IMKD: Intensity-Aware Multi-Level Knowledge Distillation for Camera-Radar Fusion

Shashank Mishra[1] Karan Patil[1,2] Didier Stricker[1,2] Jason Rambach[1]
[1]German Research Center for Artificial Intelligence (DFKI) [2]RPTU

Abstract

High-performance Radar-Camera 3D object detection can be achieved by leveraging knowledge distillation without using LiDAR at inference time. However, existing distillation methods typically transfer modality-specific features directly to each sensor, which can distort their unique characteristics and degrade their individual strengths. To address this, we introduce IMKD, a radar-camera fusion framework based on multi-level knowledge distillation that preserves each sensor's intrinsic characteristics while amplifying their complementary strengths. IMKD applies a three-stage, intensity-aware distillation strategy to enrich the fused representation across the architecture: (1) LiDAR-to-Radar intensity-aware feature distillation to enhance radar representations with fine-grained structural cues, (2) LiDAR-to-Fused feature intensity-guided distillation to selectively highlight useful geometry and depth information at the fusion level, fostering complementarity between the modalities rather than forcing them to align, and (3) Camera-Radar intensity-guided fusion mechanism that facilitates effective feature alignment and calibration. Extensive experiments on the nuScenes benchmark show that IMKD reaches 67.0% NDS and 61.0% mAP, outperforming all prior distillation-based radar-camera fusion methods. Our code and models are available at: https://github.com/dfki-av/IMKD/.

1. Introduction

Bird's Eye View (BEV) has become the dominant representation for 3D perception in autonomous systems due to its structured spatial layout and planning compatibility. BEV maps are constructed using LiDAR, cameras, and radar—each with distinct characteristics. LiDAR offers precise depth and structure, making it highly effective for 3D detection [10, 38, 48, 65], but its high cost and limited range hinder adoption. Cameras provide rich texture but struggle with depth and low light. Radar is robust in adverse weather and long-range detection but suffers from low spatial resolution and noise.

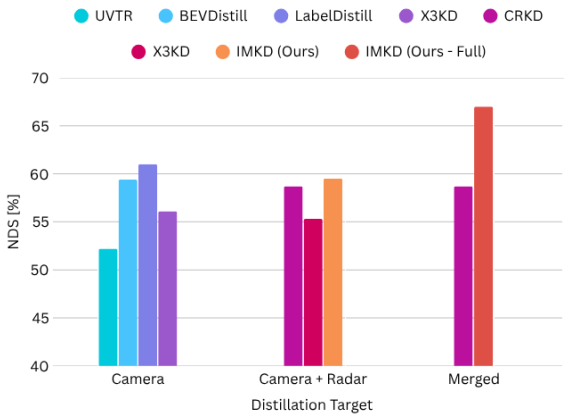


Figure 1. Comparison of KD methods grouped by distillation target. IMKD's intensity-based knowledge distillation achieves the highest performance.

To leverage cost-effective setups, recent works have explored knowledge distillation (KD) to transfer information from LiDAR to camera or radar-based models [14, 17, 71]. However, most existing approaches distill knowledge independently to each modality, often forcing radar or camera features to mimic LiDAR representations. This direct one-to-one transfer overlooks the unique characteristics of each sensor and can introduce representational conflicts, limiting the effectiveness of the distillation process. Furthermore, the intermediate representations chosen for distillation are often suboptimal, missing opportunities to enhance cross-modal fusion through better guidance signals.

In this paper, we introduce IMKD, an Intensity-Aware Multi-Level Knowledge Distillation framework that enhances camera–radar representations through cross-modal supervision. Prior methods have explored distillation either at the modality level or in the fused space, but often without accounting for sensor-specific reliability. IMKD addresses this by supervising the fused camera–radar BEV representation with LiDAR as a privileged modality, where intensity serves as a reliability prior that highlights geometrically

consistent regions. This enables structural and depth-rich cues to be transferred more effectively into the fused representation while preserving radar's robustness, leading to improved alignment, stability, and confidence in BEV features.

The intensity-guided distillation mechanism modulates supervision based on LiDAR confidence, emphasizing informative regions while down-weighting ambiguous signals. This adaptive weighting prevents overfitting to modality inconsistencies and, when applied across early, middle, and late fusion stages, yields stable feature alignment and consistent cross-modal refinement.

Beyond modality-specific distillation, we shift focus to the fused representation itself. Performing knowledge distillation in this joint feature space allows supervision to act where cross-modal interactions are already encoded. This leads to better spatial reasoning, stronger synergy between modalities, and ultimately improved detection performance.

Finally, we generalize the use of intensity-aware supervision beyond LiDAR by introducing an intensity-guided radar-camera fusion module. This module estimates confidence from both sensors to guide feature fusion. To further improve the fused representation, we incorporate structured supervision from ground truth labels, offering a reliable signal that remains robust to sensor noise and occlusions. Together, these additions strengthen cross-modal learning and reduce dependence on LiDAR-specific guidance.

To validate our approach, we conduct extensive experiments on the nuScenes dataset [1]. IMKD outperforms prior camera-radar knowledge distillation methods and establishes a new benchmark for effective cross-modal supervision in cost-efficient 3D perception.

The main contributions of this paper are listed as follows:

- We present IMKD, an Intensity-Aware Multi-Level Knowledge Distillation framework that enhances camera–radar fusion for 3D object detection, achieving state-of-the-art results among KD-based methods on the nuScenes benchmark [1].
- We design an intensity-aware distillation strategy that preserves the strengths of each sensor modality by guiding knowledge transfer based on high-confidence LiDAR cues. This is applied at multiple stages of the pipeline, enhancing both radar and fused features.
- We perform knowledge distillation in the joint fused feature space instead of individual modalities, allowing supervision to operate where cross-modal cues are already integrated, leading to better spatial reasoning and more robust predictions.
- We introduce an intensity-aware radar-camera fusion module that improves fusion using sensor confidence cues.

2. Related Work

2.1. 3D Object Detection with Multi-Sensor Fusion

Multi-modal 3D object detection combines complementary sensors to boost perception performance. LiDAR-Camera (LC) fusion remains the most accurate setup across datasets [1, 4, 39], implemented via early fusion [49, 52, 54], feature-level fusion [21, 22, 66], and BEV-based methods [25, 31]. However, LiDAR's high cost limits scalability, positioning Camera-Radar (CR) fusion as a cost-effective alternative for long-range, all-weather perception.

CR fusion is more challenging due to view misalignment and sparse radar signals. Early methods like CenterFusion [40] and RadarNet [63] established radar-image associations and multi-level fusion. Recent BEV-based models—MVFusion [61], RADIANT [33], CRAFT [15], CRN [16], and RCFusion [72]—focus on improving cross-modal alignment and radar feature aggregation. RCBEVDet [28] and RCBEVDet++ [27] further refine this pipeline with enhanced fusion strategies.

Building on these advances, our work incorporates knowledge distillation to transfer LiDAR's geometric cues into the CR pipeline, improving fused feature quality while maintaining efficiency.

2.2. Cross-modality Knowledge Distillation

In 3D object detection, traditional Knowledge Distillation (KD) approaches often maintain the same modality for teacher and student models, such as LiDAR-to-LiDAR (L2L) [57, 59, 64, 69] or Camera-to-Camera (C2C) [20, 67, 68]. However, cross-modality KD, which distills knowledge between different sensor modalities, has gained increasing attention. Established cross-modality KD paradigms include LiDAR-to-Camera (L2C) [2, 3, 5, 9, 21, 32] and Camera-to-LiDAR (C2L) [45, 51], with recent advancements exploring fusion-based KD. Methods such as UniDistill [73] and DistillBEV [58] unify features into a shared BEV space to facilitate L2C and LC-to-Camera (LC2C) distillation. Additionally, LabelDistill [14] demonstrates effective label-based distillation for camera and LiDAR models, ensuring robust supervision without relying solely on feature-space alignment. X3KD [17] and CRKD [71] extend KD to LiDAR-Camera-to-Camera-Radar (LC2CR) by introducing adaptive feature alignment, enabling radar-aware distillation while mitigating domain discrepancies.

While CRKD [71] explored fused-to-fused distillation, prior works did not explicitly model modality-to-merged transfer with reliability-aware guidance. IMKD addresses this gap by leveraging LiDAR and label supervision to enhance camera–radar fusion, introducing spatial, depth, and structural cues through intensity-guided distillation.

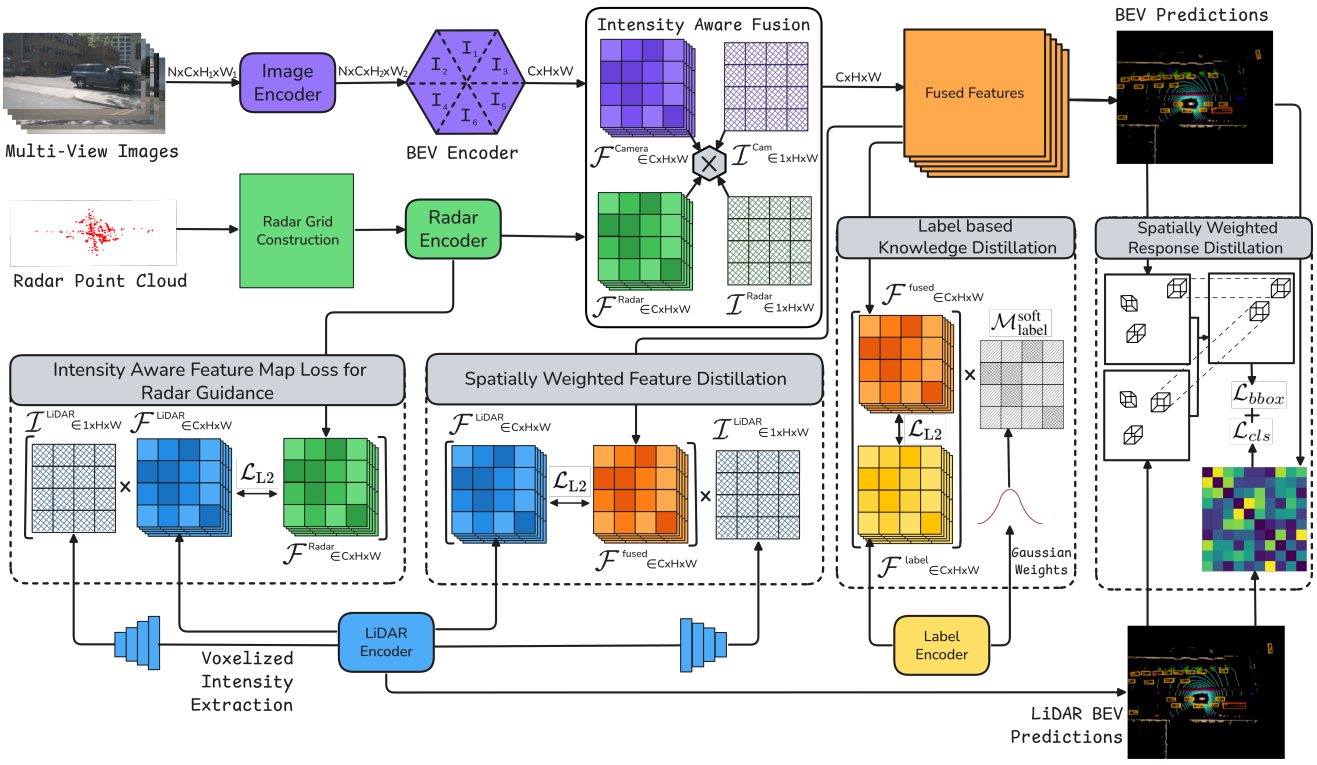


Figure 2. Overview of the proposed Intensity-Aware Multi-Level Knowledge Distillation.

3. Intensity-Aware Multi-Level Knowledge Distillation Framework

IMKD is a framework designed to overcome the shortcomings of existing knowledge distillation approaches for sensor fusion by enhancing radar representation and introducing intensity-aware, multi-level supervision.

Multi-view image features are lifted to BEV using radar-guided depth, while radar features are encoded via a learnable radar-to-grid module. Fusion is performed using intensity-aware deformable cross-attention, leveraging modality-specific intensity maps for precise alignment. LiDAR-derived features refine radar representations before distillation. LiDAR distillation injects spatial priors, while label distillation provides clean, uncertainty-free supervision. At inference, only camera and radar branches are used for efficient deployment.

3.1. Camera Feature Extraction

We extract features from N multi-view images $I_1, \cdots, I_N$ using a convolutional backbone, producing downsampled feature maps $\mathcal{F}_\mathcal{I}$ at a resolution of $1/16$ for each view. These features are refined through additional convolutional layers to generate a context-rich perspective-view feature map $\mathbf{C}_\mathcal{I}^{\mathcal{PV}} \in \mathbb{R}^{N\times C\times H\times W}$:

$$\begin{aligned} \mathbf{C}_\mathcal{I}^{\mathcal{PV}} &= Conv(\mathcal{F}_\mathcal{I}) \\ \mathcal{D}_\mathcal{I}(u,v) &= \text{Softmax}(Conv(\mathcal{F}_\mathcal{I})(u,v)), \end{aligned} \tag{1}$$

where (u, v) denotes pixel coordinates in the image plane and $\mathcal{D}_\mathcal{I} \in \mathbb{R}^{N\times D\times H\times W}$ is the predicted per-pixel depth distribution across D discrete bins. Following the depth-guided view transformation approach [43], we lift the perspective-view features into a frustum-aligned 3D representation $\mathbf{C}_\mathcal{I}^{\mathcal{FV}} \in \mathbb{R}^{N\times C\times D\times H\times W}$:

$$\mathbf{C}_\mathcal{I}^{\mathcal{FV}} = \text{Conv}(\mathbf{C}_\mathcal{I}^{\mathcal{PV}} \otimes \mathcal{D}_\mathcal{I}), \tag{2}$$

where $\otimes$ denotes the outer product between the feature maps and the depth probabilities. The resulting frustum-view features are later aggregated across views and projected into the BEV space as $\mathcal{F}^{\text{Camera}}$.

3.2. Radar Feature Extraction

Our radar processing pipeline transforms sparse, noisy point cloud measurements into dense BEV features suitable for fusion. To maximize the representational power of radar while preserving its sparsity, we design a learnable radar grid construction module followed by a compact yet effective radar encoder.

3.2.1. Radar Grid Construction

Let $\mathcal{P}_{\text{Radar}} = (x_i, y_i, z_i, v_i, \text{RCS}_i)_{i\in\{1,\dots,M\}}$ be the set of M raw radar detections per frame, where (x_i, y_i, z_i) is the 3D position, v_i is the compensated Doppler velocity, and RCS_i is the radar cross-section. Each point is first embedded using a learnable MLP ϕ:

$$\mathbf{f}_i = \phi(x_i, y_i, z_i, v_i, \text{RCS}_i) \in \mathbb{R}^C \tag{3}$$

We then project each radar point onto the BEV plane and map it to a 2D grid cell (u_i, v_i). Instead of using handcrafted statistics, we aggregate features per cell using a differentiable, channel-wise max pooling across all points within that cell:

$$\mathbf{G}_{\mathcal{R}}(u, v) = \text{MaxPool}\left(\mathbf{f}_i \mid (u_i, v_i) = (u, v)\right), \quad \mathbf{G}_{\mathcal{R}} \in \mathbb{R}^{C\times H\times W} \tag{4}$$

This learnable radar-to-grid mapping enables the model to adaptively encode semantic and spatial patterns from raw radar points, replacing brittle hand-engineered descriptors.

3.2.2. Radar Encoder

To extract higher-level features from the radar grid, we design a lightweight encoder $\mathcal{E}_{\mathcal{R}}$ that adapts sparse convolutional designs and point-based reasoning. It consists of 16 convolutional layers grouped into residual blocks with BatchNorm and ReLU activations. Two downsampling stages progressively increase channel depth while reducing spatial resolution, producing BEV features:

$$\mathcal{F}^{\text{Radar}} = \mathcal{E}_{\mathcal{R}}(\mathbf{G}_{\mathcal{R}}) \in \mathbb{R}^{C\times H\times W} \tag{5}$$

The output radar features $\mathcal{F}^{\text{Radar}}$ are resolution-aligned with the camera BEV features, ensuring seamless integration during multimodal fusion.

3.3. Intensity-Aware Feature Fusion

To demonstrate that intensity-aware mechanisms are not limited to LiDAR, we introduce an intensity-aware fusion strategy between camera and radar features. This approach enables modality-aware weighting during fusion, reducing the dominance of camera features and enhancing the contribution of radar in ambiguous or occluded regions.

Radar intensity is computed using both the Radar Cross Section (RCS) and Doppler velocity magnitude. For each radar point i, given its RCS value RCS_i and Doppler velocity components (v_{x_i}, v_{y_i}), the intensity is defined as:

$$\mathcal{I}_i^{\text{Radar}} = \sigma\left(\alpha \cdot \text{RCS}_i + \beta \cdot \sqrt{v_{x_i}^2 + v_{y_i}^2}\right) \tag{6}$$

where α and β are fixed scalar weights, and σ denotes the sigmoid activation function.

For the camera features, a spatial intensity map $\mathcal{I}^{\text{Cam}} \in \mathbb{R}^{N\times 1\times H\times W}$ is generated using a convolutional layer:

$$\mathcal{I}^{\text{Cam}} = \sigma(\text{Conv}(\mathcal{F}^{\text{Camera}})) \tag{7}$$

We adopt a deformable attention formulation in BEV space, where radar BEV features $\mathcal{F}^{\text{Radar}}$ serve as queries ($\mathbf{Q}$), and camera BEV features $\mathcal{F}^{\text{Camera}}$ serve as keys ($\mathbf{K}$) and values ($\mathbf{V}$). The fusion is computed as:

$$\mathcal{F}^{\text{fused}} = \text{DeformAttn}_{\text{intensity}}(\mathbf{Q} = \mathcal{F}^{\text{Radar}}, \mathbf{K} = \mathcal{F}^{\text{Camera}}, \mathbf{V} = \mathcal{F}^{\text{Camera}}, \mathcal{I}^{\text{Cam}}, \mathcal{I}^{\text{Radar}}) \tag{8}$$

This mechanism allows the network to learn spatially-varying cross-modal weights, reducing dominance from any single modality and improving the relevance of fused features.

3.4. Adaptive Intensity-Guided Radar Feature Enhancement

While LiDAR intensity primarily reflects surface reflectivity, it also acts as a proxy for geometric confidence; high-intensity returns often correspond to structured and reflective objects such as vehicles or road boundaries. Rather than using intensity as a semantic indicator, we leverage it to prioritize supervision from high-confidence LiDAR regions during knowledge distillation. Crucially, instead of directly relying on raw intensities, we learn a transformation over the intensity map, enabling the network to adaptively reweight spatial contributions as shown in Fig. 3.

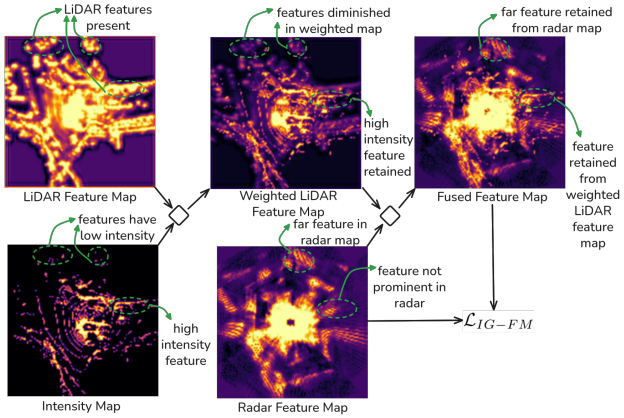


Figure 3. We illustrate a weighted LiDAR feature map, generated from intensity, is merged with the radar feature map to compute the intensity-guided feature map loss, preserving radar features and avoiding low-intensity LiDAR regions.

3.4.1. Voxelized LiDAR Intensity Extraction:

To derive intensity-aware guidance, we first voxelize LiDAR points $\mathcal{P}_{\text{LiDAR}} = \{(x_i, y_i, z_i, I_i, t_i)\}_{i=1}^{N}$, where each

point contains spatial coordinates (x_i, y_i, z_i), intensity I_i, and timestamp t_i. Using a voxelization function $\mathcal{V}$, we partition the points into discrete 3D voxels and compute the mean voxel intensity I_v as:

$$I_v = \frac{1}{N_v} \sum_{i=1}^{N_v} I_i, \tag{9}$$

where N_v denotes the number of points within a voxel. This intensity is then projected onto a BEV grid $\mathcal{I}_{\text{LiDAR}} \in \mathbb{R}^{H \times W}$, producing an intensity map aligned with the LiDAR BEV features:

$$\mathcal{I}^{\text{LiDAR}}(u, v) = \frac{\sum_i I_v \cdot \delta(u - u_i, v - v_i)}{\max(1, \sum_i \delta(u - u_i, v - v_i))}, \tag{10}$$

where (u, v) are BEV grid indices from voxel projection, and δ denotes voxel-to-grid mapping for spatial alignment.

3.4.2. LiDAR-Weighted Radar Feature Fusion:

We utilize the intensity map $\mathcal{I}^{\text{LiDAR}}$ as an adaptive weighting mechanism to guide the fusion of LiDAR and radar features. Given LiDAR feature map $\mathcal{F}^{\text{LiDAR}}$ and radar feature map $\mathcal{F}^{\text{Radar}}$, we compute intensity-based feature blending weights as:

$$w_{\text{LiDAR}} = \lambda \cdot \mathcal{I}^{\text{LiDAR}}, \quad w_{\text{Radar}} = 1 - w_{\text{LiDAR}}, \tag{11}$$

where λ is a learnable scaling factor. The fused BEV feature map is then constructed as:

$$\tilde{\mathcal{F}}^{\text{Radar}} = w_{\text{LiDAR}} \cdot \mathcal{F}^{\text{LiDAR}} + w_{\text{Radar}} \cdot \mathcal{F}^{\text{Radar}}. \tag{12}$$

3.4.3. Intensity-Aware Feature Map Loss for Radar Guidance:

To reinforce radar feature refinement, we introduce an alignment loss between radar features and the LiDAR feature map to encourage consistency between the two modalities:

$$\mathcal{L}_{\text{align}} = \left\| \mathcal{F}^{\text{Radar}} - \mathcal{F}^{\text{LiDAR}} \right\|^2. \tag{13}$$

Simultaneously, we maintain a consistency loss between the radar and fused feature maps to prevent oversuppression of radar-specific information:

$$\mathcal{L}_{\text{consist}} = \left\| \mathcal{F}^{\text{Radar}} - \tilde{\mathcal{F}}^{\text{Radar}} \right\|^2. \tag{14}$$

The final Intensity-Guided Feature Map Loss is formulated as:

$$\mathcal{L}_{\text{IG-FM}} = \alpha \mathcal{L}_{\text{align}} + (1 - \alpha) \mathcal{L}_{\text{consist}}, \tag{15}$$

where α balances alignment and consistency constraints.

3.5. LiDAR-Guided Feature Enhancement

We directly supervise the fused camera–radar BEV representation using LiDAR as a privileged modality. This process is challenged by semantic gaps between LiDAR and fused features, temporal misalignment across modalities, and instability from the high-dimensional, semantically enriched representation. To ensure stable and spatially-aware transfer, we apply LiDAR intensity as soft attention instead of binary or uniform weighting.

3.5.1. Spatially-Weighted Feature Distillation

We guide the fused BEV feature $\mathcal{F}^{\text{fused}} \in \mathbb{R}^{C \times H \times W}$ using LiDAR features $\mathcal{F}^{\text{LiDAR}}$ with spatial weighting from normalized LiDAR intensity $\mathcal{I}^{\text{LiDAR}} \in \mathbb{R}^{1 \times H \times W}$. The distillation loss is:

$$\mathcal{L}_{\text{SWFD}} = \left\langle \mathcal{I}_{ij}^{\text{LiDAR}} \cdot \left\| \mathcal{F}_{ij}^{\text{LiDAR}} - \beta(\mathcal{F}_{ij}^{\text{fused}}) \right\|_2^2 \right\rangle_{i,j} \tag{16}$$

Here, $\beta(\cdot)$ is a lightweight alignment module. This formulation ensures feature transfer is stronger in high-confidence regions while preserving full gradient flow—something not possible with handcrafted binary masks.

3.5.2. Spatially-Weighted Response Distillation

To further align the predictions of the fused detector with the LiDAR teacher, we introduce a Spatially-Weighted Response Distillation loss.

$$\begin{aligned} \mathcal{L}_{\text{SWRD}} = \big\langle \mathcal{I}_{ij}^{\text{LiDAR}} \cdot \big(& \mathcal{L}_{cls}(h^{\text{LiDAR}}, h^{\text{fused}}) \\ & + \mathcal{L}_{bbox}(b^{\text{LiDAR}}, b^{\text{fused}}) \big) \big\rangle_{i,j} \end{aligned} \tag{17}$$

Here, $\mathcal{L}_{cls}$ and $\mathcal{L}_{bbox}$ denote the standard classification and bounding-box regression losses (as in CenterPoint [65]), with h representing the predicted class heatmap and b representing predicted bounding-box coordinates. This improves upon prior works by avoiding uniform foreground masks and instead applying confidence-weighted distillation across the full spatial domain.

3.6. Label-Based Knowledge Distillation

We extend LabelDistill [14] to operate directly on fused camera-radar BEV features rather than Camera-only features. This avoids reliance on modality-specific artifacts and ensures robust supervision over the multi-modal representation.

In contrast to the binary object masks used in [14], we introduce a soft Gaussian mask centered on ground-truth boxes to enable smooth, graded supervision. The final loss is:

$$\mathcal{L}_{\text{LD}} = \frac{\sum \left\| \mathcal{F}^{\text{label}} - \mathcal{F}^{\text{fused}} \right\|^2 \cdot \mathcal{M}_{\text{label}}^{\text{soft}}}{\sum \mathcal{M}_{\text{label}}^{\text{soft}} + \epsilon} \tag{18}$$

Here, $\mathcal{F}^{\text{label}}$ is the label-encoded BEV feature and $\mathcal{M}^{\text{soft}}_{\text{label}}$ softly weights the loss around valid object regions. This fused, soft-masked formulation improves generalization while preserving clean label supervision.

3.7. Overall Loss Function and Training Strategy

The total training objective combines detection, depth estimation, and distillation terms as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{det}} + \lambda_2\mathcal{L}_{\text{depth}} + \lambda_3\mathcal{L}_{\text{IG-FM}} \\ + \lambda_4\mathcal{L}_{\text{SWFD}} + \lambda_5\mathcal{L}_{\text{SWRD}} + \lambda_6\mathcal{L}_{\text{LD}}\end{aligned} \tag{19}$$

All components are jointly optimized, with frozen LiDAR and label encoders used only during training. The modular design enables efficient fusion learning without inference-time overhead.

4. Experiments

4.1. Experimental Setup

Dataset and Evaluation Metrics

We evaluate on the nuScenes dataset [1], which includes LiDAR, radar, and camera data across 1,000 scenes (850 for training/validation, 150 for testing).
3D object detection is assessed using official nuScenes [1] metrics: mAP, NDS, and five TP metrics: mATE, mASE, mAOE, mAVE, and mAAE.

All results are reported on the nuScenes [1] validation and test sets for fair comparison with prior work.

Implementation Details

We use pretrained CenterPoint [65] as the LiDAR teacher with $(0.1\text{m}, 0.1\text{m}, 0.2\text{m})$ voxel size and adopt the label encoder from LabelDistill [14]. The camera branch is based on BEVDepth [23] with efficient depth layers. Radar inputs use multi-sweep projections with Doppler and RCS normalization, processed into a polar-to-BEV feature map. Temporal fusion follows BEVFormer [24] by accumulating four BEV frames at 1s intervals, ensuring causal inference.

We use an ImageNet-pretrained ResNet50 [8] backbone, with input size 256×704, trained using AdamW [35]. Data augmentations are applied across all modalities. Detailed architectural and training configurations are provided in the supplementary material.

4.2. Main Results and Comparison with State-of-the-Art

We evaluate IMKD on the nuScenes [1] validation and test sets, comparing against a range of distillation-based 3D object detectors, including both camera-only and camera-radar student models. Results are summarized in Tables 1 and 2.

On the validation set, IMKD achieves 61.0 NDS and 51.6 mAP, outperforming all prior KD-based methods. Notably, it surpasses the strongest baseline, CRKD, by +6.5% NDS and +10.1% mAP, along with consistent reductions in translation, scale, and orientation errors. These gains highlight the effectiveness of IMKD's fusion-level distillation and intensity-aware supervision.

On the nuScenes [1] test set, IMKD sets a new benchmark with 67.0 NDS and 61.0 mAP, significantly improving upon previous knowledge-distillation-based results. These improvements affirm that enhancing the quality and contextual relevance of supervision, rather than relying on modality-specific or heuristic KD, is key to unlocking robust camera-radar perception. Overall, these results demonstrate that IMKD offers a robust and generalizable approach to multi-modal knowledge distillation, advancing the field beyond conventional KD strategies and establishing a new standard for student models in 3D detection.

Qualitative results in Fig. 4 further illustrate IMKD's impact: knowledge distillation on fused modalities yields

Method	Input	KD	Backbone	Image Size	NDS ↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
UVTR [21]	C	L2C	R101	900×1600	45.0	37.2	0.735	0.269	0.397	0.761	0.193
BEVDistill [2]	C	LC2C	R50	640×1600	45.7	38.6	0.693	0.264	0.399	0.802	0.199
UniDistill [73]	C	L2C	R50	256×704	37.8	26.5	-	-	-	-	-
BEVSimDet [14]	C	LC2C◇2C	SwinT	256×704	45.3	40.4	0.526	0.275	0.607	0.805	0.273
LabelDistill [14]	C	LL◇2C	R50	256×704	52.8	41.9	0.582	0.258	0.413	0.346	0.220
DistillBEV [14]	C	L2C◇2C	R50	256×704	41.6	34.0	0.704	0.266	0.556	0.815	0.201
X3KD [17]	C	LC2C	R50	256×704	50.5	39.0	0.615	0.269	0.471	0.345	0.203
X3KD [17]	C+R	L2CR	R50	256×704	53.8	42.3	-	-	-	-	-
CRKD [71]	C+R	LC2CR	R50	256×704	57.3	46.7	0.446	0.263	0.408	0.331	0.162
IMKD (Ours)	C+R	LL◇2M	R50	256 × 704	**61.0**	**51.6**	**0.444**	**0.259**	**0.384**	**0.229**	**0.160**

Table 1. Comparison of Knowledge distillation (KD) methods for 3D object detection results on the nuScenes [1] val set. 'L', 'L◇' 'C', 'R' and 'M' denote LiDAR, label, camera, radar and merged (camera+radar) inputs, respectively.

Method	Input	KD	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
UVTR [21]	C	L2C	52.2	45.2	0.612	0.256	0.385	0.664	0.125
BEVDistill [2]	C	LC2C	59.4	49.8	0.472	0.247	0.378	0.326	0.125
UniDistill [73]	C	L2C	39.3	29.6	0.637	0.257	0.492	1.084	0.167
LabelDistill [14]	C	LL◊2C	61.0	52.6	0.443	0.252	0.339	0.370	0.136
X3KD [17]	C	LC2C	56.1	45.6	0.506	0.253	0.414	0.366	0.131
X3KD [17]	C+R	L2CR	55.3	44.1	-	-	-	-	-
CRKD [71]	C+R	LC2CR	58.7	48.7	0.404	0.253	0.425	0.376	0.111
IMKD (Ours)	C+R	LL◊2M	**67.0**	**61.0**	**0.401**	**0.249**	**0.305**	**0.238**	**0.102**

Table 2. Comparison of Knowledge Distillation (KD) methods for 3D object detection on the nuScenes [1] test set. 'L', 'L◊', 'C', 'R', and 'M' denote LiDAR, label, camera, radar, and merged (camera+radar) inputs, respectively. *Note: Each method uses its best reported backbone and image size; comparisons should focus on distillation strategies.*

more accurate detections and better box orientation compared to individual-modality KD, especially in ambiguous or occluded scenes.

In the following section 4.3, we present ablation studies to analyze the contributions of each IMKD component. Unless otherwise specified, all experiments are conducted on the nuScenes [1] validation set using a ResNet-50 [8] image backbone. We report standard metrics including mAP, NDS, and detailed error breakdowns to evaluate performance comprehensively.

4.3. Ablation Study

We conduct a step-wise ablation to isolate the impact of each component in IMKD. Starting from a camera-only baseline, we progressively introduce radar grid learning, intensity-aware fusion, and our distillation modules.

4.3.1. Effectiveness of Learnable Radar Grid

We assess the effect of radar input design on detection performance in Tab. 3. Introducing radar via a fixed handcrafted grid boosts performance over the camera-only setup, confirming the benefit of multi-modal fusion. Switching to a learnable radar grid further improves both mAP and NDS, validating its role in producing task-adaptive radar features.

Method	mAP ↑	NDS ↑
Camera Only	34.8	44.6
Camera+Radar (Handcrafted)	41.2	52.5
Camera+Radar (Learnable) - Baseline	43.4	53.5

Table 3. Ablation on radar representation. Learnable radar grid improves over both camera-only and handcrafted radar projection.

4.3.2. Impact of Intensity-Aware C+R Fusion

We incorporate both radar and camera intensity to guide the fusion process. This leads to notable gains over the learnable grid baseline as shown in Tab. 4, validating the benefit of confidence-aware fusion. Importantly, this shows that intensity-guided processing is beneficial beyond LiDAR and can be generalized to radar-camera fusion.

Method	mAP ↑	NDS ↑
Baseline	43.4	53.5
+ Intensity-Aware Fusion	46.5	55.3

Table 4. Comparison of non-intensity-based vs. intensity-aware fusion for merged camera-radar features.

4.3.3. Effectiveness of Proposed KD Modules

We incrementally evaluate the effectiveness of our proposed distillation objectives on top of the intensity-aware fusion baseline in Tab. 5. Each module, i.e. LiDAR feature distillation, label distillation, intensity-guided feature map supervision, and response distillation yields consistent performance gains, validating their individual contribution. When combined, they offer additive improvements, with the full IMKD model achieving a notable +11% mAP and +9.3% NDS over the baseline. This demonstrates the effectiveness of our modular, fusion-aligned KD framework in enhancing multi-modal perception.

Configuration	**mAP ↑**	**NDS ↑**
Intensity-Aware Fusion	46.5	55.3
+ LiDAR Feature Distill ($\mathcal{L}_{\text{SWFD}}$)	49.56	58.30
+ Label Distill ($\mathcal{L}_{\text{LD}}$)	49.41	58.25
+ Intensity-Guided Feature Map ($\mathcal{L}_{\text{IG-FM}}$)	49.64	58.40
+ Response Distill ($\mathcal{L}_{\text{SWRD}}$)	47.68	57.31
+ Response + LiDAR Distill.	49.72	58.51
+ Resp. + LiDAR + Label Distill.	50.22	59.14
IMKD (Full Model)	**51.65**	**61.05**

Table 5. Ablation on proposed distillation objectives over the intensity-guided fusion baseline.

4.3.4. Cross-Modal vs Uni-Modal Distillation

We compare distillation into individual modalities versus fused features. While unimodal KD achieves solid results, fusion-level KD yields +4.0% mAP and +4.6% NDS as shown in Tab. 6. These gains require mitigating gradient

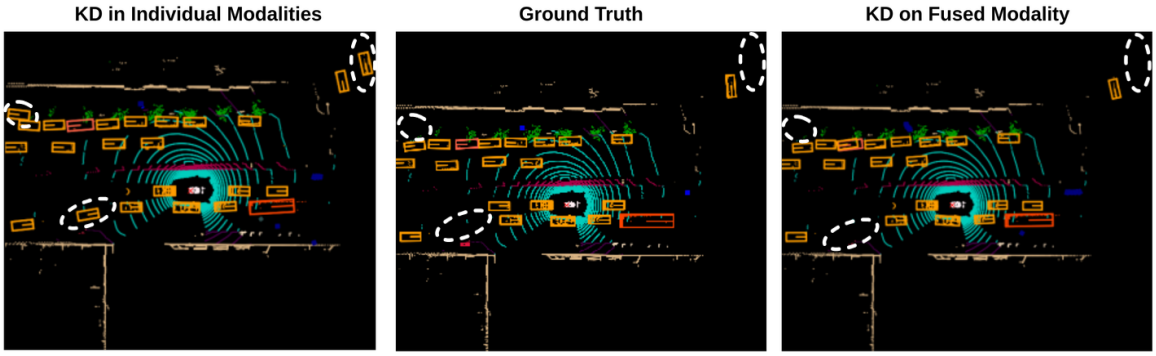


Figure 4. Comparison of distillation targets: individual modality KD yields extra false detections (white circles) and poor orientation, while merged feature KD aligns better with ground truth.

conflicts and applying pseudo-label masking and feature normalization, underscoring the complexity of fusion-level supervision. This validates our hypothesis that modality interaction within the distillation target space plays a crucial role in enhancing downstream performance.

LiDAR KD Target	Label KD Target	mAP ↑	NDS ↑
Camera & Radar	Camera & Radar	49.8	58.3
Fused	Camera & Radar	50.5	59.3
Camera & Radar	Fused	50.2	59.0
Fused	Fused	**51.6**	**61.0**

Table 6. Comparison of distillation targets using individual modalities vs. fused camera-radar features.

4.4. Robustness to Visibility and Temporal Degradation

To our knowledge, IMKD is the first distillation-based fusion framework to evaluate performance under diverse environmental conditions (e.g., rain, night) and degraded temporal input (frame drop). Prior KD-based methods such as CRKD [71], X3KD [17], and LabelDistill [14] focus solely on standard benchmarks, leaving real-world robustness unexplored. Our analysis in Tab. 7 reveals that IMKD exhibits greater stability in adverse conditions, suggesting that confidence-aware distillation leads to more reliable multi-modal fusion. Although IMKD uses LiDAR during training, it remains effective even when supervision is noisy or partially missing, as the learned guidance is intensity-adaptive and spatially grounded, rather than hard-coded. To contextualize these results, we also report robustness for non-KD fusion methods (CRN [16], RCBEV [74]). While not directly comparable, they offer a useful reference point for deployment. IMKD consistently maintains superior performance across weather and frame drop scenarios Tab. 8, demonstrating that confidence-guided distillation improves both accuracy and resilience in safety-critical conditions.

Input	Modality	Sunny	Rainy	Day	Night
RCBEV [74]	C+R	36.1	38.5	37.1	15.5
BEVDepth [23]	C	39.0	39.0	39.3	16.8
CRN [16]	C+R	54.8	57.0	55.1	30.4
IMKD (Ours)	C+R	**57.9**	**58.5**	**58.3**	**34.7**

Table 7. mAP under varying weather and lighting on nuScenes [1] val set, where IMKD outperforms other methods.

Method	Input	Modality	Drop 0	Drop 1	Drop 3	Drop 6
CRN [16]	C+R	C	47.18	40.19	22.94	-
		R		45.39	41.34	34.52
IMKD	C+R	C	**51.6**	**43.28**	**23.53**	-
		R		**49.47**	**44.40**	**36.35**

Table 8. mAP under increasing frame drops on nuScenes [1] val set, where IMKD remains more stable across sensor degradations.

5. Conclusion

We proposed IMKD, an Intensity-guided Multi-Level Knowledge Distillation framework for radar-camera 3D object detection. IMKD identifies a core limitation in existing multi-modal distillation methods: modality-specific supervision often leads to incoherent representations and suboptimal fusion. To address this, we introduce a merged feature distillation strategy and an intensity-aware refinement module that prioritizes high-confidence regions during training. Although IMKD is trained using LiDAR as a privileged modality, its design is agnostic to specific sensor pairs and may be extended to other domains where confidence-guided supervision is available. Through extensive ablations and comparisons on nuScenes [1], we demonstrated that IMKD delivers consistent improvements over state-of-the-art distillation baselines, validating both the design and effectiveness of our proposed framework. As a future direction, we plan to extend IMKD to temporal multi-frame fusion, enabling dynamic scene understanding and improved consistency in long-term predictions.

6. Acknowledgements

This work was partially funded by the European Union under Grant Agreement No. 101076360 (BERTHA) and by the German Federal Ministry of Research, Technology and Space (BMFTR) under Grant Agreement No. 01IW24009 (COPPER).

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6, 7, 8, 4, 5, 9, 10

[2] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022. 2, 6, 7

[3] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *arXiv preprint arXiv:2201.10830*, 2022. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[5] 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

[6] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 6

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 2, 3, 4

[9] 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. Springer, 2022. 2

[10] Yihan Hu, Zhuangzhuang Ding, Runzhou Ge, Wenxin Shao, Li Huang, Kun Li, and Qiang Liu. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 969–979, 2022. 1

[11] 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. 2, 5

[12] Yan Huang, Yibin Ren, and Xiaofeng Li. Deep learning techniques for enhanced sea-ice types classification in the beaufort sea via sar imagery. *Remote Sensing of Environment*, 308:114204, 2024. 1

[13] 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. 4, 5

[14] Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum. Labeldistill: Label-guided cross-modal knowledge distillation for camera-based 3d object detection. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024. 1, 2, 5, 6, 7, 8

[15] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1160–1168, 2023. 2, 6, 7

[16] 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. 2, 8, 4, 5, 6, 7

[17] 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, 7, 8

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1, 6

[19] K Lei, Z Chen, S Jia, and X Zhang. Hvdetfusion: A simple and robust camera-radar fusion framework. arxiv. 2023. *arXiv preprint arXiv:2307.11323*, 2023. 5

[20] 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. *IEEE Transactions on Intelligent Vehicles*, 9(1):2489–2498, 2023. 2

[21] 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, 6, 7

[22] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1120–1129, 2022. 2

[23] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 6, 8, 4, 5, 7

[24] Z Li, W Wang, H Li, E Xie, C Sima, T Lu, Q Yu, and J Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arxiv 2022. *arXiv preprint arXiv:2203.17270*, 2022. 6

[25] 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. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 2

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[27] Zhiwei Lin, Zhe Liu, Yongtao Wang, Le Zhang, and Ce Zhu. Rcbevdet++: Toward high-accuracy radar-camera fusion 3d perception network. *arXiv preprint arXiv:2409.04979*, 2024. 2

[28] 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. 2, 4, 5, 6

[29] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[30] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 4, 5

[31] 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 *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2

[32] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1790–1798, 2023. 2

[33] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. Radiant: Radar-image association network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1808–1816, 2023. 2

[34] Yunfei Long, Abhinav Kumar, Xiaoming Liu, and Daniel Morris. Riccardo: Radar hit prediction and convolution for camera-radar 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22276–22285, 2025. 4, 5

[35] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[37] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023. 6

[38] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3164–3173, 2021. 1

[39] 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 *European Conference on Computer Vision*, pages 53–72. Springer, 2022. 2

[40] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2021. 2, 6, 7

[41] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. 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. 5, 6

[42] 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. *arXiv preprint arXiv:2210.02443*, 2022. 5

[43] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3, 6

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 2

[46] Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. *arXiv preprint arXiv:2403.11761*, 2024. 6

[47] Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, and Muhammad Aslam. Deep learning-based change detection in remote sensing images: A review. *Remote Sensing*, 14(4):871, 2022. 1

[48] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-

rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023. 1

[49] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 2

[50] Lukas Stäcker, Shashank Mishra, Philipp Heidenreich, Jason Rambach, and Didier Stricker. Rc-bevfusion: A plug-in module for radar-camera bird's eye view feature fusion. In *DAGM German Conference on Pattern Recognition*, pages 178–194. Springer, 2023. 5

[51] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024. 2

[52] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2

[53] Biao Wang, Xialei Wu, Yijun Zhang, Lijuan Zheng, Guangliang Yang, and Zhiwei Xu. Gaussian focal loss for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14656, 2021. 3

[54] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11794–11803, 2021. 2

[55] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 5

[56] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 5

[57] Yue Wang and Justin M Solomon. Object dgcnn: 3d object detection using dynamic graphs. *Advances in Neural Information Processing Systems*, 34:20745–20758, 2021. 2

[58] 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

[59] 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 *European Conference on Computer Vision*, pages 179–195. Springer, 2022. 2

[60] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7467–7474. IEEE, 2025. 5

[61] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2766–2773. IEEE, 2023. 2, 5

[62] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[63] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *European conference on computer vision*, pages 496–512. Springer, 2020. 2, 6

[64] Jihan Yang, Shaoshuai Shi, Runyu Ding, Zhe Wang, and Xiaojuan Qi. Towards efficient 3d object detection with knowledge distillation. *Advances in Neural Information Processing Systems*, 35:21300–21313, 2022. 2

[65] 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. 1, 5, 6, 3, 4

[66] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*, pages 720–736. Springer, 2020. 2

[67] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 992–1001, 2023. 2

[68] 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

[69] Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, and Kaisheng Ma. Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21791–21801, 2023. 2

[70] 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. 2

[71] 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, 6, 7, 8

[72] Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rc-fusion: Fusing 4-d radar and camera with bird's-eye view features for 3-d object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023. 2, 6

[73] 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, 6, 7

[74] 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. 8

[75] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 7

[76] 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. 4, 5

IMKD: Intensity-Aware Multi-Level Knowledge Distillation for Camera-Radar Fusion

Supplementary Material

7. Overview

This supplementary material provides additional details on our proposed approach, including architectural design choices, implementation specifics, and extended experimental results. In Sec. 8, we describe the network architecture and key design decisions. Sec. 9 covers implementation details, including data preprocessing, hyperparameters, training configuration, and the impact of feature partitioning on fusion. In Sec. 10, we present additional experimental results, such as per-class performance analysis. Finally, Sec. 11 provides qualitative results to further illustrate the effectiveness of our method.

8. Architectural Details

8.1. Motivation for Intensity-Guided Distillation

LiDAR intensity encodes the strength of signal returns, which is closely tied to geometric reliability and boundary consistency [18, 36, 65]. In IMKD, intensity is not transferred as a raw feature; instead, it is used to guide knowledge distillation and fusion. Specifically, LiDAR supervision is intensity-weighted when transferring features to the camera–radar fused representation, while camera and radar intensities are also used to modulate their deformable fusion. This ensures that distillation emphasizes reliable LiDAR regions, aligns multi-sensor features, and sharpens fused predictions.

The motivation is threefold: intensity provides a reliability prior that (i) emphasizes consistent LiDAR features during transfer, preventing noisy regions from dominating; (ii) improves alignment of camera–radar fusion by highlighting structurally meaningful areas for cross-modal attention; and (iii) refines prediction confidence by guiding the fused BEV representation toward sharper object boundaries and more stable detections.

This design avoids directly forcing radar to mimic LiDAR, preserving radar's modality-specific robustness (e.g., under adverse weather), while still leveraging LiDAR's depth-rich supervision. Similar strategies have proven effective beyond autonomous driving. In Medical imaging, MRI and CT often leverage intensity-weighted priors to guide segmentation, where voxel intensity correlates with tissue density and boundary sharpness [29, 44]. In Remote sensing, satellite imagery, and radar backscatter intensity is used to enhance feature fusion for land-cover classification and flood detection, where high-return regions correspond to structurally reliable terrain [12, 47].

These parallels show that intensity is a widely validated proxy for reliability and structure across domains. By incorporating it into the distillation process, IMKD enhances the transfer of geometric and structural knowledge without erasing modality-specific strengths.

8.2. Architectural Design Considerations

Our architecture is designed with modularity and supervision efficiency in mind. While the main paper details the overall pipeline, here we highlight key considerations behind specific design choices that enhance robustness and enable clean integration of privileged signals.

Intensity-Aware Cross-Modality Fusion: Our architecture fuses camera and radar features using a deformable attention mechanism guided by both camera confidence and radar intensity maps. This dual-intensity guidance enables the network to adaptively align features across modalities, prioritizing reliable regions and suppressing noise. Unlike modality-agnostic or uniform fusion schemes, our design selectively emphasizes trustworthy cues from each sensor, leading to more robust representations under challenging conditions such as rain, night, or partial sensor failure.

In Eq. 8, the camera and radar intensity maps $\mathcal{I}^{\text{Cam}}, \mathcal{I}^{\text{Radar}}$ serve as modulation signals within the deformable attention module. Specifically, intensity values are concatenated with key–value embeddings and passed through a learned gating function $g(\cdot)$, which rescales both the sampling offsets and the attention weights:

$$w_{ij} = \text{softmax}\big((\mathbf{q}_i \cdot \mathbf{k}_j) \cdot g(\mathcal{I}_j)\big), \tag{20}$$

where $g(\cdot)$ is a lightweight MLP with sigmoid activation. This mechanism ensures that attention is biased toward high-intensity regions (i.e., geometrically reliable points), enabling intensity-aware feature selection and fusion.

Intensity-Guided Radar Representation: Radar intensity is used to modulate the radar branch features before fusion. Although simple, this plays a vital role in enhancing geometric priors, especially under sensor degradation (e.g., frame drops or poor lighting). This design avoids the need for radar-specific heuristics or handcrafted filters.

Late Injection of Supervision: All remaining distillation losses are injected post-fusion, reducing the risk of modality dominance and preserving the integrity of radar

features during training. This ensures that supervision acts as a guidance mechanism, not a constraint.

Drop-In Extensibility: The design is easily extendable to other sensor pairs, e.g., camera+thermal or camera+event. Our use of post-fusion supervision and intensity-aware enhancement ensures that new modalities can be added without major architectural changes.

These choices, while not architectural novelties in isolation, collectively enable IMKD to scale well under different conditions and sensor setups with minimal adjustments.

8.3. Inference Pipeline

During inference, our model operates efficiently using only camera and radar inputs, ensuring a lightweight and deployable architecture. Several components used during training are discarded, streamlining computation without compromising detection performance.

Components Removed at Inference:

LiDAR Feature Maps: Since LiDAR supervision is only utilized during training to inject spatial priors, these feature maps are not required at test time.

Label Encoder: The label encoder, responsible for transforming ground truth 3D bounding boxes into a BEV representation, is used solely for training supervision and is omitted during inference.

Efficient Operation with Camera and Radar Inputs: At inference, multi-view camera and radar features are first projected into a BEV space. These BEV features are then fused using an intensity-aware deformable fusion module, which leverages both camera confidence scores and radar intensity maps to guide spatial alignment. This design ensures robustness under adverse conditions by emphasizing high-confidence regions from each modality. Although LiDAR and label supervision are used during training, they are not required at test time. As a result, our method achieves accurate and efficient 3D detection using only camera and radar inputs, making it practical for real-world deployment.

8.4. Inference Time

We evaluate the inference speed of our IMKD framework on an RTX 3090 GPU using a single batch with FP16 precision. With a ResNet-50 [8] backbone, our method achieves real-time performance at 25 FPS, making it competitive with existing camera-radar fusion approaches. Our knowledge distillation framework is employed solely during training and introduces no additional latency during inference.

Among knowledge distillation-based 3D detection methods, only BEVSimDet [70] and UVTR-C [21] report inference speeds—11.1 FPS and 3.1 FPS, respectively—while BEVDet-Tiny [11] (a camera-only baseline) runs at 15.6 FPS. Other methods, such as UniDistill [73], LabelDistill [14], DistillBEV [2], X3KD [17] and CRKD [71], do not disclose inference performance. In contrast, our IMKD model delivers 25 FPS while outperforming these methods in detection accuracy, highlighting its strong balance between efficiency and robustness for real-world deployment.

Method	Type	FPS
BEVDet-Tiny [11]	Camera-Only	15.6
UVTR-C [21]	KD-Based	3.1
BEVSimDet [70]	KD-Based	11.1
IMKD (Ours)	KD-Based	**25.0**

Table 9. Comparison of inference speeds (FPS) across KD-based and camera-only baselines. Our method achieves real-time performance while maintaining strong accuracy.

8.5. Loss Function Weight Tuning

The weights for individual loss terms in Eq. (19) are empirically tuned to ensure balanced contributions during training. Specifically, the detection and depth losses (λ_1, λ_2) are set to 0.3, while the LiDAR- and label-based distillation losses ($\lambda_4, \lambda_5, \lambda_6$ in Eq. (16), Eq. (17), and Eq. (18)) are also weighted at 0.3 to provide auxiliary supervision without overwhelming the primary objectives. The radar distillation loss (λ_3 in Eq. (15)) is governed by a learnable scalar, initialized at 100, which allows the network to adaptively adjust its relative contribution during training and reduces manual sensitivity. Within Eq. (15), the alignment–consistency trade-off is controlled by $\alpha = 0.5$, which provides a balanced emphasis across geometric consistency and feature alignment.

Loss Term	**Symbol**	**Weight**
Detection loss ($\mathcal{L}_{\text{det}}$)	λ_1	0.3
Depth loss ($\mathcal{L}_{\text{depth}}$)	λ_2	0.3
Intensity-Guided Feature Map ($\mathcal{L}_{\text{IG-FM}}$)	λ_3	Learn., init. 100
LiDAR Feature Distill ($\mathcal{L}_{\text{SWFD}}$)	λ_4	0.3
Response Distill ($\mathcal{L}_{\text{SWRD}}$)	λ_5	0.3
Label Distill ($\mathcal{L}_{\text{LD}}$)	λ_6	0.3
Alignment-consistency trade-off	α	0.5

Table 10. Loss functions and corresponding weights used in IMKD.

These settings were chosen after preliminary sweeps to equalize the order of magnitude of gradients from each term, preventing instability from any single loss. We observed that training remained stable across all experiments without requiring further re-tuning, indicating that the framework is not overly sensitive to precise hyperparameter choices. The final values used in all experiments

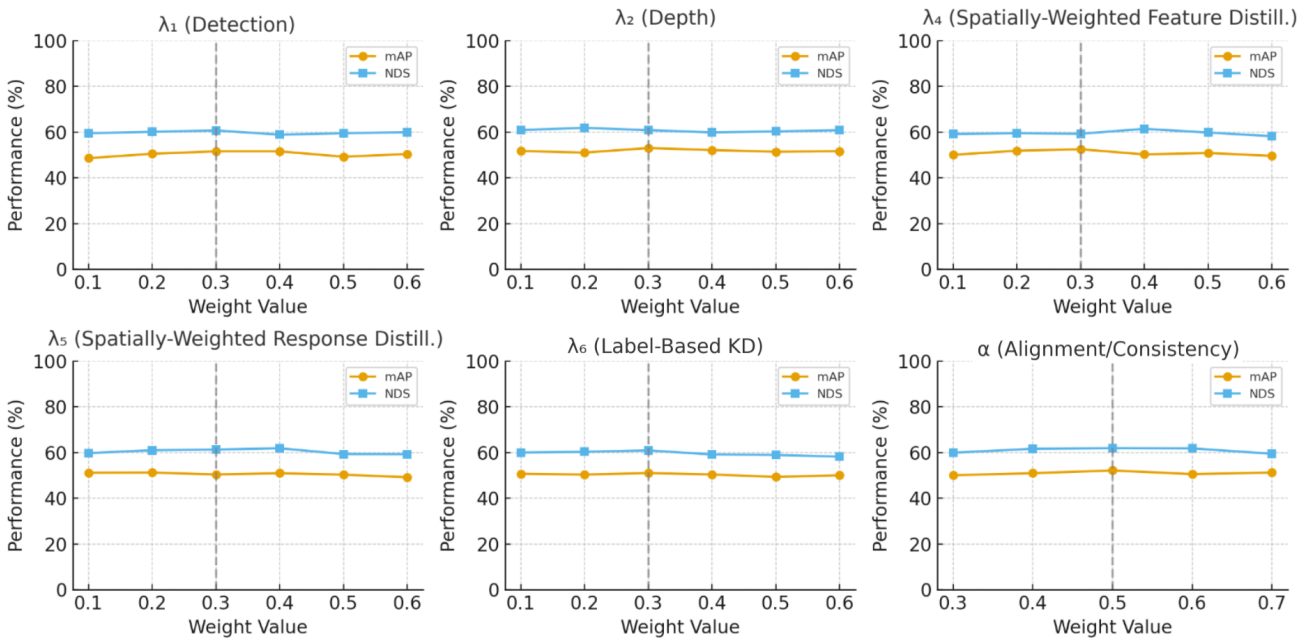


Figure 5. Sensitivity of mAP and NDS to individual loss weights λ. Each subplot reports an illustrative sweep over $\lambda \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$; dashed vertical lines mark the chosen operating points ($\lambda = 0.3$ for most terms, $\alpha = 0.5$ for alignment). The curves indicate that performance is stable near the chosen weights and degrades when weights deviate substantially.

are summarized in Tab. 10 for reproducibility. This stability is also illustrated in Fig. 5, where we plotted the values of all loss weights. The curves show that performance remains largely stable near the chosen weights, while substantial deviations can lead to degradation, confirming that the selected operating points strike a robust balance across losses.

9. Implementation Details

9.1. Pre-Processing

Pre-processing
Our method utilizes multi-modal data comprising images, radar, and LiDAR point clouds. The following pre-processing steps are applied to each modality:
Image Pre-processing: Images undergo a random resize within a scaling range of $[0.386, 0.55]$ before being cropped to a fixed resolution of 256×704. Data augmentation includes random horizontal flipping and a constrained vertical crop with no bottom percentage limit. Rotation augmentation is disabled. Six camera views are used.
Radar Pre-processing: Radar points are projected into the BEV space, with an intensity-aware transformation applied to align them with the camera features. The radar representation is downsampled using a voxelization process with a fixed BEV grid resolution.
LiDAR Pre-processing: LiDAR points are voxelized with a voxel size of $[0.1, 0.1, 0.2]$, ensuring consistent spatial resolution. The voxel encoder uses a sparse convolutional network to generate a compact feature representation while maintaining high spatial fidelity.
BEV Augmentation: BEV-space transformations include a random rotation within $[-22.5°, 22.5°]$, a scaling perturbation in the range $[0.9, 1.1]$, and a probabilistic flipping along both axes with a 50% chance.

9.2. Hyperparameters Settings

Backbone (Image Branch): A ResNet-50 [8] extracts multi-scale image features, processed via an FPN-style [26] neck with an upsampling strategy of $\{0.25, 0.5, 1, 2\}$.
Backbone (Radar & LiDAR Fusion): Point cloud features are voxelized and encoded using a SECOND-based [62] architecture, followed by a stacked CNN backbone. The features are refined via a SECONDFPN-style neck with output strides of $\{0.5, 1, 2\}$.
Detection Head: The detection follows a CenterPoint-style [65] approach, leveraging a hierarchical BEV backbone and an FPN-style [26] neck. Bounding boxes are regressed using a CenterPoint-based [65] box coder with a post-center range of $[-61.2, 61.2]$.

9.3. Training Configuration

Loss Functions: Apart from the losses mentioned in the paper, the classification loss is based on Gaussian Focal Loss [53], while regression losses include L1 Loss [7] for bounding box estimation and a smooth transition function for ori-

entation prediction. Additional loss terms are incorporated to enhance knowledge-distillation and overall detection performance.

Voxelization: The LiDAR point cloud is voxelized within a spatial range of $[-51.2, 51.2]$ meters in the XY plane and a vertical range from -5 to 3 meters.

Training Grid Settings: The BEV grid is constructed with a spatial resolution of $[512, 512]$ and an output downsampling factor of 4. For LiDAR, the grid is defined over $[1024, 1024, 40]$ points, maintaining high spatial fidelity.

Config	ResNet-50/101
Optimizer	AdamW
Base Learning Rate	$4e-4$
Backbone Learning Rate	$2e-4/1e-4$
Weight Decay	$1e-2$
Batch Size	16 / 8
Training Epochs	30
LR Schedule	Cosine
Gradient Clip	5

Table 11. Training configurations for ResNet-50/101.

10. Additional Experimental Results

10.1. Comparison with LiDAR Teacher Model

To evaluate the effectiveness of IMKD, we compare it against its LiDAR-based teacher, specifically CenterPoint [65] pretrained on the nuScenes [1] dataset. The student model consists of a BEVDepth [23] camera module and a radar encoder.

Tab. 12 summarizes the results. While the LiDAR teacher achieves strong performance, it is not the best-performing LiDAR model on the nuScenes [1] dataset. We report IMKD results with and without distillation. Although direct comparison across modalities is inherently challenging, distillation significantly improves the student, with NDS increasing by 1.8 and mAP by 2.6 compared to the teacher. This improvement arises because our multi-level distillation transfers depth cues, geometric structure, and point-density patterns from LiDAR into the fused camera–radar representation, thereby compensating for the modalities' inherent weaknesses. In addition, the prediction-level distillation between LiDAR outputs and the student predictions refines decision boundaries and reduces ambiguity in challenging cases. Together, these mechanisms allow the student to not only close the gap with the LiDAR teacher but in some settings surpass it by leveraging complementary cross-modal information absent in LiDAR alone.

Method	mAP	NDS
LiDAR Teacher [65]	58.40	65.20
IMKD w/o LiDAR Distil.	56.90	62.5
IMKD Full	**61.0**	**67.0**

Table 12. Performance comparison between our IMKD model and its LiDAR teacher on the nuScenes [1] test set.

10.2. Comparison with Camera-Radar Methods without Knowledge Distillation

To further contextualize the performance of our IMKD framework, we compare it against recent camera-radar fusion methods that do not use knowledge distillation. As shown in Table 13, we benchmark IMKD against several strong baselines including CRN [16], RCBEVDet [28], and CRT-Fusion [13], all evaluated on the nuScenes [1] validation set.

To ensure a fair and meaningful comparison, we primarily benchmark IMKD against radar-camera fusion methods that share the same foundational settings. Specifically, we focus on approaches that adopt BEVDepth [23] with a ResNet-50 [8] backbone, avoiding discrepancies introduced by stronger visual encoders. We also exclude methods that leverage CBGS [76], test-time augmentation, or future frames, as such enhancements can distort the true impact of the fusion strategy. All comparisons are conducted on the nuScenes [1] validation set, where the backbone architecture and image resolution are consistent across methods, unlike the test set, where configurations often vary. IMKD is the first distillation-driven framework to surpass the performance of standard radar-camera fusion methods, elevating knowledge distillation from a regularization tool to a core mechanism for advancing 3D detection performance.

As an exception, we additionally report results for RICCARDO [34], which employs SparseBEV [30] with a ResNet-101 [8] backbone rather than BEVDepth [23] with ResNet-50 [8]. While this setting is not strictly comparable to our fairness-controlled benchmark, it provides useful context on how IMKD scales with stronger visual encoders. To avoid misleading comparisons, we align RICCARDO's [34] results with our own ResNet-101 [8] BEVDepth [23] variant, and present these separately in Tab. 13 under a distinct block. This highlights that IMKD maintains its advantage even when evaluated under higher-capacity camera backbones, demonstrating robustness across configurations.

These improvements stem from IMKD's fusion-aware and signal-sensitive design. By incorporating intensity-aware distillation and fusion-based supervision, IMKD captures fine-grained signal reliability and cross-modal interactions that traditional fusion models overlook. As a result, IMKD not only bridges the gap between handcrafted fusion

Method	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet [11]	C	ResNet-50	256 × 704	39.2	31.2	0.691	0.272	0.523	0.909	0.247
BEVDepth [23]	C	ResNet-50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198
RC-BEVFusion [50]	C+R	ResNet-50	256 × 704	52.5	43.4	0.511	0.270	0.527	0.421	0.182
SOLOFusion [42]	C	ResNet-50	256 × 704	53.4	42.7	0.567	0.274	0.411	0.252	0.188
StreamPETR [55]	C	ResNet-50	256 × 704	54.0	43.2	0.581	0.272	0.413	0.295	0.195
SparseBEV [30]	C	ResNet-50	256 × 704	54.5	43.2	0.606	0.274	0.387	0.251	0.186
CRN [16]	C+R	ResNet-50	256 × 704	56.0	49.0	0.487	0.277	0.542	0.344	0.197
RCBEVDet [28]	C+R	ResNet-50	256 × 704	56.8	45.3	0.486	0.285	0.404	**0.220**	0.192
CRT-Fusion [13]	C+R	ResNet-50	256 × 704	57.2	50.0	0.499	0.277	0.531	0.261	0.192
IMKD (Ours)	C+R	ResNet-50	256 × 704	**61.0**	**51.6**	**0.444**	**0.259**	**0.384**	0.229	**0.160**
RICCARDO [34]	C+R	ResNet101	1408 × 512	62.2	**54.4**	0.481	0.266	**0.325**	0.237	0.189
IMKD (Ours)	C+R	ResNet101	1408 × 512	**62.7**	53.9	**0.417**	**0.255**	0.348	**0.235**	**0.158**

Table 13. Comparison of 3D object detection performance on the nuScenes [1] validation set. 'C' and 'R' denote camera and radar, respectively. Methods utilizing future frames, test-time augmentation, and CBGS [76] are excluded to ensure fairness. The upper block reports comparisons restricted to BEVDepth with ResNet-50, while the lower block extends to ResNet-101 backbones and includes RICCARDO [34] for completeness.

and learned fusion but also pushes the performance frontier for camera-radar 3D object detection.

We further report results on the nuScenes [1] test set to contextualize IMKD against the latest benchmark entries, as shown in Tab. 14. While this comparison is not strictly fair, methods employ heterogeneous camera backbones (e.g., SparseBEV [30] in RICCARDO [34]) and varying image resolutions, it provides a broader view of IMKD's standing. Despite these differences, IMKD achieves performance highly competitive with state-of-the-art methods, while remaining the only knowledge-distillation-based approach among the top-performing entries on the benchmark. This highlights both the practicality and the effectiveness of IMKD in advancing camera–radar 3D detection under challenging real-world settings.

10.3. Comparison on VoD Dataset

To evaluate the generalization of IMKD beyond the nuScenes [1] dataset, we conduct experiments on the View-of-Delft (VoD) [41] dataset, which provides synchronized LiDAR, camera, and 3+1D radar sensors, with the radar capturing elevation in addition to range, azimuth, and Doppler. This richer radar representation presents a more challenging detection scenario compared to the sparse 2+1D radar in nuScenes [1].

As reported in Tab. 15, IMKD achieves strong performance across all categories, demonstrating competitive results relative to existing camera-radar methods. In particular, IMKD maintains high AP in both the entire annotated area and the region of interest, indicating that the intensity-guided distillation framework effectively transfers LiDAR

Method	Input	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PGD [56]	C	44.8	38.6	0.626	0.245	0.451	1.509	0.127
SparseBEV [30]	C	67.5	60.3	0.425	0.239	0.311	0.172	0.116
MVFusion [61]	C+R	51.7	45.3	0.569	0.246	0.379	0.781	0.128
CRN [16]	C+R	62.4	57.5	0.416	0.264	0.456	0.365	0.130
RCBEVDet [28]	C+R	63.9	55.0	0.390	**0.234**	0.362	0.259	0.113
HyDRa [60]	C+R	64.2	57.4	0.398	0.251	0.423	0.249	0.122
HVDetFusion [19]	C+R	67.4	60.9	0.379	0.243	0.382	0.172	0.132
SparseBEV+RICCARDO [34]	C+R	**69.5**	**63.0**	**0.363**	0.240	0.311	**0.167**	0.118
IMKD (Ours)	C+R	67.0	61.0	0.401	0.249	**0.305**	0.238	**0.102**

Table 14. Comparison of 3D object detection performance on the nuScenes [1] test set. 'C' and 'R' represent camera and radar, respectively.

Method	Input	AP in Entire Annotated Area (%)				AP in Region of Interest (%)			
		Car	Pedestrian	Cyclist	mAP	Car	Pedestrian	Cyclist	mAP
PointPillars [18]	R	37.06	35.04	63.44	45.18	70.15	47.22	85.07	67.48
RadarPillarNet [63]	R	39.30	35.10	63.63	46.01	71.65	42.80	83.14	65.86
RCFusion [72]	C+R	41.70	38.95	68.31	49.65	71.87	47.50	88.33	69.23
RCBEVDet [28]	C+R	40.63	38.86	**70.48**	49.99	72.48	49.89	87.01	69.80
IMKD (Ours)	C+R	**47.55**	**45.51**	68.40	**53.81**	**89.13**	**57.10**	**89.56**	**78.59**

Table 15. Comparison of 3D object detection results on the VoD [41] validation set. The region of interest is the driving corridor near the ego-vehicle. AP thresholds are set to 0.5 for cars, 0.25 for pedestrians, and 0.25 for cyclists.

knowledge and enhances fused representations even under different radar characteristics.

These results validate that our method generalizes robustly to other datasets and radar configurations, confirming that intensity-aware multi-level knowledge distillation can consistently improve cross-modal 3D detection beyond the original nuScenes [1] setting.

10.4. BEV Segmentation

Our method leverages knowledge distillation from LiDAR and label guidance to enhance camera-radar features, enabling precise segmentation of road elements such as drivable areas, lanes, and crossings. LiDAR distillation refines spatial accuracy, improving object boundaries and structural details. We use mean Intersection over Union (mIoU) as the primary metric, following [43]. As shown in Tab. 16, our approach achieves an mIoU of 62.2, demonstrating effective segmentation with real-time performance.

Method	Input	Backbone	mIoU↑	Veh↑	D.A↑
BEVFormer-S [24]	C	R101	48.4	43.2	80.7
CRN [16]	C+R	R50	-	58.8	**82.1**
Simple-BEV++† [46]	C+R	R101	55.4	52.7	77.7
BEVGuide [37]	C+R	EffNet	60.0	59.2	76.7
BEVCar [46]	C+R	R101	61.0	57.3	81.8
IMKD (Ours)	C+R	R101	**62.2**	**60.5**	81.9

Table 16. Comparison of BEV semantic segmentation on the nuScenes [1] validation set. 'C' and 'R' represent camera and radar, respectively. 'D.A' denotes drivable area. † indicates a Simple-BEV [6] model customized by BEVCar [46].

10.5. Per-Class Performance Analysis

In Tab. 17, we compare per-class performance across different camera-radar fusion methods, using a fixed resolution of 256×704 and the ResNet-50 backbone for consistency.

In Tab. 18, we compare each camera-only network with its camera+radar variant on the nuScenes [1] validation set. The results show that radar significantly improves performance in most classes. Using the same camera-only baseline as CRN, our method outperforms previous approaches in several categories.

Our IMKD method consistently achieves the highest mAP, with notable improvements in Truck, Bus, C.V., Pedestrian, and Bicycle. This demonstrates the effectiveness of our fusion strategy in handling various object types, particularly for smaller or more dynamic objects where radar data can be especially beneficial. The improvements in classes like Pedestrian and Bicycle, where radar information is typically sparse, further validate the robustness of our approach.

Key to this performance is our knowledge distillation framework, which refines the fusion of camera and radar features through LiDAR-guided and label-based distillation, ensuring that radar signals contribute meaningfully to object detection rather than introducing noise. This structured supervision enhances detection accuracy, leading to more reliable and consistent object localization across all categories.

Overall, our results show that distilling knowledge into the fused modality improves camera-radar fusion, significantly boosting performance.

Method	Input	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	Bicycle	T.C.	Barrier	mAP
CenterFusion [40]	C+R	52.4	26.5	36.2	15.4	5.5	38.9	30.5	22.9	56.3	47.0	33.2
CRAFT [15]	C+R	69.6	37.6	47.3	20.1	10.7	46.2	39.5	31.0	57.1	51.1	41.1
CRN [16]	C+R	71.9	42.4	51.1	27.1	16.2	46.6	54.0	44.2	56.7	61.6	47.1
IMKD (Ours)	C+R	**75.3**$^{4.7\%}$	**50.9**$^{20.0\%}$	**55.6**$^{8.8\%}$	**28.6**$^{5.5\%}$	**20.6**$^{27.2\%}$	**55.1**$^{18.2\%}$	**54.5**$^{0.9\%}$	**51.1**$^{15.6\%}$	**62.2**$^{9.7\%}$	**62.1**$^{0.8\%}$	**51.6**$^{9.6\%}$

Table 17. Per-class comparisons on the nuScenes [1] validation set. 'C.V.', 'Ped.', 'M.C.', and 'T.C.' denote construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. All results are sourced from MMDetection3D and official implementations, except CRN, which was reproduced using its official GitHub repository.

Method	Input	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	Bicycle	T.C.	Barrier	mAP
CenterNet [75]	C	48.4	23.1	34.0	13.1	3.5	37.7	24.9	23.4	55.0	45.6	30.6
CenterFusion [40]	C+R	52.4$^{8.3\%}$	26.5$^{14.7\%}$	36.2$^{6.5\%}$	15.4$^{17.5\%}$	5.5$^{57.1\%}$	38.9$^{3.2\%}$	30.5$^{22.5\%}$	22.9$^{-1.4\%}$	56.3$^{2.4\%}$	47.0$^{3.0\%}$	33.2$^{0.6\%}$
CRAFT-I [15]	C	52.4	25.7	30.0	15.8	5.4	39.3	28.6	29.8	57.5	47.8	33.2
CRAFT [15]	C+R	69.6$^{32.8\%}$	37.6$^{46.3\%}$	47.3$^{57.6\%}$	20.1$^{27.2\%}$	10.7$^{98.1\%}$	46.2$^{17.5\%}$	39.5$^{38.1\%}$	31.0$^{4.0\%}$	57.1$^{-0.7\%}$	51.1$^{7.0\%}$	41.1$^{23.8\%}$
BEVDepth [23]	C	55.3	25.2	37.8	16.3	7.6	36.1	31.9	28.6	53.6	55.9	34.8
CRN [16]	C+R	71.9$^{30.0\%}$	42.4$^{67.9\%}$	51.1$^{35.2\%}$	27.1$^{66.9\%}$	16.2$^{113.2\%}$	46.6$^{29.1\%}$	54.0$^{69.2\%}$	44.2$^{54.2\%}$	56.7$^{5.8\%}$	61.6$^{10.2\%}$	47.1$^{35.6\%}$
BEVDepth [23]	C	55.3	25.2	37.8	16.3	7.6	36.1	31.9	28.6	53.6	55.9	34.8
IMKD (Ours)	C+R	75.3$^{36.6\%}$	50.9$^{101.2\%}$	55.6$^{57.3\%}$	28.6$^{75.2\%}$	20.6$^{171.1\%}$	55.1$^{52.4\%}$	54.5$^{71.8\%}$	51.1$^{78.7\%}$	62.2$^{10.5\%}$	62.1$^{9.6\%}$	51.6$^{47.6\%}$

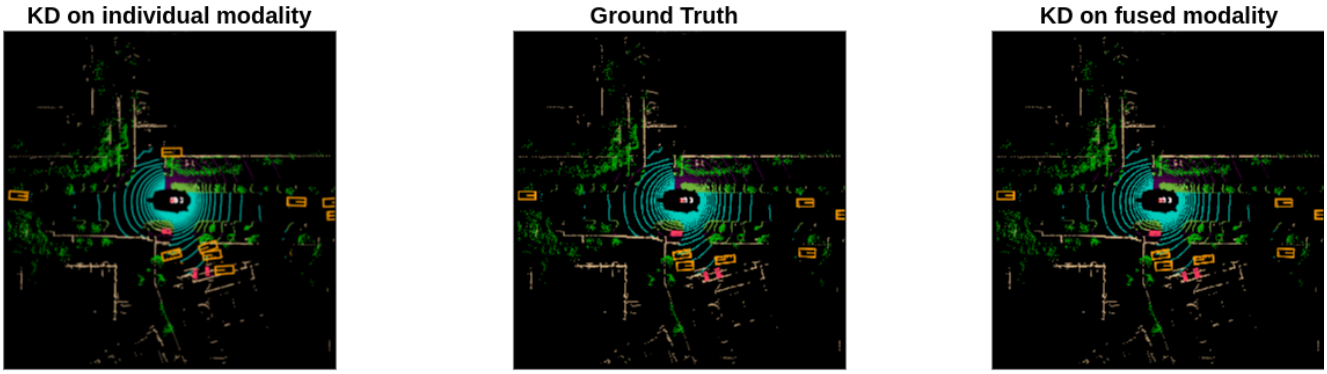
Table 18. Per-class comparisons on the nuScenes [1] validation set, evaluating each camera + radar network against its corresponding camera-only variant. 'C.V.', 'Ped.', 'M.C.', and 'T.C.' denote construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. All results are sourced from MMDetection3D and official implementations, except CRN, which was reproduced using its official GitHub repository.

11. Qualitative Analysis

We present additional qualitative results under varying weather and lighting conditions, including rainy, nighttime, and daytime scenarios, from the nuScenes [1] dataset. As shown in Figs. 6 to 8, IMKD consistently performs better than individual modality distillation baselines, particularly under challenging scenarios like rain and low light.

In these adverse conditions, conventional single-modality distillation models often fail to detect occluded or distant objects. In contrast, IMKD consistently performs better by utilizing intensity-guided fusion and merged-modality knowledge distillation. The fusion mechanism dynamically weighs radar and camera features based on signal confidence, while the distillation strategy transfers depth and structural cues from LiDAR into the joint camera-radar representation. This enables IMKD to produce more accurate and robust object detections, boxes with better translation, orientation, and scale accuracy than baselines, crucial under low visibility where conventional methods struggle to infer reliable geometry. These improvements are clearly reflected in both BEV and multi-view camera predictions.

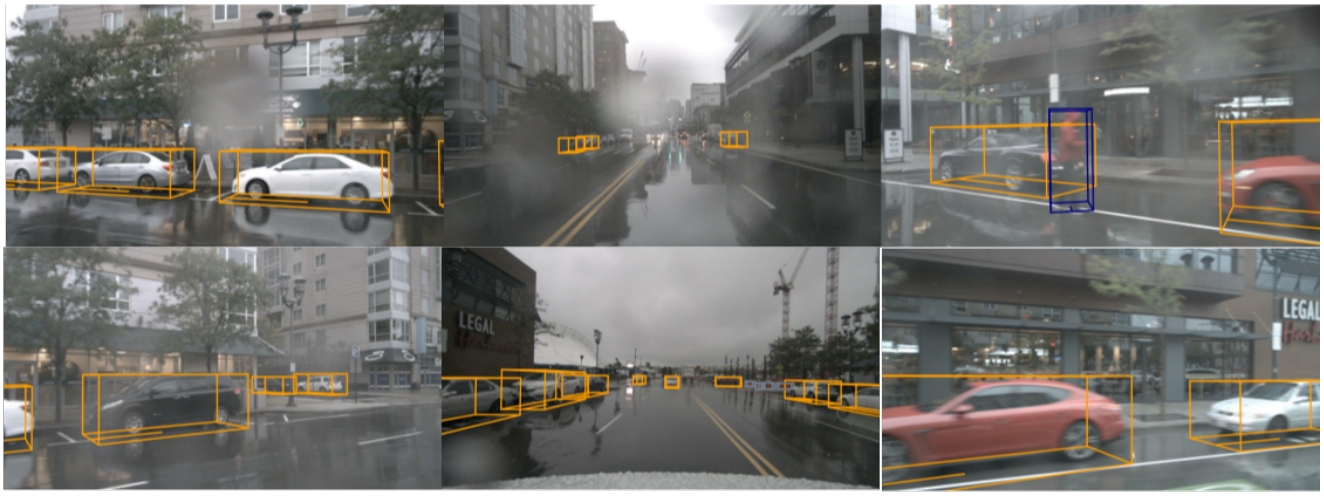
(a) Indivial Modality KD Predictions



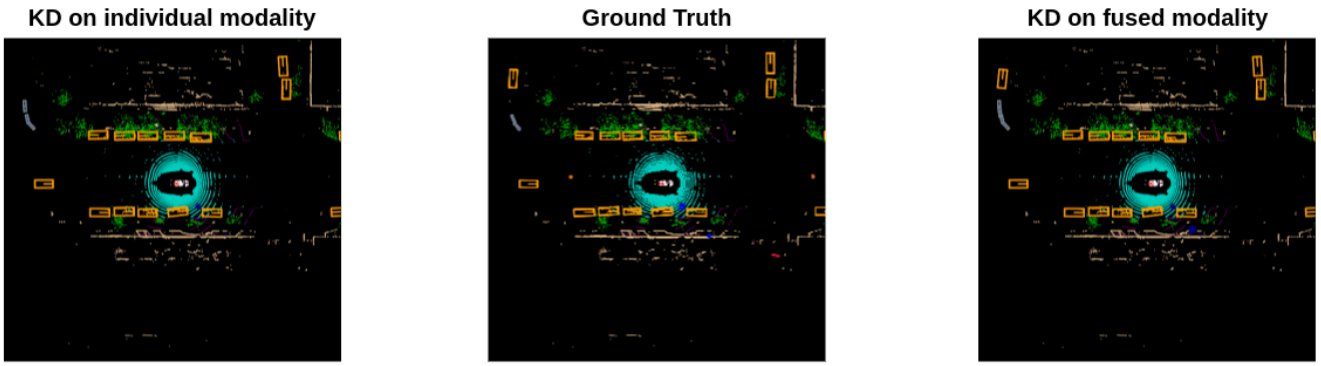
(b) BEV Predictions and Ground Truth

(c) Fused Modality KD Predictions

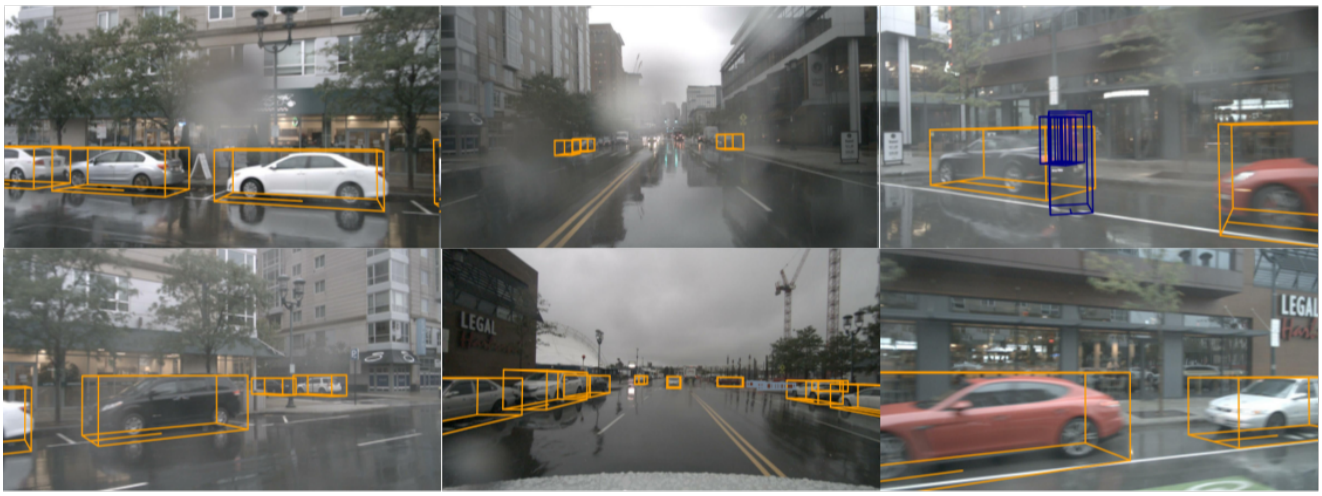
Figure 6. Qualitative results of our proposed IMKD method on night scenes from the nuScenes [1] dataset. (a) shows camera-view predictions from individual modality distillation baselines. (b) presents BEV predictions: left shows individual modality predictions, middle is the ground truth, and right shows IMKD results. (c) displays IMKD's predictions across six camera views, illustrating improved detection quality under challenging low-light conditions.



(a) Indivial Modality KD Predictions



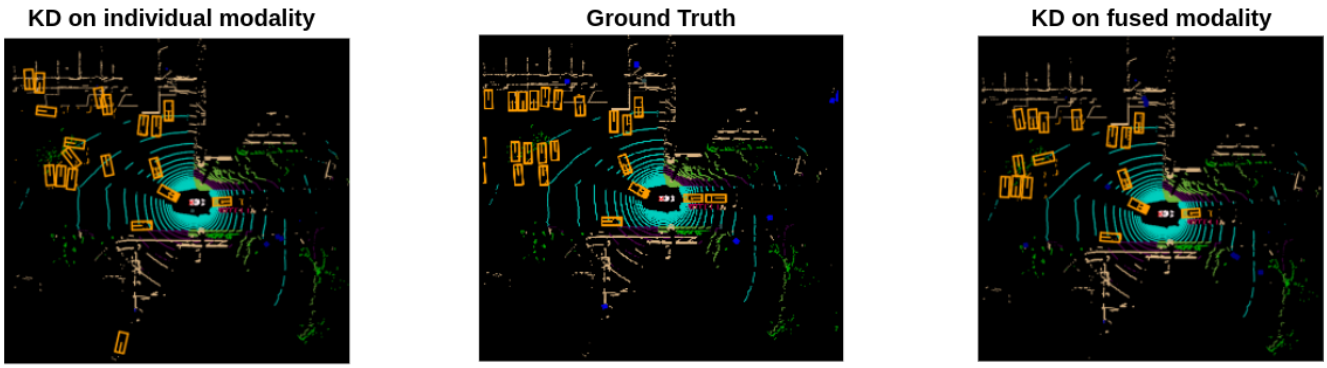
(b) BEV Predictions and Ground Truth



(c) Fused Modality KD Predictions

Figure 7. Qualitative results of our proposed IMKD method on rainy scenes from the nuScenes [1] dataset. (a) shows camera-view predictions from individual modality distillation baselines. (b) presents BEV predictions: left shows individual modality predictions, middle is the ground truth, and right shows IMKD results. (c) displays IMKD's predictions across six camera views, illustrating improved detection quality under challenging low-light conditions.

(a) Indivial Modality KD Predictions



(b) BEV Predictions and Ground Truth

(c) Fused Modality KD Predictions

Figure 8. Qualitative results of our proposed IMKD method on day scenes from the nuScenes [1] dataset. (a) shows camera-view predictions from individual modality distillation baselines. (b) presents BEV predictions: left shows individual modality predictions, middle is the ground truth, and right shows IMKD results. (c) displays IMKD's predictions across six camera views, illustrating improved detection quality under challenging low-light conditions.