Data-efficient 3D Object Detection for Autonomous Driving

Xiang Li[1], Junbo Yin[1], Yan Wang[1], Wei Li[2], and Jianbing Shen[3]

[1]Beijing Institute of Technology [2]Inceptio Tech [3]University of Macau

1 Introduction

3D object detection with LiDAR data is important for autonomous driving system because point cloud provide more accurate and more reliable information than RGB images. Fully supervised 3D object detection has been studied for a long time and achieved great progress. However, 3D annotation in the point cloud is extremely tedious, expensive and time-consuming for a new introduced self-driving system. Hence, it's meaningful to explore data-efficient methods that can learn from a limited amount of high-resolution LiDAR data.

Different from traditional LiDAR datasets such as Waymo [5], KITTI [3] and nuScenes [2], the recently proposed InnovizTwo LiDAR [1] provides 3D measurements with a better range and resolution. For example, the front-view range of InnvizTwo LiDAR data is about 280m which is about 4 times farther than Waymo [5] LiDAR dataset. The total points number of each LiDAR frame is also much larger that assures a better resolution, (*e.g.*, contains 400 to 600 thousands of points). Only a limited amount of labeled data (*e.g.*, 100 frames) is available to mine the detection performance. Our method achieves 2nd place in this InnovizTwo Challenge. In this report, we provide necessary technical details of our solution.

2 Method

Bascially, our method benefits from the *pre-training and fine-tuning* paradigm. Our method first pretrain the Centepoint [8] model on Waymo [5] dataset and then fine-tune the model on the labeled data of Innoviz dataset. Then a semi-supervised method called Noisy Student [7] is adopted to train a more powerful model with both labeled and unlabeled data. Furthermore, we proposed a box-wise contrastive learning strategy to explore representation learning based on expressive 3D boxes.

2.1 Pretraining and Fine-tuning

The most challenging part is that the label is very limited in the Innoviz dataset. To alleviate this problem, we first pre-train our base model Centerpoint [8] on Waymo [5] dataset in a supervised manner and then fine-tune the network layers

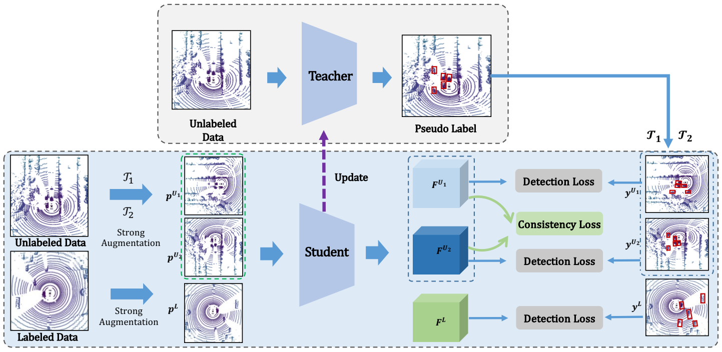

Fig. 1. The overall architecture of our method.

with limited labeled of the Innoviz dataset such that the model can also learn useful information from the large-scale Waymo dataset. The learned model also acts as the teacher model in next Noisy Student module.

2.2 Noisy Student for Semi-supervised Learning

To fully take advantage of the unlabeled data in the Innoviz dataset, we also apply a semi-supervised learning(SSL) approach, Noisy Student [7], on the 3D object detection task. As shown in Fig. 1, this algorithm contains four steps: 1) Learn teacher model which minimizes the detection loss on labeled data. 2) Use an unnoised teahcer model to generate pseudo labels for unlabeled data. 3) Learn an equal-or-larger student model which minimizes the detection loss on labeled images and unlabeled images with noise added to the student model. 4) Iterative training: Use the student as a teacher and go back to step 2. The noise add to the student model can make it more capable and robust than the teacher. In our case, we mainly consider the noise as normal data augmentation in point cloud, *e.g.*, random filp, rotation and scale.

2.3 Box-wise Constrastive Learning

During the semi-supervised learning, we learn from both the labeled and unlabeled data. For the labeled data, we directly optimize the detection loss. For the unlabeled data, we propose to learn from both the pseudo-label based detection loss and the box-wise contrastive learning. Different from typical contrastive learning used for self-supervise pre-training, our box-wise contrastive learning learn from explicit pseudo-labeled boxes, which can enforces the feature consistency of the same box instance from different augmented views. Formally, given an unlabeled point cloud P^U, we first apply two random augmentations to generate different views P^{U_1} and P^{U_2}. Then we get the pseudo labels of the two views Y^{U_1} and Y^{U_2}. After that, we transform Y^{U_1} and Y^{U_2} to the same view and build positive and negative sample pairs by a greedy matching. Afterwards, we extract box features by a point-wise interpolation, where the corner points of a box in the bird's eye view (BEV) are used to get features from the detector BEV feature maps $\mathbf{F}^U$,

$$\mathbf{B}^{U_1} = I(Y^{U_1}, \mathbf{F}^U), \mathbf{B}^{U_2} = I(Y^{U_2}, \mathbf{F}^U) \tag{1}$$

where $I(\cdot)$ is the bilinear interpolation function, $\mathbf{B}^{U_1}$ and $\mathbf{B}^{U_2}$ are box features from different views. Later, a projection head that contains two 1×1 linear layers is used to map the box features to an embedding space. Then the InfoNCE [4] loss is exploited to compute the loss on these box-wise examples, such that $L^U_{\text{Info}} = \text{InfoNCE}(\mathbf{B}^{U_1}, \mathbf{B}^{U_2})$. The purpose of InfoNCE loss is to minimize the feature distance of the same box instance, meanwhile maximize the feature distance between different box instances. Finally the overall loss of the student model is defined over the detection loss of labeled and unlabeled data L^L_{Det} and L^U_{Det}, as well as the InfoNCE loss L^U_{Info},

$$L = L^L_{\text{Det}} + L^U_{\text{Det}} + \alpha L^U_{\text{Info}} \tag{2}$$

3 Experiment

3.1 InnovizTwo dataset

A dataset with 1219 LiDAR frames from a large variety of scenarios is provided, with only 103 annotated frames . The target classes of interest are cars, trucks (including trailers and buses), two-wheelers (motorcycles/bicycles) and pedestrians. The performance is measured using average IoU_{XY}.

3.2 Implementation Details

We chose Centerpoint [8] as our baseline. To ensure that the distribution of the unlabeled data match that of the training set, we also need to balance the number of unlabeled data for each class [7]. The point cloud range alone the x-axis, y-axis and z-axis set to $[0, 281.6]$, $[-80, 80]$, $[-2, 2]$. We train the model for 60 epoch with learning rate 0.005. Other parameters are kept the same with original config [6]. We random sample 20% of the data in Waymo [5] for pretraining the model.

Table 1. The main results on *InnovizTwo* `evaliuation` dataset

Model	Pretrain	Noisy Student	BCE	$\text{IOU}_{\text{XY}}(\%)$
CenterPoint	-	-	-	49
Ours	✓	-	-	64
	✓	✓	-	65
	✓	✓	✓	66

3.3 Ablation Studies

As shown in Table 1, pretraining on Waymo brings great improvement. The Noisy Student further utilizes the unlabeled data to strengthen the student model. Finally, BCE strategy enforces consistency regularization to learn the target in a soft way.

4 Conclusion

Our experiments shows that our method is data-efficient and has good results on the on *InnovizTwo* `evaliuation` dataset. Finally we got the 2nd place in the InnovizTwo Challenge. We expect this work can encourage more works towards data-efficient 3d object detection.

References

1. Eccv-challenge. https://innoviz.tech/eccv-challenge
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
4. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
5. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
6. Team, O., et al.: Openpcdet: An open-source toolbox for 3d object detection from point clouds (2020)
7. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
8. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: CVPR (2021)