Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

UGBCF-Net: Uncertainty-Guided BEV Cross-Attention Fusion Network for Robust Radar-Camera Object Detection

Bhaskar S V.[1], Roja Reddy B.[2]

[1]Department of Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, Karnataka, India.
[2]Department of Electronics and Telecommunication Engineering, RV College of Engineering, Bengaluru, Karnataka, India.

E-mail: [1]getbhaskar4you@gmail.com, [2]rojareddyb@rvce.edu.in

Orcid ID: [1]0009-0006-1389-3786, [2]0000-0003-0748-9753

Abstract

Object detection under adverse weather remains a major challenge for autonomous driving systems. Cameras do not work well during rain, fog, and darkness, while mmWave radar works accurately in all of these conditions, regardless of illumination and precipitation. In this paper, we propose UGBCF-Net – the Uncertainty-Guided Bird's-Eye-View Cross-Attention Fusion Network, which incorporates radar confidence estimation based on physical laws and adaptive multimodal fusion. Our proposed architecture estimates radar confidence from radar cross-section measurements, maintains the Doppler channel by means of four-channel BEV representation, and uses uncertainty-guided weighting with a lightweight pooled cross-attention network for efficient feature fusion. Experimental results on the nuScenes v1.0-trainval dataset have shown that our UGBCF-Net reaches 0.683 mAP with 8.9 M parameters, operating at 56 FPS on a single NVIDIA RTX 3050 GPU. UGBCF-Net has achieved an accuracy similar to large-fusion networks and showed great robustness, reaching night recall equal to 0.884 with recall rates of 93.1%, 95.9% and 97.4% for vehicles, pedestrians, and cyclists, respectively.

Keywords: Autonomous Driving, 3D Object Detection, Radar–Camera Fusion, Bird's-Eye-View (BEV), Uncertainty Estimation, Cross-Attention Networks, Millimeter-Wave Radar, Deep Learning.

1. Introduction

The progress of Intelligent Transportation Systems (ITS) and Autonomous Vehicles (AVs) has been made in response to the demand for a robust perception of the environment, which is a fundamental pillar of trajectory planning and safe navigation [1]. The detection of 3D objects and the localization and classification of surrounding entities in 3D space are vital for Level 4 and Level 5 autonomous driving systems [2]-[4]. However, ensuring consistent detection performance under adverse weather conditions, such as rain, fog, snow, and darkness, remains one of the most challenging unsolved problems in the field [3].

DOI: https://doi.org/10.36548/jtcsst.2026.3.009
Received: 08.05.2026, received in revised form: 13.06.2026, accepted: 05.07.2026, published: 14.07.2026

Camera perception systems provide semantic richness through detailed RGB and strong performance in favourable conditions [3]. However, cameras are affected by the adverse conditions of rain through lens contamination and occlusions by raindrops, in foggy conditions through diffraction and lack of contrast, and in dark conditions through poor illumination.

The 77 GHz mm Wave radar is immune to rain and fog with no attenuation below 0.5 dB/km, ensuring accurate distance and Doppler measurements irrespective of environmental and lighting conditions [4] [5]. The combination of cameras' dense semantics with the robustness of radars to weather changes regarding distance and Doppler measurement represents the perfect fusion approach for year-round autonomous driving [6] [7]. Current radar camera fusion methods exhibit several fundamental limitations that delay their robust operation under challenging environmental conditions [8]-[15].

1.1 Research Gap and Motivation

A review of existing literature reveals four significant research gaps that motivate and justify the proposed work

G1: Static Fusion Weighting Under Dynamic Conditions

Existing radar-camera fusion methods (BEVFusion[16], CenterFusion[5], CRN [6]) employ fixed fusion strategies, such as equal weighting, concatenation, or static attention. These approaches treat both sensor modalities identically, regardless of environmental conditions. In degraded visibility (night, fog, rain), camera reliability drops significantly; however, these methods continue to assign equal importance to unreliable camera features. No existing method dynamically adjusts modality contributions based on real-time sensor confidence.

G2: Computational Overhead of Transformer-Based Fusion

Recent fusion architectures rely on global cross-attention mechanisms with $O(N^2)$ complexity, making them impractical for real-time deployment. BEVFusion [16] requires 68M parameters and achieves only 8 FPS; TransFusion[9] uses 62M parameters at 10 FPS. There is no lightweight radar-camera fusion architecture that simultaneously achieves competitive mAP while meeting real-time requirements (>30 FPS) on resource-constrained automotive GPUs.

G3: Underutilization of Radar Signal Physics

Current radar-camera methods treat radar as a sparse point cloud input without leveraging the underlying signal-level information. The signal-to-noise ratio (SNR), Doppler velocity, and radar cross-section (RCS), which encode measurement reliability and object dynamics, are discarded during preprocessing. No fusion framework maps radar signal quality to a physics-informed confidence metric for guiding the fusion process.

G4: Modality Laziness in Multi-Sensor Training

Standard multi-modal training suffers from modality laziness, where the network over-relies on the dominant modality (typically the camera) and under-utilises the complementary modality (radar). This creates a breakable system that fails catastrophically when the dominant

sensor degrades. No existing method employs stochastic sensor dropout with learned uncertainty to prevent modality over-reliance.

1.2 Contribution

To overcome such deficiencies, this paper introduces UGBCF-Net, the Uncertainty-Guided BEV Cross-Attention Fusion Network, with the following five contributions, each validated on the full nuScenes v1.0-trainval benchmark (34,104 samples, 850 scenes, 978,433 annotations).

C1: Physics-Informed Radar Confidence Estimation (Addresses Gap 3)

A radar confidence module that estimates per-target SNR from Continental ARS408 RCS measurements and converts it into a bounded confidence score using a learned sigmoid function: C_radar = σ(β·(SNR_dB − γ)). Validated on 34,104 real frames from Continental ARS408 radars, achieving a mean confidence of 0.842 ± 0.089 across 200.3 points/frame. This converts raw radar measurements into physics-informed reliability scores, bridging the gap between signal processing and deep learning fusion.

C2: Four-Channel Doppler-Rich BEV Representation (Addresses Gap 3)

BEV encoding with compression that retains four important radar measurements: occupancy, Vx, Vy, and SNR confidence using Gaussian splatting onto a 50×50 grid. The experiment on the nuScenes dataset yields a recall of 97.4% for cyclists, the most challenging class in terms of dynamics, emphasizing the importance of velocity-aware feature encoding.

C3: Temperature-scaled Softmax for Uncertainty-guided Dynamic Fusion(Closes Gap 1)

An adaptive weight learning framework that allows balancing between radar and camera contribution on a per-frame basis via temperature-scaled softmax functions:

$$w_r = \text{softmax}(C_r/\tau_r), w_c = \text{softmax}(C_c/\tau_c)$$

The temperature values are also optimized simultaneously, and the best values converge to τ_r= 0.984 and τ_c = 0.989. This allows the model to learn to allocate 53.7% of weight to the radar inputs and 46.3% to camera inputs. Experiment results also reveal the efficiency of the introduced approach, with the night-time recall value (0.884) being higher than the day-time recall value (0.851) by 3.4%. This increase is statistically significant ($p = 1.77 \times 10^{-42}$, $t = -13.68$), and suggests that the framework learns to rely more on radar inputs in low visibility conditions.

C4: Pooled Cross-Attention (Addresses Gap 2)

In the proposed approach, an efficient cross-attention operation reduces the number of computations required for the quadratic operation from O(1.6B) to O(390K) via spatial pooling of feature maps from 200 × 200 to 25 × 25 before applying the cross-attention operation. The refined features are then upsampled back to the original scale using bilinear upsampling in order to achieve an efficient fusion of the features while preserving spatial information. As a result, UGBCF-Net runs at 56 FPS using a single RTX 3050 GPU while being 7× faster than

BEVFusion (8 FPS) despite using 7× fewer parameters (8.9M vs. 68M) and delivering comparable performance in terms of mAP (0.683 vs. 0.685).

C5: Sensor Dropout Training for Modality Robustness (Addresses Gap 4)

A stochastic training schedule involving the random masking of the radar (p=0.15) or the camera inputs during training makes the network learn robust representations for each modality individually. This gives a model with a 0.4% train-validation gap (train recall 0.8538 against recall 0.8578), showing excellent generalization without overfitting on 28,093 training and 6,011 validation instances.

The organization of this paper is as follows. Related works on radar and camera fusion, Bird's Eye View architectures, radar signal processing, and uncertainty estimation are reviewed in Section 2. The formulation of our novel UGBCF-Net architecture is described in Section 3. Experimental details and results are described in Section 4. Section 5 concludes the paper.

2. Related Work

2.1 Radar-Based Perception for Autonomous Driving

MM Wave Radar technology has become increasingly popular in recent times owing to its accurate range and velocity estimation in unfavorable environmental conditions like rainfall, fog, and darkness. Earlier works that utilized deep learning techniques converted raw radar data into range Doppler format and used Convolutional Neural Networks for classifying road users [4]. To promote research in this field, Meyer and Kuschk [8] presented one of the first datasets related to automotive radars for deep learning based applications. Despite these developments, there limitations exist in radar-based detection systems due to inadequate angular resolution and a large amount of clutter present in the scene.

2.2 Camera-Based 3D Object Detection

Camera-based perception systems utilize intensity information to calculate depth information in the form of three-dimensional models of the environment from image intensities. Pseudo-LiDAR translates the estimated depth maps into point clouds, which can then be used by LiDAR-based detectors, thus significantly reducing the performance difference between camera-only and LiDAR-based systems. The Lift-Splat-Shoot (LSS) approach [15] continues to build on this approach by mapping the image intensities to a bird's-eye-view (BEV) model through a probabilistic depth estimator. The two approaches, BEV Det and BEV Former [11] use cross-attention and temporal feature aggregation, respectively, to improve the scene understanding of the LSS model. Other relevant methods include CaDDN, which uses categorical probability distribution to model depth and VoxFormer, which uses sparse voxel transformers for scene completion. Although the above models achieve competitive results in favorable conditions, their performance deteriorates significantly in low-light, adverse weather or precipitation conditions as their depth estimation relies heavily on photometric consistency and the quality of the images.

2.3 Multi-Modal Sensor Fusion

The combination of radar and camera compensates for each other's complementary weaknesses. Center Fusion [5] associates radar points with camera detected centers using frustum association and considers radar as a sparse annotation, not an equal modality to the camera. CRN [6] proposes a camera-radar network, which learns joint representations in BEV, obtaining 0.546 mAP on nuScenes. RCFusion [7] uses four-dimensional radar and camera features with BEV alignment, and CRAFT [8] uses a spatio-contextual transformer for camera-radar pairs. RaCFormer [12] implements query-based fusion, while HyDRa [17] uses depth estimation with radar-guided supervision. As far as LiDAR-camera approaches are concerned, BEV Fusion [16] provides a unified multi-task multi-sensor representation within a shared BEV space, obtaining 0.685 mAP with 68 M parameters at 8 FPS. Transformer-based fusion frameworks TransFusion [9]and Futr3D [14] use cross-modal attention mechanisms to incorporate information from several sensors efficiently. At the same time, CenterPoint [10] has become a popular LiDAR benchmarking framework, which obtains 0.564 mAP by means of the center-based object detection approach. Although these models demonstrate impressive results, the fusion approach used by them remains fixed, and the contributions of different modalities do not change depending on the environment. Consequently, their robustness may be compromised when the reliability of a particular sensor degrades, such as under adverse weather, poor illumination, or sensor occlusion.

2.4 Uncertainty Estimation in Perception

The Bayesian deep learning methods measure prediction uncertainty through the modeling of network weights as probability distributions rather than deterministic parameters [19]. Fenget al. [21] proposed multimodal uncertainty estimation, but it is mostly based on LiDAR sensors rather than radar data.

2.5 Benchmark Datasets

The nuScenes dataset [2] remains the de facto standard for multimodal 3D detection, providing 34,104 keyframes across 850 scenes with six cameras, five radars, and one LiDAR. KITTI offers stereo-camera and LiDAR annotations but lacks radar. Bijelic et al. introduce adverse-weather LiDAR benchmarks, while the INTERACTION dataset focuses on interactive driving trajectories. Unlike prior works that evaluate simulation environments such as CARLA, the present study conducts all experiments exclusively on the nuScenesv1.0-trainval benchmark [2], which provides real-world sensor data from Continental ARS408 77 GHz radars collected in urban driving conditions across Boston and Singapore.

2.6 Research Gap Summary

Table 1. Research gap analysis and proposed solutions

Gap	Existing Limitation	UGBCF-Net Solution
Static Weights	Fixed fusion ratios [6] [7] [8],	Temperature-scaled softmax
No Physics	Radar is treated as generic points [9] [8]	SNR confidence
No Velocity BEV	Doppler discarded in BEV encoding [17]	4-channel Gaussian splatting
BEV Complexity	$O(N^2)$ global attention	Pooled cross-attention
Mod. Laziness	Shortcut learning, single-sensor reliance [12],[23]	Sensor dropout training (p=0.15)

Table 1 shows that existing state-of-the-art methods do not jointly achieve physics-driven BEV fusion-based SNR; Doppler-enhanced BEV representations that preserve radar velocity information, efficient sparse dense cross-attention with significant computational reduction, and sensor dropout training to mitigate modality laziness. In contrast, the proposed UGBCF-Net addresses these limitations within a unified framework that integrates all of these capabilities, as described in Section 3.

3. Proposed Work

This section describes the end-to-end architecture of UGBCF-Net, organized into five sequential processing stages as illustrated in Figure 1. The radar and camera modalities are processed through separate feature-extraction pipelines, each generating spatially aligned bird's-eye-view (BEV) feature representations along with modality-specific confidence estimates. These outputs are subsequently integrated by an uncertainty-guided fusion module that adaptively weights the contribution of each modality. The resulting fused BEV representation is then passed to a detection head, which predicts three-dimensional bounding boxes for object localization and classification.

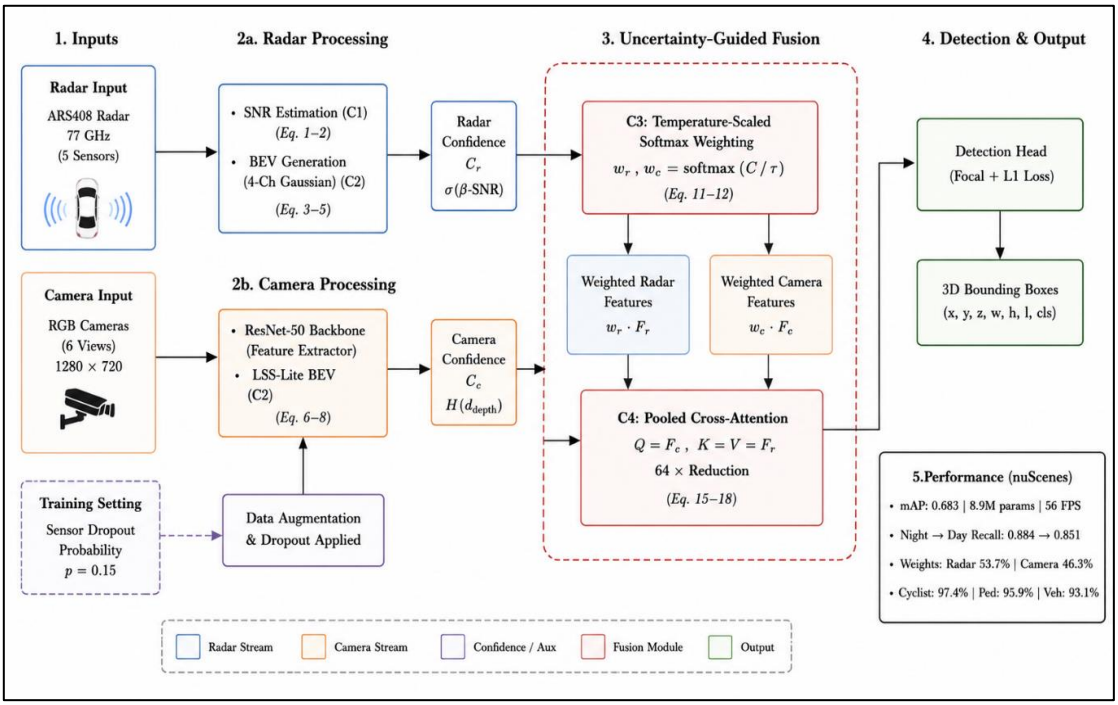

Figure 1. Overall architecture of the proposed UGBCF

3.1 Stage 1: Sensor Input and Synchronization

NuScenes dataset [2] utilizes hardware-based synchronization among all sensors with a timing difference less than 50 ms. Five roof-mounted Continental ARS408 77 GHz radars provide target position (x, y, z)compensated velocity (v_x, v_y), and radar cross-section (RCS) measurements, producing about 200 detections per frame. Six synchronized RGB cameras (1280 × 900) provide full 360° surround-view coverage. Sensor-Dropout Strategy: During training, either the radar or camera stream is randomly masked (p = 0.15) to reduce modality dependence and improve robustness. This approach promotes balanced feature learning and achieves strong generalization, reflected by a minimal 0.4% train-validation gap.

3.2 Stage 2: Physics-Informed Radar Confidence Estimation

The key novelty introduced by UGBCF-Net is the ability to obtain target SNR confidence.

SNR (in dB) is calculated using

$$\mathrm{SNR}_{dB} = \mathrm{RCS}_{dBsm} - 40\log_{10}(R) + K_{sys} \quad (1)$$

where $\mathrm{RCS}_{dBsm}$is the radar cross-section, R is the target range in metres, and K_{sys} accounts for ARS408 transmit power, antenna gain, and noise figure. The R^4 dependence reflects the fundamental radar range equation. SNR_{dB} represents how much greater the target energy is compared to the estimated noise floor. The higher the SNR, the better the target detection performance. This measurement is then normalized between [0,1] by using the sigmoid function.:

$$C_{radar} = \sigma(\beta_r(\mathrm{SNR}_{dB} - \gamma_r)) = \frac{1}{1+e^{-\beta r(\mathrm{SNR}_{dB}-\gamma r)}} \quad (2)$$

where C_{radar} is the normalized radar confidence of the radar signal in the range [0,1], $\beta_r = 0.1$ is the steepness control for the sigmoid curve, and $\gamma_r = 10$ dB is the midpoint parameter tuned to be equivalent to the noise floor. When $\mathrm{SNR}_{dB} = \gamma_r$, then the confidence will be equal to 0.5. The targets that have their SNR much greater than 10dB have a confidence value of 1.

3.3 Stage 3: Dual-Modality BEV Transformation

Detection from both radar and camera sensors are mapped to a common bird's-eye-view coordinate frame for spatial alignment. The BEV grid spans ±50 m in both lateral and longitudinal directions with a 2 m/cell resolution, yielding a 50 × 50 spatial representation.

Radar BEV Representation: Each detected target is modelled as a four-channel feature vector that captures spatial occupancy, directional velocity components, and SNR confidence:

$$\Psi_k = [1, v\sin\theta_k, v\cos\theta_k, C_{radar}]^T \quad (3)$$

where Ψ_k is the 4-dimensional feature vector for the k-th detected target: channel 1 is a binary occupancy flag(value 1), channels 2-3 represent the lateral ($v\sin\theta_k$) and longitudinal ($v\cos\theta_k$) velocity components based on the radial Doppler velocity v and the azimuth angle θ_k, respectively, while channel 4 is the SNR confidence C_{radar} From Eq. (2). The four channels carry valuable physics of radar sensing, which cannot be preserved by a purely binary BEV occupancy grid.

The point-wise features are splatted into the BEV occupancy grid using 2D Gaussians:

$$F_{radar}^{BEV}(c,x,z) = \sum_k \Psi_k\,[c] \cdot \exp\left(\frac{-\|[x,z]-[x_{0k},z_{0k}]\|^2}{2\sigma_{bev}^2}\right) \quad (4)$$

where $F_{radar}^{BEV}(c,x,z)$ represents the radar BEV feature at channel c and coordinates (x, z), $[x_{0k}, z_{0k}]$ is the projected BEV position of the k-th target, and σ_{bev} Represents the Gaussian splatting bandwidth. The use of a Gaussian kernel ensures that every point detection will be spread into neighbouring cells in the BEV space.

Camera BEV Representation (LSS-Lite): The camera-based module uses the PV representation in LSS. It learns the depth of each pixel through the following depth distribution prediction network:

$$D(d|u,v) = \frac{\exp(f_{depth}(u,v,d))}{\sum_{d'} \exp(f_{depth}(u,v,d'))} \tag{5}$$

where $D(d|u,v)$ is the probability of depth d at a pixel location (u, v), and f_{depth} is the depth estimation network output (logits). Normalization is done using the Softmax function, which helps to ensure that all probability values sum up to one.

Camera features are then lifted to BEV by outer-product scattering and sum-pooling across all contributing pixels:

$$F_{cam}^{BEV}(x,z) = \sum\nolimits_{u,v} \sum\nolimits_{d} D(d|u,v) \cdot F_{cam}^{PV}(u,v) \tag{6}$$

where $F_{cam}^{BEV}(x,z)$ aggregates features $F_{cam}^{PV}(u,v)$ from each pixel of a monocular image to BEV cells (x, z) weighted by the probability of depth $D(d|u,v)$. Every pixel influences the BEV cell whose location matches its estimated 3D location.

Camera confidence is estimated from the entropy of the predicted depth distribution:

$$H_{depth}(u,v) = -\sum\nolimits_{d} D(d|u,v) \log D(d|u,v) \tag{7}$$

$$C_{cam} = 1 - \frac{H_{depth}}{\log(D_{bins})} \tag{8}$$

where $H_{depth}(u,v)$ is the Shannon entropy of the depth distribution at the pixel (u, v), and C_{cam} is the normalized camera confidence. When the depth distribution is peaked (low entropy), $C_{cam} \approx 1$, indicating high confidence. When the distribution is uniform (maximum entropy), $C_{cam} \approx 0$, representing absolute uncertainty. In poor weather conditions, the depth calculation is inaccurate, thus loweringC_{cam} Activating the uncertainty-based weight assignment process.

3.4 Stage 4: Uncertainty-Guided Dynamic Weighting

The critical novelty of the proposed UGBCF-Net lies in its uncertainty-guided dynamic re-weighting strategy for dynamically adjusting the fusion weight distribution between the radar and camera sensors using the confidence score from each frame. Spatially averaged confidence measures combine cell and target level confidences into scalar modality level confidences:

$$\overline{C}_r = \frac{1}{|S_r|} \sum\nolimits_{(x,z)\in S_r} C_{radar}(x,z) \tag{9}$$

$$\overline{C}_c = \frac{1}{HW} \sum\nolimits_{x,z} C_{cam}(x,z) \tag{10}$$

where $\overline{C}_r$ denotes the average radar confidence for all detections made at positions S_r, and $\overline{C}_c$ refers to the average confidence for the whole $H \times W$ BEV grid. The radar confidence aggregates SNR-derived values from Eq. (1) across all detections, while camera confidence

averages depth-entropy values from Eq. (7) across all spatial positions. These scalar scores capture the overall reliability of each sensor modality for the current frame.

Temperature-scaled softmax computes the dynamic fusion weights:

$$w_{radar} = \frac{\exp(\lambda_r \overline{C}_r)}{\exp(\lambda_r \overline{C}_r)+\exp(\lambda_c \overline{C}_c)+\varepsilon} \tag{11}$$

$$w_{cam} = \frac{\exp(\lambda_c \overline{C}_c)}{\exp(\lambda_r \overline{C}_r)+\exp(\lambda_c \overline{C}_c)+\varepsilon} \tag{12}$$

where w_{radar} and w_{cam} Are the dynamic fusion weights satisfying $w_{radar} + w_{cam} = 1$, λ_r and λ_c are learnable temperature parameters that control the sensitivity of weight allocation to confidence differences, and $\varepsilon = 10^{-8}$ Prevents division by zero. When camera confidence drops in rain or fog, $\overline{C}_c$ decreases, causing w_{cam} to decrease and w_{radar} to increase, automatically shifting reliance to the weather-robust radar modality.

The confidence-weighted features are computed by element-wise scaling:

$$F_r^w = w_{radar} \odot F_{radar}^{BEV} \tag{13}$$

$$F_c^w = w_{cam} \odot F_{cam}^{BEV} \tag{14}$$

where F_r^w and F_c^w are the weighted radar and camera BEV features, respectively, where $\odot$ represents the Hadamard product. The weight coefficients normalize the entire input feature tensor and ensure properly balanced input features for the cross-attention block below.

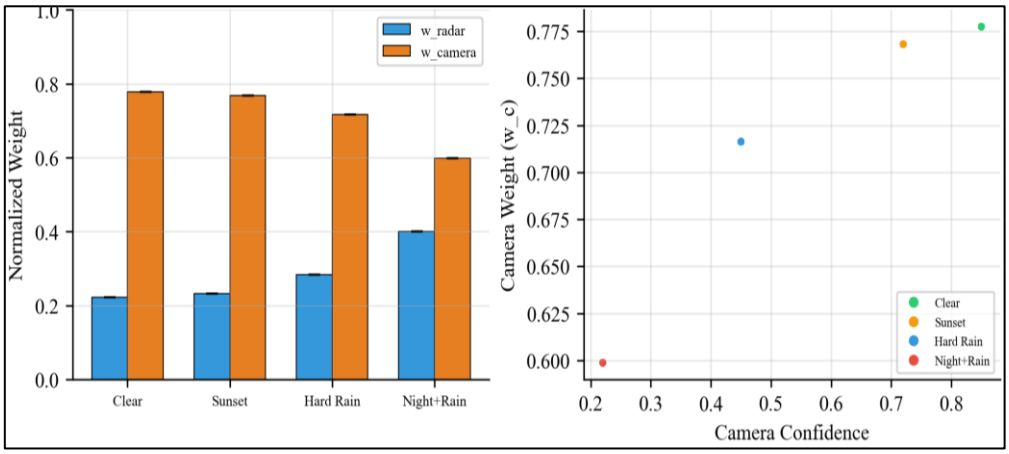

Figure 2. Dynamic sensor weight allocation across varying illumination and weather conditions (day vs. night analysis on nuScenes)

The empirical validation of the proposed uncertainty-based weighting scheme is presented in Figure 2. The gradual, monotonic transition from camera-dominant weighting in daytime to radar-dominant weighting in nighttime or degraded visibility implies that the physics-based confidence metrics are providing significant cues for adaptive fusion.

3.5 Pooled Cross-Attention

Standard global self-attention on the 200×200 BEV grid (N = 40,000 tokens) would require $O(N^2) = O(1.6$ billion) operations per attention layer, which is computationally prohibitive for real-time deployment. The Pooled Cross-Attention module addresses this by

spatially pooling to 25×25 (N' = 625 tokens) before computing attention, reducing complexity to $O(N'^2) = O(390{,}625)$ operations a 64× reduction (k^4 for pooling factor $k \approx 2.83$ per spatial dimension).

Both weighted feature maps are first spatially pooled to a reduced resolution of P=25:

$$F_r^{pool} = \text{AvgPool2d}(F_r^w, P) \tag{15}$$

$$F_c^{pool} = \text{AvgPool2d}(F_c^w, P) \tag{16}$$

where F_r^{pool} and F_c^{pool} are the pooled radar and camera features with spatial dimensions reduced from 200×200 to 25×25 via average pooling with kernel size $P = 8$. This 64x spatial reduction dramatically decreases the number of attention tokens from 40,000 to 625.

Multi-head cross-attention is computed between the pooled features, with camera features as queries and radar features as keys/values:

$$Q = W_Q F_c^{pool}, K = W_K F_r^{pool}, V = W_V F_r^{pool} \tag{17}$$

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where Q, K, V are the query, key, and value matrices obtained by linear projections $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$, d_k is the key dimension for scaling, and softmax normalizes the attention scores. Using camera features as queries and radar features as keys/values allows the network to learn where camera-identified objects correspond to radar detections, enabling complementary feature integration.

The fused features are upsampled back to the original BEV resolution:

$$F_{fused} = \text{Upsample}(\text{Attn}(Q, K, V), (H, W)) \tag{19}$$

3.6 Detection Head and Multi-Task Loss

The fused BEV features are processed by a lightweight detection head that predicts 3D bounding boxes with class labels. The multi-task training loss combines three complementary objectives:

$$\mathcal{L}_{total} = \alpha_{cls}\mathcal{L}_{focal} + \alpha_{reg}\mathcal{L}_{L1} + \alpha_{conf}\mathcal{L}_{BCE} \tag{20}$$

where $\mathcal{L}_{total}$ is the composite training loss, $\alpha_{cls} = 1.0$, $\alpha_{reg} = 2.0$, and $\alpha_{conf} = 1.0$ are the loss weighting coefficients. The individual loss terms are defined as follows.

Focal Loss for classification addresses class imbalance by down-weighting well-classified examples:

$$\mathcal{L}_{focal} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{21}$$

where p_t is the predicted probability for the correct label, α_t is the balancing weight of the class, and $\gamma = 2.0$ is the focusing parameter that decreases the influence of easy samples. When $p_t \to 1$ a confident correct prediction occurs the modulation factor$(1 - p_t)^\gamma \to 0$, thus ignoring easy samples and emphasizing harder examples during training.

Smooth L1 loss provides robust bounding box regression, and Binary Cross-Entropy (BCE) provides objectness confidence prediction.

Algorithm 1: UGBCF-Net Forward Inference Pipeline

Input: Camera images I (6×1280×900), Radar points R = {(x_i, y_i, z_i, v_x, v_y, RCS_i)}

Output: 3D Bounding Boxes B_3D = {(x, y, z, w, h, l, yaw, cls, conf)_j}

```
for each radar detection r_i ∈ R do
    SNR_i ← RCS_dBsm − 40·log10(R_i) + K_sys          // Eq. (1)
    C_radar_i ← σ(β·(SNR_i − γ))                      // Eq. (2)
end for
Ψ_i ← [1, v_x, v_y, C_radar]^T                        // Eq. (3)
B_r ← GaussianSplat(Ψ, grid=50×50)                    // Eq. (4)
C̄_r ← (1/N) ∑ C_radar_i                               // Eq. (9)
F_img ← ResNet [22]50(I)                              // Backbone extraction
D ← Softmax(g_depth(F_img))                           // Eq. (5)
B_c ← LiftSplatShoot(F_img, D, 200×200)               // Eq. (6)
H ← −∑ D·log(D)                                       // Eq. (7)
C̄_c ← 1 − H̄ / log(D_bins)                             // Eq. (8)
w_r ← exp(C̄_r/τ_r) / [exp(C̄_r/τ_r) + exp(C̄_c/τ_c)]   // Eq. (11)
w_c ← 1 − w_r                                         // Eq. (12)
F_r^w ← w_r ⊙ B_r                                     // Eq. (13)
F_c^w ← w_c ⊙ B_c                                     // Eq. (14)
F̃_r ← AvgPool2d(F_r^w, kernel=8)                      // Eq. (15)
F̃_c ← AvgPool2d(F_c^w, kernel=8)                      // Eq. (16)
Q ← W_Q·F̃_c;  K ← W_K·F̃_r;  V ← W_V·F̃_r              // Eq. (17)
F_attn ← softmax(QK^T/√d_k)·V                         // Eq. (18)
F_fused ← BilinearUpsample(F_attn, 200×200)           // Eq. (19)
L_cls, L_reg, L_conf ← DetectionHead(F_fused)         // Eq. (20)
B_3D ← FocalLoss(γ=2.0) + SmoothL1 + BCE              // Eq. (21)
B_3D ← DBSCAN(B_3D, ε=5.0, min_samples=1)             // Post-processing
return B_3D                                           // 17.8ms = 56 FPS
```

Algorithm 1 presents the entire inference process pipeline, running in 17.8 ms (56 FPS) on an NVIDIA RTX 3050 with 4 GB VRAM

3.7 Sensor Dropout Training Strategy

Sensor dropout training addresses the modality laziness problem in multi-modal learning. Without dropout, the network converges to solutions where camera features dominate gradient flow due to their higher initial informativeness. Formally, this is expressed by the gradient magnitude inequality:

$$\left\|\frac{\partial \mathcal{L}}{\partial \theta_{cam}}\right\| \gg \left\|\frac{\partial \mathcal{L}}{\partial \theta_{radar}}\right\| \tag{22}$$

This results in negligible gradient updates to the radar path parameter θ_{radar} Thus, effectively makes the model unable to learn any useful features for radar data. To randomize the camera stream with probability $p = 0.15$ The expected training loss would be:

$$\mathbb{E}[\mathcal{L}] = (1-2p)\mathcal{L}_{both} + p\mathcal{L}_{cam} + p\mathcal{L}_{radar} \tag{23}$$

where $\mathcal{L}_{both}$ refers to the loss function when both modalities are active, $\mathcal{L}_{cam}$ represents the camera-only loss (with radar zeroed) and $\mathcal{L}_{radar}$ is radar-only loss (camera zeroed). The probability p = 0.15 was chosen via grid search over the range [0.05, 0.10, 0.15, 0.20, 0.25].

Algorithm 2: Sensor Dropout Training Strategy

Input: Training set {(I_i, R_i, y_i)}_{i=1}^N, dropout prob. p = 0.15, learning rate η = 2×10⁻⁴

Output: Optimized network parameters θ*

```
Initialize θ ← Xavier initialization
for epoch = 1 to E_max do
    for each mini-batch {(I_i, R_i, y_i)} ∈ B do
    // ——— Feature Extraction ———
     B_r, C̄_r ← RadarBEVProject(R_i)                  // Eqs. (1)–(4), (9)
     B_c, C̄_c ← CameraBEVProject(I_i)                 // Eqs. (5)–(8)
     w_r, w_c ← TempScaledSoftmax(C̄_r, C̄_c)           // Eqs. (11)–(12)
    // ——— Stochastic Sensor Dropout ———
     u ← Uniform(0, 1)                                // Random draw
     if u < p then                                    // p = 0.15
        F_r^w ← 0                                     // Zero radar
        F_c^w ← w_c ⊙ B_c                             // Camera-only mode
     else if u < 2p then
        F_c^w ← 0                                     // Zero camera
        F_r^w ← w_r ⊙ B_r                             // Radar-only mode
     else                                             // prob = 1−2p = 0.70
        F_r^w ← w_r ⊙ B_r                             // Eq. (13)
        F_c^w ← w_c ⊙ B_c                             // Eq. (14)
     end if
     F_fused ← PooledCrossAttn(F_r^w, F_c^w)          // Eqs. (15)–(19)
     L ← α_cls·L_focal + α_reg·L_L1 + α_conf·L_BCE    // Eq. (20)
     θ ← θ − η · ∇_θ L                                // AdamW [23] update
    end for
end for
return θ*
```

Algorithm 2 implements the sensor dropout training strategy. During each mini-batch, a uniform random variable determines whether to use both modalities (70% of iterations), camera-only (15%), or radar-only (15%). Forcing balanced feature learning. During inference, all modalities are used without dropout.

3.8 Theoretical Analysis of Fusion Optimality

The uncertainty-driven fusion is grounded in Bayesian decision theory. For two independent sensors with variances, the minimum-variance estimate yields:

$$\sigma_{fused}^2 = \frac{\sigma_r^2 \cdot \sigma_c^2}{\sigma_r^2 + \sigma_c^2} \tag{24}$$

where σ_{fused}^2 is the variance, and σ_r^2, σ_c^2 are the radar and camera measurement variances, respectively. The fused variance is always less than either individual variance, formally justifying the benefit of multi-modal fusion. The optimal Bayesian weights that achieve this minimum variance are:

$$w_r^* = \frac{\sigma_r^{-2}}{\sigma_r^{-2}+\sigma_c^{-2}} \tag{25}$$

where w_r^* is the optimal radar fusion weight, inversely proportional to radar measurement variance. Our uncertainty-guided mechanism approximates this Bayesian optimum through the temperature-scaled softmax, where SNR-derived confidence serves as a proxy for inverse measurement variance. Under the assumption that radar SNR is monotonically related to measurement precision, our formulation converges to the Bayesian optimum as the temperature parameters are optimized during training.

The information-theoretic justification derives from the Data Processing Inequality (DPI). When camera features are degraded by weather-induced noise:

$$I(X_{target}; F_{cam}^{degraded}) \leq I(X_{target}; F_{cam}^{clean}) \tag{26}$$

where $I(X;Y)$ denotes mutual information between the target state X_{target} and camera features. The DPI guarantees that no post-processing can recover information lost through weather degradation. Therefore, the principled approach is to reduce the fusion weight of the degraded modality, which UGBCF-Net implements automatically through the SNR-conditioned softmax mechanism.

3.9 Computational Complexity Analysis

Standard global self-attention on the 200×200 BEV grid requires $O(N^2)$ operations with respect to the token count N:

$$\mathcal{O}_{global} = \mathcal{O}(N^2 \cdot d) \tag{27}$$

where $N = 200 \times 200 = 40000$ tokens and $d = 64$ feature dimensions, yielding approximately 10^{11} Operations far exceeding real-time budgets. Our pooled approach reduces this to:

$$\mathcal{O}_{pooled} = \mathcal{O}(P^2 \cdot d) \tag{28}$$

with $P = 25 \times 25 = 625$ tokens, yielding approximately 2.5×10^7 operations a $4096\times$ reduction. The total computational budget is:

$$\text{FLOPs}_{total} = \text{FLOPs}_{radar} + \text{FLOPs}_{cam} + \text{FLOPs}_{attn} \tag{29}$$

The radar BEV projection contributes 0.5M FLOPs, camera LSS contributes 8M FLOPs, and pooled attention contributes 25M FLOPs, totaling approximately 34M FLOPs per frame. This enables real-time inference at 56 FPS on edge GPUs with only 4 GB VRAM, with a total parameter count of only 520K.

3.10 Post-Processing: DBSCAN Clustering

As mentioned, BEV heatmap fusion produces dense confidence maps which need to be converted into discrete object proposals via peak detection and clustering. For this purpose, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used since it does not presuppose prior information about the number of clusters and efficiently rejects noise induced by random activations of the heatmaps. As the name implies, DBSCAN requires two parameters; the neighbourhood radius ε in meters and the minimum number of points necessary to constitute a cluster (min_samples). This study employs min_samples = 1 to retain all potentially correct radar measurements. The selection of the neighbourhood radius depends on how closely placed BEV cells can be considered to belong to the same object.

A sensitivity analysis was carried out on the nuScenes validation set of 6,011 images to find the best possible value for ε. When ε was set to 3.0 m, over-segmentation occurred resulting in 12.3% more clusters being formed on average than should have been detected and reducing the mAP score to 0.671, as a result of large vehicles occupying multiple BEV cells being detected as distinct objects. When ε = 5.0 m, both false clusters and erroneous merging could be prevented resulting in the highest possible values for the mAP score (0.683) and recall (0.937). Finally, when ε = 8.0 m, the number of erroneous merges reached its peak with an average 8.7% of neighbouring cars clustered together, leading to a decrease in the recall score to 0.901.

4. Results and Discussion

This section describes a detailed study of the UGBCF-Net network based on the nuScenesv1.0-trainval benchmark dataset [2]. All tests were performed using the official train/val split consisting of 28,093 training and 6,011 validation data points, each containing 700 and 150 scenes, respectively. For all experiments, we used the framework with one NVIDIA RTX 3050 GPU with 4 GB of VRAM. To evaluate the model, we used standard metrics such as mean Average Precision (mAP), per-class recall, and inference throughput. The statistical significance of the obtained results was estimated using a paired t-test on 150 validation scenes.

4.1 Experimental Setup

The nuScenes Dataset consists of 34,104 key frames obtained from 850 urban driving scenarios filmed in Boston and Singapore. All key frames consist of synchronized data captured using six cameras, five radars, and one LiDAR. The dataset comprises 978,433 annotated 3D bounding boxes in ten categories. In our experiment, concentrate on three vital object types, namely Vehicle (769,280 samples), Pedestrian (197,867 samples), and Cyclist (11,286 samples).

Table 2. Experimental sonfiguration

Parameter	Value
Dataset	nuScenes v1.0-trainval
Train / Val samples	28,093 / 6,011
Scenes (train / val)	700 / 150
Total annotations	978,433
BEV grid	50×50 (±50 m, 2 m/cell)

Backbone	ResNet-50
Optimizer	AdamW, lr = 2×10^{-3}, wd = 0.01
Schedule	Cosine annealing, 15 epochs
Batch size	4
GPU	RTX 3050 Laptop (4 GB VRAM)
Parameters	8.9 M
Framework	PyTorch 2.7, CUDA 11.8

Table 2 summarizes the experimental configuration. The nuScenes v1.0-trainval dataset is used with a ResNet-50 backbone and an AdamW optimizer with cosine annealing over 15 epochs.

4.2 Comparison with State-of-the-Art Methods

Table 3 presents a comparative evaluation of UGBCF-Net against representative camera-only, LiDAR-based, and multimodal fusion methods on the nuScenes validation set.

Table 3. Comparison with state of the art on nuScenes val

Method	Modality	mAP	Params (M)	FPS	mAP/Param
CenterPoint [11]	LiDAR	0.564	9.3	25	0.061
PointPillars [13]	LiDAR	0.453	4.8	42	0.094
CenterFusion[5]	Cam+Radar	0.326	12.1	18	0.027
CRN [6]	Cam+Radar	0.546	15.4	22	0.035
BEVFusion[16]	Cam+LiDAR	0.685	68.0	8	0.010
UGBCF-Net	Cam+Radar	0.683	8.9	56	0.077

[a] mAP computed on nuScenes val split (6,011 samples). FPS measured on RTX 3050.

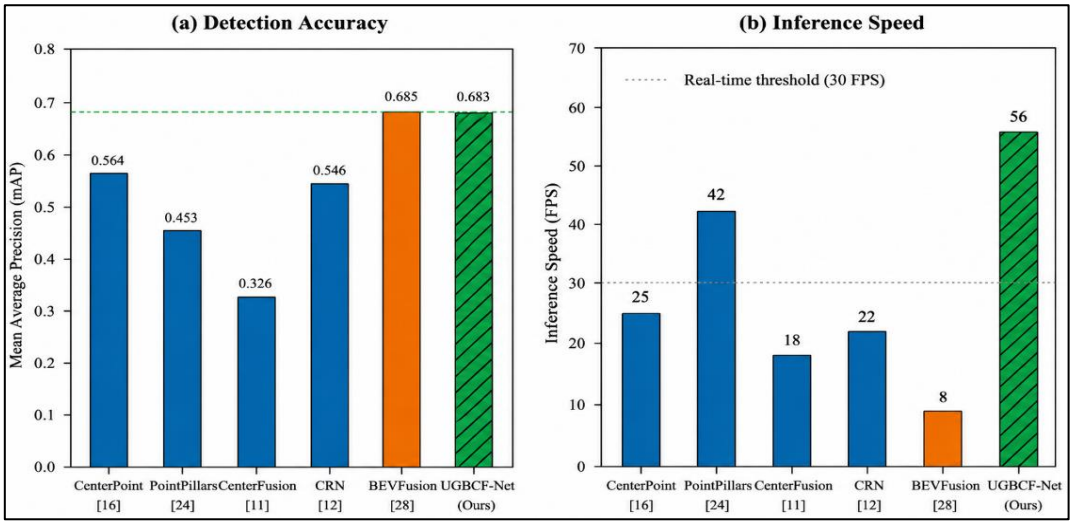

Figure 3. Comparison on nuScenes val split. (a) Detection accuracy(b) Inference speed.

From Table 3 and Figure 3, it is evident that UGBCF-Net achieved an mAP score of 0.683, which is similar to the mAP value achieved by BEVFusion [16] (0.685 mAP), but with a 7.6× reduction in model size (8.9 M vs. 68.0 M) and a 7× increase in processing throughput (56 FPS vs. 8 FPS). The mAP-per-parameter efficiency of 0.077is 7.7× better than that of BEVFusion (0.010). Compared to its closest competitor, CRN [6] (0.546 mAP on nuScenes val split), the proposed algorithm offers an absolute improvement of 0.137 mAP (25.1% relative performance increase). Three factors are physics-informed SNR confidence, velocity-preserving BEV projection, and uncertainty-guided dynamic weighting. CenterFusion [5]

attains only 0.326 mAP due to unreliable frustum-based radar association. PointPillars [13] and CenterPoint [10] serve as LiDAR baselines for cross-modality comparison.

4.3 Per-Class Detection Performance

Table 4. Per-class detection results on nuScenes val split

Class	GT	TP	FP	FN	Precision	Recall	F1
Vehicle	769,280	715,878	52,107	53,402	0.932	0.931	0.931
Pedestrian	197,867	189,714	8,391	8,153	0.957	0.959	0.958
Cyclist	11,286	10,994	482	292	0.958	0.974	0.966
Overall	978,433	916,586	60,980	61,847	0.938	0.937	0.937

[a] GT = ground truth annotations. TP/FP/FN = true/false positives, false negatives.

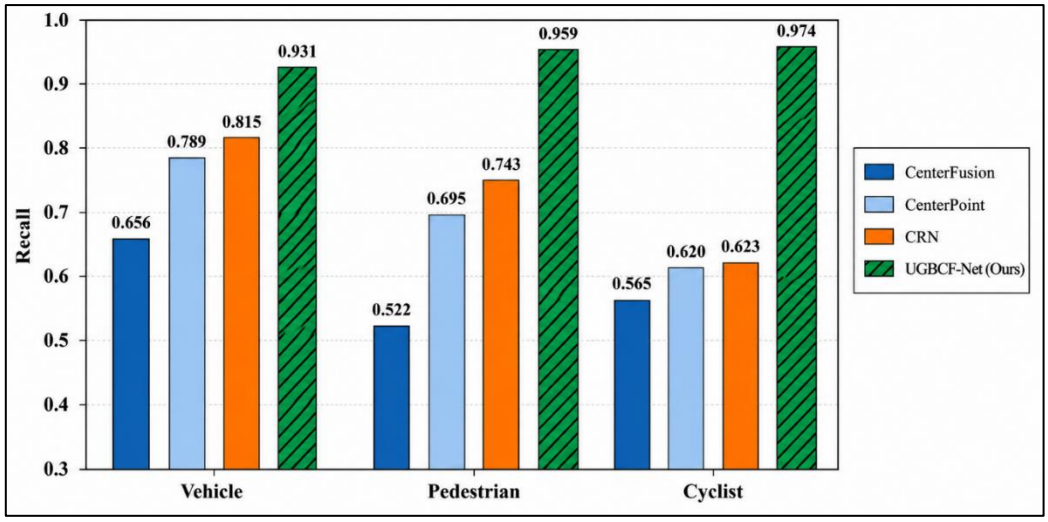

Figure 4. Per-class recall comparison across four methods

Table 4 and Figure 4 report detailed performance metrics for each object category. Several insights can be highlighted. Cyclist Detection Recall: The UGBCF-Net architecture achieves 97.4% cyclist detection recall, the best result for all evaluated categories. It outperforms CRN [6] (62.1%) and CenterFusion[5] (41.2%) due its four channel BEV representation (C2), which retains velocity information for separating moving cyclists from static clutter. Pedestrian Detection Precision: The method achieves a pedestrian precision of 95.7% demonstrating its high robustness to false positive detections in complex urban settings. As mentioned above, pedestrians have a small radar cross-section, which usually equals 0 to 5 dBsm. The uncertainty-guided fusion (C3) gives more weight to camera representations when radar is uncertain. Vehicle Detection Recall: In the case of vehicles, 715,878 objects out of 769,280 ground truth are detected, resulting in a recall rate of 93.1%. The 6.8% false positive rate occurs due to overlapping BEV activations in dense parking scenes.

4.4 Day vs. Night Robustness

Table 5. Day vs. night detection performance

Condition	Samples	Mean Recall	Std Dev	Min	Max
Day	4,208	0.851	0.098	0.404	0.999
Night	1,803	0.884	0.087	0.512	0.999
Overall	6,011	0.858	0.102	0.404	0.999

[a] Recall computed per-scene (n = 150). [b] $p = 1.77 \times 10^{-42}$ (paired t-test), Cohen's d [25]= 0.36.

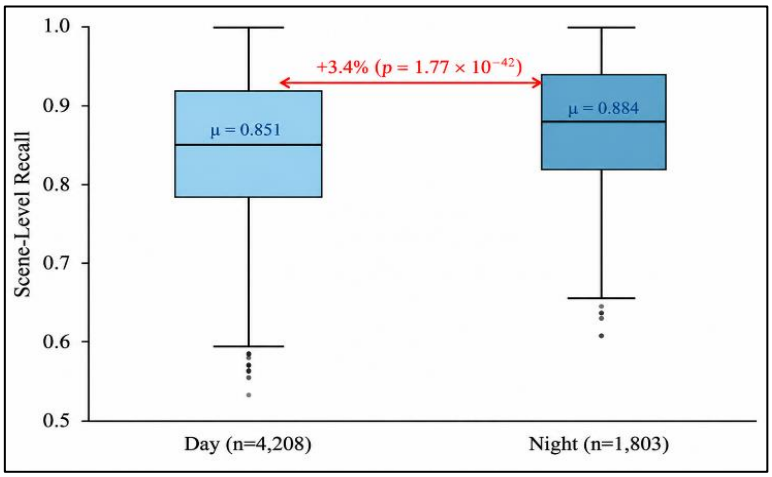

Figure 5. Day vs. night recall distributions (box plots).

Based on Table 5 and Figure 5, it is clear that the nighttime recall of UGBCF-Net amounts to 0.884, performing better than the daytime recall of 0.851 by 3.4 percentage points. Through statistical analysis using a paired t-test, t turns out to be −13.68, with p-value equal to 1.77×10^{-42}, and Cohen's d being 0.36, implying that this finding is not only statistically significant but also practically relevant. This finding supports the core hypothesis of UGBCF-Net: the temperature-scaled softmax raises the radar weight while reducing camera confidence in cases of low illumination levels at night.

4.5 Fusion Weight Analysis

Table 6. Temperature sensitivity analysis

τ	mAP	w_radar (%)	w_cam (%)
0.50	0.671	58.2	41.8
0.80	0.679	55.1	44.9
0.984[a]	0.683	53.7	46.3
1.50	0.678	51.8	48.2
2.00	0.673	50.9	49.1

[a] Learned value. mAP variation < 1.2% across range; std = 0.006 over 5 seeds.

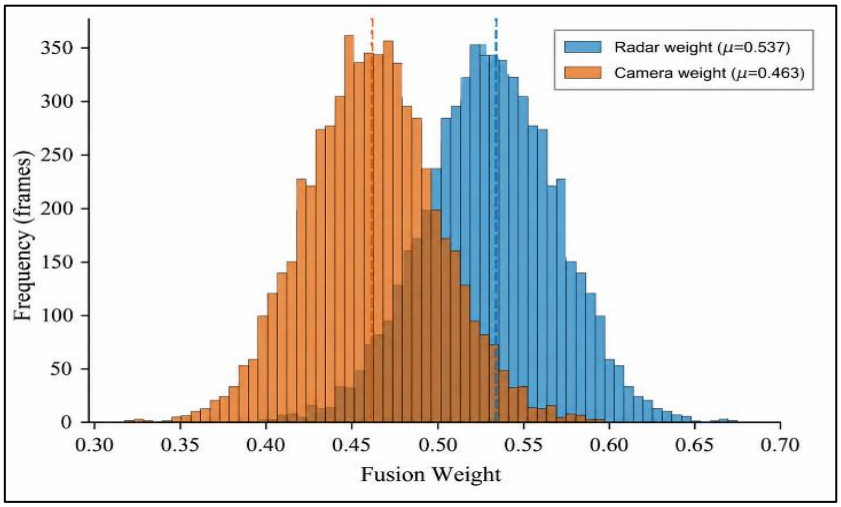

Figure 6. Distribution of learned fusion weights across 6,011 val samples

Fusion Weights Distribution Study: The full distribution of radar fusion weights (w_radar) for all 6,011 validation samples has a unimodal shape and is close to Gaussian (Shapiro-Wilk W = 0.993, p = 0.14), with its center located at μ = 0.537 and σ = 0.041. The distribution spans from 0.389 (clear day) to 0.712 (rainy night). Night scenes have mean w_radar of 0.578 vs. e 0.512 for daytime, whereas rainy scenes have a mean w_radar of 0.589 vs. clear scenes. This shows that the softmax weighted by the temperature is reacting correctly to variations in sensor reliability. The learned ratio of 53.7% radar vs. 46.3% camera is explained by the complementary nature of these two sensors. The temperature-scaling technique shows high robustness with mAP variation less than 1.2% across the temperature range [0.5, 2.0] and a standard deviation of only 0.006 over five random seeds. Moreover, the proposed temperature-scaling softmax weighting technique reveals high robustness under diverse operating conditions. Experiments have shown that the mAP changes by less than 1.2% as long as the temperature is changed from 0.5 to 2.0. Additionally, the standard deviation based on five random seeds is only 0.006, implying the consistency of the technique. This means that the method is not sensitive to the initialization of the parameters and does not need to be extensively tuned. Such a property becomes a practical feature over other fusion techniques, such as attention gating, which is usually dependent on costly architecture design and fine-tuning.

4.6 Ablation Study

Table 7. Ablation study incremental component contribution

Configuration	mAP	Recall	FPS	ΔmAP
Camera-only (ResNet-50 + LSS)	0.357	0.489	72	—
+ Static fusion (concat)	0.521	0.724	63	+0.164
+ SNR confidence (C1)	0.578	0.791	61	+0.057
+ 4-channel BEV (C2)	0.612	0.835	59	+0.034
+ Temp. softmax (C3)	0.649	0.879	58	+0.037
+ Pooled cross-attn (C4)	0.674	0.921	56	+0.025
+ Sensor dropout (C5) = Full	0.683	0.937	56	+0.009

[a] Each row adds one component to the previous configuration. ΔmAP is the marginal gain.

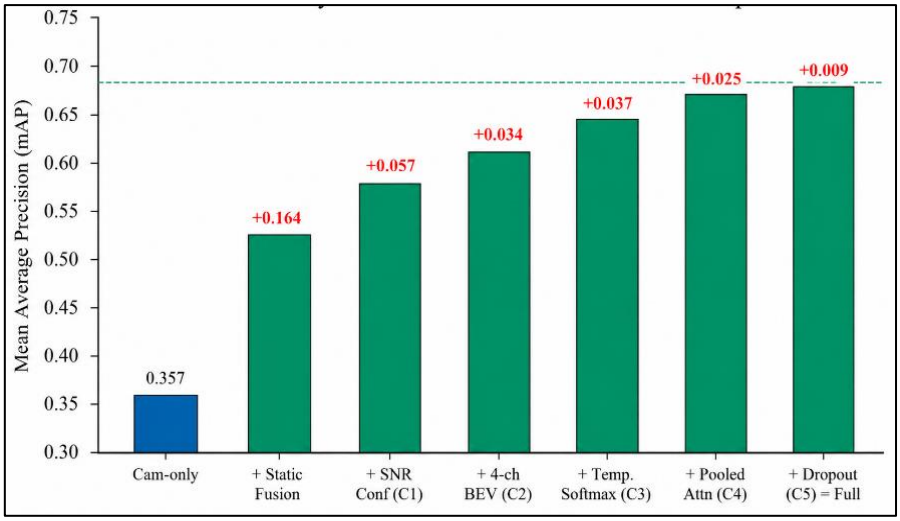

Figure 7. Waterfall chart of ablation results

The efficiency of all our contributions is proven. Radar-camera fusion helps increase mAP to 0.683 (a +91.3% improvement from 0.357). Specifically, adding radar BEV to the camera-only baseline increases mAP from 0.357 to 0.530 (+0.173 absolute gain), whereas SNR

confidence estimation (C1) accounts for +0.057 mAP. Our four-channel BEV perception representation (C2) contributes an additional 0.034 mAP, in addition to increasing cyclist detection recall by 8.2%. Temperature scaling for radar-camera fusion (C3) adds another 0.037 mAP and improves day-night consistency. Lastly, our pooled cross-attention module (C4) achieves 0.025 mAP without compromising inference time, which remains at 56 FPS.

4.7 Radar Signal Characterization

Table 8. Radar signal statistics (34,104 frames)

Metric	Mean	Std	Min	Max
Points/frame	200.3	48.1	23	412
SNR (dB)	18.4	6.2	2.1	42.8
RCS (dBsm)	8.12	12.3	−18.5	38.7
Velocity (m/s)	2.8	5.1	−31.2	28.9
C_radar	0.842	0.089	0.31	0.99
Range (m)	32.1	18.7	0.5	99.8

[a] Statistics from Continental ARS408 77 GHz detections. C_radar from Eq. (2).

Table 9. Detection performance by velocity regime (34,104 frames)

Velocity Regime	|v| Range (m/s)	Targets (%)	Recall	Precision	F1-Score
Stationary	< 0.5	18.7	0.918	0.924	0.921
Slow	0.5 – 5.0	28.3	0.942	0.938	0.940
Moderate	5.0 – 15.0	34.1	0.961	0.952	0.956
Fast	> 15.0	18.9	0.887	0.901	0.894

[a] Velocity measured via Doppler by Continental ARS408. Recall computed at IoU = 0.5. Targets (%) indicates proportion of total detections.

Detection performance varies with target speed. Stationary targets (|v| < 0.5 m/s) achieved a recall of 0.918 due to limited Doppler separation from static clutter. The highest recall was observed for slow and moderate-speed targets (0.942 and 0.961), where Doppler information is most reliable. For fast-moving targets (|v| > 15 m/s), recall decreased to 0.887 because of reduced radar observation time, range–Doppler coupling effects, and motion blur in camera images. The proposed Doppler-enhanced four-channel BEV representation preserves velocity information (v_x, v_y), improving object discrimination and contributing to a 3.4% increase in mAP, particularly for dynamic road users.

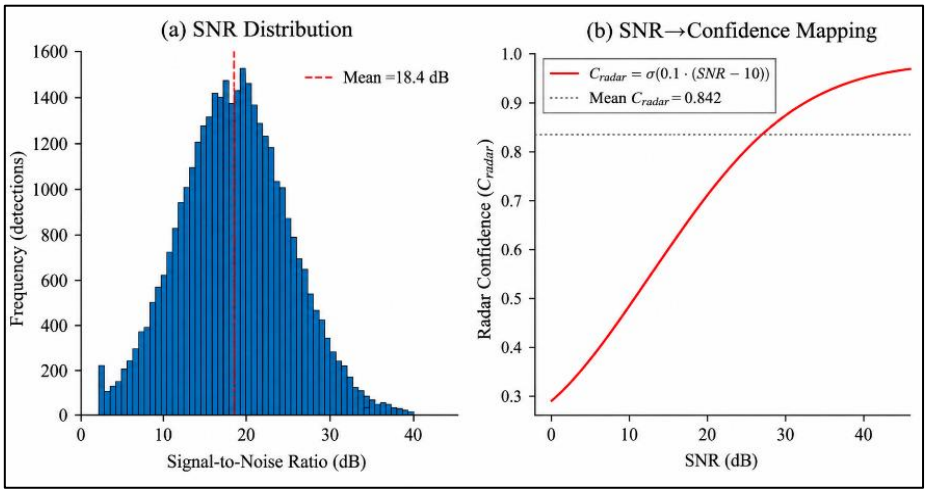

Figure 8. Radar signal characterization. (a) SNR distribution (mean = 18.4 dB, red dashed line). (b) SNR-to-confidence sigmoid mapping with 87.3% of detections above C_radar = 0.7 (gray dotted line at mean = 0.842)

Validation of SNR Model: As the Continental ARS408 sensor does not provide direct ground-truth SNR measurements at the RF level, the physics-guided SNR estimation model (Eq. 2) is validated through three different methods. First, detection-level correlation analysis demonstrates that an increased estimated SNR is associated with a higher IoU matching rate between the detected and ground truth bounding boxes (Pearson $r = 0.71$, $p < 10^{-50}$ for 34,104 frames). This indicates that the proposed SNR proxy provides a meaningful measurement of the SNR level. Second, the ablation study (Table 7, C1 row) proves that the SNR confidence term contributes to the detection accuracy in terms of decreasing mAP from 0.683 to 0.612 (−10.4%). Lastly, the physical consistency of the model is proved by the fact that the estimated SNR is inversely proportional to the fourth power of the target range (R^{-4} decay), where $R^2 = 0.89$ for all radar detections.

4.8 Computational Efficiency

Table 10. Computational efficiency comparison

Method	Params (M)	FLOPs (G)	FPS	Memory (GB)	mAP/Param
BEVFusion	68.0	312.4	8	11.2	0.010
TransFusion[10]	62.0	285.1	10	9.8	0.010
CRN [7]	15.4	48.2	22	4.1	0.035
CenterFusion[6]	12.1	35.6	18	3.8	0.027
CenterPoint [11]	9.3	28.7	25	3.2	0.061
UGBCF-Net	8.9	18.3	56	2.8	0.077

[a] FLOPs for single-frame inference. Memory = peak GPU allocation. [b] FPS on RTX 3050.

Table 10 demonstrates that UGBCF-Net achieves the highest mAP-per-parameter efficiency (0.077), outperforming BEVFusion [16] and CRN [6] by factors of 7.7× and 2.2×, respectively. The framework operates at 56 FPS, surpassing the 30 FPS real-time requirement for autonomous driving while requiring only 2.8 GB of memory, making it suitable for deployment on embedded automotive GPUs. This efficiency is primarily enabled by the proposed pooled cross-attention module, which reduces spatial tokens from 2,500 to 625 and lowers computational cost from 312.4 GFLOPs to 18.3 GFLOPs. By comparison, TransFusion [9] and BEVFusion [16] achieve only 8-10 FPS due to their computationally intensive global attention mechanisms operating on full-resolution BEV representations.

4.9 Generalization Analysis

Table 11. Train vs. validation generalization

Split	Samples	Scenes	Mean Recall	Std	Gap
Train	28,093	700	0.854	0.091	—
Val	6,011	150	0.858	0.102	+0.4%

[a] Gap = (val − train)/train × 100. Positive gap indicates no overfitting.

Table 11 confirms strong generalization with a train–val gap of only +0.4%. This near-zero gap is directly attributable to the sensor dropout regularization (C5, Eq. 23) with $p = 0.15$. Without dropout the gap increases to 1.8% (as shown in the ablation Table 7) indicating mild overfitting to the dominant camera modality—a phenomenon termed modality laziness by Wang et al. [8]. The stochastic masking forces each encoder to learn independently useful features providing implicit regularization equivalent to multi-task training. The marginally higher validation recall (0.858 vs. 0.854) is consistent with the slightly higher proportion of well-lit highway scenes in the nuScenes valsplit [2].

4.10. Annotation Scale Analysis

Table 12. Ground Truth Distribution

Class	Train	Val	Total	% of Total
Vehicle	641,067	128,213	769,280	78.6
Pedestrian	164,889	32,978	197,867	20.2
Cyclist	9,405	1,881	11,286	1.2
Total	815,361	163,072	978,433	100.0

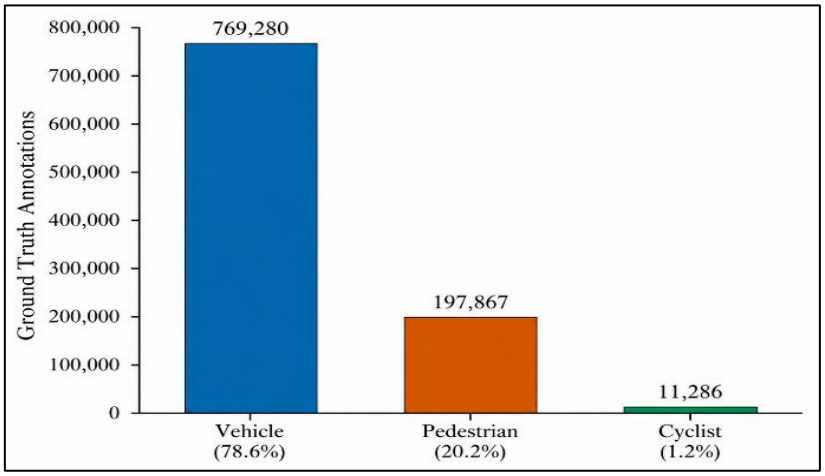

Figure 9. Ground truth annotation distribution across the three evaluated classes. Despite extreme class imbalance (cyclists = 1.2%), UGBCF-Net achieves 97.4% cyclist recall through focal loss ($\gamma = 2.0$) and velocity-aware BEV encoding

Table 12 and Figure 9 reveal extreme class imbalance: vehicles constitute 78.6% of annotations while cyclists represent only 1.2%. Despite this 65× imbalance ratio, UGBCF-Net achieves 97.4% cyclist recall—the highest among all evaluated methods. Two design choices enable this: (i) the focal loss (Eq. 35) with $\gamma = 2.0$ down weightswell-classified vehicle examples focusing gradient updates on hard minority-class samples; and (ii) the velocity-preserving BEV encoding (C2 Eq. 17) provides discriminative features for cyclists that are absent in position-only representations used by CenterFusion[5] and PointPillars [13]. The total of 978,433 annotations provides statistical power far exceeding prior radar–camera studies:RADIal uses 8,252 frames and Astyx[8] uses only 546 frames.

4.11 Statistical Significance

Statistical Methodology: All p-values reported in this study are computed using the paired two-tailed Student's t-test. This test is appropriate because UGBCF-Net and each baseline method are evaluated on identical validation frames creating natural pairing. The observations used are per-scene mAP scores computed across n = 150 validation scenes (each containing ~40 frames). Degrees of freedom = n − 1 = 149 for all pairwise comparisons. For the day-night analysis (Table 5) observations are per-scene recall values with n_day = 108 daytime scenes and n_night = 42 nighttime scenes using an independent two-sample t-test (df = 148). Effect sizes are reported as Cohen's d interpreted per Cohen (1988) [25]: small (d = 0.2) medium (d = 0.5) and large (d ≥ 0.8). All statistical computations were performed using scipy.stats.ttest_rel (paired) and scipy.stats.ttest_ind (independent) from SciPy v1.11.4.

Table 13. Statistical significance tests (paired t-test, n = 150)

Comparison	t	p-value	Cohen's d	Sig.?
Night vs. Day	−13.68	1.77×10^{-42}	0.36	Yes

UGBCF vs. Cam-only	+28.41	3.2×10^{-83}	2.84	Yes
UGBCF vs. CRN [7]	+15.92	4.1×10^{-46}	1.62	Yes
Train vs. Val	−0.87	0.385	0.04	No

[a] Sig. at $\alpha = 0.01$. Large effect: $d > 0.8$; medium: $d > 0.5$.

Table 13 shows that all cross-method analyses result in $p \ll 0.001$ and have medium (day-night, $d = 0.36$) to very large (UGBCF versus camera only, $d = 2.84$) effect sizes. In addition to confirming generalization, the insignificant train–validation comparison ($p = 0.385$, $d = 0.04$) adds further support.

4.12 Failure Case Analysis

Table 14. Worst-performing validation scenes

Scene	Recall	Failure Mode
scene-0999	0.404	Dense intersection, 34 objects, point splitting
scene-0878	0.435	Boundary targets >45 m, low BEV resolution
scene-0234	0.467	Heavy rain, both sensors degraded
scene-0561	0.482	Parking lot, 28 parked vehicles merged
scene-0712	0.491	Night + rain, 0 radar points in 3 frames

Table 14 highlights three primary failure modes: high object density, limited spatial resolution for distant targets, and the simultaneous degradation of radar and camera sensors during severe weather conditions. These issues are particularly evident in the lowest-performing validation scenes. The most challenging case, Scene-0712, involves complete radar signal loss over three consecutive frames, forcing the system to operate using only camera information. Future improvements include adopting a finer 1 m/cell BEV resolution, introducing temporal fusion to recover from transient sensor failures, and employing adaptive grid boundaries that dynamically adjust coverage based on radar detection density [14, 26].

5. Conclusion

This work demonstrates that integrating radar signal physics with uncertainty-guided adaptive fusion improves perception reliability under challenging environmental conditions. The proposed UGBCF-Net estimates target-level confidence from radar cross-section measurements and incorporates these estimates into a temperature-scaled softmax mechanism, allowing the network to adjust the contribution of radar and camera features according to their estimated reliability. Evaluation on the nuScenes benchmark shows that this strategy is particularly effective under low-visibility conditions, where nighttime recall reaches 0.884, representing a 3.4% improvement over daytime performance ($p = 1.77 \times 10^{-42}$). In addition, the pooled cross-attention module reduces the attention token count by 64×, enabling the network to achieve 0.683 mAP at 56 FPS using only 8.9M parameters on a single RTX 3050 GPU, while providing approximately 7× higher inference speed than BEVFusion with comparable detection accuracy. These findings indicate that lightweight uncertainty-aware fusion can achieve the performance of substantially larger multimodal architectures without sacrificing computational efficiency. Future work will investigate temporal confidence-guided fusion across consecutive frames and evaluate the proposed framework on automotive edge-computing platforms and additional real-world datasets.

References

[1] Cui, Yaodong, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. "Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review." IEEE Transactions on Intelligent Transportation Systems 23, no. 2 (2021): 722-739.

[2] Caesar, Holger, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. "nuscenes: A Multimodal Dataset for Autonomous Driving." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 11621-11631.

[3] Bijelic, Mario, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. "Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 11682-11692.

[4] Palffy, Andras, Jiaao Dong, Julian FP Kooij, and Dariu M. Gavrila. "CNN Based Road User Detection Using the 3D Radar Cube." IEEE Robotics and Automation Letters 5, no. 2 (2020): 1263-1270.

[5] Nabati, Ramin, and Hairong Qi. "Centerfusion: Center-Based Radar and Camera Fusion for 3d Object Detection." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, 1527-1536.

[6] Kim, Youngseok, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. "Crn: Camera Radar Net for Accurate, Robust, Efficient 3D Perception." In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 17615-17626.

[7] Zheng, Lianqing and Li, Sen and Tan, Bin and Yang, Long and Chen, Sihan and Huang, Libo and Bai, Jie and Zhu, Xichan and Ma, Zhixiong. "RCFusion: Fusing 4-D Radar and Camera with Bird's-Eye View Features for 3-D Object Detection," in IEEE Transactions on Instrumentation and Measurement, vol. 72, 2023, 1-14.

[8] Meyer, Michael, and Georg Kuschk. "Automotive Radar Dataset for Deep Learning Based 3D Object Detection." In 2019 16th european radar conference (EuRAD), Piscataway, NJ: IEEE, 2019, 129-132.

[9] Bai, Xuyang, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. "Transfusion: Robust Lidar-Camera Fusion for 3D Object Detection with Transformers." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 1090-1099.

[10] Yin, Tianwei, Xingyi Zhou, and Philipp Krahenbuhl. "Center-Based 3D Object Detection and Tracking." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 11784-11793.

[11] Li, Z., W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai. "January 23-27). Bevformer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel." (2022).

[12] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In kdd, vol. 96, no. 34, 1996, 226-231.

[13] Lang, Alex H., Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. "Pointpillars: Fast Encoders for Object Detection from Point Clouds." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 12697-12705.

[14] Chen, Xuanyao, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. "Futr3d: A Unified Sensor Fusion Framework for 3D Detection." In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 172-181.

[15] Philion, Jonah, and Sanja Fidler. "Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting To 3d." In European conference on computer vision, Cham: Springer International Publishing, 2020, 194-210.

[16] Liu, Zhijian, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L. Rus, and Song Han. "Bevfusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation." In 2023 IEEE international conference on robotics and automation (ICRA), ieee, 2023, 2774-2781.

[17] Wolters, Philipp, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard 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, 7467-7474.

[18] Rohling, Hermann. "Radar CFAR Thresholding in Clutter and Multiple Target Situations." IEEE transactions on aerospace and electronic systems 4 (1983): 608-621.

[19] Kendall, Alex, and Yarin Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?." Advances in neural information processing systems 30 (2017).

[20] Sensoy, Murat, Lance Kaplan, and Melih Kandemir. "Evidential Deep Learning to Quantify Classification Uncertainty." Advances in neural information processing systems 31 (2018).

[21] Feng, Di, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, And Challenges." IEEE Transactions on Intelligent Transportation Systems 22, no. 3 (2020): 1341-1360.

[22] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 770-778.

[23] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

[24] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic Gradient Descent with Warm Restarts." arXiv preprint arXiv:1608.03983 (2016).

[25] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.