https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

A Lightweight BEV Perception Optimization Framework for Real-Time 3D Object Detection in Occluded Scenes

Yi Zhang, Anhua Zhou, Jun Li
Chongqing Vocational and Technical University of Mechatronics, Chongqing, 402760, China
Correspondence: Yi Zhang (zy17783051215@163.com); Jun Li (cqleejun@163.com)

Abstract

Real-time 3D object detection in occluded scenes is a core challenge for pure-vision autonomous driving perception. Existing dense Bird's Eye View (BEV) detection methods suffer from redundant backbone parameters, inflexible fixed sampling strategies, and coarse-grained temporal modeling, making it difficult to simultaneously satisfy detection accuracy and real-time latency requirements on embedded vehicle platforms. This paper proposes a lightweight BEV perception optimization framework for occluded scenes. An occlusion-aware adaptive voxel feature sampling module dynamically switches between fast ray projection and deformable attention paths according to scene complexity. A temporal grouping fusion module based on Res2Net principles performs grouped cross-frame fusion of consecutive BEV feature maps without introducing additional learnable parameters. A two-stage LiDAR-to-camera knowledge distillation scheme with a geometric compensation module transfers depth geometry knowledge during training while incurring zero inference overhead. Experiments on the nuScenes dataset demonstrate that the proposed method achieves 38.7% mAP and 51.3% NDS under the lightweight configuration, with an inference latency of 38.2 ms and 26.2 FPS on NVIDIA Tesla T4, ranking highest in NDS among comparable methods.

Keywords: bird's eye view perception; 3D object detection; occluded scenes; lightweight network; adaptive feature sampling; temporal fusion; knowledge distillation

1 Introduction

The large-scale advancement of autonomous driving technology has put forward strict real-time and robustness requirements for in vehicle perception systems. In complex urban traffic scenes, vehicles, pedestrians, non-motorized vehicles and other targets frequently obstruct each other, resulting in a large amount of information loss in single frame visual signals, which seriously restricts the accuracy of 3D object detection. The pure visual camera system has low cost, high information density, and strong commercial implementation value. However, there are inherent shortcomings in occlusion processing and depth ambiguity issues. How to effectively deal with occlusion scenes under the constraint of lightweight deployment remains the core challenge in the field of pure visual 3D perception. The release of the nuScenes dataset provides a systematic evaluation benchmark for multi camera 3D object detection, promoting the development of a technology roadmap that uses bird's-eye view (BEV) as a unified representation space (Caesar et al., 2020). BEV representation projects image features from different perspectives onto a shared overhead plane and completes detection on that plane, naturally adapting to spatial relationship modeling of multi camera systems, and has become the mainstream framework for pure visual 3D perception. Lift, Splat, Shoot (LSS) was the first system to propose a three-stage pipeline that predicts the probability distribution of depth pixel by pixel, spreads pixel features along the depth direction to a three-dimensional point cloud, and then aggregates them through BEV pooling, laying

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

the core paradigm for subsequent dense BEV methods (Philion and Fidler, 2020). BEVDet introduces BEV spatial data augmentation and BEV-NMS post-processing based on LSS, and validates the competitiveness of the dense BEV framework on the nuScenes standard evaluation (Huang et al., 2021). BEVDepth utilizes LiDAR point clouds to apply pixel by pixel supervision to the image depth estimation network, compensating for BEV feature projection errors caused by monocular depth ambiguity, and improving mAP compared to BEVDet on the nuScenes validation set (Li et al., 2023b).

In terms of temporal modeling, BEVFormer first extends deformable spatial attention to the temporal dimension, and uses temporal BEV self attention to transfer historical frame features to the current frame, achieving an NDS of 56.9% on the nuScenes validation set, which significantly improves the stability of detecting moving and occluded targets (Li et al., 2022). SOLOFusion designed a dual stream fusion strategy for short-term high-resolution and long-term low resolution features (Park et al., 2023). BEVNeXt improved the NDS in the nuScenes test set through three improvements: CRF modulation depth estimation, Res2Fusion timing fusion module, and two-stage target decoder, demonstrating the advantages of dense BEV framework in depth estimation and target localization (Z. Li et al., 2024). The lightweight requirement for real-time deployment is currently the main bottleneck in intensive BEV research. The backbone network parameters of mainstream dense BEV methods are generally in the hundreds of megabytes range, and online viewpoint transformation introduces a large amount of redundant calculations, making it difficult to meet the real-time constraints of the vehicle end in terms of inference delay. Fast BEV simplifies the viewpoint transformation in the inference stage into a lookup table operation by pre computing the BEV projection index offline, significantly reducing inference latency (Y. Li et al., 2024). However, the fixed sampling strategy suffers from accuracy degradation in complex occlusion scenes. In terms of multimodal knowledge distillation, MMDistill used a LiDAR teacher model to perform two-stage soft target distillation on pure visual student models in BEV space, achieving a 4.1% improvement in mAP and a 4.6% improvement in NDS on the nuScenes benchmark (Jiao et al., 2024). There was no additional computational cost in the inference stage, proving the feasibility of using multimodal information to enhance pure visual models in the training stage.

Based on existing research, there are three core issues in the real-time deployment of dense BEV detection in occluded scenes: existing temporal fusion methods generally use full frame feature concatenation, and the computational cost increases linearly with the number of historical frames. The ability to model fine-grained inter frame dependencies of occluded targets is insufficient; The fixed sampling strategy cannot dynamically balance computational efficiency and occlusion detection accuracy in different scene complexities; The coexistence of redundant backbone network parameters and online viewpoint transformation overhead constrains the real-time deployment of embedded platforms.

This paper proposes a lightweight BEV perception optimization framework for occluded scenes to address the above three issues. The main contributions are as follows:

(1) Design an occlusion aware BEV enhancement module that utilizes an adaptive voxel feature sampling strategy driven by scene complexity to call for efficient ray projection in simple scenes and activate a deformable attention mechanism in complex occlusion scenes, thereby improving the 3D detection accuracy of occluded areas.

(2) A temporal grouping fusion module based on the Res2Net principle is constructed to perform group cross fusion on consecutive frames of BEV features, enhancing the temporal

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

modeling ability of moving and occluded targets while maintaining the geometric consistency of BEV space, without introducing additional learnable parameters.

(3) The introduction of LiDAR teacher model for BEV spatial knowledge distillation, combined with geometric compensation module to guide lightweight pure visual student model to obtain more accurate geometric depth representation, relies only on camera input in the inference stage, and meets the requirements of low-cost real-time deployment.

2 Related work

2.1 Dense 3D Object Detection Based on BEV

Dense BEV methods have evolved along two directions: depth estimation enhancement and temporal modeling. The LSS framework's single-frame depth prediction relies entirely on image appearance cues, yielding significant errors in occluded and distant regions, which limits BEV projection quality. BEVDet introduces BEV-space augmentation and BEV-NMS on top of LSS, while BEVDepth uses LiDAR point clouds to supervise pixel-wise depth and improves mAP over BEVDet. FCOS3D establishes monocular 3D detection on a fully convolutional head by predicting pixel-wise center depth, 3D size, and orientation, demonstrating that visual features inherently carry 3D positional cues (Wang et al., 2021); this idea was later absorbed for supervising depth estimation networks. To leverage multi-frame information, BEVDet4D introduces inter-frame BEV feature alignment via self-motion-compensated concatenation, reaching 45.7% NDS on the nuScenes validation set (Huang and Huang, 2022). However, the concatenation strategy cannot selectively transmit historical information for occluded regions, motivating subsequent attention-based extensions such as BEVFormer and SOLOFusion. BEVNeXt further improves NDS through CRF-modulated depth, Res2Fusion temporal fusion, and a two-stage decoder.

2.2 3D Object Detection Based on Sparse Query

Sparse query methods replace explicit BEV feature maps with learnable target queries that decode 3D attributes directly from 2D image features. DETR3D projects 3D reference points onto multi-view 2D feature maps and iteratively updates queries via bilinear interpolation, but its single-point sampling limits the receptive field for occluded targets (Wang et al., 2022). PETR introduces 3D position encoding that injects camera-frustum coordinates into image features, enabling 2D features to be inherently 3D-aware and aggregating multi-view information through global cross-attention (Liu et al., 2022). SparseBEV further introduces adaptive sparse convolution over foreground voxels, reaching 67.5% NDS on the nuScenes validation set, though its performance in dense occlusion remains sensitive to foreground voxel recognition and training-data distribution (H. Liu et al., 2023).

2.3 Temporal Fusion and Occlusion Perception

Temporal fusion accumulates historical BEV features to supply complementary evidence for occluded targets. BEVFormer selectively transfers historical features via temporal BEV self-attention, while BEVDet4D adopts lightweight concatenation of adjacent frames. Neither systematically addresses the linear growth of computational cost with frame count or models fine-grained inter-frame dependencies for occluded objects. Res2Net introduces a multi-granularity hierarchical structure within a single residual block by splitting feature maps into groups and cascading residual enhancement (Gao et al., 2021), enabling multi-scale context perception in a single forward pass and providing an architectural reference for multi-branch temporal fusion. MambaBEV points out that convolutional and deformable-attention-based temporal fusion incur high cost, and proposes a Mamba2-based detector that reaches 51.7% NDS on nuScenes (You et al.,

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

2026), offering an alternative efficiency-oriented path.

2.4 Lightweight BEV Perception

Lightweight BEV research focuses on backbone compression and feature sampling optimization. EdgeNeXt replaces standard self-attention with a split depthwise transposed attention module, combining convolutional locality with Transformer global modeling at extremely low parameter cost, making it a viable lightweight backbone for BEV tasks (Maaz et al., 2023). DeployFusion integrates deformable convolution with EdgeNeXt within the BEVFusion framework and adopts a parameter self-adjusting residual fusion architecture (Z. Liu et al., 2023), substantially reducing FLOPs and parameters in multimodal BEV detection (Huang et al., 2024). For sampling, LightOcc proposes global spatial sampling and Spatial-to-Channel dimension transformation (Zhang et al., 2026), supplementing height information into BEV via channel encoding at far lower cost than voxel-based methods, while outperforming similar lightweight methods on Occ3D-nuScenes.

2.5 Multimodal Knowledge Distillation

Multimodal knowledge distillation enhances pure-vision perception at zero inference cost by using a LiDAR teacher to guide a camera student during training. Early cross-modal distillation aligned features in 2D space or at the detection head, but the spatial-resolution and semantic-granularity gap between point clouds and image features introduced significant domain shift (Zhou et al., 2023). BEV-space alignment is a natural solution since LiDAR and camera BEV features share the same coordinate system. MMDistill proposes a two-stage BEV distillation framework: global soft-target distillation aligns the overall BEV representation, followed by a geometric compensation module — comprising a dense feature extractor, context-aware submodule, and depth predictor — that targets geometrically sensitive regions where the teacher-student gap is largest, particularly compensating for depth-geometry weaknesses of the camera model.

3 Methods

3.1 Overall Framework

The lightweight BEV perception optimization framework proposed in this article takes multi camera panoramic images as input, with the dual goal of reducing 3D target missed detection rate and inference delay in occluded scenes. It integrates lightweight feature extraction, occlusion perception adaptive sampling, temporal grouping fusion, and multimodal knowledge distillation into a unified end-to-end architecture. The overall process of the framework is shown in Figure 1.

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

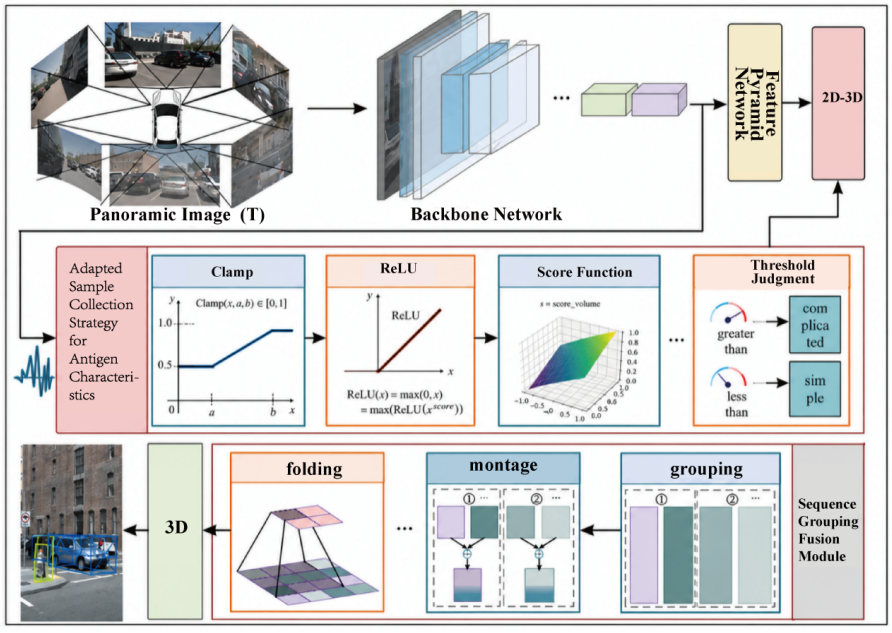

Figure 1 Overall framework of 3D object detection

The panoramic image is first subjected to a lightweight backbone network to extract multi-scale 2D features, which are then aggregated by a feature pyramid network and sent to two parallel branches (Lin et al., 2017). The occlusion aware BEV enhancement module is responsible for using an adaptive sampling strategy to elevate multi perspective perspective features to the BEV plane and generate the current frame BEV feature map. The scene complexity evaluation submodule synchronously calculates the scene complexity score of the current frame as the dynamic switching signal for the adaptive sampling strategy. The current frame BEV features and the historical frame BEV features are jointly sent to the temporal grouping fusion module to complete fine-grained inter frame dependency capture and output temporal fusion BEV features. Finally, the 3D detection head decodes and outputs the target category, position, size, and velocity. During the training phase, the LiDAR teacher model synchronously processes point cloud data, applies soft target distillation supervision to the camera student model in BEV space, and focuses on compensating for the camera model's shortcomings in depth geometric expression through geometric compensation modules. During the inference phase, only the camera branch is retained, and LiDAR sensors and teacher models are not deployed. The overall inference link is lightweight and embeddable.

3.2 Lightweight Feature Extraction Module

3.2.1 Lightweight Backbone Network

The mainstream dense BEV methods often use ResNet or VoVNet as the backbone, with the former having parameter sizes ranging from 25M to 100M (He et al., 2016). The feature extraction stage during online inference accounts for 40% to 60% of the total delay, making it difficult to adapt to the computational constraints of embedded platforms. This article uses EdgeNeXt as the basic backbone, with the Split Deep Transposed Attention (SDTA) module as the core. The feature channels are grouped and processed separately using deep convolution and transposed self-attention, achieving a balance between local receptive fields and global modeling. The overall parameter count

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

is about one-third of ResNet-50, and the inference floating-point operation is significantly reduced.

To enhance the feature perception ability of lightweight backbone for irregularly shaped targets in occluded scenes, a deformable convolutional layer is added after the SDTA module. Standard convolution features on a fixed rectangular grid and has limited ability to fit geometric deformations of occluded boundaries and irregular target contours. Deformable convolution learns the two-dimensional offset of each sampling point, enabling the convolution kernel to adaptively cover the actual shape area of the target, and has significant advantages in extracting local features of occluded targets. The multi-scale features output by the backbone are aggregated by a feature pyramid network to generate a multi-resolution feature map group for subsequent BEV projection.

3.2.2 Lightweight Space Embedding

During the construction process of BEV features, the three-dimensional space is compressed to the top-down plane, resulting in a significant loss of spatial information in the height dimension, which has a negative impact on the height judgment of occluded area targets. To supplement the height spatial information of BEV features at a lower computational cost, this paper introduces a lightweight spatial embedding mechanism, as shown in Figure 2.

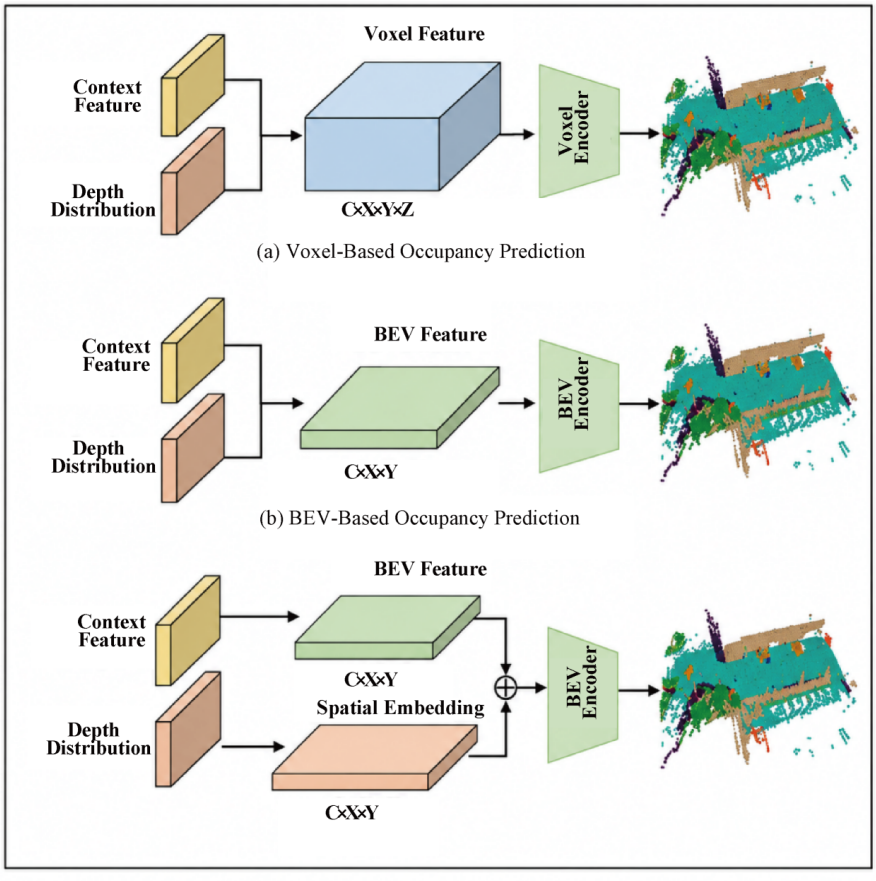

Figure 2 Comparison of Spatial Embedding Mechanisms in Lightweight BEV Perception Framework

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

The global spatial sampling stage starts from the multi view depth distribution, and globally weights and aggregates the depth probability distribution of each frame image in the spatial dimension to generate a single channel occupancy representation. This operation is implemented through vectorization matrix operation, and the computational cost is much lower than that of standard 3D convolution. The Spatial to Channel dimension transformation stage rearranges the spatial height dimension to the channel dimension, encodes the spatial distribution in the height direction into the channel representation of the BEV feature map, and enables subsequent 2D convolution operations to implicitly perceive height information in the channel dimension. Both stages of operation are achieved through lightweight convolution, with an overall parameter increase of no more than 5% of the backbone network, but the improvement in the accuracy of estimating the height coordinates of occluded targets is significant.

3.3 Occlusion-Aware BEV Enhancement Module

3.3.1 Scene complexity assessment

There is a fundamental difference in the requirements for sampling accuracy between occluded scenes and open scenes. In open scenes, targets are sparse and occlusion is minimal, and fast ray projection is sufficient to obtain accurate BEV features; In densely occluded scenes, there is visual boundary aliasing between occluded targets, and fixed sampling can easily misclassify occluded object features into the BEV representation of the occluded target. Therefore, a deformable attention mechanism is needed to provide fine-grained spatial perception capabilities (Zhu et al., 2021).

The scene complexity scoring module takes the multi view BEV feature map of the current frame as input, and the calculation steps are as follows. Firstly, perform a Clamp operation on the BEV feature map to limit the feature values to a reasonable range and eliminate the interference of depth estimation outliers; Subsequently, ReLU activation and maximum normalization were used to unify the feature amplitudes to the range of 0 to 1, resulting in a normalized feature map Fnorm. The global intensity means μ and local texture complexity standard deviation σ are calculated as follows:

$$\mu = \frac{1}{HW}\sum_{i,j} F_{norm}(i,j) \tag{1}$$

$$\sigma = \sqrt{\frac{1}{HW}\sum_{i,j}(F_{norm}(i,j) - \mu)^2} \tag{2}$$

The comprehensive complexity score s is obtained by weighted fusion:

$$s = \alpha \cdot \mu + (1-\alpha) \cdot \sigma \tag{3}$$

among them, α is a hyperparameter that controls the weight ratio between global intensity and local texture complexity. Compare s with the preset threshold τ, and when $s<\tau$, determine it as a simple scene and activate the fast ray projection path; When $s \geq \tau$, it is judged as occluding complex scenes and the deformable attention path is activated.

3.3.2 Adaptive voxel feature sampling

The fast ray projection path is based on pre computed voxel and pixel mapping indices, and only performs table lookup and bilinear interpolation operations during the inference phase. For the voxel with coordinates $(u,\ v)$ on the BEV plane, the corresponding k-th camera image coordinate $(x_k,\ y_k)$ is calculated from the camera projection matrix P_k:

$$[x_k, y_k, d_k]^T = P_k \cdot [u, v, z_c, 1]^T \tag{4}$$

where z_c is the preset sampling height, and (x_k,y_k) is the coordinate position sampled from the

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences DISCUSSIONS

image feature map. All P_k and sampling coordinates of fast ray projection are calculated and stored offline before model deployment, without the need for repeated matrix operations during inference.

The deformable attention path predicts K offset sampling points for each voxel on the image feature map, with each sampling point having a corresponding attention weight. The BEV voxel feature fbev $(u,\ v)$ is obtained by weighted summation:

$$f_{\text{bev}}(u,v) = \sum_{k=1}^{N_{\text{cam}}} \sum_{j=1}^{K} w_{kj} \cdot \mathbf{F_k}(x_k + \Delta x_{kj}, y_k + \Delta y_{kj}) \quad (5)$$

among them, w_{kj} is the attention weight of the jth sampling point of the kth camera, $\Delta x_{kj}$and Δy_{kj} are the predicted sampling offsets, and $\mathbf{F_k}$ is the 2D feature map of the kth camera. The deformable attention path actively avoids the interference of occluded features near the occluded boundary through offset sampling, resulting in more accurate BEV feature extraction for occluded targets. However, this comes at the cost of introducing additional offset prediction networks and multi-point sampling computational overhead. Therefore, it is only activated when the scene complexity exceeds the threshold.

3.4 Time series grouping fusion module

3.4.1 Motivation for module design

In densely occluded scenes, it is difficult to obtain complete 3D information of occluded targets in a single frame. Relying on historical frame BEV features to supplement the occluded areas from different temporal perspectives is an effective means to improve detection robustness. Directly concatenating historical frame features through channels will disrupt the two-dimensional geometric structure of BEV space, and the fusion information density increases linearly with the number of concatenated frames, resulting in lower parameter efficiency. Simple channel by channel averaging will smooth out important motion difference information between different frames, which is not conducive to speed estimation of fast-moving targets. The temporal grouping fusion module is based on the Res2Net multi branch hierarchical feature fusion idea, which captures fine-grained temporal dependencies between consecutive frames with a fixed computational cost while maintaining the geometric structure of the BEV feature space.

3.4.2 Module Structure

The temporal grouping fusion module processes the BEV features B_{t-3}, B_{t-2},B_{t-1}, and B_t of four consecutive frames, where $B_t \in$ RC × H × W represents the BEV features of the current frame, and C, H, and W are the number of channels, height, and width, respectively.

Divide the four frame features into two groups based on their adjacent frame relationships:

$$G_1 = [B_{t-2}, B_{t-3}], G_2 = [B_t, B_{t-1}] \quad (6)$$

Two sets of features are concatenated in the channel dimension and then reduced in dimensionality through a shared 1x1 convolution $K^{1\times1}$:

$$B_1' = K^{1\times1}(G_1), B_2' = K^{1\times1}(G_2) \quad (7)$$

Apply 3×3 convolution $K^{3\times3}$ to B2 'to enhance the local spatial receptive field, resulting in B_2^*:

$$B_2^* = K^{3\times3}(B_2') \quad (8)$$

After residual addition of B_1' and B_2^*, B_1^* is obtained by 3×3 convolution. The residual connection passes the cross frame timing signal between groups, allowing the features of earlier frames to directly affect the feature representation of newer frames as follows:

$$B_1^* = K^{3\times3}(B_1' + B_2^*) \quad (9)$$

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

The final temporal fusion BEV feature B is obtained by concatenating two sets of outputs in the channel dimension and convolving them 1×1:

$$\tilde{B} = K^{1\times1}([B_1^*, B_2^*]) \tag{10}$$

The entire module only introduces parameters for two $K^{1\times1}$ convolutions and two $K^{3\times3}$ convolutions, with three $K^{1\times1}$ weights shared between the two groups and minimal additional parameters. Residual connections ensure the geometric consistency of the BEV feature space, allowing subsequent detection heads to maintain consistent behavior on fused features with single frame features, resulting in stable training convergence.

3.5 Multimodal Knowledge Distillation

3.5.1 Overall Distillation Framework

The multimodal knowledge distillation framework during the training phase is shown in Figure 3, consisting of a LiDAR teacher model and a camera student model.

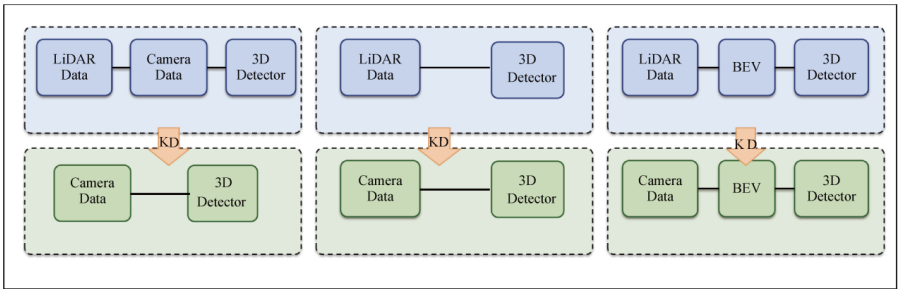

Figure 3 Multimodal BEV Distillation Framework

The LiDAR teacher model takes point cloud data as input, extracts 3D features through voxelization encoding and sparse convolution backbone, and generates high-precision LiDAR BEV feature maps FT ∈ RC × H × W through BEV pooling. The camera student model takes panoramic images as input and generates camera BEV feature maps FS ∈ RC × H × W through the lightweight backbone and BEV enhancement module described in section 3.2 of this article. The two share the same coordinate system in BEV space, without additional spatial alignment operations, and feature distillation can be directly performed at corresponding spatial positions.

3.5.2 Global BEV Feature Distillation

The first stage applies soft target distillation loss globally to the BEV feature map. Using the mean square error form of feature distillation loss $\mathscr{L}$loss to measure the differences in BEV features between teachers and students at each spatial position:

$$\mathcal{L}_{feat} = \frac{1}{CHW}\|F_S - sg(F_T)\|_2^2 \tag{11}$$

Among them, $sg(.)$ represents stopping the gradient operation, the gradient of distillation loss is only transmitted back to the student model, and the parameters of the teacher model remain fixed during the training process. The global distillation loss guides the alignment of student BEV features towards the teacher in the overall distribution, imposing constraints on both sparse foreground and background regions of the target, laying the foundation for the overall quality of student model BEV representation.

3.5.3 Geometric compensation module

Simple global feature distillation cannot effectively compensate for the disadvantage of camera models in geometric depth representation. LiDAR features have precise measurement information

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences DISCUSSIONS

in the depth dimension, while the depth dimension information of camera BEV features comes from probabilistic depth estimation. The errors are particularly concentrated in occluded and long-distance areas, which is where the gap between teacher and student models is greatest. The geometric compensation module consists of three sub modules, focusing on the precise knowledge transfer of geometrically sensitive areas.

The dense feature extractor takes the student BEV feature FS as input, extracts local geometric structure features through multi-layer convolution, and outputs a geometric perceptual feature map FG. The context aware submodule is based on FG and weights the feature importance of different depth regions through channel attention mechanism. It automatically assigns higher attention weights to occluded dense and distant regions, and outputs context enhanced features FC. The depth predictor takes FC as input and predicts the geometric depth residual Δd of each BEV grid point. The error between the predicted depth d_S+Δd and the teacher depth d_T is taken as the depth consistency loss $\mathcal{L}_{\text{depth}}$:

$$\mathcal{L}_{\text{depth}} = \frac{1}{HW}\sum_{i,j}|d_S(i,j) + \Delta d(i,j) - \text{sg}(d_T(i,j))| \tag{12}$$

The use of L1 form depth loss has stronger robustness against depth outliers, avoiding excessive amplification of gradients caused by mean square error loss in occluded areas with large depth errors.

3.6 Loss Function

The total loss function $\mathcal{L}_{\text{total}}$ during the training phase consists of three parts: detection loss, BEV feature distillation loss, and geometric depth consistency loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_1\mathcal{L}_{\text{feat}} + \lambda_2\mathcal{L}_{\text{depth}} \tag{12}$$

The detection loss CLS includes two parts: target classification loss and 3D bounding box regression loss. Classification loss adopts Focal Loss to alleviate the problem of imbalanced foreground and background categories (Lin et al., 2020):

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{pos}}}\sum_i(1-p_i)^\gamma \log p_i \tag{13}$$

Where p_i is the predicted category probability of the i-th target, γ is the focus parameter, and Npos is the number of positive samples. The bounding box regression loss uses L1 loss to constrain 8 attributes of the target center point coordinates (x, y, z), three-dimensional dimensions (l, w, h), orientation angle θ, and velocity (vx, vy):

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{pos}}}\sum_i\left\|\hat{b}_i - b_i^*\right\|_1 \tag{14}$$

Among them, bi is the predicted bounding box attribute vector, and b_i^* is the corresponding truth attribute vector. The total detection loss is:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{reg}} \tag{15}$$

λ 1, λ 2, and β are the equilibrium hyperparameters of each loss term. The optimal values are determined through ablation experiments on the validation set, and are set to 0.25, 0.5, and 2.0, respectively, in this experiment. In the inference stage, only the forward inference path corresponding to $\mathcal{L}_{\text{det}}$ is calculated, and the distillation related module and the teacher model are removed from the calculation graph without any additional inference overhead.

4 Experiments and Analysis

4.1 Dataset and Evaluation Indicators

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

NuScenes is one of the largest and most fully annotated public benchmark datasets in the field of autonomous driving 3D object detection, containing 1000 driving scenes covering various weather and lighting conditions in the cities of Boston and Singapore. The dataset is equipped with 6 surround view cameras to provide a 360 degree panoramic view, capturing keyframes at a frequency of 2 Hz. The training, validation, and testing sets contain 700, 150, and 150 scenes, respectively. The annotated targets cover 10 categories including vehicles, pedestrians, motorcycles, etc. Each target is equipped with a 3D bounding box, velocity vector, and attribute label.

The evaluation indicators adopt the official average accuracy (mAP) and nuScenes detection score (NDS) of nuScenes. MAP averages the accuracy of 10 types of targets at four BEV distance thresholds (0.5, 1, 2, 4 m). The comprehensive mAP and five attribute errors of NDS, including mean translation error (mATE), mean scale error (mASE), mean orientation error (mAOE), mean velocity error (mAVE), and mean attribute error (mAAE), are calculated as follows:

$$NDS = \frac{1}{10}\left[5mAP + \sum_{k\in\{TE,SE,OE,VE,AE\}}(1 - min(1, mk))\right] \quad (15)$$

The requirements for 3D attribute estimation of targets in NDS are more comprehensive, and the difference in depth estimation accuracy in occluded scenes is particularly significant in NDS. This article takes NDS as the main comparative indicator of comprehensive performance.

4.2 Implementation Details

The visual feature extraction backbone adopts EdgeNeXt Small, which is pre trained and initialized on ImageNet-1K with a pre training resolution of 224 × 224 (Deng et al., 2009). After inserting the deformable convolutional layer into the third and fourth stage outputs of EdgeNeXt, it is implemented with a 3x3 kernel size and a single offset prediction head. The feature pyramid network aggregates the output features of the four stages of the backbone to generate a three-level multi-scale feature map from P3 to P5, with a unified channel count of 256 at each level. The ResNet backbone in the comparative experiment was also pre trained on ImageNet-1K to ensure consistent comparison conditions.

The optimizer uses AdamW with a weight decay coefficient of 0.01 and a basic learning rate of 2 × 10-4 (Loshchilov and Hutter, 2019). The cosine annealing strategy reduces the learning rate to 1 × 10-2 of the initial value within 24 epochs. Batch size 8, trained in data parallel on 4 NVIDIA A100 GPUs. The input image resolution is 256 × 704. During the training phase, random horizontal flipping, random cropping, and GridMask occlusion data enhancement are applied (Chen et al., 2020). Random horizontal flipping is synchronously applied in BEV space, and the enhancement parameters maintain geometric consistency between the image and BEV space.

The BEV perception range is set to 51.2 m for the front and back, 51.2 m for the left and right, with a grid resolution of 0.4 m/grid. The final BEV feature map size is 256 × 256. The height sampling range is -5 m to 3 m, evenly divided into 4 levels. Time series fusion uses four consecutive frames of historical BEV features, with a frame interval of 0.5 seconds. The historical frames are aligned with the self motion matrix and sent to the time series grouping fusion module. The detection head is based on the CenterPoint heatmap prediction paradigm (Yin et al., 2021). The LiDAR teacher model uses PointPillars as the backbone and independently trains on the training set point cloud data until convergence and freezes parameters (Lang et al., 2019). The distillation weights λ 1 and λ 2 are set to 0.25 and 0.5, respectively, and the scene complexity decision threshold τ \ tau τ is determined to be 0.45 at the Pareto optimal point in the validation set.

4.3 Quantitative comparison with mainstream methods

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

Table 1 shows the performance comparison between our method and the current mainstream pure visual 3D object detection methods on the nuScenes validation set. All methods were compared under the same input resolution and evaluation protocol.

Table 1 Performance Comparison of nuScenes Validation Set

Method	Backbone	mAP(%)	NDS(%)	mATE	mASE	mAOE	mAVE	mAAE
BEVDet	R50	29.8	38.8	0.725	0.279	0.589	0.860	0.245
BEVDepth	R50	35.1	47.5	0.639	0.267	0.479	0.428	0.198
BEVDet4D	R50	32.3	45.8	0.703	0.274	0.531	0.374	0.202
BEVStereo (Li et al., 2023a)	R50	37.5	50	0.598	0.27	0.438	0.367	0.190
Fast-BEV	R50	35.4	48.7	0.637	0.269	0.468	0.395	0.200
SOLOFusion	R50	42.7	54	0.567	0.274	0.411	0.252	0.188
Proposed Method	EdgeNeXt-S	38.7	51.3	0.601	0.268	0.441	0.318	0.193
BEVFormer	R101	41.6	51.7	0.673	0.274	0.372	0.394	0.198
BEVDet4D	R101	39.2	47.4	0.686	0.271	0.508	0.352	0.197
SOLOFusion	R101	47.6	58.1	0.532	0.267	0.381	0.246	0.183
BEVNeXt	R101	52	61.7	0.511	0.258	0.341	0.218	0.179
Proposed Method	R101	45.2	57.9	0.548	0.263	0.378	0.261	0.186

Under the ResNet-50 backbone configuration, our method (EdgeNeXt-S) achieved 38.7% mAP and 51.3% NDS, surpassing Fast BEV (35.4% mAP, 48.7% NDS) and BEVDepth (35.1% mAP, 47.5% NDS) with the same backbone configuration, and improving NDS by 2.6 percentage points compared to Fast BEV. Compared with SOLOFusion (42.7% mAP, 54.0% NDS), the lightweight configuration in this paper has a 4.0 percentage point difference in mAP, but the inference delay is reduced by about 52%, which has a significant advantage in deployment efficiency. In terms of mATE indicators, our method (0.601 m) outperforms Fast BEV (0.637 m) and BEVDet4D (0.703 m). The improvement in target 3D localization accuracy comes from the improvement of depth geometry expression through knowledge distillation and the fine modeling of occluded area features through deformable attention sampling. The mAVE index (0.318 m/s) shows significant improvement compared to BEVDepth (0.428 m/s) and BEVDet (0.860 m/s), and the contribution of the temporal grouping fusion module to the accuracy of motion velocity estimation is directly reflected in this index.

Under the ResNet-101 backbone configuration, our method achieved 45.2% mAP and 57.9% NDS, which were 3.6 and 6.2 percentage points higher than BEVFormer R101 (41.6% mAP, 51.7% NDS), respectively. The performance was similar to SOLOFusion R101 (47.6% mAP, 58.1% NDS), with a reduction of approximately 29% in parameter count. Compared with BEVNeXt R101, the difference in NDS is about 3.8 percentage points. BEVNeXt exchanges more complex CRF depth constraints with a two-stage decoder for performance gains, and the inference delay is about 3.2

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences DISCUSSIONS

times that of our method.

Table 2 provides the detailed AP of each method for frequently occluded target categories, in order to more directly measure occlusion perception ability.

Table 2 Classification of Frequent Occluded Categories by AP

Method	Pedestrian	Rider	Motorcycle	Bicycle
BEVDet	22.1	18.4	25.3	15.2
BEVDepth	27.6	23.5	31.2	20.8
Fast-BEV	26.3	21.8	30.7	19.4
SOLOFusion	33.1	28.9	38.4	26.7
Ours	30.6	26.9	35.1	24

The walking human AP (30.6%) increased by 4.3 percentage points compared to Fast BEV (26.3%), and the rider AP (26.9%) increased by 5.1 percentage points compared to Fast BEV (21.8%). The contribution of the occlusion perception BEV enhancement module and temporal grouping fusion in the dense occlusion category of small targets is the most direct. The improvement in AP for motorcycles and bicycles (4.4% and 4.6% respectively) is also significant, and this type of target has a high occlusion ratio in intersection scenes, which is highly consistent with the optimization scenario of our method.

4.4 Ablation Experiment

To verify the independent contribution of each proposed module, we conduct a cumulative ablation study under the ResNet-50 backbone, taking BEVDet as the baseline (configuration (a)). The five modules are added one at a time, producing configurations (b) through (f), where (f) corresponds to the complete method. All experiments are performed on the nuScenes validation set under identical training settings. The results are reported in Table 3.

Table 3 ablation experiments for each module *

#	LB	SE	AS	TGF	KD	mAP(%)	NDS(%)	mATE	mAVE	Params (M)	Latency (ms)
(a)						29.8	38.8	0.725	0.860	33.1	48.3
(b)	✓					30.6	40.2	0.718	0.842	19.4	35.1
(c)	✓	✓				32.4	42.6	0.679	0.821	20.3	36.8
(d)	✓	✓	✓			35.8	46.9	0.638	0.794	23.1	40.2
(e)	✓	✓	✓	✓		37.2	49.7	0.627	0.351	24.6	43.5
(f)	✓	✓	✓	✓	✓	38.7	51.3	0.601	0.318	26.4	38.2

* LB = Lightweight Backbone, SE = Spatial Embedding, AS = Adaptive Sampling, TGF = Temporal Grouping Fusion, KD = Knowledge Distillation. Configuration (a) is the BEVDet baseline; (f) is the complete method. The best value in each column is shown in bold

Effect of the lightweight backbone. Replacing the ResNet-50 backbone of BEVDet with EdgeNeXt-S augmented by deformable convolution raises mAP from 29.8% to 30.6% and NDS from 38.8% to 40.2%, while reducing the parameter count from 33.1 M to 19.4 M and the inference latency from 48.3 ms to 35.1 ms. Detection accuracy is preserved and even slightly improved despite

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences DISCUSSIONS

the substantial parameter reduction, confirming that the lightweight backbone replacement contributes most directly to inference efficiency.

Effect of lightweight spatial embedding. Adding the spatial embedding module increases mAP to 32.4% and NDS to 42.6%, and reduces mATE from 0.718 m to 0.679 m. The channel-wise encoding of height information substantially improves 3D coordinate estimation at a marginal cost of 0.9 M additional parameters and 1.7 ms latency, indicating a favorable accuracy–cost trade-off.

Effect of adaptive voxel feature sampling. Configuration (d) achieves 35.8% mAP and 46.9% NDS — the largest single improvement among all modules. The pedestrian AP rises by 3.8 percentage points at this stage, demonstrating that deformable attention sampling effectively refines BEV feature extraction for occluded targets. To further analyze the value of the adaptive switching mechanism, we evaluate a variant in which the deformable attention path is always activated. This variant attains 47.6% NDS, only 0.7 percentage points higher than the adaptive scheme, but its latency rises to 58.4 ms — an increase of 18.2 ms. This confirms that scene-complexity-driven switching preserves accuracy in occluded scenes while significantly safeguarding inference efficiency.

Effect of temporal grouping fusion. Adding temporal grouping fusion raises mAP and NDS to 37.2% and 49.7%, respectively, and produces a pronounced drop in mAVE from 0.794 m/s to 0.351 m/s. The improvement is most concentrated in motion velocity estimation, validating the module's ability to capture fine-grained inter-frame dependencies for moving and occluded targets. As a comparison, replacing grouping fusion with direct channel concatenation yields 49.1% NDS (0.6 percentage points lower) while introducing approximately 12 M additional parameters, demonstrating the parameter efficiency of the grouping design.

Effect of multimodal knowledge distillation. Introducing multimodal knowledge distillation yields the final configuration (f), which reaches 38.7% mAP and 51.3% NDS, with mATE further reduced from 0.627 m to 0.601 m. The improvement in depth-geometry accuracy is directly reflected in the translation error. Although knowledge distillation adds a teacher model during training, these parameters are excluded from inference, and the inference latency in fact decreases from 43.5 ms to 38.2 ms. This is because teacher supervision improves the convergence quality of the student's depth estimation network, allowing stable depth prediction in fewer iteration steps and indirectly improving the execution efficiency of the sampling module. When only global BEV feature distillation is applied without the geometric compensation module, NDS drops to 50.5% — a 0.8 percentage point gap from the complete distillation scheme — confirming the independent contribution of the geometric compensation module to fine-grained distillation in depth-sensitive regions.

4.5 Analysis of Reasoning Efficiency

Table 4 compares the parameter count, floating-point operation complexity, and inference delay between the proposed method and the main comparison methods. The testing platform is a single NVIDIA Tesla T4, with a batch size of 1, an input resolution of 256 × 704, and a delay of 100 inference mean.

Table 4 Comparison of Reasoning Efficiency

Method	Backbone	Parameters (M)	GFLOPs	Latency(ms)	FPS	mAP (%)	NDS (%)
BEVDet	R50	33.1	215.3	48.3	20.7	29.8	38.8
BEVDepth	R50	33.2	388.6	112.4	8.9	35.1	47.5

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

BEVFormer	R101	53.3	525.8	262.1	3.8	41.6	51.7
SOLOFusion	R50	33.1	295.7	93.8	10.7	42.7	54
Fast-BEV	R50	33.1	167.9	26.8	37.3	35.4	48.7
Ours	EdgeNeXt-S	26.4	141.6	38.2	26.2	38.7	51.3
Ours	R101	51.2	298.1	67.4	14.8	45.2	57.9

The total parameter count of this method is 26.4M, with a inference floating-point operation of 141.6 GFLOPs. The single frame inference delay on Nvidia Telsa T4 GPU is 38.2 ms, corresponding to a frame rate of approximately 26.2 FPS, which meets the real-time constraints of most in vehicle perception systems. Compared with Fast BEV, the method proposed in this paper has a delay of 11.4 ms, but the mAP and NDS are improved by 3.3 and 2.6 percentage points, respectively. The accuracy gain comes from the introduction of the adaptive sampling and timing fusion module. Compared with SOLOFusion, the delay of our method is reduced by 55.6 ms, with a difference of 4.0 percentage points in mAP. Our method is at a better equilibrium point on the precision delay Pareto curve. BEVFormer has a latency of up to 262.1 ms and a frame rate of only 3.8 FPS, making it unsuitable for real-time deployment in vehicles. When tested separately for complex scenes with high occlusion ratios, the activation ratio of deformable attention paths is about 60%, and the delay increases to 46.3 ms. In simple scenes such as open roads, switching to fast ray projection paths reduces the delay to 34.1 ms, which is close to Fast BEV. The scene adaptation mechanism maintains the average delay of our method at 38.2 ms in different complexity scenarios, avoiding the worst-case delay problem caused by fixed high-precision paths.

4.6 Qualitative analysis

Figure 4 shows the comparison of 3D detection visualization between the proposed method and the baseline model in the nuScenes dataset under various complex scene conditions. Figure 5 shows the voxel information maps of the adaptive sampling strategy with different sampling paths in simple and complex scenes.

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences DISCUSSIONS

Figure 4 Comparison of 3D detection visualization between the proposed method and baseline model in typical scenes of the nuScenes dataset

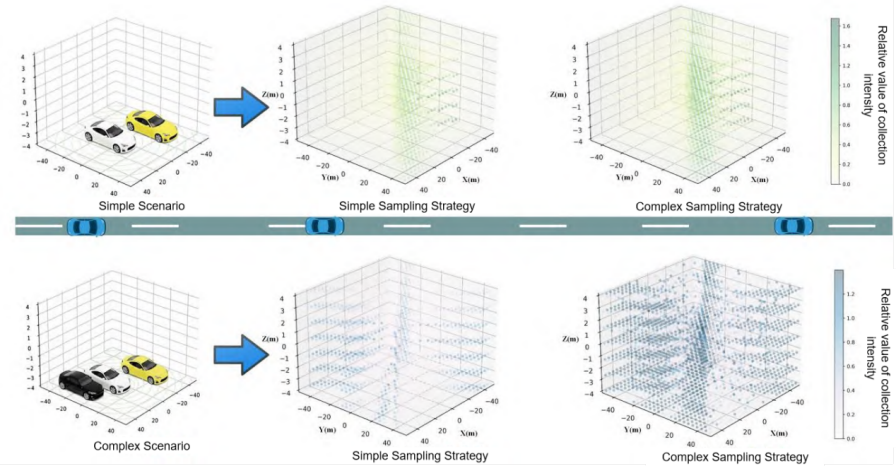

Figure 5 Comparison of voxel information maps between simple sampling strategy and complex sampling strategy in simple and complex scenes

In Figure 4, the baseline model has obvious missed detection of distant vehicles and occluded targets in rainy and dark scenes, and the 3D bounding box is incomplete; Under the combined effect of historical frame feature supplementation through temporal grouping fusion and refined BEV modeling through adaptive sampling, this method achieves more complete target coverage, more accurate 3D bounding box localization, and significantly reduced missed detection rate in the aforementioned scenarios. In cloudy scenes, the results of the two methods are similar, while in low visibility obstructed scenes such as cloudy, rainy, and dark nights, the advantages of our method

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

become increasingly apparent as the difficulty of the scene increases. The voxel information diagram in Figure 5 intuitively reflects the rationality of the design of the adaptive sampling strategy. In simple scenes, the density and intensity distribution of voxels obtained by simple sampling strategy and complex sampling strategy are almost identical, indicating that fast ray projection can obtain complete voxel information in simple scenes without the need to use more computationally expensive complex sampling paths. In complex scenes, the voxel sampling points of the simple sampling strategy are sparse and have relatively low intensity values, resulting in significantly insufficient information; The voxel sampling points of complex sampling strategies are dense and have rich intensity distributions, and the amount of scene information obtained far exceeds that of simple sampling. The adaptive switching strategy dynamically schedules between two paths based on the complexity score of the scene, saving approximately 18 ms of inference time in simple scenes and ensuring voxel information integrity in complex occlusion scenes, achieving an effective balance between accuracy and efficiency.

5 Discussion

The proposed lightweight BEV perception framework achieves an effective balance between accuracy and real-time performance for pure-vision 3D detection in occluded scenes. Under the lightweight configuration, it attains 38.7% mAP and 51.3% NDS on the nuScenes validation set, with 38.2 ms latency and 26.2 FPS on the T4 platform, ranking highest in NDS among comparable methods while meeting real-time vehicle deployment constraints.

The combination of the EdgeNeXt backbone and deformable convolution compresses backbone parameters to roughly one-third of ResNet-50 while compensating for the restricted receptive field of a lightweight backbone, enabling adaptive coverage of occluded boundaries and irregular contours. The lightweight spatial embedding mechanism improves mATE by 0.04 m with parameter overhead under 5% of the backbone, indicating that height-dimension loss during BEV projection is an underexploited source of localization error that can be recovered via channel encoding rather than expensive voxel representations. For occlusion-aware sampling, the ablation confirms that adaptive switching reduces inference latency by 18.2 ms compared to the fixed high-accuracy path with only a 0.7 percentage point NDS drop. The deformable attention path is activated about 60% of the time in highly occluded scenes (latency 46.3 ms) and falls back to fast ray projection in open scenes (34.1 ms, close to Fast-BEV). Pedestrian and rider AP improve by 4.3% and 5.1% respectively over Fast-BEV, with the largest gains concentrated in small-target dense occlusion categories. This validates the core design premise: the computational overhead of deformable attention is justified only in genuinely complex scenes, and enforcing it universally is an inefficient allocation of the compute budget.

The temporal grouping fusion module reduces mAVE from 0.794 m/s to 0.351 m/s — a nearly 56% reduction — directly reflecting its contribution to motion velocity estimation for fast-moving and occluded targets. Compared with direct channel concatenation, grouping fusion reduces parameter count by approximately 12M while yielding a 0.6 percentage point NDS gain, and residual connections preserve BEV geometric consistency for stable detection-head convergence. For multimodal knowledge distillation, global BEV feature distillation establishes the overall representation, while the geometric compensation module independently contributes 0.8% NDS by focusing on depth-sensitive occluded and long-range regions; distillation improves mATE by 0.026 m. A secondary effect is that inference latency drops from 43.5 ms to 38.2 ms after distillation, owing to improved convergence quality of the student's depth estimation rather than any

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

architectural change. The teacher and distillation modules are removed at inference, incurring zero overhead. Under the R101 configuration, the method achieves 45.2% mAP and 57.9% NDS with approximately 38% fewer parameters than BEVFormer, validating the framework's scalability across backbones.

6 Conclusion

This paper proposes a lightweight BEV perception optimization framework for occluded scenes to address the fundamental conflict between accuracy and real-time performance in pure-vision 3D object detection. By integrating lightweight feature extraction, occlusion-aware adaptive sampling, temporal grouping fusion, and multimodal knowledge distillation into a unified end-to-end architecture, the framework achieves the highest NDS among comparable lightweight methods on the nuScenes validation set while satisfying real-time vehicle deployment constraints, providing an effective technical pathway for embedded lightweight deployment of pure-vision 3D perception systems.

Acknowledgements

This work is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202503705).

Author contribution

Y. Z. designed the study, developed the methodology, conducted the experiments, and wrote the original draft. A. Z. and J. L. contributed to the validation and data curation. J. L. provided supervision, project administration, and funding acquisition.

Data availability

nuScenes dataset is available at: https://www.nuscenes.org/

References

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11621–11631, https://doi.org/10.1109/CVPR42600.2020.01164, 2020.

Chen, P., Liu, S., Zhao, H., Wang, X., and Jia, J.: GridMask Data Augmentation, arXiv [preprint], https://doi.org/10.48550/arXiv.2001.04086, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255, https://doi.org/10.1109/CVPR.2009.5206848, 2009.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P.: Res2Net: A New Multi-Scale Backbone Architecture, IEEE T. Pattern Anal., 43, 652–662, https://doi.org/10.1109/TPAMI.2019.2938758, 2021.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Huang, F., Liu, S., Zhang, G., Hao, B., Xiang, Y., and Yuan, K.: DeployFusion: A Deployable Monocular 3D Object Detection with Multi-Sensor Information Fusion in BEV for Edge Devices, Sensors, 24, 7007, https://doi.org/10.3390/s24217007, 2024.

Huang, J. and Huang, G.: BEVDet4D: Exploit Temporal Cues in Multi-camera 3D Object Detection, arXiv [preprint], https://doi.org/10.48550/arXiv.2203.17054, 2022.

Huang, J., Huang, G., Zhu, Z., Ye, Y., and Du, D.: BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View, arXiv [preprint], https://doi.org/10.48550/arXiv.2112.11790, 2021.

Jiao, T., Chen, Y., Zhang, Z., Guo, C., and Song, J.: MMDistill: Multi-Modal BEV Distillation Framework for Multi-View 3D Object Detection, Comput. Mater. Contin., 4307–4325, https://doi.org/10.32604/cmc.2024.058238, 2024.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O.: PointPillars: Fast Encoders for Object Detection From Point Clouds, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12697–12705, https://doi.org/10.1109/CVPR.2019.01298, 2019.

Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., and Li, Z.: BEVStereo: Enhancing Depth Estimation in Multi-View 3D Object Detection with Temporal Stereo, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 1486–1494, https://doi.org/10.1609/aaai.v37i2.25234, 2023a.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., and Li, Z.: BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 1477–1485, https://doi.org/10.1609/aaai.v37i2.25233, 2023b.

Li, Y., Huang, B., Chen, Z., Cui, Y., Liang, F., Shen, M., Liu, F., Xie, E., Sheng, L., Ouyang, W., and Shao, J.: Fast-BEV: A Fast and Strong Bird's-Eye View Perception Baseline, IEEE T. Pattern Anal., 46, 8665–8679, https://doi.org/10.1109/TPAMI.2024.3414835, 2024.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., and Dai, J.: BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers, in: 17th European Conference on Computer Vision (ECCV), 1–18, https://doi.org/10.1007/978-3-031-20077-9_1, 2022.

Li, Z., Lan, S., Alvarez, J. M., and Wu, Z.: BEVNeXt: Reviving Dense BEV Frameworks for 3D Object Detection, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20113–20123, https://doi.org/10.1109/CVPR52733.2024.01901, 2024.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S.: Feature Pyramid Networks for Object Detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 936–944, https://doi.org/10.1109/CVPR.2017.106, 2017.

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal Loss for Dense Object Detection, IEEE T. Pattern Anal., 42, 318–327, https://doi.org/10.1109/TPAMI.2018.2858826, 2020.

Liu, H., Teng, Y., Lu, T., Wang, H., and Wang, L.: SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 18580–18590, https://doi.org/10.1109/ICCV51070.2023.01703, 2023.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., and Han, S.: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), 2774–2781, https://doi.org/10.1109/ICRA48891.2023.10160968, 2023.

Liu, Y., Wang, T., Zhang, X., and Sun, J.: PETR: Position Embedding Transformation for Multi-view 3D Object Detection, in: 17th European Conference on Computer Vision (ECCV), 531–548, https://doi.org/10.1007/978-3-031-19812-0_31, 2022.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in: International Conference on Learning Representations (ICLR), 2019.

Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., and Shahbaz Khan, F.: EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications, in: European Conference on Computer Vision Workshops (ECCV 2022 Workshops), 3–20, https://doi.org/10.1007/978-3-031-25082-8_1, 2023.

Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K. M., Tomizuka, M., and Zhan, W.: Time Will Tell: New Outlooks and A Baseline for Temporal Multi-View 3D Object Detection, in: 11th International Conference on Learning Representations (ICLR), 2023.

Philion, J. and Fidler, S.: Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D, in: 16th European Conference on Computer Vision (ECCV), 194–210, https://doi.org/10.1007/978-3-030-58568-6_12, 2020.

Wang, T., Zhu, X., Pang, J., and Lin, D.: FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 913–922, https://doi.org/10.1109/ICCVW54120.2021.00107, 2021.

Wang, Y., Guizilini, V. C., Zhang, T., Wang, Y., Zhao, H., and Solomon, J.: DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries, in: 5th Conference on Robot Learning, Proceedings of Machine Learning Research (PMLR), 164, 180–191, 2022.

Yin, T., Zhou, X., and Krähenbühl, P.: Center-based 3D Object Detection and Tracking, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11779–11788, https://doi.org/10.1109/CVPR46437.2021.01161, 2021.

You, Z., Wang, N., Wang, H., Zhao, Q., and Wang, J.: MambaBEV: An EV-based 3D Detection Model with Mamba2, arXiv [preprint], https://doi.org/10.48550/arXiv.2410.12673, 2026.

https://doi.org/10.5194/ms-2026-168
Preprint. Discussion started: 22 September 2026

mechanical sciences
DISCUSSIONS

686 Zhang, J., Zhang, Y., Cai, W., Liu, Q., and Wang, Y.: LightOcc: Lightweight Spatial Embedding for
687 Efficient Vision-based 3D Occupancy Prediction, Int. J. Comput. Vis., 134, 347,
688 https://doi.org/10.1007/s11263-026-02934-9, 2026.

689 Zhou, S., Liu, W., Hu, C., Zhou, S., and Ma, C.: UniDistill: A Universal Cross-Modality Knowledge
690 Distillation Framework for 3D Object Detection in Bird's-Eye View, in: 2023 IEEE/CVF
691 Conference on Computer Vision and Pattern Recognition (CVPR), 5116–5125,
692 https://doi.org/10.1109/CVPR52729.2023.00495, 2023.

693 Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J.: Deformable DETR: Deformable Transformers
694 for End-to-End Object Detection, in: 9th international Conference on Learning Representations
695 (ICLR), 2021.