DSERT-RoLL: Robust Multi-Modal Perception for Diverse Driving Conditions with Stereo Event-RGB-Thermal Cameras, 4D Radar, and Dual-LiDAR

Supplementary Material

This supplemental document provides additional information on the proposed dataset, **DSERT-RoLL**, additional details, results and comparisons.

- The license for the DSERT-RoLL dataset is provided in Section 1;
- More extensive comparisons with existing datasets, along with additional details, are provided in Section 2;
- The criteria for distinguishing weather and lighting conditions are described in Section 3;
- Data statistics of the proposed dataset are presented in Section 4;
- Detailed information on the calibration between sensors of different modalities in Section 5;
- Details on the annotation procedure and the provided annotations in Section 6;
- Pixel-level alignment between cameras with different axes for the 2D detection is described in Section 7;
- The anonymization of sensitive personal information is described in Section 8;
- The implementation details of the proposed multi-modal 3D detection model are provided in Section 9;
- Additional dataset samples are presented in Section 10;
- Qualitative comparisons with other methods are provided in Section 11;
- Experiment on sensitivity to extrinsic calibration errors in Section 12;
- Evaluation results across multiple classes are presented in Section 13;

1. Dataset License

The DSERT-RoLL dataset and accompanying code are provided strictly for research purposes and are distributed under the CC BY-NC 4.0 license, allowing use solely for non-commercial activities.

2. Dataset Comparison

Table 4 provides additional comparisons with other datasets that could not be fully covered in the main paper. Although many autonomous driving datasets have been introduced recently, to the best of our knowledge there is no dataset that simultaneously includes multiple novel sensors while also covering a wide range of weather and lighting conditions. This highlights and further emphasizes the unique advantages of our proposed dataset, which provides a unified benchmark for studying different sensor modalities across diverse environmental conditions.

3. Weather and Lighting Conditions Descriptions

Table 1. Detailed criteria for weather and light conditions.

Condition	Name	Description and criteria
Weather Condition	Clear	Clear weather that does not meet the five weather conditions below.
	Fog	Weather conditions in which distant objects are dimly visible due to omnidirectional fog, corresponding to regions and periods under officially issued weather advisories.
	Light Rain	Weather conditions with a precipitation rate of up to 5 mm per hour.
	Heavy Rain	Weather conditions with a precipitation rate ranging from 10 to 15 mm per hour.
	Light Snow	Weather condition with a snowfall accumulation rate of less than 1 cm per hour.
	Heavy Snow	Weather condition with a snowfall accumulation rate of greater than 1 cm per hour.
Light Condition	Normal	Standard lighting condition with balanced illumination and no significant overexposure or underexposure.
	Low Light	Condition with insufficient illumination, typically resulting in reduced visibility and increased image noise.
	Over Expose	Condition in which excessive brightness causes loss of detail in highlighted regions.
	HDR	Condition where multiple exposure levels are combined to capture both dark and bright areas with enhanced contrast and detail.

The proposed DSERT-RoLL dataset includes a wide range of weather conditions, such as clear, fog, rain, and snow, as well as diverse light conditions, including normal, low-light, overexposed, and HDR scenarios. These variations allow the dataset to cover autonomous driving scenes across highly diverse environmental settings. All sequences are categorized according to the criteria we defined for each condition, and the dataset is organized following the detailed definitions presented in Table 1.

4. Dataset Statistics

We analyze the distributions of object classes and weather conditions across distance intervals in both the training and test sets, as shown in Figs. 1 and 2. The results show broad coverage over the full distance range and similar tendencies between the two splits, indicating a limited distribution gap between training and evaluation.

As shown in Fig. 1, the three object categories, Bike, Pedestrian, and Vehicle, are distributed across all distance bins in both splits. Although their proportions vary by interval, no class is concentrated within only a narrow distance range, which suggests balanced coverage over near, middle, and far regions.

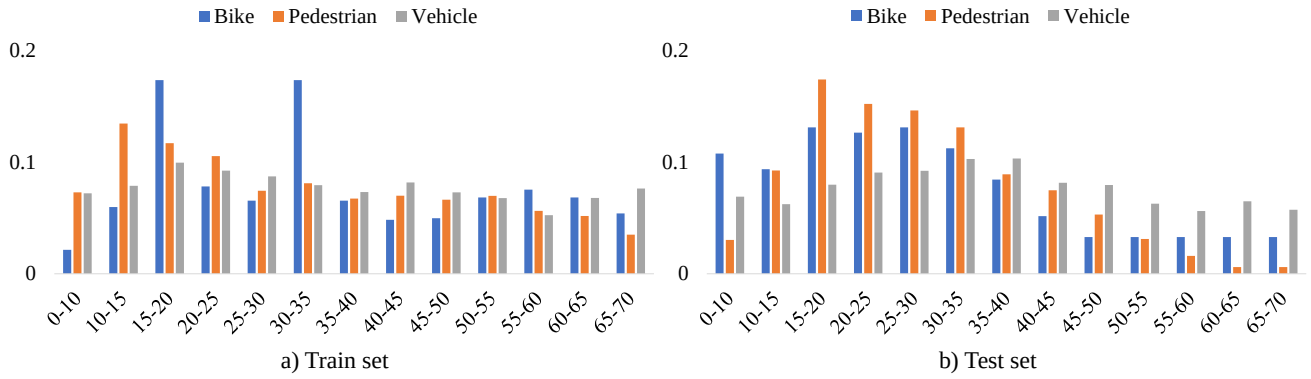


Figure 1. Class distribution across distance bins for the Bike, Pedestrian, and Vehicle categories. Panel a) shows the distribution in the training set, and panel b) shows the distribution in the test set.

A similar pattern is observed for weather conditions in Fig. 2. Clear, Fog, Heavy Rain, Heavy Snow, Light Rain, and Light Snow are all represented across the analyzed distance bins, and the train and test sets exhibit comparable trends. This indicates that the dataset remains reasonably balanced with respect to both semantic categories and environmental conditions over distance. Overall, these statistics support fair training and reliable evaluation under diverse real-world scenarios.

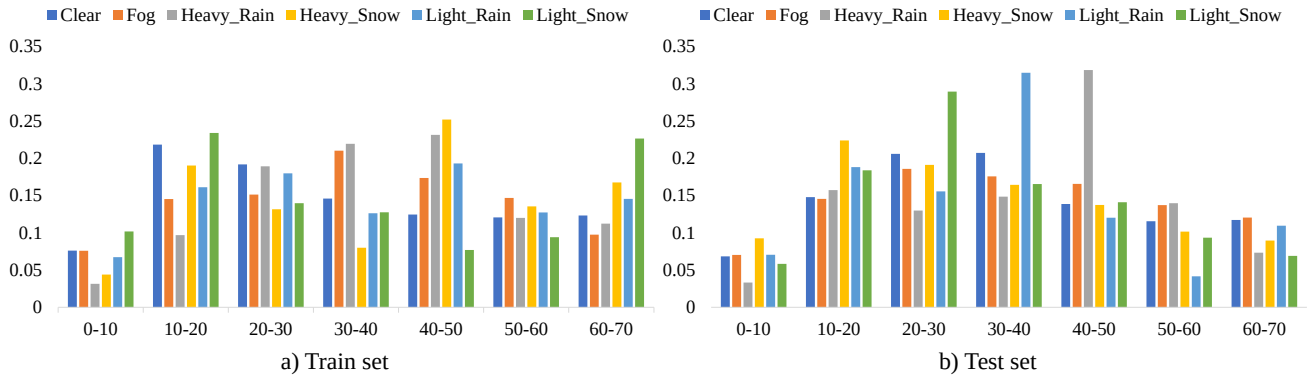


Figure 2. Weather condition distribution across distance bins. Each bar represents the proportion of Clear, Fog, Heavy Rain, Heavy Snow, Light Rain, and Light Snow in each distance interval. Panel a) shows the distribution in the training set, and panel b) shows the distribution in the test set.

5. Details of calibration

In a multi-modal sensor system, calibration is essential for fusing and jointly using measurements from different sensors. To make this process convenient and effective, we use the RGB cameras as the reference modality and perform pairwise calibration with the other sensors in the following order: [Sec. 5.1] calibration between RGB and the event stereo cameras, [Sec. 5.2] calibration between RGB and the thermal stereo cameras, [Sec. 5.3] calibration between RGB and the LiDAR, and [Sec. 5.4] calibration between RGB and the 4D radar.

5.1. Calibration of the RGB and Event Stereo Cameras

To robustly calibrate the event cameras, we first collect a large amount of data from diverse viewpoints. Following prior work [13, 29], we employ an event-to-image reconstruction model [36] to convert the raw event streams into dense intensity images. As shown in Fig. 3, this yields four synchronized image pairs from two RGB cameras and two event cameras, from which we construct multiple image pairs for calibration. Using the widely adopted calibration toolbox Kalibr [12], we estimate the intrinsic parameters of each of the four cameras and jointly optimize the extrinsic parameters between all camera pairs in a single optimization step. This procedure allows us to obtain accurate intrinsic and extrinsic calibration for the four cameras.

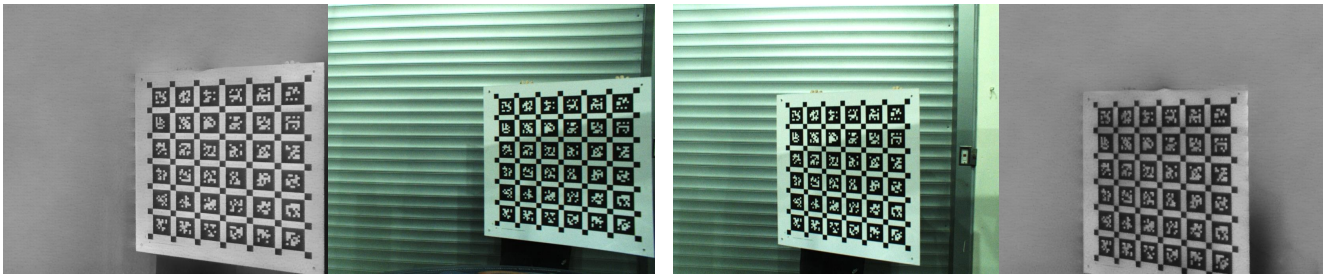


Figure 3. Sample data used for calibration, captured simultaneously by all four cameras observing the calibration pattern. The event data were reconstructed into intensity images using previous method [36].

5.2. Calibration of the RGB and Thermal Stereo Cameras

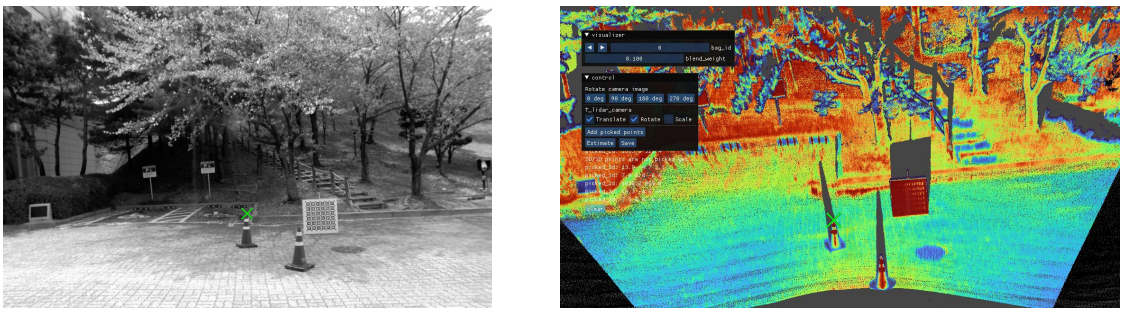
Because the standard calibration pattern is hardly visible in the thermal images, we resort to a manual RGB–thermal calibration procedure. To this end, we design a novel calibration target in which a grid of copper wires is mounted on a board so that the wires maintain a different temperature from the background surface and can be clearly captured by the thermal cameras. As shown in Fig. 4, this setup yields four synchronized images from two RGB cameras and two thermal cameras, from which we construct multiple image pairs for calibration. We then place points sequentially at the grid intersections and use them to construct the corresponding RGB–thermal point pairs.

Figure 4. Sample data used for calibration, captured simultaneously by all four cameras observing the calibration pattern. We place points sequentially at the grid intersections and use them to construct the corresponding RGB-thermal point pairs.

5.3. Calibration between RGB camera and LiDAR

We use a recent LiDAR-camera extrinsic calibration toolbox [19] to estimate accurate transformations between each LiDAR and the reference RGB camera. As shown in Fig. 5, the toolbox first constructs a dense LiDAR point cloud and derives an

initial alignment with the camera using geometric correspondences. It then refines the extrinsic parameters through error-based optimization, yielding a precise calibration. The same procedure is applied to both the long-range and short-range LiDARs in our setup.

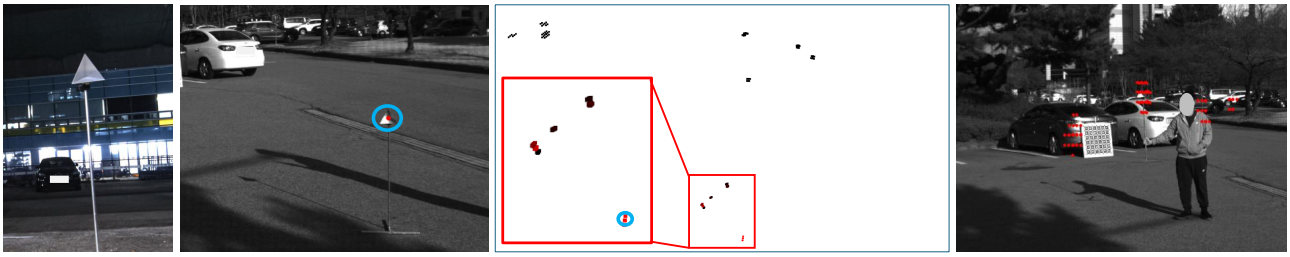


Left RGB Camera LiDAR Point Cloud

Figure 5. Example of a scene used by the calibration toolbox [19] for computing the extrinsic parameters between the RGB camera and the LiDAR.

5.4. Calibration between RGB camera and 4D Radar

We calibrate the left RGB camera and radar by placing a corner reflector target in front of the sensor rig, causing the radar returns to collapse into a single high-intensity point. We record radar point clouds with 3D location, power, and Doppler attributes together with time-synchronized RGB images. Using an in-house annotation tool, we manually click the reflector in the radar point cloud and the corresponding pixel in the RGB image for each synchronized pair, yielding a set of 3D–2D correspondences. Since the reflector concentrates the incident energy, the radar return power becomes significantly higher. We therefore select, in the 3D radar point cloud, the point with the highest return power as its 3D counterpart, as shown in Fig. 6 at the location indicated by the light-blue circle. From these correspondences, we solve for the rigid transformation between the radar and camera coordinate systems, enabling accurate projection of radar point clouds onto the image plane, while the remaining sensing modalities are tied to the radar via their pre-calibrated extrinsic parameters.



Corner Reflector 2D Image Point Select 3D Radar Point Select Projection Example

Figure 6. Example scene used for RGB–4D Radar calibration. Using the 4D Radar reflector target, we obtain 2D–3D point correspondences at various distances.

5.5. Stereo Camera Rectification

The proposed DSERT-RoLL dataset provides stereo data for all camera modalities. For stereo rectification, we process each modality (RGB, event, and thermal) independently. Given the intrinsic matrices K_L, K_R, the distortion coefficients, and the relative pose (R, t) obtained from our calibration, we compute the rectifying rotations R_L, R_R and the new projection matrices P_L, P_R using OpenCV's `stereoRectify`. We then undistort and warp the left and right images with `initUndistortRectifyMap` and `remap`, so that corresponding points in each stereo pair lie on the same scanline and the epipolar lines become approximately horizontal. The same procedure is applied to the RGB, event, and thermal stereo pairs. We provide sample stereo-rectified images in Fig. 7, where the rectification quality can be visually inspected. As shown in the figure, corresponding points between the left and right cameras of the same modality lie on the same horizontal scanline, confirming proper stereo rectification.

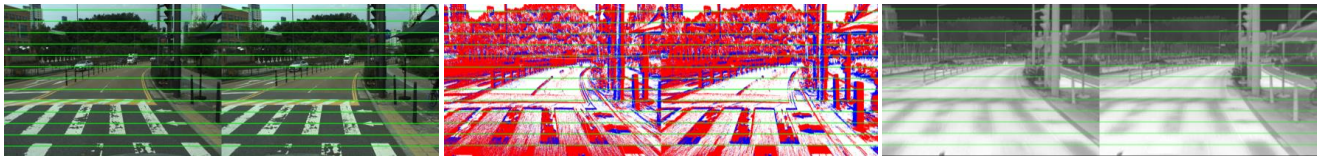


Figure 7. Example of stereo rectification results for the RGB, event, and thermal modalities. The green horizontal lines visualize the epipolar lines, which become well aligned after calibration.

5.6. Calibration Results of All Sensors

Through the calibration steps described above, we obtain a consistent extrinsic calibration that links all sensing modalities. Figure 11 shows a representative scene demonstrating the quality of this calibration: 3D points from the range sensors are projected onto the image planes of the six camera modalities (left and right images of RGB, event, and thermal), demonstrating good spatial alignment across all views.

6. Details of Annotation

6.1. Annotation Procedure

Figure 8. Multi-modal annotation tool. 4D Radar points are shown in red, and LiDAR points are color-coded by timestamp using a viridis colormap. 3D sensor data and 3D annotations can be projected onto camera sensor images, ensuring consistent annotations across modalities under diverse driving conditions..

As shown in Fig. 8, to obtain reliable annotations under diverse adverse weather and lighting conditions, we jointly exploited multiple sensor modalities during the labeling process. Specifically, we used two complementary 3D range sensors, LiDAR and 4D Radar, to provide accurate geometric information, while RGB and thermal images were employed to verify the alignment and validity of the bounding boxes. This multi-modal setup improves the accuracy of the ground-truth bounding boxes and helps ensure that no objects are missed during annotation.

All multi-modal data were imported into a dedicated professional annotation tool [1], and labeling was carried out by hired expert annotators. Fig. 8 illustrates the annotation tool interface used by annotators to label multi-modal sensor data. In the main view, LiDAR and RADAR points are plotted in different colors, providing reliable 3D information even under adverse weather conditions (*e.g.*, snow, fog). The 3D sensor data and 3D bounding boxes are projected onto the images, assisting annotators during labeling and enabling consistent annotations across modalities. Each video sequence was subsequently reviewed by at least three annotators to guarantee high-quality labels.

For 2D bounding box annotation, 3D annotations are projected onto the image plane and subsequently refined. Odometry is defined as the relative pose with respect to the first sample of each video sequence and is estimated using an IMU–LiDAR-

coupled SLAM algorithm [38], with GNSS signals used to aid pose estimation when 3D sensors are unreliable due to weather condition.

6.2. Types of Annotations

Table 2 summarizes the annotation types provided in the DSERT-RoLL dataset. Odometry is defined as the relative pose with respect to the first frame of each video sequence. The weather and lighting conditions are annotated once per sequence, meaning that a single sequence-level label is assigned rather than frame-level labels. The 3D annotations follow the Waymo Open Dataset (WOD) [39] format, while the 2D annotations are provided in the COCO [22] format.

Table 2. Provided annotation types and descriptions.

Type	Format	Description
3D	3D Bbox	$(center\_x, center\_y, center\_z, l, w, h, yaw, class)$
2D	2D Bbox	$(u_{min}, v_{min}, w, h, class)$
Odometry	pose	$\mathbf{p} = [\mathbf{R} \mid \mathbf{t}]$ where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$
Weather Condition	text	Sequence-level weather condition
Lighting Condition	text	Sequence-level lighting condition

7. Image Homography for 2D Detection

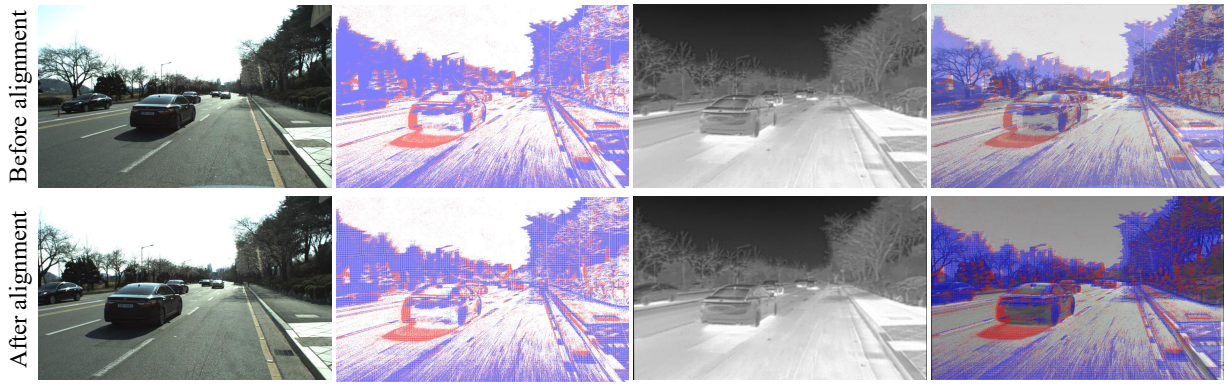


Figure 9. Example of RGB–event–thermal alignment using the homography. The top row shows the three modalities before alignment, and the bottom row shows the event and thermal images after being warped onto the RGB image plane and cropped to the common field of view, with the last column visualizing their overlap.

The proposed multi-modal 3D object detection framework operates by projecting 3D information onto each camera and sampling in the image space, and therefore does not require the images themselves to be mutually aligned. However, for fair comparison in 2D object detection and to enable a pixel-level multi-modal fusion approach, pixel-wise alignment between the different image modalities is necessary. To this end, we align thermal and event images to the RGB camera using a rotation-based image homography derived from the calibrated intrinsics and extrinsics of all sensors, following the prior work [13]. We do not perform any additional dedicated calibration for this alignment step; instead, we reuse the existing calibration parameters and treat the resulting homography as an approximate alignment. From the LiDAR-to-camera extrinsic matrices, we extract the 3×3 rotation matrices and compute the relative rotation between each source camera (thermal or event) and the RGB camera. After rescaling each camera's intrinsic matrix to the target resolution, we construct the homography using the infinite-plane approximation

$$H = K_{\mathrm{RGB}}\, R_{\mathrm{RGB,src}}\, K_{\mathrm{src}}^{-1}, \tag{1}$$

where each K denotes the intrinsic parameters of the corresponding camera, and R represents the extrinsic parameters that transform points from the source camera to the RGB camera coordinate system. We apply homography to warp the thermal and event images onto the RGB image plane, followed by cropping the overlapping field of view shared by all three modalities. An example scene is following Fig. 9. The event image exhibits a grid-like appearance because the sparse events are forward-warped and dispersed over the RGB image plane.

8. Privacy Concerns

To address privacy concerns, we ensure that all privacy-sensitive information in the dataset is properly anonymized. In particular, all human faces and vehicle license plates are blurred to prevent any form of personal identification, as illustrated in Fig. 10.

License Plate Number Human Face

Figure 10. Example images in which privacy-sensitive information, such as license plate numbers and human faces, has been blurred.

9. Implementation Details

We train our multi-modal 3D object detection model in an end-to-end manner for a total of 20 epochs. The batch size is set to 8 per GPU, and all experiments are conducted using four NVIDIA Quadro RTX 8000 GPUs. We set the 3D detection range to $[0, 75.2\,\mathrm{m}]$ along the X axis, $[-75.2\,\mathrm{m}, 75.2\,\mathrm{m}]$ along the Y axis, and $[-2\,\mathrm{m}, 4\,\mathrm{m}]$ along the Z axis. For each attention operation, we sample $Q = 4$ points (Eq. (3) in the main paper). Following [10], the grid size S for the box refinement is set to 6. For the each image branch, we use a Swin-Tiny [25] backbone together with an FPN [23]. We perform feature fusion using only the left cameras. The event data are converted into a tensor-based voxel grid representation [50] with a bin size of 5, while the thermal data are used as single-channel images. To reduce computational cost, we downsample the RGB, thermal, and event inputs to $1/4$, 1, and $1/2$ of their original resolution, respectively. During training, we adopt common data augmentation techniques: random flips along the horizontal axis, global scaling with factors drawn from $[0.95, 1.05]$, and rotations about the Z axis sampled from $[-\pi/4, \pi/4]$. After prediction, we apply non-maximum suppression (NMS) with an IoU threshold of 0.7 to filter out overlapping detections and keep a single bounding box per object.

10. More Dataset Samples

We provide additional dataset samples in Figs. 12 and 13 to illustrate the diversity of weather conditions. Figure 12 presents scenes under clear, fog, and light snow, whereas Fig. 13 includes samples from heavy snow, light rain, and heavy rain. Moreover, we illustrate dataset samples of varying light conditions in Fig. 14.

11. Qualitative Results

Fig. 15 presents qualitative 3D detection results on the DSERT-RoLL dataset. For comparison, we include Radar-based RTNH [30], LiDAR-based VoxelNeXt [8], and LiDAR–RGB fusion LoGoNet [20]. Our method leverages all available sensors (RGB, event, thermal, 4D Radar, and LiDAR), and the proposed multi-modal approach demonstrates robust performance across a wide range of weather conditions.

Normal. Under normal conditions, all methods show reasonable performance and successfully detect distant objects as well as objects in the opposite lane. However, RTNH suffers from degraded detection performance when 4D Radar returns are sparse or only partially observed. VoxelNeXt also fails to detect the bus in this scenario. LoGoNet successfully detects all objects by exploiting sensor fusion, but the estimated object sizes are inaccurate. In contrast, our model accurately predicts both the positions and sizes of all objects.

Fog. In foggy scenes, the range and quality of LiDAR measurements degrade significantly, causing VoxelNeXt and LoGoNet—both heavily dependent on LiDAR—to miss the object. RTNH, which relies on robust 4D Radar, detects the targets but produces several false positives due to sensor noise. Our model, on the other hand, leverages 4D Radar for object detection and is effectively guided by image sensors, resulting in more accurate and stable predictions.

Rain. In the rainy-condition example, only our model successfully detects the bike. Radar is inherently weak at capturing small objects, and LiDAR performance also degrades under rain, which increases noise and introduces ambiguity into the

detection, making it difficult to detect a small and fast-moving bike. In contrast, our method leverages the complementary advantages of different sensors, enabling reliable detection even under such adverse conditions.

Snow. Under snowy conditions, RTNH fails to detect stationary objects due to the inherent characteristics of radar, which is more sensitive to moving targets. VoxelNeXt detects all objects, but LiDAR noise induced by snow leads to uncertain predictions, producing multiple ambiguous bounding boxes around the vehicle's LiDAR cluster center. LoGoNet also fails to detect some objects because the RGB images are partially occluded by snow. In this snow scenario, thermal sensing provides more reliable cues than RGB and LiDAR, and our model effectively exploits high-confidence thermal responses while down-weighting noisy measurements from other sensors, resulting in accurate and stable detections.

12. Sensitivity of Extrinsic Calibration

We perturb the extrinsic calibrations between modalities by adding random noise to evaluate the sensitivity of our model to calibration errors. Table 3 shows that the performance decreases under noisy calibration as expected. However, the degradation is not substantial, indicating that our model remains fairly robust even under realistic levels of calibration noise. These results suggest that the model does not rely excessively on precise extrinsic calibration and can maintain stable performance in the presence of moderate calibration perturbations.

Table 3. Detection performance sensitivity to calibration errors.

Noise (translation, rotate)	Weather Condition						Light Condition			
	Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
0cm, 0°	**90.30**	**71.42**	**95.10**	**80.26**	**85.59**	**72.94**	**82.93**	**92.65**	**85.47**	**86.33**
3cm, 3°	89.96	70.19	94.60	79.50	84.78	72.22	82.51	92.33	84.81	85.90
5cm, 5°	88.98	68.46	93.75	79.42	83.67	71.72	82.02	92.02	84.16	85.75

13. Quantitative Results for Multiple Classes

Following previous works [3, 18, 30], the main paper focuses on reporting results for the Vehicle category. In this supplementary material, we additionally provide results for the Pedestrian and Bike categories. For 3D object detection, Table 5 presents the results of stereo-based and 3D range sensor–based approaches, while Table 6 summarizes the performance of multi-modal fusion–based methods. For 2D object detection, Table 7 reports the multi-class results of the methods discussed in the main paper.

References

[1] Segments.ai: Building high-quality training data for computer vision. https://segments.ai/. Accessed: 2025-11-20. 5

[2] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20178–20188, 2023. 12

[3] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 8, 12

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. 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. 12

[5] Y. Chae, H. Kim, and K.-J. Yoon. Towards robust 3d object detection with lidar and 4d radar fusion in various weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15162–15172, 2024. 19

[6] K. Chaney, F. Cladera, Z. Wang, A. Bisulco, M. A. Hsieh, C. Korpela, V. Kumar, C. J. Taylor, and K. Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4015–4022, June 2023. 12

[7] Y. Chen, S. Liu, X. Shen, and J. Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12536–12545, 2020. 18

[8] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 7, 17, 18

[9] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 12

[10] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1201–1209, 2021. 7

[11] C. A. Diaz-Ruiz, Y. Xia, Y. You, J. Nino, J. Chen, J. Monica, X. Chen, K. Luo, Y. Wang, M. Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21383–21392, 2022. 12

[12] P. Furgale, J. Rehder, and R. Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286. IEEE, 2013. 3

[13] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 3, 6, 12

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. URL https://api.semanticscholar.org/CorpusID:6724907. 12

[15] Z. Gu, J. Ma, Y. Huang, H. Wei, Z. Chen, H. Zhang, and W. Hong. Hgsfusion: Radar-camera fusion with hybrid generation and synchronization for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3185–3193, 2025. 19

[16] X. Guo, S. Shi, X. Wang, and H. 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. 18

[17] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen. Deim: Detr with improved matching for fast convergence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15162–15171, 2025. 20

[18] D. Kent, M. Alyaqoub, X. Lu, H. Khatounabadi, K. Sung, C. Scheller, A. Dalat, A. bin Thabit, R. Whitley, and H. Radha. Msu-4s-the michigan state university four seasons dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22658–22667, 2024. 8

[19] K. Koide, S. Oishi, M. Yokozuka, and A. Banno. General, single-shot, target-less, and automatic lidar-camera extrinsic calibration toolbox. *arXiv preprint arXiv:2302.05094*, 2023. 3, 4

[20] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023. 7, 17, 19

[21] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17182–17191, 2022. 19

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[24] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 12

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[26] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 19

[27] J. Mao, M. Niu, C. Jiang, H. Liang, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, et al. One million scenes for autonomous driving: Once dataset. *NeurIPS*, 2021. 12

[28] M. Meyer and G. Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th european radar conference (EuRAD)*, pages 129–132. IEEE, 2019. 12

[29] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza. How to calibrate your event camera. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, June 2021. 3

[30] D.-H. Paek, S.-H. Kong, and K. T. Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022. 7, 8, 12, 17, 18

[31] A. Palffy, E. Pool, S. Baratam, J. F. Kooij, and D. 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. 12

[32] E. Palladin, R. Dietze, P. Narayanan, M. Bijelic, and F. Heide. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 19

[33] A. Patil, S. Malla, H. Gang, and Y.-T. Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE, 2019. 12

[34] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin. A* 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 12

[35] M. Pitropov, D. E. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021. 12

[36] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 3

[37] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17021–17030, 2022. 12

[38] T. Shan and B. Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 6

[39] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 6, 12

[40] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017. 12

[41] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 20

[42] L. Wang, X. Zhang, B. Xv, J. Zhang, R. Fu, X. Wang, L. Zhu, H. Ren, P. Lu, J. Li, et al. Interfusion: Interaction-based 4d radar and lidar fusion for 3d object detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12247–12253. IEEE, 2022. 19

[43] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 12

[44] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 3095–3101. IEEE, 2021. 12

[45] Y. Xiao, F. Meng, Q. Wu, L. Xu, M. He, and H. Li. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5541–5549, 2024. 20

[46] G. Zhang, J. Chen, G. Gao, J. Li, S. Liu, and X. Hu. Safdnet: A simple and effective network for fully sparse 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14477–14486, June 2024. 18

[47] G. Zhang, C. Junnan, G. Gao, J. Li, and X. Hu. Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 36, 2024. 18

[48] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 20

[49] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan, et al. Tj4dradset: A 4d radar dataset for autonomous driving. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 493–498. IEEE, 2022. 12

[50] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 7

Table 4. Comparison of datasets and benchmarks for perception in autonomous driving. The upper rows of the table list datasets built with conventional sensors, while the lower rows present datasets that include novel sensors. The symbol △ indicates annotations that are not officially provided in the dataset but were annotated by other paper authors. If 3D bounding boxes are available, the column 'Num Data' reports the number of samples that include 3D bounding boxes; otherwise, it reports the total number of data samples.

Dataset	Num	Adverse Weather				3D Range Sensor		Camera Sensor			Ground-truth		
	Data	Clear	Rain	Fog	Snow	LiDAR	Radar	Frame	Event	Thermal	3D Bbox.	Tr. ID	Odom
KITTI [14]	15k	✓	✗	✗	✗	✓	✗	Stereo	✗	✗	✓	✓	✓
Waymo [39]	230k	✓	✓	✗	✗	✓	✗	Multi-view	✗	✗	✓	✓	✗
NuScenes [4]	40k	✓	✓	✗	✗	✓	3D	Multi-view	✗	✓	✓	✓	✓
H3D [33]	27k	✓	✗	✗	✗	✓	✗	Multi-view	✗	✗	✓	✓	✓
A*3D [34]	39k	✓	✗	✓	✗	✓	✗	Stereo	✗	✗	✓	✗	✗
PandaSet [44]	8.2k	✓	✗	✗	✗	✓	✗	Multi-view	✗	✗	✓	✗	✓
Once [27]	1M	✓	✓	✗	✗	✓	✗	Multi-view	✗	✗	✓	✗	✗
Argoverse 2 [43]	150k	✓	✓	✗	✓	✓	✗	Multi-view	✗	✗	✓	✓	✓
CADC [35]	8k	✓	✓	✗	✓	✓	✗	Multi-View	✗	✗	✓	✓	✓
Ihtaca365 [11]	14.8k	✓	✓	✗	✓	✓	✗	Multi-view	✗	✗	✗	✗	✓
K-Radar [30]	35k	✓	✓	✓	✓	✓	4D	Multi-view	✗	✗	✓	✓	✓
TJ4DRadSet [49]	7.8k	✓	✗	✗	✗	✓	4D	Mono	✗	✗	✓	✓	✓
VoD [31]	8.7k	✓	✗	✗	✗	✓	4D	Stereo	✗	✗	✓	✓	✓
RADIal [37]	25k	✓	✗	✗	✗	✓	3D	Mono	✗	✗	✗	✗	✓
Astyx [28]	0.5k	✓	✗	✗	✗	✓	4D	Mono	✗	✗	✓	✗	✗
ZOD [2]	100k	✓	✓	✗	✓	✓	4D	Mono	✗	✗	✓	✗	✓
SeeingThroughFog [3]	13.5k	✓	✓	✓	✓	✓	3D	Stereo	✗	Mono	✓	✗	✗
Multispectral [40]	3.0k	✓	✓	✓	✗	✗	✗	Mono	✗	Mono	✗	✗	✗
M^3FD [24]	4.2k	✓	✓	✓	✗	✗	✗	Mono	✗	Mono	✗	✗	✗
KAIST [9]	8.9k	✓	✗	✗	✗	✓	✗	Stereo	✗	Mono	✗	✗	✗
DSEC [13]	5.4k	✓	✗	✗	✗	✓	✗	Stereo	Stereo	✗	△	△	✓
M3ED [6]	122k	✓	✗	✗	✗	✓	✗	Stereo	Stereo	✗	✗	✗	✓
DSERT-RoLL (Ours)	22k	✓	✓	✓	✓	✓	4D	Stereo	Stereo	Stereo	✓	✓	✓

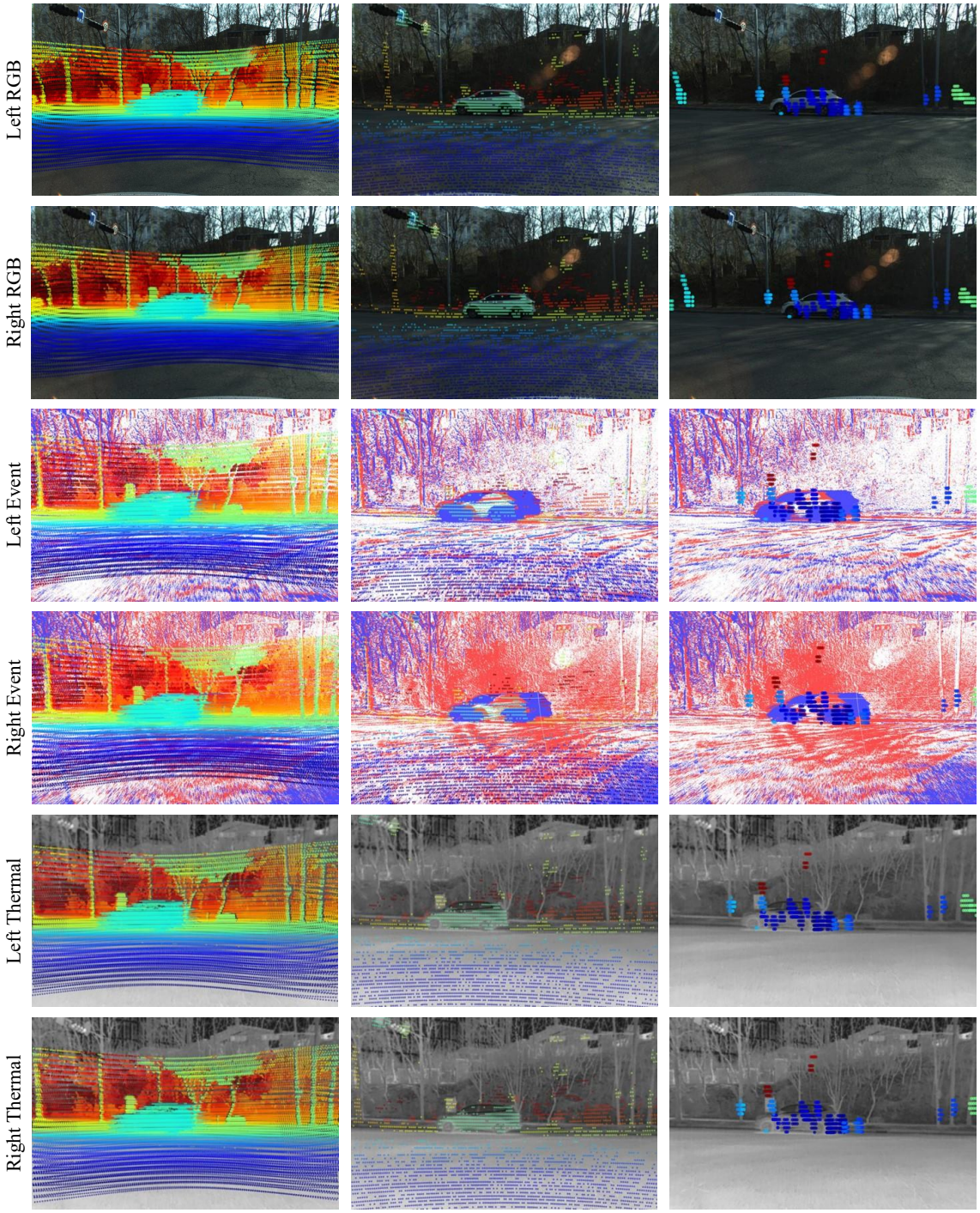


Figure 11. Calibration results for all sensors, accompanied by an example scene illustrating the projection of 3D range sensor measurements onto the image planes of all six modalities.

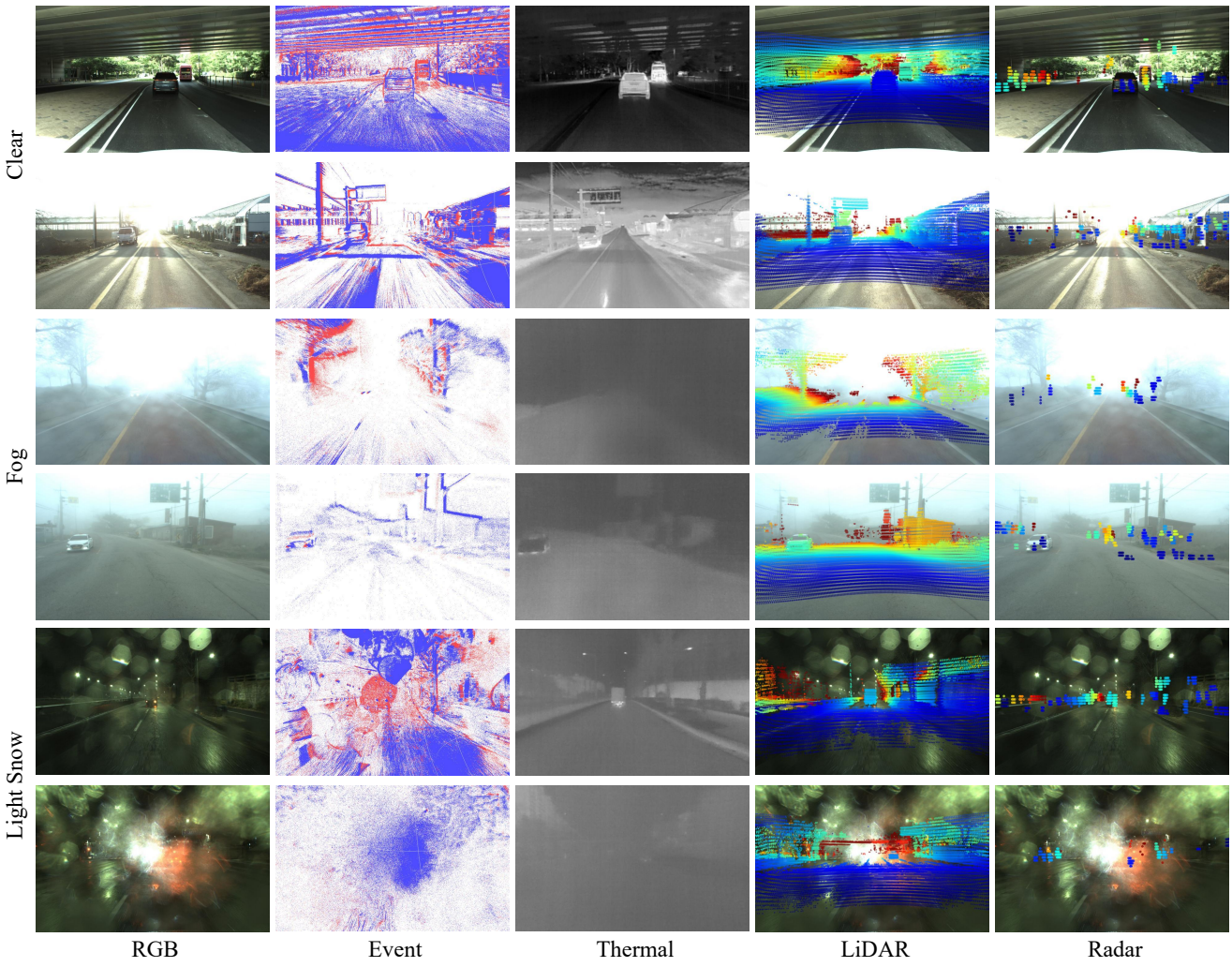


Figure 12. More data samples for clear, fog and, light snow weather conditions. All 2D camera images are from the left camera. We project 3D range sensor data onto the left RGB camera image.

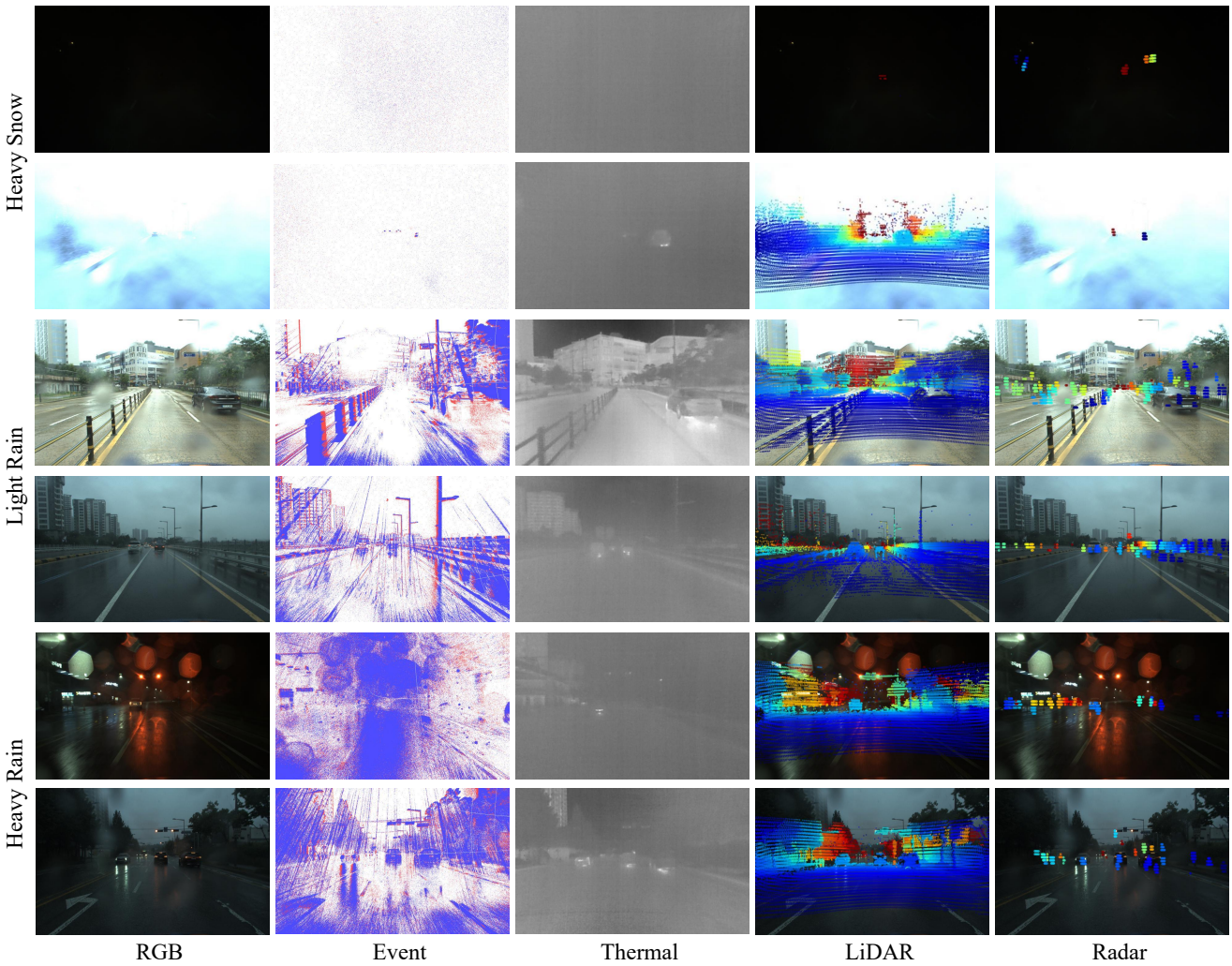


Figure 13. More data samples for heavy snow, light rain and, heavy rain weather conditions. All 2D camera images are from the left camera. We project 3D range sensor data onto the left RGB camera image.

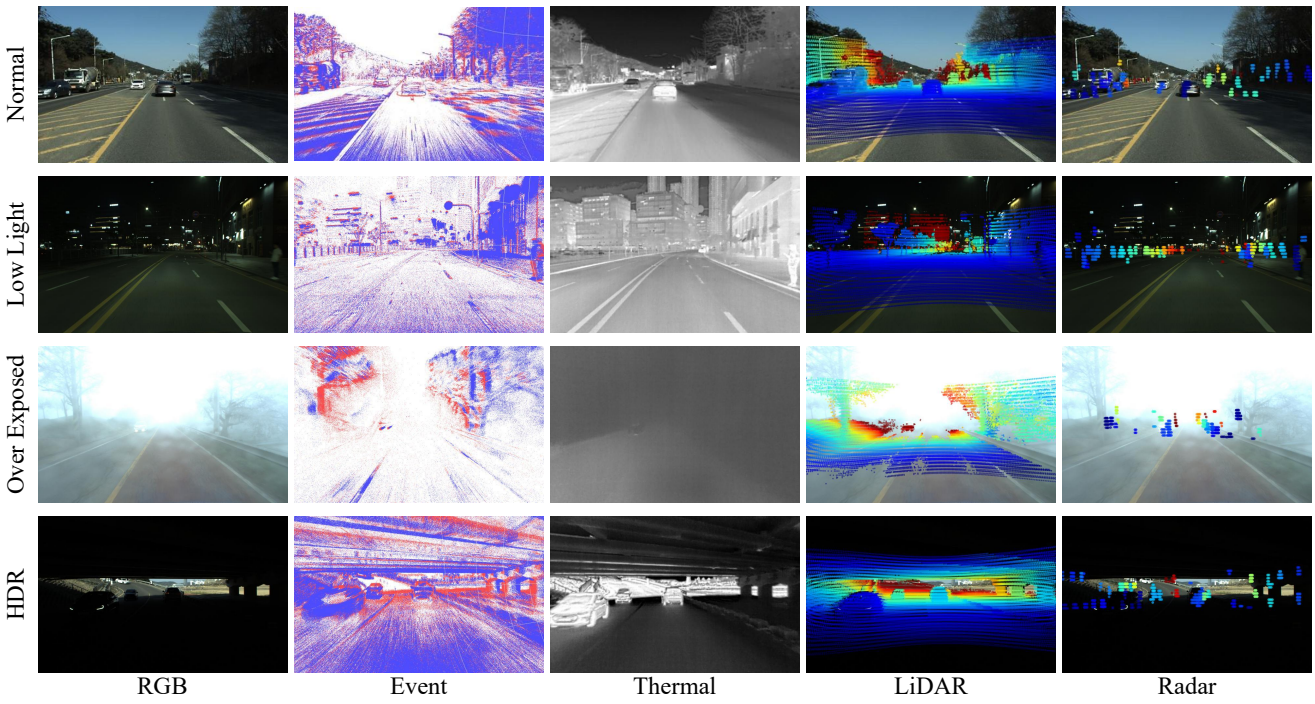


Figure 14. More data samples for light conditions. All 2D camera images are from the left camera. We project 3D range sensor data onto the left RGB camera image.

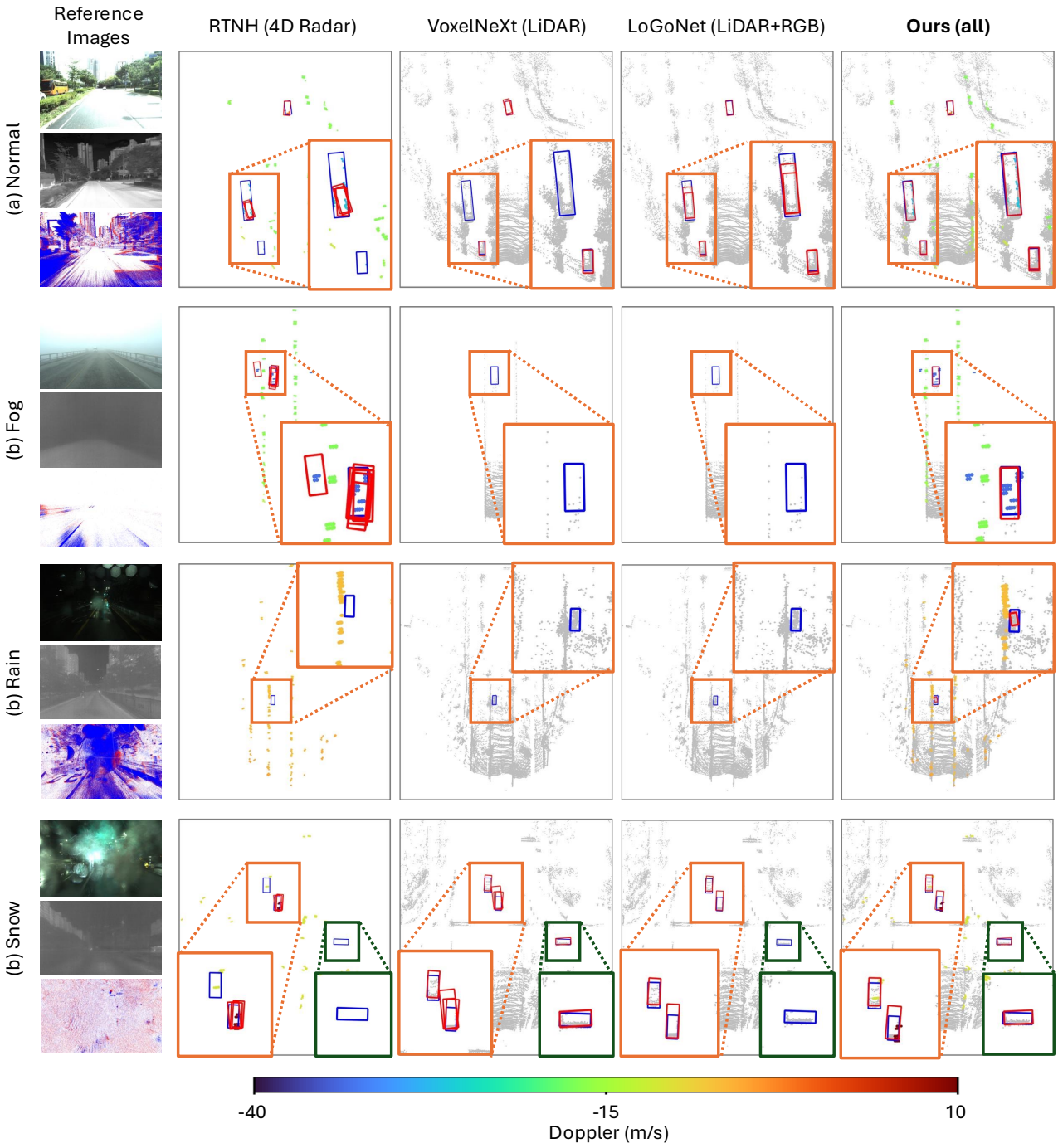


Figure 15. Qualitative comparison of 3D detection results on the DSERT-RoLL dataset in the BEV (bird's-eye-view) plane. 4D Radar-based (RTNH [30]), LiDAR-based (VoxelNeXt [8]), and LiDAR–RGB fusion (LoGoNet [20]) methods are shown for comparison, and Ours uses all available sensors (RGB, event, thermal, 4D Radar, and LiDAR). Regions containing objects are highlighted and zoomed in for better visibility. We visualize only the 3D range sensor used by each method. LiDAR measurements are shown as gray point clouds, while 4D radar points are colored according to their Doppler velocities. Blue boxes denote the 3D ground truth, and red boxes denote the predicted 3D bounding boxes.

Table 5. 3D object detection performance comparison on the DSERT-RoLL dataset for vehicle (IoU = 0.5), bike (IoU = 0.3), and pedestrian (IoU = 0.3) detection. R, E, T, 4R, and L represent the RGB, Event, and Thermal cameras, as well as the 4D Radar and LiDAR, respectively. In addition, VEH and PED denote the vehicle and pedestrian classes, respectively. N/A indicates that the corresponding class is not present under the given adverse weather condition.

Modality	Methods	Class	Weather Condition						Light Condition			
			Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
Stereo-based												
R	DSGN [7]	VEH	31.08	43.66	42.48	20.51	25.94	0.01	29.99	25.68	22.55	40.69
		BIKE	0.18	N/A	0.00	N/A	N/A	N/A	0.00	0.00	N/A	0.84
		PED	0.00	N/A	1.88	N/A	0.00	N/A	1.84	0.00	0.00	0.00
	LIGA [16]	VEH	35.52	41.67	37.52	20.57	26.02	0.00	31.31	30.06	22.44	42.80
		BIKE	0.15	N/A	0.76	N/A	N/A	N/A	0.00	0.40	N/A	0.47
		PED	0.04	N/A	0.42	N/A	0.00	N/A	0.37	0.01	0.06	0.01
E	DSGN [7]	VEH	24.23	22.06	26.93	31.38	23.12	0.01	21.41	21.42	15.58	36.44
		BIKE	0.06	N/A	0.23	N/A	N/A	N/A	0.00	0.24	N/A	0.10
		PED	0.25	N/A	0.36	N/A	0.00	N/A	0.22	0.06	0.70	0.00
	LIGA [16]	VEH	27.11	22.53	23.43	22.84	24.61	0.00	23.10	23.20	15.30	34.92
		BIKE	0.35	N/A	0.11	N/A	N/A	N/A	0.87	0.07	N/A	0.43
		PED	0.17	N/A	0.74	N/A	0.00	N/A	1.07	0.08	1.01	0.03
T	DSGN [7]	VEH	28.49	25.98	37.50	28.74	36.52	0.02	16.89	36.07	25.83	36.03
		BIKE	0.08	N/A	0.00	N/A	N/A	N/A	0.00	0.00	N/A	0.21
		PED	1.62	N/A	0.36	N/A	0.15	N/A	0.22	2.50	0.97	0.07
	LIGA [16]	VEH	28.96	31.87	36.87	25.72	39.83	0.00	17.02	34.62	23.28	40.50
		BIKE	1.02	N/A	0.00	N/A	N/A	N/A	0.48	0.02	N/A	2.77
		PED	0.77	N/A	0.84	N/A	1.98	N/A	0.55	1.98	0.50	0.23
3D Range Sensor-based												
L	VoxelNeXt [8]	VEH	86.06	59.51	90.19	71.82	82.86	54.75	78.93	88.76	71.06	80.93
		BIKE	65.24	N/A	47.12	N/A	N/A	N/A	88.50	12.73	N/A	85.14
		PED	26.51	N/A	61.14	N/A	72.54	N/A	40.15	33.56	21.39	15.90
	HEDNet [47]	VEH	79.27	48.41	84.74	68.36	70.29	55.98	71.64	83.34	63.97	73.33
		BIKE	56.76	N/A	4.28	N/A	N/A	N/A	79.59	17.12	N/A	53.86
		PED	10.91	N/A	35.26	N/A	93.33	N/A	16.39	18.10	28.69	9.24
	SAFDNet [46]	VEH	79.30	43.83	82.82	57.33	65.07	49.30	66.28	81.62	58.38	76.19
		BIKE	58.68	N/A	0.81	N/A	N/A	N/A	78.77	8.24	N/A	65.23
		PED	9.78	N/A	33.30	N/A	72.63	N/A	14.90	12.07	25.78	6.75
4R	RTNH [30]	VEH	23.49	37.30	43.40	27.86	36.96	21.70	28.70	26.28	24.69	27.00
		BIKE	0.71	N/A	0.00	N/A	N/A	N/A	6.99	0.00	N/A	0.00
		PED	0.13	N/A	1.36	N/A	1.30	N/A	1.39	0.30	0.47	0.00
	VoxelNeXt [8]	VEH	25.03	44.03	48.78	27.91	37.42	32.79	31.82	24.02	32.50	35.03
		BIKE	0.24	N/A	0.25	N/A	N/A	N/A	0.03	1.35	N/A	0.00
		PED	0.16	N/A	4.85	N/A	7.93	N/A	0.25	3.71	0.32	0.00
	HEDNet [47]	VEH	24.10	43.51	41.16	28.57	31.28	25.67	28.92	22.28	37.01	30.82
		BIKE	0.29	N/A	0.04	N/A	N/A	N/A	0.40	0.02	N/A	0.71
		PED	0.02	N/A	0.62	N/A	0.23	N/A	0.17	0.09	0.13	0.00

Table 6. 3D object detection performance comparison on the DSERT-RoLL dataset for vehicle (IoU = 0.5), bike (IoU = 0.3), and pedestrian (IoU = 0.3) detection. R, E, T, 4R, and L represent the RGB, Event, and Thermal cameras, as well as the 4D Radar and LiDAR, respectively. In addition, VEH and PED denote the vehicle and pedestrian classes, respectively. N/A indicates that the corresponding class is not present under the given adverse weather condition.

Modality	Methods	Class	Weather Condition						Light Condition			
			Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
Multi-modal Fusion-based												
R+L	LoGoNet [20]	VEH	87.18	64.96	91.41	79.12	79.74	66.20	79.01	90.56	80.49	82.78
		BIKE	58.52	N/A	2.86	N/A	N/A	N/A	64.87	10.88	N/A	79.29
		PED	24.77	N/A	48.70	N/A	81.38	N/A	36.70	30.82	44.65	11.52
	BEVFusion [26]	VEH	85.20	62.40	90.91	73.30	75.22	57.61	76.86	87.55	78.07	78.90
		BIKE	52.80	N/A	0.00	N/A	N/A	N/A	72.22	6.10	N/A	67.95
		PED	8.99	N/A	32.72	N/A	74.84	N/A	17.90	12.75	32.00	14.72
	DeepFusion [21]	VEH	87.19	63.94	91.91	75.61	81.77	57.26	79.19	90.81	78.66	80.10
		BIKE	53.36	N/A	11.55	N/A	N/A	N/A	77.15	2.55	N/A	69.04
		PED	16.33	N/A	48.06	N/A	71.34	N/A	33.14	19.94	40.21	9.19
R+4R	HGSFusion [15]	VEH	25.74	46.49	49.62	28.49	37.87	34.02	32.66	24.31	34.47	35.96
		BIKE	0.20	N/A	0.39	N/A	N/A	N/A	0.00	1.50	N/A	0.00
		PED	0.19	N/A	5.28	N/A	8.44	N/A	0.31	4.10	0.35	0.00
4R+L	InterFusion [42]	VEH	84.52	66.94	94.31	76.56	74.13	64.82	79.31	87.49	75.55	79.95
		BIKE	32.03	N/A	5.70	N/A	N/A	N/A	79.43	10.01	N/A	16.93
		PED	7.45	N/A	19.74	N/A	69.53	N/A	24.20	6.80	18.24	6.16
	RL3DOD [5]	VEH	85.05	63.15	88.39	76.17	81.41	65.87	77.77	87.26	78.32	81.50
		BIKE	59.85	N/A	13.86	N/A	N/A	N/A	75.51	13.94	N/A	69.73
		PED	19.25	N/A	55.23	N/A	51.49	N/A	42.26	22.00	35.63	13.22
R+T+4R+L	SAMFusion [32]	VEH	87.03	65.13	91.69	78.02	79.81	70.59	80.54	89.93	80.16	82.50
		BIKE	40.05	N/A	2.70	N/A	N/A	N/A	72.95	2.78	N/A	41.32
		PED	21.09	N/A	43.60	N/A	64.34	N/A	38.37	17.49	41.13	11.51
R+E+T+4R+L	**Ours**	VEH	90.30	71.42	95.10	80.26	85.59	72.94	82.93	92.65	85.47	86.33
		BIKE	66.33	N/A	19.26	N/A	N/A	N/A	83.73	17.25	N/A	85.41
		PED	28.67	N/A	55.64	N/A	55.09	N/A	50.28	25.40	42.39	45.46

Table 7. 2D object detection performance comparison on the DSERT-RoLL dataset for vehicle (IoU = 0.5), bike (IoU = 0.3), and pedestrian (IoU = 0.3) detection. R, E, and, T represent the RGB, Event, and Thermal cameras respectively. In addition, VEH and PED denote the vehicle and pedestrian classes, respectively. N/A indicates that the corresponding class is not present under the given adverse weather condition.

Modality	Methods	Class	Weather Condition						Light Condition			
			Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
Multi-modal Fusion-based												
R	YOLOv10 [41]	VEH	76.47	72.99	84.95	76.68	58.76	2.84	71.98	67.69	76.25	76.15
		BIKE	43.13	N/A	24.74	N/A	N/A	N/A	73.50	10.32	N/A	32.76
		PED	12.36	N/A	19.97	N/A	0.691	N/A	15.59	6.59	23.22	1.98
	DEIM [17]	VEH	81.85	82.99	91.48	73.60	65.07	13.37	77.76	72.74	85.14	79.50
		BIKE	68.47	N/A	4.86	N/A	N/A	N/A	87.84	2.25	N/A	74.23
		PED	22.47	N/A	38.59	N/A	13.25	N/A	39.90	19.91	32.79	25.77
E	RT-DETR [48]	VEH	73.77	83.17	83.57	58.93	47.28	0.023	69.89	58.28	78.87	77.83
		BIKE	49.99	N/A	0.983	N/A	N/A	N/A	72.26	10.07	N/A	39.25
		PED	4.88	N/A	8.79	N/A	0.204	N/A	2.35	5.28	16.15	0.346
	DEIM [17]	VEH	65.56	85.67	80.77	64.36	50.00	0.075	69.31	53.94	57.20	69.38
		BIKE	63.30	N/A	1.30	N/A	N/A	N/A	82.91	25.08	N/A	53.44
		PED	10.52	N/A	2.90	N/A	0.05	N/A	4.44	5.19	22.27	0.837
T	YOLOv10 [41]	VEH	78.31	83.84	92.16	76.75	75.30	0.619	69.48	74.15	73.93	81.03
		BIKE	54.29	N/A	78.37	N/A	N/A	N/A	64.29	59.54	N/A	67.50
		PED	43.93	N/A	30.35	N/A	43.62	N/A	27.07	49.90	31.81	51.66
	DEIM [17]	VEH	81.84	85.56	83.21	77.75	77.04	0.576	66.07	76.69	84.91	86.19
		BIKE	72.01	N/A	47.17	N/A	N/A	N/A	89.77	70.90	N/A	58.30
		PED	44.38	N/A	27.77	N/A	29.43	N/A	20.66	49.47	23.78	70.42
R+E	GM-DETR [45]	VEH	84.24	87.54	95.07	80.92	59.44	15.62	83.32	73.61	86.04	81.90
		BIKE	78.68	N/A	1.94	N/A	N/A	N/A	93.07	26.03	N/A	72.02
		PED	26.75	N/A	29.22	N/A	9.11	N/A	23.54	27.99	5.35	44.16
R+T	GM-DETR [45]	VEH	84.10	86.64	92.18	77.99	79.44	1.48	71.87	77.41	86.12	88.70
		BIKE	75.33	N/A	38.63	N/A	N/A	N/A	97.67	38.49	N/A	63.12
		PED	42.37	N/A	39.68	N/A	43.92	N/A	38.36	44.38	36.67	76.19
E+T	GM-DETR [45]	VEH	85.44	92.13	88.35	79.64	81.19	11.20	71.00	78.96	87.74	93.04
		BIKE	76.45	N/A	81.47	N/A	N/A	N/A	95.92	73.82	N/A	63.94
		PED	52.42	N/A	51.46	N/A	48.40	N/A	43.64	60.65	34.90	71.11
R+E+T	GM-DETR [45]	VEH	90.36	93.66	96.28	82.29	81.60	16.56	82.07	82.60	94.93	93.52
		BIKE	82.05	N/A	7.63	N/A	N/A	N/A	97.87	9.88	N/A	86.40
		PED	76.90	N/A	61.98	N/A	53.41	N/A	64.07	71.53	72.60	84.96