Method	Car	Bicycle	Motorcycle	Truck	Bus	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other ground	Building	Fence	Vegetation	Trunk	Terrian	Pole	Traffic sign	mIoU
MinkUNet34 [7]	96.8	55.0	81.4	83.2	70.2	79.5	89.8	7.8	94.8	54.6	82.8	1.5	92.0	68.3	87.9	69.4	72.8	66.1	52.4	68.3
HEDNet (Ours)	97.3	57.2	82.3	88.1	73.9	80.4	91.3	23.2	95.1	51.5	83.1	2.8	92.1	69.6	87.6	69.4	72.3	66.6	52.1	**70.3**

Table A1: 3D semantic segmentation results on the SemanticKiTTI validation set. Metrics: mIoU (%)↑ for the overall results, IoU (%)↑ for each category.

1 Implementation details on 3D object detection

We implemented our method with Pytorch using the open-source OpenPCDet [1].

Waymo Open dataset. We set the hyperparameter m to 2 for all SED and DED blocks and stacked 4 DED blocks for the 2D dense backbone by default. We adopted the detection head of CenterPoint [2] for HEDNet. As metioned in the main paper, we primarily followed the training and inference schemes of DSVT [3]. Specifically, the voxel size was set to (0.08m, 0.08m, 0.15m), and the detection range was set to [-75.2m, 75.2m] for X and Y axis, and [-2m, 4m] for Z axis. We trained HEDNet for 24 epochs on the entire training dataset and reported the evaluation results on the validation set to compare with previous state-of-the-art methods. For the ablation experiments, we trained all models for 30 epochs on a 20% training subset. All models were trained with a batch size of 16 on 8 RTX 3090 GPUs. We employed the Adam [4] optimizer with a one-cycle learning rate policy, and set the weight-decay to 0.05, and the max learning rate to 0.003. We also adopted the faded training strategy in the last epoch. During inference, we applied class-specific NMS with an IoU threshold of 0.75, 0.6 and 0.55 for vehicle, pedestrian, and cyclist, respectively.

nuScenes dataset. We set the hyperparameter m to 2 for all SED and DED blocks and stacked 5 DED blocks for the 2D dense backbone. We adopted the detection head of TransFusion-L [5] for HEDNet and primarily followed the training and inference schemes of TransFusion-L [5]. The voxel size was set to (0.075m, 0.075m, 0.2m), and the detection range was set to [-54m, 54m] for X and Y axis, and [-5m, 3m] for Z axis. We trained HEDNet for 20 epochs on the combined training and validation sets with a batch size of 16 on 8 RTX 3090 GPUs and reported the results on the test set to compare with other methods. We employed the Adam [4] optimizer with a one-cycle learning rate policy, and set the weight-decay to 0.1, the momentum to [0.85, 0.95], and the max learning rate to 0.001. The faded strategy was used during the last 5 epochs. For submission to the test server, we set the query number of detection head to 300 and did not use any test-time augmentation.

2 Experiments on 3D semantic segmentation

We conducted experiments on the popular LiDAR semantic segmentation dataset SemanticKiTTI [6], It provides 22 sequences with 19 semantic classes, captured by a 64-beam LiDAR sensor. Following the standard practice [7, 8], we report the Intersection-over-Union (IoU) for each category and the average score (mIoU) over all categories. For the backbone network, we employed a UNet-style structure, *i.e.,*the same designs as the first two layers of the MinkUNet34 are first adopted to extract sparse features with a spatial down-sampling ratio of 4, followed by 4 SED layers to transform the resulting features, finally, two symmetrical layers are used to recover high-resolution features following MinkUNet34. The other settings strictly followed [7]. Table A1 shows that the proposed model exhibited significant gains over its counterpart MinkUNet34 (*i.e.,*2.0% in mIoU), which demonstrated the generality of our method.

3 A step-wise ablation from VoxelNet to HEDNet

We conducted a step-wise ablation from the standard VoxelNet to our HEDNet on the Waymo Open dataset to show the effectiveness of different components (see Table A2). For the second model, we employed the training tricks used by DSVT, including IoU loss, class-specific NMS, faded strategy (disabling data augmentations in the last epoch), and a weight decay of 0.05. These training tricks can significantly boost detection accuracy. Actually, most of the training tricks have been used by

No.	VoxelNet	Tricks*	Smaller-voxel	SED-block	DED-block	Full-data	Latency	L1 mAPH	L2 mAPH
1	✓						40 ms	70.0	64.0
2	✓	✓					40 ms	75.4	69.1
3	✓	✓	✓				49 ms	76.6	70.2
4	✓	✓	✓	✓			55 ms	77.6	71.3
5	✓	✓	✓	✓	✓		67 ms	78.0	71.9
6	✓	✓	✓	✓	✓	✓	67 ms	79.5	73.4

Table A2: A step-wise ablation from VoxelNet to HEDNet. The first five models were trained on a 20% training subset and the last model was trained on the full training set. *: training tricks used in DSVT.

previous works, such as PV-RCNN++, and FSD. The codes and training configurations of the DSVT model can be found in the OpenPCDet archives. For the third model, we adopt a smaller input voxel size to keep more detailed information, which boosts the detection accuracy of pedestrian and cyclist. The 4th and 5th models sequentially incorporate our proposed SED blocks and DED blocks. The last model was trained on the full training set. Our final model (the 6th model) outperformed the previous SOTA method DSVT by 1.3% L2 mAPH while being 50% faster.

References

[1] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet`, 2020.

[2] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021.

[3] Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. Dsvt: Dynamic sparse voxel transformer with rotated sets. In *CVPR*, 2023.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[5] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022.

[6] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019.

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019.

[8] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *ICCV*, 2023.