3rd Place Solution to SSLAD Challenge - 3D Object Detection Track

Zehui Chen[1], Zhenyu Li[2], Shuo Wang[1], Jiahao Chang[1], Dengpan Fu[3], and Feng Zhao[1⋆]

[1] University of Science and Technology of China
{lovesnow,shuowang2323,changjh}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn
[2] Harbin Institute of Technology
zhenyuli17@hit.edu.cn
[3] NIO
dengpanfu@nio.com

Abstract. In this report, we present our solution to the 3D object detection of the SSLAD2022 Challenge at ECCV 2022 Workshop. We propose a simple semi-supervised 3D object detection framework, which conducts dense supervision on the student model. Instead of generating sparse pesudo-label for student imitation, we leverage the dense teacher predictions for more efficient semi-supervised learning. Our baseline is a simple CenterPoint with several simple modifications according to the last year winning solutions. The final submission is a single model with test-time augmentation and achieves 78.32 % mAP, ranking the 3rd place in the 3D object detection track.

1 Introduction

We first introduce self-supervised learning for next-generation industry-level autonomous driving challenge 3D Object Detection track, at ECCV 2022 Workshop. It introduce ONCE dataset [8], one of the currently largest autonomous driving dataset for 3D object detection, which consists of 1 million LiDAR scenes and 7 million camera images. The dataset provide three unlabeled splits with different amount of data, to facilitate the researches to conduct more experiments on semi- and self-supervised learning for 3D object detection.

2 Methods

In this section, we first describe the detailed implementation of our approach and then introduce some attempts that not work in our work.

⋆ Corresponding Author.

2.1 Baseline 3D Detector

We first reimplement CenterPoint based on MMDetection framework. In order to achieve better performance, we adopt dynamic voxelization, DCN separate head, and stronger data augmentation. Besides, we replace the RPN head with SCFA module and introduce IoU prediction branch for more accurate bounding box prediction.

Dynamic CenterPoint. Following the official benchmark, we choose CenterPoint [9] as our baseline model. We use the voxel-version of the CenterPoint. It consists of the point voxelization, 3D feature extraction, and center-based object prediction network. Considering that point feature extraction is of great importance in the object localization phase, we replace the original hard voxelization into dynamic voxelization proposed in [12]. This strategy not only improves the utility of the GPU memory but also enhance the detection performance. Deformable convolution [3] is an effective approach in enlarging dynamic receptive field of the network and greatly improves the performance. Therefore, following the practice in [2], we adopt deformable convolution for separate classification and regression branches in our model.

Data Augmentation. Data augmentation is a crucial step in achieving competitive detection performance. Therefore, we carefully tune the augmentation hyper-parameters. Firstly, we apply horizontal and vertical flip on the point clouds with a rate of 0.5. Secondly, we try more challenging augmentations like global rotation with a range of $[-\pi/4 - \pi/4]$ and randomly add offsets to the whole point clouds ranges from $[-0.2 - 0.2]$ m. We also rescale the point clouds with a ratio of $[0.95 - 1.05]$. We generate an annotation database containing labels and associated point clouds. During training, we randomly select 1, 4, 3, 2, and 2 ground truth samples for car, bus, truck, pedestrian, and cyclist and place them in the current frame, which is denoted as GT-AUG.

Backbone & Neck. We follow the implementation in the last year winning solution, where we adopt the Spatial Semantic Feature Aggregation (SSFA) module introduced in CIA-SSD [11].

IoU Head. Classification score are proven to be inconsistent with the localization accuracy [5]. In this case, obtaining a suitable score for the detected bounding boxes is important. Following the practice in [4], we introduce an additional IoU regression branch, in parallel with the classification and regression heads. The target is set to the IoU value between the model prediction and the matched GT. And we use two MLP layers and a sigmoid function to construct the IoU prediction branch.

Test-Time Augmentation. In most cases, test-time augmentation (TTA) provides the largest performance gain without taking any extra efforts. In this work, we simply adopt vertical and horizontal flip for TTA. Besides, we find that simply averaging prediction maps before re-flip model predictions yields better performance compared to vanilla TTA.

2.2 Semi-Supervised 3D Detection

In this section, we present our framework on semi-supervised 3D object detection. Following the common practice, our framework is consists of two branches: a teacher model with weak data augmentation, and a student model with strong augmentation. Then, we conduct consistency loss on their predictions, following [6]. Specifically, we perform such consistency regularization on the predictions from the teacher model and student model.

Pseodo Label Generation We first instantiate a fully-supervised model as teacher, which denotes as T, and then perform weak augmentation, such as random flip horizontally and vertically. After that, we feed the teacher model with the augmented point clouds P to obtain the model predictions. Due to the characristic of the one-stage object detectors, we can get a dense $H \times W$ size predictions, where each pixel corresponds to one possible instance.

Consistency Loss In order to avoid tuning delicate NMS threshold and boxes score threshold, we directly let the student mimic the dense predictions by the teacher model. Formally, let $P_T = (C_T, B_T)$ denotes the predictions from the teacher model, C_T and B_T are the predictions in the classification and regression branches, respectively. The same annotations are applied for the student as P_S. Therefore, we can get the consistency loss with

$$L = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \mathrm{CE}(c_T^{ij}, c_S^{ij}) + ||b_T^{ij} - b_S^{ij}||_1. \tag{1}$$

Considering that not all predictions are of the same quality, we introduce an attention mask with the guidance from the ground truths' position. Hence the final regularization loss is represented as:

$$L = \frac{1}{\sum_i^H \sum_j^W w_{ij}} \sum_{i}^{H} \sum_{j}^{W} w_{ij}(\mathrm{CE}(c_T^{ij}, c_S^{ij}) + ||b_T^{ij} - b_S^{ij}||_1). \tag{2}$$

2.3 Attempts that not Works

Since ONCE dataset also provides paired imagery data, we give an attempt on multi-modal 3D object detection with paired 2D and 3D data. We first followed the practice in MoCa [10] and reimplement AutoAlignV2 [1] for 2D & 3D feature fusion, where the image backbone is initialized from the pretrained checkpoint followed in [7]. However, the results are far from the satisfactory, actually lower than its vanilla 3D version.

3 Experiments

3.1 Implementation Details

We implement our methods based on the official repository MMDetection3D. The point cloud range is limited to [(-75.2, 75.2), (75.2, 75.2), (-5.0, 3.0)] respect to x, y, z-axis during training and testing process. The voxel size along x, y, z-axis is set to [0.2m, 0.2m, 0.2m]. AdamW with one-cycle policy is used as optimizer. We set the max learning rate to 3×10^{-3}, division factor to 10, momentum ranges from 0.95 to 0.85, fixed weight decay to 0.01, pct stat to 0.4, grad norm clip to 35. Our model is trained with 80 epochs on the training set.

3.2 Detailed Ablations

In this section, we provide detailed ablations on the effect of each component in our model. Our baseline model starts from 63.4 mAP, which follows the official practice in ONCE_BENCHMARK. When we apply the dynamic voxelization, the performance raised 0.4 mAP. Then, we carefully tune the augmentation parameters and the baseline improves to 67.4 mAP. Then, we follow the last year winning solution, replacing the backbone and the neck to SSFA module, and the performance improves 1.5 mAP. To get better score representation, we introduce IoU predictions and get another 1.3 mAP enhancement. TTA yields the largest improvement, which raises from 70.3 to 76.6 mAP. Finally, we employ the proposed SSOD framework and reach 78.7 mAP on the validation subset.

Table 1. Effect of each component in our framework. Results are reported on ONCE validation set with CenterPoint.

Dynamic Voxelization	Train Aug	SSFA	IoU Pred	TTA	SSOD	mAP
						63.42
✓						63.79
✓	✓					67.40
✓	✓	✓				68.92
✓	✓	✓	✓			70.26
✓	✓	✓	✓	✓		76.56
✓	✓	✓	✓	✓	✓	78.76

4 Conclusion

In this report, we present a competitve 3D detector and win 3rd place on the 3D object detection of the SSLAD2022 Challenge at ECCV 2022. We hope our work could inspire more researches on the large-scale ONCE Dataset in the future.

References

1. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F.: Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. arXiv preprint arXiv:2207.10316 (2022)
2. Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z.J., Wu, F.: Disentangle your dense object detector. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4939–4948 (2021)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
4. Huang, Z., Chen, Z., Li, Q., Zhang, H., Wang, N.: 1st place solutions of waymo open dataset challenge 2020–2d object detection track. arXiv preprint arXiv:2008.01365 (2020)
5. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 784–799 (2018)
6. Li, Z., Chen, Z., Li, A., Fang, L., Jiang, Q., Liu, X., Jiang, J.: Unsupervised domain adaptation for monocular 3d object detection via self-training. arXiv preprint arXiv:2204.11590 (2022)
7. Li, Z., Chen, Z., Li, A., Fang, L., Jiang, Q., Liu, X., Jiang, J., Zhou, B., Zhao, H.: Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1500–1508 (2022)
8. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al.: One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037 (2021)
9. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
10. Zhang, W., Wang, Z., Loy, C.C.: Exploring data augmentation for multi-modality 3d object detection. arXiv preprint arXiv:2012.12741 (2020)
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. vol. 35, pp. 3555–3562 (2021)
12. Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., Vasudevan, V.: End-to-end multi-view fusion for 3d object detection in lidar point clouds. In: Conference on Robot Learning. pp. 923–932. PMLR (2020)