Appendix

In this appendix, we first introduce implementation details in Sec. A. We then include additional experimental results in Sec. B. We also provide more visualizations and discussions in Sec. C and Sec. D.

A. Implementation Details

nuScenes The nuScenes dataset [1] has 1,000 drive sequences, split into 700, 150, and 150 sequences for training, validation, and testing. nuScenes is collected by a 32-beam synced LIDAR and 6 cameras. The annotations include 10 classes. In the ablation study, detection models are trained on 1/4 training data and evaluated on the full validation set.

Waymo Waymo [11] is a large-scale public autonomous driving dataset, which contains 1,150 sequences in total, with 798 for training, and 202 for validation. It was collected by one long-range LiDAR sensor at 75 meters and four near-range sensors.

Argoverse2 Argoverse2 [13] has 1000 sequences, including 700 for training, 150 for validation. The perception range is 200 radius meters, covering area of 400m × 400m. We follow FSD [3] for data processing.

Voxelization For nuScenes [1] dataset, point clouds are clipped in [-54m, 54m] for X or Y axis, and [-5m, 3m] for Z axis. Voxel size is (0.075m, 0.075m, 0.2m) by default. For VoxelNeXt-2D, the voxel size along Z axis is 8m.

For Waymo [11] dataset, point clouds are clipped into [-75.2m, 75.2m] X or Y axis, and [-2m, 4m] for Z axis. Voxel size is (0.1m, 0.1m, 0.15m) by default. For VoxelNeXt-2D, the voxel size along Z axis is 6m.

Data Augmentations

For nuScenes dataset, random flipping, global scaling, global rotation, GT sampling [14], and translation augmentations are used. Flipping is randomly conducted along X and Y axes. Rotation angle is randomly picked between -45° and 45°. Global scaling is conducted by a factor sampled between 0.9 and 1.1. The translation noise factors are sampled between 0 and 0.5. Only for test submission models, GT sampling is removed in the last 5 training epochs [12].

For Waymo dataset, data augmentations also include random flipping, global scaling, global rotation, and ground-truth (GT) sampling [14]. These settings are similar to those of nuScenes dataset and follow baseline methods [9, 16].

For Argoverse2 dataset, we use similar data augmentation to nuScenes and Waymo, except that we do not use ground-truth sampling.

Training Hyper-parameters

For nuScenes dataset, models are trained for 20 epochs with batch size 16. They are optimized with Adam [7]. Learning rate is initially 1e-3 and decays to 1e-4 in a cosine annealing. Weight decay is 0.01. Gradients are clipped by norm 35. These settings follow CenterPoint [16].

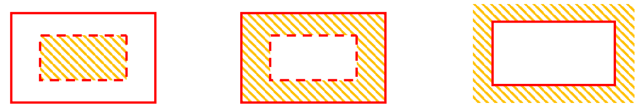

Figure A - 1. The relative positions of query voxel to the predicted boxes, *e.g.*, *near center*, *near boundary*, *outside box*, corresponding to Tab. 7 in the paper.

For Waymo dataset, models are trained for 12 epochs by default. Batch size is set as 16. Learning rate is initialized as 3e-3. They are also optimized with Adam [7].

For Argoverse2 dataset, we use similar settings to Waymo, except that only 6 epochs for training is enough.

Network Structures

We develop our VoxelNeXt network upon the widely-used residual sparse convolutional block [2, 9, 16]. We use 2D sparse convolutions in its variant of VoxelNeXt-2D. For voxel selection and box regression, we both use kernel-size-3 submanifold sparse convolutions [5] for prediction. The former convolution has 128 channels in VoxelNeXt-2D and 64 in 3D networks. Training schedules and hyper-parameters follow prior works [9, 16].

The backbone network of VoxelNeXt has 6 stages. The channels for these stages are {16, 32, 64, 128, 128, 128}. There are 2 residual submanifold sparse convolutional blocks [5] in each stage. The sparse head predicts outputs by 3×3 submainfold sparse convolutions. Following CenterPoint [16], the prediction layers are only shared among similar classes on nuScenes and shared among all classes on Waymo.

B. Experimental results

Performance on nuScenes Validation We provide the performance of VoxelNeXt on nuScenes *val* in Tab. A - 1.

Gaps between VoxelNeXt and VoxelNeXt-2D We analyze the gaps between VoxelNeXt and VoxelNeXt-2D on different amounts of training data in Tab. A - 3. These models are trained on 1/4, 1/2, and full nuScenes training set, respectively, and evaluated on the full validation set. It shows that The gap is large on the 1/4 training data, while the gaps gradually narrow as the data amount grows. Overall, the 3D network can obtain much better performance than its 2D counterpart at a small amount of data. Meanwhile, VoxelNeXt-2D has the potential on a large data amount.

Resolution of Sparse Head We make an ablation study on the resolution of prediction head in Tab. A - 2. The performance decreases as the head resolution increases from

Table A - 1. Comparison on the nuScenes validation split. This table presents detailed performance for Tab. 1 in the paper.

Method	Latency	mAP	NDS	Car	Truck	Bus	Trailer	C.V.	Ped	Mot	Byc	T.C.	Bar
SECOND [14]	64 ms	50.6	62.3	81.8	51.7	66.9	37.3	15.0	77.7	42.5	17.5	57.4	59.2
CenterPoint [16]	96 ms	58.6	66.2	85.0	58.2	69.5	35.7	15.5	85.3	58.8	40.9	70.0	67.1
VoxelNeXt	66 ms	60.0	67.1	85.6	58.4	71.6	38.6	17.9	85.4	59.7	43.4	70.8	68.1

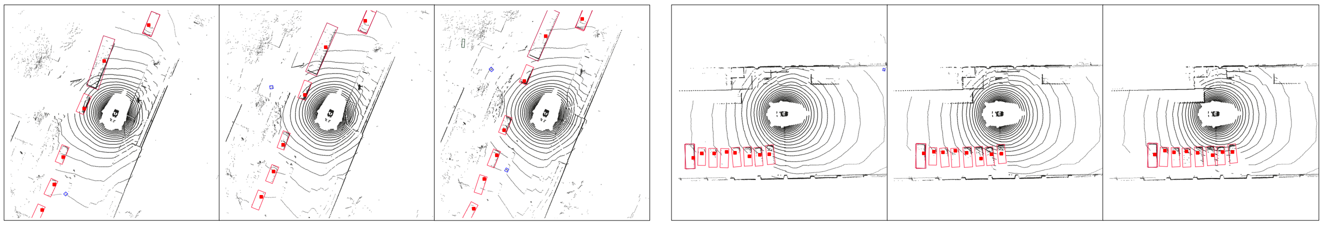

Figure A - 2. Detections of adjacent frames. We visualize predicted boxes and the corresponding query voxels, which are enlarged as red squares. This figure is best viewed by zoom-in.

Table A - 2. Effects of the feature levels for prediction. D^{3-5} and D^{1-5} contains multiple heads on various feature levels.

Method	Head resolution	mAP	NDS
D^3	8	**56.2**	**64.3**
D^4	16	52.5	60.7
D^5	32	49.0	57.9
D^{3-5}	{8, 16, 32}	55.7	63.7
D^{1-5}	{2, 4, 8, 16, 32}	53.9	62.2

Table A - 3. Gap between VoxelNeXt-2D and VoxelNet. mAP on nuScenes validation with different amounts of training data.

Method	1/4	1/2	full
VoxelNeXt-2D	53.4	56.0	58.7
VoxelNeXt	56.2	58.2	60.0

Table A - 4. Results on Vehicle detection on Waymo. * means decreasing the number of pasted instances in the ground-truth sampling augmentation and increase training epochs by 6 epochs [3].

Method	L1 AP/APH	L2 AP/APH
VoxelNeXt	78.2 / 77.7	69.9 / 69.4
VoxelNeXt*	79.1 / 79.0	70.8 / 70.5

the default setting of 8 to 32. In addition, we also evaluate the multi-head design of {8, 16, 32} and {2, 4, 8, 16, 32}, where results are combined from the multiple heads with various resolutions. These multi-head models present no better results than the single-resolution 8 network.

Performance on Waymo vehicle detection In Tab. A - 4, we follow FSD [3] to decrease the number of pasted instances in the ground-truth sampling augmentation and increase training epochs by 6 epochs. This trick leads to better results upon VoxelNeXt on the Waymo object detection.

C. Visualizations

We visualize the detections of adjacent frames in Fig. A - 2. The corresponding query voxels are depicted as red squares. We also provide a sequence of video frames, in both BEV and perspective views.

D. Discussions

Point-based Detectors Point-based 3D object detectors [8, 10, 15, 17] are fully sparse by their very nature. Point R-CNN [10] is a pioneer work and presents decent performance on KITTI [4]. Methods of SSD series [6, 15, 18, 19], including 3DSSD [15], inherit the point-based tradition and accelerate the methods with simplified pipelines. VoteNet [8] is based on center voting and studies indoor 3D object detection. However, point-based detectors are usually used in scenes with limited points. The neighborhood query operation is still unaffordable in large-scale benchmarks [1, 11], which are dominated by voxel-based detectors [9, 16].

Boarder Impacts VoxelNeXt replies on 3D data and its spatially sparse distribution. It might reflect biases in data collection, including the ones of negative societal impacts.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan,

Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020. 1, 2

[2] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: towards high performance voxel-based 3d object detection. In *AAAI*, pages 1201–1209, 2021. 1

[3] Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fully sparse 3d object detection. *CoRR*, abs/2207.10035, 2022. 1, 2

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 32(11):1231–1237, 2013. 2

[5] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 1

[6] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *CVPR*, pages 11870–11879, 2020. 2

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 1

[8] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9276–9285, 2019. 2

[9] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10526–10535, 2020. 1, 2

[10] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 2

[11] Pei Sun and et. al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2443–2451, 2020. 1, 2

[12] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pages 11794–11803, 2021. 1

[13] Benjamin Wilson and et. al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021. 1

[14] Yan Yan, Yuxing Mao, and Bo Li. SECOND: sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2

[15] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *CVPR*, pages 11037–11045, 2020. 2

[16] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1, 2

[17] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pages 18931–18940, 2022. 2

[18] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. CIA-SSD: confident iou-aware single-stage object detector from point cloud. In *AAAI*, pages 3555–3562, 2021. 2

[19] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. SE-SSD: self-ensembling single-stage object detector from point cloud. In *CVPR*, pages 14494–14503, 2021. 2