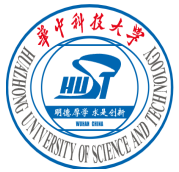
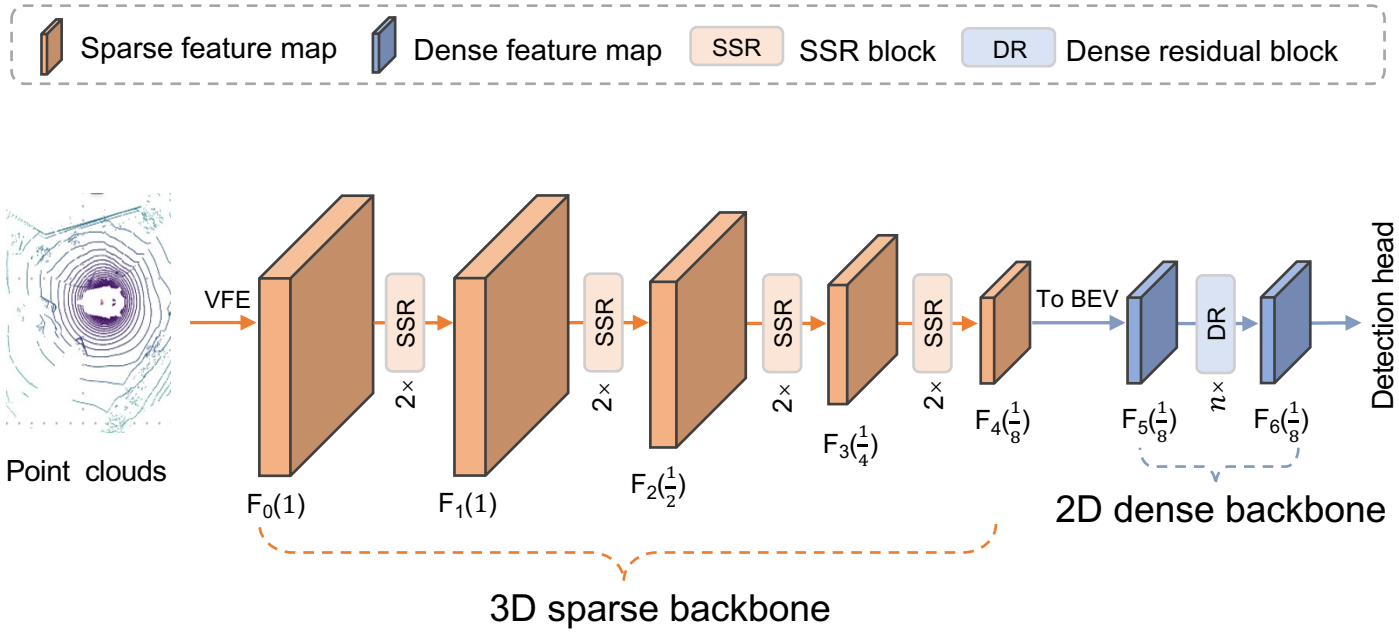
HEDNet: A Hierarchical Encoder-Decoder Network for 3D Object Detection in Point Clouds

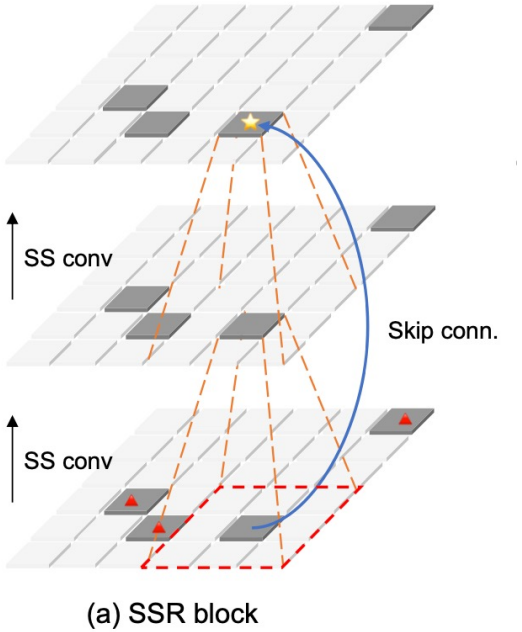
Gang Zhang[1] Junnan Chen[2] Guohuan Gao[3] Jianmin Li[1] Xiaolin Hu[1]*

Tsinghua University



Review VoxelNet





submanifold sparse (SS) residual block

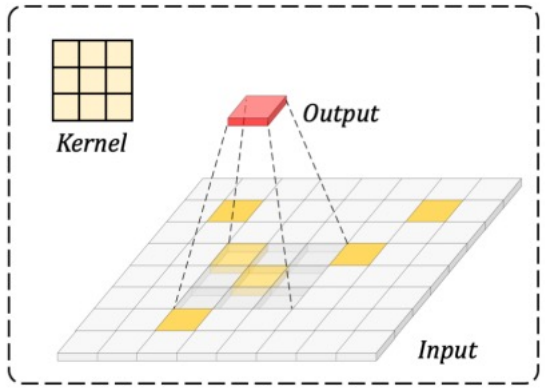
Limited receptive filed:

submanifold sparse (SS) convolution with small kernels, *i.e.* 3x3x3

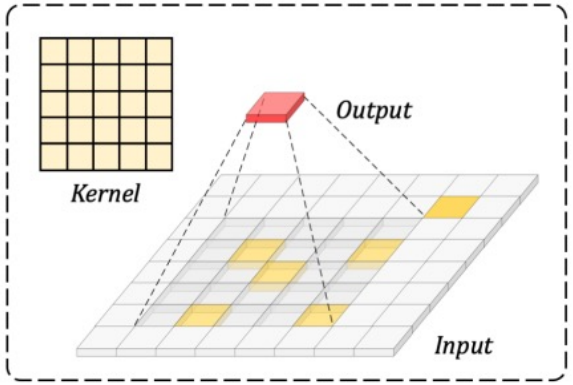
VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. [Zhou et al. CVPR 2019]

Related work

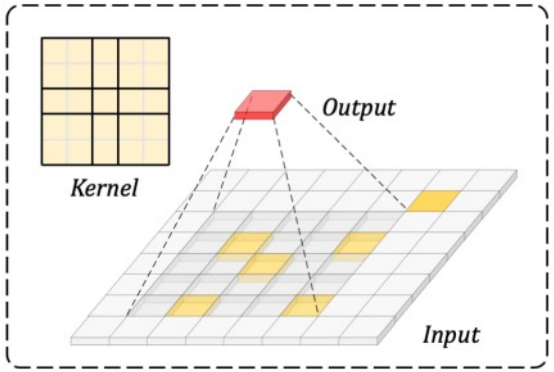
LargeKernel3D (Large convolutional kernel)



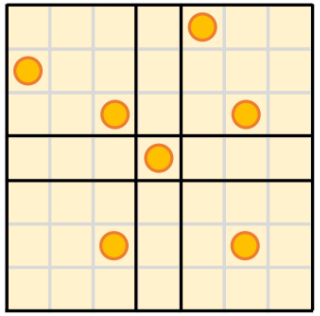
Small-kernel sparse conv



Large-kernel sparse conv



Spatial-wise partition conv


Input & Weight

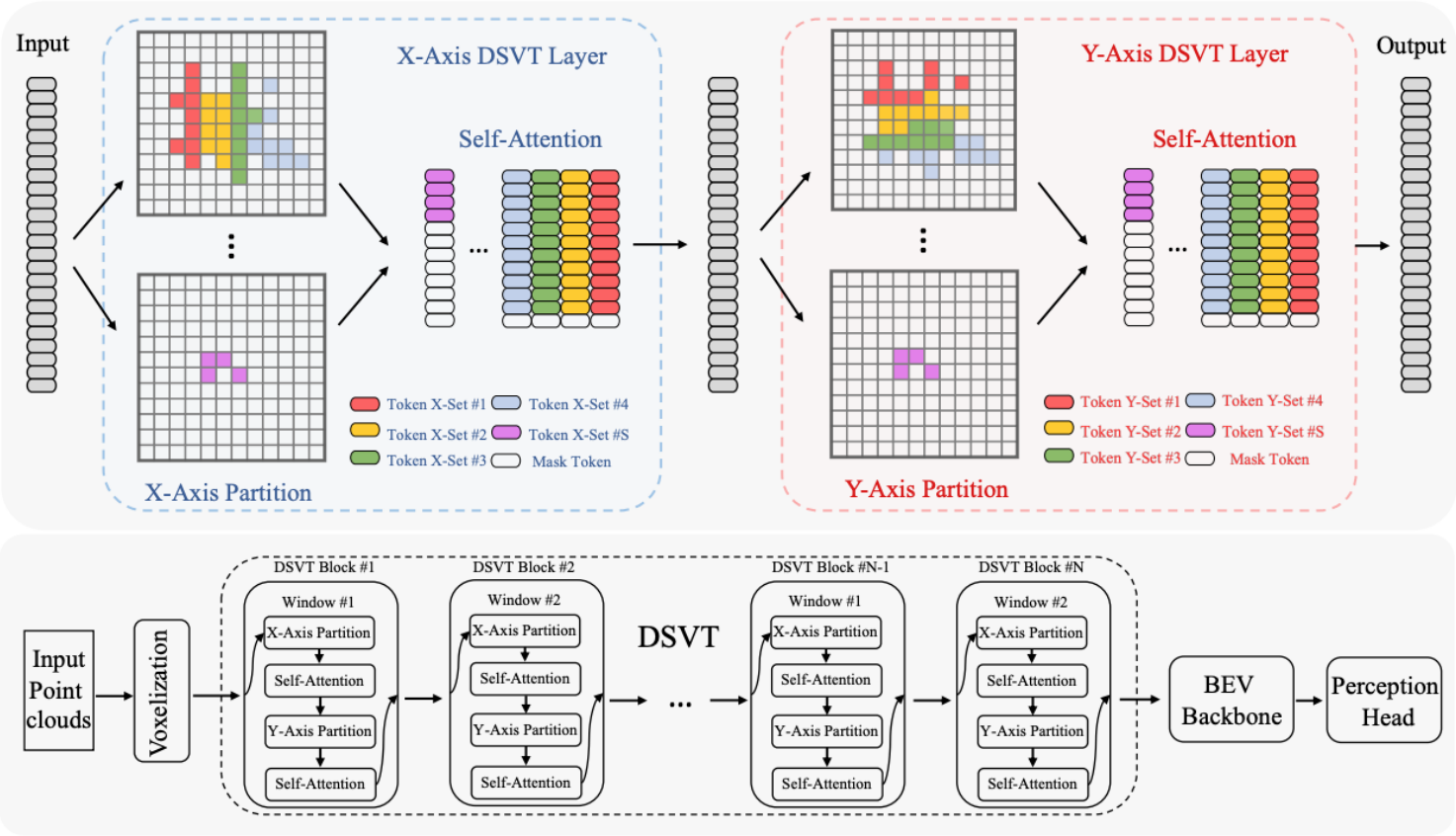
Problems:

- Overfitting
- Low efficiency

LargeKernel3D: Scaling up Kernels in 3D Sparse CNNs. [Chen et al. CVPR 2023]

Related work

DSVT (Transformer)

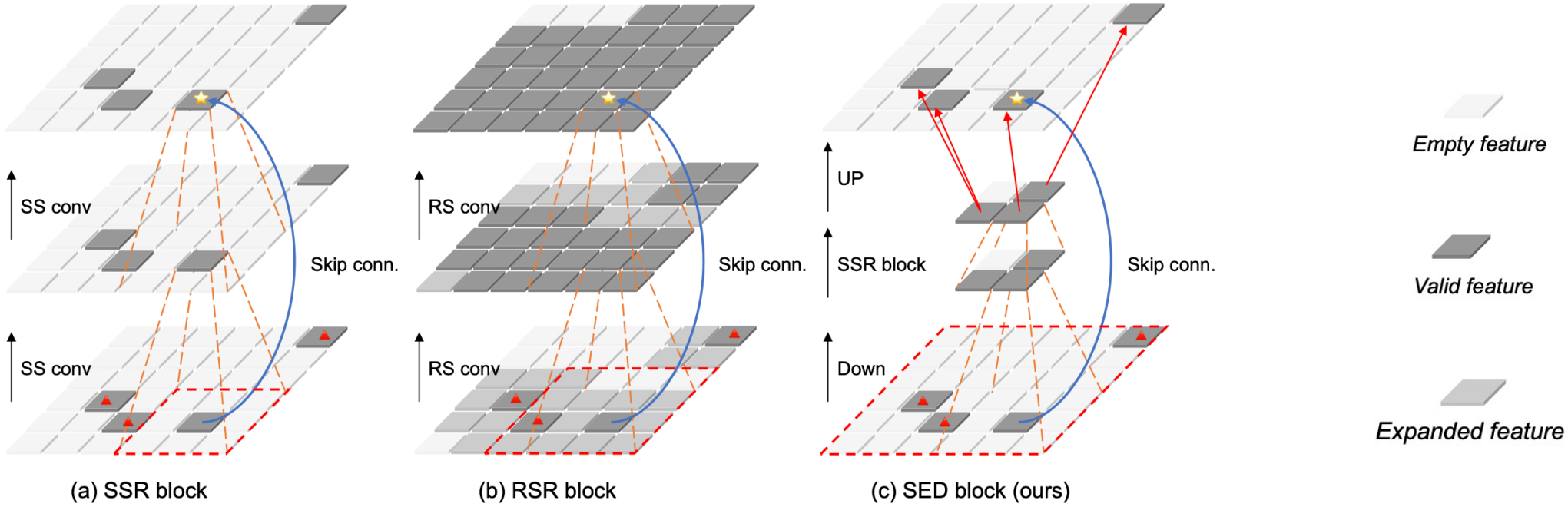


Problems:

- Low efficiency (single stride)

DSVT: Dynamic Sparse Voxel Transformer with Rotated Sets. [Wang et al. CVPR 2023]

Our solution (Encoder-Decoder Block)

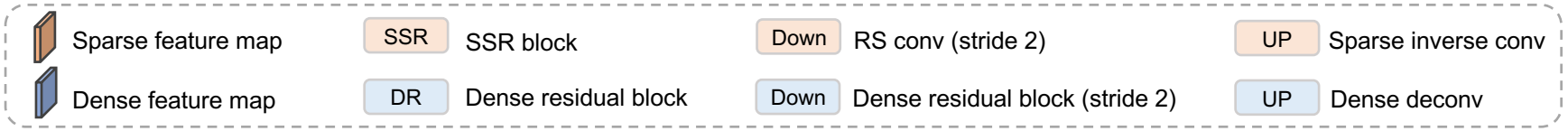


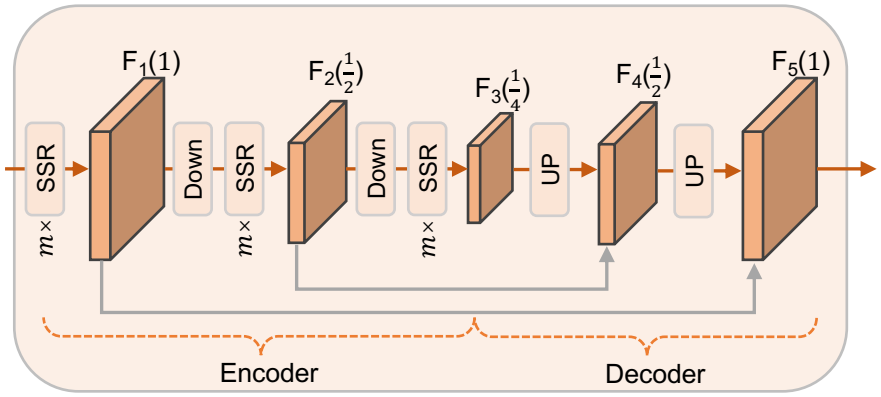
(a) SSR block (b) RSR block (c) SED block (ours)

submanifold sparse residual *regular sparse residual* sparse encoder-decoder

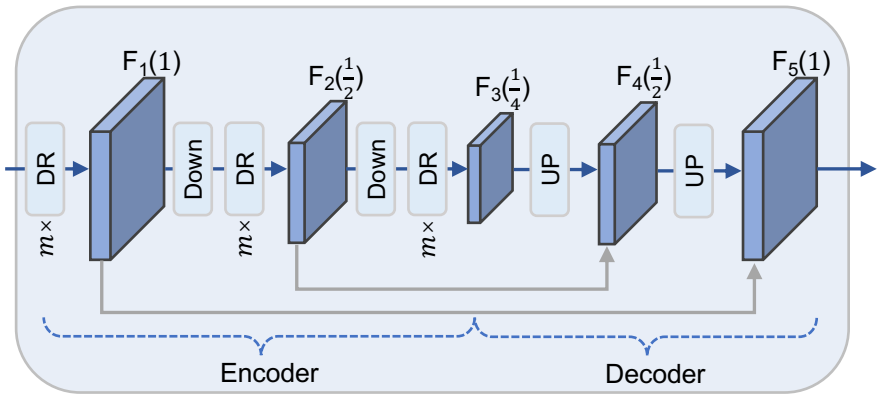
Limited receptive field Low efficiency Expanded receptive field w/. high efficiency

Our solution (Encoder-Decoder Block)





Sparse encoder-decoder (SED) block

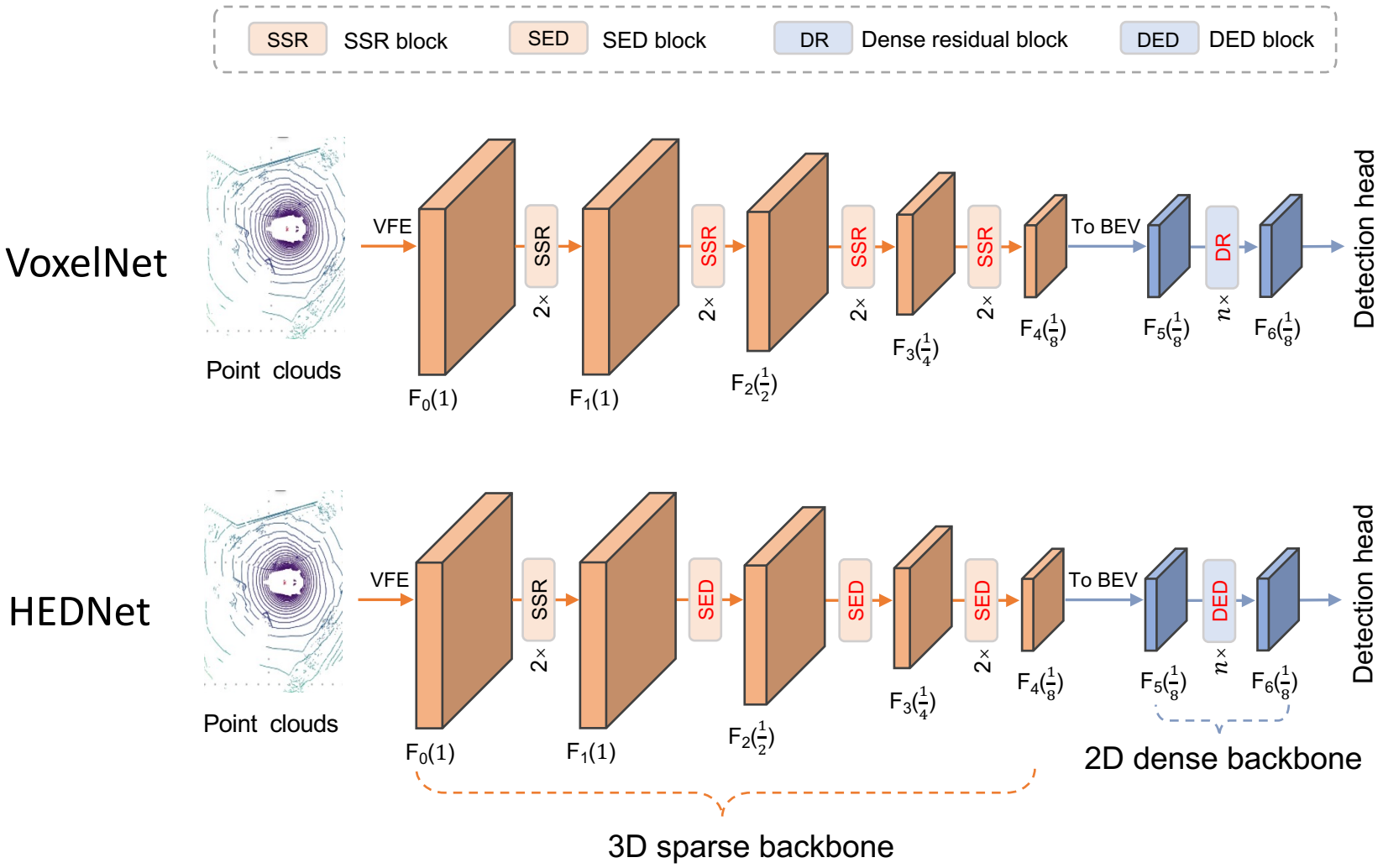


Dense encoder-decoder (DED) block

To capture long-range dependency

To expand features to object centers

Our solution (HEDNet)



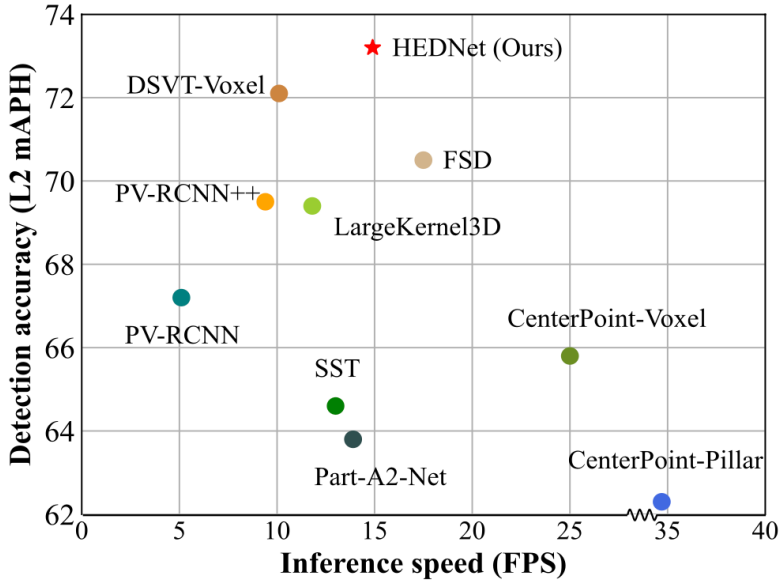
Results on Waymo Open

Results on the validation data set

Method	mAP/mAPH L2	Vehicle AP/APH L1	Vehicle AP/APH L2	Pedestrian AP/APH L1	Pedestrian AP/APH L2	Cyclist AP/APH L1	Cyclist AP/APH L2
SECOND [18]	61.0/57.2	72.3/71.7	63.9/63.3	68.7/58.2	60.7/51.3	60.6/59.3	58.3/57.0
PointPillar [19]	62.8/57.8	72.1/71.5	63.6/63.1	70.6/56.7	62.8/50.3	64.4/62.3	61.9/59.9
Lidar-RCNN [42]†	65.8/61.3	76.0/75.5	68.3/67.9	71.2/58.7	63.1/51.7	68.6/66.9	66.1/64.4
Part-A2-Net [37]†	66.9/63.8	77.1/76.5	68.5/68.0	75.2/66.9	66.2/58.6	68.6/67.4	66.1/64.9
SST [14]	67.8/64.6	74.2/73.8	65.5/65.1	78.7/69.6	70.0/61.7	70.7/69.6	68.0/66.9
CenterPoint [27]	68.2/65.8	74.2/73.6	66.2/65.7	76.6/70.5	68.8/63.2	72.3/71.1	69.7/68.5
PV-RCNN [43]†	69.6/67.2	78.0/77.5	69.4/69.0	79.2/73.0	70.4/64.7	71.5/70.3	69.0/67.8
CenterPoint [27]†	69.8/67.6	76.6/76.0	68.9/68.4	79.0/73.4	71.0/65.8	72.1/71.0	69.5/68.5
SWFormer [13]	-/-	77.8/77.3	69.2/68.8	80.9/72.7	72.5/64.9	-/-	-/-
OcTr [44]	70.7/68.2	78.1/77.6	69.8/69.3	80.8/74.4	72.5/66.5	72.6/71.5	69.9/68.9
PillarNet-34 [11]	71.0/68.5	79.1/78.6	70.9/70.5	80.6/74.0	72.3/66.2	72.3/71.2	69.7/68.7
AFDetV2 [45]	71.0/68.8	77.6/77.1	69.7/69.2	80.2/74.6	72.2/67.0	73.7/72.7	71.0/70.1
CenterFormer [46]	71.1/68.9	75.0/74.4	69.9/69.4	78.6/73.0	73.6/68.3	72.3/71.3	69.8/68.8
LargeKernel3D[33]	-/-	78.1/77.6	69.8/69.4	-/-	-/-	-/-	-/-
PV-RCNN++ [47]†	71.7/69.5	79.3/78.8	70.6/70.2	81.3/76.3	73.2/68.0	73.7/72.7	71.2/70.2
FSD [48]†	72.7/70.5	79.5/79.0	70.3/69.9	83.6/78.2	74.4/69.4	75.3/74.1	73.3/72.1
DSVT-Voxel [15]	74.0/72.1	79.7/79.3	71.4/71.0	83.7/78.9	76.1/71.5	77.5/76.5	74.6/73.7
HEDNet (ours)	**75.3/73.4**	81.1/80.6	**73.2/72.7**	84.4/80.0	76.8/72.6	78.7/77.7	75.8/74.9

Results on the test data set

Method	mAP/mAPH L2	Vehicle AP/APH L1	Vehicle AP/APH L2	Pedestrian AP/APH L1	Pedestrian AP/APH L2	Cyclist AP/APH L1	Cyclist AP/APH L2
PV-RCNN [43]	71.3/68.8	80.6/80.1	72.8/72.4	78.2/72.0	71.8/66.0	71.8/70.4	69.1/67.8
PV-RCNN++ [47]	72.4/70.2	81.6/81.2	73.9/73.5	80.4/75.0	74.1/69.0	71.9/70.8	69.3/68.2
AFDetV2 [45]	72.2/70.3	80.5/80.0	73.0/72.6	79.8/74.3	73.7/68.6	72.4/71.2	69.8/69.7
FSD [48]	74.4/72.4	82.7/82.3	74.4/74.1	82.9/77.9	75.9/71.3	75.6/74.4	72.9/71.8
HEDNet (ours)	**76.9/75.0**	84.2/83.8	**77.0/76.6**	84.1/79.7	78.3/74.0	78.2/77.0	75.4/74.3



HEDNet vs DSVT (prior SOTA)

Faster: 15 vs 10 FPS

Better: 1.3% L2 mAPH gains

Results on nuScenes

Results on the validation data set

Method	NDS	mAP	Car	Truck	Bus	T.L.	C.V.	Ped.	M.T.	Bike	T.C.	B.R.
CenterPoint [27]	66.5	59.2	84.9	57.4	70.7	38.1	16.9	85.1	59.0	42.0	69.8	68.3
VoxelNeXt [50]	66.7	60.5	83.9	55.5	70.5	38.1	21.1	84.6	62.8	50.0	69.4	69.4
TransFusion-L [28]	70.1	65.5	86.9	60.8	73.1	43.4	25.2	87.5	72.9	57.3	77.2	70.3
HEDNet (Ours)	**71.4**	**66.7**	87.7	60.6	**77.8**	**50.7**	**28.9**	87.1	74.3	56.8	76.3	66.9

Results on the test data set

Method	NDS	mAP	Car	Truck	Bus	T.L.	C.V.	Ped.	M.T.	Bike	T.C.	B.R.
PointPillars [19]	45.3	30.5	68.4	23.0	28.2	23.4	4.1	59.7	27.4	1.1	30.8	38.9
3DSSD [3]	56.4	42.6	81.2	47.2	61.4	30.5	12.6	70.2	36.0	8.6	31.1	47.9
CBGS [51]	63.3	52.8	81.1	48.5	54.9	42.9	10.5	80.1	51.5	22.3	70.9	65.7
CenterPoint [27]	65.5	58.0	84.6	51.0	60.2	53.2	17.5	83.4	53.7	28.7	76.7	70.9
FCOS-LiDAR [9]	65.7	60.2	82.2	47.7	52.9	48.8	28.8	84.5	68.0	39.0	79.2	70.7
HotSpotNet [52]	66.0	59.3	83.1	50.9	56.4	53.3	23.0	81.3	63.5	36.6	73.0	71.6
CVCNET [53]	66.6	58.2	82.6	49.5	59.4	51.1	16.2	83.0	61.8	38.8	69.7	69.7
AFDetV2 [45]	68.5	62.4	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3	80.1	71.0
UVTR-L [54]	69.7	63.9	86.3	52.2	62.8	59.7	33.7	84.5	68.8	41.1	74.7	74.9
VISTA [55]	69.8	63.0	84.4	55.1	63.7	54.2	25.1	82.8	70.0	45.4	78.5	71.4
Focals Conv [56]	70.0	63.8	86.7	56.3	67.7	59.5	23.8	87.5	64.5	36.3	81.4	74.1
VoxelNeXt [50]	70.0	64.5	84.6	53.0	64.7	55.8	28.7	85.8	73.2	45.7	79.0	74.6
TransFusion-L [28]	70.2	65.5	86.2	56.7	66.3	58.8	28.2	86.1	68.3	44.2	82.0	78.2
LargeKernel3D [12]	70.6	65.4	85.5	53.8	64.4	59.5	29.7	85.9	72.7	46.8	79.9	75.5
LinK [16]	71.0	66.3	86.1	55.7	65.7	62.1	30.9	85.8	73.5	47.5	80.4	75.5
HEDNet (Ours)	**72.0**	**67.7**	87.1	56.5	**70.4**	**63.5**	**33.6**	87.9	70.4	44.8	85.1	78.1

Analysis

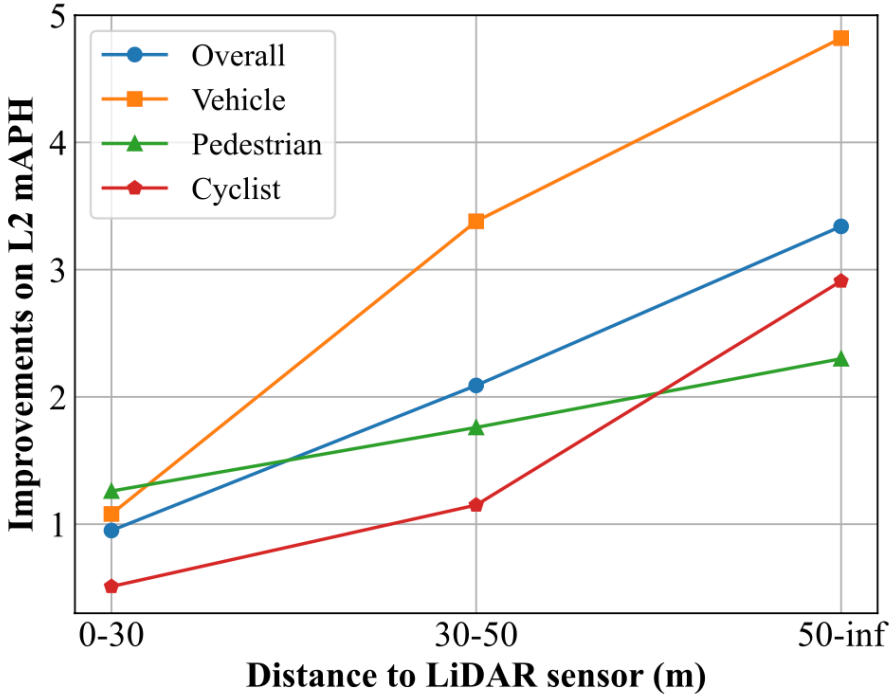
Block	Latency	L1 mAPH	L2 mAPH
RSR block	176 ms	74.61	68.30
SSR block	43 ms	74.42	67.93
SED block	48 ms	76.13	69.89
SSR block†	49 ms	76.67	70.49
SED block†	54 ms	77.39	71.37

(a) Effectiveness of the SED block.

#Scale	Latency	L1 mAPH	L2 mAPH	
1	43 ms	76.18	69.88	*(HEDNet-single)*
2	59 ms	77.61	71.44	
3	67 ms	78.02	71.92	*(HEDNet)*
4	78 ms	78.12	72.02	

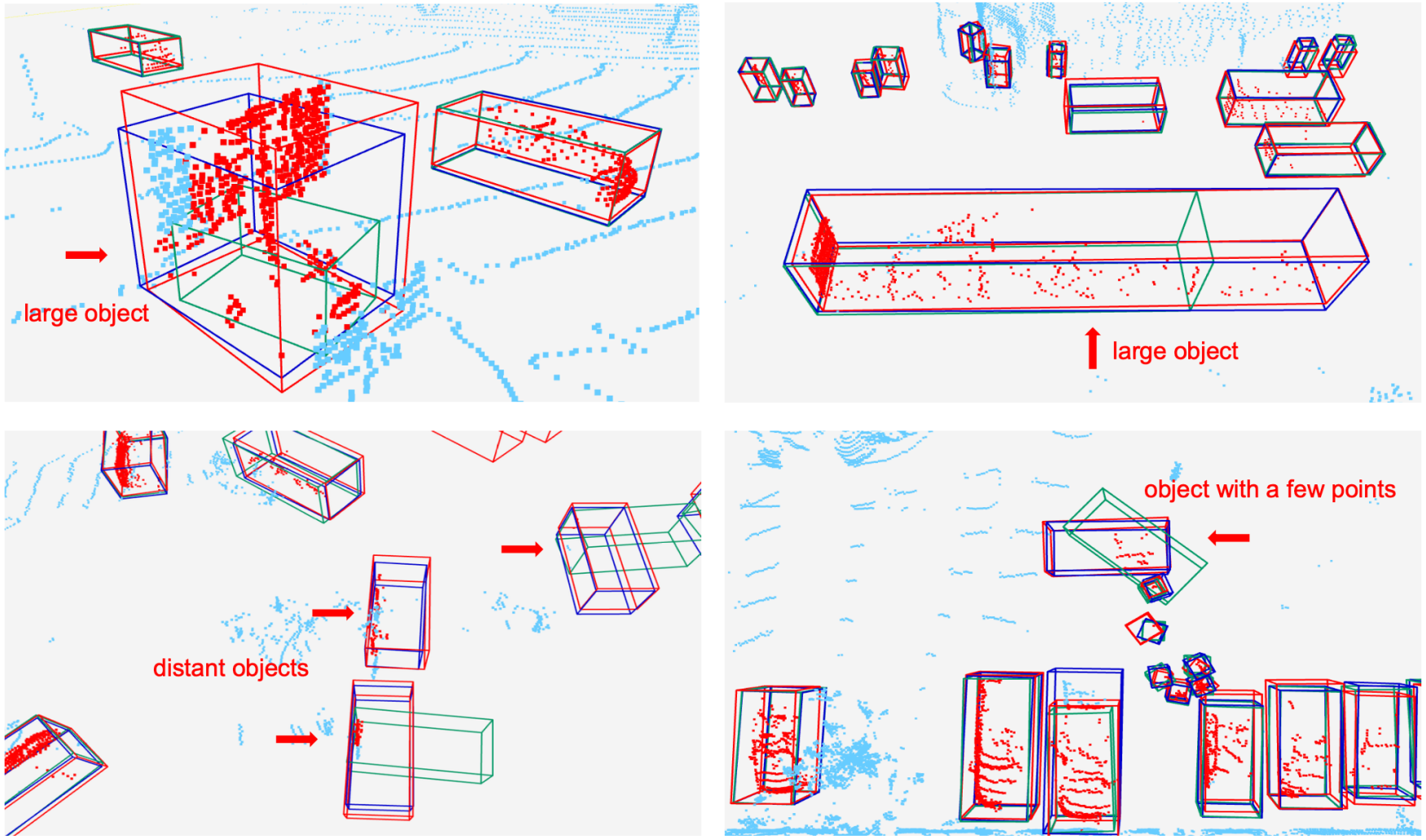
(d) Effectiveness of encoder-decoder design.

Improvements of HEDNet over HEDNet-single



More gains on large and distant objects

Visualization Examples from Waymo Open



Red boxes: human Blue boxes: HEDNet Green boxes: HEDNet-single

Thanks for your attention!

Gang Zhang[1] Junnan Chen[2] Guohuan Gao[3] Jianmin Li[1] Xiaolin Hu[1]*

zhanggangthu@gmail.com

https://github.com/zhanggang001/HEDNet