FocusBEV: Jointly Focusing on 2D Semantic Guidance and Holistic LiDAR Representation

1st Sixian Chan
College of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
sxchan@zjut.edu.cn

2nd Yaohui Li
College of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
211123120020@zjut.edu.cn

3rd Jian Tao
College of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
taojian@zjut.edu.cn

4th Xinggang Fan*
Zhijiang College
Zhejiang University of Technology
Shaoxing, China
xgfan@zjut.edu.cn

***Abstract*—Accurate 3D object detection stands as a cornerstone of autonomous driving systems. Among various methodologies, the Bird's-Eye-View (BEV) paradigm, which fuses LiDAR point clouds and surround-view images, has emerged as the predominant approach. However, existing frameworks struggle to fully exploit the semantic potential of images and construct contextually integral LiDAR representations. To address these challenges, we propose FocusBEV, a "dual-focus" framework. First, we introduce the Semantic-Focus module, a 2D semantic guidance mechanism. Leveraging a lightweight 2D detector as a "semantic prior", this module proactively enhances foregrounds on feature maps prior to view transformation, thereby improving semantic fidelity. Second, we design the Long-range Interaction Aggregator (LIA) that uses shifted-window self-attention to reconstruct coherent global dependencies from sparse LiDAR features, overcoming the locality of CNNs. Finally, extensive experiments conducted on nuScenes demonstrate that FocusBEV achieves 72.3% mAP and 74.6% NDS, explicitly validating the effectiveness of our FocusBEV.**

***Index Terms*—3D Object Detection, Multi-modal Fusion, Semantic Guidance, Bird's-Eye-View (BEV)**

I. Introduction

3D object detection serves as a pivotal component in autonomous driving perception systems, necessitating the precise identification and localization of objects within a complex environment [1]–[4]. To overcome the inherent limitations of individual sensors, the combination of LiDAR's geometric precision with the rich semantic information from cameras has emerged as a critical pathway to robust perception [4]–[8]. Currently, the Bird's-Eye-View (BEV) paradigm dominates this domain, primarily following two trajectories: (1) Transformer-based Query Fusion [6], [9]–[11]; and (2) BEV Feature Map Alignment, exemplified by BEVFusion [12]. While feature alignment offers architectural efficiency, we identify a systemic bottleneck restricting its potential: the *dual-defocus* phenomenon. This issue stems from intrinsic defects in independent encoding processes, which collectively impose an upper bound on fusion performance.

Specifically, the image branch suffers from *semantic defocus*. Explicit view transformations like LSS [13] are semantically blind, treating foreground objects and background clutter indiscriminately. This creates a "frustum smearing" effect, where semantically irrelevant background noise is projected into 3D space. This noise influx severely dilutes the Signal-to-Noise Ratio, making it difficult for downstream fusion modules to distinguish authentic objects from ambiguous projections [14].

Simultaneously, the LiDAR branch faces *contextual defocus* due to the paradigm mismatch between point cloud sparsity and the locality of traditional CNN backbones [15], [16]. LiDAR data are highly non-uniform; standard convolutions, constrained by local receptive fields, struggle to bridge the vast empty regions in sparse data. Consequently, complete objects (*e.g.*, large trucks) often manifest as spatially disjointed components separated by voids, lacking macro-level awareness of the holistic structure or boundary information [17]–[19].

To resolve these bottlenecks, we propose FocusBEV, a unified framework that incorporates two proactive mechanisms. First, to cure semantic defocus, we introduce the Semantic-Focus Module. Leveraging a lightweight 2D prior [20], it acts as a semantic filter to actively enhance foregrounds and suppress background noise before projection, ensuring high semantic fidelity. Second, to resolve contextual defocus, we design the Long-range Interaction Aggregator (LIA). LIA replaces local convolutions with Shifted-Window Self-Attention [21]. Designed as a global connectivity agent, LIA actively stitches sparse feature fragments to reconstruct a

*Corresponding author.

This work is partially supported by the the National Key Research and Development Program Project of China (Grant No. 2024YFC3306902), National Natural Science Foundation of China under (Grant No. U24A20242), the Fundamental Research Funds for the Provincial Universities of Zhejiang(Grant No.RF-A2024013), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023, LDT23F02024F02, LZ26F020008), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No.2025C01152).

contextually integral LiDAR representation.

The main contributions are summarized as follows:

- We present the FocusBEV framework to combat the dual-defocus phenomenon, integrating Semantic-Focus and LIA modules to achieve excellent performance on the nuScenes dataset.
- We propose the Semantic-Focus Module to mitigate semantic ambiguity by enhancing foreground features prior to view transformation.
- We introduce the LIA Module to resolve contextual defocus, utilizing shifted-window attention to reconstruct global context from fragmented LiDAR features.

II. Related Work

A. Camera-only 3D detection methods

Recent camera-based methods primarily leverage BEV representations to unify multi-view features. These approaches follow two evolutionary paths. LSS [13] and BEVDet [22] represent explicit depth-based methods, predicting depth distributions to lift 2D features into 3D frustums and then splatting [13] onto the BEV plane. BEVFormer [10] represents Transformer-based methods, bypassing ill-posed depth estimation by utilizing learnable queries to aggregate features via attention mechanisms. Other works, such as BEVDepth [14], further improve the quality of depth estimation to improve the projection of features.

However, camera-only solutions suffer from inherent limitations. Being passive sensors, they lack precise depth measurement capabilities, leading to geometric ambiguity. In addition, their perception reliability degrades sharply under adverse conditions—such as low light or rainy weather—which poses significant risks for safety-critical autonomous systems.

B. LiDAR-only 3D detection methods

LiDAR detectors offer precise spatial geometry and have evolved through various representation paradigms. To address efficiency, VoxelNet [3] and SECOND [15] discretize points into regular 3D grids for convolutional processing. SECOND specifically introduces sparse convolutions to effectively handle the emptiness of voxels, significantly boosting inference speed. Further, pursuing real-time performance, PointPillars [2] collapses 3D voxels into vertical columns to form pseudo-images, enabling the utilization of highly optimized 2D convolutional backbones. Recent approaches like CenterPoint [18] further refine the detection head for anchor-free prediction.

Despite their geometric precision, LiDAR-only methods encounter distinct bottlenecks. The intrinsic sparsity of point clouds results in insufficient structure for distant or small objects, and the lack of texture information, often referred to as semantic paucity, makes it challenging to distinguish semantically similar categories, distinguishing a pedestrian from a signpost.

C. Multi-modality 3D detection methods

Fusion methods seek to complement LiDAR geometry with camera semantics. Early approaches, such as PointPainting [5], pioneer data-level fusion by projecting segmentation scores onto raw LiDAR points. However, these early attempts often suffer from strict calibration dependence or sparse interactions limited to object proposals.

Consequently, BEV-level fusion has emerged as the definitive paradigm for high-performance perception. Unlike its predecessors, BEV fusion unifies heterogeneous modalities into a shared geometrically coherent coordinate system. This unification enables dense, pixel-wise interaction, preserving both the geometric structure of LiDAR and the semantic density of images. BEVFusion [12] establishes an efficient align-then-fuse strategy where camera features are transformed via LSS [13] and LiDAR features are encoded via VoxelNet [3] or PointPillars [2]. They are aligned on the same BEV grid for deep integration. Subsequent work has further explored advanced alignment mechanisms using deformable cross-attention [10], [23] to handle feature misalignment. Numerous recent studies [6], [24]–[28] have continued to refine this paradigm to improve robustness and precision. By providing a holistic representation of the scene, BEV fusion significantly outperforms previous paradigms in both accuracy and robustness.

However, these frameworks predominantly prioritize fusion architectures, often overlooking intrinsic feature defects. Specifically, semantic defocus arises from an inaccurate view transformation, and contextual defocus is caused by local CNN operations. To address this, our FocusBEV introduces a dual-focus mechanism to proactively enhance the feature fidelity at the source.

III. Methdology

In this section, we elaborate on the FocusBEV framework, as illustrated in Fig.1. Our architecture is established upon the parallel image and LiDAR encoder paradigm. To address the systemic dual-defocus bottleneck inherent in traditional BEV fusion pipelines [12], we incorporate two distinctive components: the Semantic-Focus Module, embedded within the image branch prior to view transformation, to resolve semantic defocus; and the LIA Module, which replaces the standard LiDAR backbone to cure contextual defocus.

A. Camera and LiDAR Encoding Paradigms

Camera Encoding: The image branch follows the LSS paradigm [13] to transform multi-view 2D image features into a BEV representation. First, a 2D image backbone such as ResNet equipped with a Feature Pyramid Network [20], extracts multi-scale 2D features F_{img}. Simultaneously, a lightweight depth prediction head predicts a discrete categorical depth distribution D for each pixel (u, v) on F_{img}. Here, (u, v) denotes the 2D pixel coordinates on the feature map, and d represents a specific depth interval from a pre-defined discrete set $\mathbb{D}$, which covers the range of $[4.0\text{m}, 45.0\text{m}]$. Thus,

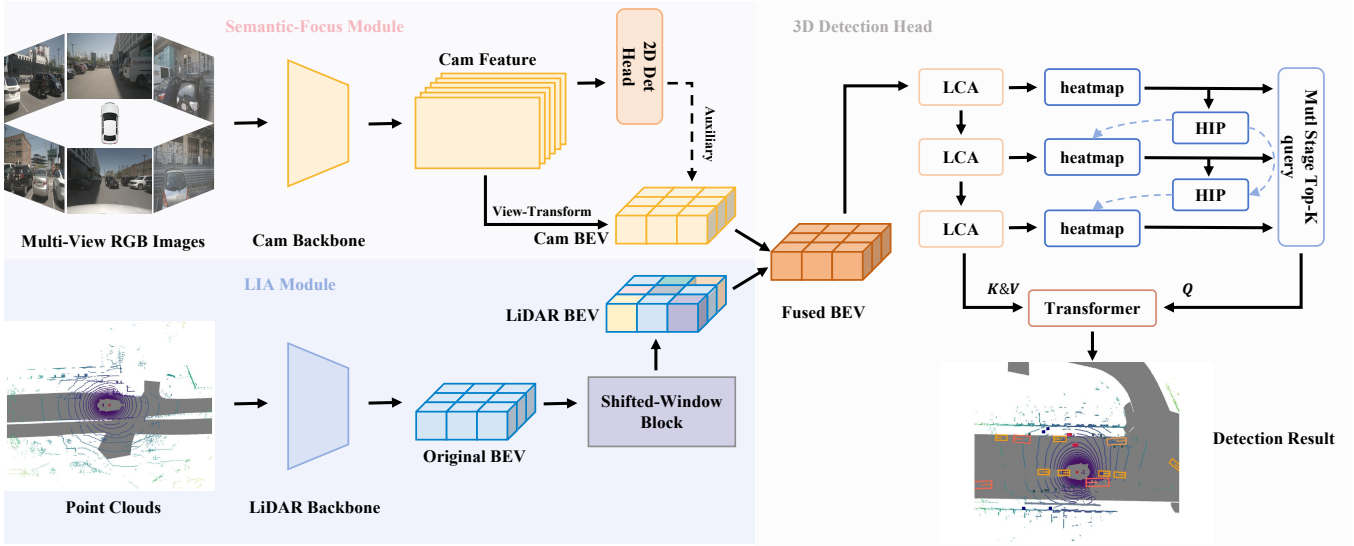


Fig. 1. The overall architecture of FocusBEV. The framework comprises three core components: the Semantic-Focus Module integrated into the image branch, the LIA module embedded in the LiDAR branch, and a Transformer-based 3D detection head for final prediction.

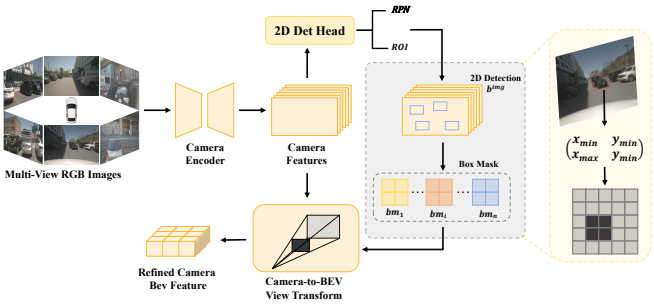


Fig. 2. Architecture of the Semantic-Focus module. It utilizes a 2D bounding box mask to reweight and refine camera features before the view transformation.

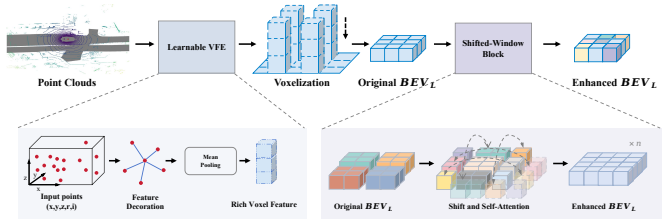


Fig. 3. Architecture of the Long-Range Interaction Aggregator (LIA). It extracts local features using a Learnable VFE and then aggregates long-range context via Shifted-Window self-attention blocks to generate enhanced LiDAR BEV features.

$D_n(u, v, d)$ indicates the probability of the feature (u, v) being located at depth d. This process can be decomposed as follows:

We compute the weighted features F_{weighted} for each frustum point (u, v, d):

$$F_{\text{weighted}}(u, v, d) = D_n(u, v, d) \cdot F_{\text{img},n}(u, v) \tag{1}$$

Next, we accumulate all weighted features F_{weighted} that map to the same BEV grid cell (i, j) to calculate the contribution C_n from camera n:

$$C_n(i, j) = \sum_{(u,v,d)\to(i,j)} F_{\text{weighted}}(u, v, d) \tag{2}$$

Finally, we generate the final BEV image feature B_I by aggregating contributions from all N_{cam} cameras:

$$B_I(i, j) = \sum_{n=1}^{N_{\text{cam}}} C_n(i, j) \tag{3}$$

LiDAR Encoding: The LiDAR branch is responsible for transforming raw sparse 3D point clouds into a BEV representation. First, the point cloud is partitioned into regular 3D voxels or pillars. A standard Voxel Feature Encoder (VFE), denoted as Φ_{vfe} and exemplified by HardSimpleVFE [3], [15], aggregates points p_k within each non-empty voxel v_i into a feature vector f_{v_i}. This is typically achieved via simple brute-force pooling operations, such as Max Pooling or Mean Pooling:

$$f_{v_i} = \Phi_{\text{vfe}}(\{p_k | p_k \in v_i\}) \tag{4}$$

These sparse voxel features f_{v_i} are subsequently processed by a 3D Sparse Convolutional Network Φ_{3d} [15]. This network efficiently performs feature extraction and downsampling in 3D space. Finally, by collapsing the Z-axis, commonly referred to as a scatter operation [2], the 3D features are projected onto the ground plane to obtain the LiDAR BEV feature map B_L.

B. Semantic-Focus Module

To mitigate the semantic defocus inherent in the standard LSS pipeline [13], where the generated BEV image feature B_I is susceptible to background noise contamination, we

introduce the Semantic-Focus Module. Embedded immediately prior to the view transformation, this module proactively refines the 2D features F_{img} to ensure that only high-value foreground features are projected into the BEV space.

As shown in Fig. 2, we append a lightweight detection head $\Phi_{\text{2D\_head}}$ on top of the FPN features F_{img} to generate a set of 2D spatial priors $\mathcal{B}_{\text{2D}}$:

$$\mathcal{B}_{\text{2D}} = \Phi_{\text{2D\_head}}(F_{\text{img}}) \tag{5}$$

where $\mathcal{B}_{\text{2D}} = \{b_1, b_2, \ldots, b_N\}$ represents a collection of N predicted 2D bounding boxes. Each box b_i contains location information, specifically the coordinates $\{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$, which serves as the spatial guidance for semantic focusing.

The generated 2D spatial prior $\mathcal{B}_{\text{2D}}$ is utilized during both training and inference to construct a spatial mask $\mathcal{M}_{\text{2D}}$. Rather than employing a hard binary mask, which might completely eliminate contextual information, we design a soft weighting mask defined as follows:

$$\mathcal{M}_{\text{2D}}(u,v) = \begin{cases} \alpha & \text{if } (u,v) \in \mathcal{B}_{\text{2D}} \\ \beta & \text{if } (u,v) \notin \mathcal{B}_{\text{2D}} \end{cases} \tag{6}$$

where $\alpha > 1$ serves as an enhancement factor to amplify the feature response of foreground regions, and $\beta \in (0,1)$ acts as a suppression factor to dampen the background noise. This formulation ensures that regions within the detected boxes are emphasized, while background regions are suppressed but not discarded.

Before the LSS projection described in Eq. (1) occurs, we apply this mask $\mathcal{M}_{\text{2D}}$ to perform element-wise re-weighting on the original 2D features F_{img}. For every pixel (u,v), the enhanced feature $F'_{\text{img}}(u,v)$ is computed as:

$$F'_{\text{img}}(u,v) = F_{\text{img}}(u,v) \cdot \mathcal{M}_{\text{2D}}(u,v) \tag{7}$$

The semantically focused feature F'_{img} subsequently replaces F_{img} as the input to the LSS module. This Focus-then-Lift strategy significantly improves the semantic clarity of B_I with negligible additional computational cost.

C. Long-Range Interaction Aggregator Module

To cure the contextual defocus caused by the locality of CNNs (Sec. III-A), we propose the LIA Module. By replacing standard backbones with a global attention mechanism, LIA re-aggregates sparse features into a contextually integral macro-representation.

LIA adopts Window-based Multi-head Self-Attention (W-MSA) [21] to capture global context, avoiding the quadratic cost of standard attention. To bridge information isolation between non-overlapping windows, we further introduce Shifted Window (SW-MSA) attention. By cyclically shifting window partitions, LIA establishes cross-window connections, expanding the receptive field to a quasi-global range. A learnable Relative Position Bias R is integrated to encode geometric relationships:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + R\right)V \tag{8}$$

As shown in Fig. 3, LIA stacks alternating W-MSA and SW-MSA blocks. For consecutive layers l and $l+1$, the process is defined as:

$$\hat{B}_L^l = \Phi_{W\text{-}Att}(\Phi_{LN}(B_L^{l-1})) + B_L^{l-1} \tag{9}$$

$$B_L^l = \Phi_{FFN}(\Phi_{LN}(\hat{B}_L^l)) + \hat{B}_L^l \tag{10}$$

$$\hat{B}_L^{l+1} = \Phi_{SW\text{-}Att}(\Phi_{LN}(B_L^l)) + B_L^l \tag{11}$$

$$B_L^{l+1} = \Phi_{FFN}(\Phi_{LN}(\hat{B}_L^{l+1})) + \hat{B}_L^{l+1} \tag{12}$$

where Φ_{LN} and Φ_{FFN} denote Layer Normalization and Feed-Forward Networks, respectively. This alternating design produces a contextually complete LiDAR BEV representation B'_L, serving as a robust foundation for fusion.

D. Detection Head and Loss

We concatenate the semantically focused image feature B'_I (Sec. III-B) and contextually focused LiDAR feature B'_L (Sec. III-C), fusing them via Local Context-Attention (LCA) [12]:

$$B'_{\text{fused}} = \Phi_{\text{LCA}}(\text{Concat}(B'_I, B'_L)) \tag{13}$$

For detection, we use a Transformer-based head enhanced with Hard Instance Probing (HIP) [29]. Unlike standard strategies, HIP explicitly targets difficult objects (e.g., small or occluded) by iteratively probing multi-stage heatmaps to initialize and refine queries, ensuring computational focus on challenging regions.

Heatmap Loss ($\mathcal{L}_{\text{hm}}$): We predict object center heatmaps using a 2D head. To handle sample imbalance caused by BEV sparsity, we employ Gaussian Focal Loss [18]:

$$\mathcal{L}_{\text{hm}} = -\frac{1}{N}\sum_{xyc} \begin{cases} (1-\hat{Y})^{\alpha}\log(\hat{Y}) & \text{if } Y=1 \\ (1-Y)^{\beta}\hat{Y}^{\alpha}\log(1-\hat{Y}) & \text{otherwise} \end{cases} \tag{14}$$

where Y and $\hat{Y}$ are the ground truth and prediction, N is the object count, and α, β are hyperparameters.

Transformer Loss ($\mathcal{L}_{\text{trans}}$): Based on bipartite matching via the Hungarian algorithm [9], this term combines focal classification and L1 regression losses:

$$\mathcal{L}_{\text{trans}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} \tag{15}$$

Total Loss. To facilitate the training of the entire framework, we use a Faster R-CNN style [30] 2D auxiliary loss $\mathcal{L}_{2D}$ alongside the primary 3D task losses. The final total Losses are defined as a weighted sum of the 3D and 2D auxiliary losses, balanced by empirical hyperparameters λ:

$$\mathcal{L}_{total} = \lambda_{hm}\mathcal{L}_{hm} + \lambda_{trans}\mathcal{L}_{trans} + \lambda_{2D}(\mathcal{L}_{RPN} + \mathcal{L}_{RoI}) \tag{16}$$

IV. Experiments

A. Dataset and Metrics

We evaluate on the nuScenes benchmark [1], which features a 360-degree sensor suite including a 32-beam LiDAR (20Hz) and six cameras (1600×900, 12Hz). Its 1,000 driving scenes are officially split into 700/150/150 for training/validation/testing, providing 3D annotations for 10 classes.

TABLE I

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART DETECTORS ON THE NUSCENES TEST SET. ABBREVIATIONS: C.V. (CONSTRUCTION VEHICLE), T.C. (TRAFFIC CONE), L (LIDAR), AND C (CAMERA). OUR MODEL IS MARKED IN GRAY. **BOLD** INDICATES THE BEST PERFORMANCE.

Method	Modality	mAP↑	NDS↑	Car	Truck	C.V.	Trailer	Bus	Barrier	Motor.	Bike	Ped.	T.C.
PointPillars [2]	L	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
CenterPoint [18]	L	65.5	70.2	86.2	56.7	28.2	58.8	66.3	78.2	68.3	44.2	86.1	82.0
TransFusion-L [23]	L	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
FocalFormer3D-L [29]	L	68.7	72.6	87.8	59.4	37.8	65.7	**73.0**	77.8	77.4	52.4	90.0	83.4
GraphBEV [25]	L+C	70.9	73.2	89.9	63.1	**40.0**	71.4	64.2	79.5	74.2	52.8	89.1	86.4
CMT [31]	L+C	72.2	74.1	88.0	**63.3**	37.3	**75.4**	65.4	78.2	79.1	60.6	87.9	84.7
SETR-Fusion [27]	L+C	71.2	73.3	**90.8**	62.3	38.5	70.6	63.5	79.6	75.3	53.9	90.5	87.0
SparseFusion [28]	L+C	72.0	73.8	88.0	60.2	38.7	72.0	64.9	79.2	78.5	59.8	**90.9**	87.9
MSMDFusion [24]	L+C	71.5	74.0	88.4	61.0	35.2	66.2	71.4	**80.7**	76.9	58.3	90.6	88.1
FocalFormer3D [29]	L+C	71.6	73.9	88.5	61.4	35.9	66.4	71.7	79.3	80.3	57.1	89.7	85.3
BEVFusion-MIT [12]	L+C	71.3	73.3	88.1	60.9	34.4	62.1	69.3	78.2	72.2	52.2	89.2	86.7
CMF-IoU [26]	L+C	69.8	72.6	87.5	59.0	36.2	69.5	65.1	78.9	74.9	48.2	90.4	**88.3**
FocusBEV (Ours)	**L+C**	**72.3**	**74.6**	89.9	63.1	37.6	66.3	70.8	76.5	**80.7**	**61.2**	90.2	86.6

TABLE II

PERFORMANCE COMPARISON WITH SOTA DETECTORS ON NUSCENES VALIDATION SET, GROUPED BY MODALITY (C: CAMERA-ONLY, L: LIDAR-ONLY, L+C: FUSION). OUR MODEL IS MARKED IN GRAY. **BOLD** INDICATES THE BEST PERFORMANCE.

Method	Modality	mAP↑	NDS↑
BEVDepth [14]	C	35.10	47.50
PETRv2 [32]	C	34.90	45.60
StreamPETR [11]	C	**44.90**	**54.60**
FocusBEV (Ours)	**C**	36.89	43.78
CenterPoint [18]	L	56.40	64.80
FSTR-L [33]	L	65.50	70.30
FocalFormer3D-L [29]	L	**66.40**	70.90
DSVT-Pillar [17]	L	**66.40**	71.10
FocusBEV (Ours)	**L**	**66.40**	**71.50**
BEVFusion(MIT) [12]	L+C	68.50	71.40
DeepInteraction [34]	L+C	69.90	72.60
CMT [31]	L+C	70.30	72.90
SparseFusion [28]	L+C	70.50	72.80
FocalFormer3D-F [29]	L+C	70.50	73.10
FocusBEV (Ours)	**L+C**	**71.32**	**73.77**

Following the official protocol, we employ mean Average Precision (mAP) and nuScenes Detection Score (NDS).

NuScenes mAP uses 2D center distance on the ground plane, averaged over thresholds $\mathbb{D} = \{0.5\text{m}, 1\text{m}, 2\text{m}, 4\text{m}\}$ across classes $\mathbb{C}$:

$$\text{mAP} = \frac{1}{|\mathbb{C}||\mathbb{D}|} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} \text{AP}_{c,d} \tag{17}$$

NDS is a weighted combination of mAP and a True Positive Metric Score (S_{TP}). S_{TP} aggregates five error types: Translation (mATE), Scale (mASE), Orientation (mAOE), Velocity (mAVE), and Attribute (mAAE). NDS is defined as:

$$S_{TP} = \sum_{mTP \in \mathbb{M}_{TP}} (1 - \min(1, mTP)) \tag{18}$$

$$\text{NDS} = \frac{1}{10}(5 \times \text{mAP} + S_{TP}) \tag{19}$$

where $\mathbb{M}_{TP}$ denotes the set of the five TP metrics. This offers a robust evaluation covering localization, size, and orientation accuracy.

B. Implementation Details

Our framework is implemented based on the mmdetection3d codebase [7], utilizing FocalFormer3D [29] as the baseline. For the image branch, we employ the Swin-Transformer [21] as the backbone equipped with FPN [20]. Notably, the Semantic-Focus Module is inserted immediately before the LSS view transformation to proactively enhance foreground features. Simultaneously, the LiDAR branch covers a detection range of $[-54\text{m}, 54\text{m}]$ for X/Y and $[-5\text{m}, 3\text{m}]$ for Z, utilizing a voxel size of $(0.075\text{m}, 0.075\text{m}, 0.2\text{m})$. We integrate the LIA module immediately after the sparse backbone to aggregate global context via Shifted-Window attention. Finally, multi-modal features are concatenated into the FocalDecoder [29] (2 layers, 8 heads), utilizing Hard Instance Probing (HIP) for query initialization.

We utilize the AdamW optimizer with Cosine Annealing. The initial learning rates are set to 1×10^{-4} for LiDAR and 1×10^{-5} for images. To ensure stable convergence, we adopt a Two-Stage training strategy: (1) Unimodal Pre-training (20 epochs): Image and LiDAR branches are trained independently. We apply CBGS [35] for LiDAR, and disable GT-Paste [15] augmentation after the 15th epoch to reduce domain gaps. (2) Multi-modal Fine-tuning (6 epochs): We freeze the image backbone and LiDAR BEV components, training only the fusion module and detection head. GT-Paste is applied only during the first epoch. All experiments are conducted on 8 NVIDIA RTX 4090 GPUs with a total batch size of 8.

Regarding the loss configuration, we set the weights to $\lambda_{\text{hm}} = 1.0$, $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{box}} = 0.15$, and $\lambda_{\text{2D}} = 0.2$ for the auxiliary task. The Gaussian Focal Loss hyperparameters are

TABLE III
ABLATION STUDY ON THE EFFECTIVENESS OF OUR CORE COMPONENT. BEST RESULTS ARE IN **BOLD**.

	Semantic-Focus	LIA	mAP(%)↑	NDS(%)↑
(a)			70.83	73.26
(b)		✓	70.98	73.56
(c)	✓		71.24	73.69
(d)	✓	✓	**71.32**	**73.77**

TABLE IV
ABLATION ANALYSIS ON THE WEIGHTING STRATEGIES OF THE SEMANTIC-FOCUS MODULE. BEST RESULTS ARE IN **BOLD**.

	FG Enhance	BG Suppress	mAP(%)↑	NDS(%)↑
(a)			35.89	43.26
(b)		✓	35.98	43.30
(c)	✓		36.25	43.56
(d)	✓	✓	**36.89**	**43.78**

set to $\alpha = 2$ and $\beta = 4$. No Test-Time Augmentation (TTA) is employed during inference.

C. Comparison With State-of-the-Art Methods

To comprehensively evaluate the effectiveness of FocusBEV, we conduct extensive quantitative comparisons with existing State-of-the-Art (SOTA) methods in both the nuScenes test and validation sets [1]. As presented in Table I, under the single-model setting on the test set, FocusBEV demonstrates superior performance. Compared to mainstream competitive approaches, our model establishes a significant lead in key metrics, achieving 72.3% mAP and 74.6% NDS. These results underscore the strong robustness and generalization capability of our framework when navigating complex, real-world driving scenarios. For a more granular performance analysis, we further benchmark our method on the validation set (Table II).

FocusBEV consistently outperforms existing methods in multiple evaluation dimensions, achieving high scores of 71.3% mAP and 73.7% NDS. These empirical results validate the efficacy of the Dual-Focus mechanism, demonstrating that overcoming the systemic bottlenecks of prior fusion approaches directly translates to consistent improvements in detection precision.

D. Qualitative Analysis

Detection Results Visualization. To first validate the perceptual capabilities in challenging real-world scenarios, Fig. 4 compares the 3D detection results of FocusBEV against the baseline. As observed in the "Back Left" and "Back" views, the baseline model fails to recall cyclists and distant vehicles. In stark contrast, FocusBEV successfully identifies these targets and generates bounding boxes in the LiDAR BEV view that are highly consistent with the ground truth. This improvement is largely attributed to the Semantic-Focus Module, which preserves critical semantic cues of small or distant objects that are often drowned out by noise in the baseline. Simultaneously, regarding hallucinations, the baseline model produces significant false positives in the open area of the "Back" view. Although it detects the presence of a vehicle, it exhibits severe orientation deviations. Conversely, FocusBEV demonstrates superior robustness in such sparse regions, effectively avoiding spurious responses to generate cleaner detection results. This validates the ability of the LIA module to construct a coherent global context, reducing confusion caused by local feature ambiguity.

TABLE V
ABLATION STUDY ON THE INTERACTION BETWEEN LIA AND MULTISTAGE HEATMAP (MH) LAYERS. BEST RESULTS ARE IN **BOLD**.

	LIA	MH Layers	mAP(%)↑	NDS(%)↑
(a)		1	65.49	70.70
(b)		2	66.31	71.19
(c)	✓	1	65.70	71.47
(d)	✓	2	**66.39**	**71.58**

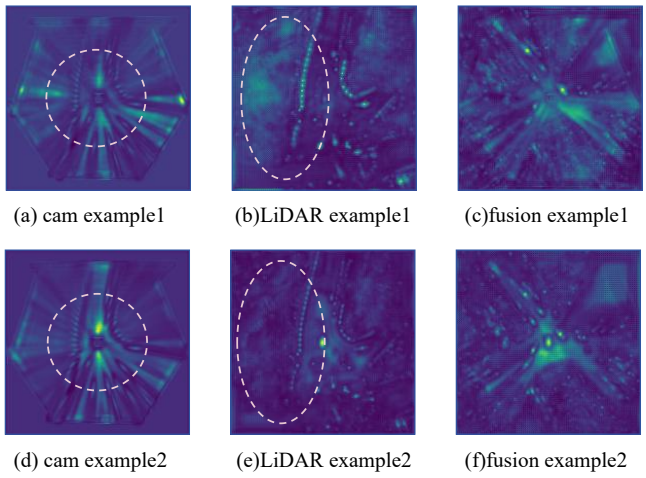
BEV Feature Visualization. To elucidate the underlying reasons for these detection improvements, we further visualize the intermediate BEV features in Fig. 5. Comparing Fig. 5(a) and (d), the baseline BEV features suffer from severe radial artifacts and background noise caused by the inherent ambiguity of the LSS mechanism. In contrast, our Semantic-Focus Module effectively suppresses these spurious noises along the viewing frustums, concentrating feature responses on the objects themselves and yielding a higher semantic Signal-to-Noise Ratio (SNR). Simultaneously, in the LiDAR branch (Fig. 5(b) vs. (e)), the LIA module reconstructs spatially continuous structures and reduces background clutter compared to the fragmented baseline features, particularly in the white dashed region. Consequently, the final fused features (Fig. 5(f)) exhibit sharper object boundaries and cleaner backgrounds compared to the baseline (Fig. 5(c)). This visual evidence strongly corroborates the effectiveness of our "dual-focus" strategy in providing high-quality representations for downstream detection.

E. Ablation Studies

We conduct ablation studies on the nuScenes validation set [1] using FocalFormer3D [29] as the baseline to verify our components.

Effectiveness of Core Components. To assess the individual and combined contributions of our core modules, we perform a component-wise analysis as detailed in Table III. Starting from the baseline performance of 70.83% mAP and 73.26% NDS, integrating only the LIA module improves the metrics to 70.98% mAP and 73.56% NDS, validating that aggregating long-range dependencies in the LiDAR branch effectively compensates for the locality limitations of CNNs. Alternatively, incorporating only the Semantic-Focus module yields a substantial gain, boosting the mAP to 71.24% and NDS to 73.69%, which underscores the critical role of proactive semantic enhancement in mitigating noise during projection. Most importantly, when both modules are jointly inte-

Fig. 4. Qualitative comparison of 3D detection results between FocusBEV (top) and the Baseline (bottom). Red dashed circles highlight targets missed by the baseline but successfully detected by FocusBEV. Green dashed circles indicate false positive predictions or orientation errors generated by the baseline in sparse areas.



(a) cam example1 (b)LiDAR example1 (c)fusion example1

(d) cam example2 (e)LiDAR example2 (f)fusion example2

Fig. 5. Qualitative Comparison of BEV Feature Maps. The top row (a-c) shows results from the baseline model, while the bottom row (d-f) presents our proposed method.

grated, FocusBEV achieves the peak performance of 71.32% mAP and 73.77% NDS. This synergistic effect demonstrates that the two modules are complementary: Sem-Focus purifies the input features while LIA reconstructs the global context, together unlocking the full potential of the fusion framework.

Impact of Weighting Strategies in Semantic-Focus. To validate the efficacy of the masking mechanism within the Semantic-Focus Module, we decouple it into two distinct strategies: Foreground Enhancement (FG Enhance) and Background Suppression (BG Suppress). The ablation results are reported in Table IV. Starting from the baseline without any weighting (Row a), incorporating solely BG Suppress yields a marginal performance gain, improving mAP from 35.89% to 35.98%. In contrast, applying only FG Enhance (Row c) leads to a more significant improvement, boosting the mAP to 36.25% and NDS to 43.56%, which suggests that highlighting critical object features is vital for view transformation. The optimal performance is achieved when both strategies are jointly deployed (Row d), reaching 36.89% mAP and 43.78% NDS. This demonstrates that simultaneously amplifying foreground signals and dampening background noise effectively maximizes the semantic signal-to-noise ratio.

Interaction between LIA and Detection Head. Finally, we analyze the structural interaction between the LIA module and the Multistage Heatmap (MH) layers in the detection head, as shown in Table V. Increasing the number of MH layers from 1 to 2 in the baseline (Row a vs. b) brings noticeable improvements, confirming that deeper heatmap supervision aids in object localization. However, the introduction of LIA (Row c vs. a) provides a more robust gain even with a single MH layer, proving its capability to extract superior context features. The best performance is observed when LIA is coupled with 2 MH layers (Row d), demonstrating that the high-quality global representations generated by LIA can be further exploited by a more capable detection head [29] to refine final predictions.

V. Conclusion

In this paper, we identified the systemic dual-defocus bottleneck in multi-modal fusion and proposed FocusBEV as a comprehensive solution. By redesigning the encoding pipelines, we introduced the Semantic-Focus Module to proactively rectify image feature distributions prior to view transformation, and the Long-range Interaction Aggregator (LIA) to reconstruct integral LiDAR context via global attention. Extensive experiments on the nuScenes dataset demonstrate that FocusBEV significantly outperforms state-of-the-art methods, exhibiting superior robustness in complex scenarios.

Limitations and Future Work. Despite promising results, limitations remain. First, the stacked self-attention layers in LIA incur higher computational overhead than pure CNNs, potentially affecting latency on edge devices. Second, the Semantic-Focus Module relies on the quality of the 2D detection prior; thus, its guidance may degrade under extreme visibility conditions. Future work will explore more efficient attention mechanisms and temporal fusion strategies to further enhance system robustness.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11 621–11 631.

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

[3] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *CVPR*, 2018, pp. 4490–4499.

[4] X. Zhang, K. Bi, S. Chan, S. Lu, and X. Zhou, "Synet: A synergistic network for 3d object detection through geometric-semantic-based multi-interaction fusion," *IEEE Transactions on Multimedia*, 2025.

[5] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *CVPR*, 2020, pp. 4604–4612.

[6] S. Chan, B. Duan, X. Fan, and J. Hu, "Dci-prnet: 3d object detection network via dual cross-modal interaction and progressive reasoning," in *IJCNN*, 2025, pp. 1–8.

[7] M. Contributors, "Mmdetection3d: Openmmlab next-generation platform for general 3d object detection," 2020.

[8] H. Yu, S. Chan, X. Zhou, and X. Zhang, "Primepsegter: Progressively combined diffusion for 3d panoptic segmentation with multi-modal bev refinement," *IEEE Transactions on Multimedia*, 2025.

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

[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *ICCV*, 2023, pp. 3621–3631.

[12] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.

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

[14] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *AAAI*, vol. 37, no. 2, 2023, pp. 1477–1485.

[15] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[16] J. Tu, P. Wang, and F. Liu, "Pp-rcnn: Point-pillars feature set abstraction for 3d real-time object detection," in *IJCNN*, 2021, pp. 1–8.

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

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

[19] Z. Huang, Y. Wang, X. Tang, and H. Sun, "Boundary-aware set abstraction for 3d object detection," in *IJCNN*, 2023, pp. 01–07.

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

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

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

[23] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *CVPR*, 2022, pp. 1090–1099.

[24] Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, "Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection," in *CVPR*, 2023, pp. 21 643–21 652.

[25] Z. Song, L. Yang, S. Xu, L. Liu, D. Xu, C. Jia, F. Jia, and L. Wang, "Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection," in *ECCV*, 2024, pp. 347–366.

[26] Z. Ning, Z. Liu, X. Gao, Y. Zuo, J. Yang, Y. Fang, and W. Liu, "Cmf-iou: Multi-stage cross-modal fusion 3d object detection with iou joint prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[27] X. Qu, K. Qin, Y. Li, S. Zhang, Y. Li, S. Shen, and Y. Gao, "Semantic-enhanced and temporally refined bidirectional bev fusion for lidar–camera 3d object detection," *Journal of Imaging*, vol. 11, no. 9, p. 319, 2025.

[28] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, and W. Zhan, "Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection," in *ICCV*, 2023, pp. 17 591–17 602.

[29] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, and J. M. Alvarez, "Focalformer3d: focusing on hard instance for 3d object detection," in *ICCV*, 2023, pp. 8394–8405.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, vol. 28, 2015.

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

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

[33] D. Zhang, Z. Zheng, H. Niu, X. Wang, and X. Liu, "Fully sparse transformer 3-d detector for lidar point cloud," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

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

[35] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *arXiv preprint arXiv:1908.09492*, 2019.