J Intel Connect Veh

Research Article

FGA-BEVFusion: Frequency-Guided Attention for LiDAR–Camera BEV Fusion in 3D Object Detection

Zechang Ye[1,2], Haiyang Yu[1,2,3,✉], Shuai Liu[2,3], Han Jiang[1,2], Yilong Ren[1,2,3]

Cite this article: https://doi.org/10.26599/JICV.2026.9210111

ABSTRACT: Multi-sensor 3D object detection is essential for autonomous driving; however, it faces three persistent challenges: scale variations across objects, structural degradation of fine-grained features during view transformation, and ambiguous object boundaries in complex scenes. To address these issues, we propose FGA-BEVFusion, a LiDAR–camera bird's-eye-view (BEV) fusion framework built upon a novel frequency-guided attention (FGA) mechanism that explicitly preserves structural information in the frequency domain. FGA operates in the view space before BEV projection: a Multi-Spectral Channel Attention (MCA) module adaptively reweights spectral responses, and a Structure-Aware Edge Injection (SEI) module enhances spatial boundary sensitivity. In the BEV space, a Sparsity-Aware BEV Refinement (SAR) module performs occupancy-guided hierarchical fusion and spatial recalibration to reduce information loss during decoding. Experiments on the nuScenes benchmark show that FGA-BEVFusion achieves 70.8% mAP and 72.3% NDS on the validation set, with consistent improvements on small and distant object categories and under challenging conditions such as rain and nighttime.

KEYWORDS: Autonomous driving, 3D object detection, bird's-eye-view (BEV), LiDAR–camera fusion, frequency-domain attention, structural preservation.

1 Introduction

With the rapid development of autonomous driving technology, the accuracy and efficiency of multi-sensor perception systems have become increasingly critical. Among various computer vision tasks, 3D object detection is particularly vital. However, single-sensor methods often suffer from inherent performance limitations, encouraging the use of multi-modal fusion. Reliable LiDAR–camera fusion detection remains highly sensitive to both spatial alignment and timestamp synchronization. Beyond the vehicle-mounted setting studied in this work, such fusion is also a key enabler for potential application contexts such as roadside perception and cooperative driving (Yu et al., 2026). The BEV paradigm has attracted substantial academic and industrial interest for its ability to provide a unified geometric representation. By projecting multi-modal data into a top-down coordinate space, BEV effectively preserves geometric structure and semantic information, thereby offering rich spatial context to support downstream tasks, such as 3D object detection, scene segmentation, and path planning. However, many existing approaches neglect frequency-domain information, which limits the efficiency of cross-modal fusion and weakens the representation of multi-scale targets. Consequently, improving the balance between global and local BEV characterization through multi-frequency feature fusion remains a critical challenge.

Moreover, robust BEV feature transformation has recently been recognized as another key factor for improving multi-sensor 3D detection under sensor perturbation and spatial inconsistency (Liu et al., 2025). The lift-splat-shoot (LSS) framework improves multi-view fusion flexibility by estimating discrete depth distributions to form a 3D feature point cloud, which is subsequently projected into the BEV space (Philion and Fidler, 2020). Nevertheless, its reliance on ground-plane assumptions restricts its ability to handle geometric distortions of non-ground objects. Algorithms such as BEVDet and BEVFormer incorporate temporal information via Transformer architecture to enhance scene understanding in dynamic environments (Huang et al., 2021; Li et al., 2022b). Despite their progress, these methods continue to face challenges in achieving accurate depth estimation and effective background noise suppression. For instance, in the LSS framework, approximately 98% of the generated BEV features represent background information, which can dominate the target features and degrade detection accuracy (Li et al., 2022b).

A central challenge in multi-sensor BEV perception is the effective fusion of heterogeneous sensor data, with most existing methods concentrating on feature-level integration of LiDAR and image modalities. For instance, BEVFusion primarily employs basic fusion operations, such as element-wise concatenation, summation, or SE modules (Liu et al., 2023). However, in practical scenarios, these approaches inadequately address the spatial misalignment between multi-view images and LiDAR inputs.

Just Accepted

1 School of Transportation Science and Engineering, Beihang University, Beijing 100191, China. 2 State Key Lab of Intelligent Transportation System, Beihang University, Beijing 100191, China. 3 Hangzhou Innovation Institute, Beihang University, Hangzhou 310056, China.

✉ Corresponding author. E-mail: hyyu@buaa.edu.cn

Received: September 1, 2026; Revised: September 26, 2026; Accepted: October 27, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

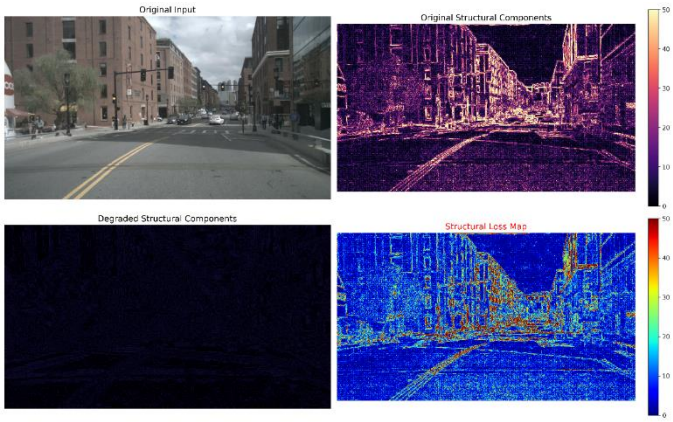


Fig. 1. Illustration of structural information loss during the view-to-BEV projection. (a) Input image from the nuScenes dataset. (b) High-frequency features extracted using a Laplacian operator. (c) The resulting feature map after standard view transformation, showing obvious blur and detail loss. (d) The absolute difference between (b) and (c). The highlighted areas (red/yellow) indicate that the projection process mainly discards edge details of foreground objects, such as vehicles and traffic poles.

While multi-view images inherently suffer from degradation factors such as motion blur and limited resolution, the 2D-to-3D transformation may introduce an additional "structural degradation". This structural degradation is visually confirmed in **Fig. 1.**, where high-pass filtering reveals that the view transformation predominantly discards the fine geometric boundaries required for precise localization. The LSS paradigm fundamentally relies on accurate depth estimation to project 2D features into 3D space (Li et al., 2024a). Due to the ill-posed nature of monocular depth estimation, pixels belonging to fine structural details (e.g., edges of poles, distant vehicles) are often assigned incorrect depths, causing them to be scattered across multiple voxels in the BEV space. This aggregation process may behave analogously to a low-pass filter. We emphasize that this interpretation is a motivating hypothesis, rather than providing direct spectral evidence for the underlying mechanism. Currently, multi-sensor BEV object detection faces two primary challenges:

1) Degradation of sensor information structure: It occurs when high-frequency details, such as object edges, corners, and fine textures, are often blurred or lost due to downsampling, interpolation, and cross-modal misalignment. This effect is especially pronounced for small or distant objects, resulting in reduced localization accuracy and detection confidence.

2) Feature scale conflict: It arises when distant targets yield sparse BEV features, while nearby objects produce redundant feature representations, making single-frequency fusion strategies inadequate for robust multi-scale detection requirements.

Multi-sensor 3D object detection is challenged by scale variation, noise, and ambiguous object boundaries, especially due to the loss of fine structural details during fusion. To address structural degradation in multi-sensor BEV fusion, we propose frequency-guided attention (FGA), which performs frequency-aware feature refinement at the view-space level before BEV projection. FGA selectively enhances high-frequency components to preserve critical edge and texture information, employing Multi-Spectral Channel Attention (MCA) for spectral recalibration followed by Structure-Aware Edge Injection (SEI) to improve spatial sensitivity. This sequential design enables progressive, modality-adaptive feature refinement while suppressing noise. To address information degradation during decoding, we introduce the Sparsity-Aware BEV Refinement module (SAR) that integrates hierarchical BEV fusion with spatial recalibration. By preserving cross-scale feature consistency, the SAR improves the precision and coherence of object localization. On the nuScenes benchmark, FGA achieves strong performance in both mean Average Precision (mAP) and nuScenes Detection Score (NDS), and improves detection accuracy for multiple object classes, including those in complex traffic scenarios.

In summary, the main contributions of this work are threefold:

1) We identify and formalize the structural degradation problem in multi-sensor BEV detection, i.e., the attenuation of high-frequency geometric information during view transformation, and analyze its impact on the detection of small and distant objects.

2) We propose FGA, a modular framework that combines MCA and SEI to preserve frequency-aware feature refinement and preserve structural boundaries at early processing stages.

3) We design a Sparsity-Aware BEV Refinement (SAR) module to address information loss during multi-modal decoding. Extensive experiments on the nuScenes benchmark demonstrate that our method achieves consistent gains on small and distant object categories and under challenging traffic scenarios.

2 Related Work

2.1. Camera-Based 3D Object Detection

BEV camera perception has emerged as a critical focus in autonomous driving research; however, it continues to face challenges in accurate depth estimation and effective background noise suppression, both of which critically affect 3D object detection and scene understanding. The seminal LSS framework introduced an explicit geometric projection by estimating pixel-wise discrete depth distributions and aggregating multi-view image features into a unified BEV representation (Philion and Fidler, 2020). Despite its strong geometric consistency, reliance on coarse depth discretization leads to pronounced long-range depth errors and excessive background feature activations (Li et al., 2024a). Building on this foundation, BEVDepth mitigated depth ambiguity by integrating camera intrinsic and extrinsic parameters into the depth estimation. (Li et al., 2023b) This architectural enhancement yielded a 10% mAP improvement on the nuScenes benchmark compared with LSS. However, these methods rely on LiDAR-based pseudo-supervision, which can introduce systematic depth bias in areas with sparse or missing LiDAR coverage.

To avoid explicit depth estimation, implicit paradigms, such as BEVFormer (Li et al., 2022b) and CVT (Zhou and Krähenbühl, 2022), learn end-to-end view transformations using deformable architectures. BEVFormer mitigates depth-induced geometric distortions via spatially adaptive deformable cross-attention that aligns multi-view features with BEV queries. However, the absence of explicit geometric priors intensifies depth ambiguity for small or distant objects, and the high computational complexity of global attention further amplifies background noise in dense urban environments. The hybrid framework BEVHeight++ (Yang et al., 2023), which jointly optimizes depth and height estimation, attempts to address these issues but struggles to balance computational

Just Accepted

efficiency with detection accuracy.

Despite recent advances, background interference remains a fundamental limitation in BEV-based perception systems, significantly impairing detection robustness and scene understanding. BEVDet reports that over 90% of 3D detection false positives arise from background feature misclassification, underscoring the necessity of adaptive noise suppression in BEV perception pipelines (Huang et al., 2021). Recent innovative approaches, such as SA-BEV, employ semantic-aware pooling to attenuate low-confidence background regions, substantially increasing target feature density by a factor of five (Zhang et al., 2023). IA-BEV introduces instance-aware depth estimation with structural priors mining and self-boosting learning, showing that explicitly emphasizing object-aware structural cues can improve BEV feature construction and downstream detection accuracy (Jiao et al., 2024). MvACon further improves query-based multi-view 3D detection by enhancing 2D-to-3D feature lifting with attentive contextualization, demonstrating that richer contextualized image features remain critical during view transformation (Liu et al., 2024). More recently, GeoBEV highlights that existing methods often overlook the geometric quality of BEV representation and proposes dedicated designs to recover fine-grained geometric information in BEV space (Zhang et al., 2025). However, their reliance on high-quality semantic annotations limits scalability due to annotation cost. Multi-task frameworks, such as BEVerse, jointly optimize detection and segmentation to exploit contextual cues but often incur performance degradation from inter-task feature competition (Zhang et al., 2022). In summary, camera-based BEV perception has progressed from explicit depth lifting toward geometry- and structure-aware designs, yet the spectral characteristics of the features being transformed are rarely modeled explicitly; this gap motivates the frequency-guided design presented in this work.

2.2. LiDAR-Based 3D Object Detection

BEV LiDAR-based perception has achieved significant progress in 3D object detection; however, intrinsic challenges related to point cloud sparsity and multi-scale feature extraction continue to limit robustness, particularly for distant and occluded objects. Early voxel-based methods discretized point clouds into uniform 3D grids to enable convolutional feature learning, but this homogeneous voxelization leads to significant information loss in sparse regions and frequent missed detections (Zhou et al., 2018; Deng et al., 2021). To address computational inefficiency, SECOND introduced sparse 3D convolution, reducing inference time by 40% while maintaining competitive accuracy (Yan et al., 2018). Nevertheless, voxel-based methods remain constrained by irregular point distributions and high memory demands in large-scale deployment scenarios.

Pillar-based approach later emerged as a lightweight alternative, projecting point clouds into 2D pseudo-images via vertical aggregation (Lang et al., 2019). Although this design improves real-time performance, the compression of vertical information degrades detection accuracy for small objects, such as non-motorized vehicles and pedestrians, due to vertical feature compression (Noh et al., 2021). A recent study enhanced multi-scale localization by integrating point-wise and voxel-level features through attention mechanisms, yielding improved detection across diverse spatial resolutions (Shi et al., 2023). Nevertheless, the pillar-based architectures remain fundamentally constrained in modeling vertical structure, limiting their ability to resolve vertical ambiguities, such as distinguishing overlapping vehicles with similar horizontal positions.

Background noise suppression also remains a major challenge, as non-target points contribute to over 70% of false alarms in LiDAR-based detection systems (Wang et al., 2020). SAFDNet proposes a fully sparse detector to alleviate the inefficiency of dense feature construction in long-range settings (Zhang et al., 2024). On the other hand, FSHNet enhances fully sparse detection by strengthening long-range interactions and refining downsampled sparse voxels, which is particularly beneficial for preserving fine-grained details of small objects (Liu et al., 2025). UniMamba explores state-space modeling for LiDAR detection and demonstrates that jointly modeling local structure and global context is beneficial for sparse and occluded point clouds (Jin et al., 2025). Although Pyramid R-CNN introduced an adaptive RoI head to prioritize sparse foreground features, its effectiveness in cluttered scenes is limited by insufficient background discrimination (Shi et al., 2022). Overall, LiDAR-based detectors highlight the importance of preserving fine-grained structure in sparse point clouds, a principle that our SAR module extends to the fused BEV space.

2.3. Multi-Modal Fusion

Early multi-modal fusion frameworks (Sindagi et al., 2019; Yoo et al., 2020) combined LiDAR and camera data at the point or voxel level, but they were often susceptible to sensor calibration inaccuracies. BEVFusion projects both modalities into a unified BEV space through optimized pooling, thereby achieving strong performance on the nuScenes benchmark (Liang et al., 2022). However, its dependence on accurate extrinsic calibration limits system robustness in dynamic and real-world environments. Recent work has also explored LiDAR-camera fusion under feature degradation conditions, showing that useful cross-modal cues may be weakened during deep hierarchical abstraction (Ye et al., 2025).

To mitigate modality-specific biases, TransFusion adaptively fuses LiDAR BEV features with semantically rich camera cues using Transformer-based cross-attention, resulting in lower false positive rates in complex environments (Bai et al., 2022). Temporal fusion paradigms, such as SOLOFusion, further enhanced occlusion reasoning by integrating long- and short-term BEV features, though they remain affected by depth errors propagated from single-frame predictions (Park et al., 2022). BEVDepth introduced explicit depth supervision via LiDAR-derived pseudo-labels, yet it continues to struggle in LiDAR-blind regions, such as the sky and other occluded areas (Li et al., 2023b). IS-Fusion explicitly integrates instance-level and scene-level multi-modal context, showing that object-centric fusion can complement conventional BEV scene fusion (Yin et al., 2024). GAFusion introduces multiple LiDAR-guided cues, including sparse depth guidance and LiDAR occupancy guidance, to enhance cross-modal interaction and adaptive fusion (Li et al., 2024b). More recently, EVT points out that conventional view transformation often suffers from insufficient geometric guidance and ray-direction misalignment, and proposes a more efficient transformation design for multi-sensor BEV detection (Lee et al., 2025). This issue is closely related to the longstanding challenge of LiDAR-camera calibration,

Just Accepted

which remains a critical prerequisite for reliable cross-modal perception in practical autonomous systems. BEVHeight++ jointly optimized object height regression and depth estimation to improve long-range detection performance, yet it still demands substantial computational resources (Yang et al., 2023). In contrast, HeightFormer explicitly captures vertical spatial relationships via height-aware attention mechanisms, effectively reducing ground clutter on the nuScenes benchmark (Wu et al., 2023). Nonetheless, background interference remains a fundamental challenge in BEV-based perception. Multi-modal self-supervised approaches, which seek to eliminate dependence on annotated data by aligning voxel features across different modalities, frequently suffer from unstable training when inter-modal overlap is limited (Xu et al., 2021; Li et al., 2022a). Taken together, these methods establish BEV-space fusion as a strong paradigm for multi-modal 3D detection; however, fusion is typically performed without distinguishing the spectral content of the features being merged, which directly motivates our frequency-guided design.

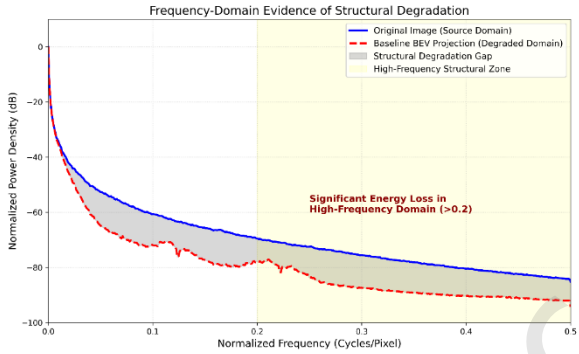


Fig. 2. Frequency-domain comparison between the original image and the baseline BEV projection. The BEV projection exhibits a clear spectral energy drop, particularly in the high-frequency region, suggesting a substantial loss of structural details that may be associated with the view transformation.

3 Method

The premise of our approach is that the standard view transformation paradigm may compromise high-frequency geometric details. To better understand the representation bottleneck in BEV generation, we analyze the frequency-domain characteristics of the original image and its baseline BEV projection. As shown in Fig. 2, the projected BEV representation exhibits a consistent energy drop over the full spectrum, with a notably larger degradation in the high-frequency region. This suggests that fine-grained structural information, such as edges, contours, and local textures, may be weakened during view transformation. To address this, we introduce FGA-BEVFusion, a framework designed to preserve structural integrity across the 2D-to-3D manifold. As illustrated in Fig. 3, the pipeline operates in three distinct phases: (A) View-Space Frequency Enhancement, (B) Geometry-Aware View Transformation, and (C) Sparsity-Aware BEV Refinement.

The overall pipeline proceeds in the order of sensor input, feature extraction, viewpoint transformation, BEV fusion, and detection output. Specifically, multi-view images and LiDAR point clouds are first encoded by modality-specific backbones; the image features are then refined by the FGA modules (MCA and SEI) in the view space; the refined features are lifted and projected onto the BEV plane through a geometry-aware view transformation, in parallel with the voxelized LiDAR features; the two BEV representations are pre-fused and further refined by the SAR module; and the resulting dense BEV feature map is finally fed into the detection head to produce 3D bounding boxes.

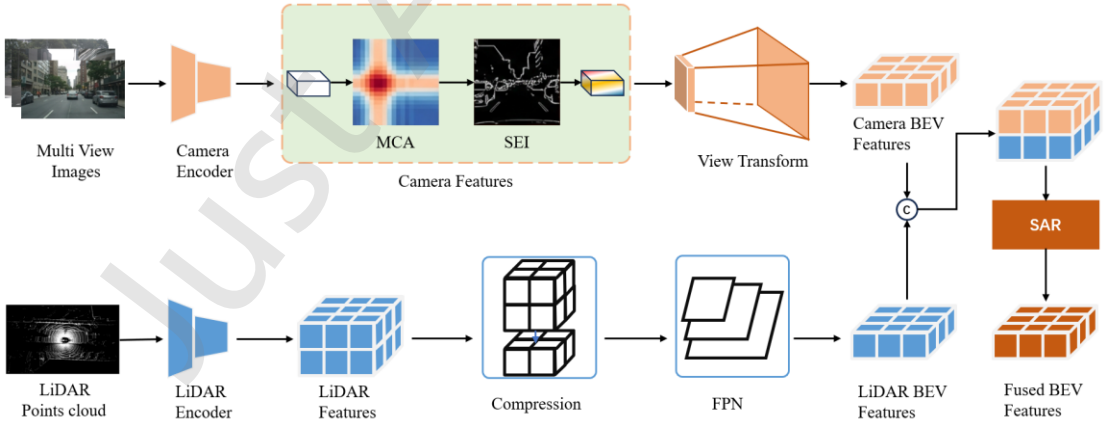


Fig. 3. Pipeline of the proposed FGA-BEVFusion method. Multi-view image features are first enhanced by the MCA and SEI modules and then transformed into BEV representations. These representations are concatenated and further refined by the SAR module to produce the final fused features for robust 3D object detection.

3.1. Structural Degradation in view-to-BEV Transformation

Let $\boldsymbol{I} \in \mathbb{R}^{3\times H_{img}\times W_{img}}$ denote a multi-view RGB image and $\mathcal{E}(\cdot)$ be the encoder. The extracted feature map often retains semantic abstractions (low-frequency components) while attenuating structural boundaries (high-frequency components) as Eq. (1):

$$\boldsymbol{F} = \mathcal{E}(\boldsymbol{I}) \in \mathbb{R}^{C\times H\times W} \tag{1}$$

In lift-splat-based pipelines, each pixel feature is lifted along the camera ray according to a predicted depth distribution and then aggregated into BEV voxels. Due to depth ambiguity, features associated with high-frequency structures are often scattered across multiple depth bins. The subsequent voxel aggregation and height-wise pooling can be expressed as a smoothing operator $\mathcal{A}(\cdot)$ yielding the BEV feature as Eq. (2):

$$\boldsymbol{F}_{BEV} = \mathcal{A}(\boldsymbol{F}) \tag{2}$$

To characterize the loss of structural information, we consider the relative high-frequency energy ratio as Eq. (3):

$$r_{\text{high}}(\mathbf{F}) = \frac{\int_{\Omega_{\text{high}}} |\mathcal{F}(\mathbf{F})(u,v)|^2 \, du\, dv}{\int_{\Omega} |\mathcal{F}(\mathbf{F})(u,v)|^2 \, du\, dv} \tag{3}$$

Just Accepted

where $\mathcal{F}(\cdot)$ denotes the Fourier transform. Ω_{high} denotes the high-frequency domain. If the structural cues are smoothed out in $\boldsymbol{F}$, the resulting 3D projection will suffer from "spatial aliasing," leading to blurred object boundaries in the BEV space. r_{high} decreases after view lifting and BEV pooling, especially for thin structures and distant objects. This observation leads us to introduce a frequency-aware compensation step before view projection.

3.2. View-Space Frequency Enhancement

To pre-compensate for the aforementioned feature loss, we propose the FGA module operating in the 2D view space. It consists of two orthogonal mechanisms: Multi-Spectral Channel Attention (MCA) for restoring frequency sensitivity, and Structure-Aware Edge Injection (SEI) for recovering spatial gradients.

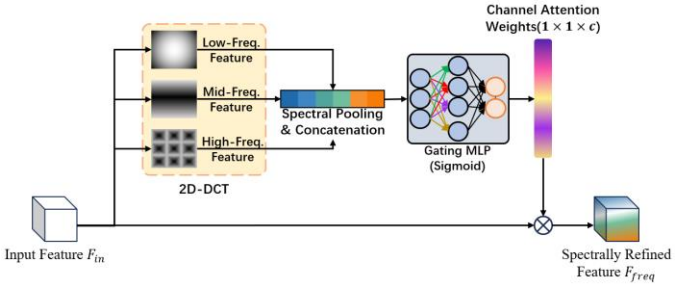


Fig. 4. Illustration of the proposed frequency-guided channel attention mechanism. The input feature $\boldsymbol{F}_{in}$ is decomposed by 2D-DCT into low-frequency, mid-frequency, and high-frequency components, which are further fused by spectral pooling and concatenation. A gating MLP is then used to generate channel-wise attention weights for adaptive feature reweighting, yielding the spectrally refined feature $\boldsymbol{F}_{freq}$.

Multi-Spectral Channel Attention (MCA): Given the view-space feature map $\boldsymbol{F}_{in} \in \mathbb{R}^{C\times H\times W}$, conventional channel attention applies Global Average Pooling (GAP) as Eq. (4):

$$g_c = \frac{1}{HW}\sum_{i,j} \boldsymbol{F}_{in}(c,i,j) \tag{4}$$

Standard GAP reduces spatial dimensions to a single scalar, treating all frequencies equally. This is suboptimal for geometric perception. Instead, we employ the 2D Discrete Cosine Transform (2D-DCT) to compress spatial information while preserving spectral characteristics. The 2D-DCT basis function $B_{i,j}^{h,w}$ for frequency component (h,w) is defined as Eq. (5):

$$B_{i,j}^{h,w} = \cos\left(\frac{\pi h(2i+1)}{2H}\right)\cos\left(\frac{\pi w(2j+1)}{2W}\right) \tag{5}$$

We select a fixed set of K frequency indices $\{(\mathrm{h_k}, \mathrm{w_k})\}_{\mathrm{k=1}}^{\mathrm{K}}$ covering low-to-mid bands. The vector for channel c is computed by projecting the feature onto the bases as Eq. (6):

$$z_c(k) = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \boldsymbol{F}_{in}(c,i,j) B_{i,j}^{h_k,w_k} \tag{6}$$

A shared MLP maps $z_c \in \mathbb{R}^K$ to a scalar channel weight as Eq. (7):

$$\mathrm{a_c} = \sigma\big(\mathrm{W_2}\,\mathrm{ReLU}(\mathrm{W_1 z_c})\big) \tag{7}$$

yielding frequency-recalibrated features as Eq. (8):

$$\boldsymbol{F}_{freq}(\mathrm{c,i,j}) = \mathrm{a_c}\cdot \boldsymbol{F}_{in}(\mathrm{c,i,j}) \tag{8}$$

Through this process, MCA compresses features while explicitly preserving frequency responses relevant to geometric structures. Specifically, the reweighting in Eq. (8) is an element-wise product broadcast along the spatial dimensions: the gating weights are scalars normalized by a sigmoid function and therefore operate across channels.

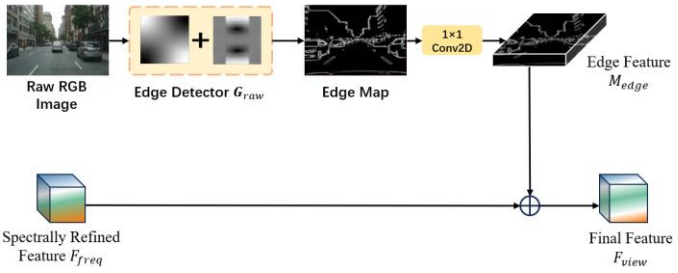


Fig. 5. Structure of the proposed edge-enhancement module. The raw RGB image is converted into an edge map by the edge detector $\boldsymbol{G}_{raw}$, and a $\mathbf{1}\times\mathbf{1}$ Conv2D layer is used to generate the edge feature $\boldsymbol{M}_{edge}$. The edge feature is then fused with the spectrally refined feature $\boldsymbol{F}_{freq}$ to produce the final feature $\boldsymbol{F}_{view}$. This design explicitly injects structural boundary cues into the visual representation, thereby improving contour preservation and fine-grained object characterization in BEV perception.

Structure-Aware Edge Injection (SEI): MCA performs global spectral recalibration but does not explicitly enforce spatial discontinuities. We thus introduce an edge prior computed from the raw image using Sobel filters as Eq. (9):

$$\boldsymbol{G}_{raw} = \sqrt{\big(\mathcal{S}_x(\boldsymbol{I})\big)^2 + \big(\mathcal{S}_y(\boldsymbol{I})\big)^2} \tag{9}$$

The edge map is normalized and downsampled as Eq. (10):

$$\boldsymbol{M}_{edge} = \mathrm{Norm}\big(\mathrm{Down}(\boldsymbol{G}_{raw})\big) \tag{10}$$

where $\mathrm{Norm}(\cdot)$ denotes per-image normalization with clipping. For alignment of this prior with the semantic feature $\boldsymbol{F}_{freq}$, adaptive max-pooling $\mathcal{P}_{max}$ is first applied to preserve peak gradients, followed by a $\mathbf{1}\times\mathbf{1}$ projection ϕ_{proj}. The aligned structural map $\boldsymbol{M}_{edge}$ is then injected via residual gating as Eq. (11):

$$\boldsymbol{F}_{view} = \boldsymbol{F}_{freq} + \beta \boldsymbol{F}_{freq} \odot \big(2\sigma\big(\boldsymbol{M}_{edge}\big) - 1\big) \tag{11}$$

where, β is a learnable scalar and $\boldsymbol{F}_{view}$ represents the final enhanced 2D features, rich in both semantic context and high-frequency structural cues, ready for view transformation. By injecting explicit boundary priors, SEI prevents the structural washout of small distant instances, thereby mitigating the downsampling-induced scale conflicts before multi-modal integration.

3.3. Geometry-Aware View Transformation

Image-to-BEV: For camera features, we adopt a lift-splat transformation. A depth estimation head predicts a categorical depth distribution along each pixel ray. The enhanced view-space feature map is lifted into a 3D frustum by weighting features with the predicted depth probabilities. A depth head predicts a categorical depth distribution as Eq. (12):

$$\sum_{k=1}^{D} \mathbf{D}(k,i,j) = 1 \tag{12}$$

where $\mathbf{D} \in R^{D\times H\times W}$. View-space features are lifted into a frustum volume via as Eq. (13):

$$\mathbf{P}(c,k,i,j) = \mathbf{D}(k,i,j)\cdot \boldsymbol{F}_{view}(c,i,j) \tag{13}$$

Each frustum feature is projected to BEV coordinates using camera intrinsics and extrinsics. Features falling into the same voxel are aggregated as Eq. (14):

$$\mathbf{V}(c,z,y,x) = \sum_{(k,i,j)\to(x,y,z)} \mathbf{P}(c,k,i,j) \tag{14}$$

Finally, features falling into the same BEV voxel are aggregated, and the vertical dimension is collapsed to obtain a sparse BEV feature map derived from image inputs. Collapsing the vertical dimension yields a

Just Accepted

sparse BEV feature as Eq. (15):

$$F_{bev}^{img}(c,y,x) = \sum_{z} \mathbf{V}(c,z,y,x) \tag{15}$$

Importantly, because structural cues have been reinforced in the view space, the lifted features retain sharper boundaries and more coherent spatial organization compared to conventional pipelines.

LiDAR-to-BEV: For LiDAR inputs, raw point clouds are transformed into the BEV space through standard voxelization or pillarization. Each point is mapped to a BEV cell based on its planar coordinates, and point-wise features are aggregated to form a LiDAR BEV feature map as Eq. (16).

$$\mathcal{L} = \{(x_n, y_n, z_n, f_n)\}_{n=1}^{N} \tag{16}$$

Given a LiDAR point cloud points are discretized into BEV cells according to planar coordinates (x_n, y_n). For a BEV grid cell (y, x), the aggregated feature is computed as Eq. (17):

$$\boldsymbol{F}_{bev}^{lidar}(c,y,x) = \mathrm{Agg}(\{f_n^c \mid (x_n, y_n) \in \mathrm{cell}(y,x)\}) \tag{17}$$

While LiDAR features provide accurate geometric measurements, their BEV representations are inherently sparse and unevenly distributed, especially at long ranges. This sparsity motivates a careful fusion strategy rather than naive feature summation.

BEV-Level Alignment and Preliminary Fusion: After projection, both modalities are represented in the same BEV grid as Eqs. (18) and (19):

$$\boldsymbol{F}_{bev}^{img} \in R^{C_i \times H_{bev} \times W_{bev}} \tag{18}$$

$$\boldsymbol{F}_{bev}^{lidar} \in R^{C_l \times H_{bev} \times W_{bev}} \tag{19}$$

To enable cross-modal interaction, we first align channel dimensions through linear projections as Eqs. (20) and (21):

$$\tilde{\boldsymbol{F}}_{bev}^{img} = \phi_i\left(\boldsymbol{F}_{bev}^{img}\right) \tag{20}$$

$$\tilde{\boldsymbol{F}}_{bev}^{lidar} = \phi_l\left(\boldsymbol{F}_{bev}^{lidar}\right) \tag{21}$$

where ϕ_i and ϕ_l are $\mathbf{1} \times \mathbf{1}$ convolutions mapping both modalities to a common feature dimension C. A preliminary BEV fusion is then performed via element-wise summation as Eq. (22):

$$\boldsymbol{F}_{bev}^{pre} = \tilde{\boldsymbol{F}}_{bev}^{lidar} + \tilde{\boldsymbol{F}}_{bev}^{img} \tag{22}$$

This operation aligns features from different modalities with limited additional computation. However, due to uneven sparsity and modality-specific noise, naive fusion may amplify unreliable regions.

Therefore, instead of directly feeding $\boldsymbol{F}_{bev}^{pre}$ into the detection head, we introduce a dedicated refinement module in the next stage to perform sparsity-aware selective integration and structural reconstruction.

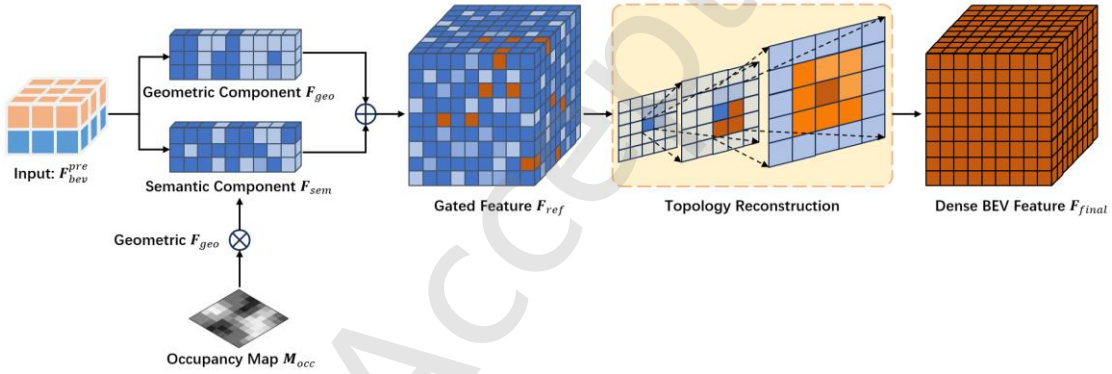


Fig. 6. Structure of the proposed SAR module. The pre-fused BEV feature $\boldsymbol{F}_{bev}^{pre}$ is decomposed into geometric and semantic components, where the occupancy map $\boldsymbol{M}_{occ}$ guides feature gating to generate the refined feature $\boldsymbol{F}_{ref}$. A topology reconstruction step is then applied to obtain the dense BEV feature $\boldsymbol{F}_{final}$, improving BEV representation quality for 3-D object detection.

3.4. Sparsity-Aware BEV Refinement

After geometric projection and preliminary BEV fusion, the resulting representation $\boldsymbol{F}_{bev}^{pre} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H_{bev}}\times\mathrm{W_{bev}}}$ remains structurally imperfect. This imperfection arises from two sources: 1) Spatial sparsity caused by discrete voxelization and uneven LiDAR sampling; and 2) Fragmented geometric responses due to depth uncertainty and modality-specific noise.

A naive convolutional decoder treats all BEV cells uniformly, implicitly assuming spatial homogeneity. However, BEV representations in multi-modal perception are inherently heterogeneous: some cells contain reliable geometric evidence, while others correspond to empty or weakly supported regions.

To address this issue, we design the Sparsity-Aware BEV Refinement module (SAR) as a unified operator that simultaneously acts as: 1) A sparsity-aware refiner, selectively integrating reliable geometric details; and 2) A BEV decoder, reconstructing a dense and structurally coherent feature map from fragmented inputs.

Formalizing BEV Sparsity: Let $\boldsymbol{F}_{bev}^{pre}(c,y,x)$ denote the fused BEV feature at location (y,x). We define an occupancy-driven reliability map as Eq. (23):

$$\boldsymbol{M}_{occ}(\mathrm{y},\mathrm{x}) = \min\left(1, \frac{\mathrm{count}(\mathrm{y},\mathrm{x})}{\tau}\right) \tag{23}$$

where $\mathrm{count}(y,x)$ denotes the number of LiDAR points or frustum samples projected into cell (y,x), and τ is a normalization constant. This map can be interpreted as an empirical estimator of geometric support. The reliability map is min–max normalized to [0, 1] and gates spatial locations, i.e., the gating operates across spatial positions rather than across channels or modalities.

Sparsity-Aware Geometric Refinement: We first decompose the preliminary BEV feature into two conceptual components as Eq. (24):

$$\boldsymbol{F}_{bev}^{pre} = \boldsymbol{F}_{geo} + \boldsymbol{F}_{sem} \tag{24}$$

where $\boldsymbol{F}_{geo}$ captures geometry-sensitive responses (primarily from LiDAR and lifted image features), and $\boldsymbol{F}_{sem}$ encodes contextual semantic information.

Rather than directly summing or concatenating features, we introduce a gated refinement operator as Eq. (25):

$$\boldsymbol{F}_{ref}(\mathrm{c},\mathrm{y},\mathrm{x}) = \boldsymbol{F}_{sem}(\mathrm{c},\mathrm{y},\mathrm{x}) + \boldsymbol{M}_{occ}(\mathrm{y},\mathrm{x})\,\boldsymbol{F}_{geo}(\mathrm{c},\mathrm{y},\mathrm{x}) \tag{25}$$

In regions where sufficient geometric evidence is available ($\boldsymbol{M}_{occ} \approx 1$), fine structural patterns are retained and further strengthened. By contrast, areas with limited occupancy support ($\boldsymbol{M}_{occ} \approx 0$) remain largely guided by the surrounding contextual features rather than unreliable geometric responses.

From an optimization perspective, this gating mechanism prevents noisy geometric artifacts from propagating into empty BEV cells,

Just Accepted

stabilizing gradient flow and reducing overfitting to sparsely populated regions.

Structural Decoding as Topology Reconstruction: Although sparsity-aware gating suppresses unreliable responses, the refined feature map $\boldsymbol{F}_{ref}$ still contains local discontinuities due to discrete projection.

We interpret BEV decoding as a topology reconstruction problem: given a fragmented spatial field, infer a continuous and geometrically consistent representation. Let $\mathcal{K}_d^{(\ell)}$ denote a dilated convolution operator at layer l with dilation factor d_ℓ. The decoding process is formulated recursively as Eqs. (26) and (27):

$$\boldsymbol{F}^{(l+1)} = \phi_l\left(\mathcal{K}_{d_\ell}^{(\ell)}\left(\boldsymbol{F}^{(l)}\right)\right) \tag{26}$$

$$\boldsymbol{F}^{(0)} = \boldsymbol{F}_{ref} \tag{27}$$

where $\phi_l(\cdot)$ denotes normalization and nonlinearity. The dilation enlarges the effective receptive field without reducing spatial resolution, enabling information propagation across non-adjacent BEV cells as Eq. (28):

$$\mathcal{R}_{eff} \propto \sum_{\ell} d_\ell \tag{28}$$

This mechanism allows the network to infer local topology by aggregating signals from structurally consistent neighboring regions, effectively bridging micro-discontinuities caused by projection. The final decoded representation is as Eq. (29):

$$\boldsymbol{F}_{final} = \boldsymbol{F}^{(L)} \tag{29}$$

which serves as a dense, high-resolution BEV feature map for downstream detection.

Unified Interpretation of SAR: The proposed SAR can be interpreted as a composite operator as Eq. (30):

$$\boldsymbol{F}_{final} = \mathcal{D}\left(\mathcal{G}\left(\boldsymbol{F}_{bev}^{pre}, \boldsymbol{M}_{occ}\right)\right) \tag{30}$$

where $\mathcal{G}(\cdot)$ is a sparsity-aware gating operator; $\mathcal{D}(\cdot)$ is a topology-preserving decoder.

Unlike conventional BEV decoders that treat sparsity implicitly, SAR explicitly models spatial reliability before structural reconstruction. Geometric cues are therefore integrated in a spatially selective manner, guided by occupancy confidence. Subsequent contextual aggregation restores local continuity and produces a dense BEV representation. This dual-role design directly addresses the structural degradation problem from both reliability filtering and topological recovery perspectives.

4 Experiment and Analysis

This section presents a comprehensive experimental evaluation of the proposed FGA-BEVFusion method for multi-sensor 3D object detection. We first introduce the dataset and implementation details, compare performance against representative methods, and conduct ablation studies to assess individual components. Qualitative results further demonstrate the model's capability in capturing boundary cues under challenging conditions.

4.1. Dataset and Evaluation Metrics

Datasets: We evaluate our method on a large-scale multi-modal dataset containing 1000 driving scenes with synchronized and calibrated camera images and LiDAR point clouds. The dataset includes 3D bounding box annotations across 10 object classes. Experiments follow the standard training and validation split (700/150 scenes), and test performance is reported using the nuScenes test server.

Evaluation Metrics: We assess detection performance using the official nuScenes evaluation protocol (Caesar et al., 2020), namely the mean Average Precision (mAP) and the nuScenes Detection Score (NDS). Unlike the intersection-over-union matching used in 2D detection, mAP on nuScenes is computed by matching predictions to ground-truth objects according to the 2D center distance on the ground plane, averaged over the thresholds of 0.5 m, 1 m, 2 m, and 4 m. NDS is a weighted combination of mAP and five true-positive quality metrics measuring translation, scale, orientation, velocity, and attribute errors. All absolute score differences are reported as percentage-point (pp) gains.

Implementation Details: Our model is implemented in PyTorch and trained on eight NVIDIA RTX 3090 GPUs with a total batch size of 48 (6 per GPU). We use AdamW optimizer with a cosine learning rate scheduling with an initial lr = 2e-4 for 24 epochs on nuScenes. Data augmentation includes random flipping, scaling, rotation, and color jittering applied to both images and point clouds.

The BEV grid spans [−50 m, 50 m] at a resolution of 0.1 m. The FGA module is positioned after the backbone stage, where MCA processes three frequency bands, and SEI applies adaptive edge detection with τ = 0.3. The SAR employs depth-wise convolutions with a 3 × 3 kernel size and a dilation rate of 2.

The principal FGA hyperparameters were configured for efficiency with C_e=64, C_{min}=32, H_{min}=W_{min}=8, $\gamma=\frac{1}{2}$, and r=16, while multi-scale and multi-frequency feature fusion was governed by S = 3 and K = 16, respectively.

4.2. Results and Analysis

4.2.1. Results on nuScenes Validation Set

Table 1. Performance comparison on the nuScenes validation set. All results are reported in percentage (%) † denotes our reproduction under the unified training configuration, against which all reported gains are computed.

Method	Modality	mAP (val)	NDS (val)
BEVDet (Huang et al., 2021)	C	-	-
BEVFormer (Li et al., 2022b)	C	41.7	47.0
BEVStereo (Li et al., 2023a)	C	51.3	59.4
PointPillars (Lang et al., 2019)	L	52.5	60.8
CVCNet (Chen et al., 2020)	L	52.3	61.4
CenterPoint (Yin et al., 2021)	L	59.5	65.7
TransFusion (Bai et al., 2022)	C+L	67.6	71.3
BEVFusion (Liang et al., 2022)	C+L	67.9	70.4
CL-FusionBEV (Shi et al., 2024)	C+L	70.3	71.3
BEVFusion (Liu et al., 2023)†	C+L	68.5	71.7
Ours	C+L	70.8	72.3

Table 1 compares FGA-BEVFusion with representative LiDAR-only, camera-only, and fusion-based methods on the nuScenes validation set. FGA-BEVFusion achieves 70.8% mAP and 72.3% NDS, outperforming the reproduced BEVFusion baseline (Liu et al., 2023) by 2.3 percentage points in mAP and 0.6 percentage points in NDS. The results show that the proposed frequency-guided design can consistently improve

Just Accepted

detection performance in BEV-based multi-sensor fusion.

4.2.2. Results on nuScenes Test Set

Table 2. Performance comparison on the nuScenes test set. All results are reported in percentage (%)† denotes our reproduction under the unified training configuration, against which all reported gains are computed.

Method	Modality	mAP (test)	NDS (test)
BEVDet (Huang et al., 2021)	C	42.5	48.1
BEVFormer (Li et al., 2022b)	C	44.5	53.5
BEVStereo (Li et al., 2023a)	C	52.5	61.0
PointPillars (Lang et al., 2019)	L	-	-
CVCNet (Chen et al., 2020)	L	55.3	64.7
CenterPoint (Yin et al., 2021)	L	60.3	67.3
TransFusion (Bai et al., 2022)	C+L	68.9	71.7
BEVFusion (Liang et al., 2022)	C+L	69.7	70.2
CL-FusionBEV (Shi et al., 2024)	C+L	71.3	72.9
BEVFusion (Liu et al., 2023)†	C+L	70.2	72.9
Ours	C+L	71.7	73.5

Table 2 reports the official nuScenes test set results. FGA-BEVFusion achieves 71.7% mAP and 73.5% NDS, improving upon the reproduced BEVFusion baseline (Liu et al., 2023) by 1.5 and 0.6 percentage points, respectively. These results confirm that the proposed method generalizes well to unseen driving scenes.

Table 3. Per-category AP comparison on the nuScenes validation set. Cons. Veh.: construction vehicle; Ped.: pedestrian; Motor.: motorcycle; T.C.: traffic cone. All results are reported in percentage (%). The Improvement row denotes the absolute percentage-point gain of FGA-BEVFusion over the reproduced BEVFusion baseline (Liu et al., 2023).

Method	Car	Truck	Bus	Trailer	Cons. Veh.	Ped.	Motor.	Bicycle	T.C.	Barrier	mAP
BEVFusion	86.5	60.4	72.2	45.3	35.4	88.0	75.2	61.9	78.0	80.0	68.5
FGA-BEVFusion	88.2	62.0	74.3	47.0	37.2	90.3	78.7	64.7	81.2	82.0	70.8
Improvement	+1.7	+1.6	+2.1	+1.7	+1.8	+2.3	+3.5	+2.8	+3.2	+2.0	+2.3

Table 4. Scenario-wise performance comparison on the nuScenes validation set. ΔmAP denotes the absolute percentage-point gain in mAP over the baseline under the same scenario.

Scenario	BEVFusion mAP	BEVFusion NDS	FGA-BEVFusion mAP	FGA-BEVFusion NDS	ΔmAP
Day-Sunny	68.2	72.6	72.1	73.7	+3.9
Day-Rainy	64.9	68.5	70.8	70.1	+5.9
Night	52.8	64.2	61.4	66.0	+8.6
Overall	65.5	71.4	70.2	72.8	+4.7

4.2.3. Per-Category Performance Analysis

Table 3 presents the AP results for all 10 nuScenes object categories. FGA-BEVFusion improves the overall mAP from 68.5% to 70.8%, with gains across most categories. The largest improvements appear on small or structurally complex objects such as pedestrians, motorcycles, bicycles, and traffic cones.

These results indicate that the proposed method is particularly effective for categories that rely on fine-grained boundary and edge information. These gains may be associated with the preservation of structural cues before BEV projection; we note, however, that the present experiments validate the effectiveness of the overall design rather than isolating a specific frequency-domain mechanism.

Compared with the closest related approaches, IS-Fusion (Yin et al., 2024), GAFusion (Li et al., 2024b), and CL-FusionBEV share with our method the general paradigm of BEV-space feature alignment and attention-based cross-modal fusion. The key differences are that FGA-BEVFusion performs explicit frequency-domain decomposition in the view space (MCA and SEI) and occupancy-guided refinement in the BEV space (SAR), rather than relying solely on instance-level or guidance-based feature interaction. These differences are validated by our experiments: the ablation results in Tables 5 and 6 confirm that both frequency-domain modeling and boundary injection contribute to the gains, and the scenario-wise results in Table 4 indicate that the benefits are most pronounced under degraded visual conditions.

Just Accepted

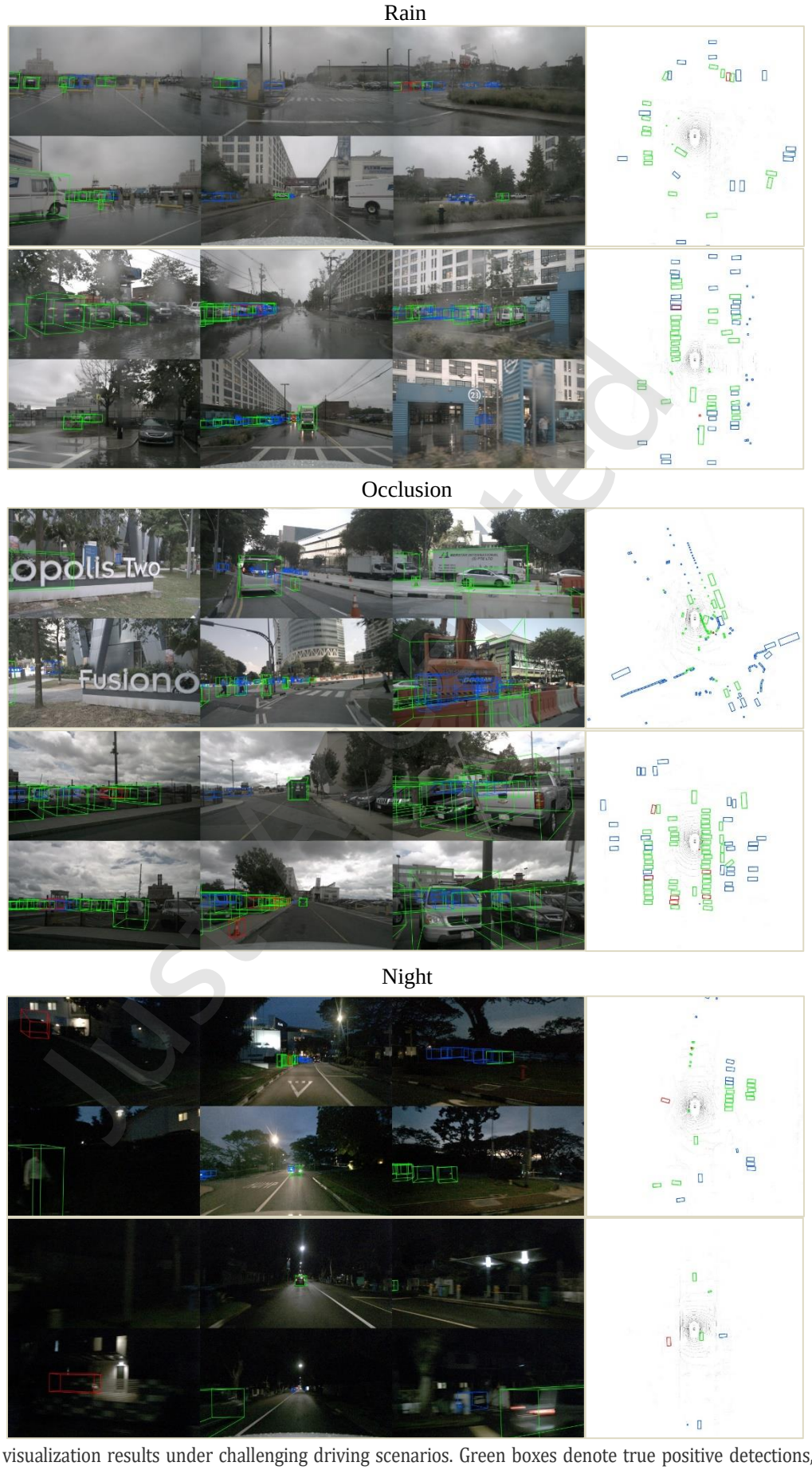


Fig. 7. Qualitative visualization results under challenging driving scenarios. Green boxes denote true positive detections, blue boxes denote missed detections, and red boxes denote false positive detections. The visualizations include representative samples from rainy, occluded, and night-time scenes. The subfigures (a), (b), and (c) correspond to the rain, occlusion, and night scenarios, respectively.

Just Accepted

4.2.4. Performance under Different Scenarios

Table 4 evaluates performance under different driving scenarios, including sunny, rainy, night, and complex urban scenes. FGA-BEVFusion consistently outperforms BEVFusion in all scenarios, with especially clear gains in rainy and nighttime conditions.

These results suggest that the proposed frequency-guided modules may improve robustness under challenging environments. In particular, MCA and SEI help recover useful image structure when visual quality is degraded, while SAR further improves fused BEV representations.

4.3. Ablation Studies

4.3.1. Effect of Each Component

Table 5. Ablation study of the MCA, SEI, and SAR components on the nuScenes validation set. All results are reported in percentage (%)

	MCA	SEI	SAR	mAP	NDS
Baseline	×	×	×	65.7	66.9
I	√	×	×	67.5	68.3
II	×	√	×	67.9	68.1
III	√	√	×	69.0	70.1
IV	√	√	√	70.8	72.3

We perform extensive ablation studies to assess the contribution of individual architectural components by selectively removing or substituting MCA, SEI, and SAR. As reported in **Table 5**, the baseline model attains 65.7% mAP and 66.9% NDS. Incorporating the MCA module alone improves performance to 67.5% mAP and 68.3% NDS, highlighting the effectiveness of frequency-domain feature modeling. The addition of SEI yields larger gains (+2.2% mAP and +1.2% NDS), particularly for boundary-sensitive categories, such as pedestrians and traffic signs. The full model, integrating all components, achieves the best performance at 70.8% mAP and 72.3% NDS, indicating that frequency modeling, boundary enhancement, and hierarchical refinement contribute to performance in different ways. Eliminating the SAR results in the largest performance drop (−1.8% mAP and −2.2% NDS), suggesting that this component plays a non-negligible role in preserving structural consistency. The largest gains arise from jointly applying MCA and SEI, which improve performance by +3.3% mAP and +3.2% NDS over baseline, indicating that frequency-domain and spatial-domain processing provide complementary cues.

4.3.2. Frequency Band Selection in MCA

Table 6. Effect of different frequency band selections in MCA

Method	Selected Bands	mAP	NDS
w/o MCA	-	65.0	71.8
Low-frequency only	L	62.2	65.0
Mid-frequency only	M	63.5	65.9
High-frequency only	H	62.3	64.2
Low + Mid	L + M	66.8	69.4
Mid + High	M + H	67.0	69.6
Low + High	L + H	68.2	70.3
All bands equally weighted	L + M + H	69.6	71.2
Proposed adaptive selection	Adaptive L + M + H	70.8	72.3

Table 6 compares different frequency band selections in MCA on the nuScenes validation set. Results show that using a single band provides limited improvement, while combining multiple bands yields better performance. Among the fixed settings, Mid + High achieves the strongest result, suggesting that structural cues are mainly concentrated in mid- and high-frequency responses.

The proposed adaptive band selection performs best, reaching 70.8% mAP and 72.3% NDS. This indicates that adaptively reweighting different frequency bands is more effective than using a manually fixed combination. By jointly preserving semantic context and boundary details, MCA can produce more discriminative features for BEV fusion.

4.4. Qualitative Results

Fig. 7 presents qualitative our method on challenging nuScenes scenarios, where our FGA-BEVFusion method consistently preserves the geometric boundaries of distant and small objects that are often fragmented or missed by other methods. Under adverse conditions, such as rain and partial occlusion scenarios, our approach maintains coherent object representations. In nighttime scenes, although low illumination and strong contrast degradation lead to several missed detections and occasional false positives, the model still identifies a considerable number of objects in both camera views and the BEV space. These results suggest that frequency-aware refinement can help preserve object boundaries in BEV feature representations, especially in crowded scenes.

5 Conclusions

High-frequency structural cues, such as edges and textures in camera and LiDAR features, are often degraded during view transformation and fusion, thereby impairing 3D detection performance. To address this issue, we have presented a novel FGA framework for multi-modal 3D object detection that explicitly mitigates structural loss in BEV feature learning. The framework mainly relies on two components: MCA and SEI. MCA analyzes the spectral distribution of feature maps and adjusts channel responses, giving relatively higher weights to high-frequency components, so that these cues are less affected by subsequent view transformation and aggregation. SEI incorporates multi-scale context to improve boundary responses in spatial features. Results on nuScenes demonstrate consistent performance gains, particularly in preserving detection accuracy for small and distant objects under structural degradation. These findings confirm that explicit modeling of frequency-domain characteristics and edge-aware spatial patterns is an effective strategy for robust 3D perception. The proposed FGA method offers a generalizable solution for preserving structural fidelity in multi-modal fusion systems. We also note that the present study focuses on vehicle-mounted multi-modal 3D object detection evaluated on nuScenes; the applicability of the proposed framework to roadside perception and cooperative driving scenarios remains a potential application context to be validated in future work.

Future work will proceed along three specific directions. First, we will integrate FGA-BEVFusion with vehicle-to-infrastructure (V2I) cooperative perception, in which roadside units provide complementary viewpoints and the frequency-guided design may help compensate for the structural degradation caused by infrastructure-side sensing and communication constraints. Second, we will investigate adaptive reward mechanisms, e.g., reinforcement-learning-based policies that dynamically reweight modalities and frequency bands according to environmental conditions such as illumination and weather. Third, we will study real-time deployment challenges, including lightweight frequency-decomposition operators, latency optimization on edge computing platforms, and model quantization, and we will extend the evaluation to BEV segmentation and multi-object tracking. Overall, the results indicate that modeling high-frequency information loss during view transformation can be

Just Accepted

beneficial for LiDAR–camera BEV fusion.

Author contributions

Zechang Ye: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Haiyang Yu:** Writing – review & editing, Supervision, Funding acquisition. **Shuai Liu:** Writing – review & editing, Formal Analysis, Methodology. **Han Jiang:** Formal Analysis, Software, Data curation. **Yilong Ren:** Formal Analysis, Methodology.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020008 and the Fundamental Research Funds for the Central Universities.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of generative Al and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT 5.6 for checking grammar and spelling errors in the paper. After utilizing this tool/service, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

Replication and data sharing

The source code and reproduction package is available on the ETS-Data at https://doi.org/10.26599/ETSD.2026.9190110.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Just Accepted

References

Bai, X.Y., Hu, Z.Y., Zhu, X.G., Huang, Q.Q., Chen, Y.L., Fu, H.B., et al., 2022. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1090–1099.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., et al., 2020. nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11621–11631.

Chen, Q., Sun, L., Cheung, E., Yuille, A., 2020. Every view counts: Cross-view consistency in 3D object detection with hybrid-cylindrical-spherical voxelization. In: Advances in Neural Information Processing Systems, 33, 21297–21308.

Deng, J.J., Shi, S.S., Li, P.W., Zhou, W.G., Zhang, Y.Y., Li, H.Q., 2021. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, 35, 1201–1209.

Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D., 2021. BEVDet: High-performing multi-camera 3D object detection in bird-eye-view. arXiv:2112.11790.

Jiao, Y., Jie, Z., Chen, S., Cheng, L., Chen, J., Ma, L., et al., 2024. Instance-aware multi-camera 3D object detection with structural priors mining and self-boosting learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2598–2606.

Jin, X., Su, H., Liu, K., Ma, C., Wu, W., Hui, F., et al., 2025. UniMamba: Unified spatial-channel representation learning with group-efficient Mamba for LiDAR-based 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1407–1417.

Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O., 2019. PointPillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12697–12705.

Lee, Y., Jeong, H.-M., Jeon, Y., Kim, S., 2025. EVT: Efficient view transformation for multi-modal 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 26632–26642.

Li, H., Sima, C., Dai, J., Wang, W., Lu, L., Wang, H., et al., 2024a. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(4), 2151–2170.

Li, X., Fan, B., Tian, J., Fan, H., 2024b. GAFusion: Adaptive fusing LiDAR and camera with multiple guidance for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 21209–21218.

Li, Y., Bao, H., Ge, C., Liu, J., Sun, J., 2023a. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In: Proceedings of the AAAI Conference on Artificial Intelligence, 1486–1494.

Li, Y.H., Ge, Z., Yu, G.Y., Yang, J.R., Wang, Z.R., Shi, Y.K., et al., 2023b. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 1477-1485.

Li, Y.W., Chen, Y.L., Qi, X.J., Li, Z.M., Sun, J., Jia, J.Y., 2022a. UVTR: Unifying voxel-based representation with transformer for 3D object detection. In: Proceedings of Advances in Neural Information Processing Systems 35, 18442-18455

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., et al., 2022b. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Proceedings of the European Conference on Computer Vision, 13669, 1-18

Liang, T.T., Xie, H.W., Yu, K.C., Xia, Z.Y., Lin, Z.W., Wang, Y.T., et al., 2022. BEVFusion: A simple and robust LiDAR-camera fusion framework. In: Proceedings of Advances in Neural Information Processing Systems, 10421-10434.

Liu, B., Wu, Y., 2025. BFT3D: A robust BEV feature transformation module for multi-sensor 3D object detection. IEEE Sensors Journal 25, 30175–30185.

Liu, S., Cui, M., Li, B., Liang, Q., Hong, T., Huang, K., et al., 2025. FSHNet: Fully sparse hybrid network for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8900–8909.

Liu, X., Zheng, C., Qian, M., Xue, N., Chen, C., Zhang, Z., et al., 2024. Multi-view attentive contextualization for multi-view 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16688–16698.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., et al., 2023. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the IEEE International Conference on Robotics and Automation, 2774-2781.

Noh, J., Lee, S., Ham, B., 2021. HVPR: Hybrid voxel-point representation for single-stage 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14600–14609.

Park, J., Xu, C.f., Yang, S.j., Keutzer, K., Kitani, K.M., Tomizuka, M., et al., 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection. In: Proc. Int. Conf. Learn. Represent. ICLR.

Philion, J., Fidler, S., 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of the European Conference on Computer Vision, 12359.194-210.

Shi, G.S., Li, R.F., Ma, C.X., 2022. PillarNet: Real-time and high-performance pillar-based 3D object detection. In: Proceedings of the European Conference on Computer Vision, 13670. 35-52.

Shi, P., Liu, Z., Dong, X., Yang, A., 2024. CL-fusionBEV: 3D

Just Accepted

object detection method with camera-LiDAR fusion in Bird's Eye View. Complex & Intelligent Systems, 10(8): 12457-12473.

Shi, S., Wang, Z., Wang, X., Li, H., 2023. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. Int. J. Comput. Vis. 131(2). 531-551.

Sindagi, V.A., Zhou, Y., Tuzel, O., 2019. MVX-Net: Multimodal voxelnet for 3D object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation, 7276-7282.

Wang, Z., Jia, K., 2020. InfoFocus: 3D object detection for autonomous driving with dynamic information modeling. In: Proceedings of the European Conference on Computer Vision, 12335, 405-420.

Wu, Y., et al., 2023. HeightFormer: Explicit height modeling without extra data for camera-only 3D object detection in bird's-eye view. arXiv:2307.13510.

Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., Zhang, L., 2021. FusionPainting: Multimodal fusion with adaptive attention for 3D object detection. In: Proceedings of the IEEE International Conference on Intelligent Transportation Systems, 3047-3054

Yan, Y., Mao, Y., Li, B., 2018. SECOND: Sparsely embedded convolutional detection. Sensors 18, 3337.

Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., et al., 2023. BEVHeight++: Toward robust visual-centric 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 47(6). 5094-5111.

Ye, T., Liu, R., Min, C., Li, Y., Li, X., 2025. DFNet: A dual LiDAR-camera fusion 3-D object detection network under feature degradation condition. IEEE Sensors Journal. doi:10.1109/JSEN.2025.3551149.

Yin, J., Shen, J., Chen, R., Li, W., Yang, R., Frossard, P., et al., 2024. IS-Fusion: Instance-scene collaborative fusion for multimodal 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14905–14915.

Yin, T., Zhou, X., Krahenbuhl, P., 2021. Center-based 3D object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14783–14792.

Yoo, J.H., Kim, Y., Kim, J., Choi, J.W., 2020. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection. In: Proceedings of the European Conference on Computer Vision, 12372, 720-736.

Yu, D.R., Qu, A.Q., You, J., Wang, D.Y., Li, R.S., Jia, S.C., et al., 2026 Which2comm: An Efficient Collaborative Perception Framework with Connected and Automated Vehicles. Communications in Transportation Research. https://doi.org/10.26599/COMMTR.2026.9640048

Zhang, G., Chen, J., Gao, G., Li, J., Liu, S., Hu, X., 2024. SAFDNet: A simple and effective network for fully sparse 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14477–14486.

Zhang, J., Zhang, Y., Liu, Q., Wang, Y., 2023. SA-BEV: Generating semantic-aware bird's-eye-view feature for multi-view 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3325–3334.

Zhang, J., Zhang, Y., Qi, Y., Fu, Z., Liu, Q., Wang, Y., 2025. GeoBEV: Learning geometric BEV representation for multi-view 3D object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 9960–9968.

Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., et al., 2022. BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv:2205.09743.

Zhou, B., Krähenbühl, P., 2022. Cross-view transformers for real-time map-view semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13750-13759.

Zhou, Y., Tuzel, O., 2018. VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4490–4499.

Just Accepted

Author biography



Zechang Ye received the B.S. degree from School of Automobile, Chang'an University, Xi'an, China in 2019. He is currently pursuing the Ph.D. degree with the School of Transportation Science and Engineering, Beihang University. His research interests include intelligent transportation systems and multi-modal information fusion.



Haiyang Yu received his Ph.D. degree in traffic environment and safety technology from Jilin University, China, in 2012. He is currently a Professor with the School of Transportation Science and Engineering, Beihang University, China. His research interests include traffic big data, traffic control and intelligent vehicle infrastructure cooperative systems.



Shuai Liu received the Ph.D. degree in traffic and transportation engineering from Beihang University, Beijing, China. He is currently a Postdoctoral Researcher with the Hangzhou Innovation Institute of Beihang University, Hangzhou, China. His research interests include deep learning, spatiotemporal Big Data modeling, and intelligent transportation systems.



Han Jiang received the B.S. and Ph.D. degrees in traffic information engineering and control from Beihang University in 2017 and 2025, respectively. He is currently a postdoctoral research fellow in Beihang University. His research interests include agentic AI, pervasive and crowd computing in intelligent transportation systems



Yilong Ren received B.S. and Ph.D. degrees from Beihang University in 2010 and 2017, respectively. He is currently a Professor with the School of Transportation Science and Engineering, Beihang University, China. His research interests include vehicle to everything, traffic control and vehicular crowd sensing

Just Accepted