

RCTDistill: Cross-Modal Knowledge Distillation Framework for Radar-Camera 3D Object Detection with Temporal Fusion

Geonho Bang^{1*} Minjae Seong^{2*} Jisong Kim^{2*} Geunju Baek¹
 Daye Oh³ Junhyung Kim³ Junho Koh³ Jun Won Choi^{1†}

¹ Seoul National University ²Hanyang University ³Hyundai Motor Company

{ghbang, gjbaek}@snu.ac.kr {mjseong, jskim}@snu.hanyang.ac.kr
 {daye.oh, univjun, junhkoh}@hyundai.com junwchoi@snu.ac.kr

Abstract

Radar-camera fusion methods have emerged as a cost-effective approach for 3D object detection but still lag behind LiDAR-based methods in performance. Recent works have focused on employing temporal fusion and Knowledge Distillation (KD) strategies to overcome these limitations. However, existing approaches have not sufficiently accounted for uncertainties arising from object motion or sensor-specific errors inherent in radar and camera modalities. In this work, we propose RCTDistill, a novel cross-modal KD method based on temporal fusion, comprising three key modules: Range-Azimuth Knowledge Distillation (RAKD), Temporal Knowledge Distillation (TKD), and Region-Decoupled Knowledge Distillation (RDKD). RAKD is designed to consider the inherent errors in the range and azimuth directions, enabling effective knowledge transfer from LiDAR features to refine inaccurate BEV representations. TKD mitigates temporal misalignment caused by dynamic objects by aligning historical radar-camera BEV features with current LiDAR representations. RDKD enhances feature discrimination by distilling relational knowledge from the teacher model, allowing the student to differentiate foreground and background features. RCTDistill achieves state-of-the-art radar-camera fusion performance on both the nuScenes and View-of-Delft (VoD) datasets, with the fastest inference speed of 26.2 FPS.

1. Introduction

Radar-camera-based 3D detection serves as a crucial component in autonomous driving systems. While various Radar-Camera fusion approaches exist, recent methods predominantly focus on feature-level fusion, with particular emphasis

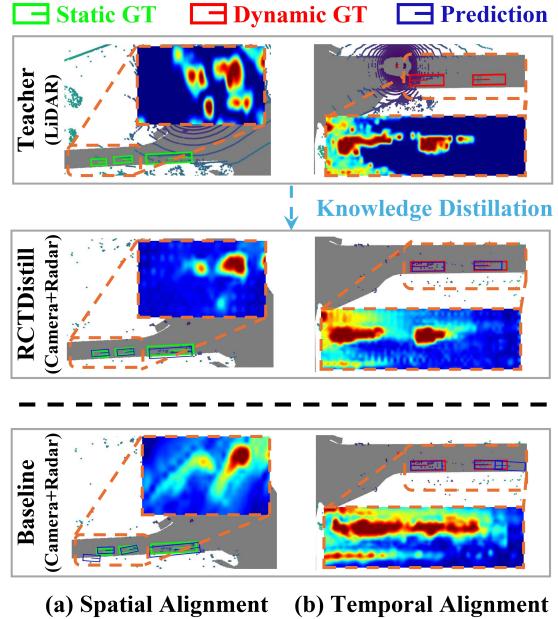


Figure 1. RCTDistill enhances the quality of BEV features for 3D object detection by aligning spatial features in the range-azimuth direction and temporal features in dynamic object trajectories.

on Bird’s Eye View (BEV) domain [13, 15, 22]. Although these methods effectively address the inherent limitations of individual sensors, there remains a notable performance gap when compared to LiDAR-based 3D detectors, as illustrated in Figure 1, where radar-camera fusion models show spatially and temporally misaligned BEV representations.

In an effort to overcome these performance gaps, Knowledge Distillation (KD) [16, 40] and temporal fusion approaches [14, 15, 20] have recently emerged as promising solutions in 3D object detection. KD-based approaches employ detectors based on low-fidelity sensors (e.g., cameras, radar, or camera-radar fusion) as student models while lever-

*Equal contributions

†Corresponding author

aging detectors based on higher-fidelity sensor inputs (e.g., LiDAR or LiDAR-camera fusion) as teacher models to enhance the student model’s performance. In parallel, temporal fusion methods aim to improve performance by incorporating additional information from past frames and merging it with current features to compensate for information that may be missed at a single time point.

However, several significant challenges remain in these approaches. First, many existing KD methods do not fully account for sensor-specific characteristics [16, 40], suggesting that explicitly incorporating these characteristics could further improve KD effectiveness. Cameras suffer from depth ambiguity, introducing uncertainties in object distance estimation during BEV feature generation [26, 38], whereas radar provides reliable depth but has low angular resolution, leading to imprecise object localization in the horizontal plane [33, 43]. While these sensors appear complementary in nature, integrating both modalities remains challenging, especially when it comes to mitigating their inherent uncertainties. Second, despite recent advances [14], addressing error propagation caused by object movements remains a challenge for temporal fusion models. The independent motion of dynamic objects leads to temporal misalignment, requiring explicit motion estimation to align features across frames. This process increases latency, ultimately degrading real-time detection performance.

In this paper, we propose RCTDistill, a novel cross-modal knowledge distillation method that transfers knowledge from a LiDAR model to a temporal radar-camera fusion model. RCTDistill performs knowledge distillation in three directions: Range-Azimuth Knowledge Distillation (RAKD), Temporal Knowledge Distillation (TKD), and Region-Decoupled Knowledge Distillation (RDKD). RAKD mitigates feature uncertainties in the fused BEV representation using an elliptical Gaussian mask, which accounts for varying uncertainties along the range and azimuth directions. By focusing on elliptical regions with Gaussian mask intensities above a threshold, it performs targeted feature distillation to effectively transfer knowledge from the teacher model. TKD addresses temporal feature misalignment in dynamic object detection by employing HA-Net to align historical BEV features. It generates trajectory-aware Gaussian regions for distillation, effectively capturing temporal dynamics while suppressing misaligned features. RDKD leverages an affinity map to distill relational knowledge between features, enabling the student model to learn discriminative feature relationships between foreground and background regions from the teacher model. RCTDistill achieves state-of-the-art performance and latency on both the nuScenes [3] and View-of-Delft [27] radar-camera 3D detection benchmarks, outperforming all previous methods.

The main contributions of this paper are as follows:

- Our RCTDistill is the first method to introduce a knowl-

edge distillation technique specifically designed for radar-camera temporal fusion. Previous methods [16, 40] focused solely on a single-frame setup, without explicitly utilizing temporal information. In this paper, we propose a novel cross-modal knowledge distillation approach that aligns BEV features over time, enhancing 3D object detection performance.

- We propose three knowledge distillation techniques, RAKD, TKD, and RDKD designed to address misalignment issues in radar-camera temporal 3D object detection. RAKD incorporates sensor-specific characteristics into the distillation process, while TKD focuses on temporal fusion by compensating for dynamic object motion. RDKD enhances the model’s discriminative power by distilling the relationships between foreground and background features from the teacher model.
- RCTDistill achieves improvements of a 4.7% mean Average Precision (mAP) and 4.9% NuScenes Detection Score (NDS) over the student model. Furthermore, it outperforms existing state-of-the-art radar-camera fusion methods on both the nuScenes and VoD datasets, while maintaining a real-time inference speed of 26.2 FPS.

2. Related Work

2.1. Radar-Camera 3D Object Detection

Recent radar-camera fusion models have adopted view transformation-based methods, which can be categorized into two directions: implicitly mapping features through attention mechanisms [21, 23, 31], or explicitly using depth distributions obtained from camera features [19, 30]. RCM-Fusion [13], following the former approach, integrated radar and camera features into a unified bird’s-eye-view (BEV) space using deformable attention. Similarly, SpaRC [33] enhances the positional information of sparse object queries and 2D camera features through an attention mechanism based on radar point features. Conversely, CRN [15] and CRT-Fusion [14] adopted the latter strategy to enhance camera BEV features via radar-guided view transformation before feature-level fusion. RCBEVDet [22] introduced a dual-path radar backbone combining point-based and transformer-based modules, leveraging deformable cross-attention for effective radar-camera feature fusion. To incorporate temporal modeling, some works [14, 15, 22, 33] have extended their methods by integrating multi-frame radar and camera features. Notably, CRT-Fusion [14] introduces dynamic object modeling to mitigate temporal misalignment.

2.2. Knowledge Distillation in 3D Object Detection

Cross-modal KD approaches have been widely adopted to facilitate knowledge transfer across heterogeneous sensor modalities to overcome sensing limitations [4, 5, 8, 10, 16, 18, 32, 41, 42]. BEVDistill [4] proposed feature- and

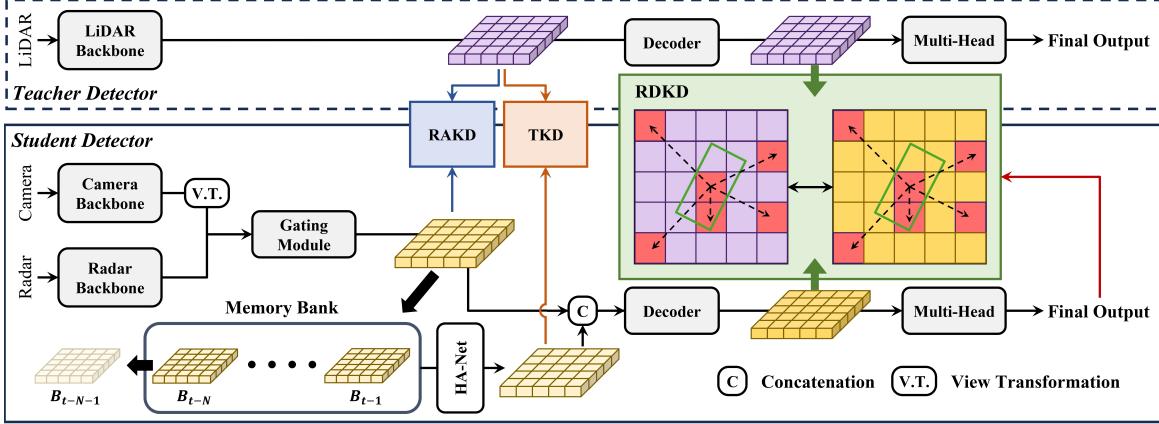


Figure 2. **Overall architecture of RCTDistill.** The student model fuses radar and camera features using a gating module to generate low-level BEV features, which are stored in a memory bank. These historical features are then temporally aggregated with the current BEV features, followed by a decoder that produces high-level BEV representations. RAKD and TKD enhance the low-level features by accounting for sensor uncertainty and object motion, respectively, while RDKD learns discriminative feature relationships between foreground and background regions from the teacher detector. Note that the teacher model is utilized only during training and omitted during inference.

instance-level distillation in the BEV space, and UniDistill [42] employed flexible KD pathways—such as LiDAR-to-camera, camera-to-LiDAR, and fusion-to-single modality—by projecting both teacher and student features into a unified BEV representation.

Growing interest in leveraging radar for 3D object detection has led to increased attention on cross-modal KD, where data sparsity remains a key challenge. RadarDistill [2] and CRKD [40] addressed this by transferring rich knowledge from LiDAR or LiDAR-camera models using feature- and relation-level supervision. LEROjD [28] and SCKD [35] further explored 4D radar-based KD; LEROjD employed a multi-stage training and distillation framework to leverage the spatial density of LiDAR, while SCKD adopted a semi-supervised approach using a LiDAR-camera teacher.

In parallel, temporal KD approaches have emerged to address object misalignment across time, a critical issue in dynamic driving scenes. VCD [11] projected the historical positions of ground truth into the current frame and applied KD at these aligned positions. STXD [12] leveraged similarity maps to align past frames of the teacher model with both the current frame of the teacher and the student.

3. Method

The overall architecture of RCTDistill is illustrated in Figure 2. Our model leverages a cross-modal knowledge distillation framework to enhance the performance of radar-camera fusion models. In Section 3.1, we provide a comprehensive overview of our teacher and student model architectures. Sections 3.2–3.4 present the detailed descriptions of our proposed knowledge distillation methods, RAKD, TKD, and RDKD.

3.1. Baseline Model

Student model. We adopt BEVFusion-R as our student model, which extends the BEVFusion [25] framework for radar-camera fusion. The model extracts radar BEV features and camera features using separate backbones. These camera features are then transformed into BEV representation via a View Transformation (VT) module. Subsequently, the radar and camera BEV features are concatenated and fused through a 1×1 convolution, followed by 2D CNN blocks and a detection head to produce the final predictions.

We improve the performance of the baseline by incorporating several modifications to BEVFusion-R. First, for multi-modal feature fusion, we replace 1×1 convolution layer with an adaptive gating network [40] that adaptively adjusts the influence of each modality. This gating network outputs low-level fused BEV features $B_{\text{low}}^{\text{RC}} \in \mathbb{R}^{D \times H \times W}$. These BEV features are aggregated using a streaming-based temporal fusion, where a memory bank retains historical BEV features obtained from previous time steps. These historical features are transformed into the current coordinate and then concatenated channel-wise with the current BEV features. Finally, the temporally fused features pass through a decoder consisting of 2D CNN layers to produce high-level BEV features $B_{\text{high}}^{\text{RC}} \in \mathbb{R}^{D' \times H \times W}$.

Teacher model. We employed CenterPoint, a widely adopted LiDAR-based 3D object detector, as our teacher model. CenterPoint groups raw LiDAR point clouds into 3D voxels, then extracts low-level BEV features $B_{\text{low}}^L \in \mathbb{R}^{C \times H \times W}$ through a 3D sparse convolutional backbone. These low-level features are further processed by a decoder consisting of 2D CNN layers to generate high-level BEV features, $B_{\text{high}}^L \in \mathbb{R}^{C' \times H \times W}$.

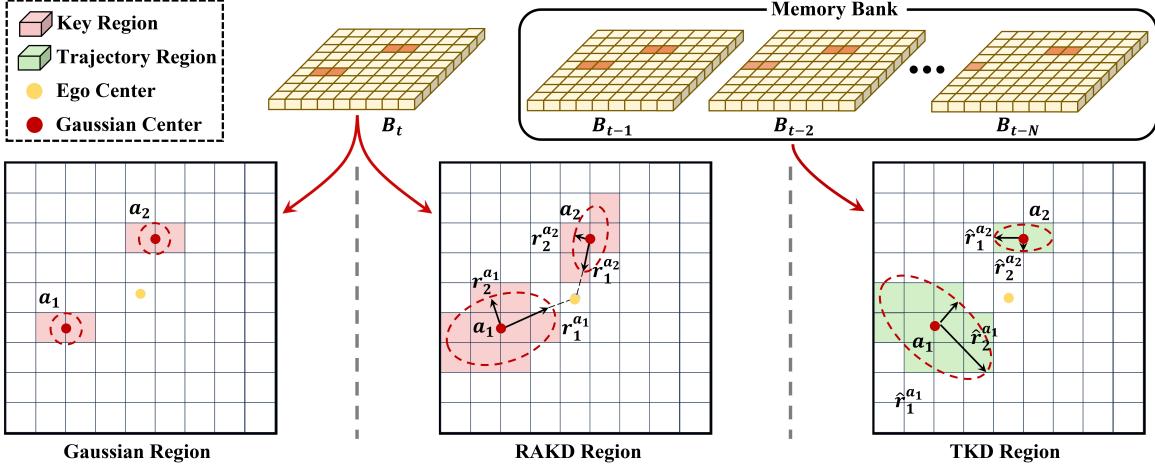


Figure 3. **Elliptical Gaussian Mask Regions for RAKD and TKD.** Elliptical masks in the RAKD (center) and TKD (right) regions are used for knowledge distillation, targeting range-azimuth uncertainty and temporal misalignment, respectively.

3.2. Range-Azimuth Knowledge Distillation

Cross-modal knowledge distillation aims to transfer the desirable representations from the teacher modality to the student modality. Earlier methods [4, 39] have used ground truth (GT) box-centered Gaussian masks to determine the region for feature distillation, as shown in Figure 3 (left). However, these approaches fail to capture modality-specific characteristics, such as depth ambiguity in cameras and range-azimuth uncertainties in radar. Gaussian masks apply weights that decrease uniformly from the object center, which makes it difficult to adapt to the unique distribution of each modality. To address these challenges, we propose Range-Azimuth Knowledge Distillation (RAKD), which utilizes an elliptical Gaussian mask. The shape of the ellipse is determined by the inherent uncertainties along range and azimuth directions, as shown in Figure 3 (middle).

For each i -th GT 3D box projected to bird’s eye view, we obtain a 2D box defined by center position (p_x^i, p_y^i) , heading angle θ^i , and dimensions (l^i, w^i) . Inspired by DORN [6], we determine the radius (r_1^i, r_2^i) of an elliptical region in the major and minor axes based on the object size and distance from the ego-vehicle as

$$r_1^i = l^i \cdot \left(\frac{\alpha_l}{l^i} \right)^\beta, \quad r_2^i = w^i \cdot \left(\frac{\alpha_w}{w^i} \right)^\beta, \quad (1)$$

where α_l and α_w denote the hyperparameters that control the scale of r_1^i and r_2^i , and $\beta (< 1)$ denotes the normalized distance from the ego-vehicle to the center position of the object. Notice that r_1^i and r_2^i are proportional to the size l^i and w^i of the 2D box, respectively. Using these parameters, we generate an Elliptical Gaussian mask $E_{i,x,y}$ of the i -th GT box at the position (x, y) as

$$E_{i,x,y} = \exp \left(-\frac{1}{2} \left(\frac{(x')^2}{(r_1^i)^2} + \frac{(y')^2}{(r_2^i)^2} \right) \right), \quad (2)$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta^i & -\sin \theta^i \\ \sin \theta^i & \cos \theta^i \end{bmatrix} \begin{bmatrix} x - p_x^i \\ y - p_y^i \end{bmatrix}. \quad (3)$$

We then merge the elliptical Gaussian masks $E \in \mathbb{R}^{N \times H \times W}$, generated from the N GT boxes, into $\bar{E} \in \mathbb{R}^{H \times W}$. When elliptical masks overlap between objects, $\bar{E}_{x,y}$ is obtained by taking the maximum magnitude in those areas. Next, we obtain a Distillation Mask $W_{RA} \in \mathbb{R}^{H \times W}$ by assigning the corresponding value from $\bar{E}_{x,y}$ to each spatial position (x, y) if it surpasses the threshold τ , and zero otherwise. RAKD aims to align the low-level features of the student model with those of the teacher model within the Distillation Mask. To this end, we introduce a RAKD loss function

$$L_{RA} = \frac{1}{|N_{RA}|} \sum_{j=1}^H \sum_{k=1}^W W_{RA,j,k} \|B_{low,j,k}^L - \bar{B}_{low,j,k}^{RC}\|_2, \quad (4)$$

where N_{RA} represents the number of non-zero elements in W_{RA} . The student feature map \bar{B}_{low}^{RC} is obtained by applying a 1×1 convolution to the original features B_{low}^{RC} to match its channel dimension with that of $B_{low,i,j}^L$.

3.3. Temporal Knowledge Distillation

While previous temporal fusion methods [15, 22] in 3D object detection have achieved significant performance improvements through simple concatenation of historical BEV features followed by CNN layers, they often struggle to adequately capture the movement of dynamic objects. To address this issue, we propose Temporal Knowledge Distillation (TKD), a novel framework that effectively takes object motion information into account for knowledge distillation.

As illustrated in Figure 2, we first concatenate historical BEV feature maps and process them through Historical

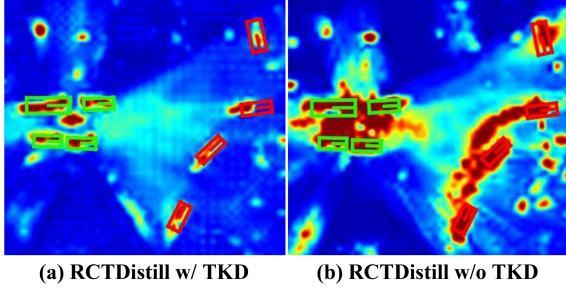


Figure 4. **Visualization of the BEV feature map.** Green boxes denote static object Ground Truth (GT) boxes, Red boxes indicate dynamic object GT boxes.

Alignment Network (HA-Net), a 2D CNN-based architecture. In the presence of rapid object motion, BEV features exhibit significant variations between consecutive frames. HA-Net aggregates these BEV features through large-scale kernel CNNs to alleviate inter-frame variations. The detailed architecture of HA-Net is provided in the Supplementary Material.

We identify the spatial regions in temporal features for knowledge transfer. To achieve this, we introduce temporal elliptical Gaussian masks $T \in \mathbb{R}^{N \times H \times W}$ whose shapes are determined based on object trajectories as shown in Figure 3 (right). These masks are designed to transfer temporal representations within regions that encompass the trajectories of dynamic objects. For the i -th object, we determine the elliptical Gaussian mask T_i by revising the center position and major axis radius of the elliptical function E_i in Equation 2. Given the velocity vector $\mathbf{v}^i = (v_x^i, v_y^i)$ and the center point $\mathbf{p}^i = (p_x^i, p_y^i)$ for the i -th object, the ellipse center $\hat{\mathbf{p}}^i = (\hat{p}_x^i, \hat{p}_y^i)$ is computed as

$$\hat{\mathbf{p}}^i = \left\{ \mathbf{p}^i - \frac{t_s}{2} \mathbf{v}^i, \quad \text{if } \|\mathbf{v}^i\|_2^2 > \tau_v \right. \quad (5)$$

where t_s is the temporal duration chosen to be less than the total duration of the historical frame. This implies that when the magnitude of the velocity vector exceeds a threshold τ_v , the center of the ellipse is obtained by shifting the object center in the direction opposite to its motion. Subsequently, the radius \hat{r}_1^i and \hat{r}_2^i in the major and minor axis is calculated as

$$\hat{r}_1^i = l^i + \sqrt{(\hat{p}_x^i - p_x^i)^2 + (\hat{p}_y^i - p_y^i)^2} \quad (6)$$

$$\hat{r}_2^i = w^i, \quad (7)$$

to allow the masks to cover a historical trajectory of the i -th dynamic object. Using the computed $\hat{\mathbf{p}}^i$, \hat{r}_1^i and \hat{r}_2^i , the temporal mask $T_{i,x,y}$ is given by

$$T_{i,x,y} = \exp \left(-\frac{1}{2} \left(\frac{(x'')^2}{(\hat{r}_1^i)^2} + \frac{(y'')^2}{(\hat{r}_2^i)^2} \right) \right), \quad (8)$$

where

$$\begin{bmatrix} x'' \\ y'' \end{bmatrix} = \begin{bmatrix} \cos \theta^i & -\sin \theta^i \\ \sin \theta^i & \cos \theta^i \end{bmatrix} \begin{bmatrix} x - \hat{p}_x^i \\ y - \hat{p}_y^i \end{bmatrix}. \quad (9)$$

When the elliptical masks overlap, we obtain $\bar{T} \in \mathbb{R}^{H \times W}$ by taking the maximum value of $T_{i,x,y}$ over i . Similarly to RAKD, the Temporal Distillation Mask $W_T \in \mathbb{R}^{H \times W}$ is computed by retaining $\bar{T}_{x,y}$ at position (x, y) if it exceeds the threshold τ , or setting it to zero otherwise. To facilitate temporal knowledge transfer in these regions, we introduce a temporal distillation loss function

$$L_T = \frac{1}{|N_T|} \sum_{j=1}^H \sum_{k=1}^W W_{T,j,k} \left\| B_{\text{low},j,k}^L - \hat{B}_{\text{low},j,k}^{RC} \right\|^2, \quad (10)$$

where N_T denotes the number of non-zero elements in W_T , $\hat{B}_{\text{low}}^{RC}$ represents the temporally aggregated feature maps obtained through the HA-Net, and B_{low}^L denotes the current low-level feature map from the teacher model. This distillation process aims to align the historical camera-radar features with the current LiDAR features within the regions defined by W_T . As shown in Figure 4, the BEV feature maps produced with our proposed TKD module exhibit clearer object boundaries and reduced temporal artifacts for dynamic objects. In contrast, the baseline displays noticeable trailing effects around moving objects. The enhanced temporal alignment achieved by TKD leads to more accurate localization of dynamic objects at the current timestamp.

3.4. Region-Decoupled Knowledge Distillation

High-level feature maps—especially those from LiDAR-based models—are effective at capturing rich semantic information that distinctly separates foreground objects from background regions. To transfer this discriminative capability to radar-camera models, we propose RDKD, which aligns the relational structure between foreground and background regions in the high-level feature maps. Guided by the teacher model, the student model is encouraged to maintain high similarity among foreground features while effectively distinguishing them from background features.

The high-level feature map B_{high}^{RC} obtained from the student detector is processed through the detection head to generate a classification score map B_{cls}^{RC} . We obtain the confidence score map $\bar{B}_{\text{cls}}^{RC} \in \mathbb{R}^{H \times W}$ by taking the maximum value of B_{cls}^{RC} across all classes. From $\bar{B}_{\text{cls}}^{RC}$, we identify all positions with confidence scores above a predefined threshold τ . At these K selected positions, we extract the corresponding high-level features from $\bar{B}_{\text{cls}}^{RC}$, resulting in a set of K feature vectors, denoted as $\mathbf{f}_1^{RC}, \mathbf{f}_2^{RC}, \dots, \mathbf{f}_K^{RC}$. Using these extracted features, the $K \times K$ affinity map is computed as

$$S_{jk}^{RC} = \frac{\mathbf{f}_j^{RC} \cdot \mathbf{f}_k^{RC}}{\|\mathbf{f}_j^{RC}\|_2 \|\mathbf{f}_k^{RC}\|_2}, \quad \text{for } j, k = 1, 2, \dots, K, \quad (11)$$

| Dataset | Methods | Input | KD | Backbone | Image Size | Frames | NDS | mAP | mATE | mASE | mAOE | mAVE | mAAE | FPS |
|---------|------------------------------|-------|----|----------|------------|--------|-------------|-------------|-------|-------|-------|-------|-------|-------------|
| Val. | Teacher [37] | L | - | - | - | 1 | 67.1 | 60.3 | 0.298 | 0.251 | 0.288 | 0.281 | 0.190 | - |
| | BEVDepth [†] [19] | C | - | R50 | 256×704 | 2 | 47.5 | 35.1 | 0.639 | 0.267 | 0.479 | 0.428 | 0.198 | 11.6 |
| | SOLOFusion [†] [29] | C | - | R50 | 256×704 | 16+1 | 53.4 | 42.7 | 0.567 | 0.274 | 0.411 | 0.252 | 0.188 | 11.4 |
| | X3KD [†] [16] | C+R | ✓ | R50 | 256×704 | 1 | 53.8 | 42.3 | 0.487 | 0.277 | 0.542 | 0.344 | 0.197 | - |
| | CRKD [†] [40] | C+R | ✓ | R50 | 256×704 | 1 | 54.9 | 43.2 | 0.450 | 0.267 | 0.442 | 0.339 | 0.176 | - |
| | RCTDistill-S [‡] | C+R | ✓ | R50 | 256×704 | 1 | 55.5 | 46.4 | 0.488 | 0.262 | 0.462 | 0.381 | 0.176 | 28.0 |
| | CRN [15] | C+R | - | R50 | 256×704 | 4 | 56.0 | 49.0 | 0.487 | 0.277 | 0.542 | 0.344 | 0.197 | 20.4 |
| | RCBEVDet [†] [22] | C+R | - | R50 | 256×704 | 2 | 56.8 | 45.3 | 0.486 | 0.285 | 0.404 | 0.220 | 0.192 | 21.3 |
| | CRT-Fusion [†] [14] | C+R | - | R50 | 256×704 | 6+1 | 59.7 | 50.8 | 0.461 | 0.264 | 0.419 | 0.234 | 0.186 | 14.5 |
| | RCTTrans [†] [20] | C+R | - | R50 | 256×704 | 4 | 58.6 | 50.9 | 0.537 | 0.269 | 0.491 | 0.203 | 0.183 | 19.2 |
| | SpaRC [33] | C+R | - | R50 | 256×704 | 8 | 62.0 | 54.5 | 0.496 | 0.269 | 0.403 | 0.177 | 0.181 | 19.1 |
| | Baseline | C+R | - | R50 | 256×704 | 8+1 | 57.3 | 50.5 | 0.491 | 0.271 | 0.560 | 0.280 | 0.189 | 27.8 |
| | RCTDistill | C+R | ✓ | R50 | 256×704 | 8+1 | 62.2 | 55.2 | 0.434 | 0.241 | 0.432 | 0.263 | 0.174 | 26.2 |
| Test | Teacher [37] | L | - | - | - | 1 | 67.3 | 60.6 | 0.286 | 0.251 | 0.315 | 0.261 | 0.185 | - |
| | BEVDepth [†] [19] | C | - | R101 | 512×1408 | 2 | 53.5 | 41.2 | 0.565 | 0.266 | 0.358 | 0.331 | 0.190 | 5.0 |
| | SOLOFusion [†] [29] | C | - | R101 | 512×1408 | 16+1 | 58.2 | 48.3 | 0.503 | 0.264 | 0.381 | 0.246 | 0.207 | 11.4 |
| | CRN [15] | C+R | - | R101 | 512×1408 | 4 | 59.2 | 52.5 | 0.460 | 0.273 | 0.443 | 0.352 | 0.180 | 7.2 |
| | CRT-Fusion [14] | C+R | - | R101 | 512×1408 | 6+1 | 62.1 | 55.4 | 0.425 | 0.264 | 0.433 | 0.237 | 0.193 | 4.9 |
| | SpaRC [33] | C+R | - | R101 | 512×1408 | 8 | 64.4 | 57.1 | 0.484 | 0.264 | 0.308 | 0.175 | 0.178 | 7.2 |
| | Baseline | C+R | - | R101 | 512×1408 | 8+1 | 60.9 | 54.2 | 0.433 | 0.272 | 0.469 | 0.252 | 0.198 | 8.7 |
| | RCTDistill | C+R | ✓ | R101 | 512×1408 | 8+1 | 65.5 | 59.0 | 0.378 | 0.250 | 0.346 | 0.236 | 0.186 | 8.4 |
| | BEVDepth [19] | C | - | ConvX-B | 640×1600 | 2 | 60.9 | 52.0 | 0.445 | 0.243 | 0.352 | 0.347 | 0.128 | - |
| | SOLOFusion [29] | C | - | ConvX-B | 640×1600 | 16+1 | 61.9 | 54.0 | 0.453 | 0.257 | 0.376 | 0.276 | 0.148 | - |
| Test | CRN [‡] [15] | C+R | - | ConvX-B | 640×1600 | 4 | 62.4 | 57.5 | 0.416 | 0.264 | 0.456 | 0.365 | 0.130 | - |
| | RCBEVDet [22] | C+R | - | V2-99 | 640×1600 | 2 | 63.9 | 55.0 | 0.390 | 0.234 | 0.362 | 0.259 | 0.113 | - |
| | CRTFusion [14] | C+R | - | ConvX-B | 512×1408 | 6+1 | 64.9 | 58.3 | 0.365 | 0.261 | 0.405 | 0.262 | 0.132 | 3.7 |
| | RCTTrans [20] | C+R | - | V2-99 | 640×1600 | 4 | 64.7 | 57.8 | 0.459 | 0.245 | 0.392 | 0.198 | 0.121 | - |
| | SpaRC [33] | C+R | - | V2-99 | 640×1600 | 8 | 67.1 | 60.0 | - | - | - | - | - | - |
| | RCTDistill | C+R | ✓ | ConvX-B | 512×1408 | 8+1 | 67.4 | 60.3 | 0.334 | 0.245 | 0.334 | 0.251 | 0.113 | 5.0 |

Table 1. Performance comparisons with 3D object detectors on the nuScenes validation (Val.) and test set. Teacher denotes CenterPoint [37]. ‘C’, and ‘R’ represent camera and radar, respectively. RCTDistill-S indicates the use of single-frame input. [†]: trained with CBGS [44]. [‡]: use Test Time Augmentation.

where \cdot denotes the inner product operation. From the same K selected positions, we also extract K high-level feature vectors from the high-level feature map B_{high}^L obtained by the teacher model. Then, we compute the teacher affinity map S^L in the same manner as Equation 11.

We define the RDKD loss L_{RD} to guide the student model to learn the ability to discriminate between the foreground and background regions from the teacher model as

$$L_{RD} = \frac{1}{|K^2|} \sum_{j=1}^K \sum_{k=1}^K |S_{jk}^L - S_{jk}^{RC}|. \quad (12)$$

3.5. Loss Function

The total loss function comprises a 3D detection loss L_{det} and three KD loss terms. The total loss L_{total} is obtained as

$$L_{total} = L_{det} + \lambda_{RA} L_{RA} + \lambda_T L_T + \lambda_{RD} L_{RD}, \quad (13)$$

where λ_{RA} , λ_T , and λ_{RD} are hyperparameters that control the relative importance of each loss term.

| Methods | RAKD | RDKD | TKD | NDS↑ | mAP↑ |
|------------|------|------|-----|------|------|
| Baseline | | | | 55.0 | 47.1 |
| RCTDistill | ✓ | | | 57.2 | 49.3 |
| | | ✓ | | 56.8 | 49.0 |
| | | | ✓ | 57.0 | 48.5 |
| | ✓ | ✓ | | 57.5 | 50.1 |
| | | ✓ | ✓ | 57.4 | 49.3 |
| | ✓ | ✓ | ✓ | 57.6 | 49.9 |
| | ✓ | ✓ | ✓ | 58.4 | 51.0 |

Table 2. Ablation study for evaluating the main components.

4. Experiments

4.1. Experimental Setup

Datasets and metrics. We conducted the experiments on the nuScenes [3] and View-of-Delft (VoD) [27] datasets.

The nuScenes dataset consists of 1,000 driving scenes, each scene is approximately 20 seconds long with 360-degree coverage from six cameras, five radars, and one LiDAR sensor. Keyframes are annotated at 2Hz, covering 10 object classes. Our evaluation follows the official nuScenes benchmark metrics, including mean Average Precision (mAP) and nuScenes Detection Score (NDS).

The VoD dataset comprises 8,693 frames, each capturing forward view using a single 4D radar, a stereo camera, and a LiDAR. Evaluation is performed using mAP measured over the Entire Annotated Area (EAA) and within the Driving Corridor (RoI) for three object classes (Car, Pedestrian, Cyclist). Since the VoD test server is not yet available, evaluation is conducted on the validation set.

Implementation details. For the student model, we utilized ResNet [9] and ConvNeXt [24] as the image backbone networks and a modified Pillar Feature Network [14] building on PointPillars [17] as the radar backbone network. The teacher model employs the SECOND [36] backbone for processing LiDAR point clouds. Following the SOLOFusion [29], we incorporate a streaming-based parallel temporal fusion mechanism that leverages BEV features from the preceding eight frames. During the training, the teacher model weights are frozen while the student model is trained for 60 epochs. Detailed training configurations and hyperparameter settings are provided in the supplementary material.

4.2. Comparison to the state-of-the-art

Table 1 compares RCTDistill with existing camera-radar fusion and camera-only 3D object detectors on both the nuScenes validation and test sets. RCTDistill sets a new state-of-the-art performance across various backbone configurations, consistently surpassing previous methods. With ResNet-50, RCTDistill achieves 62.2% NDS and 55.2% mAP while maintaining the highest efficiency at 26.2 FPS. For the ResNet-101 backbone configuration, RCTDistill exhibits significant performance gains of 1.1% in NDS and 1.9% in mAP compared to SpaRC [33]. Moreover, on the test set, our method outperforms all existing radar-camera fusion models. Even when a single frame input is used (RCTDistill-S), the proposed method achieves superior results. Notably, RCTDistill achieves the highest inference speed (up to 28.0 FPS) across all evaluated configurations.

4.3. Ablation Studies

We conducted ablation studies on both the nuScenes and VoD validation sets. All experiments were conducted using a ResNet-50 image backbone and trained for 20 epochs.

Components analysis. Table 2 presents an ablation study of the RCTDistill, focusing on the contributions of RAKD, RDKD, and TKD. Each component, when applied individually, improves performance by at least 1.4% in mAP and 1.8% in NDS. Pairwise combinations further boost results,

| Methods | BEVDistill [4] | CRKD [40] | Ours | NDS↑ | mAP↑ |
|----------|----------------|-----------|------|------|------|
| Baseline | | | | 55.0 | 47.1 |
| RAKD | ✓ | | | 56.4 | 48.3 |
| | | ✓ | | 56.7 | 48.6 |
| | | | ✓ | 57.2 | 49.3 |

(a) Performance Comparison Based on Mask Types.

| Methods | VCD [11] | STXD [12] | Ours | NDS↑ | mAP↑ |
|----------|----------|-----------|------|------|------|
| Baseline | | | | 55.0 | 47.1 |
| TKD | ✓ | | | 55.9 | 48.3 |
| | | ✓ | | 56.5 | 48.1 |
| | | | ✓ | 57.0 | 48.5 |

(b) Performance Comparison Based on Temporal Knowledge Distillation Methods.

| Methods | MonoDistill [5] | CRKD [40] | Ours | NDS↑ | mAP↑ |
|----------|-----------------|-----------|------|------|------|
| Baseline | | | | 55.0 | 47.1 |
| RDKD | ✓ | | | 56.0 | 47.8 |
| | | ✓ | | 55.9 | 48.1 |
| | | | ✓ | 56.8 | 49.0 |

(c) Performance Comparison by Relation KD Application Scope.

Table 3. Ablation studies of proposed KD methods.

| Methods | Input | EAA AP↑ | RoI AP↑ |
|----------------|-------|--------------|--------------|
| BEVFusion [25] | C+R | 49.25 | 68.52 |
| LXL [34] | C+R | 56.31 | 72.93 |
| RCBEVDet [22] | C+R | 49.99 | 69.80 |
| HGSFusion [7] | C+R | 58.96 | 79.46 |
| SGDet3D [1] | C+R | 59.75 | 77.42 |
| RCTDistill | C+R | 62.37 | 82.25 |

Table 4. Performance Comparison on VoD [27] Dataset. RoI denotes the driving corridor area. EAA: Entire Annotated Area.

while integrating all three components achieves the highest performance, reaching 58.4% NDS and 51.0% mAP. This demonstrates the effective synergy between spatial, temporal, and relational distillation techniques.

Impact of KD region selection on model performance. Table 3 presents an ablation study evaluating the performance impact of applying KD to different regions for each module. In Table 3a, we compare our proposed elliptical Gaussian mask with previous methods that adopt different strategies for determining knowledge distillation regions. BEVDistill [4] employs a circular Gaussian distribution centered on the object. In contrast, CRKD [40] proposes MSFD, which expands the ground-truth bounding box region based on object velocity and range thresholds. We replaced RAKD with these methods in our RCTDistill framework for comparison. While the previous approaches outperform the baseline, our method achieves greater performance gains—specifically, a 2.2% improvement in both NDS and mAP. This improvement may be attributed to the explicit modeling of sensor-

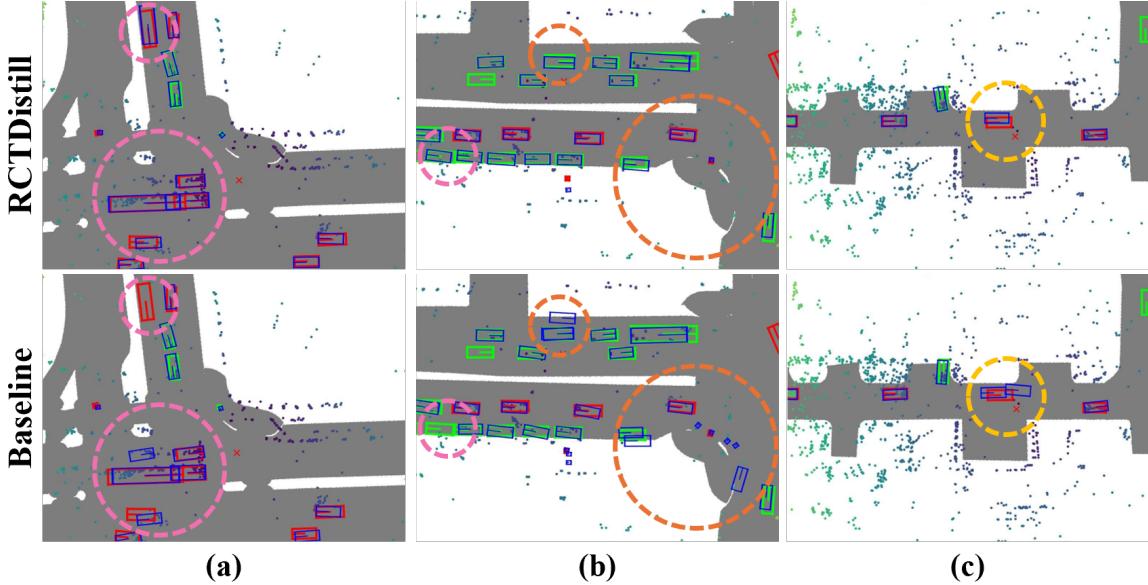


Figure 5. Qualitative results comparing RCTDistill and Baseline. Blue, green, and red boxes denote predictions, static ground truth, and dynamic ground truth, respectively. Pink circles highlight areas where the baseline model fails to detect objects, but RCTDistill successfully identifies them. Orange and yellow circles indicate regions where false positives have been corrected.

specific uncertainty.

Table 3b demonstrates the effectiveness of our proposed TKD method compared to existing temporal KD approaches. While STXD [12] relies on constructing temporal similarity maps between past and current BEV features for KD, and VCD [11] applies distillation over circular Gaussian masks across all time steps, our method employs a simpler yet more effective approach. TKD leverages only current-frame ground truth boxes and velocity information to identify adaptive distillation regions that reflect potential vehicle movement patterns. Our method outperforms existing temporal KD approaches, achieving significant improvements of 2.0% in NDS and 1.4% in mAP over the baseline. Additional analysis is provided in the Supplementary Material.

Table 3c highlights the effectiveness of our RDKD compared to conventional approaches that apply relation KD across the entire region. MonoDistill [5] proposes distillation between feature maps, and CRKD [40] presents a ReID strategy that utilizes distillation on affinity maps. In contrast, our method selectively targets specific regions. Our approach outperforms the existing methods, improving mAP by at least 0.9% and NDS by 0.8%.

Performance comparison on VoD dataset. Table 4 presents a performance comparison between RCTDistill and existing models based on the VoD [27] validation set. RCTDistill achieves state-of-the-art performance by outperforming the latest SGDet3D method [1], achieving a 2.62% increase in EAA AP and a 4.83% improvement in RoI AP. These results demonstrate that RCTDistill consistently delivers superior

performance across both the nuScenes [3] and VoD [27] datasets, underscoring its robust performance across diverse environments and sensor characteristics.

Qualitative results. Figure 5 presents a qualitative comparison between the proposed RCTDistill and the baseline model. The highlighted circles illustrate RCTDistill’s enhanced ability to detect missed objects (pink), reduce false positives caused by range and azimuth uncertainties (orange), and resolve false positives caused by dynamic object movement (yellow). These results demonstrate its effectiveness in addressing the limitations of previous methods.

5. Conclusions

In this paper, we presented RCTDistill, a novel cross-modal knowledge distillation framework designed to improve temporal radar-camera fusion for 3D object detection. RCTDistill effectively addresses key challenges, including sensor-specific uncertainty and temporal misalignment, leading to significant performance gains. The proposed RAKD module mitigates sensor-specific uncertainty by applying elliptical Gaussian regions, thereby enhancing the quality of BEV representations. TKD alleviates temporal misalignment caused by dynamic objects by aligning historical BEV features more accurately with current LiDAR features. Lastly, RDKD improves feature discrimination by transferring relational knowledge from the teacher model, enabling the student model to more effectively distinguish foreground from background features. Overall, RCTDistill establishes a new state-of-the-art in radar-camera-based 3D object detection.

6. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129), and the Technology Innovation Program (20018112, Development of autonomous manipulation and gripping technology using imitation learning based on visualtactile sensing) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

References

- [1] Xiaokai Bai, Zhu Yu, Lianqing Zheng, Xiaohan Zhang, Zili Zhou, Xue Zhang, Fang Wang, Jie Bai, and Hui-Liang Shen. Sgdet3d: Semantics and geometry fusion for 3d object detection using 4d radar and camera. *IEEE Robotics and Automation Letters*, 2024. [7](#), [8](#)
- [2] Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15491–15500, 2024. [3](#)
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Lioung, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. [2](#), [6](#), [8](#)
- [4] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qin-hong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. In *International Conference on Learning Representations*, 2023. [2](#), [4](#), [7](#)
- [5] 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 *International Conference on Learning Representations*, 2022. [2](#), [7](#), [8](#)
- [6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. [4](#)
- [7] Zijian Gu, Jianwei Ma, Yan Huang, Honghao Wei, Zhanye Chen, Hui Zhang, and Wei Hong. Hgsfusion: Radar-camera fusion with hybrid generation and synchronization for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3185–3193, 2025. [7](#)
- [8] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3153–3163, 2021. [2](#)
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. [7](#)
- [10] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 87–104, 2022. [2](#)
- [11] Linyan Huang, Zhiqi Li, Chonghao Sima, Wenhui Wang, Jingdong Wang, Yu Qiao, and Hongyang Li. Leveraging vision-centric multi-modal expertise for 3d object detection. *Advances in Neural Information Processing Systems*, 36:38504–38519, 2023. [3](#), [7](#), [8](#)
- [12] Sujin Jang, Dae Ung Jo, Sung Ju Hwang, Dongwook Lee, and Daehyun Ji. Stxd: structural and temporal cross-modal distillation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 36:29323–29342, 2023. [3](#), [7](#), [8](#)
- [13] Jisong Kim, Minjae Seong, Geonho Bang, Dongsuk Kum, and Jun Won Choi. Rcm-fusion: Radar-camera multi-level fusion for 3d object detection. In *IEEE International Conference on Robotics and Automation*, pages 18236–18242. IEEE, 2024. [1](#), [2](#)
- [14] Jisong Kim, Minjae Seong, and Jun Won Choi. CRT-Fusion: Camera, radar, temporal fusion using motion information for 3d object detection. *Advances in Neural Information Processing Systems*, 37:108625–108648, 2024. [1](#), [2](#), [6](#), [7](#)
- [15] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17615–17626, 2023. [1](#), [2](#), [4](#), [6](#)
- [16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13343–13353, 2023. [1](#), [2](#), [6](#)
- [17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. [7](#)
- [18] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. [2](#)
- [19] Yinhan 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. [2](#), [6](#)
- [20] Yiheng Li, Yang Yang, and Zhen Lei. Rctrans: Radar-camera transformer via radar densifier and sequential decoder for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5048–5056, 2025. [1](#), [6](#)

- [21] Zhiqi Li, Wenhui 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2
- [22] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14928–14937, 2024. 1, 2, 4, 6, 7
- [23] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2
- [24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7
- [25] 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 *IEEE International Conference on Robotics and Automation*, pages 2774–2781. IEEE, 2023. 3, 7
- [26] Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Toward real-world bev perception: Depth uncertainty estimation via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17124–17133, 2025. 2
- [27] Andras Palfy, Ewoud Pool, Srimannarayana Baratam, Julian FP Kooij, and Dariu M Gavrila. Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. 2, 6, 7, 8
- [28] Patrick Palmer, Martin Krüger, Stefan Schütte, Richard Altenedorfer, Ganesh Adam, and Torsten Bertram. Lerojd: Lidar extended radar-only object detection. In *European Conference on Computer Vision*, pages 379–396. Springer, 2024. 3
- [29] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *International Conference on Learning Representations*, 2023. 6, 7
- [30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 2
- [31] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2
- [32] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8637–8646, 2023. 2
- [33] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Felix Fent, and Gerhard Rigoll. Sparc: Sparse radar-camera fusion for 3d object detection. *arXiv preprint arXiv:2411.19860*, 2024. 2, 6, 7
- [34] Weiyi Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. *IEEE Transactions on Intelligent Vehicles*, 9(1):79–92, 2023. 7
- [35] Ruoyu Xu, Zhiyu Xiang, Chenwei Zhang, Hanzhi Zhong, Xijun Zhao, Ruina Dang, Peng Xu, Tianyu Pu, and Eryun Liu. Sckd: Semi-supervised cross-modality knowledge distillation for 4d radar object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8933–8941, 2025. 3
- [36] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 7
- [37] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021. 6
- [38] Songsong Yu, Yifan Wang, Yunzhi Zhuge, Lijun Wang, and Huchuan Lu. Dme: unveiling the bias for better generalized monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6817–6825, 2024. 2
- [39] Haimei Zhao, Qiming Zhang, Shanshan Zhao, Zhe Chen, Jing Zhang, and Dacheng Tao. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7460–7468, 2024. 4
- [40] Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15470–15480, 2024. 1, 2, 3, 6, 7, 8
- [41] Wu Zheng, Mingxuan Hong, Li Jiang, and Chi-Wing Fu. Boosting 3d object detection by simulating multimodality on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13638–13647, 2022. 2
- [42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5116–5125, 2023. 2, 3
- [43] 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. *IEEE Transactions on Intelligent Vehicles*, 8(2):1523–1535, 2023. 2
- [44] 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. 6