JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

TUE-PM-098

OcTr: Octree-based Transformer for 3D Object Detection

Chao Zhou[1,2], Yanan Zhang[1,2], Jiaxin Chen[2] , Di Huang[1,2,3]*

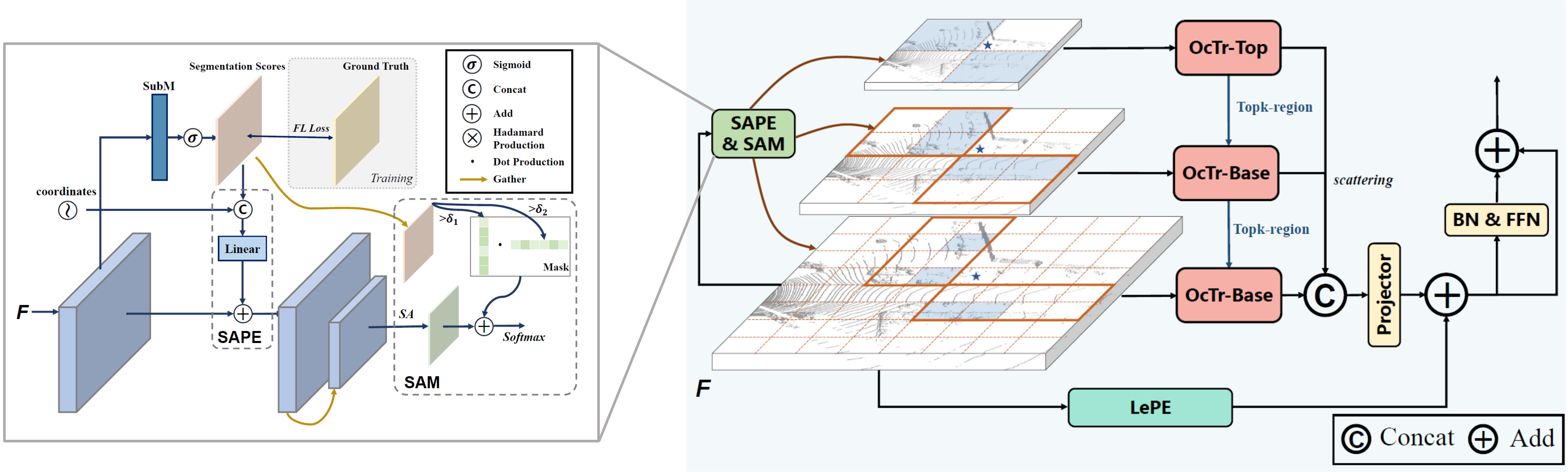
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Hangzhou Innovation Institute, Beihang University, Hangzhou, China

Highlight

Transformer in 3D Object Detection

High-resolution requirement

Limited Receptive Field
or
Limited Representations



Motivation

Transformers

- long-range dependencies modeling
- dynamic aggregation

3D object detection

- sparse data input
- high resolutions feature map

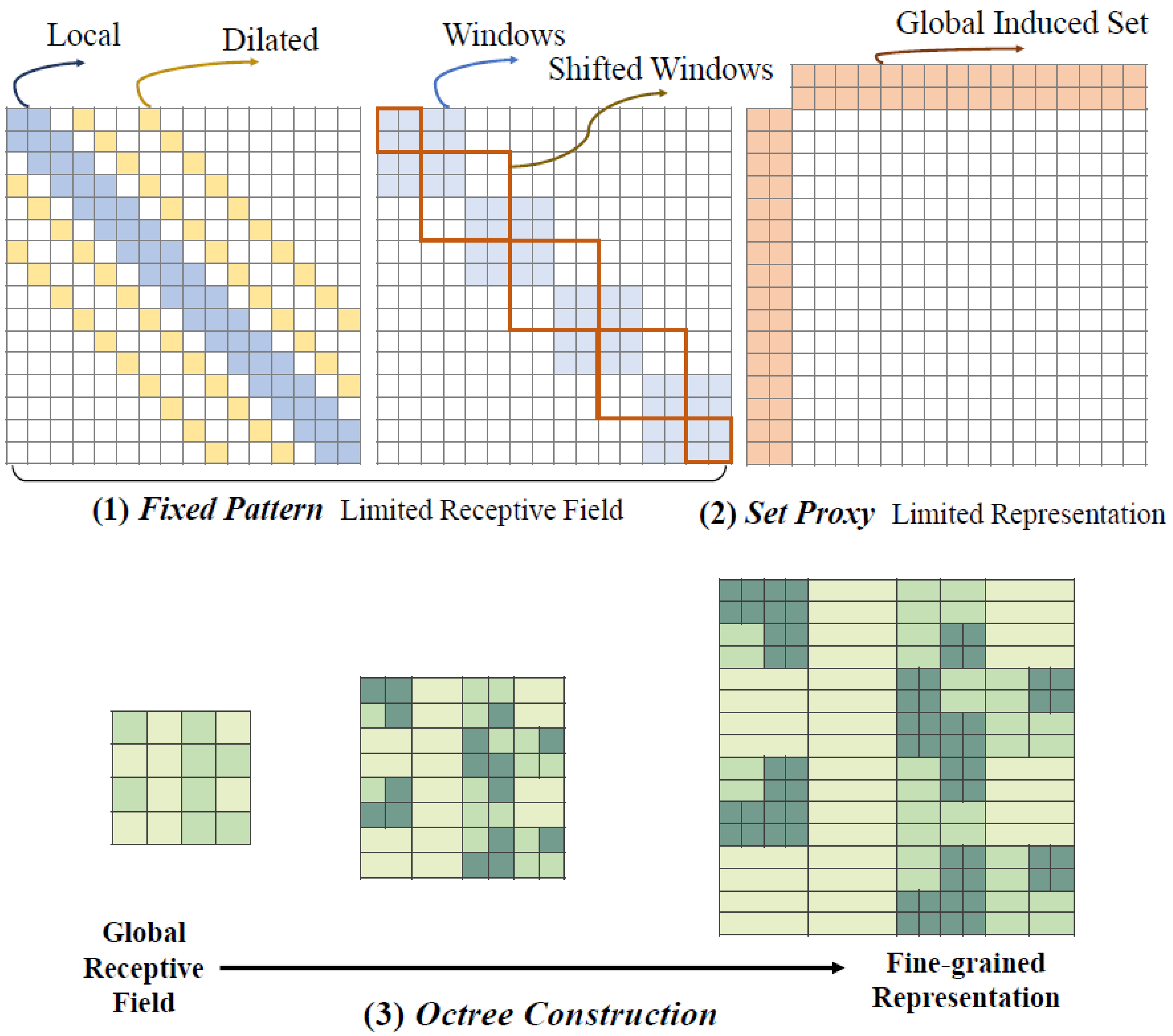
Dillema of heavy computations

Motivation

Limited Receptive Fields or
Limited Representations.

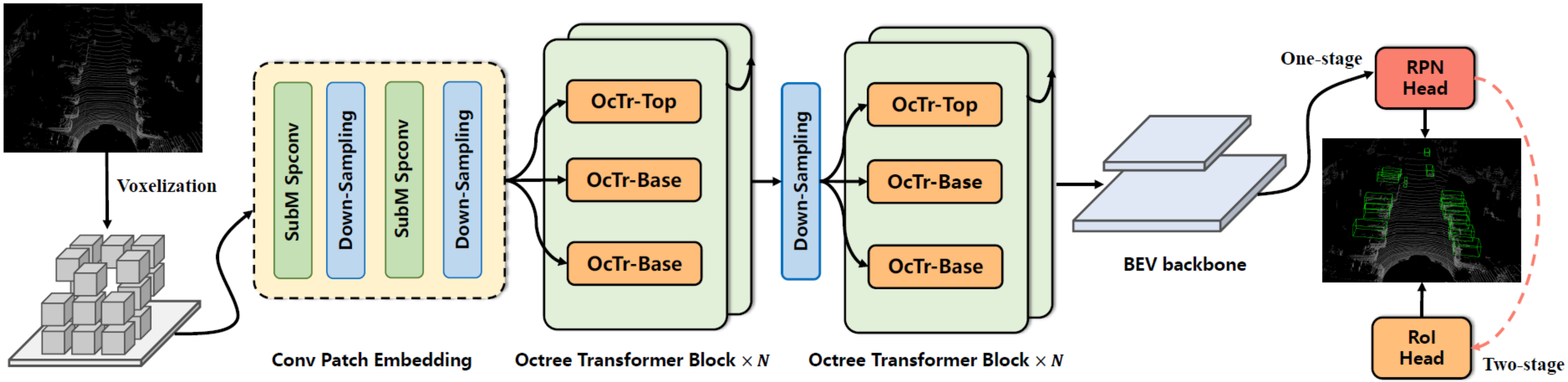
↓

Global Receptive Fields and
Fine-grained Representations.

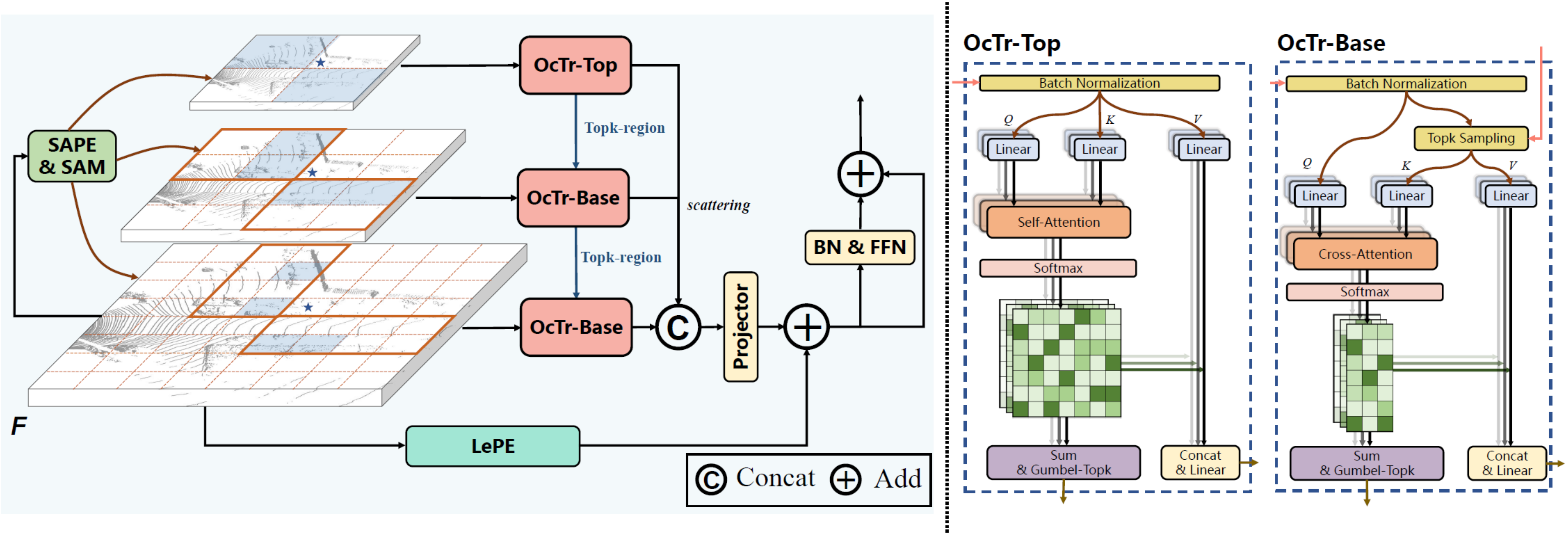


Method

The overall framework of the proposed OcTr:

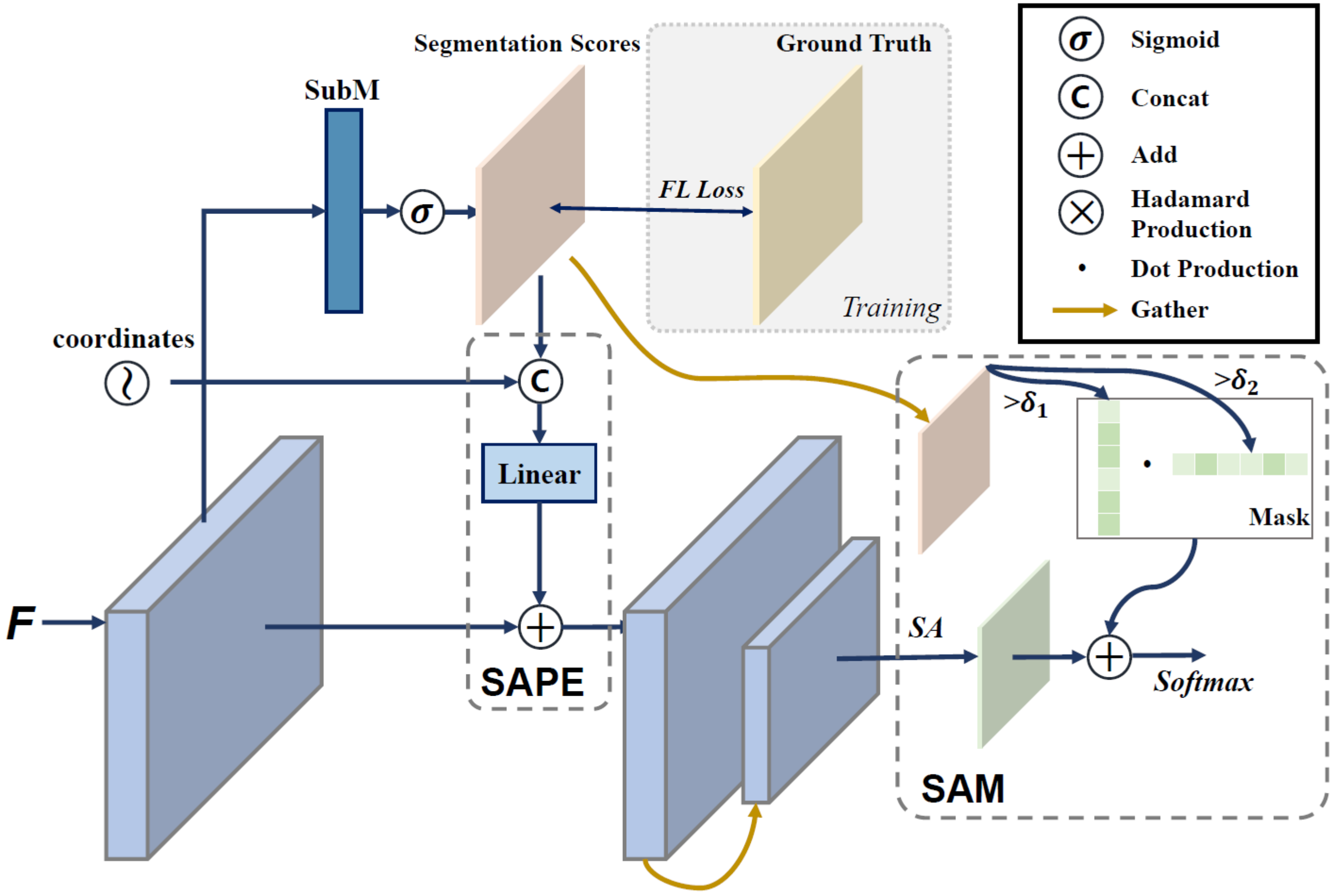


Octree-Attention



- *Construct an octree from hierarchical feature map*
- *Perform global self-attention on top layer*

Semantic Positional Embedding



Semantic Absolute Positional Embedding(SAPE):

- *Embed both **semantic and position** information*

Semantic Attention Mask(SAM):

- *High-quality tokens guide inferior ones*

Experiments

Comparison with state-of-the-art approaches on the WOD *val* split:

Model	Vehicle (L1) mAP/mAPH	Vehicle (L2) mAP/mAPH	Pedes. (L1) mAP/mAPH	Pedes. (L2) mAP/mAPH	Cyclist (L1) mAP/mAPH	Cyclist (L2) mAP/mAPH
SECOND [52]	70.96/70.34	62.58/62.02	65.23/54.24	57.22/47.49	57.13/55.62	54.97/53.53
PointPillar [18]	70.43/69.83	62.18/61.64	66.21/46.32	58.18/40.64	55.26/51.75	53.18/49.80
PartA^2Net [41]	74.82/74.32	65.88/65.42	71.76/63.64	62.53/55.30	67.35/66.15	65.05/63.89
PVRCNN [38]	75.41/74.74	67.44/66.80	71.98/61.24	63.70/53.95	65.88/64.25	63.39/61.82
CenterPoint [55]	71.33/70.76	63.16/62.65	72.09/65.49	64.27/58.23	68.68/67.39	66.11/64.87
LiDAR-RCNN [20]	73.5/73.0	64.7/64.2	71.2/58.7	63.1/51.7	68.6/66.9	66.1/64.4
Voxel-RCNN [6]	75.59/-	66.59/-	-/-	-/-	-/-	-/-
PVRCNN++ [39]	77.82/77.32	69.07/68.62	77.99/71.36	69.92/63.74	71.80/70.71	69.31/68.26
SST† [9]	76.22/75.79	68.04/67.64	81.39/74.05	72.82/65.93	-/-	-/-
PDV [15]	76.85/76.33	69.30/68.81	74.19/65.96	65.85/58.28	68.71/67.55	66.49/65.36
Ours	**78.12/77.63**	**69.79/69.34**	**80.76/74.39**	**72.48/66.52**	**72.58/71.50**	**69.93/68.90**

Experiments

Comparison on the WOD *val* by Distance:

Model	mAP_{3D} (L1)@Vehicle Overall	0-30m	30m-50m	50m-inf
PV-RCNN [38]	70.30	91.92	69.21	42.17
Voxel-RCNN [6]	75.59	92.49	74.09	53.15
VoTR-TSD [24]	74.95	92.28	73.36	51.09
CT3D [37]	76.30	92.51	75.07	55.36
Pyramid_PV [25]	76.30	92.67	74.91	54.54
PDV [15]	76.85	**93.13**	75.49	54.75
VoxSeT [12]	77.82	92.78	77.21	54.41
Ours	**78.82**	92.99	**77.66**	**58.02**

Model	mAP_{3D} (L2)@Vehicle Overall	0-30m	30-50m	50m-inf
PV-RCNN [38]	65.36	91.58	65.13	36.46
Voxel-RCNN [6]	66.59	91.74	67.89	40.80
CT3D [37]	69.04	91.76	68.93	42.60
PDV [15]	69.30	**92.41**	69.36	42.16
VoxSeT [12]	70.21	92.05	70.10	43.20
Ours	**70.50**	91.78	**71.28**	**45.46**

Comparison on the KITTI *test* :

Model	mAP_{3D}@Car on test				mAP_{3D}@Car on val			
	Easy	Mod.	Hard	Mean	Easy	Mod.	Hard	Mean
SECOND [52]	83.34	72.55	65.82	73.90	88.61	78.62	77.22	81.48
PointPillars [18]	82.58	74.31	68.99	75.29	86.62	76.06	68.91	77.20
STD [54]	87.95	79.71	75.09	80.92	89.70	79.80	**79.30**	82.93
SA-SSD [13]	88.75	79.79	74.16	80.90	**90.15**	79.91	78.78	82.95
3DSSD [53]	88.36	79.57	74.55	80.83	89.71	79.45	78.67	82.61
PV-RCNN [38]	90.25	81.43	76.82	82.83	89.35	83.69	78.70	83.91
Voxel-RCNN [6]	**90.90**	81.62	77.06	83.19	89.41	84.52	78.93	84.29
CT3D [37]	87.83	81.77	77.16	82.25	89.54	86.06	78.99	<u>84.86</u>
VoTR-TSD [24]	89.90	82.09	**79.14**	<u>83.71</u>	89.04	84.04	78.68	83.92
VoxSeT [12]	88.53	82.06	77.46	82.68	89.21	<u>86.71</u>	78.56	84.83
Focals Conv [4]	90.55	<u>82.28</u>	77.59	83.47	89.52	84.93	79.18	84.54
Ours	<u>90.88</u>	**82.64**	<u>77.77</u>	**83.76**	<u>89.80</u>	**86.97**	<u>79.28</u>	**85.35**

Ablation Study

Extensions to different detectors:

Detector	Veh. mAP (L1/L2)	Pedes. mAP (L1/L2)
SECOND [52]	70.96/62.58	65.23/57.22
Ours	**73.28/65.05**	**68.08/60.36**
PV-RCNN [38]	75.41/67.44	71.98/63.70
Ours	**76.77/68.31**	**73.22/64.30**
PV-RCNN++ [39]	77.82/69.07	77.99/69.92
Ours	**78.01/69.60**	**80.75/72.45**

Comparison with different mechanism:

Attention	Veh. mAP (L1/L2)	Pedes. mAP (L1/L2)
Ours (*OctAttn*)	**73.3/65.1**	**68.1/60.4**
Performer [5]	71.4/63.6	65.7/57.9
ACT [27]	71.7/63.5	64.3/56.1
VoTr [24]	69.4/61.5	65.0/57.0
Nearest K	68.2/59.8	64.9/56.7

Ablation on Semantic Positional Embedding:

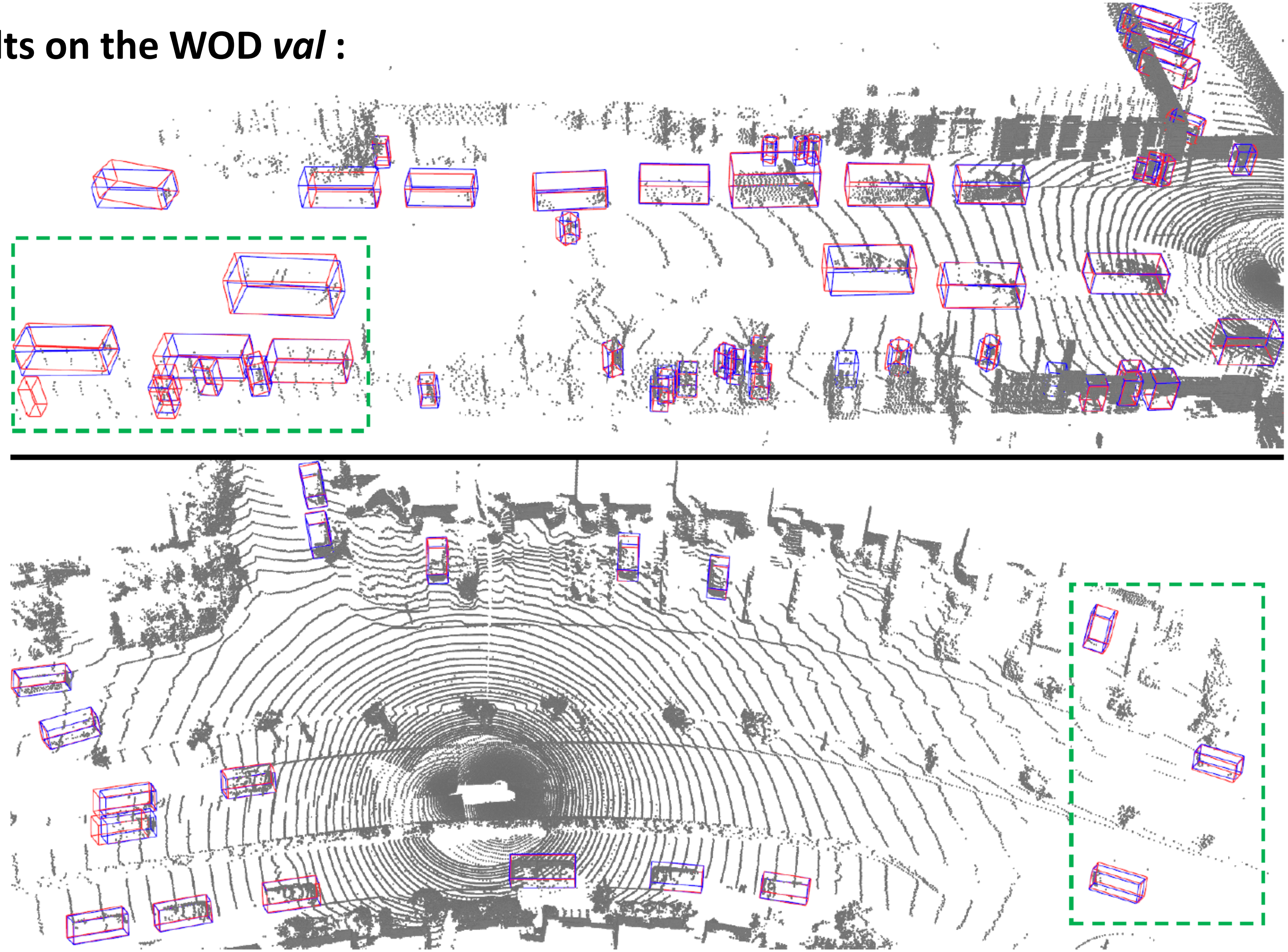
LEPE	SAPE	SAM	Veh. mAP (L1/L2)	Pedes. mAP (L1/L2)
			71.35/63.30	65.75/57.89
✓			72.34/64.32	66.56/58.62
✓	✓		72.64/64.46	66.62/58.83
✓		✓	72.86/64.40	67.79/59.90
✓	✓	✓	**73.28/65.05**	**68.08/60.36**

Resource Costs:

Method	#Param. (M)	Latency (ms)	Memory (GB)
SECOND [52]	5.3	48	**2.3**
VoTR-SSD [24]	4.8	67	3.0
VoxSeT-SSD [12]	3.0	**37**	3.6
OcTr-SSD	**2.9**	64	2.5

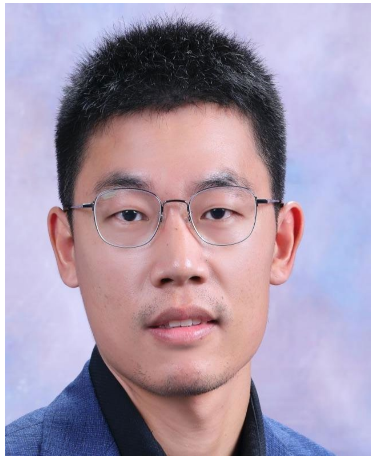
Visualization Results

Visualization results on the WOD *val* :



Thanks






Chao Zhou
Beihang University
zhouchaobeing@buaa.edu.cn

Yanan Zhang
Beihang University
zhangyanan@buaa.edu.cn

Jiaxin Chen
Beihang University
jiaxinchen@buaa.edu.cn

Di Huang
Beihang University
dhuang@buaa.edu.cn