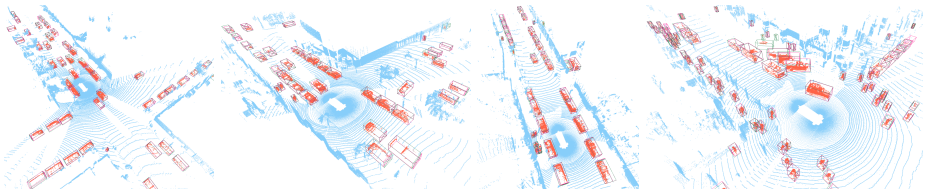

Fig. 1. Qualitative results of Graph-Ce on the WOD validation sequences. We show the raw point cloud in blue, ground truth in pink bounding boxes, our detected objects in green bounding boxes, and points inside our detected bounding boxes in orange.

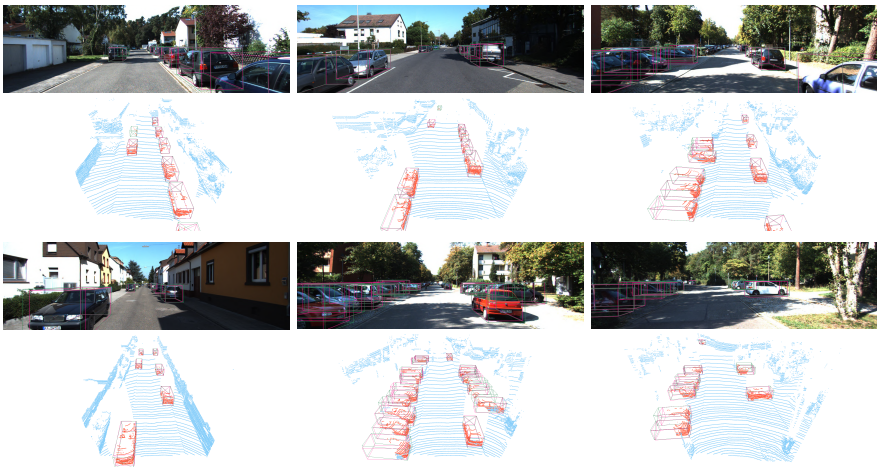

Fig. 2. Qualitative results of Graph-Po on the KITTI *val* set. We show detection results in six scenes, where the upper is the scene image, and the lower is the front view detection results.

Table 1. Ablation study of σ to enlarge each proposal.

σ	0.2	0.4	0.8	1.6
Vehicle	71.95	72.10	72.12	**72.16**
Pedestrian	68.95	**69.02**	68.97	68.95

Table 2. Ablation study of λ and δ in DFVS.

λ	0.14	0.16	0.18	0.18
δ	50	50	50	60
APH	70.78	70.81	**70.83**	70.81

This supplementary material contains the following sections. Sec. A illustrates additional visualization results. Sec. B presents more ablation studies. Sec. C provides potential improvements and more discussions.

A Qualitative Results

Fig. 1 demonstrates the qualitative results of our method on the Waymo dataset. Fig. 2 shows predicted 3D bounding boxes and projected 2D bounding boxes on the KITTI dataset. Our model can predict distant cars (2*nd* col of 1*st* row)

and highly occluded cars (1st col of 1st row), which are even not labeled. These demonstrate the high-quality prediction results of our model.

B More Ablation Studies

Analysis of the Dynamic Point Aggregation. We ablate the hyperparameter σ in Table 1 and find that enlarging σ to wrap contextual points is beneficial for detection. We also study the sensitivity of the hyperparameter λ and δ of DFVS in Table 2. We keep its efficiency similar (by randomly selecting a similar number of non-empty voxels) and study the detection accuracy. We find DFVS is robust to the hyperparameters.

C Potential Improvements and More Discussions

Multi-frame Fusion. Our model may benefit from multi-frame data as denser point clouds can help improve detection performance. As the number of point clouds in 3D proposals increases, DFVS can better balance accuracy and efficiency. Specifically, the random point sampling may become less accurate to sample nearby objects, as it will face a more unevenly distributed point cloud. At the same time, as the number of original point clouds grows, the sampling efficiency of the farthest point sampling will decrease.

Feature Reuse. Reusing features from region proposal networks (RPN), e.g., the voxel features and point features, may achieve more stability and better performance. For the sake of generality, we currently only take the raw point cloud as input, which makes it not tied to a specific RPN. We can also design our model for a specific RPN to reuse the features. For example, we can use DPA to aggregate voxels and their voxel features from sparse convolutions [2], or points and their point features from PointNet++ [1], and then fully utilize them to produce robust RoI features. Especially, DPA is a differentiable operation, which supports the back propagation of the aggregated features.

References

1. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems (2017)
2. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10) (2018)