CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation

Supplementary Material

We provide this supplementary material with additional details to support the main paper.

A. Implementation Details

In this section, we provide implementation details of the CRKD framework to enable cross-modality Knowledge Distillation (KD) from a LiDAR-Camera (LC) teacher detector to a Camera-Radar (CR) student detector.

A.1. Data Augmentation

Our data processing pipeline is mainly adopted from the open-source implementation of BEVFusion [8]. The pipeline of processing the camera image is the same in the teacher and student models. The images of the cameras from 6 Perspective View (PV) are loaded and resized to 256×704. During the training process, data augmentation is applied to the images. We resize the image with scaling factors in the range of $[0.38, 0.55]$. We also set the random rotation in the range of $[-5.4°, 5.4°]$. The images are normalized following the default practice in [3]. For the LiDAR input, the keyframe point cloud is loaded along with 9 previous sweeps. During training, random resizing is applied with the scaling factor in $[-0.9, 1.1]$. Translation augmentation is also applied with a limit of 0.5m. For the radar input, the keyframe is loaded with 6 previous sweeps. We follow BEVFusion [8] to select radar data dimensions. The training time augmentation of radar points is the same as LiDAR data for consistency. We also apply the class-balanced grouping and sampling (CBGS) strategy during training [17]. We do not apply any test-time augmentation for any of our models.

A.2. Teacher Model

As mentioned in the main paper, we add a gated network to the BEVFusion-LC [8] and denote it as BEVFusion-LC*. We use the CenterHead [14] as the detector head in the BEVFusion-LC*. There are two streams in the teacher model for LiDAR and cameras. The LiDAR point cloud is encoded as the Bird's Eye View (BEV) feature map through the LiDAR encoder and BEV reduction module (flatten over z dimension). For the camera stream, the images are loaded and pre-processed to a resolution of 256×704. We use the SwinT [7] backbone to process the images from 6 cameras separately. The PV features are transformed to BEV by taking advantage of the efficient PV-to-BEV transformation module in BEVFusion [8]. The BEV feature maps from the LiDAR stream and camera stream are passed into the gated network to obtain gated feature maps with attentional relative importance between the input features. These gated feature maps are further fused by the original convolutional fusion module in BEVFusion [8]. The fused feature map is then fed into a decoder and the CenterHead [14] to generate object predictions. We train the teacher model for 20 epochs using an AdamW optimizer [10]. The initial learning rate is set as 2e-4 with a cosine annealing schedule [3, 9]. The batch size is set as 16. The object sampling strategy [12] is applied for the first 15 epochs. During distillation, the pre-trained teacher weight is loaded and frozen.

A.3. Student Model

Similar to the LC teacher model, the gated network is also applied to the CR student model, which is denoted as BEVFusion-CR*. The stream to process the camera images is the same as the teacher model. The input radar data is processed by a PointPillar-based backbone [3, 5] to obtain the BEV feature map for the radar stream. The feature maps from these two streams are fused via the gated network and the convolutional fusion module in BEVFusion [8]. To maintain the consistency between the teacher and student models, the student model also uses the CenterHead [14] as the detector head. We train the student model following the same setting as the teacher model. During distillation, we load the pre-trained BEVFusion-CR* model and operate cross-modality distillation with the proposed CRKD framework.

A.4. Base Loss Choice

In general, $\mathcal{L}_2$ is more common for feature KD. However, for CSRD, we have to consider the sensor properties. Due to radar's sparse measurements, some objects may be missed, causing radar features at corresponding locations to be outliers when computing loss with the objectness heatmap. We use $\mathcal{L}_1$ to downplay this effect as it penalizes large errors less heavily than $\mathcal{L}_2$, which leads to 0.4% improvement in mAP than using $\mathcal{L}_2$. For MSFD, we follow common practice ($\mathcal{L}_2$) as the domain gap is relatively small (shared camera modality). For RelD, we agree with reviewer vB2P that applying $\mathcal{L}_1$ between similarity matrices is appropriate. For RespD, we mainly follow existing works (e.g., CMKD [4], BEVSimDet [15]) to choose the base loss. Our method is fairly robust to base loss choice, while the final design aligns with our design consideration and brings the best performance.

Method	Modality	NDS↑			mAP↑		
		$[0\text{m}, 20\text{m}]$	$[20\text{m}, 30\text{m}]$	$[30\text{m}, 50\text{m}]$	$[0\text{m}, 20\text{m}]$	$[20\text{m}, 30\text{m}]$	$[30\text{m}, 50\text{m}]$
Teacher	L+C	76.71	68.63	50.57	77.11	62.37	38.25
Student	C+R	63.03	52.87	38.86	58.54	38.64	19.50
CRKD	C+R	**65.53(+2.50)**	**53.52(+0.65)**	**39.21(+0.35)**	**61.59(+3.05)**	**39.04(+0.40)**	**20.53(+1.03)**

Table 1. Performance breakdown by range evaluated on the nuScenes *val* split. We quantitatively show the improvement made by CRKD over the student model.

B. Supplementary Experiments

B.1. CRKD

After loading the pre-trained weights for the teacher and student models, we add the four proposed KD loss terms to the normal detection loss and start the KD training process. We disable the object sampling strategy [12] during distillation. We set the learning rate as 1e-4 with the cosine annealing strategy and train the model for 20 epochs. The batch size is set as 8. For the loss weights, we set the hyper-parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ as $100, 10, 0.25, 1, 1$, respectively. More specifically, we set 1 as the weight for $\mathcal{L}_{respd}$ and $\mathcal{L}_{det}$ since they both compute loss for box regression and classification. The weight of $\mathcal{L}_{reld}$ is 0.25 as it sums up the loss of 4 downsampled affinity map pairs. We empirically select the weights of $\mathcal{L}_{csrd}$ and $\mathcal{L}_{msfd}$ (i.e., 100 and 10) to balance with other loss modules. For the Mask-Scaling Feature Distillation (MSFD), we set r_1 and r_2 as 20m and 30m. The mask-scaling factors α and β are set as 0.25 and 0.5. For the velocity threshold, we set v_1 and v_2 as 0.3m/s and 0.8m/s. We also clip the object size expanding value within $[0.5\text{m}, 4\text{m}]$ to balance between different sizes of objects.

B.2. CRKD Improvement Analysis

Since CRKD is performing a novel KD path (LC to CR), we conduct more experiments to break down the improvement brought by CRKD to provide further insight. As the camera sensor is shared in both the teacher and student models, we narrow down our focus to the difference between LiDAR and radar integration. Radars have better long-range detection capability and weather robustness than LiDAR [6, 13, 16]. In practice, we group objects by their range to the ego-vehicle and the weather of the scene they belong to. We show mAP and NDS of the teacher model, the student model and CRKD. We highlight the quantitative improvement KD brings over the student model. As shown in Tab. 1, we can see that the most improvement comes from the short-range group. This finding demonstrates that CRKD helps the CR student detector to refine its detections in the short-range group, which can be considered as one of LiDAR's strengths as LiDAR has satisfying density for objects that are near to the LiDAR. We are also surprised to see that for mAP, the improvement in long-range group is more than the medium-range group. This finding can provide evidence that cross-modality KD can also enhance the strength of the student detector. In addition, we group different scenes according to the weather and lighting conditions. Table 2 demonstrates increased performance from CRKD across all weather conditions, compared to the baseline student model. Notably, we see a more significant increase in improvement from CRKD in rainy weather. This finding supports that cross-modality KD can help the student to learn and leverage radar's robustness to the varying weather for better results.

B.3. Radar Distillation Design

CRKD presents a novel distillation path to a CR detector. We specifically design the KD module for radars, which has not been previously studied. We present more ablation studies to justify our design choice. We hope our work can bring more insights for future KD frameworks that leverage the radar sensor. In the proposed Cross-Stage Radar Distillation (CSRD) module, we design a calibration module to account for the noisy radar measurements. We conduct an ablation study to understand the effect of the calibration module. Table 3 demonstrates that the calibration module helps to further improve the performance of the student detector.

In addition to the ablation study in the main paper, we show another ablation study of the best distillation source for the CSRD module. Specifically, we compare between using the ground truth heatmap or the heatmap predicted by the teacher model. The results in Tab. 4 show that the objectness heatmap predicted by the teacher detector is a better distillation source for radar distillation.

We additionally compare taking the max or mean pooling along the class dimension of the objectness heatmap Y^T predicted by the teacher detector. Table 5 shows that taking the mean value along different classes of the source heatmap brings more improvement.

Method	Modality	NDS↑				mAP↑			
		Sunny	Rainy	Day	Night	Sunny	Rainy	Day	Night
Teacher	L+C	70.22	71.01	70.54	44.92	66.02	65.48	66.25	41.12
Student	C+R	55.60	57.56	56.37	33.40	44.73	47.27	45.78	23.94
CRKD	C+R	**56.56(+0.96)**	**59.97(+2.41)**	**57.59(+1.22)**	**34.22(+0.82)**	**45.95(+1.22)**	**49.59(+2.32)**	**47.16(+1.38)**	**24.14(+0.20)**

Table 2. Performance breakdown by weather and lighting evaluated on the nuScenes *val* split. We quantitatively show the improvement made by CRKD over the student model.

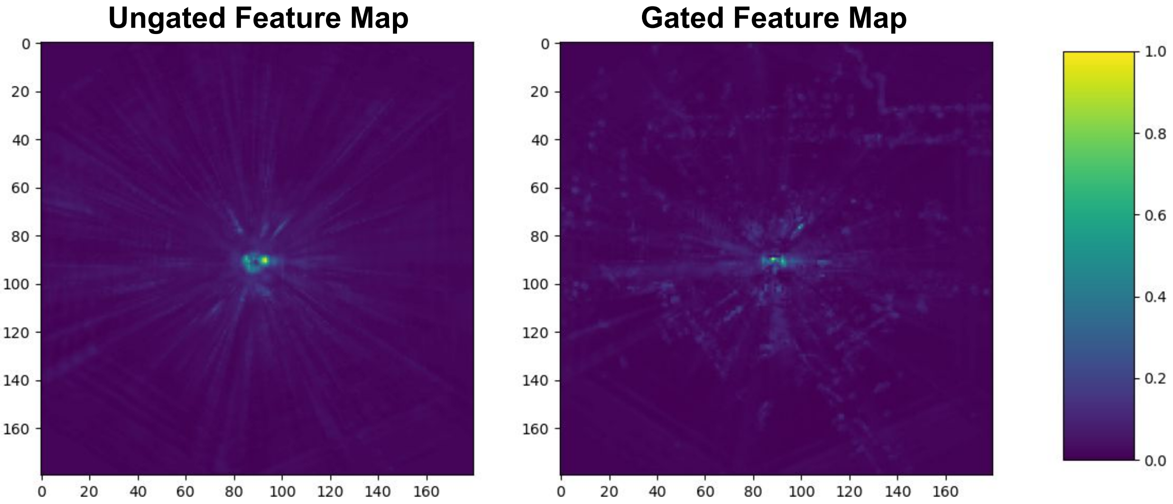


Figure 1. Visualization of the ungated and gated camera feature maps in the teacher detector. The scene geometry can be more easily interpreted from the gated feature map, as it has encoded information from the LiDAR point cloud. Best viewed in color.

Module	w/o calib	w/ calib	mAP↑	NDS↑
CSRD	✓		45.9	56.9
		✓	46.0	57.0

Table 3. Ablation study of CSRD with the radar calibration module.

Module	GT	Teacher Heatmap	mAP↑	NDS↑
CSRD	✓		45.9	56.8
		✓	46.0	57.0

Table 4. Ablation study of CSRD with the different distillation sources.

Module	max	mean	mAP↑	NDS↑
CSRD	✓		45.6	56.8
		✓	46.0	57.0

Table 5. Ablation study of CSRD with different channel-wise pooling methods on the heatmap predicted by the teacher model.

Module	Ungated	Gated	mAP↑	NDS↑
MSFD	✓		45.5	56.8
		✓	45.7	56.9

Table 6. Ablation study of MSFD with the gated camera feature.

Module	Cam	Fused	Cam&Fused	mAP↑	NDS↑
MSFD	✓			45.7	56.9
		✓		45.8	56.7
			✓	46.2	56.9

Table 7. Ablation study of MSFD with different distillation locations.

Module	Dense	Gaussian	Ours	mAP↑	NDS↑
MSFD	✓			45.7	56.8
		✓		45.5	56.7
			✓	46.0	57.0

Table 8. Ablation study of MSFD with different feature masking algorithms.

B.4. Feature Distillation Location

We also experiment with introducing feature distillation at different locations. Since we introduce the gated network to the original BEVFusion [8] model, we design an ablation experiment justifying the introduction of the gated feature map to improve the feature distillation. Specifically, we compare using the gated camera feature map or the ungated camera feature map as the feature distillation source. The results shown in Tab. 6 demonstrate that the gated feature map serves as a more effective distillation source. We additionally show a qualitative example in Fig. 1 to demonstrate the benefits of using the gated feature map. The gated feature map has more informative scene-level geometry thanks to the gated network and learned relative importance weight.

Since the teacher and student models are both fusion-based, we have multiple options of feature distillation locations (e.g., camera feature, fused feature). For the proposed Mask-Scaling Feature Distillation (MSFD) module, we experiment between different locations. As shown in Tab. 7, the most effective design of MSFD is to perform the distillation of the gated camera feature map and fused feature map together. Moreover, we conduct an experiment testing alternative foreground mask generation methods. We experiment with not including any foreground mask compared to methods that include a foreground mask [1, 8, 16]. To complement the ablation study in the main paper, we compare the proposed CRKD module against the same instance without any foreground mask (denoted as dense). In addition, we try CRKD with a Gaussian-style heatmap [2, 15]. The results are shown in Tab. 8. This table demonstrates that though there are certain papers reporting having a Gaussian heatmap is helpful [2, 15], the most effective masking strategy in our scenario is still to apply the proposed mask-scaling strategy.

B.5. Response Distillation: Strength Amplification or Weakness Mitigation?

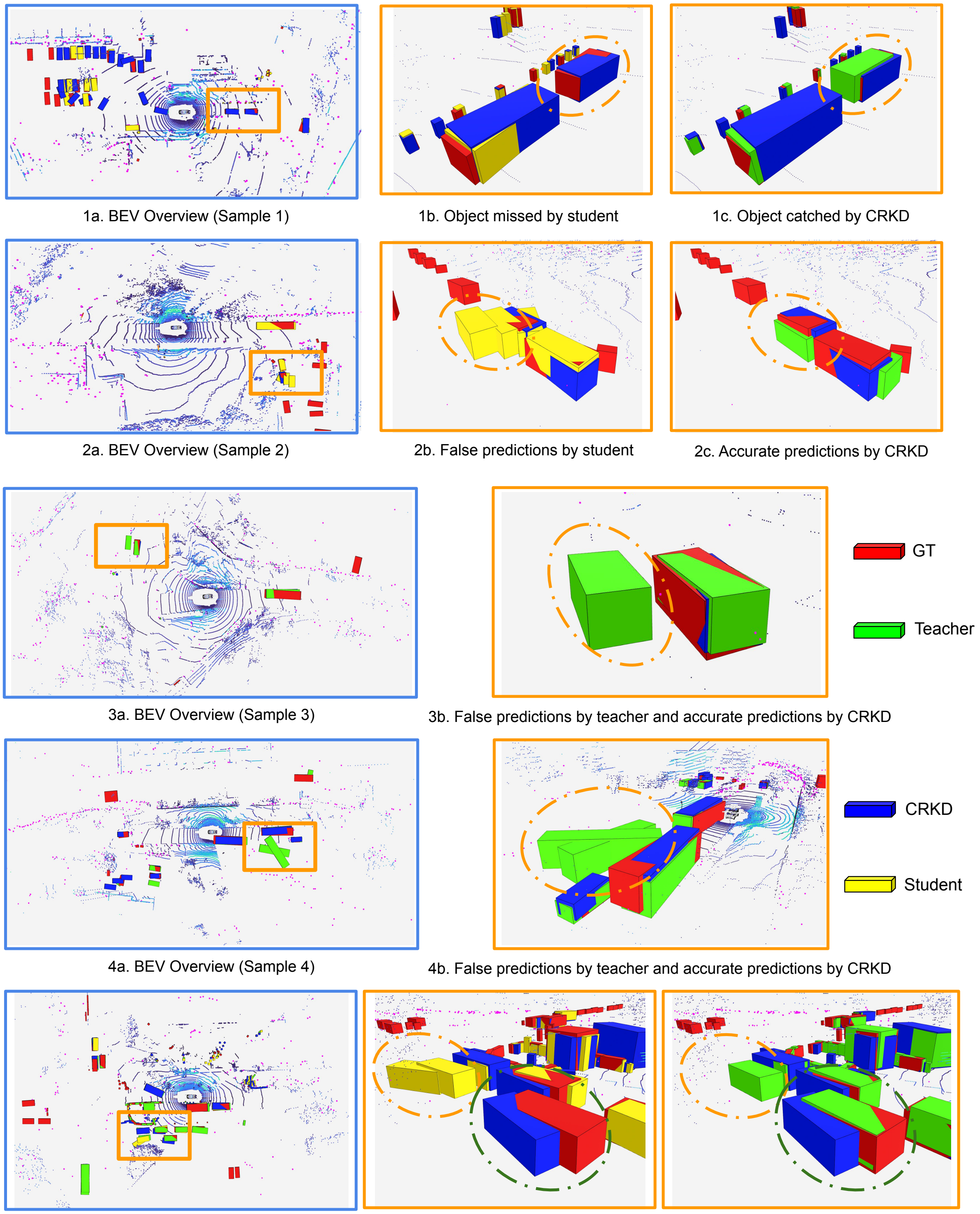
To better study the most suitable choice for the Response Distillation (RespD) module, we design an experiment trying to answer an insightful question: is the cross-modality distillation most helpful in amplifying the strength of the student or mitigating the weakness of the student? It is widely recognized that radars are more capable of perceiving dynamic objects [6, 11, 16]. Therefore, the CR student may benefit from the radar's strength. As we have the flexibility of varying the loss weight w_i for different classes in RespD, we experiment with different loss weight settings. In addition to the ablation study of RespD reported in the main paper, we conduct an experiment with static setting where the loss weights for static classes are set to 2 while the weights for the other classes are set to 1. In the static setting, priority is given to the static classes, which radars are less capable of detecting. As shown in Tab. 9, the RespD module works better when we prioritize the learning of dynamic objects, which indicates that the RespD module is more effective when designed to be amplifying the strength of the student detector. The results complement the ablation study we show in the main manuscript, demonstrating the effectiveness of the proposed dynamic RespD module. We hope this interesting finding could provide some guidance to future study about designing cross-modality distillation to leverage the strength of different modalities effectively.

Module	Vanilla	Static	Dynamic	mAP↑	NDS↑
RespD	✓			45.3	56.7
		✓		45.4	56.6
			✓	45.7	56.7

Table 9. Ablation study of Response Distillation (RespD) with different weight settings.

B.6. Additional Qualitative Results

We show additional qualitative results of CRKD in Fig. 2. In the first two samples (sample 1 and 2), we firstly show that CRKD is able to outperform the student model since its predictions are more aligned with the ground truth. We credit this improvement to the effective design of CRKD. We also show additional examples (sample 3 and 4) where CRKD can even outperform the teacher detector thanks to the long-range detection capability of radars. In the last sample frame (sample 5), we show that CRKD is capable of capturing the object that is missed by the student model. In addition, it is also demonstrated that CRKD is able to maintain accurate predictions where the teacher and student models generate false predictions.



1a. BEV Overview (Sample 1)

1b. Object missed by student

1c. Object catched by CRKD

2a. BEV Overview (Sample 2)

2b. False predictions by student

2c. Accurate predictions by CRKD

3a. BEV Overview (Sample 3)

3b. False predictions by teacher and accurate predictions by CRKD

4a. BEV Overview (Sample 4)

4b. False predictions by teacher and accurate predictions by CRKD

5a. BEV Overview (Sample 5)

5b. Object missed and false predictions by student

5c. Object catched and accurate predictions by CRKD

Figure 2. More Qualitative results on nuScenes. We show zoomed-in views in panel b and c for the highlighted regions in panel a, with the border dash as the correspondence. The highlighted regions are enclosed with border dash in ellipse. We show the ground truth annotation in red, teacher prediction in green, student prediction in yellow, CRKD prediction in blue, and radar points in magenta. In (1a) to (1c), we show an example frame where CRKD can capture the object missed by the student with the guidance of the teacher. In (2a) to (2c) we show an example frame where CRKD can reject false predictions by the student model. In (3a) to (3b) and (4a) to (4b), we show two examples where CRKD rejects false predictions by the teacher model and generates more accurate predictions. In (5a) to (5c), we show an example where CRKD outperforms both the teacher and student models by capturing missed objects and generating less false predictions. Best viewed on screen and in color.

References

[1] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *CVPR Workshop*, 2023. 4

[2] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. In *ICLR*, 2023. 4

[3] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`, 2020. 1

[4] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, 2022. 1

[5] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1

[6] Yao Li, Jiajun Deng, Yu Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. ezfusion: A close look at the integration of lidar, millimeter-wave radar, and camera for accurate 3d object detection and tracking. In *IEEE RAL*, 2022. 2, 4

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[8] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. 1, 4

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *WACV*, 2021. 4

[12] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In *MDPI Sensors*, 2018. 1, 2

[13] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *ECCV*, 2020. 2

[14] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 1

[15] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Jing Zhang, and Dacheng Tao. Bevsimdet: Simulated multi-modal distillation in bird's-eye view for multi-view 3d object detection. *arXiv preprint arXiv:2303.16818*, 2023. 1, 4

[16] Yi Zhou, Lulu Liu, Haocheng Zhao, Miguel López-Benítez, Limin Yu, and Yutao Yue. Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. In *MDPI Sensors*, 2022. 2, 4

[17] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 1