Representation Disparity-aware Distillation for 3D Object Detection

Yanjing Li[1†], Sheng Xu[1†], Mingbao Lin[3], Jihao Yin[1], Baochang Zhang[1,2,4], Xianbin Cao[1*]
[1] Beihang University [2] Zhongguancun Laboratory [3] Tencent
[4] Nanchang Institute of Technology

Abstract

In this paper, we focus on developing knowledge distillation (KD) for compact 3D detectors. We observe that off-the-shelf KD methods manifest their efficacy only when the teacher model and student counterpart share similar intermediate feature representations. This might explain why they are less effective in building extreme-compact 3D detectors where significant representation disparity arises due primarily to the intrinsic sparsity and irregularity in 3D point clouds. This paper presents a novel representation disparity-aware distillation (RDD) method to address the representation disparity issue and reduce performance gap between compact students and over-parameterized teachers. This is accomplished by building our RDD from an innovative perspective of information bottleneck (IB), which can effectively minimize the disparity of proposal region pairs from student and teacher in features and logits. Extensive experiments are performed to demonstrate the superiority of our RDD over existing KD methods. For example, our RDD increases mAP of CP-Voxel-S to 57.1% on nuScenes dataset, which even surpasses teacher performance while taking up only 42% FLOPs.

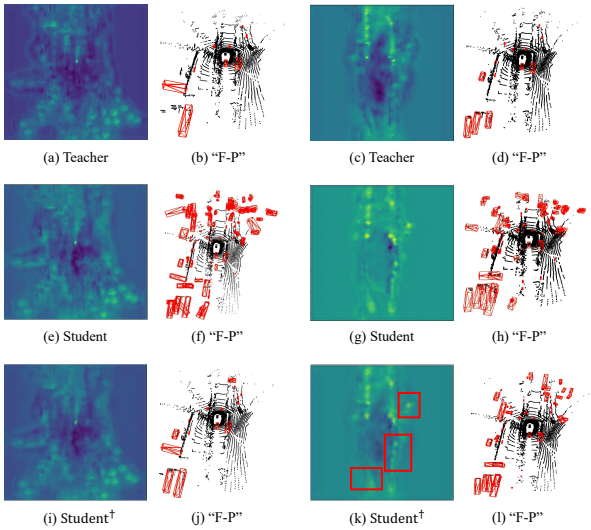


Figure 1. Visualization of intermediate neck features from teacher CP-Voxel [44] and student CP-Voxel-XXS [41]. Student† denotes CP-Voxel-XXS distilled by [41]. "F-P" denotes false positive predictions from the detector. The second and fourth column images show false positives on the original inputs where the red boxes denote false positives from the detector. Off-the-shelf implementation fails to tackle false positives if significant disparity exists between teacher (c) and student (g) feature maps.

1. Introduction

3D object detection in point clouds [25, 40, 29] is a fundamental perception task with broad applications on autonomous driving, robotics and smart city, *etc*. Beneficial from the large-scale 3D perception datasets [8, 2, 32] as well as advanced point [25], pillar [18, 37] and voxel based [9, 21] representations of sparse and irregular LiDAR point cloud scenes, 3D detection has achieved remarkable progress [28, 44]. Unfortunately, stronger performance is often accompanied with heavier computation burden, therefore the adoption in real-world applications still remains a challenging problem.

Recent attempts to improve efficiency focus on developing specified architectures for point-based 3D object detectors [5, 46], not generalizable to a wide spectrum of pillar/voxel-based methods [47, 18, 28, 44, 7, 40]. Here, we aim at a model-agnostic framework for obtaining efficient and accurate 3D object detectors with knowledge distillation (KD). Due to its effectiveness, generality and simplicity, KD has become a popular strategy to develop efficient models in a variety of 2D tasks [12, 21, 6, 13], which improves the performance of a lightweight student model by harvesting knowledge from an accurate yet computationally heavy teacher model.

The recent art [41] employs pivotal position logit KD to enhance the performance of compact 3D detectors. However, as we analyze in this paper, the intrinsic representation

† Equal contribution.
* Corresponding author: xbcao@buaa.edu.cn
[1] Code: https://github.com/YanjingLi0202/RDD

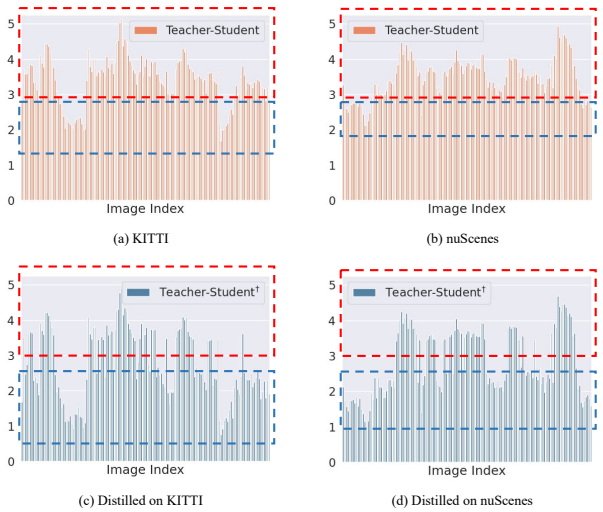


Figure 2. Histogram of *mean square distance* between feature maps of CP-Voxel & CP-Voxel-XXS (orange), and CP-Voxel & distilled CP-Voxel-XXS-PP (blue) on KITTI [8] and nuScenes [2].

disparity, stemming from region distances between compact students and pre-trained teachers, is crucial to 3D student detectors and solely neglected in existing methods [41].

For an in-depth analysis, in Fig. 1, we visualize the feature maps and predictions of pre-trained heavy CP-Voxel [44], a state-of-the-art 3D LiDAR-based detector, and these of compact CP-Voxel-XXS and its distilled version by [41]. The visual results indicate that off-the-shelf KD methods manifest their efficacy only when the teacher model and student counterpart share similar feature maps as in Fig. 1a & Fig. 1e. Otherwise, false positives are overmuch if significant representation disparity arises as like Fig. 1c & Fig. 1g, which greatly deteriorate the performance of compact 3D detectors. For a comprehensive verification, in Fig. 2, we calculate the *mean square distance* between feature maps of teacher CP-Voxel and student CP-Voxel-XXS as a metric to reflect if teacher knowledge can be well transferred to student. It is intuitive that a large distance indicates a higher representation disparity. We perform upon two datasets including KITTI [8] and nuScenes [2]. Fig. 2 manifests the statistics where the Fig. 2a & Fig. 2b represents distance histogram between teacher and vanilla student while the Fig. 2c & Fig. 2d represents distance histogram between teacher and distilled student. Each histogram can be separated into 1) the first one in compliance with small distance (blue area) and 2) the second one in line with large distance (orange area). We observe that the small-distance one is further reduced after distilling, which indicates an efficient distillation. While the large-distance one almost remains unchangeable, which indicates an inefficient distillation. Therefore, it remains an open issue to tackle the representation disparity in existing methods.

Therefore, in this paper we propose a novel 3D detector oriented representation disparity-aware distillation (RDD) method to address the above issue and reduce performance gap between compact students and over-parameterized teachers. Framework of our RDD is illustrated in Fig. 3, where the distillation objective is actually formulated under the principle of information bottleneck (IB) to maximize the mutual information between intermediate features of teacher and students. To this end, for each region proposal in teacher (student) model, our RDD first pair it by cropping a counterpart region in the same location of student (teacher) model. We measure representation disparity in each pair with mutual information under the IB framework and then learn to weight the region pairs to better bilaterally transfer information between teacher and student. In contrast to off-the-shelf pivotal position logit KD or simply involving ground truths [36, 41], the weighted information is transferred by a feature-level representation disparity-aware distillation as well as logit-level representation disparity-aware distillation loss.

We compare our RDD against state-of-the-art 2D and 3D KD methods [6, 34, 24, 41, 45] on datasets of KITTI [8] and large-scale nuScenes [2]. Extensive results reveal that our method outperforms the others by a considerable margin. For instance, on nuScenes, the CP-Voxel-S [41] distilled by our RDD obtains 57.1% mAP with only 42% FLOPs of CP-Voxel [44], achieving a new state-of-the-art.

2. Related work

3D LiDAR-based Object Detection targets to localize and classify 3D objects from point clouds. Point-based methods [29, 5, 43, 46] take raw point clouds and leverage PointNet++ [25] to extract sparse point features and generate point-wise 3D proposals. Pillar-based works [18, 37] finish voxelization in bird eye's view and extract pillar-wise features with PointNet++. Voxel-based methods [47, 40, 28] voxelize point clouds and obtain voxel-wise features with 3D sparse convolutional networks, which has become one of the most popular data treatment. Besides, range-based works [1, 33] were proposed for long-range and fast detection. Recently, designing efficient 3D detectors has drawn some attentions [5, 46] with raw point data treatment. In this work, we focus on exploring model-agnostic knowledge distillation methods to boost the performance of lightweight 3D detectors.

Knowledge Distillation aims to transfer knowledge from a large teacher model to a lightweight student network, which has become a thriving area in efficient deep learning. The simple-yet-common used KD method [12] distills knowledge between teacher and student on the output prediction logits. Another line of research proposed to help student's optimization with hints stored in informative intermediate features from teacher [27, 14, 17, 11, 16, 3].

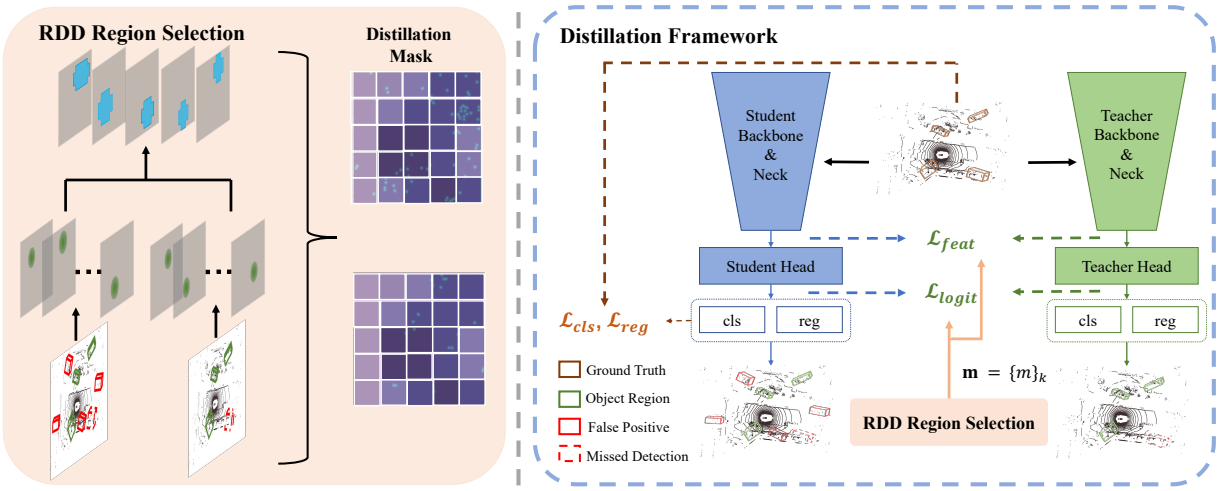


Figure 3. Overview of the proposed representation disparity-aware distillation (RDD) framework. We first select representative region pairs based on the representation disparity and formulate a balanced mask for each region pair. Then we distill student in both feature-level and logit-level to effectively eliminate the representation disparity.

In 2D detection task, some works attempt distillation techniques [19, 36, 6, 26, 42, 24] by emphasizing instance-wise distillation and feature knowledge. Mimic [19], FG [36] and GID [6] sample local region features with box proposals or custom indicators for foreground-aware feature imitation. Label KD [24] utilizes teacher's information for label assignment of student. Moreover, recent methods have also been proposed to distill the 3D LiDAR-based detectors. PointDistiller [45] captures and makes usage of the semantic information in the local geometric structure of point clouds for compressing the student. [41] leverages cues in teacher prediction to determine the important areas for distillation. Nevertheless, existing 3D LiDAR-based KD only leverage the information from well-trained teachers but neglect the distillation-needed areas in the students. In the contrary, in this work, we propose an enhanced 3D detection KD method which takes into consideration representation disparity and effectively transfer the comprehensive information from well-trained teachers to compact students.

3. Representation Disparity-aware Distillation

Fig. 3 illustrates the framework of our RDD for 3D object detection. To complete our landscape, in Sec. 3.1, we first model our distillation objective under the principle of information bottleneck (IB) [30, 39, 35]. Then, Sec. 3.2 depicts the generation of region proposal pairs between teacher and student, and representation disparity in each pair. Lastly, we detail our distillation losses including feature-level representation disparity-aware distillation and logit-level representation disparity-aware distillation.

3.1. Knowledge Distillation Objective

Usually, knowledge distillation involves a to-be-trained student detector and a pre-trained teacher detector, and we distinguish them with scripts $\mathcal{S}$ and $\mathcal{T}$, respectively. We start with a novel perspective of information bottleneck (IB) principle [30] to explore KD in 3D detectors. As discussed in [39], IB is commensurate with the best compression hypothesis that advocates minimizing data misfitting and model complexity in concert such that the task-irrelevant information can be well diminished in the compressed model for a better performance. Considering the facts in 3D object detection that the point clouds are overwhelmingly sparse and the extreme imbalance keeps going between informative instances and redundant background, an efficient information extraction is therefore particularly important.

Given a set of point clouds X, KD objective in the IB principle is written as:

$$\min_{\theta_{\boldsymbol{B}}^{\mathcal{S}}, \theta_{\boldsymbol{D}}^{\mathcal{S}}} [I(X; f^{\mathcal{S}}) - \delta I(f^{\mathcal{S}}; y^{GT})] - \beta I(f^{\mathcal{S}}; f^{\mathcal{T}}), \quad (1)$$

where $f^{\mathcal{T}}$ and $f^{\mathcal{S}}$ are the high-level feature maps from the neck/backbone of the teacher and the student, respectively. y^{GT} denotes ground-truth. $\theta_{\boldsymbol{B}}^{\mathcal{S}}$ and $\theta_{\boldsymbol{D}}^{\mathcal{S}}$ are the parameters of backbone and detection part in student respectively. Meanwhile, δ, β are Lagrange multipliers [30]. $I()$ returns the mutual information between its two input variables. With $I(f^{\mathcal{S}}; y^{GT})$ maximizing the mutual information between the features and ground-truth, the first item $I(X; f^{\mathcal{S}})$ minimizes the mutual information between the point cloud data and the high-level feature maps of student to control the noise introduction. This part can be treated as the original detection loss of the detector [44, 18]. With teacher model's features as guidance, the second item $\beta I(f^{\mathcal{S}}; f^{\mathcal{T}})$

maximizes the mutual information to preserve more teacher information in student. The collective cooperation between the two items guides student to focus more on beneficial information and less on noise information [35, 39, 30].

3.2. Region Pairs with Representation Disparity

Region Pairs. We denote $\{R_i^{\mathcal{T}}|(p_{reg,i}^{\mathcal{T}}, p_{cls,i}^{\mathcal{T}})\}_{i=1}^{M}$ and $\{R_i^{\mathcal{S}}|(p_{reg,i}^{\mathcal{S}}, p_{cls,i}^{\mathcal{S}})\}_{i=M+1}^{M+N}$ as the outputs of teacher and student where the i-th region proposal $R_i^{\mathcal{T}}/R_i^{\mathcal{S}} \in \mathbb{R}^{C\times H\times W}$ contains two information including a regression coordinate $p_{reg,i}$ to model the proposal position and a classification probability $p_{cls,i}$ to tell the proposal category. Note that in the center-based 3D detectors, *e.g.*, CenterPoints [44], each proposal corresponds to a Gaussian area in the heatmaps, while in the anchor-based 3D detectors, *e.g.*, PointPillars [18], each proposal corresponds to a region in the intermediate features.

As shown in the left of Fig. 4, the recent study [41] unilaterally passes on teacher proposals of higher classification probability to the corresponding regions of student. On the one hand, it ignores the efficacy of student information; on the other hand, the representation disparity is neglected as discussed in Sec. 1. In this paper, we propose to bilaterally transfer information between teacher and student. Specifically, as depicted in the right of Fig. 4, for each single region in teacher (student) model, we crops a counterpart feature map patch in the same location of student (teacher) model to form a total of $M+N$ region pairs $\{(R_i^{\mathcal{T}}, R_i^{\mathcal{S}})\}_{i=1}^{M+N}$. Here, $\{R_i^{\mathcal{S}}\}_{i=1}^{M}$ accords with cropped region patches in student and $\{R_i^{\mathcal{T}}\}_{i=M+1}^{M+N}$ is in tune with these patches in teacher. For ease of representation, the superscripts $\mathcal{T}$ and $\mathcal{S}$ will be dropped from time to time in the following contents.

Then, our distillation considers representation disparity issue by weighting the patch-level distance under the IB principle. Before diving into details, we first channel-wise normalize proposal R_i considering the large-scale magnitude gap between the pre-trained teacher and the to-be-trained student, as:

$$\hat{R}_{i;c,:,:} = \frac{\exp(\frac{R_{i;c,:,:}}{\tau})}{\sum_{c'\in\{1,2,...,C\}}\exp(\frac{R_{i;c',:,:}}{\tau})}, \tag{2}$$

where $R_{i;c,:,:}$ denotes the c-th channel of R_i and $\tau = 4$ in this paper denotes a hyper-parameter controlling the statistical attributions of the channel-wise alignment operation

Representation Disparity. Under the framework of IB principle, we define and evaluate representation disparity as mutual information between student patch $\hat{R}_i^{\mathcal{S}}$ and teacher patch $\hat{R}_i^{\mathcal{T}}$. This is formulated as:

$$I(\hat{R}_i^{\mathcal{S}}; \hat{R}_i^{\mathcal{T}}) = H(\hat{R}_i^{\mathcal{S}}) - H(\hat{R}_i^{\mathcal{S}}|\hat{R}_i^{\mathcal{T}}), \tag{3}$$

where $H()$ returns the information entropy. A smaller $I(\hat{R}_i^{\mathcal{S}}; \hat{R}_i^{\mathcal{T}})$ indicates higher disparity between $\hat{R}_i^{\mathcal{S}}$ and $\hat{R}_i^{\mathcal{T}}$.

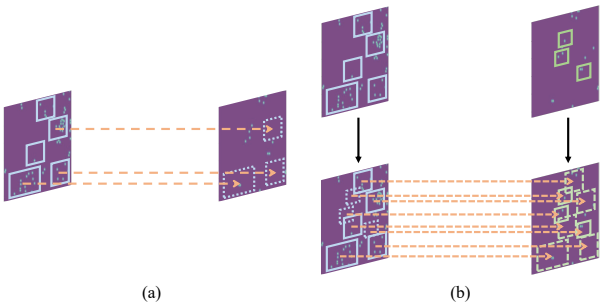


Figure 4. Illustration for the generation of the region pairs. Each single region in one model generates a counterpart feature map patch in the same location of the other model.

Then, a naive way to measure mutual information between teacher and student $I(f^{\mathcal{S}}; f^{\mathcal{T}})$ in the KD objective of Eq. (1) is to sum up mutual information of all patch pairs as:

$$I(\hat{R}_i^{\mathcal{S}}; \hat{R}_i^{\mathcal{T}}) = \sum_{i=1}^{M+N} m_i I(\hat{R}_i^{\mathcal{S}}; \hat{R}_i^{\mathcal{T}}). \tag{4}$$

Nevertheless, it does not take into account the representation disparity among different region pairs. Given this, prior to a formal distillation, we propose to optimize a weighting vector $\boldsymbol{m} = [m_1, m_2, ..., m_{M+N}] \in \mathbb{R}^{M+N}$ to identify disparate region pairs, leading to our learning objective as:

$$\min_{\mathbf{m}} \underbrace{\sum_{i=1}^{M+N} m_i I(\hat{R}_i^{\mathcal{S}}; \hat{R}_i^{\mathcal{T}})}_{I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m})} + \lambda\|\mathbf{m}\|_1, \tag{5}$$

where the minimization leads $m_i \in \boldsymbol{m}$ to be large to penalize high disparity stemming from pair $\hat{R}_i^{\mathcal{S}}$ and $\hat{R}_i^{\mathcal{T}}$. The term $\min_{\mathbf{m}} \|\mathbf{m}\|_1$ is involved to prevent the model to equally distill all alternative region pairs. Also, to indicate the disparity degree. we clip m_i to 1 or 0 if its value is beyond $[0, 1]$. $\lambda > 0$ is a hyper-parameter to determine the sparsity of $\boldsymbol{m}$.

Moreover, the introduction of $\boldsymbol{m}$ leads our format of $I(f^{\mathcal{S}}; f^{\mathcal{T}})$ to $I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m})$ as manifested in Eq. (5). Together with our representation disparity, the KD objective under IB framework finally changes from Eq. (1) to:

$$\begin{aligned} &\min_{\theta_{\boldsymbol{B}}^{\mathcal{S}}, \theta_{\boldsymbol{D}}^{\mathcal{S}}} [I(X; f^{\mathcal{S}}) - \delta I(f^{\mathcal{S}}; y^{GT})] - \beta I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m}^*), \\ &\text{s.t. } \mathbf{m}^* = \arg\min_{\mathbf{m}} I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m}) + \lambda\|\boldsymbol{m}\|_1. \end{aligned} \tag{6}$$

Our objective involve two sub-problems. In each training iteration, we first perform inner-level optimization to derive a current optimal $\boldsymbol{m}^*$; and then solves the upper-level optimization to conduct distillation based on explicit distillation losses in Sec. 3.3. Notice the inner-level optimization causes negligible costs compared to the upper-level one

since the size of region pairs is not large. On the contrary, it derives distillation to focus more on disparate region pairs for a better performance.

3.3. Knowledge Transferring

We present the upper-level optimization in Eq. (6). Recall in Sec. 3.1 we analyze that the first item $[I(X; f^{\mathcal{S}}) - \delta I(f^{\mathcal{S}}; y^{GT})]$ is in compliance with the original detection loss such as proposal classification and coordinate regression. Our central is to hand over the specific format of the second term $I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m}^*)$, which according to Eq. (3) and Eq. (5) can be explicitly derived as:

$$I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m}^*) = \sum_{i=1}^{M+N} m_i^*(H(\hat{R}_i^{\mathcal{S}}) - H(\hat{R}_i^{\mathcal{S}}|\hat{R}_i^{\mathcal{T}})). \quad (7)$$

Considering the intrinsic entropy of $\hat{R}_i^{\mathcal{S}}$ remains unchanged within each iteration, therefore $H(\hat{R}_i^{\mathcal{S}})$ is regarded as a constant. Maximizing $I(f^{\mathcal{S}}; f^{\mathcal{T}}|\boldsymbol{m}^*)$ turns to minimizing $H(\hat{R}_i^{\mathcal{S}}|\hat{R}_i^{\mathcal{T}})$. Unfortunately, it is hard to directly minimize $H(\hat{R}_i^{\mathcal{S}}|\hat{R}_i^{\mathcal{T}})$. Instead, we choose to minimize norm distance between $\hat{R}_i^{\mathcal{S}}$ and $\hat{R}_i^{\mathcal{T}}$ as a substitute since both of them reach optimal when $\hat{R}_i^{\mathcal{S}} = \hat{R}_i^{\mathcal{T}}$.

In view of that feature pyramid network (FPN) [20] has been adopted in most 3D detectors for robustness of multi-scale detection [44, 18, 40], it is natural to choose the neck feature maps after FPN for distillation. After forming the region pairs, the feature-level representation disparity-aware distillation loss is computed as:

$$\mathcal{L}_{feat} = \frac{1}{M+N} \sum_{i=1}^{M+N} m_i^* \|\varphi(\psi(\hat{R}_i^{\mathcal{S}})) - \hat{R}_i^{\mathcal{T}}\|_2, \quad (8)$$

where φ indicates the RoI Align [10]. ψ is 1×1 convolution followed by batch normalization [15] and ReLU [23] to align channel-wise discrepancy between teacher region $\hat{R}_i^{\mathcal{T}}$ and and student region $\hat{R}_i^{\mathcal{S}}$ [6, 36].

In addition to FPN outputs, another neglected special kind of feature maps come to the logits in the classification and regression branches. Previous 2D and 3D methods [41, 36, 12] conduct distillation on the whole or pivotal part of the detection head outputs or on the conventional feature maps, degrading the student performance [6]. The probable cause can be attributed to the extreme imbalance between instances and backgrounds in 3D LiDAR-based detection task. Therefore, based on the weighting vector $\boldsymbol{m}$, we also take into consideration the regression coordinate and classification probability of each region proposal $\hat{R}_i$, and form our logit-level representation disparity-aware distillation loss as:

$$\mathcal{L}_{logit} = \frac{1}{M+N} \sum_{i=1}^{M+N} m_i^* (\|p_{cls,i}^{\mathcal{S}} - p_{cls,i}^{\mathcal{T}}\|_1 + \|(p_{reg,i}^{\mathcal{S}} - p_{reg,i}^{\mathcal{T}})\|_1), \quad (9)$$

in which $p_{cls,i}$ and $p_{reg,i}$ denote the classification probability and proposal coordinates for the i-th region proposal R_i as introduced in Sec. 3.2. It should be noted that, for the center-based 3D detector [44], the regression loss item do not exist.

Finally, the exact definition of our KD objective in Eq. (1) is given as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{cls} + \gamma \mathcal{L}_{reg}}_{I(X;f^{\mathcal{S}}) - \delta I(f^{\mathcal{S}}; y^{GT})} + \underbrace{\alpha_1 \mathcal{L}_{feat} + \alpha_2 \mathcal{L}_{logit}}_{-\beta I(f^{\mathcal{S}}; f^{\mathcal{T}})}, \quad (10)$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ are the original detection losses supervised by the ground-truth labels. The γ, α_1 and α_2 are trade-off parameters between different objectives.

4. Experiments

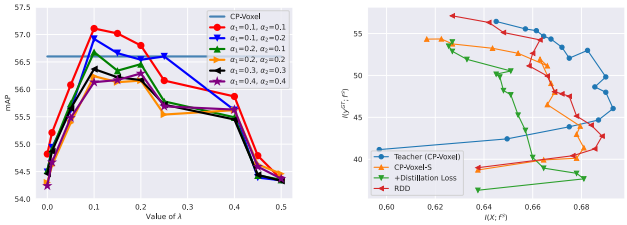
Comprehensive experiments are conducted to evaluate our proposed method on two datasets for object detection: nuScenes [2] and KITTI [8]. First, we introduce the datasets, metrics, implementation details and compact model architecture in Sec. 4.1. Then we select the hyper-parameters, validate the effectiveness of the components, and analyze the information of our method through ablation studies in Sec. 4.2. Finally, in Sec. 4.3 and Sec. 4.4, we compare our method with other image-based 2D distillation methods implemented on 3D detectors, and other 3D distillation methods to demonstrate the superiority of RDD.

4.1. Experimental Settings

Datasets and Evaluation Metrics. For nuScenes dataset [2], metrics are mean average precision (mAP) and the nuScenes detection score (NDS). These metrics are computed in the physical unit. For KITTI dataset [8], we report the average precision calculated by 40 sampling recall positions for 3D object detection on the validation split. Following the typical protocol, the IoU threshold is set as 0.7 for class Car and 0.5 for class Pedestrians and Cyclists.

Training & **Validation**. For experiments conducted on nuScenes dataset, we follow the same setups as the original CenterPoint [44]. We use AdamW [22] to train the model. The weight decay for AdamW is $1e-2$. Following a cyclic schedule [31], the learning rate is initially $1e-4$ and gradually increased to $1e-3$, and finally decreased to $1e-8$. We train for 20 epochs on 8 NVIDIA Tesla V100 GPUs. During inference, we take the top 100 highest-scored objects as the final predictions. We do not use any post-processing such as non maximum suppression (NMS). We use the toolkit provided with the nuScenes dataset for evaluation.

For experiments conducted on KITTI dataset, we use AdamW [22] to train the model. The weight decay for AdamW is $1e-2$. Following a cyclic schedule, the learning rate is initially $1e-4$ and gradually increased to $1e-3$, which is finally decreased to $1e-8$. We train for 80 epochs



(a) Effect of λ, α_1 and α_2 (b) Mutual information curves

Figure 5. (a) Influence of λ, α_1 and α_1 using CP-Voxel-S [41] on nuScenes [2]. (b) The mutual information curves of $I(X;\mathbf{E})$ and $I(\boldsymbol{y}^{GT};\mathbf{E},\mathbf{q})$ (Eq. 1) on the information plane. The red curves represent the teacher CP-Voxel [44]. The blue, orange, green, and red lines represent the teacher (CP-Voxel [44]), the vanilla student (CP-Voxel-S [41]), the vanilla student with distillation loss, and CP-Voxel-S trained with our RDD.

Table 1. Effects of different components in RDD with CP-Voxel-S and CP-Pillar-v0.4 [41] on nuScenes [2]. All experiments use ℓ_2 loss for distillation. Teacher models are marked in grey.

Model	Region Selection	mAP	NDS
CP-Voxel	-	56.6	64.7
CP-Voxel-S	-	54.0	62.1
	Hint	52.9	61.1
	FG	54.3	62.6
	RDD	**57.1**	**65.0**
CP-Pillar	-	49.1	59.7
CP-Pillar-v0.4	-	46.5	55.5
	Hint	45.6	55.1
	FG	47.6	57.2
	RDD	**50.0**	**58.9**

on 8 NVIDIA Tesla V100 GPUs. We apply axis aligned non maximum suppression (NMS) with an overlap threshold of 0.5 IoU, following [18] for inference. We evaluate the model performance using the toolkit provided with the KITTI dataset.

Compact Model Architecture. We apply CenterPoint-Voxel and CenterPoint-Pillar [44] on nuScenes, and SECOND [40] and PointPillars [18] on KITTI.

For CP-Voxel and CP-Pillar [44], we follow [41] to adopt width and input resolution compression. Specifically, we compress the channels of {encoder, backbone&neck, head} (*i.e.*, {Pillar Feature Encoding (PFE), Bird eye's view Feature Encoding (BFE), Head} in [41]) into {1×, 0.5×, 0.5×}, {0.75×, 0.5×, 0.5×} and {0.5×, 0.25×, 0.25×}, forming the CP-Voxel-S, CP-Voxel-XS and CP-Voxel-XXS. We follow [41] to change the voxel size of CP-Pillar from 0.32 to 0.4, 0.48 and 0.64, forming the CP-Pillar-v0.4, CP-Pillar-v0.48 and CP-Pillar-v0.64. For SECOND [40], we also follow [41] to compress the width of SECOND [40], where the channels of {encoder, backbone&neck} (*i.e.*, {PFE, BFE} in [41]) are reduced to {0.75×, 0.5×} and {0.5×, 0.5×}, forming the SECOND-S and SECOND-XS. For PointPillars [18], we compress the width of PointPillars [18] forming the PointPillars-S and PointPillars-XS, in which the channels of {encoder, backbone&neck} are reduced to {0.75×, 0.5×} and {0.5×, 0.5×}.

Table 2. Evaluating the components of RDD based on CP-Voxel-S and CP-Pillar-v0.4. RDD-F and RDD-L denote w/ or w/o distillation loss in Eq. (8) and Eq. (9), respectively.

Student	RDD-F	RDD-L	mAP	NDS
CP-Voxel-S	-	-	54.0	62.1
	✓		56.8	64.0
		✓	57.0	64.3
	✓	✓	**57.1**	**65.0**
CP-Pillar-v0.4	-	-	46.5	55.5
	✓		49.7	58.0
		✓	49.2	57.9
	✓	✓	**50.0**	**58.9**

4.2. Ablation Study

Hyper-parameter Selection. We first select hyper-parameters λ of Eq. (5) and α_1, α_2 of Eq.(10) in this part, with experiments conducted on nuScenes dataset. Note that γ in Eq.(10) is sett following CenterPoint [44], SECOND [40], and Pointpillar [18] when conducted our expeirments on these detectors. We show the model performance (mAP) with different setups of hyper-parameters λ in Fig. 5a, in which the performances increase first and then decrease with the uplift of λ from left to right. Since λ controls the decline of $\|\mathbf{m}\|_1$ related to the proportion of selected distillation-desired region pairs, $\lambda = 0$ denotes all the alternative region pairs are equally distilled leading to a unsatisfactory performance. While with λ increasing to more than 0.1, the proportion of selected distillation-desired region pairs is not large enough for effectively eliminating the representation disparity, which also affects the performance. In conclusion, the CP-Voxel-S [41] obtains better performances with λ set as 0.1. The figure also shows that the CP-Voxel-S [41] obtains best performances with α_1 and α_2 set as 0.2 and 0.2. However, enlarging α_1 and α_2 degenerates the performance of the detectors. Based on the ablative study above, we set hyper-parameters $\{\lambda, \alpha_1, \alpha_2\}$ as $\{0.1, 0.2, 0.2\}$ for the experiments in this paper.

Component Ablation. We first compare our representation disparity-aware (RDD) region selecting method with other methods to select regions: Hint [4] (using the neck feature without region mask) and FG [36]. We show the effectiveness of RDD with center-based 3D detectores [44] on nuScenes dataset [2] in Tab. 1. On the CP-Voxel-S [41],

the introducing of RDD achieves improvements of the mAP and NDS by $\{$ 3.1%, 4.2%, 2.8%$\}$ and $\{$2.9%, 3.9%, 2.4%$\}$ compared to non-distillation, Hint [4], and FGFI [36], under the same student-teacher framework. We also evaluate the RDD on the Pillar-based 3D detector, *i.e.*, CP-Pillar-v0.4 [41]. Our RDD selection method improves the mAP and NDS by $\{$3.5%, 4.4%, 2.4%$\}$ and $\{$3.4%, 3.8%, 1.7%$\}$ compared to non-distillation, Hint [4], and FGFI [36], supervised by the same teacher (CP-Pillar [44]).

Then we evaluate the proposed distillation losses, *i.e.*, $\mathcal{L}_{feat}$ in Eq. (8) (RDD-F) and $\mathcal{L}_{logit}$ in Eq. (9) (RDD-L), in Tab. 2. As listed, RDD-F and RDD-L improve the performance when used alone, and the two losses further boost the performance considerably when combined together. For example, the RDD-F improve the mAP of CP-Voxel-S [41] by 2.8% and the RDD-L achieves 3.0% mAP improvement. While combining the RDD-F and RDD-L together, the performance improvement achieves 3.1%. These ablative experiments further validates the effectiveness of our method.

Information analysis. We further show the information plane following [38] in Fig. 5b, where we adopt the test mAP to quantify $I(y^{GT}; f^{\mathcal{S}})$. We employ a reconstruction decoder to decode the encoded feature $f^{\mathcal{S}}$ to reconstruct the input feature of the backbone and quantify $I(X; f^{\mathcal{S}})$. Detailed settings are given in the supplementary material. As shown in Fig. 5b, the curve of the larger teacher CP-Voxel is usually on the right of the curve of small student models, which indicates greater ability of information representation. Likewise, the red line (CP-Voxel-S with RDD) is usually on the right of the three left curves, showing the information representation improvements with the proposed methods.

4.3. Results on nuScenes

We first compare our method with image-based 2D and 3D distillation methods for the task of 3D object detection on nuScenes [2]. Note that for width compressed students, we use the same input voxel size as [44], *i.e.* 0.1. We mainly discuss the mAP and NDS (default nuScenes metric) in the following. We evaluate the proposed RDD on CenterPoint detectors [44] in Tab. 3. For CP-Voxel-S, compared to non-distillation of GID-L [6], GID-F [6], FG [36] and LAD [24], our RDD boosts the performance of mAP and NDS by $\{$3.1%, 3.0%, 3.7%, 2.8%, 4.1%$\}$ and $\{$2.9%, 2.9%, 3.2%, 2.4%, 3.3%$\}$, respectively. Moreover, our RDD improves the mAP and NDS of CP-Voxel-S by 1.4% and 0.7% respectively, compared with previous state-of-the-art 3D distillation method (PP-logit-KD). It is worth noting that our RDD trained CP-Voxel-S even surpasses the teacher detectors by 0.5% mAP and 0.3% NDS but taking up only 41.6% FLOPs and 51.3% parameters of the teacher, which is a significant achievement. For CP-Voxel-XS, our RDD improves the performance of mAP and NDS by $\{$1.0%, 0.8%, 1.1%, 0.7%, 1.0%$\}$ and $\{$0.3%, 0.5%, 0.8%, 0.3%, 0.6%$\}$, compared to non-distillation of GID-L [6], GID-F [6], FG [36] and LAD [24]. In addition, our RDD improves the mAP and NDS of CP-Voxel-XS by 0.5% and 0.4%, compared with previous state-of-the-art 3D distillation method (PP Logit KD). For CP-Voxel-XXS, our RDD

Table 3. Experimental results on nuScenes [2]. # F and # P indicate float operations (FLOPs) and parameters of the detector. Teacher models are marked in gray shadow.

Detector	# F/# P (G/M)	Method	mAP (↑)	NDS (↑)
CP-Voxel	114.8 / 7.8	-	56.6	64.7
CP-Voxel -S	47.8 / 4.0	No Distill	54.0	62.1
		GID-L	54.1	62.1
		GID-F	53.4	61.8
		FG	54.3	62.6
		LAD	53.0	61.7
		PP Logit KD	55.7	64.3
		RDD	**57.1**	**65.0**
CP-Voxel -XS	36.0 / 2.8	No Distill	53.0	61.8
		GID-L	53.2	61.6
		GID-F	52.9	61.3
		FG	53.3	61.8
		LAD	53.0	61.5
		PP Logit KD	53.5	61.7
		RDD	**54.0**	**62.1**
CP-Voxel -XXS	12.0 / 1.0	No Distill	46.7	55.5
		GID-L	46.8	56.5
		GID-F	47.0	56.4
		FG	47.1	56.2
		LAD	46.6	55.3
		PP Logit KD	47.9	57.8
		RDD	**49.4**	**57.9**
CP-Pillar	333.9 / 5.2	-	49.1	59.7
CP-Pillar -v0.4	212.9 / 5.2	No Distill	46.5	55.5
		GID-L	47.3	56.4
		GID-F	47.6	56.8
		FG	47.7	57.2
		LAD	46.9	55.7
		PP Logit KD	48.6	57.5
		RDD	**50.0**	**58.9**
CP-Pillar -v0.48	149.4 / 5.2	No Distill	45.3	54.2
		GID-L	45.4	54.5
		GID-F	46.1	55.3
		FG	47.0	56.2
		LAD	45.2	54.4
		PP Logit KD	47.3	57.5
		RDD	**48.8**	**58.5**
CP-Pillar -v0.64	85.1 / 5.2	No Distill	44.0	52.3
		GID-L	44.2	52.6
		GID-F	44.4	53.9
		FG	44.7	53.7
		LAD	43.9	53.2
		PP Logit KD	45.0	55.9
		RDD	**45.8**	**56.1**

surpasses non-distillation of GID-L [6], GID-F [6], FG [36] and LAD [24] by {2.7%, 2.6%, 2.4%, 2.3%, 2.8%} mAP and {2.4%, 1.4%, 1.5%, 1.7%, 1.6%} NDS. Moreover, our RDD boosts the mAP and NDS of CP-Voxel-XXS by 1.5% and 0.1% respectively, compared with previous state-of-the-art 3D distillation method (PP Logit KD). Above experiments well validates the effectiveness of our method.

Besides, our method generates convincing results on CP-Pillar based detectors [44]. As shown in the 24-th to 45-th rows of Tab. 3, the performance of the proposed RDD with CP-Pillar-v0.4, CP-Pillar-v0.48 and CP-Pillar-v0.64 outperforms the non-distillation baseline by {3.5%, 3.5%, 1.8%} and {3.4%, 4.3%, 3.8%} on mAP and NDS, a large margin. Compared with previous state-of-the-art 3D distillation methods, our RDD achieves {1.4%, 1.5%, 0.8%} and {1.4%, 1.0%, 0.2%} improvement on mAP and NAS respectively with CP-Pillar-v0.4, CP-Pillar-v0.48 and CP-Pillar-v0.64, which well validates the efficacy of our method. The above experimental results prove the superiority of our RDD method on both Voxel-based and Pillar-based CenterPoint [44].

4.4. Results on KITTI

We further show that our RDD can generalize well to KITTI [8] dataset with anchor-based detectors SECOND [40] and PointPillar [18]. As shown in Tab. 4, for SECOND-S, compared to non-distillation, GID-L [6], GID-F [6], FG [36] and LAD [24], our RDD boosts the performance of moderate mAP@R40 by {2.6%, 1.9%, 1.4%, 1.6%, 1.2%}. Moreover, our RDD improves the moderate mAP@R40 of SECOND-S by 0.4% and 0.5%, compared with PointDistiller [45] and PP Logit KD [41]. And SECOND-S trained with our RDD even surpasses its teacher model by 1.0%. And for SECOND-XS, our RDD surpasses previous image-based 2D distillation methods GID-L [6], GID-F [6], FG [36] and LAD [24] by by {2.8%, 2.0%, 1.8%, 1.7%, 1.3%} in moderate mAP@R40. SECOND-XS trained with our RDD also surpasses previous 3D distillation method, PoinrDistiller and PP Logit KD, by 0.7% and 0.6% on the moderate mAP@R40, which well validates the effectiveness of our method. Moreover, our method can also be generalized to PointPillars [18]. As shown in the 20-th to 36-th rows of Tab. 4, the performance of the proposed RDD with PointPillars-S and PointPillars-XS outperforms the non-distillation baseline by 4.5% and 2.0% on the moderate mAP@R40, a large margin. Compared with PoinrDistiller, our RDD achieves 1.7% and 0.9% improvement on the moderate mAP@R40 with PointPillars-S and PointPillars-XS. Our RDD also surpasses PP Logit KD by 1.8% and 0.7% improvement on the moderate mAP@R40 with PointPillars-S and PointPillars-XS, which well validates the efficacy of our method. The above results are of great significance in the 3D LiDAR-based object detection.

Table 4. Experimental results for 3D detection on KITTI [8]. # F and # P indicate float operations (FLOPs) and parameters of the detector. Teacher models are marked in gray shadow.

Detector	# F (G)	# P (M)	Method	3D Moderate mAP@R40(↑)
SECOND	80.5	5.3	-	67.2
SECOND-S	23.0	1.6	No Distill	65.6
			GID-L	66.3
			GID-F	66.8
			FG	66.6
			LAD	67.0
			PointDistiller	67.8
			PP Logit KD	67.7
			RDD	**68.2**
SECOND-XS	20.5	1.4	No Distill	64.2
			GID-L	65.0
			GID-F	65.2
			FG	65.3
			LAD	65.7
			PointDistiller	66.3
			PP Logit KD	66.4
			RDD	**67.0**
PointPillars	34.3	4.8	-	60.3
PointPillars-S	9.8	1.5	No Distill	58.6
			GID-L	58.9
			GID-F	59.1
			FG	59.4
			LAD	59.2
			PointDistiller	62.3
			PP Logit KD	62.2
			RDD	**63.0**
PointPillars-XS	8.7	1.3	No Distill	58.9
			GID-L	59.2
			GID-F	59.6
			FG	59.4
			LAD	59.7
			PointDistiller	60.0
			PP Logit KD	60.2
			RDD	**60.9**

5. Conclusion

This paper presents a novel method for training compact 3D LiDAR-based detectors with knowledge distillation to eliminate the representation disparity (RDD). RDD employs a information bottleneck (IB) principle to select the regions with maximum representation disparity and proposes effective distillation losses to supervise the representation disparity. As a result, our RDD significantly boosts the performance of compact 3D detectors. Extensive experiments show that RDD surpasses state-of-the-art compact 3D detectors and other knowledge distillation methods in 3D LiDAR-based object detection.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62076016, under Grant 61827901, "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268, Foundation of China Energy Project GJNY-19-90.

References

[1] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. *arXiv preprint arXiv:2005.09927*, 2020. 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. of CVPR*, pages 11621–11631, 2020. 1, 2, 5, 6, 7

[3] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proc. of AAAI*, pages 7028–7036, 2021. 2

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proc. of NeurIPS*, 2017. 6, 7

[5] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *Proc. of ICCV*, pages 9775–9784, 2019. 1, 2

[6] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proc. of CVPR*, pages 7842–7851, 2021. 1, 2, 3, 5, 7, 8

[7] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proc. of AAAI*, pages 1201–1209, 2021. 1

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. of CVPR*, pages 3354–3361, 2012. 1, 2, 5, 8

[9] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, pages 1–10, 2017. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of ICCV*, pages 2961–2969, 2017. 5

[11] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proc. of AAAI*, pages 3779–3787, 2019. 2

[12] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 5

[13] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proc. of ICCV*, pages 1013–1021, 2019. 1

[14] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 2

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*, pages 448–456, 2015. 5

[16] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *Proc. of ICCV*, pages 1345–1354, 2019. 2

[17] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *Proc. of ICLR*, pages 1–13, 2017. 2

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proc. of CVPR*, pages 12697–12705, 2019. 1, 2, 3, 4, 5, 6, 8

[19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proc. of CVPR*, pages 6356–6364, 2017. 3

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. of CVPR*, pages 2117–2125, 2017. 5

[21] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proc. of CVPR*, pages 2604–2613, 2019. 1

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of ICLR*, pages 1–18, 2017. 5

[23] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of ICML*, pages 1–8, 2010. 5

[24] Chuong H Nguyen, Thuy C Nguyen, Tuan N Tang, and Nam LH Phan. Improving object detection by label assignment distillation. In *Proc. of WACV*, pages 1005–1014, 2022. 2, 3, 7, 8

[25] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proc. of NeurIPS*, page 5099–5108, 2017. 1, 2

[26] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *Proc. of CVPR*, pages 14443–14453, 2021. 3

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[28] 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 *Proc. of CVPR*, pages 10529–10538, 2020. 1, 2

[29] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proc. of CVPR*, pages 770–779, 2019. 1, 2

[30] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 3, 4

[31] Leslie N Smith. Cyclical learning rates for training neural networks. In *Proc. of WACV*, pages 464–472. IEEE, 2017. 5

[32] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. of CVPR*, pages 2446–2454, 2020. 1

[33] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *Proc. of CVPR*, pages 5725–5734, 2021. 2

[34] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020. 2

[35] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. *arXiv preprint arXiv:2210.06458*, 2022. 3, 4

[36] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proc. of CVPR*, pages 4933–4942, 2019. 2, 3, 5, 6, 7, 8

[37] Yue Wang, Alireza Fathi, Abhijit Kundu, David A Ross, Caroline Pantofaru, Tom Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *Proc. of ECCV*, pages 18–34, 2020. 1, 2

[38] Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Revisiting locally supervised learning: an alternative to end-to-end training. In *Proc. of ICLR*, pages 1–21, 2021. 7

[39] Ziwei Wang, Ziyi Wu, Jiwen Lu, and Jie Zhou. Bidet: An efficient binarized object detector. In *Proc. of CVPR*, pages 2049–2058, 2020. 3, 4

[40] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2, 5, 6, 8

[41] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. In *Proc. of NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[42] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proc. of CVPR*, pages 4643–4652, 2022. 3

[43] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proc. of ICCV*, pages 1951–1960, 2019. 2

[44] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proc. of CVPR*, pages 11784–11793, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[45] Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. *arXiv preprint arXiv:2205.11098*, 2022. 2, 3, 8

[46] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proc. of CVPR*, pages 18953–18962, 2022. 1, 2

[47] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proc. of CVPR*, pages 4490–4499, 2018. 1, 2