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

Hoonhee Cho* Jae-Young Kang* Yuhwan Jeong* Yunseo Yang
Wonyoung Lee Youngho Kim Kuk-Jin Yoon

Visual Intelligence Lab, KAIST

* equal contribution

https://jeongyh98.github.io/dsert-roll

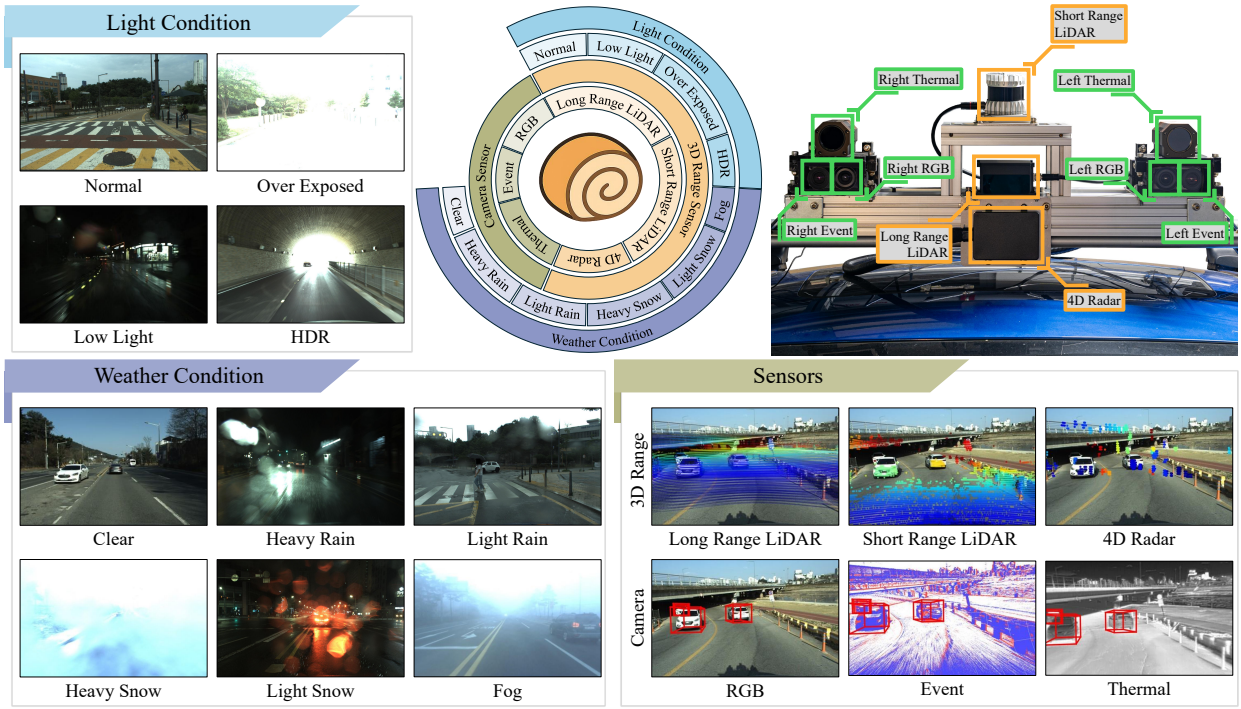

Figure 1. The proposed DSERT-RoLL dataset comprises stereo event, RGB, and thermal cameras, together with 4D radar and dual LiDAR, collected in on-road driving across a wide range of weather and illumination conditions, and provided with precise 3D annotations.

Abstract

In this paper, we present DSERT-RoLL, a driving dataset that incorporates stereo event, RGB, and thermal cameras together with 4D radar and dual LiDAR, collected across diverse weather and illumination conditions. The dataset provides precise 2D and 3D bounding boxes with track IDs and ego vehicle odometry, enabling fair comparisons within and across sensor combinations. It is designed to alleviate data scarcity for novel sensors such as event cameras and 4D radar and to support systematic studies of their behavior. We establish unified 3D and 2D benchmarks that enable direct comparison of characteristics and strengths across sensor families and within each family. We report baselines for representative single modality and multimodal methods and provide protocols that encourage research on different fusion strategies and sensor combinations. In addition, we propose a fusion framework that integrates sensor specific cues into a unified feature space and improves 3D detection robustness under varied weather and lighting.

1. Introduction

Perception underpins autonomous driving, enabling safe decision-making and control. Early unimodal systems [61, 135] suffered from limited accuracy and incomplete scene

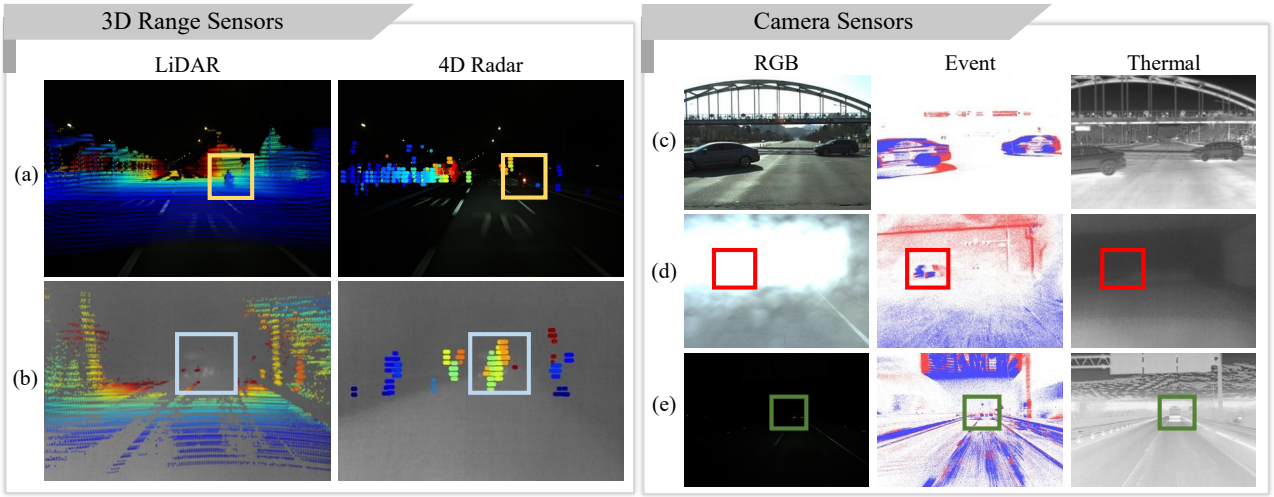

Figure 2. Complementary scenarios across sensor families. (a–b) 3D range sensors: (a) LiDAR-dominant, effective in clear conditions and at long range with accurate geometry; (b) 4D radar-dominant, reliable in adverse weather (*e.g.*, fog and snow) using Doppler. (c–e) Camera-based sensors: (c) RGB-dominant, strong in daylight and textured scenes; (d) Event-dominant, responsive to small and rapid motions and robust in high dynamic range; (e) Thermal-dominant, informative at night or in low light. Together, (a)–(e) illustrate complementary strengths across sensor types.

coverage, motivating multimodal approaches. Fusing cameras and LiDAR [75, 119] has become standard, combining image semantics with point-cloud geometry to boost robustness and accuracy. Nonetheless, extending these gains to varied weather and lighting remains challenging.

Conventional RGB cameras are widely used in autonomous driving due to their rich semantic information [52, 71, 87, 101, 115, 117]. However, they are sensitive to illumination changes and often degrade under high dynamic range or low light. To mitigate these issues, two alternative sensing modalities have gained attention:

• **Thermal Cameras** operate in the infrared spectrum and are effective in night-time environments where RGB cameras often struggle. They provide complementary cues beyond visible light, thereby enhancing perception robustness under challenging conditions [12, 59, 97].

• **Event Cameras** [34] asynchronously capture changes in brightness at the pixel level, enabling high temporal resolution and low-latency perception. Unlike frame-based cameras, they naturally handle high dynamic range scenes and fast motions, providing complementary information that enhances robustness in challenging environments.

From the perspective of 3D sensing, LiDAR operates independently of illumination but remains sensitive to weather conditions [2, 42, 68], where sensing range may decrease, and measurements can become noisy. To mitigate these issues, alternative sensors have been explored.

• **4D Radar** is a 3D range sensor that emits radio waves and can acquire data even through adverse conditions such as heavy rain or snow. In contrast to LiDAR, which often suffers from reduced range and noise in adverse weather, 4D radar offers more stable perception [83, 92], serving as a useful complement for all-weather operation.

Research leveraging novel sensors, including event cameras and 4D radar, and, to a lesser extent, thermal cameras, for perception under adverse conditions has accelerated in recent years. However, existing benchmarks [24, 35, 83] that include these sensors are typically modality-specific and primarily compare against conventional RGB cameras and LiDAR, leaving direct cross-sensor comparisons and systematic studies of their fusion relatively underexplored.

To advance research in this direction, we present the **DSERT-RoLL** dataset: **D**riving with **S**tereo **E**vent-**R**GB-**T**hermal Cameras, 4D **R**adar, and Dua**L** **L**iDAR/ As shown in Fig. 1, we collected multi-modal data across diverse driving scenarios, including night, high dynamic range (HDR), rain, snow, fog, and other challenging conditions. Through DSERT-RoLL, we provide a comprehensive multi-modal benchmark that supports robust perception and fusion studies under diverse driving conditions. Our dataset and approach differ from existing works in the following key aspects:

- While recent datasets have analyzed the advantages of novel sensors, the evaluation of various sensors in the same environment remains largely unexplored. In contrast, the DSERT-RoLL dataset includes widely researched emerging sensors and data collected from extreme environments. By providing a fair benchmark for training and evaluation across multiple sensors in the same setting, DSERT-RoLL enables a deeper analysis of each sensor's strengths and characteristics.
- Although there are numerous benchmarks and datasets based on widely used sensors like frame cameras and LiDAR, benchmarks based on emerging sensors, such as event cameras, thermal cameras, and 4D radar, are relatively scarce. The proposed DSERT-RoLL dataset con-

Table 1. Comparison of object detection datasets in autonomous driving. Upper rows use conventional sensors; lower rows include novel sensors. The symbol △ marks annotations not officially provided but added by other authors. If 3D boxes exist, 'Num Data' counts samples with 3D boxes; otherwise, it counts all samples. Additional comparisons with more datasets are provided in the *supple*.

Dataset	Num	Adverse Weather				3D Range Sensor		Camera Sensor			Ground-truth		
	Data	Clear	Rain	Fog	Snow	LiDAR	Radar	RGB	Event	Thermal	3D Bbox.	Tr. ID	Odom
KITTI [37]	15k	✓	✗	✗	✗	✓	✗	Stereo	✗	✗	✓	✓	✓
Waymo [98]	230k	✓	✓	✗	✗	✓	✗	Multi-view	✗	✗	✓	✓	✗
NuScenes [7]	40k	✓	✓	✗	✗	✓	3D	Multi-view	✗	✓	✓	✓	✓
Argoverse 2 [110]	150k	✓	✓	✗	✓	✓	✗	Multi-view	✗	✗	✓	✓	✓
K-Radar [83]	35k	✓	✓	✓	✓	✓	4D	Multi-view	✗	✗	✓	✓	✓
TJ4DRadSet [131]	7.8k	✓	✗	✗	✗	✓	✓	Mono	✗	✗	✓	✓	✓
DSEC [35]	5.4k	✓	✗	✗	✗	✓	✗	Stereo	Stereo	✗	△	△	✓
1Mpx [89]	32M	✓	✓	✗	✗	✗	✗	Mono	Mono	✗	✗	✗	✗
SeeingThroughFog [4]	13.5k	✓	✓	✓	✓	✓	3D	Stereo	✗	Mono	✓	✗	✗
KAIST [24]	8.9k	✓	✗	✗	✗	✓	✗	Stereo	✗	Mono	✗	✗	✗
DSERT-RoLL (Ours)	22k	✓	✓	✓	✓	✓	4D	Stereo	Stereo	Stereo	✓	✓	✓

tributes to enhancing data richness and is expected to support a wide range of studies, including domain adaptation, domain generalization, and more.

- While differences between camera sensors and 3D range sensors are well documented, Fig. 2 shows that complementary strengths also exist within each sensor type, for example among cameras such as RGB, event, and thermal, and among 3D range sensors such as LiDAR and 4D radar. Each type excels under different failure modes, which suggests synergy rather than simple substitution. Building on this, we propose a framework that leverages these complementary strengths to achieve robust 3D object detection under varying weather and illumination. By integrating cues into a unified feature space, our method improves perception reliability across conditions.

2. Related Works

Traditional Driving Datasets. Numerous datasets [7, 25, 37, 98, 110] have demonstrated the potential for safe autonomous driving perception [11, 66, 111, 125] through the use of RGB and LiDAR-based data [78, 90, 113], human annotations, and extensive training with large-scale datasets. Subsequent research [29, 91] has incorporated more diverse driving scenes under various conditions, and by expanding the scale of data, it has enabled the development of more robust perception algorithms. However, due to the inherent limitations of RGB and LiDAR sensors, their robustness in extreme environments (*e.g.*, fog, snow) remains insufficient. Consequently, research in this area is advancing with the emergence of novel sensors [4, 85, 128, 132], fostering additional fusion studies.

Thermal Camera-based Driving Datasets. Thermal imaging captures emitted infrared radiation rather than reflected visible light, providing illumination-invariant cues that complement RGB, especially in darkness and adverse weather. This is particularly valuable for object detection in road scenes, where nighttime and inclement conditions demand robust perception across diverse environments. Recent RGB–thermal benchmarks [24, 38, 48, 96, 97, 102] have catalyzed research on multimodal detection, segmentation [41, 60], and tracking [3, 31] under challenging illumination and weather [4, 73, 100]. Alongside dataset growth, fusion methodologies have matured from early feature concatenation to more principled designs [28, 30, 49, 62, 82, 104, 114, 133].

Event Camera-based Driving Datasets. Event cameras asynchronously report per-pixel brightness changes with microsecond latency and extremely high dynamic range, producing motion-blur-free signals [21, 58] that complement frame-based RGB and LiDAR under fast ego-motion [6, 43, 79, 81], low light [22, 57, 63, 72, 112, 118], and glare [80, 138]. These properties make them well suited for autonomous-driving perception in road scenes, where rapid maneuvers and abrupt illumination transitions demand temporally precise, HDR sensing. Recent event-driven driving benchmarks with single [5, 26] or stereo sensors [23, 35, 88, 136], aligned RGB cameras, and vehicle telemetry have supported progress across a broad range of perception tasks [10, 19, 20, 23, 36, 50, 54, 55, 89, 93, 99, 108]. Despite these advantages, multimodal fusion beyond fusion with RGB [17, 18, 129] remains underexplored in event-based vision, and annotations for 3D perception are still scarce. By providing additional modalities alongside events and 3D annotations, this work serves as a strong foundation for subsequent research on event cameras.

4D Radar-based Driving Datasets. Automotive 4D radar measures range, azimuth, elevation, and radial velocity (Doppler), delivering long-range, illumination- and weather-robust cues that complement RGB and LiDAR with direct motion observables [33]. These properties are attractive for perception [83, 116] in adverse conditions, such as fog, rain, snow, and nighttime, where temporally stable velocity estimates and extended detection ranges are critical. Recent 4D radar driving datasets [1, 46, 84, 94, 122, 122, 131] provide synchronized radar point clouds together with LiDAR and cameras, and have demonstrated success-

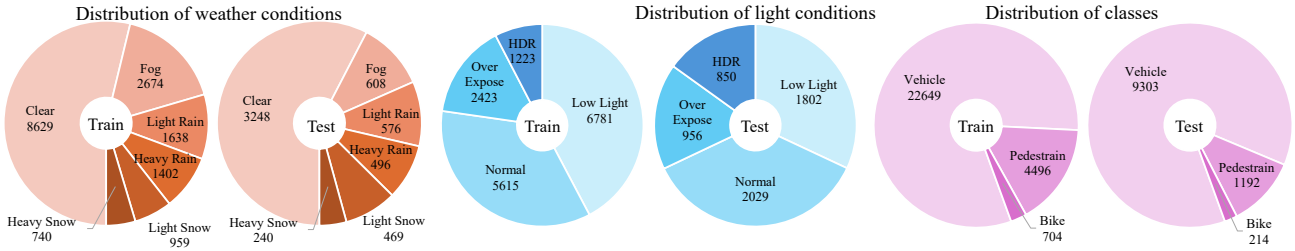

Figure 3. Distribution of training and testing data with respect to weather conditions, lighting conditions, and object classes.

Table 2. Sensor suit details.

Sensors	Model Name	Resolution	FoV	FPS
RGB	2 × BFS-U3-51S5C	2448 × 2048	82.2° × 66.5°	10
Event	2 × Prophesee EVK4	1280 × 720	76.7° × 65.5 °	>10k
Thermal	2 × FLIR A65	640 × 512	90° × 69°	30
Sensors	Model Name	Max. Range	FoV	FPS
4D Radar	RETINA-4FN	100m	100° × 24°	20
Long-range LiDAR	Livox HAP	150m	120° × 25°	10
Short-range LiDAR	os0-128	100m	360° × 90°	20
GPS/IMU	Microstrain 3DM-GX5-45	N/A	N/A	10/100

ful perception under these adverse conditions. As LiDAR datasets have grown, a wide range of sensor combination studies have emerged. By contrast, 4D radar is relatively recent and has primarily been paired with cameras or LiDAR. Our work aims to provide a foundation for exploring richer 4D radar-based multimodal configurations, including event cameras and thermal cameras, enabling broader research on these combinations.

Multi-modal 3D Object Detection. 3D object detection [51, 69, 86, 106] aims to estimate 3D bounding boxes and object orientations. Unimodal LiDAR approaches [16, 53, 67, 76, 77, 124, 126] leverage the depth accuracy of point clouds to regress 3D boxes. Recently, multimodal fusion has been actively explored to exploit the complementary strengths of different sensors under diverse conditions. The most common setup fuses RGB images with LiDAR [13, 32, 109, 120]. This pairing adds color and texture to precise depth and improves small object recall and 3D localization. To improve robustness in adverse weather and low light, radar [47, 70, 134] has also been incorporated. With the advent of a 4D radar that provides range, Doppler, azimuth, and elevation information, camera and radar fusion [8, 9] has advanced further. Moreover, recent architectures [13, 85] enable fusion of two or more modalities within a unified framework.

3. DSERT-RoLL Dataset

3.1. Sensor Configuration

As illustrated in Fig. 1, we equipped the vehicle with the multi-modal sensor setting described in Table 2. We first mounted LiDAR sensors, which are widely adopted 3D range sensors for object detection, including a long-range LiDAR and a high-resolution short-range LiDAR. The long-range LiDAR is used to obtain reliable object annotations over extended distances, whereas the short-range LiDAR provides high-resolution fine-grained point measurements with an enlarged vertical field of view, thereby offering dense geometric coverage. To further ensure perception performance under extreme weather such as fog or snow, we additionally equipped the vehicle with a 4D Radar. All cameras were deployed in stereo configurations to fully cover the frontal field of view, and we extended the setup beyond RGB sensors by incorporating a thermal camera and an event camera. These complementary modalities enhance robustness against challenging lighting conditions and motion blur, providing reliable perception in adverse environments. Finally, a GPS antenna and IMU sensors were mounted near the camera suite on the vehicle to enable precise localization of the ego-vehicle.

3.2. Data Distribution

The DSERT-RoLL dataset encompasses diverse driving scenarios, including highways, urban streets, suburban roads, and narrow alleys. In total, the data collection process results in 22K frames of multi-modal sensor measurements captured under various environmental conditions.

The DSERT-RoLL dataset is categorized into six weather conditions: clear, fog, light rain, heavy rain, light snow, and heavy snow. This design allows a fair comparison of the strengths of different sensor modalities under diverse weather conditions. A key advantage of DSERT-RoLL is that multiple modalities are captured simultaneously in the same environment, enabling deeper research on multi-sensor fusion, particularly for 3D range sensors.

While 3D range sensors (*e.g.*, LiDAR and Radar) are largely unaffected by illumination, camera-based perception can vary significantly depending on lighting. To facilitate fusion research in such scenarios, we further categorize the data into four lighting conditions: normal, low light, overexposed, and HDR.

The dataset defines three object categories, namely *vehicle*, *pedestrian*, and *bike*, which represent the most common types of driving datasets. We split the dataset into training and test sets with a 7:3 ratio. As shown in Fig. 3, the distributions of weather, lighting, and object classes are balanced

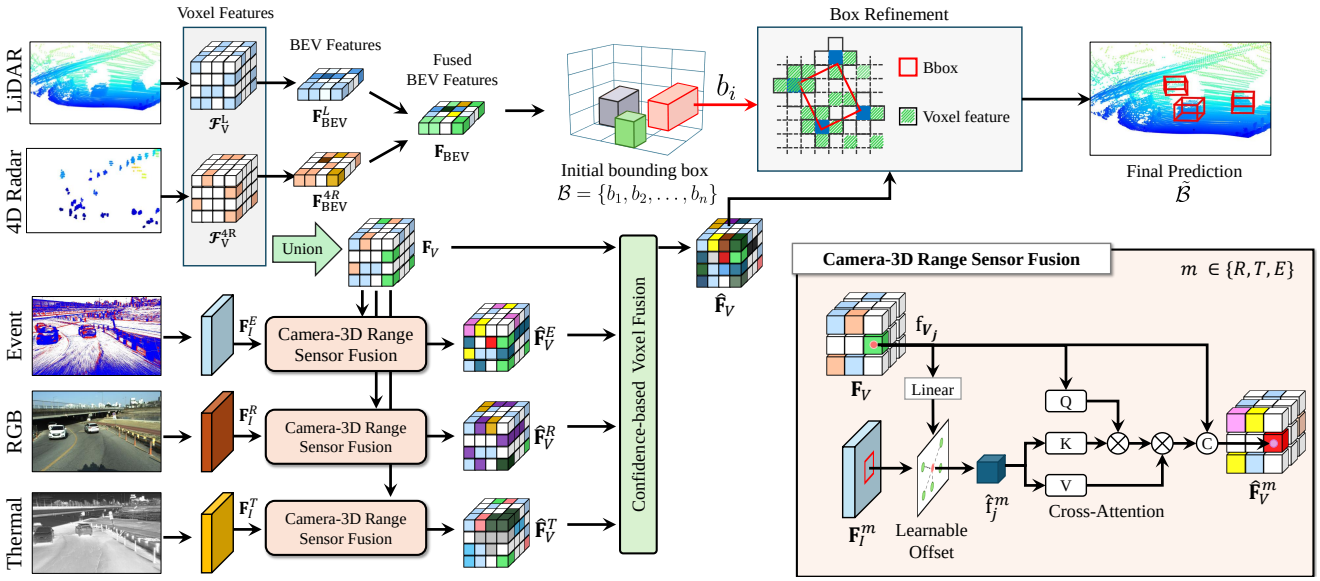

Figure 4. Overview of the proposed multi-modal 3D detection framework. LiDAR and 4D Radar features are voxelized and fused to generate initial 3D box proposals. RGB, thermal, and event features are then projected into 3D space via voxel-centric sampling and integrated through confidence-based fusion. The refined fused features are used for final bounding box prediction.

across both splits. This balanced coverage across diverse scenarios and environmental conditions makes DSERT-RoLL a reliable benchmark for evaluating the robustness of perception models in challenging real-world settings.

4. Multi-modal Approach on 3D Detection

To demonstrate the benefits of leveraging multi-modality, we propose a method that effectively fuses diverse sensor modalities into a unified feature space. This in-depth methodology is made possible by the strength of our DSERT-RoLL benchmark dataset, which incorporates multiple modalities within a single collection. In this section, we elaborate on the design and implementation of the proposed approach and highlight how multi-modal fusion contributes to robust perception performance.

Overall Framework. As illustrated in Fig. 4, the inputs to our framework consist of 3D range sensors and multi-modal single-view images. For the 3D range sensor, we use a LiDAR (L) and a 4D Radar ($4R$) providing point sets $\mathcal{P}^L = \{(x_i, y_i, z_i)|f_i^L\}_{i=1}^{N_L}$ and $\mathcal{P}^{4R} = \{(x_j, y_j, z_j)|f_j^{4R}\}_{j=1}^{N_{4R}}$. Here, (x, y, z) denotes the 3D spatial coordinate of each point, and $f \in \mathbb{R}^{C_p}$ represents point-wise features such as intensity (LiDAR) or Doppler velocity (4D Radar). N_L and N_{4R} denote the number of points from LiDAR and 4D Radar, respectively. Each point cloud is processed by a 3D voxel-based backbone [135] to obtain voxel features $\mathcal{F}_V^L$ and $\mathcal{F}_V^{4R}$, with $\mathcal{F}_V \in \mathbb{R}^{X \times Y \times Z \times C_V}$, where C_V denotes the number of channels and (X, Y, Z) represents the voxel grid size. For the camera sensors, we incorporate three modalities: an RGB image $\mathbf{I}^R \in \mathbb{R}^{H \times W \times 3}$, a thermal image $\mathbf{I}^T \in \mathbb{R}^{H \times W \times 1}$, and a voxel grid [137] of events $\mathbf{I}^E \in \mathbb{R}^{H \times W \times 5}$, which are processed by a 2D backbone [74] to extract features $\mathbf{F}_I^R, \mathbf{F}_I^T, \mathbf{F}_I^E \in \mathbb{R}^{H/4 \times W/4 \times C_I}$, where C_I denotes the number of channels for RGB, thermal, and event features.

We first generate initial 3D bounding boxes from the range sensors and then refine them by incorporating complementary cues from the camera sensors, where confidence is taken into account during the fusion. Finally, all heterogeneous inputs are processed through our proposed multi-modal fusion strategy to produce unified representations for 3D object detection.

Initial 3D Box Proposal from Range Sensors. To perform effective and efficient computation in 3D space, the voxel features $\mathcal{F}_V^L$ and $\mathcal{F}_V^{4R}$ are collapsed along the vertical axis and transformed by 2D convolutions on the ground plane into bird's-eye-view (BEV) features $\mathbf{F}_{\text{BEV}}^L, \mathbf{F}_{\text{BEV}}^{4R} \in \mathbb{R}^{\frac{X}{s} \times \frac{Y}{s} \times C_B}$, where X and Y are the voxel grid dimensions in the horizontal plane, s is the stride, and C_B is the channel dimension. Given the BEV features $\mathbf{F}_{\text{BEV}}^L$ and $\mathbf{F}_{\text{BEV}}^{4R}$, we concatenate the two modality features along the channel dimension and apply a convolutional layer to fuse them. This simple yet effective fusion produces cross-modally enriched BEV representations while maintaining computational efficiency. Through the detector [121], we obtain the initial set of bounding boxes $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$, where n is a pre-defined box number.

Camera-3D Range Sensor Fusion. We aim to leverage the camera features, which contain multiple strengths and rich semantic information, to gain additional performance. To integrate image features into the 3D space in a multi-

Table 3. Ablation study across sensor modalities on the DSERT-RoLL dataset for 3D detection. For the modalities, we use the following notation: R: RGB, E: Event, T: Thermal, 4R: 4D Radar, and L: LiDAR.

Modality	Weather Condition						Light Condition			
	Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
L	82.90	65.67	89.62	62.97	77.26	54.14	74.71	86.10	75.82	74.51
R+L	84.67	66.14	90.29	72.82	78.40	59.43	76.26	87.41	77.55	79.31
4R+L	88.26	67.41	91.43	67.41	77.43	69.96	79.43	88.73	82.85	82.98
R+4R+L	88.35	67.38	91.79	79.03	84.39	70.26	81.31	91.04	80.34	83.93
R+E+4R+L	88.70	**71.45**	92.94	80.11	82.75	71.64	81.92	91.43	83.37	**86.55**
R+T+4R+L	89.48	71.00	93.94	79.77	84.02	71.32	82.26	92.20	83.20	85.66
R+E+T+4R+L	**90.30**	71.42	**95.10**	**80.26**	**85.59**	**72.94**	**82.93**	**92.65**	**85.47**	86.33

modal manner, we propose a voxel-centric sampling strategy. Specifically we extract non-empty voxel indices from the LiDAR and 4D Radar voxel features $\mathcal{F}_V^L$ and $\mathcal{F}_V^{4R}$ to obtain $\mathbf{F}_V^L$ and $\mathbf{F}_V^{4R}$. We then combine the features on the union of these indices to form a unified sparse voxel feature space. We denote the resulting set as

$$\mathbf{F}_V \;=\; \mathbf{F}_V^L \cup \mathbf{F}_V^{4R} \;=\; \{(V_j, \mathbf{f}_{V_j})\}_{j=1}^{N_V}, \tag{1}$$

where V_j is the voxel index, $\mathbf{f}_{V_j} \in \mathbb{R}^{C_V}$ is the fused feature of a non-empty voxel, and N_V is the number of non-empty voxels. Let Ω^L and Ω^{4R} denote the sets of non-empty voxels for the LiDAR and 4D Radar, respectively. For each V_j in the union $\Omega = \Omega^L \cup \Omega^{4R}$, the fused feature is defined by

$$\mathbf{f}_{V_j} = \begin{cases} \mathbf{f}_{V_j}^L & \text{if } V_j \in \Omega^L \setminus \Omega^{4R}, \\ \mathbf{f}_{V_j}^{4R} & \text{if } V_j \in \Omega^{4R} \setminus \Omega^L, \\ \mathbf{P}\big[\mathbf{f}_{V_j}^L \mid \mathbf{f}_{V_j}^{4R}\big] & \text{if } V_j \in \Omega^L \cap \Omega^{4R}, \end{cases} \tag{2}$$

where $[\cdot \mid \cdot]$ denotes channel-wise concatenation and $\mathbf{P}$ is a per-scale linear projector that maps the concatenated $2C_V$ channels back to C_V. Thus, voxels with neither modality remain absent, voxels with exactly one modality keep that feature as-is, and voxels with both modalities are concatenated and projected to preserve dimensionality while enabling cross-modal fusion.

For each non-empty voxel V_j, we obtain modality-specific projections onto the image planes of the RGB, thermal, and event cameras as $u_j^R = \mathbf{M}^R \cdot V_j, \quad u_j^T = \mathbf{M}^T \cdot V_j, \quad u_j^E = \mathbf{M}^E \cdot V_j$, where $\mathbf{M}^R$, $\mathbf{M}^T$, and $\mathbf{M}^E$ are the projection matrices, which are the products of the intrinsic and extrinsic matrices for each modality. The projected locations u_j^R, u_j^T, u_j^E are used to sample nearby image features from $\mathbf{F}_I^R$, $\mathbf{F}_I^T$, and $\mathbf{F}_I^E$, respectively. Given the modality-specific projections u_j^R, u_j^T, u_j^E, aggregated image features are obtained by sampling feature values in the neighborhood of each projection. For modality $m \in R, T, E$, the sampling process for the number of sampled point, Q, yields the following aggregated features.

$$\hat{\mathbf{f}}_j^m = \sum_{q=1}^{Q} w_q \cdot \mathbf{F}_I^m(u_j^m + \Delta u_j^{m,q}), \tag{3}$$

where $\Delta u_j^{m,q}$ and w_q are the learnable offset and aggregation weight for the q-th sampling point. Both the offsets $\Delta u_j^{m,q}$ and the weights w_q are predicted from the voxel feature, $\mathbf{f}_{V_j}$. Each voxel feature $\mathbf{f}_{V_j}$ is treated as the query, while the aggregated image features $\hat{\mathbf{f}}_j^m$ serve as keys and values. The deformable cross-attention [103] for voxel V_j is formulated as $\hat{\mathbf{f}}_{V_j}^m = \text{Attn}(\mathbf{Q} = \mathbf{f}_{V_j},\ \mathbf{K} = \hat{\mathbf{f}}_j^m,\ \mathbf{V} = \hat{\mathbf{f}}_j^m)$.

Confidence-based Voxel Fusion. We concatenate the image-enhanced voxel features from all modalities ($m \in \{R, T, E\}$) to form

$$\hat{\mathbf{F}}_V^{\text{cam}} = \big[\, \hat{\mathbf{F}}_V^R \mid \hat{\mathbf{F}}_V^T \mid \hat{\mathbf{F}}_V^E \,\big] \in \mathbb{R}^{N_V \times (KC_V)}, \quad K{=}3, \tag{4}$$

where $\hat{\mathbf{F}}_V^R$, $\hat{\mathbf{F}}_V^T$, and $\hat{\mathbf{F}}_V^E$ denote the RGB, thermal, and event branches, respectively. For camera-axis attention, we view $\hat{\mathbf{F}}_V^{\text{cam}}$ as a 3D tensor, $\mathbb{R}^{N_V \times K \times C_V}$. We compute the camera-wise gates using a global summary as

$$\mathbf{w} = \sigma\left(\frac{1}{N_V C_V} \sum_{i=1}^{N_V} \sum_{c=1}^{C_V} \hat{\mathbf{F}}_V^{\text{cam}}(i, :, c) \right) \in [0,1]^{1 \times K \times 1}, \tag{5}$$

where σ denotes the sigmoid activation. Each camera branch is reweighted by its corresponding scalar gate as:

$$\bar{\mathbf{F}}_V^{cam} = \mathbf{w} \odot \hat{\mathbf{F}}_V^{cam} \in \mathbb{R}^{N_V \times K \times C_V},$$

where $\odot$ denotes element-wise multiplication. Finally, we concatenate the image-enhanced voxel features with the original voxel features. A feed-forward network is then applied to reduce the channel dimension and yield the final fused features $\tilde{\mathbf{F}}_V \in \mathbb{R}^{N_V \times C_V}$.

Bounding Box Refinement. Given the initial bounding box proposals $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$ from the range sensors, we further perform refinement using the final fused voxel features $\tilde{\mathbf{F}}_V \in \mathbb{R}^{N_V \times C_V}$. For each proposal b_i, we divide the 3D region into $S \times S \times S$ regular sub-voxels and apply ROI pooling [27, 44] to extract proposal-aligned features from both the fused image-enhanced voxel features $\tilde{\mathbf{F}}_V$ and the original voxel features. This produces grid features $\tilde{\mathbf{F}}_V^i \in \mathbb{R}^{S^3 \times C_V}$ for each initial bounding box, which are then passed through a multi-layer perceptron (MLP) to estimate the refined boxes, $\tilde{\mathcal{B}} = \{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n\}$.

Table 4. 3D object detection performance comparison on the DSERT-RoLL dataset. We categorize the methods into three groups: stereo-based, 3D range sensor–based, and multi-modal fusion-based approaches. For the modalities, we use the following notation: R: RGB, E: Event, T: Thermal, 4R: 4D Radar, and L: LiDAR.

Modality	Methods	Weather Condition						Light Condition			
		Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
Stereo-based											
R	DSGN [14]	31.08	43.66	42.48	20.51	25.94	0.01	29.99	25.68	22.55	40.69
R	LIGA [40]	35.52	41.67	37.52	20.57	26.02	0.00	31.31	30.06	22.44	42.80
E	DSGN [14]	24.23	22.06	26.93	31.38	23.12	0.01	21.41	21.42	15.58	36.44
E	LIGA [40]	27.11	22.53	23.43	22.84	24.61	0.00	23.10	23.20	15.30	34.92
T	DSGN [14]	28.49	25.98	37.50	28.74	36.52	0.02	16.89	36.07	25.83	36.03
T	LIGA [40]	28.96	31.87	36.87	25.72	39.83	0.00	17.02	34.62	23.28	40.50
3D Range Sensor-based											
L	VoxelNeXt [15]	86.06	59.51	90.19	71.82	82.86	54.75	78.93	88.76	71.06	80.93
L	HEDNet [127]	79.27	48.41	84.74	68.36	70.29	55.98	71.64	83.34	63.97	73.33
L	SAFDNet [123]	79.30	43.83	82.82	57.33	65.07	49.30	66.28	81.62	58.38	76.19
4R	RTNH [83]	23.49	37.30	43.40	27.86	36.96	21.70	28.70	26.28	24.69	27.00
4R	VoxelNeXt [15]	25.03	44.03	48.78	27.91	37.42	32.79	31.82	24.02	32.50	35.03
4R	HEDNet [127]	24.10	43.51	41.16	28.57	31.28	25.67	28.92	22.28	37.01	30.82
Multi-modal Fusion-based											
R+L	LoGoNet [64]	87.18	64.96	91.41	79.12	79.74	66.20	79.01	90.56	80.49	82.78
R+L	BEVFusion [75]	85.20	62.40	90.91	73.30	75.22	57.61	76.86	87.55	78.07	78.90
R+L	DeepFusion [65]	87.19	63.94	91.91	75.61	81.77	57.26	79.19	90.81	78.66	80.10
R+4R	HGSFusion [39]	25.74	46.49	49.62	28.49	37.87	34.02	32.66	24.31	34.47	35.96
4R+L	InterFusion [107]	84.52	66.94	94.31	76.56	74.13	64.82	79.31	87.49	75.55	79.95
4R+L	RL3DOD [8]	85.05	63.15	88.39	76.17	81.41	65.87	77.77	87.26	78.32	81.50
R+T+4R+L	SAMFusion [85]	87.03	65.13	91.69	78.02	79.81	70.59	80.54	89.93	80.16	82.50
R+E+T+4R+L	**Ours**	**90.30**	**71.42**	**95.10**	**80.26**	**85.59**	**72.94**	**82.93**	**92.65**	**85.47**	**86.33**

Loss Functions. We train the entire framework in an end-to-end manner. The overall loss consists of three terms: the RPN loss [27, 44] $\mathcal{L}_{\text{RPN}}$, the confidence prediction loss [27] $\mathcal{L}_{\text{conf}}$, and the box regression loss [27, 95] $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L} = \mathcal{L}_{\text{RPN}} + \lambda_1 \mathcal{L}_{\text{conf}} + \lambda_2 \mathcal{L}_{\text{reg}}. \tag{6}$$

5. Experiments on Multi-modal Approach

5.1. Experimental Settings

We train the entire framework in an end-to-end manner using four NVIDIA Quadro RTX 8000 GPUs. The loss weights for both λ_1 and λ_2 are set to 1. For evaluation, to align with the front camera's field of view, we constrain the point cloud along the X-axis to the range [0, 70] meters. We set the sampled points K as 4 for feature aggregation in the camera-3D range sensor fusion module. Following prior work [27, 44], the bounding-box refinement grid size, S, is set to 6. We evaluate all models using the official Waymo Open Dataset metrics [98]. We report Average Precision (AP) with a 3D IoU threshold of 0.5. Following prior work [4, 56, 83] under the standard challenge conditions for 3D perception, our main tables emphasize the *vehicle* class. The results for additional classes are provided in the supplementary material. We use a long-range LiDAR for the LiDAR modality, and for all camera modalities, we use only the left camera from the stereo setup.

5.2. 3D Object Detection Results Across Modalities

The proposed method operates adaptively across diverse modality combinations, providing a framework that highlights the strengths of multi-modal fusion. To study these effects, we conduct ablations over different sensor combinations and report the results in Table 3.

We begin with the most fundamental 3D range sensor, LiDAR, which serves as the base for initial bounding box estimation. From this foundation, we incrementally incorporate additional modalities to evaluate how each sensor contributes to detection robustness and accuracy under various environmental conditions. Adding RGB introduces richer semantics and contextual cues, improving category discrimination and boundary localization under moderate conditions. However, its impact is limited in extreme lighting and weathers, so the overall gains remain modest in the most challenging scenarios. Introducing 4D Radar further enhances spatial consistency and stability, especially in adverse weather (*e.g.* heavy snow) where LiDAR signals may degrade. The fusion of LiDAR and 4D Radar yields noticeable gain across most conditions, confirming the complementary nature of their geometric cues. When event and thermal modalities are integrated, the model becomes more

Table 5. 2D object detection performance on the DSERT-RoLL dataset, focusing on camera-based methods. For the modalities, we use the following notation: R: RGB, E: Event, and T: Thermal

Modality	Methods	Weather Condition						Light Condition			
		Clear	Fog	Light Rain	Heavy Rain	Light Snow	Heavy Snow	Normal	Low Light	Over Expose	HDR
R	YOLOv10 [105]	76.47	72.99	84.95	76.68	58.76	2.84	71.98	67.69	76.25	76.15
R	DEIM [45]	81.85	82.99	91.48	73.60	65.07	13.37	77.76	72.74	85.14	79.50
E	RT-DETR [130]	73.77	83.17	83.57	58.93	47.28	0.023	69.89	58.28	78.87	77.83
E	DEIM [45]	65.56	85.67	80.77	64.36	50.00	0.075	69.31	53.94	57.20	69.38
T	YOLOv10 [105]	78.31	83.84	92.16	76.75	75.30	0.619	69.48	74.15	73.93	81.03
T	DEIM [45]	81.84	85.56	83.21	77.75	77.04	0.576	66.07	76.69	84.91	86.19
R+E	GM-DETR [114]	84.24	87.54	95.07	80.92	59.44	15.62	**83.32**	73.61	86.04	81.90
R+T	GM-DETR [114]	84.10	86.64	92.18	77.99	79.44	1.48	71.87	77.41	86.12	88.70
T+E	GM-DETR [114]	85.44	92.13	88.35	79.64	81.19	11.20	71.00	78.96	87.74	93.04
R+T+E	GM-DETR [114]	**90.36**	**93.66**	**96.28**	**82.29**	**81.60**	**16.56**	82.07	**82.60**	**94.93**	**93.52**

resilient to dynamic illumination changes and low-visibility environments. In particular, R+E+T+4R+L, which leverages all modalities, achieves the highest performance overall, demonstrating the framework's ability to adaptively fuse heterogeneous inputs and fully exploit the advantages of multi-modal perception.

6. Benchmarks on the DSERT-RoLL Dataset

DSERT-RoLL enables fair, like-for-like evaluation across multiple modalities on the same scenes. With both 3D and 2D manual annotations and diverse weather and lighting conditions, it offers a rigorous testbed for robustness and generalization. Accordingly, we establish benchmarks for both 3D Object Detection and 2D Object Detection on DSERT-RoLL dataset.

6.1. 3D Object Detection Benchmark Results

The selected 3D detector models, organized by sensor-configuration groups, are categorized into three types: stereo-based, 3D range sensor–based, and multi-modal fusion–based approaches.

Stereo camera-based methods. We select two existing methods, LIGA [40] and DSGN [14]. To isolate the effect of the sensor type, we keep the architectures and training settings identical and simply replace the camera input with event or thermal data.

3D range sensor-based methods. We select 3D range sensor–based methods, VoxelNeXt [15], HEDNet [127], and SAFDNet [123], for our evaluation, and additionally adopted RTNH [83] for the 4D Radar modality. Since the LiDAR pipeline naturally extends to 4D Radar, these methods enable fair and consistent evaluation across 3D range-sensor modalities.

Multi-modal fusion methods. We include the following multi-modal methods as comparisons: LoGoNet [64], BEV-Fusion [75], DeepFusion [65], HGSFusion [39], InterFusion [107], RL3DOD [8], and SAMFusion [85].

As shown in Table 4, we observe that limited information settings, specifically stereo without explicit 3D depth and 4D radar only, tend to underperform, whereas LiDAR is generally strong thanks to accurate geometric cues. However, LiDAR performance drops in adverse weather such as fog and snow. In contrast, multi-modal methods compensate: cameras provide missing semantic detail and 4D radar adds weather robustness. Our approach adaptively fuses all available sensor types, delivering consistently strong results across weather and illumination conditions.

6.2. 2D Object Detection Benchmark Results

Although this paper focuses on 3D detection, DSERT-RoLL also supports 2D detection research and provides a foundation for future multi-modal studies. To facilitate subsequent work, we establish a 2D benchmark and report baseline results using existing methods only. Specifically, we evaluate the camera-based methods YOLOv10 [105], RT-DETR [130], and DEIM [45] for the single-modality setting, and GM-DETR [114] for the multi-modal setting. Consistent with our 3D evaluation, we report Average Precision (AP) at a 2D IoU threshold of 0.5 with an emphasis on the Vehicle category; results for additional categories are available in the supplementary material. For 2D detection, we use the left camera images for each modality. As shown in Table 5, and in line with our 3D results, the multi-modal setting demonstrates greater robustness across diverse weather and illumination conditions.

7. Conclusion

We present DSERT-RoLL, a comprehensive multi-modal perception dataset featuring stereo Event–RGB–Thermal cameras, 4D radar, and dual LiDAR sensors. We establish benchmarks on DSERT-RoLL for both 3D and 2D object detection and introduce a modality-adaptive fusion baseline that strengthens detection under challenging weather and lighting conditions. We believe DSERT-RoLL will serve as a valuable foundation for future research, promoting progress in robust multi-modal perception and enabling more reliable 3D and 2D understanding under diverse real-world conditions.

8. Acknowledgments.

This work was supported by the InnoCORE program of the Ministry of Science and ICT(N10250156), and by the Technology Innovation Program (2410013617, 20024355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea).

References

[1] 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. 3

[2] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2

[3] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015. 3

[4] 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. 3, 7

[5] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 3

[6] L. Burner, A. Mitrokhin, C. Fermüller, and Y. Aloimonos. Evimo2: an event camera dataset for motion segmentation, optical flow, structure from motion, and visual inertial odometry in indoor scenes with monocular or stereo algorithms. *arXiv preprint arXiv:2205.03467*, 2022. 3

[7] 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. 3

[8] 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. 4, 7, 8

[9] Y. Chae, H. Park, H. Kim, and K.-J. Yoon. Doppler-aware lidar-radar fusion for weather-robust 3d detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27197–27208, 2025. 4

[10] 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. 3

[11] G. Chang, J. Lee, D. Kim, J. Kim, D. Lee, D. Ji, S. Jang, and S. Kim. Unified domain generalization and adaptation for multi-view 3d object detection. *Advances in Neural Information Processing Systems*, 37:58498–58524, 2024. 3

[12] T. Chen, Z. Tan, Q. Chu, Y. Wu, B. Liu, and N. Yu. Tciformer: Thermal conduction-inspired transformer for infrared small target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1201–1209, 2024. 2

[13] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 172–181, 2023. 4

[14] 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. 7, 8

[15] 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, 8

[16] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023. 4

[17] H. Cho and K.-J. Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 454–462, 2022. 3

[18] H. Cho and K.-J. Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 3

[19] H. Cho, J. Cho, and K.-J. Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 3

[20] H. Cho, Y. Jeong, T. Kim, and K.-J. Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 3

[21] H. Cho, T. Kim, Y. Jeong, and K.-J. Yoon. A benchmark dataset for event-guided human pose estimation and tracking in extreme conditions. *Advances in Neural Information Processing Systems*, 37:134826–134840, 2024. 3

[22] H. Cho, S.-H. Yoon, H. Kweon, and K.-J. Yoon. Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In *European Conference on Computer Vision*, pages 266–286. Springer, 2024. 3

[23] H. Cho, J.-y. Kang, Y. Kim, and K.-J. Yoon. Ev-3dod: Pushing the temporal boundaries of 3d object detection with event cameras. In *Proceedings of the Computer Vision*

and Pattern Recognition Conference, pages 27197–27210, 2025. 3

[24] 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. 2, 3

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[26] P. De Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 3

[27] 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. 6, 7

[28] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[29] 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. 3

[30] W. El Ahmar, Y. Massoud, D. Kolhatkar, H. AlGhamdi, M. Alja'Afreh, R. Hammoud, and R. Laganiere. Enhanced thermal-rgb fusion for robust object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 365–374, 2023. 3

[31] W. El Ahmar, A. Sappa, and R. Hammoud. Thermal pedestrian multiple object tracking challenge (tp-mot). In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4602–4609, 2025. 3

[32] B. Fan, X. Li, Y. Zhou, C. Xia, H. Fan, F. Xu, and J. Tian. Mgaf: Lidar-camera 3d object detection with multiple guidance and adaptive fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 4

[33] L. Fan, J. Wang, Y. Chang, Y. Li, Y. Wang, and D. Cao. 4d mmwave radar for autonomous driving perception: A comprehensive survey. *IEEE Transactions on Intelligent Vehicles*, 9(4):4606–4620, 2024. 3

[34] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[35] 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. 2, 3

[36] M. Gehrig, M. Millhäusler, D. Gehrig, and D. Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021. 3

[37] 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. 3

[38] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. M. López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6):820, 2016. 3

[39] 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. 7, 8

[40] 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. 7, 8

[41] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 3

[42] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15283–15292, 2021. 2

[43] J. Hidalgo-Carrió, G. Gallego, and D. Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2022. 3

[44] J. S. Hu, T. Kuai, and S. L. Waslander. Point density-aware voxels for lidar 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8469–8478, 2022. 6, 7

[45] 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. 8

[46] X. Huang, J. Wang, Q. Xia, S. Chen, B. Yang, X. Li, C. Wang, and C. Wen. V2x-r: Cooperative lidar-4d radar fusion with denoising diffusion for 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27390–27400, 2025. 3

[47] J.-J. Hwang, H. Kretzschmar, J. Manela, S. Rafferty, N. Armstrong-Crews, T. Chen, and D. Anguelov. Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *European conference on computer vision*, pages 388–405. Springer, 2022. 4

[48] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on com-*

puter vision and pattern recognition, pages 1037–1045, 2015. 3

[49] J. Jang, C. Park, H. Kim, J. Lee, and J. Paik. Multispectral object detection enhanced by cross-modal information complementary and cosine similarity channel resampling modules. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9437–9446. IEEE, 2025. 3

[50] Y. Jeong, H. Cho, and K.-J. Yoon. Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In *European Conference on Computer Vision*, pages 286–306. Springer, 2024. 3

[51] H. Ji, P. Liang, and E. Cheng. Enhancing 3d object detection with 2d detection-guided query anchors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21178–21187, 2024. 4

[52] X. Jiang, S. Li, Y. Liu, S. Wang, F. Jia, T. Wang, L. Han, and X. Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2561–2569, 2024. 2

[53] X. Jin, H. Su, C. Ma, K. Liu, W. Wu, F. Hui, and J. Yan. Geoformer: Geometry point encoder for 3d object detection with graph-based transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26879–26889, 2025. 4

[54] J.-Y. Kang, H. Cho, and K.-J. Yoon. Temporal stereo matching from event cameras via joint learning with stereoscopic flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[55] J.-Y. Kang, H. Cho, and K.-J. Yoon. Unleashing the temporal potential of stereo event cameras for continuous-time 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6869–6881, 2025. 3

[56] 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. 7

[57] T. Kim, J. Jeong, H. Cho, Y. Jeong, and K.-J. Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pages 433–451. Springer, 2024. 3

[58] Y. Kim, H. Cho, and K.-J. Yoon. From sharp to blur: Unsupervised domain adaptation for 2d human pose estimation under extreme motion blur using event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9406–9417, 2025. 3

[59] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon. Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504, 2021. 2

[60] Z. Kütük and G. Algan. Semantic segmentation for thermal images: A comparative survey. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 286–295, 2022. 3

[61] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1

[62] C. Li, D. Song, R. Tong, and M. Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, 2018. 3

[63] H. Li, J. Wang, J. Yuan, Y. Li, W. Weng, Y. Peng, Y. Zhang, Z. Xiong, and X. Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3250–3259, 2024. 3

[64] 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, 8

[65] 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. 7, 8

[66] Y. Li, L. Fan, Y. Liu, Z. Huang, Y. Chen, N. Wang, and Z. Zhang. Fully sparse fusion for 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7217–7231, 2024. 3

[67] Z. Li, S. Lan, J. M. Alvarez, and Z. Wu. Bevnext: Reviving dense bev frameworks for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20113–20123, 2024. 4

[68] Y. Liao, J. Xie, and A. Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 2

[69] H. Lin, Y. Zhang, S. Niu, S. Cui, and Z. Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, pages 96–114. Springer, 2024. 4

[70] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14928–14937, 2024. 4

[71] F. Liu, T. Huang, Q. Zhang, H. Yao, C. Zhang, F. Wan, Q. Ye, and Y. Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 200–217. Springer, 2024. 2

[72] H. Liu, S. Peng, L. Zhu, Y. Chang, H. Zhou, and L. Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 3

[73] 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. 3

[74] 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. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. URL https://api.semanticscholar.org/CorpusID:232352874. 5

[75] 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. 2, 7, 8

[76] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, and X. Bai. Lion: Linear group rnn for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 37:13601–13626, 2024. 4

[77] Z. Liu, J. Hou, X. Ye, T. Wang, J. Wang, and X. Bai. Seed: A simple and effective 3d detr in point clouds. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 4

[78] 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. 3

[79] A. Mitrokhin, C. Ye, C. Fermüller, Y. Aloimonos, and T. Delbruck. Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. in 2019 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6105–6112. 3

[80] M. Mostafavi, L. Wang, and K.-J. Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, 129(4):900–920, 2021. 3

[81] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International journal of robotics research*, 36(2): 142–149, 2017. 3

[82] F. Munir, S. Azam, and M. Jeon. Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 206–213. IEEE, 2021. 3

[83] 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. 2, 3, 7, 8

[84] 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. 3

[85] 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. 3, 4, 7, 8

[86] S. Park, M. Lee, J. Choi, and J. Choi. Selectively dilated convolution for accuracy-preserving sparse pillar-based embedded 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8104–8113, 2024. 4

[87] L. Peng, J. Xu, H. Cheng, Z. Yang, X. Wu, W. Qian, W. Wang, B. Wu, and D. Cai. Learning occupancy for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10281–10292, 2024. 2

[88] S. Peng, H. Zhou, H. Dong, Z. Shi, H. Liu, Y. Duan, Y. Chang, and L. Yan. Cosec: A coaxial stereo event camera dataset for autonomous driving. *arXiv preprint arXiv:2408.08500*, 2024. 3

[89] E. Perot, P. De Tournemire, D. Nitti, J. Masci, and A. Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33: 16639–16652, 2020. 3

[90] 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. 3

[91] 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. 3

[92] K. Qian, S. Zhu, X. Zhang, and L. E. Li. Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 444–453, 2021. 2

[93] 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

[94] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler. Radarscenes: A real-world radar point cloud data set for automotive applications. In *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2021. 3

[95] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. 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. 7

[96] U. Shin and J. Park. Deep depth estimation from thermal image: Dataset, benchmark, and challenges. *arXiv preprint arXiv:2503.22060*, 2025. 3

[97] U. Shin, J. Park, and I. S. Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1043–1053, 2023. 2, 3

[98] 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. 3, 7

[99] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 3

[100] 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. 3

[101] Y. Tang, Z. Meng, G. Chen, and E. Cheng. Simpb: A single model for 2d and 3d object detection from multiple cameras. In *European conference on computer vision*, pages 1–17. Springer, 2024. 2

[102] Teledyne FLIR. Thermal datasets for adas algorithm training. https://oem.flir.com/solutions/automotive/dataset/. Accessed: 2025-09-24. 3

[103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[104] V. Vs, D. Poster, S. You, S. Hu, and V. M. Patel. Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1412–1423, 2022. 3

[105] 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. 8

[106] J. Wang, Q. Meng, G. Liu, L. Yan, K. Wang, M.-M. Cheng, and Q. Hou. Towards stable 3d object detection. In *European conference on computer vision*, pages 197–213. Springer, 2024. 4

[107] 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. 7, 8

[108] X. Wang, S. Wang, C. Tang, L. Zhu, B. Jiang, Y. Tian, and J. Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19248–19257, 2024. 3

[109] Z. Wang, Z. Huang, Y. Gao, N. Wang, and S. Liu. Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 4

[110] 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. 3

[111] Q. Xia, W. Ye, H. Wu, S. Zhao, L. Xing, X. Huang, J. Deng, X. Li, C. Wen, and C. Wang. Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15321–15330, 2024. 3

[112] R. Xia, C. Zhao, M. Zheng, Z. Wu, Q. Sun, and Y. Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21572–21581, 2023. 3

[113] 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. 3

[114] 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. 3, 8

[115] L. Yan, P. Yan, S. Xiong, X. Xiang, and Y. Tan. Monocd: Monocular 3d object detection with complementary depths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10257, 2024. 2

[116] L. Yang, X. Zhang, J. Li, C. Wang, J. Ma, Z. Song, T. Zhao, Z. Song, L. Wang, M. Zhou, et al. V2x-radar: A multi-modal dataset with 4d radar for cooperative perception. *arXiv preprint arXiv:2411.10962*, 2024. 3

[117] L. Yang, T. Tang, J. Li, K. Yuan, K. Wu, P. Chen, L. Wang, Y. Huang, L. Li, X. Zhang, et al. Bevheight++: Toward robust visual centric 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[118] Z. Yao and M. C. Chuah. Event-guided low-light video semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3330–3341. IEEE, 2025. 3

[119] J. Yin, J. Shen, R. Chen, W. Li, R. Yang, P. Frossard, and W. Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14905–14915, 2024. URL https://api.semanticscholar.org/CorpusID:268667232. 2

[120] J. Yin, J. Shen, R. Chen, W. Li, R. Yang, P. Frossard, and W. Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024. 4

[121] T. Yin, X. Zhou, and P. Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 5

[122] A. Zhang, F. E. Nowruzi, and R. Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102. IEEE, 2021. 3

[123] 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. 7, 8

[124] 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*, pages 14477–14486, 2024. 4

[125] G. Zhang, J. Fan, L. Chen, Z. Zhang, Z. Lei, and L. Zhang. General geometry-aware weakly supervised 3d object detection. In *European Conference on Computer Vision*, pages 290–309. Springer, 2024. 3

[126] G. Zhang, L. Fan, C. He, Z. Lei, Z. Zhang, and L. Zhang. Voxel mamba: Group-free state space models for point cloud based 3d object detection. *Advances in Neural Information Processing Systems*, 37:81489–81509, 2024. 4

[127] 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. 7, 8

[128] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 3

[129] F. Zhao, Q. Zhou, and J. Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pages 190–205. Springer, 2024. 3

[130] 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. 8

[131] 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. 3

[132] X. Zheng, Y. Lyu, and L. Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pages 146–165. Springer, 2024. 3

[133] K. Zhou, L. Chen, and X. Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European conference on computer vision*, pages 787–803. Springer, 2020. 3

[134] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang. Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection. *IEEE Transactions on Intelligent Vehicles*, 8(2):1523–1535, 2023. 4

[135] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 1, 5

[136] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 3

[137] 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. 5

[138] Y. Zou, Y. Fu, T. Takatani, and Y. Zheng. Eventhdr: From event to high-speed hdr videos and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3