

# EVT: Efficient View Transformation for Multi-Modal 3D Object Detection

Yongjin Lee<sup>1</sup>, Hyeon-Mun Jeong<sup>1</sup>, Yurim Jeon<sup>2</sup>, Sanghyun Kim<sup>1,2\*</sup>

<sup>1</sup>ThorDrive Co., Ltd, South Korea <sup>2</sup>Seoul National University, South Korea

{dydwls462, hyeonmun123}@gmail.com {fabioisyo01, shyun613}@snu.ac.kr

## Abstract

Multi-modal sensor fusion in Bird’s Eye View (BEV) representation has become the leading approach for 3D object detection. However, existing methods often rely on depth estimators or transformer encoders to transform image features into BEV space, which reduces robustness or introduces significant computational overhead. Moreover, the insufficient geometric guidance in view transformation results in ray-directional misalignments, limiting the effectiveness of BEV representations. To address these challenges, we propose Efficient View Transformation (EVT), a novel 3D object detection framework that constructs a well-structured BEV representation, improving both accuracy and efficiency. Our approach focuses on two key aspects. First, Adaptive Sampling and Adaptive Projection (ASAP), which utilizes LiDAR guidance to generate 3D sampling points and adaptive kernels, enables more effective transformation of image features into BEV space and a refined BEV representation. Second, an improved query-based detection framework, incorporating group-wise mixed query selection and geometry-aware cross-attention, effectively captures both the common properties and the geometric structure of objects in the transformer decoder. On the nuScenes test set, EVT achieves state-of-the-art performance of 75.3% NDS with real-time inference speed.

## 1. Introduction

LiDAR-camera fusion is essential for 3D object detection, as it leverages the complementary strengths [1, 3, 5, 6, 21, 25, 31, 39, 45, 46, 49, 53]. LiDAR provides precise geometric information for accurate object localization, while cameras capture rich semantic details such as color and texture. However, their integration remains challenging due to the differences in sensing modalities.

Recently, the dominant multi-modal fusion methods are categorized as implicit or explicit fusion. Implicit fusion employs cross-attention within transformers, where object queries iteratively interact with independently processed

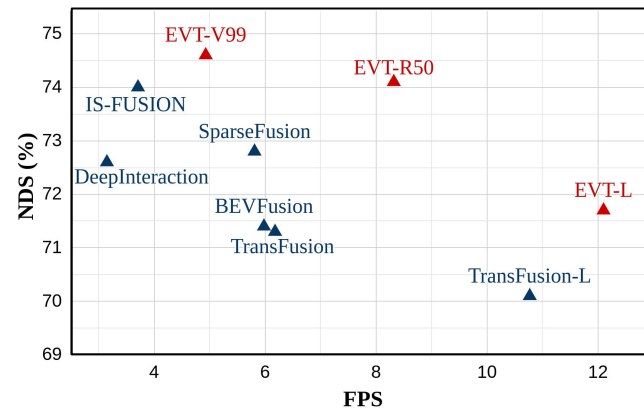


Figure 1. Performance comparison of EVT and other methods on the nuScenes validation set. The FPS of all methods is measured in FP32 on a single Tesla A100 GPU using the official implementations, excluding voxelization time.

sensor features [1, 5, 46, 49, 50]. Implicit fusion offers flexibility and simplicity by removing explicit feature alignment across sensor modalities but incurs high computational costs and struggles to extract complementary features. Explicit fusion, on the other hand, directly aligns and integrates sensor data in BEV space using view transformation (VT), which enhances complementary feature fusion while reducing computational costs. Although its performance heavily depends on the accuracy of VT, conventional methods have inherent limitations. Depth-based VT methods [20, 22, 25, 31, 34, 56] lift image features into BEV using pixel-wise depth estimation, but their sensitivity to depth errors compromises robustness. In contrast, query-based VT methods [1, 3, 7, 8, 12, 24] refine BEV queries via attention mechanisms, incurring high computational costs. Furthermore, both methods lack geometric guidance, resulting in ray-directional misalignment. This misalignment reduces spatial accuracy, leading to unintended information capture along the ray direction, ultimately degrading BEV representation accuracy.

To address the limitations of explicit fusion methods and improve detection performance, we propose Efficient View Transformation (EVT), a novel 3D object detection framework aimed at enhancing both accuracy and efficiency by

\*Corresponding Author.

constructing a well-structured BEV representation. EVT introduces two key innovations: (1) Adaptive Sampling and Adaptive Projection (ASAP), a novel VT method which leverages LiDAR guidance, and (2) an improved query-based object detection framework incorporating group-wise mixed query selection and geometry-aware cross-attention, enabling effective multi-modal BEV feature decoding for accurate 3D object detection.

ASAP consists of two key modules: Adaptive Sampling (AS) and Adaptive Projection (AP). AS generates 3D sampling points from LiDAR features to effectively represent image features in BEV space while focusing more on high-relevance areas in the image. AP refines BEV representations using adaptive kernels generated from LiDAR features to enhance structurally meaningful 3D information. As a result, ASAP improves BEV feature representation and eliminates ray-directional misalignment. Unlike existing methods, it does not rely on depth estimation or attention mechanism, ensuring both efficiency and robustness.

Additionally, we improve the query-based object detection framework by introducing group-wise mixed query selection and geometry-aware cross-attention. The mixed query selection generates object queries using group-wise learnable parameters and heatmaps, allowing them to capture the common properties of each group and initialize at high-confidence positions. Then, the geometry-aware cross-attention refines object queries by integrating corner-aware sampling for precise feature selection and position-aware feature mixing for spatially aware feature decoding. These enhancements improve more robust and accurate detection while maintaining computational efficiency.

We evaluate EVT on the nuScenes dataset in terms of accuracy and efficiency. As shown in Fig. 1, EVT achieves 74.1% NDS and 8.3 FPS with ResNet-50 [11], 74.6% NDS and 4.9 FPS with V2-99 [18], and 71.7% NDS and 12.1 FPS with the LiDAR-only model EVT-L on the nuScenes validation set, outperforming other methods in both accuracy and inference speed. On the nuScenes test set, EVT achieves 75.3% NDS and 72.5% mAP using single-frame raw data, without model ensemble or test-time augmentation, surpassing previous state-of-the-art methods.

In summary, the main contributions are as follows:

- We introduce EVT, a novel 3D object detection framework that improves both accuracy and efficiency through ASAP and an improved query-based framework.
- ASAP, our view transformation method, utilizes LiDAR guidance to generate BEV feature maps while improving efficiency by eliminating the need for depth estimators and transformer encoders.
- Our improved query-based detection framework further enhances detection accuracy and robustness.
- EVT achieves state-of-the-art performance of 75.3% NDS and 72.6% mAP on the nuScenes test set.

## 2. Related Work

### 2.1. Query-based Object Detection Framework

DETR [9] introduces transformers into 2D object detection, eliminating hand-designed components like non-maximum suppression, but suffering from slow convergence. To address this, Deformable DETR [60] proposes deformable cross-attention to accelerate training. DAB-DETR [28] enhances query representation by modeling queries as anchor boxes, while DN-DETR [19] stabilizes training with query denoising and auxiliary supervision.

DETR-like approaches have been extended to 3D object detection [5, 16, 21, 27, 29, 30, 33, 42, 43, 46, 49]. PETR [29] defines object query features and positions using learnable parameters, leveraging position embeddings within multi-head attention for feature refinement. DETR3D [43] and BEVFormer [24] project BEV queries onto image planes, refining them with bilinear interpolation or deformable cross-attention. CenterFormer [58] samples features from high-scoring heatmap keypoints for query initialization, while TransFusion [1] enhances sampled features with category embeddings.

While extensive research has explored query initialization, DINO [54] notes that direct feature sampling in methods [1, 44, 51, 55, 58] limits performance due to inaccurate initial feature representations. Furthermore, despite advancements in attention-based query refinement, these approaches insufficiently leverage the geometric structure of object queries, limiting 3D spatial understanding.

### 2.2. Implicit Multi-modal Fusion in Transformer

Implicit multi-modal fusion integrates each sensor’s data through cross-attention with object queries in the transformer decoder, without relying on BEV representation for feature alignment. FUTR3D [5] introduces a modality-agnostic feature sampler that aggregates features from different sensors using deformable cross-attention [60]. DeepInteraction [49] and DeepInteraction++ [50] preserve modality-specific information by using a modality interaction strategy. Meanwhile, TransFusion [1] updates object queries in a sequential manner by applying cross-attention separately to LiDAR and camera features. And CMT [46] constructs input tokens by adding 3D position embeddings to each sensor’s data before concatenating them.

These implicit fusion methods rely on cross-attention between object queries and sensor features within the transformer decoder, providing a flexible and generalizable framework for multi-modal fusion. However, because it depends entirely on query-driven attention mechanisms, it struggles to effectively extract complementary features across different sensors. Furthermore, the iterative execution of cross-attention over the entire sensor data incurs high computational overhead and memory inefficiency.

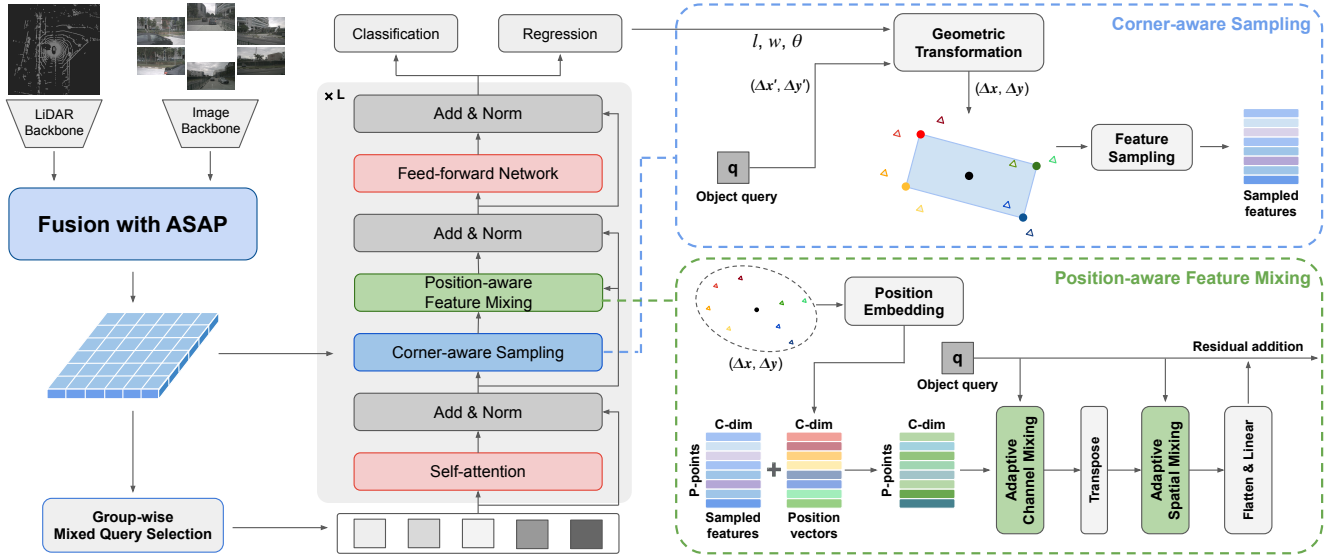


Figure 2. Overall architecture of EVT. Each backbone extracts either image features or LiDAR features. The proposed ASAP module fuses these two features in BEV space. The group-wise mixed query selection stage generates queries from the fused feature map. In the transformer decoder, corner-aware sampling leverages the geometric properties of the object queries to sample multi-modal features, and position-aware feature mixing decodes the sampled features to update the queries. These queries then predict the 3D bounding boxes.

### 2.3. Explicit Multi-modal Fusion in BEV Space

Explicit multi-modal fusion focuses on view transformation of 2D image features into BEV space [1, 3, 14, 15, 21–25, 31, 34, 35, 48]. For view transformation, depth-based methods predict pixel-wise depth distributions to lift multi-view image features into BEV space [14, 15, 21, 22, 25, 31, 34]. While these methods effectively incorporate spatial priors through depth estimation, their heavy reliance on the depth estimators significantly limits their overall robustness.

In contrast, query-based view transformation methods [3, 12, 24, 48] project predefined 3D points onto image planes and extract features using deformable cross-attention [60]. This approach eliminates the need for depth estimation; however, predefined positions of sampling points fail to align accurately with regions where objects are located, leading to suboptimal feature representation. Furthermore, it suffers from high computational overhead due to the extensive use of multi-layer transformers and ray-directional misalignment due to insufficient geometric guidance.

Aforementioned challenges cause the inherent difficulty in establishing precise correspondences between 2D and 3D spaces, which is crucial for effective BEV representation. Existing approaches often suffer from feature misalignment due to depth estimation errors and also struggle with maintaining spatial consistency in transformer-based methods, while their high computational cost significantly limits real-time deployment. Therefore, a more efficient and geometrically grounded approach is essential for achieving accurate and robust multi-modal BEV representations.

## 3. Methodology

The overall pipeline of EVT is illustrated in Fig. 2. First,  $N_s$ -scale perspective-view image features  $\{\mathbf{PV}_j\}_{j=1}^{N_s}$  and BEV LiDAR features  $\mathbf{BEV}_{\text{lidar}} \in \mathbb{R}^{C \times H \times W}$  are extracted from separate backbone networks, where  $C$  is the feature dimension and  $H \times W$  is the size of the BEV feature map.

To fuse 2D image and LiDAR features in BEV space, the Adaptive Sampling and Adaptive Projection (ASAP) module transforms 2D image features into BEV space and then fuses them with  $\mathbf{BEV}_{\text{lidar}}$  (Sec. 3.1). For query-based 3D detection with multi-modal BEV features, group-wise mixed query selection initializes object queries based on heatmap-guided locations (Sec. 3.2), and geometry-aware cross-attention refines them in the transformer decoder to predict 3D bounding boxes (Sec. 3.3).

### 3.1. Adaptive Sampling and Adaptive Projection

The proposed ASAP module consists of two stages: Adaptive Sampling (AS) and Adaptive Projection (AP). In the first stage, AS selectively extracts and aggregates multi-scale perspective-view image features  $\{\mathbf{PV}_j\}_{j=1}^{N_s}$  into an initial BEV representation  $\mathbf{BEV}_{\text{as}}$ . In the second stage, AP refines  $\mathbf{BEV}_{\text{as}}$  to obtain the final image BEV feature map  $\mathbf{BEV}_{\text{camera}}$ . The structure of this module is shown in Fig. 3.

**Adaptive Sampling** To transform multi-scale image features into BEV space, the Adaptive Sampling (AS) module predicts optimal sampling heights for each BEV grid cell using LiDAR features, allowing effective feature extraction from high-relevance areas in the image (see Fig. 4).

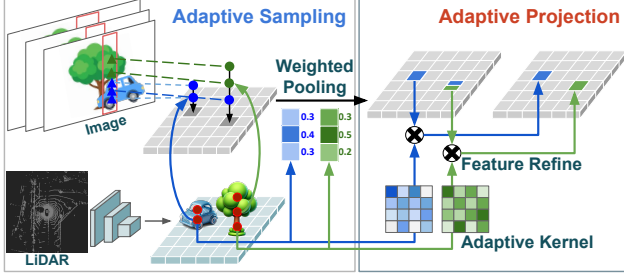


Figure 3. Overview of AS and AP. In AS, the LiDAR BEV feature map generates 3D sampling points and their corresponding weights. The sampling points are projected onto the image plane, where features are sampled and combined by weighted pooling in each BEV grid cell. In AP, the image BEV feature map produced by AS is refined channel-wise using the adaptive kernels generated from the LiDAR features.

First, multiple height values for each grid cell are generated from LiDAR features:

$$\{Z_i\}_{i=1}^{N_h} = \text{Conv}(\mathbf{BEV}_{\text{lidar}})(u, v), \quad (1)$$

where  $(u, v)$  denotes the coordinates of a grid cell in BEV space, and  $N_h$  denotes the number of generated heights. Consequently, the 3D sampling points  $P = \{(X, Y, Z_i)\}_{i=1}^{N_h}$  are defined based on real-world coordinates  $(X, Y)$  corresponding to the grid cell at  $(u, v)$  and the set of heights  $\{Z_i\}$ .

Next, the generated points  $P$  are projected onto the image plane. Each projected point samples the multi-scale features  $\{\mathbf{PV}_j\}_{j=1}^{N_s}$  across  $N_s$  different image scales with downsampled strides  $\{S_j\}_{j=1}^{N_s}$ . Consequently,  $N_h \times N_s$  projected points  $(x_i^j, y_i^j)$  and sampled features  $f_i^j$  are obtained via projection:

$$(x_i^j, y_i^j) = \frac{\text{Proj}(X, Y, Z_i)}{S_j} \quad (2)$$

$$f_i^j = \text{B}(\mathbf{PV}_j, (x_i^j, y_i^j)) \in \mathbb{R}^C, \quad (3)$$

where  $\text{Proj}(\cdot)$  denotes the projection of 3D points onto the image plane, and  $\text{B}(\cdot)$  denotes bilinear interpolation.

To aggregate the  $N_s \times N_h$  sampled features  $\{f_i^j\}$ , adaptive sampling weights  $W_{\text{as}} \in \mathbb{R}^{N_s \times N_h}$  are derived from  $\mathbf{BEV}_{\text{lidar}}$ .  $W_{\text{as}}$  determines the importance of the heights and image feature scales for each grid cell. The multi-scale image features in BEV space are obtained as follows:

$$W_{\text{as}} = \sigma(\text{Conv}(\mathbf{BEV}_{\text{lidar}})(u, v)) \quad (4)$$

$$\mathbf{BEV}_{\text{as}}(u, v) = \sum_{j=1}^{N_s} \sum_{i=1}^{N_h} W_{\text{as}}(j, i) \cdot f_i^j \in \mathbb{R}^C, \quad (5)$$

where  $\sigma(\cdot)$  denotes the softmax function applied to all  $N_s \times N_h$  elements.  $\mathbf{BEV}_{\text{as}}$  denotes the image BEV feature map.

**Adaptive Projection** The Adaptive Projection (AP) module refines the BEV feature map  $\mathbf{BEV}_{\text{as}}$ , produced by the AS module, by applying adaptive kernels to each grid cell. The overall process is represented by the following equations:

$$K_{\text{ap}} = \text{Conv}(\mathbf{BEV}_{\text{lidar}})(u, v) \in \mathbb{R}^{C \times C} \quad (6)$$

$$\mathbf{BEV}_{\text{camera}}(u, v) = \mathbf{BEV}_{\text{as}}(u, v) \times K_{\text{ap}} \in \mathbb{R}^C. \quad (7)$$

First, an adaptive kernel  $K_{\text{ap}} \in \mathbb{R}^{C \times C}$  is derived from  $\mathbf{BEV}_{\text{lidar}}$  for each BEV grid cell. Then,  $\mathbf{BEV}_{\text{as}}$  is refined by applying a channel-wise linear projection using  $K_{\text{ap}}$  to obtain the image BEV feature map  $\mathbf{BEV}_{\text{camera}}$ .

Unlike static transformations, our LiDAR-guided feature refinement leverages spatial information to effectively mitigate ray-directional misalignment, which primarily results from occlusions and empty 3D spaces (see Fig. 5).

**Multi-modal Fusion in BEV Space** The multi-modal BEV feature map  $\mathbf{BEV}_{\text{fuse}} \in \mathbb{R}^{C \times H \times W}$  is obtained by concatenating the image and LiDAR BEV feature maps ( $\mathbf{BEV}_{\text{camera}}$  and  $\mathbf{BEV}_{\text{lidar}}$ ) along the channel dimension, followed by a convolution operation:

$$\mathbf{BEV}_{\text{fuse}} = \text{Conv}(\text{Concat}(\mathbf{BEV}_{\text{camera}}, \mathbf{BEV}_{\text{lidar}})), \quad (8)$$

where  $\text{Concat}(\cdot)$  denotes channel-wise concatenation.

### 3.2. Group-wise Mixed Query Selection

Our proposed group-wise mixed query selection generates object queries for transformer-based detection frameworks. First, group-wise heatmaps are predicted from the multi-modal BEV feature map  $\mathbf{BEV}_{\text{fuse}}$  (from Sec. 3.1), where each group consists of similarly sized object classes. The predicted heatmaps have scores ranging from 0 to 1, representing the likelihood that each BEV pixel corresponds to the center of an object. The heatmap head is supervised by 2D Gaussian distributions centered at each object's location. Next, the top-k keypoints are selected from each heatmap group, and their positions are used as reference points for queries in BEV space.

Inspired by DINO [54], we initialize query features solely with group-wise learnable parameters, unlike query positions. DINO defines query features as instance-wise learnable parameters without using any categorical priors. In contrast, our approach allows all queries within the same group to share these parameters, effectively capturing the common properties of their group. Experimentally, the group-wise shared initial embeddings outperform instance-wise embeddings for object query representation. Moreover, our approach outperforms traditional approaches that either define both query features and positions as learnable parameters [12, 27, 29, 30, 46] or obtain both from heatmap keypoints [1, 44, 51, 55, 58]. Further details and analysis of our method are provided in Sec. 4.4.

### 3.3. Geometry-aware Cross-Attention

To enhance query representations in transformer decoders, we refine deformable cross-attention [60] with corner-aware sampling for improved feature sampling and position-aware feature mixing for better feature aggregation.

**Corner-aware Sampling** The conventional deformable cross-attention samples features around the centers of object queries. However, this approach struggles with objects of varying sizes and fails to capture fine-grained boundary details and spatial extent effectively. To address this limitation, corner-aware sampling explicitly incorporates object geometry, ensuring precise spatial alignment of sampling points with the object’s true structure.

First, the initial sampling offsets  $\{(\Delta x'_i, \Delta y'_i)\}$  are generated from the query feature  $\mathbf{q}$  using a linear layer:

$$\{(\Delta x'_i, \Delta y'_i) \mid i \in 0, 1, \dots, N_p - 1\} = \text{Linear}(\mathbf{q}), \quad (9)$$

where  $N_p$  denotes the number of sampling points.

Next, the final sampling offsets  $\{(\Delta x_i, \Delta y_i)\}$  and sampling points  $\{(x_i, y_i)\}$  are determined via a geometric transformation, which relocates sampling points to object corners and aligns them with the object’s heading, as follows:

$$\begin{bmatrix} \Delta x_i \\ \Delta y_i \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} I_j \cdot \frac{l}{2} + \Delta x'_i \\ I'_j \cdot \frac{w}{2} + \Delta y'_i \end{bmatrix} \quad (10)$$

$$\begin{bmatrix} x_i & y_i \end{bmatrix} = \begin{bmatrix} x_c & y_c \end{bmatrix} + \begin{bmatrix} \Delta x_i & \Delta y_i \end{bmatrix}, \quad (11)$$

where  $(I_j, I'_j) \in \{(1, 1), (1, -1), (-1, 1), (-1, -1)\}$ .  $j$  denotes the index of each corner and is the remainder of  $i$  divided by four, the number of corners. The coordinates  $(x_c, y_c)$  denote the object’s center.  $l$ ,  $w$ , and  $\theta$  denote the length, width, and yaw of the predicted bounding box, respectively, obtained from the regression head of the previous transformer layer. In the first decoder layer, we initialize  $l$ ,  $w$ ,  $\theta$  to zero.

The features are sampled using bilinear interpolation at each sampling point on the multi-modal BEV feature map  $\text{BEV}_{\text{fuse}}$ , as follows:

$$g_i = \text{B}(\text{BEV}_{\text{fuse}}, (x_i, y_i)) \in \mathbb{R}^C, \quad (12)$$

where  $\text{B}(\cdot)$  denotes the bilinear interpolation, and  $g_i$  represents the sampled features at  $(x_i, y_i)$ .

**Position-aware Feature Mixing** Given the sampled features from different corners, the key challenge is how to decode them while maintaining their spatial relationships. While AdaMixer [10] is designed for feature decoding, it struggles with the geometric transformations introduced during corner-aware sampling, which complicate the association between sampled features and their transformed

sampling locations (see Tab. 4 (e)). To address this, we propose position-aware feature mixing, which incorporates positional embeddings that encode the transformed sampling offsets, allowing for more structured feature aggregation.

First, the sampling offsets are embedded as position vectors using sinusoidal position encoding [38], followed by a linear layer. Then, the position-aware sampled features  $G \in \mathbb{R}^{N_p \times C}$  are obtained by element-wise addition of the sampled feature  $g_i$  and the position vector  $e_i$ :

$$e_i = \text{Linear}(\Phi_{\text{pos}}((x_i, y_i))) \in \mathbb{R}^C \quad (13)$$

$$G = \{(g_i + e_i) \mid i \in 0, 1, \dots, N_p - 1\}, \quad (14)$$

where  $\Phi_{\text{pos}}(\cdot)$  denotes sinusoidal position encoding.

Subsequently, adaptive channel mixing is applied to  $G$  using the dynamic weights  $W_c$  generated from the query feature  $\mathbf{q}$  to obtain the channel-mixed feature  $G_c$ :

$$W_c = \text{Linear}(\mathbf{q}) \in \mathbb{R}^{C \times C} \quad (15)$$

$$G_c = \text{ReLU}(\text{LN}(G \times W_c)). \quad (16)$$

Next, adaptive spatial mixing is applied to the spatial dimensions of  $G_c$  using dynamic weights  $W_s$  to obtain the spatial-mixed feature  $G_{cs}$ :

$$W_s = \text{Linear}(\mathbf{q}) \in \mathbb{R}^{N_p \times N_p} \quad (17)$$

$$G_{cs} = \text{ReLU}(\text{LN}(G_c^T \times W_s)). \quad (18)$$

Finally, the query feature is formulated as follows:

$$\mathbf{q}' = \mathbf{q} + \text{Linear}(\text{Flatten}(G_{cs})), \quad (19)$$

where  $\mathbf{q}'$  represents the refined query feature. By integrating positional embeddings at the feature level, this formulation enhances spatial awareness in query representation, ensuring more robust geometric reasoning in object queries.

## 4. Experiments

### 4.1. Implementation Details

For the image backbone, we use ResNet-50 [11] with a resolution of  $704 \times 256$  or V2-99 [18] with  $1600 \times 640$  with FPN [26]. The LiDAR backbone is VoxelNet [57] with an ROI of  $[-54.0m, 54.0m]$  in  $(X, Y)$  and  $[-5.0m, 3.0m]$  in  $Z$ , with a voxel size of  $(0.075m, 0.075m, 0.2m)$ . In ASAP, we use four sampling points for each grid cell. The multi-modal BEV feature map size is  $180 \times 180$ . Following [52], object groups in 3.2 are defined as: (1) car, (2) truck, construction vehicle, (3) bus, trailer, (4) barrier, (5) motorcycle, bicycle, (6) pedestrian, traffic cone. Each group contains 150 queries, resulting in a total of 900 queries, and the corner-aware sampling generates 16 sampling points per query, both of which were empirically determined. The transformer decoder has six layers, and the feature dimension is set to 256.

Method	Modality	NDS ( <i>val</i> )	mAP ( <i>val</i> )	NDS ( <i>test</i> )	mAP ( <i>test</i> )
UVTR-L [21]	L	67.7	60.9	69.7	63.9
TransFusion-L [1]	L	70.1	65.1	70.2	65.5
FocalFormer3D-L [6]	L	-	-	<b>72.6</b>	<b>68.7</b>
CMT-L [46]	L	68.6	62.4	70.1	65.3
<b>EVT-L (Ours)</b>	L	<b>71.7</b>	<b>66.4</b>	72.1	67.7
MVP [53]	LC	70.8	67.1	70.5	66.4
UVTR [21]	LC	70.2	65.4	71.1	67.1
AutoAlignV2 [7]	LC	71.2	67.1	72.4	68.4
TransFusion [1]	LC	71.3	67.5	71.7	68.9
DeepInteraction [49]	LC	72.6	69.9	73.4	70.8
BEVFusion [31]	LC	71.4	68.5	72.9	70.2
Objectfusion [4]	LC	72.3	69.8	73.3	71.0
FocalFormer3D [6]	LC	71.1	66.5	73.9	71.6
CMT [46]	LC	72.9	70.3	74.1	72.0
BEVFusion4D-S [3]	LC	72.9	70.9	73.7	71.9
SparseFusion [45]	LC	72.8	70.4	73.8	72.0
MSMDFusion [17]	LC	-	-	74.0	71.5
UniTR [40]	LC	73.3	70.5	74.5	70.9
FusionFormer [12]	LCT	74.1	71.4	75.1	72.6
<b>EVT (Ours)</b>	LC	<b>74.6</b>	<b>72.1</b>	<b>75.3</b>	<b>72.6</b>

Table 1. Performance comparison on the nuScenes validation and test sets. The results are obtained without model ensemble or test-time augmentation. ‘L’, ‘C’ and ‘T’ denote LiDAR, camera and temporal fusion, respectively.

Our model is trained on 8 RTX 3090 GPUs with a batch size of 16. The model is trained end-to-end for 10 epochs using CBGS [59], whereas GT sample augmentation [47] is applied for the first 9 epochs. The query denoising strategy [19] is also adopted. Gaussian Focal loss [41], Focal loss [36] and L1 loss are used for heatmap prediction, classification and regression, respectively. The AdamW [32] optimizer is adopted with a learning rate of  $1 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-2}$  with a cyclical learning rate policy [37]. No model ensemble or test-time augmentation is applied during inference.

## 4.2. Dataset and Metric

Consistent with previous works [3, 13, 20, 51, 55, 56], we conduct extensive experiments on the nuScenes dataset [2], a large-scale benchmark for evaluating 3D object detection in autonomous driving. It consists of 1,000 scenes, each lasting 20s, divided into training, validation, and testing sets (700, 150, and 150 scenes, respectively). This dataset contains multimodal sensor data, including point clouds from a 32-beam LiDAR at 20 frames per second (fps), images from six cameras with a resolution of 1600×900 pixels at 12 fps, and data from five radars, providing a 360-degree view, with annotations are provided every 0.5s, resulting in 1.4 million annotated objects across 10 traffic categories.

Performance on this dataset is assessed using metrics like mean Average Precision (mAP), calculated over distance thresholds of 0.5, 1, 2, and 4m across all classes, and the

nuScenes Detection Score (NDS), which offers a holistic evaluation by combining mAP with measures of translation, scale, orientation, velocity, and attribute errors.

## 4.3. Comparison with State-of-the-art Methods

As shown in Tab. 1, we compare EVT and its LiDAR-only model, EVT-L, with existing methods on the nuScenes validation and test sets. The multi-modal EVT achieves 75.3% NDS and 72.6% mAP, surpassing all previous approaches on both the nuScenes validation and test sets. In particular, it surpasses recent methods, such as UniTR [40] by 0.8% NDS and 1.7% mAP, MSMDFusion [17] by 1.3% NDS and 1.1% mAP, and SparseFusion [45] by 1.5% NDS and 0.6% mAP. Furthermore, EVT shows a performance improvement of 3.2% NDS and 4.9% mAP compared with the LiDAR-only model EVT-L. In contrast, TransFusion [1] shows only a 1.5% NDS and 3.4% mAP improvement over its LiDAR-only model, TransFusion-L. This indicates that EVT effectively utilizes camera data through the proposed view transformation method.

We also compare EVT-L with the LiDAR-only versions of other multi-modal 3D object detectors. On the nuScenes validation set, EVT-L surpasses TransFusion [1] and CMT [46] by 1.6% and 3.1% NDS, respectively. EVT-L also demonstrates competitive performance on the test set. These results validate the effectiveness of the proposed query initialization and cross-attention mechanisms in enhancing 3D object detection.

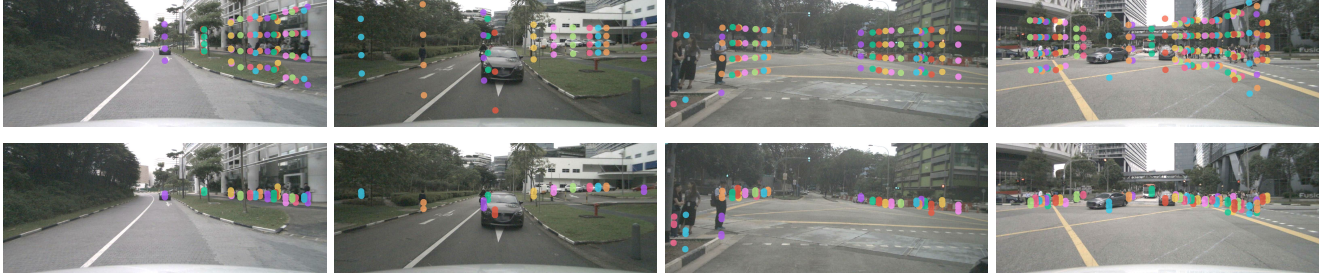


Figure 4. Visualization of projected sampling points for each object. The top row shows the projections of predefined 3D points, and the bottom row shows the projections of the points generated by the AS module. Points of each object are denoted by different colors.

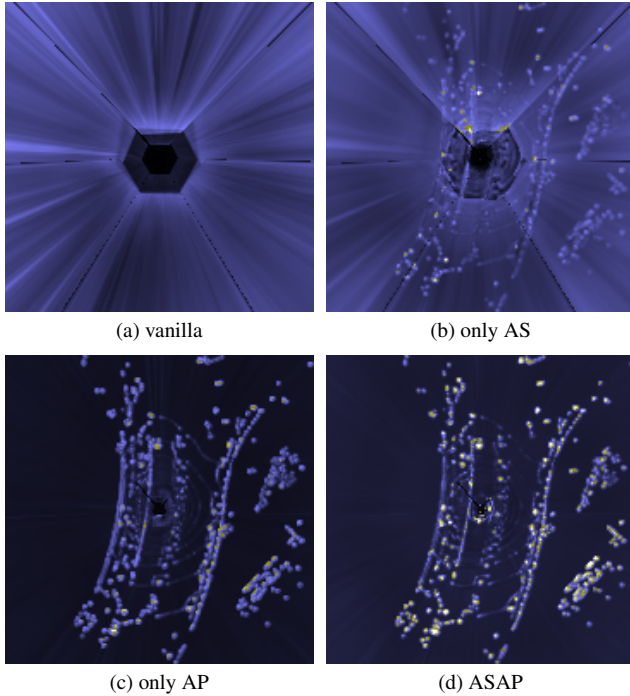


Figure 5. Comparison of BEV feature maps from different methods: (a) Vanilla, (b) AS, (c) AP, and (d) ASAP. Vanilla shows less informative and unaligned features. AS improves feature representation using adaptive sampling. AP corrects ray-directional misalignment. ASAP integrates both AS and AP, leading to the most refined and well-aligned feature representation.

#### 4.4. Ablation Studies

In this section, we describe the validation of each component of the proposed method on the nuScenes validation set. Unless otherwise specified, all experiments are conducted using the proposed LiDAR-only model EVT-L, trained for 10 epochs with CBGS [59] and the denoising strategy [19].

**Adaptive Sampling and Adaptive Projection** In Tab. 2, the ablation results for ASAP are obtained by retraining the entire model. In these experiments, ResNet [11] is used as

	LiDAR	Camera	AS	AP	NDS	mAP	FPS
(a)	✓				71.7	66.4	12.1
(b)	✓	✓			72.7	69.1	8.5
(c)	✓	✓	✓		73.5	70.6	8.5
(d)	✓	✓	✓	✓	74.1	71.1	8.3

Table 2. Ablation study of the proposed VT method. The proposed ASAP shows a significant performance improvement compared to the vanilla VT method (b), which projects predefined 3D points.

Method	# Decoder Layers & NDS (%)					
	1	2	3	4	5	6
(a) Learnable Init.	56.8	64.4	67.5	68.7	68.9	69.6
(b) Heatmap Init.	<b>70.4</b>	70.5	70.5	70.3	70.5	70.7
(c) Mixed init.	69.2	70.3	70.5	70.5	70.7	71.1
+ Group-wise	69.8	<b>71.1</b>	<b>71.3</b>	<b>71.2</b>	<b>71.4</b>	<b>71.7</b>

Table 3. Comparison of query initialization strategies on the nuScenes validation set. (a) Fully learnable initialization, (b) Fully heatmap-based initialization, (c) Our mixed initialization strategy: the first row denotes instance-wise mixed query selection and the second row denotes group-wise mixed query selection.

the backbone network with a resolution of  $704 \times 256$ . (a) shows the performance of EVT-L, the LiDAR-only model. In (b), the vanilla view transformation (VT) method, without LiDAR guidance, employs a 3D-to-2D projection of predefined 3D points for feature sampling.

In (c), AS achieves improvements of 0.8% NDS and 1.5% mAP compared to the vanilla VT method. The sampling points of the vanilla VT and AS are visualized in Fig. 4. The sampling points of AS are adaptively generated in highly object-relevant regions of the image.

In (d), the entire ASAP achieves improvements of 1.4% NDS and 2.0% mAP compared to (b), while maintaining high efficiency with only a 3ms latency increase. The BEV feature maps of each component are visualized in Fig. 5. ASAP effectively transforms image features into BEV space and resolves ray-directional misalignment.

	Attention	Reference	scale	rotate	feature mixing	position-aware	NDS	mAP
(a)	Standard	Center					69.6	65.3
(b)	Deformable	Center					70.0	64.8
(c)			✓	✓			70.5	65.2
(d)	Ours	Corner		✓			71.3	65.7
(e)				✓	✓		71.2	65.6
(f)				✓	✓	✓	✓	<b>71.7</b>

Table 4. Comparison of attention methods within the transformer decoder. The proposed geometry-aware cross-attention, which includes corner-aware sampling and position-aware feature mixing, achieves significant performance improvements.

Method	NDS	mAP
ResNet-50 baseline (24 epochs)	47.8	37.2
+ geometry-aware cross-attention	<b>49.1</b>	<b>37.8</b>
ResNet-50 baseline (90 epochs)	53.5	42.7
+ geometry-aware cross-attention	<b>54.7</b>	<b>43.4</b>

Table 5. Impact of geometry-aware cross-attention on StreamPETR [42]. All models are trained in our experiments.

**Group-wise Mixed Query Selection** As shown in Tab. 3, we conduct an ablation study on query initialization strategies. (a) fully learnable initialization, where both features and positions are learnable [12, 27, 29, 30, 46]. (b) fully heatmap-based initialization, where positions are obtained from high-score heatmap keypoints, and features are sampled at those locations [1, 44, 51, 55, 58]. (c) our mixed initialization strategy, which combines heatmap-derived positions with either instance-wise or group-wise embeddings.

These results highlight the importance of query initialization. Notably, fully learnable initialization without any prior information consistently yields the lowest performance. While (b) achieves the best performance in a single-layer transformer decoder, (c) outperforms all other strategies in multi-layer settings, improving NDS by over 1% at the last layer. Additionally, group-wise embeddings allow each group to learn more generalized feature representations, leading to meaningful improvements.

**Geometry-aware Cross-Attention** We ablate the corner-aware sampling method, as shown in Tab. 4 (a)-(d). (a) employs multi-head attention [38], and (b)-(c) use deformable attention [60]. Specifically, in (c), the sampling offsets are scaled based on the bounding box size and rotated according to the heading. (d), the proposed corner-aware sampling method, samples features from bounding box corners aligned with the heading, achieving improvements of 0.8% NDS and 0.5% mAP compared to (c).

In (e) and (f), AdaMixer [10] fails to improve performance, as it does not effectively preserve the structured sampling introduced by corner-aware sampling. In contrast, our position-aware feature mixing explicitly incorporates positional embeddings, leading to improvements of 0.4%

Initial Query Formulation	NDS	mAP
Bbox from regression head	71.4	66.3
Bbox from learnable parameters	70.9	66.0
BEV Reference Point	<b>71.7</b>	<b>66.4</b>

Table 6. Comparison of the initial query formulations for the first layer of the transformer decoder.

NDS and 0.7% mAP compared to (d). As a result, our proposed geometry-aware cross-attention achieves overall improvements of 1.2% NDS and 1.2% mAP.

Additionally, as shown in Tab. 5, we validate the applicability of geometry-aware cross-attention by integrating it into a camera-only 3D detector. Modifying only the cross-attention layers in StreamPETR [42] improves performance without further adjustments.

**Initial Query Formulation for Transformer** As described in Sec. 3.3, the corner-aware sampling method is applied starting from the second transformer layer. We explore two different approaches to extend the corner-aware sampling to all transformer layers. The first approach involves adding regression heads during query initialization in Sec. 3.2 to predict bounding boxes, whereas the second approach uses learnable bounding boxes. However, as shown in Tab. 6, neither approach resulted in any noticeable performance gains, and the experiment with learnable bounding boxes even led to performance degradation.

## 5. Conclusion

We propose EVT, a novel multi-modal 3D object detector based on BEV representation, enhancing both efficiency and accuracy. Our method introduces ASAP, an efficient LiDAR-camera fusion method that leverages LiDAR guidance for accurate view transformation. Additionally, the proposed group-wise mixed query selection improves initial feature representation through shared embeddings. The geometry-aware cross-attention refines queries using geometric properties and can be easily extended to other models. We expect EVT to provide valuable insights into multi-modal 3D object detection.

## References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089, 2022. 1, 2, 3, 4, 6, 8
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6
- [3] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, Wenbo Ding, and Jiuhua Zhao. Bevfusion4d: Learning lidar-camera fusion under bird’s-eye-view via cross-modality guidance and temporal aggregation. *arXiv preprint arXiv:2303.17099*, 2023. 1, 3, 6
- [4] Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Objectfusion: Multi-modal 3d object detection with object-centric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18067–18076, 2023. 6
- [5] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 172–181, 2023. 1, 2
- [6] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023. 1, 6
- [7] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, and Feng Zhao. Deformable feature aggregation for dynamic multi-modal 3d object detection. In *European conference on computer vision*, pages 628–644. Springer, 2022. 1, 6
- [8] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*, 2022. 1
- [9] Gopi Krishna Erabati and Helder Araujo. Li3detr: A lidar based 3d detection transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4250–4259, 2023. 2
- [10] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5364–5373, 2022. 5, 8
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 7
- [12] Chunyong Hu, Hang Zheng, Kun Li, Jianyun Xu, Weibo Mao, Maochun Luo, Lingxuan Wang, Mingxia Chen, Qihao Peng, Kaixuan Liu, et al. Fusionformer: A multi-sensory fusion in bird’s-eye-view and temporal consistent transformer for 3d object detection. *arXiv preprint arXiv:2309.05257*, 2023. 1, 3, 4, 6, 8
- [13] Haotian Hu, Fanyi Wang, Jingwen Su, Yaonong Wang, Laifeng Hu, Weiye Fang, Jingwei Xu, and Zhiwang Zhang. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. *arXiv preprint arXiv:2303.17895*, 2023. 6
- [14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 3
- [15] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3
- [16] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2561–2569, 2024. 2
- [17] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. 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*, pages 21643–21652, 2023. 6
- [18] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13903–13912, 2019. 2, 5
- [19] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 2, 6, 7
- [20] Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21209–21218, 2024. 1, 6
- [21] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. 1, 2, 3, 6
- [22] Yin hao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1, 3
- [23] Yangguang Li, Bin Huang, Zeren Chen, Yufeng Cui, Feng Liang, Mingzhu Shen, Fenggang Liu, Enze Xie, Lu Sheng, Wanli Ouyang, et al. Fast-bev: A fast and strong bird’s-eye view perception baseline. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

- [24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2, 3
- [25] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 1, 3
- [26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5
- [27] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 2, 4, 8
- [28] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2
- [29] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2, 4, 8
- [30] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petr2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023. 2, 4, 8
- [31] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 3, 6
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [33] Zhipeng Luo, Changqing Zhou, Gongjie Zhang, and Shijian Lu. Detr4d: Direct multi-view 3d object detection with sparse attention. *arXiv preprint arXiv:2212.07849*, 2022. 2
- [34] Jonah Philion and Sanja Fidler. 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*, pages 194–210. Springer, 2020. 1, 3
- [35] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 3
- [36] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 6
- [37] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017. 6
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 8
- [39] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 1
- [40] Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhen-guo Li, Bernt Schiele, and Liwei Wang. 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*, pages 6792–6802, 2023. 6
- [41] Jian Wang, Fan Li, and Haixia Bi. Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 6
- [42] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023. 2, 8
- [43] Yue Wang, Vitor Campanholo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 2021. 2
- [44] Zitian Wang, Zehao Huang, Yulu Gao, Naiyan Wang, and Si Liu. Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection. *arXiv preprint arXiv:2408.05945*, 2024. 2, 4, 8
- [45] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17591–17602, 2023. 1, 6
- [46] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023. 1, 2, 4, 6, 8
- [47] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 6
- [48] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, 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*, pages 17830–17839, 2023. 3
- [49] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via

- modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022. [1](#), [2](#), [6](#)
- [50] Zeyu Yang, Nan Song, Wei Li, Xiatian Zhu, Li Zhang, and Philip HS Torr. Deepinteraction++: Multi-modality interaction for autonomous driving. *arXiv preprint arXiv:2408.05075*, 2024. [1](#), [2](#)
- [51] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024. [2](#), [4](#), [6](#), [8](#)
- [52] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. [5](#)
- [53] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multi-modal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021. [1](#), [6](#)
- [54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. [2](#), [4](#)
- [55] Hongcheng Zhang, Liu Liang, Pengxin Zeng, Xiao Song, and Zhe Wang. Sparselif: High-performance sparse lidar-camera fusion for 3d object detection. In *European Conference on Computer Vision*, pages 109–128. Springer, 2024. [2](#), [4](#), [6](#), [8](#)
- [56] Yun Zhao, Zhan Gong, Peiru Zheng, Hong Zhu, and Shaohua Wu. Simplebev: Improved lidar-camera fusion architecture for 3d object detection. *arXiv preprint arXiv:2411.05292*, 2024. [1](#), [6](#)
- [57] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. [5](#)
- [58] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer, 2022. [2](#), [4](#), [8](#)
- [59] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. [6](#), [7](#)
- [60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020. [2](#), [3](#), [5](#), [8](#)