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

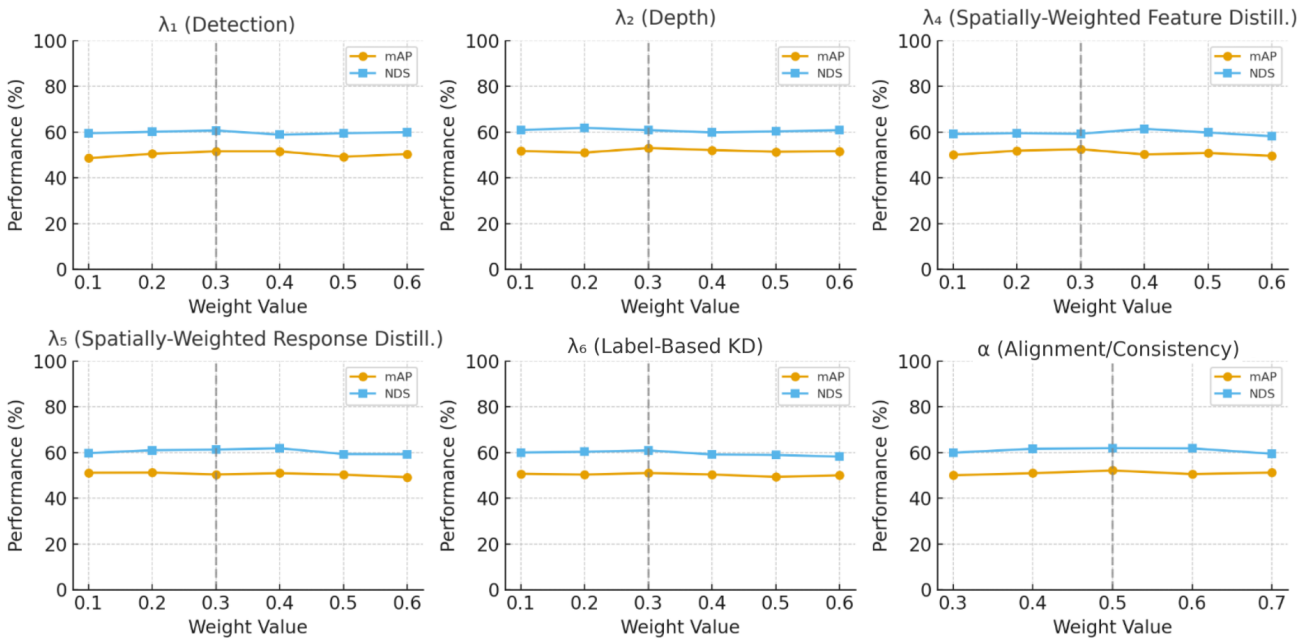


Figure 5. Sensitivity of mAP and NDS to individual loss weights λ. Each subplot reports an illustrative sweep over $\lambda \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$; dashed vertical lines mark the chosen operating points ($\lambda = 0.3$ for most terms, $\alpha = 0.5$ for alignment). The curves indicate that 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$^{47.6\%}$

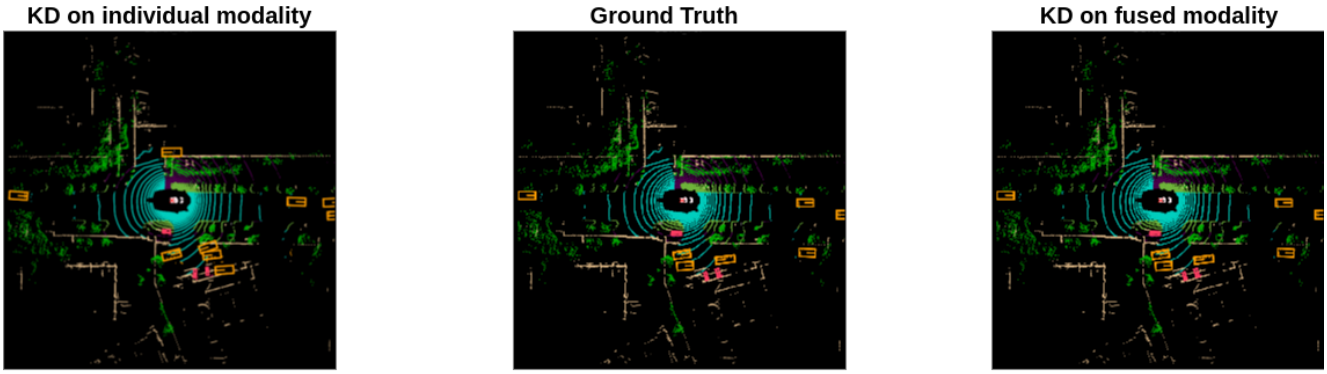
Table 18. Per-class comparisons on the nuScenes [1] validation set, evaluating each camera + radar network against its corresponding camera-only variant. 'C.V.', 'Ped.', 'M.C.', and 'T.C.' denote construction vehicle, pedestrian, motorcycle, and traffic cone, respectively. All results are sourced from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

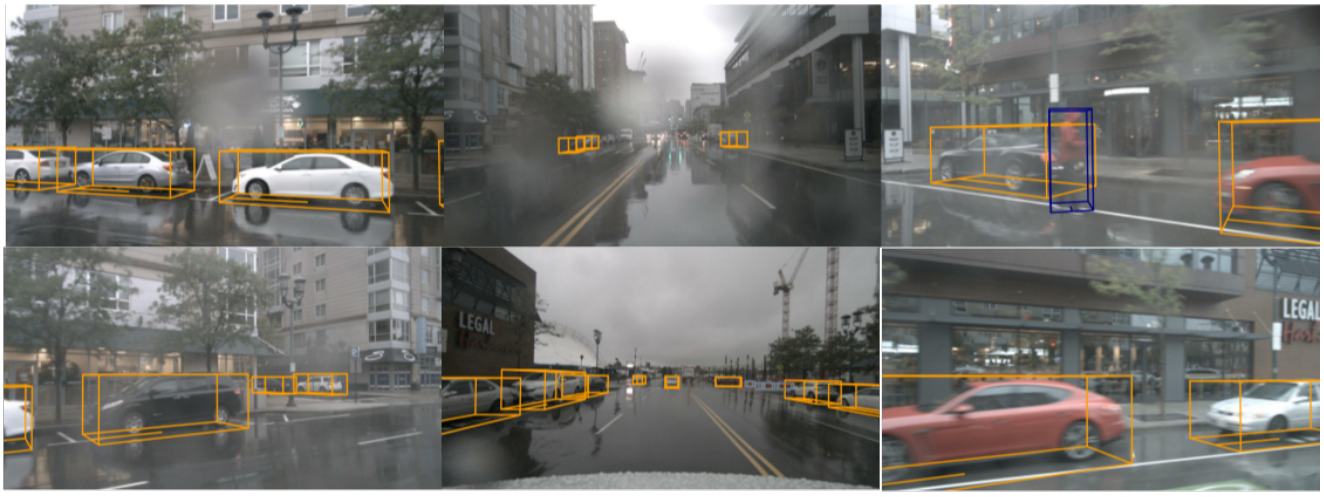
(a) Indivial Modality KD Predictions



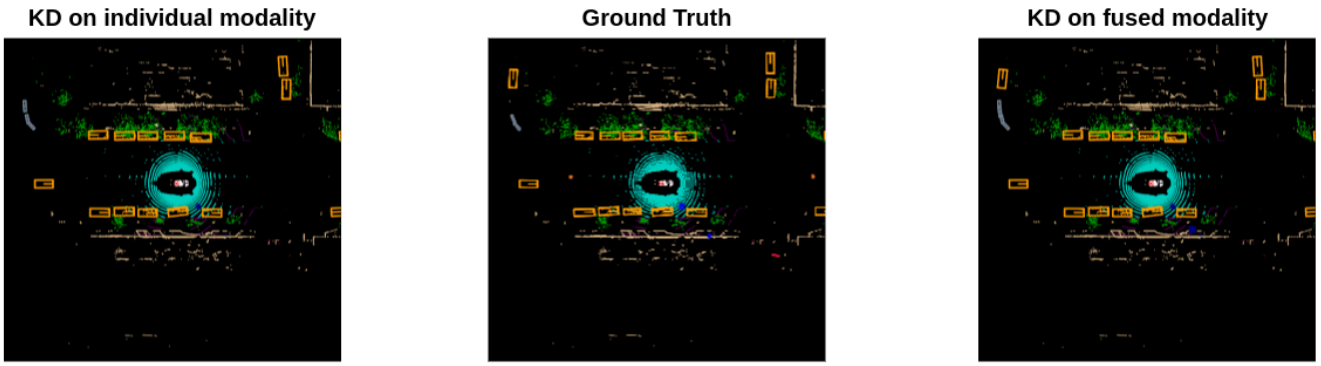
(b) BEV Predictions and Ground Truth

(c) Fused Modality KD Predictions

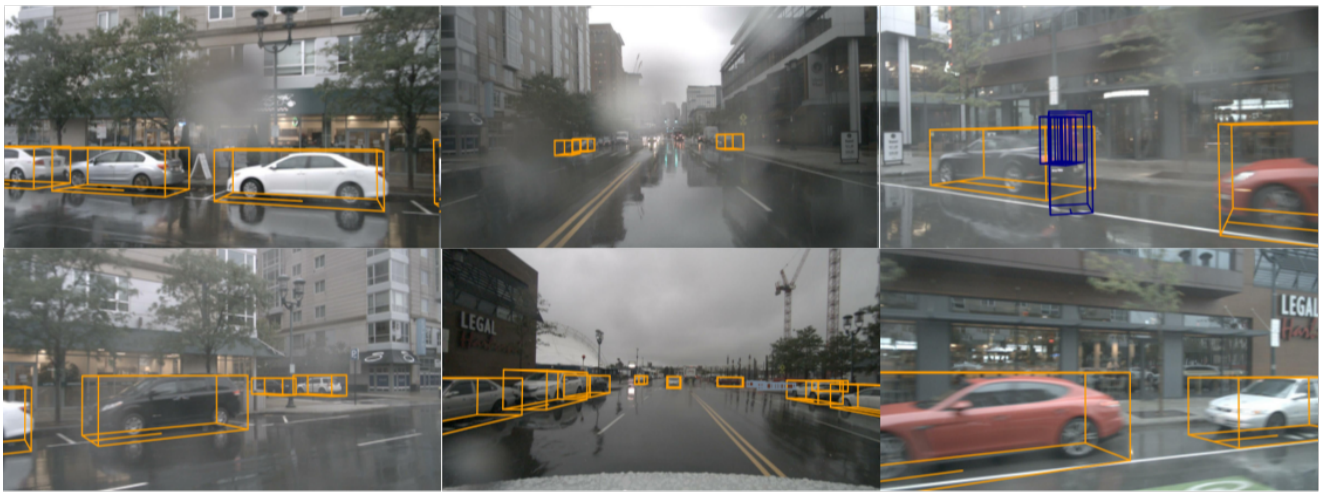
Figure 6. Qualitative results of our proposed IMKD method on night scenes from the nuScenes [1] dataset. (a) shows camera-view predictions from individual modality distillation baselines. (b) presents BEV predictions: left shows individual modality predictions, middle is the ground truth, and right shows 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



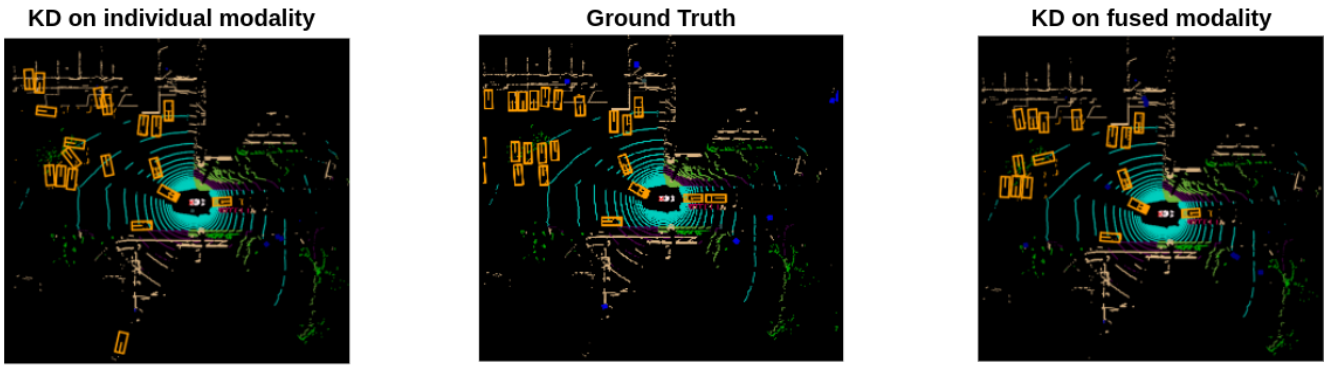
(b) BEV Predictions and Ground Truth



(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.