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

Lingjun Zhao*, Jingyu Song*†, Katherine A. Skinner
University of Michigan, Ann Arbor, MI USA
{lingjunz,jingyuso,kskin}@umich.edu

Abstract

In the field of 3D object detection for autonomous driving, LiDAR-Camera (LC) fusion is the top-performing sensor configuration. Still, LiDAR is relatively high cost, which hinders adoption of this technology for consumer automobiles. Alternatively, camera and radar are commonly deployed on vehicles already on the road today, but performance of Camera-Radar (CR) fusion falls behind LC fusion. In this work, we propose Camera-Radar Knowledge Distillation (CRKD) to bridge the performance gap between LC and CR detectors with a novel cross-modality KD framework. We use the Bird's-Eye-View (BEV) representation as the shared feature space to enable effective knowledge distillation. To accommodate the unique cross-modality KD path, we propose four distillation losses to help the student learn crucial features from the teacher model. We present extensive evaluations on the nuScenes dataset to demonstrate the effectiveness of the proposed CRKD framework. The project page for CRKD is `https://song-jingyu.github.io/CRKD`.

1. Introduction

Perception is an important module for achieving safe and effective autonomous driving [3, 15, 55]. 3D object detection is an essential task in perception as it is of great significance for subsequent tasks [44, 47, 50, 54]. Among various perceptual sensors used by autonomous driving researchers, LiDAR, camera and radar are the most common choices to enable autonomy on the road [50]. Sensor fusion techniques are usually used to improve the detector's performance and robustness. LiDAR-Camera (LC) fusion has been widely demonstrated as the top-performing sensor configuration for 3D object detection [1, 2, 7, 14, 38, 61]. However, the high cost of LiDAR constrains the wide application of this configuration. Though Camera-Only (CO) detectors have demonstrated impressive performance in recent Bird's-Eye-View (BEV) based frameworks [16, 17, 33, 34], the camera's vulnerability to lighting conditions and lack of accurate depth measurements motivates researchers to turn to other sensors such as radar. Radar is robust to varying weather and lighting conditions and features automotive-grade design and low cost. Radars are already highly accessible on most cars equipped with driver assistance features. Compared with LiDAR, the radar measurements are sparse and noisy, which makes designing a Camera-Radar (CR) detector challenging. Recent CR detectors have leveraged the advancement brought by BEV-based Camera-Only (CO) detectors [16, 17, 33, 34] to achieve further improvement in accuracy and robustness to weather and lighting changes [28, 70].

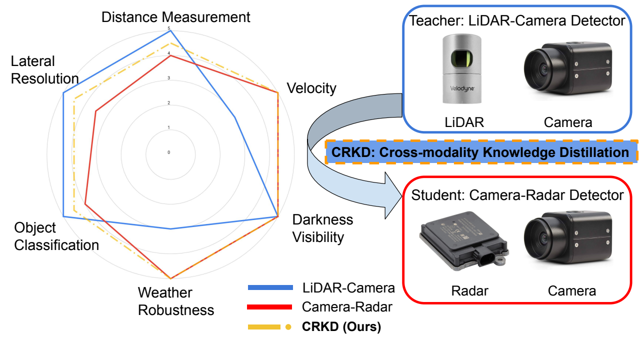

Figure 1. We propose CRKD to conduct a novel cross-modality knowledge distillation path from a LiDAR-camera teacher to a camera-radar student. We present a radar chart to illustrate the complementary nature of these sensing configurations and the improvement that CRKD can enable.

Despite the advancement in architecture design, there is still a distinct performance gap when comparing LiDAR-Only (LO) and LC detectors against CO and CR detectors. Recent research has focused on applying the Knowledge Distillation (KD) technique to alleviate this gap [12, 13, 26, 52, 68]. Generally, KD features a teacher-student framework that aims to propagate the informative knowl-

*Equal contribution.
†Corresponding author.

This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N028603.

edge from a well-performing teacher model to facilitate the learning process of the student model. This usually leads to improved performance compared to simply training the student model on the same task. The KD technique has been employed in either intra-modal [29, 53, 64, 66] or cross-modal [5, 6, 13, 18, 26, 68] configurations for 3D object detection. Though many cross-modal methods use a single-modality detector as the teacher model to leverage the privileged LiDAR data that is widely available in open-source datasets, they mainly focus on distilling knowledge to a LiDAR-based or camera-based student detector. We argue the importance of designing a distillation path from an LC teacher detector to a CR student detector, which could benefit from the existing superior design of LC detectors and the shared point cloud representation between measurements of LiDAR and radars [59].

Inspired by the above observations, we propose **CRKD**: an enhanced **C**amera-**R**adar 3D object detector with cross-modality **K**nowledge **D**istillation (Fig. 1) that distills knowledge from an LC teacher detector to a CR student detector. To our best knowledge, CRKD is the first KD framework that supports a fusion-to-fusion distillation path. As the LiDAR sensor is used only during training, we emphasize the value of CRKD as it could facilitate the practical application of perceptual autonomy with a low-cost and robust CR sensor configuration.

To summarize, our main contributions are as follows:

- We propose a novel cross-modality KD framework to enable LC-to-CR distillation in the BEV feature space. With the transferred knowledge from an LC teacher detector, the CR student detector can outperform existing baselines without additional cost during inference.
- We design four KD modules to address the notable discrepancies between different sensors to realize effective cross-modality KD. As we operate KD in the BEV space, the proposed loss designs can be applied to other KD configurations. Our improvement also includes adding a gated network to the baseline model for adaptive fusion.
- We conduct extensive evaluation on nuScenes [2] to demonstrate the effectiveness of CRKD. CRKD can improve the mAP and NDS of student detectors by 3.5% and 3.2% respectively. As our method focuses on a novel KD path with a large modality gap, we provide thorough study and analysis to support our design choices.

2. Related Works

2.1. Multi-modality 3D Object Detection

The multi-modality 3D object detectors generally outperform single-modality detectors in accuracy and robustness as the perceptual sensors (e.g., LiDAR, camera and radar) can complement each other [50]. Among the common sensor combinations, LC is the best-performing modality configuration on most existing datasets [2, 7, 9, 40]. In general, LiDAR and camera are fused in different ways. One trend is to augment LiDAR points with features from cameras [19, 46, 48, 49, 57], which is usually referred to as early fusion. Other solutions apply deep feature fusion in a shared representation space [31, 32, 63]. One trending choice is to leverage the BEV space to deliver impressive improvement [36, 38]. There are also methods that fuse information at a later stage. In [1, 4, 58], features are independently extracted and aggregated via proposals or queries in the detection head, while some methods [42, 43] combine the output candidates from single-modality detectors.

Nevertheless, the LC configuration is less accessible to consumer cars due to the high cost of LiDAR. Thus, the potential for CR detectors stands out due to the robustness brought by radars and the potential of large-scale deployment. One of the key challenges facing CR detectors is how to handle the discrepancy in sensor views and data returns. CenterFusion [41] applies feature-level fusion by associating radar points with image features via a frustum-based method. Following this, more feature-level fusion methods are proposed in [8, 23, 56, 69]. Recently, as many camera-based methods [33, 34] have started to leverage the unified BEV space by transforming camera features from perspective view (PV), many CR detectors also explore fusion of camera and radar features in the BEV space [22, 25]. Though LiDAR data is not available in real inference and deployment, its wide appearance in open-source dataset [2] has motivated researchers to leverage LiDAR data to guide the feature transformation process [28]. Motivated by the aforementioned works, CRKD also unifies features in the BEV space and leverages LiDAR data in a KD-based framework. To our best knowledge, CRKD is the first framework that improves CR detectors with cross-modality KD from a top-performing LC teacher detector.

2.2. Cross-modality Knowledge Distillation

The idea of KD is initially proposed in [12] for model compression in an image classification task. It is then extended to the field of object detection for model compression and performance improvement [35]. Specifically, in the field of 3D object detection, a group of KD methods requires that the teacher and student models use the same modality such as LO-to-LO (L2L) [21, 51, 53, 60, 66] and CO-to-CO (C2C) [29, 64, 65]. In contrast, cross-modality KD focuses on KD with different modality configurations. Typical paths include LiDAR-to-Camera (L2C) [5, 6, 10, 13, 18, 31, 39] and Camera-to-LiDAR (C2L) [45, 68]. Recently, new cross-modality KD paths including a fusion-based modality have been explored. In UniDistill [68], a universal framework that supports multiple KD paths is proposed. By unifying the features from different modalities to the shared BEV space, it supports L2C, C2L, LC-to-LO (LC2L) and LC-to-Camera (LC2C). DistillBEV [52] also supports L2C and LC2C by leveraging the shared BEV space. X3KD [26]

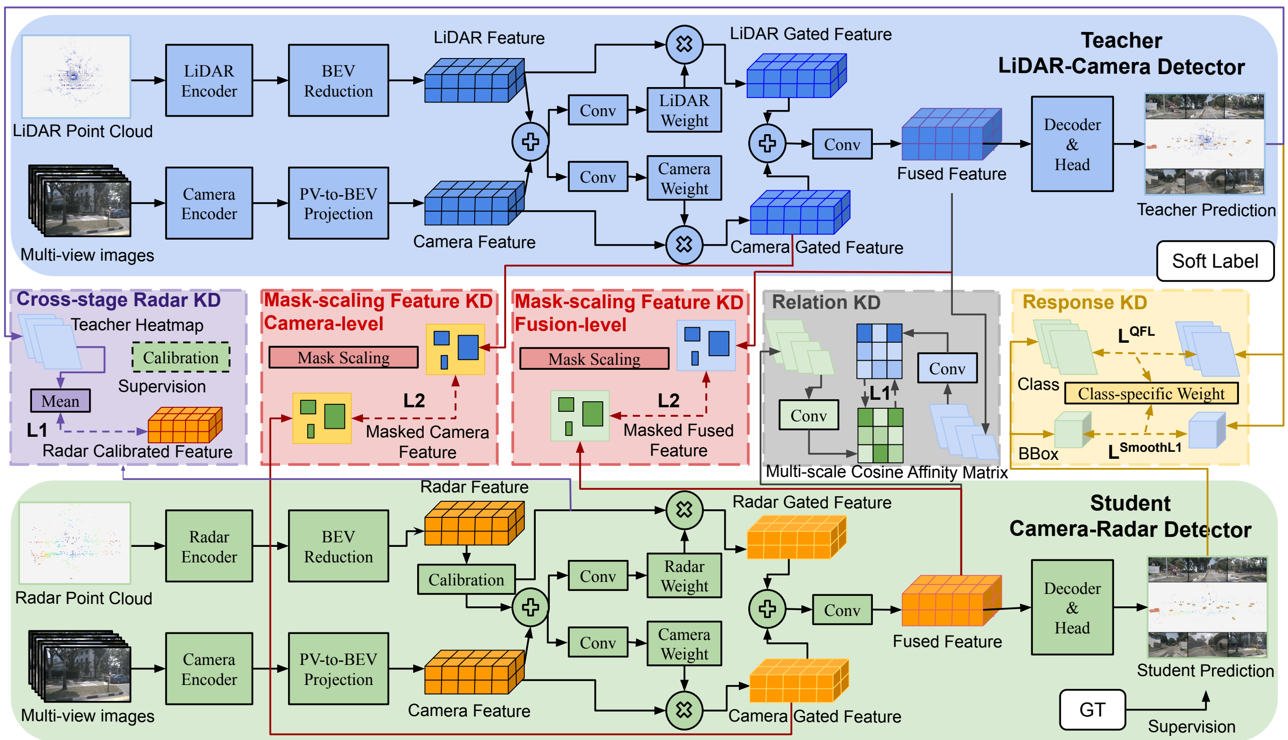

Figure 2. An overview of the proposed cross-modality (LC-to-CR) distillation framework CRKD. We narrow down the modality gap by unifying both the teacher and the student into the BEV space with similar 3D object detector structure. We refine the model design to enable adaptive fusion and design four novel distillation losses for effective cross-modality KD. During inference, only CR input is needed.

proposes a cross-modal and cross-task KD framework for L2C KD. It is evaluated on an LO-to-CR (L2CR) KD path as a supplementary task. However, it lacks specific consideration for the large domain discrepancies with radars and further experiments and analysis with the L2CR KD path. Among the existing prior works, we observe the lack of KD methods that can support a fusion-to-fusion distillation path and handle domain differences with radars. We argue the importance of conducting distillation from an LC teacher to a CR student to leverage the shared BEV feature space of camera-based detectors and the shared point cloud representation of LiDAR and radar measurements. According to our best knowledge, we are the first to investigate a KD framework with a fusion-to-fusion path. We demonstrate that our novel framework improves the detection performance of CR detectors comprehensively.

3. Method

We show an overview of CRKD in Fig. 2. We set up the teacher and student models with a similar BEV-based encoder-decoder head architecture. Taking advantage of the shared BEV feature space, we build CRKD based on the highly optimized BEVFusion [38] codebase. We use BEVFusion-LC as the teacher model and BEVFusion-CR as the baseline student model. The detector head in both models is set as CenterHead [62] for response KD.

To account for the challenging cross-modality fusion-to-fusion KD, we design several KD modules. We propose cross-stage radar distillation with a learning-based calibration module to enable the radar encoder to learn a more accurate scene-level object distribution. A mask-scaling feature KD is designed for feature imitation on foreground regions while accounting for inaccurate view transformation to BEV features for objects that are far from the sensor and dynamic. We apply a relation KD to maintain relation consistency in scene-level geometry. In addition, we improve the response KD design with class-specific loss weights to better leverage the CR model's ability to capture dynamic objects. Details of the proposed KD modules will be discussed in the following sections.

3.1. Model Architecture Refinement

We add a gated network [20, 47, 63] to BEVFusion [38] to enable the model to learn to generate attention weights on the single-modality feature maps to fuse the complementary modalities adaptively. Specifically, the gated network learns the gated features as follows:

$$\tilde{F}_{M_1} = F_{M_1} \times \sigma(\mathrm{Conv}_{M_1}(\mathrm{Concat}(F_{M_1}, F_{M_2}))), \quad (1)$$

$$\tilde{F}_{M_2} = F_{M_2} \times \sigma(\text{Conv}_{M_2}(\text{Concat}(F_{M_1}, F_{M_2}))), \quad (2)$$

where $\tilde{F}_{M_1}$ and $\tilde{F}_{M_2}$ are the gated features for modality M_1 and M_2, F_{M_1} and F_{M_2} are the input feature maps to the gated network from the backbone and view transform module of modality M_1 and M_2, respectively, σ denotes the sigmoid function, and Conv_{M_1} and Conv_{M_2} are two separate convolution layers for M_1 and M_2 that could learn channel-wise attention weights for the input features. The output of the gated network is further fused by a convolutional fusion module in BEVFusion [38]. We apply the adaptive gated network to our teacher and student models to learn the relative importance between input modalities. This modification improves the detection performance of the teacher and student models, and also makes feature-based distillation more effective since the gated feature maps encode informative scene geometry from both input modalities.

In the LC teacher model, we denote the camera feature map as $F_c^T \in \mathbb{R}^{C_c^T \times H \times W}$ and the LiDAR feature map as $F_l^T \in \mathbb{R}^{C_l^T \times H \times W}$. Similarly, the camera and radar features in the student model are denoted as $F_c^S \in \mathbb{R}^{C_c^S \times H \times W}$ and $F_r^S \in \mathbb{R}^{C_r^S \times H \times W}$, respectively. We denote the fused feature map in the teacher model and the student model as F_f^T and F_f^S. We keep the spatial dimension $H \times W$ consistent for all feature maps, and also set $C_c^T = C_c^S$ and $C_l^T = C_l^S$ for feature mimicking across the feature dimension.

3.2. Cross-Stage Radar Distillation (CSRD)

Though the measurements of radar and LiDAR are both represented as point clouds, the physical meaning behind them are slightly different. Compared with LiDAR, radar points are much sparser and can be interpreted as a list of object-level points with velocity measurements [50, 59], while LiDAR is denser and captures geometry-level information. Observing this gap, we argue the common method of direct feature imitation may not work well in this scenario. Instead, as radar measurements are sparse and represent scene-level object distribution, we propose a novel Cross-Stage Radar Distillation (CSRD) method. Specifically, we design a distillation path between the radar feature map and the scene-level objectness heatmap predicted by the LC teacher model, which is denoted as $Y^T \in \mathbb{R}^{K \times H \times W}$, where K is the number of classes. Since radars are generally believed to be noisy in the range and azimuth angle measurements, we design a calibration module to learn to compensate the noise. Specifically, we pass F_r^S to three blocks of convolutional layer, batch normalization and ReLU activation with kernel size 3×3. We add another 1×1 convolution layer to project the calibrated feature map to $\hat{F}_r^S \in \mathbb{R}^{H \times W}$. The CSRD loss $\mathcal{L}_{csrd}$ is formed as follows:

$$\mathcal{L}_{csrd} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \| \hat{Y}^T_{i,j} - \hat{F}^S_{r\ i,j} \|_1, \quad (3)$$

where $\hat{Y}^T \in \mathbb{R}^{H \times W}$ is obtained by taking the mean along the K dimension of $\sigma(Y^T)$.

3.3. Mask-Scaling Feature Distillation (MSFD)

We propose feature distillation for aligning the camera feature maps and the fused feature maps. It has been acknowledged in many works [5, 6, 67, 68] that direct feature imitation between teacher and student models may not work effectively in 3D object detection tasks due to the notable imbalance between the foreground and background. Therefore, a common fix to this issue is to generate a mask $M \in \mathbb{R}^{H \times W}$ to only distill information from the foreground region. Meanwhile, more works have demonstrated that the boundary region of the foreground can also contribute to effective KD [5, 67]. We follow this finding and propose Mask-Scaling Feature Distillation (MSFD) that is aware of object range and movement. For the student CR model, the detection performance is mainly dependent on the depth prediction for images and the geometric accuracies of radar points. Since the range and object movement can cause extra challenges for view transformation to BEV space, we scale up the area of the foreground region to account for the potential misalignment. We increase the width and length of the mask by α and β if the objects are in the range groups $[r_1, r_2]$ and $[r_2, \infty]$. We also increase the width and length by α and β if the velocities along that axis are within $[v_1, v_2]$ and $[v_2, \infty]$. In practice, We clip the increase of object size within a pre-defined range to balance between different sizes of objects. We form the MSFD loss as follows:

$$\mathcal{L}_{msfd} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} M_{i,j} \| F^T_{i,j} - F^S_{i,j} \|_2, \quad (4)$$

where F^T and F^S represent the feature maps in the teacher and student model, respectively. We compute MSFD loss for the gated camera feature maps ($\tilde{F}_c^T$ and $\tilde{F}_c^S$) and the fused feature maps (F_f^T and F_f^S).

3.4. Relation Distillation (RelD)

While the aforementioned CSRD and MSFD can handle feature-level distillation effectively, we follow MonoDistill [6] to highlight the importance of maintaining similar geometric relations in the scene level between the teacher and student models. We compute the affinity matrix describing cosine similarity of the fused feature map. We propose the RelD loss as follows:

$$C_{i,j} = \frac{F_i^\top F_j}{\|F_i\|_2 \cdot \|F_j\|_2}, \quad (5)$$

where $C_{i,j}$ denotes the cosine similarity value at (i, j) in the affinity matrix, and F_i and F_j represent the i^{th} and j^{th} feature map separately. The scene-level information gap between the student and teacher models can be computed as the $\mathcal{L}_1$ norm between their respective affinity matrices. We refer to the RelD loss as $\mathcal{L}_{reld}$, as shown below:

$$\mathcal{L}_{reld} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \| C_{i,j}^T - C_{i,j}^S \|_1, \tag{6}$$

where H and W represent the BEV spatial size, and C^T and C^S denote the affinity matrix of the student and teacher model, respectively. In CRKD, we compute $\mathcal{L}_{reld}$ between the fused feature maps of the teacher and student models since they are the input to the decoder and detector heads. The refined feature maps with distilled relation information could improve the detection performance. Moreover, in order to distill the scene-level relation information of different scales, we apply a downsampling operation and a convolutional block. Then we use these multi-level feature maps to calculate the multi-scale RelD losses and take the average value as the final loss term.

3.5. Response Distillation (RespD)

Response Distillation has been proven effective in image classification [12] and 3D object detection [13, 52, 68]. The predictions inferred by the teacher are served as the soft labels for the student. The soft labels and the hard labels are combined to supervise the learning of the student model. We refer to the RespD design in CMKD [13] and improve it to be aware of modality strength. Since radar has the unique advantage of direct velocity measurements due to the Doppler effect [59, 70], we set larger weights for dynamic classes in RespD to allow higher priority for dynamic objects to leverage the student CR model's strength. The loss for dynamic response distillation is denoted as $\mathcal{L}_{resp}$, consisting of the classification loss $\mathcal{L}_{cls}$ and the regression loss $\mathcal{L}_{reg}$. $\mathcal{L}_{cls}$ is for object categories and is computed using the Quality Focal Loss (QFL) [30]. $\mathcal{L}_{reg}$ is for 3D bounding box regression and can be obtained by calculating the $SmoothL1$ loss. We compute these two losses as:

$$\mathcal{L}_{cls} = \sum_{i=1}^{K} QFL(P_{C_i}^T, P_{C_i}^S) \times w_i, \tag{7}$$

$$\mathcal{L}_{reg} = \sum_{i=1}^{K} Smooth\mathcal{L}1(P_{B_i}^T, P_{B_i}^S) \times w_i, \tag{8}$$

where $P_{C_i}^T$ and $P_{C_i}^S$ denote the classification predictions of the i^{th} task generated by the teacher and student model, and $P_{B_i}^T$ and $P_{B_i}^S$ denote the regression predictions in the teacher and student models. K is the number of tasks in the Centerhead [62], and w_i represents the class-specific weights.

3.6. Overall Loss Function

We combine the proposed KD loss and standard 3D object detection loss $\mathcal{L}_{det}$. The overall loss function we apply in the training stage is:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{csrd} + \lambda_2 \cdot \mathcal{L}_{msfd} + \lambda_3 \cdot \mathcal{L}_{reld} + \lambda_4 \cdot \mathcal{L}_{respd} + \lambda_5 \cdot \mathcal{L}_{det}, \tag{9}$$

where λ_1 through λ_5 are hyperparameters we set for weighting different loss components.

4. Experiments

4.1. Experimental Setup

We evaluate our method on the nuScenes dataset [2] as the three modalities (i.e., LiDAR, camera and radar) are all available. We follow the official split that has 700 scenes for training and 150 scenes for validation. We set $[-54m, -54m, -5m] \times [54m, 54m, 3m]$ as the region to conduct object detection. We use the mean-Average-Precision (mAP) and NuScenes Detection Score (NDS) [2] as the main evaluation metrics. We also report the True Positive (TP) metrics [2] for comprehensive evaluation.

Our implementation is based on the MMDetection3D codebase [7]. All of our experiments are conducted using 4 NVIDIA A100 GPUs. We set the default BEVFusion-LC with Centerhead [62] as the teacher model. As mentioned in Sec. 3.1, we add the adaptive gated network to the baseline BEVFusion-CR model and denote it as BEVFusion-CR*. We set BEVFusion-CR* as the student model. We set the camera backbone as SwinT [37] and image resolution as 256×704 for both the teacher and student models. We also test CRKD with a ResNet R50 backbone [11] for comprehensive evaluation. The BEV spatial size is set to 180×180. We use PointPillars [27] as the backbone of the radar branch in BEVFusion-CR*. We set the same Centerhead [62] as the detector head for the student model. During distillation, we freeze the teacher model and train the student model for 20 epochs. We set the batch size as 8 and learning rate as 1e-4. We include more implementation details and results in the supplementary material.

4.2. Quantitative Results

We give an overall comparison of CRKD with existing CO and CR detectors with single-frame image input on nuScenes [2]. We follow common practices [5, 26, 52, 68] to show a complete comparison on both the *val* and *test* splits. As shown in Tab. 1, CRKD is the top-performing model in most metrics on the *val* set of nuScenes. We also present the performance of CRKD on the *test* split. The results show that CRKD has the best or second best performance on most metrics without using any test-time optimization techniques (e.g., test-time augmentation, larger

Set	Method	Modality	Backbone	Resolution	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
val	BEVFormer-S [34]	C	R101	900 × 1600	37.5	44.8	0.725	0.272	**0.391**	0.802	0.200
	BEVDet [17]	C	R50	256 × 704	29.8	37.9	0.725	0.279	0.589	0.860	0.245
	RCM-Fusion [22]	C+R	R101	900 × 1600	44.3	52.9	-	-	-	-	-
	CenterFusion [41]	C+R	DLA34	450 × 800	33.2	45.3	0.649	**0.263**	0.535	0.540	**0.142**
	CRAFT [24]	C+R	DLA34	448 × 800	41.1	51.7	0.494	0.276	0.454	0.486	0.176
	RCBEV [69]	C+R	SwinT	256 × 704	37.7	48.2	0.534	0.271	0.558	0.493	0.209
	BEVFusion [38]	C+R	SwinT	256 × 704	43.2	54.1	0.489	0.269	0.512	**0.313**	0.171
	UVTR (L2C) [31]	C♢	R101	900 × 1600	37.2	45.0	0.735	0.269	0.397	0.761	0.193
	BEVDistill (BEVFormer-S) [5]	C♢	R50	640 × 1600	38.6	45.7	0.693	0.264	0.399	0.802	0.199
	UniDistill (LC2C) [68]	C♢	R50	256 × 704	26.5	37.8	-	-	-	-	-
	BEVSimDet [67]	C♢	SwinT	256 × 704	40.4	45.3	0.526	0.275	0.607	0.805	0.273
	X3KD (LC2C) [26]	C♢	R50	256 × 704	39.0	50.5	0.615	0.269	0.471	0.345	0.203
	DistillBEV (BEVDet) [52]	C♢	R50	256 × 704	34.0	41.6	0.704	0.266	0.556	0.815	0.201
	X3KD (L2CR) [26]	C+R♢	R50	256 × 704	42.3	53.8	-	-	-	-	-
	CRKD	C+R♢	R50	256 × 704	43.2	54.9	0.450	0.267	0.442	0.339	0.176
	CRKD	C+R♢	SwinT	256 × 704	**46.7**	**57.3**	**0.446**	**0.263**	0.408	0.331	0.162
test	BEVFormer-S [34]	C	R101	900 × 1600	40.9	46.2	0.650	0.261	0.439	0.925	0.147
	BEVDet† [17]	C	SwinT	640 × 1600	42.4	48.2	0.528	**0.236**	0.395	0.979	0.152
	RCM-Fusion [22]	C+R	R101	900 × 1600	**49.3**	58.0	0.485	0.255	0.386	0.421	0.115
	CenterFusion† [41]	C+R	DLA34	450 × 800	32.6	44.9	0.631	0.261	0.516	0.614	0.115
	CRAFT† [24]	C+R	DLA34	448 × 800	41.1	52.3	0.467	0.268	0.456	0.519	0.114
	RCBEV [69]	C+R	SwinT	256 × 704	40.6	48.6	0.484	0.257	0.587	0.702	0.140
	UVTR (L2C) [31]	C♢	V2-99	900 × 1600	45.2	52.2	0.612	0.256	**0.385**	0.664	0.125
	X3KD (LC2C) [26]	C♢	R101	640 × 1600	45.6	56.1	0.506	0.253	0.414	**0.366**	0.131
	UniDistill (LC2C) [68]	C♢	R50	256 × 704	29.6	39.3	0.637	0.257	0.492	1.084	0.167
	X3KD (L2CR) [26]	C+R♢	R50	256 × 704	44.1	55.3	-	-	-	-	-
	CRKD	C+R♢	SwinT	256 × 704	48.7	**58.7**	**0.404**	0.253	0.425	0.376	**0.111**

Table 1. Comparison on the nuScenes dataset. We group methods based on modality and whether KD is used. We include existing SOTA works that use single-frame image input for fair comparison. Methods [5, 52, 67] missed in the *test* split group do not report their results with single-frame image input. The proposed CRKD outperforms the baseline methods in most metrics. ♢ denotes the distillation-based methods. † denotes using test time augmentation. The best is **bolded** and the second best is underlined.

image resolution). Overall, CRKD is the most consistent in achieving high performance across all of the baselines.

We also show a complete comparison of per-class AP in Tab. 2 to break down the improvement brought by CRKD. The results show that CRKD achieves consistent improvement in AP of all the classes. There is an interesting finding that larger gains of CRKD come from dynamic classes, which indicates that CRKD successfully helps the student model leverage its strength on dynamic object detection more effectively due to the availability of direct velocity measurements from radar.

4.3. Ablation Studies

To further break down the improvement brought from each module we design, extensive experiments are conducted to discuss and validate our design choices. We firstly present the main ablation study in Tab. 3. It can be observed that all the proposed modules have contributed to the superior performance of CRKD. Among the four proposed KD losses, we see the most improvement comes from RespD, which indicates the significance of RespD in cross-modality KD. The other three losses contribute more to the improvement of mAP. This finding validates our design objectives in improving object localization as CSRD and MSFD supervise on the feature maps, and RelD tries to align the scene-level geometric relation.

We next demonstrate the experiments we conduct to validate the design choices of each module. As mentioned before, response distillation (RespD) brings the most improvement among all the proposed KD modules. Our empirical finding indicates that for other KD modules, the best-performing design alone may not be the best choice that works with RespD and other modules. We conjecture that this inconsistency in performance gain comes from the considerably large domain discrepancies between LC and CR. We would like to highlight the importance of RespD in the cross-modality KD work and use the experiments of combining different modules with RespD to guide the overall design of CRKD. Therefore, for the following ablation study in Tab. 4, unless otherwise mentioned, we show results of experiments with using RespD together. All the ablation study instances are trained using the same setting as the full model.

4.3.1 Effect of CSRD

We conduct an ablation study to validate the proposed CSRD module. As mentioned in Sec. 3.2, the radar points

Model	Modality	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bicycle	TC	Barrier	mAP↑
Teacher	L+C	88.4	62.4	73.8	40.6	29.2	78.7	75.3	65.8	74.9	72.3	66.1
Baseline	C+R	72.1	37.8	48.9	18.3	12.6	48.4	42.0	33.8	58.8	59.6	43.2
Student	C+R	72.2	41.3	51.0	19.2	15.2	49.0	46.2	35.5	59.1	**60.1**	44.9
CRKD	C+R	**74.8(+2.7)**	**44.1(+6.3)**	**53.6(+4.7)**	**20.6(+2.3)**	**16.9(+4.3)**	**50.6(+2.2)**	**46.8(+4.8)**	**38.2(+4.4)**	**61.5(+2.7)**	**60.1(+0.5)**	**46.7(+3.5)**

Table 2. Comparison of the per-class AP results of the BEVFusion-LC (teacher), BEVFusion-CR (baseline), BEVFusion-CR* (student) and CRKD models on the nuScenes *val* split. We quantitatively show the improvement made by CRKD over the baseline.

Model	Gated	RespD	CSRD	MSFD	RelD	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
Baseline						43.2	54.1	0.489	0.269	0.512	0.313	0.171
CRKD	✓					44.9	55.9	0.464	0.267	0.458	0.304	0.165
	✓	✓				45.7	56.7	0.448	0.262	0.409	0.330	0.166
	✓	✓	✓			46.0	57.0	0.445	0.261	0.407	0.326	0.163
	✓	✓	✓	✓		46.2	57.2	0.439	0.260	0.394	0.332	0.166
	✓	✓	✓	✓	✓	46.7	57.3	0.446	0.263	0.408	0.331	0.162

Table 3. Ablation study of the proposed modules in CRKD evaluated on the nuScenes *val* split. The baseline is the BEVFusion-CR model.

Module	LiDAR	Heatmap	mAP↑	NDS↑
CSRD	✓		44.9	56.3
		✓	46.0	57.0

(a) Ablation study of Cross-stage Radar Distillation (CSRD).

Module	Mask	Mask-scaling	mAP↑	NDS↑
MSFD	✓		46.0	56.7
		✓	46.2	56.9

(b) Ablation study of Mask-scaling Feature Distillation (MSFD).

Module	Vanilla	Adapt	mAP↑	NDS↑
RelD	✓		45.9	56.9
		✓	46.2	57.0

(c) Ablation study of Relation Distillation (RelD).

Module	Vanilla	Dynamic	mAP↑	NDS↑
RespD	✓		45.3	56.7
		✓	45.7	56.7

(d) Ablation study of Response Distillation (RespD).

Table 4. Ablation Study of single distillation modules on the nuScenes *val* split.

Fuser	*In Channels*	*Out Channels*	*mAP↑*	*NDS↑*
Conv	80 + 256	256	43.2	54.1
Gated	64 + 64	64	44.2	54.3
	128 + 128	256	44.4	54.7
	80 + 256	256	44.9	55.9

Table 5. Ablation study of fusion module design and number of channels of the student model on the nuScenes *val* split.

represent object-level information. Therefore, the common practice to distill information at the same stage of the network may not work well for radar feature maps. We propose CSRD to add cross-stage supervision on the object-level information. We demonstrate the model with CSRD outperforms that using LiDAR feature maps as the distillation source in Tab. 4a. The significant improvement from CSRD validates that the higher-level objectness heatmap provides more suitable guidance to distill radar features.

4.3.2 Effect of MSFD

We conduct a comparison between the proposed mask-scaling strategy accounting for object range and velocity and the common foreground mask of ground truth bounding boxes. In Tab. 4b, the improvement achieved with the proposed mask-scaling strategy verifies that our mask-scaling strategy in MSFD helps to achieve more effective KD.

4.3.3 Effect of RelD

We study the effect of applying a convolutional block after the downsampling operation for the feature maps used in RelD. We name the instance with the convolution layer as Adapt and the baseline instance as Vanilla in Tab. 4c. The results verify our design choice.

4.3.4 Effect of RespD

Though RespD has been a widely used loss term in several KD works, we are the first to weight the RespD loss differently based on different classes. In practice, we set w_i as 2 for the dynamic classes and 1 for static classes. This design allows the training supervision to prioritize dynamic classes, which radars are more capable of detecting. In the vanilla setting we set w_i as 1 for all classes. As shown in Tab. 4d, applying the class-specific weights helps to improve the overall performance of the student detector.

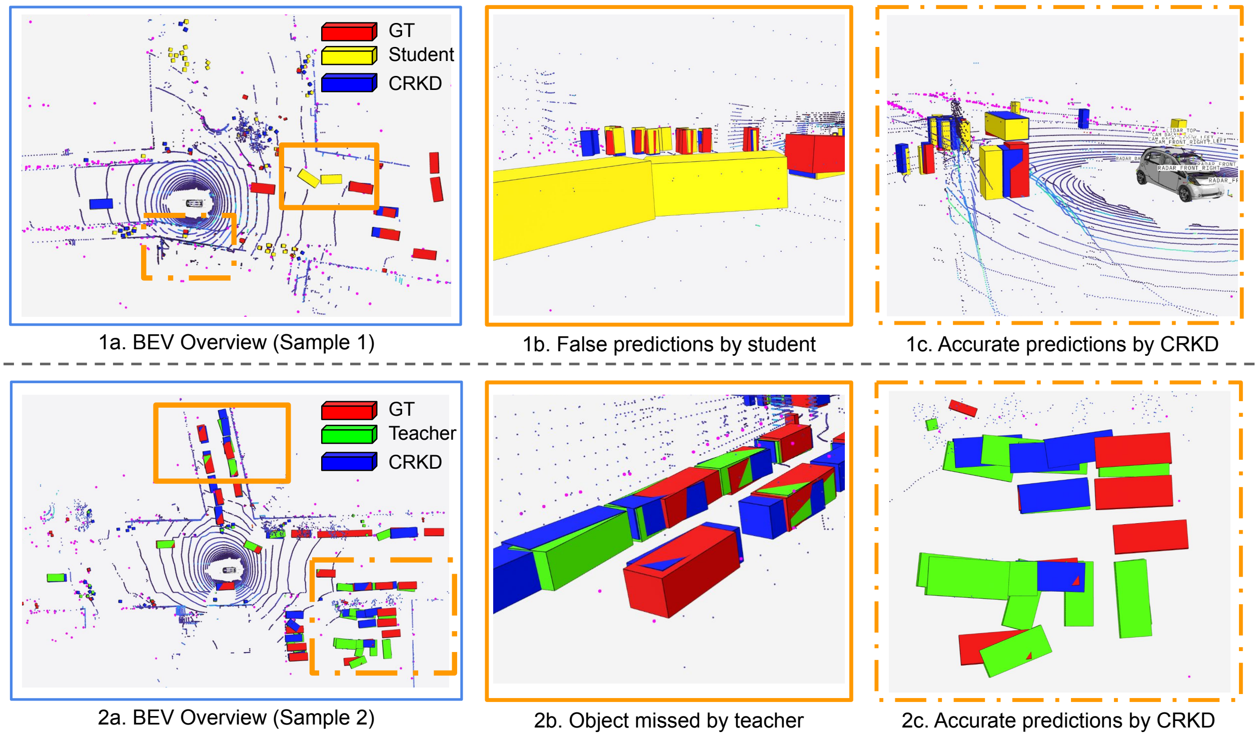

Figure 3. Qualitative results on nuScenes. We show zoomed-in views in panel b and c for the highlighted regions in panel a, with the border dash as the correspondence. We show the ground truth annotation in red, teacher prediction in green, student prediction in yellow, CRKD prediction in blue, and radar points in magenta. In (1a) to (1c), we show an example frame where CRKD has more accurate predictions and fewer false predictions than the student model. In (2a) to (2c) we show another example frame where CRKD even outperforms the LC teacher by detecting a missed car and rejecting several false predictions. Best viewed on screen and in color.

4.3.5 Effect of Model Architecture Refinement

We study the effect of adding the adaptive gated network to the original fusion module in BEVFusion [38] and tuning the number of channel dimensions. As shown in Tab. 5, the addition of the gated network brings notable improvement over the default fusion module in the baseline BEVFusion-CR model. We also notice that matching the input and the output channel dimension of the fusion module to the teacher model (Camera: 80, LiDAR: 256) brings additional improvement. The following KD operation also gets benefit from the same channel dimension setting between the teacher and student models since no additional channel-wise projection is needed for feature-level KD.

4.4. Qualitative Results

We show the visualization of the 3D object detection results to highlight the effectiveness of CRKD in Fig. 3. With CRKD, the detector is able to predict fewer false positive predictions and localize the objects better. We also show an impressive comparison between the teacher LC model and CRKD to demonstrate that CRKD can even outperform the teacher model with the help of radar measurements. This qualitative example validates the effectiveness of the CRKD framework and the value of radars for modern perception for autonomous driving. We will include more qualitative examples and discussion in the supplementary material.

5. Conclusion

We have proposed CRKD, a novel KD framework that supports a cross-modality fusion-to-fusion KD path for 3D object detection. We leverage the BEV space to design a novel LC-to-CR KD framework. We design four distillation losses to address the significant domain gap and facilitate the distillation process in this cross-modality setting. We also introduce the adaptive gated network to learn the relative importance between two expert feature maps. Extensive experiments show the effectiveness of CRKD in improving the detection performance of CR detectors. We hope CRKD will inspire future research to leverage our proposed KD framework to further explore the potential of CR detectors to improve the reliability of this widely accessible sensor suite. In future work, we plan to extend the proposed CRKD framework to other perception tasks such as occupancy mapping.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022. 1, 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 *CVPR*, 2020. 1, 2, 5

[3] Spencer Carmichael, Austin Buchan, Mani Ramanagopal, Radhika Ravi, Ram Vasudevan, and Katherine A Skinner. Dataset and benchmark: Novel sensors for autonomous vehicle perception. *arXiv preprint arXiv:2401.13853*, 2024. 1

[4] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *CVPR Workshop*, 2023. 2

[5] 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. 2, 4, 5, 6

[6] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022. 2, 4

[7] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1, 2, 5

[8] Florian Drews, Di Feng, Florian Faion, Lars Rosenbaum, Michael Ulrich, and Claudius Gläser. Deepfusion: A robust and modular 3d object detector for lidars, cameras and radars. In *IROS*, 2022. 2

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[10] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*, 2021. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

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

[13] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, 2022. 1, 2, 5

[14] Haotian Hu, Fanyi Wang, Jingwen Su, Yaonong Wang, Laifeng Hu, Weiye Fang, Jingwei Xu, and Zhiwang Zhang. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. *arXiv preprint arXiv:2303.17895*, 2023. 1

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 1

[16] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1

[17] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 6

[18] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. *arXiv preprint arXiv:2212.13979*, 2022. 2

[19] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *ECCV*, 2020. 2

[20] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. In *Neural computation*, 1991. 3

[21] Bo Ju, Zhikang Zou, Xiaoqing Ye, Minyue Jiang, Xiao Tan, Errui Ding, and Jingdong Wang. Paint and distill: Boosting 3d object detection with semantic passing network. In *ACM MM*, 2022. 2

[22] Jisong Kim, Minjae Seong, Geonho Bang, Dongsuk Kum, and Jun Won Choi. Rcm-fusion: Radar-camera multi-level fusion for 3d object detection. *arXiv preprint arXiv:2307.10249*, 2023. 2, 6

[23] Youngseok Kim, Jun Won Choi, and Dongsuk Kum. Grif net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image. In *IROS*, 2020. 2

[24] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *AAAI*, 2023. 6

[25] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *ICCV*, 2023. 2

[26] Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, and Fatih Porikli. X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. In *CVPR*, 2023. 1, 2, 5, 6

[27] 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. 5

[28] Kai Lei, Zhan Chen, Shuman Jia, and Xiaoteng Zhang. Hvdetfusion: A simple and robust camera-radar fusion framework. *arXiv preprint arXiv:2307.11323*, 2023. 1, 2

[29] Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Yuan Du, Li Du, and Shanghang Zhang. Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection. In *IEEE IV*, 2023. 2

[30] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020. 5

[31] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with

transformer for 3d object detection. In *NeurIPS*, 2022. 2, 6

[32] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, 2022. 2

[33] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, 2023. 1, 2

[34] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 2, 6

[35] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. When object detection meets knowledge distillation: A survey. In *IEEE TPAMI*, 2023. 2

[36] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *NeurIPS*, 2022. 2

[37] 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. 5

[38] 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, 2, 3, 4, 6, 8

[39] Zhe Liu, Xiaoqing Ye, Xiao Tan, Errui Ding, and Xiang Bai. Stereodistill: Pick the cream from lidar for distilling stereo-based 3d object detection. In *AAAI*, 2023. 2

[40] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022. 2

[41] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *WACV*, 2021. 2, 6

[42] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *IROS*, 2020. 2

[43] Su Pang, Daniel Morris, and Hayder Radha. Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection. In *WACV*, 2022. 2

[44] Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *CVPR*, 2023. 1

[45] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *CVPR*, 2022. 2

[46] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *ICRA*, 2019. 2

[47] Jingyu Song, Lingjun Zhao, and Katherine A Skinner. Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection. *arXiv preprint arXiv:2402.11735*, 2024. 1, 3

[48] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, 2020. 2

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

[50] Li Wang, Xinyu Zhang, Ziying Song, Jiangfeng Bi, Guoxin Zhang, Haiyue Wei, Liyao Tang, Lei Yang, Jun Li, Caiyan Jia, and Lijun Zhao. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. In *IEEE IV*, 2023. 1, 2, 4

[51] Yue Wang and Justin M Solomon. Object dgcnn: 3d object detection using dynamic graphs. In *NeurIPS*, 2021. 2

[52] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, 2023. 1, 2, 5, 6

[53] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In *ECCV*, 2022. 2

[54] Joey Wilson, Jingyu Song, Yuewei Fu, Arthur Zhang, Andrew Capodieci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. Motionsc: Data set and network for real-time semantic mapping in dynamic environments. In *IEEE RAL*, 2022. 1

[55] Joey Wilson, Yuewei Fu, Arthur Zhang, Jingyu Song, Andrew Capodieci, Paramsothy Jayakumar, Kira Barton, and Maani Ghaffari. Convolutional bayesian kernel inference for 3d semantic mapping. In *ICRA*, 2023. 1

[56] Zizhang Wu, Guilian Chen, Yuanzhu Gan, Lei Wang, and Jian Pu. Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion. In *ICRA*, 2023. 2

[57] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *IEEE ITSC*, 2021. 2

[58] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *ICCV*, 2023. 2

[59] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *ECCV*, 2020. 2, 4, 5

[60] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. In *NeurIPS*, 2022. 2

[61] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. In *NeurIPS*, 2022. 1

[62] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 3, 5

[63] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features us-

ing cross-view spatial feature fusion for 3d object detection. In *ECCV*, 2020. 2, 3

[64] Jia Zeng, Li Chen, Hanming Deng, Lewei Lu, Junchi Yan, Yu Qiao, and Hongyang Li. Distilling focal knowledge from imperfect expert for 3d object detection. In *CVPR*, 2023. 2

[65] Linfeng Zhang, Yukang Shi, Hung-Shuo Tai, Zhipeng Zhang, Yuan He, Ke Wang, and Kaisheng Ma. Structured knowledge distillation towards efficient and compact multi-view 3d detection. *arXiv preprint arXiv:2211.08398*, 2022. 2

[66] Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In *CVPR*, 2023. 2

[67] 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. 4, 6

[68] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *CVPR*, 2023. 1, 2, 4, 5, 6

[69] Taohua Zhou, Junjie Chen, Yining Shi, Kun Jiang, Mengmeng Yang, and Diange Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. In *IEEE IV*, 2023. 2, 6

[70] 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. 1, 5