CBCenterDet: 2^{nd} Place Solution to the 3D Object Detection of the SSLAD2022 Challenge

Weiping Xiao, Yiqiang Wu, Jiantao Gao, Xiaomao Li

The Research Institute of Unmanned Surface Vehicle (USV) Engineering,
Shanghai University

Abstract. In this report, we present our 2^{nd} place solution for the ECCV 2022 Workshop SSLAD Track 2 - 3D Object Detection. We focus on the class imbalance problem and classification-localization misalignment problem. To solve these problems, we implement adaptive object augmentation in data pre-processing and incorporate the IoU branch in the detection head of the CenterPoint 3D detection framework, respectively. In addition, several improvements are achieved for the feature extraction network and post-processing techniques are adopted to boost the model accuracy. Our final model achieves 79.7 mAP on the ONCE 3D object detection *test* set.

1 Introduction

The SSLAD 3D Object Detection Challenge at ECCV 2022 is a 3D object detection task for autonomous driving. A large-scale dataset named ONCE (**O**ne millio**N** s**C**en**E**s) [5] for 3D object detection in the autonomous driving scenario is provided. The ONCE dataset consists of one million LiDAR scenes and seven million corresponding camera images, which are collected across a range of different areas, periods, and weather conditions. The 5K, 3K, and 8K scenes in the ONCE dataset are annotated for training, validation, and testing respectively. The remaining data in the ONCE dataset is not annotated and can be used for semi- and self-supervised learning for 3D object detection.

2 Method

In this section, we present the details of our simple yet effective 3D object detector. As shown in Fig. 1, our detector is composed of four parts, namely data pre-processing, feature encoding network, detection head, and post-processing.

2.1 Data Pre-processing

Point Cloud Voxelization: The input point cloud is first converted into voxel presentation with predefined voxel size across the x, y and z-axes. The voxel feature is defined as the mean feature of the points belonging to the same voxel.

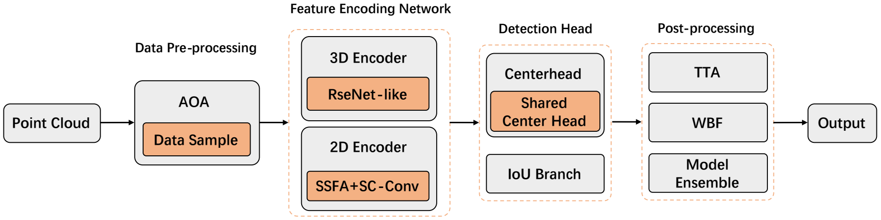

Fig. 1. Illustration of the overall architecture of our method, which consists of four parts: data pre-processing, feature encoding network, detection head, and post-processing.

After that, we obtain a 3D voxel feature representation as the input for the following 3D feature extractor.

Data Sampling and Object Augmentation: Based on the statistical analysis of data, we found that there is a class imbalance problem in the ONCE dataset. In particular, the number of vehicles significantly exceeds the number of pedestrians and cyclists. To alleviate this problem, we adjusted the sample number of pedestrians and cyclists in each scene. Besides, we adopt Adaptive Object Augmentation (AOA) [9] to augment pedestrians and cyclists. Specifically, AOA utilizes the vertical distribution characteristics (VDCs) of point clouds to search for suitable ground regions to paste the virtual instances for augmentation. Different from existing data augmentation methods [10], AOA can avoid collision conflicts occurring in new scenarios, and will not place objects into occluded areas.

2.2 Feature Encoding Network

3D Feature Encoder: 3D feature encoder is designed to hierarchically extract informative semantic feature representations from voxel inputs. Based on the classic 2D object detector backbone ResNet [2], we design a ResNet-like 3D feature encoder. As illustrated in Fig. 2, the 3D feature encoder consists of four stages and each stage is stacked with two residual blocks. Each convolutional layer in the residual block is replaced by the SubMConv layer [1], which is then followed by a batch normalization (BN) and rectified linear unit (ReLU).

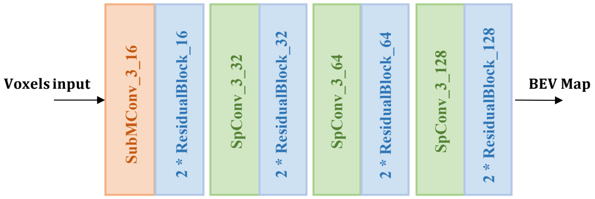

Fig. 2. Illustration of the architecture of our 3D feature extractor.

2D Feature Encoder: After extracting 3D feature maps from voxel inputs by the 3D feature encoder, we flatten them into the Bird's-Eye-View (BEV) as the input of the 2D feature encoder. As illustrated in Fig. 3, we adopt the Spatial-Semantic Feature Aggregation (SSFA) paradigm proposed by CIA-SSD [11] to adaptively fuse high-level abstract semantic features and low-level spatial features of the BEV map. Furthermore, SCBottleNeck [4] is also adopted in the 2D feature encoder for feature enhancement.

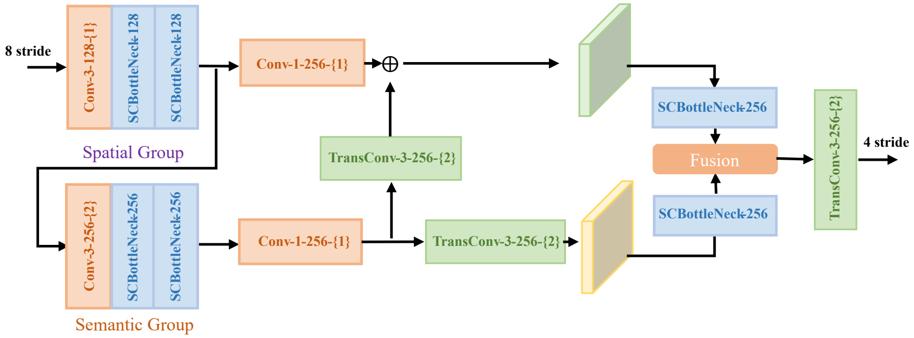

Fig. 3. Illustration of the architecture of our 2D feature extractor. *Conv* stands for convolutional layer and *TransConv* stands for transposed convolutional layer. The format of the layer setting follows *kernel size-channels-*{*strides*}, i.e. k-C-{s}.

2.3 Detection Head

The traditional detection head in 3D object detectors usually regards the classification scores as the final prediction output. However, the classification scores lack the localization information of the regressed bounding box. To alleviate the misalignment, we follow the AFDetV2 [3] to adopt an IoU prediction branch for the incorporation of IoU information into classification confidence scores as Equation 1. Furthermore, we shared the detection head for all classes to improve the performance.

$$f = score^{1-\alpha} \times iou^{\alpha} \tag{1}$$

where *score* is denoted as the classification scores while *iou* is the predicted IoU. The α represents the hyper-parameter from the interval [0,1] that controls the contributions from the classification scores and predicted IoU.

2.4 Post-processing

In the inference phase, we adopt the Test Time Augmentation (TTA), Weighted Boxes Fusion (WBF) [8], and Model Ensemble for the boost of the model accuracy.

Test Time Augmentation: The multiple test time augmentation is conducted on the inference stage, such as global rotation, scaling, and flip. We perform the global rotation for the point cloud in the range $[0,\ \pi/8,\ \pi,\ 7\pi/8,\ 3\pi/4]$, and global flip along both the x and y-axes. Also, we scale the point cloud in [0.95, 0.975, 1.025, 1.05].
Weighted Boxes Fusion: After using the TTA to generate the results of different augmentation, the Weighted Boxes Fusion (WBF) is adopted to merge them for better model accuracy. We set the box filtering IoU threshold to [0.7, 0.7, 0.7, 0.3, 0.5] and the score threshold to [0.3, 0.25, 0.25, 0.25, 0.35] for the different categories ['Car', 'Bus', 'Truck', 'Pedestrian', 'Cyclist'] in the WBF process.
Model Ensemble: Furthermore, we also apply Weighted Box Fusion (WBF) to merge the results of different models. Here we ensemble 5 models to get our final results, including different ways of sub-task designs and their two-stage versions. The box filtering threshold and the IoU threshold are set the same as those for TTA.

3 Experiments

3.1 Implementation Details

Training Details: We implement our method based on the official repository of the ONCE-Benchmark [5]. The scene ranges are limited to $[-75.2m,\ 75.2m]$ for the x and y-axes, while $[-5.0m,\ 3.0m]$ for the z-axis. The voxel size in the data pre-processing is set as $[0.1m,\ 0.1m,\ 0.2m]$ and the max number of the voxels in one scene is limited to 60000. The object augmentation range in AOA is set as $[0,\ 2\pi]$ and the copy-paste times of each augmentation instance is 16. In addition to AOA, we employ traditional data augmentation methods as done in [10, 7, 6], including randomly flipping along the x-axis, globally scaling with a scaling factor randomly chosen from the interval [0.95,1.05], and globally rotating around the z-axis with an angle randomly sampled from the interval $[-\pi/4, \pi/4]$. We use the Adam optimizer as our optimizer. The batch size is set to 4 for the total 100 epochs. The initial learning rate is set as 0.003 and decreases by 1% after every epoch.
Inference Details: The classification score threshold and the NMS IoU threshold are both set as 0.1 to filter the predicted bounding boxes. The hyper-parameter α in the detection head at test time is 0.55 for all categories and the max predict size from the detection head is set as 500. For our final submission, we train our model on both the training and validation splits.

3.2 Ablation Study

We conduct an ablation study of our models on the *val* set. As shown in Table 1, we ablate the improvement of our models based on CenterPoint. Data Sampling and Object Augmentation strategies bring a 1.27 mAP improvement. The Shared Center Head further leads to a significant improvement of 5.66 mAP.

DSOA	SCH	EFEN	IoUDH	Vehicle	Pedestrian	Cyclist	mAP
				65.70	**48.72**	**63.64**	**59.35**
✓				**66.93**	**49.76**	**65.17**	**60.62**
✓	✓			**78.02**	**52.52**	**68.32**	**66.28**
✓	✓	✓		**79.65**	**54.63**	**71.25**	**68.51**
✓	✓	✓	✓	**80.65**	**56.73**	**72.75**	**70.04**

Table 1. Ablation studies for 3D object detection on ONCE *val* set. We ablate each component of our submission compared to our baseline model CenterPoint. SCH, DSOA, EFEN, and IoUDH refer to the Shared Center Head, Data Sampling and Object Augmentation, Enhanced Feature Encoding Network, and IoU-based Detect Head, respectively.

The Enhanced Feature Encoding Network yields an improvement of 2.23 mAP. Besides, Adopting the IoU branch in Detect Head brings a 1.53 mAP improvement. For the post-processing techniques, we validate their effect on several baseline models. As shown in Table 2, the post-processing techniques TTA & WBF can significantly boost the model accuracy.

Method	Vehicle	Pedestrian	Cyclist	mAP
Ours	**87.59**	**72.17**	**78.38**	**79.38**
+ TTA & WBF	**89.20**	**77.42**	**79.75**	**82.12**
Ours (two-stage)	**89.65**	**70.79**	**77.43**	**79.29**
+ TTA & WBF	**91.78**	**75.54**	**81.30**	**82.87**

Table 2. Effect of the post-processing on our proposed models. TTA and WBF refer to the Test Time Augmentation and Weighted Boxes Fusion, respectively.

3.3 Main Results

Table 3 shows the ECCV 2022 Workshop SSLAD Track 2 - 3D Object Detection Challenge Leaderboard. Our submission ranked second among all entries.

4 Conclusion

In this report, we present a class-balanced 3D object detector and prove its effectiveness on the ONCE dataset. The proposed detector consists of data preprocessing, feature encoding network, detection head, and post-processing. To alleviate the class imbalance problem, we achieve improvements in each module of the proposed detection model. Based on the improved model, we win the second place in the 3D object detection of the SSLAD2022 Challenge at ECCV 2022.

Method	mAP	Vehicle	Pedestrian	Cyclist
galvatron	**85.13**	**88.38**	**84.48**	**82.52**
SHULab	**79.70**	**86.64**	**74.71**	**77.75**
lovesnowbest	**78.32**	**80.92**	**76.41**	**77.64**

Table 3. State-of-the-art comparisons for 3D object detection on the ONCE leaderboard of the SSLAD2022 Challenge at ECCV 2022. We show the mean average precision weighted by heading accuracy (mAP).

References

1. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
3. Hu, Y., Ding, Z., Ge, R., Shao, W., Huang, L., Li, K., Liu, Q.: Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)
4. Liu, J.J., Hou, Q., Cheng, M.M., Wang, C., Feng, J.: Improving convolutional networks with self-calibrated convolutions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
5. Mao, J., Niu, M., Jiang, C., Liang, H., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., et al.: One million scenes for autonomous driving: Once dataset. NeurIPS (2021)
6. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
7. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
8. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing (2021)
9. Xiao, W., Li, X., Liu, C., Gao, J., Luo, J., Peng, Y., Zhou, Y.: 3d-vdnet: Exploiting the vertical distribution characteristics of point clouds for 3d object detection and augmentation. Image and Vision Computing (2022)
10. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors (2018)
11. Zheng, W., Tang, W., Chen, S., Jiang, L., Fu, C.W.: Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In: Proceedings of the AAAI conference on artificial intelligence (2021)