

Appendix. The appendix is organized as follows. Section A provides detailed explanations of the sensor fusion framework, including specific implementations of sensor-specific encoders and the detection head. Section B presents additional t-SNE visualizations that demonstrate the effectiveness of the unified canonical space. Section C analyzes attention score distributions across different weather conditions and distance ranges. Section D includes comprehensive performance results across various sensor combinations and alternative stereo camera configurations.

A Details of Sensor Fusion Framework

As shown in Fig. 4, the sensor fusion framework is divided into three stages: (1) sensor-specific encoders (Liu et al., 2021; Phillion and Fidler, 2020; Yan et al., 2018; Shi et al., 2020; Paek et al., 2022; Kong et al., 2024) that extract feature maps (FMs) from different types of multi-modal sensors (e.g., camera, LiDAR, and 4D Radar), (2) a fusion method that integrates these FMs, and (3) a task-dependent detection head. Stage (1) has evolved independently for each sensor (Dosovitskiy et al., 2021; Lang et al., 2019; Paek et al., 2022), while stage (3) employs a task-dependent method (e.g., YOLO head (Redmon et al., 2016) for object detection) regardless of sensor type. Therefore, in this work, we propose a fusion method focusing on stage (2) rather than stages (1) and (3). Details of both existing and proposed fusion methods are illustrated in Fig. 1. In this section, we describe both the sensor-specific encoders and the detection head of the proposed availability-aware sensor fusion (ASF).

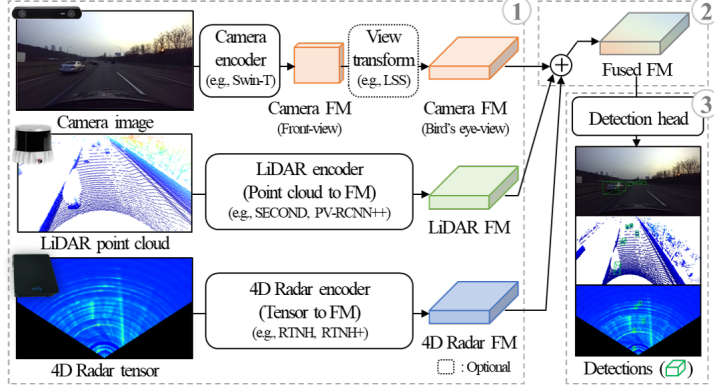


Figure 4: Overall sensor fusion framework for camera, LiDAR, and 4D Radar. ‘FM’ denotes ‘feature map’.

A.1 Sensor-specific Encoders

Sensor-specific encoders extract FMs from data of each sensor. Because different sensors provide data in different modalities—cameras produce 2D RGB images, LiDAR generates 3D point clouds, and 4D Radar outputs low-resolution tensors with power values—research on these encoders has evolved independently for each sensor type.

For ASF, we adopt established backbone architectures for each sensor and pre-train these backbones before ASF training. Each encoder is trained for 11 epochs following the K-Radar (Paek et al., 2022) training protocol. During ASF training, we freeze all sensor-specific encoders to ensure that only the ASF components and detection head are learned. Specifically, ‘UCP’ and ‘CASAP’ are the ASF components updated during ASF training, which denote *unified canonical projection* and *cross-attention across sensors along patches*, respectively. Details for each encoder are as follows:

Camera. Camera-based sensor encoders for autonomous driving typically require transforming front-view (FV) images into a bird’s-eye-view (BEV) format for 3D object detection. Current approaches can be broadly categorized into two methods: (1) learnable BEV transformations such as Lift-Splat-Shoot (LSS) (Phillion and Fidler, 2020; Liu et al., 2023; Liang et al., 2024), or (2) maintaining FV while using transformer decoders (Liu et al., 2022) that incorporate depth information as positional embeddings in BEV space. Although methods like PETR (Liu et al., 2022) have improved 3D object detection by injecting depth embeddings, they inevitably increase computational complexity due to the additional depth dimension. These two approaches form the foundation of camera-based fusion (Xiong et al., 2023; Zheng et al., 2023; Liu et al., 2023; Liang et al., 2024; Yan et al., 2023), commonly seen in deeply coupled fusion (DCF) and sensor-wise cross-attention fusion (SCF).

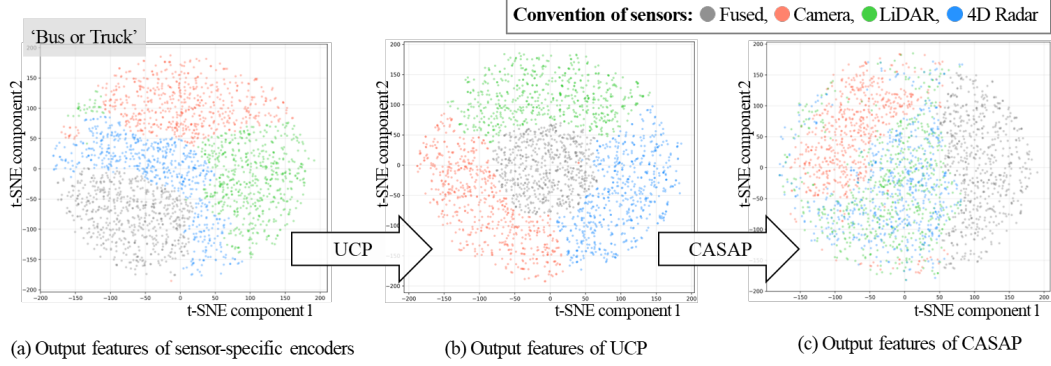


Figure 5: Visualization of feature representations through t-SNE (van der Maaten and Hinton, 2008) at different stages of the proposed ASF for the ‘Bus or Truck’ class. ‘UCP’ and ‘CASAP’ denote *unified canonical projection* and *cross-attention across sensors along patches*, which are the two main components of the proposed ASF.

To develop a computationally efficient method, we opt for the LSS approach and adopt BEVDepth (Li et al., 2023), which supervises the LSS depth estimation using LiDAR point clouds. For encoding images, we use the Swin-S model of SwinTransformer as the 2D backbone, achieving 83% top-1 accuracy on ImageNet (Russakovsky et al., 2015). We resize front camera images to 704×256 pixels for encoding. The encoded 2D FMs are projected into BEV FMs via LSS with a grid size of 0.2 m. To match the 0.4 m grid size used in K-Radar, we add a convolutional layer with stride 2. During training, we employ the depth information projected from LiDAR point clouds onto the image and, inspired by (Yang et al., 2023), we use the LiDAR backbone’s (SECOND) BEV FMs for distillation. As shown in Fig. 3, camera images have limited utility in adverse weather, so we train the camera encoder only on normal or overcast scenes (sequences 1–20) from K-Radar.

A limitation of this approach is that promising stereo or multi-view detection models cannot be fully leveraged, because the K-Radar dataset (Paek et al., 2022)—the only one offering camera, LiDAR, and 4D Radar data in adverse weather—provides object annotations only for the front view. Moreover, unlike the KITTI dataset with its longer stereo baseline of 0.54 m, K-Radar uses waterproof ZED cameras with a baseline of just 0.12 m, limiting stereo-based detection for distant objects. Consequently, we employ a monocular camera method despite inherent performance constraints.

LiDAR. LiDAR provides precise depth information in the form of dense point clouds, often used as a reference for labeling other sensors (Paek et al., 2022; Sun et al., 2024). We adopt the SECOND (Yan et al., 2018) model as the LiDAR encoder, which is open-source in K-Radar and widely used in various benchmarks. Leveraging LiDAR’s high spatial resolution, we set the voxel size to 0.05 m and apply multiple layers of sparse convolution to generate BEV FMs with a 0.4 m grid size.

4D Radar. Unlike LiDAR, which uses Time-of-Flight (ToF) measurement, Radar detects spatial information through signal processing, which can introduce sidelobe noise. Traditional 3D Radar typically filters out noisy measurements to produce a sparse point cloud focusing on object presence (Paek and Guan, 2024). In contrast, 4D Radar captures additional elevation information, yielding higher-density point clouds (Zhang et al., 2023). This property is exemplified by K-Radar (Paek et al., 2022, 2023). We adopt RTNH (Paek et al., 2022), the baseline provided in K-Radar, as our 4D Radar encoder, following the original network settings and design.

A.2 Detection Head

Following the K-Radar baseline, we use an SSD detection head (Liu et al., 2016; Simon et al., 2018) for the 3D object detection task. Our implementation uses two anchor boxes per object class, oriented at 0° and 90° . Each anchor box is parametrized by eight values: the 3D center coordinates (x_c, y_c, z_c), 3D size (x_l, y_l, z_l), and orientation ($\cos(\text{yaw}), \sin(\text{yaw})$), as in (Simon et al., 2018).

To address the severe class imbalance in object detection (negative samples far outnumber positives), we use the focal loss (Lin et al., 2017) for classification. For bounding box regression, we use smooth L1 loss to measure differences between predictions and ground truth, improving training stability. During inference, we select the class with the highest logit score for each proposal, then apply non-maximum suppression to remove redundant overlapping detections, producing the final predictions.

B t-SNE Visualizations

B.1 Additional t-SNE Visualizations

This section provides additional feature visualizations to complement those in Fig. 2, which only shows results for the ‘Sedan’ class. Fig. 5 presents t-SNE (van der Maaten and Hinton, 2008) visualizations of the ‘Bus or Truck’ class at different stages of the ASF. These visualizations further demonstrate how ASF unifies feature representations from different sensors.

Consistent with Fig. 2, we observe three main trends in the t-SNE visualizations. First, the initial features from the sensor-specific encoders are scattered without clear alignment across different sensors. Next, after UCP, sensor-specific features become better aligned with the fused features. Finally, following CASAP, features from available sensors form cohesive clusters (in this case, LiDAR and 4D Radar).

These findings underscore the effectiveness of ASF in creating a unified canonical space for multi-sensor features. Note that in K-Radar, the ‘Sedan’ class has 37,213 instances, whereas the ‘Bus or Truck’ class has 7,449 instances, leading to sparser data in Fig. 5 compared to Fig. 2.

B.2 Details of t-SNE Visualizations

To visualize object-specific features in each sensor’s feature map (FM), we use RoIAlign (He et al., 2017) to crop features within bounding boxes. RoIAlign extracts fixed-size FMs, making it suitable for t-SNE (which requires consistent dimensions). We then visualize these cropped features in the unified canonical space. For those extracted before projection (where channel dimensions may differ), we downsample as needed to ensure consistent feature dimensions.

C Analysis of Attention Scores across Sensors

Table 5: Attention scores ratios [%] across different weather conditions

Sensor	Normal	Overcast	Fog	Rain	Sleet	Light Snow	Heavy Snow
Camera	11.6	11.1	11.3	11.1	11.1	10.9	11.0
LiDAR	39.5	37.5	37.5	37.3	35.9	36.3	34.7
4D Radar	48.9	51.4	51.2	51.6	53.0	52.8	54.3

Table 6: Attention scores ratios [%] across different distance ranges

Sensor	0-8m	8-16m	16-24m	24-32m	32-40m	40-48m	48-56m	56-64m	64-72m
Camera	9.8	10.1	10.4	10.9	11.1	11.3	11.9	12.7	13.4
LiDAR	48.4	40.2	37.2	36.0	34.9	34.8	35.0	35.5	36.0
4D Radar	41.8	49.7	52.4	53.1	54.0	53.9	53.1	51.8	50.6

Tables 5 and 6 show the attention score ratios for each sensor according to weather conditions and distance from the ego-vehicle. Intuitively, a higher attention score ratio indicates greater utilization of the corresponding sensor in ASF. Note that the sensor attention maps (SAMs) in Fig. 3 visualize the attention scores for each sensor.

As shown in Tab. 5, the attention score ratios of camera and LiDAR decrease under severe adverse weather conditions such as sleet and heavy snow (camera: 11.6% to 11.0%, LiDAR: 39.5% to 34.7%), while 4D Radar shows the opposite trend, increasing from 48.9% to 54.3%. This demonstrates

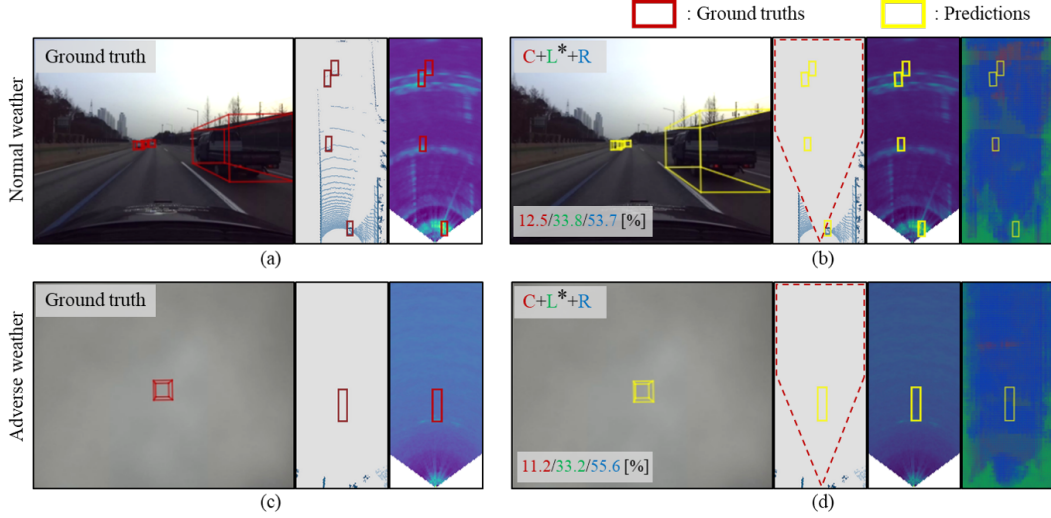


Figure 6: Qualitative results of ASF with C+L*+R (C: Camera, L*: damaged LiDAR, R: 4D Radar). We show results under (a–b) normal and (c–d) adverse weather conditions. Each grid shows the camera image, LiDAR point cloud, 4D Radar tensor, and a sensor attention map (SAM) illustrating attention scores from CASAP (except for (a) and (c), which only show sensor data for ground truth visualization). In the SAM, red, green, and blue represent attention to camera, LiDAR, and 4D Radar, respectively. For example, a predominantly blue SAM indicates that 4D Radar has the highest attention in that scene. The bottom-left corner of each grid shows attention score proportions (C/L/R[%]). Yellow and red boxes denote predictions and ground truth, respectively.

that 4D Radar is more robust than camera and LiDAR in adverse weather conditions and provides available measurements when other sensors degrade.

Furthermore, as shown in Tab. 6, LiDAR’s attention score ratio decreases as distance increases (from 48.4% at 0–8m to 36.0% at 64–72m), while camera’s attention score increases (from 9.8% to 13.4%). This indicates that ASF relies more on camera measurements at longer ranges where LiDAR becomes sparse. In contrast, 4D Radar maintains a relatively consistent attention ratio (41.8% to 50.6%) regardless of distance. Notably, 4D Radar shows the highest absolute attention scores compared to camera and LiDAR across most conditions. While camera lacks depth information and LiDAR provides very sparse measurements in 3D space, 4D Radar leverages dense tensor data, which enables more reliable feature extraction for sensor fusion.

D Additional Detection Results

This section presents additional detection results for ASF. First, we provide qualitative results demonstrating ASF’s robustness when the LiDAR sensor is damaged, shown in the same road environment as Fig. 3, along with additional results from diverse road scenarios. Second, we demonstrate the robustness of ASF (C+L+R) with respect to random seeds by presenting experimental results on the K-Radar (Paek et al., 2022) benchmark v1.0 using 10 different random seeds (0 to 8 and 2025), complementing Tab. 1 which reports results only for seed 2025. Third, we evaluate alternative camera backbone architectures, including stereo camera-based methods, to assess the impact of improved camera encoders on the overall fusion performance. Finally, we provide comprehensive quantitative performance on K-Radar benchmark v2.0, including AP_{3D} and AP_{BEV} for both ‘Sedan’ and ‘Bus or Truck’ classes at IoU=0.3 and 0.5, complementing Tab. 3 which reports only AP_{3D} at IoU=0.3.

D.1 Additional Qualitative Results

Damaged LiDAR. When LiDAR surfaces are damaged or contaminated, LiDAR measurements can be lost (Schlager et al., 2022). In adverse conditions such as heavy snow, accumulation on the

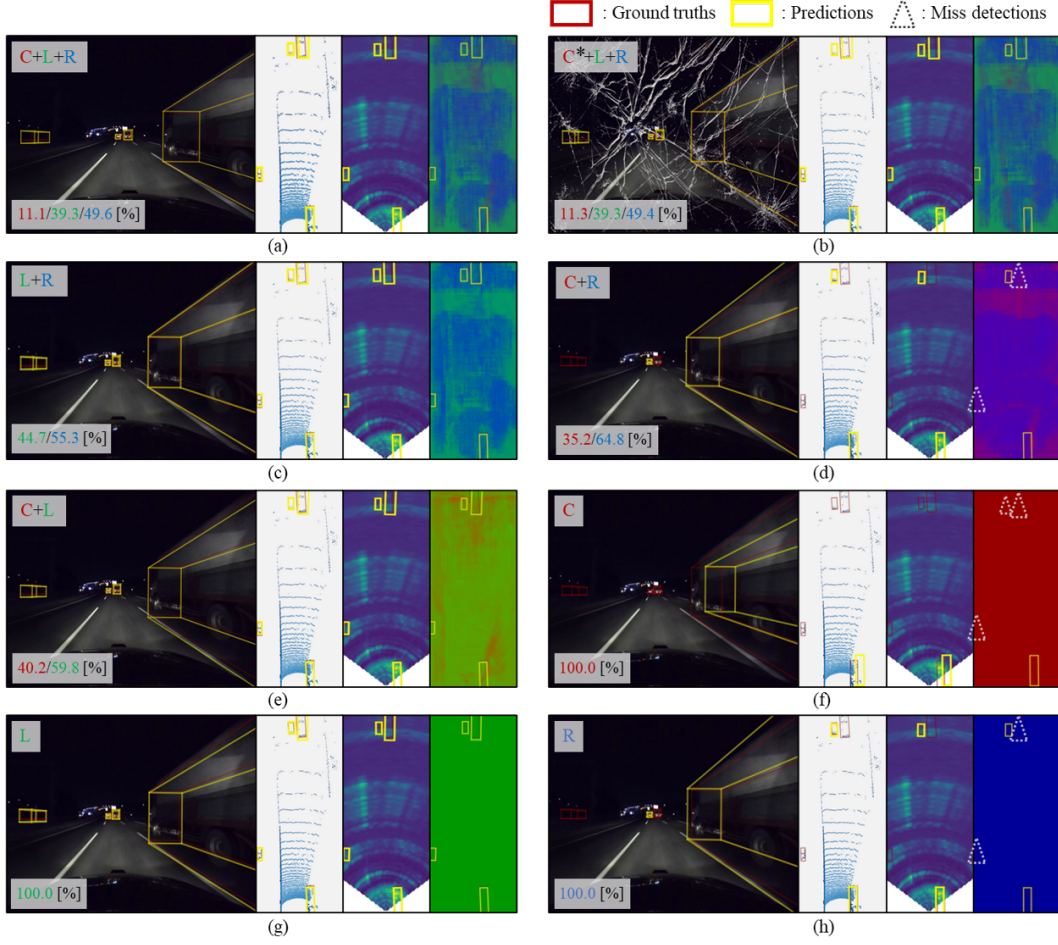


Figure 7: Qualitative results of ASF for various sensor combinations. We show results low-light nighttime roads. Each grid shows the camera image, LiDAR point cloud, 4D Radar tensor, and a SAM illustrating attention scores from CASAP. In the SAM, red, green, and blue represent attention to camera, LiDAR, and 4D Radar, respectively. The bottom-left corner of each grid shows attention score proportions (C/L/R[%]). Yellow and red boxes denote predictions and ground truth, respectively, which are displayed in all panels. Dotted triangles indicate miss detections and are marked in the SAM. Note that predictions are visualized on all sensor data, even when a sensor is not employed for detection (e.g., predictions from L+R are also visualized on the camera image).

LiDAR sensor can block forward measurements, as illustrated by the red dashed line in Fig. 6-(d), even though vehicles are present. We also simulate total loss of forward LiDAR returns from surface damage (e.g., broken sensor), as seen in Fig. 6-(b). Despite missing LiDAR data, ASF still detects all objects using only camera and 4D Radar inputs. As reported in Tab. 3, ASF with C+L*+R often performs similarly or even better than C+R alone. Compared to Fig. 3-(b), the sensor attention map (SAM) in Fig. 6-(b) confirms the increased reliance on camera or 4D Radar when LiDAR is unavailable.

Additional Results Visualization. Fig. 7 shows the inference results for multiple vehicles in low-light road environments. When detection is performed using camera-only in low-illumination conditions, numerous miss detections occur, as shown in Fig. 7-(f). Additionally, as shown in Fig. 7-(h), in the case of 4D Radar, one of the distant vehicles aligned side by side is covered by the reflection of another vehicle, resulting in miss detection. Furthermore, the leftmost vehicle experiences miss detection as the radio waves are blocked by the road barrier. In the case of Fig. 7-(a), which utilizes all of 4D Radar, LiDAR, and camera, stable and accurate detection is possible without miss detections.

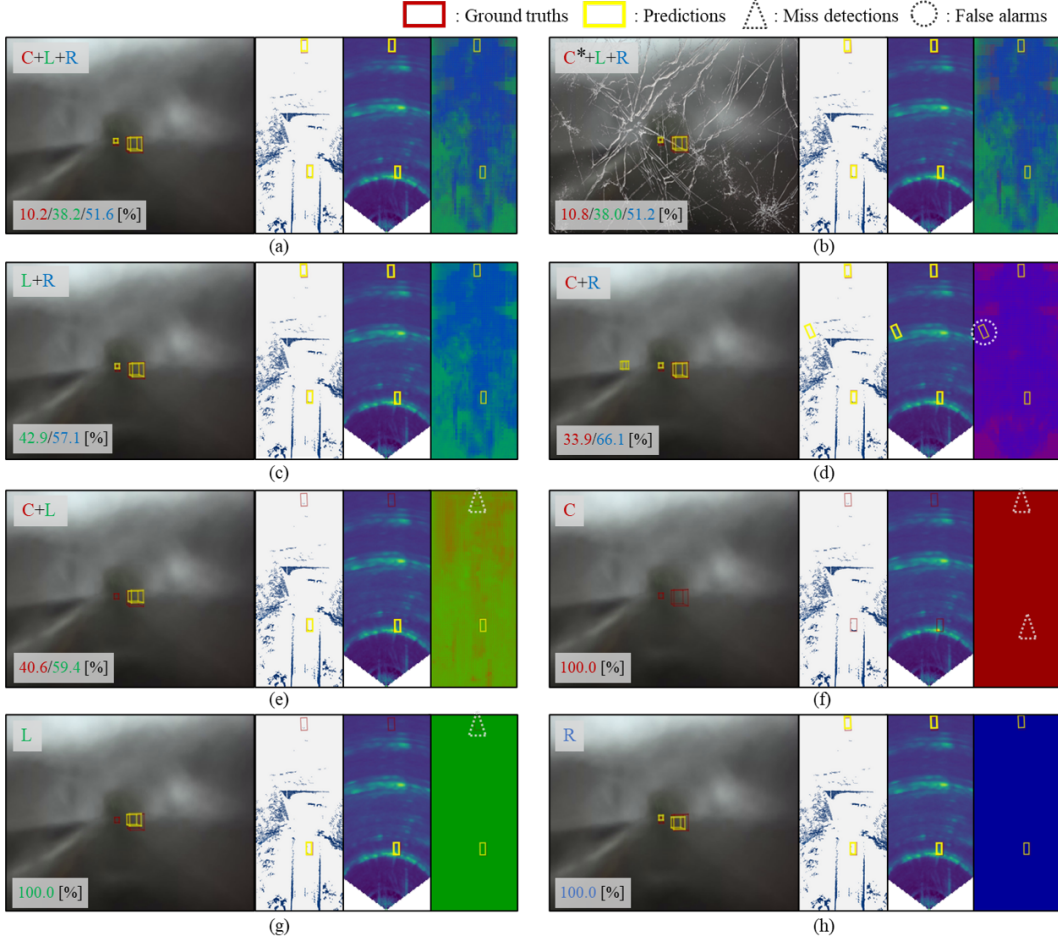


Figure 8: Qualitative results of ASF for various sensor combinations. We show results under adverse weather conditions. Each grid shows the camera image, LiDAR point cloud, 4D Radar tensor, and a SAM illustrating attention scores from CASAP. In the SAM, red, green, and blue represent attention to camera, LiDAR, and 4D Radar, respectively. The bottom-left corner of each grid shows attention score proportions (C/L/R[%]). Yellow and red boxes denote predictions and ground truth, respectively, which are displayed in all panels. Dotted triangles and circles indicate miss detections and false alarms, respectively, and are marked in the SAM. Note that predictions are visualized on all sensor data, even when a sensor is not employed for detection (e.g., predictions from L+R are also visualized on the camera image).

Fig. 8 shows the inference results for multiple vehicles in adverse weather conditions. As shown in Fig. 8-(e), when only LiDAR is utilized, nearby vehicles can be detected due to the presence of measurements, but distant vehicles experience miss detection due to the absence of measurements. In contrast, as shown in Fig. 8-(h), 4D Radar has measurements regardless of distance, enabling stable detection without miss detections. Similar to Fig. 7-(a), in the case of Fig. 8-(a), which utilizes all of 4D Radar, LiDAR, and camera, stable and accurate detection is possible without miss detections and false alarms.

D.2 Additional Quantitative Results

Multiple Random Seeds. Tab. 7 presents the mean and standard deviation of ASF (C+L+R) performance across 10 random seeds on the K-Radar benchmark v1.0, demonstrating the robustness and generality of our approach. The low standard deviations (mostly below 0.3) indicate stable performance across different random initializations. Notably, the sleet and heavy snow condition shows higher variance (± 2.85 for AP_{BEV} and ± 2.33 for AP_{3D} , respectively), suggesting this

Table 7: Additional performance comparison of ASF (C+L+R) under various weather conditions on the K-Radar (Paek et al., 2022) benchmark v1.0. We report both AP_{BEV} and AP_{3D} at IoU = 0.3 for the ‘Sedan’ class, showing mean and standard deviation across 10 random seeds (0 to 8 and 2025).

Metric	Total	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow
AP_{BEV}	88.64 ± 0.12	88.24 ± 0.12	90.29 ± 0.09	98.97 ± 0.27	88.93 ± 0.17	86.24 ± 2.85	89.25 ± 0.11	78.44 ± 0.16
AP_{3D}	87.48 ± 0.21	86.96 ± 0.17	89.94 ± 0.20	90.67 ± 0.08	88.28 ± 0.20	80.46 ± 0.21	88.92 ± 0.22	76.03 ± 2.33

Table 8: Performance comparison of monocular and stereo camera-based networks on the K-Radar (Paek et al., 2022) benchmark v2.0. We report both AP_{BEV} and AP_{3D} at IoU = 0.3 for the ‘Sedan’ and ‘Bus or Truck’ classes. We note that camera performance is evaluated on sequences 1 to 20, which consist of normal weather conditions.

Metric	Class	BEVDepth (Mono)	DSGN++ (Stereo)	ASF (BEVDepth)	ASF (Stereo)
AP_{BEV}	Sedan	19.9	24.3	82.5	82.4
	Bus or Truck	13.4	24.8	70.2	67.2
AP_{3D}	Sedan	17.8	20.8	79.4	79.3
	Bus or Truck	11.5	22.5	60.4	59.5

weather condition presents more challenging and variable detection scenarios. The consistently high performance in Fog conditions (98.97 AP_{BEV} and 90.67 AP_{3D}) with minimal variance demonstrates ASF’s particular effectiveness in low-visibility scenarios where 4D Radar excels.

Stereo Camera. As shown in Tab. 8, the stereo camera-based DSGN++ (Chen et al., 2022) achieves improved performance on K-Radar (24.8% and 22.5% AP_{3D} for Sedan and Bus or Truck, respectively) compared to the monocular camera-based BEVDepth (Li et al., 2023). We integrate DSGN++ as ASF’s camera-specific encoder. However, as shown in Tab. 8, ASF using DSGN++ shows comparable performance to ASF using BEVDepth. This may be attributed to two factors: (1) the detection performance of LiDAR and 4D Radar is significantly superior to that of the camera, and (2) K-Radar’s ZED 2i camera has a 12cm baseline compared to KITTI’s 54cm, limiting the improvement potential of stereo-based methods.

Performance on Benchmark v2.0. Tab. 9, 10, 11, and 12 present comprehensive evaluations of ASF performance across various sensor configurations and weather conditions on the K-Radar benchmark v2.0. Results demonstrate that multi-sensor fusion approaches consistently outperform single-sensor methods. The C+L+R configuration achieves the highest performance in most scenarios for both ‘Sedan’ and ‘Bus or Truck’ classes, with C*+L+R often ranking second, which indicates that a damaged camera does not significantly affect performance.

Performance analysis across weather conditions reveals that 4D Radar-inclusive combinations (C+L+R, L+R) exhibit superior performance in adverse weather (demonstrating 4D Radar’s effectiveness under challenging environmental conditions), particularly in fog and snow. While camera-LiDAR fusion (C+L) occasionally outperforms in normal and overcast conditions, its performance deteriorates significantly in precipitation.

References

- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- 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.
- Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In *Sensors*, volume 18, page 3337. MDPI, 2018.

Table 9: Performance comparison of ASF under various sensor combinations evaluated on the K-Radar (Paek et al., 2022) benchmark v2.0. We indicate the employed sensors (C: Camera, L: LiDAR, R: 4D Radar) and report AP_{3D} at IoU = 0.3 under various weather conditions for the ‘Sedan’ and ‘Bus or Truck’ classes. The **bold** and underlined values indicate the best and the second-best, respectively. All ten ASF models share the same neural network weights trained with R+L+C.

Class	Sensors	Total	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow
Sedan	R	47.3	40.7	58.8	84.4	41.0	45.9	66.1	56.5
	L	73.0	73.0	86.1	92.9	64.3	64.9	86.0	54.5
	C	14.8	14.9	7.7	-	-	-	-	-
	C*	3.7	3.7	3.2	-	-	-	-	-
	L+R	77.3	77.7	<u>87.3</u>	95.5	71.0	74.4	91.0	65.4
	C+R	52.7	49.1	<u>62.4</u>	84.6	44.6	46.0	66.5	57.2
	C+L	76.4	<u>78.3</u>	86.5	93.4	69.4	64.2	87.3	57.1
	C+L+R	79.3	78.8	86.1	93.7	72.9	<u>74.2</u>	91.2	65.8
	C*+L+R	<u>77.6</u>	78.2	87.7	<u>93.8</u>	<u>71.3</u>	74.4	<u>91.1</u>	<u>65.4</u>
	C+L*+R	58.9	59.2	64.0	92.4	47.7	65.4	66.5	58.2
Bus or Truck	R	34.2	22.9	41.0	-	0.2	21.1	83.5	51.2
	L	54.9	53.7	<u>74.8</u>	-	5.1	69.1	85.2	37.8
	C	9.6	9.0	17.2	-	-	-	-	-
	C*	3.7	3.7	0.0	-	-	-	-	-
	L+R	59.9	52.5	71.6	-	6.8	68.2	88.1	68.9
	C+R	36.2	24.4	41.4	-	0.3	23.4	82.1	56.5
	C+L	53.0	49.1	60.1	-	4.4	72.1	85.1	39.6
	C+L+R	60.4	<u>52.7</u>	77.4	-	8.0	69.2	<u>87.9</u>	68.9
	C*+L+R	<u>60.1</u>	52.1	72.0	-	8.4	<u>70.9</u>	<u>87.9</u>	69.1
	C+L*+R	40.0	31.5	38.2	-	0.4	28.9	81.0	54.8

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020.

Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022.

Seung-Hyun Kong, Dong-Hee Paek, and Sangyeon Lee. Rtnh+: Enhanced 4d radar object detection network using two-level preprocessing and vertical encoding. *IEEE Transactions on Intelligent Vehicles*, pages 1–14, 2024. doi: 10.1109/TIV.2024.3428696.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.

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.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

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.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: a simple and robust lidar-camera fusion framework. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Table 10: Performance comparison of ASF under various sensor combinations evaluated on the K-Radar (Paek et al., 2022) benchmark v2.0. We indicate the employed sensors (C: Camera, L: LiDAR, R: 4D Radar) and report AP_{BEV} at IoU = 0.3 under various weather conditions for the ‘Sedan’ and ‘Bus or Truck’ classes. The **bold** and underlined values indicate the best and the second-best, respectively. All ten ASF models share the same neural network weights trained with R+L+C.

Class	Sensors	Total	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow
Sedan	R	55.9	48.7	70.1	92.0	50.8	52.6	76.8	61.1
	L	76.7	78.1	93.8	93.6	70.0	67.8	88.4	58.2
	C	17.6	17.7	8.9	-	-	-	-	-
	C*	5.3	5.3	3.5	-	-	-	-	-
	L+R	81.8	80.7	94.3	98.1	75.7	81.0	93.6	69.6
	C+R	58.7	54.7	71.1	91.9	52.7	52.4	74.6	61.3
	C+L	79.3	81.0	94.9	95.1	72.5	68.3	89.7	59.5
	C+L+R	82.5	81.8	94.9	98.0	76.2	81.0	93.8	<u>69.5</u>
	C*+L+R	<u>82.1</u>	<u>81.1</u>	<u>94.7</u>	98.1	<u>76.0</u>	80.2	93.8	69.4
	C+L*+R	64.5	64.6	76.1	92.5	57.2	56.9	75.9	60.6
Bus or Truck	R	43.5	28.4	42.9	-	0.3	23.7	92.5	77.2
	L	63.1	57.0	75.5	-	6.2	72.3	88.7	61.3
	C	11.4	11.2	18.1	-	-	-	-	-
	C*	3.9	4.0	0.0	-	-	-	-	-
	L+R	69.5	58.3	75.2	-	7.3	71.3	<u>96.2</u>	89.7
	C+R	44.6	29.2	43.4	-	0.3	26.0	91.1	76.9
	C+L	61.8	53.1	68.0	-	5.3	75.5	90.0	64.7
	C+L+R	70.2	58.3	81.5	-	8.0	72.2	96.1	89.6
	C*+L+R	<u>70.0</u>	<u>58.2</u>	<u>75.7</u>	-	9.2	<u>73.8</u>	96.4	88.4
	C+L*+R	48.0	36.2	42.6	-	0.4	33.6	91.0	77.2

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

Weiyi Xiong, Jianan Liu, Tao Huang, Qing-Long Han, Yuxuan Xia, and Bing Zhu. Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. *IEEE Transactions on Intelligent Vehicles*, 2023.

Lianqing Zheng, Sen Li, Bin Tan, Long Yang, Sihan Chen, Libo Huang, Jie Bai, Xichan Zhu, and Zhixiong Ma. Rcfusion: 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. doi: 10.1109/TIM.2023.3280525.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023.

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*, volume 37, pages 1477–1485, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Yiran Yang, Dongshuo Yin, Xue Rong, Xian Sun, Wenhui Diao, and Xinming Li. Beyond the limitation of monocular 3d detector via knowledge distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9052, 2023. doi: 10.1109/ICCV51070.2023.00833.

Min-Hyeok Sun, Dong-Hee Paek, Seung-Hyun Song, and Seung-Hyun Kong. Efficient 4d radar data auto-labeling method using lidar-based object detection network. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2616–2621, 2024. doi: 10.1109/IV55156.2024.10588669.

Table 11: Performance comparison of ASF under various sensor combinations evaluated on the K-Radar (Paek et al., 2022) benchmark v2.0. We indicate the employed sensors (C: Camera, L: LiDAR, R: 4D Radar) and report AP_{3D} at IoU = 0.5 under various weather conditions for the ‘Sedan’ and ‘Bus or Truck’ classes. The **bold** and underlined values indicate the best and the second-best, respectively. All ten ASF models share the same neural network weights trained with R+L+C.

Class	Sensors	Total	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow
Sedan	R	20.4	13.2	26.6	59.5	18.1	20.9	28.7	30.7
	L	41.1	38.5	43.1	67.8	37.1	34.1	52.4	35.4
	C	5.9	5.9	2.8	-	-	-	-	-
	C*	1.0	1.0	0.8	-	-	-	-	-
	L+R	52.1	49.0	56.9	<u>81.4</u>	48.2	45.1	59.6	50.1
	C+R	27.5	22.6	29.8	<u>66.2</u>	23.1	27.2	30.2	39.3
	C+L	<u>52.3</u>	52.3	50.0	79.4	45.7	40.1	65.9	42.7
	C+L+R	52.9	<u>50.5</u>	55.8	81.5	49.1	44.6	60.1	<u>50.0</u>
	C*+L+R	<u>52.3</u>	49.3	<u>56.8</u>	79.6	<u>48.9</u>	<u>45.0</u>	<u>60.3</u>	<u>50.0</u>
	C+L*+R	33.7	29.9	32.3	67.5	29.0	29.1	37.6	38.7
Bus or Truck	R	16.5	8.5	22.6	-	0.1	8.7	65.3	24.6
	L	29.1	22.1	38.0	-	1.1	44.2	66.2	20.5
	C	2.6	2.5	6.3	-	-	-	-	-
	C*	2.6	2.6	0.0	-	-	-	-	-
	L+R	<u>31.7</u>	<u>24.2</u>	52.6	-	2.8	44.4	<u>70.9</u>	<u>33.2</u>
	C+R	18.1	9.7	18.3	-	0.2	8.7	65.1	28.0
	C+L	31.2	22.5	33.6	-	0.9	51.4	69.6	21.2
	C+L+R	31.6	23.7	47.0	-	<u>2.5</u>	<u>45.0</u>	72.3	33.0
	C*+L+R	32.1	24.5	<u>48.2</u>	-	2.3	44.8	70.6	33.3
	C+L*+R	19.2	13.5	15.6	-	0.0	9.8	62.5	27.9

Dong-Hee Paek and Junfeng Guan. Introduction to 4D radar: Hardware, MIMO, signal processing, dataset, and AI. Tutorial at 2024 IEEE Intelligent Vehicles Symposium, 6 2024. URL <https://www.ieee-iv-4dradar.org/>.

Xinyu Zhang, Li Wang, Jian Chen, Cheng Fang, Lei Yang, Ziyang Song, Guangqi Yang, Yichen Wang, Xiaofei Zhang, and Jun Li. Dual radar: A multi-modal dataset with dual 4d radar for autonomous driving. *arXiv preprint arXiv:2310.07602*, 2023.

Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. Enhanced k-radar: Optimal density reduction to improve detection performance and accessibility of 4d radar tensor-based object detection. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–6. IEEE, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 197–209, 2018. doi: 10.1007/978-3-030-11009-3_11. URL <https://arxiv.org/abs/1803.06199>.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Birgit Schlager, Thomas Goelles, Stefan Muckenhuber, and Daniel Watzenig. Contaminations on lidar sensor covers: Performance degradation including fault detection and modeling as potential

Table 12: Performance comparison of ASF under various sensor combinations evaluated on the K-Radar (Paek et al., 2022) benchmark v2.0. We indicate the employed sensors (C: Camera, L: LiDAR, R: 4D Radar) and report AP_{BEV} at IoU = 0.5 under various weather conditions for the ‘Sedan’ and ‘Bus or Truck’ classes. The **bold** and underlined values indicate the best and the second-best, respectively. All ten ASF models share the same neural network weights trained with R+L+C.

Class	Sensors	Total	Normal	Overcast	Fog	Rain	Sleet	Light snow	Heavy snow
Sedan	R	40.8	30.0	54.7	85.6	36.9	39.6	65.2	48.4
	L	66.8	65.7	79.0	90.7	60.1	54.6	83.4	52.7
	C	8.8	8.9	4.3	-	-	-	-	-
	C*	2.1	2.1	1.1	-	-	-	-	-
	L+R	74.0	72.6	84.4	<u>96.1</u>	67.3	<u>67.4</u>	90.8	63.4
	C+R	45.8	39.7	57.2	<u>87.4</u>	40.7	<u>41.6</u>	66.7	<u>53.9</u>
	C+L	72.3	72.8	87.2	93.3	62.5	55.4	87.2	55.9
	C+L+R	74.5	74.8	<u>85.1</u>	96.2	<u>67.0</u>	67.5	90.8	63.5
	C*+L+R	<u>74.1</u>	<u>72.9</u>	85.0	<u>96.1</u>	<u>66.9</u>	66.7	90.7	63.2
	C+L*+R	53.1	54.5	57.0	61.5	50.2	50.3	68.5	55.2
Bus or Truck	R	23.8	14.2	30.0	-	0.2	12.7	87.3	64.3
	L	45.8	35.7	61.1	-	2.3	58.2	81.9	53.3
	C	11.4	11.2	18.1	-	-	-	-	-
	C*	2.7	2.8	0.0	-	-	-	-	-
	L+R	53.1	37.5	59.8	-	6.1	56.6	88.2	<u>80.1</u>
	C+R	31.9	14.4	25.9	-	0.2	13.4	90.1	64.7
	C+L	47.3	32.4	40.7	-	3.3	63.1	85.3	55.6
	C+L+R	53.9	<u>36.5</u>	54.7	-	<u>7.9</u>	59.3	<u>90.4</u>	80.4
	C*+L+R	<u>53.6</u>	36.1	57.5	-	8.0	<u>59.7</u>	90.8	<u>80.1</u>
	C+L*+R	33.2	20.9	24.7	-	0.2	18.3	88.0	61.4

applications. *IEEE Open Journal of Intelligent Transportation Systems*, 3:738–747, 2022. doi: 10.1109/OJITS.2022.3214094.

Yilun Chen, Shijia Huang, Shu Liu, Bei Yu, and Jiaya Jia. Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4416–4429, 2022.