Robotics, Artificial Intelligence
and Embedded Systems

Mamba-based 3D Object Detection with Object Relation

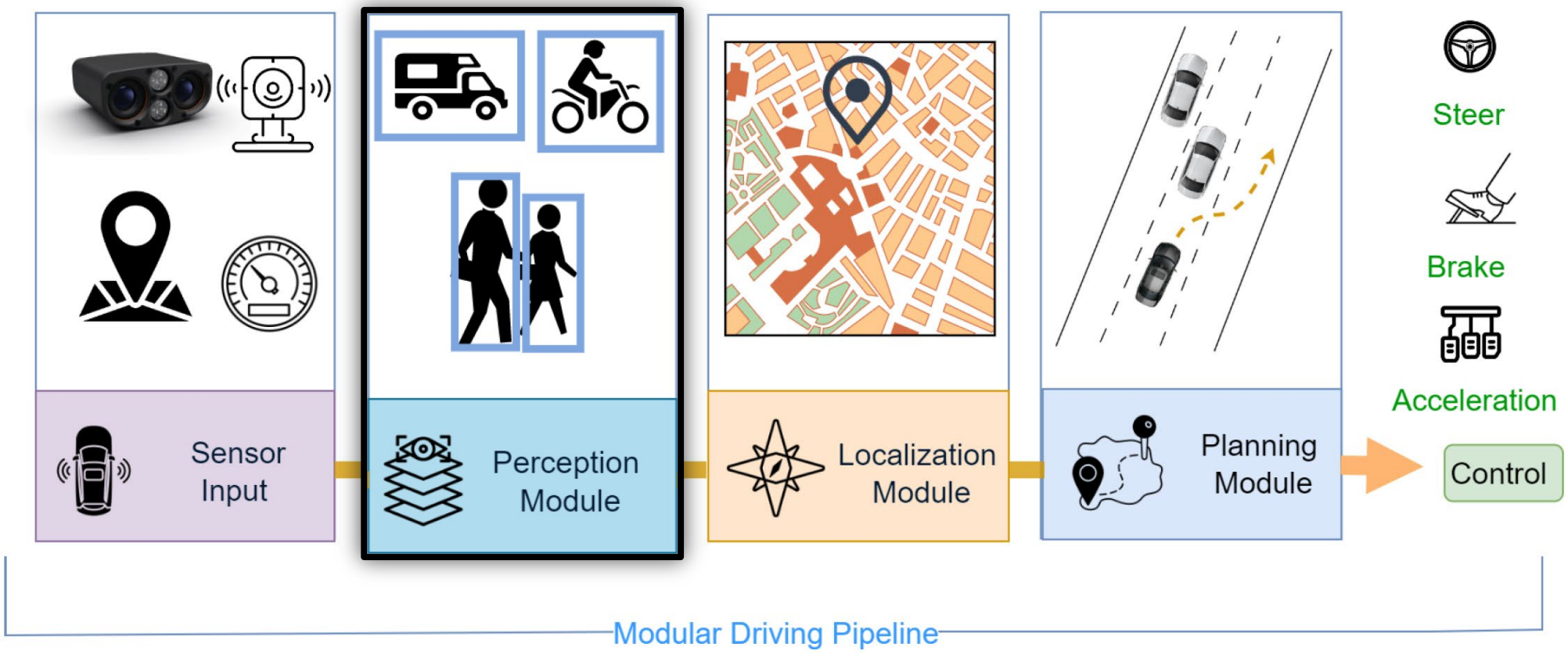
Supervisor: Prof. Dr.-Ing. habil. Alois C. Knoll
Advisor: Mingyu Liu, M.Sc.; Bare Luka Žagar, M.Sc.
Name: Jun Meng

Date: 28.06.2024

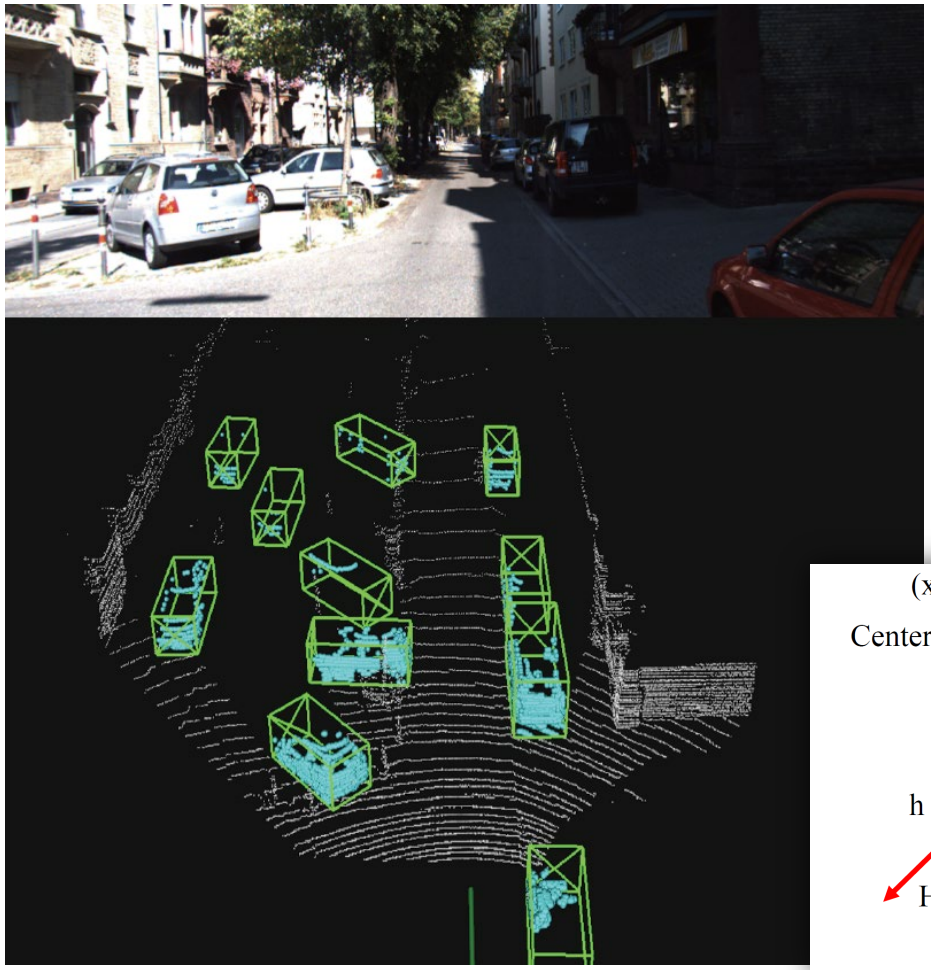
Agenda

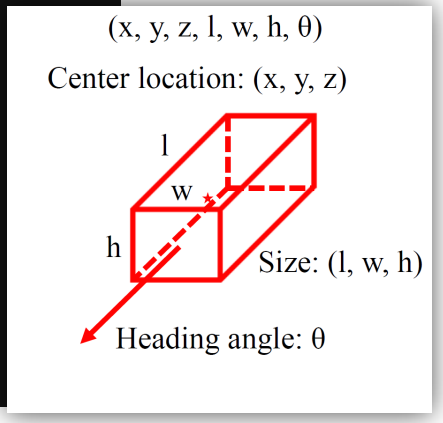
1. Motivation & Background
2. Related Works
3. Methodology
4. Experiments & Results
5. Conclusions & Future Works

Modular pipeline of Autonomous Driving：



3D Object Detection：

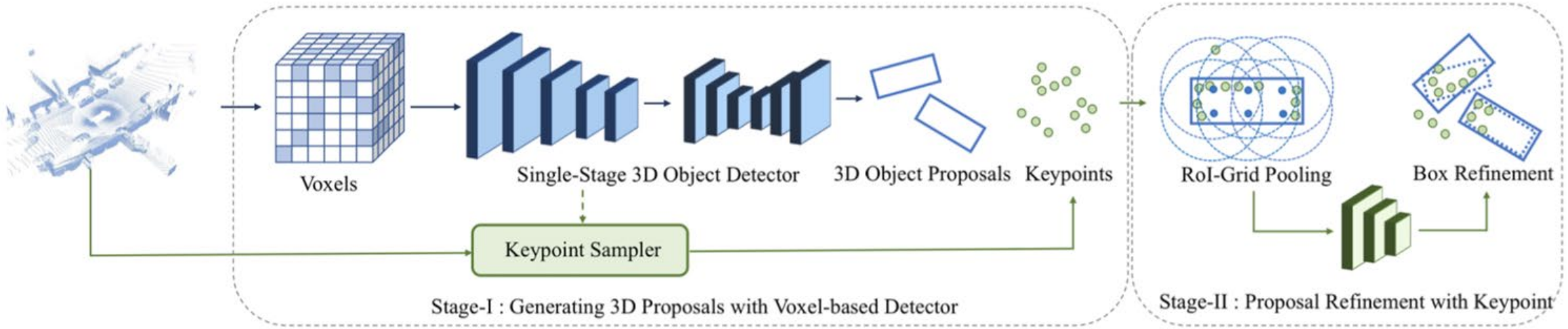




Two-stage Point-voxel based Detector

- Stage I: Proposal generation
- Stage II: Proposal refinement

➔ Each proposal is processed individually in the refinement stage



Source: Mao, J., Shi, S., Wang, X., and Li, H. 3D Object Detection for Autonomous Driving: A Comprehensive Survey. 2023

Motivation:

Safety, reliability

→ Higher detection accuracy

Real-time capability

→ Lower computational complexity

Motivation: Object Relation

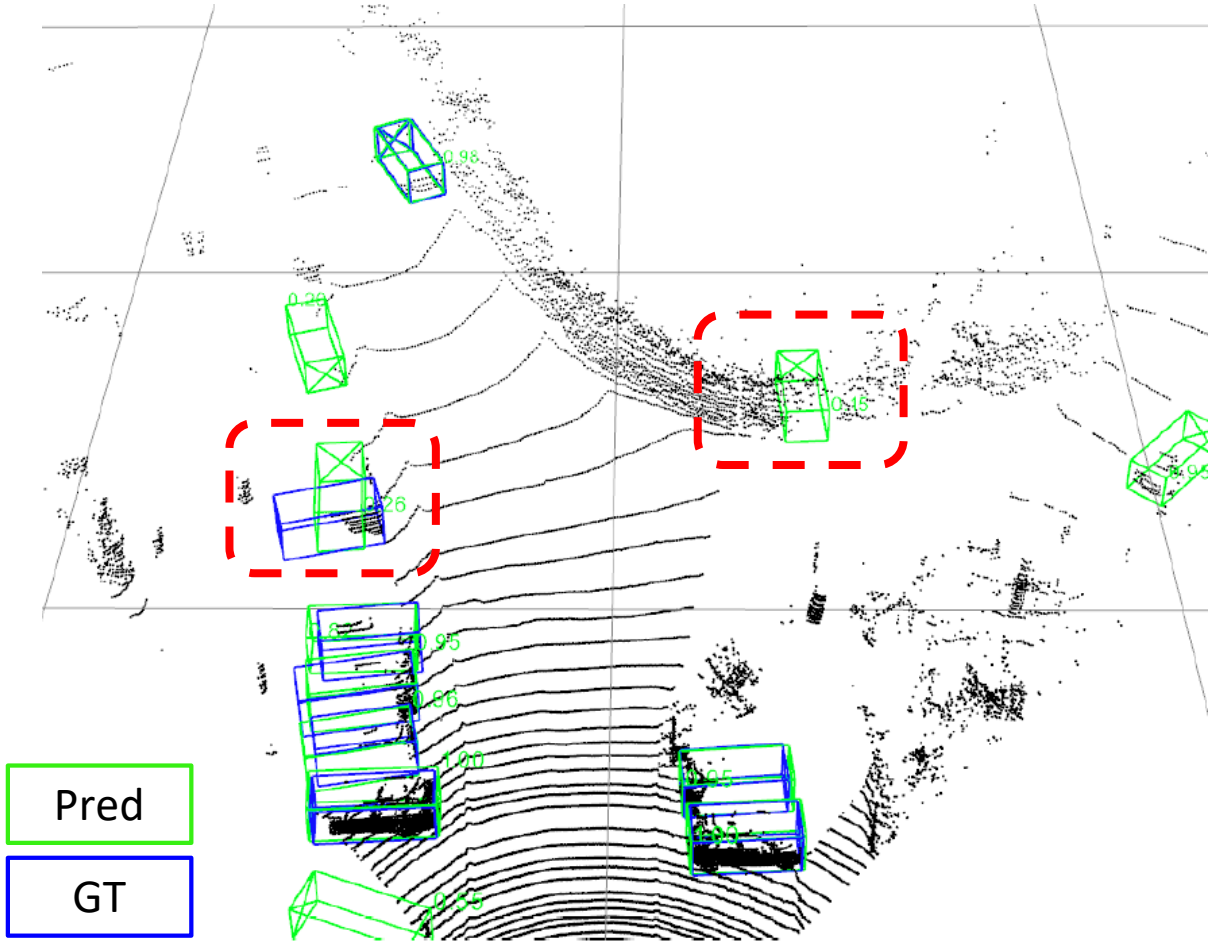
Limitations of current detectors:

Incomplete representation due to

Occlusions & Sparsity of distant objects

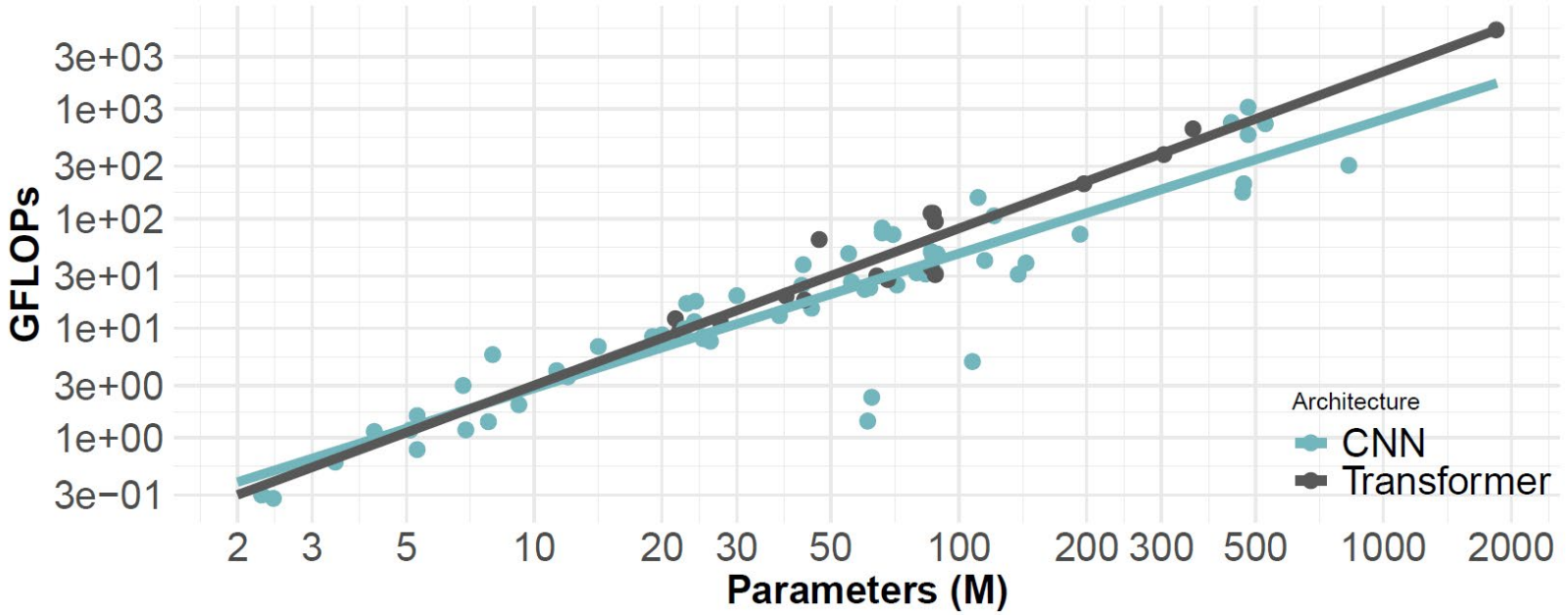
Exploit Object Relation explicitly:

- Utilize contextual information beyond receptive field
- Alleviate negative effects from occlusions
- Exploring traffic patterns:
  traffic jam, parking lot, etc.



Motivation: lower FLOPs

Source: Compute and Energy Consumption Trends in Deep Learning Inference



FLOPs (Floating Point Operations): the total number of computations for a given task/model

Lower FLOPs → faster inference

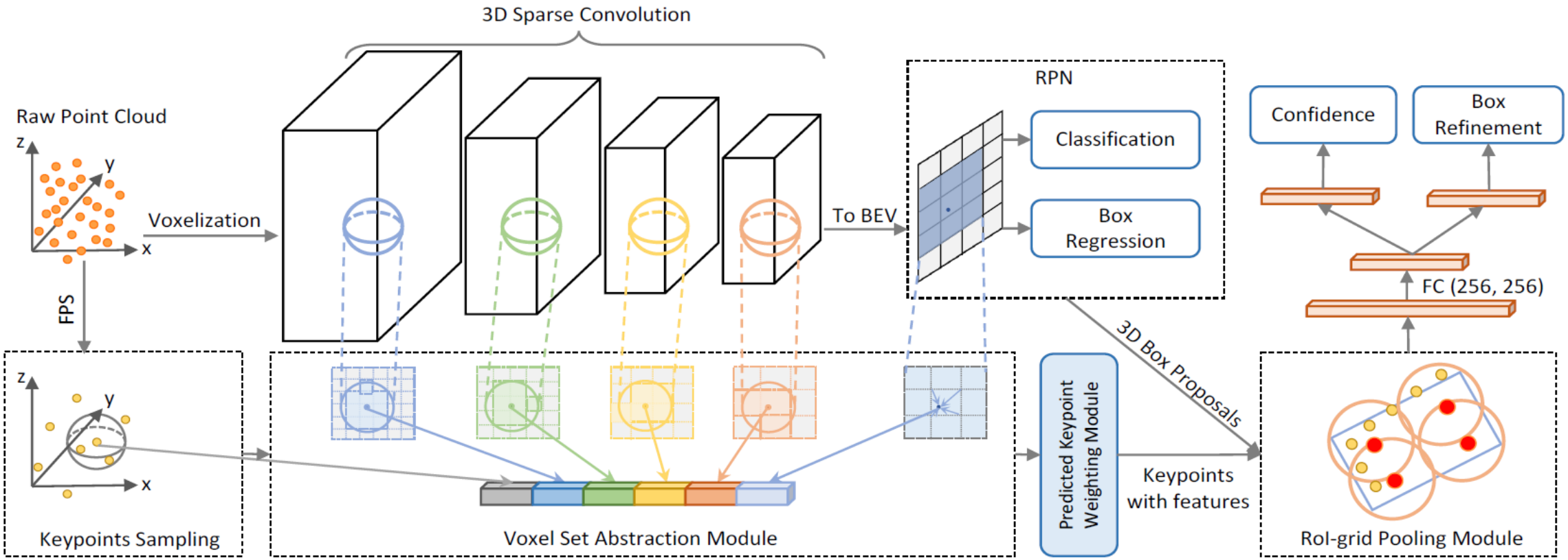
Related Works

Agenda

• PV-RCNN • PartA2	Baseline 3D Detectors
• GACE • Relation Graph, Edge Convolution	Object Relation
• SSM, Mamba • VMamba, VSS-Block • SS2D	State Space Model

Related Work: PV-RCNN

Point-Voxel RCNN

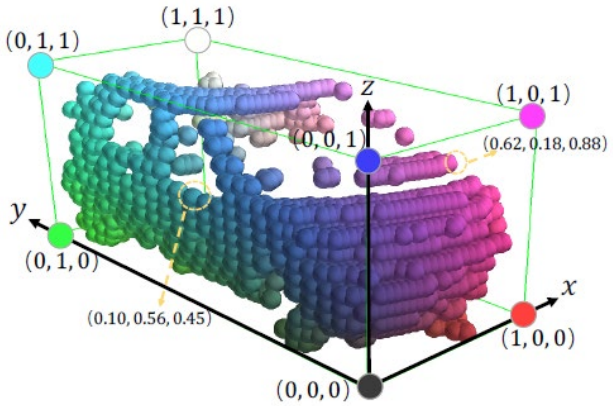


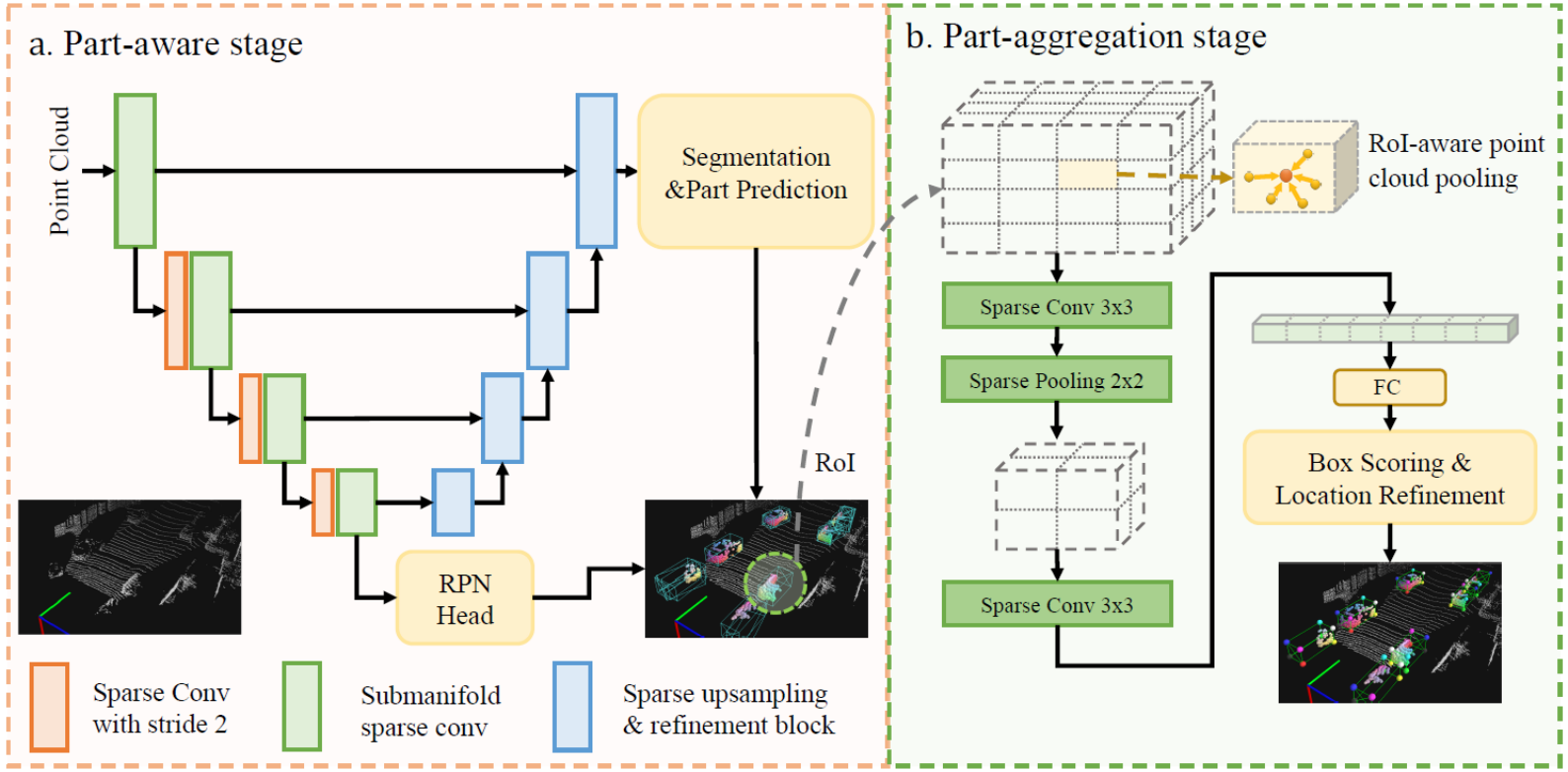
Source: Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., and Li, H. "PV- RCNN: Point-voxel feature set abstraction with local vector representation for 3D object detection". In: *International Journal of Computer Vision* 131.2 (2023), pp. 531–551.

Related Work: PartA2

Part-aware and Part-aggregation network

segments foreground points and predicts their intra-object part locations, even when objects are partially occluded





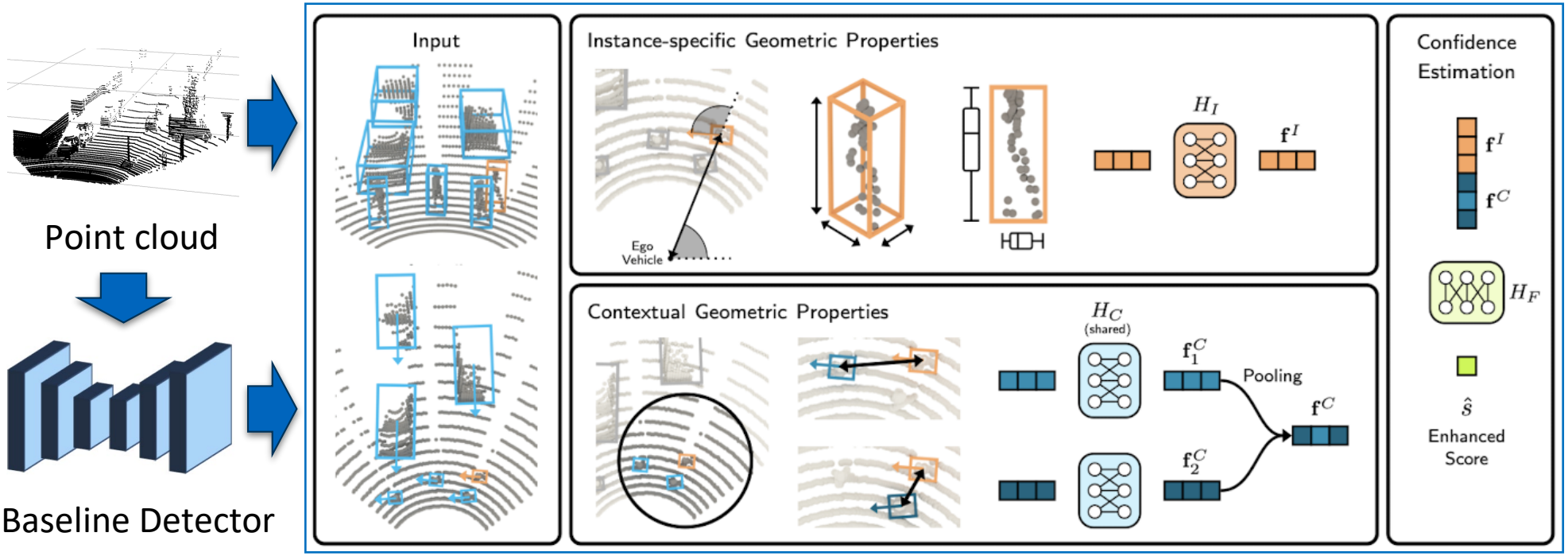
Source: Shi, S., Wang, Z., Shi, J., Wang, X., and Li, H. “From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network”. In: IEEE transactions on pattern analysis and machine intelligence 43.8 (2020), pp. 2647–2664

Related Work: GACE

Geometry Aware Confidence Enhancement (GACE)

- Takes the raw point cloud and detection results of base detector as input. Baseline model as black-box.
- Optimizes confidence scores for a given detection set by re-evaluating the results as TP and FP.

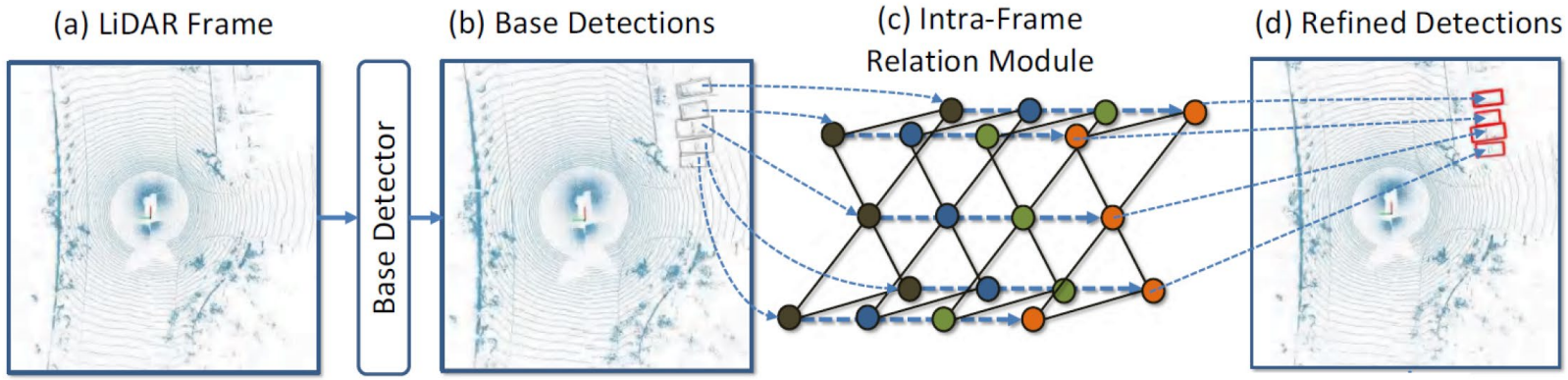
→ An independent network. Not compatible for end-to-end training.



Source: Schinagl, D., Krispel, G., Fruhwirth-Reisinger, C., Possegger, H., and Bischof, H. "GACE: Geometry Aware Confidence Enhancement for Black-Box 3D Object Detectors on LiDAR-Data". In: *Proceedings of the IEEE/CVF International Confer- ence on Computer Vision*. 2023, pp. 6566–6576.

Related Work: Ret3D and DGCNN

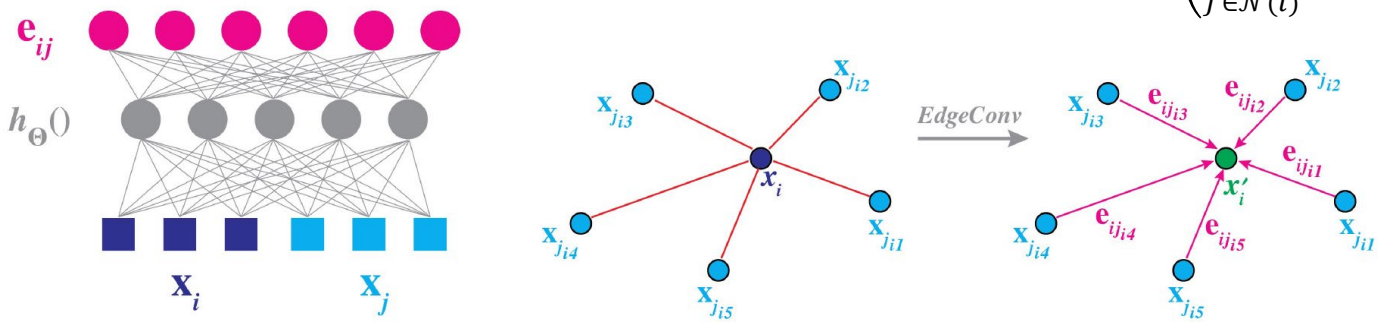
Ret3D: Intra-Frame relation graph



Source:. Wu, Y.-H., Zhang, D., Zhang, L., Zhan, X., Dai, D., Liu, Y., and Cheng, M.-M. Ret3D: Rethinking Object Relations for Efficient 3D Object Detection in Driving Scenes. 2022. arXiv: 2208.08621

DGCNN: Edge Convolution

$$x_i^{(t+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} W \cdot [x_i^{(t)}, x_j^{(t)}, e_{ij}]\right)$$



Source:. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic Graph CNN for Learning on Point Clouds. 2019. arXiv: 1801.07829

Related Work: SSM & Mamba

State Space Model (SSM)

$$h'(t) = Ah(t) + Bx(t)$$
$$y(t) = Ch(t)$$

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t$$
$$y_t = Ch_t$$

$$\overline{K} = (C\overline{B}, C\overline{AB}, \dots, C\overline{A}^k\overline{B}, \dots)$$
$$y = x * \overline{K}$$

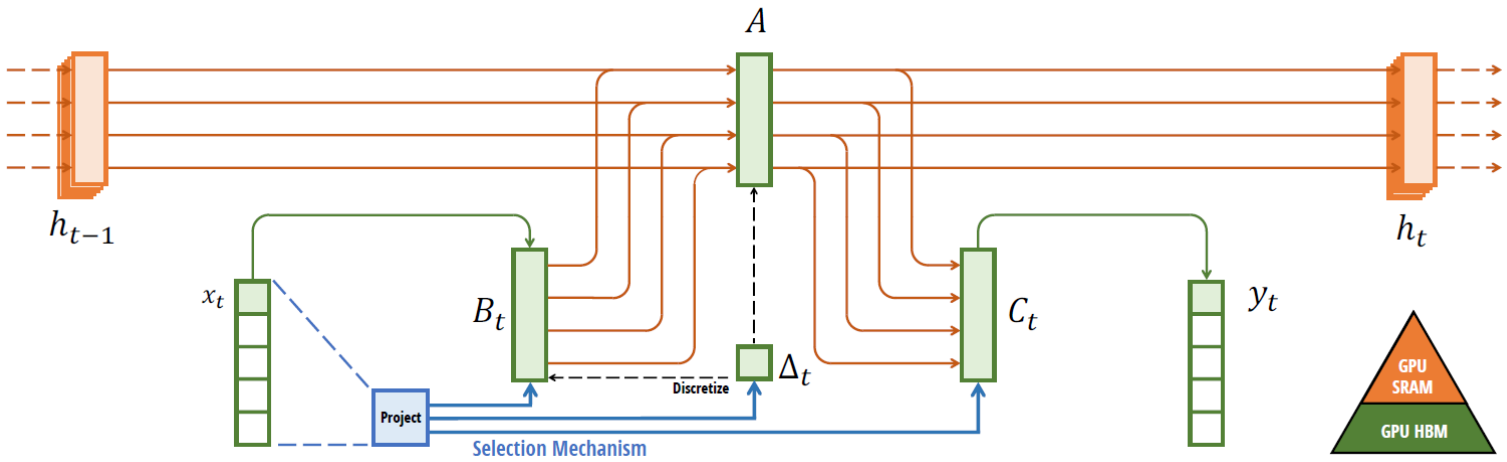
Discretization

zero-order hold (ZOH) rule

$$\overline{A} = \exp(\Delta A) \qquad \overline{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$$

A, ***B***: time-invariant

Mamba: Selective SSM

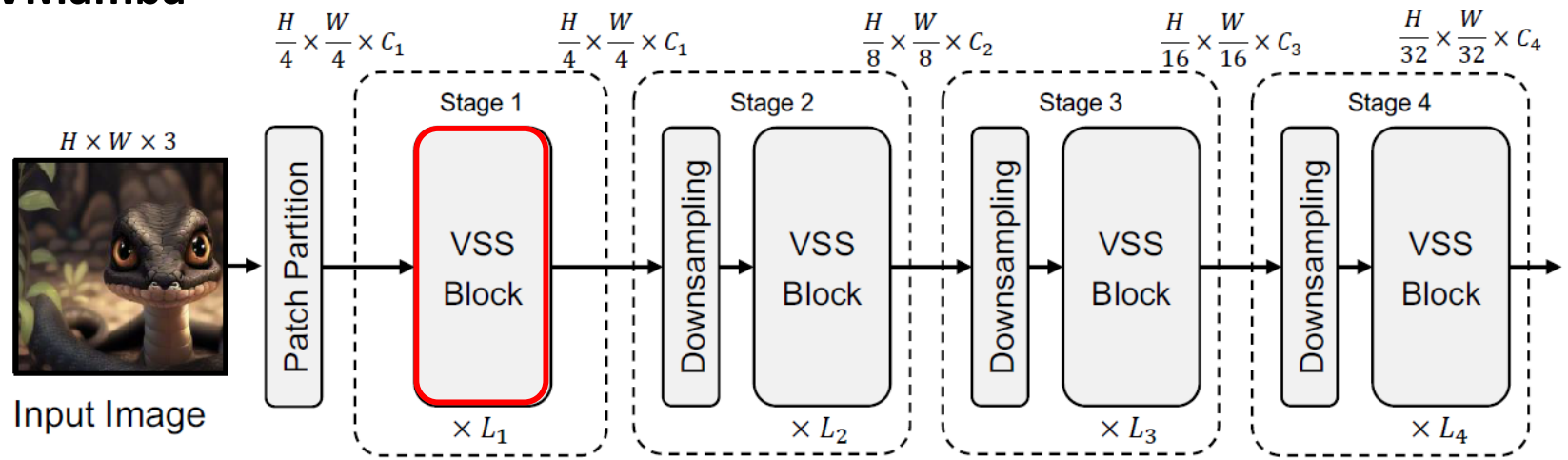


A(x), ***B(x)***: time-varying

Source:. Gu, A. and Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. 2023. arXiv: 2312.00752

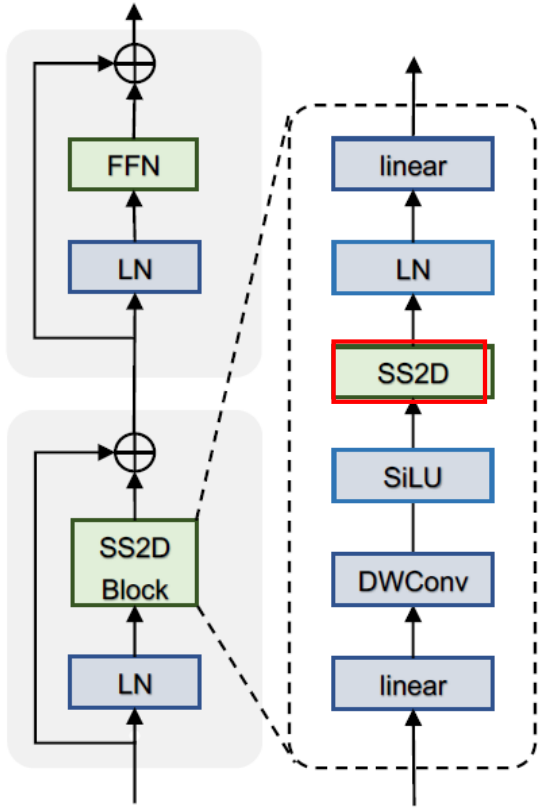
Methodology: VMamba, VSS-Block

VMamba

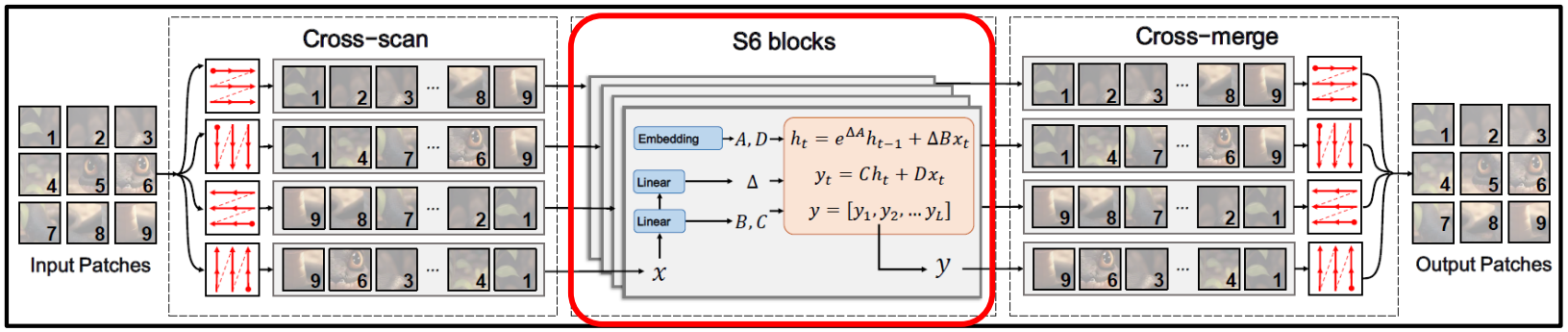


Source:. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., and Liu, Y. VMamba: Visual State Space Model. 2024. arXiv: 2401.10166

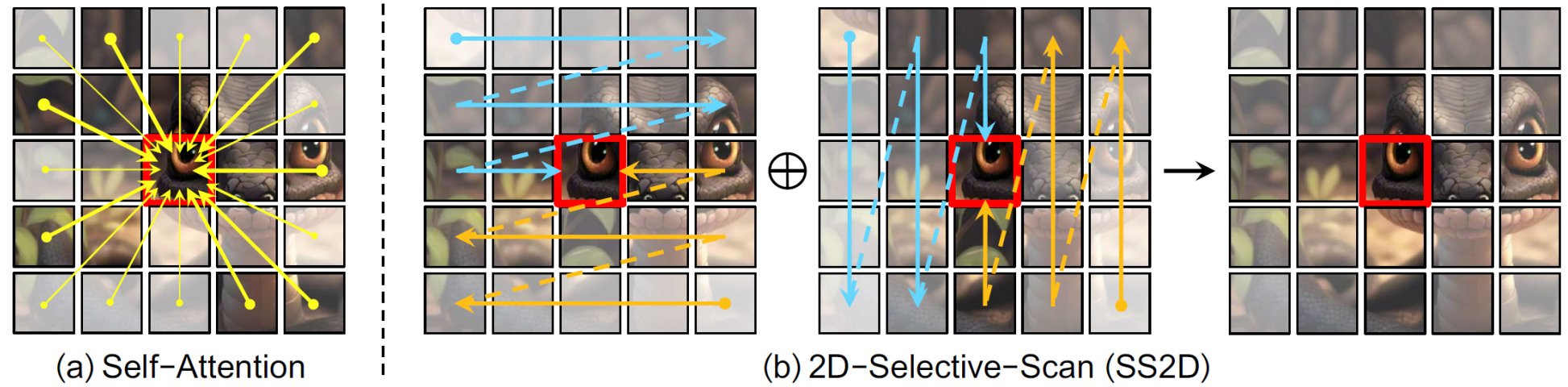
Visual State Space Block



2D Selective Scan



Related Work: SS2D

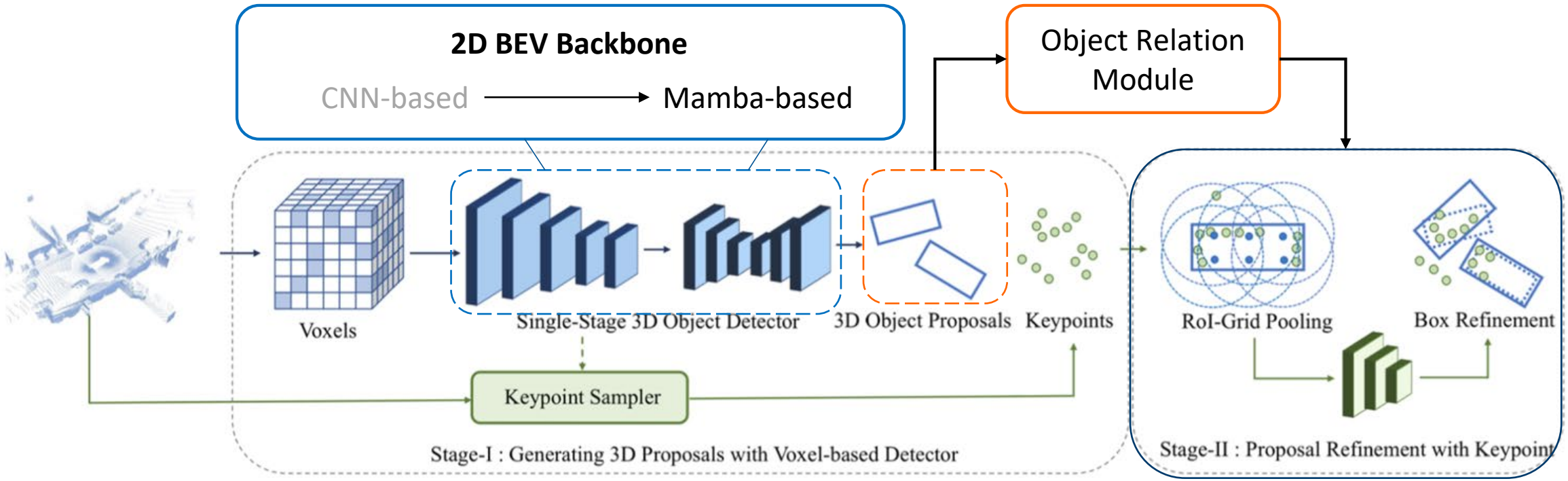


quadratic attention complexity

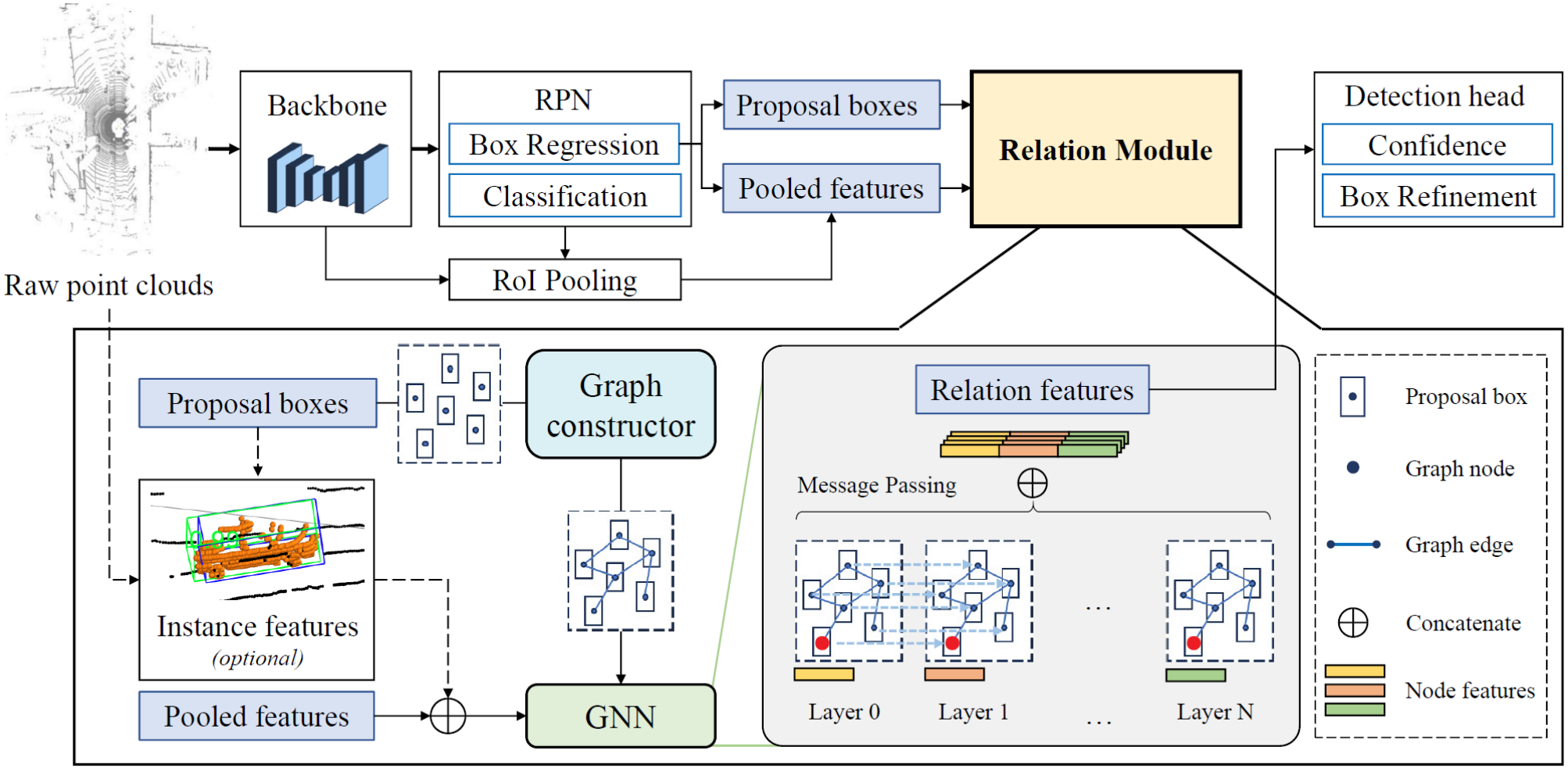
linear space and time complexity

Methodology:

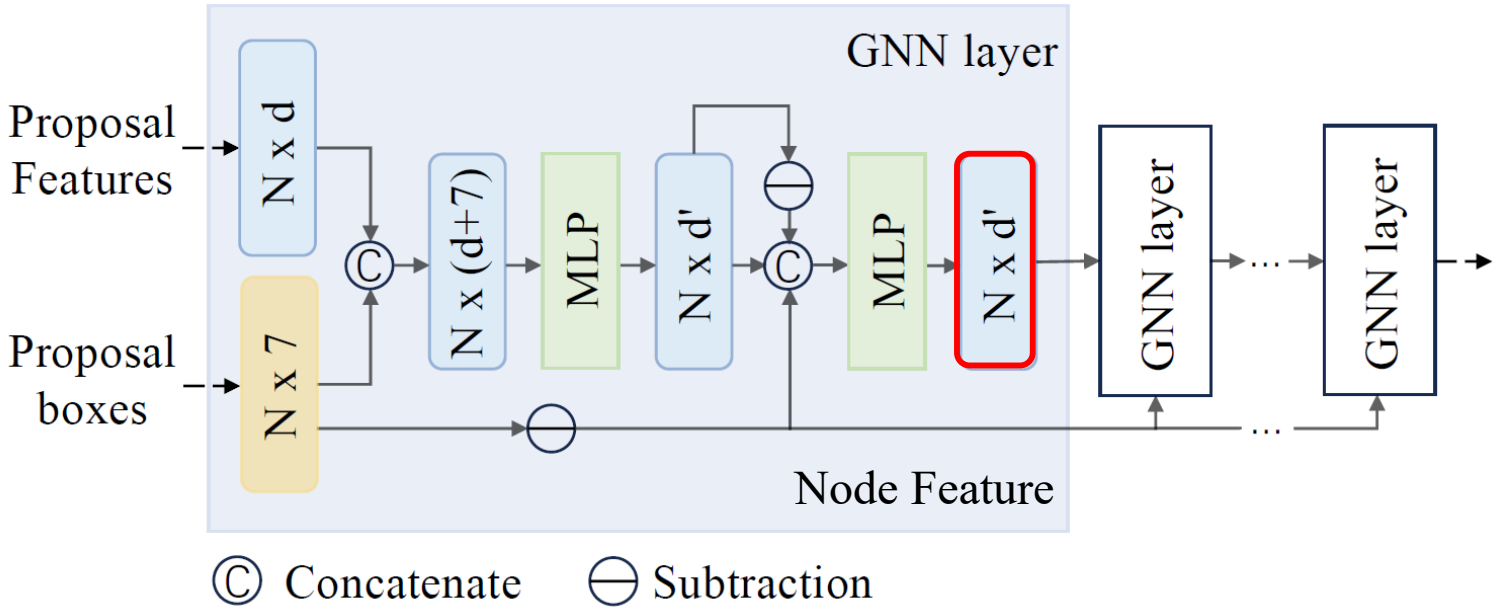
1. Integrate a GNN-based object relation module
2. Substitute CNN-based encoders in the 2D BEV backbone with Mamba-based ones



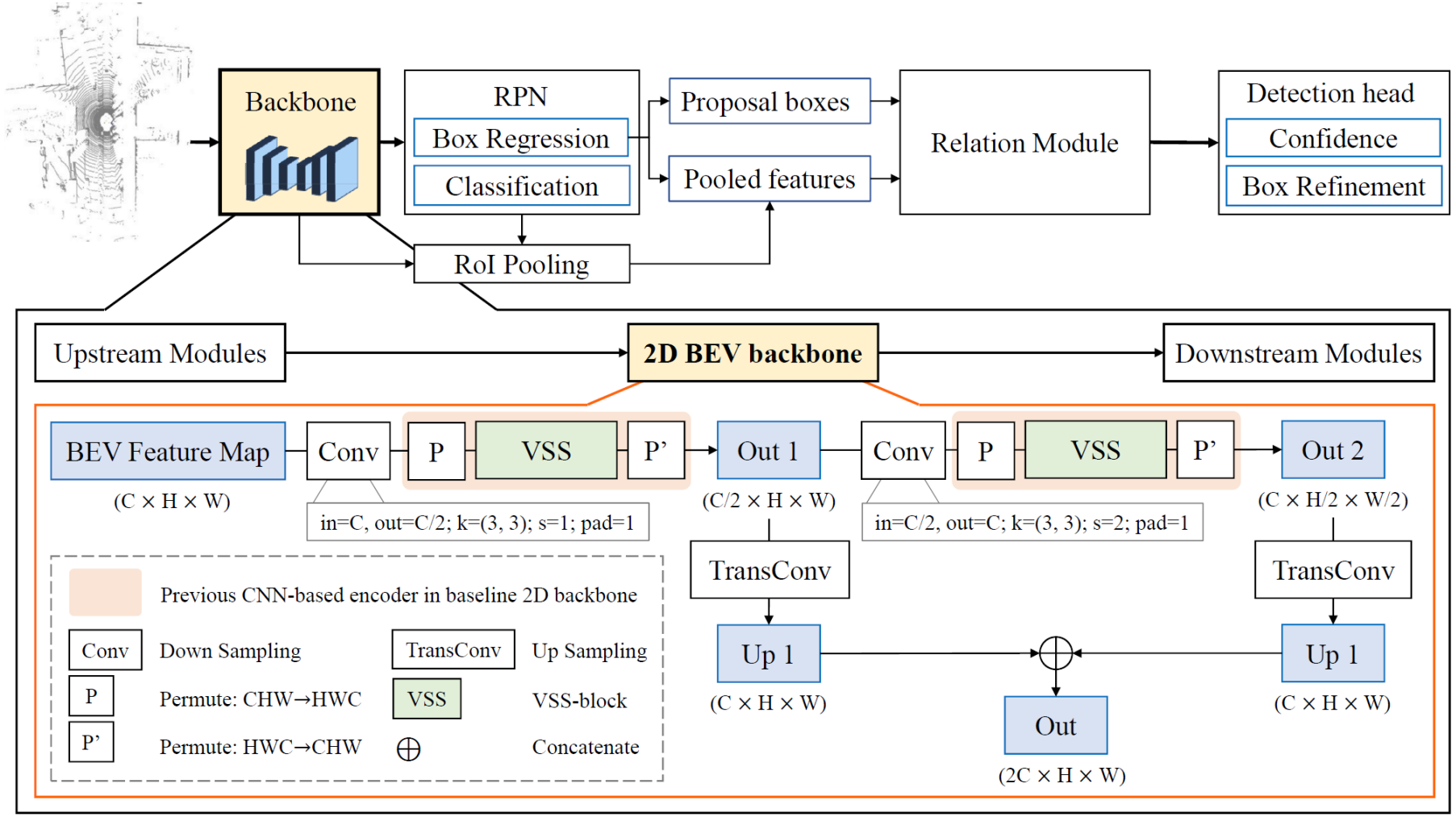
Methodology: Object Relation Module



Methodology: GNN



Methodology: Mamba-based 2D Backbone



Methodology: Training Loss

PV-RCNN

Region proposal loss

$$L_{rpn} = L_{cls} + \beta \sum \mathcal{L}_{smooth-L1}(b - b^{gt})$$

Proposal refinement loss

$$L_{rcnn} = L_{iou} + \sum \mathcal{L}_{smooth-L1}(b^{res} - b^{gt})$$

PartA2

Part-aware Stage I

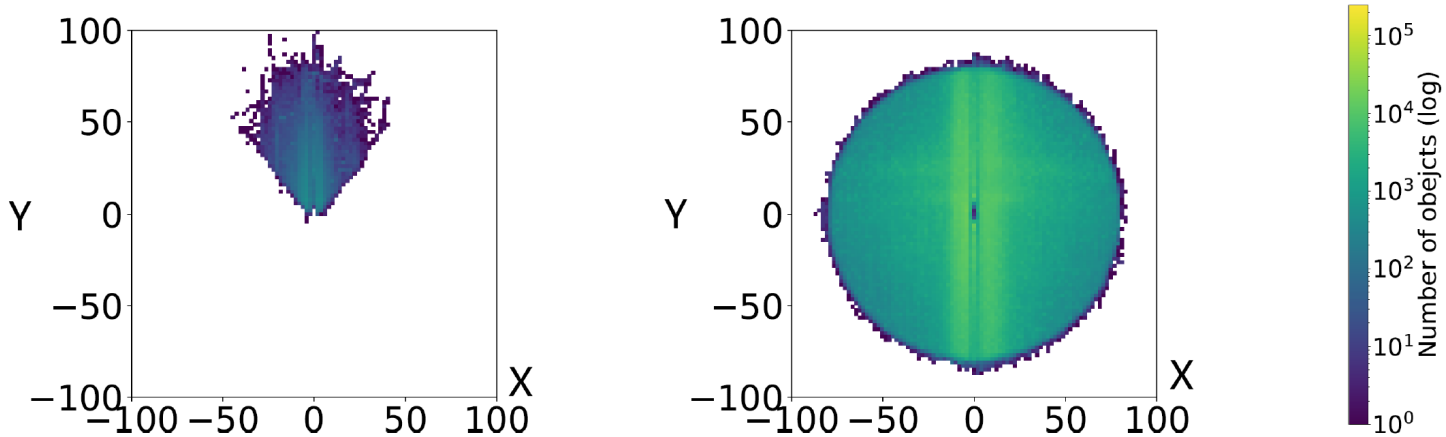
$$L_{aware} = L_{seg} + \frac{1}{N_{pos}} L_{part} + \lambda \frac{1}{M_{pos}} L_{box}$$

Part-aggregation Stage II

$$L_{aggregation} = L_{score} + \frac{1}{T_{pos}} L_{box\_refine}$$

Experiments: Datasets

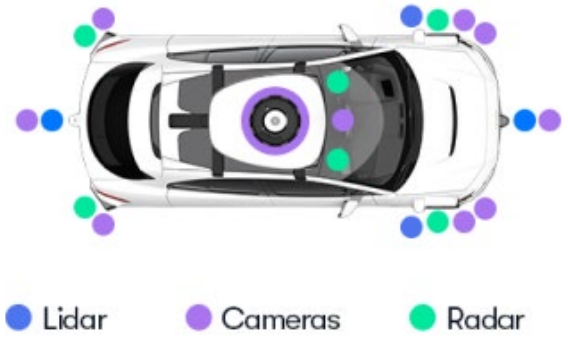
	KITTI	Waymo
Scenes	22	1150
Ann. LiDAR Fr.	15K	230K
Hours	1.5	6.4
3D Boxes	80K	12M
2D Boxes	80K	9.9M
LiDARs	1	5
Cameras	4	6
Avg. Points/Fr.	120K	177K
LiDAR Features	1	2
Maps	No	Yes
Released Year	2013	2020



(a) KITTI (b) Waymo

Source:. Liu, M., Yurtsever, E., Fossaert, J., Zhou, X., Zimmer, W., Cui, Y., Zagar, B. L., and Knoll, A. C. A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook. 2024. arXiv: 2401.01454

Source: https://www.cvlibs.net/datasets/kitti/setup.php

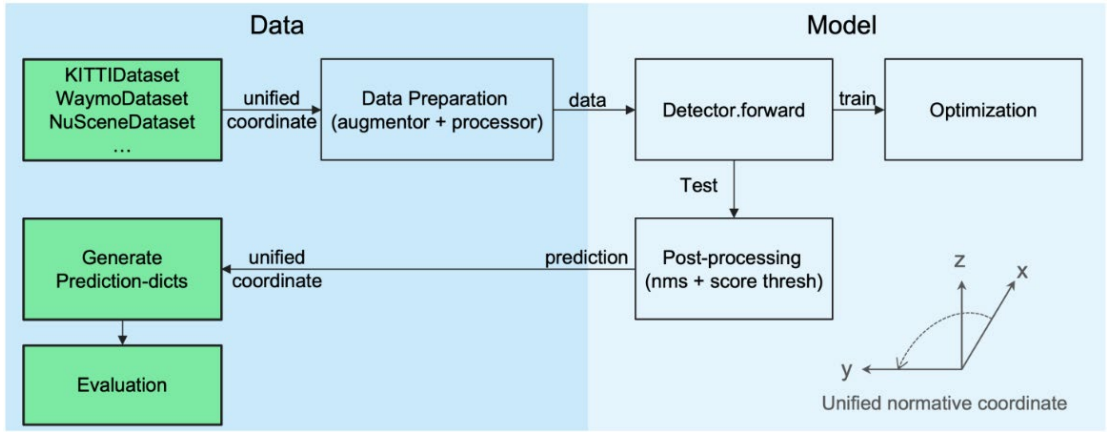


Source: https://support.google.com/waymo/answer/9190838?hl=en

Experiments: Setup

Dataset	KITTI	Waymo
GPU	NVIDIA GTX3090 (24GB)	NVIDIA GTX4080 (16GB)
Training Set	3712 samples	20% set (31618 samples)
Validation Set	3769 samples	100% set (39987 samples)
Batch size	2	2
Epoch	80	30
LR	0.01	0.01
Optimizer	Adam OneCycle	Adam OneCycle

OpenPCDet Workflow



Experiments:

FLOPs & Params.

Method	avg. 3D AP@R40 (%) ↑		FLOPs (G) ↓		Params (M) ↓	
PV-RCNN (*Baseline*)	73.61		178.5		12.41	
PV-RCNN_Mamba	**74.86**	(+1.25)	110.1	(−38.3%)	9.87	(−20.5%)
PV-RCNN_Rel	73.99		180.3		13.47	
PV-RCNN_Rel_Mamba	74.13	(+0.14)	111.9	(−37.9%)	10.94	(−18.8%)
PartA2 (*Baseline*)	72.88		165.1		61.60	
PartA2_Mamba	**74.63**	(+1.75)	96.7	(−41.4%)	59.11	(−4.0%)
PartA2_Rel	70.02		168.7		62.71	
PartA2_Rel_Mamba	71.33	(+1.31)	100.2	(−40.6%)	60.22	(−4.0%)

Mamba reduces FLOPs by around **40%**, and meanwhile improves detection accuracy.

Experiments: Results on KITTI

3D

Method	Car 3D AP 0.7 (%) ↑			Pedestrian 3D AP 0.5 (%) ↑			Cyclist 3D AP 0.5 (%) ↑		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PV-RCNN	91.87	84.53	82.41	**65.38**	**57.99**	**53.17**	89.95	70.81	66.37
PV-RCNN_Rel_Mamba	**92.41**	**85.27**	**82.93**	62.86	56.17	50.95	**92.02**	**74.68**	**69.90**
Improvements	+0.54	+0.74	+0.52	-2.52	-1.82	-2.22	+2.07	+3.87	+3.53
PartA2	91.84	**82.53**	80.17	**64.13**	**58.04**	**53.08**	88.24	71.12	66.76
PartA2_Rel_Mamba	91.84	82.48	**80.35**	59.94	51.82	47.39	**88.36**	**71.54**	**68.22**
Improvements	+0.00	-0.05	+0.18	-4.19	-6.22	-5.69	+0.12	+0.42	+1.46

BEV

Method	Car BEV AP 0.7 (%) ↑			Pedestrian BEV AP 0.5 (%) ↑			Cyclist BEV AP 0.5 (%) ↑		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PV-RCNN	94.58	91.00	88.51	**68.47**	**61.37**	**56.51**	91.26	73.94	69.53
PV-RCNN_Rel_Mamba	**95.56**	**91.55**	**89.23**	66.49	58.77	53.64	**94.94**	**76.91**	**71.96**
Improvements	+0.98	+0.55	+0.72	-1.98	-2.60	-2.87	+3.68	+2.97	+2.43
PartA2	94.05	88.33	88.04	**66.73**	**61.47**	**56.71**	89.14	72.87	69.59
PartA2_Rel_Mamba	**95.20**	**90.76**	**88.83**	62.66	54.65	50.42	**93.84**	**76.39**	**73.19**
Improvements	+1.15	+2.43	+0.79	-4.07	-6.82	-6.29	+4.70	+3.52	+3.60

Experiments: Results on KITTI

Car only

Method	Car - 3D AP (%) ↑			Car - BEV AP (%) ↑		
	Easy	Mod.	Hard	Easy	Mod.	Hard
SECOND [YML18]	87.43	76.48	69.10	89.96	87.07	79.66
CIA-SSD [Zhe+21]	90.04	79.81	78.80	-	-	-
SSL Point-GNN [Erç+22]	91.43	82.85	80.12	93.55	89.79	87.23
PointRCNN [SWL19]	88.88	78.63	77.38	-	-	-
PartA2	92.28	82.70	80.41	93.33	**89.69**	88.40
PartA2_Rel (ours)	**92.53**	**83.15**	**80.88**	**95.89**	89.45	**89.19**
PartA2_Rel_Mamba (ours)	89.62	82.35	80.39	93.45	88.96	86.94
PV-RCNN	91.91	84.78	82.63	93.17	90.72	88.73
PV-RCNN_Rel (ours)	**92.73**	**85.52**	**83.21**	**95.48**	**91.25**	89.09
PV-RCNN_Rel_Mamba (ours)	92.12	83.12	82.81	93.43	91.04	**89.16**

- Mamba reduces detection accuracy for single-class training

Ablation Studies:

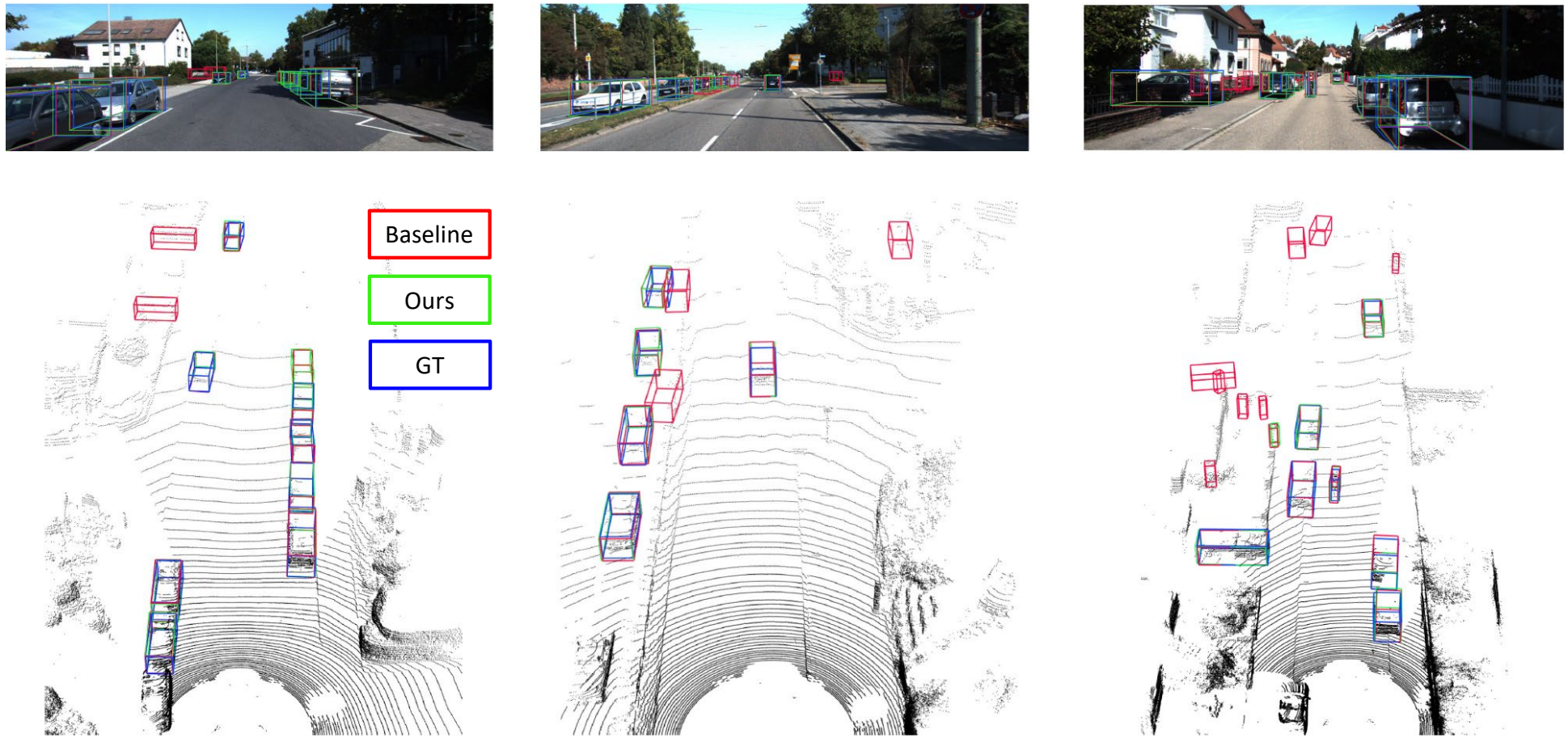
Effect of Mamba-based 2D Backbone

Method		Model Config		Car 3D AP (%) ↑			Pedestrian 3D AP (%) ↑			Cyclist 3D AP (%) ↑		
		Relation	Mamba	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PV-RCNN	1)	-	-	91.87	84.53	82.41	65.38	57.99	53.17	89.95	70.81	66.37
	2)	-	✓	91.48	84.12	82.28	**66.28**	**59.03**	**53.70**	**93.04**	73.99	69.82
	3)	✓	-	**92.45**	**85.36**	**83.00**	65.93	58.13	51.81	91.52	71.13	66.59
	4)	✓	✓	92.41	85.27	82.93	62.86	56.17	50.95	92.02	**74.68**	**69.90**
PartA2	1)	-	-	**91.84**	82.53	80.17	64.13	58.04	53.08	88.24	71.12	66.76
	2)	-	✓	**91.84**	**84.00**	**82.10**	**67.72**	**60.72**	**54.64**	90.20	**72.02**	**68.39**
	3)	✓	-	91.72	82.59	80.43	54.10	48.37	43.63	**90.78**	70.93	67.60
	4)	✓	✓	**91.84**	82.48	80.35	59.94	51.82	47.39	88.36	71.54	68.22

PV-RCNN: both Mamba-based 2D backbone and Object Relation Module improves detection accuracy;

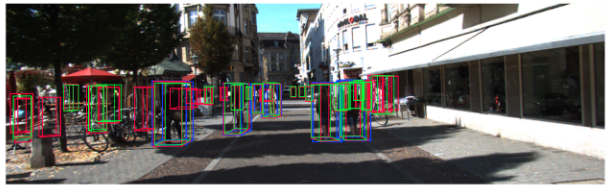
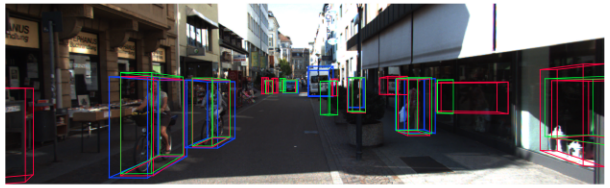
PartA2: Object Relation Module leads to worse detection accuracy in Pedestrian class;
barely implementing Mamba-based 2D backbone gets the best performance

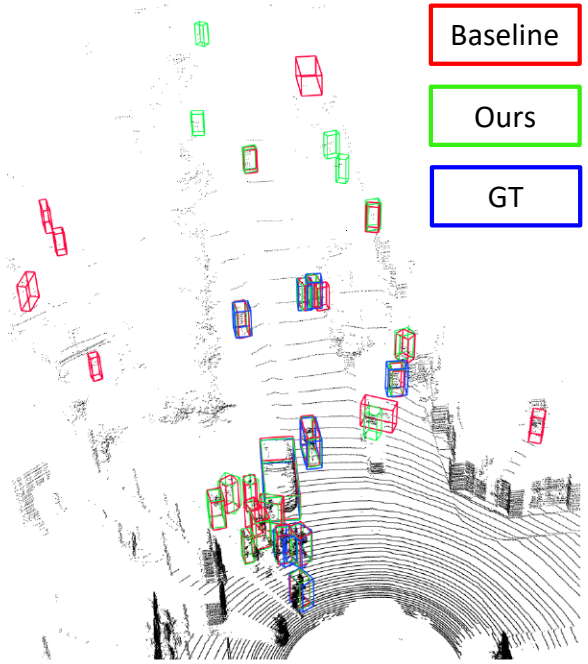
Experiments: Qualitative Results on KITTI



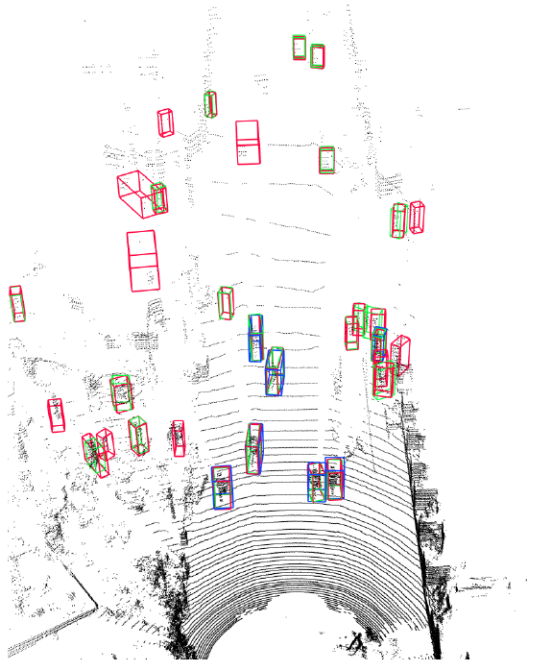
(a) frame 159 **(b)** frame 216 **(c)** frame 335

Experiments: Qualitative Results on KITTI

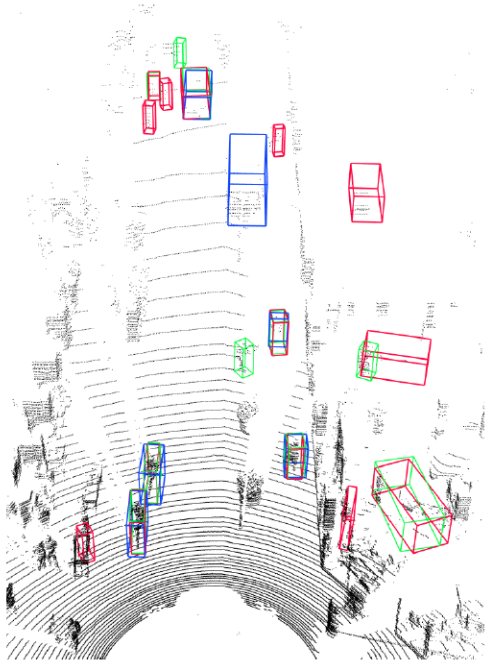






(d) frame 6908



(e) frame 7318



(f) frame 7469

Experiments: Results on Waymo

PV-RCNN vs. PV-RCNN_Mamba

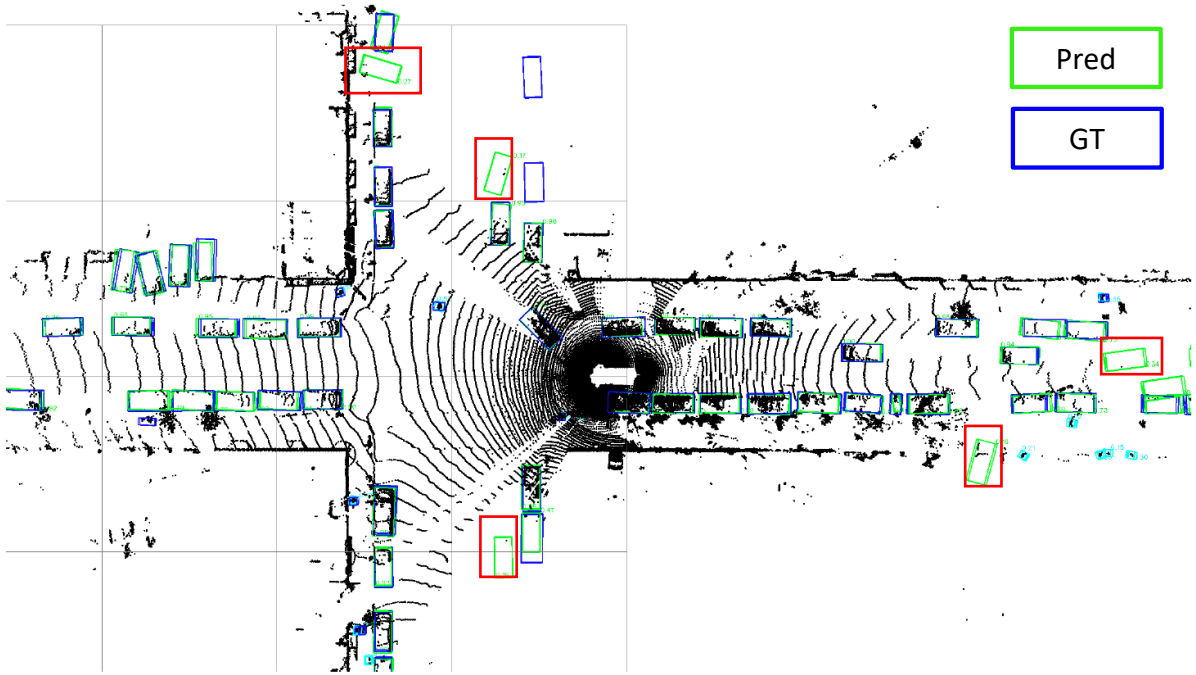
Methods	Vehicle (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN	75.19	74.53	66.59	65.99
PV-RCNN_Mamba	**76.31**	**75.71**	**67.79**	**67.25**
Improvements	+1.12	+1.18	+1.20	+1.26
	Pedestrian (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN	71.46	34.88	62.51	30.53
PV-RCNN_Mamba	**74.00**	**37.27**	**65.33**	**32.93**
Improvements	+2.54	+2.39	+2.82	+2.40
	Cyclist (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN	50.54	**30.46**	48.61	**29.30**
PV-RCNN_Mamba	**50.92**	28.22	**48.98**	27.15
Improvements	+0.38	-2.24	+0.37	-2.15

Experiments: Results on Waymo

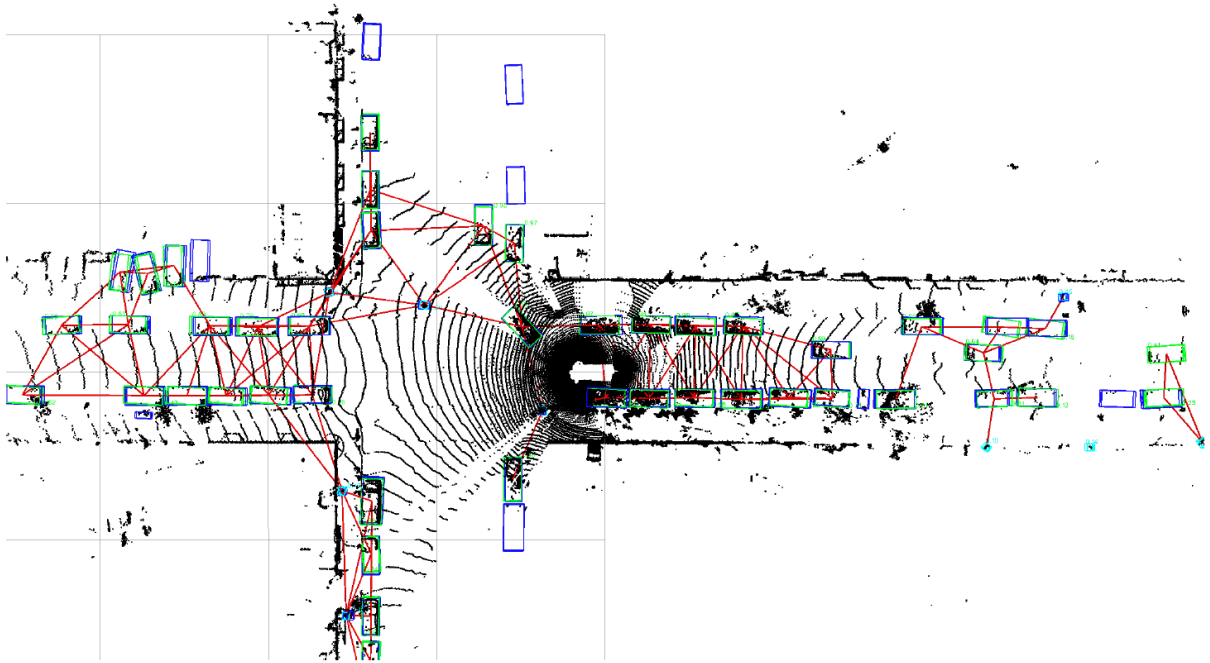
PV-RCNN_Rel vs. PV-RCNN_Rel_Mamba

Methods	Vehicle (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN_Rel	70.28	69.54	61.60	60.94
PV-RCNN_Rel_Mamba	**74.86**	**74.28**	**66.15**	**65.63**
Improvements	+4.58	+4.74	+4.55	+4.69
	Pedestrian (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN_Rel	61.87	31.53	53.14	27.09
PV-RCNN_Rel_Mamba	**70.92**	**34.41**	**61.98**	**30.10**
Improvements	+9.05	+2.88	+8.84	+3.01
	Cyclist (%) ↑			
	Level 1 mAP	Level 1 mAPH	Level 2 mAP	Level 2 mAPH
PV-RCNN_Rel	33.66	14.8	32.37	14.24
PV-RCNN_Rel_Mamba	**44.34**	**24.88**	**42.64**	**23.92**
Improvements	+10.68	+10.08	+10.27	+9.68

Experiments: Qualitative Results on Waymo



PV-RCNN

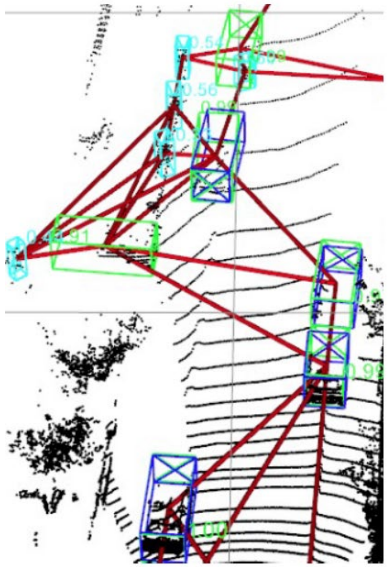


PV-RCNN_Rel

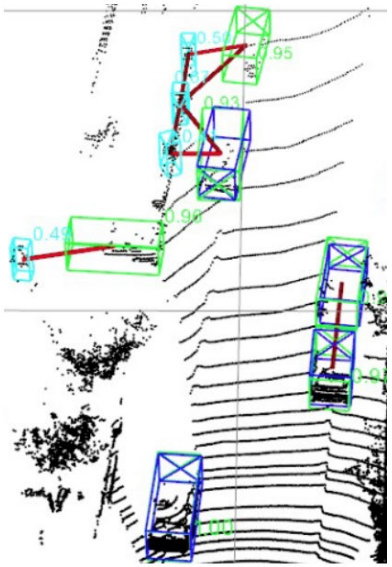
FP eliminated by object relation module

Ablation Studies:

Effect of Different Graph constructors



(a) KNN (k=16)

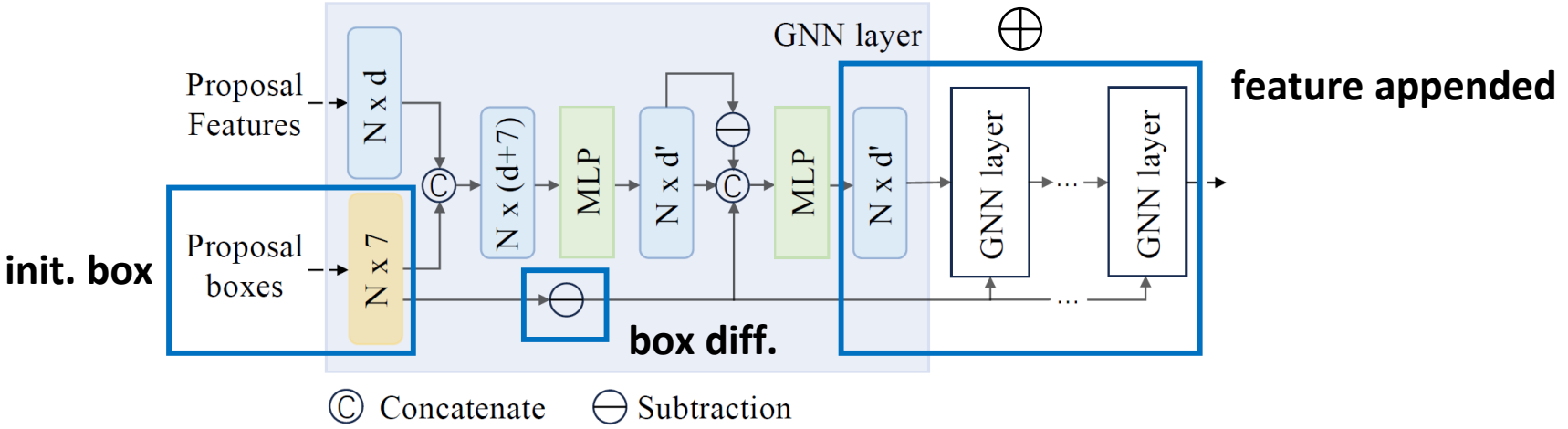


(b) Radius (r=6m)

Method	3D AP (Mod.) (%) ↑		
	Car	Pedestrian	Cyclist
PV-RCNN Relation (Radius)	83.25	**59.28**	**74.29**
PV-RCNN Relation (KNN)	**84.68**	57.31	73.29

Ablation Studies:

Effect of GNN Components

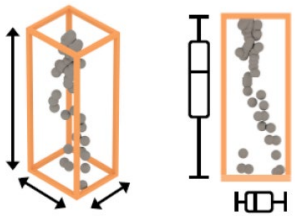


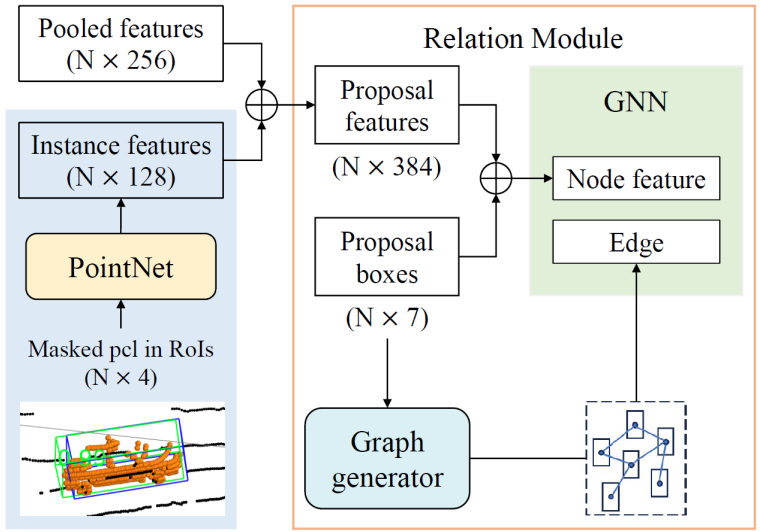
Method	init. box	box diff.	feature appended	3D AP (Mod.) (%) ↑
Exp. 1	-	-	-	82.91
Exp. 2	✓	-	-	84.07
Exp. 3	✓	✓	-	84.86
Exp. 4	✓	✓	✓	**85.36**

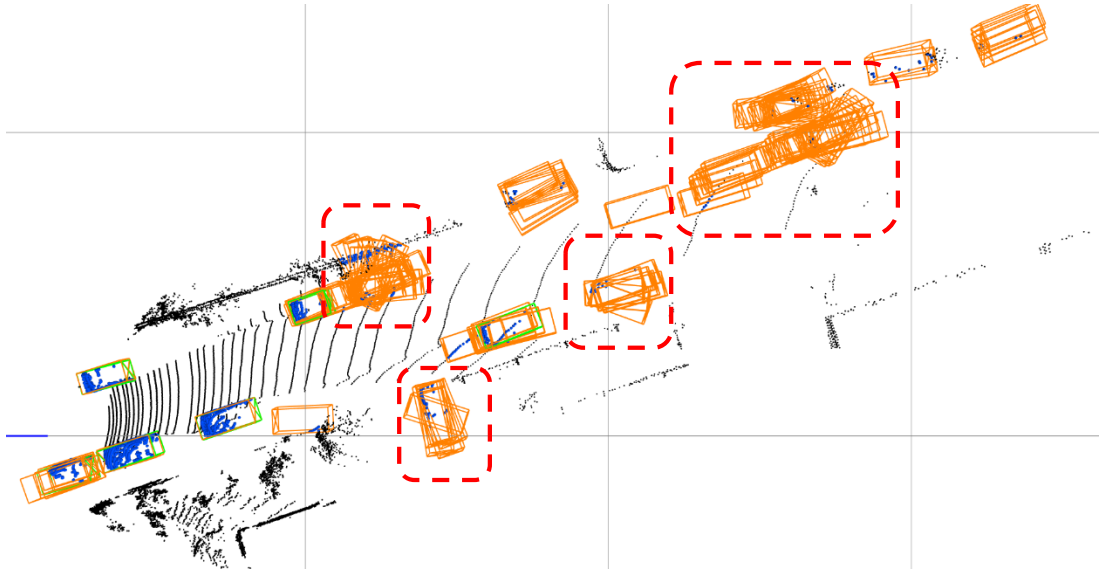
Ablation Studies:

Effect of Additional Instance Feature

GACE







Method	Car 3D AP (%) ↑			Pedestrian 3D AP (%) ↑			Cyclist 3D AP (%) ↑		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PV-RCNN	91.87	84.53	82.41	65.38	57.99	**53.17**	89.95	70.81	66.37
PV-RCNN_Rel	**92.45**	**85.36**	**83.00**	**65.93**	**58.13**	51.81	**91.52**	**71.13**	**66.59**
PV-RCNN_Rel+IF	92.28	83.50	82.74	60.41	52.70	48.77	89.88	70.96	66.58
Decrements	-0.17	-1.86	-0.26	-5.52	-5.43	-3.04	-1.64	-0.17	-0.01

Conclusions & Future Work

Conclusions

- Compared to CNN-based feature encoders, Mamba-based encoders reduce the **amount of parameters** by **20%** for **PV-RCNN**, **4%** for **PartA2**, and reduce **FLOPs** by around **40%** for both baseline models.
- Involving object relation shows **improvements** in detection performance of **car** and **cyclist** classes, but shows **decrements** in **pedestrian** class for both PV-RCNN and PartA2 on KITTI dataset.
- Qualitatively, involving contextual information via object relation **reduces FP** effectively.

Future work

- Apply Mamba-based 2D backbone and object relation in more 3D object detection networks, train and evaluate on more datasets.
- Explore to apply Mamba-based methods in multi-modal detection.

Thanks for your attention

Q&A