Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Pure-Vision BEV 3D Perception with Hybrid Data-Feature Optimization

Aitong Mao
Guangdong University of Technology
Beihai Tan
bhtan@gdut.edu.cn
Guangdong University of Technology
Rong Yu
Guangdong University of Technology
Guoliang Cheng
Southern University of Science and Technology
Xiaoxu Wang
South China Normal University
Wuchang Zhong
Guangdong University of Technology

Research Article

Keywords: Bird's-eye view , Data augmentation, 3D object detection, Multi-view fusion, Semantic segmentation

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8898319/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Enhancing Pure-Vision BEV 3D Perception with Hybrid Data-Feature Optimization

Aitong Mao[1], Beihai Tan[1*], Rong Yu[2], Guoliang Cheng[3], Xiaoxu Wang[4], Wuchang Zhong[2]

[1*]School of Integrated Circuits, Guangdong University of Technology, Huanchengxi Road, Guangzhou, 510006, Guangdong, China.
[2]School of Automation, Guangdong University of Technology, Huanchengxi Road, Guangzhou, 510006, Guangdong, China.
[3]Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Xueyuan Avenue, Shenzhen, 518055, Guangdong, China.
[4]School of Electronic Science and Engineering, South China Normal University, Wanjin Road, Foshan, 528225, Guangdong, China.

*Corresponding author(s). E-mail(s): bhtan@gdut.edu.cn;
Contributing authors: 15159888503@163.com; yurong@gdut.edu.cn;
chenggl2024@mail.sustech.edu.cn; wangxiaoxu@m.scnu.edu.cn;
domainarcher618@gmail.com;

Abstract

Bird's eye view (BEV)-based approaches relying solely on cameras offer a cost-effective alternative to LiDAR-based solutions for 3D perception, attracting significant research attention. However, existing methods face limitations in data diversity, contextual modeling across multi-camera views, and multi-scale feature fusion. To address these challenges, this paper proposes HDO-BEV, a Hybrid Data-Feature Optimization-enhanced architecture for pure-vision BEV 3D perception. HDO-BEV integrates three novel modules: RandomFlip for data augmentation, HS-FPN for optimized multi-scale feature fusion, and ContextBlock for context enhancement. Experiments on the nuScenes-mini dataset demonstrate that HDO-BEV achieves 0.388 mAP and 0.472 NDS, outperforming the baseline SA-BEV by 0.9%. These results validate that targeted architectural enhancements can significantly advance pure-vision BEV 3D environmental sensing for scalable autonomous driving systems.The source code

for this study is open-sourced at: https://github.com/zaixianbaipiao/Enhancing-Pure-Vision-BEV-3D-Perception-with-Hybrid-Data-Feature-Optimization [DOI: 10.5281/zenodo.18587515]

Keywords: Bird's-eye view ; Data augmentation; 3D object detection; Multi-view fusion; Semantic segmentation

1 Introduction

In recent years, 3D object detection based on BEV representation has become a prominent research focus due to its cost-effectiveness and efficiency in the sensor-driven perception systems of autonomous driving. Recent advances in BEV (Bird's-Eye-View) representation have unified multi-camera views into an occlusion-free scene representation, becoming the mainstream framework for vision-sensor-based 3D detection. Representative works have achieved remarkable progress in multi-camera geometric reasoning and 3D object localization by constructing BEV features from camera sensors. However, in resource-constrained scenarios, existing sensor-based methods still face three core challenges: First, current data augmentation strategies for sensor inputs remain limited. Most conventional augmentation techniques are restricted to single-frame geometric transformations (e.g., flipping and cropping) and fail to adequately simulate the diversity of camera viewpoints [1–4]. Studies have shown that models trained solely on fixed sensor viewpoints exhibit significantly degraded generalization capabilities in dynamic scenarios involving vehicle steering, vertical jolting, etc. Second, multi-sensor feature fusion often suffers from severe local isolation. Most existing methods employ fixed connection patterns between views, which hinder dynamic interaction across views and make it difficult to capture long-range cross-view dependencies among camera streams [5–7]. For example, the Feature Pyramid Network (FPN) structure of BEVDet [1] directly concatenates multi-scale sensor features, leading to loss of semantic associations for distant targets. Third, failure of multi-scale sensor representation. Traditional FPN architectures struggle to adapt to the perspective projection characteristics in BEV space, where high-frequency details of small objects are diluted during recursive upsampling of camera-derived features. Although SOLOFusion [8] introduces temporal fusion, it lacks an effective mechanism to address cross-scale misalignment in dense regions using purely visual sensor data.

Therefore, to address the aforementioned issues, this paper proposes an enhanced hybrid architecture, HDO-BEV, which achieves improvements in pure-vision sensor-based BEV detection performance. Specifically, three key innovations are introduced: (1) The RandomFlip sensor data augmentation module, which applies controllable geometric flip perturbations and synchronously transforms multi-modal sensor annotations. Based on the invariance learning theory, it significantly enhances the model's robustness to camera viewpoint variations at low cost. (2) The HS-FPN for optimized multi-scale sensor feature representation. Its high-frequency perception and spatial dependency components are dedicated to reinforcing edge features of small objects and correcting cross-scale sensor feature misalignment, thereby generating higher-quality,

scale-aware BEV feature maps. (3) A redesigned ContextBlock module, first adapted and applied to the feature neck of BEV models, to tackle context fragmentation in multi-camera feature fusion. By integrating channel attention and global context aggregation, it enhances semantic consistency and long-range dependency modeling across visual sensor views.

Comprehensive experiments on the NuScenes-mini dataset demonstrate that the HDO-BEV model outperforms baseline methods including BEVDet, BEVFormer, and the original SA-BEV in both detection accuracy and comprehensive score. Detailed ablation studies further validate the effectiveness of each proposed sensor-oriented module. This work provides new insights and technical solutions for constructing high-performance, robust vision-sensor-based BEV 3D detection systems under resource-constrained conditions.

2 Related Work

2.1 Vision-Based 3D Object Detection

In early vision-based 3D object detection methods, the most prevalent approach involves extracting features from 2D images and directly predicting 3D object attributes such as position and size. These methods typically rely on traditional convolutional neural network (CNN) architectures for object detection, with 3D information obtained via regression or classification[9]. For instance, the 2D detector CenterNet[10] can be adapted to predict 3D objects with minimal modifications[11]. RTM3D[12] recovers complete 3D information by predicting projections of 9 key points of a 3D bounding box onto the image plane and constructing geometric constraints. Building upon 2D detection results, MonoGRNet [13] directly regresses 3D attributes through a neural network. The advantages of these methods lie in their lightweight structures, low computational complexity, and fast image processing capabilities. However, their inherent limitations are prominent: relying solely on 2D cues from monocular or binocular images to infer 3D attributes introduces inherent ambiguity in depth estimation. Particularly for distant objects, their extremely low pixel occupancy in images severely restricts the accuracy of depth and size estimation[11].

To address this core issue, subsequent research has sought to deeply integrate geometric information with advanced network architectures to enhance the reliability of depth estimation and 3D localization[14], alongside the introduction of more sophisticated network designs. Recent methods exemplify this trend: the Categorical Depth Distribution Network (CaDDN)[15] predicts a categorical distribution of depths for each pixel. This projects rich image context features into discrete 3D spatial depth intervals, enabling the construction of bird's-eye-view (BEV) features for detection. MonoDLE[16], through intensive diagnostic experiments, was the first to systematically quantify and proposed a loss function tailored for 3D Intersection over Union (IoU) to optimize size estimation. These methods attempt to more reliably solve the lifting problem from 2D to 3D from different perspectives.

With the increasing maturity of Transformer architectures in computer vision[17], several works have explored their application in vision-based 3D detection, leveraging their powerful global context modeling capability to enhance feature representation.

For example, the PETRv2 method[2] incorporates temporal information by utilizing features from consecutive frames to improve detection accuracy and efficiency. BEVFormer [18] employs spatial cross-attention for view transformation and temporal self-attention for temporal feature fusion.BEVDet[1] adopts the CenterPoint[19] detection head to predict 3D objects directly from BEV features. This paradigm achieves comparable accuracy without cumbersome operations and is easily extensible. BEVDet4D[20] processes multiple keyframes to introduce temporal information. Our model, based on SA-BEV, enhances feature extraction and BEV spatial representation capabilities—thereby improving detection robustness and accuracy—by introducing three key components: a random flip augmentation module, a context feature enhancement module, and an efficient HS-FPN (High-frequency and Spatial perception FPN) structure.

2.2 Multi-Scale Feature Fusion in BEV Perception

In 3D object detection, FPN and its variants are primarily designed to construct multi-scale feature representations that integrate high-resolution details and strong semantic information[21]. However, the original FPN[22] was tailored for 2D image planes; its direct application to 3D detection lacks targeted optimizations for 3D geometric structures and Bird's-Eye-View (BEV) space characteristics.

Keypoint detector-based FPNs, exemplified by CenterNet[10], treat object centers as keypoints and regress 3D attributes at these locations. The VPN method[18] adopts a ResNet+FPN backbone to extract multi-scale features from multi-camera images, providing inputs for cross-view attention fusion. The Pseudo-LiDAR approach[23] leverages FPN to extract multi-scale features from monocular images, enhancing depth estimation robustness. Furthermore, the pure vision branch of DCF[24] employs FPN to fuse multi-view image features and mitigate scale discrepancies. These methods share a common limitation: the features generated by FPN struggle to precisely conform to spatial geometric constraints during 3D transformation, with insufficient integration of semantic and geometric information.

To address these issues, several enhanced FPN structures have recently been proposed. For instance, Lite-FPN[25] departs from dense full-feature-map fusion, instead performing multi-scale sampling and fusion only at sparse keypoint locations. BEVDet[1] uses a standard FPN to extract multi-resolution features from multi-camera images, which are projected into BEV space and concatenated to form a multi-scale BEV feature pyramid balancing high-resolution details and low-resolution context. Fast-BEV [26] employs FPN for multi-scale image encoding paired with a lightweight view transformation module to maintain perception at reduced computational cost. For multi-task unified BEV representation, the FPN in M^{2}BEV[27] ensures multi-scale consistency of object boundaries and semantic regions in BEV space. Notably, SOLOFusion[8] first applies FPN to single-frame multi-scale features, then temporally aligns and concatenates multi-frame features, followed by a secondary FPN fusion to enhance perception stability in long-term sequences.

These improved FPN variants enhance BEV detection performance through multi-scale fusion, lightweight design, or temporal alignment. In this work, we introduce the HS-FPN[28] into our BEV model. Specifically, HS-FPN utilizes a high-pass filter to

generate high-frequency responses, thereby enriching and highlighting features of small objects. Additionally, it incorporates a Spatial Dependency Perception (SDP) module to capture spatial relationships that are typically underrepresented in standard FPN architectures.

2.3 Lightweight Feature Enhancement for BEV Models

Effective contextual feature enhancement provides crucial spatial and semantic correlations between objects and scenes[29], as well as among objects themselves. Early feature enhancement methods in 3D detection largely adopted established 2D vision techniques. The milestone LSS[3] introduced a "lift-and-project" paradigm, enhancing BEV feature expressiveness through explicit depth estimation and feature projection. OFT[22] employed an orthographic feature transform, projecting BEV grids onto images to generate scale-consistent BEV features via regional averaging. VPN [30] extracted multi-scale features from multiple camera views, followed by weighted cross-view attention fusion. PON[21] utilized a multi-scale feature pyramid to directly learn the perspective transformation matrix from the image to the BEV plane. These early approaches, however, faced two primary limitations. First, their performance heavily relied on unstable geometric priors (e.g., LSS's depth estimation, OFT's planar assumption)[31], making them susceptible to feature misalignment from geometric errors in complex scenes. Second, their fusion processes lacked explicit modeling of long-range spatial contexts (e.g., PON's global transformation, VPN's local attention), limiting holistic scene understanding.

As BEV evolved into a unified representation space for autonomous driving perception, research shifted towards more intelligent image-to-BEV mapping, spatiotemporal feature integration, and 3D prior injection. BEVFormer[18] fused multi-level features via a Feature Pyramid Network (FPN) to perceive objects across scales simultaneously. BEVDet [1] used FPN for multi-scale feature extraction from multi-camera images, projecting them into BEV space to form a multi-scale BEV feature pyramid[32]. The PETR series[2, 33, 34] implicitly encoded 3D positional information into 2D features before transforming them into BEV via Transformers. SOLOFusion[8] focused on long-term temporal fusion, aligning and concatenating multi-frame 2D features temporally prior to BEV encoding. Pursuing efficiency, Fast-BEV[26] employed a shared backbone, multi-scale encoding, and a lightweight view transformer for rapid BEV feature generation.

Early methods overemphasized geometric priors at the expense of long-range context, causing feature misalignment and semantic fragmentation[35]. Contextual feature enhancement requires balancing semantic correlation and computational efficiency. Recent studies have optimized feature focusing through dynamic attention mechanisms, for instance, the DDA module [36], which leverages dynamic offsetting of object queries and weight constraints to enhance the perception of small objects and complex contours, enabling efficient cross-feature-map fusion without increasing computational cost.

Inspired by this, we introduce a lightweight ContextBlock module into the image feature fusion neck of the SA-BEV framework. This module draws on the idea of dynamic attention by integrating channel attention and global context aggregation,

enabling dynamic calibration of cross-view and cross-scale dependencies while avoiding significant computational overhead. This enables adaptive fusion that boosts the semantic richness and spatial consistency of BEV features, providing a more reliable foundation for subsequent detection heads.

3 METHODS

We propose an enhanced hybrid architecture named HDO-BEV.The overall framework is illustrated in FIGURE 1.

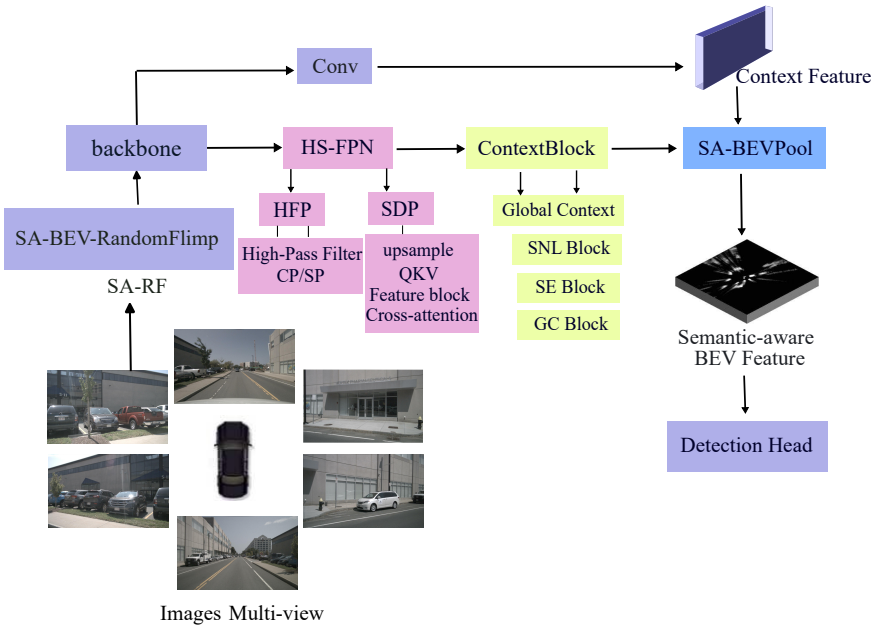


Fig. 1 Overall framework of HDO-BEV. The framework takes multi-view images as input, initially extracting foundational features via a backbone network. These foundational features are subsequently processed along two parallel pathways: one pathway enhances the features, while the other generates contextual features. Features from multiple pathways are integrated within the SA-BEVPool to derive semantic-aware bird's-eye-view (BEV) features. Finally, object detection is accomplished by the detection head.

Based on the SA-BEV model as the baseline, HDO-BEV systematically enhances model performance through targeted improvements at three levels: data augmentation, feature extraction optimization, and feature fusion upgrading.First, at the data input level, we introduce a random flip augmentation strategy to enhance the model's generalization capability for scene symmetry and diversity in a cost-effective manner.Second, at the feature extraction level, we integrate the HS-FPN module and the

ContextBlock context enhancement module. By explicitly modeling global context, these modules generate higher-quality, scale-aware bird's-eye-view (BEV) features.

3.1 Data Augmentation: Random Flip

In pure vision-based BEV-based 3D detection, models are susceptible to limitations in training data distribution, leading to insufficient generalization capability against viewpoint perturbations in real-world driving scenarios. To address this issue, this study introduces random flip augmentation, whose effectiveness is grounded in the theory of invariance learning for data augmentation: by applying an invertible transformation T to original samples to generate new samples T(x), if the model learns transformation invariance such that$f(T(x)) = T(f(x))$, it can generalize to unseen transformed scenarios. Taking horizontal flip as an example, with the origin of the coordinate system at the image center, let the original image be $I(x, y)$; the flipped image is $I'(x, y) = I(x, -y)$. Correspondingly, the 3D bounding box parameter transformation is

$$T : (x, y, z, \theta) \mapsto (x, -y, z, -\theta). \tag{1}$$

If the model-predicted 3D box is $\hat{B} = f(x)$, the prediction for the flipped sample should satisfy

$$\hat{B}' = T(\hat{B}) = f(T(I)), \tag{2}$$

indicating the model possesses flip invariance. For multi-camera systems, it is imperative to ensure the synchronization and global consistency of view-wise augmentations. Let the projection matrix of the u-th camera be P_u, and its image transformation under horizontal flipping be $I'_u = T^h_{\text{in}}$, where T^h_{in} denotes the pixel-level flipping operation. The global 3D transformation T^h_{3D} must be compatible with the image transformations of all views to satisfy projection consistency:

$$\pi(P_u, T^h_{3D}(X)) = T^h_{\text{in}}(\pi(P_u, X)) \quad \forall X \in \mathbb{R}^3, \tag{3}$$

where π denotes the projection function. In this study, we approximately satisfy this constraint by applying the same pixel-level flipping T^h_{in} to all rectified images and simultaneously updating 3D annotations via the corresponding T^h_{3D}, ensuring that the generated multi-view samples remain geometrically consistent.

indicate the transformations applied to inputs and labels, respectively. Invariance learning is typically achieved by minimizing the joint loss before and after augmentation:

$$L_{\text{total}} = \mathbb{E}_{(I,B)} \left[\mathcal{L}(f(I), B) + \lambda \cdot \mathcal{L}\left(f(T(I)), T(B)\right) \right], \tag{4}$$

Among them, λ is the weighting coefficient (set to 1 in this study), and $\mathcal{L}$ denotes the detection loss. This objective compels the model to fit both original and transformed samples simultaneously, thereby learning robust features invariant to transformation T.

Random Flip, as a classical data augmentation technique, effectively expands data diversity and enhances model robustness to pose variations by simulating symmetry transformations of scenes. In this study, we design the RandomFlip module tailored

for the SA-BEV model, whose structure is illustrated in FIGURE 2. A key feature of this module is its specific adaptation and optimization for the pure visual multi-view input paradigm, relying solely on RGB image sequences captured by multiple cameras to form a panoramic visual representation of driving scenes. The core objectives of this module are: (1) simulating viewpoint perturbations in real-world driving (e.g., vehicle left/right turning, vertical jolting); (2) generating virtual samples via symmetry transformations to mitigate overfitting risks; and (3) maintaining physical consistency of data (synchronous transformation of annotations and images) to ensure augmentation validity.

The RandomFlip augmentation module is implemented within the PIPELINES framework of MMDetection3D, integrated into the data loading pipeline. During the training phase, it randomly applies horizontal or vertical flipping to input images and annotations, thereby simulating viewpoint perturbations in real driving through controllable random flipping. Specific operations are as follows: we first specify the flip probability to control augmentation intensity and the flip direction, validate the legitimacy of the direction, then use a random number generator to decide whether to execute flipping—ensuring each sample is augmented with probability flip_ratio to avoid distortion from over-augmentation. A differentiated flipping strategy is designed for multimodal data to guarantee physical consistency. Tensor flip operations enable dimension reversal: for horizontal flipping, the width dimension (axis$=-1$) is reversed; for vertical flipping, the height dimension (axis$=-2$) is reversed. For 3D bounding boxes (gt_bboxes_3d) in the LiDAR coordinate system (x: forward, y: leftward, z: upward, θ: yaw angle), coordinate transformations are applied: during horizontal flipping, the y-coordinate is inverted ($y \to -y$) and the yaw angle θ is negated ($\theta \to -\theta$), mathematically expressed as transforming original parameters ;

$$(x, y, z, l, w, h, \theta) \to (x, -y, z, l, w, h, -\theta). \tag{5}$$

During vertical flipping, the z-coordinate is inverted ($z \to -z$) and the yaw angle is adjusted to $\pi - \theta$.This transformation is implemented in code by directly modifying the NumPy arrays of the bounding box tensors. Depth maps (gt_depth) and semantic maps (gt_semantic) follow the same flipping logic as images, maintaining spatial alignment with image pixels via dimension reversal. After flipping, samples are tagged to facilitate subsequent modules in tracing augmentation status, forming a a closed-loop pipeline encompassing initialization, condition judgment, multimodal flipping, and tagging.This ensures strict synchronization between augmented data and annotations. Consequently, the model is provided with diverse and pose-robust training samples.

Compared with conventional random flip, the RandomFlip module in this paper supports joint flipping of images, 3D boxes, depth maps, and semantic maps. Without introducing additional data, it significantly enhances the original model's generalization capability against viewpoint perturbations, laying a high-quality input foundation for subsequent feature enhancement and fusion.

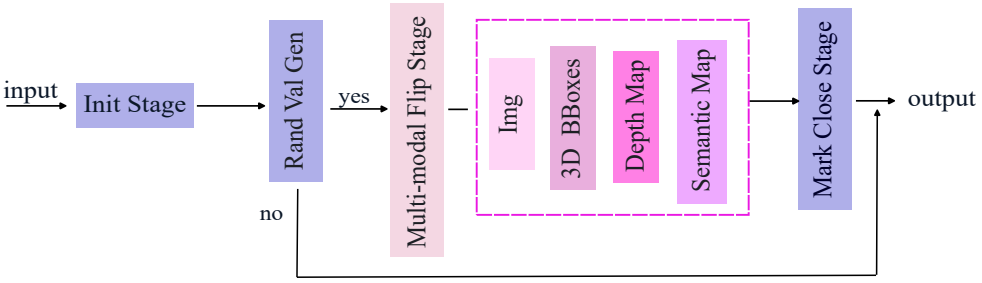


Fig. 2 Structure of the RandomFlip module: The module employs random value generation to determine whether to proceed to the multimodal flipping stage, where it flips images, 3D bounding boxes, depth maps, and semantic maps, and outputs the result after tagging to close the loop.

3.2 Multi-Scale Feature Fusion for BEV Detection

The BEV model performs detection by projecting multi-view image features onto a unified top-down 2D plane, and its performance fundamentally hinges on the quality of the generated BEV feature map. However, due to the inherent properties of perspective projection, objects in the scene exhibit significant scale variation on the image plane: nearby objects occupy a substantial number of pixels, containing rich texture details but relatively local semantic contexts; whereas distant objects may be represented by merely a few pixels, with highly abstract semantic information but lacking discernible details. HS-FPN directly addresses these issues by leveraging high-frequency perception and spatial dependency modeling: it enriches features of small objects through high-frequency responses and captures spatial correlations via cross-attention. This significantly enhances the discriminative power of BEV features for multi-scale objects, constituting a crucial upgrade for robust BEV detection.

As illustrated in FIGURE 3, HS-FPN enhances the traditional FPN by introducing two core components: the High-Frequency Perception module (HFP) and the Spatial Dependency Perception module (SDP), which are specifically designed to tackle the challenge of small target detection. The overall processing pipeline of HS-FPN can be formalized as follows. Given a set of channel-regularized feature maps C_i extracted from the backbone network, HS-FPN aims to generate an enhanced feature pyramid P_i. For each level, the computation of its output features integrates the operations of HFP and SDP:

$$P_i = \mathrm{Conv}_{3\times3}\left(F_{\mathrm{HFP}}\left(\mathbf{C}_i\right) + F_{\mathrm{SDP}}\left(\mathbf{C}_i,\ \mathbf{P}_{i+1}^{\uparrow}\right)\right). \tag{6}$$

The HFP module aims to enrich features for small objects by accentuating high-frequency components—such as edges and textures—that are crucial for their recognition.

As illustrated in FIGURE 4 of HS-FPN, the High-Frequency Perception module (HFP) first applies a high-pass filter to the input feature C_i: by controlling the frequency band via the hyperparameter α, it generates the high-frequency response F_i.

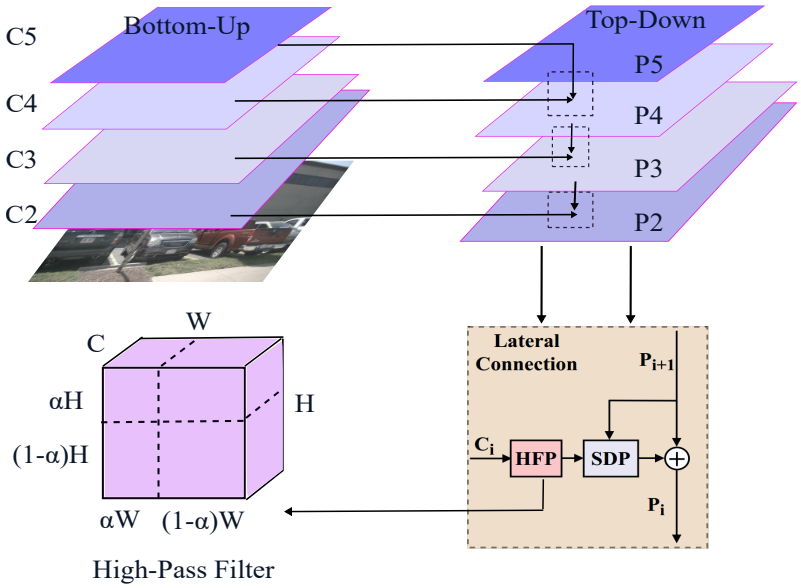


Fig. 3 Structure of the HS-FPN Module: The module fuses multi-scale features via lateral connections. At its core, it integrates the High-Frequency Perception module (HFP) to enhance high-frequency details of small targets and the Spatial Dependency Perception module (SDP) to model cross-scale pixel-level associations, achieving multi-scale feature alignment and enhanced discriminative power, thereby providing high-quality fused features for BEV detection.

Let $Z \in \mathbb{R}^{C\times H\times W}$ denote a binary mask filter:

$$Z^{(c)}(u,v) = \begin{cases} 0, & \text{if } u < \alpha H \text{ and } v < \alpha W \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

where (u, v) denote the coordinates in the frequency domain (after DCT transformation), and $\alpha \in [0, 1]$ controls the size of the filtered low-frequency region. Through this filter, we obtain the high-frequency response feature F_i

$$F_i = \mathcal{DCT}^{-1}\left(Z \odot \mathcal{DCT}(\mathcal{C}_i)\right). \tag{8}$$

This process filters out the uniform background and amplifies features of small targets. Subsequently, HFP splits into two paths: the Channel Path (CP) and the Spatial Path (SP). To identify channels rich in small target information, CP applies global average pooling (GAP) and global max pooling (GMP) to F_i, generating two 1D vectors:

$$\mathbf{f}_{\text{gap}} = \text{GAP}(F_i) \quad \mathbf{f}_{\text{gmp}} = \text{GMP}(F_i), \tag{9}$$

The channel attention weights are output after these vectors are processed by grouped convolution $\mathbf{u}_{\text{cp}} \in \mathbb{R}^{C\times1\times1}$

$$\mathbf{u}_{\text{cp}} = \sigma\left(\text{GroupConv}\left(\delta(\mathbf{f}_{\text{gap}})\right) + \text{GroupConv}\left(\delta(\mathbf{f}_{\text{gmp}})\right)\right), \tag{10}$$

where δ denotes the ReLU activation function, and σ represents the Sigmoid function.These weights are used to recalibrate the input feature channels.

The Spatial Path (SP) focuses on highlighting spatial regions likely containing small objects. It generates a spatial attention mask $U_{\text{sp}} \in \mathbb{R}^{1\times H\times W}$via a 1×1 convolution followed by a Sigmoid activation:

$$U_{\text{sp}} = \sigma\left(\text{Conv}_{1\times1}(F_i)\right). \tag{11}$$

This mask acts as a soft spatial selector for regions with strong high-frequency responses.

Finally, the outputs from the CP and SP are combined and refined to produce the enhanced feature map C_i^{HFP}

$$C_i^{\text{HFP}} = \text{Conv}_{3\times3}\left((\mathbf{u}_{\text{cp}} \cdot C_i) + (U_{\text{sp}} \odot C_i)\right). \tag{12}$$

The Spatial Dependency Perception module (SDP) addresses the spatial perception deficiency of FPN by modeling pixel-level associations across adjacent scales. For features : $\{C_i, P_{i+1}\}$ (P_{i+1}) is the upsampled high-level feature), SDP employs a pixel-wise cross-attention mechanism to learn spatial dependencies, computes attention weights based on feature similarity, achieves cross-scale alignment of small objects features, and reduces misalignments caused by recursive upsampling. SDP is applied to shallow-to-intermediate features, where small targets are most prominent

The HS-FPN design in HDO-BEV is intended to fuse two key feature maps from the image neck: the low-level 1024-channel and high-level 2048-channel feature maps. They are first unified via concatenation to unify channel dimensions, followed by entry into the self-attention-enhanced HFP module. A simplified HFP is applied to the low-level features, which mildly suppresses weak high-frequency noise while preserving spatial details; an enhanced HFP is applied to the high-level features, which undergoes cross-attention fusion (via the SDP module) and is upsampled to the low-level feature size via nearest-neighbor interpolation. The SDP performs pixel-wise cross-attention on the upsampled high-level features and low-level features, learning spatial associations to align cross-scale small target features. Finally, after output refinement, a final 512-channel BEV feature map is generated, providing high-quality input for subsequent view transformation and detection tasks.

Traditional FPN and its variants typically rely on fixed, top-down lateral connections for feature fusion; this unidirectional and static path easily leads to the dilution of fine-grained information in shallow high-resolution features during transmission. Compared with CustomFPN, which treats all features equally and fails to emphasize key high-frequency details for small targets, the HFP module of HS-FPN explicitly extracts high-frequency responses via an adaptive filter, enriching features of small targets. In our implementation, the HFP for high-level features focuses on local high-frequency

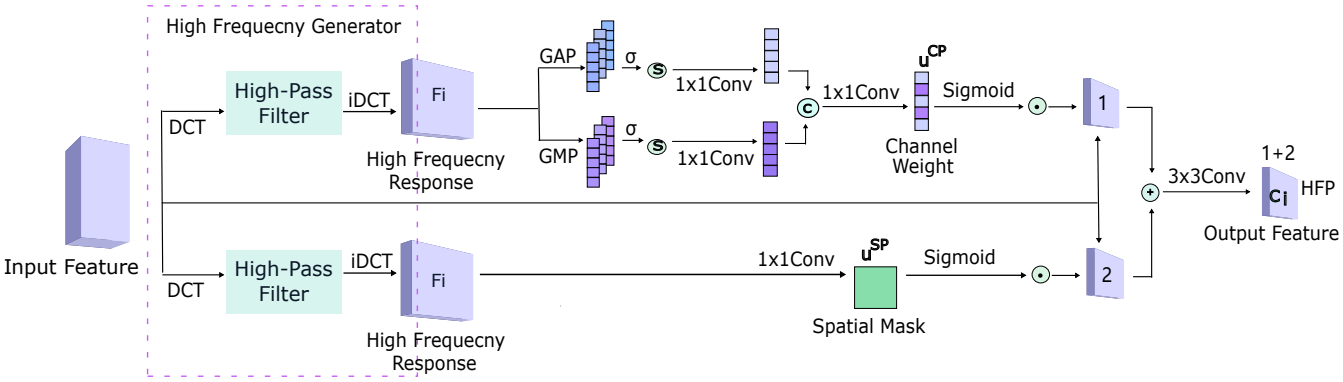


Fig. 4 The design philosophy of HFP consists of a high-frequency feature response module, a Channel Path (CP), and a Spatial Path (SP). Input features generate dual-branch high-frequency feature responses. The upper branch extracts channel-dimensional information via global average pooling (GAP) and global max pooling (GMP), generates channel weights, and achieves feature enhancement for the Channel Path. The lower branch generates a spatial mask via 1×1 convolution and Sigmoid activation, and achieves feature enhancement for the Spatial Path through pixel-wise summation ($\oplus$). The enhanced features from both branches are fused to output the target feature (Output Feature). Note: $\oplus$ denotes pixel-wise summation, σ represents the activation function (e.g., ReLU), channel homology (c) indicates shared channels, and other operations correspond to contextual semantics.

blocks, while the HFP for low-level features preserves global spatial contexts. Moreover, element-wise addition in CustomFPN causes cross-scale spatial misalignments of small targets, whereas the SDP module of HS-FPN learns pixel-level associations via cross-attention to correct such misalignments. Unlike the original HS-FPN, which processes 4 feature layers, our implementation simplifies to 2 layers, matching the image feature fusion neck output of HDO-BEV, reducing parameters while retaining core functionality. For image features corresponding to dense regions in BEV space, HS-FPN can adaptively adjust its weighting to retain and propagate more fine edge and texture information from high-resolution shallow networks, preventing these small yet critical features from being smoothed out during pooling. This enables the finally projected BEV feature map to achieves higher feature discriminability in dense regions as well, directly benefiting the 3D detection head in more accurate localization and classification of crowded or small targets.

3.3 Contextual Feature Enhancement in BEV Models

Context enhancement can improve a model's semantic understanding, helping it better understand inter-object relationships and spatial dependencies. In 3D detection, especially with pure visual inputs, strong context perception effectively enhances detection accuracy and robustness. Pure visual BEV models rely on multi-camera inputs with high-dimensional image features; if an enhancement module is parameter-heavy and computationally complex, it becomes an inference bottleneck, leading to excessive latency and making the overall model difficult to deploy on in-vehicle or edge devices.

Compared to complex self-attention mechanisms, ContextBlock is a lightweight feature enhancement module that optimizes spatial and semantic correlations in feature maps by aggregating global contextual information without introducing significant computational overhead:

$$z_i = F\left(x_i, \delta\left(\sum_{j=1}^{N_p} \alpha_j x_j\right)\right). \tag{13}$$

Architecturally streamlined, unlike traditional non-local blocks (NL) that model long-range dependencies via query-key-value matrix operations, ContextBlock replaces complex matrix computations with bottleneck transformations and aggregates context via global attention pooling. It integrates global context features to original features via element-wise summation, thereby injecting broader semantics while preserving local details and avoids the potential over-suppression of subtle features by multiplicative fusion.

This paper introduces the ContextBlock module into the SA-BEV model for the first time. Targeting the insufficient feature context in pure visual multi-view BEV perception, the module is inserted after the image feature neck to achieve global context enhancement of multi-view features. This application adapts a single-view lightweight context module to the multi-view BEV feature generation pipeline, enhancing the semantic consistency and spatial perception capability of BEV features through cross-view semantic correlation and long-range dependency modeling. As shown in FIGURE 5, ContextBlock serves as a lightweight feature enhancement module designed for this purpose. Its core design integrates the long-range dependency modeling capability of the Simplified Non-local Block (SNL) with the lightweight computational advantage of the Squeeze-and-Excitation (SE) Block. This synergy addresses the parameter redundancy in traditional non-local blocks and the insufficient long-range dependency capture in SE blocks.

The structure of ContextBlock comprises three components: (1) Global Attention Pooling, which aggregates features from all positions using the attention weights:

$$\alpha_j = \frac{e^{W_k x_j}}{\sum_m e^{W_k x_m}}, \tag{14}$$

(2) Bottleneck Transformation, which employs a two-layer structure of "1×1 convolution + Layer Normalization (LN) + ReLU + 1×1 convolution". This design compresses the parameter count from c^2 to $\frac{2c^2}{r}$ (default r=16 reduces parameters to 1/8 of the original SNL block) and simplify optimization through LN; (3) Broadcast Element-wise Addition, which fuses the transformed global context features with the original features via element-wise summation. The complete mathematical expression of the module is given by

$$z_i = x_i + W_{v2}\mathrm{ReLU}\left(\mathrm{LN}\left(W_{v1}\sum_{j=1}^{N_p} \alpha_j x_j\right)\right) \tag{15}$$

where

$$\delta(\cdot) = W_{v2}\mathrm{ReLU}\left(\mathrm{LN}\left(W_{v1}(\cdot)\right)\right)$$

denotes the bottleneck transformation.

ContextBlock achieves unified long-range dependency modeling and channel attention calibration at low cost via "global attention pooling + bottleneck transformation + additive fusion", becoming a general-purpose component for context enhancement in vision tasks.

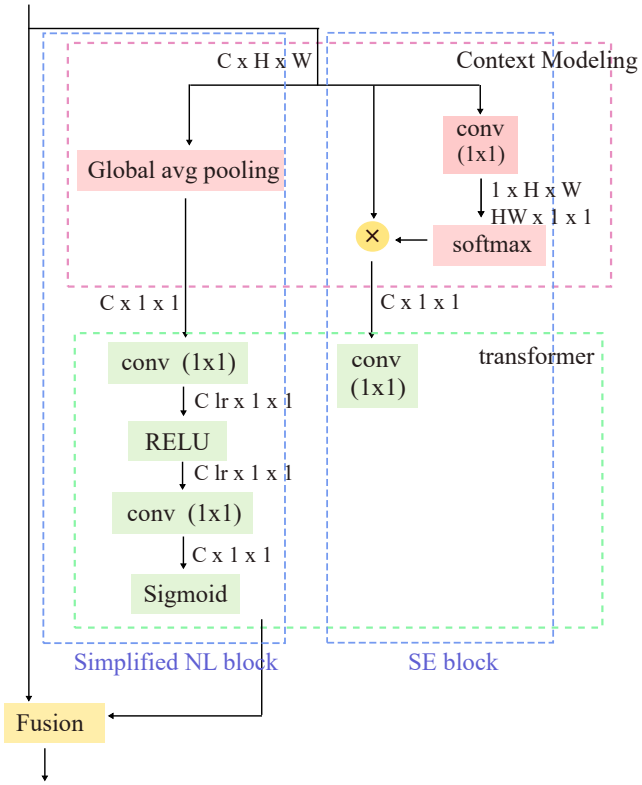


Fig. 5 As shown, the architecture of the ContextBlock module is presented, where C×H×W denotes an input feature map with channel number C, height H, and width W. In this architecture, the Context Modeling module generates contextual weights and weights the features; the Transformer module performs feature transformation.

Originally designed for single-view images, this work adapts the single-view lightweight ContextBlock into a multi-view context enhancer and integrates it into the BEV model. It takes features from the image feature neck (image feature fusion neck) as input, enhances their context, then passes the enhanced features to the view transformation module. For multi-view adaptation: (1) Channel Attention recalibrates weights of 512 channels in multi-view features to identify and emphasize semantically

important channels common across views and suppress noisy channels; (2) Global attention pooling aggregates global context of multi-view features, encodes it into a spatial memory vector via bottleneck transformation (two-layer 1×1 convolution + LN), and injects it into each local feature point via additive fusion, capturing long-range correlations between multi-view features. This design resolves the issues of local isolation and semantic fragmentation in multi-view features. Applying ContextBlock to the BEV model enhances context representation of single-view features, promotes semantic alignment of multi-view features, and yields more discriminative BEV features for subsequent view transformation.

4 EXPERIMENTS

In this section, we first outline the datasets and experimental setups utilized in our research. We then compare our method against state-of-the-art (SOTA) multi-view 3D object detection techniques reported in existing literature. To further substantiate the efficacy of our design, a series of ablation experiments and comprehensive analyses are performed, specifically targeting the roles of three core modules: RandomFlip, ContextBlock, and HS-FPN.

4.1 Experimental Setup

NuScenes-Mini is a lightweight subset of the large-scale autonomous driving benchmark dataset NuScenes, officially released to accommodate resource-constrained research and development scenarios. NuScenes-Mini is a lightweight subset of the original dataset. While preserving the core data distribution and sensor configuration, it significantly reduces the number of scenes, with a data volume approximately 1/5 to 1/3 of the full set. It retains key challenges such as multi-view coverage and multi-scale objects. Models trained on this subset typically exhibit a performance gap of less than 5% compared to those trained on the complete dataset, making it well-suited for preliminary research and rapid iteration of BEV models under computationally constrained conditions.

This study employs mAP and NDS as evaluation metrics. NDS is a comprehensive indicator that integrates five types of errors—translation, scale, orientation, velocity, and attribute—on the basis of mAP, providing a holistic reflection of detection performance. We adopt this dual-metric framework in our experiments to balance comprehensiveness and conciseness.

4.2 Results and Analysis

4.2.1 Comparative Experiments

Due to the limited scale of nuScenes-Mini, all compared methods exhibit significant performance degradation when contrasted with the full nuScenes benchmark results. On the nuScenes-Mini dataset, under a unified configuration with an input image resolution of 256 × 704 and ResNet-50 as the backbone network, we compare the proposed method (HDO-BEV) with representative approaches including BEVDet, BEVFormer, BEVDet4D, BEVDepth, SA-BEV and BEVHeight++. Experimental results 1 show

that HDO-BEV achieves excellent performance with 0.388 mAP and 0.472 NDS, demonstrating comprehensive leadership across all metrics: both its mAP and NDS are significantly higher than those of other compared methods, while it also exhibits balanced and outstanding performance on error-type metrics measuring spatial consistency (mATE, mASE, mAOE). These results fully validate the detection accuracy of HDO-BEV and highlight the advantages of our proposed enhancements.

Table 1 Comparison with previous state-of-the-art multi-view 3D detectors on the nuScenes-Mini dataset.

Method	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
BEVDet	0.339	0.414	0.760	0.291	0.572
BEVFormer	0.352	0.444	0.72	0.280	0.556
BEVDet4D	0.370	0.460	0.668	0.269	0.540
BEVDepth	0.371	0.470	0.663	0.265	0.499
SA-BEV	0.379	0.463	0.660	0.262	0.486
BEVHeight++	0.389	0.471	0.660	0.264	0.483
HDO-BEV	**0.388**	**0.472**	**0.660**	**0.260**	**0.480**

4.2.2 Visualization Analysis

We visualize the detection results of the images in FIGURES 6 and 7. Compared with SA-BEV, HDO-BEV enables more precise predictions with the assistance of semantically aware BEV features. FIGURE 6 illustrates the detection performance of SA-BEV on a street view from six viewpoints (CAM_BACK_LEFT to CAM_FRONT_RIGHT), with vehicles marked by yellow bounding boxes, pedestrians by blue ones, and the background containing elements such as buildings, traffic lights, and Starbucks. FIGURE 7 presents the detection results of HDO-BEV: under the same viewpoints, the bounding boxes of HDO-BEV align more accurately with target contours, exhibiting reduced shape distortion, clear straight edges of vehicle bodies, and no box distortion caused by potential viewpoint overlap in the original model. This effectively reduces category misclassification, yields a more orderly spatial distribution of multiple targets, avoids potential box overlap or hierarchical confusion in the original model, and enhances the algorithm's robustness.

The convergence behavior of the proposed method during training is further visualized in FIGURE 8, where the training and validation losses demonstrate stable descent across 34 epochs. The small gap between the two curves indicates negligible overfitting, validating the rationality of our training setup .

4.2.3 Results and Analysis of Ablation Experiments

To investigate the respective impacts of RandomFlip, ContextBlock, and HS-FPN, we conducted evaluations using SA-BEV as the baseline architecture. These outcomes

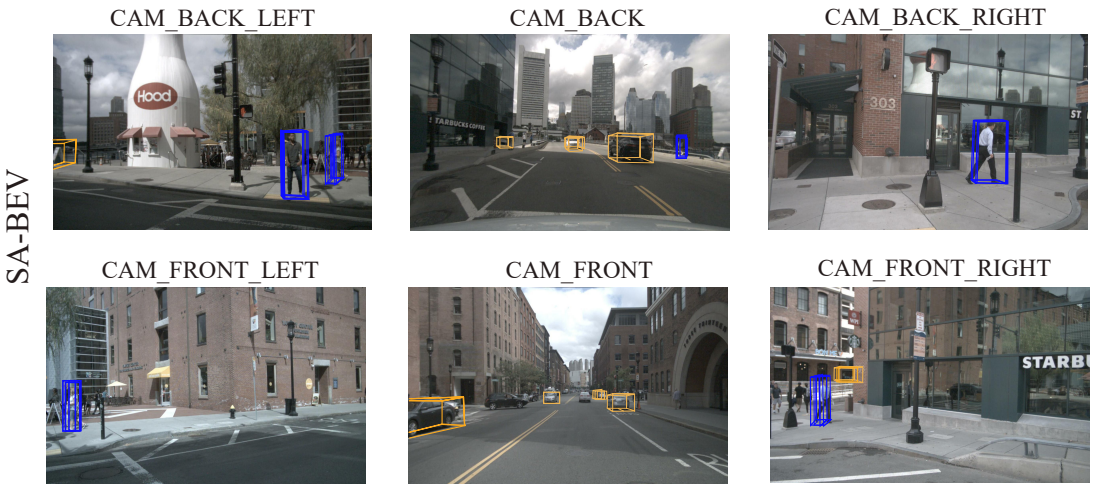


Fig. 6 Visualization of SA-BEV detection results.

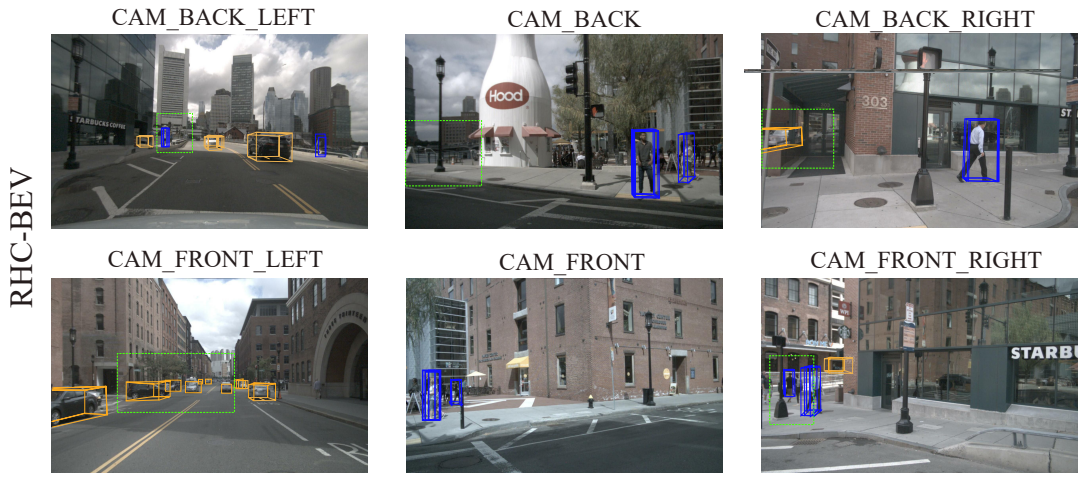


Fig. 7 Visualization of HDO-BEV detection results.

are summarized in the accompanying table. After applying RandomFlip(RF), the performance of mAP and NDS increased by 0.002 and 0.004, respectively. Upon incorporating ContextBlock on the basis of RF, the performance was further improved by 0.002 (mAP) and 0.002 (NDS). Finally, with the introduction of the HS-FPN module, the model's mAP and NDS metrics were enhanced by an additional 0.005 and 0.003, respectively. The complete model integrating all three modules achieved 0.388

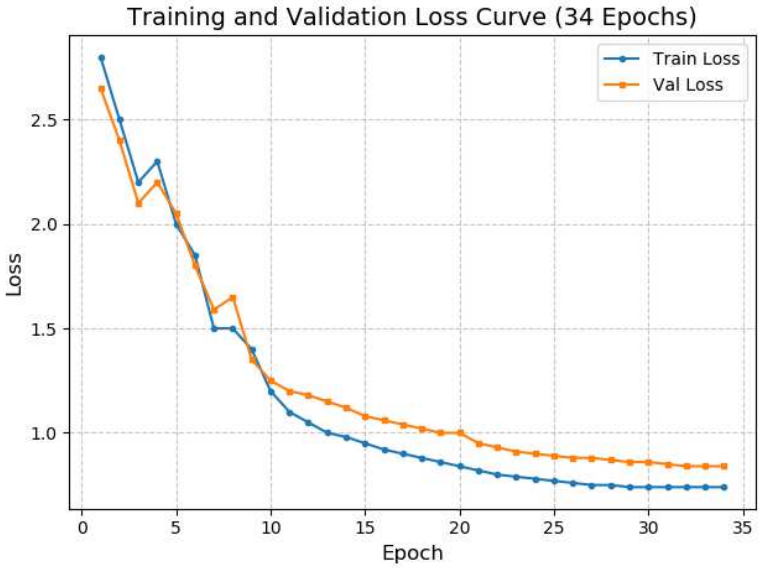


Fig. 8 Training and Validation Loss Curves over 34 Training Epochs.

mAP and 0.472 NDS. Compared to the baseline SA-BEV, this represents a cumulative improvement of 0.009 in mAP and 0.009 in NDS, validating the effectiveness of each module and the overall integration scheme.The results are shown in Table 2.

Table 2 Ablation study on components of the SA-BEV model on the nuScenes-mini dataset and the table presents the performance of mAP and NDS under different module combinations.

RandomFlip	ContextBlock	HS-FPN	mAP ↑	NDS ↑
			0.379	0.463
✓			0.381	0.467
✓	✓		0.383	0.469
✓	✓	✓	0.388	0.472
	✓		0.380	0.464
	✓	✓	0.384	0.470
		✓	0.382	0.467

4.2.4 Ablation Experiments on the RandomFlip

To validate the effectiveness of the RF module, an ablation study was conducted with the baseline model (without this module) as the control. The results are shown in the table: without the module, mAP = 0.379, NDS = 0.463, mATE = 0.660, mASE =

0.262, and mAOE = 0.486; after incorporating the module, mAP increased to 0.381 (+0.2%), NDS increased to 0.467 (+0.4%), mATE decreased to 0.655 (-0.8%), mASE decreased to 0.261 (-0.4%), and mAOE decreased to 0.485 (-0.2%). These results indicate that the module effectively enhances detection accuracy and reduces spatial errors by augmenting data diversity through viewpoint perturbation, serving as a key support for the model's robustness.The results are shown in Table 3.

Table 3 Ablation study on the RandomFlip module on the nuScenes-mini dataset and the table presents the performance of mAP, NDS, mATE, mASE, and mAOE under different module combinations.

RandomFlip	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
	0.379	0.463	0.660	0.262	0.486
✓	**0.381**	**0.467**	**0.655**	**0.261**	**0.485**

4.2.5 Ablation Experiments on the HS - FPN

The ablation study on the HS-FPN module validates the efficacy of multi-scale feature fusion. For the baseline model (without the module), mAP=0.379, NDS=0.463, mATE=0.660, mASE=0.262, and mAOE=0.486. After incorporating the module, mAP increased to 0.382 (+0.3%), NDS increased to 0.467 (+0.4%), mATE decreased to 0.655 (-0.8%), mASE decreased to 0.261 (-0.4%), and mAOE decreased to 0.483 (-0.6%). These results indicate that the module optimizes the feature representation of targets of different sizes through cross-scale feature enhancement and propagation, significantly improving detection accuracy and stability.The results are shown in Table 4.

Table 4 Ablation study on the HS-FPN module on the nuScenes-mini dataset and the table presents the performance of mAP, NDS, mATE, mASE, and mAOE under different module combinations.

HS-FPN	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
	0.379	0.463	0.660	0.262	0.486
✓	**0.382**	**0.467**	**0.655**	**0.261**	**0.483**

4.2.6 Ablation Experiments on the ContextBlock

The ablation study on the ContextBlock module focuses on its context information extraction capability. For the baseline model (without the module), mAP=0.379, NDS=0.463, mATE=0.660, mASE=0.262, and mAOE=0.486. After incorporating the

module, mAP increased to 0.381 (+0.2%), NDS increased to 0.467 (+0.4%), and mATE decreased to 0.658 (-0.3%). Although mASE (0.264) and mAOE (0.487) slightly increased, the core accuracy metrics showed significant improvement. These results indicate that the module optimizes semantic associations through global context modeling, compensates for the defect of single-view feature fragmentation, and is crucial for improving detection consistency.The results are shown in Table 5.

Table 5 Ablation study on the HS-FPN module on the nuScenes-mini dataset and the table presents the performance of mAP, NDS, mATE, mASE, and mAOE under different module combinations.

ContextBlock	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
	0.379	0.463	0.660	0.262	0.486
✓	**0.381**	**0.467**	**0.658**	**0.264**	**0.487**

5 Conclusions

This paper proposed HDO-BEV, an enhanced architecture built upon SA-BEV, which improved pure-vision BEV 3D detection through systematic upgrades in data augmentation, feature extraction, and multi-scale fusion. First, the RandomFlip module jointly flipped multi-view images and annotations to enhance robustness to viewpoint changes. Second, a ContextBlock module aggregated cross-view contexts via channel attention and global pooling, mitigating feature fragmentation. Third, HS-FPN replaced traditional FPN, using high-frequency perception and spatial dependency modules to refine small-object features and align cross-scale representations. Experiments on the nuScenes-Mini dataset showed that integrating all three modules raised mAP to 0.388 and NDS to 0.472, a cumulative gain of 0.9% over the baseline. Ablation studies confirmed the individual contributions of each component. Comparative results demonstrated advantages over methods such as BEVDet and BEVFormer. Limitations included evaluation on a reduced dataset and increased computational cost due to multi-module fusion. Future work would focus on model lightweighting and efficiency-accuracy trade-offs.

Declarations

- **Funding**: This work received support from National Natural Science Foundation of China (No. U22A2054) and the Key-Area Research and Development Program of Guangdong Province (No.2022B0701180001).
- **Conflict of interest**:The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate**:Not applicable.
- **Consent for publication**:Not applicable.

- **Data availability**:The datasets analysed during the current study are publicly available from the nuScenes dataset.
- **Materials availability**: Not applicable.
- **Code availability**: The code is available from the corresponding author on reasonable request.
- **Author contribution**:
  - **AiTong Mao**: Conceptualization, Methodology, Data curation, Software, Writing - original draft, Writing - review & editing.
  - **BeiHai Tan**: Funding acquisition, Writing - review & editing.
  - **Rong Yu**: Funding acquisition, Writing - review & editing.
  - **Guoliang Cheng**: Writing - review & editing.
  - **Xiaoxu Wang**: Writing - review & editing.
  - **Wuchang Zhong**: Writing - review & editing.

References

[1] Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021) https://doi.org/10.48550/arXiv.2112.11790

[2] Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images, 3262–3272 (2023) https://doi.org/10.1109/ICCV51070.2023.00302

[3] Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, 194–210 (2020) https://doi.org/10.1007/978-3-030-58568-6_12

[4] Wang, H., Chen, X., Yuan, Q., Liu, P.: A review of 3d object detection based on autonomous driving. The Visual Computer **41**(3), 1757–1775 (2025) https://doi.org/10.1007/s00371-024-03480-6

[5] Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints, 734–750 (2018) https://doi.org/10.1007/978-3-030-01264-9_45

[6] Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection, 4490–4499 (2018) https://doi.org/10.1109/CVPR.2018.00472

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[8] Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., Zhan, W.: Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443 (2022) https://doi.org/10.48550/arXiv.2210.02443

[9] Feng, Q., Zhao, C., Liu, P., Zhang, Z., Jin, Y., Tian, W.: Lst-bev: Generating a long-term spatial–temporal bird's-eye-view feature for multi-view 3d object detection. Sensors **25**(13), 4040 (2025) https://doi.org/10.3390/s25134040

[10] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection, 6569–6578 (2019) https://doi.org/10.1109/ICCV.2019.00667

[11] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019) https://doi.org/10.48550/arXiv.1904.07850

[12] Li, P., Zhao, H., Liu, P., Cao, F.: Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving, 644–660 (2020) https://doi.org/10.1007/978-3-030-58580-8_38 . Springer

[13] Qin, Z., Wang, J., Lu, Y.: Monogrnet: A geometric reasoning network for monocular 3d object localization **33**(01), 8851–8858 (2019) https://doi.org/10.1609/aaai.v33i01.33018851

[14] Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., Alvarez, J.M.: Fb-bev: Bev representation from forward-backward view transformations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6919–6928 (2023). https://doi.org/10.1109/ICCV51070.2023.00637

[15] Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection, 8555–8564 (2021) https://doi.org/10.1109/CVPR46437.2021.00845

[16] Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection, 4721–4730 (2021) https://doi.org/10.1109/CVPR46437.2021.00469

[17] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022) https://doi.org/10.1145/3505244

[18] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) https://doi.org/10.1109/TPAMI.2024.3515454

[19] Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11784–11793 (2021). https://doi.org/10.1109/CVPR46437.2021.01161

[20] Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d

object detection. arXiv preprint arXiv:2203.17054 (2022) https://doi.org/10.48550/arXiv.2203.17054

[21] Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks, 11138–11147 (2020) https://doi.org/10.1109/CVPR42600.2020.01115

[22] Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. arXiv preprint arXiv:1811.08188 (2018) https://doi.org/10.48550/arXiv.1811.08188

[23] Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8445–8453 (2019). https://doi.org/10.1109/CVPR.2019.00864

[24] Qiu, Q., Cheng, X., Calderbank, R., Sapiro, G.: Dcfnet: Deep neural network with decomposed convolutional filters. arXiv preprint arXiv:1802.04145 (2018) https://doi.org/10.48550/arXiv.1802.04145

[25] Yang, L., Zhang, X., Li, J., Wang, L., Zhu, M., Zhu, L.: Lite-fpn for keypoint-based monocular 3d object detection. Knowledge-Based Systems **271**, 110517 (2023) https://doi.org/10.1016/j.knosys.2023.110517

[26] Li, Y., Huang, B., Chen, Z., Cui, Y., Liang, F., Shen, M., Liu, F., Xie, E., Sheng, L., Ouyang, W., *et al.*: Fast-bev: A fast and strong bird's-eye view perception baseline. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 8665–8679 (2024) https://doi.org/10.1109/TPAMI.2024.3414835

[27] Xie, E., Yu, Z., Zhou, D., Philion, J., Anandkumar, A., Fidler, S., Luo, P., Alvarez, J.M.: M^2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. arXiv preprint arXiv:2204.05088 (2022) https://doi.org/10.48550/arXiv.2204.05088

[28] Shi, Z., Hu, J., Ren, J., Ye, H., Yuan, X., Ouyang, Y., He, J., Ji, B., Guo, J.: Hs-fpn: High frequency and spatial perception fpn for tiny object detection **39**(7), 6896–6904 (2025) https://doi.org/10.1609/aaai.v39i7.32740

[29] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving, 17853–17862 (2023) https://doi.org/10.1109/CVPR52729.2023.01712

[30] Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters **5**(3), 4867–4873 (2020) https://doi.org/10.1109/LRA.2020.3004325

[31] Yao, J., Lai, Y., Kou, H., Wu, T., Liu, R.: Qe-bev: Query evolution for bird's eye view object detection in varied contexts, 2927–2935 (2024) https://doi.org/10.1145/3664647.3680807

[32] Fang, S., Wang, Z., Zhong, Y., Ge, J., Chen, S.: Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1368–1378 (2023). https://doi.org/10.1109/CVPR52729.2023.00138

[33] Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection, 531–548 (2022) https://doi.org/10.1007/978-3-031-19812-0_31 . Springer

[34] Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection, 3621–3631 (2023) https://doi.org/10.1109/ICCV51070.2023.00335

[35] Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., Zhu, X.: Vision-centric bev perception: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) https://doi.org/10.1109/TPAMI.2024.3449912

[36] Zhang, G., Zhang, W., Li, W., Wang, L., Cui, H.: A dynamic attention mechanism for object detection in road or strip environments. The Visual Computer **41**(6), 4171–4181 (2025) https://doi.org/10.1007/s00371-024-03653-3