Horizon3D: Sparse Radar-Camera Fusion for Long-Range 3D Perception in Autonomous Driving

Geonho Bang, Geunju Baek, Dongyoung Lee, Wonjun Jeong, and Jun Won Choi*

Seoul National University, Seoul, Republic of Korea
{ghbang,gjbaek,dylee,wjjeong}@adr.snu.ac.kr, junwchoi@snu.ac.kr

Abstract. Long-range 3D object detection is critical for safe autonomous driving at highway speeds, yet existing radar-camera fusion methods face notable limitations at extended ranges. BEV-based approaches effectively encode scene-level context but incur rapidly growing computational cost and struggle to preserve fine-grained object-level detail, while query-based methods provide efficient object-centric encoding but lack sufficient scene-level context. Temporal fusion introduces additional challenges: distant objects produce only a few radar returns and occupy only a few image pixels, requiring scene-level accumulation, while high-speed motion causes large inter-frame displacements that require object-level motion modeling. BEV-based aggregation alleviates sparsity through multi-frame accumulation but is less suited to individual object motion, whereas query-based modeling captures object-level motion but provides limited scene-level temporal context. In this paper, we propose Horizon3D, a sparse radar-camera fusion framework for long-range 3D object detection that jointly captures object-level detail and scene-level context in both spatial and temporal dimensions through a hybrid representation that combines Gaussian primitives with sparse BEV features. Horizon3D first employs Keypoint-Guided Gaussian Initialization (KGGI) to initialize Gaussian primitives at object keypoints estimated from radar and camera features. Object-Centric Sparse Fusion (OCSF) aggregates cross-modal features around these primitives and splats the refined Gaussians onto the BEV plane, where they are fused with sparse radar BEV features to combine object-level detail with scene-level context. Finally, Dual-Path Temporal Fusion (DPTF) aggregates temporal cues through a BEV path for multi-frame feature accumulation and a Gaussian path for propagating primitives across frames to encode per-object motion. Extensive evaluations on TruckScenes demonstrate that Horizon3D achieves state-of-the-art performance for radar-camera 3D object detection. On the validation set, our approach outperforms the previous best method by **+3.0** NDS and **+1.6** mAP while maintaining a sparse representation with competitive inference speed.

Keywords: Autonomous Driving · 3D Object Detection · Sensor Fusion · Radar Sensor · Camera Sensor · Long-range 3D Object Detection

* Corresponding author.

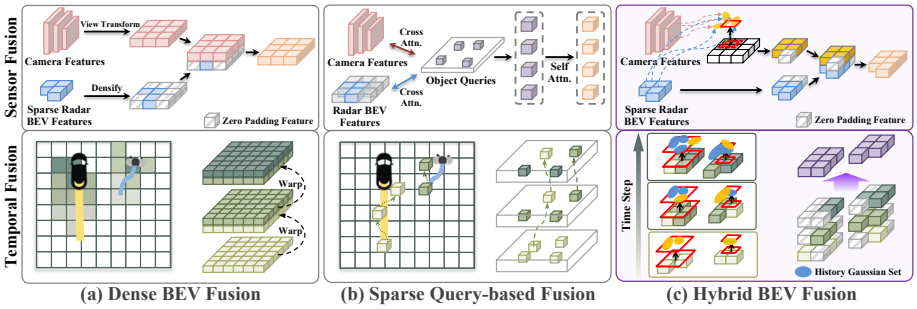

Fig. 1: Comparison of radar-camera fusion paradigms. (a) Dense BEV fusion encodes scene-level context but incurs growing cost with range, and its uniform temporal aggregation struggles to capture per-object motion. (b) Sparse query-based fusion achieves efficient object-centric encoding but lacks scene-level context in the spatiotemporal domain. (c) Our hybrid BEV fusion employs Gaussian primitives and sparse BEV features for object- and scene-level encoding, and uses dual-path temporal fusion.

1 Introduction

Reliable long-range perception is essential for safe autonomous driving, as substantially longer braking distances at highway speeds require accurate detection of objects beyond 150 m [1, 10, 31, 36]. Prior work on long-range 3D object detection has predominantly focused on LiDAR-based methods [22, 28, 40]. However, the high deployment and maintenance costs of LiDAR, along with its vulnerability to adverse weather, limit its practicality for large-scale deployment. As a cost-effective alternative, radar-camera fusion methods [4, 15–17, 19, 21] have gained significant attention. Recent work has increasingly adopted 4D radar, which provides spatial measurements including elevation, remains robust to adverse weather, and offers velocity cues for dynamic objects [2, 35, 44]. Despite this progress, widely used benchmarks (e.g., nuScenes, VoD, and TJ4D) [5, 26, 43] primarily target short- to mid-range urban perception. Although TruckScenes [10] was introduced for long-range, high-speed scenarios, fully exploiting 4D radar-camera complementarity in this setting remains largely unexplored.

Existing radar-camera 3D object detection frameworks can be broadly categorized into BEV-based and query-based methods. As illustrated in Fig. 1, both paradigms exhibit clear limitations in long-range detection, especially in sensor fusion and temporal fusion. For sensor fusion, BEV-based methods [2, 15–17, 21, 35, 44] project multimodal features into a unified dense BEV representation, which effectively captures scene-level context but incurs rapidly increasing computational cost as the detection range grows. Moreover, the dense BEV grid encodes the scene uniformly at a fixed spatial resolution, limiting fine-grained object-level detail (Fig. 1(a)). Query-based methods [7, 19, 32] instead perform object-centric encoding around learnable queries, enabling efficient long-range perception. However, the queries are optimized to represent foreground objects, and the resulting representation lacks sufficient scene-level context [20, 29] (Fig. 1(b)).

In this paper, we tackle the challenges of modeling both scene-level and object-level information for temporal fusion in long-range 3D perception. First, long-range objects produce sparse radar returns and occupy only a few image pixels, making single-frame modeling insufficient for robust perception. Consequently, an effective temporal fusion strategy is required to aggregate complementary information across multiple frames. Second, dynamic objects exhibit substantial motion between consecutive frames, particularly at highway speeds. Therefore, temporal feature propagation should be performed at the object level to accurately capture the motion of individual objects over time.

Several temporal fusion methods have been proposed for 3D perception. BEV-based methods [17, 21] aggregate dense BEV features over the entire spatial grid, effectively mitigating the sparsity of single-frame observations through multi-frame fusion. Although methods such as CRT-Fusion [16] introduce local motion compensation, they still rely on dense BEV representations, resulting in high computational cost, particularly for long-range perception. In contrast, query-based methods [19, 32] propagate compact object queries across frames, enabling each query to capture the temporal evolution of an individual object. While this design efficiently models object motion, it provides limited scene-level temporal context. Therefore, effective long-range 3D perception requires a temporal fusion method that jointly models scene-level context and object-level dynamics while maintaining high computational efficiency for extended-range perception.

In this paper, we propose Horizon3D, a sparse radar-camera fusion framework for long-range 3D object detection that addresses these challenges through a hybrid representation combining Gaussian primitives and sparse BEV features (Fig. 1(c)). Our approach comprises three key modules. The Keypoint-Guided Gaussian Initialization (KGGI) module estimates object keypoints from both radar and camera features and initializes a compact set of Gaussian primitives at these locations to guide object-centric encoding. Unlike existing Gaussian-based methods [12, 13] that locate numerous primitives across the scene, KGGI places primitives only at estimated object locations, enabling efficient computation even at extended detection ranges. Given the initialized Gaussians, the Object-Centric Sparse Fusion (OCSF) module aggregates cross-modal features around these primitives and iteratively refines their parameters to align with object structures. The refined Gaussians are then splatted onto the BEV plane to produce a sparse object-centric BEV representation, which is fused with sparse radar BEV features to encode fine-grained object-level detail and scene-level context. The Dual-Path Temporal Fusion (DPTF) module integrates temporal information through two complementary paths. The BEV path accumulates foreground features over multiple frames to compensate for the per-frame sparsity of distant objects, while the Gaussian path propagates past Gaussians to the current frame and encodes per-object motion across frames. In both paths, predicted velocities are used to compensate for the motion of dynamic objects, mitigating temporal misalignment in high-speed scenarios.

We evaluate Horizon3D on the TruckScenes benchmark [10], which features high-speed and long-range driving scenarios. On the validation split, Horizon3D

outperforms the previous best radar-camera fusion method by **+3.0** NDS and **+1.6** mAP, achieving state-of-the-art performance while running faster than existing BEV-based fusion methods.

The main contributions of this paper are as follows:

- We propose **Horizon3D**, a sparse radar-camera fusion framework for long-range 3D object detection that combines Gaussian primitives and sparse BEV features into an efficient hybrid representation.
- We introduce **KGGI** and **OCSF**. KGGI initializes Gaussian primitives at estimated object keypoints, while OCSF iteratively refines them to construct a sparse BEV representation that jointly captures object-level structure and scene-level context without requiring dense BEV construction. As a result, our approach significantly reduces computational cost.
- We propose **DPTF**, a dual-path temporal fusion that simultaneously addresses per-frame sparsity at long range and object motion at high speed. It employs velocity-based compensation in both paths to maintain temporal consistency.
- Extensive experiments on **TruckScenes** demonstrate that Horizon3D achieves state-of-the-art performance among radar-camera fusion methods, outperforming the previous best method by **+3.0** NDS and **+1.6** mAP while maintaining a sparse representation with competitive inference speed.

2 Related Work

2.1 Radar-Camera 3D Object Detection

Radar-camera fusion for 3D object detection combines camera semantics with radar's robust range and velocity measurements. Recent work has increasingly adopted 4D radar, which additionally provides elevation information [2, 35, 44]. Prior methods can be broadly categorized into BEV-based and query-based approaches.

BEV-based methods project multimodal features into a unified BEV space to align cross-sensor features and capture scene-level context [15–17, 21]. Radar measurements can serve as depth priors for view transformation [17,35], and cross-attention or multi-level refinement can be applied for flexible feature alignment [15, 21]. In contrast, query-based methods aggregate multimodal features using a limited number of object queries, constraining computational growth at longer ranges [7, 19, 32]. Recent methods use range-aware circular query initialization [7], densify sparse radar features for query-centric fusion [19], or perform alignment directly in a sparse feature space [32]. However, BEV-based methods incur rapidly growing computational cost as the detection range extends, while query-based methods often lack sufficient scene-level context.

Temporal fusion has been widely adopted to improve detection by aggregating consecutive frames [16, 19, 21, 32], and can further compensate for per-frame sparsity and large inter-frame displacements at extended ranges. BEV-based methods accumulate ego-motion-compensated BEV features over multiple frames

[16, 21]. CRT-Fusion [16] further addresses the misalignment of moving objects by explicitly modeling object-level dynamics, but still relies on dense BEV aggregation. Query-based methods propagate object queries across frames to aggregate temporal information [19, 32], yet their object-centric aggregation provides limited scene-level temporal context. However, temporal fusion that efficiently addresses both per-frame sparsity and object-level motion at long range without dense BEV construction remains unexplored.

2.2 Long-range 3D Object Detection

Long-range 3D object detection faces compounded challenges as the sensing range extends, since observations degrade while representation costs grow. Recent progress in long-range detection has been driven largely by LiDAR-based methods, which adopt a fully sparse paradigm that focuses computation on non-empty regions [6, 8, 39]. Sparse convolutions selectively process occupied voxels to mitigate computational growth [6, 8]. To compensate for the limited context of sparse representations, recent methods employ feature diffusion [39], hierarchical aggregation [38], sparse transformers [28], and linear-complexity architectures [24, 40]. Camera-based methods face different challenges at long range, as depth uncertainty increases and small-object observations degrade with distance. Far3D [14] constructs long-range queries from 2D priors, and LR3D [36] leverages 2D box supervision to compensate for the lack of 3D annotations at long distances.

Existing multimodal long-range fusion work has mainly focused on LiDAR-camera settings. SparseFusion [34] aligns object-level features in 3D space with lightweight attention, and SparseLIF [41] constructs 3D queries from image-based geometric cues with uncertainty-aware fusion. Self-supervised approaches for learning sparse fusion representations have also been explored [27]. However, these methods rely on relatively accurate LiDAR geometry for proposal generation and alignment, making them difficult to extend to 4D radar-camera settings where observations are substantially sparser and noisier.

3 Method

The overall architecture of Horizon3D is illustrated in Fig. 2. Horizon3D consists of three key modules: **Keypoint-Guided Gaussian Initialization (KGGI)**, **Object-Centric Sparse Fusion (OCSF)**, and **Dual-Path Temporal Fusion (DPTF)**. Section 3.1 introduces KGGI, which initializes sparse Gaussian primitives at object keypoints predicted from radar and camera features. Section 3.2 presents OCSF, which iteratively aggregates cross-modal features and refines the Gaussians to produce an object-centric BEV representation. Section 3.3 describes DPTF, which jointly aggregates temporal cues through a BEV path that accumulates multi-frame features and a Gaussian path that propagates primitives across frames to encode per-object motion.

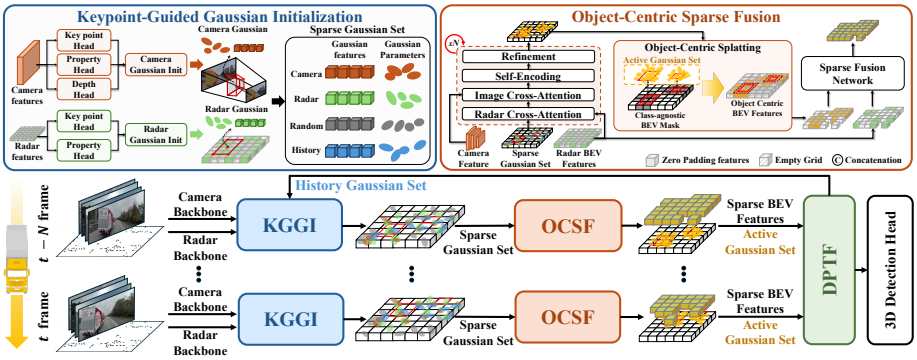

Fig. 2: Overall architecture of Horizon3D. Multi-view camera images and 4D radar points are encoded by their respective backbones. The KGGI module estimates object keypoints from both modalities and initializes sparse Gaussian primitives at these locations. The OCSF module aggregates cross-modal features around the Gaussians and produces a sparse BEV representation that captures both object-level detail and scene-level context. The DPTF module then fuses temporal information through two complementary paths—Gaussian and BEV—to produce the fused feature map, which is passed to a 3D detection head.

3.1 Keypoint-Guided Gaussian Initialization

Gaussian primitives have been explored as a flexible scene representation for 3D perception [3,12,13,42]. These methods typically maintain a large set of Gaussians across the scene and iteratively refine their parameters to concentrate primitives in foreground regions. However, sufficient spatial coverage requires a substantial number of primitives, and the computational cost scales with the detection range. To address this limitation, we propose Keypoint-Guided Gaussian Initialization, which estimates object keypoints from radar and camera features and anchors a compact set of Gaussians at these locations, enabling object-centric feature aggregation over a wide detection range without redundant spatial coverage.

We process 4D radar points, which provide spatial measurements, radar cross-section (RCS), and radial velocity, using VoxelNeXt [6]. VoxelNeXt maintains a fully sparse representation while capturing broad scene-level context through its large receptive field, producing sparse radar BEV features $F_R \in \mathbb{R}^{C \times H_R \times W_R}$, where C, H_R, and W_R denote the channel dimension, BEV height, and BEV width, respectively. Given N surround-view camera images, we extract image features $F_I \in \mathbb{R}^{N \times C_I \times H_I \times W_I}$ using a camera backbone.

For the radar branch, we apply a lightweight MLP to the per-cell features in F_R to predict objectness scores and select the top-K_r scoring cells as radar keypoints. These keypoints are used to initialize K_r Gaussian primitives. For the camera branch, inspired by Far3D [14], we feed F_I into a 2D keypoint head to produce a per-pixel objectness score map $S_n \in \mathbb{R}^{1 \times H_I \times W_I}$ for each view. In parallel, a depth head predicts a depth distribution $D_n \in \mathbb{R}^{N_d \times H_I \times W_I}$, where N_d is the number of depth bins. At each pixel (u, v) in view n, we obtain the

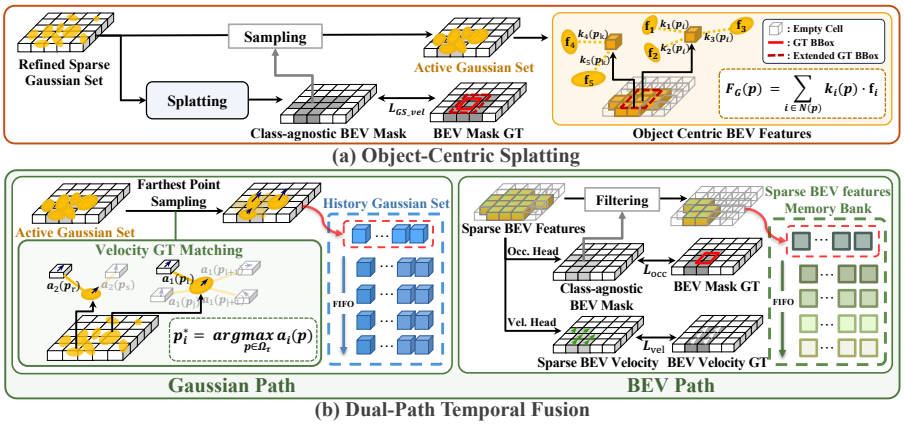

Fig. 3: Details of Object-Centric Splatting and Velocity-Guided Temporal Alignment (a) Refined Gaussians are splatted onto BEV cells to produce a supervised occupancy mask. The Active Gaussian Set is pooled via Gaussian-weighted aggregation into object-centric BEV features. (b) For the Gaussian path (left), the Active Gaussian Set is subsampled via FPS to form the History Gaussian Set, with per-Gaussian velocity supervised by the most-contributed occupied cell. Stored Gaussians are warped via predicted velocities and ego-motion, then fed into the next frame's KGGI. For the BEV path (right), occupancy and a velocity head predict per-cell occupancy and velocity from the fused sparse BEV features, selecting foreground cells for memory storage and velocity-guided temporal alignment.

estimated depth d and its confidence from D_n, and compute a candidate score by combining the objectness score $S_n(u, v)$ with the depth confidence. The top-K_c scoring candidates are then lifted to 3D using the estimated depth and camera calibration and projected onto the BEV plane as camera keypoints. These keypoints are also used for K_c Gaussian primitives.

Each Gaussian is parameterized as

$$\mathbf{g}_i = (\mathbf{m}_i \in \mathbb{R}^2,\ \mathbf{s}_i \in \mathbb{R}^2,\ \theta_i \in \mathbb{R},\ o_i \in \mathbb{R},\ \mathbf{v}_i \in \mathbb{R}^2), \tag{1}$$

where $\mathbf{m}_i$, $\mathbf{s}_i$, θ_i, o_i, and $\mathbf{v}_i$ denote the mean, scale, rotation, opacity, and velocity of the i-th Gaussian, respectively. Each Gaussian is also associated with a feature vector $\mathbf{f}_i \in \mathbb{R}^C$. The K_r and K_c Gaussians are initialized using the radar key points and camera key points, respectively. Specifically, we initialize $\mathbf{m}_i$ with the corresponding keypoint location. The remaining attributes $(\mathbf{s}_i, \theta_i, o_i, \mathbf{v}_i)$ and the feature $\mathbf{f}_i$ are initialized as learnable parameters, which are refined by the OCSF module. In addition to $K_r + K_c$ Gaussians, we add K_{rand} randomly initialized Gaussians with learnable parameters to improve coverage in regions not covered by either branch. We also include K_{hist} history Gaussians propagated from previous frames to provide object-level temporal cues for the current frame (Sec. 3.3). The resulting sparse Gaussian set is defined as $\mathcal{G}_0 = \{\mathbf{g}_i\}_{i=1}^{N_g}$, where $N_g = K_r + K_c + K_{\text{rand}} + K_{\text{hist}}$.

3.2 Object-Centric Sparse Fusion

As discussed in Sec. 1, dense BEV-based methods incur prohibitive computational cost at extended ranges, whereas query-based methods provide limited scene-level context. To address these limitations, OCSF takes the initialized Gaussian set $\mathcal{G}_0$ and progressively aggregates cross-modal features around the Gaussian primitives. This process iteratively refines the Gaussian parameters, allowing the primitives to better align with the underlying 3D object geometry. The refined Gaussians are then splatted onto the BEV plane to generate a sparse object-centric BEV representation, which is fused with sparse radar BEV features to jointly capture fine-grained object-level information and scene-level context.

Multimodal Gaussian Encoder. Motivated by recent progress in Gaussian-based representations [12, 13], we propose a Multimodal Gaussian Encoder that refines the initialized primitives over L layers, where each layer updates both Gaussian features and parameters. At layer l, each Gaussian gathers modality-specific contexts through deformable cross-attention [45]. For each Gaussian, its mean $\mathbf{m}_i$ serves as the reference point. Around this reference point, we generate K sampling points $\mathcal{Q}_i = \{\mathbf{m}_i + \boldsymbol{\Delta}_k\}_{k=1}^{K}$. The offsets $\boldsymbol{\Delta}_k$ are derived from the scale $\mathbf{s}_i$ and rotation θ_i to distribute the sampling points along the Gaussian's spatial extent. Additional learnable offsets predicted from $\mathbf{f}_i$ further adapt the sampling locations to regions of interest. The Gaussian feature is then updated as

$$\hat{\mathbf{f}}_i^{(l)} = \mathbf{f}_i^{(l)} + \sum_{M \in \{R,I\}} \mathrm{DCA}_M\left(\mathbf{f}_i^{(l)},\, F_M,\, \phi_M\right), \tag{2}$$

where DCA_M is deformable cross-attention over modality M, and F_R and F_I are the radar BEV and multi-view image features, respectively. The projection ϕ_M maps the sampling points $\mathcal{Q}_i$ to feature coordinates in modality M, where the projected locations serve as sampling locations for deformable cross-attention. The projection ϕ_M maps the sampling points $\mathcal{Q}_i$ to modality-specific feature coordinates used by deformable cross-attention. Specifically, ϕ_R maps the 2D sampling points to BEV cell coordinates for radar feature sampling. For image feature sampling, each 2D sampling point in $\mathcal{Q}_i$ is lifted from the BEV plane to 3D by predicting an additional height coordinate from $\mathbf{f}_i$, and ϕ_I projects the resulting 3D points onto each camera view.

We quantize the 2D mean $\mathbf{m}_i$ of each Gaussian onto the BEV grid and apply 2D sparse convolution over the occupied BEV cells, enabling interactions among nearby Gaussians. The mean is refined as

$$\mathbf{m}_i^{(l+1)} = \mathbf{m}_i^{(l)} + \mathbf{r}_i^{(l)}, \quad \mathbf{r}_i^{(l)} = \mathrm{MLP}_m\left(\hat{\mathbf{f}}_i^{(l)}\right), \tag{3}$$

where $\mathbf{r}_i^{(l)}$ is the predicted residual. The remaining parameters ($\mathbf{s}_i$, θ_i, o_i) and per-Gaussian velocity $\mathbf{v}_i$ for temporal fusion (Sec. 3.3) are also regressed from $\hat{\mathbf{f}}_i^{(l)}$ at each layer. The output feature after sparse convolution is used as $\mathbf{f}_i^{(l+1)}$ and passed to the next layer. After L layers, we obtain the refined Gaussian set $\mathcal{G}_L$.

Object-Centric Gaussian Splatting. The refined Gaussian set $\mathcal{G}_L$ is splatted onto BEV cells to produce an object-centric BEV representation (Fig. 3(a)). For each Gaussian $\mathbf{g}_i$, we construct its 2D covariance matrix as

$$\Sigma_i = R(\theta_i)\,\mathrm{diag}(s_{ix}^2,\, s_{iy}^2)\,R(\theta_i)^\top, \tag{4}$$

where $R(\theta_i)$ is the 2D rotation matrix. At a BEV cell center $\mathbf{p}$, the Gaussian kernel weight is given by $k_i(\mathbf{p}) = \exp\big(-\frac{1}{2}(\mathbf{p}-\mathbf{m}_i)^\top \Sigma_i^{-1}(\mathbf{p}-\mathbf{m}_i)\big)$, and the occupancy contribution is given by $a_i(\mathbf{p}) = \sigma(o_i)\,k_i(\mathbf{p})$. The class-agnostic occupancy probability at $\mathbf{p}$ is then

$$\alpha(\mathbf{p}) = 1 - \prod_{i=1}^{N_g}\big(1 - a_i(\mathbf{p})\big). \tag{5}$$

The resulting occupancy map $\alpha \in \mathbb{R}^{H_R \times W_R}$ is supervised using a ground-truth BEV mask derived from rasterized 3D bounding boxes, with each box enlarged by a margin ϵ to encourage encoding of surrounding spatial context.

BEV cells exceeding an occupancy threshold, $\Omega_\tau = \{\mathbf{p} \mid \alpha(\mathbf{p}) > \tau\}$, form a sparse object-centric BEV representation. For each selected cell, we define the Active Gaussian Set $\mathcal{N}(\mathbf{p})$ as the set of Gaussians contributing to that cell, and compute its BEV feature by Gaussian-weighted pooling as

$$F_G(\mathbf{p}) = \sum_{i \in \mathcal{N}(\mathbf{p})} k_i(\mathbf{p})\,\mathbf{f}_i, \qquad \mathbf{p} \in \Omega_\tau. \tag{6}$$

The object-centric BEV features F_G are then fused with the sparse radar BEV features F_R to produce the final sparse BEV representation. We take the union of non-empty BEV cells in F_G and F_R and apply a lightweight gating network to adaptively weight both features. The gated features are processed by a sparse encoder with submanifold sparse convolutions, yielding the fused sparse BEV feature map F_{fuse}. We provide the full formulation in the supplementary material.

3.3 Dual-Path Temporal Fusion

As discussed in Sec. 1, long-range objects produce sparse per-frame observations that demand multi-frame accumulation, while fast-moving objects undergo large inter-frame displacements that require object-level motion modeling. To address both challenges, we propose DPTF, which integrates temporal cues through two complementary paths. The BEV path compensates for sparsity by accumulating BEV features over multiple frames, while operating only on object-occupied cells rather than on the entire dense grid. The Gaussian path warps past Gaussians to the current frame and merges them with current-frame primitives to capture object-level temporal dynamics. In both paths, predicted velocities are used alongside ego-motion to handle large displacements of fast-moving objects.

BEV Path. The BEV path accumulates foreground BEV features over multiple frames to compensate for the sparse observations of distant objects (Fig. 3(b),

right). Given the fused sparse BEV feature map F^t_{fuse}, we predict an occupancy score $\hat{o}^t(\mathbf{p})$ and a planar velocity vector $\hat{\mathbf{u}}^t(\mathbf{p})$ at each non-empty BEV cell $\mathbf{p}$ as

$$\hat{o}^t(\mathbf{p}) = \sigma\big(h_{\text{occ}}\big(F^t_{\text{fuse}}(\mathbf{p})\big)\big)\,, \qquad \hat{\mathbf{u}}^t(\mathbf{p}) = h_{\text{vel}}\big(F^t_{\text{fuse}}(\mathbf{p})\big)\,, \tag{7}$$

where h_{occ} and h_{vel} are small stacks of submanifold sparse convolutions. While the occupancy mask in Sec. 3.2 uses enlarged bounding boxes to encourage contextual encoding, $\hat{o}^t(\mathbf{p})$ is supervised with a tight mask from the original boxes, yielding a compact set of object-occupied cells for efficient memory storage and temporal aggregation. The predicted velocity $\hat{\mathbf{u}}^t(\mathbf{p})$ is supervised by per-cell ground-truth velocities, where each cell within a ground-truth bounding box is assigned the velocity of that box. We select non-empty cells whose occupancy scores exceed a threshold τ_{mem} as foreground cells and store only their features $F^t_{\text{fuse}}(\mathbf{p})$ and predicted velocities $\hat{\mathbf{u}}^t(\mathbf{p})$ in the BEV memory bank.

For temporal fusion, we align each historical frame $(t{-}k)$ to the current frame by combining ego-motion with velocity-guided warping. The warped BEV cell index is computed as

$$\mathbf{p}' = \pi\big(T_{t-k\to t}\big(\mathbf{x}(\mathbf{p}) + \hat{\mathbf{u}}^{t-k}(\mathbf{p})\,\Delta t_k\big)\big)\,, \tag{8}$$

where $\mathbf{p}'$ is the warped BEV cell index in the current frame, $T_{t-k\to t}$ is the ego-motion transformation from frame $t{-}k$ to t, Δt_k is the elapsed time between the two frames, $\mathbf{x}(\cdot)$ maps a BEV cell center to world coordinates, and $\pi(\cdot)$ quantizes the warped coordinates back to BEV indices. The aligned historical BEV features are stacked with the current fused BEV feature map F^t_{fuse} along the temporal axis. Each temporal feature map is augmented with a time embedding derived from Δt_k, and sparse convolutions fuse the stacked features into a temporally fused sparse BEV feature map for the detection head. We provide the detailed architecture of this aggregation module in the supplementary material.

Gaussian Path. The Gaussian path captures object-level temporal dynamics by selecting and propagating past Gaussians to the current frame (Fig. 3(b), left). After splatting, we gather Gaussians that contribute to occupied cells via the Active Gaussian Set $\mathcal{N}(\mathbf{p})$ and apply farthest point sampling (FPS) to form a spatially diverse History Gaussian Set for memory storage. Compared to top-K selection by opacity, which tends to concentrate on a few high-response regions, FPS samples Gaussians more evenly across objects, improving object-level temporal modeling after merging with current-frame primitives.

At highway speeds, dynamic objects can move substantially between consecutive frames, and ego-motion compensation alone leaves significant spatial misalignment. The velocity $\mathbf{v}_i$ predicted by the OCSF refinement module (Sec. 3.2) is supervised using the ground-truth velocity of the occupied cell to which each Gaussian contributes most. The target cell and velocity are defined as

$$\mathbf{p}^*_i = \arg\max_{\mathbf{p}\in\Omega_\tau} a_i(\mathbf{p}), \qquad \mathbf{v}^{\text{gt}}_i = \mathbf{v}^{\text{gt}}(\mathbf{p}^*_i). \tag{9}$$

Each Gaussian's predicted velocity is then employed alongside ego-motion to warp past Gaussians to the current frame. The warped History Gaussian Set

is incorporated into the sparse Gaussian set of KGGI (Sec. 3.1) and jointly processed with current-frame Gaussians through the OCSF pipeline (Sec. 3.2). Since residual misalignment can remain after velocity compensation, we add temporal self-attention within the Gaussian encoder over the merged set, enabling cross-frame interaction beyond the local receptive field of sparse convolution.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate our proposed method on the TruckScenes [10] dataset. The TruckScenes dataset consists of 747 driving scenes, each approximately 20 seconds long, captured by four cameras, six LiDARs, and six 4D radars mounted on a heavy-duty commercial truck. Unlike existing urban-driving benchmarks, TruckScenes features high-speed driving scenarios with an evaluation range of up to 150m. Our evaluation follows the official benchmark metrics, including mean Average Precision (mAP) and nuScenes Detection Score (NDS).
Implementation Details. We employ VoVNet-99 [18] initialized from a FCOS3D [30] checkpoint following SpaRC [32] and ResNet-50 [11] pretrained on ImageNet as camera backbones. In the radar branch, we accumulate the past 10 sweeps from six radar sensors as input point clouds and use VoxelNeXt [6] as the backbone network. Training proceeds in two phases: we first train the single-frame model (KGGI and OCSF) for 24 epochs, then train the full model including DPTF for an additional 24 epochs. For temporal fusion, the BEV path aggregates features from the past 8 frames, while the Gaussian path maintains history Gaussians from the past 4 frames. We use AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. All experiments are conducted on NVIDIA RTX PRO 5000 Blackwell and RTX 3090 GPUs, and inference speed is measured on a single RTX 3090. Detailed training configurations and hyperparameter settings are provided in the supplementary material.

4.2 Comparison to the state of the art

Table 1 compares Horizon3D with existing methods on the TruckScenes dataset. On the validation split with a ResNet-50 backbone, Horizon3D achieves the highest performance among radar-camera methods, outperforming both BEV-based methods [16, 25] and the query-based method [19] by a large margin. Notably, Horizon3D also surpasses CenterPoint-V [37], a LiDAR-based method, by $+\mathbf{2.1}$ NDS and $+\mathbf{1.0}$ mAP. With the V2-99 backbone, Horizon3D exceeds SpaRC [32], the previous state-of-the-art radar-camera method, by $+\mathbf{3.0}$ NDS and $+\mathbf{1.6}$ mAP. On the test split, Horizon3D further outperforms SpaRC by $+\mathbf{4.4}$ NDS and $+\mathbf{0.5}$ mAP while using lower-resolution image inputs, demonstrating the effectiveness of our hybrid representation. Horizon3D also surpasses CenterPoint-V [37] by $+\mathbf{0.8}$ NDS and $+\mathbf{1.0}$ mAP, achieving competitive performance against LiDAR-based methods.

Table 1: Performance comparison on the TruckScenes dataset. 'L', 'C', and 'R' denote LiDAR, Camera, and Radar, respectively. * indicates results from the official TruckScenes baseline. † indicates methods reproduced using officially released code following their original training configurations.

Methods	Input	Backbone	Img Size	Split	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-V* [37]	L	Voxel	–	val	35.3	22.6	0.461	0.405	0.468	3.028	0.261
RCTrans† [19]	C+R	R50	256×864	val	22.9	12.6	1.025	0.507	0.608	0.894	0.331
CRT-Fusion† [16]	C+R	R50	256×864	val	28.8	16.9	0.833	0.512	0.444	0.852	0.322
BEVFusion† [25]	C+R	R50	256×864	val	30.4	18.2	0.941	0.442	0.389	0.892	0.208
Horizon3D (Ours)	C+R	R50	256×864	val	**37.4**	**23.6**	0.876	0.410	0.331	0.625	0.198
Far3D [14]	C	V2-99	640×960	val	21.4	10.7	0.883	0.507	0.671	1.352	0.338
SpaRC [32]	C+R	V2-99	640×960	val	35.4	22.5	0.798	0.449	0.476	0.613	0.248
Horizon3D (Ours)	C+R	V2-99	640×960	val	**38.4**	**24.1**	0.833	0.404	0.355	0.561	0.208
CenterPoint-V* [37]	L	Voxel	–	test	41.0	26.7	0.409	0.352	0.277	2.730	0.201
RadarGNN* [9]	R	–	–	test	10.7	7.0	0.892	0.809	1.132	8.003	0.571
PETR* [23]	C	V2-99	300×800	test	12.1	2.2	1.125	0.686	0.647	1.499	0.564
HyDRa [33]	C+R	V2-99	928×1952	test	22.4	12.8	0.725	0.544	0.744	1.180	0.388
SpaRC [32]	C+R	V2-99	928×1952	test	37.4	27.2	0.759	0.413	0.411	0.814	0.227
Horizon3D (Ours)	C+R	V2-99	640×960	test	**41.8**	**27.7**	0.779	0.354	0.269	0.637	0.167

Table 2: Ablation study of main components of Horizon3D.

Input	KGGI	OCSF	DPTF	mAP↑	NDS↑
R				12.8	27.1
C+R		✓		17.7	32.5
	✓	✓		20.3	35.4
		✓	✓	23.0	36.4
	✓	✓	✓	**23.6**	**37.4**

Table 3: Ablation study of Gaussian initialization strategies in KGGI.

	mAP↑	NDS↑
Random	17.7	32.5
Keypoint only	19.2	34.3
Random + Keypoint	**20.3**	**35.4**

Table 4: Ablation study of bounding box enlargement for occupancy supervision.

Enlarged GT	mAP↑	NDS↑
	17.2	32.9
✓	**20.3**	**35.4**

Table 5: Ablation study for the BEV and Gaussian paths in DPTF.

BEV	Gaussian	mAP↑	NDS↑
✓		22.5	36.3
	✓	21.3	36.0
✓	✓	**23.6**	**37.4**

4.3 Ablation Studies

We conduct ablation studies on the TruckScenes validation set using the ResNet-50 backbone. For variants without DPTF, we report results from the first training phase (single-frame model) for efficiency.

Component analysis. To assess the contribution of each module, we incrementally add components and report the results in Table 2. The radar-only baseline achieves 12.8 mAP and 27.1 NDS. Adding OCSF with randomly initialized Gaussians improves mAP by +4.9 and NDS by +5.4, showing that Gaussian-based cross-modal fusion effectively injects object-level detail into the BEV space. Replacing random initialization with KGGI further improves mAP by +2.6 and NDS by +2.9, demonstrating that placing Gaussians at estimated object keypoints leads to more effective object-centric encoding. Adding DPTF

Table 6: Ablation study for evaluating the effect of object velocity compensation.

Path	Object Vel. Compensation	mAP↑	NDS↑
BEV		21.6	36.1
	✓	**23.0**	**36.7**
Gaussian		20.6	34.8
	✓	**21.3**	**36.0**

Table 7: Ablation study on sampling strategies for Gaussian path.

	mAP↑	NDS↑
Random	20.7	35.8
Top-K	21.1	35.9
FPS	**21.3**	**36.0**

Table 8: Comparison of accuracy and efficiency with existing radar–camera fusion methods.

Method	mAP↑	NDS↑	FPS↑	Mem(GB)↓
BEVFusion [25]	18.2	30.4	4.2	6.520
CRTFusion [16]	16.9	28.8	2.9	7.580
RCTrans [19]	12.6	22.9	**9.5**	**2.620**
Horizon3D	**23.6**	**37.4**	8.5	3.236

Table 9: Ablation study on distance-wise performance.

	0-25m		25-50m		50-100m		100-150m	
	mAP	NDS	mAP	NDS	mAP	NDS	mAP	NDS
RCTrans [19]	19.9	28.4	12.5	24.1	7.9	19.9	6.2	15.6
CRTFusion [16]	26.2	34.4	16.5	27.3	11.0	26.3	7.0	11.4
BEVFusion [25]	28.8	35.7	16.8	25.2	11.7	20.1	7.4	13.9
Horizon3D	**36.5**	**43.0**	**23.3**	**37.8**	**17.0**	**32.7**	**8.9**	**18.4**

without KGGI improves over the OCSF-only variant by +5.3 mAP and +3.9 NDS, showing the effectiveness of dual-path temporal fusion. Combining all three modules achieves the best performance, with an mAP of 23.6 and an NDS of 37.4, confirming that the modules provide complementary benefits.

Initialization strategies for KGGI. Table 3 compares Gaussian initialization strategies. Keypoint-only initialization outperforms random initialization by +1.5 mAP and +1.8 NDS, confirming the importance of sensor-guided placement. Combining keypoint and random Gaussians further improves both metrics by +1.1, as random Gaussians help cover regions missed by sensor-guided keypoints.

Effect of Bounding Box Enlargement. Table 4 examines the effect of bounding box enlargement for occupancy supervision in OCSF (Sec. 3.2). Without enlargement, the model achieves 17.2 mAP and 32.9 NDS. Enlarging the bounding boxes by a margin ϵ before rasterization improves mAP by +3.1 and NDS by +2.5, as the expanded supervision encourages Gaussians to encode not only object interiors but also surrounding spatial context.

Effect of Dual-Path Temporal Fusion. Tables 5 and 6 analyze the design choices in DPTF (Sec. 3.3). In Table 5, the BEV path alone achieves 22.5 mAP and 36.3 NDS, while the Gaussian path alone yields 21.3 mAP and 36.0 NDS. Combining both paths improves over the BEV-only variant by +1.1 mAP and +1.1 NDS, demonstrating that object-level and scene-level temporal fusion provide complementary benefits. Table 6 evaluates the effect of object velocity compensation in each path. Applying velocity compensation improves the BEV path by +1.4 mAP and +0.6 NDS, and the Gaussian path by +0.7 mAP and +1.2 NDS, confirming that explicit motion compensation is essential for both paths in high-speed scenarios.

Effect of Gaussian Sampling Strategy. Table 7 compares strategies for selecting Gaussians to store in the temporal memory bank. Random selection

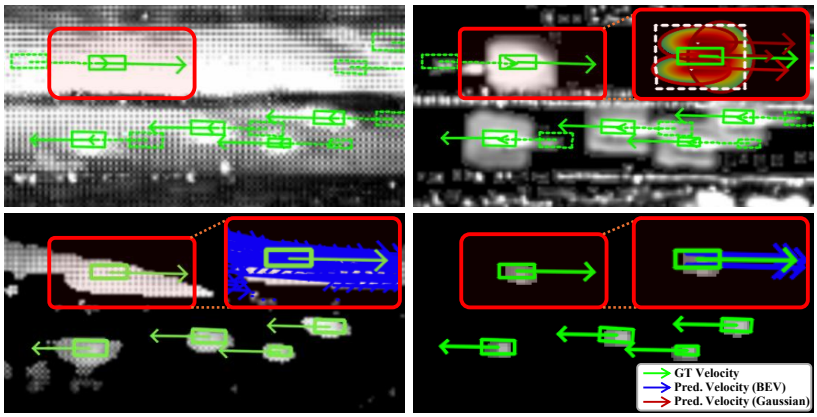

Fig. 4: Qualitative comparison of temporally aggregated BEV feature maps (top) and velocity predictions with foreground regions (bottom). Green and blue arrows denote ground-truth and predicted per-cell velocities, respectively. Red arrows indicate per-Gaussian velocities predicted by the Gaussian path. Dashed green boxes represent object positions from previous frames.

achieves 20.7 mAP and 35.8 NDS, top-K selection by opacity reaches 21.1 mAP and 35.9 NDS, and farthest point sampling (FPS) performs best with 21.3 mAP and 36.0 NDS. Top-K selection tends to store Gaussians from a few high-response regions, whereas FPS selects a spatially diverse subset that better covers multiple objects, leading to more consistent temporal fusion after merging.

Efficiency comparison. Table 8 compares accuracy and efficiency with existing radar-camera fusion methods. Dense BEV-based methods [16, 25] suffer from high memory consumption and low inference speed due to the cost of dense BEV construction at extended ranges. RCTrans [19] runs efficiently through query-based encoding but yields substantially lower detection accuracy. Horizon3D achieves the highest accuracy while maintaining competitive speed and moderate memory usage, demonstrating that the hybrid Gaussian-BEV representation effectively balances accuracy and efficiency for long-range perception.

Distance-wise performance. Table 9 reports performance across different distance ranges. Horizon3D achieves the highest mAP at every range, with particularly notable gains at 25-50 m (+6.5 mAP over BEVFusion) and 100-150 m (+1.5 mAP and +4.5 NDS over BEVfusion), confirming the effectiveness of our approach for long-range detection.

Qualitative results. Fig. 4 compares CRT-Fusion and Horizon3D in terms of temporally aggregated BEV feature maps (top) and velocity predictions with foreground regions (bottom). CRT-Fusion shows widespread activations across the BEV grid, as temporal misalignment is further diffused by dense convolution. As shown in the bottom-left, foreground regions are inaccurately predicted and

velocities are estimated over broad areas including background cells, introducing erroneous motion compensation. In contrast, Horizon3D preserves activations only around object regions (top-right). The object-centric Gaussian representation enables accurate foreground and velocity prediction in the fused BEV features (bottom-right), resulting in precise temporal alignment. The top-right zoom-in shows that Gaussians are distributed along the enlarged bounding box regions, extending activations slightly beyond object boundaries to capture surrounding context.

5 Conclusion

We presented Horizon3D, a sparse radar-camera fusion framework for long-range 3D object detection that captures both fine-grained object detail and scene-level context through a hybrid representation combining Gaussian primitives and sparse BEV features. KGGI initializes a compact set of Gaussians at sensor-estimated object keypoints, and OCSF iteratively refines them to produce a sparse, object-centric BEV representation that captures both fine-grained object detail and scene-level context without dense BEV construction. DPTF integrates temporal information through dual paths with explicit velocity compensation, maintaining both object-level dynamics and scene-level temporal context in high-speed scenarios. Experiments on TruckScenes demonstrate that Horizon3D outperforms prior radar-camera fusion methods by $+\mathbf{3.0}$ NDS and $+\mathbf{1.6}$ mAP while maintaining competitive inference speed.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00421129), and in part by Hyundai Rotem Company.

References

1. Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindström, C., Motorniuk, D., Fu, J., Widahl, J., Petersson, C.: Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20178–20188 (2023)
2. Bai, X., Yu, Z., Zheng, L., Zhang, X., Zhou, Z., Zhang, X., Wang, F., Bai, J., Shen, H.L.: SGDet3D: Semantics and geometry fusion for 3d object detection using 4d radar and camera. IEEE Robotics and Automation Letters **10**(1), 828–835 (2025)

3. Bai, X., Zhou, C., Zheng, L., Liu, J., Cao, S.Y., Zhang, X., Li, Y., Zhang, Z., Shen, H.L.: RaGS: Unleashing 3D gaussian splatting from 4D radar and monocular cue for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4983–4992 (2026)
4. Bang, G., Seong, M., Kim, J., Baek, G., Oh, D., Kim, J., Koh, J., Choi, J.W.: RCT-Distill: Cross-modal knowledge distillation framework for radar-camera 3D object detection with temporal fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25315–25324 (2025)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
6. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: VoxelNeXt: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21674–21683 (2023)
7. Chu, X., Deng, J., You, G., Duan, Y., Li, H., Zhang, Y.: RaCFormer: Towards high-quality 3d object detection via query-based radar-camera fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17081–17091 (2025)
8. Fan, L., Wang, F., Wang, N., Zhang, Z.X.: Fully sparse 3d object detection. Advances in Neural Information Processing Systems **35**, 351–363 (2022)
9. Fent, F., Bauerschmidt, P., Lienkamp, M.: RadarGNN: Transformation invariant graph neural network for radar-based perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 182–191 (2023)
10. Fent, F., Kuttenreich, F., Ruch, F., Rizwin, F., Juergens, S., Lechermann, L., Nissler, C., Perl, A., Voll, U., Yan, M., Lienkamp, M.: MAN TruckScenes: a multimodal dataset for autonomous trucking in diverse conditions. Advances in Neural Information Processing Systems **37**, 62062–62082 (2024)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Huang, Y., Thammatadatrakoon, A., Zheng, W., Zhang, Y., Du, D., Lu, J.: GaussianFormer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27477–27486 (2025)
13. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: GaussianFormer: Scene as gaussians for vision-based 3D semantic occupancy prediction. In: European Conference on Computer Vision. pp. 376–393 (2024)
14. Jiang, X., Li, S., Liu, Y., Wang, S., Jia, F., Wang, T., Han, L., Zhang, X.: Far3D: Expanding the horizon for surround-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2561–2569 (2024)
15. Kim, J., Seong, M., Bang, G., Kum, D., Choi, J.W.: RCM-Fusion: Radar-camera multi-level fusion for 3d object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 18236–18242 (2024)
16. Kim, J., Seong, M., Choi, J.W.: CRT-Fusion: Camera, radar, temporal fusion using motion information for 3d object detection. Advances in Neural Information Processing Systems **37**, 108625–108648 (2024)
17. Kim, Y., Shin, J., Kim, S., Lee, I.J., Choi, J.W., Kum, D.: CRN: Camera radar net for accurate, robust, efficient 3d perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17615–17626 (2023)

18. Lee, Y., Hwang, J.w., Lee, S., Bae, Y., Park, J.: An energy and GPU-computation efficient backbone network for real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 752–760 (2019)
19. Li, Y., Yang, Y., Lei, Z.: RCTrans: Radar-camera transformer via radar densifier and sequential decoder for 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5048–5056 (2025)
20. Li, Z., Lan, S., Alvarez, J.M., Wu, Z.: BEVNeXt: Reviving dense BEV frameworks for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20113–20123 (2024)
21. Lin, Z., Liu, Z., Xia, Z., Wang, X., Wang, Y., Qi, S., Dong, Y., Dong, N., Zhang, L., Zhu, C.: 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. pp. 14928–14937 (2024)
22. Liu, S., Cui, M., Li, B., Liang, Q., Hong, T., Huang, K., Shan, Y., Huang, K.: FSHNet: Fully sparse hybrid network for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8900–8909 (2025)
23. Liu, Y., Wang, T., Zhang, X., Sun, J.: PETR: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. pp. 531–548 (2022)
24. Liu, Z., Hou, J., Wang, X., Ye, X., Wang, J., Zhao, H., Bai, X.: LION: Linear group RNN for 3D object detection in point clouds. Advances in Neural Information Processing Systems **37**, 13601–13626 (2024)
25. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 2774–2781 (2023)
26. Palffy, A., Pool, E., Baratam, S., Kooij, J.F., Gavrila, D.M.: 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)
27. Palladin, E., Brucker, S., Ghilotti, F., Narayanan, P., Bijelic, M., Heide, F.: Self-supervised sparse sensor fusion for long range perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27498–27509 (2025)
28. Wang, H., Shi, C., Shi, S., Lei, M., Wang, S., He, D., Schiele, B., Wang, L.: DSVT: Dynamic sparse voxel transformer with rotated sets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13520–13529 (2023)
29. Wang, H., Gao, J., Hu, W., Zhang, Z.: Height-fidelity dense global fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26664–26674 (2025)
30. Wang, T., Zhu, X., Pang, J., Lin, D.: FCOS3D: Fully convolutional one-stage monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 913–922 (2021)
31. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021)

32. Wolters, P., Gilg, J., Teepe, T., Herzog, F., Fent, F., Rigoll, G.: SpaRC: Sparse radar-camera fusion for 3D object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 6589–6596 (2026)
33. Wolters, P., Gilg, J., Teepe, T., Herzog, F., Laouichi, A., Hofmann, M., Rigoll, G.: Unleashing HyDRa: Hybrid fusion, depth consistency and radar for unified 3d perception. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 7467–7474 (2025)
34. Xie, Y., Xu, C., Rakotosaona, M.J., Rim, P., Tombari, F., Keutzer, K., Tomizuka, M., Zhan, W.: SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17591–17602 (2023)
35. Xiong, W., Liu, J., Huang, T., Han, Q.L., Xia, Y., Zhu, B.: LXL: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. IEEE Transactions on Intelligent Vehicles **9**(1), 79–92 (2024)
36. Yang, Z., Yu, Z., Choy, C., Wang, R., Anandkumar, A., Alvarez, J.M.: Improving distant 3D object detection using 2D box supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14853–14863 (2024)
37. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11784–11793 (2021)
38. Zhang, G., Chen, J., Gao, G., Li, J., Hu, X.: HEDNet: A hierarchical encoder-decoder network for 3d object detection in point clouds. Advances in Neural Information Processing Systems **36**, 53076–53089 (2023)
39. Zhang, G., Chen, J., Gao, G., Li, J., Liu, S., Hu, X.: 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. pp. 14477–14486 (2024)
40. Zhang, G., Fan, L., He, C., Lei, Z., Zhang, Z., Zhang, L.: Voxel Mamba: Group-free state space models for point cloud based 3d object detection. Advances in Neural Information Processing Systems **37**, 81489–81509 (2024)
41. Zhang, H., Liang, L., Zeng, P., Song, X., Wang, Z.: SparseLIF: High-performance sparse lidar-camera fusion for 3d object detection. In: European Conference on Computer Vision. pp. 109–128 (2024)
42. Zhao, X., Liu, C., Zhu, M., Zhang, Z., Song, L., Luo, Q., Guo, C., Su, K.: GaussianFusion: Unified 3D gaussian representation for multi-modal fusion perception. In: International Conference on Learning Representations (2026)
43. Zheng, L., Ma, Z., Zhu, X., Tan, B., Li, S., Long, K., Sun, W., Chen, S., Zhang, L., Wan, M., et al.: TJ4DRadSet: A 4d radar dataset for autonomous driving. In: IEEE International Conference on Intelligent Transportation Systems. pp. 493–498 (2022)
44. Zhong, H., Xiang, Z., Xu, R., Fu, J., Xu, P., Wang, S., Yang, Z., Pu, T., Liu, E.: CVFusion: Cross-view fusion of 4D radar and camera for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28188–28197 (2025)
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)