DVPE: Divided View Position Embedding for Multi-View 3D Object Detection

Jiasen Wang[1] , **Zhenglin Li**[1*] , **Ke Sun**[1] , **Xianyuan Liu**[2] and **Yang Zhou**[1]
[1]Shanghai University
[2]University of Sheffield
{wangjiasen, zhenglin_li, ke_sun}@shu.edu.cn, xianyuan.liu@sheffield.ac.uk, saber_mio@shu.edu.cn

Abstract

Sparse query-based paradigms have achieved significant success in multi-view 3D detection for autonomous vehicles. Current research faces challenges in balancing between enlarging receptive fields and reducing interference when aggregating multi-view features. Moreover, different poses of cameras present challenges in training global attention models. To address these problems, this paper proposes a divided view method, in which features are modeled globally via the visibility cross-attention mechanism, but interact only with partial features in a divided local virtual space. This effectively reduces interference from other irrelevant features and alleviates the training difficulties of the transformer by decoupling the position embedding from camera poses. Additionally, 2D historical RoI features are incorporated into the object-centric temporal modeling to utilize high-level visual semantic information. The model is trained using a one-to-many assignment strategy to facilitate stability. Our framework, named DVPE, achieves state-of-the-art performance (57.2% mAP and 64.5% NDS) on the nuScenes test set. Codes will be available at https://github.com/dop0/DVPE.

1 Introduction

Recently, vision-based multi-view 3D object detection in autonomous driving has garnered considerable attention due to its low cost and remarkable performance. Existing research can be categorized into two main groups: Bird's Eye View (BEV)-based and sparse query-based paradigms. The former [Philion and Fidler, 2020; Li *et al.*, 2022b; Li *et al.*, 2023b] conducts view transformation to explicitly construct BEV feature maps from multi-view images, providing a unified intermediate representation for various downstream tasks. However, achieving a trade-off between perception capabilities and computational complexity proves challenging due to the dense representation of BEV features and intricate view transformation. Comparatively, sparse query-based methods [Wang *et al.*, 2022b; Liu *et al.*, 2022] typically predefine sparse query embedding or reference points in 3D space and iteratively aggregate image features, which has been proven effective in 3D object detection.

Existing sparse query-based methods can be summarized into two branches based on the way of aggregating image features: sparse sampling and global attention. Sparse sampling-based approaches extract features through linear interpolation at specific projected points within image feature maps, as illustrated in Figure 1a. DETR3D [Wang *et al.*, 2022b] pioneered this group of methods, which projects 3D reference points onto vision feature maps based on camera parameters to sample features. While sparse sampling is efficient in terms of inference speed, the limited number of sampling points makes it challenging to achieve accurate global modeling, thereby hindering optimal performance. PETR [Liu *et al.*, 2022], as a representative work of global attention-based approaches (as illustrated in Figure 1b), encodes 3D coordinates as position embedding based on camera rays, thereby enabling image features to become 3D position-aware. Subsequently, global cross-attention is performed between queries and multi-view image features, which is largely redundant. For instance, looking up image features of the rear view does not contribute to detecting targets in front, but will lead to interference from irrelevant objects and unnecessary computation. Moreover, the existing PETR series [Liu *et al.*, 2022; Liu *et al.*, 2023b; Wang *et al.*, 2023b], initializes 3D reference points and outputs predictions in the 3D world space (i.e., the ego coordinate system). This compels models to adapt queries to account for camera poses through position embedding, thereby heightening the complexity of learning.

To address the aforementioned challenges, we propose a novel end-to-end framework based on divided view position embedding (DVPE) encoded in local virtual spaces. In contrast to typical position embedding methods that encode 3D coordinates of camera rays in the global world space, the proposed approach divides the 3D world space into multiple local virtual spaces. By transforming both 3D reference points and camera 3D coordinates into these virtual spaces, it decouples position embedding from camera poses and positions, thereby reducing the model's learning complexity. Subsequently, as illustrated in Figure 1c, we introduce a visibility attention module, performing cross-attention separately within each virtual space. This enables efficient global modeling by only interacting with image features 'visible' to each

[*] Corresponding author.

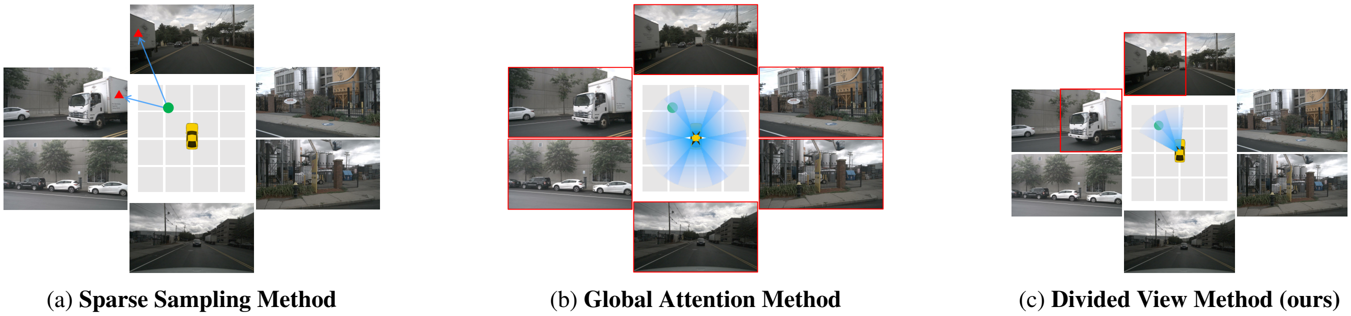
(a) **Sparse Sampling Method**
(b) **Global Attention Method**
(c) **Divided View Method (ours)**

Figure 1: Illustration of the image feature aggregation process of sparse sampling, global attention, and our proposed divided view approach, all belonging to sparse query-based paradigms. The green dot represents the 3D reference point of a query. Solid lines indicate projection onto the image plane, while triangles represent projected sampling points on images. The portion enclosed by a red frame represents the image features that need to interact with queries through cross-attention. The blue frustums represent the regions confirmed by camera rays, where the 3D coordinates are encoded to position embedding for cross-attention operations.

query, instead of the entire 360° view, thus effectively avoiding interference from irrelevant features.

Additionally, we introduce an object-centric temporal modeling approach enhanced by 2D vision prior. Specifically, historical Region of Interest (RoI) features from a 2D detector are incorporated to enrich the contextual information with high-level semantic features beyond decoder embedding, thereby facilitating the detection performance. Moreover, we explore a one-to-many assignment strategy in the training of 3D multi-view object detection frameworks, effectively enhancing the training stability by enriching the supervision of the decoder. The main contributions of this paper can be summarized as follows:

- We propose a multi-view 3D object detection framework named DVPE. It divides the global world 3D space into several local virtual spaces and performs visibility cross-attention based on the divided view position embedding, which effectively reduces the interference from irrelevant features and mitigates the learning difficulty by decoupling position embedding from camera poses.
- A 2D visual information-enhanced object-centric temporal modeling approach is proposed, which caches historical instance-level information from 2D detectors to enrich 3D features for temporal fusion.
- We introduce a one-to-many assignment training strategy to the framework for 3D object detection, effectively stabilizing the training of the decoder.
- Our model has been extensively evaluated on the nuScenes dataset, demonstrating superior performance over other camera-based 3D object detection methods and achieving state-of-the-art (SOTA) performance of 57.2% mAP and 64.5% NDS.

2 Related Work

2.1 Transformer-Based Object Detection

DETR [Carion *et al.*, 2020] represents a pioneering work in applying the transformer decoder [Vaswani *et al.*, 2017] for object detection, where learnable queries interact with both image features and object information. Each query generates one prediction and is assigned one-to-one to the ground truth, thereby avoiding the need for non-maximum suppression (NMS). Many subsequent works [Meng *et al.*, 2021; Liu *et al.*, 2021] focus on accelerating the convergence speed of DETR training. Specifically, Deformable DETR [Zhu *et al.*, 2020] proposes a deformable attention module to localize sparse features. DN-DETR [Li *et al.*, 2022a] and DINO [Zhang *et al.*, 2022] mitigate the instability of bipartite matching by introducing a denoising training method. Recent studies [Jia *et al.*, 2023; Chen *et al.*, 2023; Chen *et al.*, 2023] achieve notable performance by introducing a one-to-many assignment method to enhance training efficiency.

2.2 Multi-View 3D Object Detection

In contrast to monocular 3D object detection methods [Wang *et al.*, 2021; Wang *et al.*, 2022a], multi-view detection frameworks perceive multiple images from different cameras and predict bounding boxes in 3D world space by fully utilizing the correlation of multi-view images and geometric information of surrounding cameras. Based on whether vision features are transformed into BEV, these methods can be broadly categorized into two branches: BEV-based paradigm and sparse query-based paradigm.

BEV-based Paradigm. LSS [Philion and Fidler, 2020] lifts multi-view features into 3D space through depth estimation and splats them into BEV representation using voxel pooling. BEVDepth [Huang *et al.*, 2021] improves LSS by conducting depth supervision via LiDAR point cloud projection. BEVStereo [Li *et al.*, 2023a] and VideoBEV [Han *et al.*, 2023] perform BEV temporal fusion to enhance the performance. Different from LSS, many frameworks project image features back to BEV, typically using a transformer decoder. BEVFormer [Li *et al.*, 2022b] predefines a grid of reference points in BEV and projects them onto images in PV to obtain corresponding features with a transformer. PolarFormer [Jiang *et al.*, 2023] improves BEVFormer by replacing the Cartesian coordinate system with the polar coordinate system. Despite the unified and structured representation, BEV-based paradigms suffer from high computational cost due to the complex view transformation operations.

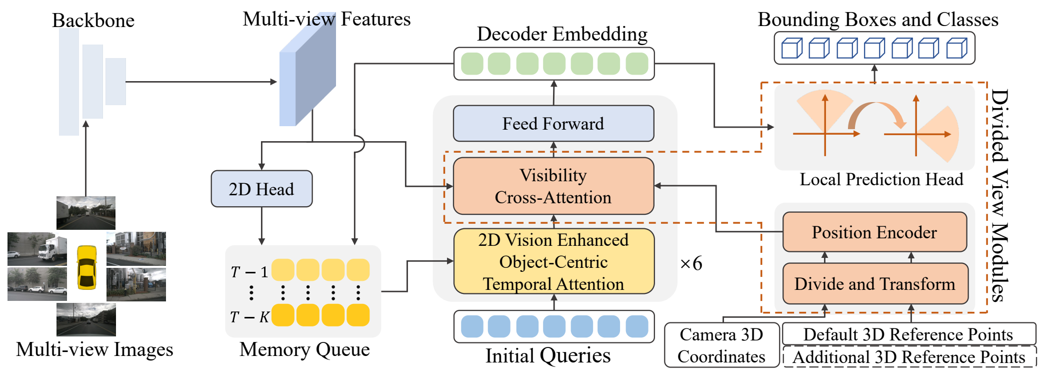

Figure 2: **Overall architecture of DVPE.** The framework is based on the transformer decoder, where the initial queries update iteratively through temporal attention and visibility cross-attention. In temporal attention, object queries interact with themselves as well as historical decoder embedding and 2D RoI embedding stored in the memory queue. Before visibility cross-attention, object queries and image features are grouped based on their 3D coordinates and then transformed into several local virtual spaces to obtain divided view position embedding. Isolated cross-attention is performed between queries and image features within different spaces. Subsequently, predictions are made in local virtual spaces and then transformed back to the 3D world coordinate system as final detection results, following which the memory queue is updated. Additional 3D reference points are used in conjunction with default 3D reference points for one-to-many assignment during training.

Sparse Query-based Paradigm. Sparse sampling and global attention are two key branches of methods belonging to the sparse query-based paradigm. Sparse sampling based methods suffer from limited information due to sparse sampling points and constrained receptive fields. Following DETR3D [Wang *et al.*, 2022b], several works increase the number of sampling points and propose approaches for the integration of their sampled features to compensate for the inadequate information resulting from single-point sampling [Lin *et al.*, 2022; Liu *et al.*, 2023a]. Alternatively, global attention methods have a global receptive field, but they are also susceptible to interference from irrelevant features. Among them, MV2D [Wang *et al.*, 2023c] generates reference points through 2D detection priors, while 3DPPE [Shu *et al.*, 2023] employes 3D point encoding with the help of depth estimation. CAPE [Xiong *et al.*, 2023] introduces camera view position embedding method to reduce the difficulty of view transformation learning, but it interacts with all surrounding image features through a bilateral cross-attention approach. To improve temporal modeling, PETRv2 [Liu *et al.*, 2023b] extends 3D position encoding to align 2D image features of different frames. StreamPETR [Wang *et al.*, 2023b] develops an efficient object-centric temporal fusion mechanism that propagates historical 3D decoder embedding as contextual cues.

3 Method

3.1 Overall Architecture

As illustrated in Figure 2, N surrounding images are initially processed by the 2D backbone network (e.g. ResNet [He *et al.*, 2016], VoVNet [Lee *et al.*, 2019]) to extract multi-view image features $\mathbf{F} = \{F_i \in \mathbb{R}^{C \times H_F \times W_F}, i = 1, 2, \ldots, N\}$, where C, H_F and W_F represent the channel, height, and width of the feature maps, respectively. Meanwhile, the feature maps $\mathbf{F}$ are fed into a 2D detector to obtain instance-level semantic features, which are then stored in the memory queue to enhance the object-centric temporal modeling.

Each decoder layer consists of two key components: temporal attention and visibility cross-attention. In the temporal attention module, object queries interact not only with themselves but also with cached historical object-centric embedding generated from both 2D and 3D features. Subsequently, visibility cross-attention is independently performed in each divided local virtual space, aided by the DVPE. Predictions made in virtual spaces are transformed back to the global 3D world coordinate system as final detection. The decoder embedding also updates the cached memory queue for processing future frames.

3.2 Divided View Method

Global cross-attention enables queries to interact with image features from all views, effectively extending the receptive field. However, it also induces high computational costs, some of which are unnecessary. Specifically, queries do not need to interact with irrelevant features whose areas are far from their corresponding targets. Considering this, we aim to achieve the same performance as global attention by having queries interact only with the aggregated image features surrounding them. To minimize interference computation with view locality, the 3D world space is divided into multiple frustum-shaped spaces based on the field of view, where visibility cross-attention can be independently performed within each local space. Furthermore, encoding the poses and positions of different cameras through global position embedding increases the difficulties of model training. Therefore, we transform the partitioned spaces into a unified virtual coordinate system, generating divided view position embedding. This enables the model to decouple position embedding from camera poses, thereby reducing the complexity of model learning. Finally, predictions of bounding boxes are made in the virtual coordinate system, which will be transformed back to the world coordinate system later.

Divided View Position Embedding. We discretize the camera rays starting from each pixel at (u, v) in a feature

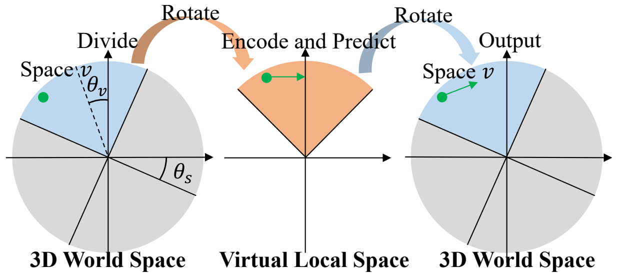

Figure 3: Illustration of space division (left) and transformation between the divided space and local virtual space in BEV. We only partition the global space into four, using the v-th space for illustration. The green dot denotes a 3D reference point, while the green arrows in the middle and on the right indicate the predicted yaws in the local virtual space and the world coordinate system.

map into 3D coordinates along the predefined D depth bins $\{p_d \in \mathbb{R}^3, d = 1, 2, \ldots, D\}$. The image 3D coordinates are transformed to 3D world space as:

$$p_i^c = H_i^{-1} p_{d,i} \tag{1}$$

where $H_i \in \mathbb{R}^{4\times4}$ denotes the matrix transforming the coordinates in 3D world space to the camera frustum space corresponding to the i-th view , and $p_{d,i}$ represents the homogeneous coordinates of a point in the i-th view at the d-th depth.

The furthest point of each ray is selected to represent the angle (in BEV) of each image token for the subsequent view division operation. The obtained ray-shaped camera 3D coordinates of all image tokens of N views are denoted as $\mathbf{P}^r \in \mathbb{R}^{N\cdot H_F\cdot W_F\times D\times3}$, while $\mathbf{P}^k \in \mathbb{R}^{N\cdot H_F\cdot W_F\times3}$ represents all the furthest points of each image ray. Points are partitioned into V groups based on their angles along the Z-axis (as illustrated on the left in Figure 3) as follows

$$g(p) = \lfloor \frac{(V \times \arctan(y/x) + \theta_s)}{2\pi} \rfloor \bmod V \tag{2}$$

where $g(p)$ denotes the group index of the point (x, y, z), θ_s is the starting shift angle (as shown in Figure 3), and mod represents the modulo operation. Based on the grouping results $g(\mathbf{P}^q)$ and $g(\mathbf{P}^k)$, 3D reference points $\mathbf{P}^q$, 3D camera coordinates $\mathbf{P}^r$, object queries $\mathbf{Q}$ and flattened image features $\mathbf{F}$ are grouped into different divided views. M_q learnable 3D reference points $\mathbf{P}^q \in \mathbb{R}^{M_q\times3}$ are randomly initialized within a cylindrical space to adapt to the rotational coordinate transformation needed for divided view position embedding.

Subsequently, the divided space coordinate systems are rotated along the z-axis to align with a unified virtual coordinate system with the following matrix:

$$R_v = \begin{bmatrix} \cos(\theta_v) & -\sin(\theta_v) & 0 \\ \sin(\theta_v) & \cos(\theta_v) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where θ_v is the angle difference between the v-th divided space and the unified virtual space. As illustrated from left to middle in Figure 3, a 3D reference point in space v is transformed into the local virtual coordinate system by rotating θ_v. Accordingly, divided view position embedding for queries and the keys, denoted as $\mathbf{Q}_v^{pe}$ and $\mathbf{K}_v^{pe}$ are obtained through respective position encoder:

$$\mathbf{Q}_v^{pe} = \psi(R_v \mathbf{P}_v^q) \tag{4}$$

$$\mathbf{K}_v^{pe} = \xi(R_v \mathbf{P}_v^r) \tag{5}$$

where ψ is the query position encoding function consists of sine-cosine position encoding function [Vaswani *et al.*, 2017] and multi-layer perceptron (MLP) ξ is the position encoding function for keys as PETR [Liu *et al.*, 2022]. $\mathbf{P}_v^q$ and $\mathbf{P}_v^r$ represent the 3D reference points and 3D camera coordinates of the v-th divided view, respectively.

Visibility Cross-Attention. As discussed above, similar performance to global attention can be achieved when cross-attention is conducted within each local view:

$$\begin{aligned} \mathbf{Q}_v \leftarrow \text{CrossAttn}(\text{Q} = \mathbf{Q}_v,\ \text{K,V} = \mathbf{F}_v, \\ \text{Q}^{\text{PE}} = \mathbf{Q}_v^{pe}, \text{K}^{\text{PE}} = \mathbf{K}_v^{pe}) \end{aligned} \tag{6}$$

Since the number of queries or keys in each view is not identical, we pad them to the same size for parallel computation. The padded invalid image features are masked during processing. Different views are treated as separate batches for isolation from each other. To ensure that the receptive field of edge queries is not affected, the starting shift angle θ_s in Eq.(2) is gradually increased with different decoder layers.

Local Prediction Head. Since queries aggregate image features in the virtual coordinate system, predictions are made in the same space as well. Therefore, the predicted bounding boxes need to be transformed back to the 3D world coordinate system. The center c, yaw α, and velocity ν of the bounding box in the world space are transformed based on the rotation matrix R_v and 3D reference points $\mathbf{P}_v^q$ as follows:

$$c = R_v^{-1}(R_v \mathbf{P}_v^q + \hat{c}), \quad \alpha = R_v^{-1}\hat{\alpha}, \quad \nu = R_v^{-1}\hat{\nu} \tag{7}$$

where $\hat{c}, \hat{\alpha}, \hat{\nu}$ represent the center offset, yaw and velocity of the bounding box predicted in the local virtual space, respectively. The transformation of yaw between two coordinate systems is shown from middle to right in Figure 3.

3.3 Enhanced Object-Centric Temporal Modeling

Current object-centric temporal modeling typically represent historical object features with query embedding O^q output by the decoder. To fully utilize the semantic information of multi-view images, we aim to incorporate historical RoI features from 2D detectors together with query embedding.

The bounding boxes with top-k class scores are selected, of which the RoI features $\mathbf{F}^e$ are extracted using RoIAlign [He *et al.*, 2017]. To match the shape of query embedding, RoI features $\mathbf{F}^e$, together with the class predictions, are encoded as $\mathbf{O}^e$ using a network. Based on the camera's intrinsic parameters I and RoI features, RoI points are generated to provide positional information for RoI embedding as follows:

$$\mathbf{P}^e = H^{-1}\text{MLP}([\text{Conv}(\mathbf{F}^e); I]). \tag{8}$$

Temporal attention conducts both self-attention and temporal interaction as follows:

$$\begin{aligned} \mathbf{Q} \leftarrow \text{CrossAttn}(\text{Q} = \mathbf{Q}, \\ \text{K, V} = [\mathbf{Q}\ \mathbf{O}_{t-1:t-k}^q\ \mathbf{O}_{t-1:t-k}^e]) \end{aligned} \tag{9}$$

where $[\cdot\,\cdot]$ represents the concatenation along the sequence dimension, and $\mathbf{O}^e_{t-1:t-k}$ represents the RoI embedding of last k frames. It is noteworthy that a standard position encoding is utilized, with cross-attention conducted in the global space.

Following the StreamPETR [Wang *et al.*, 2023b], the memory queue stores the RoI embedding $\mathbf{O}^e$, RoI points $\mathbf{P}^e$, ego transformation matrices $\mathbf{E}$, and timestamps $\mathbf{t}$ for motion embedding prediction. The historical RoI points are transformed in response to the ego movement of the vehicle as follows (taking time $t-k$ as an example):

$$\mathbf{P}^e_t = \mathbf{E}^t_{t-k}\mathbf{P}^e_{t-k} \tag{10}$$

where $\mathbf{E}^t_{t-k}$ transforms the coordinates of RoI points $\mathbf{P}^e_{t-k}$ in the ego coordinate system at time $t-k$ to the one at t.

3.4 One-to-Many Assignment and Loss

DETR-based detection methods are trained using a one-to-one assignment strategy, where each ground-truth object is assigned to a single prediction. It results in fewer positive samples compared to the traditional paradigm, leading to insufficient supervision information. To address this issue, Group-DETR [Chen *et al.*, 2023] introduces a one-to-many assignment approach in 2D detection to enrich the supervision for the decoder and achieves satisfactory performance. Therefore, we aim to explore the effectiveness of the one-to-many assignment strategy in 3D object detection.

Additional G groups of queries are introduced in addition to the default set in the proposed framework. During training, queries are matched one-to-one with GTs within each group. These groups of queries are separated in the temporal attention module to preserve the duplicate prediction suppression ability of self-attention. Only the default group is activated during inference, which is identical to the models trained with the one-to-one assignment.

Unlike the default group undergoing temporal attention, the additional groups involve self-attention at the temporal fusion stage. A denoising training method [Zhang *et al.*, 2022] is also employed. The overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{2d} + \lambda_1\mathcal{L}_{3d} + \lambda_2\mathcal{L}^{dn}_{3d} + \lambda_3\mathcal{L}^{a}_{3d} \tag{11}$$

where $\mathcal{L}_{2d}$ denotes the 2D detection loss, $\mathcal{L}_{3d}$, $\mathcal{L}^{dn}_{3d}$ and $\mathcal{L}^{a}_{3d}$ represent the 3D detection loss for default queries, denoising queries and additional queries, with weight coefficients λ_1, λ_2, λ_3. $\mathcal{L}_{3d}$ includes an ℓ_1 loss for bounding box regression and a focal loss [Lin *et al.*, 2017] for classification.

4 Experiments

4.1 Dataset and Metircs

Our framework is evaluated on the nuScenes dataset [Caesar *et al.*, 2020]. It contains 1k driving scenes, each with a duration of 20 seconds. The dataset is split into three groups: 750 for training, 150 for validation, and 150 for testing. Each scene includes 6 RGB images covering a 360° field of view, sampled at a frequency of 2Hz. Official metrics used in this paper include: mean Average Precision (mAP), nuScenes Detection Score (NDS), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE).

4.2 Implementation Details

We select StreamPETR [Wang *et al.*, 2023b] as our baseline. Following previous methods [Liu *et al.*, 2023a; Wang *et al.*, 2023b], for the backbone network, we use ResNet [He *et al.*, 2016] on the validation dataset and VoVNet [Lee *et al.*, 2019] on the test dataset. The 2D auxiliary supervision head from Focal-PETR [Wang *et al.*, 2023a] is employed to provide 2D RoI bounding boxes. The AdamW [Loshchilov and Hutter, 2017] optimizer with a cosine annealing policy is used for training DVPE. The learning rate and batch size are set to 4×10^{-4} and 16, respectively. Our models for performance comparison are trained for 60 epochs, whereas in ablation studies they are trained for 24 epochs. All experiments of DVPE are conducted without employing the CBGS [Zhu *et al.*, 2019] strategy or future frames.

For the proposed framework, the 3D world space is divided into 6 spaces and the shift angle is incremented by 20 degrees at each layer. By default, the top 128 2D RoI features are cached in a memory queue with a length of 4 frames. We adopt one additional group of queries to perform one-to-many assignment training, and the number of 3D object queries and additional ones are both set to 900.

4.3 State-of-the-art Comparison

NuScenes Val Split. Experimental results of the proposed framework and existing methods on the validation split of nuScenes are shown in Table 1. When adapting ResNet50 pre-trained on ImageNet-1K [Deng *et al.*, 2009], DVPE surpasses Sparse4Dv2 by 1.1% mAP and 1.6% NDS. When switching to backbone with nuImages pretraining, we reduce the number of queries to match StreamPETR and outperforms the baseline by 1.6% mAP and 0.9% NDS. With a larger backbone Resnet101 and resolution of 1408×512, DVPE also attains superior performance with 52.1% mAP and 60.3% NDS, exceeding the SOTA by 1.6% mAP and 0.9% NDS.

NuScenes Test Split. In Table 2, we provide the experimental results on the nuScenes test split and compare the proposed framework with previous SOTA methods . Following the common practice, we use VoVNet-99 as the backbone pretrained by DD3D [Park *et al.*, 2021]. In visual 3D object detection, DVPE achieves state-of-the-art performance with 57.2% mAP and 64.5% NDS, exceeding Sparse4Dv2 by 1.6% mAP and 0.7% NDS.

4.4 Ablation Study & Analysis

We conduct comprehensive experiments and analyze the effectiveness of each module in our work. All experiments are conducted with a ResnNet50 pretrained on nuImage. The number of queries is set to 900, training for 24 epochs. Table 3 provides detailed results of ablation experiments.

Divided View Method. From experiments (a) to (b) in Table 3, the divided view method significantly enhances the model with increase of 2.4% mAP and 2.5% in NDS and decrease of 4.9% mATE. In addition, the comparison between experiments (h) and (g) demonstrates that the removal of divided view method results in a drop of 2.7% mAP and 3.6% NDS, highlighting the method's substantial contribution to the DVPE. To explore the impact of the number of divided

Method	Backbone	Resolution	Epochs	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FB-BEV	R50	704×256	90‡	0.378	0.498	0.620	0.273	0.444	0.374	0.200
SOLOFusion	R50	704×256	90‡	0.427	0.534	0.567	0.274	0.511	0.252	0.181
Sparse4Dv2	R50	704×256	100	0.439	0.539	0.598	0.270	0.475	0.282	0.179
SparseBEV	R50†	704×256	36	0.448	0.558	0.581	0.271	0.373	0.247	0.190
StreamPETR	R50†	704×256	60	0.450	0.550	0.613	0.267	0.413	0.265	0.196
DVPE	R50	704×256	90	0.450	0.555	0.606	0.271	0.366	0.269	0.187
DVPE*	R50†	704×256	60	**0.466**	**0.559**	0.608	0.271	0.386	0.274	0.202
BEVDepth	R101	1408×512	90‡	0.412	0.535	0.565	0.266	0.358	0.331	0.190
BEVFormer	R101-D†	1600×900	24	0.416	0.517	0.673	0.274	0.372	0.394	0.198
AeDet	R101	1408×512	90‡	0.449	0.561	0.501	0.262	0.347	0.330	0.194
SOLOFusion	R101	1408×512	90‡	0.483	0.582	0.503	0.264	0.381	0.246	0.207
SparseBEV	R101†	1408×512	24	0.501	0.592	0.562	0.265	0.321	0.243	0.195
StreamPETR	R101†	1408×512	60	0.504	0.592	0.569	0.262	0.315	0.257	0.199
Sparse4Dv2	R101†	1408×512	100	0.505	0.594	0.548	0.268	0.348	0.239	0.184
DVPE	R101†	1408×512	60	**0.521**	**0.603**	0.544	0.262	0.318	0.248	0.200

Table 1: 3D object detection results on the nuScenes `val` split. † benefits from perspective pretraining. ‡ indicates methods with CBGS which will elongate 1 epoch into 4.5 epochs. * halves the number of queries for fair comparison.

Method	Backbone	Resolution	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
AeDet†	ConvNeXt-B	1600×640	0.531	0.620	0.439	0.247	0.344	0.292	0.130
SOLOFusion	ConvNeXt-B	1600×640	0.540	0.619	0.453	0.257	0.376	0.276	0.148
StreamPETR	VoVNet-99	1600×640	0.550	0.636	0.479	0.239	0.317	0.241	0.119
VideoBEV	ConvNeXt-B	1600×640	0.554	0.629	0.457	0.249	0.381	0.266	0.132
SparseBEV	VoVNet-99	1600×640	0.556	0.636	0.485	0.244	0.332	0.246	0.117
Sparse4Dv2	VoVNet-99	1600×640	0.556	0.638	0.462	0.238	0.328	0.264	0.115
DVPE	VoVNet-99	1600×640	**0.572**	**0.645**	0.466	0.240	0.319	0.266	0.126

Table 2: 3D object detection on the nuScenes `test` split. † uses test time augmentation. All methods in the table do not use future frames.

Exp	DVM	TM	OM	mAP↑	NDS↑	mATE↓
a				0.418	0.517	0.659
b	✓			0.442	0.542	0.610
c		✓		0.423	0.520	0.657
d			✓	0.428	0.521	0.653
e	✓	✓		0.446	0.549	0.604
f	✓		✓	0.451	0.551	**0.603**
g		✓	✓	0.428	0.518	0.651
h	✓	✓	✓	**0.455**	**0.554**	0.604

Table 3: Ablation study on nuScenes `val` split. DVM, TM and OM denote the divided view method, enhanced object-centric temporal modeling and one-to-many assignment, respectively.

#Views	θ_i (°)	mAP↑	NDS↑	mATE↓
4	30	0.452	0.550	0.616
6	20	**0.455**	**0.554**	0.604
8	15	0.453	0.552	**0.601**
6	60	0.449	0.548	0.614
6	30	0.451	0.549	0.614

Table 4: Results of models with different numbers of divided views and incremental steps of the shift angle θ_s at each decoder layer.

views and spatial division resulted from the shift angle, we conduct several experiments in Table 4. More variations in shift angles across different layers provides a more comprehensive receptive field for queries situated at the edges of the view, contributing to improve performance.

Applicability of Divided View Method. We also apply the proposed DVPE to PETR, which is a single-frame framework. As illustrated in Table 5, the mAP and NDS are enhanced by 1.5% and 1.2%, respectively, and the mAOE decreases significantly by 8.6% after incorporating our divided view method. This indicates that our DVPE work is not limited to specific models and can be applied to many sparse query-based multi-view object detection frameworks.

Visualization of Divided View Method. The visualization of image regions in a divided space and attention maps of a query is illustrated in Figure 4. Each divided space covers only a portion of six camera images. Instead of interacting with features from all views, the query looks up features within a divided virtual space (blue shaded regions), thus effectively reducing interference and redundant computation. It can be observed from the attention maps that the model

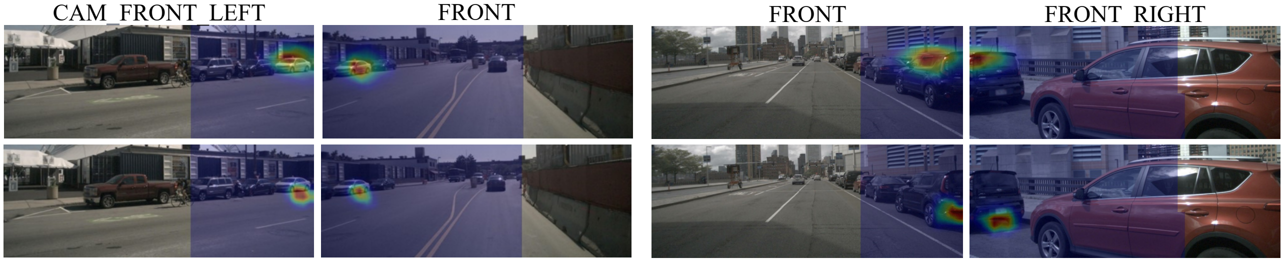

Figure 4: Visualization of image regions within one of the divided spaces and the corresponding attention maps of a query. Attention maps are from two heads of the last decoder layer.

Model	mAP↑	NDS↑	mATE↓	mAOE↓
PETR	0.299	0.342	0.766	0.679
PETR+DVM	**0.327**	**0.369**	**0.757**	**0.593**

Table 5: Application of divided view method to PETR. Models are trained for 24 epochs with Resnet50 pretrained on ImageNet-1K.

#RoIs	#Frames	mAP↑	NDS↑	mATE↓
32	4	0.450	0.552	0.616
64	4	0.455	0.553	0.617
128	4	0.455	0.554	**0.604**
128	2	0.449	0.550	0.615
128	6	**0.456**	**0.556**	0.614

Table 6: Results of experiments with different numbers of RoI proposals and cached historical frames.

#Add'l groups	#Add'l queries	mAP↑	NDS↑	mATE↓
1	450	0.440	0.541	0.632
1	900	0.455	**0.554**	**0.604**
2	450	0.449	0.548	0.617
2	900	**0.456**	**0.554**	0.605

Table 7: Results of experiments with different numbers of additional groups and additional queries.

can focus on targets even if they appear twice in overlapping fields of adjacent views, demonstrating the effectiveness of DVPE. Features of targets are highlighted even when they only partially appear in both views, as it can be treated as a single view for perception with the aid of divided view position embedding, validating the remarkable localization capability of the proposed framework.

Enhanced Object-Centric Temporal Modeling. Comparing experiment (a) with experiment (c) in Table 3, the addition of RoI features contributes to a gain of 0.5% mAP and 0.3% NDS. This indicates that RoI features and query features are complementary, collectively providing historical cues. In Table 6, we investigated the impact of the number of RoI proposals and cached historical frames on performance. As the number of RoI proposals increases to 128 and frames to 6, performance tends towards saturation.

One-to-Many Assignment. From experiments (a) and (d) in Table 3, one-to-many assignment brings an improvement of 1.0% mAP and 0.4% NDS. It demonstrates that even with denoising training, one-to-many assignment remains effective due to supplementary supervision. Additionally, based on experiments (c), (d), and (g), joint enhancement are not obtained from the combination of temporal modeling and one-to-many assignment. We argue that this phenomenon can be attributed to the divergent functionalities between the temporal attention of the default queries and self-attention of the additional queries. The shared parameters of temporal attention may impede the capacity of model for temporal fusion. In Table 7, we further analyze the impacts of the number of additional groups and queries on the one-to-many training strategy, finding that the number of queries has a larger impact than the number of groups.

5 Conclusion

This paper proposes a divided view position embedding approach to effectively aggregate image features for 3D object detection. The global 3D world space is partitioned into multiple local virtual spaces, utilizing the proposed DVPE to perceive the relative spatial locations of queries and image features. After subsequent visibility cross-attention, predictions are made in the local virtual coordinate system to decouple the position embedding from camera poses. To better utilize historical information to assist detection, we establish object-centric temporal modeling by incorporating 2D RoI features in addition to 3D query embedding. Furthermore, we investigate the role of a one-to-many assignment training strategy in stabilizing the training of the proposed framework. Experiments show that the DVPE achieves state-of-the-art performance on the vision-only nuScenes benchmark. We hope our work offers new insights into the interaction between queries and multi-view features for embodied agents.

Limitation and Future Work. While visibility cross-attention reduces the computational cost of global attention, the frequent partition and restoration of space negate this advantage. Therefore, there is a need for improvement to address this issue. Additionally, the misalignment between the one-to-many assignment and temporal attention might lead to sub-optimal results, which expects further exploration.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (Grant 62203289, 62225308), Shanghai Sailing Program (No. 22YF1413800).

References

[Caesar *et al.*, 2020] 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 multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2023] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group DETR: Fast DETR training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642, 2023.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[Han *et al.*, 2023] Chunrui Han, Jianjian Sun, Zheng Ge, Jinrong Yang, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3D perception. *arXiv preprint arXiv:2303.05970*, 2023.

[He *et al.*, 2016] 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.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[Huang *et al.*, 2021] 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.

[Jia *et al.*, 2023] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. DETRs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023.

[Jiang *et al.*, 2023] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. PolarFormer: Multi-camera 3D object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023.

[Lee *et al.*, 2019] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and GPU-computation efficient backbone network for real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 0–0, 2019.

[Li *et al.*, 2022a] 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.

[Li *et al.*, 2022b] 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 *Proceedings of the European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[Li *et al.*, 2023a] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1486–1494, 2023.

[Li *et al.*, 2023b] Yinhao 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*, volume 37, pages 1477–1485, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Lin *et al.*, 2022] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4D: Multi-view 3D object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[Liu *et al.*, 2021] 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. In *Proceedings of International Conference on Learning Representations*, 2021.

[Liu *et al.*, 2022] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *Proceedings of the European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[Liu *et al.*, 2023a] 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.

[Liu *et al.*, 2023b] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. PETRv2: 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.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Meng *et al.*, 2021] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.

[Park *et al.*, 2021] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3D object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021.

[Philion and Fidler, 2020] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *Proceedings of the European Conference on Computer Vision*, pages 194–210. Springer, 2020.

[Shu *et al.*, 2023] Changyong Shu, Jiajun Deng, Fisher Yu, and Yifan Liu. 3DPPE: 3D point positional encoding for transformer-based multi-camera 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3589, 2023.

[Vaswani *et al.*, 2017] 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.

[Wang *et al.*, 2021] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.

[Wang *et al.*, 2022a] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of Conference on Robot Learning*, pages 1475–1485. PMLR, 2022.

[Wang *et al.*, 2022b] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Proceedings of the Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[Wang *et al.*, 2023a] Shihao Wang, Xiaohui Jiang, and Ying Li. Focal-PETR: Embracing foreground for efficient multi-camera 3D object detection. *IEEE Transactions on Intelligent Vehicles*, 2023.

[Wang *et al.*, 2023b] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. *arXiv preprint arXiv:2303.11926*, 2023.

[Wang *et al.*, 2023c] Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Object as query: Equipping any 2D object detector with 3D detection ability. *arXiv preprint arXiv:2301.02364*, 2023.

[Xiong *et al.*, 2023] Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, and Xiang Bai. CAPE: Camera view position embedding for multi-view 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21570–21579, 2023.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations*, 2022.

[Zhu *et al.*, 2019] 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.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of International Conference on Learning Representations*, 2020.