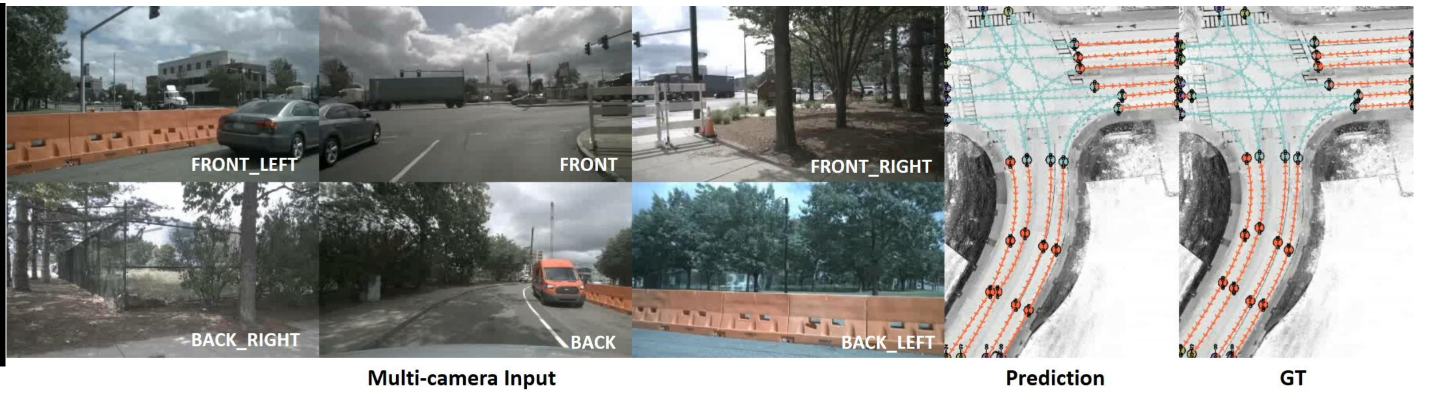
大规模自动驾驶仿真系统研究

张力

复旦大学

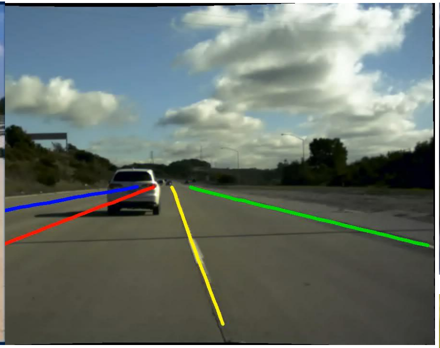
Visual segmentation – CVPR20, ECCV20, CVPR 2021, CVPR23



Autonomous driving – CVPR20, ECCV20, ICCV 2023

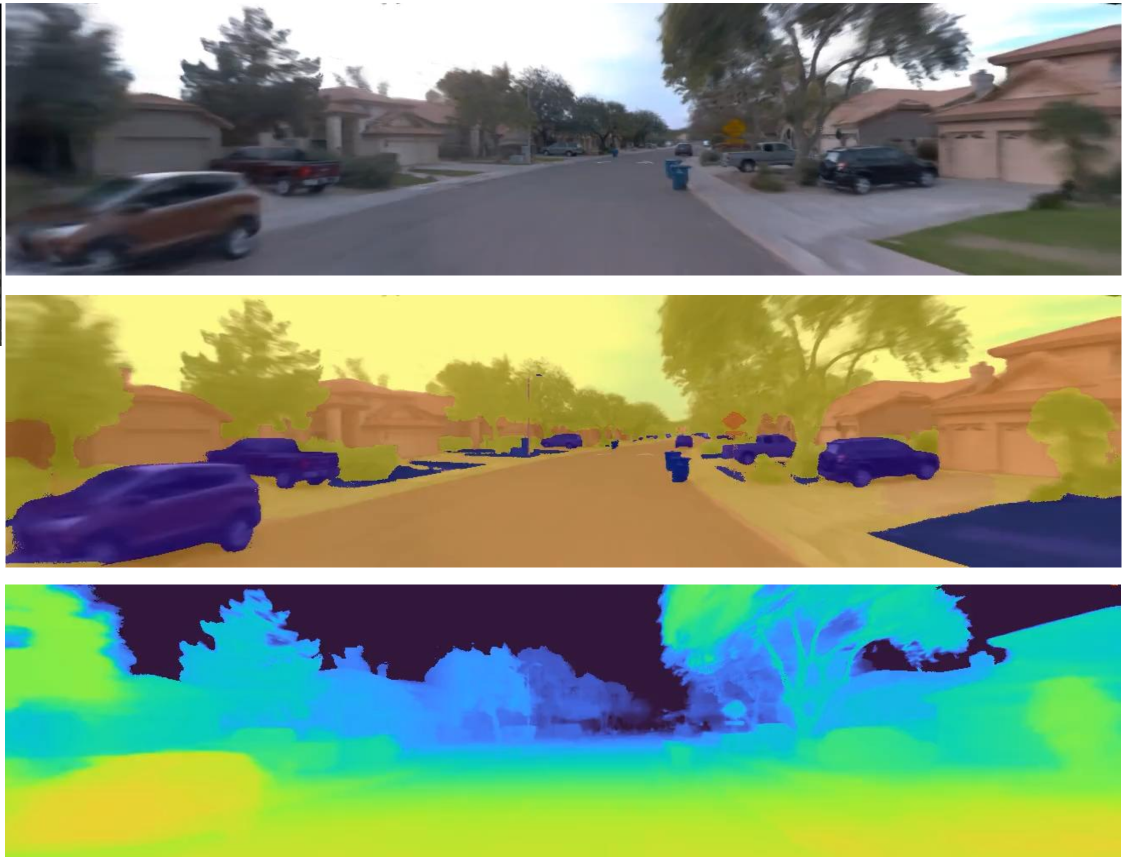
Dynamic network
CVPR20, TPAMI22

Video object tracking
CVPR19, TPAMI22



Lane prediction
CVPR22, ECCV22

3D detection – CVPR21, ICCV21, NeurIPS21, ECCV22, NeurIPS22, AAAI 23



Neural rendering for self-driving simulation, ICLR23

Large-scale self-driving simulation

Graphics-based simulation

- ☹ Cost $ 1million /km.
- ☹ Need Professional developers
- ☹ Not realistic. It is far from real world.
- ☹ Algorithms developed on it can't be directly used in real world.

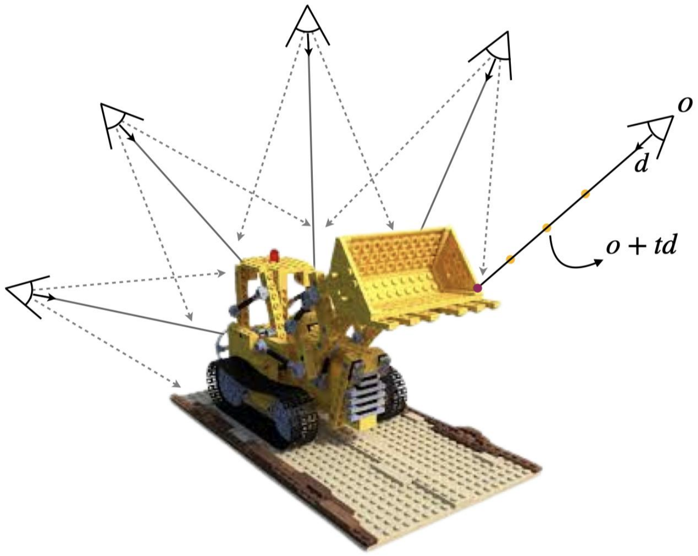
Large-scale self-driving simulation

S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang*** **–** ***ICLR 2023***

Motivation: NeRF Designed for Street Views

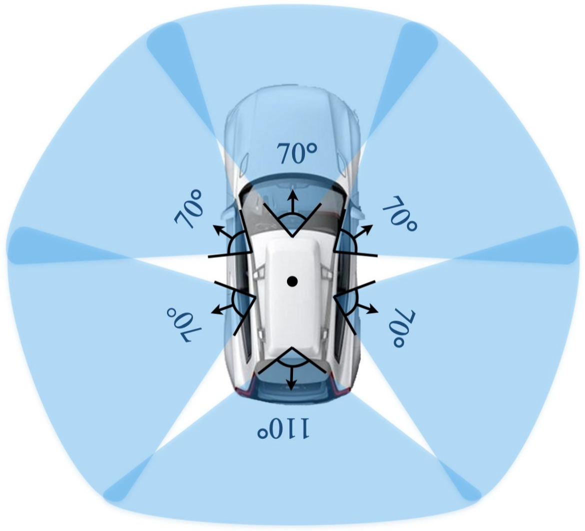
1. Vanilla NeRFs Camera Settings

- Object Centric Camera Settings
- Dense Image Overlap
- Meticulously Captured by Human Specialists



2. Camera Settings for Street Views

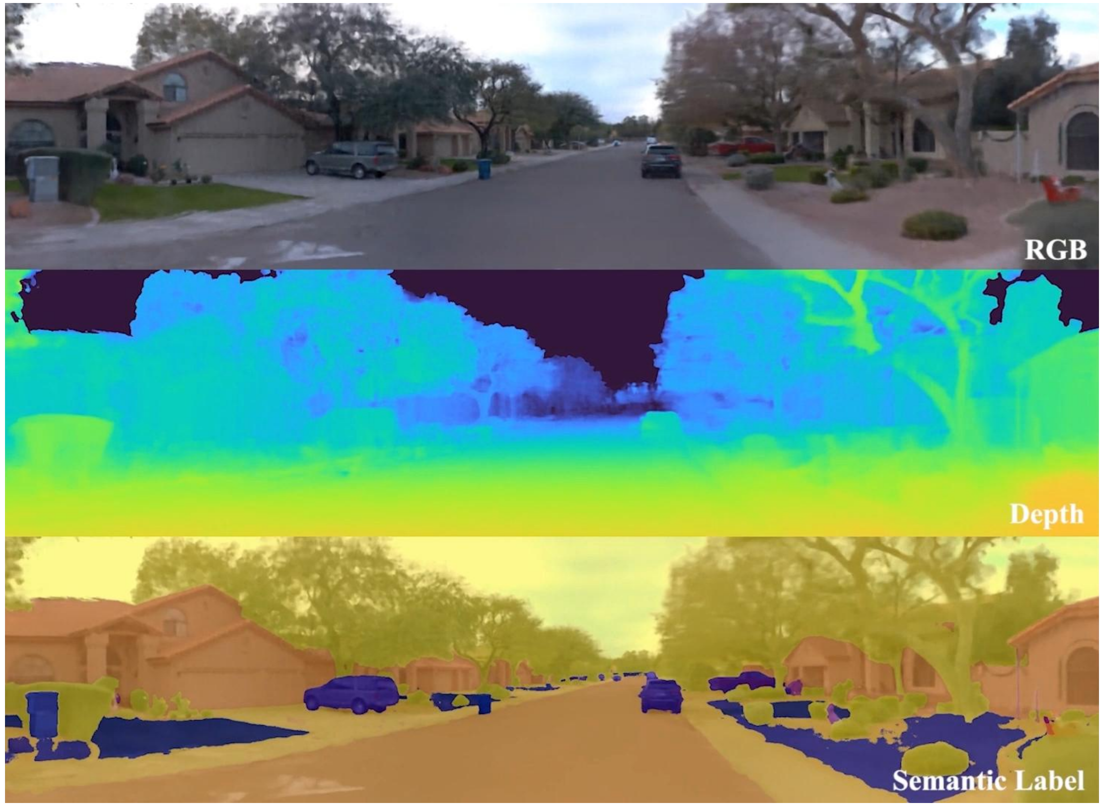
- Ego Centric Camera Settings
- Sparse Image Overlap
- Captured by Vehicles in Transit with lots of noise



S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

S-NeRF: Robust NeRF System for Street Views

We introduce S-NeRF a robust NeRF system for high-quality street view reconstruction for both the large-scale background scenes and foreground vehicles



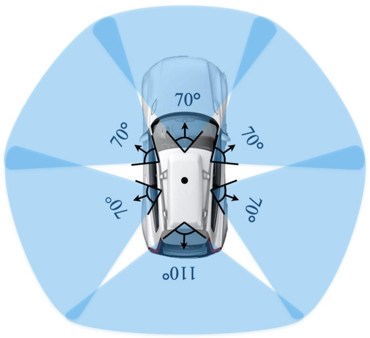
Background Scene Reconstruction

Foreground Vehicle Reconstruction

S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

Augment NeRF model with LiDAR signals

Limited RGB overlap

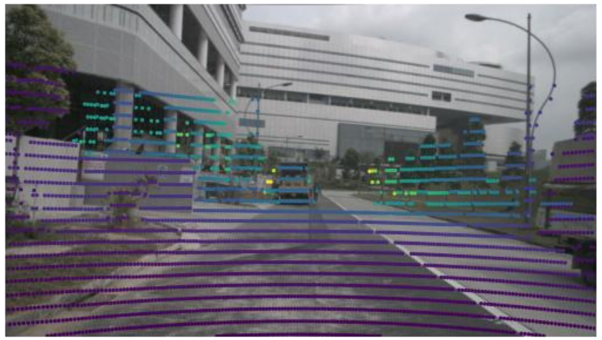


→

LiDAR Sensor

→

Sparse Point Cloud



Auxiliary Depth Supervision:

Given Camera Rays: $R(t) = \{\mathbf{o} + t\mathbf{d} | t \in \mathcal{R}^+\}$

NeRF Termination Depth Rendering: $\hat{D}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(t)\mathrm{d}t.$

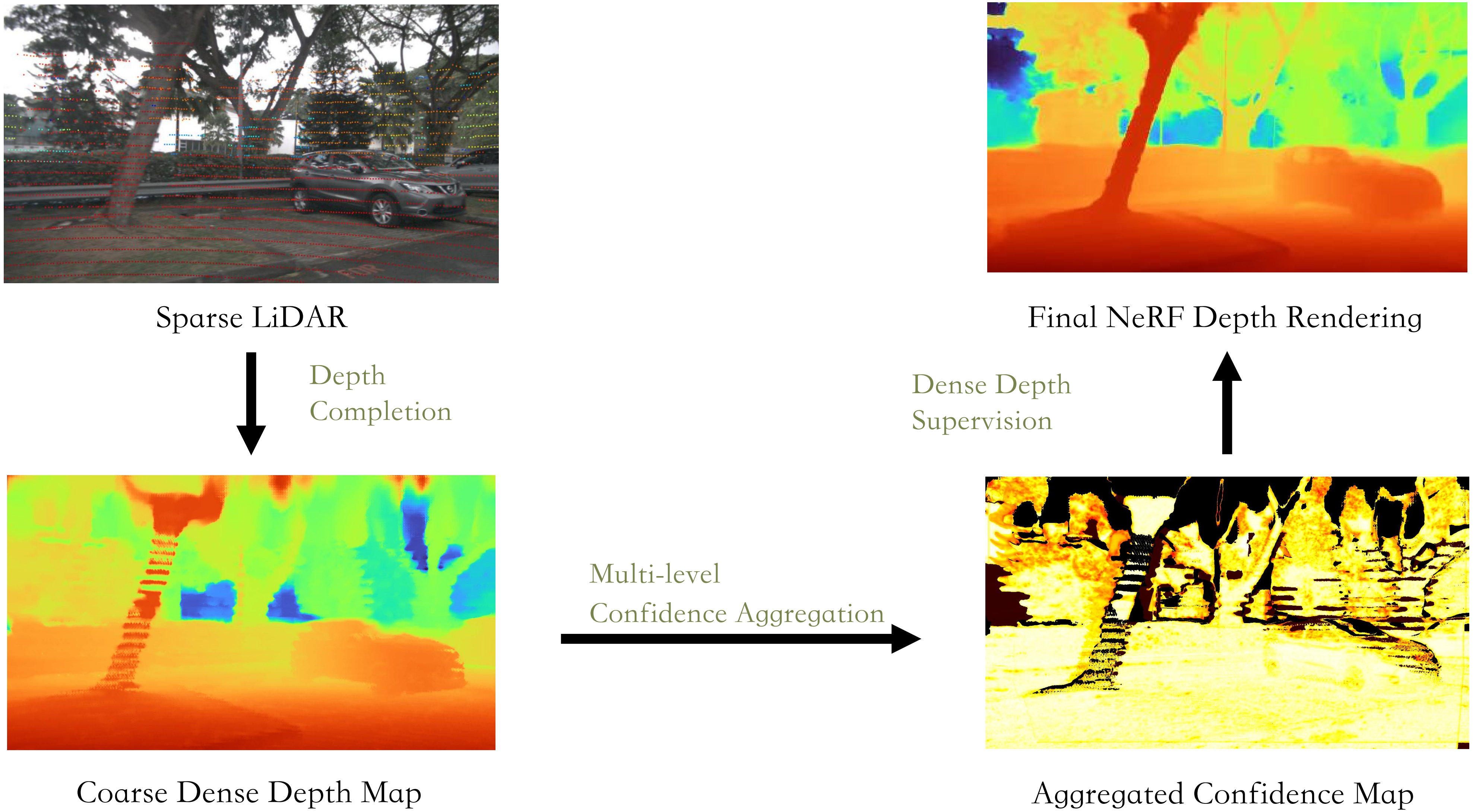
Depth Supervision: $\mathcal{L}_{depth} = \sum_{r \in R} \|D(r) - \hat{D}(r)\|$

Issues:

1. Limited Operational Ranges
2. Sparse Point Cloud Signals

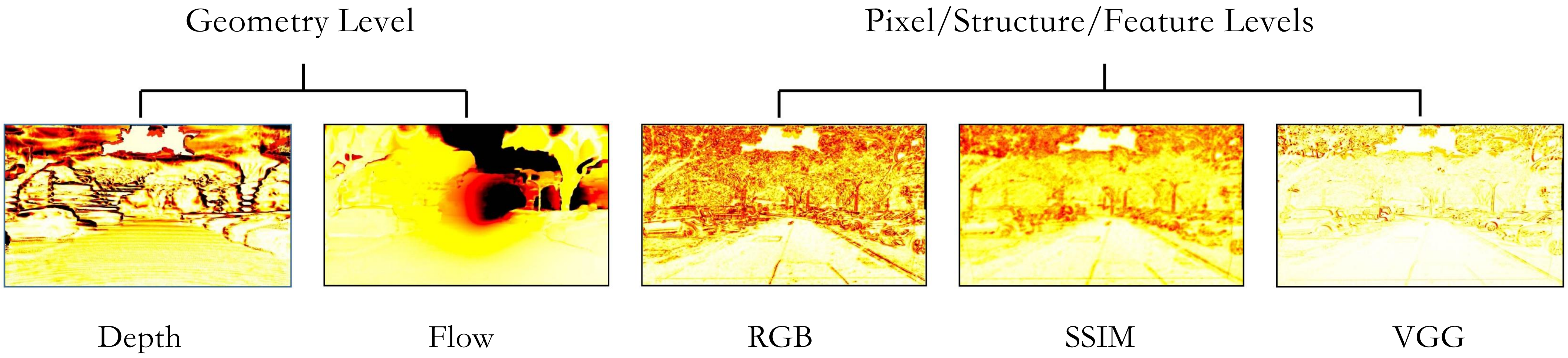
S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

Robust dense depth supervision



S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

Multi-level confidence aggregation



1. Multi-Level Confidence Map

$$\mathcal{C}_{depth} = \gamma(|d_t - \hat{d}_t)|/d_s), \quad \gamma(x) = \begin{cases} 0, & \text{if } x \geq \tau, \\ 1 - x/\tau, & \text{otherwise.} \end{cases}$$

$$\mathcal{C}_{flow} = \gamma(\frac{\|\Delta_{x,y} - f_{s \to t}(x_s, y_s)\|}{\|\Delta_{x,y}\|}), \quad \Delta_{x,y} = (x_t - x_s, y_t - y_s).$$

$$\mathcal{C}_{\text{rgb}} = 1 - |\mathcal{I}_s - \hat{\mathcal{I}}_s|, \quad \mathcal{C}_{\text{ssim}} = \text{SSIM}(\mathcal{I}_s, \hat{\mathcal{I}}_s)), \quad \mathcal{C}_{\text{vgg}} = 1 - ||\mathcal{F}_s - \hat{\mathcal{F}}_s||.$$

2. Learnable Confidence Aggregation

Final Confidence Map $\hat{\mathcal{C}} = \sum_i \omega_i \mathcal{C}_i$

$\sum_i \omega_i = 1 \quad i \in \{depth, flow, rgb, ssim, vgg\}$

3. Confidence based Dense Depth supervision

$$\mathcal{L}_{depth} = \sum_{r \in R} \hat{\mathcal{C}}(r) \|D(r) - \hat{D}(r)\|$$

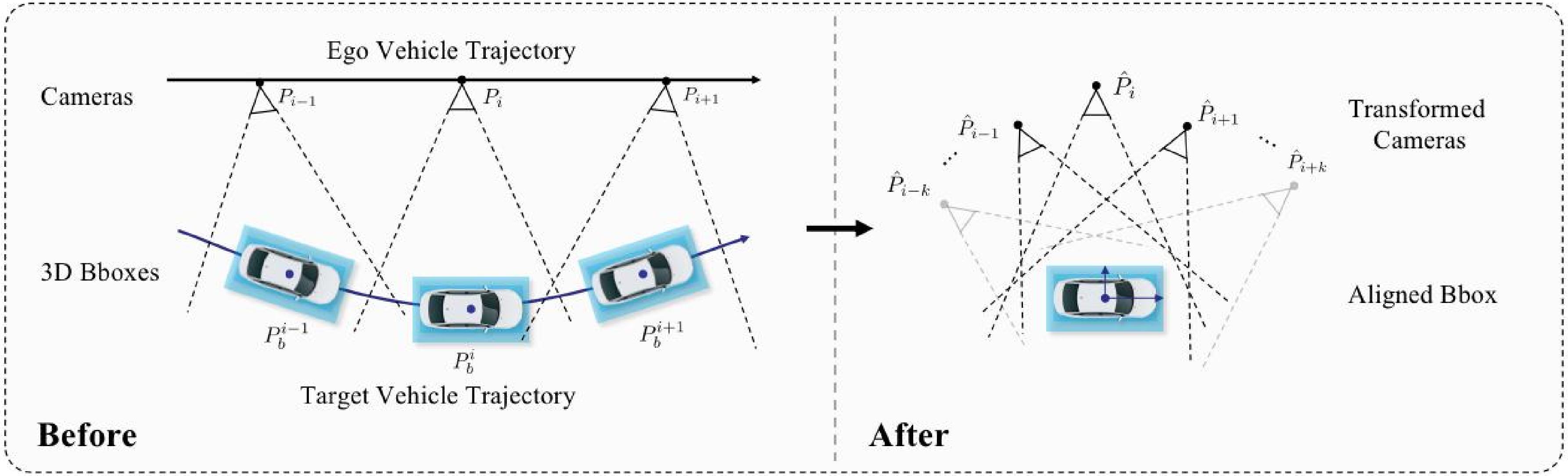
S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

Camera transformation for moving vehicles

Given original camera pose P_i and the 3D bounding box of the target vehicle P_b

The transformed camera system treats the target car (moving object) as static and then compute the relative camera poses for the ego car's camera.

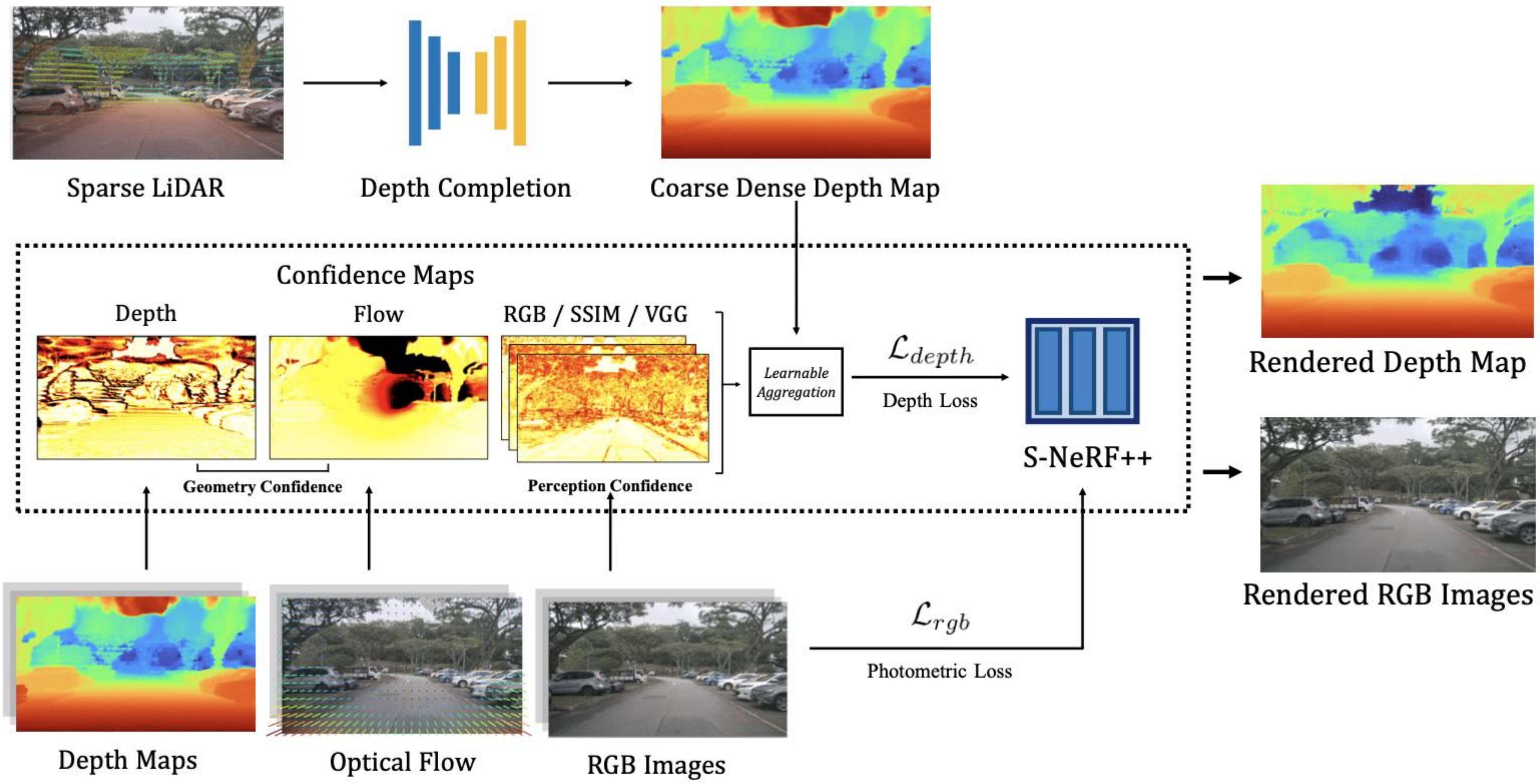
$$\hat{P}_i = (P_i P_b^{-1})^{-1} = P_b P_i^{-1}, \quad P^{-1} = \begin{bmatrix} R^T & -R^T T \\ \mathbf{0}^T & 1 \end{bmatrix}$$



Camera Transformation Process

S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

S-NeRF: Robust NeRF System for Street Views



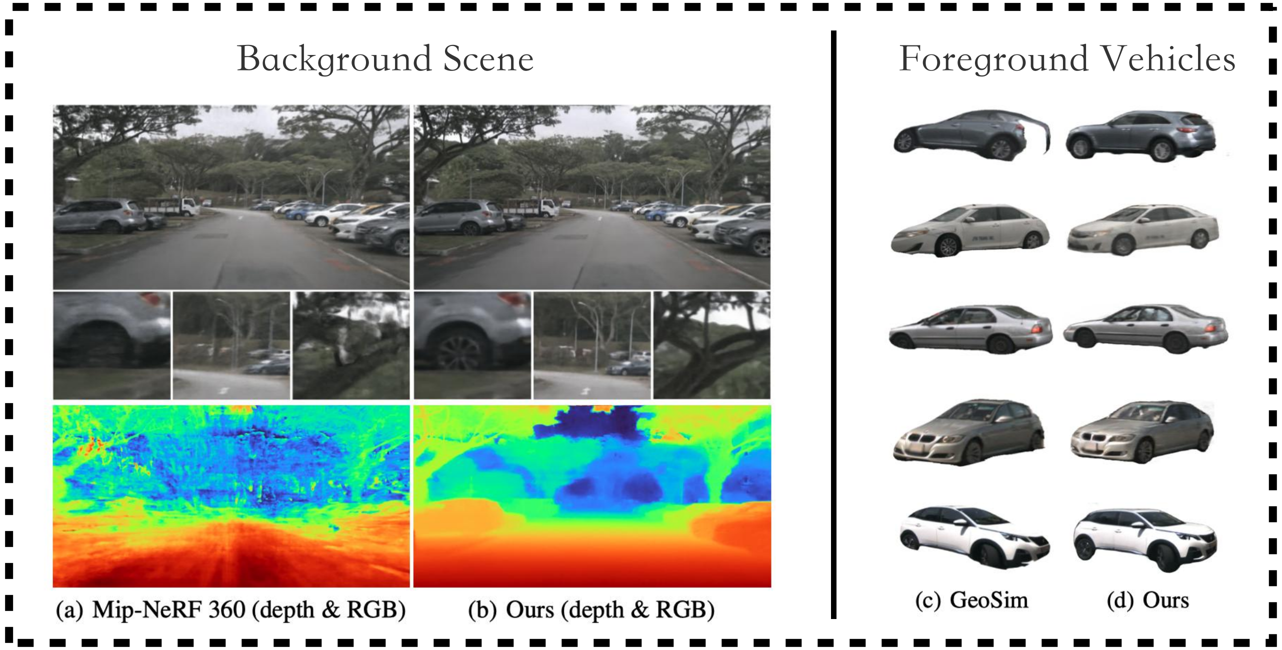
S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang*** *–* ***ICLR 2023***

Results

S-NeRF significantly outperforms previous State-of-the-Art Models by a large margin

In both the background reconstruction and foreground vehicle reconstruction

qualitatively and quantatively



Large-scale Scenes Synthesis			
Methods	PSNR↑	SSIM↑	LPIPS↓
Mip-NeRF (Barron et al., 2021)	18.22	0.655	0.421
Mip-NeRF360 (Barron et al., 2021)	24.37	0.795	0.240
Urban-NeRF (Rematas et al., 2022)	21.49	0.661	0.491
Ours	**26.21**	**0.831**	**0.228**

	Static Vehicles			Moving Vehicles		
Methods	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	11.78	0.539	0.444	–	–	–
GeoSim	11.58	0.602	0.367	12.24	0.623	0.322
Ours	**18.81**	**0.785**	**0.194**	**18.00**	**0.736**	**0.226**

S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang*** ***– ICLR 2023***

Further applications

S-NeRF offers a range of powerful features that enable advanced scene manipulation.

Input video for reconstruction

Rendering with lightning variations

Rendering novel views and insert new cars

Rendering novel trajectory

S-NeRF: Neural Radiance Fields for Street Views, *Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang,* ***Li Zhang – ICLR 2023***

S-NeRF: Neural Radiance Fields for Street Views

Ziyang Xie*, Junge Zhang*, Wenye Li, Feihu Zhang, Li Zhang
Fudan University

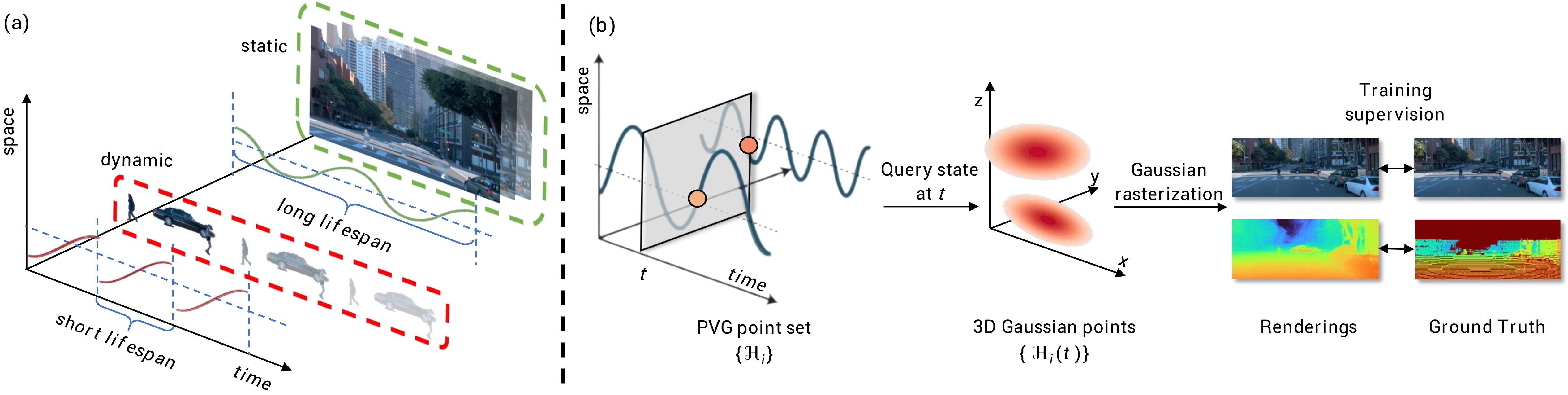
Periodic Vibration Gaussian: *Dynamic* Urban Scene Reconstruction and *Real-time* Rendering

Yurui Chen[1*] Chun Gu[1*] Junzhe Jiang[1] Xiatian Zhu[2] Li Zhang[1✉]

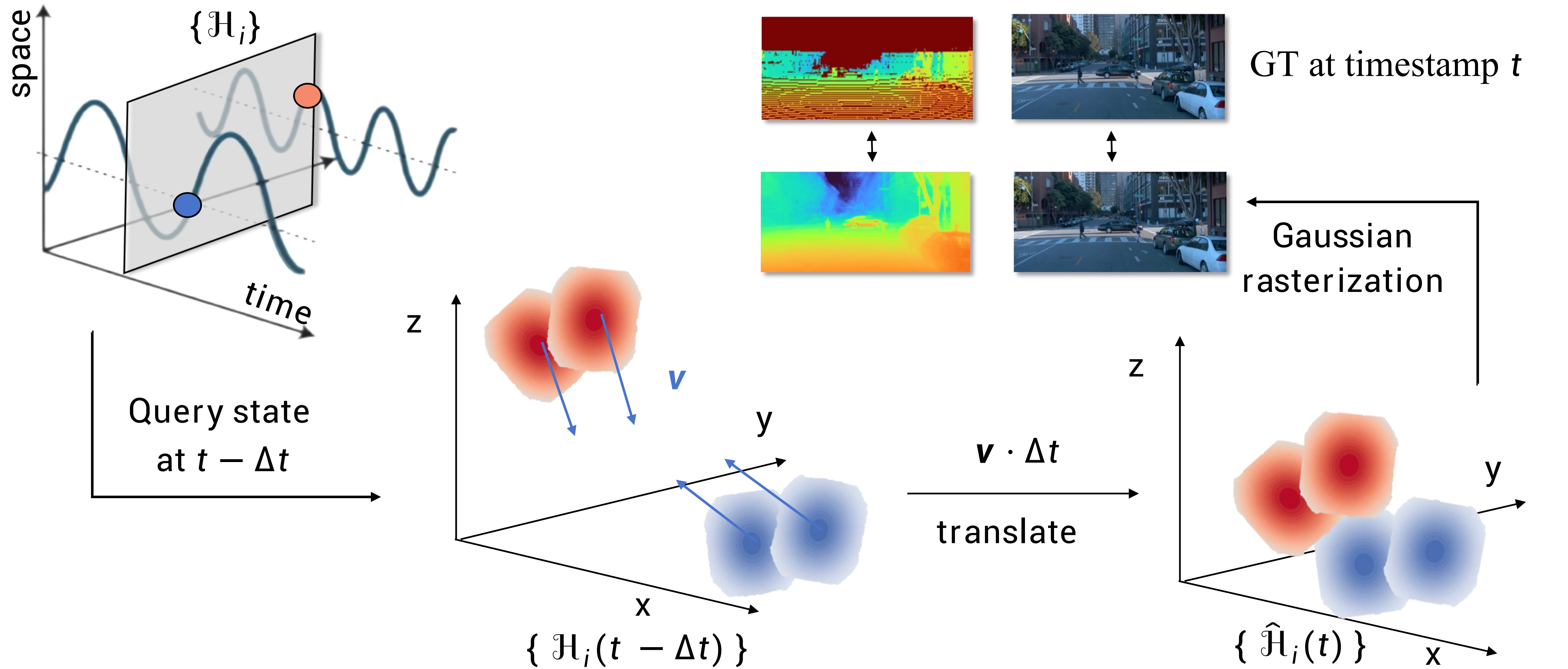
[1] Fudan University [2] University of Surrey

*Equally contributed

We present: **Periodic Vibration Gaussian** (PVG), a model adept at capturing the diverse characteristics of various objects and materials within dynamic urban scenes in a unified formulation.



To enhance temporally coherent and large scene representation learning with sparse data, we introduce a novel flow-based temporal smoothing mechanism and a position-aware adaptive control strategy respectively.



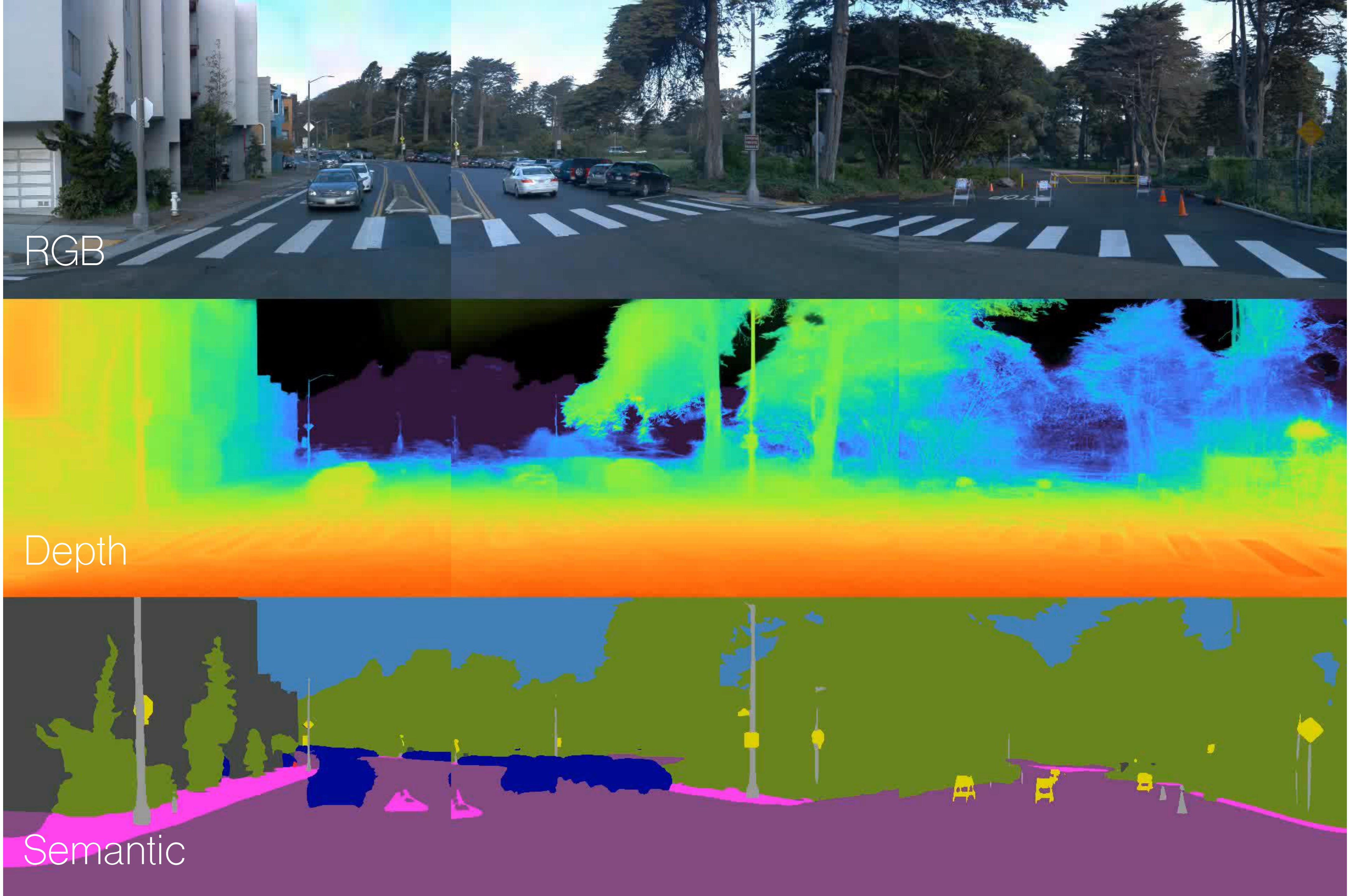
Without relying on manually labeled object bounding boxes or expensive optical flow estimation, PVG exhibits 50/6000-fold acceleration in training/rendering over the best alternative.

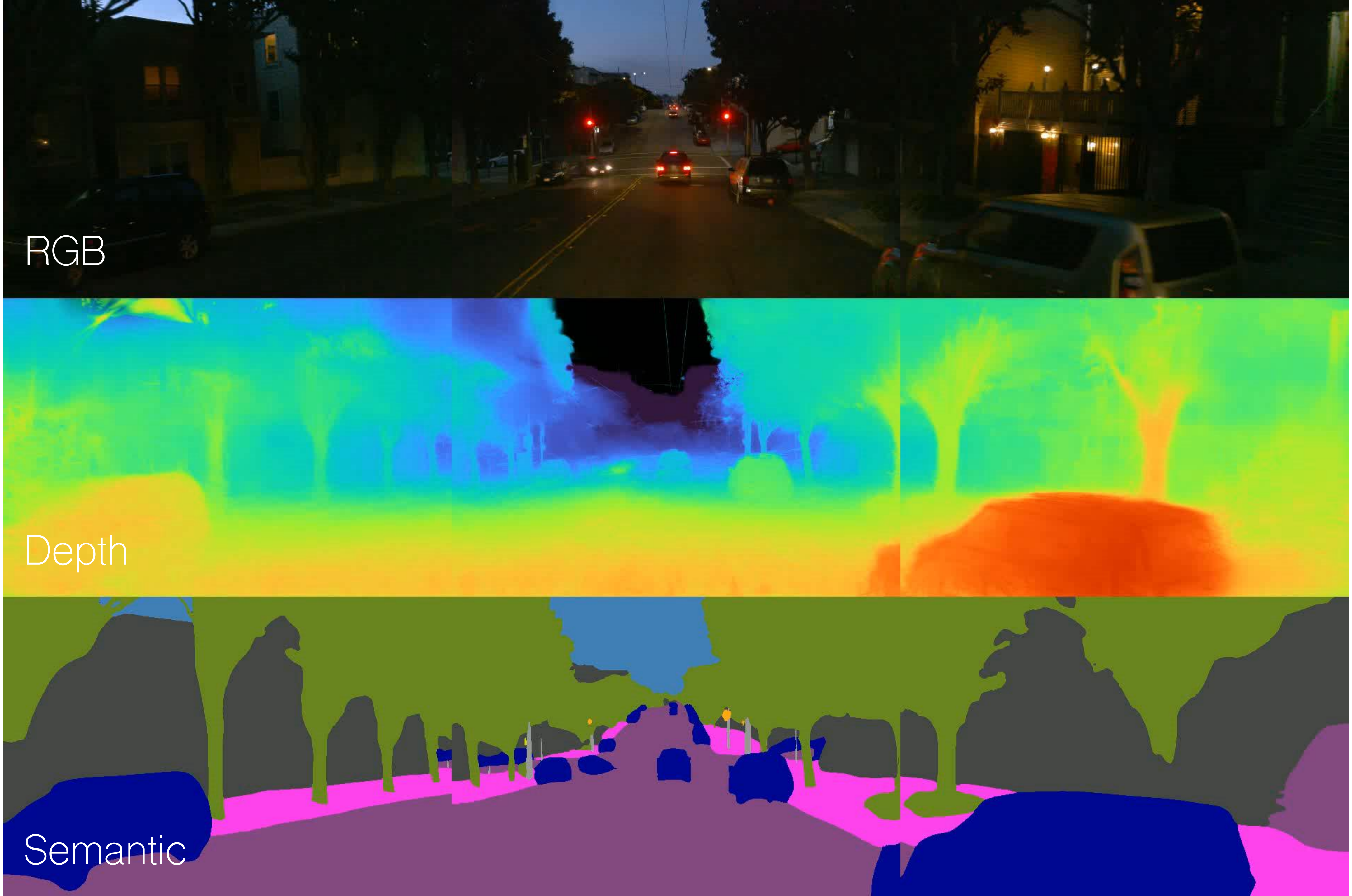
S-NeRF (ICLR'23)
0.0014 FPS

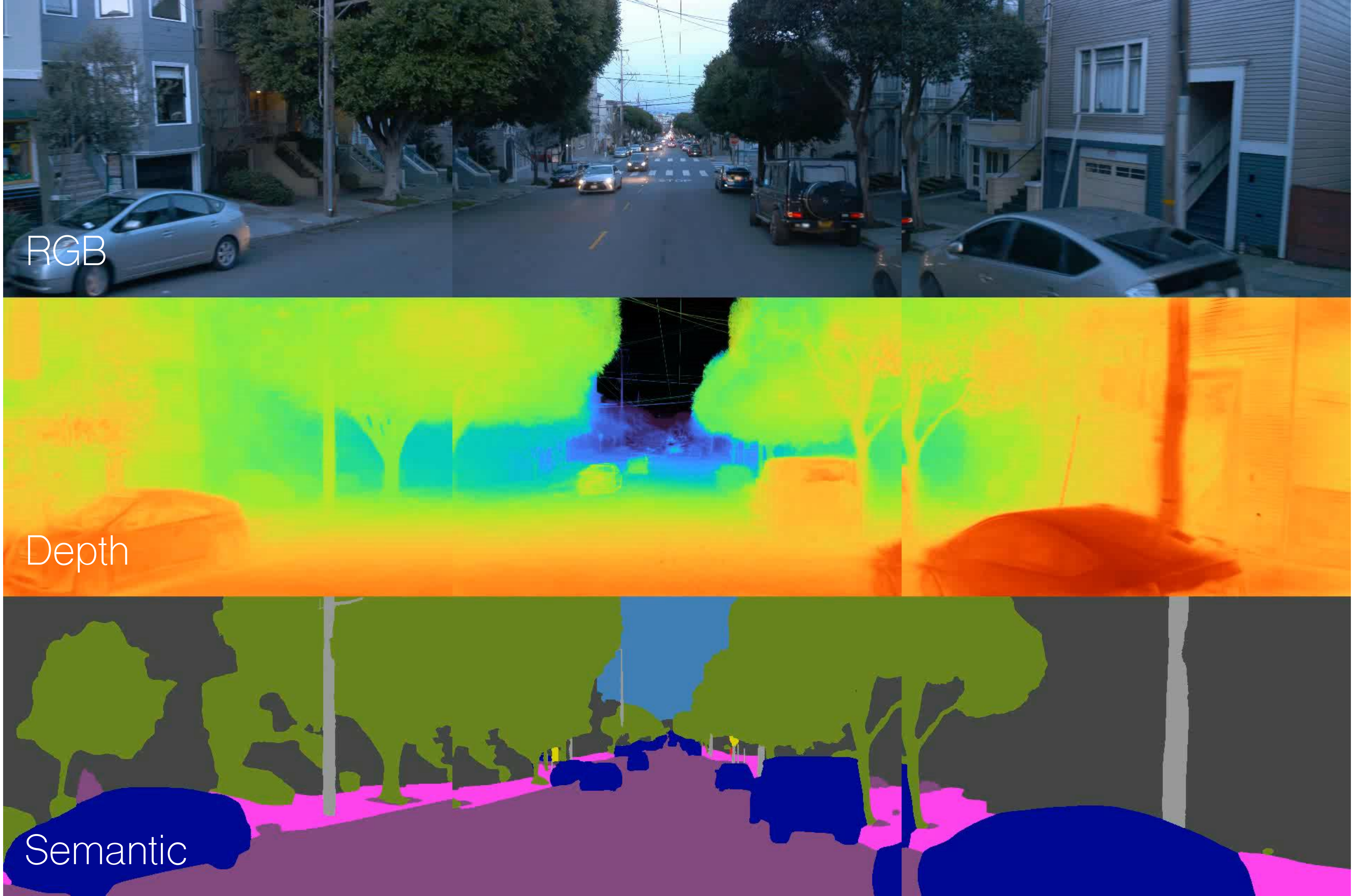
PVG (Ours)
50 FPS

Image Reconstruction on Waymo

Rendered RGB, Depth and Semantic


RGB
Depth
Semantic


RGB
Depth
Semantic


RGB
Depth
Semantic

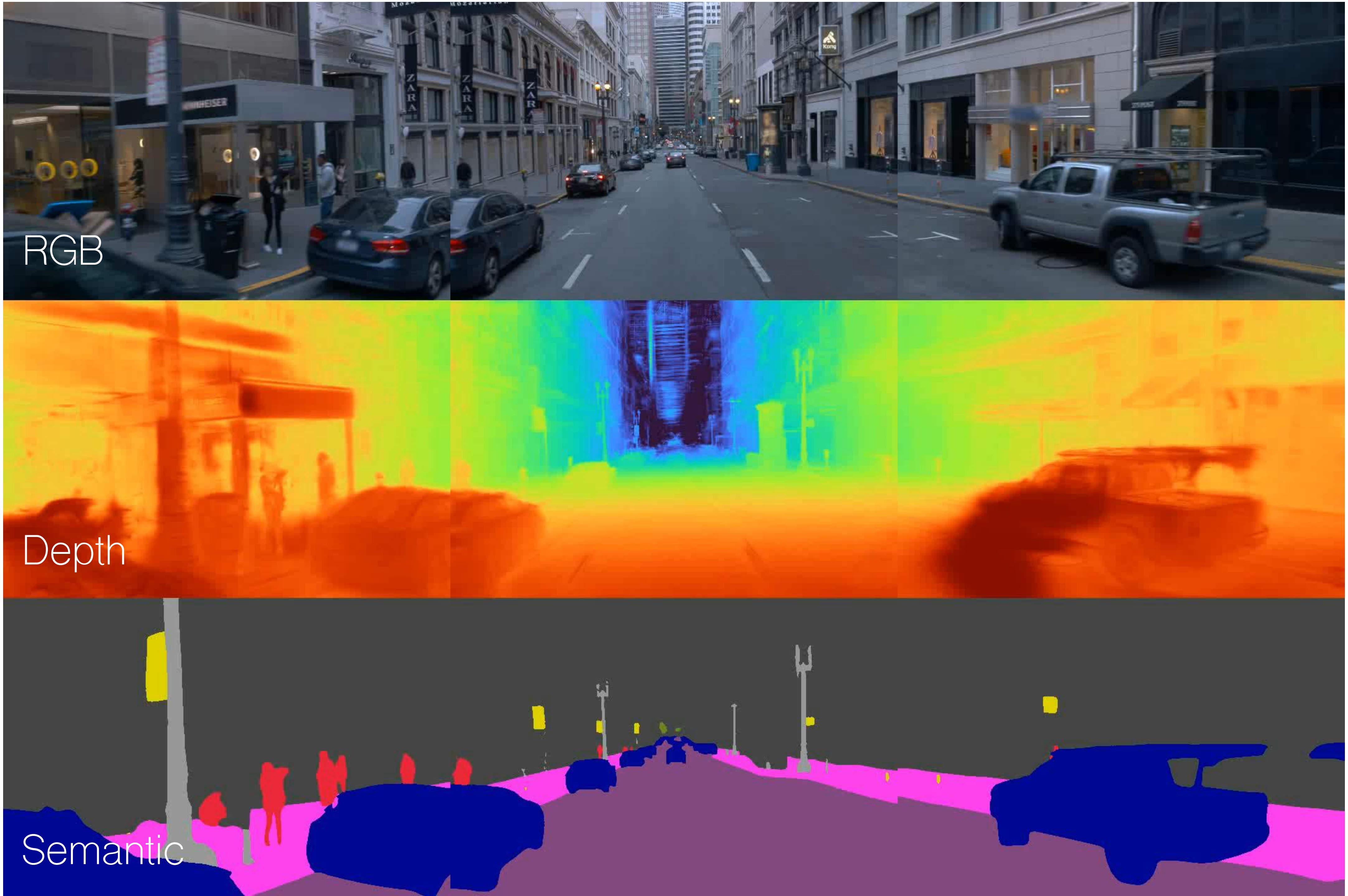
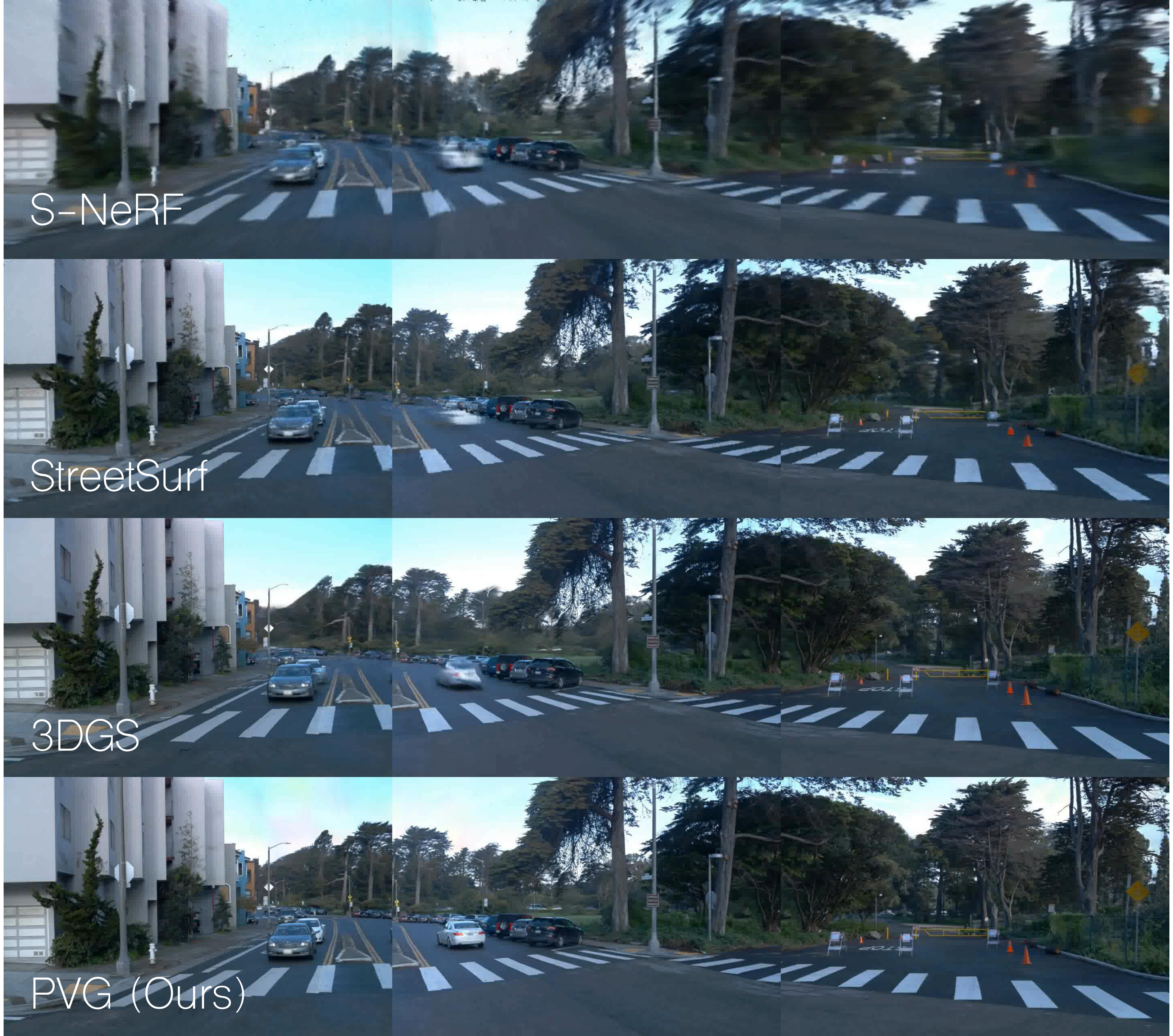

RGB
Depth
Semantic

Image Reconstruction on Waymo

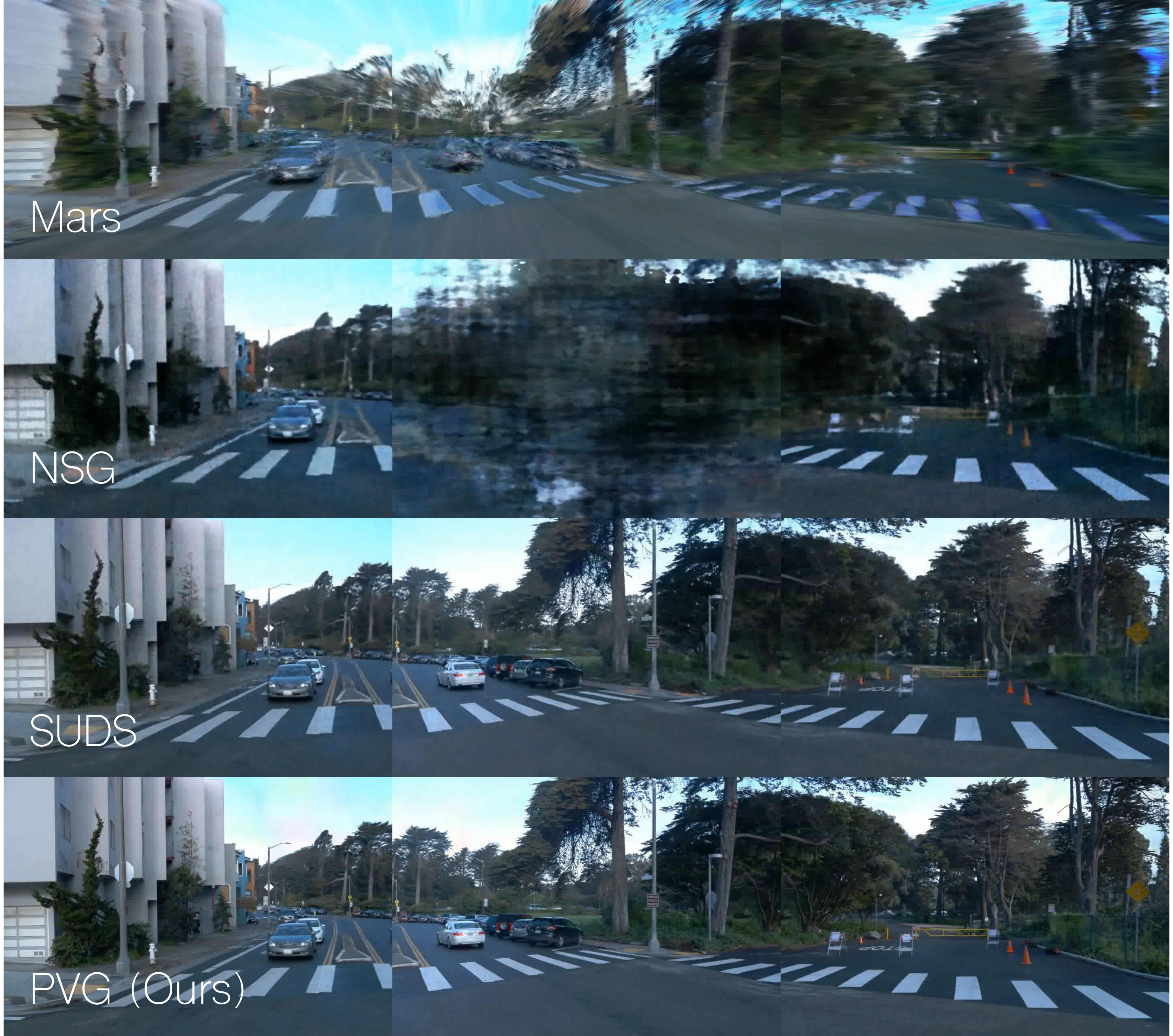
Comparison with static methods


S-NeRF
StreetSurf
3DGS
PVG (Ours)


S-NeRF
StreetSurf
3DGS
PVG (Ours)

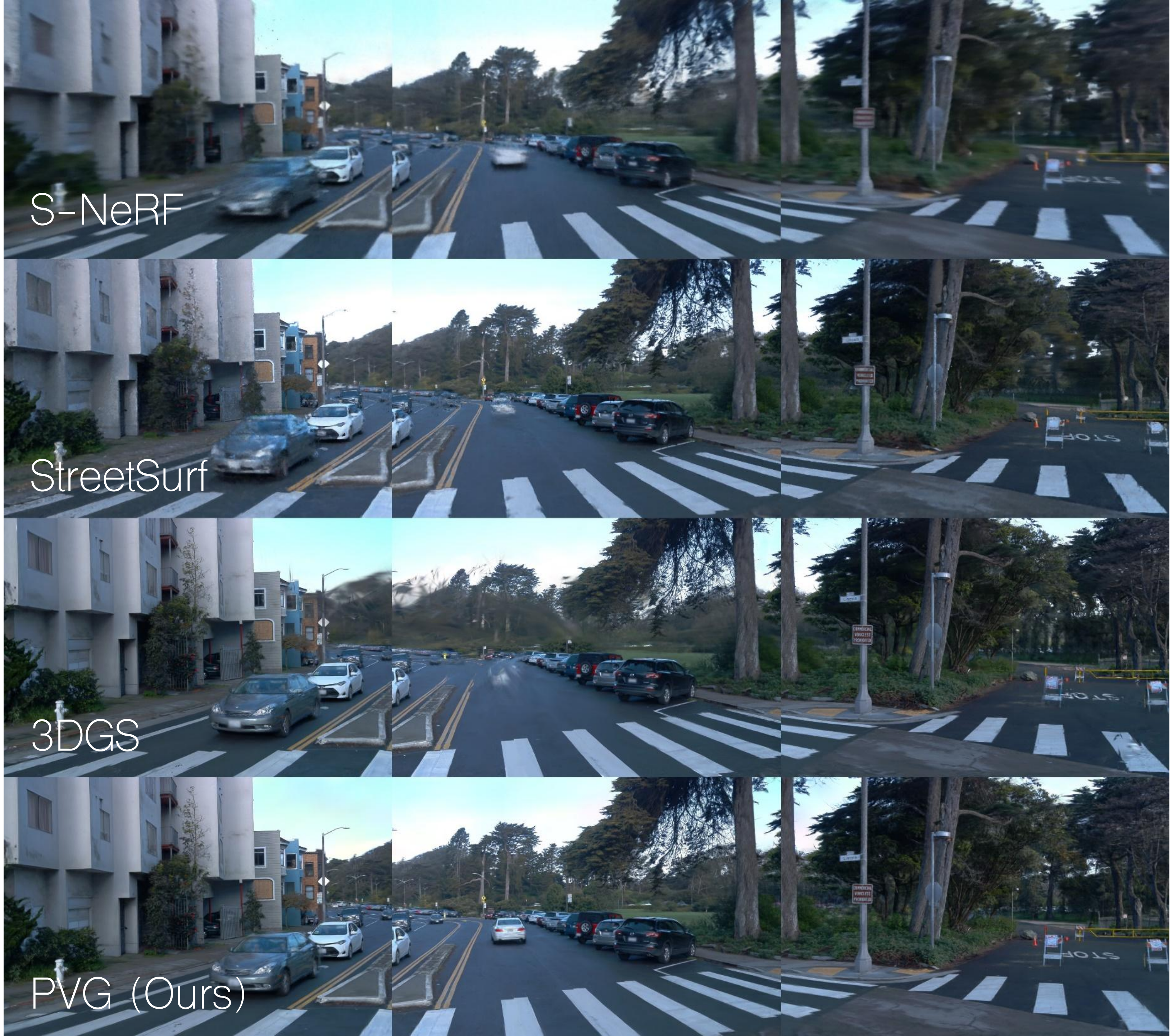
Image Reconstruction on Waymo

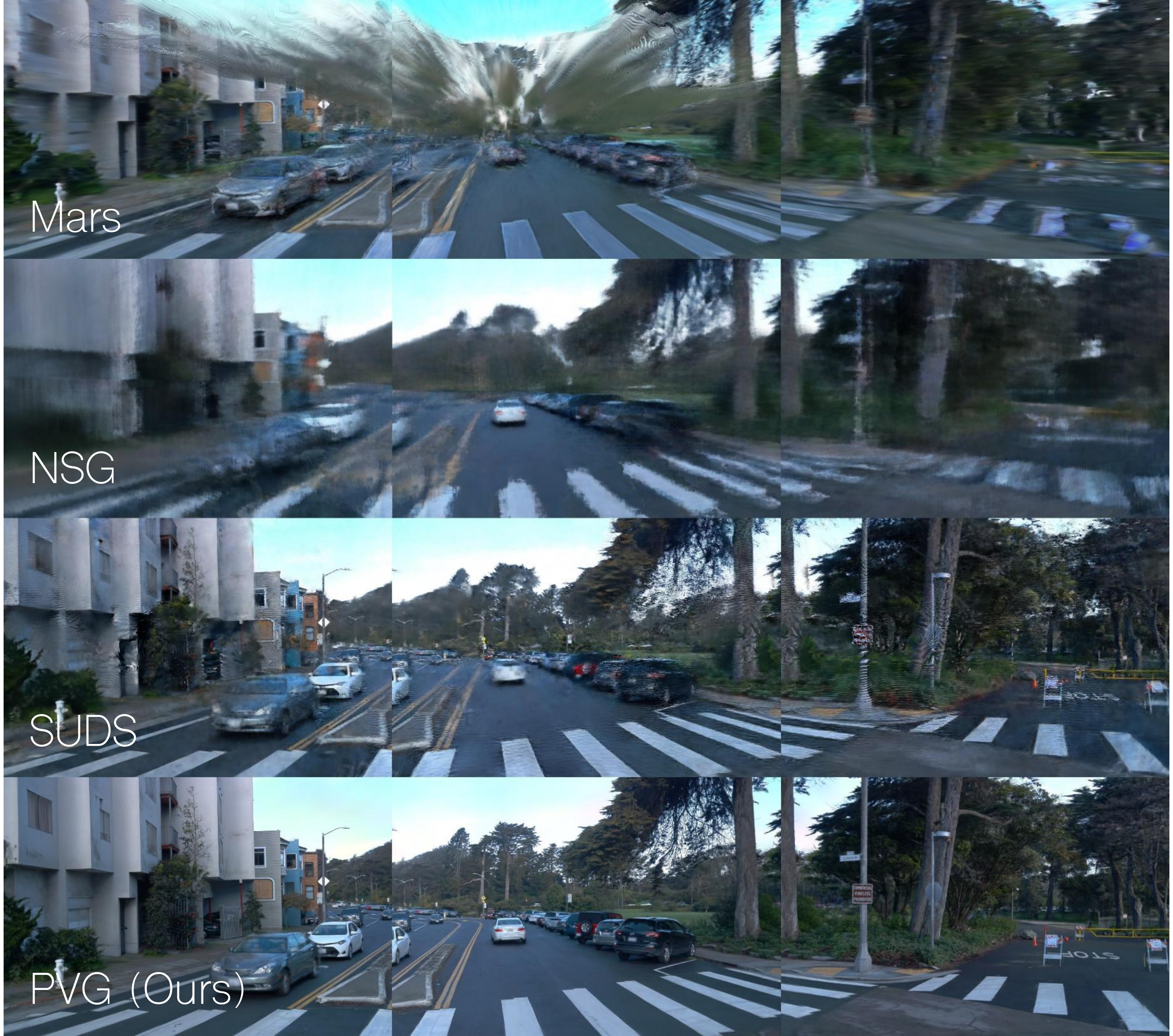
Comparison with dynamic methods


Mars
NSG
SUDS
PVG (Ours)


Mars
NSG
SUDS
PVG (Ours)

Novel View Synthesis on Waymo


S-NeRF
StreetSurf
3DGS
PVG (Ours)



Mars
NSG
SUDS
PVG (Ours)


S-NeRF
StreetSurf
3DGS
PVG (Ours)


Mars
NSG
SUDS
PVG (Ours)

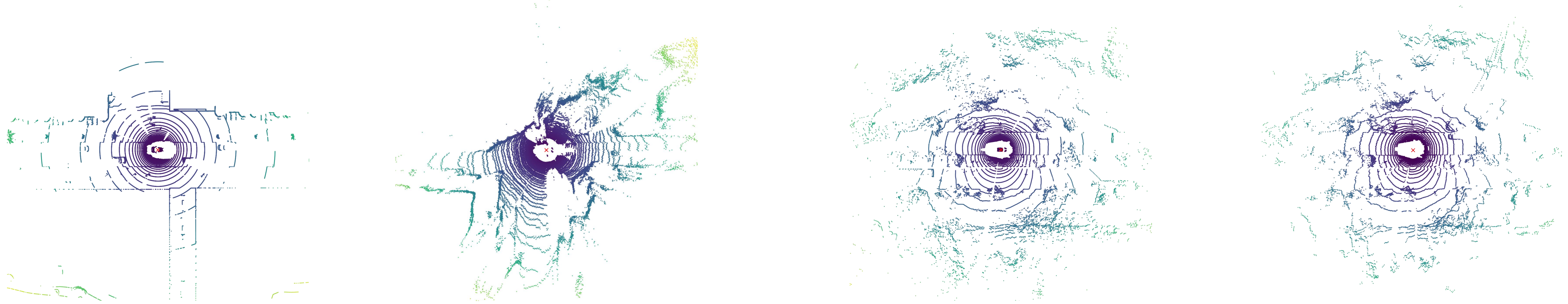

S-NeRF
StreetSurf
3DGS
PVG (Ours)


Mars
NSG
SUDS
PVG (Ours)

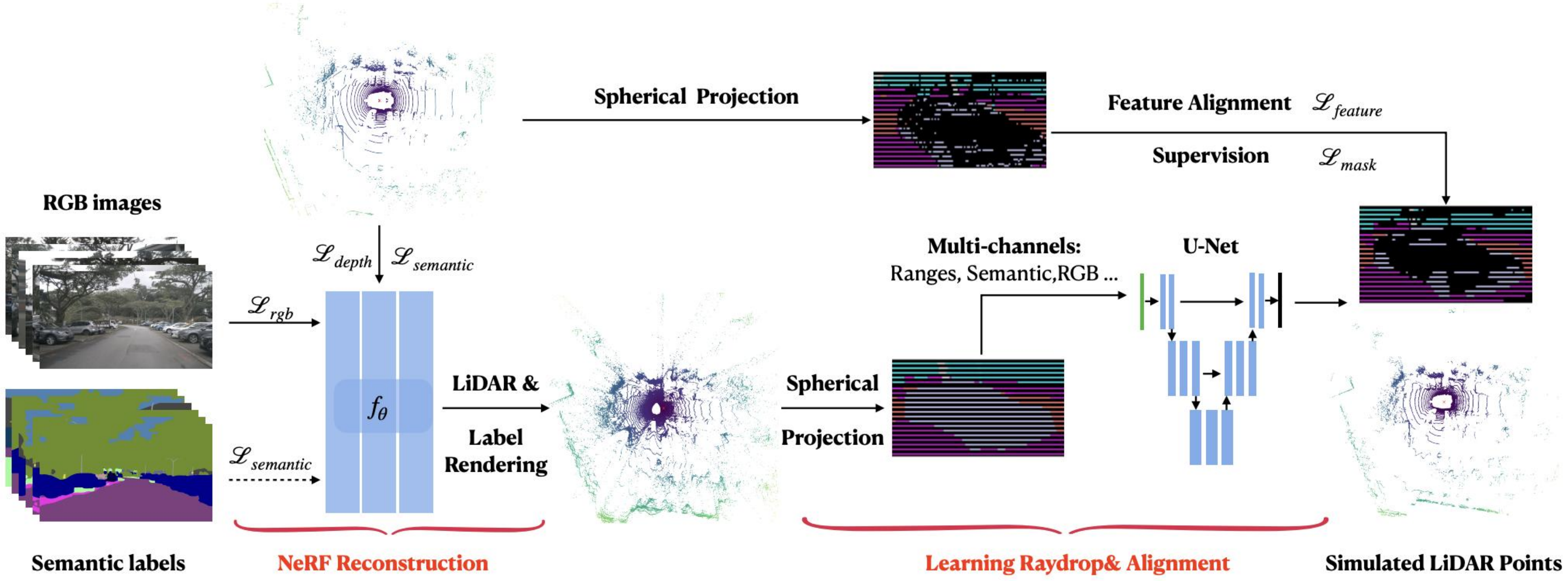

S-NeRF
StreetSurf
3DGS
PVG (Ours)
STOP
STOP
STOP
STOP

Mars
STOP
NSG
SUDS
STOP
STOP
PVG (Ours)
STOP
STOP
STOP

NeRF-LiDAR



(a) CARLA (b) LiDARGen (c) Our NeRF-LiDAR (d) Real LiDAR sensor



NeRF-LiDAR: Generating Realistic LiDAR Point Clouds with Neural Radiance Fields, Junge Zhang, Feihu Zhang, Shaochen Kuan, ***Li Zhang* – *AAAI*** *2024*

Driving Scene Generation

Large-scale 3D urban scene generation

WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

Driving Scene Generation

Large-scale 3D urban scene generation

WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

Driving Scene Generation

Large-scale 3D urban scene generation

Unseen camera intrinsic:
changing focal length.

WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

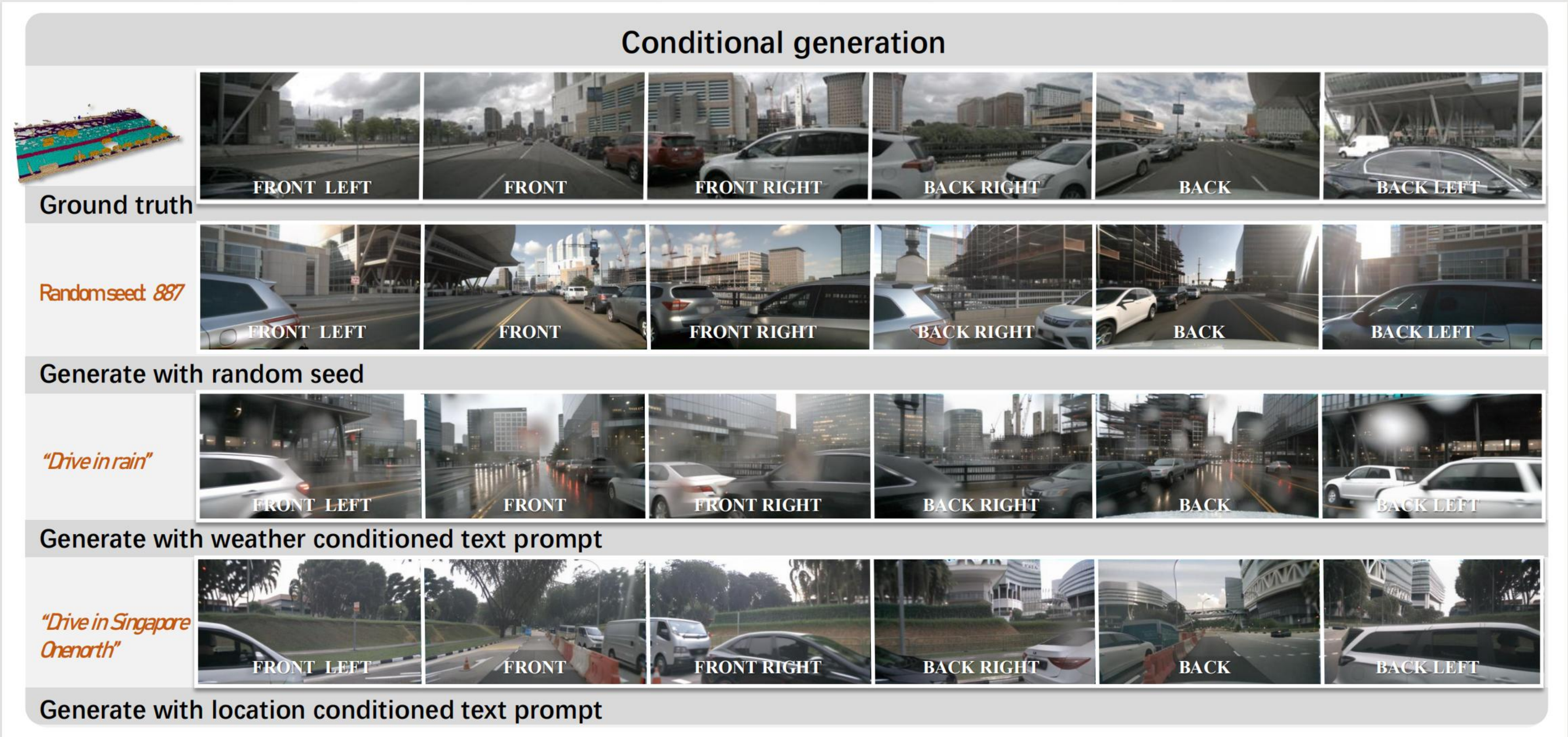
Driving Scene Generation

Large-scale 3D urban scene generation

WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

Driving Scene Generation

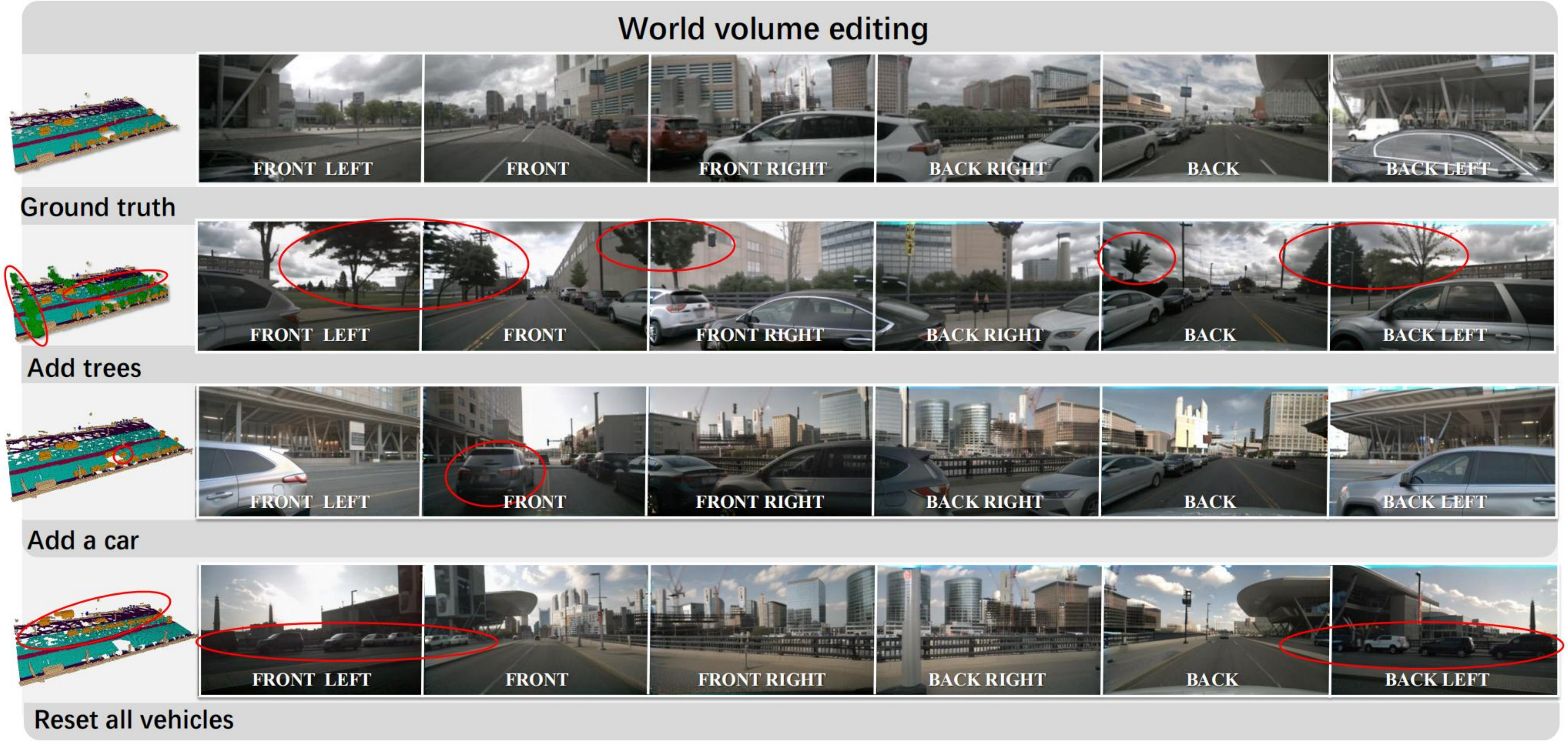
Large-scale 3D urban scene generation



WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

Driving Scene Generation

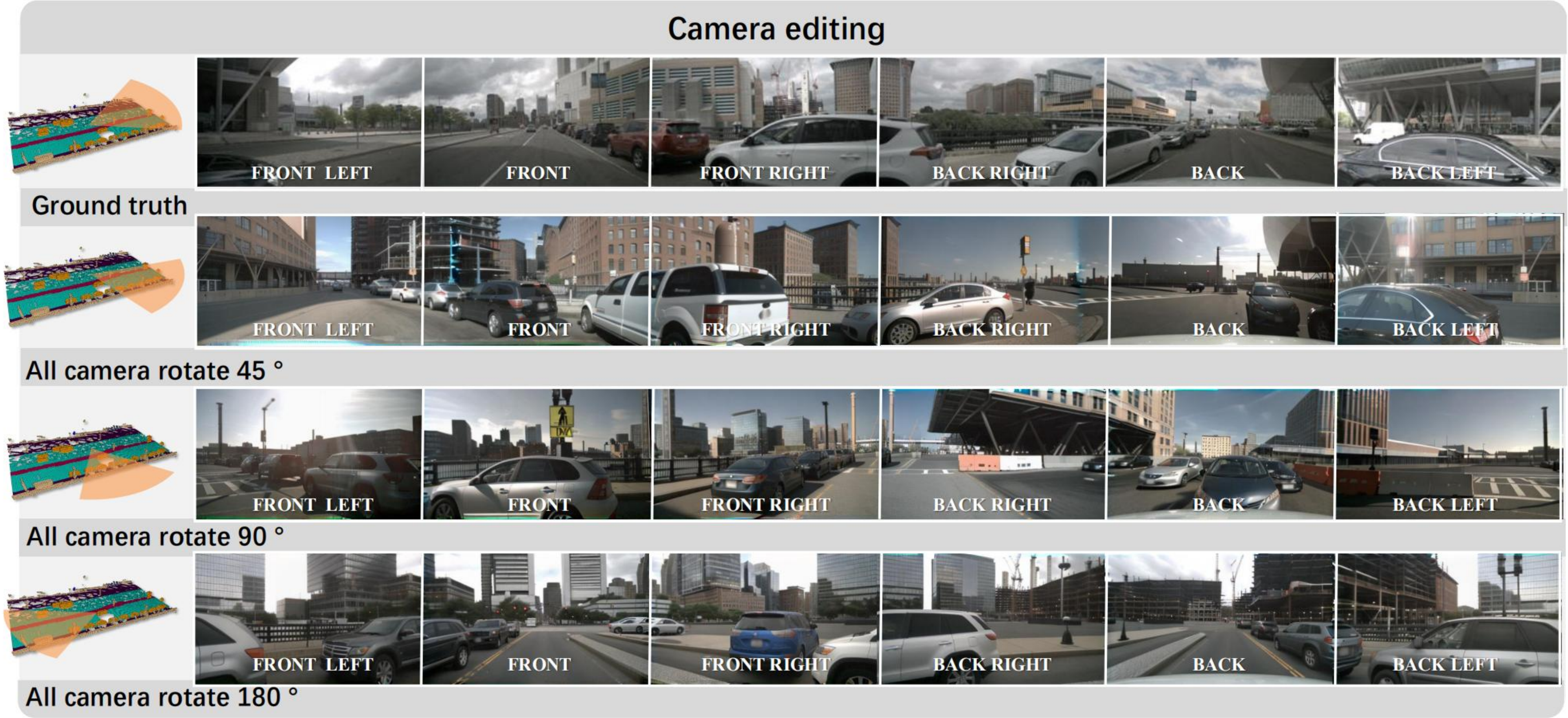
Large-scale 3D urban scene generation



WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

Driving Scene Generation

Large-scale 3D urban scene generation



WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation,*Jiachen Lu, Ze Huang, Jiahui Zhang, Li Zhang - 2024*

路网监测 (RoadNet)

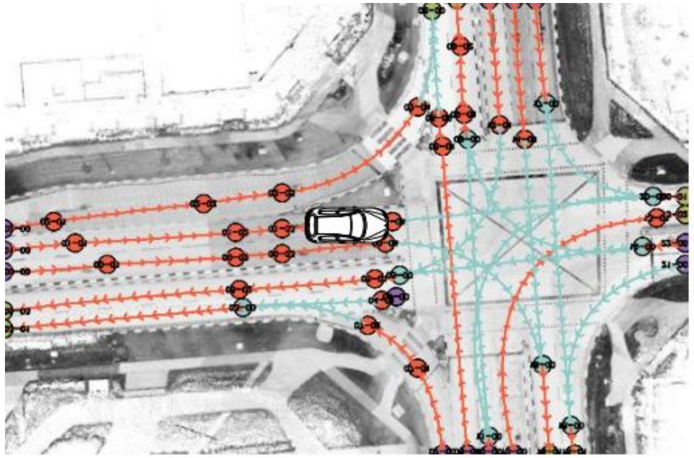
更新慢
成本高

离线高精地图构建 → 在线道路结构感知

实时性
成本低

路网：道路关键点位置、中心线曲线形状、中心线连接关系

环视相机图像

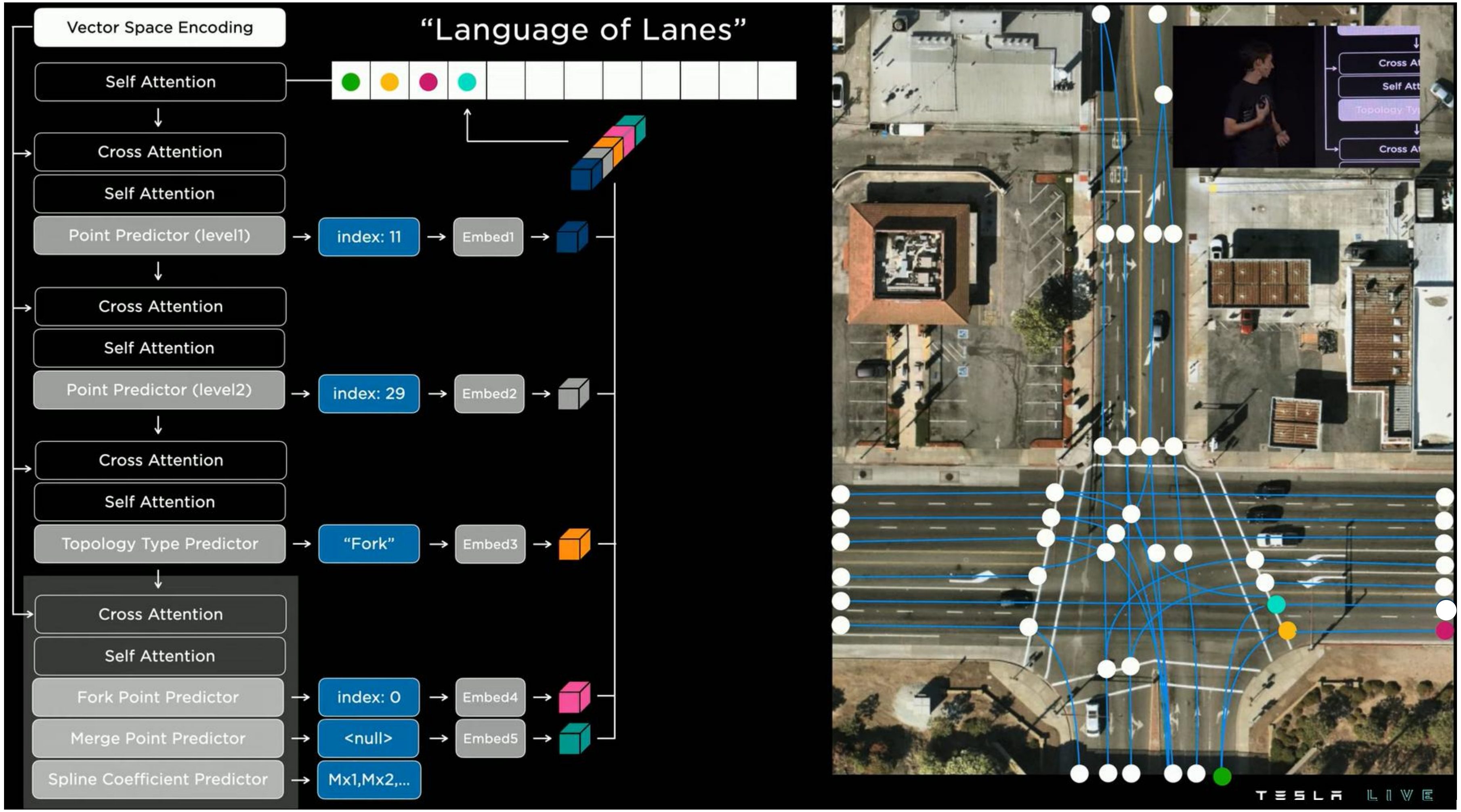


路网

Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach, Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang, ICCV 2023

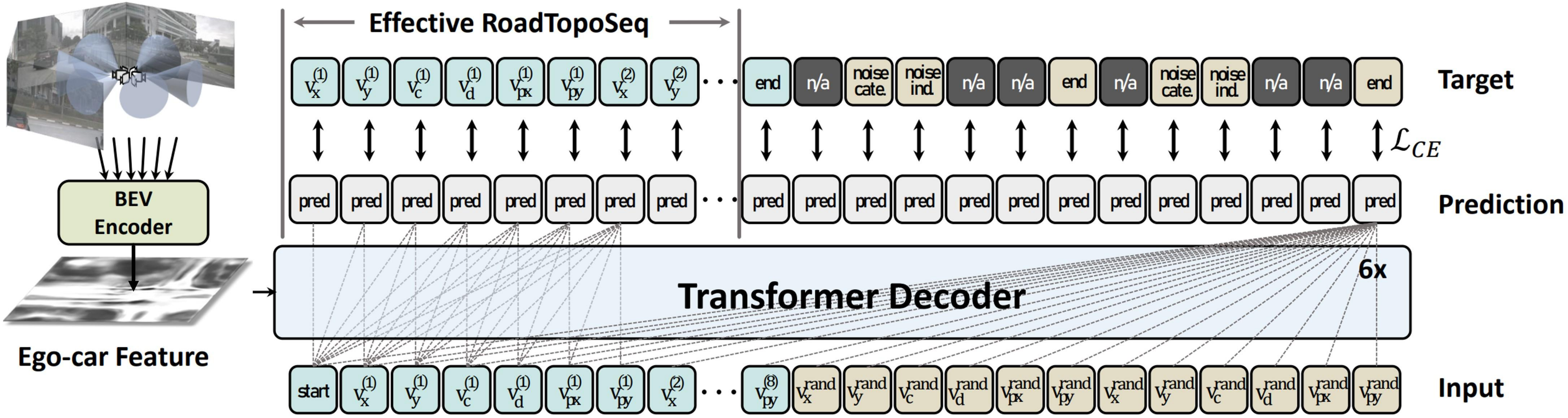
路网监测 (RoadNet)

Tesla “Language of Lanes”



路网监测 (RoadNet)

模型结构（Auto-Regressive）

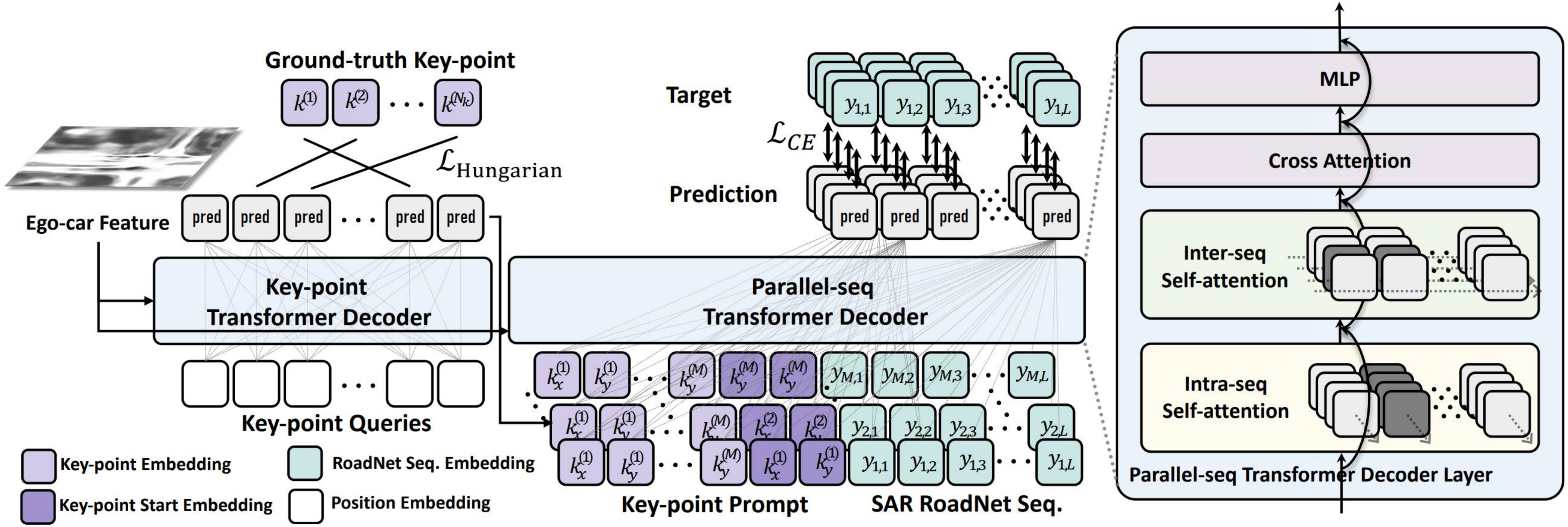


能够实现路网检测，但自回归推理速度很慢，无法满足自动驾驶的需求

Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach, Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang, ICCV 2023

路网监测 (RoadNet)

改进——半自回归：模型结构(Semi-Autoregressive)



1、DETR结构检测有明显视觉特征的点（起始点和分叉点）

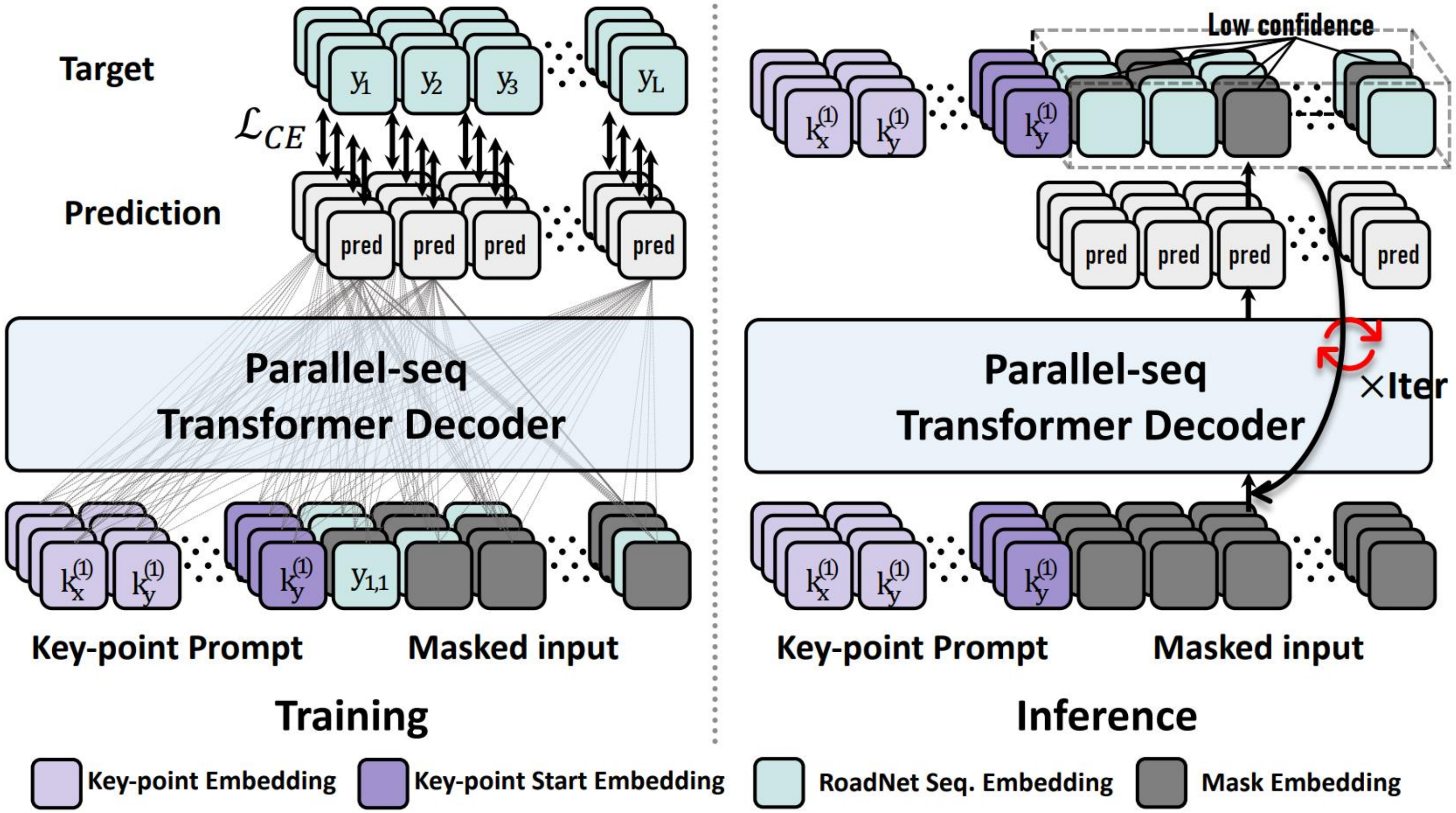
2、半自回归网络结构，子序列内自回归预测，子序列间并行预测

速度比完全自回归提升六倍，且效果优于完全自回归方法

Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach, Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang, ICCV 2023

路网监测 (RoadNet)

进一步改进——masked language modeling 完全并行 (None-Autoregressive)



效果相比半自回归略有下降，但仍优于完全自回归，且速度大幅提升

Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach, Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang, ICCV 2023

路网监测 (RoadNet)

Multi-camera Input

Prediction

GT

Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach, Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang, ICCV 2023

ECCV
TEL AVIV 2022

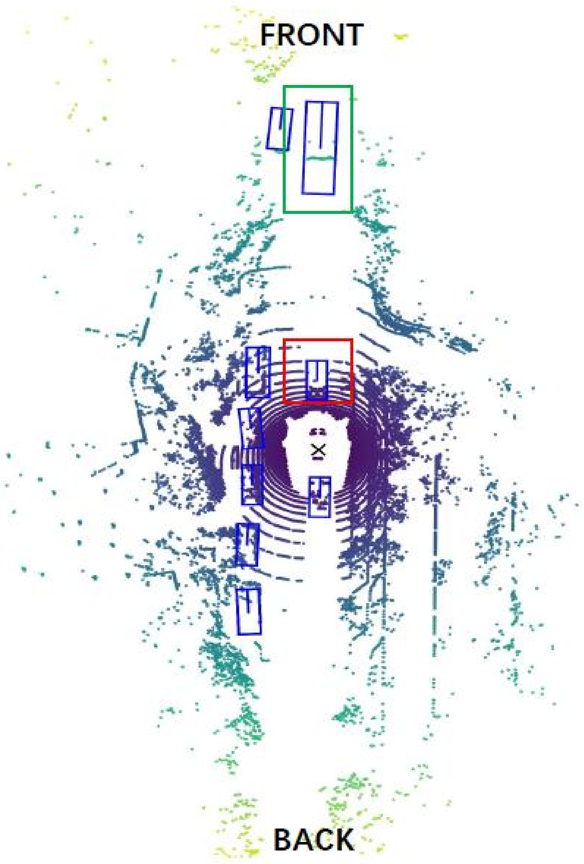
Ego3RT: Learning Ego 3D Representation as Ray Tracing

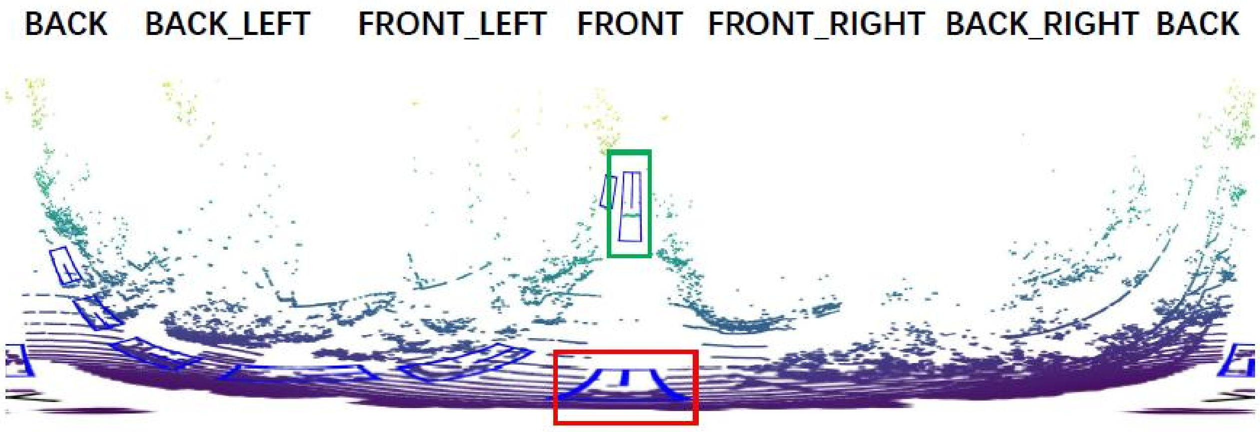
https://fudan-zvg.github.io/Ego3RT

Jiachen Lu[1] Zheyuan Zhou[1] Xiatian Zhu[2] Hang Xu[3] Li Zhang[1]

[1]Fudan University [2]University of Surrey [3]Huawei Noah's Ark Lab

Multi-camera 3D detection (PolarFormer)



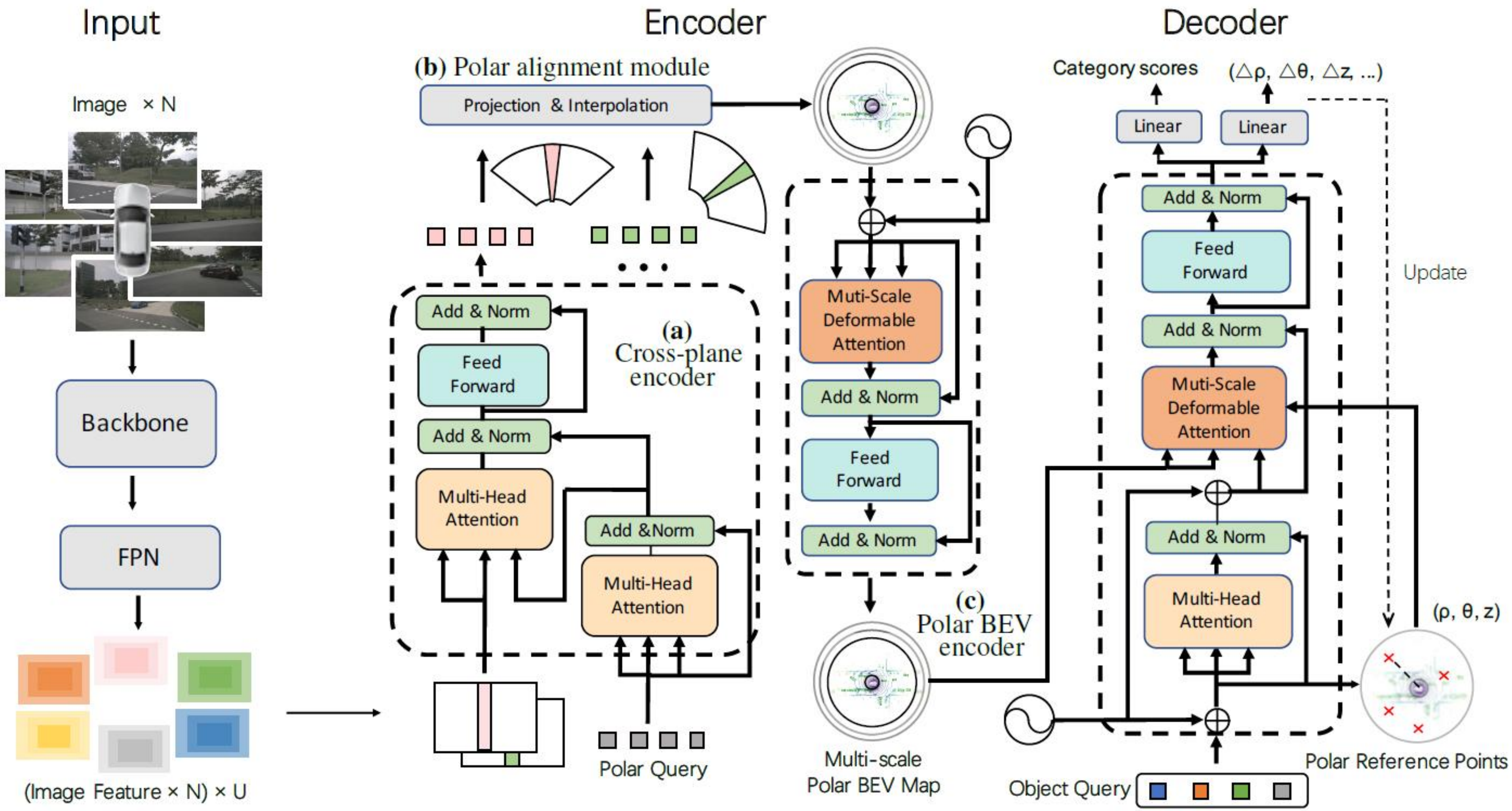


- Disadvantages of Cartesian
  - Down-sampling in non-far region leads to information loss
  - Over-sampling in far region is useless

So we choose **Polar coordinate consistent with imaging process**

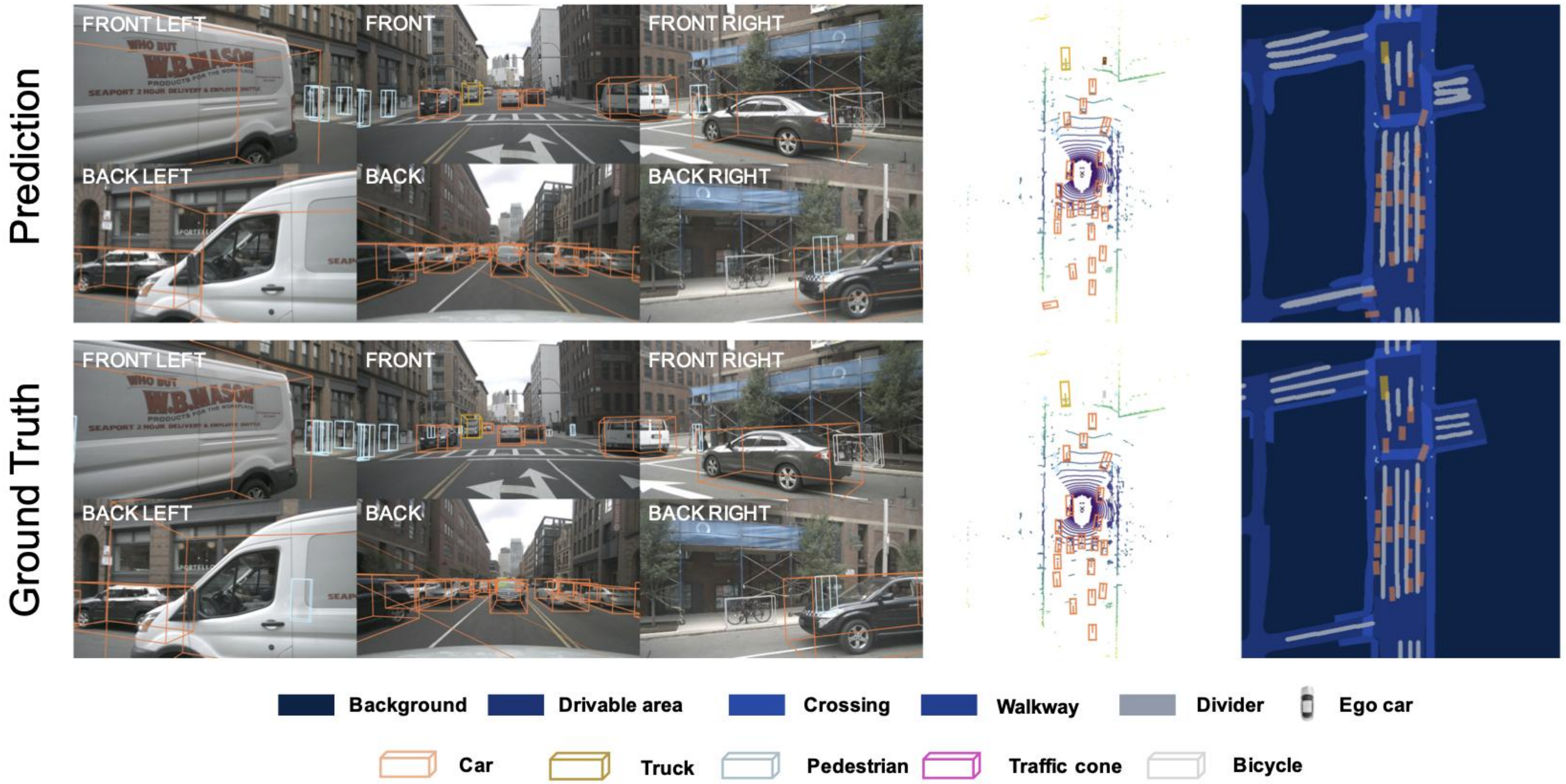
PolarFormer: Multi-camera 3D object detection with polar transformers – *Yanqin Jiang, **Li Zhang**, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, Yu-Gang Jiang* – ***AAAI*** *2023* ***Oral***

Multi-camera 3D detection (PolarFormer)



PolarFormer: Multi-camera 3D object detection with polar transformers – *Yanqin Jiang, **Li Zhang**, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, Yu-Gang Jiang – **AAAI** 2023 **Oral***

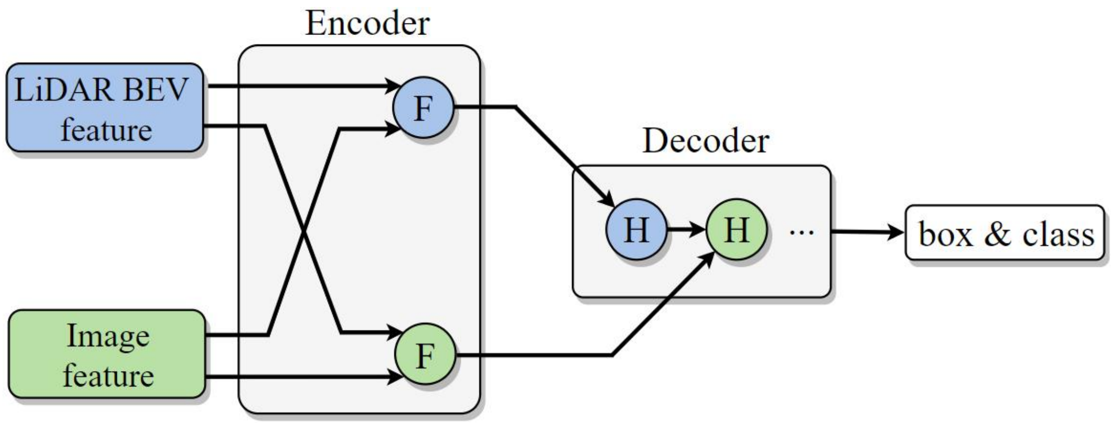
Multi-camera 3D detection (PolarFormer)



PolarFormer: Multi-camera 3D object detection with polar transformers – *Yanqin Jiang, **Li Zhang**, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, Yu-Gang Jiang* – ***AAAI*** *2023* ***Oral***

Multi-sensor 3D detection (DeepInteraction)

- **Encoder-decoder architecture**
  - Set prediction
  - Bipartite matching
- **Modality interaction** in both the encoder and decoder
  - **Representational interaction** in the encoder
  - **Predictive interaction** in the decoder



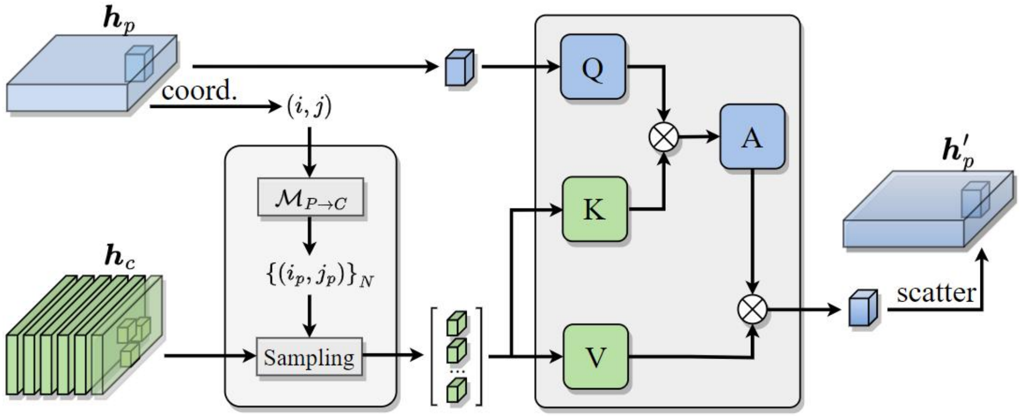
DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Multi-sensor 3D detection (DeepInteraction)

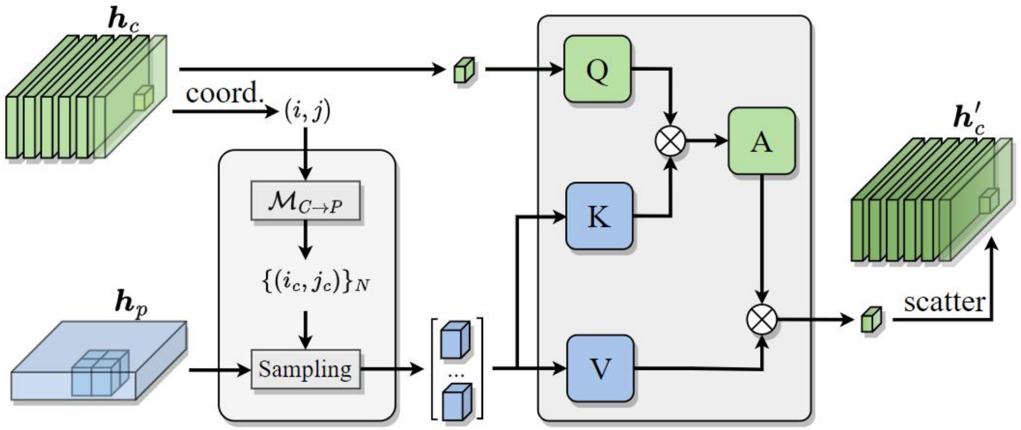
- Encoder aims at constructing more powerful representations via multi-modal representational interaction.

- Hierarchical structure
  - Composed by stacking multiple encoder layers which takes as input two modality-specific scene representations and produces two enhanced representations as output.

- Extensive interaction
  - Multi-modal representational interaction (MMRI)
  - Intra-modal representational learning (IML)
  - Representational interaction



(a) Representational interaction from image to LiDAR

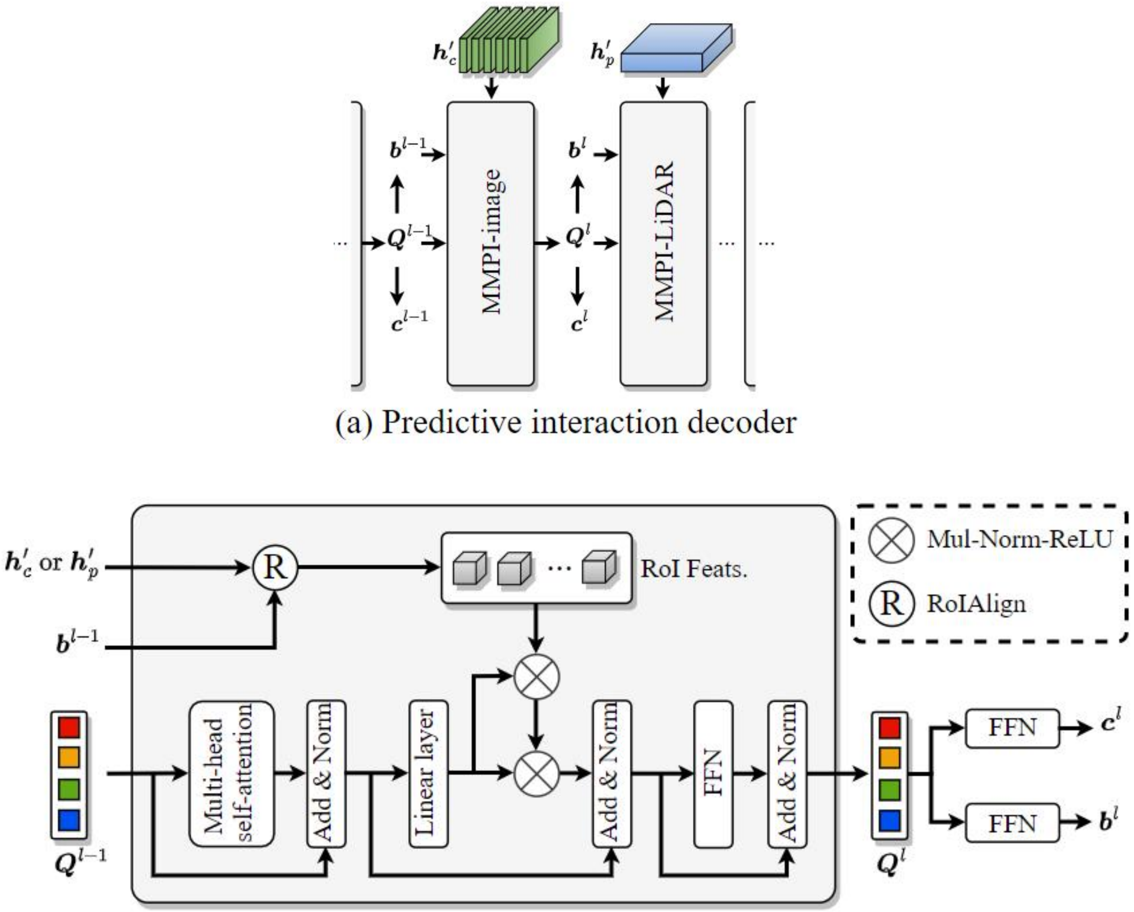


(b) Representational interaction from LiDAR to image

DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Multi-sensor 3D detection (DeepInteraction)

- Iterative query refinement
- Aggregating information from hetero-genous scene representations in an unified manner – **MMPI**
  - MMPI-Image
  - MMPI-LiDAR
- Alternating interaction of two modalities.



(a) Predictive interaction decoder

(b) Multi-modal predictive interaction layer (MMPI)

DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Multi-sensor 3D detection (DeepInteraction)

- We achieve the state-of-the-art performance on the highly competitive nuScenes 3D detection benchmark.
- The ensemble version of DeepInteraction-e now ranks 1st on the nuScenes leaderboard.

Table 1: Comparison with state-of-the-art methods on the nuScenes `test` set. `Metrics`: mAP(%), NDS(%). 'L' and 'C' represent LiDAR and camera, respectively. † denotes test-time augmentation is used. § denotes that test-time augmentation and model ensemble both are applied for testing.

Method	Modality	Backbones		*validation*		*test*	
		Image	LiDAR	mAP↑	NDS↑	mAP↑	NDS↑
BEVDet4D [17]	C	Swin-Base	-	42.1	54.5	45.1	56.9
BEVFormer [25]	C	V99	-	-	-	48.1	56.9
Ego3RT [31]	C	V99	-	47.8	53.4	42.5	47.9
PolarFormer [19]	C	V99	-	50.0	56.2	49.3	57.2
CenterPoint [45]	L	-	VoxelNet	59.6	66.8	60.3	67.3
Focals Conv [8]	L	-	VoxelNet-FocalsConv	61.2	68.1	63.8	70.0
Transfusion-L [1]	L	-	VoxelNet	65.1	70.1	65.5	70.2
LargeKernel3D [9]	L	-	VoxelNet-LargeKernel3D	63.3	69.1	65.3	70.5
FUTR3D [7]	L+C	R101	VoxelNet	64.5	68.3	-	-
PointAugmenting [39]†	L+C	DLA34	VoxelNet	-	-	66.8	71.0
MVP [46]	L+C	DLA34	VoxelNet	67.1	70.8	66.4	70.5
AutoAlignV2 [10]	L+C	CSPNet	VoxelNet	67.1	71.2	68.4	72.4
TransFusion [1]	L+C	R50	VoxelNet	67.5	71.3	68.9	71.6
BEVFusion [26]	L+C	Swin-Tiny	VoxelNet	67.9	71.0	69.2	71.8
BEVFusion [30]	L+C	Swin-Tiny	VoxelNet	68.5	71.4	70.2	72.9
DeepInteraction-base	L+C	R50	VoxelNet	**69.9**	**72.6**	**70.8**	**73.4**
Focals Conv-F [8]†	L+C	R50	VoxelNet-FocalsConv	67.1	71.5	70.1	73.6
LargeKernel3D-F [9]†	L+C	R50	VoxelNet-LargeKernel	-	-	71.1	74.2
DeepInteraction-large†	L+C	Swin-Tiny	VoxelNet	**72.6**	**74.4**	**74.1**	**75.5**
BEVFusion-e [30]§	L+C	Swin-Tiny	VoxelNet	73.7	74.9	75.0	76.1
DeepInteraction-e§	L+C	Swin-Tiny	VoxelNet	**73.9**	**75.0**	**75.6**	**76.3**

DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Leaderboard test

nuScenes 3D detection

NUSCENES by Motional — Home | nuPlan

nuScenes detection task

Leaderboard

Search: | Export as JSON

Date	Name	Modalities	Map data	External data	mAP
		Camera	All	All	
2022-02-08	FudanZVG-TPD-e	Camera	no	yes	0.4
2021-12-19	BEVDet	Camera	no	yes	0.424
2021-10-13	DETR3D	Camera	no	yes	0.412
2021-06-15	DD3D	Camera	no	yes	0.418
2021-12-18	BEVDet-pure	Camera	no	no	0.398
2022-01-18	FudanZVG-TPD	Camera	no	no	0.401
2021-11-12	IPD3D	Camera	no	no	0.387

NUSCENES by Motional — Home | nuPlan | nuScen

nuScenes detection task

Leaderboard

Search: | Export as JSON

Date	Name	Modalities	Map data	External data	mAP	mATE (m)
		Camera	All	All		
2022-04-11	bevdepth	Camera	no	yes	0.503	0.445
2022-06-01	PETRv2	Camera	no	yes	0.490	0.561
2022-05-18	PolarFormer	Camera	no	yes	0.493	0.556
2022-04-18	BEVDet4D	Camera	no	no	0.451	0.511
2022-03-10	BEVFormer	Camera	no	yes	0.481	0.582
2022-05-11	UVTR-Camera	Camera	no	yes	0.472	0.577
2022-05-16	PolarFormer-pure	Camera	no	no	0.456	0.610

NUSCENES by Motional — Home | nuPlan | nuScenes | nuImages | nuReality

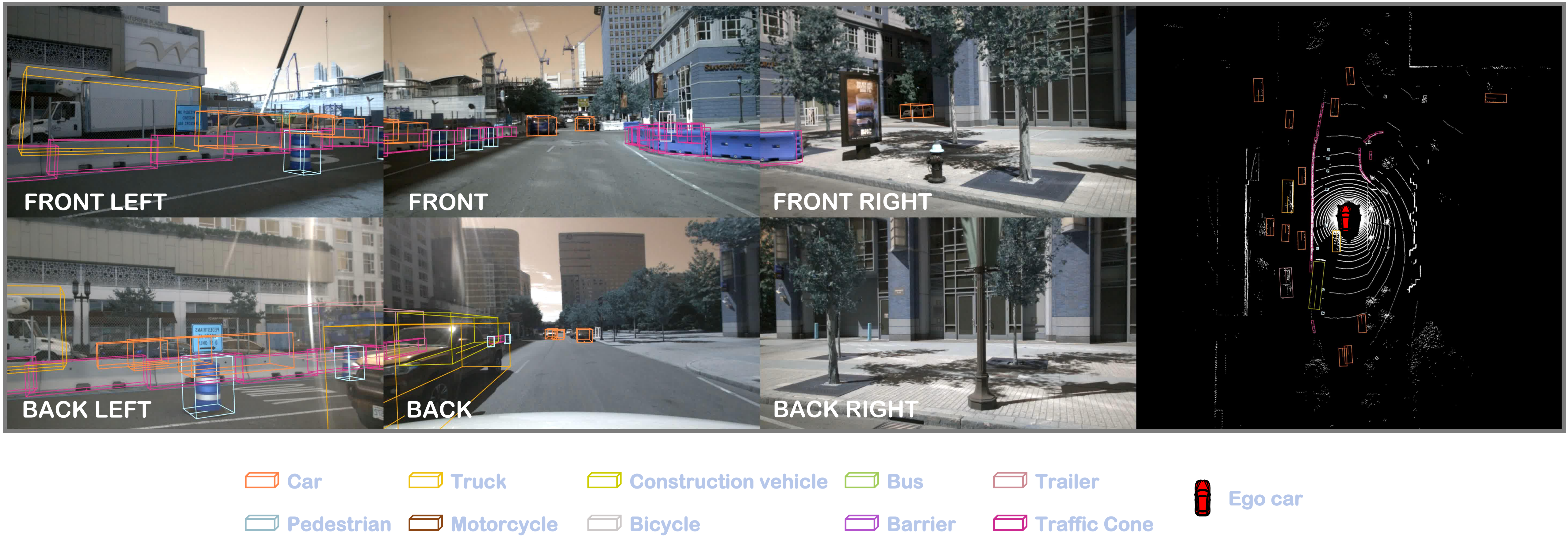
nuScenes detection task

Leaderboard

Search: | Export as JSON | Lidar track | Vision trac

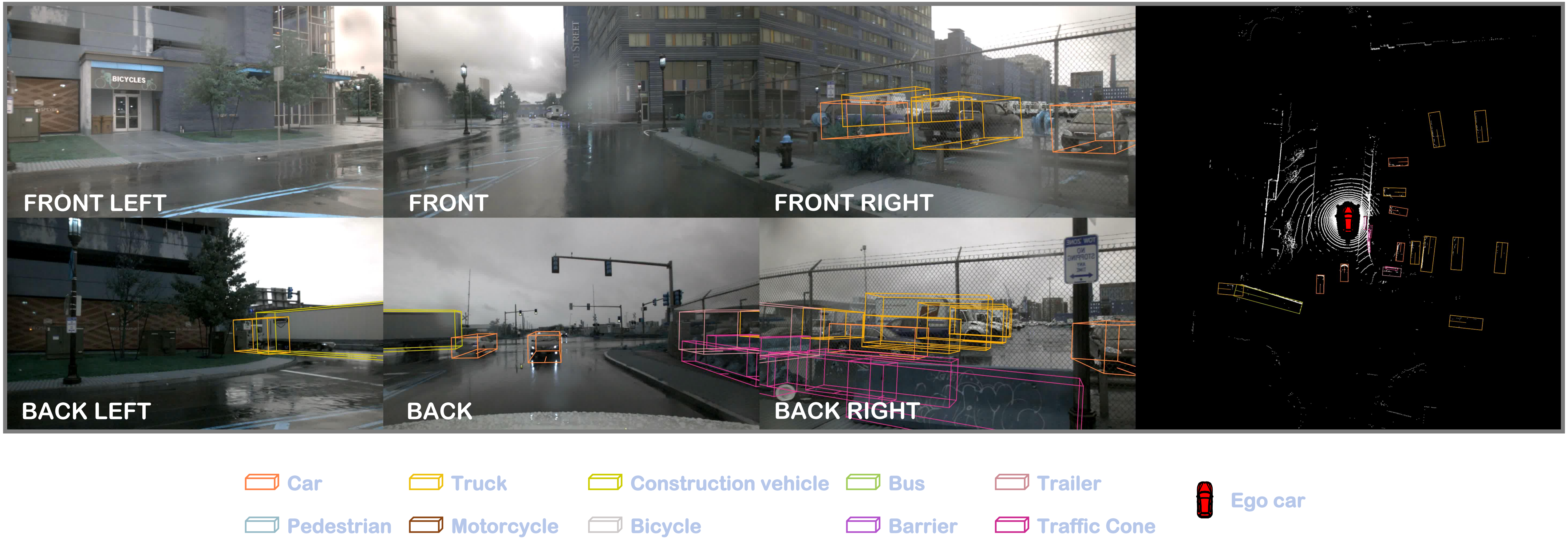
Method					Metrics					
Date	Name	Modalities	Map data	External data	mAP	mATE (m)	mASE (1-IOU)	mAOE (rad)	mAVE (m/s)	mAAE (1-acc)
		Any	All	All						
2022-06-27	DeepInteraction-e	Camera, Lidar	no	no	0.756	0.235	0.233	0.328	0.226	0.130
2022-06-03	BEVFusion-e	Camera, Lidar	no	no	0.750	0.242	0.227	0.320	0.222	0.130
2022-06-26	DeepInteraction-lai	Camera, Lidar	no	no	0.741	0.244	0.232	0.322	0.223	0.133
2022-01-13	FusionVPE	Camera, Lidar	no	no	0.733	0.235	0.227	0.284	0.243	0.128
2021-05-25	Centerpoint-Fusion	Camera, Lidar, R...	no	yes	0.724	0.237	0.227	0.318	0.211	0.133
2022-06-16	LargeKernel-F	Camera, Lidar	no	no	0.711	0.236	0.228	0.298	0.241	0.131

Multi-sensor 3D detection (DeepInteraction)



DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Multi-sensor 3D detection (DeepInteraction)



DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

Multi-sensor 3D detection (DeepInteraction)

DeepInteraction: 3D Object Detection via Modality Interaction – *Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu,* ***Li Zhang*** – ***NeurIPS*** *2022* ***Spotlight***

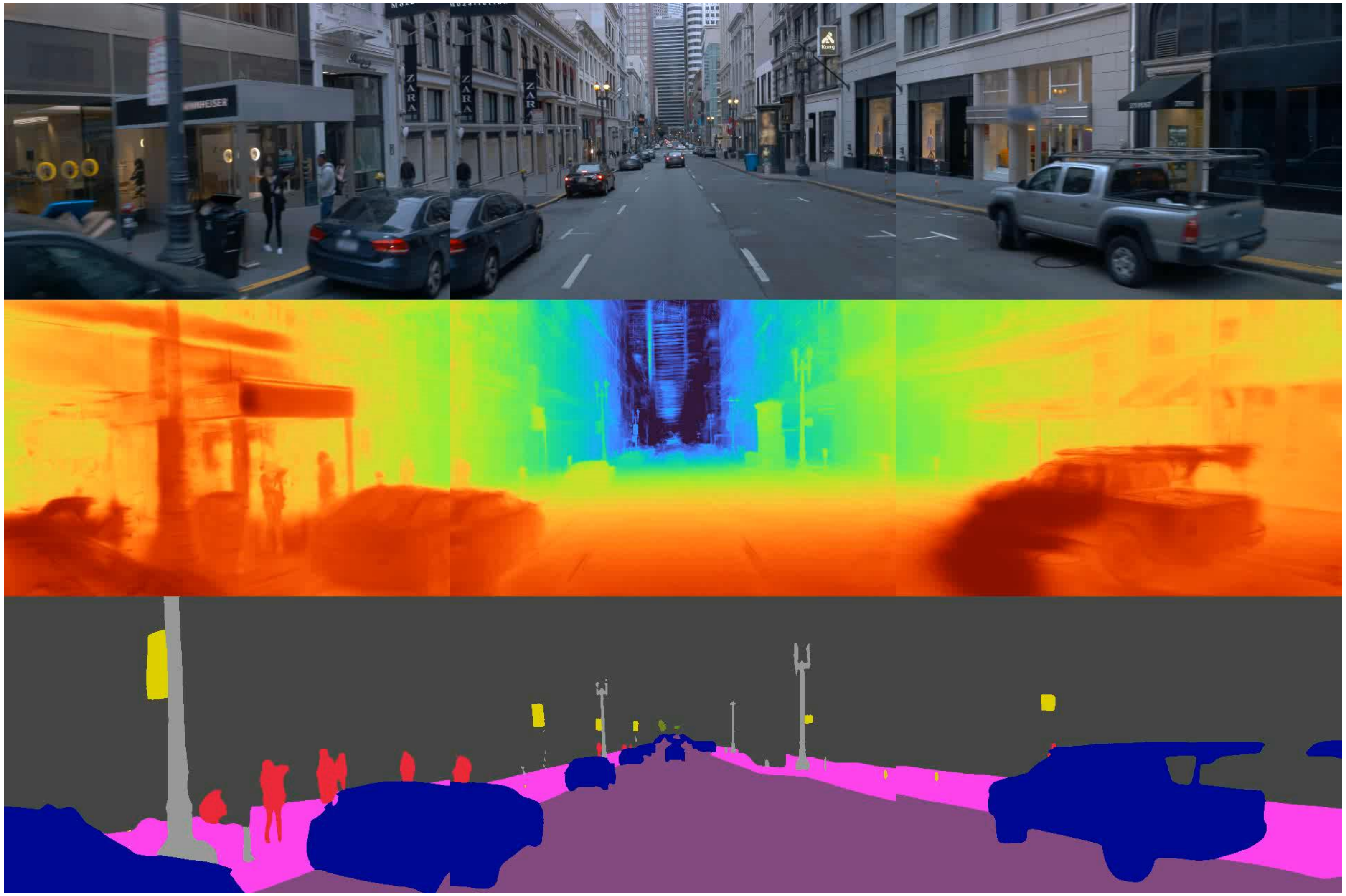
S-Agents: Self-organizing Agents in Open-ended Environments

Jiaqi Chen*   Yuxian Jiang*   Jiachen Lu   Li Zhang

Fudan University

https://github.com/fudan-zvg/S-Agents

THANK YOU!



Code

Project