State Space Model Meets Transformer: A New Paradigm for 3D Object Detection

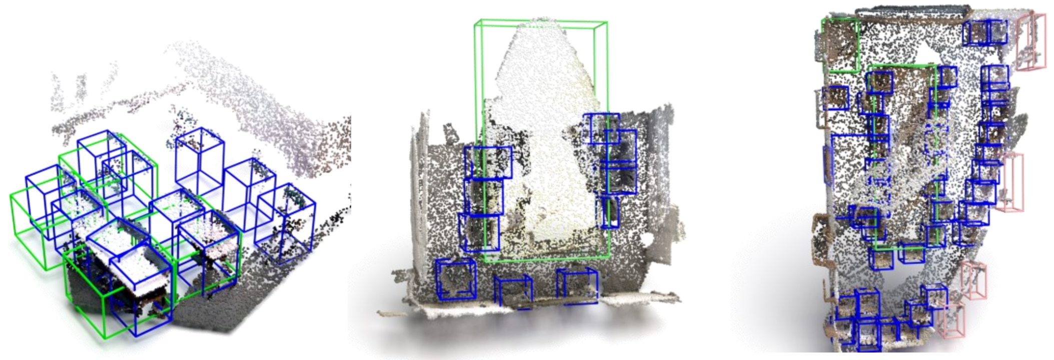
Chuxin Wang[1,2] Wenfei Yang[1,2] Xiang Liu[4] Tianzhu Zhang[1,2,3]

[1]University of Science and Technology of China [2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory

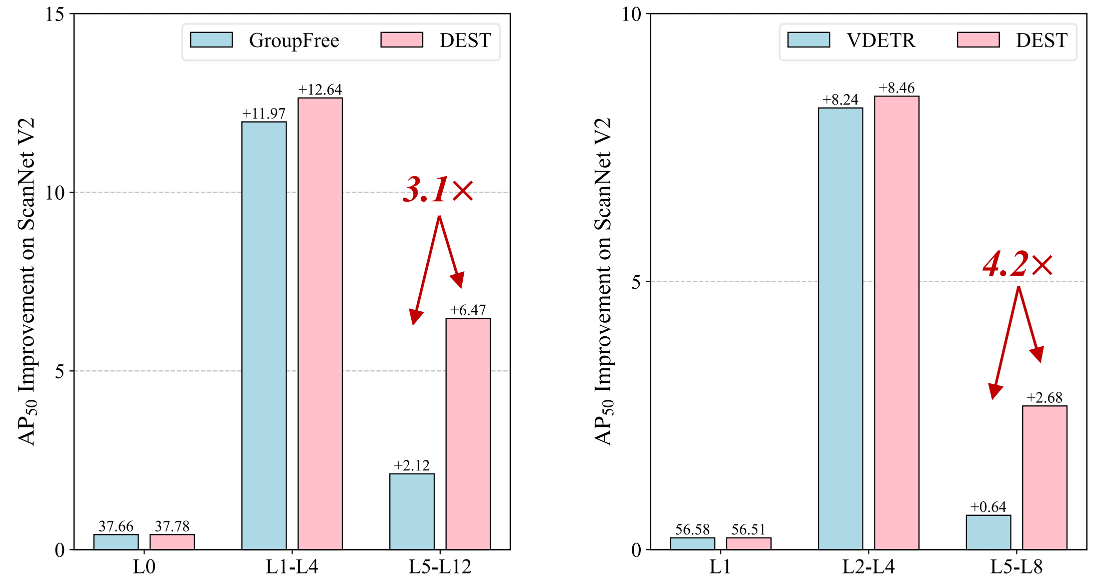
[3]Hainan Aerospace Technology Innovation Center [4]Dongguan University of Technology

Indoor 3D Object Detection

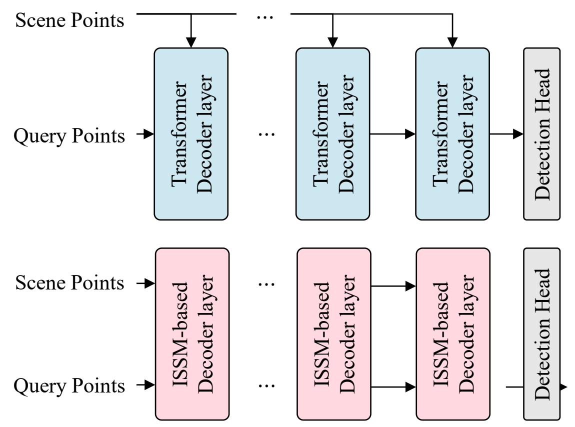
Perceive and locate 3D objects in the real world



Challenges and Contributions

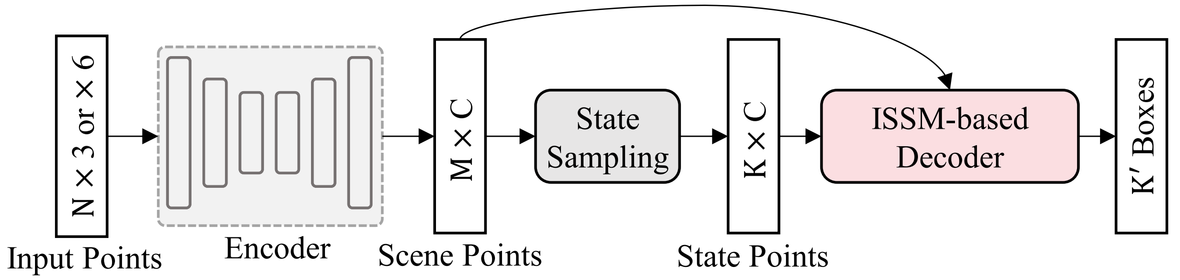


DETR-based models show **limited improvement** in later layers due to **fixed scene point features**, while our DEST dynamically updates them, achieving significant gains.



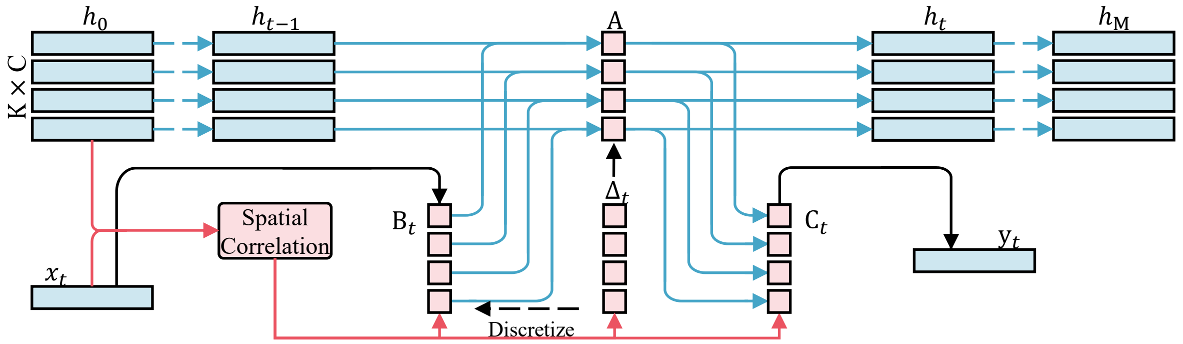
- Transformer decoder **solely** updates query point features.
- Can we design a **State Space Model** to replace it, enabling **simultaneous updates** of scene and query point features?

Methods



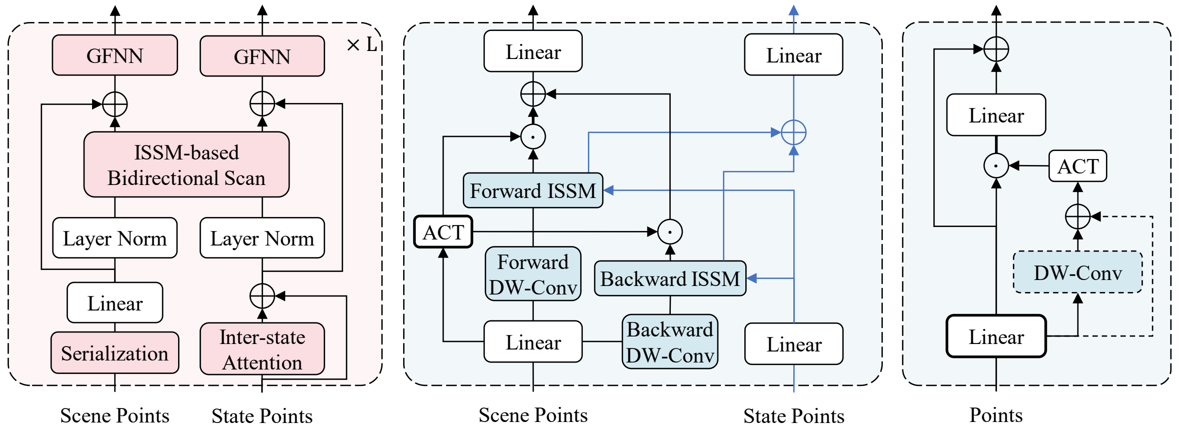
DEST-based framework emphasizes **innovative decoder design**.

(A) Interactive State Space Model (ISSM)



- Query point features are modeled as **system states**, while scene point features serve as **system inputs** at different time steps.
- ISSM modifies SSM parameters (Δ_t, B_t, C_t) to be dependent on system states and introduces a **spatial correlation** module to capture relationships between state points and scene points.

(B) ISSM-based Decoder Block



(a) ISSM-based Decoder (b) ISSM-based Bidirectional Scan (c) GFFN

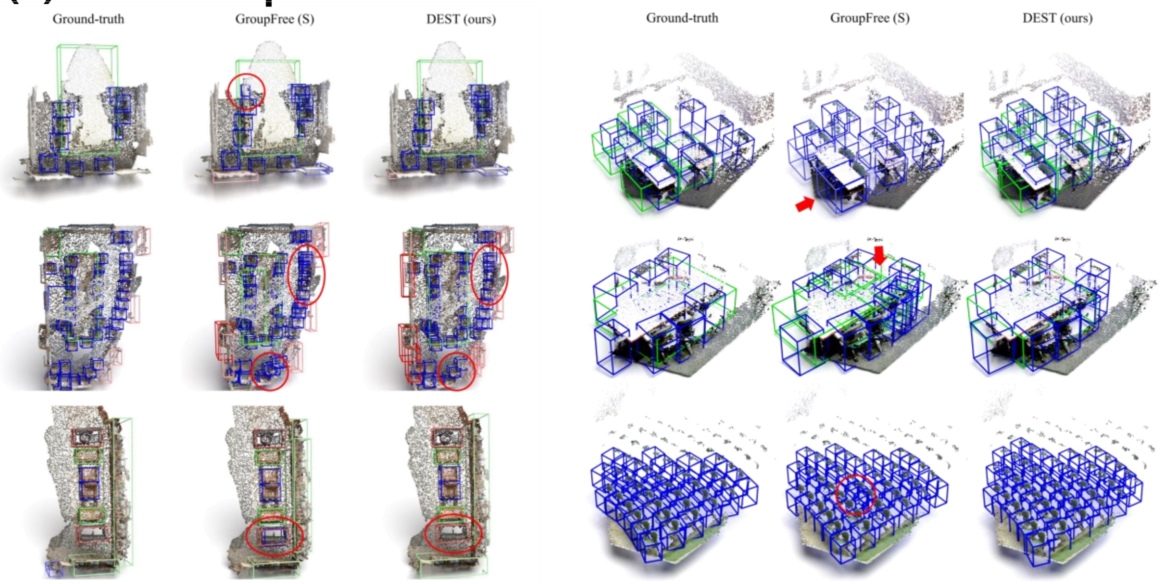
- We design the ISSM-based decoder tailored to the characteristics of 3D point clouds, **fully harnessing the potential of the ISSM** for point cloud object detection.

Experiments

(A) Results on ScanNet V2 and SUN RGB-D datasets

Method	RGB	ScanNet V2(H)		ScanNet V2(A)		SUN RGB-D(H)		SUN RGB-D(A)	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}	AP_{25}	AP_{50}
VoteNet (Qi et al., 2019)	✗	62.9	39.9	-	-	57.7	-	-	-
HGNet (Chen et al., 2020)	✗	61.3	34.4	-	-	61.6	-	-	-
3D-MPA (Engelmann et al., 2020)	✗	64.2	49.2	-	-	-	-	-	-
MLCVNet (Xie et al., 2020)	✗	64.5	41.4	-	-	59.8	-	-	-
GSDN (Gwak et al., 2020)	✗	62.8	34.8	-	-	-	-	-	-
H3DNet (Zhang et al., 2020)	✗	64.4	43.4	-	-	60.1	39.0	-	-
BRNet (Cheng et al., 2021)	✗	66.1	50.9	-	-	61.1	43.7	-	-
3DETR (Misra et al., 2021)	✗	65.0	47.0	-	-	59.1	32.7	-	-
VENet (Xie et al., 2021)	✗	67.7	-	-	-	62.5	39.2	-	-
GroupFree(S)(Liu et al., 2021)	✗	67.3	48.9	66.3	48.5	63.0	45.2	62.6	44.4
GroupFree(L)(Liu et al., 2021)	✗	69.1	52.8	68.6	51.8	-	-	-	-
RBGNet (Wang et al., 2022b)	✗	70.6	55.2	69.9	54.7	64.1	47.2	63.6	46.3
HyperDet3D (Zheng et al., 2022)	✗	70.9	57.2	-	-	63.5	47.3	-	-
LeadNet (Wang et al., 2023)	✗	68.0	51.3	-	-	63.4	45.8	-	-
FCAF3D (Rukhovich et al., 2022)	✓	71.5	57.3	70.7	56.0	64.2	48.9	63.8	48.2
TR3D (Rukhovich et al., 2023)	✓	72.9	59.3	72.0	57.4	67.1	50.4	66.3	49.6
CAGroup3D (Wang et al., 2022a)	✓	75.1	61.3	74.5	60.3	66.8	50.2	66.4	49.5
VDETR (Shen et al., 2024)	✓	77.4	65.0	76.8	64.5	67.5	50.4	66.8	49.7
VDETR(TTA) (Shen et al., 2024)	✓	77.8	66.0	77.0	65.3	68.0	51.1	67.5	50.0
GroupFree(S)(Liu et al., 2021)	✗	67.3	48.9	66.3	48.5	63.0	45.2	62.6	44.4
+ DEST(ours)	✗	68.8(+1.5)	53.2(+4.3)	67.9(+1.6)	52.7(+4.2)	65.3(+2.3)	48.4(+3.2)	64.7(+2.1)	47.6(+3.2)
GroupFree(L)(Liu et al., 2021)	✗	69.1	52.8	68.6	51.8	-	-	-	-
+ DEST(ours)	✗	71.3(+2.2)	58.1(+5.3)	70.5(+1.9)	56.8(+5.0)	-	-	-	-
VDETR (Shen et al., 2024)	✓	77.4	65.0	76.8	64.5	67.5	50.4	66.8	49.7
+ DEST(ours)	✓	78.5(+1.1)	66.6(+1.6)	77.8(+1.0)	66.2(+1.7)	68.4(+0.9)	51.8(+1.4)	67.4(+0.8)	50.9(+1.2)
VDETR(TTA) (Shen et al., 2024)	✓	77.8	66.0	77.0	65.3	68.0	51.1	67.5	50.0
+ DEST(ours)	✓	78.8(+1.0)	67.9(+1.9)	78.3(+1.3)	66.9(+1.6)	69.2(+1.2)	52.2(+1.1)	68.8(+1.3)	51.6(+1.6)

(B) Visual Comparison on ScanNet V2 and SUN RGB-D datasets



Our DEST-based methods **significantly outperform** the baseline methods on both ScanNet V2 and SUN RGB-D datasets.