})$
return D, S

1) Accumulated transmittance:

$$T_i = \exp(\sum_{j=1}^{i-1} \sigma(p_j)\delta_j),$$

2) Depth rendering:

$$\hat{d}(u,v) = \sum_{i=1}^{N_p} T_i(1-\exp(-\sigma(p_i)\delta_i))d(p_i),$$

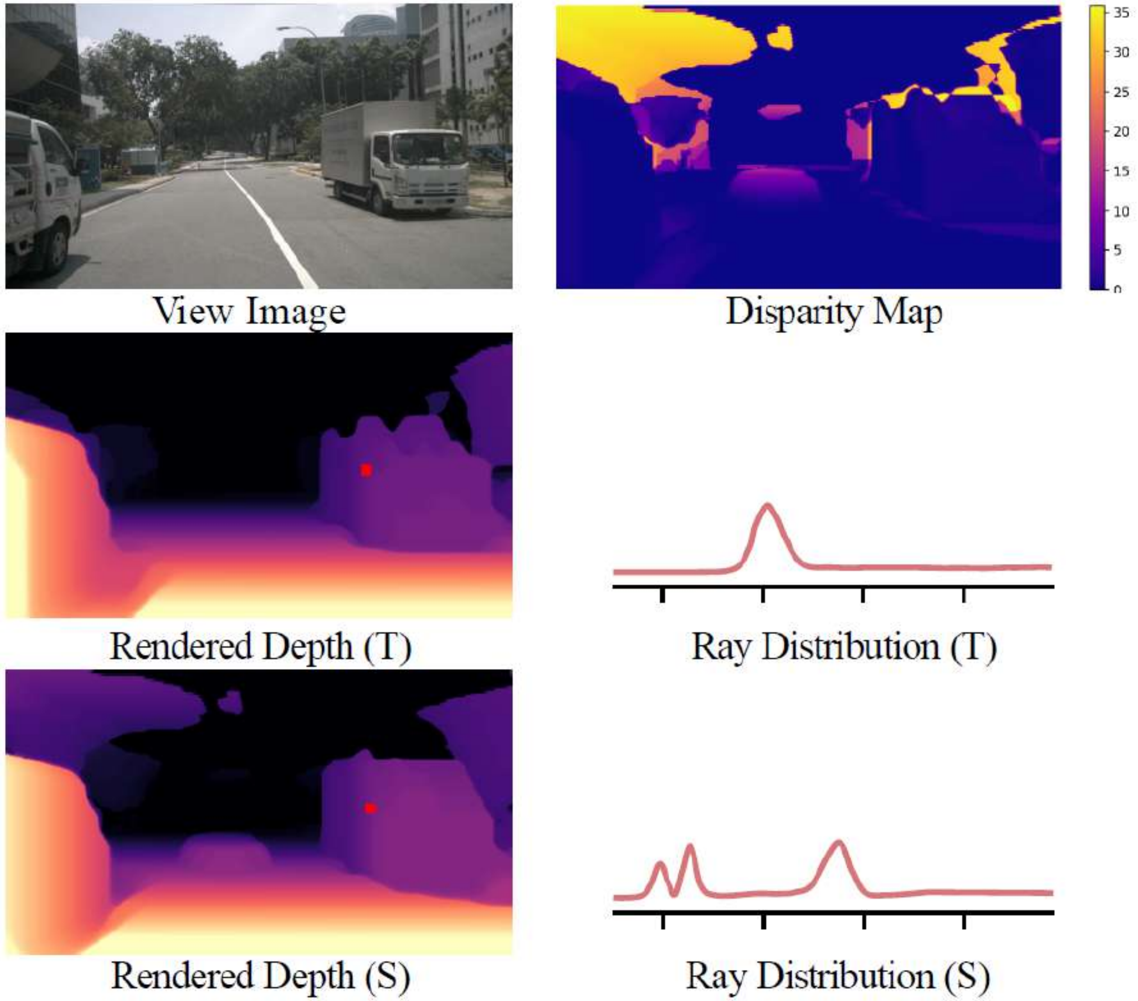
3) Semantic rendering:

$$\hat{s}(u,v) = \sum_{i=1}^{N_p} T_i(1-\exp(-\sigma(p_i)\delta_i))s(p_i),$$

RadOcc

Rendered Depth Consistency (RDC)

Ray termination distribution aligning



1) Findings:

- Directly align depth maps is a hard constraint;
- Similar depths between teacher and student models show great discrepancy in ray distribution;

2) RDC Loss:

$$\mathcal{R}^{(\cdot)}_{(u,v)} = \{T_i(1-\exp(-\sigma(p_i)\delta_i))\}_{i=1}^{N_p},$$

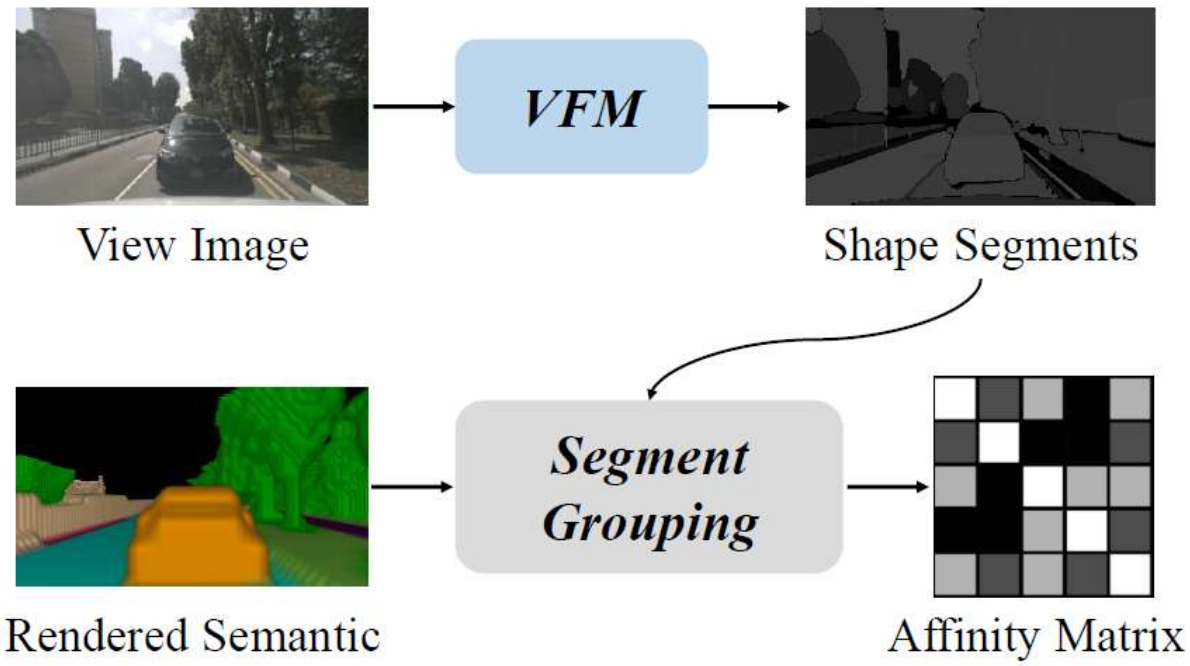
$$\mathcal{L}_{rdc} = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W}\boxed{\mathcal{D}_{KL}}(\mathcal{R}^{\text{teacher}}_{(u,v)}||\mathcal{R}^{\text{student}}_{(u,v)})$$

KL divergence

RadOcc

Rendered Semantic Consistency (RSC)

Segment-guided affinity distillation (SAD)



1) Pipeline:

- Utilizing **VFM** (i.e. SAM) to segment view images;
- **Grouping** the rendered semantic logits based on the segmentation patches;
- Applying **average pooling** within each group to extract *M* **semantic embeddings**;
- Computing affinity matrix for student and teacher model based on the semantic embeddings:

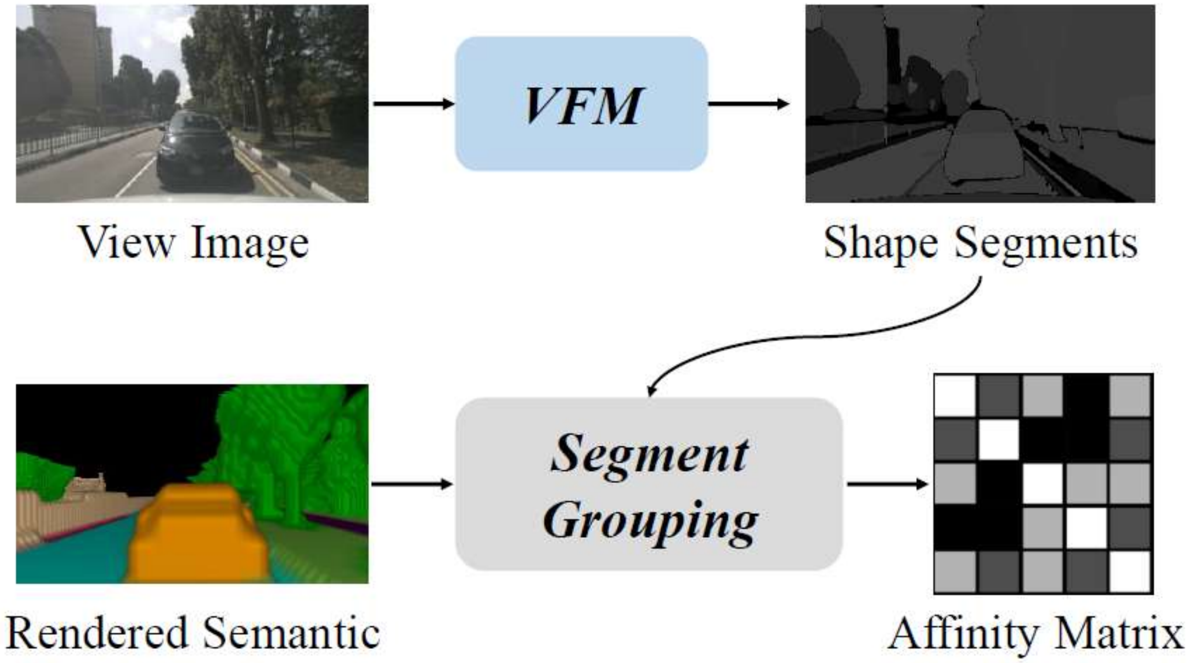
$$\mathcal{C}_{i,j,r} = \frac{\boxed{\mathcal{E}}(i,r), \mathcal{E}(j,r)}{||\mathcal{E}(i)||_2||\mathcal{E}(j)||_2}.$$

Semantic embedding with shape (M, C)

RadOcc

Rendered Semantic Consistency (RSC)

Segment-guided affinity distillation (SAD)



2) RSC Loss:

$$\mathcal{L}_{sad} = \sum_{r=1}^{C} \sum_{i=1}^{M} \sum_{j=1}^{M} ||C^T_{i,j,r} - C^S_{i,j,r}||^2_2,$$

$$\mathcal{L}_{rsc} = \mathcal{L}_{sad}/CM^2 + \omega \boxed{\mathcal{D}_{KL}(\mathcal{S}^T||\mathcal{S}^S)}$$

KL divergence between rendered semantics

Results

Datasets

- Occ3D: dense 3D occupancy prediction dataset;
- nuScenes-lidarseg: sparse LiDAR semantic segmentation prediction dataset;

Experimental Settings

- **Dense prediction**: BEVDet as baseline, Swin-Transformer base as image backbone;
- **Sparse prediction**: TPVFormer as baseline, R101-DCN as image backbone;

Results

3D occupancy prediction performance on the Occ3D

Method	Image Backbone	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Performances on Validation Set																			
MonoScene	R101-DCN	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
CTF-Occ	R101-DCN	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
BEVFormer	R101-DCN	39.24	10.13	47.91	24.90	47.57	54.52	20.23	28.85	28.02	25.73	33.03	38.56	81.98	40.65	50.93	53.02	43.86	37.15
PanoOcc	R101-DCN	42.13	11.67	50.48	29.64	49.44	55.52	23.29	**33.26**	30.55	30.99	34.43	42.57	**83.31**	44.23	54.40	56.04	45.94	40.40
BEVDet†	Swin-B	42.02	12.15	49.63	25.10	52.02	54.46	27.87	27.99	28.94	27.23	36.43	42.22	82.31	43.29	54.62	57.90	48.61	43.55
Baseline (ours)	Swin-B	44.14	**13.39**	52.20	**31.43**	52.01	56.70	**30.66**	32.95	31.56	**31.31**	39.87	44.64	82.98	**44.97**	**55.43**	**58.90**	48.43	42.99
RadOcc (ours)	Swin-B	**46.06**	9.78	**54.93**	20.44	**55.24**	**59.62**	30.48	28.94	**44.66**	28.04	**45.69**	**48.05**	81.41	39.80	52.78	56.16	**64.45**	**62.64**
Teacher (ours)	Swin-B	49.38	10.93	58.23	25.01	57.89	62.85	34.04	33.45	50.07	32.05	48.87	52.11	82.9	42.73	55.27	58.34	68.64	66.01
Performances on 3D Occupancy Prediction Challenge																			
BEVFormer	R101-DCN	23.70	10.24	36.77	11.70	29.87	38.92	10.29	22.05	16.21	14.69	27.44	33.13	48.19	33.10	29.80	17.64	19.01	13.75
SurroundOcc†	R101-DCN	42.26	11.7	50.55	32.09	41.59	57.38	27.93	38.08	30.56	29.32	48.29	38.72	80.21	48.56	53.20	47.56	46.55	36.14
BEVDet†	Swin-B	42.83	18.66	49.82	31.79	41.90	56.52	26.74	37.31	30.01	31.33	48.18	38.59	80.95	50.59	53.87	49.67	46.62	35.62
PanoOcc-T⋆	Intern-XL	47.16	**23.37**	50.28	36.02	47.32	59.61	31.58	39.59	34.58	33.83	52.25	43.29	**83.82**	**55.81**	**59.41**	53.81	53.48	43.61
Baseline-T (ours)	Swin-B	47.74	22.88	50.74	**41.02**	**49.39**	55.40	33.41	45.71	38.57	35.79	48.94	44.40	83.19	52.26	59.09	**55.83**	51.35	43.54
RadOcc-T (ours)	Swin-B	**49.98**	21.13	**55.17**	39.31	48.99	**59.92**	**33.99**	**46.31**	**43.26**	**39.29**	**52.88**	**44.85**	83.72	53.93	59.17	55.62	**60.53**	**51.55**
Teacher-T (ours)	Swin-B	55.09	25.94	59.04	44.93	57.95	63.70	38.89	52.03	53.21	42.16	59.90	50.45	84.79	55.70	60.83	58.02	67.66	61.40

- † denotes the performance reproduced by official codes;
- * means the results provided by authors;
- '-T' represents results through test-time augmentation (TTA);

Results

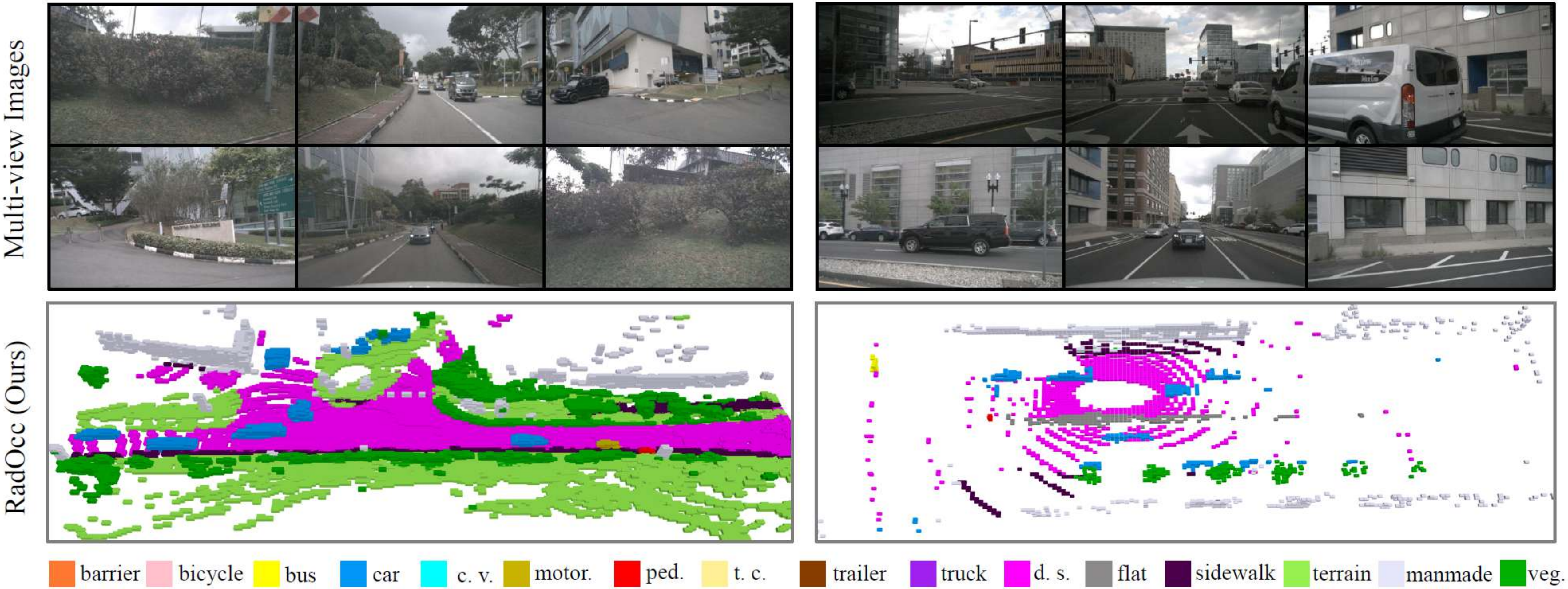
LiDAR semantic segmentation results on nuScenes test benchmark

Method	Input Modality	Image Backbone	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
PolarNet	LiDAR	-	69.4	72.2	16.8	77.0	86.5	51.1	69.7	64.8	54.1	69.7	63.5	96.6	67.1	77.7	72.1	87.1	84.5
Cylinder3D	LiDAR	-	77.2	82.8	29.8	84.3	89.4	63.0	79.3	77.2	73.4	84.6	69.1	97.7	70.2	80.3	75.5	90.4	87.6
2DPASS	LiDAR	-	80.8	81.7	55.3	92.0	91.8	73.3	86.5	78.5	72.5	84.7	75.5	97.6	69.1	79.9	75.5	90.2	88.0
TPVFormer	Camera	R50-DCN	59.2	65.6	15.7	75.1	80.0	48.8	43.1	44.3	26.8	72.8	55.9	92.3	53.7	61.0	59.2	79.7	75.6
BEVDet†	Camera	Swin-B	65.2	31.3	**63.9**	74.6	79.1	51.5	59.8	63.4	56.2	74.7	59.8	92.8	61.4	69.5	65.7	84.1	82.9
TPVFormer (BL)	Camera	R101-DCN	69.4	**74.0**	27.5	**86.3**	85.5	**60.7**	68.0	62.1	49.1	**81.9**	68.4	94.1	59.5	66.5	63.5	83.8	79.9
RadOcc (ours)	Camera	R101-DCN	**71.8**	49.1	34.2	84.5	**85.8**	59.2	**70.3**	**71.4**	**62.5**	79.7	**69.0**	**95.4**	**66.2**	**75.1**	**72.0**	**87.4**	**86.0**
Teacher (ours)	Cam+Li	R101-DCN	75.2	62.7	33.2	88.7	88.8	64.6	78.1	74.1	65.0	83.1	72.2	96.5	68.3	77.6	74.4	88.7	87.1

- † denotes the performance reproduced by official codes;

Results

Visualizations



(a) Dense 3D Occupancy Prediction

(b) Sparse 3D Occupancy Prediction

Ablation

Ablation Study on Occ3D

Method	RDC(-)	RDC	SAD	RSC	mIoU
BEVDet					36.10
Model A	✓				35.08
Model B		✓			36.76
Model C			✓		37.13
Model D				✓	37.42
RadOcc (ours)		✓		✓	**37.98**

- RDC(-) denotes directly aligning the rendered depth map with Scale-Invariant Logarithmic loss;

- *Latent ray distribution alignment is useful.*
- *Sorely aligning depth maps is not a good choice.*
- *RSC loss (including SAD and KL divergence) obviously improve performance.*

Discussion

How about different kinds of knowledge distillation paradigms?

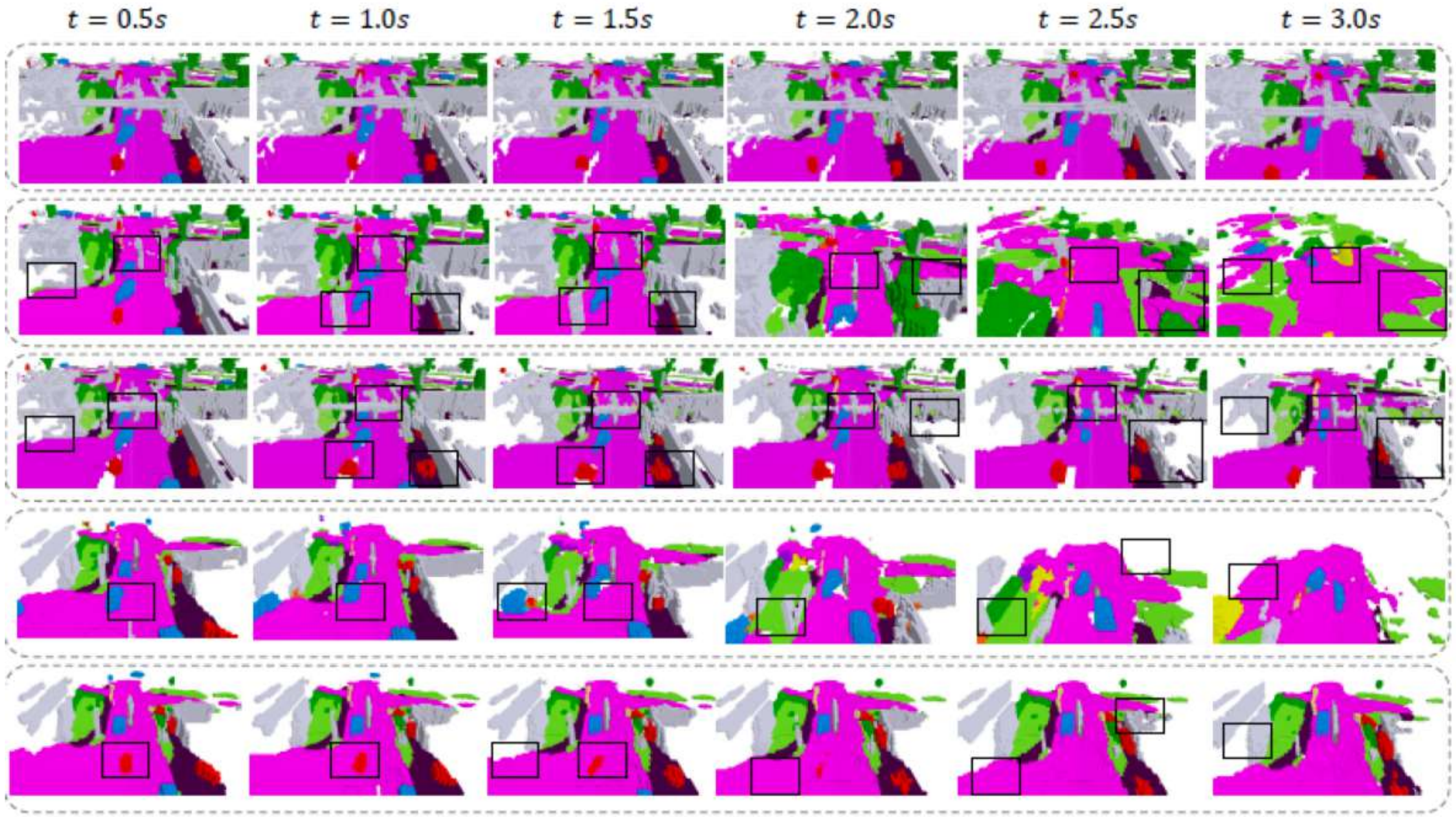
Method	Consistency	mIoU	Gains
BEVDet (baseline)	-	36.10	-
Hinton *et al.*	Prob.	37.00	+0.90
Hinton *et al.*	Feature	35.89	-0.21
BEVDistill	Prob. + Feature	35.95	-0.15
RadOcc (ours)	Render	37.98	+1.88
RadOcc (ours)	Prob. + Render	**38.53**	+2.43

- *Sorely align the occupancy probabilistic logits improve performance slightly.*
- *Aligning the volume feature hinder the performance.*
- *Our rendering-assisted knowledge distillation, combined with logits alignment enhance performance a lot.*

Conclusion

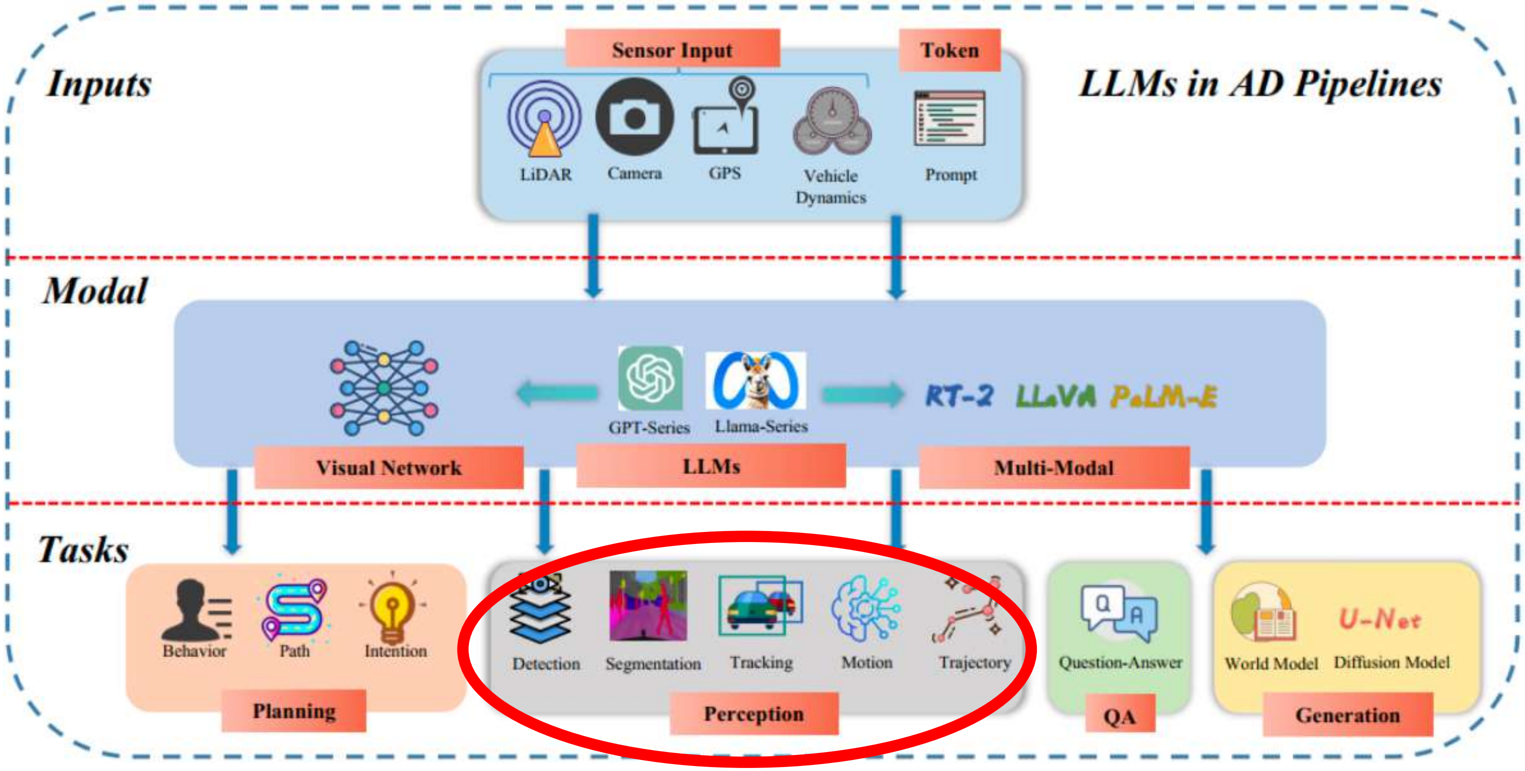
RadOcc: Learning Cross-Modality Occupancy Knowledge through Rendering Assisted Distillation:

- We propose a rendering assisted distillation paradigm, **RadOcc**, for 3D occupancy prediction;
- Two novel **consistency losses** are introduced to achieve better alignment between the rendered outputs;
- The proposed RadOcc achieves **state-of-the-art** performance on the **Occ3D** and **nuScenes** benchmarks for dense and sparse occupancy prediction, respectively.



World model based 4D occupancy forecasting and planning prediction

Outline



Enhanced 3D Perception for AD

- Monocular (front/ego view)/Multimodality 3D Lane Detection
- 3D/4D Occupancy (world model)
- **Semantic Segmentation and Semantic Completion**

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds

Xu Yan[1†], Jiantao Gao[2†], Chaoda Zheng[1†],
Chao Zheng[3], Ruimao Zhang[1], Shuguang Cui[1], Zhen Li[1⋆]

[1]The Chinese University of Hong Kong (Shenzhen), The Future Network of Intelligence Institute, Shenzhen Research Institute of Big Data,
[2]Shanghai University, [3]Tencent Map, T Lab

2DPASS

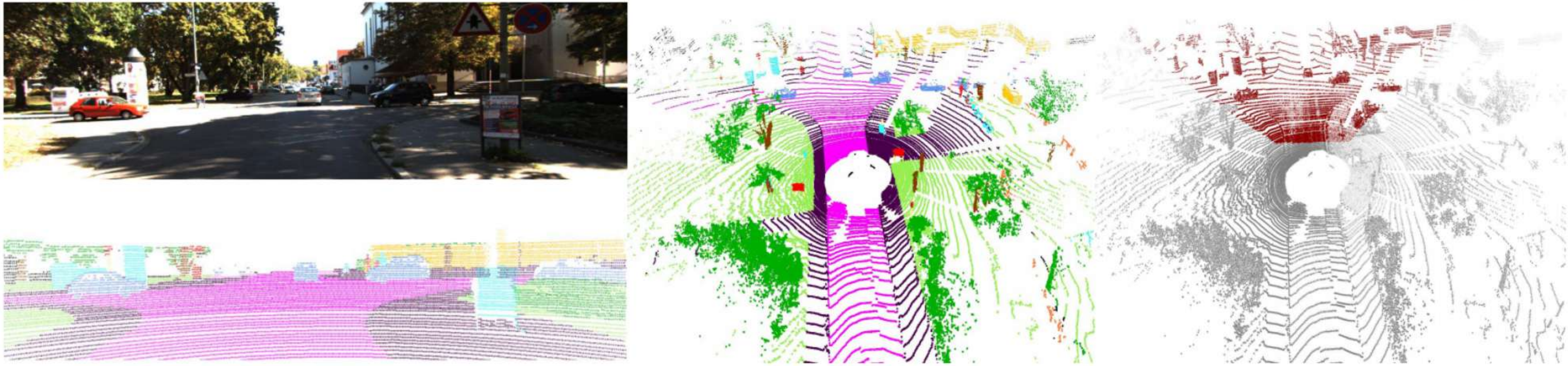
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Motivation

Front-Camera Image and Perspective Projection

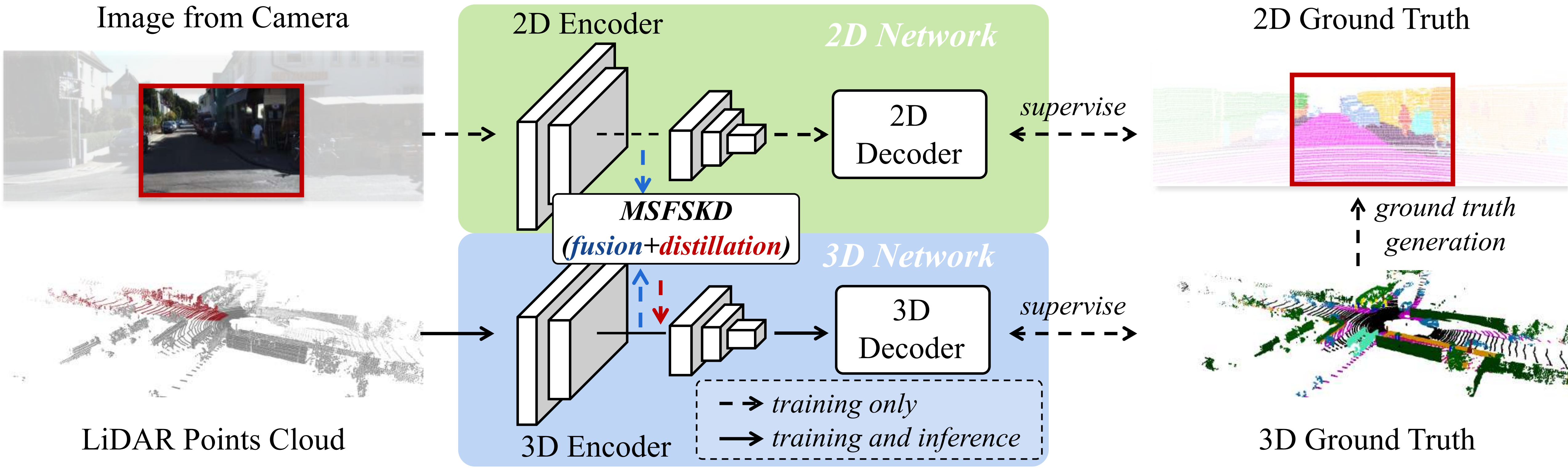
360° LiDAR Point Cloud

Point Cloud in Camera Perspective



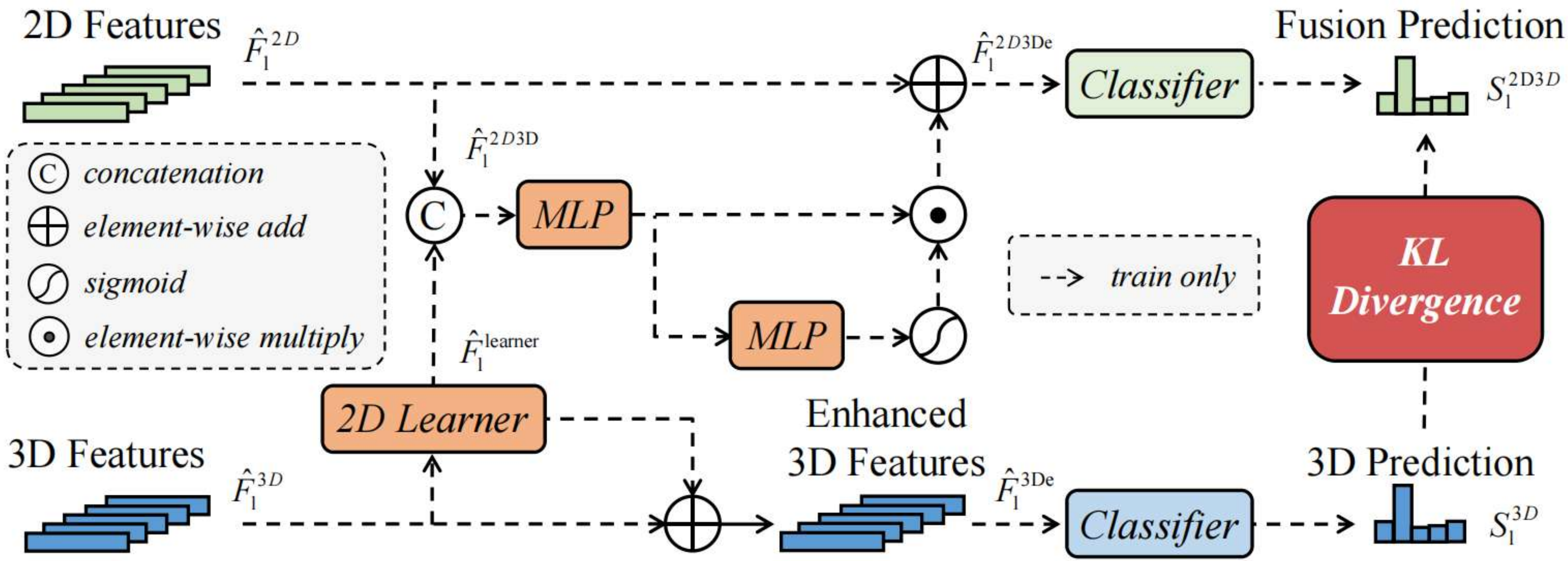
2DPASS

Architecture



2DPASS

Multi-Scale Fusion-to-Single Knowledge Distillation (MSFSKD)



2DPASS

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

SemanticKITTI-SingleScan

Results

#	User	Entries	Date of Last Entry	mIoU	accuracy	road	sidewalk	parking	other-ground	building	car	car (moving)	truck	truck (moving)	bicycle	motorcycle	other-vehicle	other-vehicle (moving)	vegetation	trunk	terrain	person	person (moving)	bicyclist	bicyclist (moving)	motorcyclist	motorcyclist (moving)	fence	pole	traffic-sign
1	Point-2DPASS	2	02/28/22	0.729 (1)	0.917 (14)	0.897 (159)	0.747 (84)	0.674 (19)	0.400 (4)	0.935 (4)	0.970 (12)	- (-)	0.611 (1)	- (-)	0.636 (12)	0.634 (10)	0.615 (5)	- (-)	0.862 (11)	0.739 (5)	0.710 (18)	0.779 (5)	- (-)	0.813 (5)	- (-)	0.741 (3)	- (-)	0.729 (3)	0.650 (7)	0.704 (10)
2	SVQNet	1	11/16/21	0.716 (2)	0.927 (3)	0.922 (13)	0.789 (6)	0.716 (5)	0.354 (10)	0.936 (3)	0.972 (5)	- (-)	0.536 (14)	- (-)	0.653 (9)	0.562 (29)	0.615 (3)	- (-)	0.873 (2)	0.762 (1)	0.721 (2)	0.789 (3)	- (-)	0.743 (16)	- (-)	0.528 (13)	- (-)	0.724 (4)	0.682 (2)	0.727 (3)
3	Point-Voxel-KD	9	11/18/21	0.712 (3)	0.921 (8)	0.918 (27)	0.775 (12)	0.709 (8)	0.410 (2)	0.924 (16)	0.970 (13)	- (-)	0.535 (15)	- (-)	0.679 (4)	0.693 (4)	0.602 (8)	- (-)	0.865 (6)	0.738 (6)	0.719 (5)	0.751 (12)	- (-)	0.735 (19)	- (-)	0.505 (16)	- (-)	0.694 (13)	0.649 (9)	0.658 (39)
4	PV-KD	3	11/20/21	0.711 (4)	0.921 (7)	0.918 (20)	0.777 (10)	0.714 (6)	0.406 (3)	0.923 (18)	0.970 (11)	- (-)	0.509 (24)	- (-)	0.681 (3)	0.686 (6)	0.594 (10)	- (-)	0.865 (5)	0.740 (4)	0.720 (4)	0.758 (10)	- (-)	0.745 (14)	- (-)	0.496 (19)	- (-)	0.693 (15)	0.652 (6)	0.659 (37)
5	huanghui	6	10/28/21	0.710 (5)	0.924 (4)	0.929 (6)	0.797 (5)	0.730 (2)	0.271 (87)	0.918 (27)	0.973 (3)	- (-)	0.493 (32)	- (-)	0.735 (1)	0.721 (3)	0.585 (15)	- (-)	0.869 (3)	0.758 (2)	0.720 (3)	0.798 (2)	- (-)	0.828 (1)	- (-)	0.236 (81)	- (-)	0.685 (20)	0.700 (1)	0.751 (1)
6	AF2S3Net	2	03/21/21	0.708 (6)	0.900 (108)	0.920 (16)	0.762 (28)	0.668 (28)	0.458 (1)	0.925 (12)	0.943 (94)	- (-)	0.402 (94)	- (-)	0.630 (14)	0.814 (1)	0.400 (83)	- (-)	0.786 (192)	0.680 (77)	0.631 (196)	0.764 (8)	- (-)	0.817 (3)	- (-)	0.777 (2)	- (-)	0.696 (11)	0.640 (20)	0.733 (2)
7	DRINet_PLUSPLUS	1	11/17/21	0.707 (7)	0.917 (15)	0.898 (155)	0.746 (85)	0.662 (36)	0.301 (39)	0.923 (17)	0.969 (17)	- (-)	0.593 (2)	- (-)	0.658 (7)	0.580 (23)	0.610 (7)	- (-)	0.873 (1)	0.730 (14)	0.725 (1)	0.804 (1)	- (-)	0.827 (2)	- (-)	0.463 (22)	- (-)	0.696 (12)	0.661 (4)	0.716 (6)
8	NickForever	10	02/09/22	0.705 (8)	0.911 (28)	0.894 (176)	0.741 (121)	0.636 (86)	0.342 (14)	0.920 (22)	0.968 (21)	- (-)	0.588 (3)	- (-)	0.638 (11)	0.605 (18)	0.601 (9)	- (-)	0.860 (19)	0.725 (19)	0.712 (10)	0.780 (4)	- (-)	0.814 (4)	- (-)	0.533 (10)	- (-)	0.683 (23)	0.646 (13)	0.707 (9)
9	HRI-ADLab-HZ	3	03/17/21	0.703 (9)	0.927 (2)	0.934 (2)	0.807 (2)	0.703 (9)	0.333 (16)	0.935 (5)	0.976 (1)	- (-)	0.442 (58)	- (-)	0.684 (2)	0.687 (5)	0.611 (6)	- (-)	0.865 (4)	0.751 (3)	0.717 (8)	0.759 (9)	- (-)	0.744 (15)	- (-)	0.434 (30)	- (-)	0.721 (6)	0.648 (10)	0.614 (84)
10	GuidedContrast	5	06/17/21	0.702 (10)	0.914 (17)	0.930 (4)	0.798 (4)	0.719 (4)	0.356 (9)	0.940 (1)	0.968 (20)	- (-)	0.483 (39)	- (-)	0.655 (8)	0.620 (13)	0.505 (37)	- (-)	0.816 (151)	0.709 (36)	0.644 (178)	0.643 (52)	- (-)	0.755 (11)	- (-)	0.715 (4)	- (-)	0.750 (1)	0.647 (12)	0.691 (13)

2DPASS

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

SemanticKITTI-MultiScan

Results

#	User	Entries	Date of Last Entry	mIoU	accuracy	road	sidewalk	parking	other-ground	building	car	car (moving)	truck	truck (moving)	bicycle	motorcycle	other-vehicle	other-vehicle (moving)	vegetation	trunk	terrain	person	person (moving)	bicyclist	bicyclist (moving)	motorcyclist	motorcyclist (moving)	fence	pole	traffic-sign
1	Point-2DPASS	3	02/28/22	0.624 (1)	0.914 (8)	0.897 (67)	0.747 (31)	0.674 (11)	0.400 (2)	0.936 (3)	0.962 (3)	0.821 (5)	0.482 (5)	0.161 (7)	0.636 (11)	0.637 (7)	0.527 (3)	0.038 (31)	0.862 (8)	0.739 (6)	0.710 (9)	0.354 (1)	0.803 (4)	0.079 (15)	0.712 (5)	0.620 (2)	0.731 (4)	0.729 (1)	0.650 (7)	0.705 (8)
2	DRINet_PLUSPLUS	1	11/17/21	0.613 (2)	0.924 (2)	0.923 (6)	0.791 (2)	0.696 (5)	0.309 (10)	0.937 (2)	0.972 (1)	0.854 (1)	0.468 (6)	0.159 (8)	0.636 (12)	0.533 (21)	0.646 (1)	0.263 (4)	0.868 (3)	0.758 (2)	0.712 (4)	0.305 (4)	0.848 (1)	0.000 (71)	0.731 (2)	0.000 (66)	0.769 (2)	0.728 (2)	0.680 (2)	0.735 (1)
3	SVQNet	8	11/16/21	0.605 (3)	0.927 (1)	0.932 (1)	0.805 (1)	0.716 (1)	0.370 (3)	0.937 (1)	0.961 (5)	0.805 (7)	0.404 (24)	0.039 (23)	0.644 (9)	0.603 (11)	0.609 (2)	0.075 (22)	0.873 (2)	0.767 (1)	0.723 (2)	0.274 (6)	0.847 (2)	0.000 (71)	0.724 (4)	0.000 (68)	0.910 (1)	0.726 (3)	0.684 (1)	0.710 (5)
4	CPGNet	1	01/24/22	0.601 (4)	0.915 (7)	0.929 (2)	0.781 (4)	0.680 (9)	0.246 (41)	0.927 (6)	0.956 (8)	0.800 (9)	0.489 (4)	0.273 (1)	0.629 (14)	0.611 (9)	0.492 (4)	0.349 (1)	0.846 (31)	0.729 (8)	0.702 (14)	0.283 (5)	0.720 (7)	0.323 (1)	0.738 (1)	0.014 (38)	0.430 (18)	0.711 (4)	0.645 (9)	0.719 (2)
5	NickForever	4	02/09/22	0.589 (5)	0.909 (14)	0.894 (75)	0.740 (45)	0.636 (34)	0.342 (4)	0.920 (15)	0.955 (9)	0.804 (8)	0.420 (17)	0.145 (10)	0.631 (13)	0.617 (8)	0.490 (5)	0.067 (23)	0.859 (12)	0.725 (12)	0.707 (11)	0.321 (2)	0.737 (6)	0.084 (14)	0.675 (12)	0.357 (3)	0.571 (8)	0.683 (12)	0.646 (8)	0.707 (6)
6	PVKD	8	12/10/21	0.582 (6)	0.919 (6)	0.924 (5)	0.774 (7)	0.699 (4)	0.315 (9)	0.927 (8)	0.962 (2)	0.843 (2)	0.500 (1)	0.209 (3)	0.649 (7)	0.648 (4)	0.464 (6)	0.190 (9)	0.864 (6)	0.741 (5)	0.702 (15)	0.166 (15)	0.685 (12)	0.000 (71)	0.692 (9)	0.020 (35)	0.505 (12)	0.703 (7)	0.669 (3)	0.706 (7)
7	PV-KD	8	12/01/21	0.582 (7)	0.920 (3)	0.926 (3)	0.782 (3)	0.706 (3)	0.331 (5)	0.929 (5)	0.962 (4)	0.833 (4)	0.467 (7)	0.145 (9)	0.692 (1)	0.675 (2)	0.445 (9)	0.131 (14)	0.864 (7)	0.744 (3)	0.707 (12)	0.168 (13)	0.703 (9)	0.000 (71)	0.695 (8)	0.000 (60)	0.564 (9)	0.705 (5)	0.668 (4)	0.701 (9)
8	PVD-KD	6	12/06/21	0.577 (8)	0.919 (4)	0.922 (7)	0.779 (5)	0.707 (2)	0.276 (20)	0.927 (7)	0.960 (6)	0.835 (3)	0.492 (3)	0.161 (6)	0.665 (4)	0.673 (3)	0.438 (11)	0.133 (13)	0.866 (5)	0.743 (4)	0.711 (8)	0.162 (18)	0.701 (10)	0.000 (71)	0.697 (7)	0.044 (30)	0.473 (14)	0.698 (8)	0.667 (5)	0.691 (11)
9	Kyber_HW	8	05/18/21	0.569 (9)	0.881 (66)	0.913 (22)	0.725 (66)	0.688 (7)	0.535 (1)	0.879 (64)	0.918 (26)	0.653 (23)	0.157 (79)	0.056 (19)	0.654 (6)	0.868 (1)	0.275 (47)	0.039 (30)	0.751 (84)	0.646 (44)	0.574 (94)	0.164 (17)	0.676 (13)	0.151 (5)	0.664 (17)	0.671 (1)	0.596 (7)	0.632 (44)	0.626 (18)	0.710 (4)
10	Henry_Wang	9	11/18/21	0.548 (10)	0.906 (18)	0.917 (11)	0.761 (9)	0.655 (19)	0.242 (48)	0.901 (40)	0.938 (18)	0.714 (19)	0.442 (12)	0.105 (13)	0.593 (18)	0.599 (12)	0.381 (21)	0.299 (3)	0.849 (24)	0.709 (23)	0.701 (16)	0.137 (29)	0.618 (18)	0.231 (3)	0.622 (24)	0.234 (8)	0.151 (29)	0.635 (39)	0.622 (20)	0.640 (24)

2DPASS

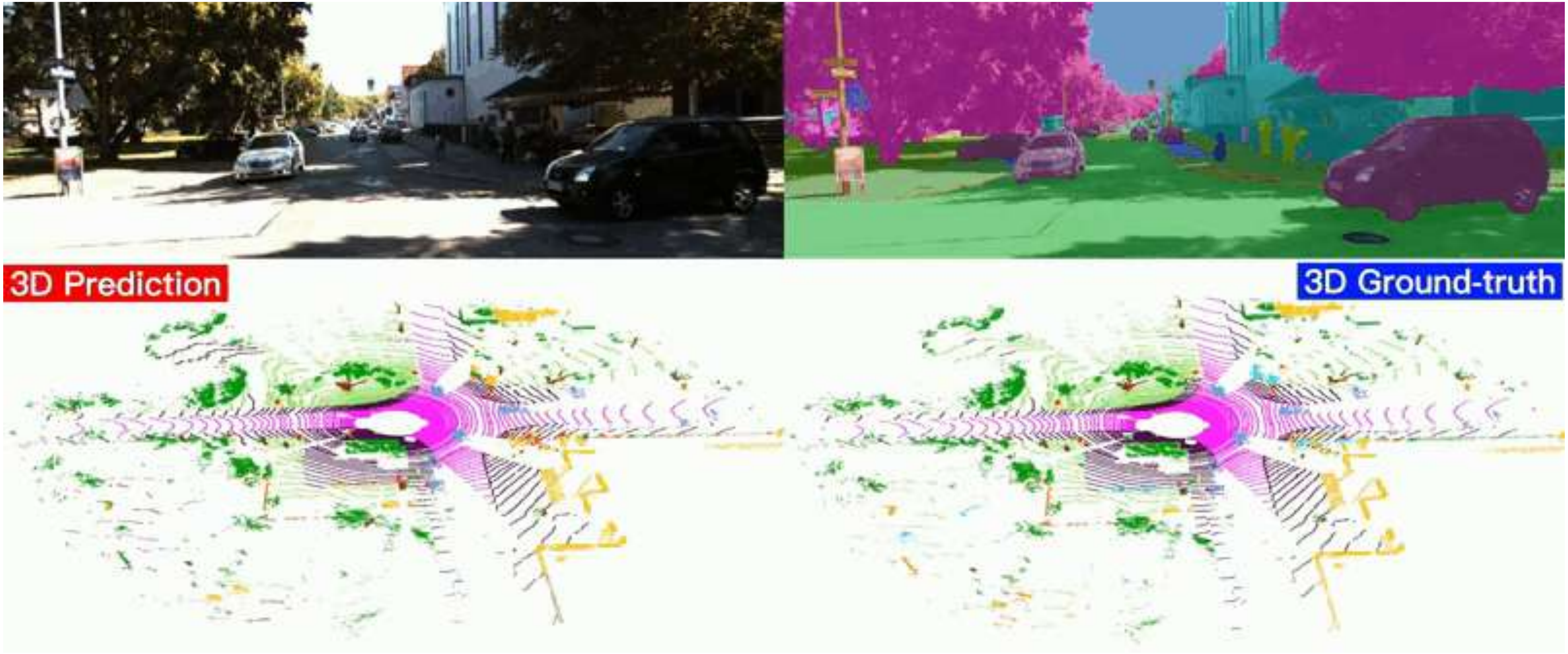
香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

NuScenes

Rank	Participant team	mIOU (↑)	barrier (↑)	bicycle (↑)	bus (↑)	car (↑)	constr_vehicle (↑)	motorcycle (↑)	pedestrian (↑)	traffic_cone (↑)	trailer (↑)
1	SVQNet (SVQNet)	0.81	0.85	0.42	0.93	0.93	0.69	0.86	0.84	0.78	0.85
2	MIT HAN Lab (SPVCNN++)	0.81	0.86	0.43	0.92	0.92	0.76	0.76	0.83	0.77	0.87
3	2DPASS	0.81	0.82	0.55	0.92	0.92	0.73	0.86	0.79	0.72	0.85
4	DRINet++ (DRINet++: Efficient Voxel-as-p)	0.80	0.86	0.43	0.90	0.92	0.65	0.86	0.83	0.73	0.84
5	Uisee-FR (GU-Net)	0.80	0.85	0.33	0.87	0.91	0.74	0.85	0.81	0.78	0.88
6	2D3DNet (2D3DNet)	0.80	0.83	0.59	0.88	0.85	0.64	0.84	0.82	0.76	0.85
7	Kyber (AF2S3Net)	0.78	0.79	0.52	0.90	0.84	0.77	0.74	0.77	0.72	0.84
8	Cylinder3D++ (Cylinder3D++)	0.78	0.83	0.34	0.84	0.89	0.70	0.79	0.77	0.73	0.85
9	RH (CPFusion)	0.78	0.84	0.37	0.89	0.86	0.70	0.77	0.78	0.75	0.83
10	MIT HAN LAB (SPVNAS)	0.77	0.80	0.30	0.92	0.91	0.65	0.79	0.76	0.71	0.81

2DPASS

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen



Challenge

ICCV2021 2nd Place Award，ECCV2022 3rd Place Award

URBAN3D

Urban3D: First Challenge on Large-Scale Point Cloud Analysis for Urban Scenes Understanding

The International Conference on Computer Vision (ICCV), October 16, 2021

presents

2nd Place Award

in

ICCV 2021 Challenge on Urban Scenes Understanding

to the Authors

X. Yan[1], Z. Li[1], C. Zheng[1], H. Zhang[1], J. Gao[2]
W. Zhou[1], Y. Liao[1], Z. Yuan[1], S. Wang[3], S. Cui[1]

The Chinese University of Hong Kong, Shenzhen[1], Shanghai University[2], Southern University of Science and Technology[3].

ICCV VIRTUAL 2021 OCTOBER 11-17 · UNIVERSITY OF OXFORD · sensat · THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學 · Newcastle University · CodaLab

URBAN3D

Urban3D: The 2nd Challenge on Large Scale Point-cloud Analysis for Urban Scenes Understanding

European Conference on Computer Vision, TEL AVIV 2022

ECCV TEL AVIV 2022

presents

3rd Place Award

in

ECCV 2022 Challenge on Urban Scenes Understanding (Semantic Segmentation Track)

to the Authors

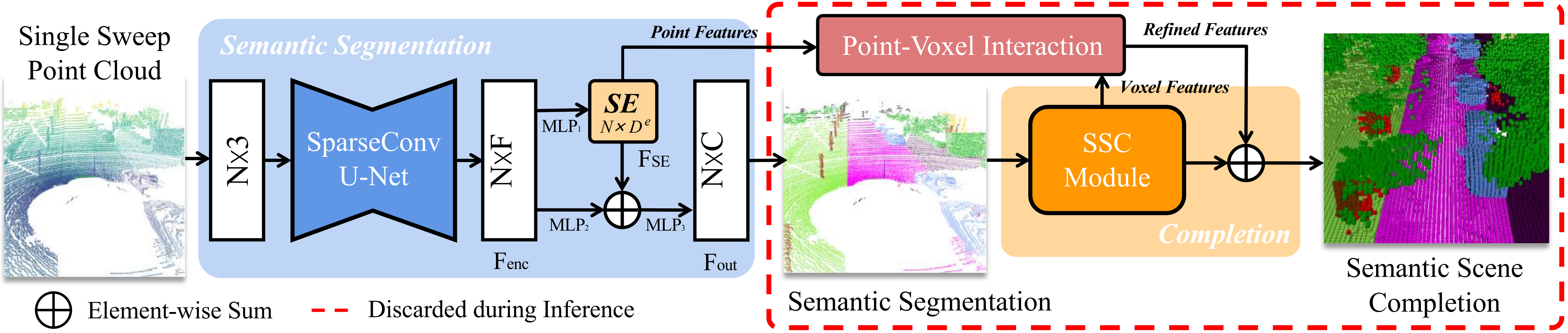
Xu Yan[1], Jiantao Gao[2], Zhuo Li[1], Zhen Li[1], Yan Peng[2], Shuguang Cui[1]

[1]The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen),
[2]Research Institute of Unmanned Surface Vehicle (USV) Engineering, Shanghai University

USC Institute for Creative Technologies · UNIVERSITY OF OXFORD · THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學 · sensat · CodaLab

JS3C-Net

Joint single sweep LiDAR point cloud Semantic Segmentation by exploiting learned shape prior form Scene Completion network



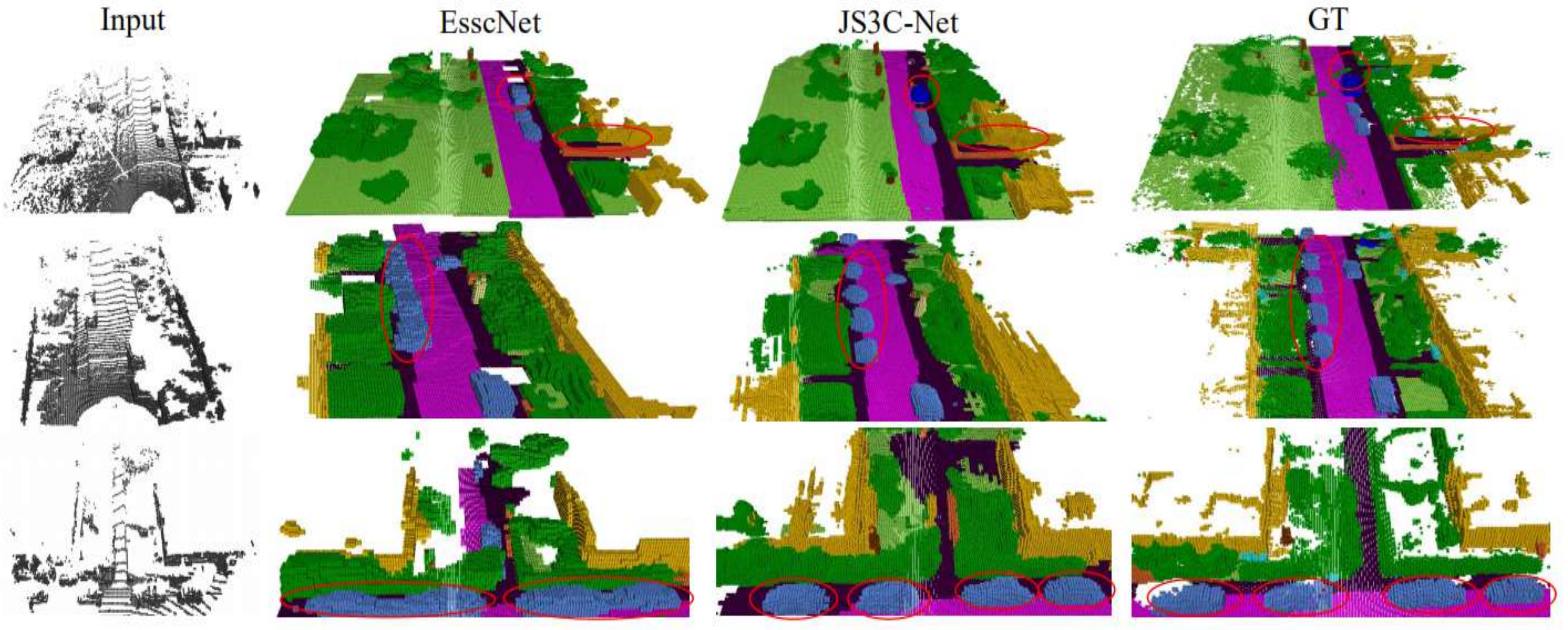
Results

Table 2: Semantic scene completion results on the *SemanticKITTI* benchmark. Only the recent published approaches are compared.

Method	Scene Completion			Semantic Scene Completion																			
	precision	recall	IoU	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic sign	mIoU
SSCNet (Song et al. 2017)	31.7	83.4	29.8	27.6	17.0	15.6	6.0	20.9	10.4	1.8	0.0	0.0	0.1	25.8	11.9	18.2	0.0	0.0	0.0	14.4	7.9	3.7	9.5
TS3D (Garbade et al. 2019)	31.6	84.2	29.8	28.0	17.0	15.7	4.9	23.2	10.7	2.4	0.0	0.0	0.2	24.7	12.5	18.3	0.0	0.1	0.0	13.2	7.0	3.5	9.5
TS3D[2] (Garbade et al. 2019; Behley et al. 2019)	25.9	**88.3**	25.0	27.5	18.5	18.9	6.6	22.1	8.0	2.2	0.1	0.0	4.0	19.5	12.9	20.2	2.3	0.6	0.0	15.8	7.6	6.7	10.2
EsscNet (Zhang et al. 2018)	62.6	55.6	41.8	43.8	28.1	26.9	10.3	29.8	26.4	5.0	0.3	5.4	9.1	35.8	**20.1**	28.7	2.9	2.7	0.1	23.3	16.4	**16.7**	17.5
TS3D[3] (Garbade et al. 2019; Behley et al. 2019; Liu et al. 2018)	**80.5**	57.7	50.6	62.2	31.6	23.3	6.5	34.1	30.7	4.9	0.0	0.0	0.1	40.1	21.9	33.1	0.0	0.0	0.0	24.1	16.9	6.9	17.7
JS3C-Net (Ours)	70.2	74.5	**56.6**	**64.7**	**39.9**	**34.9**	**14.1**	**39.4**	**33.3**	**7.2**	**14.4**	**8.8**	**12.7**	**43.1**	19.6	**40.5**	**8.0**	**5.1**	**0.4**	**30.4**	**18.9**	15.9	**23.8**

1 http://www.semantic-kitti.org/tasks.html

Results



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

AAAI
Association for the Advancement of Artificial Intelligence

Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion

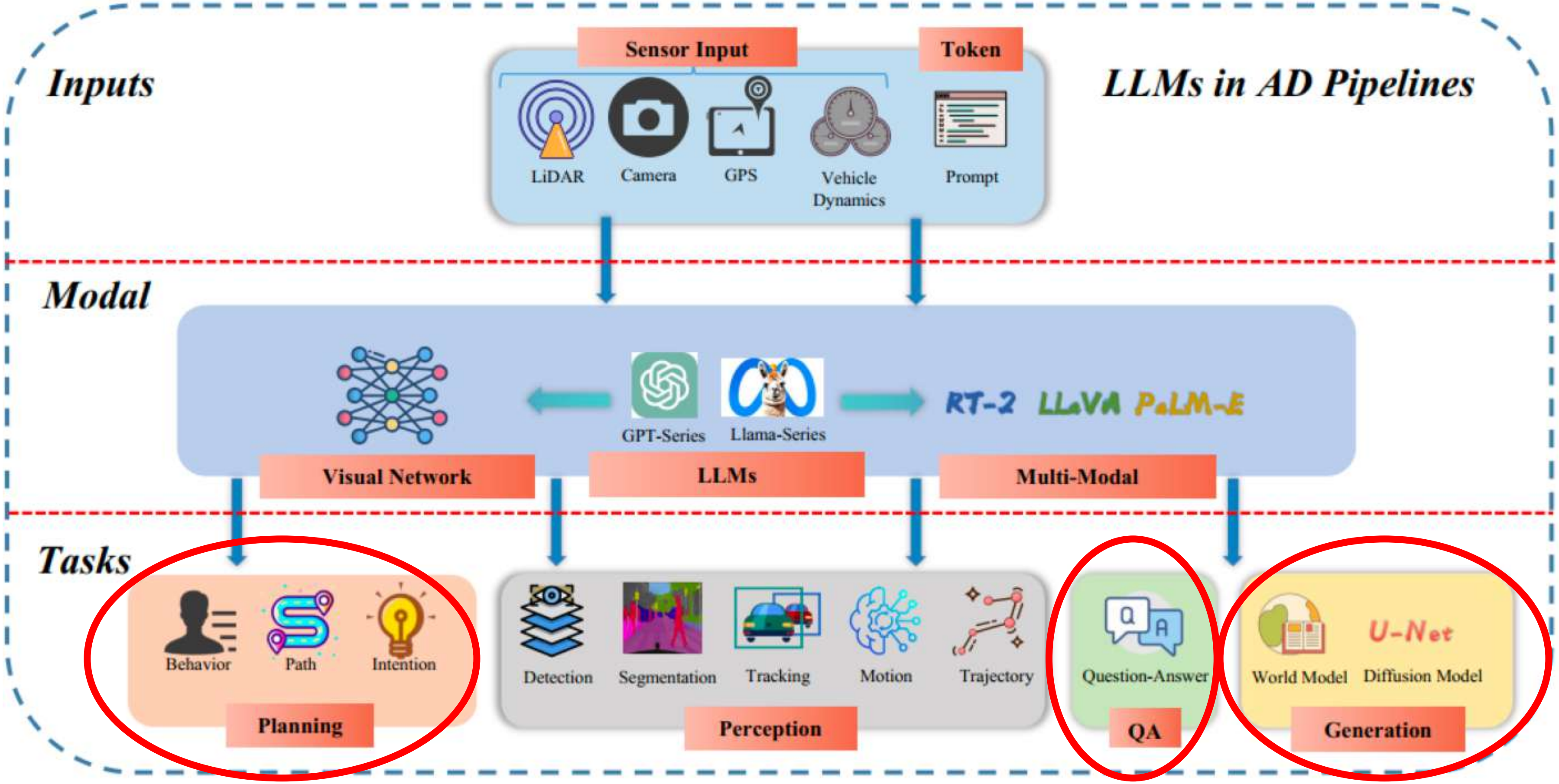
Thanks for watching !

GitHub

Xu Yan [1,2,†], *Jiantao Gao* [2,4,†], *Jie Li* [1,3], *Ruimao Zhang* [1,2]
Zhen Li [1,2,*], *Rui Huang* [1,3], *and Shuguang Cui* [1,2]

[1] The Chinese University of Hong Kong (Shenzhen),
[2] Shenzhen Research Institute of Big Data (SRIBD),
[3] Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS),
[4] Shanghai University

Outline



Enhanced 3D Reasoning

- ➢ Visual programming for open-world grounding using LLM
- ➢ 3D VQA without data bias

Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding

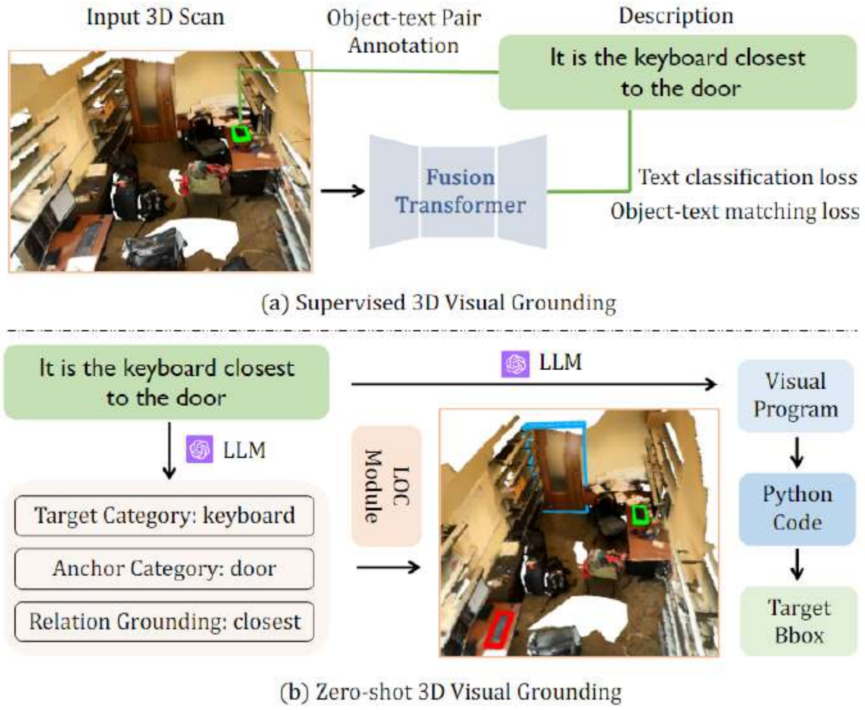
Zhihao Yuan [1,2], Jinke Ren [1,2], Chun-Mei Feng[4], Hengshuang Zhao[3], Shuguang Cui [2,1],
***Zhen Li** [2,1†]*

1 The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen),
2 School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen),
3 The University of Hong Kong
4 IHPC, A*STAR, Singapore

(CVPR 2024)

3D Zero-shot Grounding

- **Problems**
  - **Explore zero-shot 3DVG using LLMs, without the need of object-text pair annotation**
  - **Solve spatial relationships between objects explicitly**



Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

3D Zero-shot Grounding

- **Dialog with LLM: A Naive Approach**
  - **Detect all objects in the scene, describe object's location and size**
  - **Given the text prompt to ChatGPT to find the correct object**

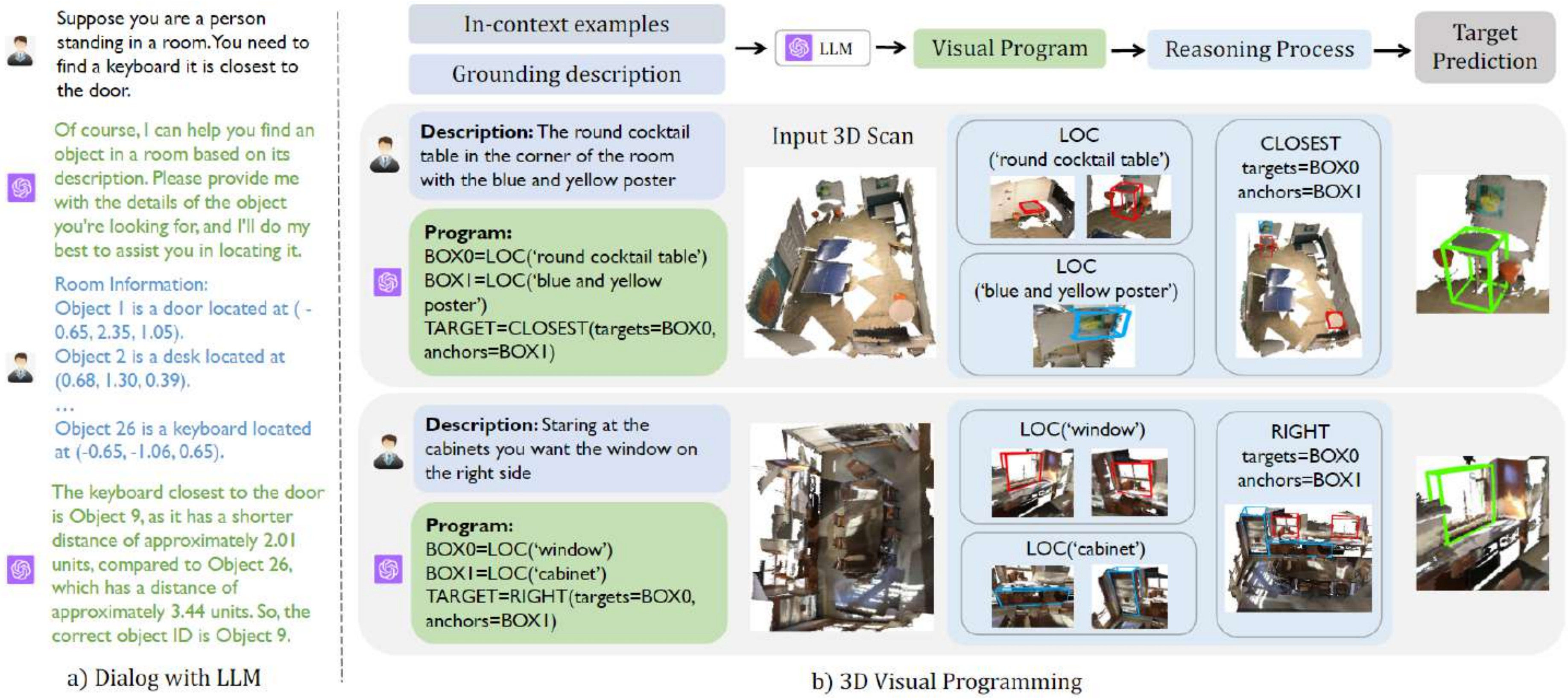
Query: This is a storage shelf. It is white in color and is above a desk with a chair in front of it.

Answer: Based on the description, we are looking for a storage shelf that is white in color and is above a desk with a chair in front of it.
Looking at the room information, we can see that there are several objects identified as shelves, chairs, and desks. However, the description also mentions that the shelf is above a desk, which means we need to look at the coordinates to determine the vertical positioning of these objects.
The objects identified as shelves are: <obj 24>, <obj 75>, and <obj 76>. The objects identified as chairs are: <obj 1>, <obj 36>, <obj 77>. The objects identified as desks are not explicitly mentioned, but we can consider objects like tables or nightstands as potential desks: <obj 18>, <obj 58>.
Now, we need to find a combination of these objects where the shelf is above a desk and there is a chair in front of it. This requires the shelf to have a higher z-coordinate (vertical position) than the desk, and the chair to have similar x and y coordinates (horizontal position) as the desk.
After analyzing the coordinates, none of the objects seem to fit the description perfectly. However, the closest match could be <obj 24> (shelf), <obj 36> (chair), and <obj 18> (table as desk).
Please note that this is an approximation as the description does not perfectly match any set of objects in the room.
Answer: <obj 24>

Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

3D Zero-shot Grounding

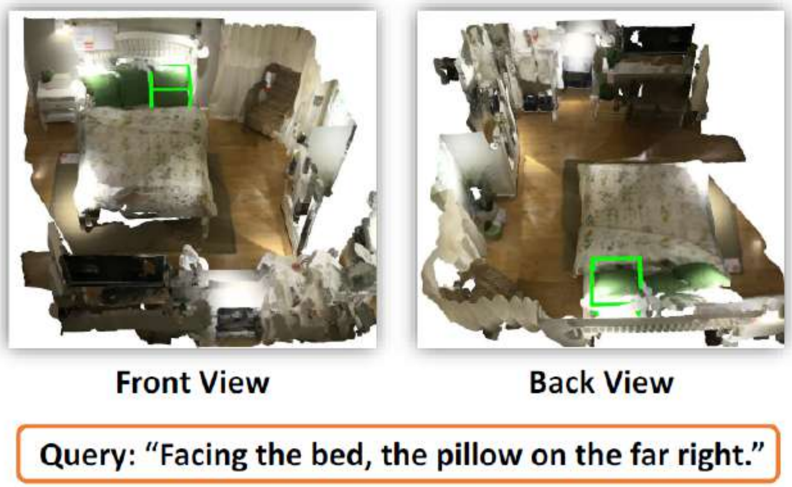
- **Method**
  - **Two zero-shot approaches for 3DVG**

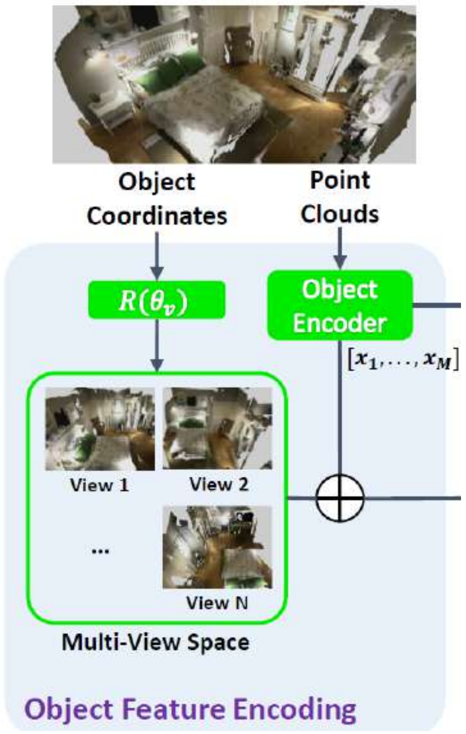


Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

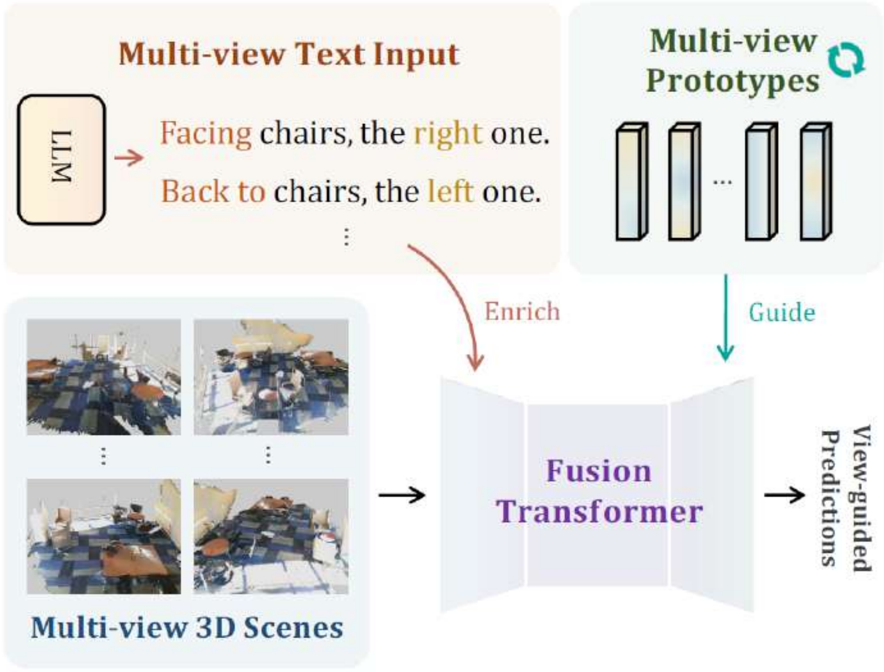
3D Zero-shot Grounding

- How to solve the view problem in 3D space?
  - Previous methods use data augmentation on text or object features





Multi-view transformer for 3d visual grounding (CVPR 22)



ViewRefer: Grasp the Multi-view Knowledge for 3D Visual Grounding (ICCV 23)

3D Zero-shot Grounding

- **Method**
  - **Addressing view-dependent relations: A shift to 2D egocentric view.**
  - **Addressing view-independent relations: using 3D coordinates.**

View-independent	*near, close, next to, far, above, below, under, top, on, opposite, middle*
View-dependent	*front, behind, back, right, left, facing, leftmost, rightmost, looking, across, between*
Functional	*min, max, size, length, width*

Table 1. Common relations in 3DVG.

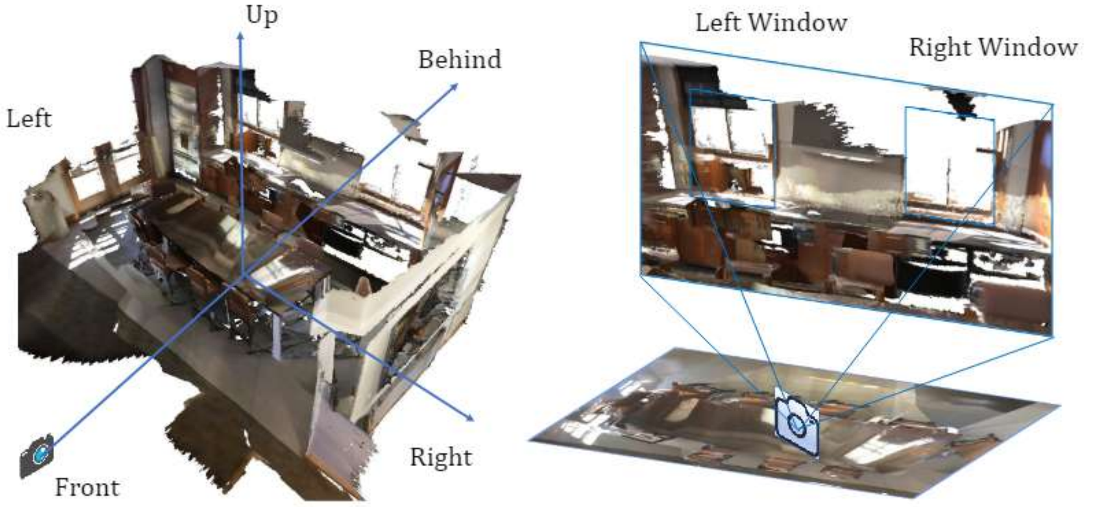


Figure 3. Addressing view-dependent relations: A shift to 2D egocentric view.

3D Zero-shot Grounding

- **Method**
  - **LOC module: extend the scope of existing 3D object detectors into open-vocabulary scenarios.**

BOX0=LOC(object='round cocktail table')

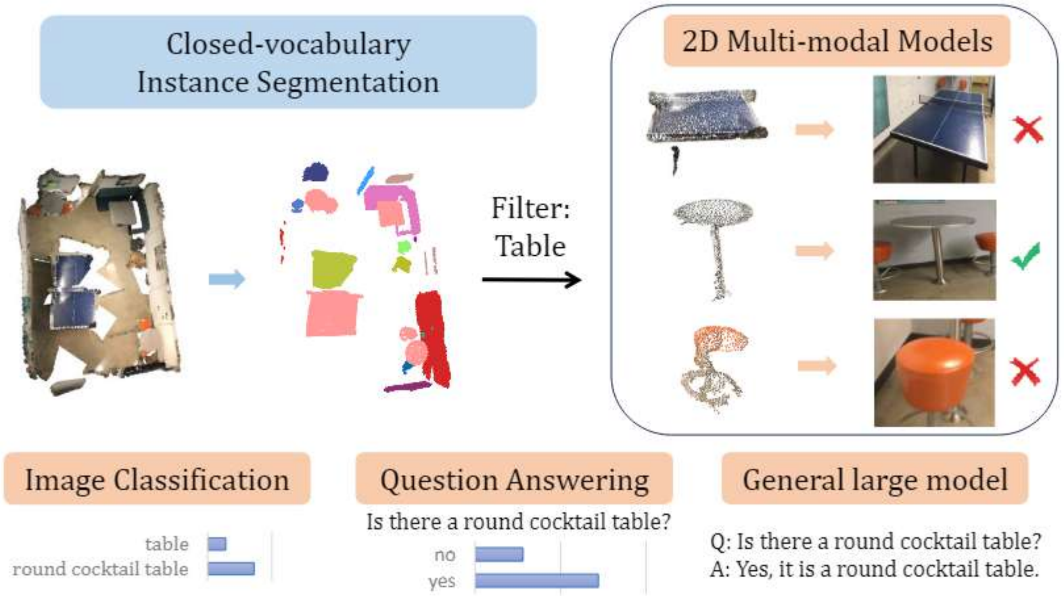
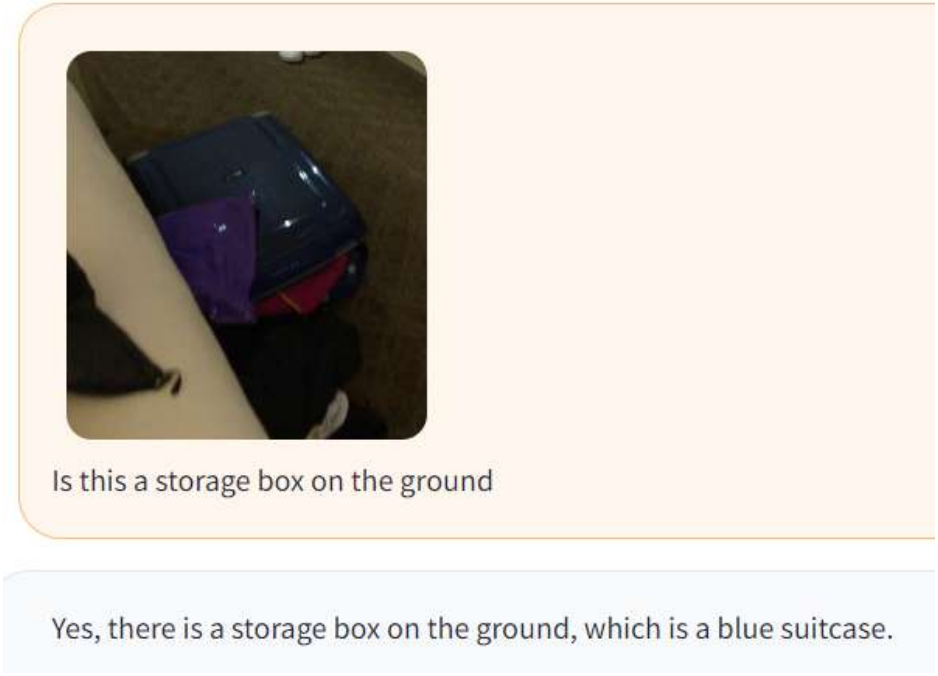


Figure 4. Illustration of the language-object correlation module.

BOX0=LOC(object=storage box on the ground)



Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

3D Zero-shot Grounding

- **Results**
  - **Our zero-shot approach can outperform some supervised baselines**

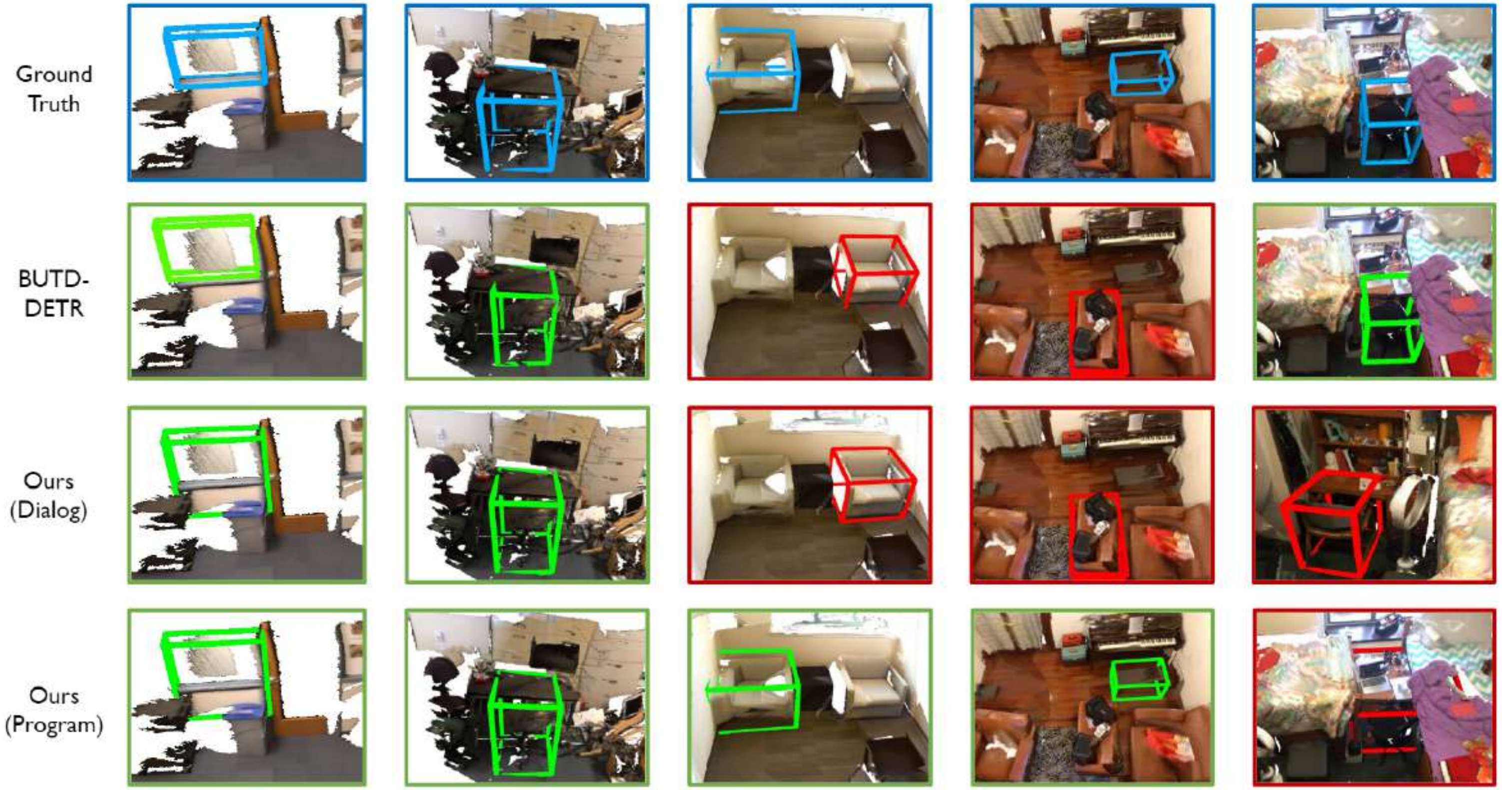
		Unique		Multiple		Overall	
Methods	Supervision	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [4]	fully	65.0	43.3	30.6	19.8	37.3	24.3
TGNN [17]	fully	64.5	53.0	27.0	21.9	34.3	29.7
InstanceRefer [60]	fully	77.5	66.8	31.3	24.8	40.2	32.9
3DVG-Transformer [65]	fully	81.9	60.6	39.3	28.4	47.6	34.7
BUTD-DETR [20]	fully	84.2	66.3	46.6	35.1	52.2	39.8
LERF [23]	-	-	-	-	-	4.8	0.9
OpenScene [34]	-	20.1	13.1	11.1	4.4	13.2	6.5
Ours (2D only)	-	32.5	27.8	16.1	14.6	20.0	17.6
Ours (3D only)	-	57.1	49.4	25.9	23.3	33.1	29.3
Ours	-	**63.8**	**58.4**	**27.7**	**24.6**	**36.4**	**32.7**

Table 2. 3DVG results on ScanRefer validation set. The accuracy on the "unique" subset, "multiple" subset, and whole validation set are all provided. Following [4], we label the scene as "unique" if it only contains a single object of its class. Otherwise, we label it as "multiple".

Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

3D Zero-shot Grounding

- **Visualization results**



It is a window. It is located above a recycle bin that has a blue top.
(a)

The rolling office chair. The chair is under the desk.
(b)

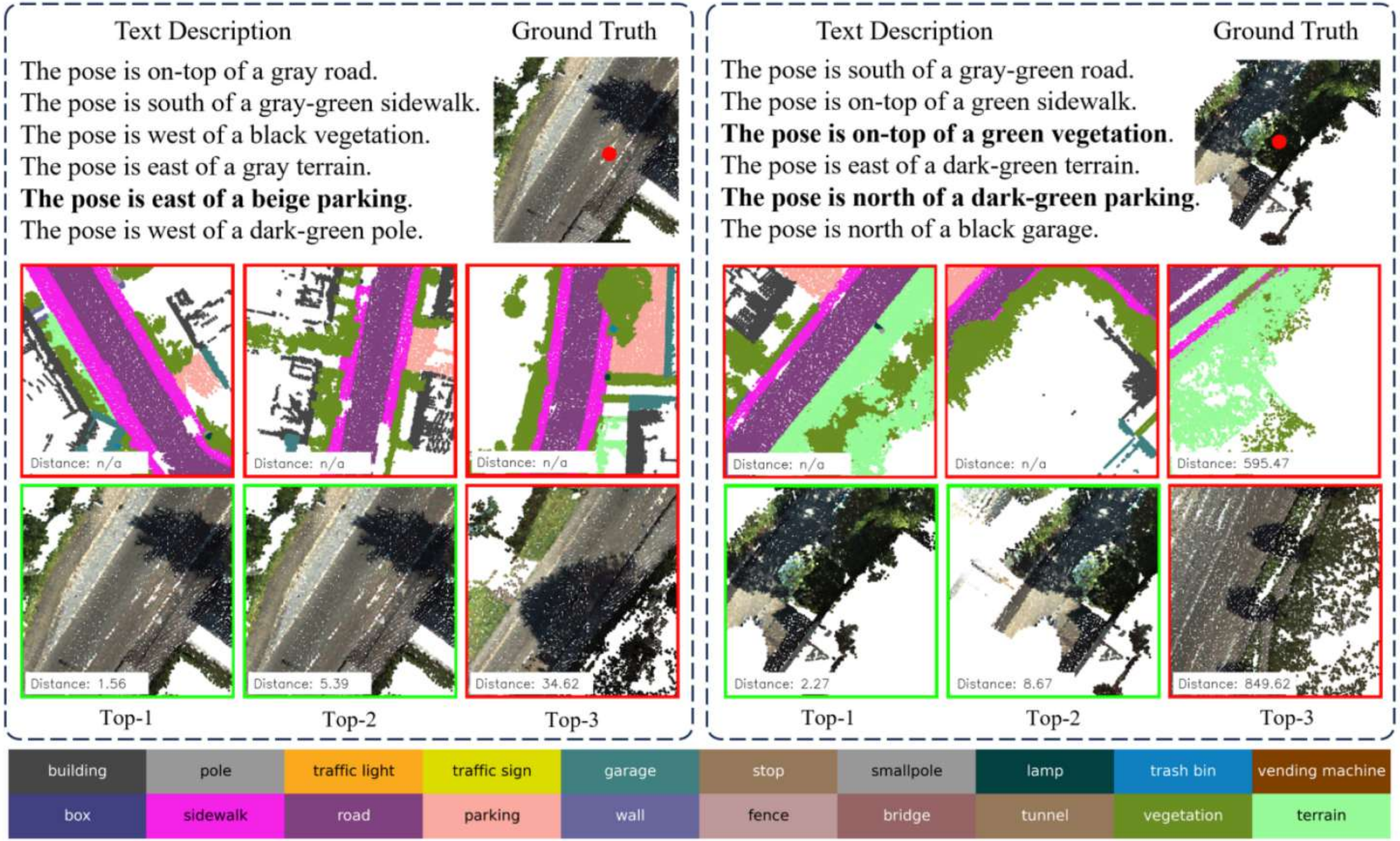
There is a square beige armchair. It is left of a square table.
(c)

This is a brown piano bench. It is in front of the piano.
(d)

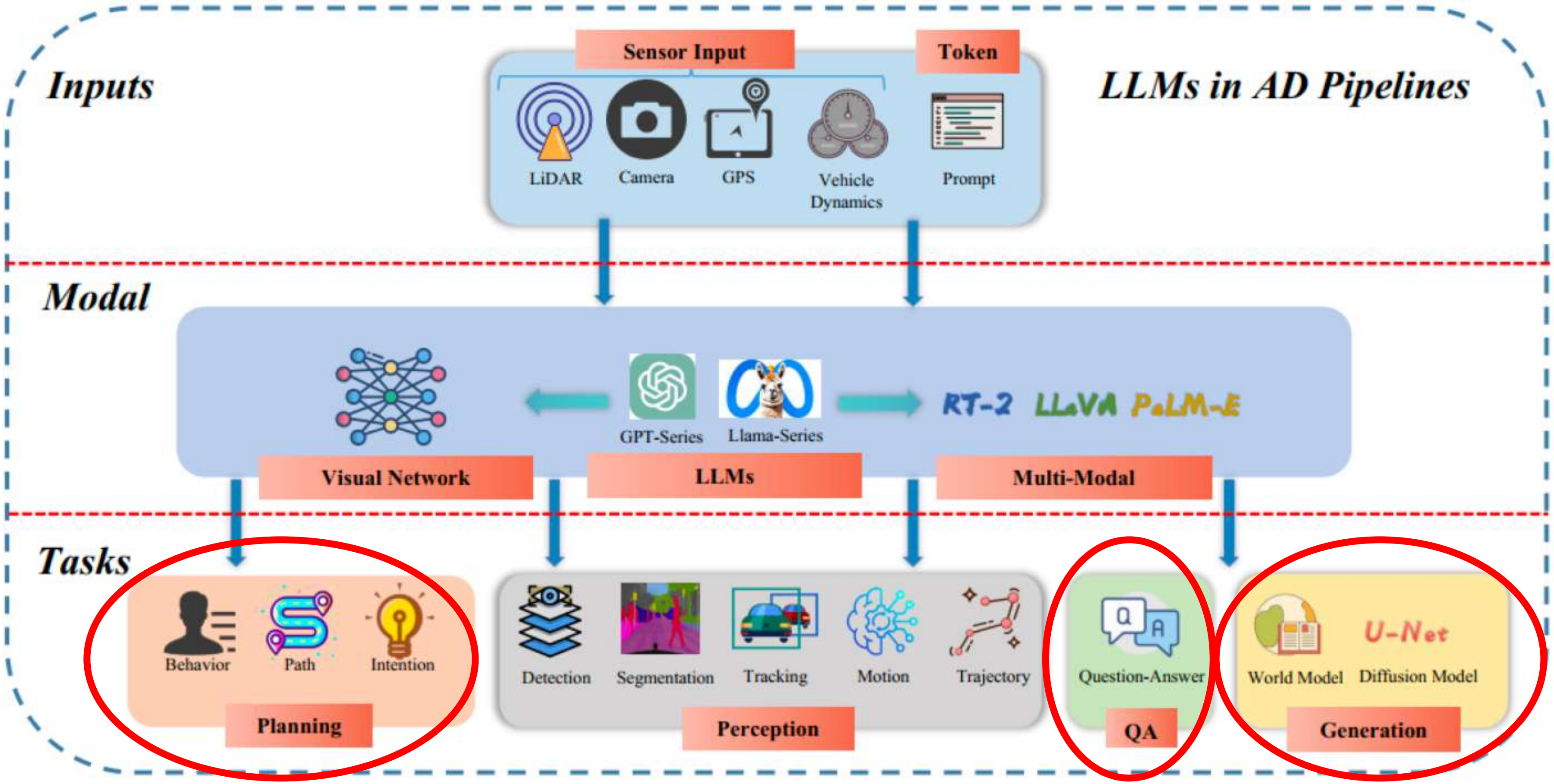
A desk chair is pushed into a small computer desk. The chair has wheels .
(e)

Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding (CVPR 24)

3D Visual localization for AD



Outline

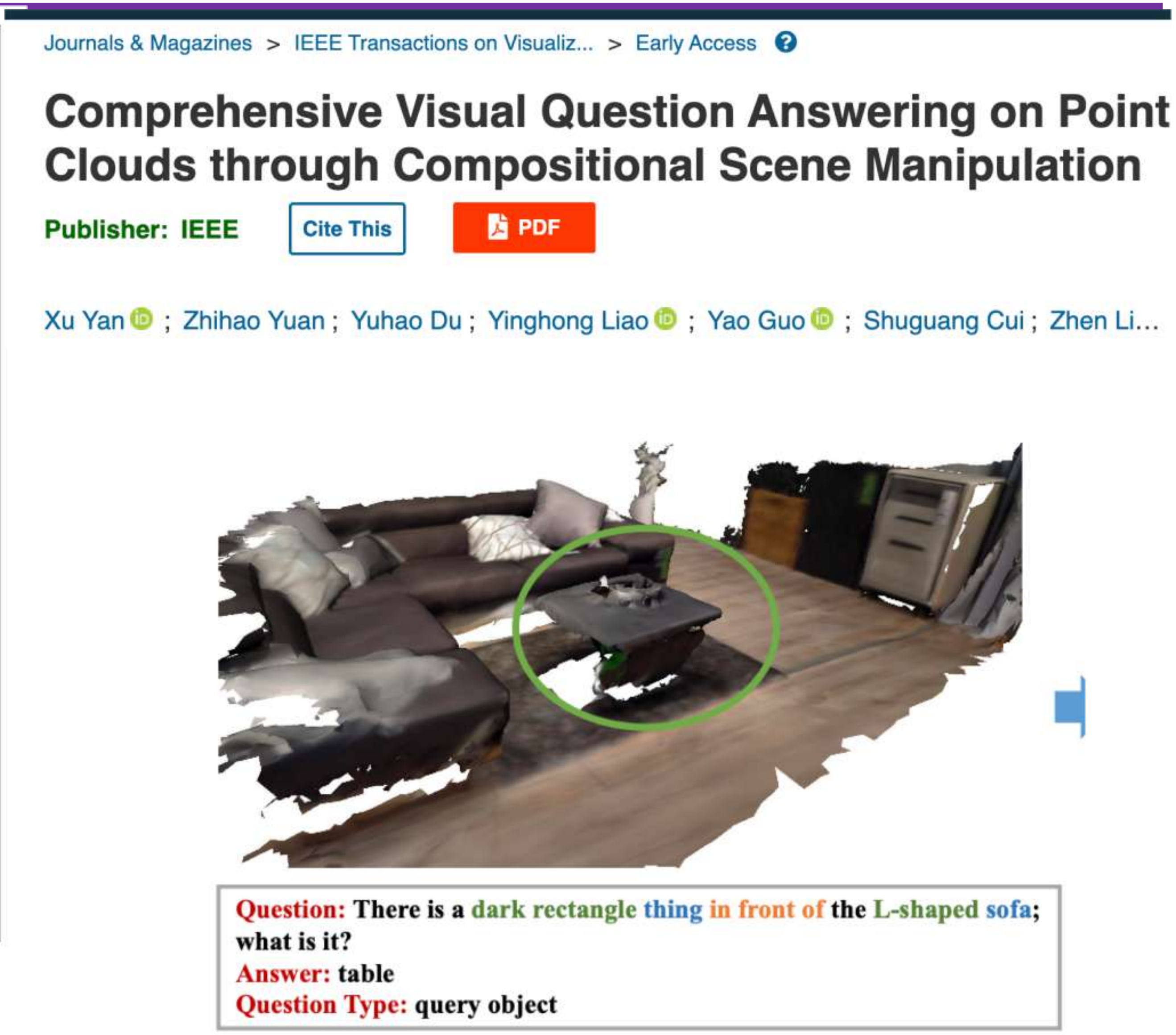


Enhanced 3D Reasoning

- ➢ Visual programming for open-world grounding using LLM
- ➢ 3D VQA without data bias for AD

3D VQA with manipulation to overcome data bias

LLVM-AD



CLEVR3D-REAL

3D VQA with manipulation to overcome data bias

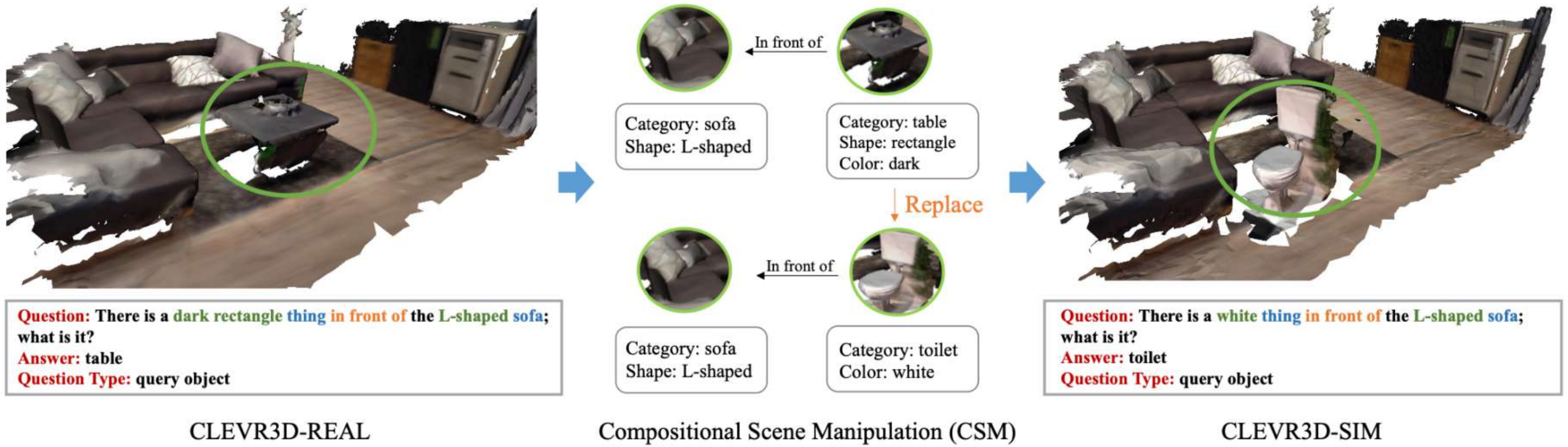


Fig. 1. Visual Question Answering on 3D Point Cloud (VQA-3D). In this paper, we introduce a new dataset CLEVR3D, which consists of CLEVR3D-REAL and CLEVR3D-SIM sub-datasets. Selected questions from CLEVR3D-REAL (left) test aspects of visual reasoning in 3D scenes such as counting, object identification, query attribute, and attribute comparison. Each question contains objects, attributes, and relationships. CLEVR3D-SIM dataset is obtained through the compositional scene manipulation (CSM) shown in right for common-sense-independent VQA-3D.

3D VQA with manipulation to overcome data bias

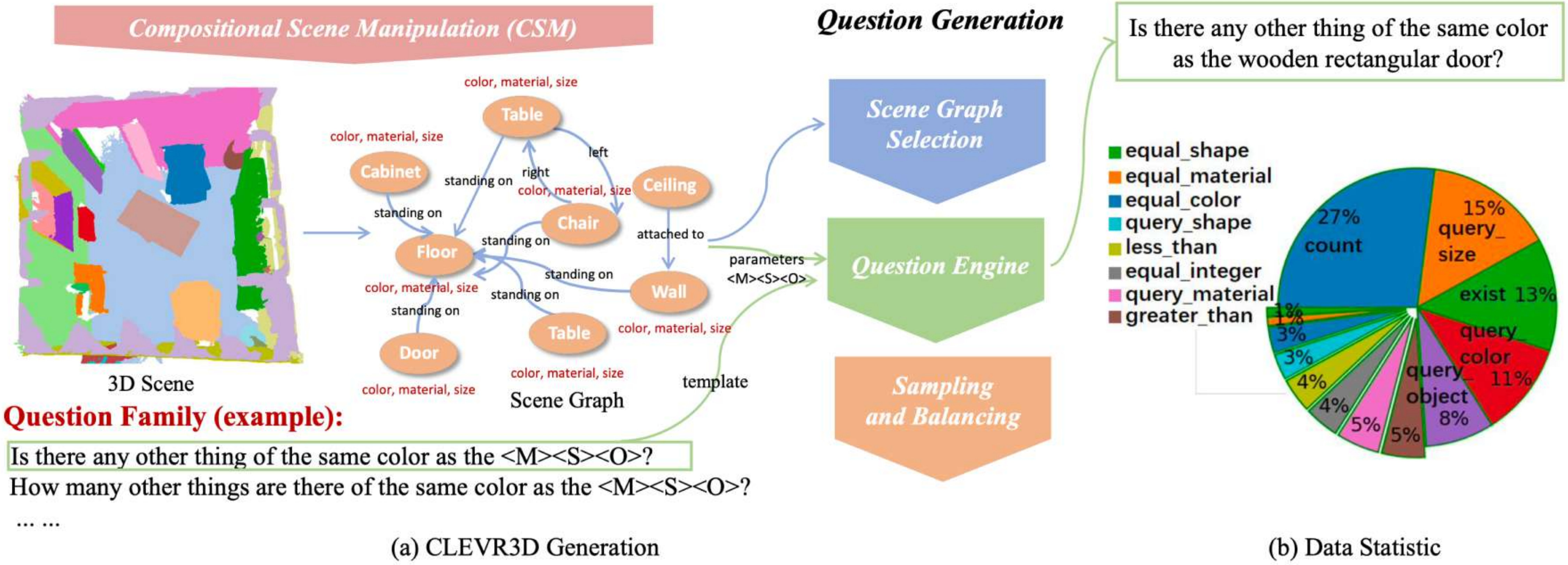
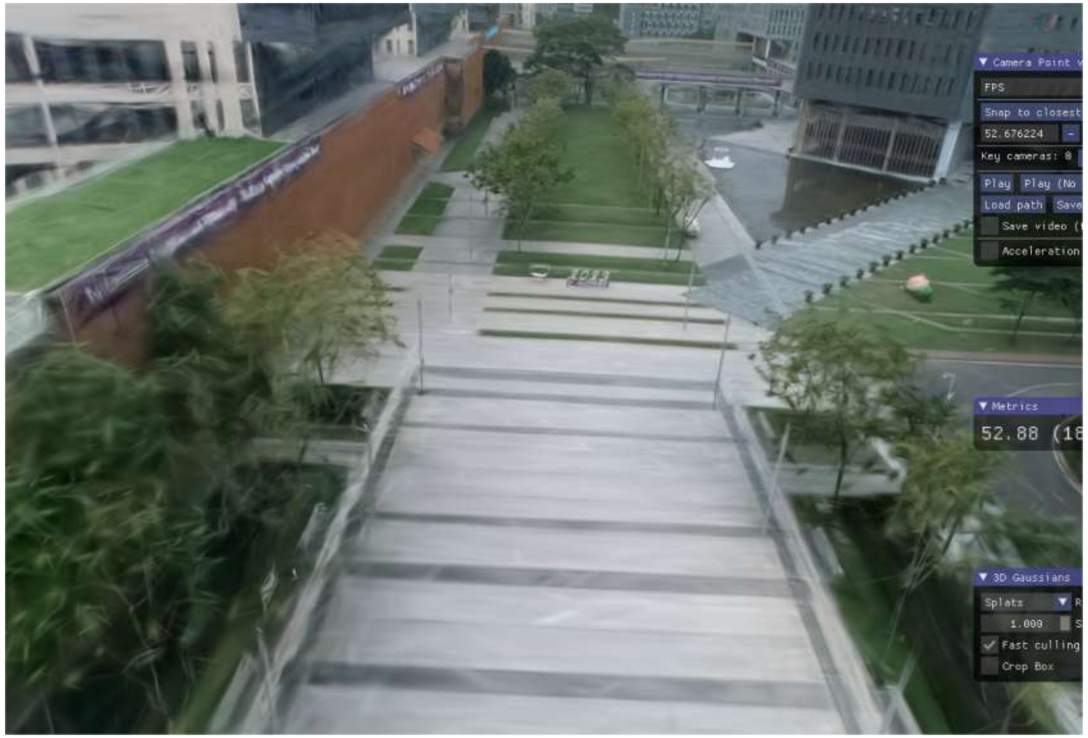


Fig. 2. **Overview of the CLEVR3D.** Part (a) illustrates the data generation process of CLEVR3D-REAL, where the whole process contains three steps: scene graph selection, question engine design, and sampling. Besides, we can further exploit the compositional scene manipulation (CSM) strategy to generate more simulated common-sense-independent 3D scenes and corresponding scene graphs for the CLEVR3D-SIM dataset. Part (b) shows the data statistics of question length and proportions. CLEVR3D contains more question types compared with the CLEVR dataset.

3D manipulation for AD

Synthetic 3D scene



Real 3D scene

Conclusion and Discussion

- ***Perception is still important, especially in open-world, but maybe precise perception limitation can be relaxed.***

- ***Reasoning is important for planning, especially for long-tailed scenes, but maybe large models can help.***

- ***Close-loop evaluation is really important for end-to-end AD, sim-to-real and real-to-sim dual-view can help.***

Acknowledge

Dr. Zhen Li

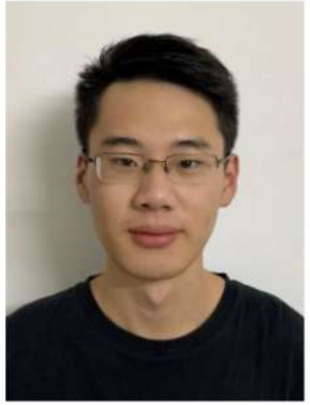
Assistant Professor

Deep Bit Lab

YAN Xu

Chaoda Zheng

Haiming Zhang



Zhihao Yuan

Yueru Luo

Yinghong Liao

Deep Bit Lab

Thanks You !

Q&A

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen