DeGuNet: Depth-Guided Ultra-Compact Backbones for Efficient LiDAR-Camera 3D Detection

Haifa Zhang[1], Yijing Wang[1⋆], Peixi Peng[2], and Zhiqiang Zuo[1⋆]

[1] Tianjin University, Tianjin, China
{zhanghaifa,yjwang,zqzuo}@tju.edu.cn
[2] Peking University, Beijing, China
pxpeng@pku.edu.cn

Abstract. In autonomous driving perception, the fusion of LiDAR and camera modalities has become the dominant paradigm for 3D object detection. However, current multi-modal frameworks heavily rely on massive visual backbones pretrained on 2D semantic tasks. This reliance introduces substantial parameter redundancy and a structural misalignment, as 2D priors are ill-equipped to handle the extreme sparsity of LiDAR projections required for Bird's-Eye-View geometry. To address this, we present DeGuNet, an ultra-compact and plug-and-play image backbone explicitly designed for depth-guided representation learning. By incorporating sparsity-aware feature extraction mechanisms, DeGuNet effectively aligns multi-view images with unstructured LiDAR depth while strictly preventing invalid-region contamination. Extensive experiments on the nuScenes dataset demonstrate DeGuNet's broad plug-and-play applicability and superior efficiency. When integrated into established baselines, it fundamentally eliminates architectural redundancy, reducing GPU memory consumption by up to 66.5% and achieving a 1.16× inference speedup. Concurrently, DeGuNet delivers up to a 6.20 absolute mAP gain, establishing a new paradigm for parameter-efficient multi-modal 3D perception.

Keywords: 3D Object Detection · LiDAR-Camera Fusion · Sparse Representation · Lightweight Backbone

1 Introduction

In autonomous driving perception, multi-modal 3D object detection [3,7,22,41] fusing LiDAR and camera data has become the dominant paradigm [1,4,7,30,52]. While LiDAR provides precise geometric structure [6,10,12,14,36], cameras offer crucial high-resolution textures and semantic details [25,27,59]. Current state-of-the-art frameworks achieve remarkable performance by projecting features from both modalities into a unified Bird's-Eye-View (BEV) representation. However,

⋆ Corresponding authors.

this success heavily relies on large-scale visual backbones, such as large ResNets or Transformers, which are pretrained on 2D semantic tasks like image classification or object detection [17, 29, 44]. Consequently, the image branch often dominates the system's parameter count. Empirical profiling of prevailing multi-modal frameworks indicates that standard visual backbones consume between 30M and 78M parameters, accounting for 50% to 86% of the total model size. This highlights a severe system-level resource redundancy for a modality that primarily serves a complementary role.

Beyond computational overhead [7, 48], this reliance on parameter-heavy 2D backbones leads to a structural inconsistency. Features optimized for 2D semantic tasks lack the vital 3D spatial awareness required for BEV localization, which inherently depends on accurate metric depth estimation to lift pixels into 3D space. To bridge this gap, a natural intuition is to adopt a depth-centric pretraining paradigm. However, our empirical analysis (detailed in Sec. 5) reveals that simply pretraining standard lightweight backbones (such as a tiny ResNet [17] or MobileViT [32]) on depth completion tasks yields marginal improvements in downstream 3D detection, typically plateauing around 62.1 mAP. We attribute this limited efficacy to the inability of standard convolutional architectures to process extreme spatial sparsity. When projecting sparse LiDAR point clouds onto dense camera planes (over 98% of the pixels remain empty, as counted in our nuScenes [2] analysis), standard receptive fields become heavily contaminated by these invalid, zero-value regions. Thus, establishing an effective depth-guided representation requires not only a geometry-centric pretraining objective but also a dedicated, sparsity-aware architecture.

To address these challenges, we propose the Depth-Guided Network (DeGuNet), an ultra-compact, plug-and-play image backbone containing only 0.31M parameters. Instead of inheriting fixed 2D priors, DeGuNet is purposefully pretrained on a depth completion task to generate features intrinsically aligned with the target BEV geometry. To handle sparse geometric guidance, we propose Masked Partial Inverted Residual (MPIR) blocks and progressive Guide modules. These components enforce mask-aware feature extraction to prevent invalid regions from degrading the learned representations. By extracting depth-aware features directly aligned with LiDAR geometry before cross-modal fusion, DeGuNet serves as a general-purpose infrastructure capable of seamlessly replacing heavyweight image backbones across various multi-modal frameworks.

Our main contributions are summarized as follows:

- We reveal the critical feature misalignment in current multi-modal 3D detectors and demonstrate that resolving it requires not just depth-centric pretraining, but a specialized sparsity-aware architecture capable of handling extreme sparsity levels over **98%** invalid regions.
- We propose DeGuNet, a parameter-efficient (0.31M parameters) and plug-and-play image backbone. It integrates MPIR blocks and progressive Guide modules to effectively fuse multi-view images with sparse LiDAR geometry without invalid-region contamination.

– Extensive cross-baseline experiments demonstrate DeGuNet's broad plug-and-play applicability. When integrated as a drop-in replacement into established frameworks including BEVFusion, GraphBEV, EA-LSS, and IS-Fusion, it reduces GPU memory consumption by up to **66.5%** and achieves a **1.16**× inference speedup, while concurrently delivering up to **6.20** absolute mAP gain in 3D detection accuracy.

2 Related Work

2.1 Multi-modal 3D Object Detection

The fusion of LiDAR and camera data has become a widely adopted paradigm for high-performance 3D object detection. A prevailing approach projects features from both modalities into a unified BEV representation [26, 30, 34], providing a shared, geometrically consistent grid for cross-modal alignment [25]. Concurrently, query- and transformer-based methods model multi-modal interactions through attention mechanisms [1, 49, 54, 55]. Recently, to address computational bottlenecks in dense BEV maps, several schemes such as SparseFusion [48] and Fully Sparse Fusion [23] have been proposed to operate on sparse, instance-level features. Despite these advances in the fusion stage, both lines of research remain tethered to bulky 2D-pretrained visual backbones. Their image features are not intrinsically aligned with the geometry-aware BEV space, as BEV localization requires metric depth ordering and cross-view geometric consistency that are not directly encouraged by 2D pretraining.

2.2 Limitations of Visual Backbones in 3D Perception

A primary bottleneck exacerbating the aforementioned geometry-semantic misalignment is the conventional visual feature extraction paradigm. Current state-of-the-art multi-modal detectors heavily rely on large-scale visual backbones, such as deep ResNets [17] and Swin Transformers [29], which are pretrained on 2D tasks like ImageNet classification [37] or 2D object detection. While these heavy backbones provide rich semantic priors, they incur considerable parameter redundancy. Furthermore, the semantic features inherited from fixed 2D pretraining are misaligned with the geometry-aware representations required for 3D prediction, which depend on metric depth cues and cross-view consistency rather than purely appearance semantics. The backbone's original 2D objective is not specifically optimized for 3D supervision, resulting in underutilized feature capacity. In contrast, our work discards the conventional large-scale visual backbone. We propose a parameter-efficient architecture that serves as a task-aligned drop-in alternative.

2.3 Depth-Guided Representation Learning

Depth information is essential for bridging the modality gap between 2D images and 3D space. Many existing 3D detection frameworks incorporate auxiliary

depth estimation heads and dense depth losses during detector training (e.g., BEVDepth) [24,57] to actively guide the projection process [20,46]. On a parallel track, depth completion networks [8,47,51] address the generation of dense depth maps from sparse LiDAR projections. To handle the irregular distribution of sparse points, specialized operations such as partial convolutions [28] have been explored to apply convolutional filters only to valid pixels. Our approach differs from standard auxiliary depth strategies used during detector training. Instead of adding depth prediction burdens to the downstream 3D detector, we shift the geometric alignment to a standalone pretraining stage via a depth completion task. By integrating mask-aware partial convolutions [28] into our MPIR blocks, we strictly exclude invalid zero-value pixels, preventing sparse-data artifacts from contaminating the representation. This ensures that the extracted features are intrinsically geometry-aligned prior to integration into any downstream multi-modal pipeline.

3 Problem Analysis

While multi-modal 3D detectors achieve state-of-the-art performance, their reliance on bulky visual backbones introduces a critical yet often overlooked inefficiency [7, 30]. Standard frameworks typically deploy large 2D-pretrained networks (e.g., Swin Transformers [29] or deep ResNets [17]). For instance, in the established BEVFusion framework [26], the CBSwin-T visual backbone consumes 78.1M parameters, which accounts for 86.6% of the 90.2M total model capacity. Furthermore, a deep-rooted structural inconsistency exists: the semantic features optimized for 2D classification or detection are intrinsically misaligned with the geometry-aware BEV representations demanded by 3D localization. Consequently, these over-parameterized backbones merely act as heavy, inefficient parameter initializers whose capacity for geometry-aware feature extraction remains largely underutilized.

To mitigate this misalignment, an intuitive strategy is to replace 2D semantic pretraining with LiDAR-guided depth completion [8, 47]. However, altering the pretraining objective alone does not overcome the inherent structural limitations of standard network operators. Specifically, conventional convolutional layers operate with spatial invariance, applying identical local aggregation rules across the entire feature map. When processing highly irregular and sparse LiDAR projections, these standard kernels indiscriminately blend valid geometric depth cues with the surrounding empty pixels. Consequently, as the network deepens, the sparse valid signals are rapidly diluted, leading to significant feature degradation within the receptive field. This mechanistic analysis indicates that bridging the semantic-geometric gap demands more than a task-level adjustment; it inherently requires a purpose-built, sparsity-aware architecture.

This insight motivates the design of DeGuNet. Rather than relying on large-scale 2D backbones or naively applying standard lightweight convolutions to sparse data, we propose a purpose-built, streamlined architecture. By incorporating MPIR blocks [28] and progressive Guide modules, DeGuNet selectively

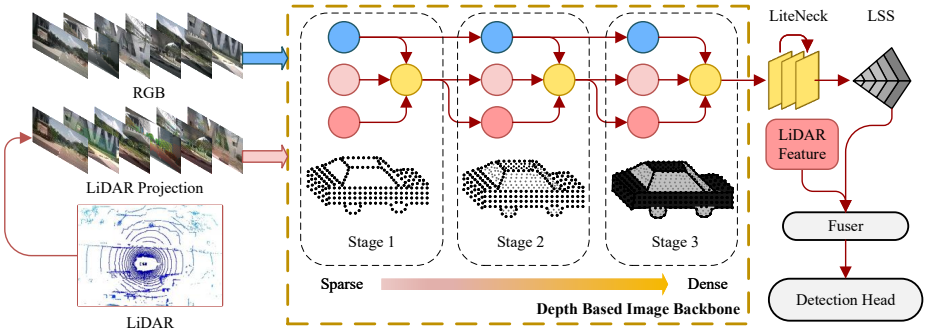

Fig. 1: Overview of the proposed detection framework incorporating DeGuNet as a drop-in module. The architecture features a dual-stream design. The depth-based image backbone processes multi-view RGB images and sparse LiDAR projections through progressive stages, generating densified, geometry-aligned features. These features are subsequently decoupled via a LiteNeck and projected into the BEV space for fusion with LiDAR features.

processes valid geometric signals while isolating invalid regions. This targeted, sparsity-aware design allows a compact backbone to inherently align image features with 3D LiDAR geometry, addressing both the parameter redundancy and the structural misalignment.

4 Methodology

Our primary objective is to bridge the structural misalignment between standard 2D-pretrained image features and BEV representations demanded by multi-modal 3D object detection. To this end, we propose the DeGuNet, a lightweight, plug-and-play image backbone. Unlike conventional backbones that passively inherit fixed 2D semantic priors, DeGuNet is deliberately pretrained on a LiDAR-guided depth completion task. This specific pretraining ensures that the extracted image features are depth-aware and better aligned with the 3D space prior to any cross-modal fusion. In this section, we first detail the overall macro-architecture and its two-stage training lifecycle. Subsequently, we elaborate on the geometry-guided pretraining paradigm and the core sparsity-aware micro-components designed to mitigate the contamination from invalid projection regions.

4.1 Overall Architecture

As illustrated in Fig. 1, the multi-modal detection pipeline is constructed as a dual-stream architecture. It comprises a primary LiDAR processing stream, which can utilize any standard voxel-based [6, 16, 19, 38, 50, 58] or point-based encoder [11, 35, 36, 39, 40, 53], and a complementary image stream featuring the

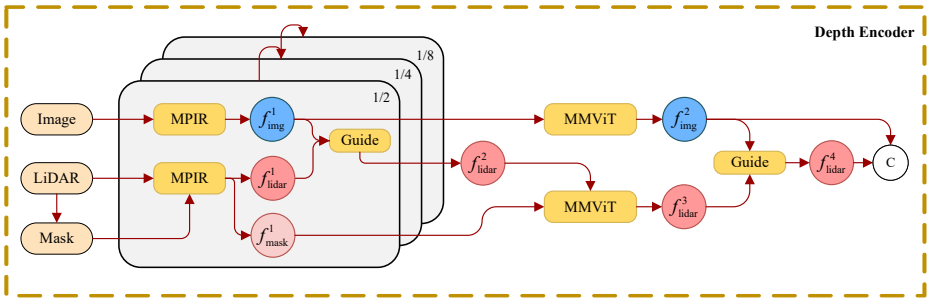

Fig. 2: Architecture of the geometry-guided pretraining encoder. The network processes dense multi-view images and sparse LiDAR projections in parallel across three progressive scales. Mask-aware operations ensure that only valid geometric cues guide the representation learning.

proposed DeGuNet as its backbone. Operating as a general-purpose feature extractor, DeGuNet takes multi-view RGB images and sparse LiDAR depth projections as parallel inputs. Through progressive stages of cross-modal guidance, it outputs a densified, depth-aware feature representation.

To ensure generalizability across different LSS-based (Lift-Splat-Shoot) frameworks, the deployment of DeGuNet follows a two-phase lifecycle:

- **Phase 1: Geometry-Guided Pretraining.** An encoder-decoder architecture is constructed to perform depth completion. The encoder (DeGuNet) learns to extract geometry-aligned features from dense RGB and sparse LiDAR inputs, while the decoder upsamples these features solely for depth loss computation.
- **Phase 2: End-to-End Detection Integration.** Upon completion of pretraining, the depth decoder is discarded, while the pretrained encoder is retained and integrated as a drop-in image backbone within the multi-modal 3D detection framework. The features from both sensor streams are then projected into a unified BEV grid and fused for final 3D bounding box prediction.

This decoupled lifecycle guarantees that the image features entering the BEV projection module are aligned with the LiDAR geometry, mitigating the semantic-geometric gap at its source.

4.2 Geometry-Guided Pretraining Paradigm

To bridge the spatial awareness gap of 2D-pretrained backbones, we shift the pretraining objective to LiDAR-guided depth completion, forcing the network to directly map 2D textures to 3D geometry. During this phase, the architecture functions as an encoder-decoder network. As shown in Fig. 2, the encoder (the core DeGuNet) processes multi-camera RGB images and sparse LiDAR projections through three progressive downsampling stages (at 1/2, 1/4, and 1/8

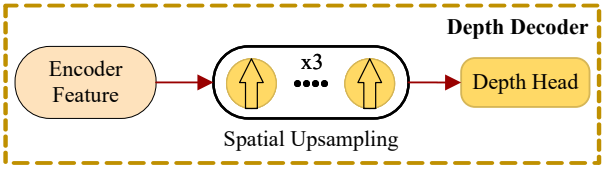

Fig. 3: The Depth Decoder, utilized exclusively during the pretraining phase to compute the depth completion loss.

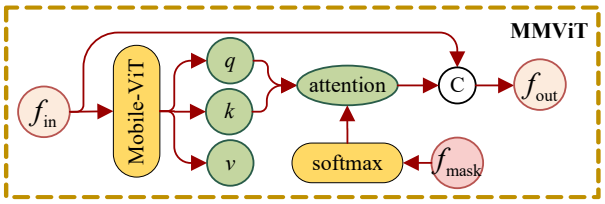

Fig. 4: Architecture of the MMViT block, which applies mask-aware attention to prevent invalid depth from polluting global context.

resolutions). To handle the inherent sparsity of LiDAR, the encoder employs specialized mask-aware components to prevent invalid projection regions from corrupting the learned features.

A lightweight spatial decoder (Fig. 3) then upsamples these multi-scale features to predict dense depth maps. Given the ground-truth depth Z and a validity mask $M = \mathbb{I}[Z > 0]$, the pretraining is supervised exclusively on valid LiDAR-projected pixels via a multi-scale loss:

$$\mathcal{L}_{dc} = \sum_{i=1}^{K} \gamma^{K-i} \left(0.5 \|(\hat{Z}^{(i)} - Z) \odot M\|_1 + 0.5 \|(\hat{Z}^{(i)} - Z) \odot M\|_2^2 \right), \quad (1)$$

where $\{\hat{Z}^{(i)}\}_{i=1}^{K}$ are multi-scale depth predictions, γ is a scale-weighting factor, and $\odot$ denotes element-wise multiplication. Crucially, this decoder is purely auxiliary. Upon completing the pretraining, it is entirely discarded. The encoder, now possessing robust geometry-aware feature extraction capabilities, will be seamlessly transferred to the downstream 3D detector, providing aligned features without incurring any auxiliary depth prediction overhead during inference.

4.3 Sparsity-Aware Feature Extraction

Standard convolutional layers apply spatial filters uniformly across their receptive fields. When applied to unstructured LiDAR projections, which predominantly consist of empty pixels, standard kernels inevitably aggregate zeros [15,43]. This leads to significant feature degradation and the dilution of valid geometric cues. To prevent this invalid-region degradation, DeGuNet employs a dedicated sparsity-aware design throughout both its local and global feature extraction stages.

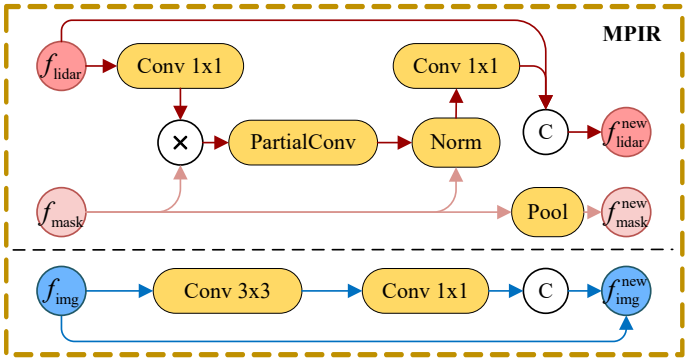

Fig. 5: Architecture of the MPIR. It employs mask-guided partial convolutions on the LiDAR stream to prevent invalid region contamination during local feature extraction.

To capture long-range spatial dependencies at lower resolutions, DeGuNet utilizes the Masked Mobile Vision Transformer (MMViT) block, as shown in Fig. 4. Standard self-attention mechanisms compute global token interactions, which would erroneously allow invalid background nodes to influence the representations of valid geometric structures. To avert this, MMViT introduces a mask-aware attention mechanism [9, 31]. The binary mask f_{mask} is dynamically downsampled, reshaped, and injected into the attention matrix calculation. By masking out invalid tokens with a large negative value prior to the softmax normalization, MMViT mathematically guarantees that the attention weights for empty regions approach zero. Consequently, the global context aggregation is confined exclusively to valid geometric nodes, preserving the absolute integrity of the spatial representations.

Complementary to the global context aggregation at deeper stages, local feature extraction is handled by the MPIR block, as illustrated in Fig. 5. The MPIR block processes the multi-modal inputs through a decoupled dual-branch architecture. The image branch processes dense RGB features (f_{img}) using standard depth-wise separable convolutions to capture continuous semantic textures. Conversely, the sparse LiDAR branch (f_{lidar}) incorporates mask-guided partial convolutions [28]. Given the binary mask f_{mask} indicating the locations of valid depth values, the partial convolution restricts its weight updates and feature aggregations to these valid pixels. A subsequent pooling operation dynamically updates f_{mask} for the next layer. This mechanism ensures that the geometric representation remains uncontaminated by the void regions, even as the receptive field expands in deeper layers.

4.4 Progressive Cross-Modal Guidance and Integration

While the MPIR and MMViT blocks ensure the purity of geometric features by preventing invalid-region pollution, relying solely on sparse LiDAR projections limits the representational density of the network. To address this, DeGuNet

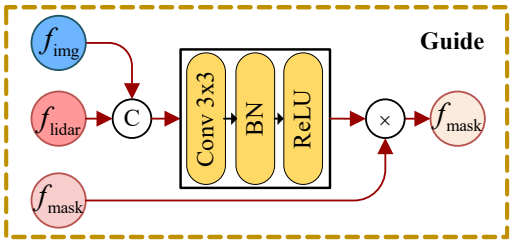

Fig. 6: Architecture of the progressive Guide module. It injects dense image semantics into the sparse LiDAR stream.

incorporates a progressive cross-modal guidance mechanism, operating at the early network stages (e.g., 1/2 and 1/4 input resolutions). This early fusion strategy injects high-resolution, dense RGB semantics into the sparse geometric stream, effectively densifying the geometry-aware representations.

To ensure strict alignment between geometric validity and feature resolution throughout the guidance process, the binary mask is recursively updated during downsampling via a max-pooling operation:

$$f_{\text{mask}}^{(l+1)} = \text{MaxPool}_{s\times s}(f_{\text{mask}}^{(l)}), \tag{2}$$

where $f_{\text{mask}}^{(l)}$ denotes the binary validity mask at layer l, $f_{\text{mask}}^{(0)} = M$ represents the initial valid mask derived from the raw LiDAR projection, and s is the downsampling stride.

Specifically, we design a lightweight Guide module (illustrated in Fig. 6) to perform this cross-modal injection. Let f_{img} and f_{lidar} denote the extracted image and LiDAR features at a given scale, respectively. The Guide module concatenates these multi-modal features and processes them through a standard convolutional block. Crucially, to prevent the dense image semantics from erroneously diffusing into the invalid geometric regions, the fused output is tightly bounded by the geometric mask f_{mask}. The operation is formulated as:

$$f_{\text{out}} = \sigma(\text{BN}(\text{Conv}([f_{\text{img}}, f_{\text{lidar}}]))) \odot f_{\text{mask}}, \tag{3}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation, Conv represents a 3×3 convolution, BN is Batch Normalization, σ is the ReLU activation function, and $\odot$ denotes element-wise multiplication. By enforcing this mask multiplication, the Guide module ensures that semantic injection is rigorously confined to valid geometric pixels.

Finally, to enable plug-and-play compatibility with various LSS-based architectures, the densified features are processed by a LiteNeck. It structurally decouples the representation into Depth Logits for categorical depth distribution and Context Features for downstream 3D detection. This design allows DeGuNet to be directly integrated into standard Lift-Splat-Shoot modules without requiring any architectural modifications to the baseline.

Table 1: Comparison of different backbones pretrained on the KITTI depth dataset [13] and fine-tuned for downstream 3D detection.

Backbone	Depth RMSE (m) ↓ *(Pre-train Task)*	3D Det (mAP) ↑ *(Downstream)*	Img Params ↓
ResNet	3.44	62.11	∼ 0.32M
MobileViT-XS	2.54	63.27	∼ 2.36M
MobileViT-XXS	3.71	62.09	∼ 1.33M
Swin-T	1.83	64.01	∼ 31.82M
DeGuNet (Ours)	**0.74**	**69.40**	**∼ 0.31M**

5 Experiments

5.1 Experimental Setup

Dataset and Evaluation Metrics. We evaluate our method on the nuScenes dataset [2]. It contains 1,000 driving scenes divided into 28,130 training and 6,019 validation samples, providing 360° LiDAR points and synchronized 6-camera images. Standard evaluation metrics include mean Average Precision (mAP) and nuScenes Detection Score (NDS). To comprehensively assess the efficacy of the proposed lightweight module, we also evaluate system-level metrics, including the total parameter count and inference Frames Per Second (FPS).

Implementation Details. The framework is implemented based on MMDetection3D [33]. Point clouds are voxelized within [-54m, 54m] for the X and Y axes, and [-5m, 3m] for the Z axis, with a grid resolution of 0.075 m × 0.075 m × 0.2 m. Multi-view input images are resized to a resolution of 256×704. DeGuNet is then integrated as a drop-in replacement for the original image backbone in the selected frameworks, keeping the subsequent fusion and detection heads unchanged.

For all baseline comparisons, we follow the official open-source implementations and their official configurations, including pretraining recipes, learning rates, optimizers, and training schedules. For inference speed evaluation, FPS is measured on a single NVIDIA A100 GPU with batch size 1, using each method's recommended best-performing inference setup.

5.2 The Superiority of Depth-Guided Architecture

The Necessity of Sparsity-Aware Architecture. A seemingly intuitive approach to bridge the modality gap is to pre-train standard networks on a depth completion task. However, empirical results in Table 1 reveal the inherent limitations of this naive substitution. When pretrained on KITTI depth data, standard convolutional networks and lightweight transformers with comparable parameter counts (e.g., ResNet and MobileViT-XXS) exhibit high Root Mean Square Errors (3.44 m and 3.71 m, respectively) and yield limited downstream 3D detection mAP (62.11 and 62.09). Even when scaling up the parameter count significantly—such as with Swin-T (31.82M parameters)—the geometric alignment

Table 2: Comparison between standard Auxiliary Depth training strategies and the proposed DeGuNet across different baselines.

Framework	mAP ↑	FPS ↑	Params ↓
BEVFusion [26]+ BEVDepth	64.8	1.1	20.7M
BEVFusion [30] + BEVDepth	62.8	4.2	19.6M
GraphBEV [42] + BEVDepth	68.7	4.7	12.7M
BEVFusion [26] + **Ours**	<u>70.3</u>	4.8	15.3M
BEVFusion [30] + **Ours**	69.4	**5.1**	**11.1M**
GraphBEV [42] + **Ours**	**70.4**	<u>5.0</u>	<u>12.2M</u>

remains sub-optimal, yielding an RMSE of 1.83 m and an mAP of 64.01. This indicates that standard spatial operators are ill-equipped to process unstructured LiDAR projections without extensive zero-value contamination. In contrast, DeGuNet, incorporating dedicated MPIR and MMViT blocks, prevents invalid-region pollution. With merely 0.31M parameters, it achieves a substantially lower RMSE of 0.74 m and a superior downstream mAP of 69.40, proving that a structurally sparsity-aware design is a prerequisite for effective geometric alignment.

Plug-and-Play Generalization vs. Auxiliary Depth. Rather than shifting the pretraining objective, many existing frameworks employ an auxiliary depth paradigm (e.g., BEVDepth), which introduces additional depth supervision during the downstream detector training. As shown in Table 2, forcibly appending auxiliary depth losses to existing frameworks yields suboptimal gains and leaves the system burdened by the heavyweight 2D image backbone. For instance, BEVFusion [26] with BEVDepth only reaches 64.8 mAP at 1.1 FPS. Conversely, replacing the original backbone entirely with the pretrained DeGuNet functions establishes a direct, parameter-efficient geometric injection mechanism. Across established baselines (BEVFusion and GraphBEV), integrating DeGuNet consistently elevates the mAP beyond 69.0 while accelerating the inference speed and reducing the overall parameter footprint (e.g., dropping GraphBEV's total parameters to 12.2M).

5.3 Main Results and System Efficiency

Performance and Parameter Efficiency. We evaluate DeGuNet against state-of-the-art multi-modal 3D detectors on the nuScenes validation set (Table 3). To rigorously quantify the architectural overhead, we introduce Parameter Efficiency (Eff.), defined as the absolute mAP improvement per million image-side parameters. It is calculated as $\frac{\text{Multi-modal mAP} - \text{LiDAR-only mAP}}{\text{Image Parameters (M)}}$. In this context, the absolute improvement over each method's respective baseline is denoted as Gain. Furthermore, the absolute capacity of the image branch (comprising both backbone and neck components) and its proportion relative to the entire system are explicitly reported as Img Params and Img Param Ratio, respectively.

As detailed in Table 3, conventional frameworks rely on parameter-intensive 2D pretrained backbones consuming roughly 29M to 78M parameters. While

Table 3: Performance comparison on the nuScenes validation set. Detailed definitions for Parameter Efficiency (Eff.) and Gain are provided in Sec. 5.3. All performance gains are computed against each method's own LiDAR-only baseline implementation.

Method	Img Backbone	Img Params ↓	Total Params ↓	Img Param Ratio ↓	mAP ↑	NDS ↑	Eff. / Gain ↑
BEVFusion'22 [26]	CBSwin-T	78.1M	90.2M	86.61%	69.60	72.1	0.06 / 4.70
DeepInteraction++'25 [55]	Swin-T	33.2M	62.8M	52.95%	70.63	**73.3**	0.08 / 2.60
EA-LSS'23 [18]	CBSwin-T	78.33M	153.7M	50.96%	70.90	72.8	0.08 / 6.00
SparseFusion'23 [48]	Swin-T	30.3M	43.8M	69.18%	70.40	72.8	0.09 / 2.70
FocalFormer3D'23 [5]	InternImage	44.3M	65.0M	68.15%	70.50	73.1	0.09 / 4.10
UVTR'22 [21]	R101	47.7M	88.8M	53.70%	65.40	70.2	0.09 / 4.60
BEVFusion'23 [30]	Swin-T	31.8M	40.8M	77.94%	68.50	71.4	0.12 / 3.84
CMT'23 [49]	VoVNet	70.6M	86.7M	81.41%	70.30	72.9	0.12 / **8.16**
ObjectFusion'23 [3]	Swin-T	31.8M	43.9M	72.43%	69.80	72.3	0.14 / 4.60
DeepInteraction'22 [54]	Swin-T	28.7M	57.9M	49.61%	69.85	72.7	0.17 / 4.80
GraphBEV'24 [42]	Swin-T	31.8M	42.8M	74.30%	70.10	72.9	0.17 / 5.44
IS-Fusion'24 [56]	Swin-T	29.1M	48.8M	59.63%	71.00	72.7	0.20 / 5.90
UniTR'23 [45]	Shrunk ViT	9.4M	12.2M	76.49%	70.00	73.1	0.29 / 3.60
AutoAlignV2'22 [7]	CSPNet [44]	4.0M	13.2M	30.30%	67.10	71.2	0.92 / 3.69
BEVFusion'22 [26] + **Ours**	DeGuNet	1.11M	15.3M (**-74.9M**)	7.25%	70.30 (**+0.70**)	72.5 (**+0.4**)	4.86 / 5.40
EA-LSS'23 [18] + **Ours**	DeGuNet	17.0M	92.1M (**-61.6M**)	18.46%	71.10 (**+0.20**)	72.9 (**+0.1**)	0.36 / 6.20
GraphBEV'24 [42] + **Ours**	DeGuNet	0.81M	12.2M (**-30.6M**)	6.64%	70.40 (**+0.30**)	72.9 (**+0.0**)	7.09 / 5.74
BEVFusion'23 [30] + **Ours**	DeGuNet	**0.31M**	**11.1M** (**-29.7M**)	2.79%	69.40 (**+0.90**)	72.7 (**+1.3**)	15.29 / 4.74
IS-Fusion'24 [56] + **Ours**	DeGuNet	**0.31M**	19.7M (**-29.1M**)	**1.57%**	**71.30** (**+0.30**)	73.0 (**+0.3**)	**20.00** / 6.20

these heavy architectures provide absolute mAP gains, their parameter efficiency remains exceedingly low, generally falling below 0.30. This indicates a systematic underutilization of parameters when forcing 2D semantic extractors to align with 3D geometric spaces.

In contrast, substituting standard visual backbones with DeGuNet in established baselines yields consistently higher detection performance while systematically eliminating parameter redundancy. On EA-LSS [18], DeGuNet reduces the total/image-side parameters from 153.7M/78.33M to 92.1M/17.0M, while improving mAP/NDS from 70.9/72.8 to 71.1/72.9. On IS-Fusion [56], it reduces the total/image-side parameters from 48.8M/29.1M to 19.7M/0.31M, while improving mAP/NDS from 71.0/72.7 to 71.3/73.0. The variation in reported image-side parameters comes from baseline-specific pre-fusion camera modules and the adaptive dimension transformation within LiteNeck, while the core DeGuNet backbone remains strictly at 0.31M parameters. Consequently, DeGuNet achieves positive gains across all tested baselines and reaches an Eff. score of 20.00 on IS-Fusion, validating the effectiveness of depth-guided representations. Beyond parameter reduction, this structural streamlining translates to significant computational gains (Table 4). Compared to the BEVFusion baseline, DeGuNet accelerates inference to 5.1 FPS (a 1.16× speedup) and drastically reduces GPU memory consumption from 20.46GB to 6.86GB (a 66.5% reduction), proving its superior deployment efficiency.

5.4 Ablation Studies

Effectiveness of Core Components. We dissect the individual contributions of DeGuNet's structural components through incremental addition and targeted

Table 4: Computational efficiency comparison on an A100 GPU. The table details DeGuNet's performance in inference speed, memory consumption, and image feature extraction time.

Method	FPS ↑	Memory ↓	Img Backbone ↓	Total Img ↓	Speedup ↑	Memory Reduction ↑
DeepInteraction [54]	0.2	43.11GB	1201.6ms	2162.7ms	0.045×	-110.8%
BEVFusion [26]	0.3	41.01GB	1258.1ms	2236.3ms	0.068×	-100.5%
FUTR3D [4]	2.2	24.03GB	221.8ms	221.8ms	0.50×	-17.4%
UVTR [21]	2.9	23.36GB	109.1ms	110.7ms	0.66×	-14.2%
TransFusion-LC [1]	4.0	22.17GB	18.04ms	**18.04ms**	0.91×	-8.4%
BEVFusion [30]	4.4	20.46GB	21.12ms	51.58ms	1.0×	-
DeGuNet (Ours)	**5.1**	**6.86GB**	**7.59ms**	40.76ms	**1.16×**	**66.5%**

Table 5: Incremental addition of DeGuNet components starting from the LiDAR-only baseline.

Components	mAP ↑	ΔmAP ↑	Params ↓
LiDAR Baseline	64.66	–	**10.30M**
+ MPIR blocks	66.14	+1.48	10.32M
+ Guide modules	67.24	+2.58	10.41M
+ MMViT	68.38	+3.72	10.61M
+ LiteNeck	**69.40**	**+4.74**	11.10M

removal, using the established **BEVFusion'23** [30] framework as our foundational baseline. As presented in Tables 5 and 6, starting from the LiDAR-only baseline (64.66 mAP), integrating MPIR blocks yields a 1.48 mAP improvement, validating the necessity of mask-guided partial convolutions in preserving local geometric features from zero-value contamination. The subsequent addition of progressive Guide modules and MMViT blocks further increases mAP to 67.24 and 68.38, respectively. This confirms that both dense semantic injection and mask-aware global attention are critical for robust representation learning. Finally, the LiteNeck module decouples the output dimensions for LSS projection, finalizing the mAP at 69.40. The removal analysis corroborates these findings; systematically stripping any sparsity-aware module incurs significant performance degradation. Notably, reverting the entire backbone to standard convolutions (Full Standard Conv.) results in a 1.39 mAP drop, reaffirming that standard spatial operators cannot adequately handle sparse geometric projections.

Progressive Fusion Strategy. Table 7 ablates the cross-modal guidance mechanism across different encoder stages. The baseline configuration (66.14 mAP), which utilizes MPIR blocks but lacks semantic injection, serves as the lower bound. Activating the Guide module at single isolated stages (e.g., 1/2 or 1/4 resolution) provides marginal improvements (+0.44 and +0.75 mAP, respectively). However, employing the progressive fusion strategy across all three hierarchical stages yields the optimal performance of 67.24 mAP (+1.10 over the unguided baseline). This demonstrates that multi-scale, continuous injection of dense high-resolution textures is mathematically essential for maximizing the structural densification of the sparse LiDAR stream.

Table 6: Component removal analysis demonstrating the performance drop when removing modules from the full DeGuNet.

Components	mAP ↑	ΔmAP ↑	Params ↓
Full w/o LiteNeck	**68.38**	**-1.02**	**10.61M**
Full w/o MMViT	67.24	-2.16	10.95M
Full w/o Guide	66.14	-3.26	11.05M
Full w/o MPIR	66.59	-2.81	11.13M
Full (Standard Conv.)	68.01	-1.39	11.10M

Table 7: Ablation study on our progressive fusion strategy. Fusing features at all three encoder stages yields the best performance.

Fusion Stages Active	mAP ↑	ΔmAP ↑	Params (K)
None (Baseline)	66.14	–	0
Stage 1/2 only	66.58	+0.44	+4.7
Stage 1/4 only	66.89	+0.75	+18.6
Stage 1/8 only	66.74	+0.60	+74.0
Stages 1/2 + 1/4	67.03	+0.89	+23.3
Stages 1/4 + 1/8	67.14	+1.00	+92.6
All (Ours)	**67.24**	**+1.10**	+97.2

6 Conclusion

In this work, we address a critical inefficiency and structural misalignment in multi-modal 3D object detection: the heavy reliance on heavyweight, 2D-pretrained visual backbones. Through empirical analysis, we demonstrate that standard 2D semantic priors are intrinsically misaligned with 3D geometry, and that simply shifting to a depth-centric pretraining task is insufficient without a specialized architecture to handle unstructured sparse data. To resolve this, we propose DeGuNet, a modular image backbone explicitly designed for geometry-guided representation learning. By integrating MPIR blocks and progressive Guide modules, DeGuNet fuses multi-view images with sparse LiDAR geometry while rigorously preventing invalid-region contamination. Extensive experiments on the nuScenes dataset demonstrate that integrating DeGuNet into established baselines effectively eliminates parameter redundancy and accelerates inference speed, while simultaneously achieving superior 3D detection accuracy. While efficient, the minimalist design inherently limits its capacity for fine-grained semantic abstraction on rare long-tail categories, presenting a valuable direction for future exploration. We believe this work highlights the necessity of structurally aligned, sparsity-aware feature extraction, offering valuable insights for the future development of multi-modal perception systems.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In: CVPR. pp. 1090–1099 (2022)

2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: Nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
3. Cai, Q., Pan, Y., Yao, T., Ngo, C.W., Mei, T.: ObjectFusion: Multi-modal 3D object detection with object-centric fusion. In: ICCV. pp. 18067–18076 (2023)
4. Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H.: Futr3d: A unified sensor fusion framework for 3D detection. In: CVPRW. pp. 172–181 (2023)
5. Chen, Y., Yu, Z., Chen, Y., Lan, S., Anandkumar, A., Jia, J., Alvarez, J.M.: Focalformer3d : Focusing on hard instance for 3D object detection. In: ICCV. pp. 8394–8405 (2023)
6. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: VoxelNeXt: Fully sparse VoxelNet for 3D object detection and tracking. In: CVPR. pp. 21674–21683 (2023)
7. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F.: AutoAlignV2: Deformable feature aggregation for dynamic multi-modal 3D object detection. In: ECCV. pp. 628–644 (2022)
8. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE TPAMI **42**(10), 2361–2379 (2019)
9. Fan, L., Pang, Z., Zhang, T., Wang, Y.X., Zhao, H., Wang, F., Wang, N., Zhang, Z.: Embracing single stride 3D object detector with sparse transformer. In: ICCV. pp. 8458–8468 (2022)
10. Fan, L., Wang, F., Wang, N., Zhang, Z.: Fully sparse 3D object detection. In: NeurIPS. pp. 351–363 (2023)
11. Feng, T., Quan, R., Wang, X., Wang, W., Yang, Y.: Interpretable3D: An ad-hoc interpretable classifier for 3D point clouds. In: AAAI. pp. 1761–1769 (2024)
12. Feng, T., Wang, W., Ma, F., Yang, Y.: LSKNet: Towards effective and efficient 3D perception with large sparse kernels. In: CVPR. pp. 14916–14927 (2024)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR. pp. 3354–3361 (2012)
14. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: ICML. pp. 1263–1272 (2017)
15. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: CVPR. pp. 9224–9232 (2018)
16. He, C., Li, R., Li, S., Zhang, L.: Voxel set transformer: A set-to-set approach to 3D object detection from point clouds. In: CVPR. pp. 8417–8427 (2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
18. Hu, H., Wang, F., Su, J., Wang, Y., Hu, L., Fang, W., Xu, J., Zhang, Z.: EA-LSS: Edge-aware lift-splat-shot framework for 3D BEV object detection (2023)
19. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: CVPR. pp. 12697–12705 (2019)
20. Li, X., Yin, J., Shi, B., Li, Y., Yang, R., Shen, J.: LWSIS: LiDAR-guided weakly supervised instance segmentation for autonomous driving. In: AAAI. pp. 1433–1441 (2023)
21. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3D object detection. In: NeurIPS. pp. 18442–18455 (2022)
22. Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., et al.: Deepfusion: LiDAR-camera deep fusion for multi-modal 3D object detection. In: CVPR. pp. 17182–17191 (2022)

23. Li, Y., Fan, L., Liu, Y., Huang, Z., Chen, Y., Wang, N., Zhang, Z.: Fully sparse fusion for 3D object detection. IEEE Trans. Pattern Anal. Mach. Intell. **46**(11), 7217–7231 (2024)
24. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In: AAAI. pp. 1477–1485 (2023)
25. 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 TPAMI **47**(3), 2020–2036 (2025)
26. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: BEVFusion: A simple and robust LiDAR-camera fusion framework. In: NeurIPS. pp. 10421–10434 (2022)
27. Liu, F., Huang, T., Zhang, Q., Yao, H., Zhang, C., Wan, F., Ye, Q., Zhou, Y.: Ray denoising: Depth-aware hard negative sampling for multi-view 3D object detection. In: ECCV. pp. 200–217 (2024)
28. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: ECCV. pp. 85–100 (2018)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
30. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA. pp. 2774–2781 (2023)
31. Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., Xu, C.: Voxel transformer for 3D object detection. In: ICCV. pp. 3164–3173 (2021)
32. Mehta, S., Rastegari, M.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
33. MMDetection3D Contributors: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, accessed 25 June 2026 (2020)
34. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: ECCV. pp. 194–210 (2020)
35. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: CVPR. pp. 652–660 (2017)
36. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS. pp. 5099–5108 (2017)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
38. Shi, G., Li, R., Ma, C.: PillarNet: Real-time and high-performance pillar-based 3D object detection. In: ECCV. pp. 35–52 (2022)
39. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In: CVPR. pp. 10529–10538 (2020)
40. Shi, W., Rajkumar, R.: Point-GNN: Graph neural network for 3D object detection in a point cloud. In: CVPR. pp. 1711–1719 (2020)
41. Song, Z., Wei, H., Bai, L., Yang, L., Jia, C.: GraphAlign: Enhancing accurate feature alignment by graph matching for multi-modal 3D object detection. In: ICCV. pp. 3358–3369 (2023)

42. Song, Z., Yang, L., Xu, S., Liu, L., Xu, D., Jia, C., Jia, F., Wang, L.: GraphBEV: Towards robust BEV feature alignment for multi-modal 3D object detection. In: ECCV. pp. 347–366 (2024)
43. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant CNNs. In: 3DV. pp. 11–20 (2017)
44. Wang, C.Y., Mark Liao, H.Y., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: CSPNet: A new backbone that can enhance learning capability of CNN. In: CVPRW. pp. 1571–1580 (2020)
45. Wang, H., Tang, H., Shi, S., Li, A., Li, Z., Schiele, B., Wang, L.: UniTR: A unified and efficient multi-modal transformer for bird's-eye-view representation. In: ICCV. pp. 6792–6802 (2023)
46. 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: CVPR. pp. 8445–8453 (2019)
47. Wang, Y., Li, B., Zhang, G., Liu, Q., Gao, T., Dai, Y.: LRRU: Long-short range recurrent updating networks for depth completion. In: CVPR. pp. 9422–9432 (2023)
48. Xie, Y., Xu, C., Rakotosaona, M.J., Rim, P., Tombari, F., Keutzer, K., Tomizuka, M., Zhan, W.: SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. In: ICCV. pp. 17545–17556 (2023)
49. Yan, J., Liu, Y., Sun, J., Jia, F., Li, S., Wang, T., Zhang, X.: Cross modal transformer via coordinates encoding for 3D object detection. In: ICCV. pp. 18222–18232 (2023)
50. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
51. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: RigNet: Repetitive image guided network for depth completion. In: ECCV. pp. 214–230 (2022)
52. Yang, L., Tang, T., Li, J., Yuan, K., Wu, K., Chen, P., Wang, L., Huang, Y., Li, L., Zhang, X., Yu, K.: BEVHeight++: Toward robust visual centric 3D object detection. IEEE TPAMI **47**(6), 5094–5111 (2025)
53. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: CVPR. pp. 11040–11048 (2020)
54. Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L.: DeepInteraction: 3D object detection via modality interaction. In: NeurIPS. pp. 1992–2005 (2022)
55. Yang, Z., Song, N., Li, W., Zhu, X., Zhang, L., Torr, P.H.: DeepInteraction++: Multi-modality interaction for autonomous driving. IEEE TPAMI **47**(8), 6749–6763 (2025)
56. Yin, J., Shen, J., Chen, R., Li, W., Yang, R., Frossard, P., Wang, W.: Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14905–14915 (2024)
57. Zhou, W., Yan, X., Liao, Y., Lin, Y., Huang, J., Zhao, G., Cui, S., Li, Z.: BEV@DC: Bird's-eye view assisted training for depth completion. In: CVPR. pp. 9233–9242 (2023)
58. Zhou, Y., Tuzel, O.: VoxelNet: End-to-end learning for point cloud based 3D object detection. In: CVPR. pp. 4490–4499 (2018)
59. Zong, Z., Jiang, D., Song, G., Xue, Z., Su, J., Li, H., Liu, Y.: Temporal enhanced training of multi-view 3D object detector via historical object prediction. In: ICCV. pp. 3758–3767 (2023)