SparseLIF: High-Performance Sparse LiDAR-Camera Fusion for 3D Object Detection

Hongcheng Zhang[1*], Liu Liang[1*], Pengxin Zeng[1,2*], Xiao Song[1†], and Zhe Wang[1]

[1] SenseTime Research
[2] College of Computer Science, Sichuan University
{zhanghongcheng,liangliu1}@senseauto.com, zengpengxin.gm@gmail.com, {songxiao,wangzhe}@senseauto.com

Abstract. Sparse 3D detectors have received significant attention since the query-based paradigm embraces low latency without explicit dense BEV feature construction. However, these detectors achieve worse performance than their dense counterparts. In this paper, we find the key to bridging the performance gap is to enhance the awareness of rich representations in two modalities. Here, we present a high-performance fully sparse detector for end-to-end multi-modality 3D object detection. The detector, termed SparseLIF, contains three key designs, which are (1) Perspective-Aware Query Generation (PAQG) to generate high-quality 3D queries with perspective priors, (2) RoI-Aware Sampling (RIAS) to further refine prior queries by sampling RoI features from each modality, (3) Uncertainty-Aware Fusion (UAF) to precisely quantify the uncertainty of each sensor modality and adaptively conduct final multi-modality fusion, thus achieving great robustness against sensor noises. By the time of paper submission, SparseLIF achieves state-of-the-art performance on the nuScenes dataset, ranking ***1st*** on both validation set and test benchmark, outperforming all state-of-the-art 3D object detectors by a notable margin.

Keywords: 3D Object Detection · Sparse Detector · LiDAR-Camera Fusion

1 Introduction

LiDAR-camera-based 3D detection is essential for accurate and robust autonomous driving systems. The two modalities naturally provide complementary information, *i.e.*, the camera offers high-resolution semantic information while LiDAR provides accurate geometric information. Therefore, camera and LiDAR sensors have been simultaneously deployed for reliable 3D object detection.

Various approaches have been proposed to thoroughly explore the compensating information in LiDAR and camera modalities. Conventional multi-modality

[*] Equal Contribution

3D object detection approaches typically transform two modalities into a unified space for feature fusion. For example, PointPainting [62] and its variants [63, 70, 78] decorate raw point clouds with image pixel features. BEVFusion [34, 44] transforms image view features into dense BEV space to fuse with LiDAR features. The dense paradigm has achieved remarkable success in recent years but suffers from cumbersome view transformation, resulting in high latency, limited detection distance, and limited upper-bound performance. Recent works introduce a sparse query-based paradigm without explicit view transformation. Some pioneering sparse detectors aggregate multi-modality features in one [64, 71] or two [1] stages using global attention. However, the exhaustive global attention buries the advantages of the sparse paradigm and makes it difficult to benefit from long-term temporal information. Lately, a stream of works explores the fully sparse paradigm, which is free from the usage of global attention and dense BEV queries. For example, works like FUTR3D [7] and DeepInteraction [75] sample features from two modalities using reference points. Despite the huge advances, these methods still lag behind their dense counterparts. Thus, whether fully sparse multi-modality detectors can achieve superior performance compared to dense detectors remains an open question.

This paper presents SparseLIF, a high-performance fully sparse multi-modality 3D object detector that outperforms all other dense counterparts and sparse detectors. SparseLIF bridges the performance gap by enhancing the awareness of rich LiDAR and camera representations in three aspects, *i.e.*, query generation, feature sampling and multi-modality fusion. **First**, we argue that the convention [68], which randomly generates queries, will suffer from extra efforts in learning to move the query proposals towards ground-truth targets. Here, we propose the Perspective-Aware Query Generation (PAQG) module to ease learning. In particular, PAQG injects a lightweight perspective detector composed of the coupled 2D and monocular-3D sub-networks on image features to predict and transform top-scored 3D proposals into query proposals. These input-dependent proposals will narrow the learning path toward ground-truth targets, thus enhancing the awareness of rich contexts in high-resolution images. **Second**, these queries with perspective priors will interact with features from two modalities via the RoI-Aware Sampling (RIAS) module. Instead of resorting to cumbersome global attention, the module locates the region of interest and then sample complementary features at merely several reference points under the guidance of prior queries, thus conforming to the fully sparse paradigm and enjoying low latency. **Third**, we observe that in realistic scenarios, LiDAR and camera usually suffer from various sensor problems as shown in Fig. 3, which will make sensor inputs unreliable and uncertain, thus degrading the performance of multi-modality detectors. Hence, we propose the Uncertainty-Aware Fusion (UAF) module to precisely quantify the uncertainty of each modality and guide our model to focus on the trustworthy modality in multi-modality fusion, thus achieving great robustness against sensor noises. Our contributions are summarized as follows.

- We point out that the key to bridging the performance gap between sparse detectors and their dense counterparts is to enhance the awareness of rich

representations from LiDAR and camera feature spaces in three aspects, *i.e.*, query generation, feature sampling, and multi-modality fusion.
- We present a high-performance fully sparse detector for LiDAR-camera-based 3D object detection. The proposed framework contains three key designs: (1) Perspective-Aware Query Generation (PAQG), which enhances the perspective awareness of query proposals on rich contexts in high-resolution images; (2) RoI-Aware Sampling (RIAS), which effectively refine prior queries by sampling complementary RoI features across two modalities; (3) Uncertainty-Aware Fusion (UAF), which conducts final multi-modality fusion under the guidance of quantified modality uncertainty.
- We conduct comprehensive experiments to demonstrate the effectiveness of our proposed method. As can be seen, SparseLIF outperforms all state-of-the-art 3D object detectors on the nuScenes dataset, ranking ***1st*** on both the validation set and test benchmark.

2 Related Work

This section briefly reviews the most related works on three topics: LiDAR-, Camera- and LiDAR-Camera-based 3D object detection.

2.1 LiDAR-based 3D Object Detection

LiDAR provides accurate geometric information, attracting much attention for single-modality 3D detection. Earlier methods [8,32,51–53,58,74] directly extract features from raw point clouds to predict 3D bounding boxes, but suffers from the complexity when processing large-scale point clouds. Modern approaches transform unordered points into structured formats such as range-view maps [2, 14,29,35,46,60], pillars [27], voxels [13,57,72,82]. Then, main-stream approaches apply 2D/3D convolution-based head [27, 77, 81] to predict 3D bounding boxes. Inspired by the huge success made by transformers, some recent works adopt transformer blocks in feature encoder [45, 56, 80] and 3D detection head [1].

2.2 Camera-based 3D Object Detection

Camera-based 3D object detection [21,22,31,33,42,49,68] has witnessed remarkable progress over the past few years since camera-based approaches have lower deployment cost compared with the LiDAR-based counterparts.

Inspired by the huge success made by LiDAR-based 3D detection methods, Pseudo-LiDAR [67] transforms images into pseudo-LiDAR point clouds via depth estimation, then conducts 3D object detection on those pseudo points with LiDAR-based approaches. A line of works (*e.g.* DD3D [48], FCOS3D [66] and CenterNet [81]) further propose end-to-end, single stage 3D object detectors by attaching extra 3D bounding box regression head to 2D detector. Those methods attempt to explicitly estimate depth to assist in 3D detection but show limited performance due to inaccurate depth estimation.

To implicitly incorporate depth information, another line of works [49,54,55] perform 3D detection in BEV space. LSS [50] predicts the categorical depth distribution for each pixel to lift pixel features into a frustum, then splats all frustums into BEV grids. Based on LSS [50], BEVDet [22] and BEVDepth [31] substantially boost performance. Inspired by transformer, BEVformer [33, 73] and VideoBEV [17] directly extract spatial features from camera views using cross-attention. Without reliance on depth information, the explicit construction of dense BEV features still limits inference speed and effective detection distance.

Another stream of works [25] employ a top-down manner that does not suffer from the explicit construction of dense BEV features. Inspired by DETR [5], DETR3D [68] manipulates predictions directly in 3D space by indexing 2D features with a sparse set of 3D object queries. PETR [42] further eases the overhead of the indexing operation. PETRv2 [43] and Stream PETR [65] utilize the temporal information of previous frames to boost 3D object detection but adopt the global cross attention, which is computationally expensive. Sparse4D [37–39] and SparseBEV [40] sparsely sample multi-frame/view/scale features for 4D reference points then fuse hierarchically, thus achieving 3D detection without relying on dense view transformation and global attention.

2.3 LiDAR-Camera-based 3D Object Detection

Recently, LiDAR-Camera-based 3D detection [6] has achieved great success in leveraging semantic and geometric information to reach impressive performance. Early approaches [24,59,62,63,78] decorate raw point clouds with image features but compromise rich context information. FrustumPointNet [51], FrustumConvNet [69], and CenterFusion [47] lift image proposals into 3D frustums with explicit depth estimation but show limited performance due to depth inaccuracy.

Lately, motivated by LSS [50], BEVFusion [34,44] ease the reliance on depth estimation by projecting fine-grained image features into BEV space then conducting fusion with LiDAR features. AutoAlign [10,11] further preserves instance-wise semantic consistency by feature alignment across two modalities. However, the explicit and dense view transformation from image to BEV space is cumbersome (*i.e.*, high latency and limited detection distance) and sensitive to sensor misalignment. BEVFusion4D [4] further improves performance by incorporating temporal information. EA-LSS [20] enhances depth estimation at the edge of objects.

Recent works utilize the sparse query-based paradigm without explicit view transformation. Transfusion [61] obtains object queries from LiDAR points and then fuses queries with rich image features using a transformer block. CMT [71] further develops an end-to-end feature interaction framework for multi-modality fusion. UniTR [64] introduces a modality-agnostic transformer encoder to proceed with unified modeling and shared parameters. Despite great success, the expensive global attention buries the advantages of the sparse paradigm and makes it difficult to benefit from long-term temporal information.

Another stream of works explores the fully sparse paradigm. SparseFusion [83] poses detectors on each modality and fuses features of detected instances. How-

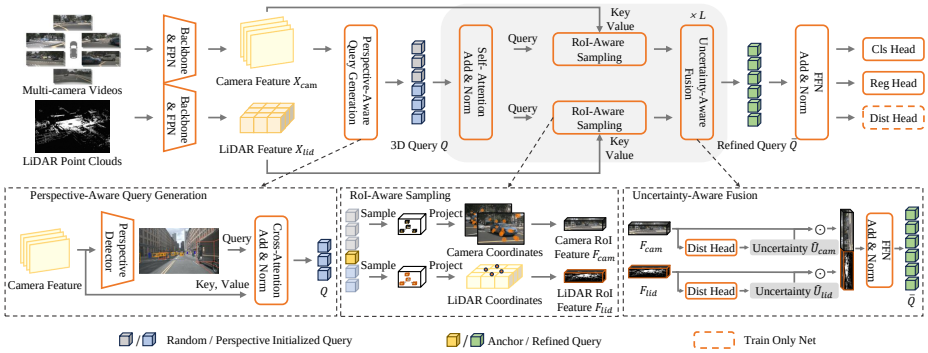

Fig. 1: The overall architecture of SparseLIF, a fully sparse LiDAR-camera-based 3D object detector. The framework contains a camera backbone to process multi-view videos and a LiDAR backbone to encode raw point clouds. We then feed the image features into the Perspective-Aware Query Generation (PAQG) module to generate queries. The queries will interact with the camera and LiDAR features via the RoI-Aware Sampling (RIAS) module to extract complementary features for further refinement. Next, the Uncertainty-Aware Fusion (UAF) module quantifies the uncertainty of RoI features from two modalities and adaptively conducts final multi-modality fusion. The decoder repeats L times.

ever, the two-stage paradigm suffers from the limited performance of modality-specific detectors. FUTR3D [7] generalizes the fully sparse paradigm by initializing 3D reference points and projecting them into all available modalities to sample features. Although methods have recently achieved good performance, there remains notable performance gap compared to dense counterparts.

3 SparseLIF

SparseLIF is a sparse query-based multi-modality detector. We use common image backbone (*e.g.* ResNet [18], V2-99 [28]) and FPN [36] to extract multi-view/scale/frame camera features, denoted as $X_{\text{cam}} = \{\mathcal{X}_{\text{cam}}^{vmt}\}_{v=1,m=1,t=1}^{V,M,T}$, where V, M, and T denote the number of camera views, feature scales, and temporal frames respectively. Based on our proposed framework, rich temporal information can be easily and sufficiently incorporated. In parallel, we use common 3D LiDAR backbone (*e.g.* VoxelNet [82]) and FPN [36] to extract multi-scale LiDAR features, denoted as $X_{\text{lid}} = \{\mathcal{X}_{\text{lid}}^{r}\}_{r=1}^{R}$, where R denotes the number of LiDAR feature scales. Taking camera features as input, the Perspective-Aware Query Generation (PAQG) module (Sec. 3.1) adopts the coupled 2D and monocular-3D image detectors to predict and generate high-quality 3D queries with perspective priors. These queries will then interact with the camera and LiDAR features via the RoI-Aware Sampling (RIAS) module to extract RoI features for further refinement. Next, the Uncertainty-Aware Fusion (UAF) module (Sec. 3.3) quantifies the uncertainty of RoI features from two modalities and adaptively conducts multi-modality fusion for final 3D object predictions.

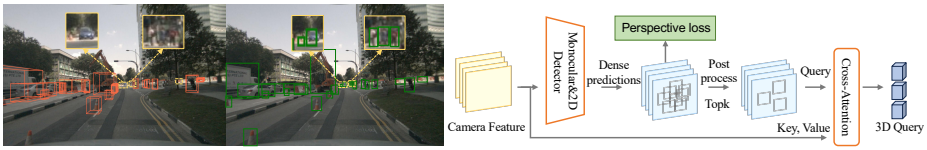

Fig. 2: Motivations and details of our proposed PAQG module. **(a)** 3D detectors struggle with low sensitivity when detecting distant and small objects. **(b)** 2D detectors demonstrate excellent pixel-wise perception capabilities on such objects. **(c)** the PAQG module adopts the coupled 2D and monocular-3D sub-networks to predict dense boxes under the supervision of a perspective loss. We pick top-ranked boxes to propose high-quality queries, and then interact with camera features via a cross-attention module.

3.1 Perspective-Aware Query Generation

Recent works typically generate queries based on randomly distributed reference points [7, 71], anchor boxes [37] or pillars [40] in 3D space and optimize as net parameters, regardless of input data. However, it has already been proved in 2D detection [76] that such input-independent queries will take extra effort in learning to move the query proposals towards ground-truth object targets. As shown in Fig. 2, we visualize the predictions of a query-based 3D detector and a 2D detector, where the 2D detector usually exhibits excellent perception capability on distant and small objects. Motivated by the strength of 2D detection, our PAQG module fully utilizes the perception capability to generate 3D queries, thereby assisting ultimate 3D detection.

The lightweight perspective detector in the PAQG module consists of the coupled 2D (*e.g.* FCOS [61]) and monocular-3D (*e.g.* FCOS3D [66]) sub-networks. Taken the multi-view/scale image features X_{cam} as input, the monocular-3D sub-network predicts raw 3D attributes, *i.e.*, depths $\mathbf{d}$, rotation angles, sizes, and velocities throughout different views. Simultaneously, the 2D sub-network predicts corresponding 2D attributes, *i.e.*, center coordinates $[\mathbf{c_x}, \mathbf{c_y}]$, confidence scores, and category labels. For each view v, we project the box centers into 3D space based on corresponding camera extrinsic E_v and intrinsic I_v, *i.e.*,

$$\mathbf{c}^{3D} = E_v^{-1} I_v^{-1} [\mathbf{c_x}\mathbf{d}, \mathbf{c_y}\mathbf{d}, \mathbf{d}, \mathbf{1}]. \tag{1}$$

The 3D center $\mathbf{c}^{3D}$ will combine with the predicted size, rotation angle, and velocity to form 3D boxes. Then, we perform non-maximum suppression in 3D space to filter intersecting boxes and pick the top N_k boxes ranked by confidence scores, to initialize queries with image features interacted via an efficient cross-attention module. Formally,

$$q_i = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{m=1}^{M} \mathcal{BS}(\mathcal{X}_{\text{cam}}^{vm}, \mathcal{P}_{\text{cam}}^{v}(c_i^{3D})), \tag{2}$$

where $\mathcal{P}_{\text{cam}}^{v}(c_i^{3D})$ projects the 3D center c_i^{3D} to v-th image using corresponding camera parameters. Besides, $\mathcal{V}$ denotes the set of hit views. $\mathcal{BS}(\cdot)$ denotes

the bilinear sampling function. Since some objects may be overlooked, we preserve N_r randomly initialized query boxes. Finally, our PAQG module generates total $N_q = N_k + N_r$ query proposals. This way, our PAQG module provides input-dependent query proposals to elevate the understanding of comprehensive perspective priors (2D and 3D attributes) for 3D detectors, thereby aiding in detecting distant and small objects.

3.2 RoI-Aware Sampling

RoI-Aware Sampling (RIAS) module is responsible for sampling RoI features from each modality to refine the queries $Q = \{q_i\}_{i=1}^{N_q} \subset \mathbb{R}^C$ initialized with perspective priors via PAQG module. We aim at locating the region of interest (RoI) to sample features without resorting to cumbersome global attention, thus enjoying low complexity and benefiting from long-term temporal information.

LiDAR Branch Inspaired by Deformable Attention [84], we merely sample $K = 4$ reference points to retrieval RoI features $\{F_{\text{lid}}^{ik}\}_{k=1}^{K}$ from LiDAR feature map $\mathcal{X}_{\text{lid}}$ for each query q_i. Formally,

$$F_{\text{lid}}^{ik} = \sum_{r=1}^{R} \mathcal{BS}\left(\mathcal{X}_{\text{lid}}^{r}, \mathcal{P}_{\text{lid}}\left(c_i + \Delta_{\text{lid}}^{irk}\right)\right) \cdot \sigma_{\text{lid}}^{irk}, \tag{3}$$

where c_i is the bounding box center of query q_i in global 3D space and $\mathcal{P}_{\text{lid}}$ projects the center into LiDAR BEV space. $\mathcal{BS}(\cdot)$ denotes the bilinear sampling function. Besides, $\Delta_{\text{lid}}^{irk}$ and $\sigma_{\text{lid}}^{irk}$ are predicted sampling offsets and attention weights using query q_i to cover the RoI on sensitive objects. Note that, different from global attention, we merely interact with several features mapped to reference points, thus embracing a fully sparse paradigm.

Camera Branch As for the camera branch, we also sample $K = 4$ reference points to retrieval RoI features from the hit views $\mathcal{V}$ of camera feature map $\mathcal{X}_{\text{cam}}$, *i.e.*,

$$F_{\text{cam}}^{itk} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{m=1}^{M} \mathcal{BS}\left(\mathcal{X}_{\text{cam}}^{vmt}, \mathcal{P}_{\text{cam}}^{vt}\left(c_i + \Delta_{\text{cam}}^{ivmtk}\right)\right) \cdot \sigma_{\text{cam}}^{ivmtk}, \tag{4}$$

where $\mathcal{P}_{\text{cam}}^{vt}(\cdot)$ is the project function from global 3D space to feature coordinate using camera parameters and temporal alignment [40]. Besides, $\Delta_{\text{cam}}^{ivmtk}$ and $\sigma_{\text{cam}}^{ivmtk}$ are also predicted sampling offsets and attention weights using query feature.

Channel-Spatial Correlation Aware Mixing To enrich the awareness of the correlation in spatial and channel dimensions of query q_i, we inject AdaMixer [15] on the retrieved features. For convenience, we organize those retrieved RoI features to $f \in \mathbb{R}^{S \times C}$, where $S = K$ or $S = T \times K$ for LiDAR or camera feature.

(a) Limited FoV (b) Object Failure (c) Camera Occlusion

Fig. 3: Visualizations of sensor noises in 3D object detection for autonomous driving. **(a) Limited FOV**: LiDAR installed in a front-facing manner yields a limited FOV, *e.g.* 120°. **(b) Object Failure**: the reflection rate of some objects (*e.g.* the black car) is below the threshold of LiDAR thus without LiDAR points reflected. **(c) Camera Occlusion**: the camera module is usually vulnerable to occlusions (*e.g.* by dust).

First, we model the channel correlation based on query q_i and transform features f to enhance channel semantics:

$$W_c = \text{Linear}(q_i) \in \mathbb{R}^{C\times C} \tag{5}$$

$$\text{M}_c(f) = \text{ReLU}\left(\text{LayerNorm}\left(fW_c\right)\right), \tag{6}$$

where W_c is the channel correlation shared across different timestamps and different sampling points. Next, we then transpose the feature and model the spatial correlation to the spatial dimension of it, *i.e.*,

$$W_s = \text{Linear}(q_i) \in \mathbb{R}^{S\times S} \tag{7}$$

$$\text{M}_s(f) = \text{ReLU}\left(\text{LayerNorm}\left(f^T W_s\right)\right), \tag{8}$$

where W_s is the spatial correlation shared across different channels. After channel-spatial correlation aware mixing, the features are flattened and aggregated by a linear layer.

3.3 Uncertainty-Aware Fusion

Given the RoI features F_{cam} and F_{lid} from two modalities, the Uncertainty-Aware Fusion (UAF) module aims to endow our fusion module with the robustness against sensor noise illustrated in Fig. 3. To this end, we inject the awareness of the uncertainty of each modality into our fusion module, *i.e.*,

$$\bar{Q} = f_{UA}(F_{\text{cam}}, U_{\text{cam}}, F_{\text{lid}}, U_{\text{lid}}), \tag{9}$$

where $\bar{Q} = \{\bar{q}_i\}_{i=1}^{N_q} \subset \mathbb{R}^C$ and f_{UA} are the refined query feature and uncertainty-aware fusion function, respectively. Besides, U_{cam} and U_{lid} are the uncertainty of two modalities.

Inspired by the unquestionable importance of accurate localization in autonomous driving, we formulate the uncertainty as a function of the Euclidean

distance between the predicted and the ground-truth bounding boxes B. For convinience, let $s \in \{\text{cam}, \text{lid}\}$ represents one modality. We have

$$U_s = 1 - exp\left(-D^{xy}\left(f_{\text{reg}}(F_{\text{s}}), B\right)\right), \tag{10}$$

where f_{reg} is the regression function for bounding boxes, and D^{xy} is the Euclidean distance function in BEV space. However, the ground-truth bounding boxes are unavailable for models. Thus, we inject a distance predictor on RoI features of each modality, then rewrite Eq. (10) as

$$\hat{U}_s = 1 - exp\left(-f_{\text{dist}}(F_{\text{s}})\right), \tag{11}$$

where f_{dist} is the distance predictor consisting of MLPs.

As for the uncertainty-aware fusion function f_{UA}, we simply formulate it as the concatenation of features weighted by uncertainty and rewrite Eq. (9) as

$$\bar{q}_i = FFN\left(Cat\left(F_{\text{cam}} \cdot (1 - \hat{U}_{\text{cam}}), F_{\text{lid}} \cdot (1 - \hat{U}_{\text{lid}})\right)\right), \tag{12}$$

where Cat and FFN denote concatenation function and feedforward networks, respectively. In this way, our UAF module quantifies the uncertainty U of each modality and guides our model to focus on trustworthy modality, thus enjoying robustness against sensor noises.

4 Experiments

This section provides the experimental settings and results. We conduct detailed ablation studies to verify our design choices in SparseLIF. Meanwhile, we also demonstrate that our multi-modality detector achieves excellent robustness against sensor noises. Above all, we compare SparseLIF with other state-of-the-art 3D object detectors on the popular nuScenes benchmark. The results show that our SparseLIF achieves superior performance, ranking ***1st*** on both the validation set and test benchmark.

4.1 Experimental Setups

Implementation Details We implement SparseLIF using the open-source MMDetection3D [12] based on PyTorch. The detection range is $[-54m, 54m]$ and $[-5m, 3m]$ for the XY- and the Z-axis. We adopt V2-99 [28] pretrained by FCOS3D [66] as the image backbone with input image size 1600×640. We adopt VoxelNet [82] as LiDAR backbone with voxel size $(0.075m, 0.075m, 0.2m)$. The total query number N_q is 900, including $N_k = 200$ queries generated by the PAQG module. The perspective detector is implemented by the coupled FCOS [61] and FCOS3D [66] sub-networks. The lightweight distance predictor f_{dist} is implemented by a two-layer FFN. The decoder repeats $L = 6$ times. In the following experiments, we report the state-of-the-art performance of two SparseLIF detectors: the single-frame detector ***SparseLIF-S*** ($V = 6$, $M = 4$,

Table 1: Quantitative comparisons of SparseLIF with all state-of-the-art 3D detectors on the nuScenes test benchmark. The notion of modality: Camera (C), LiDAR (L), and Temporal (T). †: using external training data; ‡: using TTA and complex model ensemble (*e.g.* models with different voxel sizes, BEV sizes, backbones/FPNs/heads); §: we only use very simple self-model ensemble without TTA for *SparseLIF-T*.

Method	Modality	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	mAP↑	NDS↑
TransFusion [1]	LC	25.9	24.3	35.9	28.8	12.7	68.9	71.7
FUTR3D [7]	LC	28.4	24.1	31.0	30.0	12.0	69.4	72.1
AutoAlignV2 [11]	LC	24.5	23.3	31.1	25.8	13.3	68.4	72.4
BEVFusion [44]	LC	26.1	23.9	32.9	26.0	13.4	70.2	72.9
BEVFusion [34]	LC	25.0	24.0	35.9	25.4	13.2	71.3	73.3
DeepInteraction [75]	LC	25.7	24.0	32.5	24.5	12.8	70.8	73.4
BEVFusion4D-S [4]	LC	-	-	-	-	-	71.9	73.7
SparseFusion [83]	LC	-	-	-	-	-	72.0	73.8
MSMDFusion [26]	LC	25.5	23.8	31.0	24.4	13.2	71.5	74.0
CMT [71]	LC	27.9	23.5	30.8	25.9	11.2	72.0	74.1
EA-LSS [20]	LC	24.7	23.7	30.4	25.0	13.3	72.2	74.4
UniTR [64]	LC	24.1	22.9	**25.6**	24.0	13.1	70.9	74.5
FocalFormer3D-F [9]	LC	25.1	24.2	32.8	22.6	12.6	72.4	74.5
BEVFusion4D [4]	LCT	-	-	-	-	-	73.3	74.7
DAL [23]	LC	25.3	23.8	33.4	17.4	12.0	72.0	74.8
FusionFormer† [19]	LCT	26.7	23.6	28.6	22.5	**10.5**	72.6	75.1
SparseLIF-T	LCT	**24.1**	**22.9**	27.8	**15.4**	11.8	**74.4**	**77.0**
PAI3D‡ [41]	LC	24.5	23.3	30.8	23.3	13.1	71.4	74.2
Lift-Attend-Splat‡ [16]	LC	24.3	23.8	34.5	32.8	13.3	75.5	74.9
BEVFusion‡ [44]	LC	24.2	22.7	32.0	22.2	13.0	75.0	76.1
DeepInteraction‡ [75]	LC	23.5	23.3	32.8	22.6	13.0	75.6	76.3
CMT‡ [71]	LC	23.3	**22.0**	**27.1**	21.2	12.7	75.3	77.0
BEVFusion4D‡ [4]	LCT	**22.9**	22.9	30.2	22.5	13.5	**76.8**	77.2
EA-LSS‡ [20]	LC	23.4	22.8	27.8	20.4	12.4	76.6	77.6
SparseLIF-T§	LCT	24.3	23.1	28.4	**15.2**	**11.7**	75.9	**77.7**

$R = 4$, and $T = 1$), the temporal multi-frame detector ***SparseLIF-T*** ($V = 6$, $M = 4$, $R = 4$, and $T = 13$).

Each model is trained end-to-end using the AdamW optimizer on eight NVIDIA A100 GPUs with a total batch size of 8. For fair comparisons, we apply the query-denoising strategy [30], commonly used in sparse detection heads, to address the unstable matching problem. Each model is trained for 24 epochs with a learning rate of $2e-4$.

Datasets and Evaluation Metrics We conduct experiments on the popular nuScenes dataset [3] to evaluate the performance of our proposed method for 3D object detection in autonomous driving. The nuScenes dataset has 1.4 million 3D detection annotation boxes from 40,157 samples distributed in 1000 scenes collected in Boston and Singapore. Each sample is collected with six cameras and

Table 2: Quantitative comparisons of SparseLIF with all state-of-the-art 3D detectors on the nuScenes validation set. The notion of modality: Camera (C), LiDAR (L), and Temporal (T). †: with extra CBGS training strategy. Note that all methods use corresponding best single model without TTA or model ensemble for comparisons.

Method	Modality	Image Backbone	LiDAR Backbone	mAP↑	NDS↑
FUTR3D [7]	LC	ResNet-101	VoxelNet	64.2	68.0
AutoAlignV2 [11]	LC	CSPNet	VoxelNet	67.1	71.2
TransFusion [1]	LC	ResNet-50	VoxelNet	67.5	71.3
BEVFusion [44]	LC	SwinT	VoxelNet	68.5	71.4
BEVFusion [34]	LC	SwinT	VoxelNet	69.6	72.1
DeepInteraction [75]	LC	ResNet-50	VoxelNet	69.9	72.6
CMT [71]	LC	V2-99	VoxelNet	70.3	72.9
BEVFusion4D-S [4]	LC	SwinT	VoxelNet	70.9	72.9
SparseFusion [83]	LC	SwinT	VoxelNet	71.0	73.1
EA-LSS [20]	LC	SwinT	VoxelNet	71.2	73.1
FusionFormer-S† [19]	LC	V2-99	VoxelNet	70.0	73.2
SparseLIF-S	LC	V2-99	VoxelNet	**71.2**	**74.6**
BEVFusion4D [4]	LCT	SwinT	VoxelNet	72.0	73.5
FusionFormer† [19]	LCT	V2-99	VoxelNet	71.4	74.1
SparseLIF-T	LCT	V2-99	VoxelNet	**74.7**	**77.5**

Table 3: Ablation studies of SparseLIF on the nuScenes validation set.

SparseLIF-S	Modality	Image Backbone	LiDAR Backbone	PAQG	RIAS	UAF	mAP↑	NDS↑
#1	LC	V2-99	VoxelNet				66.2	69.0
#2	LC	V2-99	VoxelNet		✓		69.8	73.3
#3	LC	V2-99	VoxelNet	✓	✓		71.0	74.3
#4	LC	V2-99	VoxelNet		✓	✓	70.5	74.1
#5	LC	V2-99	VoxelNet	✓	✓	✓	**71.2**	**74.6**
#6	LC	V2-99	VoxelNet	✓		✓	68.0	70.8

a 32-beam LiDAR sensor. We adopt the nuScenes detection evaluation metrics NDS and mAP over ten classes for our experiments.

4.2 Comparisons with State-of-the-Art 3D Object Detectors

As shown in the top part of Tab. 1, without using any test-time augmentation (TTA) or model ensemble, our *SparseLIF-T* achieves state-of-the-art single-model performance, reaching 77.0% NDS on the nuScenes test benchmark, significantly outperforming all other 3D detectors by a notable margin. In particular, we outperform the most competitive method FusionFormer [19] by 1.9% NDS without using any external training data. Regarding the test benchmark leaderboard in the bottom part of Tab. 1, many competitive methods [4, 16, 20, 41, 44, 71, 75] adopt very complex model ensemble (*e.g.* assembling models with different voxel sizes, BEV sizes, backbones/FPNs/heads) and TTA, to strive for top ranking on the test leaderboard. Contrarily, we only use very

Table 4: Performance analysis of our PAQG module on detection distances and small object classes on the nuScenes validation set, based on *SparseLIF-S*. The *AP* scores of traffic cone (T.C.) and barrier at 30*m*- are missing since corresponding annotations are unavailable.

	PAQG	0-10m	10-20m	20-30m	30m-
mAP		75.1	74.3	65.5	56.9
mAP	✓	**76.2**	**74.8**	**66.8**	**58.5**
T.C.		82.5	80.6	69.1	-
T.C.	✓	**83.8**	**82.7**	**70.4**	-
Barrier		72.1	77.5	54.5	-
Barrier	✓	**76.4**	**80.0**	**63.1**	-

simple self-model ensemble without TTA for *SparseLIF-T* (*i.e.*, 0.7% NDS improvement), achieving the best performance of 77.7% NDS and ranking ***1st*** on the test leaderboard by the time of paper submission.

SparseLIF is one of the first LiDAR-camera-based 3D detectors [4, 19] with temporal awareness, while most methods are ignorant or incapable of integrating temporal information, resulting in sub-optimal performance. For fair comparisons, we also compare our single-frame detector *SparseLIF-S* with other temporal-ignorant methods on the nuScenes validation set. As shown in the top part of Tab. 2, *SparseLIF-S* also outperforms the best competitor by a notable margin (1.4% NDS). Furthermore, as presented in the bottom part of Tab. 2, our multi-frame detector *SparseLIF-T* achieves the NDS of 77.5%, significantly outperforming all other methods by at least 3.4% on the validation set.

We also conduct latency analysis on the nuScenes dataset. We implement SparseLIF using Pytorch without any acceleration operations. The overall latency of *SparseLIF-S* is 340ms on a single NVIDIA A100 GPU. In detail, the detection head (including the PAQG module, the RIAS module, and the UAF module, *etc.*) only takes about 40ms, while the camera and LiDAR backbones take the rest of the time, which demonstrates the efficiency of our detector. We can further speed up our detector by configuring the backbones.

4.3 Ablation Studies

In Tab. 3, we conduct ablation studies on the nuScenes validation set to evaluate the key components in our multi-modality framework, based on the state-of-the-art single-frame detector *SparseLIF-S*, which yields highly convincing proofs. The RIAS module plays a vital role in our multi-modality detector. Adopting the PAQG module to generate high-quality query proposals, the mAP and NDS are improved by 1.2% and 1.0% respectively. Adopting the UAF module to conduct multi-modality fusion, the mAP and NDS are improved by 0.7% and 0.8% respectively. When *SparseLIF-S* assembles all modules, the best performance is reached: 71.2% mAP and 74.6% NDS.

We further present an in-depth analysis of our proposed PAQG module to reveal its effectiveness on detection distances and small object classes, based

Table 5: Robustness studies of SparseLIF on the nuScenes validation set, under challenging scenarios: LiDAR malfunction, camera malfunction and unsynchronization. Abbreviations: construction vehicle (C.V.), pedestrian (Ped.), and traffic cone (T.C.).

Setting		UAF	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor	Bike	T.C.	Barrier	mAP	NDS
FOV	120°		69.4	49.9	62.6	23.2	20.0	56.6	55.7	50.4	53.2	55.4	49.7	62.1
		✓	**74.5**	**55.1**	**66.5**	**33.9**	**21.7**	**60.1**	**61.3**	**57.3**	**60.6**	**62.7**	**55.4**	**65.2**
	180°		73.6	55.9	64.2	27.3	22.8	65.0	61.1	54.8	59.0	59.0	54.3	65.3
		✓	**77.5**	**59.6**	**68.5**	**35.8**	**23.1**	**67.7**	**64.6**	**60.3**	**65.4**	**65.1**	**58.8**	**67.6**
Object Failure			84.9	64.2	74.4	37.8	27.9	82.1	75.5	70.3	74.1	**70.1**	66.1	72.2
		✓	**86.2**	**67.4**	**75.2**	**41.6**	**29.2**	**82.4**	**76.4**	**71.4**	**75.4**	69.8	**67.5**	**73.0**
Front Occlusion			82.2	59.9	65.3	35.3	29.0	**86.4**	64.8	69.1	78.2	66.2	63.7	71.4
		✓	**84.1**	**62.8**	**66.7**	**36.5**	**30.3**	84.8	**66.5**	**70.2**	**78.5**	**66.3**	**64.7**	**72.0**
Stuck			90.2	69.9	81.5	43.8	33.0	90.7	82.9	**77.1**	82.6	74.9	72.7	75.9
		✓	**90.7**	**72.3**	**82.0**	**46.3**	**33.1**	**90.8**	**83.9**	76.9	**83.0**	**75.0**	**73.4**	**76.5**

on the state-of-the-art single-frame detector *SparseLIF-S*. As shown in Tab. 4, the PAQG module substantially facilitates distant object detection, *e.g.* 1.6% mAP improvement for objects beyond $30m$. Regarding small objects, the PAQG module also significantly improves the AP scores of traffic cone and barrier across all distances, *e.g.* 8.6% AP improvement for barrier in 20-30m. We attribute the performance gains to the enhanced awareness of rich context and perspective priors boosted by the proposed PAQG module.

4.4 Robustness Studies

To validate the robustness of our multi-modality framework, we evaluate SparseLIF under LiDAR/camera malfunction and unsynchronization scenarios (see [79] for more details):

- **Limited FOV**. We simulate the limited FOV angles of 120° and 180° by filtering out LiDAR points.
- **Object Failure**. Following BEVFusion [34], we simulate this scenario by selecting 50% frames to drop points of objects, where 50% objects are dropped for each selected frame.
- **Front Occlusion**. Following BEVFusion [34], we simulate such an occlusion scenario by filling the entire front-camera image with zero value.
- **Stuck**. The timestamps of two sensors might not always be synchronized, yielding stuck data, *e.g.* the detector wrongly receives data with timestamp $t-1$ at time t. Following BEVFusion [34], we simulate such an unsynchronized scenario on 50% frames.

We directly evaluate our *SparseLIF-T* model under these scenarios without any adaption or fine-tuning. As shown in Tab. 5, the UAF module boosts the robustness performance by 3.1% NDS at the most challenging LiDAR malfunction

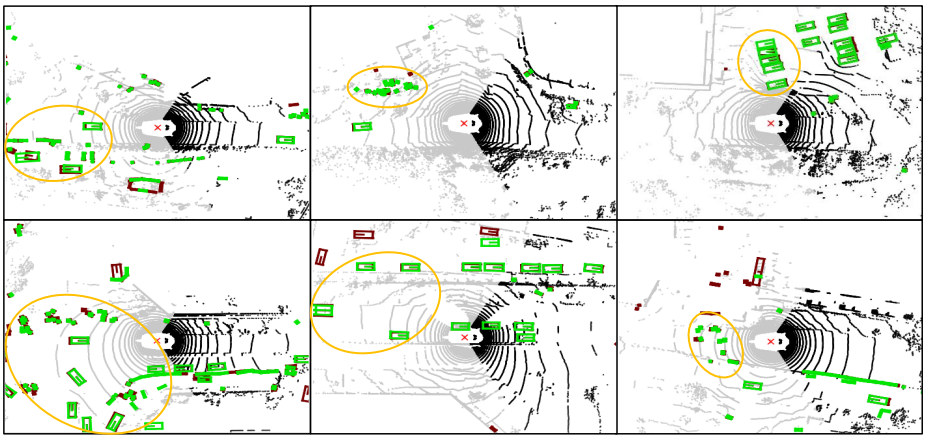

Fig. 4: Robustness visualizations under the scenario of limited LiDAR FOV angle of 120°. We color each box with green and red for prediction and ground truth.

scenario (top), *i.e.*, limited FOV angle of 120°. Simultaneously, our SparseLIF also gains robustness improvement by 0.6% NDS under camera malfunction (middle) and unsynchronization (bottom) scenarios. The experimental results convincingly demonstrate the capability of our detector against sensor noises.

We further visualize the predictions of our SparseLIF under the most challenging scenario of limited LiDAR FOV angle of 120°. As presented in Fig. 4, our SparseLIF precisely detects objects in golden circles with LiDAR input malfunctioned, showing the remarkable robustness of our multi-modality detector attributed to the proposed UAF module.

5 Conclusion

This paper proposes a high-performance fully sparse detector termed SparseLIF for LiDAR-camera-based 3D object detection. Our SparseLIF achieves state-of-the-art performance by enhancing the awareness of rich representations in two modalities. In particular, SparseLIF consists of (1) the PAQG module, which generates high-quality 3D queries with perspective priors to facilitate the perception of small and distant objects; (2) the RIAS module, which further refines prior queries by RoI feature sampling to embrace the fully sparse paradigm with the capability of low latency and integration of more temporal frames; (3) the UAF module, which quantifies the uncertainty of each modality for multi-modality fusion to enhance robustness against sensor noises. The experimental results demonstrate the superiority of our proposed method over all state-of-the-art 3D object detectors on the nuScenes benchmark. In the future, we will explore applications of SparseLIF on other tasks, such as occupancy prediction.

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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022)
2. Bewley, A., Sun, P., Mensink, T., Anguelov, D., Sminchisescu, C.: Range conditioned dilated convolutions for scale invariant 3d object detection. arXiv preprint arXiv:2005.09927 (2020)
3. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Cai, H., Zhang, Z., Zhou, Z., Li, Z., Ding, W., Zhao, J.: Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. arXiv preprint arXiv:2303.17099 (2023)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
7. Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H.: Futr3d: A unified sensor fusion framework for 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 172–181 (2023)
8. Chen, Y., Liu, S., Shen, X., Jia, J.: Fast point r-cnn. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9775–9784 (2019)
9. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8394–8405 (2023)
10. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F., Zhou, B., Zhao, H.: Autoalign: pixel-instance feature aggregation for multi-modal 3d object detection. arXiv preprint arXiv:2201.06493 (2022)
11. Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F.: Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. arXiv preprint arXiv:2207.10316 (2022)
12. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d` (2020)
13. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1201–1209 (2021)
14. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: In defense of range view for lidar-based 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2918–2927 (2021)
15. Gao, Z., Wang, L., Han, B., Guo, S.: Adamixer: A fast-converging query-based object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5364–5373 (2022)

16. Gunn, J., Lenyk, Z., Sharma, A., Donati, A., Buburuzan, A., Redford, J., Mueller, R.: Lift-attend-splat: Bird's-eye-view camera-lidar fusion using transformers. arXiv preprint arXiv:2312.14919 (2023)
17. Han, C., Sun, J., Ge, Z., Yang, J., Dong, R., Zhou, H., Mao, W., Peng, Y., Zhang, X.: Exploring recurrent long-term temporal fusion for multi-view 3d perception. arXiv preprint arXiv:2303.05970 (2023)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hu, C., Zheng, H., Li, K., Xu, J., Mao, W., Luo, M., Wang, L., Chen, M., Liu, K., Zhao, Y., et al.: Fusionformer: A multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3d objection. arXiv preprint arXiv:2309.05257 (2023)
20. 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. arXiv preprint arXiv:2303.17895 **2** (2023)
21. Huang, J., Huang, G.: Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054 (2022)
22. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
23. Huang, J., Ye, Y., Liang, Z., Shan, Y., Du, D.: Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. arXiv preprint arXiv:2311.07152 (2023)
24. Huang, T., Liu, Z., Chen, X., Bai, X.: Epnet: Enhancing point features with image semantics for 3d object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 35–52. Springer (2020)
25. Jiang, X., Li, S., Liu, Y., Wang, S., Jia, F., Wang, T., Han, L., Zhang, X.: Far3d: Expanding the horizon for surround-view 3d object detection. arXiv preprint arXiv:2308.09616 (2023)
26. Jiao, Y., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21643–21652 (2023)
27. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
28. Lee, Y., Park, J.: Centermask: Real-time anchor-free instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13906–13915 (2020)
29. Li, B., Zhang, T., Xia, T.: Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916 (2016)
30. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
31. 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: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1477–1485 (2023)
32. Li, Z., Wang, F., Wang, N.: Lidar r-cnn: An efficient and universal 3d object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7546–7555 (2021)

33. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022)
34. 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. Advances in Neural Information Processing Systems **35**, 10421–10434 (2022)
35. Liang, Z., Zhang, M., Zhang, Z., Zhao, X., Pu, S.: Rangercnn: Towards fast and accurate 3d object detection with range image representation. arXiv preprint arXiv:2009.00206 (2020)
36. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
37. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. arXiv preprint arXiv:2211.10581 (2022)
38. Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z.: Sparse4d v2: Recurrent temporal fusion with sparse model. arXiv preprint arXiv:2305.14018 (2023)
39. Lin, X., Pei, Z., Lin, T., Huang, L., Su, Z.: Sparse4d v3: Advancing end-to-end 3d detection and tracking. arXiv preprint arXiv:2311.11722 (2023)
40. Liu, H., Teng, Y., Lu, T., Wang, H., Wang, L.: Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18580–18590 (2023)
41. Liu, H., Xu, Z., Wang, D., Zhang, B., Wang, G., Dong, B., Wen, X., Xu, X.: Pai3d: Painting adaptive instance-prior for 3d object detection. In: European Conference on Computer Vision. pp. 459–475. Springer (2022)
42. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. pp. 531–548. Springer (2022)
43. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3262–3272 (2023)
44. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., 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). pp. 2774–2781. IEEE (2023)
45. Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., Xu, C.: Voxel transformer for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3164–3173 (2021)
46. Meyer, G.P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C.K.: Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12677–12686 (2019)
47. Nabati, R., Qi, H.: Centerfusion: Center-based radar and camera fusion for 3d object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1527–1536 (2021)
48. Park, D., Ambrus, R., Guizilini, V., Li, J., Gaidon, A.: Is pseudo-lidar needed for monocular 3d object detection? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3142–3152 (2021)
49. Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., Zhan, W.: Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443 (2022)

50. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 194–210. Springer (2020)
51. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018)
52. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
53. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
54. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
55. Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. arXiv preprint arXiv:1811.08188 (2018)
56. Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.S., Zhao, M.J.: Improving 3d object detection with channel-wise transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2743–2752 (2021)
57. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
58. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
59. Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7276–7282. IEEE (2019)
60. Sun, P., Wang, W., Chai, Y., Elsayed, G., Bewley, A., Zhang, X., Sminchisescu, C., Anguelov, D.: Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5725–5734 (2021)
61. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
62. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4604–4612 (2020)
63. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11794–11803 (2021)
64. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6792–6802 (2023)
65. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. arXiv preprint arXiv:2303.11926 (2023)

66. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 913–922 (2021)
67. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8453 (2019)
68. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
69. Wang, Z., Jia, K.: Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1742–1749. IEEE (2019)
70. Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., Zhang, L.: Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 3047–3054. IEEE (2021)
71. Yan, J., Liu, Y., Sun, J., Jia, F., Li, S., Wang, T., Zhang, X.: Cross modal transformer: Towards fast and robust 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18268–18278 (2023)
72. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**, 3337 (2018)
73. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al.: Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17830–17839 (2023)
74. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11040–11048 (2020)
75. Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L.: Deepinteraction: 3d object detection via modality interaction. Advances in Neural Information Processing Systems **35**, 1992–2005 (2022)
76. Yao, Z., Ai, J., Li, B., Zhang, C.: Efficient detr: improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318 (2021)
77. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
78. Yin, T., Zhou, X., Krähenbühl, P.: Multimodal virtual point 3d detection. Advances in Neural Information Processing Systems **34**, 16494–16507 (2021)
79. Yu, K., Tao, T., Xie, H., Lin, Z., Liang, T., Wang, B., Chen, P., Hao, D., Wang, Y., Liang, X.: Benchmarking the robustness of lidar-camera fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3187–3197 (2023)
80. Yuan, Z., Song, X., Bai, L., Wang, Z., Ouyang, W.: Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2068–2078 (2021)
81. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
82. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)

83. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023)
84. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)