2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SpaRC: Sparse Radar-Camera Fusion for 3D Object Detection

Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Felix Fent, Gerhard Rigoll
Technical University of Munich, Germany

***Abstract*— In this work, we present SpaRC, a novel sparse fusion transformer for 3D perception that integrates multi-view image semantics with Radar and Camera point features. The fusion of radar and camera modalities has emerged as an efficient perception paradigm for autonomous driving systems. While conventional approaches utilize dense Bird's Eye View (BEV)-based architectures for depth estimation, contemporary query-based transformers excel in camera-only detection through object-centric methodology. However, these query-based approaches exhibit limitations in false positive detections and localization precision due to implicit depth modeling. We address these challenges through three key contributions: (1) sparse frustum fusion (SFF) for cross-modal feature alignment, (2) range-adaptive radar aggregation (RAR) for precise object localization, and (3) local self-attention (LSA) for focused query aggregation. In contrast to existing methods requiring computationally intensive BEV-grid rendering, SpaRC operates directly on encoded point features, yielding substantial improvements in efficiency and accuracy. Empirical evaluations on the nuScenes and TruckScenes benchmarks demonstrate that SpaRC significantly outperforms existing dense BEV-based and sparse query-based detectors. Our method achieves state-of-the-art performance of 67.1 NDS and 63.1 AMOTA. The code is available at `https://phi-wol.github.io/sparc/`.**

I. Introduction

Developing efficient, robust, and scalable perception systems for autonomous driving is a challenging task. Autonomous vehicles must accurately perceive their surroundings and make informed decisions in real-time to ensure safe operation in complex dynamic environments like crowded urban scenarios and fast-paced highways. This requires precise localization and classification of other traffic participants [1,2]. Multi-modal sensor fusion of LiDAR, camera, and radar has made significant progress in recent years due to large-scale, diverse datasets [3]–[6] and advances in deep learning architectures [2,7]–[9]. While LiDAR-based methods achieve impressive performance [8,10,11], their high cost and maintenance requirements limit widespread deployment. This has motivated research into more cost-effective sensor combinations, particularly camera-radar fusion [12]–[15].

Cameras provide high-resolution semantic information and capture rich texture details but struggle with depth estimation and perform poorly in adverse lighting and weather conditions like night, fog, or snow [16]. In contrast, Millimeter-wave radar sensors offer sparse, metric range sensing and Doppler-based velocity measurements even under adverse lighting and weather conditions. Combined in a complementary architecture, they have the potential to unlock reliable and affordable 3D perception for autonomous driving [17]. The main barrier to radar-based perception has been the lack of high-quality and large-scale sensor recordings. Out of the traditional autonomous driving datasets [3]–[6], only nuScenes offers a limited and noisy radar sensor suite. With new datasets like TruckScenes [18] and the BSD dataset [19] this is going to change. They offer modern 4D imaging radar sensors, long-range annotations up to 150m, and a diverse set of scenarios, including various weather conditions. However, bridging the view disparity between the dense camera images and the sparse radar representation remains a key challenge due to the unique characteristics of the radar sensor [17]: low angular resolution, only a few reflected points per object, noise, and clutter due to multi-path reflections capture the intricacies of radar-based perception and require an adaptive fusion design [20].

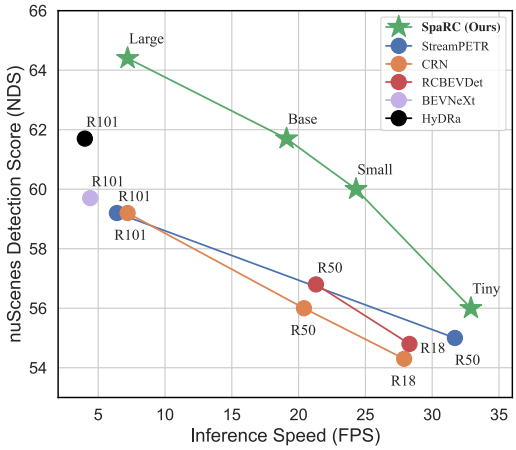

Fig. 1. Compared to previous real-time models our **SpaRC** model family achieves state-of-the-art performance in accuracy and inference speed (measured with a single consumer-grade RTX3090 GPU on nuScenes val)

Most existing methods are based on LiDAR-centric architectures that utilize dense point-cloud processing backbones like PointPillars [21] with Bird's-Eye-View (BEV) feature extraction and fusion mechanisms, which have become the default choice for 3D object detection [8,22]–[25]. However, directly applying these dense BEV representations to sparse radar data leads to computational inefficiency, as most grid cells remain empty. Recent work has focused on adapting these LiDAR-centric designs for camera-radar fusion through various strategies: radar-aided depth estimation [13,14,26], modified grid-rendering backbones [15,27], and adaptive fusion mechanisms [13]–[15]. Even with recent improvements in BEVFusion-style architectures, this fundamental mismatch between dense representations and sparse radar signals remains a key limitation for efficient optimization.

979-8-3315-8160-2/26/$31.00 ©2026 IEEE

In contrast, we propose a novel query-based fusion transformer for 3D object detection that concentrates computational resources on salient regions of the radar modality. We disregard the BEV-grid representation due to its sparseness in feature representation and opt for an object-centric paradigm. Introducing **SpaRC**, we achieve a new state-of-the-art in camera-radar perception with strong robustness, high accuracy, and real-time inference speed. Our contributions are:

- We utilize a modality-specific sparse feature set representation for radar encoding.
- We design a multi-scale but Sparse Frustum Fusion for efficient cross-modal feature alignment, improving the projection-based representation and explicit depth estimation.
- We propose a range-adaptive radar refinement and a local self-attention mechanism to model the intuitive object-to-point interactions and improve the implicit depth learning.
- **SpaRC** achieves state-of-the-art performance on the nuScenes benchmark ($+2.9$ NDS and $+2.6$ mAP). Moreover, our findings generalize to the long-range and adverse conditions on the new TruckScenes benchmark and match the LiDAR-based baseline.

II. Related Work

A. Dense BEV-based 3D Perception

Since the seminal work of LSS [28], vision-centric 3D perception has moved from the perspective view [7,29,30] to a unified Bird's-Eye-View (BEV) space [28,31]–[33]. The 3D space representation has been proven to be beneficial for unified multi-view and point-cloud fusion, as well as downstream tasks such as mapping, tracking and planning [1, 34,35]. Several differentiable lifting strategies have been proposed to transform 2D image features into the BEV [34,36]. Most prominent, the convolution-based BEVDet Series [32, 37]–[39] introduce efficient forward view-transformation, explicit depth prediction, and ego-motion-based temporal modeling. Contrary, BEVFormer [31,40] queries the image features using 3D-to-2D cross-attention, modeling the inverse and implicit camera un-projection.

B. Developing Radar-Camera Fusion Systems

Radar-camera fusion addresses the core challenge of vision-centric systems: precise and robust depth estimation. Incorporating mmWave radar features into different stages of the detection architecture, the sparse range, and Doppler measurements reduce the overall localization errors and improve velocity estimation. Hence, bridging the view disparity between the two feature spaces is an active area of research: Early works [26,41]–[44] focus on projective fusion in the image space. These methods first project 3D radar points into the 2D image plane and then perform late-stage feature association through high-level feature concatenation or Region-of-Interest (ROI) pooling. The fused features are used to refine image-based object proposals by incorporating radar's precise range measurements. More recent works have explored alternative radar feature extraction methods. RadarGNN [45] models point-pair relationships through graph neural networks, while X3KD [46], RadarDistill [12], and CRKD [47] leverage cross-modal knowledge distillation to enhance radar feature learning.

Following the success of BEVFusion [8,22,34,48,49], dense fusion in BEV space through concatenation, summation, or SE-Blocks has emerged as the dominant paradigm. These methods typically pair dense BEV-based 3D object detectors with grid-based radar feature encoders like PointPillars [21,50]. The fused features are decoded into 3D object proposals using dense detection heads like CenterPoint [35]. Originally designed for LiDAR-centric perception, these methods have been adapted to better handle sparsity, calibration errors, and noise interference. RCM-Fusion [51] relies on an instance-level refinement within the dense BEV-grid. While CRN [13] and RCBEVDet [15] upgrade the BEV-fusion with deformable cross-attention for increased receptive fields, HyDRa [14] introduces a hybrid fusion, leveraging multi-modal depth estimation and a radar-guided backpropagation for refinement.

Despite their success, current BEV-based fusion methods face several key challenges. First, the effectiveness of BEV feature maps deteriorates significantly with distance - only about 50% of grid cells receive valid projected image features [52]. This sparsity is even more pronounced for radar features, where point-pillar encoders typically populate just 1-5% of the grid cells with radar points, leading to inefficient dense representations of inherently sparse information. Second, state-of-the-art approaches [13]–[15] rely heavily on ego-motion-based temporal feature warping and require large receptive fields to compensate for the sparse nature of the features. This becomes particularly problematic for long-range perception [6,18], where computational complexity increases quadratically with range. In contrast, our method addresses the limitations through an object-centric set-to-set fusion. By operating directly on sparse point-based representations rather than dense grids, we maintain information density while reducing computational overhead. Our point-to-point interaction focuses only on the local neighborhood of object queries, enabling stable optimization and long-range perception without the quadratic scaling of grid-based methods.

C. Sparse Query-based Perception

Sparse query-based methods have been inspired by the DEtection TRansformer (DETR) [53]–[55] and emerge as a powerful and efficient alternative to grid-based methods. The PETR-Series [9,56] models a small set of object queries with a 3D position embedding and encodes them with cross-attention. Multi-head self-attention exhibits the role of the BEV encoder [57]. SparseBEV [57] and Sparse4D [58] enable spatio-temporal sampling from 3D queries in 2D feature maps by projecting deformable sampling offsets onto the 2D feature maps. Instead of stacking ego-motion compensated BEV-grids [13]–[15], StreamPETR [59] follows up with a temporal propagation module to iteratively refine the object queries from history queries. Far3D [60] shows that the

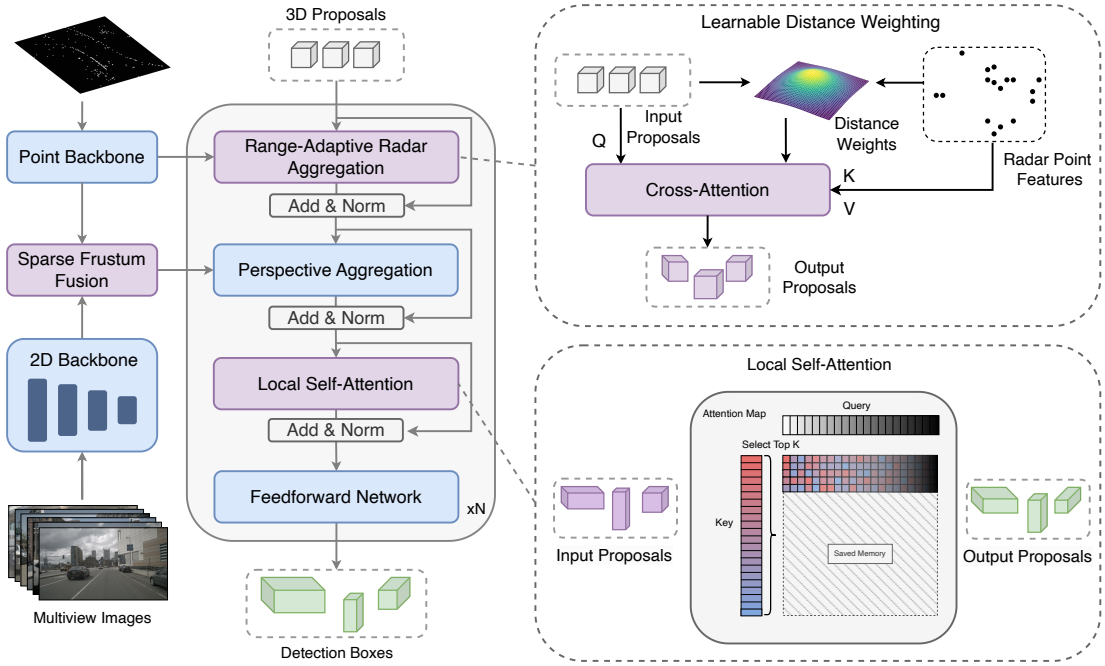

Fig. 2. **Architecture of SpaRC**. A fully-sparse object-centric radar-camera fusion 3D object detector. Radar points are sparsely encoded as a 3D-embedded feature set. Objects are modeled as a set of queries. Range-adaptive Radar Attention encodes the queries to salient region proposals, introducing strong spatial priors and inducing object-specific Doppler velocity. The radar-guided deformable cross-attention is applied to frustum-fused perspective features. Local top K self-attention focuses the object filtering and encoding on the local region around the 3D location. The fused features are decoded into 3D object proposals using a sparse detection head.

sparse design is also beneficial for long-range detection with strong object recall when employing a perspective 2D object head with a depth network for dynamic query initialization. Follow-up works [61]–[63] concentrate on different denoising strategies and positional encodings to increase the robustness and reduce false positives of ambiguous feature sampling along projection rays. Due to the implicit depth modeling of the 3D-located queries, these methods achieve strong recall but suffer from false positives and localization errors. Our object-centric radar fusion addresses the limitations of implicit localization by backprojecting and sampling from salient radar points. This reduces false correspondences between 3D and 2D space. The temporal and spatial filtering of queries can be reduced to focused local regions. Doppler velocity has a synergistic effect and is naturally suited for object-level motion modeling and compensation.

III. SpaRC Architecture

We introduce **SpaRC**, a novel **Spa**rse fusion transformer for 3D perception from **R**adar and **C**amera. Our model processes multi-view RGB images and radar point clouds in parallel streams: images are encoded by convolutional feature extractors with FPN [64], while radar points are processed by a transformer-based point encoder. The resulting features are fused in two stages: First, radar features are projected into image space and associated with semantic feature maps. Second, a sparse set of 3D object queries initialized from perspective proposals and spatially distributed 3D queries, aggregates multi-modal information through cross-attention. Range-adaptive radar refinement guides the object-radar interaction based on distance, while deformable attention in perspective space captures the fused semantic features. The model maintains a high recall through its implicit design while increasing precision through strong spatial cues in the query decoder.

A. Radar Point Encoder

We employ a lightweight point transformer [65,66] to extract features from radar point clouds. The encoder transforms unstructured radar points into a sparse but information-dense representation through space-filling curves and serialized neighbor mapping. By grouping points into non-overlapping patches and performing within-patch attention, we efficiently model spatial relationships without constructing and processing dense grids.

The 3D points are encoded in the same positional embedding space as the object queries, enabling direct interaction in later fusion stages. To prevent overfitting on the small radar point cloud while maintaining real-time performance, we adopt a reduced version of the backbone with implementation details provided in the code repository.

B. Sparse Frustum Fusion

As visualized in Fig. 3, we propose a Sparse Frustum Fusion (SFF) module to efficiently associate radar and image features in perspective space. The encoded radar feature vectors are first projected into the camera frustum space and filtered per view. We embed the projected coordinates (depth and horizontal pixel position) in a learnable positional

Methods	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-V [7]	L	Voxel	-	65.3	56.9	0.285	0.253	0.323	0.272	0.186
RCM-Fusion [51]	C+R	R50	450 × 800	52.9	44.3	-	-	-	-	-
X3KD [46]	C+R	R50	256 × 704	53.8	42.3	-	-	-	-	-
StreamPETR [59]	C	R50	256 × 704	54.0	43.2	0.581	0.272	0.413	0.295	0.195
CRN [13]	C+R	R50	256 × 704	56.0	49.0	0.487	0.277	0.542	0.344	0.197
RCBEVDet [15]	C+R	R50	256 × 704	56.8	45.3	0.486	0.285	0.404	0.220	0.192
HyDRa [14]	C+R	R50	256 × 704	58.5	49.4	0.463	0.268	0.478	0.227	0.182
SpaRC	C+R	R50	256 × 704	**62.0**	**54.5**	0.496	0.269	0.403	0.177	0.181
MVFusion [67]	C+R	R101	900 × 1600	45.5	38.0	0.675	0.258	0.372	0.833	0.196
FUTR3D [49]	C+R	R101	900 × 1600	50.8	39.9	-	-	-	0.561	-
StreamPETR [59]	C	R101	512 × 1408	59.2	50.4	0.569	0.262	0.315	0.257	0.199
CRN [13]	C+R	R101	512 × 1408	59.2	52.5	0.460	0.273	0.443	0.352	0.180
Far3D [60]	C	R101	512 × 1408	59.4	51.0	0.551	0.258	0.372	0.238	0.195
BEVNeXt [68]	C	R101	512 × 1408	59.7	50.0	0.487	0.260	0.343	0.245	0.197
HyDRa [14]	C+R	R101	512 × 1408	61.7	53.6	0.416	0.264	0.407	0.231	0.186
SpaRC	C+R	R101	512 × 1408	**64.4**	**57.1**	0.484	0.264	0.308	0.175	0.178

TABLE I

3D OBJECT DETECTION ON NUSCENES VAL SET. 'L', 'C', AND 'R' REPRESENT LIDAR, CAMERA, AND RADAR, RESPECTIVELY.

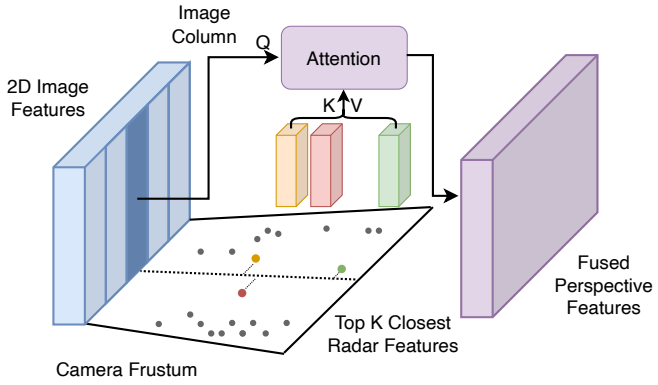

Fig. 3. **Sparse Frustum Fusion Visualization.** Encoded radar points are projected into the camera frustum space and associated with semantic image feature columns. Per image column, only the K nearest points and embeddings are used for cross-attention. This enables the subsequent perspective 3D head to also benefit from the radar-fusion.

encoding, while image features are embedded using their downsampled pixel positions. For each image column [14], we query the K nearest radar points along the vertical dimension and fuse them through cross-attention. This enables soft association between modalities without requiring noisy depth maps or pillar representations. While we explicitly project radar points using their range measurements, the cross-attention mechanism allows the model to handle uncertainties from noisy measurements and missing height information. This can be leveraged for each feature map level of the multi-scale feature extractor.

The computational complexity of SFF is $\mathcal{O}(WK)$, where W is the downsampled image width, and K is a hyperparameter defining the number of nearest radar points per column that are considered. With typical values of $W, K < 100$, this leads to efficient parallel computation. By operating in perspective space, radar features benefit from fine-grained image supervision while preserving their 3D spatial information. The learnable depth and positional embeddings allow the model to explicitly align features and handle noise, establishing semantically meaningful associations between radar points and semantically connected image regions. The subsequent perspective head can now dynamically allocate stronger 3D proposals, which initialize 3D object queries.

C. Range Adaptive Radar Aggregation

Objects are modeled as 3D reference points and semantic context features, which later are decoded into 3D bounding boxes with localization offsets, size, orientation, and velocity. Drawing inspiration from multi-scale self-attention mechanisms [57], we propose a Range-Adaptive Radar (RAR) aggregation decoder layer that dynamically adjusts feature interactions based on spatial relationships. As each object is represented by a reference point in 3D space and a context embedding that encodes semantic information, we can directly associate radar points through a set-to-set interaction with the object queries and update the context embedding.

Specifically, we formulate a distance-aware attention mechanism that adaptively weights radar features based on their proximity to object centers:

$$\text{Attn}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}} - \alpha \frac{\|\mathbf{p}_q - \mathbf{p}_k\|_2}{r_{\text{max}}}\right)\mathbf{v} \quad (1)$$

where query features $\mathbf{q} \in \mathbb{R}^{N_q \times d}$ interact with radar-derived key-value pairs $\mathbf{k}, \mathbf{v} \in \mathbb{R}^{N_k \times d}$ through scaled dot-product attention, modulated by a learnable distance penalty. Here, $\mathbf{p}_q \in \mathbb{R}^{N_q \times 3}$ and $\mathbf{p}_k \in \mathbb{R}^{N_k \times 3}$ represent the 3D positions of queries and radar points respectively, normalized by the maximum detection range r_{max}. The learnable parameter α controls the strength of the spatial bias.

By embedding both queries and radar points in a shared continuous 3D space, RAR enables direct point-to-point interactions without requiring dense and discretized grid

representations. This distance-guided attention mechanism naturally focuses on locally relevant radar features while maintaining the ability to capture longer-range dependencies when needed. The radar-aware query features provide strong spatial priors that guide subsequent cross-modal fusion through deformable attention, effectively highlighting image regions that align with reliable and salient radar reflections.

The RAR module provides two key benefits: First, it reduces local uncertainty by incorporating precise radar depth measurements before queries interact with image features This helps resolve depth ambiguities that typically plague pure vision-based approaches. Second, it enables object-level motion modeling by directly associating radar Doppler measurements with object queries rather than trying to infer motion from grid-based feature maps. This sequential fusion approach - first establishing strong spatial priors through radar, then refining with dense image features - leads to more robust 3D object detection compared to methods that rely solely on back-projection or dense grid-based fusion.

D. Local Self-Attention

Traditional DETR-like architectures employ global self-attention in their decoder blocks, where each object query attends to all other queries. However, we observe that for 3D object detection, queries primarily need to interact with their spatial neighbors that represent the same or nearby objects. We propose a Local Self-Attention (LSA) mechanism that significantly improves both efficiency and effectiveness.

Our decoder processes three types of object queries: dynamically allocated queries from the perspective head, randomly initialized 3D queries, and temporal history queries from the memory queue. Moreover, we restructure the decoder block to apply self-attention at the end, after cross-modal feature aggregation, updating the queries from all modalities before associating them with history queries. This allows queries to first gather relevant features before determining their spatial relationships and filtering out the duplicates and false positives.

The key innovation of LSA is to restrict each query's attention to only its k-nearest neighbors in 3D space. For each query, we compute distances to all other queries and select the top-k closest ones as its attention context. This local neighborhood typically contains queries that project onto the same image regions or represent temporally consistent and propagated proposals. By focusing on spatially proximate queries, LSA helps establish more meaningful relationships between queries that likely correspond to the same physical object. It significantly reduces computational complexity from $\mathcal{O}(N^2)$ for global attention to $\mathcal{O}(NK)$, where N is the total number of queries and K is the number of neighbors. The distance computation is performed only once in the first decoder layer, and the same local attention pattern can be reused in subsequent layers. The receptive field can still grow across multiple decoder layers as information propagates through overlapping local neighborhoods.

Methods	Input	Backbone	NDS↑	mAP↑
CenterPoint [7]	L	Voxel	67.3	60.3
KPConvPillars [69]	R	Pillars	13.9	4.9
RadarDistill [12]	R	Pillars	43.7	20.5
CenterFusion [43]	C+R	DLA34	44.9	32.6
MVFusion [67]	C+R	V2-99	51.7	45.3
CRAFT [44]	C+R	DLA34	52.3	41.1
BEVFormerV2 [40]	C	InternImage-B	62.0	54.0
CRN [13]	C+R	ConvNeXt-B	62.4	57.5
StreamPETR [59]	C	V2-99	63.6	55.0
RCBEVDet [15]	C+R	V2-99	63.9	55.0
HyDRa [14]	C+R	V2-99	64.2	57.4
SpaRC	C+R	V2-99	**67.1**	**60.0**

TABLE II

3D OBJECT DETECTION ON THE NUSCENES TEST SET.

Methods	Input	Backbone	AMOTA↑	AMOTP↓
CenterPoint [7]	L	Voxel	63.8	0.555
ByteTrackV2 [70]	C	V2-99	56.4	1.005
StreamPETR [71]	C	ConvNeXt-B	56.6	0.975
CRN [13]	C+R	ConvNeXt-B	56.9	**0.809**
HyDRa	C+R	V2-99	58.4	0.950
SpaRC	C+R	V2-99	**63.1**	0.901

TABLE III

3D OBJECT TRACKING ON NUSCENES TEST SET.

E. Real-time Object Detection

To achieve real-time performance, we introduce a family of models with varying complexity and speed-accuracy trade-offs. Our Tiny model (ResNet-18 backbone, 4 decoder layers) achieves over 30 FPS, while our Large model (ResNet-101, 6 layers) operates at 7 FPS with state-of-the-art accuracy (Fig. 1). We employ several key optimizations: First, we drop the perspective head from [60], which significantly reduces inference time with only marginal accuracy impact. Second, we leverage CUDA streams to parallelize radar and camera backbone processing. Third, our Local Self-Attention mechanism reduces computational complexity while maintaining detection quality. Detailed architecture specifications (number of decoder layers, queries, ResNet backbones, PointTransformer settings) can be found in the Code Repository.

IV. EXPERIMENTS

A. Datasets

We use the two large-scale radar datasets, nuScenes [4] and the new TruckScenes [18] to explore, generalize, and validate the findings on our SpaRC model architecture.

The CVPR **nuScenes** dataset [4] is the traditional research benchmark for radar-fusion-based 3D perception. In the urban scenario of Boston and Singapore, 1000 scenes of 20s are captured with six cameras, five radar, and one LiDAR sensor. The annotation range is 50 m.

Recently, the NeurIPS **TruckScenes** [18] benchmark was introduced to provide high-quality and modern 4D radar point clouds and diverse scenes for autonomous trucking.

Methods	Input	Backbone	Image Size	Split	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
CenterPoint-V* [7]	L	Voxel	-	val	35.3	22.6	0.461	0.405	0.468	3.028	0.261
Far3D [60]	C	V2-99	640 × 960	val	21.4	10.7	0.883	0.507	0.671	1.352	0.338
SpaRC	C+R	V2-99	640 × 960	val	**35.4**	**22.5**	0.798	0.449	0.476	0.613	0.248
CenterPoint-V* [7]	L	Voxel	-	test	41.0	26.7	0.409	0.352	0.277	2.730	0.201
RadarGNN* [45]	R	-	-	test	10.7	7.0	0.892	0.809	1.132	8.003	0.571
PETR* [9]	C	V2-99	300 × 800	test	12.1	2.2	1.125	0.686	0.647	1.499	0.564
HyDRa [14]	C+R	V2-99	928 × 1952	test	22.4	12.8	0.725	0.544	0.744	1.180	0.388
SpaRC	C+R	V2-99	928 × 1952	test	**37.4**	**27.2**	0.759	0.413	0.411	0.814	0.227

TABLE IV

3D OBJECT DETECTION ON TRUCKSCENES VAL AND TEST SETS. THE DETECTION RANGE IS UP TO 150 M (*DENOTES THE OFFICIAL BASELINES).

Four cameras, six LiDAR, and six 4D imaging radar sensors capture 740 scenes of 20s in 360 degree coverage. The biggest differentiation is the annotation range of 150 m, dynamic faster speeds of highway driving, and diversity of adverse splits, making it challenging for single-modal perception systems.

B. Implementation Details

We adopt StreamPETR [59] and Far3D [60] as our baseline for the camera stream and follow standard practices for training and hyperparameters [59]. For a fair comparison, we employ pretrained ResNet [72] and V2-99 [73] backbone encoders. For the radar stream, we utilize a downscaled PointTransformerV3 [66] with randomly initialized weights on multiple sweeps of radar, on RCS and Doppler features. We use the 16 closest radar points in the frustum space whereas the LSA module leverages a local neighborhood of 32 queries (instead of the default setting of 644 normal + 256 temporal + 600 denoising queries [59,60]).

Like CRN and RCBEVDet, the inference time is measured on an RTX3090 GPU (single batch, FP16 precision). No test-time augmentations, CBGS or future frames are used.

C. Main Results

We compare SpaRC to the previous state-of-the-art methods on the val and test sets of nuScenes and TruckScenes.

nuScenes Val. As reported in Tab. I, SpaRC consistently outperforms both BEV-based as well as query-based methods in terms of NDS and mAP (+5.1 for R50 and +3.5 for R101). Notably, the object-level motion modeling can benefit greatly from the Doppler velocity, as shown by the large improvement in mAVE. Especially in the small scale and low resolution (real-time) scenarios, the performance gain over the vision-based methods is significant.

nuScenes Test. When scaling up to the V2-99 backbone and evaluating on the test server (Tab. II and Tab. III), SpaRC introduces a new state-of-the-art in 3D object detection on nuScenes, with an NDS of 67.1 (+2.9) and mAP of 60.0 (+2.6), surpassing the previous best camera- or radar-based methods. Capitalizing on the high accuracy, strong motion modeling and paired with a velocity-based greedy tracker [7], SpaRC achieves also the best tracking-by-detection performance, increasing the AMOTA to 63.1 (+4.7).

SFF	RAR	LSA	Input	NDS↑	mAP↑	mATE↓	mAVE↓
			C	48.6	37.5	0.672	0.309
✓			C+R	51.4	40.7	0.644	0.252
	✓		C+R	53.5	43.8	0.587	0.199
		✓	C	49.8	38.9	0.645	0.301
	✓	✓	C+R	53.9	44.8	0.590	0.198
✓	✓		C+R	54.5	45.1	0.581	0.205
✓	✓	✓	C+R	**54.9**	**46.4**	**0.580**	**0.195**

TABLE V

ABLATION OF SPARC COMPONENTS ON NUSCENES VAL SET

TruckScenes Val. Our architecture generalizes well to the new domain and sensor setup of TruckScenes, achieving a competitive NDS of 37.4 on the validation set (Tab. IV). With more adverse conditions and longer detection ranges of up to 150 meters, the adoption of radar becomes more important, as SpaRC doubles the mAP (+11.8) over the current vision-only state-of-the-art. We provide more detailed information about the ranges, conditions, and classes in the Appendix.

TruckScenes Test. SpaRC surpasses all single-modal baselines and achieves competitive mAP scores to the LiDAR model (. Tab. IV). We set the state-of-the-art for vision-centric and radar-based methods on TruckScenes, with a long-range NDS score of 37.4, demonstrating the effectiveness of our object-centric architecture and the inclusion of 4D radar. Moreover, we emphasize the importance of Doppler information in high-speed and dynamic scenarios. While radar-only methods struggle with overall detection performance, camera-radar fusion effectively leverages Doppler measurements to achieve strong velocity prediction accuracy.

D. Ablation Studies

Component Analysis. We conduct extensive ablation studies to analyze the effectiveness of each component in our model. As shown in Tab. V, we start from a camera-only baseline using Far3D [60] trained for 24 epochs. Adding our Local Self-Attention (LSA) module improves performance by 1.2 NDS and 1.4 mAP, demonstrating the benefits of focusing attention on locally relevant features even without radar input. The efficiency gains from LSA enable incorporating additional fusion components. Introducing the Range-Adaptive Radar (RAR) module yields substantial improvements of 4.1 NDS

and 5.9 mAP. By leveraging Doppler velocity information through motion-aware layer normalization, RAR reduces velocity errors (mAVE) by 34% while also improving localization accuracy (mATE). Finally, the Sparse Frustum Fusion (SFF) module further boosts performance by 1.0 NDS and 1.6 mAP by effectively incorporating radar point features in the perspective view. The ablation results validate that each component contributes meaningfully to the final performance, with the full model achieving significant gains of 6.3 NDS and 8.9 mAP over the camera-only baseline. This demonstrates the effectiveness of our sparse fusion design in leveraging complementary radar information.

In Tab. VI, we analyze the impact of the range guidance of the RAR module. When excluding the range-modulation (-2.9 mAP) or the hierarchical point structuring (-1.9 mAP), both components show a significant decrease in performance. The best performance is achieved when including the locality bias and point-based modeling.

Achieving Real-time Speed. We analyze the inference latency vs. performance trade-off in Fig. 1 across backbone architectures (R18, R50, R101). The results demonstrate that SpaRC achieves superior accuracy at lower computational cost compared to existing methods, validating the efficiency advantages of our sparse object-centric fusion.

Methods	Input	NDS↑	mAP↑	mATE↓	mAVE↓
Camera Baseline [60]	C	48.6	37.5	0.672	0.309
SpaRC RAR V1	C+R	52.3	41.9	0.612	0.209
SpaRC RAR V2	C+R	52.3	42.9	0.608	0.246
SpaRC RAR V3	C+R	**53.9**	**44.8**	**0.590**	**0.198**

TABLE VI

ABLATION VARIANTS OF SPARC RANGE ADAPTIVE RADAR AGGREGATION (RAR) MODULE ON NUSCENES VAL SET (V1 - WITHOUT FOCUSED RANGE GUIDANCE; V2 - WITHOUT SERIALIZED POINT ENCODING; V3 - THE FULL RAR MODULE).

V. CONCLUSION

In this paper, we introduce **SpaRC**, a novel camera-radar fusion transformer for 3D object detection. We overcome the limitations of dense BEV-based fusion methods and address key challenges of query-based architectures. By introducing sparse but strong local cues into the decoder, we concentrate computational resources on salient regions of the radar modality and reduce the uncertainty of implicit 3D decoding. This information-rich but sparse representation achieves superior performance in accuracy and robustness over all existing vision-centric and radar-based methods. We achieve a new state-of-the-art in camera-radar fusion on the nuScenes and TruckScenes benchmarks in both short-range urban environments and dynamic long-range highway scenarios. Our real-time capable architecture provides an efficient and scalable solution for autonomous driving perception, bridging the gap to LiDAR-centric methods.

REFERENCES

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023, pp. 17 853–17 862.

[2] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE TPAMI*, 2023.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*. IEEE, 2012, pp. 3354–3361.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[5] 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 *CVPR*, 2020, pp. 2446–2454.

[6] 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.

[7] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *CVPR*, 2021, pp. 11 784–11 793.

[8] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *NeurIPS*, vol. 35, 2022.

[9] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *ECCV*, 2022, pp. 531—548.

[10] 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 *CVPR*, 2023, pp. 8394–8405.

[11] H. Wang, C. Shi, S. Shi, M. Lei, S. Wang, D. He, B. Schiele, and L. Wang, "Dsvt: Dynamic sparse voxel transformer with rotated sets," in *CVPR*, 2023, pp. 13 520–13 529.

[12] G. Bang, K. Choi, J. Kim, D. Kum, and J. W. Choi, "Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features," in *CVPR*, 2024, pp. 15 491–15 500.

[13] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in *ICCV*, 2023.

[14] P. Wolters, J. Gilg, T. Teepe, F. Herzog, A. Laouichi, M. Hofmann, and G. Rigoll, "Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 7467–7474.

[15] 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 *CVPR*, 2024, pp. 14 928–14 937.

[16] K. Yoneda, N. Suganuma, R. Yanase, and M. Aldibaja, "Automated driving recognition technologies for adverse weather conditions," *IATSS research*, vol. 43, no. 4, pp. 253–262, 2019.

[17] 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*, 2023.

[18] F. Fent, F. Kuttenreich, F. Ruch, F. Rizwin, S. Juergens, L. Lechermann, C. Nissler, A. Perl, U. Voll, M. Yan, *et al.*, "Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions," *arXiv preprint arXiv:2407.07462*, 2024.

[19] K. Armanious, M. Quach, M. Ulrich, T. Winterling, J. Friesen, S. Braun, D. Jenet, Y. Feldman, E. Kosman, P. Rapp, *et al.*, "Bosch street dataset: A multi-modal dataset with imaging radar for automated driving," *arXiv preprint arXiv:2407.12803*, 2024.

[20] 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*, 2024.

[21] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019, pp. 12 697–12 705.

[22] 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," in *ICRA*, 2023.

[23] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, and X. Zhang, "Cross modal transformer: Towards fast and robust 3d object detection," in *ICCV*, 2023, pp. 18 268–18 278.

[24] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "Deepinteraction: 3d object detection via modality interaction," *NeurIPS*, vol. 35, pp. 1992–2005, 2022.

[25] J. Huang, Y. Ye, Z. Liang, Y. Shan, and D. Du, "Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection," *arXiv preprint arXiv:2311.07152*, 2023.

[26] Y. Long, A. Kumar, D. Morris, X. Liu, M. Castro, and P. Chakravarty, "Radiant: Radar-image association network for 3d object detection," in *AAAI*, 2023.

[27] A. Musiat, L. Reichardt, M. Schulze, and O. Wasenmüller, "Radarpillars: Efficient object detection from 4d radar point clouds," *arXiv preprint arXiv:2408.05020*, 2024.

[28] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *ECCV*, 2020, pp. 194–210.

[29] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *ICCV*, 2021, pp. 913–922.

[30] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[31] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *ECCV*, 2022.

[32] J. Huang, G. Huang, Z. Zhu, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," in *arXiv preprint arXiv:2112.11790*, 2021.

[33] T. Teepe, P. Wolters, J. Gilg, F. Herzog, and G. Rigoll, "Earlybird: Early-fusion for multi-view tracking in the bird's eye view," in *WACV*, 2024, pp. 102–111.

[34] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-bev: What really matters for multi-sensor bev perception?" in *ICRA*, 2023, pp. 2759–2765.

[35] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[36] T. Teepe, P. Wolters, J. Gilg, F. Herzog, and G. Rigoll, "Lifting multi-view detection and tracking to the bird's eye view," in *CVPRW*, 2024, pp. 667–676.

[37] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *AAAI*, 2023.

[38] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo," in *AAAI*, 2023.

[39] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection," in *ICLR*, 2023.

[40] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *CVPR*, 2023.

[41] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, 2019.

[42] Y. Kim, J. W. Choi, and D. Kum, "Grif net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2020.

[43] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *WACV*, 2021, pp. 1527–1536.

[44] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer," in *AAAI*, 2023.

[45] F. Fent, P. Bauerschmidt, and M. Lienkamp, "Radargnn: Transformation invariant graph neural network for radar-based perception," in *CVPR*, 2023, pp. 182–191.

[46] M. Klingner, S. Borse, V. R. Kumar, B. Rezaei, V. Narayanan, S. Yogamani, and F. Porikli, "X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection," in *CVPR*, 2023.

[47] L. Zhao, J. Song, and K. A. Skinner, "Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation," in *CVPR*, 2024.

[48] Y. Man, L.-Y. Gui, and Y.-X. Wang, "Bev-guided multi-modality fusion for driving perception," in *CVPR*, 2023.

[49] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *CVPR*, 2022.

[50] J. Li, C. Luo, and X. Yang, "Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds," in *CVPR*, 2023, pp. 17 567–17 576.

[51] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, "Rcm-fusion: Radar-camera multi-level fusion for 3d object detection," in *ICRA*. IEEE, 2024, pp. 18 236–18 242.

[52] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, "Fb-bev: Bev representation from forward-backward view transformations," in *ICCV*, 2023.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[54] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[55] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *ICLR*, 2023.

[56] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, and X. Zhang, "Petrv2: A unified framework for 3d perception from multi-camera images," in *ICCV*, 2023, pp. 3262–3272.

[57] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "Sparsebev: High-performance sparse 3d object detection from multi-camera videos," in *ICCV*, 2023.

[58] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion," *arXiv preprint arXiv:2211.10581*, 2022.

[59] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *ICCV*, 2023.

[60] 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 *AAAI*, vol. 38, no. 3, 2024, pp. 2561–2569.

[61] 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 *ECCV*. Springer, 2025, pp. 200–217.

[62] J. Hou, T. Wang, X. Ye, Z. Liu, S. Gong, X. Tan, E. Ding, J. Wang, and X. Bai, "Open: Object-wise position embedding for multi-view 3d object detection," in *ECCV*. Springer, 2025.

[63] Y. Tang, Z. Meng, G. Chen, and E. Cheng, "Simpb: A single model for 2d and 3d object detection from multiple cameras," in *ECCV*. Springer, 2025.

[64] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[65] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *ICCV*, 2021, pp. 16 259–16 268.

[66] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler faster stronger," in *CVPR*, 2024, pp. 4840–4851.

[67] Z. Wu, G. Chen, Y. Gan, L. Wang, and J. Pu, "Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion," in *ICRA*, 2023.

[68] Z. Li, S. Lan, J. M. Alvarez, and Z. Wu, "Bevnext: Reviving dense bev frameworks for 3d object detection," in *CVPR*, 2024, pp. 20 113–20 123.

[69] M. Ulrich, S. Braun, D. Köhler, D. Niederlöhner, F. Faion, C. Gläser, and H. Blume, "Improved orientation estimation and detection with hybrid object detection networks for automotive radar," in *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2022.

[70] Y. Zhang, X. Wang, X. Ye, W. Zhang, J. Lu, X. Tan, E. Ding, P. Sun, and J. Wang, "Bytetrackv2: 2d and 3d multi-object tracking by associating every detection box," *arXiv preprint arXiv:2303.15334*, 2023.

[71] J. Yang, E. Yu, Z. Li, X. Li, and W. Tao, "Quality matters: Embracing quality clues for robust 3d multi-object tracking," *arXiv preprint arXiv:2208.10976*, 2022.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[73] Y. Lee, J.-w. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and gpu-computation efficient backbone network for real-time object detection," in *CVPRW*, 2019, pp. 0–0.