香 港 中 文 大 學（深 圳）
The Chinese University of Hong Kong, Shenzhen

NEURAL INFORMATION PROCESSING SYSTEMS

Towards Flexible 3D Perception: Object-Centric Occupancy Completion Augments 3D Object Detection

NeurIPS 2024

Chaoda Zheng[2,1], Feng Wang[3], Naiyan Wang[4], Shuguang Cui[2,1], Zhen Li[2,1]

[1]FNii-Shenzhen , [2]SSE, CUHK-Shenzhen, [3]TuSimple, [4]Xiaomi EV

未来智联网络研究院

香 港 中 文 大 學（深 圳）
The Chinese University of Hong Kong, Shenzhen

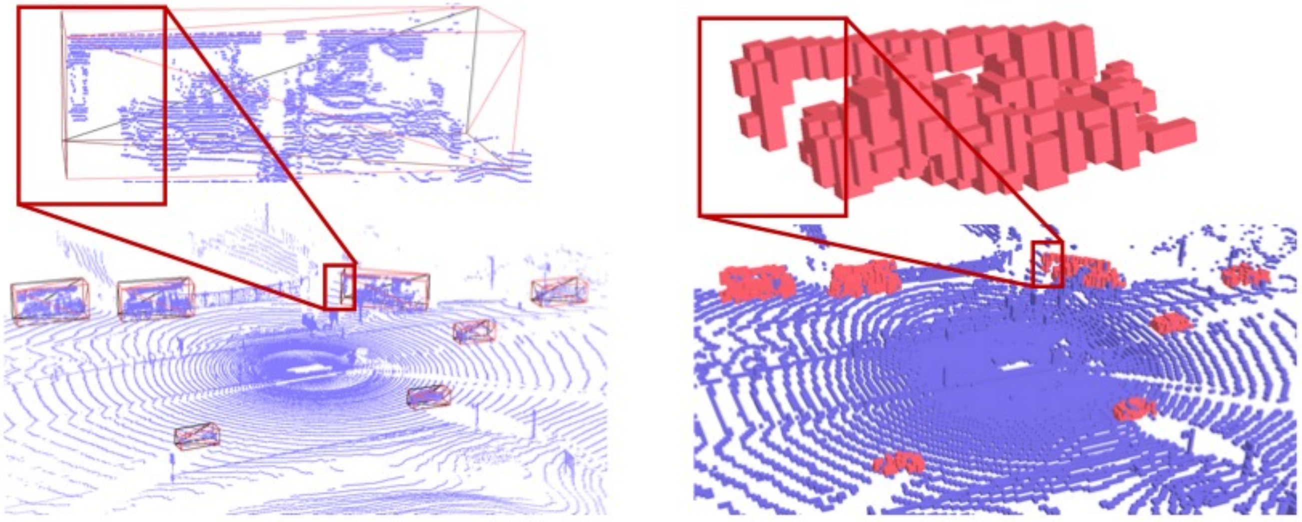
Background: Bounding Box vs. Occupancy

Scene Level Occupancy Only:

- **Low** resolution in real-time applications due to computational constrains.
- **Jagged** voxels due to coordinate misalignments

Detection Bbox Only:

- fail to capture the **intricate** shape details.



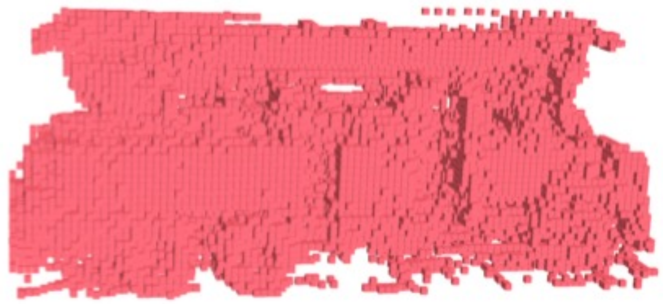
(a) Bounding Boxes (b) Scene-Level Occupancy

Object-Centric Occupancy Representation

Object-Centric Occupancy Representation

- Only focus on **foreground** objects
- No jagged voxels

- Bbox + Object-Centric Occupancy
  - **Lower** Computational Cost
    - Support **High Resolution** Occupancy inside Bboxes
    - No occupancy outside Bboxes
  - **Flexible** Shape Representation



(f) Object-Centric Occupancy

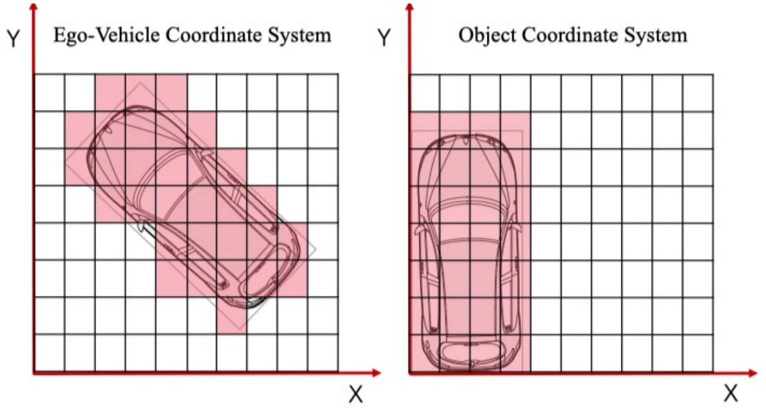
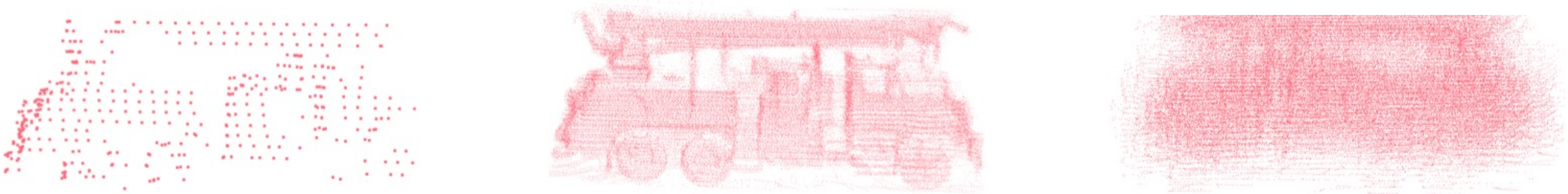


Figure 3: Occupancy grids defined in the ego-vehicle (left) and object-centric (right) coordinate systems. The object shape is jagged in the ego-vehicle occupancy grid due to coordinate misalignment.

How to generate object-centric occupancy ?

Generating Occupancy online is **non-trivial** even with LiDAR.

- Single-Scan LiDAR is too **sparse**.
- Dense Occupancy for dynamic objects relies on **accurate** detection & tracking.



(c) Single Object LiDAR Scan (d) Aggregated Scans using GT Boxes (e) Aggregated Scans using Noisy boxes

Object-centric occupancy completion via a neural network

1. Building the Object-Centric Occupancy Dataset

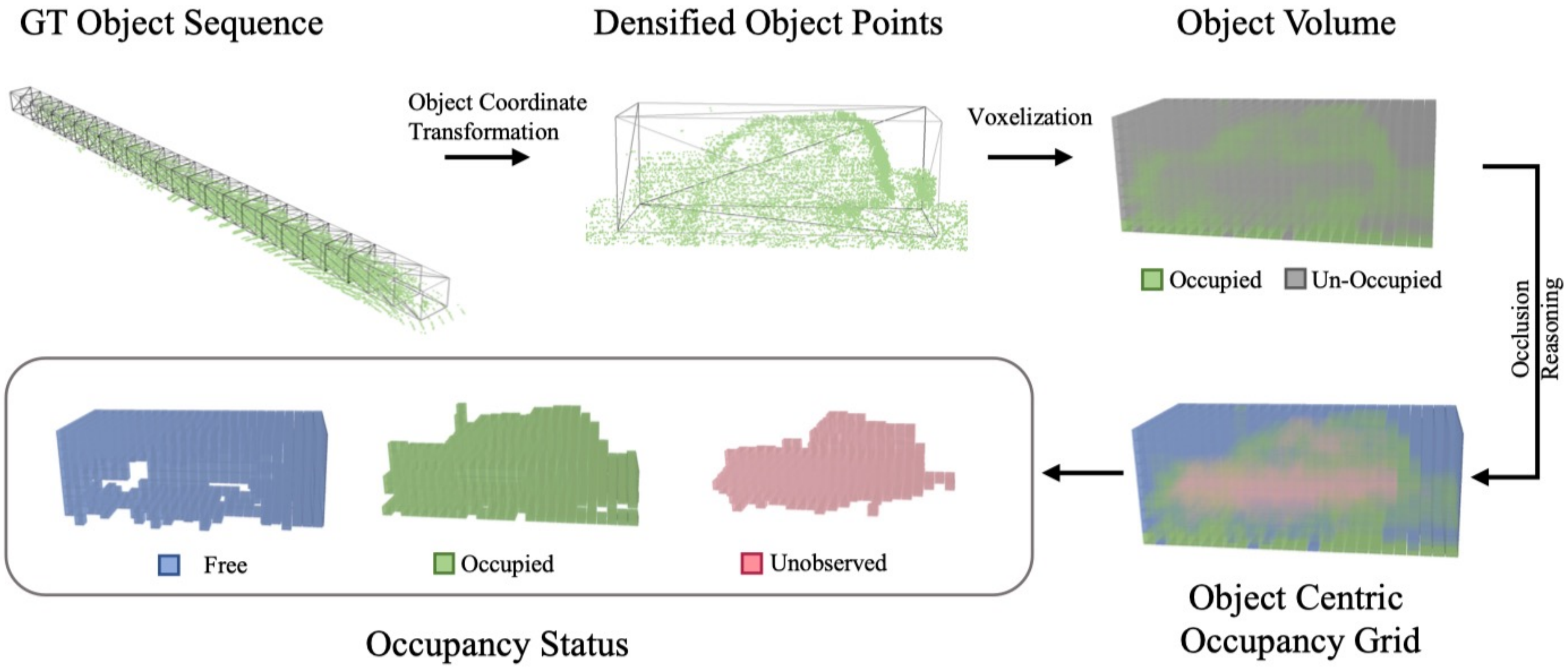


Figure 6: Our object-centric occupancy annotation pipeline.

1. Building the Object-Centric Occupancy Dataset

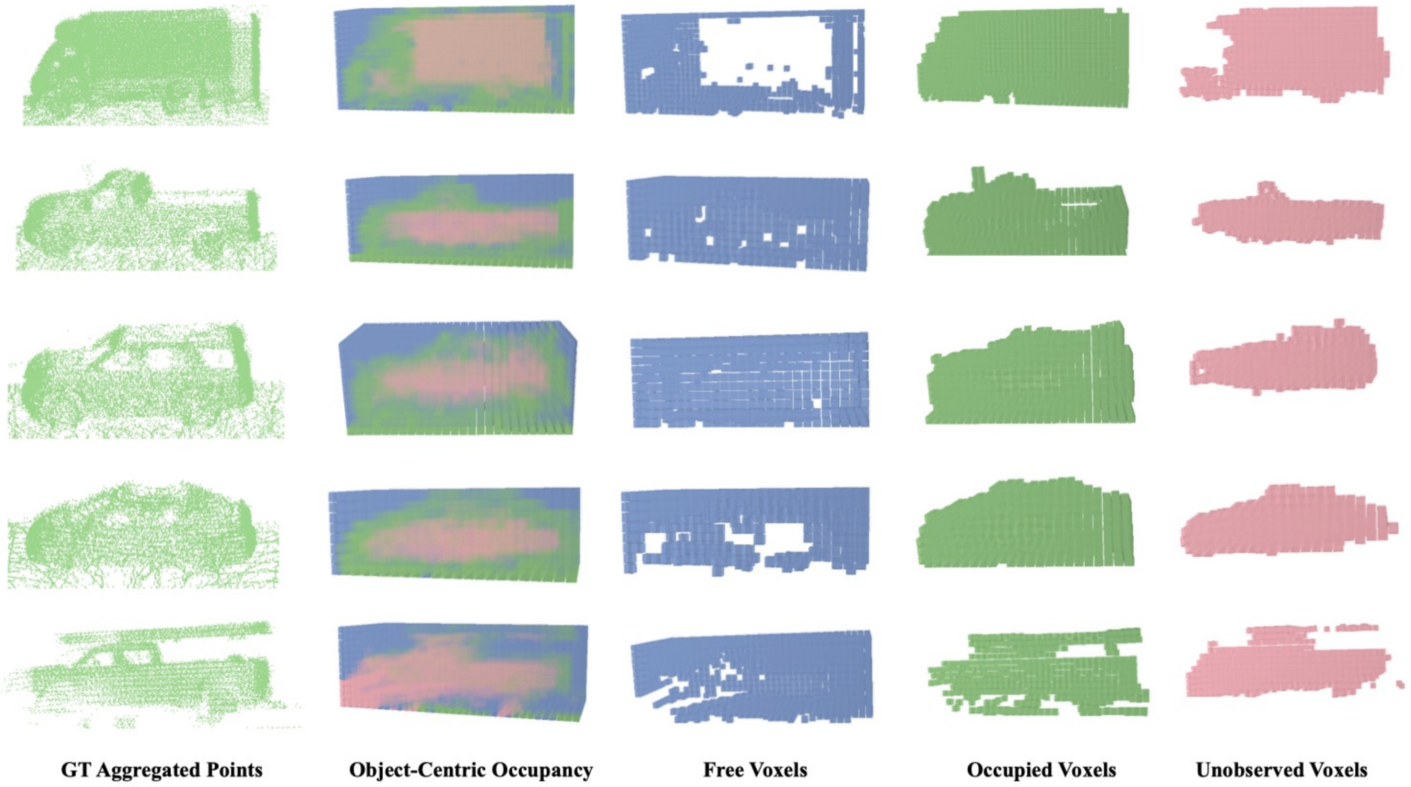


Figure 7: Visualization of our object-centric occupancy annotations. The first colume shows the GT-aggregated LiDAR points. The second column shows our annotated object-centric occupancy volume. The last three columns respectively show the occupancy at free, occupied and unobserved status.

2. Sequence-based Occupancy Completion Network

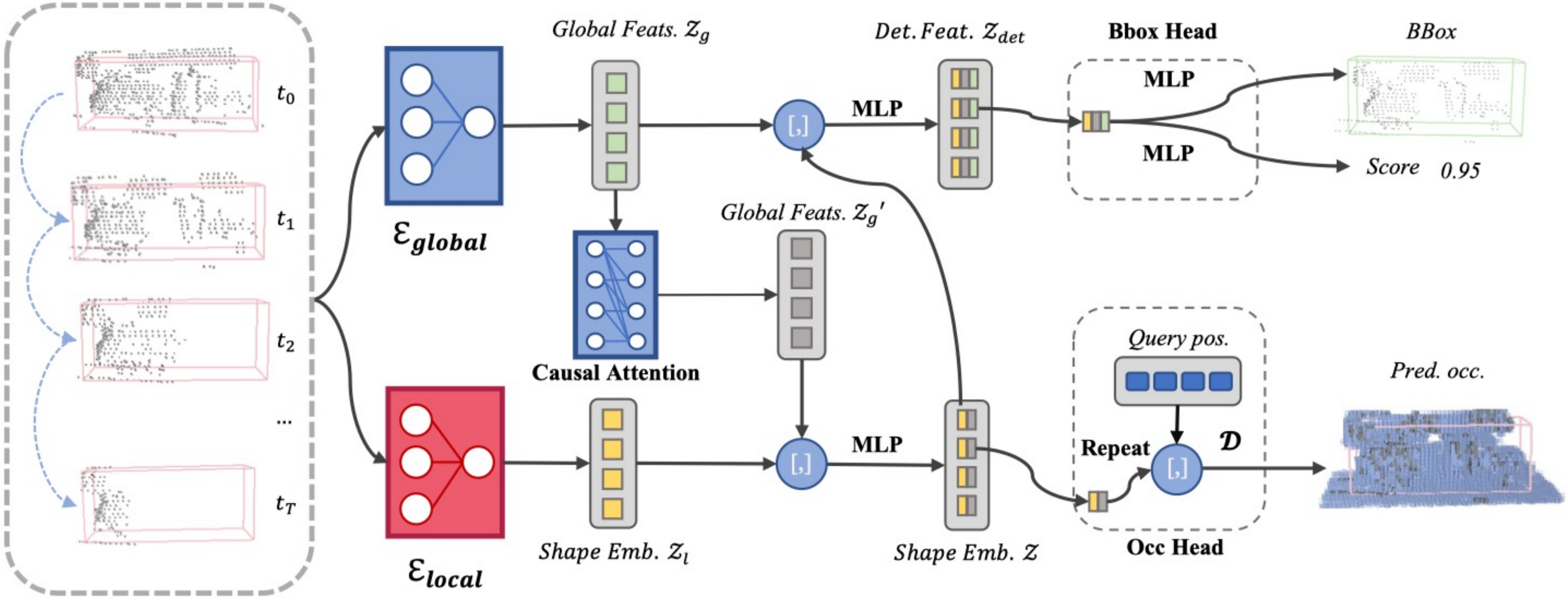


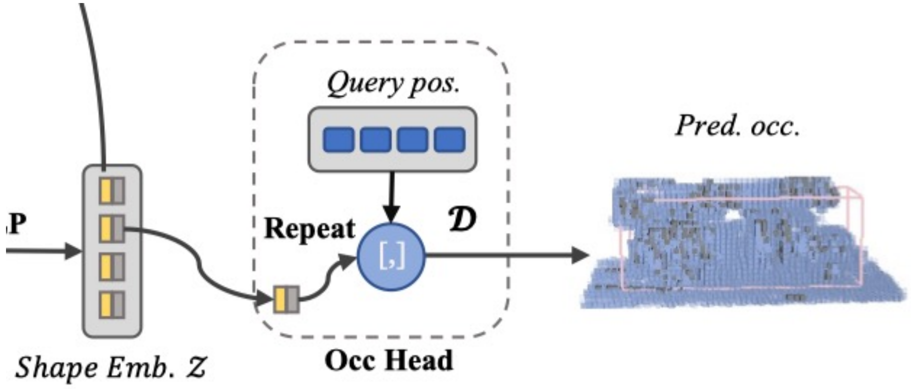
Figure 4: Architecture overview. The network takes a noisy object sequence as input and outputs the complete object-centric occupancy volume and refined bounding box for each proposal. The notation [,] denotes the concatenation operation. 'Global'/'local' indicates features from global/local coordiante system.

2. Dynamic-Size Occupancy Generation

Implicit shape decoder supports dynamic-size occupancy generation:

$$\mathcal{D} : \mathbb{R}^e \times \mathbb{R}^3 \mapsto \mathbb{R}_{[0,1]}$$

$$p = \mathcal{D}(z, q),$$



- z: a fixed-length embedding depicting the geometrics within the RoI.
- q: a query position
- p: the occupancy status at position q

Experiments – Shape Completion

Tracklet Inputs	Method	IoU %	mIoU (track) %	mIoU (box) %
GT track	Baseline	61.35	62.19	63.46
	Ours	**69.15**	**64.05**	**67.91**
GT track + noise	Baseline	50.39	45.21	48.59
	Ours	**64.92**	**60.70**	**63.78**
	Ours-E	*69.30*	*64.11*	*68.04*
FSD track	Baseline	44.28	34.77	42.61
	Ours	**62.84**	**54.12**	**61.58**
	Ours-E	*68.38*	*60.96*	*67.22*
CP track	Baseline	40.45	26.69	37.29
	Ours	**57.99**	**44.94**	**55.10**
	Ours-E	*65.80*	*56.81*	*64.29*

Table 1: Shape completion results on WOD val set. "-E" denotes using extrapolated results outside the RoIs.

Experiments – Shape Completion

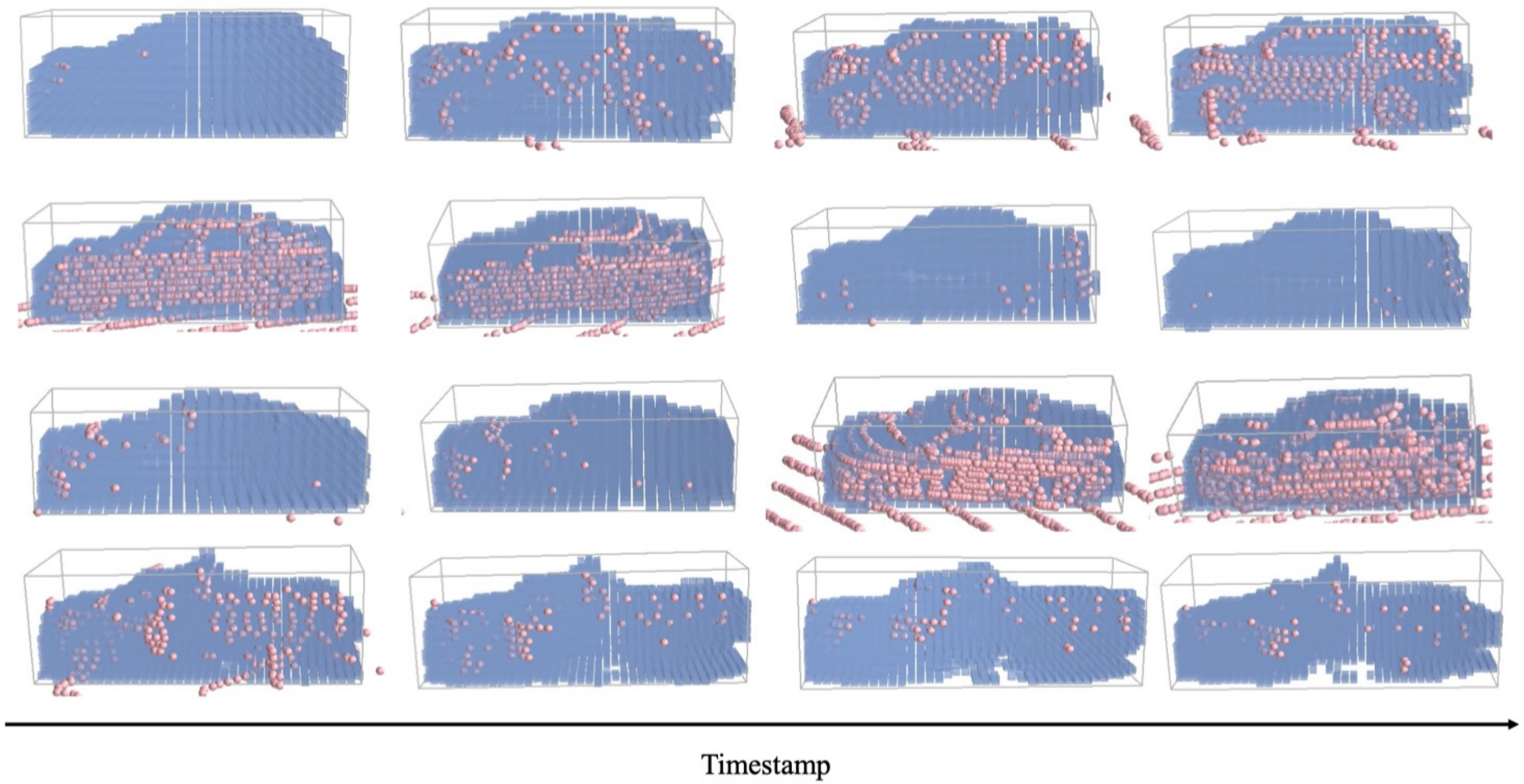


Figure 10: Visualization of the object-centric occupancy prediction. Different rows denote different object instances. Pink points indicate LiDAR points. Blue cubes represent the predicted occupied voxels.

Experiments – Shape Completion

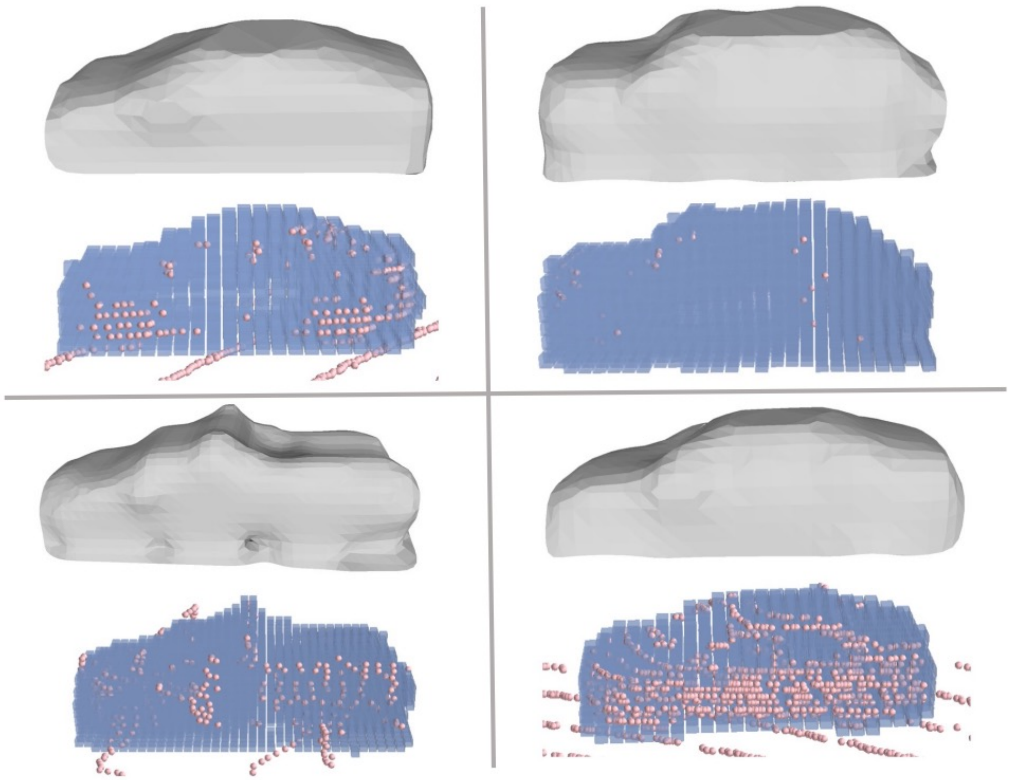


Figure 1: The renderings of predicted occupancy decoded from the shape codes for common vehicles. Top: extracted mesh from the occupancy using marching cube. Bottom: predicted occupancy and point cloud input.

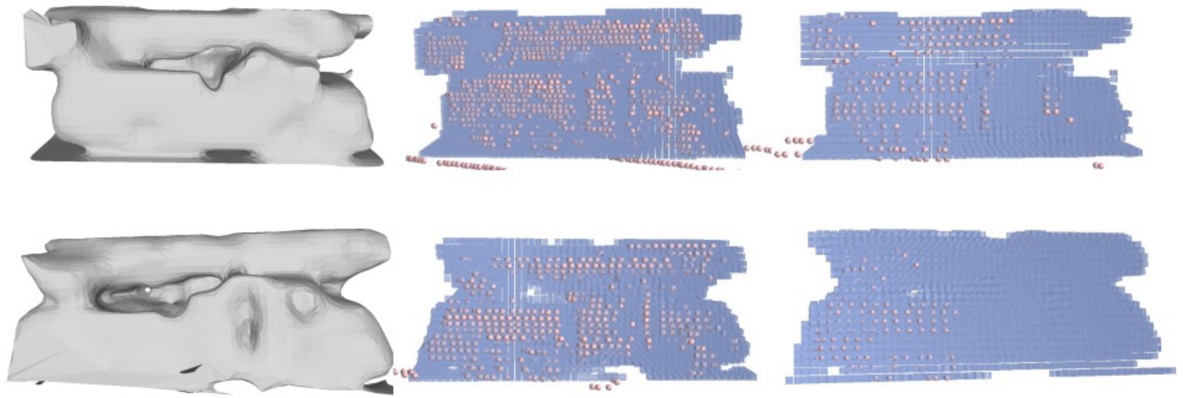


Figure 2: The renderings of complex vehicles. Each row shows the rendering, the corresponding predicted occupancy and input point cloud, and another predicted occupancy with fewer input points. These results demonstrate that the predicted object occupancy can better represent complex shape structures than bounding boxes.

Experiments – Detection

Method	Frame [-p,+f]	Vehicle L1 3D AP	Vehicle L1 3D APH	Vehicle L2 3D AP	Vehicle L2 3D APH
3D-MAN [37]	[-15, 0]	74.5	74.0	67.6	67.1
CenterFormer [41]	[-3, 0]	78.1	77.6	73.4	72.9
CenterFormer [41]	[-7, 0]	78.8	78.3	74.3	73.8
MPPNet [3]	[-3, 0]	81.5	81.1	74.1	73.6
MPPNet [3]	[-15, 0]	82.7	82.3	75.4	75.0
FSD++ [9]	[-6, 0]	81.4	80.9	73.3	72.9
MVF++ [21]	[-4, 0]	79.7	-	-	-
VoxelNeXt [4]	[0, 0]	78.2	77.7	69.9	69.4
HEDNet [40]	[0, 0]	81.1	80.6	73.2	72.7
CenterPoint* [38]	[0, 0]	72.9	72.3	64.7	64.2
+MoDAR [14]	[-91, 0]	76.1 (+3.2)	75.6 (+3.3)	68.9 (+4.2)	68.4 (+4.2)
CenterPoint‡ [38]	[0, 0]	73.2	72.7	65.2	64.6
+Ours	[−∞, 0]	81.8 **(+8.6)**	81.3 **(+8.6)**	73.6 **(+8.4)**	73.2 **(+8.6)**
SWFormer* [28]	[0, 0]	77.0	76.5	68.3	67.9
+MoDAR [14]	[-91, 0]	80.6 (+3.6)	80.1 (+3.6)	72.8 (+4.5)	72.3 (+4.4)
SWFormer* [28]	[-2, 0]	78.5	78.1	70.1	69.7
+MoDAR [14]	[-91, 0]	81.0 (+2.5)	80.5 (+2.4)	73.4 (+3.3)	72.9 (+3.2)
FSD‡ [6]	[0, 0]	78.7	78.3	70.1	69.7
+Ours	[−∞, 0]	82.8 **(+4.1)**	82.3 **(+4.0)**	74.8 **(+4.7)**	74.4 **(+4.7)**
FSD‡ [6]	[-6, 0]	80.9	80.5	73.1	72.7
+Ours	[−∞, 0]	**83.3** (+2.4)	**82.9** (+2.4)	**75.7**(+2.6)	**75.2** (+2.5)
FSDv2 [7]	[0, 0]	79.8	79.3	71.4	71.0
+Ours(no train)	[−∞, 0]	**83.2** (+3.4)	**82.7** (+3.4)	**75.2**(+3.8)	**74.7** (+3.7)

Table 2: Detection results on WOD val set. *: reported by MoDAR [14] . ‡ : our re-implementation. The Frame column illustrates the indices of the frames that are used. Blue indicates the improvement over the baseline.

Model	[0,30)	[30,50)	[50,+inf)
FSD [4]	90.97	70.87	46.04
+ Ours	92.55 (+1.58)	75.83 (+4.96)	53.85 **(+7.81)**
CenterPoint [31]	89.26	65.72	37.53
+ Ours	92.33 (+3.07)	74.88 (+9.06)	51.47 **(+13.94)**

Table 3: Detection with range breakdown. L2 mAP is reported.

Model	IoU	Vehicle 3D AP/APH L1	Vehicle 3D AP/APH L2
Ours	**62.84**	**82.80/82.31**	**74.83/74.36**
Single-Branch	62.13	80.51/80.05	72.26/71.82
Explicit Occ.	61.50	80.20/79.71	71.93/71.48
No Occ. Dec.	-	81.10/80.40	73.00/72.30

Table 4: Analysis of different designs.

Thanks

Code : https://github.com/Ghostish/ObjectCentricOccCompletion