Semi-CenterPoint submission to the ICCV 2021 Workshop SSLAD Track 2 - 3D Object Detection

Zehan Zhang[*], Yang Ji[*], Wei Cui, Yulong Wang, Xian Zhao
Hikvision Research Institute
zhangzehan, jiyang5, cuiwei6, wangyulong13, zhaoxian@hikvision.com

Abstract

In this report, we present our "Semi-CenterPoint" solution for the ICCV 2021 Workshop SSLAD Track 2 - 3D Object Detection. Our submission builds upon the CenterPoint 3D detection framework. CenterPoint is a state of the art method and its validity has been verified on several datasets. Based on the ONCE dataset, we have made some targeted improvements. First, we improve the 2D region proposal network to obtain a supervised baseline with good performance. Then, we design a semi-supervised learning framework with a combination of mean-teacher and pseudo-labeling. Our final model achieves 82.89 mAPH on the ONCE 3D detection test set.

1. Introduction

The SSLAD 3D Object Detection Challenge at ICCV 2021 is a very exciting competition in the field of automatic driving. In this challenge, it provide a large-scale dataset, called ONCE [8], with 1 million point clouds and 7 million images. ONCE annotates 5K, 3K and 8K scenes for training, validation and testing set respectively and leaves the other scenes unlabeled. ONCE provides 3D bounding boxes for car, cyclist, pedestrian, truck and bus. ONCE also has diverse environments such as day/night, sunny/rainy, and urban/suburban areas.

The input to the 3D detection [6, 10, 4] is an orderless pointcloud whose purpose is to predict a set of 3D object bounding boxes in 3D space. Each bounding box u, v, d, w, l, h consists of a center location u, v, d, relative to the objects ground plane, and 3D size w, l, h, and rotation expressed by yaw α.

2. Methods

In this section, we present the details of our 3D detector in the challenge. Since the dataset contains a large amount of unlabeled data, our method uses a semi-supervised 3D detection scheme. Before using semi supervised scheme, a supervised network with excellent performance plays a vital role. Therefore, this section first introduces the specific methods of supervision, and then introduces our semi supervision scheme. The supervised method is improved based on CenterPoint [11], and the semi-supervised scheme uses a combination of the pseudo-label method and the Mean-Teacher [1] method.

2.1. Supervised 3D Object Detection

Based on CenterPoint, we have made the following improvements: improved Region Proposal Network (RPN) network, reduced voxel grid, Test Time Augmentation (TTA) and multi-class NMS.

CenterPoint. Except for the RPN network, the network we use is basically the same as that of CenterPoint. We use CenterPoint-Voxel as the backbone network. The voxelization method, 3D feature extraction network, center heatmap head, regression head, one-stage loss function, two-stage network, and two-stage loss function are all consistent with the original paper. It is worth noting that since the first stage and the second stage are trained separately, and the weight of the first stage is fixed during the second stage training, we only use the one-stage network for training in the subsequent semi-supervised framework. Finally, we use the two stage network for the final fine-tuning.

Input and Data Augmentation. Although the ONCE dataset gives a series of sequence data and relative pose information, we still only use one frame of point cloud as the input. The reason is that the time interval is too large (500 ms) and the speed information of the moving object is lacking, when multiple frames are superimposed, the moving objects in the preceding and following frames basically do not overlap. Figure 2 shows the non-overlapping situation of moving objects when the three frames are superimposed.

We use the following data augmentation strategy [5, 14]. 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,

[*] These authors contributed equally to this work.

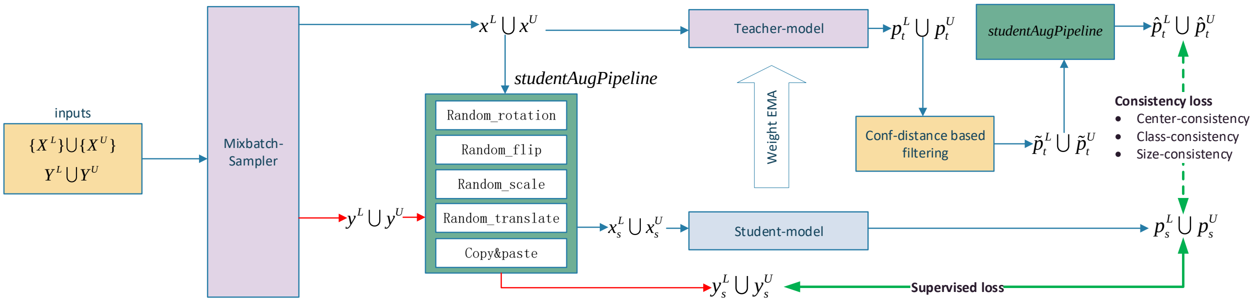


Figure 1: Semi supervised learning scheme for joint optimization of mean teacher and pseudo label.

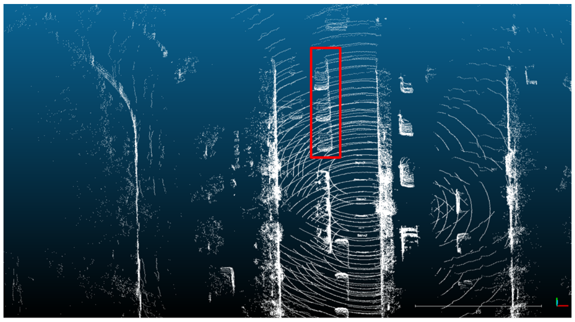


Figure 2: Three-frame superimposed point cloud visualization. The three vehicles in the red box are the same object at different times.

pedestrian and cyclist respectively, and place them in the current frame. We use random flipping along both X and Y axis, and global scaling with a random factor from [0.95, 1.05] and a random global rotation between$[-\pi/4, \pi/4]$.

RPN. We replace the RPN network with the Spatial-Semantic Feature Aggregation (SSFA) [13] structure to adaptively fuse highlevel abstract semantic features and low-level spatial features for more accurate predictions of bounding boxes and classification confidence. And, we replace the basic convolutional blocks with self-calibrated convolutions (SC-Conv) [3, 9] as shown in Fig. 2a and 2c, which helps to enlarge the receptive field for spatial locations and introduces channel-wise and spatial-wise attention with costefficiency. The improved RPN structure boosts the detection accuracy with a similar number of parameters.

Higher Spatial Resolution. In order to further improve the detection performance, we reduced the grid size from $0.1m$, $0.1m$ to $0.05m$, $0.05m$ along X, Y axis respectively to convert the raw point cloud into voxel presentation. At this time, the batchsize is 2, and the grid size cannot be further reduced.

TTA and Multi-class NMS. Empirically, we should perform several different test-time augmentations, including point cloud rotation around pitch, roll and yaw axis, point cloud global scaling and point cloud translation along z-axis, which is similar to the data augmentation in the training process. However, we find that only flipping along the Y axis can bring the greatest gain, so we only use flipping along the Y axis in the test time. This challenge uses 0.7, 0.3 and 0.5 IoU thresholds for vehicle, pedestrian and cyclist classes respectively for evaluation. Therefore, when we are doing NMS, the IoU thresholds for vehicle, pedestrian and cyclist classes are also set to 0.7, 0.3 and 0.5 respectively.

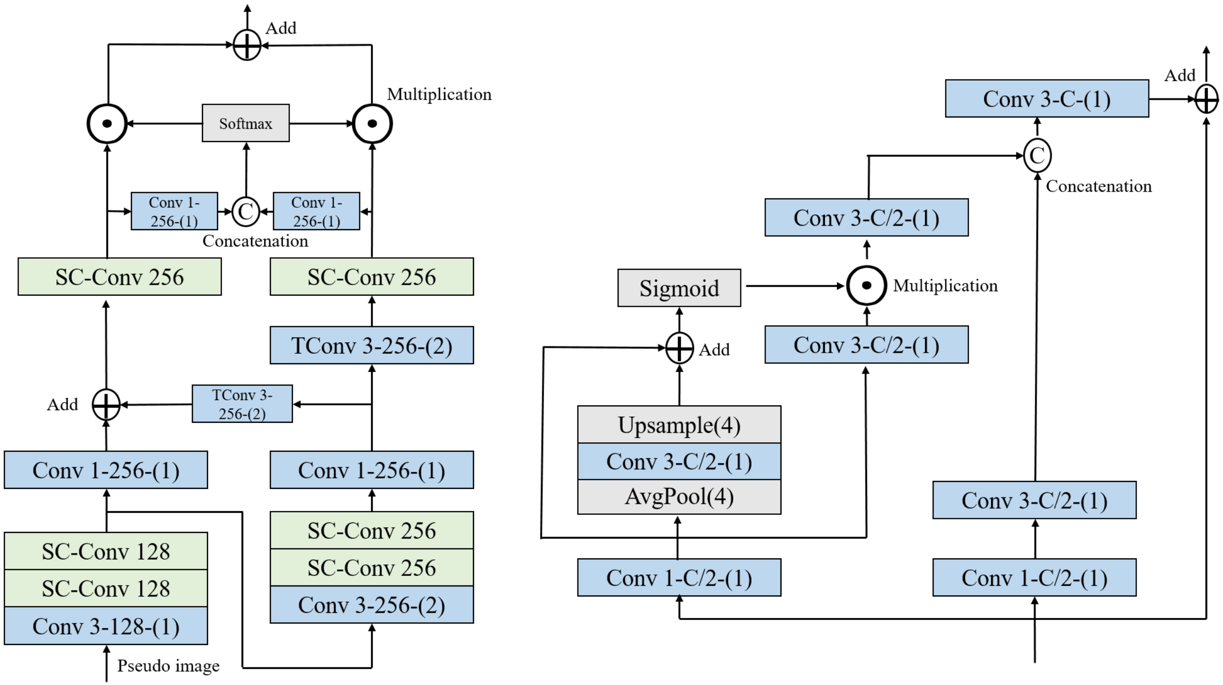
2.2. Semi-Supervised 3D Object Detection

In the semi-supervised framework, we use a combination of pseudo-label and Mean-Teacher methods: first use offline pseudo-label data for training, and then use Mean-Teancher for secondary training.

2.2.1 Pseudo Label

By analyzing the prediction results of the validation set, we find that there are a lot of false detections in the supervised model, as shown in Figure 4. In the pseudo-label scheme, a large number of false detections will introduce a lot of noise, which will degrade the performance of the model.

In order to overcome the above problem, we analyze the false detection and the positive detection for different classes at different distances ([0m, 30m], [30m, 50m], [50m, inf]), Figure 5 shows the 0-30m false detection and the positive detection. Then, we select the score thresholds for different categories and different distances in accordance with the principle of maintaining positive detections and reducing false detections. The final selection thresholds are shown in Table 1. According to these thresholds, the offline pseudo-label data is generated for network training, and the training process is the same as that of the supervised network.



(a) SSFA SC-Conv RPN

(b) SC-Conv module

Figure 3: (a) and (b) denote the SSFA SC-Conv RPN and SC-Conv module. ‘Conv’ stands for convolutional layer and ‘TConv’ stands for transposed convolutional layer. The format of the layer setting follows ‘kernel size-channels-(strides)’, i.e. k-C-(s).

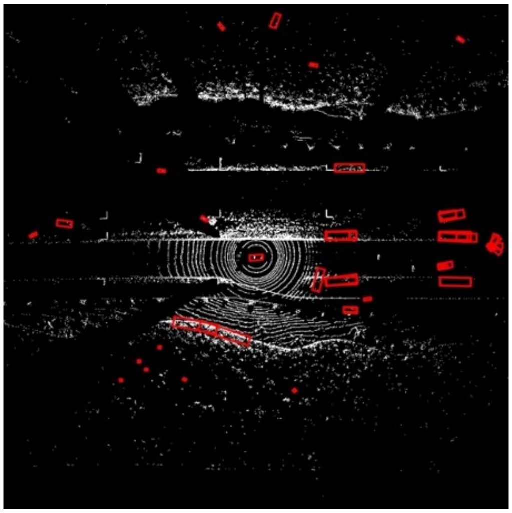


Figure 4: Visual analysis of false detection in the validation set. In the BEV, the red boxes are all false detections.

Distance/Class	Car	Bus	Truck	Pedestrian	Cyclist
0-30m	0.6	0.6	0.6	0.4	0.4
30-50m	0.4	0.7	0.7	0.3	0.3
50m-inf	0.35	0.6	0.6	0.3	0.25

Table 1: Score thresholds for different categories and distances

2.2.2 Mean-Teacher

Data selection. Offline pseudo-label training is used to obtain the initial model for MT training. Figure 1 shows the overall learning framework of MT [15, 1]. As shown in the flow chart, first, the training data is divided into labeled point cloud data X^L and unlabeled laser point cloud data X^U. Among them, there are labeled data with artificially labeled truth labels Y^L, and unlabeled data has no corresponding manual labels. We use the offline pseudo-label optimization scheme to generate pseudo-label results for unlabeled point cloud data Y^U.

Among the labeled data, we only use the training set data in the experimental stage. As for the choice of unlabeled data, considering that the training time using one million unlabeled data is too long, we extract the data every other frame to ensure that it can cover all kinds of scenes contained in 1 million data. Through the analysis of the weather distribution of the test set, it is found that the distribution ratio of rainy day data to sunny day data is basically 1:1, while among the 1 million data, there are only about 50000 rainy

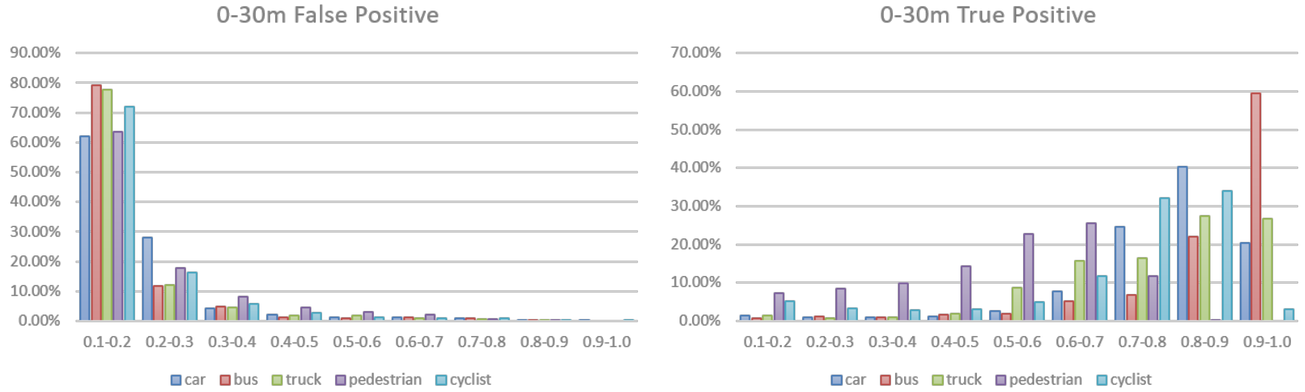


Figure 5: Distribution diagram of false detection and positive detection at 0-30m.

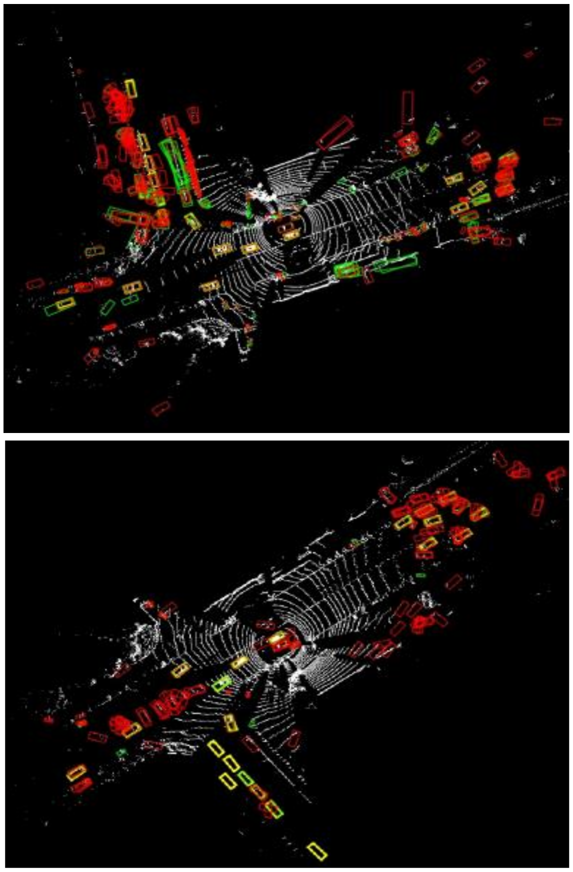


Figure 6: Comparison of the teacher-model before (first line) and after (second line) filtering the prediction results in the validation set. Yellow boxes are the ground truth, green boxes are the results predicted by teacher-model, red boxes are the results predicted by student-model.

day data. In order to maximize the proportion of rainy day data in the training set, all 50000 rainy day data are retained, and the remaining data are extracted every two frames to obtain 470000 data. Finally, a total of 520000 unlabeled data were obtained.

Then the two parts of data are mixed sampled. When sampling, we do not use the method of random sampling after mixing the two parts of data, but sample the two parts of data separately to control the sampling proportion of L set (labeled data) and U set (unlabeled data) in each sampled Mini batch. After experimental analysis, we finally use the sampling ratio of L:U = 1:1.

Semi supervised learning scheme. As shown in Figure 1, after the sampling is completed through the sampling method described above, first, x^L and x^U without data enhancement will be sent to the teacher model for prediction to obtain the prediction result p_t^L and p_t^U. The same sampling data will be enhanced through a series of random data operations, including random global rotation between[$-\pi/4$, $\pi/4$], random flipping along both X and Y axis, random scaling (scaling coefficient between 0.95 and 1.05), random translation (translation distance standard deviation is 1.0), and copy & paste operation (randomly selecting 1, 4, 3, 2, and 2 ground truth samples for car, bus, truck, pedestrian and cyclist respectively, and place them in the current frame). With the above enhancement operations, we can get the enhanced point cloud data x_s^L and x_u^L with the corresponding enhanced labels y_s^L and y_u^L. The enhanced point cloud data x_s^L and x_u^L are sent to the student model for forward reasoning to obtain the prediction results p_s^L and p_s^U. The predicted results of the student model, p_s^L and p_s^U, are supervised using the transformed truth labels y_s^L and pseudo labels y_u^L, respectively, to generate supervised losse L_{sup}.

For unsupervised loss, it is necessary to constrain the consistency of the teacher model prediction results and the student model prediction results. Since the prediction results p_t^L and p_t^U of the teacher model have many false de-

voxel 0.1m	SSFA	SC-Conv	voxel 0.05m	Yflip	Multiclass NMS	Pseudo Label	Mean Teacher	Two-Stage	mAPH↑
✓									65.15
✓	✓								66.94
✓	✓	✓							68.97
✓	✓	✓	✓						70.46
✓	✓	✓	✓	✓					76.76
✓	✓	✓	✓	✓	✓				77.70
✓	✓	✓	✓	✓	✓	✓			80.58
✓	✓	✓	✓	✓	✓	✓	✓		81.60
✓	✓	✓	✓	✓	✓	✓	✓	✓	**82.18**

Table 2: Ablation studies for 3D detection on ONCE validation set. We ablate each component of our submission compared to a single-frame single-stage CenterPoint baseline.

tections, to improve the learning performance, we perform confidence filtering on the prediction results of the teacher model, and obtain the filtered teacher model prediction results $\tilde{p}_t^L$ and $\tilde{p}_t^U$. By using the confidence threshold setting in Table 1, we filter many low confidence detection results to avoid introducing a lot of misleading information to the student model when using consistency constraints. Figure 6 shows the comparison of the teacher model prediction results on the bird's-eye view before and after filtering operation.

After completing the confidence filtering of the teacher model, the optimized detection results $\tilde{p}_t^L$ and $\tilde{p}_t^U$ are mapped to $\hat{p}_t^L$ and $\hat{p}_t^U$ by using the same data augmentation strategy used in the student model (under the same reference system as the prediction results of the student model). The consistency constrained loss is generated by $\hat{p}_t^L \cup \hat{p}_t^U$ and $p_s^L \cup p_s^U$. The consistency constraint loss contains three parts: the center position of the object, the object size (length, width, and height), and the confidence of the object category:

$$L_{ctr} = \frac{1}{n}\sum_{i=1}^{n} \|ctr(\hat{p}_t^L) - ctr(\hat{p}_s^L)\| + \|ctr(\hat{p}_t^U) - ctr(\hat{p}_s^U)\| \qquad (1)$$

$$L_{size} = \frac{1}{n}\sum_{i=1}^{n} \|size(\hat{p}_t^L) - size(\hat{p}_s^L)\| + \|size(\hat{p}_t^U) - size(\hat{p}_s^U)\|_2 \qquad (2)$$

$$L_{cls} = \frac{1}{n}\sum_{i=1}^{n} \|cls(\hat{p}_t^L) - cls(\hat{p}_s^L)\|_2 + \|cls(\hat{p}_t^U) - cls(\hat{p}_s^U)\|_2 \qquad (3)$$

$$L_{consis} = \alpha L_{ctr} + \beta L_{size} + \gamma L_{cls} \qquad (4)$$

where $\alpha, \beta and \gamma$ are 0.5, 1 and 5 respectly.

The final loss of semi-supervised learning is:

$$L = L_{sup} + \delta L_{consis} \qquad (5)$$

where δ is the consistency constraint weight, and we find that gradually increasing method δ can get better training results.

2.2.3 Two-stage CenterPoint

The above methods are performed in the first-stage CenterPoint, and after obtaining the first-stage network we then use the second-stage CenterPoint to make the final adjustments to the prediction results. Through the experiments, the second-stage CenterPoint can steadily improve the mAP by 0.4-0.6 percentage regardless of the supervised network or semi-supervised network.

3. Experiments

3.1. Implementation Details

For all of our models, we use a detection range of [-75.2m, 75.2m] for the X and Y axis, and [-5m, 3m] for the Z axis. The voxel size is (0.05m, 0.05m, 0.1m). The maximum number of objects is set to 500. We set the max point per voxel to 5, max voxel num to 150000 during training and 200000 at inference.

Supervised Networks. We train the model using AdamW [7] optimizer with onecycle [2] learning rate policy, with max learning rate 1e-3, weight decay 0.01, and momentum 0.85 to 0.95. We use a batch size of 4 evenly distributed across 2 V100 GPUs. We train the model for 80 epochs which takes about 30.5 hours.

Semi-Supervised Networks. We train the model using AdamW [7] optimizer with exponential decay [12] learning rate policy, with initial learning rate 1e-3, decay length 0.05, and decay factor 0.8. We use a batch size of 8 evenly distributed across 8 V100 GPUs. We train the model for 12 epochs which takes about 10 days.

3.2. Ablation Study

We only apply the training set to train the model and see the performance improvement of each entry in the validation set. Table 2 ablates the improvement of our entry based on original CenterPoint [11]. SSFA-SC-Conv brings a 3.8 mAPH improvement, flipping along the Y-axis during testing brings a 6.3mAPH improvement, and the semi-supervised method also brings a 3.9mAPH improvement. For our final submission, we train our model on the joint dataset of ONCE training and validation splits. This gives a improvement for test set accuracy (82.18 vs. 82.89).

Methods	Vehicle	Pedestrian	Cyclist	mAPH
basedet	88.38	84.45	82.52	85.12
zzhxyw	85.45	83.54	79.69	82.89
FangJin	83.96	78.22	76.48	79.55

Table 3: State-of-the-art comparisons for 3D detection on ONCE leaderboard. We show the mean average precision weighted by heading accuracy (mAPH).

3.3. Main Results

Table 3 shows the ICCV 2021 Workshop SSLAD Track 2 - 3D object detection challenge leaderboard. Our submission ranked second among all entries.

4. Conclusion

In this report, we demonstrate a semi-supervised CenterPoint 3D detection method and prove the effectiveness in the ONCE dataset. This learning framework consists of a supervised 3D network, a pseudo-labeling scheme, and a Mean-Teacher semi-supervised learning scheme. Based on the above framework, we win the second place in the ONCE dataset challenge.

References

[1] Tarvainen Antti and Valpola Harri. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. in advances in neural information processing systems. In *In Advances in neural information processing systems*, page 1195–1204, 2017.

[2] Sylvain Gugger. The 1cycle policy. `https://sgugger.github.io/the-1cycle-policy.html`. 2018.

[3] Liu Jiang-Jiang, Hou Qibin, Cheng Ming-Ming, Wang Changh, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. *CVPR*, 2020.

[4] Hongwu Kuang, Bei Wang, Jianping An, Ming Zhang, and Zehan Zhang. Voxel-fpn: Multi-scale voxel feature aggregation for 3d object detection from lidar point clouds. *Sensors*, 20(3):704, 2020.

[5] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[6] Zhidong Liang, Zehan Zhang, Ming Zhang, Xian Zhao, and Shiliang Pu. Rangeioudet: Range image based real-time 3d object detector optimized by intersection over union. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7140–7149, 2021.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

[9] Ge Runzhou, Ding Zhuangzhuang, Hu Yihan, Shao Wenxin, Huang Li, Li Kun, and Liu Qiang. 1st place solutions to the real-time 3d detection and the most efficient model of the waymo open dataset challenges 2021. 2021.

[10] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[11] Yin Tianwei, Zhou Xingyi, and JPhilipp Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021.

[12] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *arXiv preprint arXiv:1705.08292*, 2017.

[13] Zheng Wu, Tang Weiliang, Chen Sijin, Jiang Li, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *AAAI*, 2021.

[14] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[15] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020.