Focal Distillation From High-Resolution Data to Low-Resolution Data for 3D Object Detection

Jiawei Shan, Gang Zhang, Chufeng Tang, Hujie Pan, Qiankun Yu, Guanhao Wu, and Xiaolin Hu, *Senior Member, IEEE*

***Abstract*— LiDAR-based 3D object detection plays an essential role in autonomous driving. Although the detector trained on high-resolution data has much better performance than the same detector trained on low-resolution data, the high-resolution LiDAR cannot be widely used due to its high price. In this work, we propose a new distillation method called Focal Distillation to bridge the gap between high-resolution detector (teacher model) and low-resolution detector (student model). It consists of three essential components: focal classification distillation (FCD), focal regression distillation (FRD) and focal feature distillation (FFD). Taking the low-resolution data as input, the student model can learn discriminative features and produce more accurate results with the assistance of the teacher model trained on high-resolution data. We conducted extensive experiments to validate the effectiveness of Focal Distillation. Evaluated on the KITTI validation set, a typical SECOND model trained with Focal Distillation outperformed its non-distilled counterpart by 3.37%, 7.52%, 11.35% mAP on the category Car, Pedestrian, and Cyclist of moderate level, respectively. Moreover, the remarkable improvements observed on different models and different datasets further demonstrate the generalization ability of our proposed method.**

***Index Terms*— 3D object detection, knowledge distillation, 3D point cloud analysis.**

I. Introduction

LIDAR-BASED 3D object detection is an important task in autonomous driving. Autonomous vehicles need to accurately detect target objects (*e.g.*, people, cars) in real-time with limited on-board computational resources. Most existing benchmarks (*e.g.*, KITTI [1], Waymo [2]) collected data with 64-channel LiDARs. However, in practice, the application of the high-resolution LiDAR is limited by its high price and high computation complexity. For example, the 64-channel LiDAR is about 10 times more expensive than the 16-channel LiDAR. Based on the advantages of low price and low computation complexity, the low-resolution LiDAR is more practical for autonomous vehicles instead.

The resolution of the obtained point clouds depends on the number of LiDAR channels. The more the number of channels, the higher the resolution of the point cloud (*i.e.*, the more points on objects). In the past few years, detectors [3], [4], [5], [6] trained on high-resolution point clouds achieved remarkable performance on popular benchmarks [1], [2]. However, the reduction of point cloud resolution usually leads to a severe performance drop for 3D object detection [7], [8]. There are two main reasons: (1) For object classification, the context information in low-resolution point clouds is insufficient to yield discriminative features. (2) For box regression, it is difficult to regress the precise coordinates of objects when the point clouds are too sparse. We were naturally interested in *whether the performance of low-resolution detectors could be boosted through the guidance of high-resolution detectors*. In this work, we attempted to answer this question from the perspective of knowledge distillation.

Knowledge distillation, originally proposed by Hinton et al. [9], is widely used for model compression [10], [11], [12]. In most cases, the knowledge of a heavy model (*i.e.*, the teacher) is adopted as a soft target to guide the learning of a light-weight model (*i.e.*, the student), so as to compress the heavy model while maintaining the performance. We consider another purpose of knowledge distillation in this work, which is to distill the knowledge of high-resolution detectors to improve the performance of low-resolution detectors. The high-performance detectors trained on high-resolution data can serve as teacher models to produce discriminative features and high-quality results to guide the learning of low-resolution detectors. Note that we focus on the distillation between different data domains (from high-resolution data to low-resolution data), instead of different models (*e.g.*from a heavy model to a lightweight model).

Manuscript received 8 July 2022; revised 11 April 2023; accepted 12 July 2023. This work was supported in part by the Beijing Science and Technology Planning Project under Grant Z191100007419011, in part by the National Natural Science Foundation of China under Grant U19B2034, and in part by the Shanghai Automotive Industry Corporation (SAIC) Intelligent Technology under Contract CGHT-202112218. The Associate Editor for this article was J. Li. *(Corresponding authors: Guanhao Wu; Xiaolin Hu.)*

Jiawei Shan and Guanhao Wu are with the State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China (e-mail: sjw19@tsinghua.org.cn; guanhaowu@mail.tsinghua.edu.cn).

Gang Zhang and Chufeng Tang are with the Department of Computer Science and Technology, Institute for Artificial Intelligence, and the State Key Laboratory of Intelligent Technology and Systems, BNRist, Tsinghua University, Beijing 100084, China (e-mail: zhang-g19@mails.tsinghua.edu.cn; tcf18@mails.tsinghua.edu.cn).

Hujie Pan and Qiankun Yu are with SAIC Intelligent Technology, Shanghai 200041, China (e-mail: panhujie@saicmotor.com; yuqiankun@saicmotor.com).

Xiaolin Hu is with the Department of Computer Science and Technology, Institute for Artificial Intelligence, and the State Key Laboratory of Intelligent Technology and Systems, BNRist, Tsinghua University, Beijing 100084, China, and also with the Chinese Institute for Brain Research (CIBR), Beijing 102206, China (e-mail: xlhu@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TITS.2023.3304837

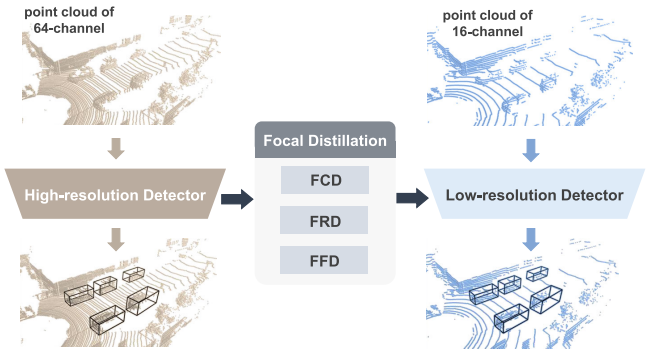


Fig. 1. **Illustration of our proposed Focal Distillation.** The high-resolution detector takes the high-resolution point clouds data as input, and the low-resolution detector takes the low-resolution point clouds data as input. Focal Distillation is utilized to boost performance of the low-resolution detector via the assistance of the high-resolution detector.

To achieve this goal, we propose a novel distillation method named **Focal Distillation (FD)**, as illustrated in Figure 1, which promotes the learning of the low-resolution detector via the assistance of the high-resolution detector. Specifically, the proposed method consists of three components as follows. (1) *Focal Classification Distillation (FCD)* adopts the output classification probabilities of the teacher model as soft targets for training the student model. (2) *Focal Regression Distillation (FRD)* utilizes the regression output of the teacher model as an intermediate instruction to help the student model localize objects more precisely. (3) *Focal Feature Distillation (FFD)* guides the feature learning of the student model based on the generated features from the teacher model. Besides, we introduce *focal weights* into each component mentioned above to lay more emphasis on those samples that are hard to classify, because performing knowledge distillation is less meaningful for easier samples that the student model has already handled well. Enhancing each component with focal weights enables the distillation process to be performed in an adaptive manner. Experiments demonstrate that the introduced focal weights are critical for the success of Focal Distillation.

To the best of our knowledge, this is the first work to perform knowledge distillation from high-resolution detectors to low-resolution detectors for 3D object detection in autonomous driving. The proposed Focal Distillation method can help the detectors trained on low-resolution point clouds to learn more discriminative representations and predict more precise bounding boxes. During inference, fed with the low-resolution point clouds only, the student model trained with Focal Distillation can produce promising detection results without any extra computation.

We evaluated the proposed method on two popular benchmarks of 3D object detection (KITTI [1] and Waymo [2]) and remarkable improvements were achieved compared to various baselines. In addition, we also conducted experiments to distill knowledge from higher resolution data to 64-channel data to demonstrate the effectiveness and the generalization ability of our method, where the higher resolution point clouds were obtained through a novel data augmentation method.

II. RELATEDWORK

A. 3D Object Detection

For 3D object detection based on point clouds [3], [4], [13], [14], most current methods project the disordered point clouds into ordered representation like pseudo-image or voxels, which can be fed into 2D or 3D CNN for further processing. Some previous methods [15], [16], [17], [18] represent point clouds as 2D image view like bird's-eye view and LiDAR front view, and then use Region Proposal Network [19] to generate positive proposals based on the predefined 3D anchors. But these works use RGB images as well. Taking point cloud data as input, the pioneer voxel-based single-stage 3D detector VoxelNet [20] groups point clouds into 3D voxels and then performs voxel-based feature extraction. To make it more efficient, SECOND [4] introduces the 3D sparse convolution for real-time applications. PointPillar [21] converts point clouds to vertical columns and encodes features on pillars. Point-based methods [22], [23], [24] learn point-wise features directly and can achieve impressive results. However, they require high computation and memory. CenterPoint [25] uses 3D centerness target and implements anchor-free regression detectors. CIA-SSD [14] optimizes confidence calibration based on IoU perception to make confidence and positioning accuracy more consistent. Some methods [3], [26], [27], [28], [29] combine advantages of both voxel-based and point-based methods to perform effective detection. Considering the degraded 3D detection performance when the object is occluded or the target is far away, Associate-3Ddet [30] uses augmented scenes to guide detection of real point clouds. SE-SSD [31] improves detector performance with self-ensembling. STD [26] generates points to refine objects and BtcDet [32] predicts the probability of object occupancy to improve detection performance. Our FD method can be applied to 3D detectors to improve its performance when the point cloud is extremely sparse.

B. Knowledge Distillation

Knowledge distillation is a model compression method introduced by [9], which usually transfers knowledge from a larger model to a smaller model to improve the performance of the latter. FitNets [33] proposes to add constraints to the middle layer of the thin and deep student model, and to transfer knowledge by imitating the middle layer features of the teacher model. Powered by the pioneer works, knowledge distillation has been applied to image classification [11], [34], object detection [35], [36], human pose estimation [37], [38], semantic segmentation [39], [40], [41], *etc*.. For the multi-class object detection task, Chen et al. [42] first propose an effective framework for learning compact object detectors with knowledge distillation. Wang et al. [43] find that global feature imitation can result in degraded performance for the student model and proposes to distill the feature knowledge under the areas near objects. By supervising the region of interests in a large network, Li et al. [44] simplify and accelerate the detector significantly. Yang et al. [45] calculate the attention of different pixels and channels in teacher's feature and propose to distill the relation between different

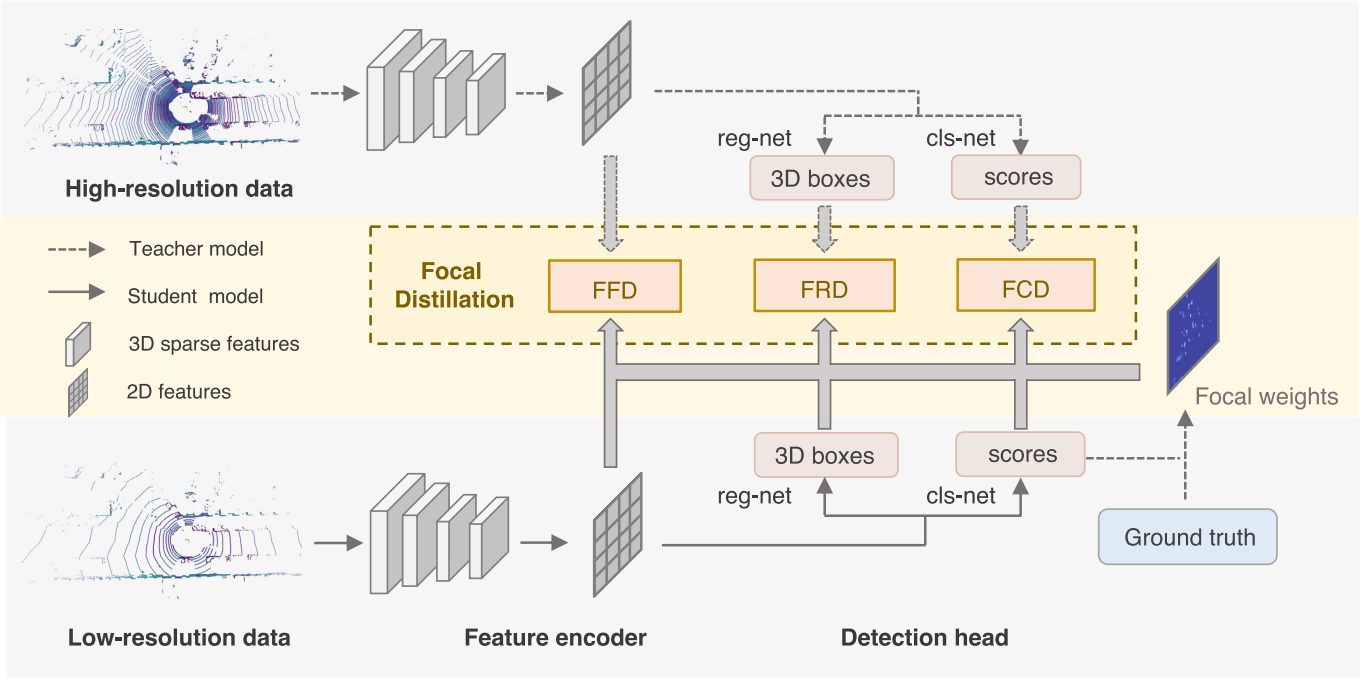


Fig. 2. **Framework of Focal Distillation**. Based on 3D object detectors, the FD method is introduced to boost the performance of the student model (the bottom row) via assistance of the teacher model (the top row). The teacher model is trained with high-resolution LiDAR data, and the student model is fed with low-resolution LiDAR data only. The Focal Distillation consists of three components, *i.e.*FCD, FRD, and FFD. The classification probabilities, regression results, and the last feature maps of the RPN from the teacher model are fed into FCD, FRD and FFD, respectively, serving as soft constraints for the learning of the student model. Reg-net and cls-net denote the regression head and the classification head, respectively.

pixels to make feature distillation more effective. In this work, we focus on the distillation between different data domains and propose a knowledge distillation method to facilitate the low-resolution detectors with high-resolution detectors.

C. Boosting Low-Resolution Model With High-Resolution Model

To improve the performance of low-resolution visual recognition, knowledge distillation are commonly used to boost classification accuracy [46]. Wang et al. [47] adopt the multi-kernel maximum mean discrepancy to narrow the data distribution discrepancy. Ge et al. [48] selectively transfer the most essential facial features of the teacher model to the student model by an optimized sparse graph. Xiao et al. [49] propose to integrate the high and low resolution features using resolution-aware transformations to promote the detection model for low-resolution pedestrians. For 3D object detection, Wang et al. [50] use knowledge distillation to narrow the gap between the model trained on aggregated multi-frame inputs and the model trained on single-frame inputs. In this work, we achieve significant improvements on 3D detection with low-resolution data. Our FD method performs knowledge distillation between high-resolution detectors and low-resolution detectors.

III. Method

An overview of the proposed framework is shown in Figure 2. Based on a 3D object detector that takes point cloud data as input, such as SECOND [4] and PV-RCNN [3], the Focal Distillation method is proposed to boost the performance of the student model via the guidance of the teacher model. We first introduce the overall framework in Section III-A, and then delve into the details of the Focal Distillation.

A. Overall Framework

Suppose there exists a high-performance detector trained on high-resolution data that can serve as a teacher model to produce discriminative features and high-quality results. Our FD method aims to leverage the teacher detector to guide the learning of the student detector trained on low-resolution data.

A 3D object detector usually consists of a feature extractor as well as an object classification head and a box regression head. As shown in Figure 2, the top row is the teacher model, and the bottom row is the student model. The student model can share the same architecture as the teacher model or use a more lightweight architecture. The proposed *Focal Distillation* method consists of three components. Moreover, in order to emphasize more on those hard samples, *focal weights* are introduced into all three components of Focal Distillation, which enables adaptive distillation and improves efficiency. We introduce the proposed focal weights and the three components of Focal Distillation as follows.

B. Focal Weights

Although the resolution reduction of point cloud data leads to a severe performance drop, there are still plenty of easy samples that the detector already handled well due to the intrinsic sparsity of point cloud data. An intuitive but reasonable idea is that performing knowledge distillation on easy samples is less meaningful than on hard samples. As a

result, we expect that the student model can focus on learning more discriminative representation and predicting more precise detection results for those hard samples with the assistance of the teacher model. Inspired by Focal Loss [51], which is originally proposed to solve the extreme imbalance problem in single-stage 2D object detectors, we introduce *focal weights* into our distillation components to adaptively attend to samples of different learning difficulties. For example, as those objects away from the sensor usually include fewer points, it is hard to learn discriminative features for classification and regression. For those samples around objects away from the sensor, we should put more emphasis on them during the distillation process.

The *focal weights* is defined as:

$$p = \begin{cases} \hat{p}_s, & \text{if } p_s = 1 \\ 1 - \hat{p}_s, & \text{otherwise,} \end{cases} \tag{1}$$

$$\mathcal{W}_s = \alpha(1-p)^{\gamma}, \tag{2}$$

where $\hat{p}_s$, p_s are the classification probability distribution produced by the student model and the classification target, respectively. A sample is positive if $p_s = 1$ and negative otherwise. α is the weight for balancing the foreground and the background categories and is defined analogously to how we defined p. When the sample is foreground, it is defined as α, otherwise it is defined as $1-\alpha$. And γ is a tunable focusing parameter to adjust the weights for different examples. We apply the modulation factor $\mathcal{W}_s$ to various components of the proposed Focal Distillation to adjust the weights for different samples adaptively as shown in Figure 3. With *focal weights*, the student model can focus more on the learning of hard samples, making the distillation process more effective.

C. Focal Classification Distillation

The goal of the classification head (cls-net in Figure 2) is to classify each sample (*e.g.*, pre-defined anchor) into a specific foreground category or the background. Due to the resolution reduction of point clouds, there are some cases where it is difficult for the student model to distinguish the foreground from the background, thus the guidance of the teacher model is more critical. In order to enhance the distillation performance for hard samples, the *focal weights* are utilized to adaptively adjust the weight for each sample according to the classification confidence of the student model. More attention is paid to the samples that are poorly classified in the student model.

In classification distillation, the student model learns to mimic the soft probability distribution produced by the teacher model to produce more accurate category predictions. The loss function for the classification distillation can be formulated as follows:

$$\mathcal{L}_{cls} = \mathcal{W}_s KL(\hat{p}_t, \hat{p}_s), \tag{3}$$

where KL indicates the KL-divergence loss, $\hat{p}_t$ is the probability distribution produced by the teacher model. Moreover, the general form of focal loss is demonstrated to be an entropy-regularized KL divergence upper bound between the predicted distribution and the target distribution [52].

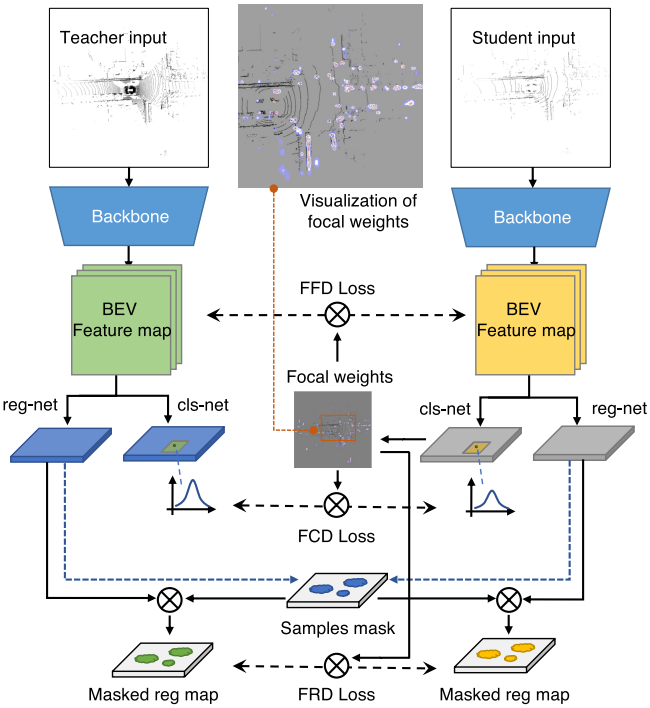


Fig. 3. **Visualization of the focal weights and the illustration of how FCD, FRD, and FFD work in the training process.** The input to the teacher model is a high-resolution point cloud, and the input to the student model is a low-resolution point cloud. Focal weights are determined by the classification probability distribution of the student model and applied in each component of the proposed FD method.

Therefore, the classification distillation component can also encourage the student model to predict class distribution with higher entropy, which can better handle the high uncertainty of bounding box regression in the sparse scene of low-resolution point clouds.

D. Focal Regression Distillation

It is hard for the student model to achieve precise localization since the low-resolution point clouds are too sparse. We propose to use the regression output of the teacher model as auxiliary guidance for the student model. We believe that the regression output of the teacher model can make it easier for the student model to learn from the regression targets.

Since the regression outputs of the teacher model are unbounded and some samples with poor results may introduce additional noise to the student model. Inspired by [42], we only perform regression distillation for those samples that the student model predicts worse results than the teacher model. Besides, we assign different weights to the samples with different difficulties using *focal weights*. The loss function can be formulated as follows:

$$\mathcal{L}_{reg} = \frac{1}{|P|} \sum_{i \in P} \sum_{j} \mathcal{M}(1 + \beta \mathcal{W}_s) |\hat{b}^i_{s_j} - \hat{b}^i_{t_j}|, \tag{4}$$

$$\mathcal{M} = \begin{cases} 1, & \text{if } \delta^i_j > m \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

$$\delta^i_j = ||\hat{b}^i_{s_j} - b^i_j||^2 - ||\hat{b}^i_{t_j} - b^i_j||^2, \tag{6}$$

where $\hat{b}^i_{s_j}$, $\hat{b}^i_{t_j}$, b^i_j are the regression output from the student model, the regression output from the teacher model and the regression target for the $j^{th} \in \{1, 2, \ldots, 7\}$ dimension of the 3D bounding box of the i^{th} positive samples, respectively. Specifically, the 3D bounding box is defined by $(x, y, z, l, w, h, \theta)$. Here, (x, y, z) represent the center coordinates of the 3D bounding box, (l, w, h) represent the length, width, and height of the box, respectively, and θ represents the yaw angle of the box. And m is the threshold to decide which samples need to be distilled.

E. Focal Feature Distillation

As stated in some recent works [43], [44], feature distillation is an effective way to improve the informativeness of features for the student model. However, it is difficult to optimize due to the feature redundancy if the student model utilizes the features of the teacher model inappropriately. Therefore, it is important to find the most discriminative regions for feature distillation. To this end, we measure the importance of different regions in the feature map by *focal weights*, which adaptively focus on the crucial regions. We perform feature distillation on the last feature map before the classification head and the regression head. Through feature distillation, the low-resolution detector learns as much discriminative information as possible under the guidance of the high-resolution detector, thereby assisting the sub-networks for classification and regression. The loss function for feature distillation can be formulated as follows:

$$\mathcal{L}_{feat} = \mathcal{W}_s ||f_s - f_t||_2^2, \tag{7}$$

where f_t denotes the features generated by the teacher model, and f_s denotes the features of the student model.

F. Traininng of Focal Distillation

The overall training objective for the student model can be formulated as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{cls} + \eta \mathcal{L}_{reg} + \sigma \mathcal{L}_{feat} + \mathcal{L}_{detector} \tag{8}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, $\mathcal{L}_{feat}$ denote the losses for classification distillation, regression distillation and features distillation, respectively. $\mathcal{L}_{detector}$ is the loss for the original detector. λ, η and σ are the weights for each component.

G. A Data-Augmentation Method

For training the teacher model, the higher the resolution of the point clouds the better. However, LiDARs with more than 64 channels are difficult to access in practice. Is there a way to obtain higher resolution data with 64-channel LiDAR? We propose a data augmentation method for this purpose, which enables the teacher model to produce more accurate results to guide the learning of the student model via the proposed FD method. The main idea is to augment the points for objects by utilizing a sequence of data for the same objects. This is particularly useful for distant objects as they usually have sparse points. Please refer to chapter IV-E for details.

IV. Experiment

A. Datasets

We performed experiments on two popular datasets: KITTI [54] and Waymo [2] datasets to validate the effectiveness of our proposed method.

KITTI Dataset KITTI dataset [1] is one of the most popular datasets of 3D object detection consisting of three categories named *Car*, *Pedestrian*, and *Cyclist*, respectively. It contains 7481 training samples. Following [55], we divided the whole dataset into two subsets: 3712 samples for training and 3769 samples for validation. The top LiDAR in KITTI is a high-resolution LiDAR Velodyne HDL-64E of 64-channel. We adopted the official evaluation tools, which calculate the mean average precision (mAP) using 40 recall positions. Three difficulty levels of easy, moderate, and hard are defined by the height, occlusion, and truncation of the bounding boxes. The rotated IoU threshold is set to 0.7 for *Car* and 0.5 for *Pedestrian* and *Cyclist*.

Waymo Dataset Waymo Open Dataset [2] is currently the largest dataset with a LiDAR sensor for autonomous driving, which contains 798 training sequences and 202 validation sequences. The sensor suite for Waymo consists of a top 64-channel LiDAR, which captured 180,000 lidar points every 0.1 seconds. We adopted the officially released evaluation tools, which calculate the mean average precision (mAP) and the mean average precision weighted by heading accuracy (mAPH). The rotated IoU threshold is set to 0.7 for *vehicle* and 0.5 for *pedestrian* and *cyclist*. The dataset is split into two difficulty levels: LEVEL_1 and LEVEL_2. LEVEL_1 includes object boxes with at least five inside points, while LEVEL_2 includes object boxes with at least one inside point.

B. Implementation Details

We implemented our method based on the official code released by PV-RCNN [3] and followed the default configurations. Take the anchor-based methods as examples, we validated the effectiveness of our method on four baseline models: SECOND [4], PiontPillars, Part-A^2 [5], and PV-RCNN [3]. Among them, SECOND [4] and PointPillar [21] are very popular one-stage detectors in autonomous driving. Part-A^2 [5] and PV-RCNN [3] are two-stage detectors with better detection performance. Both the teacher model and the student model were trained from scratch under the same configurations in an end-to-end manner using the Adam optimizer except for the new introduced parts. We trained the SECOND model and the PointPillar model with the batch size of 4 per GPU, and the learning rate of 0.003 for 80 epochs. For Part-A^2 and PV-RCNN, the learning rate was 0.01 for 80 epochs and the batch size was 4 and 2, respectively. For FD method, we set β, λ, η and σ to 2, 5, 1, 2 respectively. And m in RFD was 0.00001. Besides, the hyperparameters α and γ in focal weights are set to 0.25 and 2, the same as in [51]. Noted that since Part-A^2 and PV-RCNN are two-stage frameworks, we only applied the knowledge distillation for the first proposal generation stage, without any additional modifications in the second stage for proposal refinement.

TABLE I
DATA CONFIGURATIONS FOR KITTI. HR DENOTES THE HIGH-RESOLUTION DATA FROM VELODYNE HDL-64E. LR DENOTES THE SIMULATED LOW-RESOLUTION DATA OF 16 CHANNELS

Attributes	HR	LR
Channel Number	64	16
Vertical FOV	26.8°	12°
Vertical resolution	0.4°	0.8°
Azimuthal FOV	360°	360°

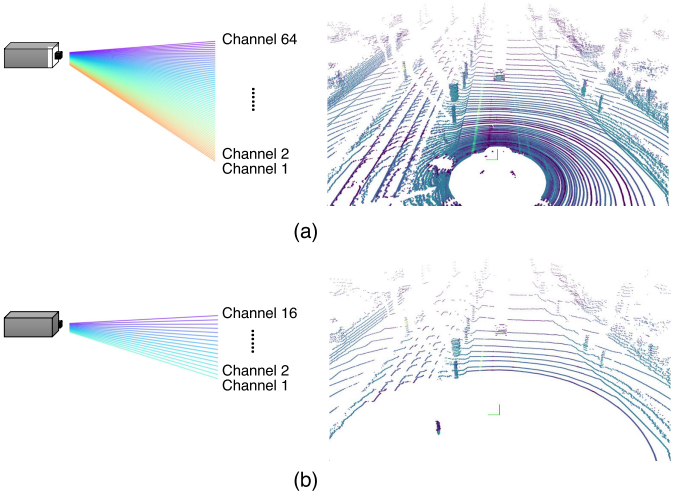


Fig. 4. **Visualization of KITTI point cloud data.** (a) The scanning laser distribution of 64-channel LiDAR and the high-resolution point clouds. (b) The scanning laser distribution of the simulated 16-channel LiDAR and the low-resolution point clouds.

C. Main Results

1) Experimental Settings: The top LiDARs in KITTI [1] and Waymo [2] are both high-resolution LiDARs of 64 channels. To evaluate our proposed FD method, we selected the suitable subsets of the channels based on the LiDAR configurations to generate the low-resolution data of 16 channels from the raw data. For KITTI, the details of data configurations are shown in Table I. Since the point clouds in KITTI are unorganized, we first reconstructed the data according to the rules of data arrangement and then assigned the channel number for each point. An example of the generated low-resolution data is shown in Figure 4. For Waymo, we labeled the channel number for each point according to the range number in range images. Since the parameter of vertical resolution of Waymo's top LiDAR is unknown, we simply simulate 16-channel point clouds by reserving one channel for every four channels out of raw data as shown in Figure 5. Note that the ground truth that does not contain any points was removed during training and evaluation on the simulated 16-channel data.

2) Quantitative Results on KITTI Dataset: We evaluated our method on several models [3], [4], [5], [21]. The results are shown in Table II. Note that the "baseline" indicates the model trained on low-resolution data (*i.e.*, 16-channel data) without using our FD method.

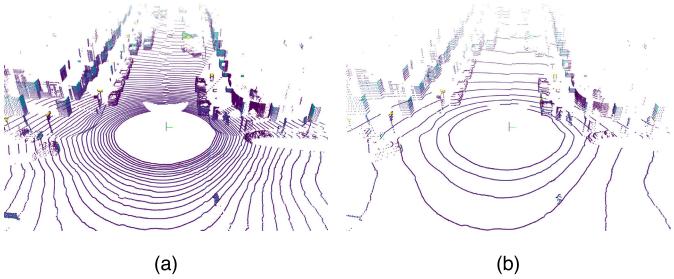


Fig. 5. **Visualization of Waymo point cloud data.** (a) The high-resolution point clouds of 64-channel from the top LiDAR. (b) The simulated low-resolution point clouds of 16-channel.

Tested on the 16-channel data, the proposed FD consistently improved the results of different models on all three categories of different difficulty levels. Note that the PV-RCNN model trained on the 16-channel data even yields superior results in the category *Pedestrian* than its 64-channel counterpart. In addition, our proposed FD successfully decreased the large performance gap from the reduced resolution of the input data of easy level. For example, our FD method narrowed 84.84%, 70.64%, 63.91% of the performance drops on the SECOND, Part-A^2, and PV-RCNN models on the category Car of easy level, respectively.

We also reported the performance of the most important car category on the BEV view on the KITTI *val* set with mAP. Similarly, as shown in Table III, the model with our FD method yielded remarkable gains in three categories, which further demonstrates the effectiveness of our method.

3) Inference Speed on KITTI Dataset: We evaluated the inference speed of our FD method on a single NVIDIA 2080Ti GPU and observed that low-resolution detectors achieved a higher Frames Per Second(FPS) than high-resolution detectors, as shown in Table II. This is because less computational costs are required to process the input data of lower resolution. It is worth noting that our FD distillation method does not increase the inference time of the student model, as the teacher and student models share the same network architecture.

4) Quantitative Results on Waymo Dataset: We evaluated our FD method on the Waymo Open Dataset with 202 validation sequences for detection on *Vehicle*, *Pedestrian* and *Cyclist*. Table IV shows that our proposed FD method improves the mAP and mAPH metrics on both two difficulty levels LEVEL_1 and LEVEL_2 for all three categories.

5) Qualitative Results: Detection results are shown in Figure 6. Compared with the normal low-resolution detectors, the model trained with FD produces fewer false positive objects and outputs more accurate bounding boxes. These results indicates that the model with the FD method has a stronger ability to distinguish target objects from background.

D. Ablation Study

We conducted ablation experiments to analyze our proposed FD method. All student models utilized the SECOND model as the baseline and were trained on the simulated low-resolution data generated from KITTI.

TABLE II

3D DETECTION RESULTS ON THE KITTI *val* SET. FD IS OUR PROPOESD METHOD. 16-CHANNEL REPRESENTS THE SIMULATED LOW-RESOLUTION DATA. * DENOTES THE RE-IMPLEMENTATION BY USING THE [53].**T** AND **S** REFER TO THE TEACHER AND STUDENT DETECTOR, RESPECTIVELY.NOTE THAT THE RESULTS WERE EVALUATED BY THE MAP WITH 40 RECALL POSITIONS

Data	Method	Car			Pedestrian			Cyclist			FPS
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard	
64-channel	SECOND [4]* (T)	90.55	81.61	78.61	55.94	51.14	46.17	82.96	66.74	62.78	17.9
16-channel	SECOND (S)	87.45	75.77	73.13	45.89	40.49	36.88	70.43	48.60	45.52	23.3
	SECOND (S) w/ FD	90.08	79.14	76.18	52.99	48.01	44.24	77.31	59.95	56.06	23.3
	Improvement	**+2.63**	**+3.37**	**+3.05**	**+7.10**	**+7.52**	**+7.36**	**+6.88**	**+11.35**	**+10.54**	
64-channel	PointPillar [21]* (T)	87.75	78.40	75.18	57.30	51.41	46.87	81.57	62.81	58.83	23.3
16-channel	PointPillar (S)	84.78	73.54	69.93	49.31	44.02	40.47	70.54	50.21	47.51	30.0
	PointPillar (S) w/ FD	85.77	75.62	72.70	51.04	45.7	42.33	75.41	52.69	49.46	30.0
	Improvement	**+0.99**	**+2.08**	**+2.77**	**+1.73**	**+1.68**	**+1.86**	**+4.87**	**+2.48**	**+1.95**	
64-channel	Part-A^2 [5]* (T)	92.15	82.91	82.00	66.89	59.68	54.62	90.34	70.14	66.93	8.8
16-channel	Part-A^2 (S)	88.88	76.78	74.37	59.70	53.51	48.92	79.68	55.76	52.32	11.2
	Part-A^2 (S) w/ FD	91.19	79.07	76.74	62.88	56.47	51.59	81.84	58.08	54.50	11.2
	Improvement	**+2.31**	**+2.29**	**+2.37**	**+3.18**	**+2.96**	**+2.67**	**+2.16**	**+2.32**	**+2.18**	
64-channel	PV-RCNN [3] (T)	92.57	84.83	82.69	64.26	56.67	51.91	88.88	71.95	66.78	7.7
16-channel	PV-RCNN (S)	90.27	79.87	77.47	62.18	55.25	51.00	79.49	57.59	54.15	11.1
	PV-RCNN (S) w/ FD	91.74	80.26	79.17	65.54	58.03	53.55	82.98	60.85	56.93	11.1
	Improvement	**+1.47**	**+0.39**	**+1.70**	**+3.36**	**+2.78**	**+2.55**	**+3.49**	**+3.26**	**+2.78**	

TABLE III

DETECTION RESULTS ON THE KITTI *val* SET ON BEV VIEW. ALL RESULTS WERE EVALUATED ON THE SIMULATED LOW-RESOLUTION DATA

Method	FD	Car		
		Easy	Mod	Hard
SECOND		91.45	86.13	84.34
	✓	**92.41**	**88.09**	**85.51**
PointPillar		91.28	85.63	83.43
	✓	**92.13**	**87.22**	**85.02**
Part-A^2		92.55	85.16	82.92
	✓	**92.98**	**86.52**	**84.86**
PV-RCNN		92.63	87.66	85.69
	✓	**94.93**	**88.15**	**86.06**

1) Effectiveness of the Three Main Components: We analyzed the effect of each component of the FD method by applying each component to the original SECOND model individually. Results in Table V show that every component has a positive contribution to the detection performance. We believe that the teacher model has valuable content in all three components that can be emulated by the student model. Our FD method enables the student model to effectively capture this content. With FCD, the student model's ability to distinguish foreground from background in low-resolution data was enhanced, since the insufficient context information in low-resolution point clouds was compensated by the guidance of a high-resolution detector. With FRD, the student model's ability to regress the target location was enhanced, since the regression output of the teacher model can serve as an auxiliary guide for the student model. With FFD, the low-resolution detector can learn discriminative features with guidance from the high-resolution detector, which in turn, benefits the classification and regression sub-networks by improving their overall performance.

2) Effectiveness of Focal Weights: Results in Table VI show that using focal weights for distillation performs much better than treating all samples and features equally. We believe the reason is that the focal weights can adaptively make the distillation process attend to the hard samples for the student model, which encourages the learning of the student model more effective. In addition, a large number of easy samples are down-weighted in distillation, so the student model can learn as much valuable knowledge as possible from the teacher.

3) Effectiveness of the Bounded Distillation in FRD: We conducted a comparative experiment to evaluate the performance of the student model trained using the unbounded regression outputs of the teacher model versus the bounded regression outputs. Results in Table VII show that filtering out the poor-performing samples from the regression outputs of the teacher model can significantly improve the performance of the student model. This is a crucial step to ensure that the student model is not adversely affected by the noise introduced by poor-performing samples.

4) Comparison of LiDAR Configurations: To demonstrate the generality of our FD method, we conducted experiments on data of different LiDAR configurations. Since the

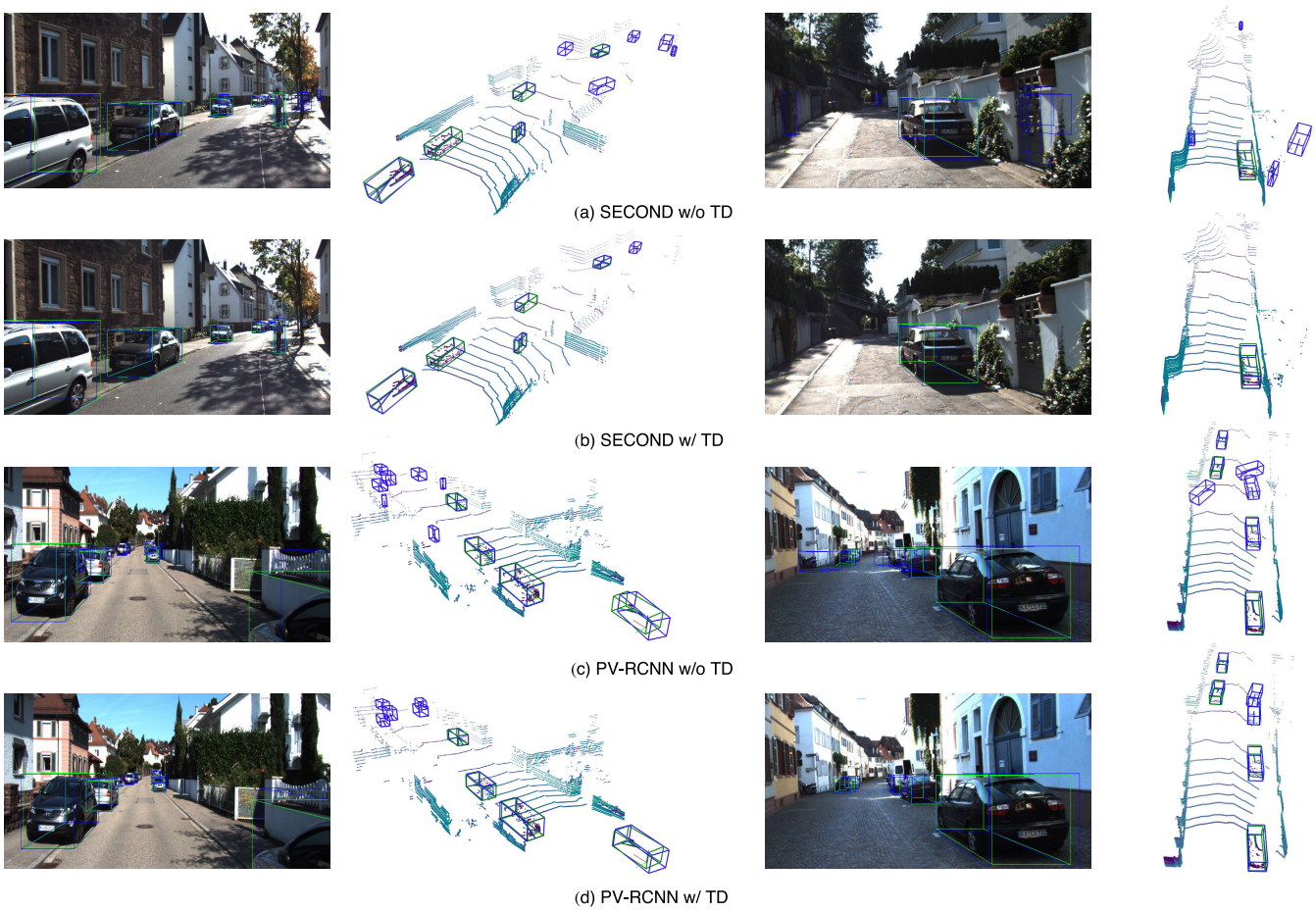


Fig. 6. **Visualization of KITTI examples.** The first row (a) and the second row (b) show the detection results generated by the SECOND model trained on the simulated low-resolution data without and with Focal-Distillation respectively. The last two row (c) and (d) show the results generated by the PV-RCNN model trained on the simulated low-resolution data without and with Focal-Distillation respectively. Blue box and green box are the predicted bounding box and the ground truth respectively. For better understanding the scene, 3D boxes detected using LiDAR are also projected on to corresponding images.

TABLE IV

3D DETECTION RESULTS ON THE WAYMO *val* SET. 16-CHANNEL REPRESENTS THE GENERATED LOW-RESOLUTION DATA.**T** AND **S** ARE ABBREVIATIONS FOR THE TEACHER AND STUDENT DETECTORS, RESPECTIVELY. L_1 AND L_2 DENOTE THE TWO DIFFICULTY LEVELS LEVEL_1 AND LEVEL_2

Data	Method	Veh.(L_1)		Veh.(L_2)		Ped.(L_1)		Ped.(L_2)		Cyc.(L_1)		Cyc.(L_2)	
		mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
64-channel	SECOND (T)	73.57	72.98	65.61	65.07	69.99	59.48	62.37	52.85	61.86	60.59	59.59	58.36
16-channel	SECOND (S)	62.63	61.84	57.09	56.37	61.45	46.21	56.01	42.01	47.16	45.27	46.36	44.50
	SECOND (S) w/ FD	65.60	64.74	60.27	59.46	64.21	48.13	58.78	43.94	52.42	50.59	51.56	49.75
	Improvement	**+2.97**	**+2.90**	**+3.18**	**+3.09**	**+2.76**	**+1.92**	**+2.77**	**+1.93**	**+5.26**	**+5.32**	**+5.20**	**+5.25**

vertical angular resolution of the Velodyne HDL-64 in KITTI is 0.4 degrees, we compare the results for the 16-channel point clouds with the vertical angular resolution of 0.8, 1.2, and 1.6 degrees, respectively. As shown in Table VIII, the proposed FD method brought great positive effects on the student models fed with point clouds of different LiDAR configurations.

E. Results of the Data-Augmentation Method

For training a better teacher model, we proposed a data augmentation method for the teacher model based on our FD method and conducted experiments. All experiments were performed on Waymo [2] dataset.

1) Obtaining Higher-Resolution Point Clouds: Point cloud data is usually dense in the near and sparse in the distance, as shown in Figure 7. But objects can be observed at different perspectives and distances in a sequence of point clouds. The nearby dense points for objects from a sequence of data can be used to augment the sparse points for distant objects to obtain high-resolution data. The main process of the data augmentation is shown in Figure 8. Different objects in consecutive frames can be distinguished by *object ID*, which

TABLE V
EFFECTIVENESS OF THE THREE MAIN COMPONENTS. FCD DENOTES FOCAL CLASSIFICATION DISTILLATION, FRD DENOTES FOCAL REGRESSION DISTILLATION, FFD DENOTES FOCAL FEATURE DISTILLATION

FCD	FRD	FFD	Car		
			Easy	Mod	Hard
			87.45	75.77	73.13
✓			88.76	78.51	75.21
	✓		87.94	77.80	74.83
		✓	87.53	77.29	74.51
✓	✓	✓	**90.08**	**79.14**	**76.18**

TABLE VI
EFFECTIVENESS OF FOCAL WEIGHTS

Setting	Car		
	Easy	Mod	Hard
Baseline	87.45	75.77	73.13
FD w/o focal weights	88.03	77.10	73.61
FD	**90.08**	**79.14**	**76.18**

TABLE VII
EFFECTIVENESS OF THE BOUNDED DISTILLATION IN FRD. FRD DENOTES FOCAL REGRESSION DISTILLATION. THE RESULTS ARE EVALUATED ON THE EASY DIFFICULTY OF THE KITTI DATASET

Setting	Car	Ped.	Cyc.
Baseline	87.45	45.89	70.43
FD w/ unbounded FRD	88.76	52.19	75.55
FD w/ bounded FRD	**90.08**	**52.99**	**77.31**

TABLE VIII
EFFECTS OF FOCAL DISTILLATION FOR POINT CLOUDS WITH DIFFERENT VERTICAL RESOLUTIONS

Vertical Resolution	FD	Car		
		Easy	Mod	Hard
0.8°		87.45	75.77	73.13
	✓	90.08	79.14	76.18
		+2.63	**+3.37**	**+3.05**
1.2°		84.58	69.91	65.68
	✓	87.03	71.41	68.04
		+2.45	**+1.50**	**+2.36**
1.6°		84.70	70.01	66.97
	✓	86.33	73.73	69.40
		+1.63	**+3.72**	**+2.43**

can be provided by dataset annotations or obtained by object tracking models, but is usually not available during inference. Assume there are a series of point clouds $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$ in the i^{th} object. The ground truth 3D bounding boxes of the i^{th} object in each frame is $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ and $\mathbf{b} \in \mathbb{R}^7$. We first extract the points for the object from different frames

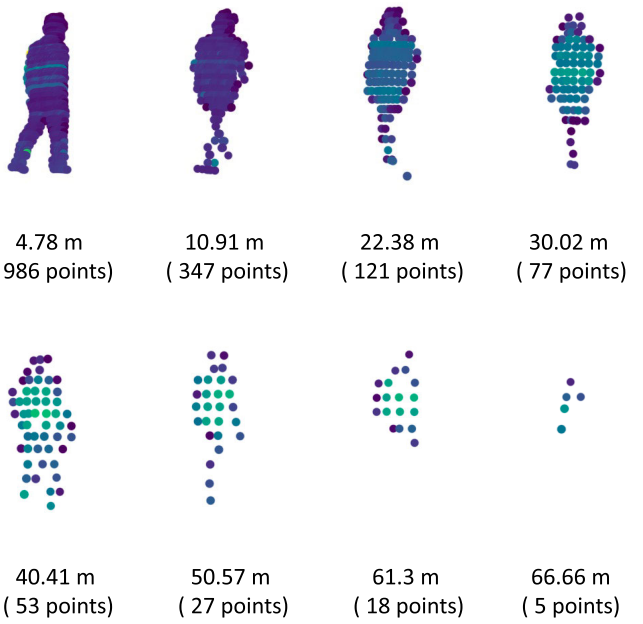


Fig. 7. **Visualization of a pedestrian instance at different distances in the Waymo dataset.**

based on the ground-truth 3D bounding boxes in each frame and rotated them into the same pose. All points for the object can be concatenated to get a larger point set. We randomly sample N points from the point set as the augmented points. Finally, the objects with fewer than N points are augmented by combining the augment points and raw points for the i^{th} object in each frame.

After data augmentation, the number of points on the most of objects increases significantly, especially for objects in distance, which enables the teacher model to produce more accurate results to guide the learning of the student model via the proposed FD method.

2) Experiment Settings: We validated the effectiveness of our method on the SECOND [4] model and the PV-RCNN++ [56] model with an anchor-based head. Note that we implemented the PV-RCNN++ model with an anchor-based head for proposal generation by using the official code of [56]. In the training process, the teacher model was trained on the point cloud data generated by 64-channel data according to the augmentation method described in Section IV-E.1. The number of points for each object is set to 1000, and the student model was trained on original point clouds. For FD method, we set β, λ, η and σ as 2, 2, 1, 2 respectively, and the hyper-parameters α and γ in focal weights were set to 0.25 and 2.

3) Results: Table IX shows that our FD method combined with the data-enhanced teacher model has obvious improvement for the SECOND models trained on both 16-channel data and 64-channel data. For 16-channel data, the FD method with the teacher model using the augmented data brought slightly less improvement than using 64-channel data directly. This is because the augmented data is not perfect, and the domain gap between it and 16-channel data is significant.

We listed the results of SECOND and PV-RCNN++ combined with the data-augmented teacher model in Table X, together with the results of some recent methods. By using

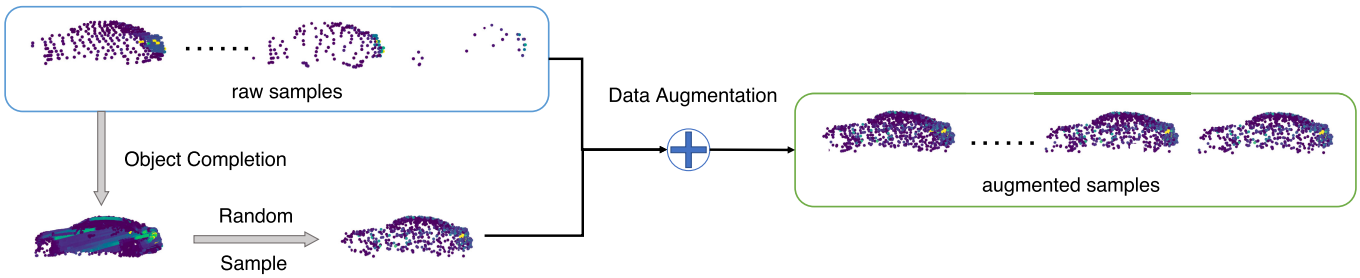


Fig. 8. **Illustration of the proposed data augmentation method**. The samples for the same objects from a sequence data are shown in the blue box. Object completion denotes aggregating all these samples to get a larger point set. Points are randomly sampled to augment the samples inside the blue box that have fewer points. The samples in the green box are the augmented samples.

TABLE IX

3D DETECTION RESULTS ON THE WAYMO *val* SET. AUGMENTED REPRESENTS THE AUGMENTED POINT CLOUDS OBTAINED BY THE PROPOSED DATA AUGMENTATION METHOD.64-CHANNEL DENOTES THE ORIGINAL POINT CLOUDS. 16-CHANNEL REPRESENTS THE GENERATED LOW-RESOLUTION DATA. **T** AND **S** ARE ABBREVIATIONS FOR THE TEACHER AND STUDENT DETECTORS, RESPECTIVELY. L_1 AND L_2 DENOTE THE TWO DIFFICULTY LEVELS LEVEL_1 AND LEVEL_2

Data	Method	Veh.(L_1)		Veh.(L_2)		Ped.(L_1)		Ped.(L_2)		Cyc.(L_1)		Cyc.(L_2)	
		mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
Augmented	SECOND (T)	88.55	87.90	83.80	83.16	88.25	79.38	84.55	75.84	79.38	78.30	78.60	77.53
16-channel	SECOND (S)	62.63	61.84	57.09	56.37	61.45	46.21	56.01	42.01	47.16	45.27	46.36	44.50
	SECOND (S) w/ FD	65.02	64.11	59.70	58.84	63.86	46.72	58.42	42.62	50.98	48.83	50.12	48.00
	Improvement	**+2.39**	**+2.27**	**+2.61**	**+2.47**	**+2.41**	**+0.51**	**+2.41**	**+0.61**	**+3.82**	**+3.56**	**+3.76**	**+3.50**
64-channel	SECOND (S)	72.27	71.69	63.85	63.33	68.70	58.18	60.72	51.31	60.62	59.28	58.34	57.05
	SECOND (S) w/ FD	74.85	74.24	67.02	66.45	73.23	62.00	65.12	55.00	67.58	66.12	65.17	63.76
	Improvement	**+2.58**	**+2.55**	**+3.17**	**+3.12**	**+4.53**	**+3.82**	**+4.40**	**+3.69**	**+6.96**	**+6.84**	**+6.83**	**+6.71**

TABLE X

COMPARISONS FOR 3D DETECTION ON THE WAYMO *val* SET. THE EVALUATION METRIC IS THE MAP IN TERMS OF LEVEL 1 DIFFICULTY AND THE TOP RESULTS ARE HIGHLIGHTED. * DENOTES THE RESULTS ARE RE-IMPLEMENTED BY [56].NOTE THAT THE PV-RCNN++ IS EQUIPPED WITH AN ANCHOR-BASED RPN HEAD

Method	Published	Veh.	Ped.	Cyc.
*SECOND [4]	Sensors 2018	72.27	68.70	60.62
MVF [57]	CoRL 2019	62.93	65.33	
Pillar-OD [58]	ECCV 2020	69.80	72.51	-
*Part-A^2 [5]	TPAMI 2020	77.05	75.24	68.60
*PV-RCNN [3]	CVPR 2020	77.51	75.01	67.81
CenterPoint-Voxel [25]	CVPR 2021	76.7	79.0	-
PV-RCNN++ [56]	IJCV 2022	79.19	77.97	72.10
PV-RCNN++ w/ FD	-	**79.47**	**79.75**	**74.06**

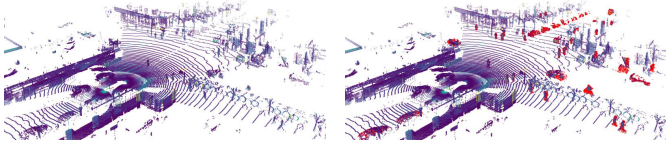


Fig. 9. **Visualization of point clouds with data augmentation.** (left) The original point cloud, and (right) the augmented point cloud. The red points represent the augmented points.

FD method, The PV-RCNN++ with an anchor-based RPN head achieved the state-of-the-art performance.

V. CONCLUSION

In this work, we propose a method named Focal Distillation method to bridge the performance gap between high-resolution detectors and low-resolution detectors. Specifically, three components named Focal Classification Distillation, Focal Regression Distillation and Focal Feature Distillation are introduced to guide the training process of the student model, which takes the low-resolution data as input. By focusing on the crucial samples and the crucial features that may have an essential impact on detection performance, the student model learns discriminative features and predicts more accurate results. We believe that our method can promote future research on 3D object detection based on low-resolution point cloud data.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[2] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2443–2451.

[3] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10526–10535.

[4] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[5] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2647–2664, Aug. 2021.

[6] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 770–779.

[7] Z. Wang et al., "Range adaptation for 3D object detection in LiDAR," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) Workshops*, Oct. 2019, pp. 2320–2328.

[8] R. Théodose, D. Denis, T. Chateau, V. Frémont, and P. Checchin, "A deep learning approach for LiDAR resolution-agnostic object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14582–14593, Sep. 2022.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Workshops*, 2014.

[10] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 441–449.

[11] W. Chen, C.-C. Chang, C.-Y. Lu, and C.-R. Lee, "Knowledge distillation with feature maps for image classification," in *Proc. Asian Conf. Comput. Vis.*, Dec. 2018, pp. 200–215.

[12] J. Wang, W. Bao, L. Sun, X. Zhu, B. Cao, and P. S. Yu, "Private model compression via knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 1190–1198.

[13] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards high performance voxel-based 3D object detection," 2020, *arXiv:2012.15712*.

[14] W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, "CIA-SSD: Confident IoU-aware single-stage object detector from point cloud," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 3555–3562.

[15] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534.

[16] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7337–7345.

[17] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 641–656.

[18] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[20] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[21] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12689–12697.

[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.

[23] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 652–660.

[24] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11037–11045.

[25] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11779–11788.

[26] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "STD: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1951–1960.

[27] P. Bhattacharyya and K. Czarnecki, "Deformable PV-RCNN: Improving 3D object detection with learned deformations," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2020.

[28] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3D object detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11870–11879.

[29] J. Noh, S. Lee, and B. Ham, "HVPR: Hybrid voxel-point representation for single-stage 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14600–14609.

[30] L. Du et al., "Associate-3Ddet: Perceptual-to-conceptual association for 3D point cloud object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13326–13335.

[31] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14489–14498.

[32] Q. Xu, Y. Zhong, and U. Neumann, "Behind the curtain: Learning occluded shapes for 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, pp. 2893–2901.

[33] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[34] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2016, pp. 3560–3566.

[35] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 16797–16807.

[36] Y. Wei, X. Pan, H. Qin, W. Ouyang, and J. Yan, "Quantization mimic: Towards very tiny CNN for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.

[37] M. R. U. Saputra, P. Gusmao, Y. Almalioglu, A. Markham, and N. Trigoni, "Distilling knowledge from a deep pose regressor network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 263–272.

[38] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3512–3521.

[39] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2599–2608.

[40] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 578–587.

[41] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, and C. C. Loy, "Inter-region affinity distillation for road marking segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12483–12492.

[42] G. Chen, W. Choi, X. Yu, T. X. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 742–751.

[43] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4928–4937.

[44] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7341–7349.

[45] Z. Yang et al., "Focal and global knowledge distillation for detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4633–4642.

[46] M. Zhu, K. Han, C. Zhang, J. Lin, and Y. Wang, "Low-resolution visual recognition via deep feature distillation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3762–3766.

[47] M. Wang, R. Liu, N. Hajime, A. Narishige, H. Uchida, and T. Matsunami, "Improved knowledge distillation for training fast low resolution face recognition model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) Workshops*, Oct. 2019, pp. 2655–2661.

[48] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[49] L. Xiao, W. Yaonan, Z. Xuanyu, Z. Qiuli, and L. Zhigang, "Generalizing pedestrian detector to low resolution images by integrating high and low resolution features," in *Proc. World Congr. Intell. Control Automat. (WCICA)*, Jul. 2018, pp. 1300–1305.

[50] Y. Wang, A. Fathi, J. Wu, T. Funkhouser, and J. Solomon, "Multi-frame to single-frame: Knowledge distillation for 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2020.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[52] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 15288–15299.

[53] OD Team. (2020). *OpenPCDet: An Open-Source Toolbox for 3D Object Detection From Point Clouds*. [Online]. Available: https://github.com/open-mmlab/OpenPCDet

[54] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[55] X. Chen et al., "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 424–432.
[56] S. Shi et al., "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 531–551, Feb. 2023.
[57] Y. Zhou et al., "End-to-end multi-view fusion for 3D object detection in LiDAR point clouds," in *Proc. Conf. Robot Learn.*, 2020, pp. 923–932.
[58] Y. Wang et al., "Pillar-based object detection for autonomous driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 18–34.

Jiawei Shan received the B.E. degree in information management and information systems from the University of International Relations, Beijing, China, in 2019, and the M.E. degree in instrument and meter engineering from Tsinghua University, Beijing, in 2022. Her current interests include deep learning and computer vision, especially 3D object detection.

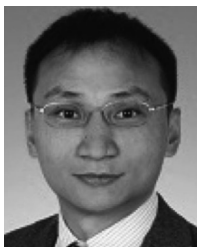
Gang Zhang received the B.E. degree from the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include deep learning and computer vision, especially 2D and 3D general object detection and segmentation.

Chufeng Tang received the B.E. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2018, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2023. His current interests include deep learning and computer vision, especially object detection and image segmentation.

Hujie Pan received the B.E. degree in mechanical engineering and the Ph.D. degree in automotive engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014 and 2021, respectively. He is currently a Senior Manager with the SAIC Intelligent Technology and in charge of the development of the perception system for Robotaxi. His current research interests include single- and multi-modal algorithms for object detection, scene understanding and reconstruction, and motion prediction.

Qiankun Yu received the B.S. degree in measurement-control technology and instrument and the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2013 and 2018, respectively. Currently, he is the Director of the Autopilot Algorithm Laboratory, SAIC Intelligent Technology (Shanghai) Company Ltd. His current research interests include autopilot algorithms and software in perception, prediction, planning, and control.

Guanhao Wu received the Ph.D. degree in optical engineering from Tsinghua University in 2008. He was a Visiting Scholar with the National Metrology Institute of Japan (NMIJ) from 2011 to 2012. He is currently an Associate Professor with the Department of Precision Instruments, Tsinghua University. His current research interests include frequency comb, comb-based distance measurement, light detection and ranging (Lidar), and other precise dimensional measurement technology.



Xiaolin Hu (Senior Member, IEEE) received the B.E. and M.E. degrees in automotive engineering from the Wuhan University of Technology, Wuhan, China, in 2001 and 2004, respectively, and the Ph.D. degree in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong, China, in 2007. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include deep learning and computational neuroscience. He is an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING.