Boosting 3D Object Detection by Simulating Multimodality on Point Clouds

Wu Zheng[1] Mingxuan Hong[1,2] Li Jiang[3] Chi-Wing Fu[1,2]
[1]Department of Computer Science and Engineering, CUHK
[2]Shun Hing Institute of Advanced Engineering, CUHK [3]Max Planck Institute
{wuzheng, cwfu}@cse.cuhk.edu.hk lijiang@mpi-inf.mpg.de

Abstract

This paper presents a new approach to boost a single-modality (LiDAR) 3D object detector by teaching it to simulate features and responses that follow a multi-modality (LiDAR-image) detector. The approach needs LiDAR-image data only when training the single-modality detector, and once well-trained, it only needs LiDAR data at inference. We design a novel framework to realize the approach: response distillation to focus on the crucial response samples and avoid most background samples; sparse-voxel distillation to learn voxel semantics and relations from the estimated crucial voxels; a fine-grained voxel-to-point distillation to better attend to features of small and distant objects; and instance distillation to further enhance the deep-feature consistency. Experimental results on the nuScenes dataset show that our approach outperforms all SOTA LiDAR-only 3D detectors and even surpasses the baseline LiDAR-image detector on the key NDS metric, filling ∼72% mAP gap between the single- and multi-modality detectors.

1. Introduction

State-of-the-art 3D object detectors, *e.g.*, [11, 33, 48, 51, 54], widely adopt LiDAR-produced point clouds as the major input modality, since point clouds offer precise depth information and are robust to varying weather condition and illumination. Yet, due to the laser-ray divergence, the sparsity of point clouds increases with distance. So, there are only few points in small and distant objects, making it very hard to predict their object boundaries and semantic classes.

On the other hand, camera-produced images are a popular modality for monocular and stereo 3D object detection [7, 16, 21, 22, 31]. As images offer clear appearance and texture with dense pixels, image-based detectors can easily recognize the object boundaries and classify even small and distant objects. Yet, images have no depth information and the visibility of objects depends on the environment conditions, *e.g.*, lighting, so image-based detectors usually cannot be as accurate and robust as the LiDAR-based ones.

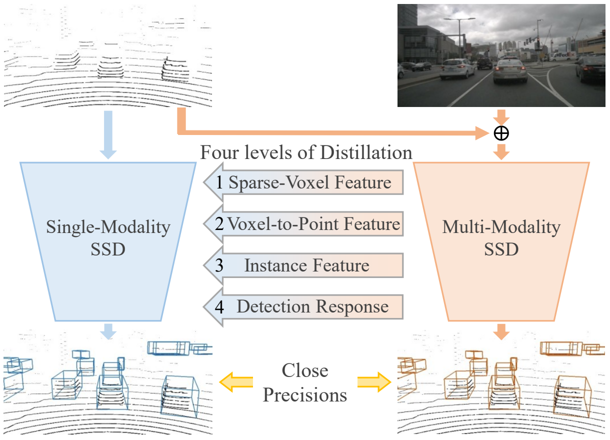

Figure 1. Overview of our Simulated Single-to-Multi-Modality Single-Stage 3D object Detector (S2M2-SSD) framework, by which we can train a single-modality SSD to learn from a multi-modality SSD and to achieve a high precision close to the multi-modality SSD but with only single-modality input at inference.

Some recent works [17, 27, 41, 44, 52] started to explore the fusion of 3D point clouds and 2D RGB images for improving the feature quality. While higher precisions can often be attained, multi-modality detectors unavoidably sacrifice the inference efficiency for processing the extra modality. Also, it is tedious to calibrate and synchronize different sensors, spatially and temporally, for high-quality data fusion. Last, a breakdown of any modality sensor will cause a detector failure, thus reducing the system's fault tolerance.

In this work, we propose *to teach a single-modality network to produce simulated multi-modality features and responses from only LiDAR input* by training the network to learn from a multi-modality LiDAR-image detector. Our approach needs multi-modality data only when training the single-modality network, and once it is well-trained, it can detect 3D objects without image inputs. This approach perfectly meets the sheer practical need of autonomous driving and boosts single-modality 3D object detection for (i) *high efficiency*, since our approach needs to process only LiDAR data at inference; (ii) *high precision*, since our network outperforms the SOTA LiDAR-only detectors; and

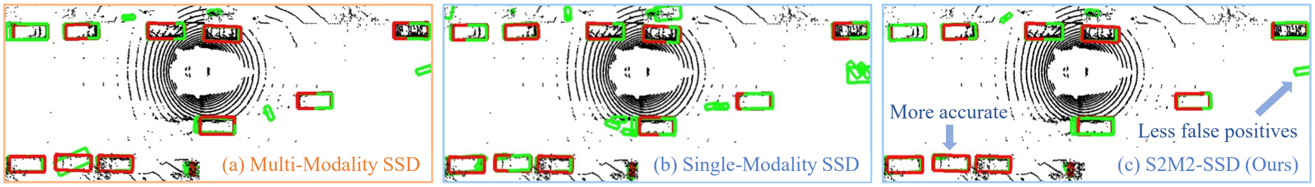

Figure 2. Even with only LiDAR input, our S2M2-SSD (c) is able to predict accurate bounding boxes (green) that well match the ground truths (red). Its precision is close to that of the multi-modality SSD (a) and outperforms the SOTA LiDAR-only SSD [25] (b).

(iii) *high robustness*, since our approach is capable of simulating LiDAR-image features for detecting objects in varying lighting conditions, even at night time.

To realize the approach, the single-modality network has to effectively learn from the multi-modality network. Yet, there are several technical challenges. First, the vast 3D space is dominated by background samples, which hinder the transfer of foreground knowledge. Second, 3D detectors involve massive points and/or voxels; naively distilling all pairs of voxel/point features from multi- to single-modality is computationally infeasible. Last, it remains challenging to effectively transfer knowledge for objects of various sizes and shapes, particularly for the small and distant ones.

We address the challenges by designing a novel Simulated Single-to-Multi-Modality Single-Stage 3D object Detector (S2M2-SSD) framework (see Figure 1) to effectively train the single-modality network to learn from the multi-modality one. This work has the following technical contributions. First, we design the crucial response mining strategy and formulate the response distillation for focusing on the crucial responses while avoiding most background ones. Second, we extend the strategy to voxels and formulate consistency constraints on voxel features and voxel relations to enhance the single-modality intermediate features. Third, we formulate a fine-grained voxel-to-point distillation on the crucial foreground points for enhancing the features of objects with sparse points or of small sizes. Last, we further correct the single-modality predictions by learning on the last-layer bird's eye view (BEV) features to improve the instance-level consistency in the deep-layer features.

The above techniques enable us to train and produce a single-modality network that takes only point clouds as input, yet capable of achieving a high performance close to a multi-modality LiDAR-image network; see Figure 2. The evaluation on the nuScenes test set also shows that our S2M2-SSD outperforms all state-of-the-art single-modality 3D object detectors; our NDS metric even surpasses the multi-modality SSD and our mAP fills more than 70% of the gap between the single- and multi-modality SSDs.

2. Related Work

Mainstream 3D object detectors can generally be divided into two categories: (i) single-modality detectors with either point clouds [9, 15, 19, 23, 24, 32, 46, 55] or RGB images [7, 16, 21, 22, 31] as input, and (ii) multi-modality detectors [6, 17, 28] with both point clouds and RGB images as input. Multi-modality detectors often have higher precisions benefited from the complementarity of point clouds and RGB images, while single-modality detectors usually have higher efficiency due to less computation overhead.

Among the single-modality detectors, the LiDAR-only two-stage detectors [33–35] focus on enhancing the region-proposal-aligned features to boost the precision. Recently, LiDAR-only single-stage detectors [11, 47, 51, 53, 54] gradually surpass the two-stage ones with higher precisions. SE-SSD [54] employs a self-ensembling framework to exploit hard and soft targets for model optimization. CenterPoint [51] regresses a confidence heatmap for anchor-free 3D detection. Besides, some recent monocular/stereo 3D detectors [7, 20–22, 31, 36] use only RGB images as input and obtain significant improvements, yet their performance is still lower than those of the LiDAR-only detectors.

Among the multi-modality detectors, the fusion of image and point cloud is the most popular. F-PointNet [28] projects 2D region proposals detected on RGB images to 3D frustums for filtering point clouds for 3D detection. 3D-CVF [52] fuses semantics from multi-view images adaptively with point features. CLOCs PVCas [27] refines the predicted confidence with features from images and point clouds. PointPainting [41] combines the segmentation scores of images with LiDAR points as input. PointAugmenting [42] performs late fusion between point and image segmentation features. So far, only a few studies [2, 25, 26] attempt to fuse radar and image data, yet the performance still cannot surpass that of the LiDAR-image methods.

Unlike previous works, we fuse LiDAR and image data only in training and design four levels of knowledge distillation to effectively train a single-stage 3D object detector to learn to produce/simulate LiDAR-image features and responses from LiDAR-only data. Knowledge distillation [13] is first proposed for model compression and is widely applied in image classification [12, 39, 49]. Recently, a few 2D detectors [4, 8, 10, 30] explore decoupling or enriching the BEV features for knowledge distillation. To the best of our knowledge, this work is the first attempt in 3D object detection on distilling knowledge from multi- to single-modality, and we are able to train a single-modality SSD whose performance is close to a multi-modality one.

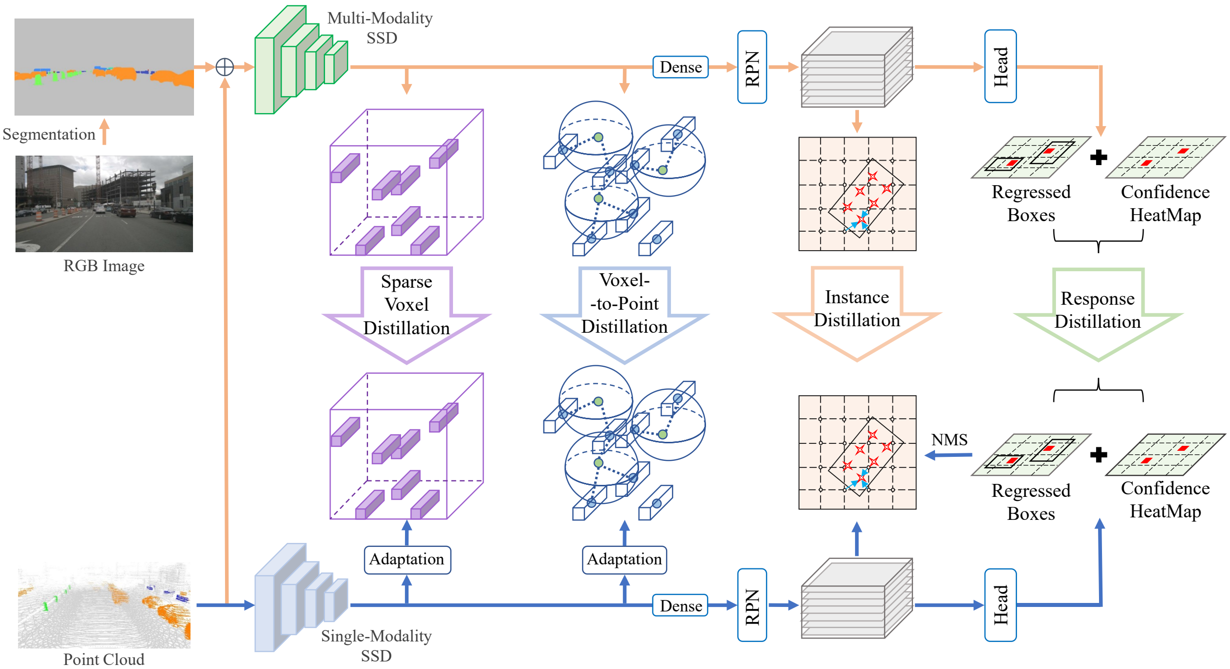

Figure 3. The pipeline of our S2M2-SSD framework. The single-modality SSD (bottom) takes only a point cloud as input, whereas the multi-modality SSD (top) further takes a segmented image. The training has two phases: first, we pre-train the multi-modality SSD (orange arrows). Then, we train the single-modality SSD (orange and blue arrows) to learn to effectively produce features and responses comparable with the pre-trained multi-modality SSD by designing four levels of knowledge distillation in the framework: response distillation (Section 3.2), sparse-voxel distillation (Section 3.3), voxel-to-point distillation (Section 3.4), and instance distillation (Section 3.5). During the testing (only the blue arrows), we only need a point cloud as the input of the well-trained single-modality SSD for the object detection.

3. Simulated Single-to-Multi-Modality SSD

3.1. Overall Framework

Figure 3 shows the pipeline of our S2M2-SSD framework. We pre-train a multi-modality SSD (top) on point clouds and segmented images, following [41], then train a single-modality SSD (bottom) only on point clouds. To effectively train the single-modality SSD to produce features and responses comparable with those of the multi-modality SSD, we design four levels of knowledge distillation:

- Response distillation (Section 3.2) exploits the knowledge in multi-modality responses to correct the single-modality responses based on the crucial response mining we designed for focusing the distillation on the responses that are crucial for precision calculation.
- Sparse-voxel distillation (Section 3.3) extends the mining strategy from responses to voxels, with consistency constraints formulated on voxel features and relations to distill semantics and relation knowledge in crucial voxels from multi- to single-modality SSD.
- Voxel-to-point distillation (Section 3.4) aims to simulate fine-grained features for objects with sparse points or of small sizes, by transforming coarse-grained voxel features to fine-grained point features then distilling the fined-grained features in a point-wise manner.
- Instance distillation (Section 3.5) helps to correct the single-modality predictions by learning the deep-layer BEV features in the NMS-filtered bounding boxes.

Note also that both our single- and multi-modality SSDs adopt the open-source SOTA CenterPoint [51] as backbone.

3.2. Response Distillation

One major challenge to transfer knowledge from multi- to single-modality responses is the *imbalance between foreground and background samples*, as background responses easily dominate the distillation. Hence, many 2D detectors either perform an indiscriminate distillation only on the positive responses [10, 38, 40] or simply ignore the response distillation [4,18,30,43] due to the sample imbalance issue. In 3D detection, such issue is even more severe than the 2D cases, as objects are much more sparse in the 3D space.

In our framework, we design the *crucial response mining* strategy to estimate the responses that are crucial for the detection precision of the single-modality SSD (subscript s), *i.e.*, true positives (TP_s), false positives (FP_s), and false negatives (FN_s). We ignore true negatives, as they are very likely background samples. For time efficiency, we find the crucial responses by comparing the confidence heatmap predicted by the single-modality SSD (denoted by h_s) and the one obtained from the ground truth (denoted by h_g).

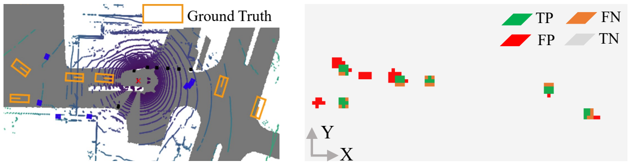

Figure 4. Left: a sample point cloud in bird's eye view (BEV) with the associated ground-truth bounding boxes (in orange). Right: our estimated crucial responses for response distillation.

The mechanism mimics how the IoU or the center distance metric measures the sample reliability:

$$\begin{aligned} TP_s &= (h_s > \tau)\ \&\ (h_g > \tau)\\ FP_s &= (h_s > \tau)\ \&\ (h_g < \tau)\\ FN_s &= (h_s < \tau)\ \&\ (h_g > \tau)\\ \text{and}\ \ h_d &= \max\left(h_d^1, \cdots, h_d^K\right), d \in \{s, g\}, \end{aligned} \tag{1}$$

where h_d^i is the confidence heatmap of the i-th class's bounding boxes from single-modality SSD or ground truth; K is the number of object classes in the detection head; and τ is a threshold. By focusing the distillation on the crucial responses, we can avoid most background samples and make effective the response distillation; see, *e.g.*, Figure 4.

With the estimated crucial responses, we can then formulate a weighted classification response distillation from the multi- to single-modality SSD by imposing a discriminative attention on different response classes:

$$\begin{aligned} &\mathcal{L}^r_{cls} = \frac{w_1^r}{|TP_s|}\sum_{i\in TP_s}\mathcal{L}^r_{\delta_h,i} + \frac{w_2^r}{|FP_s \cup FN_s|}\sum_{i\in FP_s\cup FN_s}\mathcal{L}^r_{\delta_h,i}\\ &\text{and}\ \ \mathcal{L}^r_{\delta_h,i} = \mathcal{L}_{sml1}|h_{i,s} - h_{i,m}|, \end{aligned} \tag{2}$$

where h_m is the multi-modality confidence heatmap corresponding to h_s; $h_{i,m}$ is the i-th sample of h_m; $\mathcal{L}_{sml1}$ is the smooth-L_1 loss; and w_1^r and w_2^r are weights; we empirically set a larger w_2^r to focus more on the false predictions.

Last, we perform regression distillation only on the true positive and false negative responses since they have associated ground-truth objects, while ignoring the false positive ones. As image features offer clear object boundaries, the multi-modality SSD may predict bounding boxes more accurately in some attributes, *e.g.*, size. Hence, we impose differentiated weights on different bounding box attributes in the regression distillation:

$$\begin{aligned} &\mathcal{L}^r_{loc} = \frac{1}{|TP_s \cup FN_s|}\sum_{i\in TP_s\cup FN_s}\sum_{e} w_e^r \mathcal{L}^r_{\delta_e,i}\\ &\text{and}\ \ \mathcal{L}^r_{\delta_e,i} = \mathcal{L}_{sml1}|e_{i,s} - e_{i,m}|, \end{aligned} \tag{3}$$

where attribute $e \in \{x, y, z, w, l, h, v_x, v_y, sin\theta, cos\theta\}$ and $\{w_e^r\}$ are attribute-wise weights. Combining Eqs (2) and (3), we obtain the overall response distillation loss:

$$\mathcal{L}_{rsp} = \mathcal{L}^r_{cls} + \mathcal{L}^r_{loc}. \tag{4}$$

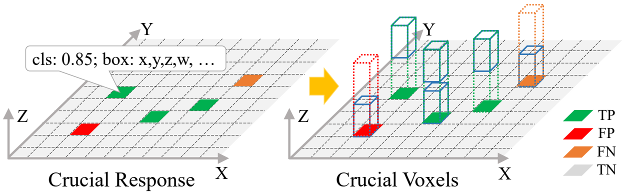

Figure 5. Illustration of crucial voxel mining. We extend each crucial response sample into a pillar and take the active voxels in each pillar as crucial voxels to effectively avoid background voxels.

3.3. Sparse-Voxel Distillation

Next, we design the sparse-voxel distillation to further enhance the single-modality SSD by exploring the voxel features in the last sparse convolution layer. This layer has rich semantics and keeps the original 3D spatial information. Compared to response distillation, sparse-voxel distillation can better *promote the consistency between the high-dimensional features* in single- and multi-modality SSDs.

To distill intermediate features, one may naively impose consistency constraints on all pairs of student and teacher features, as in 2D detection methods [3, 4, 10, 30]. Yet, 3D detection involves much more sparse target objects, so background features dominate the distillation and hinder the foreground knowledge transfer. Also, calculating massive voxels is very time- and resource-consuming. Hence, to consider the computing efficiency and avoid the background features, we leverage the crucial responses estimated from response distillation to find active nonempty *crucial voxels* in pillars extended from the crucial response samples; see Figure 5. As intermediate voxel features are shared by six detection heads for different class groups, we concatenate voxels of TP_s, FN_s, and FP_s from different detection heads as TP_s^v, FN_s^v, and FP_s^v, respectively.

With the crucial voxels, we then formulate a weighted voxel-wise consistency constraint between the single- and multi-modality SSDs (see the purple arrow in Figure 6) with differentiated weights imposed on different voxel classes:

$$\begin{aligned} &\mathcal{L}^v_{fea} = \frac{w_1^v}{|TP_s^v|}\sum_{i\in TP_s^v}\mathcal{L}^v_{\delta_{f_i}} + \frac{w_2^v}{|FP_s^v \cup FN_s^v|}\sum_{i\in FP_s^v\cup FN_s^v}\mathcal{L}^v_{\delta_{f_i}}\\ &\text{and}\ \ \mathcal{L}^v_{\delta_{f_i}} = \frac{1}{C}\mathcal{L}_{sml1}|f^s_{v_i} - f^m_{v_i}|, \end{aligned} \tag{5}$$

where $f^s_{v_i}$ and $f^m_{v_i}$ are C-dimensional single- and multi-modality features, respectively, of the i-th crucial voxel; w_1^v and w_2^v are weights; a larger w_2^v is empirically set to focus more on voxel features associated with false predictions.

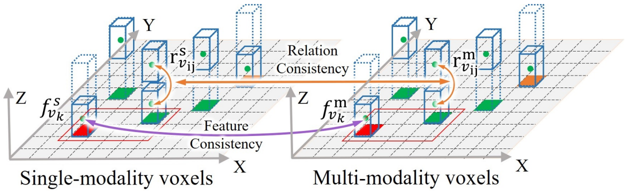

Figure 6. Our sparse-voxel distillation encourages voxel-feature consistency between the single- and multi-modality SSDs in two aspects: voxel features (purple) and voxel relations (orange).

High-dimensional voxel features contain rich semantics and also voxel relations, *e.g.*, different parts of the same object can have very different features, whereas similar parts in different objects can have similar features. By exploiting the relation knowledge between voxels in multi-modality SSD, we can train the single-modality voxel features to be more discriminative; see the orange arrows in Figure 6. So, we formulate the voxel relation consistency loss as

$$\mathcal{L}^{v}_{rel} = \frac{1}{|V_s|^2} \sum_{i \in V_s} \sum_{j \in V_s} ||r^{s}_{v_{ij}} - r^{m}_{v_{ij}}||^2_2$$
$$\text{and} \;\; r^{d}_{v_{ij}} = \frac{f^{d}_{v_i} \cdot f^{d}_{v_j}}{||f^{d}_{v_i}||_2 \cdot ||f^{d}_{v_j}||_2}, d \in \{s, m\}, \tag{6}$$

where the crucial voxel set $V_s = TP^v_s \cup FP^v_s \cup FN^v_s$, and $r^{d}_{v_{ij}}$ is the cosine similarity between voxel features $f^{d}_{v_i}$ and $f^{d}_{v_j}$. Hence, our sparse-voxel distillation loss is

$$\mathcal{L}_{vxl} = \mathcal{L}^{v}_{fea} + \mathcal{L}^{v}_{rel}. \tag{7}$$

3.4. Voxel-to-Point Distillation

For objects with sparse points or of small sizes, the low-resolution coarse-grained voxels in the last layer may not be able to capture their fine features, so sparse-voxel distillation may not be effective for these objects. Alternatively, using high-resolution voxels in shallow layers is infeasible, due to computing inefficiency and insufficient semantics.

To circumvent the issue, we design the voxel-to-point feature distillation module for better handling objects with sparse points or of small sizes. Our key idea is to interpolate the voxel features to raw point clouds and perform *fine-grained feature distillation on points*. Yet, calculating all raw points is computationally infeasible. So, we filter the points that are inside the ground-truth bounding boxes as *crucial foreground points* (denoted as P=$\{p_i\}^{M}_{i=1}$) for distillation. Specifically, denoting p_{v_j} as the center coordinate of voxel v_j, we employ the feature propagation layer [29] to obtain the interpolated point feature at p_i using the inverse distance-weighted average to fuse nearby voxel features:

$$f_{p_i} = \frac{\sum_{j=1}^{k} w_{ij} f_{v_j}}{\sum_{j=1}^{k} w_{ij}}, \;\; \text{where} \;\; w_{ij} = \frac{1}{||p_i - p_{v_j}||_2}; \tag{8}$$

f_{p_i} denotes p_i's point feature; and k denotes the number of voxels in the neighborhood of p_i. With f_{p_i}, we can then perform the point-wise feature distillation between single- and multi-modality SSDs with the consistency loss:

$$\mathcal{L}^{p}_{fea} = \frac{w^{p}_{f}}{|P|} \sum_{i \in P} \frac{1}{C} \mathcal{L}_{sml1} |f^{s}_{p_i} - f^{m}_{p_i}|, \tag{9}$$

where w^{p}_{f} is weight and C is the number of channels in f_{p_i}. Further, we can exploit the rich relation knowledge among points by formulating the point relation distillation loss:

$$\mathcal{L}^{p}_{rel} = \frac{1}{|P'|^2} \sum_{i \in P'} \sum_{j \in P'} ||r^{s}_{p_{ij}} - r^{m}_{p_{ij}}||^2_2$$
$$\text{and} \;\; r^{d}_{p_{ij}} = \frac{f^{d}_{p_i} \cdot f^{d}_{p_j}}{||f^{d}_{p_i}||_2 \cdot ||f^{d}_{p_j}||_2}, d \in \{s, m\}, \tag{10}$$

where $r^{d}_{p_{ij}}$ denotes the cosine similarity between point features $f^{d}_{p_i}$ and $f^{d}_{p_j}$; and P' is the set of $\min(M, 4500)$ points randomly selected from P to avoid excessive GPU computation. Combining Eqs. (9) and (10), we formulate the voxel-to-point distillation loss as

$$\mathcal{L}_{pts} = \mathcal{L}^{p}_{fea} + \mathcal{L}^{p}_{rel}. \tag{11}$$

3.5. Instance Distillation

While intermediate 3D feature distillations greatly promote the knowledge transfer from multi- to single-modality SSD, the single-modality features may still deviate from the multi-modality ones, as the layer goes deeper. Distillation on low-dimensional responses also cannot guarantee consistent high-dimensional deep features. To this end, we design the instance distillation module *on the last-layer BEV features to promote the consistency of the deep features*, which have a direct impact on the object prediction.

This module first uses NMS to remove redundant bounding boxes predicted on the single-modality BEV features, then performs the rotated RoI-grid pooling to crop the last-layer BEV features in each filtered bounding box. In detail, we treat each filtered bounding box as an instance, fit a uniform grid (G=5×5) in each bounding box, interpolate BEV features at each grid point, and transfer knowledge from multi- (superscript m) to single-modality (superscript s) through the 5×5 interpolated features:

$$\mathcal{L}_{ins} = \frac{w^{I}}{B} \sum_{i=1}^{B} \frac{1}{G} \sum_{j=1}^{G} \mathcal{L}_{sml1} |f^{s}_{I_{ij}} - f^{m}_{I_{ij}}|, \tag{12}$$

where B is the number of NMS-filtered bounding boxes; $f_{I_{ij}}$ is the interpolated feature at the j-th grid point of the i-th instance; and w^{I} is a hyper-parameter.

Table 1. Comparison with SOTA LiDAR-only detectors on the nuScenes test set. Our S2M2-SSD attains the *highest NDS and mAP*, as well as *highest AP consistently for all ten object classes*. The percentages in () mean the proportions of S2M2-SSD's gains on the single-modality SSD relative to the single- and multi-modality metric gap. '*' means the SSD is built on our improved version of CenterPoint.

Method	Modality	NDS	mAP	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bicycle	TC	Barrier
WYSIWYG [14]	LiDAR	41.9	35.0	79.1	30.4	46.6	40.1	7.1	65.0	18.2	0.1	28.8	34.7
PointPillars [15]	LiDAR	45.3	30.5	68.4	23.0	28.2	23.4	4.1	59.7	27.4	1.1	30.8	38.9
PointPainting [41]	LiDAR	58.1	46.4	77.9	35.8	36.2	37.3	15.8	73.3	41.5	24.1	62.4	60.2
CVCNet [5]	LiDAR	64.4	55.3	82.7	46.1	46.6	49.4	22.6	79.8	59.1	31.4	65.6	69.6
PMPNet [50]	LiDAR	53.1	45.4	79.7	33.6	47.1	43.1	18.1	76.5	40.7	7.9	58.8	48.8
SSN [57]	LiDAR	58.1	46.4	80.7	37.5	39.9	43.9	14.6	72.3	43.7	20.1	54.2	56.3
CBGS [56]	LiDAR	63.3	52.8	81.1	48.5	54.9	42.9	10.5	80.1	51.5	22.3	70.9	65.7
CenterPoint [51]	LiDAR	65.5	58.0	84.6	51.0	60.2	53.2	17.5	83.4	53.7	28.7	76.7	70.9
Multi-modality SSD*	*LiDAR+RGB*	*69.1*	*64.0*	*86.2*	*55.4*	*65.6*	*58.2*	*28.3*	*85.0*	*65.1*	*40.0*	*79.8*	*75.9*
Single-modality SSD*	LiDAR	67.3	60.1	85.2	51.9	63.6	55.9	21.7	83.1	55.7	33.1	75.7	74.7
S2M2-SSD (Ours)	LiDAR	**69.3**	**62.9**	**86.3**	**56.0**	**65.4**	**59.8**	**26.2**	**84.6**	**61.6**	**36.4**	**77.7**	**75.1**
Improvement	-	*+2.0 (111%)*	*+2.8 (72%)*	*+1.1*	*+4.1*	*+1.8*	*+3.9*	*+4.5*	*+1.5*	*+5.9*	*+3.3*	*+2.0*	*+0.4*

3.6. Overall Loss Function

We train the single-modality SSD end-to-end (while fixing the pre-trained multi-modality SSD; see Figure 3) with the following supervision losses and distillation losses:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} + \mu(\mathcal{L}_{rsp} + \mathcal{L}_{vxl} + \mathcal{L}_{pts} + \mathcal{L}_{ins}), \quad (13)$$

where L_{cls} is classification loss and L_{reg} is regression loss, following [51]; and λ and μ are hyper-parameters.

4. Experiments

4.1. Experimental Setup

Dataset. We use nuScenes [1], a popular large-scale multimodal dataset, in our experiments. The dataset comprises 1,000 driving sequences (700/150/150 for train/val/test), each 20 seconds long, and every ten frames are fully annotated with 3D object bounding boxes. The LiDAR sensor scans at a frequency of 20 FPS, producing 400 frames per sequence and around 30k points per frame. For each frame, the dataset also provides RGB images from all six cameras to realize a complete 360° coverage. The nuScenes detection task involves ten classes of objects of various sizes and shapes, including 28k, 6k, and 6k samples for training, validating, and testing, respectively. The goal of the task is to determine the following parameters of each bounding box: x, y, z, width, length, height, velocity, and yaw angle.

Metrics. We adopt the main official evaluation metrics of nuScenes, *i.e.*, the nuScenes detection score (NDS) and mean Average Precision (mAP), which are calculated by averaging over all the object classes. NDS is calculated halfly based on mAP and halfly based on the true-positive qualities, *i.e.*, translation, velocity, orientation, etc., whereas mAP is calculated based on matching boxes by thresholding the 2D center distance (0.5m, 1m, 2m, 4m) on the ground.

Network and Training. We follow CenterPoint [51] to build our single- and multi-modality SSDs and improve it by more data augmentation. The detection range is [-51.2m, 51.2m] for the X and Y axes and [-5m, 3m] for the Z axis, and the point clouds are discretized by a voxel size of [0.1m, 0.1m, 0.2m]. To obtain a high precision, we employ the open-source PointPainting [41] to pre-train the multi-modality SSD with both point clouds and RGB images. Also, we use a linear layer with ReLU activation as the adaptation layer for sparse-voxel distillation, and a submanifold sparse convolution layer with ReLU activation as the adaptation layer for voxel-to-point distillation. We set τ as 0.1 (Eq. (1)), w_1^r as 1.0 and w_2^r as 5.0 (Eq. (2)), $\{w_e^l\}$ as 0, 0, 0, 0.1, 0.1, 0.1, 0.1, 0.1, 0, and 0 (Eq. (3)), w_1^v as 2.0 and w_2^v as 8.0 (Eq. (5)), w_f^p as 2.0 (Eq. (9)), w^I as 8.0 (Eq. (12)), λ as 0.25 and μ as 0.5 (Eq. (13)).

4.2. Comparison with the State-of-the-Arts

We evaluated our S2M2-SSD on the nuScenes test set by submitting the predicted results to the nuScenes server. Table 1 reports the resulting NDS and mAP for comparison with the state-of-the-art LiDAR-only methods. As shown in the table, our S2M2-SSD attains the highest NDS and mAP among all LiDAR-only detectors, with significant gains of +2.8 points on mAP and +2.0 points on NDS over the single-modality SSD. Also, the mAP gain (+2.8) of our S2M2-SSD fills ~72% mAP gap (3.9) between the single- and multi-modality SSDs, while the NDS of our S2M2-SSD even surpasses that of the multi-modality SSD. We think that our higher NDS is due to the further improvement in the true-positive qualities based on the mAP gain.

Also, our S2M2-SSD attains consistent improvements for all ten object classes over the state-of-the-art methods. Compared with the multi-modality SSD, our S2M2-SSD attains comparable or even higher APs on 'car', 'truck', 'trailer', 'bus', and 'pedestrian', since larger objects can be

Table 2. Evaluation on the nuScenes validation set with models trained on 30% training data, serving as ablation study baselines.

Method	Modality	NDS	mAP
Multi-modality SSD	LiDAR+RGB	*55.1*	*48.0*
Single-modality SSD	LiDAR	51.0	42.0
S2M2-SSD	LiDAR	**55.6**	**46.2**
Improvement	-	*+4.6*	*+4.2*

Table 3. Ablation study on our proposed modules: "response", "voxel", "point", and "instance" denote the response, sparse-voxel, voxel-to-point, and instance distillation modules, respectively. All results are based on model training on 30% train set, and the NDS and mAP are reported on the nuScenes val split.

response	voxel	point	instance	NDS	mAP
				51.0	42.0
✓				53.0	43.9
	✓			53.8	44.2
		✓		52.7	43.8
			✓	52.6	43.2
✓	✓			54.3	44.9
✓	✓	✓		55.0	45.6
✓	✓	✓	✓	**55.6**	**46.2**

Table 4. Ablation study on response distillation, in which we show the effect of distilling different kinds of (crucial) responses, as well as the effect of regression distillation (denoted by "loc").

Distilled responses	NDS	mAP
baseline	51.0	42.0
all	51.4	42.3
true positives	52.1	43.0
+ false positives	51.6	42.7
+ false negatives	51.8	42.6
+ false positives & negatives	52.8	43.7
+ false positives & negatives + loc	**53.0**	**43.9**

easily benefited from all of our distillation modules. On the other hand, on 'construction vehicle', 'motorbike', 'bicycle', and 'traffic cone', our method is able to largely narrow down the AP gap between the single- and multi-modality SSDs. As for 'barrier', these objects have high variations in shapes and sizes, so are very hard for consistent distillation.

Besides, since S2M2-SSD only needs point clouds as input, it has high efficiency at inference, compared to multi-modality detectors that need to process both LiDAR and image inputs; see the details in Section 4.4. Also, S2M2-SSD is able to produce simulated multi-modality features, even at night time, manifesting its robustness to the environment's lighting condition. Lastly, note that while there are some higher-precision LiDAR-only and LiDAR-image methods on the nuScenes leaderboard, they do not have peer-reviewed papers and public code for us to experiment our method with. So, we resort to use the open-source CenterPoint and PointPainting as our framework backbones.

4.3. Ablation Study

For efficiency, we follow [37, 45] to conduct all ablation studies by training on 30% train samples and evaluating on the complete val split. Table 2 shows the NDS and mAP of our S2M2-SSD and also the multi- and single-modality SSDs, which serve as base results in the ablation studies. Table 3 shows the effects of each module in S2M2-SSD.

Effect of response distillation. As the first two rows in Table 3 show, our response distillation module boosts the NDS by 2.0 points and mAP by 1.9 points, showing that our approach can effectively improve the single-modality responses for higher consistency with the multi-modality ones. Also, this module consumes much less computation than others, yet delivering impressive improvements.

Table 4 shows further ablation studies on our crucial response mining strategy. From the table, we can see that performing response distillation on all samples only boosts NDS by 0.4 and mAP by 0.3, while performing response distillation only on the true positives already improves NDS by 1.1 and mAP by 1.0. These results show that the dominated background samples seriously hinder the transfer of knowledge for the foreground samples. On the other hand, adding either false positives or false negatives reduces the performance, as doing so may significantly change the confidence relations between the distilled responses and the undistilled crucial responses. When distilling all crucial responses, both NDS and mAP can further be improved by 0.7 (comparing 3rd and 6th rows). Lastly, the regression distillation further improves both metrics by 0.2.

Effect of sparse-voxel distillation. Comparing the first and third rows in Table 3 shows that our sparse-voxel distillation improves the NDS by 2.8 points and mAP by 2.2 points, which are the largest single-module improvements compared with others. These results manifest the necessity of enhancing the high-dimensional intermediate features. Also, comparing the second and sixth rows in Table 3 shows that this module further increases the NDS by 1.3 points and mAP by 1.0 points on top of the response distillation, showing the complementary strengths of the two modules.

Table 5 shows further ablation results conducted on the crucial voxel mining strategy and the consistency losses on voxel features and voxel relations. By comparing the results between the 2nd-4th and 5th-7th rows, we can see that although our crucial voxel mining strategy keeps only 984 crucial voxels out of all the 7718 active voxels on average per 3D scene, it reduces the number of FLOPs significantly from 9.15×10^{10} to 1.49×10^{9}, and boosts the precisions (both NDS and mAP) obviously, as it helps avoid the background voxels and overfitting them. Note that the FLOPs for the relation loss has a time complexity of O(V^2), where V is the number of voxels. Also, we study the effects of each item in the consistency losses; from 2nd-4th

Table 5. Ablation study on sparse-voxel distillation: "filter" means the crucial voxel mining; "cons" and "rel" denote the consistency losses on voxel features and relations, respectively; "Voxels" means the average number of voxels for distillation per point cloud; and "FLOPs" is calculated based on Eqs. (5) and (6).

Method	Voxels	FLOPs	NDS	mAP
baseline	-	-	51.0	42.0
cons w/o filter	7718	1.98×10^{6}	52.7	43.2
rel w/o filter		9.15×10^{10}	52.0	42.4
cons&rel w/o filter		9.15×10^{10}	53.3	43.4
cons w/ filter	984	2.52×10^{5}	53.1	44.1
rel w/ filter		1.49×10^{9}	52.5	42.7
cons&rel w/ filter		1.49×10^{9}	**53.8**	**44.2**

Table 6. Ablation study on voxel-to-point distillation, manifesting the effects of consistency losses on point features and relations, especially for objects with sparse points or of small sizes.

Method	ped	motor	bicycle	TC	barrier	NDS	mAP
baseline	72.2	38.6	11.7	49.6	49.5	51.0	42.0
cons loss	75.6	38.0	13.2	51.9	50.7	52.2	43.2
rel loss	75.1	39.1	13.4	52.0	51.0	52.6	43.3
cons&rel loss	**76.1**	**40.7**	**13.7**	**52.3**	**51.7**	**52.7**	**43.8**

and 5th-7th rows, we can see that the voxel-feature loss contributes most, since it promotes direct knowledge transfer between the coarse-grained voxels. Further, the voxel-relation loss can effectively boost the NDS (+0.7) more than mAP (+0.1), since it improves the true-positive metrics more by making the features more discriminative.

Effect of voxel-to-point distillation. Comparing the first and forth rows in Table 3 shows that our voxel-to-point distillation improves NDS by 1.7 and mAP by 1.8. Comparing the sixth and seventh rows shows that it still increases both NDS and mAP by 0.7 on top of the previous two modules, showing the importance of simulating fine-grained features.

Table 6 further shows the module's effects on objects with sparse points or of small sizes. It can be seen that the 'pedestrian' AP can be significantly improved by 3.9 points, and the APs of 'motorbike', 'bicycle', 'TC', and 'barrier' are also improved by around 2∼3 points, validating our motivation of promoting the fine-grained feature consistency. Also, we can see that the point-relation loss contributes more than the point-feature loss, opposite to the effects of sparse-voxel distillation, as the point-relation loss can help produce more discriminative single-modality features by exploiting and contrasting the point features.

Effect of instance distillation. Comparing the first and fifth rows in Table 3 shows that instance distillation improves the NDS by 1.6 and mAP by 1.2. Also, instance distillation still improves both two metrics by 0.6 points (see the last two rows) over the previous modules, showing its effect in enhancing the deep features and correcting the single-modality predictions. Table 7 shows more ablation results conducted with different definitions of instances, including the ground truths, NMS-filtered crucial responses, and NMS-filtered bounding boxes predicted by the single-modality SSD. We can see that our method attains better results than the other settings, since our focused deep features have direct association with the final detected boxes.

Table 7. Ablation study on instance distillation, in which "gt boxes", "crucial-nms", and "all-nms" denote the instance defined by the ground-truth boxes, NMS-filtered crucial responses, and all predicted bounding boxes filtered by NMS, respectively.

Method	baseline	gt boxes	crucial-nms	all-nms (ours)
NDS	51.0	51.8	52.4	**52.6**
mAP	42.0	42.5	42.7	**43.2**

Table 8. Comparing the runtime (in millisecond) of our S2M2-SSD against the multi-modality SSD for processing each modality data, showing the high computational efficiency of our approach.

Method	image	point cloud	total
Multi-modality SSD	425.5	107.7	533.2
S2M2-SSD (ours)	-	81.9	**81.9**

4.4. Runtime Analysis

Table 8 compares the runtime of our S2M2-SSD and the multi-modality SSD to show the high efficiency of our approach. As shown in the table, S2M2-SSD only needs 81.9ms for detecting objects in a 3D point cloud, whereas the multi-modality SSD needs 425.5ms and 107.7ms to process the input image and detect objects in the point cloud, respectively. Note that we ignore the time for calibration, synchronization, and data fusion, which are hard to be quantified. All evaluations were done on an Intel Xeon Silver CPU and a TITAN Xp GPU with a batch size of four.

5. Conclusion

We presented a novel framework capable of producing a single-stage 3D object detector with high precision, efficiency, and robustness, by simulating multi-modality (LiDAR-image) on single-modality (LiDAR) input. Our S2M2-SSD framework consists of four levels of knowledge distillation: response distillation to focus on the crucial responses and avoid most background samples; sparse-voxel distillation to learn the voxel semantics and relation knowledge; voxel-to-point distillation to attend to features of small and distant objects; and instance distillation to promote deep-layer feature consistency. Experimental results on nuScenes shows that our S2M2-SSD achieves SOTA LiDAR-only performance and surpasses the baseline multi-modality SSD on the key NDS metric, filling ∼72% mAP gap between the single- and multi-modality SSDs.

Acknowledgments. This research is supported by project MMT-p2-21 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

References

[1] 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 *CVPR*, 2020. 6

[2] Simon Chadwick, Will Maddern, and Paul Newman. Distant vehicle detection using radar and vision. In *ICRA*, 2019. 2

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

[4] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 2, 3, 4

[5] Qi Chen, Lin Sun, Ernest Cheung, Kui Jia, and Alan Yuille. Every view counts: Cross-view consistency in 3D object detection with hybrid-cylindrical-spherical voxelization. *NeurIPS*, 2020. 6

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *CVPR*, 2017. 2

[7] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. DSGN: Deep stereo geometry network for 3D object detection. In *CVPR*, 2020. 1, 2

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *CVPR*, 2021. 2

[9] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. RangeDet: In defense of range view for LiDAR-based 3D object detection. *ICCV*, 2021. 2

[10] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, 2021. 2, 3, 4

[11] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3D object detection from point cloud. In *CVPR*, 2020. 1, 2

[12] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 2

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[14] Peiyun Hu, Jason Ziglar, David Held, and Deva Ramanan. What you see is what you get: Exploiting visibility for 3d object detection. *CVPR*, 2020. 6

[15] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 2, 6

[16] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN based 3D object detection for autonomous driving. In *CVPR*, 2019. 1, 2

[17] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3D object detection. In *CVPR*, 2019. 1, 2

[18] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. 3D-to-2D distillation for indoor scene parsing. In *CVPR*, 2021. 3

[19] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. TANet: Robust 3D object detection from point clouds with triple attention. In *AAAI*, 2020. 2

[20] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. AutoShape: Real-time shape-aware monocular 3D object detection. In *ICCV*, 2021. 2

[21] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3DSSD: Monocular 3D single stage object detector. In *CVPR*, 2021. 1, 2

[22] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3D object detection. In *CVPR*, 2021. 1, 2

[23] Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. Pyramid R-CNN: Towards better performance and adaptability for 3D object detection. In *ICCV*, 2021. 2

[24] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3D object detection. In *ICCV*, 2021. 2

[25] Ramin Nabati and Hairong Qi. CenterFusion: Center-based radar and camera fusion for 3D object detection. In *WACV*, 2021. 2

[26] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2019. 2

[27] Su Pang, Daniel Morris, and Hayder Radha. CLOCs: Camera-lidar object candidates fusion for 3D object detection. In *IROS*, 2020. 1, 2

[28] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3D object detection from RGB-D data. In *CVPR*, 2018. 2

[29] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 5

[30] 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 *CVPR*, 2021. 2, 3, 4

[31] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3D object detection. In *CVPR*, 2021. 1, 2

[32] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3D object detection with channel-wise transformer. In *ICCV*, 2021. 2

[33] 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 *CVPR*, 2020. 1, 2

[34] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *CVPR*, 2019. 2

[35] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3D object detec-

tion from point cloud with part-aware and part-aggregation network. *PAMI*, 2020. 2

[36] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3D object detection. *ICCV*, 2021. 2

[37] Shaoshuai Shi, et al. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *arXiv preprint arXiv:2102.00463*, 2021. 7

[38] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020. 3

[39] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 2

[40] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *CVPR*, 2021. 3

[41] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. PointPainting: Sequential fusion for 3D object detection. In *CVPR*, 2020. 1, 2, 3, 6

[42] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. PointAugmenting: Cross-modal augmentation for 3D object detection. In *CVPR*, 2021. 2

[43] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019. 3

[44] Zhixin Wang and Kui Jia. Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection. In *IROS*, 2019. 1

[45] Chunwei Wang, et al. PointAugmenting: Cross-modal augmentation for 3D object detection. In *CVPR*, 2021. 7

[46] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2

[47] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3DSSD: Point-based 3D single stage object detector. In *CVPR*, 2020. 2

[48] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: Sparse-to-dense 3D object detector for point cloud. In *ICCV*, 2019. 1

[49] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 2

[50] Junbo Yin, Jianbing Shen, Chenye Guan, Dingfu Zhou, and Ruigang Yang. Lidar-based online 3D video object detection with graph-based message passing and spatiotemporal transformer attention. *CVPR*, 2020. 6

[51] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *CVPR*, 2021. 1, 2, 3, 6

[52] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection. In *ECCV*, 2020. 1, 2

[53] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. CIA-SSD: Confident iou-aware single-stage object detector from point cloud. *AAAI*, 2021. 2

[54] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. SE-SSD: Self-ensembling single-stage object detector from point cloud. In *CVPR*, 2021. 1, 2

[55] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3D object detection. In *CVPR*, 2018. 2

[56] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3D object detection. *NeurIPS*, 2019. 6

[57] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. SSN: Shape signature networks for multi-class object detection from point clouds. *ECCV*, 2020. 6