Multi-View Attentive Contextualization for Multi-View 3D Object Detection

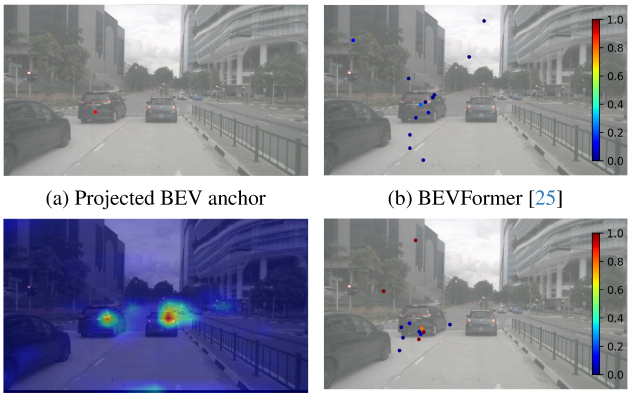
Xianpeng Liu[1], Ce Zheng[2], Ming Qian[3], Nan Xue[3], Chen Chen[2], Zhebin Zhang[4], Chen Li[4], Tianfu Wu[1]
[1]North Carolina State University [2]University of Central Florida
[3]Ant Group [4]OPPO U.S. Research Center
https://xianpeng919.github.io/mvacon

Abstract

We present Multi-View Attentive Contextualization (MvACon), a simple yet effective method for improving 2D-to-3D feature lifting in query-based multi-view 3D (MV3D) object detection. Despite remarkable progress witnessed in the field of query-based MV3D object detection, prior art often suffers from either the lack of exploiting high-resolution 2D features in dense attention-based lifting, due to high computational costs, or from insufficiently dense grounding of 3D queries to multi-scale 2D features in sparse attention-based lifting. Our proposed MvACon hits the two birds with one stone using a representationally dense yet computationally sparse attentive feature contextualization scheme that is agnostic to specific 2D-to-3D feature lifting approaches. In experiments, the proposed MvACon is thoroughly tested on the nuScenes benchmark, using both the BEVFormer and its recent 3D deformable attention (DFA3D) variant, as well as the PETR, showing consistent detection performance improvement, especially in enhancing performance in location, orientation, and velocity prediction. It is also tested on the Waymo-mini benchmark using BEVFormer with similar improvement. We qualitatively and quantitatively show that global cluster-based contexts effectively encode dense scene-level contexts for MV3D object detection. The promising results of our proposed MvACon reinforces the adage in computer vision – "(contextualized) feature matters".

1. Introduction

Camera-based 3D object detection is a pivotal area of research in computer vision, particularly owing to its appealing applications in cost-effective autonomous systems, such as autonomous driving and robot autonomy. Recently, significant progress has been witnessed in the field of multi-view 3D (MV3D) object detection, especially with the advent of end-to-end MV3D detection approaches [38]. *One crucial component in the end-to-end MV3D detection system is the 2D-to-3D feature lifting module*, which converts perspective 2D multi-view image feature maps to 3D feature representation. It aims to counter the complete loss of depth information in individual 2D images by exploiting multi-view clues. However, a significant challenge in practical applications like autonomous driving is the often insufficient field-of-view overlap across views, making it difficult to effectively address the loss of depth information.



(a) Projected BEV anchor (b) BEVFormer [25]

(c) Learned clusters (d) Our MvACon

Figure 1. The effects of our proposed MvACon in the 2D-to-3D feature lifting. Consider a point (red) on the car in (a), which is projected from a 3D BEV anchor point. In lifting 2D features to ground the 3D BEV anchor point, vanilla BEVFormer [25] utilizes a predefined number of deformable points, with offsets learned through a 6-layer cross-attention module relative to the projection point. (b) shows the deformed points after the final cross-attention layer, most of which have low attention weights, indicating the model's uncertainty or inability to consolidate contributions effectively for good lifting. Our MvACon tackles this issue with clustering-based attention, as visualized in (c). (d) shows the deformed points, where we observe not only high-confidence points on the car but also on the building. We further observe that the points on the building remain stable across encoding layers (see Fig. 4) and consecutive frames (see suppl.). With those high-confident deformed points in a spatiotemporally stable configuration, our MvACon may induce a local object-context aware coordinate system that helps the overall performance, especially the estimation of velocity and orientation, as we quantitatively observed in experiments. See text for details.

To perform 2D-to-3D feature lifting, recent methods

[15, 25, 29, 57] aim to learn a unified 3D space representation using 3D anchors that are either sparsely or uniformly sampled. These methods generally fall into two categories subject to the interaction of 3D anchors with 2D features and the feature aggregation strategy. (1) The Lift-Splat-Shoot (LSS) method [14, 15, 23, 24, 45] first lifts 2D features into 3D (pseudo-LiDAR) space using the outer product with the estimated depth, then assigns them to the nearest 3D anchors. (2) In contrast, the query-based design [25, 29, 57], pioneered by the DETR method [4] for end-to-end 2D object detection, adopts 3D anchors as queries and uses 2D image features as keys and values. They interact and aggregate via spatial cross-attention in the expressive Transformer architecture [50]. These two paradigms have been widely used in downstream tasks like map segmentation [30] and occupancy prediction [25]. *This paper primarily focuses on the query-based detection paradigm.* One reason is that LSS-based methods often encounter excessive computational complexity and issues with error propagation and depth estimation magnification post-lifting, potentially capping their performance. However, the query-based design also grapples with heavy computation costs or limited 3D information awareness, depending on their Transformer design.

In this paper, we focus on addressing limitations of two main paradigms of query based MV3D object detectors in a unified way (elaborated in Sec. 3). In particular, we introduce multi-view attentive contextualization (MvACon) to address limitations of decoder-only dense attention methods like PETR [29], which lack high-resolution features due to computational constraints, and to simultaneously address the issue of sparsely grounded 3D anchors in encoder-decoder 2D/3D deformable attention methods such as BEVFormer [25] and DFA3D [19]. *Our proposed MvACon aims to be representationally dense while computationally sparse*. To achieve this, we expand the conventional three-component paradigm of MV3D object detection to a four-component setup (Fig. 2): (1) 2D image representation learning through a feature backbone shared across views, (2) *MvACon for attentive contextualization of the 2D features*, (3) 2D-to-3D feature lifting, and (4) a 3D object detection head or decoder that utilizes these lifted features. This modular design allows our MvACon to remain agnostic to specific 2D-to-3D feature lifting strategies and aligns with the classic adage in representation learning and computer vision: —'(contextualized) feature matters'.

More specifically, our approach contextualizes original feature maps extracted from the backbone network using a cluster-attention operation. This builds upon the recently proposed patch-to-cluster attention (PaCa) [12]. For perspective-based decoder-only detectors like PETR, we apply cluster contextualization before the feature maps are fed into the decoder. For encoder-decoder based detectors, such as BEVFormer and DFA3D, we incorporate cluster contextualization within the spatial cross-attention operation. Through extensive experiments, we demonstrate that our proposed MvACon effectively and consistently enhances query-based MV3D object detectors by encoding more useful contexts, thereby facilitating better 2D-to-3D feature lifting. Rigorously controlled experiments reveal that, for perspective-based decoder-only detectors, the cluster attention contextualization significantly improves localization and velocity prediction. In the case of encoder-decoder based detectors, it effectively reduces errors in location, orientation, and velocity.

In summary, our main contributions are:

- We analyze and address the limitation of 2D-to-3D feature lifting in the prior art, that is the lack of sufficient 3D representational power due to their local 3D awareness.
- We propose MvACon (Multi-view Attentive Contextualization) to induce the global 3D awareness in an easy-to-integrate way to enhance the 2D-to-3D feature lifting in both decoder-only based MV3D object detectors and encoder-decoder based ones.
- We show consistent performance improvement of our MvACon on the challenging NuScenes [3] dataset using three baseline query-based MV3D object detectors, as well as on the Waymo-mini [11] benchmark.

2. Related Work

Camera-based 3D Object Detection. Camera-based 3D object detection can be primarily categorized into two settings: single-view and multi-view. In the realm of monocular 3D object detection research, addressing the challenge of inaccurate object localization [37] is critical. Researchers have exerted considerable effort to utilize monocular depth cues. This includes transforming inputs into pseudo-lidar point clouds [35, 36, 56] and explicitly incorporating depth into models [8, 9, 16]. Another significant research direction involves the explicit use of geometric priors, encompassing approaches like key-point constraints [6, 20, 33], shape projection relationships [22, 34, 40, 64], and temporal depth estimation [54]. Innovations in monocular 3D object detection also include novel loss modules [7, 47, 65], 3D-aware backbones [2, 17, 18], and second-stage detection paradigms [28, 44]. In multi-view settings, the configuration often closely resembles that of monocular setups due to the limited field-of-view overlap between the different camera views. Therefore, MV3D object detection focuses on addressing the challenge of learning universal representation for multi-view sensors. It has benefited from advancements in various techniques such as view lifting [5, 19, 25, 29, 45, 57], depth encoding [19, 23, 24, 46], and temporal modeling [14, 25, 26, 30, 43, 52]. In our multi-view approach, we aim at addressing the challenge of multi-view representation learning with focus on en-

hancing the view lifting module within query-based detection methods by empowering original features with cluster-based contextual features.

Representation Learning in Camera-based 3D Object Detection. Camera-based 3D object detection is inherently a data-intensive task due to its ill-posed nature, the expansive search space in 3D, and the scarcity of labeled data in scenes. Consequently, developing robust representations for this task is both critical and challenging. Early research in monocular 3D object detection has demonstrated the utility of depth contexts [41, 42] and projection contexts [27] in enhancing detection capabilities. Recent advances also highlight the effectiveness of scene-level representations, such as density fields [39], in improving 3D representation learning [59]. In the domain of multi-view research, most leading methods utilize backbones pre-trained with projection contexts (e.g., FCOS3D [53] weights) or depth contexts (e.g., DD3D [41] weights). However, these pre-trained weights may not fully leverage the capabilities of newer backbone network designs. Recent studies have begun to explore alternatives to this pre-trained paradigm, including the integration of an auxiliary projection context branch in end-to-end training [60]. Our work aims to enhance network representation by explicitly incorporating scene-level cluster context as supplementary information during the view lifting stage in query-based MV3D object detectors.

Vision Transformers. Since the pioneer work of ViT [10], extensive research [31] has been dedicated to enhancing the representational abilities of neural networks for visual tasks. It has been established that CNNs and Transformers can mutually augment each other's capabilities, as evidenced in designs like Transformer-enhanced CNNs [48, 58] and CNN-enhanced Transformers [21, 49, 61]. Additionally, a significant branch of visual Transformer research focuses on developing new attention mechanisms tailored to the locality bias in vision tasks. Notable examples include HaloNet [51], SWin [32], Deformable Attention [67], and VOLO [62], all introducing innovative local attention mechanisms to mitigate the quadratic computational cost associated with visual inputs. Concurrently, models like TNT [13], ViL [63], PVTv2-linear [55], POTTER [66], and PaCa [12] explore the integration of local and global contexts. Inspired by the progress of Vision Transformers, our work addresses the limitations of two prevalent paradigms in query-based MV3D detectors caused by their attention mechanisms.

3. Approach

Given a set of images $I_i \in \mathbb{R}^{3\times H\times W}$ from N cameras with known extrinsics $T_i \in \mathrm{SE}(3)$ and intrinsics $K_i \in \mathbb{R}^{3\times 3}$, MV3D object detection aims to infer the label (e.g., Car, Pedestrian, Barrier) and the 3D bounding box for each object instance in the scene. In this section, we first delve into the pipeline of query-based MV3D object detection in Sec. 3.1. We then analyze the pros and cons of the core 2D-to-3D feature lifting component in two state-of-the-art MV3D object detection methods in Sec. 3.2. Finally, we present our proposed MvACon in Sec. 3.3.

3.1. Query-based MV3D Object Detection

For better understanding, we explain the query-based MV3D object detection pipeline, shown in Fig. 2, in a reverse manner. The 3D detection head typically builds upon DETR3D [57], which is based on the original DETR [4]. Initially, it defines a sufficiently large number, O, of latent C-dimensional 3D object queries, $Q_{O,C}$. These object queries update using Keys and Values derived from multi-view inputs, followed by a classification head predicting object labels and a bounding box regression head determining the 3D bounding boxes. Obviously, the key challenge lies in how to transform / lift multi-view 2D inputs into 3D-aware Keys and Values.

To this end, a feature backbone is trained to extract deep 2D features from the multi-view input images. The complexity arises from different design choices for feature lifting. It mainly involves two aspects: representation and computation. From the representational perspective, multi-scale feature pyramids, crucial in 2D object detection, become even more essential in 3D object detection. Computationally, handling multi-view inputs is already demanding. Adding multi-scale feature pyramids without careful optimization can significantly increase the computational load. There are two strategies in the state-of-the-art development of query-based MV3D object detection.

Decoder-Only Architectures: Single-Scale Multi-View 2D Features with Dense Attention. These designs are among the most straightforward. They involves using multi-view 2D feature maps from the last layer of the feature backbone, which are then concatenated and flattened along the spatial dimensions to form Keys and Values. The latent object queries, $Q_{O,C}$, are updated using a vanilla Transformer (i.e., each object query attends to every 2D location in the multi-view inputs, termed dense attention). However, this basic approach often fails as it does not encode any 3D-aware information. To address this, The PETR [29] introduces a physically-meaningful 3D position transformation as positional encoding, added to the multi-view 2D feature maps before concatenation and flattening. It has proven effective for query-based MV3D object detection, as illustrated in Option 1 in Fig. 2).

Encoder-Decoder Architectures: Multi-Scale Multi-View 2D Features to Latent BEV Queries with Sparse Attention. The BEV (Bird's Eye View) representation acts as a unified, grid-based and ego-centric scene representation with predefined grid sizes (e.g., 200×200) on the XZ

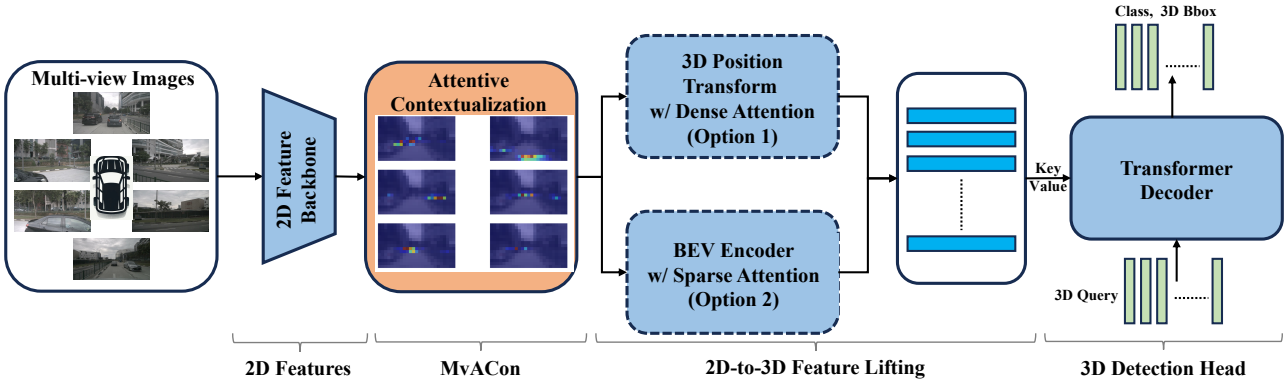


Figure 2. Overview of a query-based MV3D object detection pipeline with our proposed MvACon. Our proposed MvACon is a plug-and-play module for two state-of-the-art query-based MV3D object detection paradigms (e.g., PETR [29] and BEVFormer [25] respectively), which computes attentively contextualized features to facilitate better 2D-to-3D feature lifting in the two paradigms. See text for details.

plane. Geometrically, the BEV grid can be treated as a pillar-based point-cloud representation, collapsed along the Y-axis. A predefined number of points along the pillar (Y-axis) direction in the BEV grid will be uniformly sampled. These points form a uniform geometry prior for the underlying 3D scene and serve as BEV anchors to elevate 2D features into the BEV space. With known camera poses, these sampled points can be projected to each view at multiple scales. However, due to the uniform geometry prior in the projection, the projected points require deformability to better align with data observations. The BEVFormer [25] addresses this by introducing a sparse deformable attention mechanism (see Option 2 in Fig. 2). To counter the uniform geometry prior further, it learns a small predefined number of offsets, rather than directly deforming the projected points on each view, to lift 2D features from those deformed points with attentive weights. Latent BEV queries are introduced in learning these offsets and attentive weights. The BEV encoder's role is to refine the BEV queries, enabling them to provide meaningful offsets and attentive weights for lifting 2D features to 3D BEV anchors. Finally, embedded BEV queries as Keys/Values in the 3D detection head (i.e., decoder) update the latent object queries, $Q_{O,C}$, e.g., through sparse deformable attention as in the BEVFormer, before predicting the 3D object detection results.

3.2. The Limitation of 2D-to-3D Feature Lifting in the Prior Art

Although they both have shown remarkable progress for MV3D object detection, the PETR pipeline and the BEVFormer pipeline have a common limitation in their 2D-to-3D feature lifting, that is the local or shallow 3D awareness, rather than the desirable counterpart, the global and semantic meaningful 3D awareness.

In the PETR pipeline, consider the 2D feature map of the n-th view, $F^n_{h\times w\times c}$, where (h, w) are the spatial sizes, height and width respectively, and c the feature dimension of the backbone. The 3D position transform converts the (shared) camera frustum discretized as a mesh grid of sizes (h, w, D) to the 3D space based on the known camera poses, where D is the discretized depth levels. After the conversion, each 3D point is represented by a normalized 3D coordinate in the homogeneous form, i.e., $(x, y, z, 1)$. So the 3D position transformation results in the positional encoding $P^n_{h\times w\times 4\cdot D}$. Both $F^n_{h\times w\times c}$ and $P^n_{h\times w\times 4\cdot D}$ are projected into the space of the same dimensionality, d using a linear layer and a Multi-layer Perceptron (MLP) with the ReLU nonlinearity respectively, we have $F^n_{h\times w\times d}$ and $P^n_{h\times w\times d}$ which are summed in an element-wise way. Consider the latent representation of the 3D position in $P^n_{h\times w\times d}$ with the depth grid fused by the MLP, it is grounded to one feature point in $F^n_{h\times w\times d}$, leading to the local 3D awareness.

In the BEVFormer pipeline, as we show in Fig. 1, the projected BEV anchor is often grounded to some low-confidence and scattered deformed points. Although the grounding may not be spatially local, they are often semantically shallow.

3.3. Our Proposed MvACon Method

Our goal is to address the local or shallow 3D awareness stated above by introducing an easy-to-integrate (mostly plug-and-play) module to learn the global and semantically meaningful 3D-awareness, as illustrated in Fig. 2. The basic idea of our MvACon is to attentively contextualize the 2D features in the 2D-to-3D lifting.

In the PETR pipeline, our idea is to contextualize the individual 2D feature map, $F^n_{h\times w\times d}$,

$$\mathbf{F}^n_{h\times w\times d} = \text{MvACon}(F^n_{h\times w\times d}), \tag{1}$$

where after the contextualization every feature point in $\mathbf{F}^n_{h\times w\times d}$ can connect to the entire map $F^n_{h\times w\times d}$, inducing the global 3D awareness for grounding the positional encoding $P^n_{h\times w\times d}$.

In the BEVFormer pipeline, our idea is to contextualize the multi-scale feature maps, e.g., the l-th layer of the fea-

ture pyramid of the n-th view, $F^{n,l}_{h\times w\times c}$,

$$\mathbf{F}^{n,l}_{h\times w\times c} = \text{MvACon}(\{F^{n,l}_{h\times w\times c}\}^{L}_{l=1}), \quad (2)$$

where after the contextualization every feature point in $\mathbf{F}^{n,l}_{h\times w\times c}$ can connect to the entire L-layer feature pyramid $\{F^{n,l}_{h\times w\times c}\}^{L}_{l=1}$, inducing the global 3D awareness for grounding projected BEV anchors on the n-the view.

To achieve the global contextualization effect, we adapt the recently proposed Patch-to-Cluster attention (PaCa) [12] method. The core idea of PaCa is to leverage a learnable clustering module to cluster a feature map into a predefined number M of clusters. For notional simplicity, consider a feature map $F_{h\times w\times c}$ as $N = h \times w$ tokens $F_{N\times c}$, the clustering assignment is computed by,

$$\mathcal{C}_{N,M} = \text{Softmax}(\text{Clustering}(F_{N,c})), \quad (3)$$

where Clustering() can be implemented in different ways (see our ablation studies in Tab. 6), and the Softmax is along the token dimension. Then, we compute M clusters by,

$$z_{M,c} = \text{LN}(\mathcal{C}^{\top}_{N,M} \cdot F_{N,c}), \quad (4)$$

where LN() is the layer normalization [1].

Then, the PaCa-based MvAcon is defined by,

$$F'_{N,c} = \text{Softmax}(\frac{Q_{N,c} \cdot K^{\top}_{M,c}}{\sqrt{c}}) \cdot V_{M,c} + F_{N,c}, \quad (5)$$

where $Q_{N,c}$ is the linear projection of $F_{N,c}$, $K_{M,c}$ and $V_{M,c}$ are from the clusters $z_{M,c}$. The second term is the shortcut. The multi-head PaCa can be straightforwardly defined. For Eqn. 2, we concatenate the clusters from all the pyramid layers before computing the Key and the Value. Here, the PaCa module is of linear complexity.

4. Experiments

4.1. Experimental Setup

Dataset and Metrics We evaluate our MvAcon on the challenging large-scale NuScenes dataset [3] and Waymo dataset [11]. The **NuScenes** dataset includes 1,000 scene sequences, which are divided into training, validation, and testing subsets in a 700/150/150 split. Each sequence in the NuScenes dataset is a 20-second video clip, annotated at a rate of 2 frames per second (FPS). The NuScenes dataset employs a comprehensive suite of evaluation metrics for assessing detection performance. These metrics comprise mean Average Precision (mAP), mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribution Error (mAAE), and the NuScenes Detection Score (NDS). The **Waymo** dataset contains 798 training sequences and 202 validation sequences. We use a subset of the training set (Waymo-mini) by sampling every third frame from the training sequences following [25].

Implementation Details We leverage open-source code bases (PETR [29], BEVFormer [25], and DFA3D [19]) in our experiments. To ensure a fair and stringent comparison, we maintain all original configurations of these methods, making only one modification: the addition of an attentive contextualization module. We conduct qualitative analysis and ablation study on the BEVFormer-base model. We train all models for 24 epochs using 8 NVIDIA Tesla A100 GPUs, following the configurations and settings outlined in previous works [19, 25, 29].

4.2. The Effectiveness of MvACon across Different Methods

To demonstrate the effectiveness of our proposed MvACon method, we first apply it to two typical query-based MV3D object detection paradigms: the perspective-based decoder-only detector (PETR [29]) and the encoder-decoder based detector (BEVFormer [25]). We choose these as our baselines because state-of-the-art (SOTA) query-based MV3D detectors [19, 30, 52, 60] primarily follow these two paradigms. We also test our method on DFA3D [19] to demonstrate its generalizability to SOTA methods.

On the NuScenes dataset, Table 1 shows that our proposed MvACon consistently improves performance across different detectors. Specifically, for the perspective-based decoder-only detector PETR, it improves the baseline by 0.8 NDS. For the encoder-decoder based detector BEVFormer, our method achieves an improvement of 1.3 in NDS on average. On a more advanced, depth-context enhanced BEVFormer (DFA3D), our method further improves performance by up to 0.5 NDS. Notably, our MvACon achieves significant improvement in location (mAP, mATE), orientation (mAOE), and velocity prediction (mAVE) for encoder-decoder based detectors. It also markedly enhances performance in location (mAP, mATE) and velocity (mAVE) prediction for the perspective-based decoder-only detector.

On the Waymo dataset, since there are few released codes for MV3D detectors on Waymo except for the BEVFormer, we only test BEVFormer on Waymo-mini following its settings with results shown in Table 2. Our MvACon shows consistent improvement on Waymo metrics.

4.3. How Does Our MvACon Work?

We elaborate on the effects of our MvACon by providing detailed analyses during the 2D-to-3D feature lifting process. We first demonstrate **what the learned cluster contexts encode**, then show **how these contexts affect the behavior of deformable points during feature lifting**. Lastly, we illustrate **how our MvACon improves detection results** by presenting a qualitative comparison within a scene. We select BEVFormer-base as our analysis target due to its incorporation of six layers of deformable attention modules in the encoder. More qualitative analysis is provided in the supplementary materials.

What do the Learned Cluster Contexts Encode? We vi-

Method	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
PETR-VovNet-99 [29]	37.8	74.6	27.2	**48.8**	90.6	**21.2**	42.6
PETR-VovNet-99-MvACon	**38.2 (+0.5)**	**73.9**	**27.0**	50.5	**84.1**	21.3	**43.4 (+0.8)**
BEVFormer-t [25]	25.2	90.0	29.4	65.5	65.7	**21.6**	35.4
BEVFormer-t-MvACon	**25.9 (+0.7)**	**88.4**	**28.8**	**64.6**	**60.5**	22.5	**36.5 (+1.1)**
BEVFormer-s [25]	37.0	72.1	28.0	40.7	43.6	22.0	47.9
BEVFormer-s-MvACon	**39.3 (+2.3)**	**71.3**	**27.7**	**40.1**	**42.0**	**19.7**	**49.6 (+1.7)**
BEVFormer-b [25]	41.6	67.3	**27.4**	37.2	39.4	**19.8**	51.7
BEVFormer-b-MvACon	**42.6 (+1.0)**	**66.4**	27.6	**35.0**	**36.2**	20.0	**52.8 (+1.1)**
DFA3D-s [19]	**40.1**	72.1	27.9	41.1	39.1	**19.6**	50.1
DFA3D-s-MvACon	**40.1**	**71.0**	**27.4**	**38.3**	**37.2**	20.8	**50.6 (+0.5)**
DFA3D-b [19]	43.0	**65.4**	**27.1**	37.4	34.1	**20.5**	53.1
DFA3D-b-MvACon	**43.2 (+0.2)**	66.4	27.5	**34.4**	**32.3**	20.7	**53.5 (+0.4)**

Table 1. Comparisons of our method with baselines on the NuScenes validation set. BEVFormer-t/s/b refers to BEVFormer-tiny/small/base in the BEVFormer's open-source codes. 'DFA3D' refers to the adaptation of 2D deformable attention into a (depth-weighted) 3D deformable attention within the BEVFormer model, as adopted in [19].

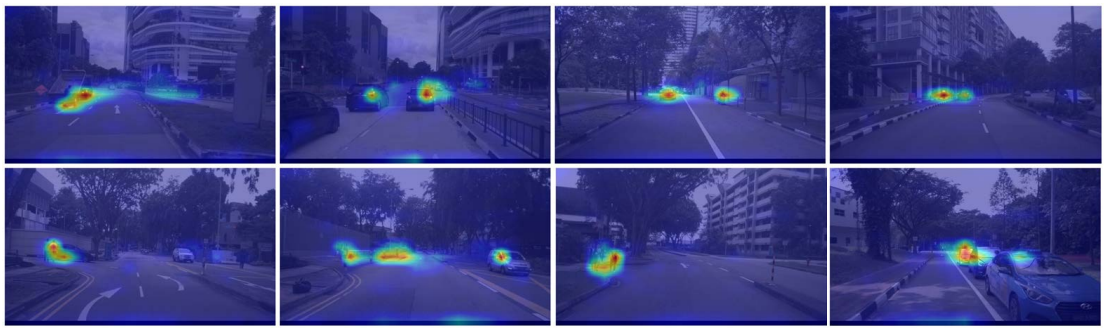


Figure 3. Visualization results of learned cluster contexts in our MvACon on the NuScenes validation set. We sum all the learned clusters along the channel and upsample it to the original image resolution through bilinear interpolation. We observed that the learned cluster context encodes abundant context information in the scene. We provide details with raw images in the supplementary.

Method	LET-mAPL↑	LET-mAPH↑
BEVFormer-ResNet101 [25]	34.9	46.3
BEVFormer-ResNet101-MvACon	**35.7 (+0.8)**	**47.5 (+1.2)**

Table 2. Comparisons on the Waymo-mini.

sualize the learned cluster context in a heatmap, by summing all clusters along the channel and then upsampling to the original image resolution using bilinear interpolation. The results, shown in Fig. 3, reveal that despite the low resolution of the original summed heatmap, it can accurately locate foreground objects in the scene after upsampling. This suggests that the cluster contexts encode the foreground layouts in a scene. Furthermore, we note that the upsampled heatmap shows a high response to various foreground objects, even when they are spatially close in 2D. This is attributed to the effectiveness of our MvACon in dense scenes, as demonstrated and analyzed later in our experiments.

How do Cluster Contexts Affect the Behavior of Deformable Points? We demonstrate the dynamics of deformable points in Fig. 4. These points originates from one 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder. As the encoding layers deepen in the vanilla BEVFormer, we observe that most deformable points have low attention weights, suggesting the model's uncertainty about the relevance of the selected context. Consequently, these contexts contribute minimally during the encoding of BEV query features in the feature lifting process. In contrast, deformable points predicted by our method maintain high confidence weights on foreground objects (e.g., cars), as well as on surrounding buildings. We also note that points on buildings remain stable across encoding layers and consecutive frames (see the supplementary). These high-confidence deformed points in our MvACon may foster a local object-context aware coordinate system, enhancing overall performance, including the estimation of velocity and orientation. This observation aligns with our quantitative findings in Table 1.

Qualitative Comparisons with the Baseline Method. We provide qualitative comparison with BEVFormer in Fig. 5. Our method exhibits superior performance in dense scenarios where objects are crowded and challenging to localize. We attribute this enhanced performance to the rich foreground layout context in the scene, facilitated by our attentive contextualization module. Additional qualitative comparisons are available in the supplementary.

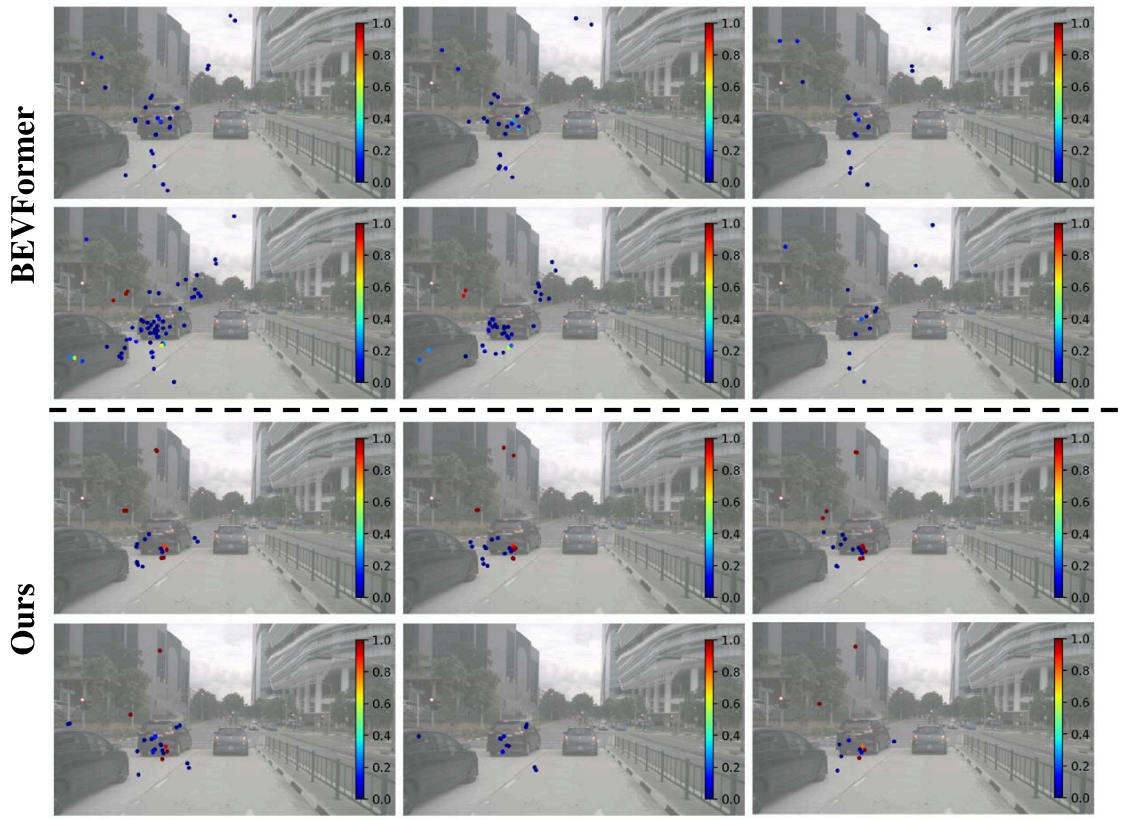


Figure 4. Visualization results of the deformable points originating from a 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder, on NuScenes validation set. We utilize the same BEV anchor point as demonstrated in Fig. 1. From left to right and up to bottom, we display the deformable points output from each layer (#1-#6) in the encoder, respectively.

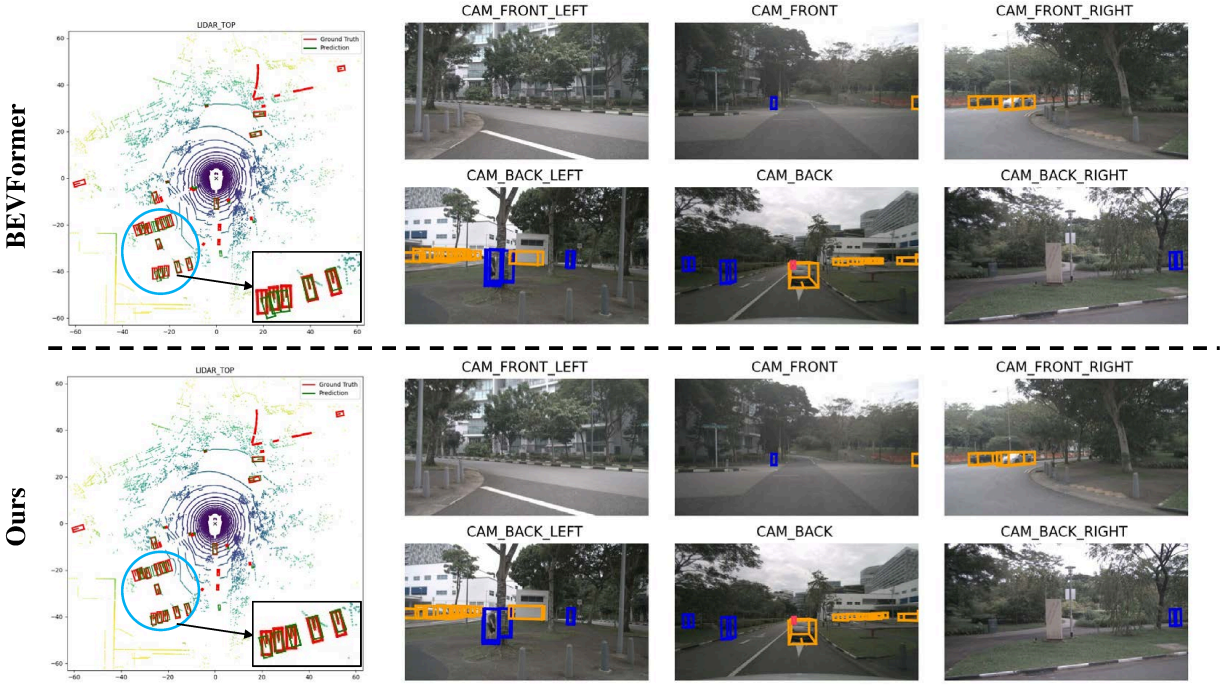


Figure 5. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

4.4. Ablation Studies

Effectiveness of Different Contextualization Methods. Table 3 demonstrates the impact of various contextual methods on detection performance. We selected three representative contexts: local shift window-based context (SWin [32]), global pooling-based context (PVTV2-linear [55]), and global cluster-based context (PaCa [12]). The results show that the local shift window-based context offers minimal improvements in detection performance, which can be attributed to similar local contexts already provided by convolutional backbone networks. Conversely, enhancing the original feature with global contexts, as observed in PVTV2-linear and PaCa experiments, leads to better performance. Notably, orientation and velocity show significant improvements with these global contexts. The inclusion of cluster contexts further enhances improvements in location, orientation, and velocity prediction, as evidenced in the global cluster experiment.

Method	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
BEVFormer-b [25]	41.6	67.3	**27.4**	37.2	39.4	19.8	51.7
Shift Window (SWin [32])	41.5 (-0.1)	67.1	27.8	37.8	39.6	20.1	51.5 (-0.2)
Global Pooling (PVTV2-linear [55])	41.6 (+0.0)	66.5	27.5	36.8	37.6	**19.6**	52.0 (+0.3)
Global Cluster (PaCa [12])	**42.6 (+1.0)**	**66.4**	27.6	**35.0**	**36.2**	20.0	**52.8 (+1.1)**

Table 3. Ablation study on different contextualization methods.

Context	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
Local (BEVFormer-b [25])	41.6	67.3	27.4	37.2	39.4	19.8	51.7
Global Cluster	41.6	66.4	27.4	38.3	35.1	19.5	52.1
Local + Global Cluster	**42.6 (+1.0)**	**66.4**	27.6	**35.0**	**36.2**	20.0	**52.8 (+1.1)**

Table 4. Ablation study on the relationship between local contexts and global cluster-based contexts.

Method	#layers	#clusters	Cross-level	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
BEVFormer-b [25]	-	-	-	41.6	67.3	27.4	37.2	39.4	19.8	51.7
Baseline 1	3	100	✓	41.1	**65.6**	**27.1**	37.7	36.4	18.9	52.0
Baseline 2	6	50	✓	41.3	66.3	27.3	36.1	34.1	**18.8**	52.4
Baseline 3	6	100	-	42.4	66.8	27.6	37.1	36.3	19.1	52.5
MvACon	6	100	✓	**42.6**	66.4	27.6	**35.0**	**36.2**	20.0	**52.8**

Table 5. Ablation study on the structure of our attentive contextualization module.

Clustering Method	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
BEVFormer-b [25]	41.6	67.3	27.4	37.2	39.4	19.8	51.7
Linear	41.9	**65.6**	**27.2**	38.2	**34.8**	19.3	52.4
MLP	42.4	66.4	27.3	38.2	35.0	**19.2**	52.6
Conv	**42.6**	66.4	27.6	**35.0**	36.2	20.0	**52.8**

Table 6. Ablation study on clustering operations in the attentive contextualization module.

Relationship between Local Contexts and Global Cluster Contexts. Table 4 reveals the relationship between local and global cluster contexts in enhancing feature learning for view lifting. The exclusive use of global cluster context results in improved velocity prediction, while local attention contributes to better orientation prediction results. Combining these two contexts enhances predictions in location, orientation, and velocity. This highlights the complementary role that global cluster contexts play in feature encoding for view lifting.

Structure of the Attentive Contextualization Module. Table 5 shows the impact of the structure of our attentive contextualization method. Baseline 1 demonstrates that attentive contextualization can efficiently encode feature information. With only three layers, MvACon achieves an improvement of 0.3 NDS compared to the vanilla BEVFormer. Baseline 2 indicates that attentive contextualization requires a sufficient number of clusters to extract abundant cluster contexts in the scene. Baseline 3 suggests that attending to clusters across the feature map aids in improving orientation, velocity, and mAP prediction. Table 6 illustrates that the convolution operation for clustering yields better results in terms of orientation, mAP, and NDS prediction. Conversely, point-based operations (such as a linear layer or multi-layer perceptron) demonstrate superior performance in location and velocity prediction.

5. Conclusion

This paper presents Multi-View Attentive Contextualization (MvACon) for improving query-based multi-view 3D (MV3D) object detection. It addresses the limitations of two main paradigms of query based MV3D object detector in a unified way: decoder-only dense attention methods like PETR, which lack high-resolution features due to computational constraints, and encoder-decoder sparse 2D/3D deformable attention methods such as BEVFormer and DFA3D. Our MvACon contextualizes the original feature maps extracted from the backbone network using a cluster-attention operation built on the recently proposed patch-to-cluster attention (PaCa). In experiments, we show that our MvACon effectively and consistently enhances query-based MV3D object detectors by encoding more useful contexts, thereby facilitating better 2D-to-3D feature lifting. Rigorously controlled experiments reveal that, for decoder-only detectors, the cluster attention contextualization significantly improves localization and velocity prediction. For encoder-decoder based detectors, it effectively reduces errors in location, orientation, and velocity.

Acknowledgments

X. Liu and T. Wu were partly supported by NSF IIS-1909644, ARO Grant W911NF1810295, ARO Grant W911NF2210010, NSF CMMI-2024688, NSF IUSE-2013451, and a research gift fund from the Innopeak Technology, Inc. (an affiliate of OPPO). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ARO, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, pages 9287–9296, 2019. 2

[3] 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. 2, 5

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3

[5] Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9213–9222, 2023. 2

[6] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2781–2790, 2022. 2

[7] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *CVPR*, pages 12093–12102, 2020. 2

[8] Yi-Nan Chen, Hang Dai, and Yong Ding. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *CVPR*, 2022. 2

[9] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPR Workshops*, pages 1000–1001, 2020. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 2, 5

[12] Ryan Grainger, Thomas Paniagua, Xi Song, Naresh Cuntoor, Mun Wai Lee, and Tianfu Wu. Paca-vit: Learning patch-to-cluster attention in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18568–18578, 2023. 2, 3, 5, 8

[13] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 3

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2

[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. 2

[16] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 2

[17] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8973–8983, 2021. 2

[18] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *ECCV*, 2022. 2

[19] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 2, 5, 6

[20] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *European Conference on Computer Vision*, pages 644–660. Springer, 2020. 2

[21] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 3

[22] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. Densely constrained depth estimator for monocular 3d object detection. *ECCV*, 2022. 2

[23] 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*, pages 1486–1494, 2023. 2

[24] 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*, pages 1477–1485, 2023. 2

[25] 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, 4, 5, 6, 8

[26] 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. 2

[27] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1810–1818, 2022. 3

[28] Xianpeng Liu, Ce Zheng, Kelvin B Cheng, Nan Xue, Guo-Jun Qi, and Tianfu Wu. Monocular 3d object detection with bounding box denoising in 3d by perceiver. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6436–6446, 2023. 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, 3, 4, 5, 6, 1

[30] 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. 2, 5

[31] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 8

[33] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[34] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. *arXiv:2107.13774*, 2021. 2

[35] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, pages 6851–6860, 2019. 2

[36] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *ECCV*, pages 311–327. Springer, 2020. 2

[37] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *CVPR*, pages 4721–4730, 2021. 2

[38] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 1

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[40] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *CVPR*, pages 7074–7082, 2017. 2

[41] 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. 3

[42] Dennis Park, Jie Li, Dian Chen, Vitor Guizilini, and Adrien Gaidon. Depth is all you need for monocular 3d detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7024–7031, 2023. 3

[43] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 2

[44] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 2

[45] 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. 2

[46] Changyong Shu, Fisher Yu, and Yifan Liu. 3d point positional encoding for multi-camera 3d object detection transformers. *arXiv preprint arXiv:2211.14710*, 2022. 2

[47] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, pages 1991–1999, 2019. 2

[48] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021. 3

[49] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[50] 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. 2

[51] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021. 3

[52] 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. 2, 5

[53] 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. 3

[54] Tai Wang, Pang Jiangmiao, and Lin Dahua. Monocular 3d object detection with depth from motion. In *ECCV*, 2022. 2

[55] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3, 8

[56] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pages 8445–8453, 2019. 2

[57] 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 *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2, 3

[58] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[59] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 3

[60] 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, 5

[61] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021. 3

[62] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6575–6586, 2022. 3

[63] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3008, 2021. 3

[64] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, pages 3289–3298, 2021. 2

[65] Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Dalong Du, Jie Zhou, and Jiwen Lu. Dimension embeddings for monocular 3d object detection. In *CVPR*, 2022. 2

[66] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. 3

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

Multi-View Attentive Contextualization for Multi-View 3D Object Detection

Supplementary Material

Overview

In this supplementary material, we provide more details on the following aspects that are not presented in the main paper due to space limit:

- *Computation and memory cost* are provided in Sec. 1.
- *Supplementary qualitative results on NuScenes validation split* are provided in Sec. 2.

1. Computation and Memory Cost

Method	Speed (FPS)	GPU Mem (MB)	#Param (M)	NDS↑
PETR-VovNet-99 [29]	9.8	3638	83.07	42.6
PETR-VovNet-99-MvACon	9.6	3638	84.75	**43.4 (+0.8)**
BEVFormer-b [25]	3.9	6928	69.14	51.7
BEVFormer-b-MvACon-lite	3.2	6936	70.75	**52.5 (+0.8)**
BEVFormer-b-MvACon	3.0	11452	70.75	**52.8 (+1.1)**

Table 7. Efficiency and resource consumption of MvACon on PETR and BEVFormer. MvACon-lite refers to the model without using the concatenation of cluster contexts from all feature pyramids. This will greatly reduce extra GPU memory consumption, with only 0.3 NDS droppped compared with our full model.

Computation and memory cost of our MvACon is provided in Tab. 7. We use the parameter calculation script provided by BEVFormer's open source codebase: https://github.com/fundamentalvision/BEVFormer.

Our MvACon is able to improve PETR with negligible computation cost. We tested two versions of BEVFormer-b-MvACon: a lite version and a full model. In the lite version, we enforce cluster attention within each feature pyramid level instead of using clusters from all levels. This will largely reduce the computation cost. It shows that our lite version is able to improve the baseline with only 8 MB extra GPU memory cost with 0.8 NDS improvement. Using our full model, we are able to improve the baseline with 1.1 NDS improvement. These results clearly demonstrates the effectiveness and necessity of incorporating useful contexts before feature lifting.

2. More Qualitative Results on NuScenes

Qualitative results for deformable points across consecutive frames. We visualize the deformable points across 3 consecutive frames in Fig. 6. We observe that our MvACon is able to learn stable and meaningful high-response deformable points on both cars and surrounding buildings.

Supplementary qualitative results for deformable points in different scenes. We visualize the deformable points in different scenes in Fig. 7 and Fig. 8. We observe that our MvACon is able to learn meaningful high-response deformable points on cars and surrounding references, which could be helpful in improving the prediction of object location, orientation and velocity.

Supplementary qualitative comparison for detection results on NuScenes validation set. We visualize prediction results on NuScenes validation set and compare it with BEVFormer in Fig. 9, Fig. 10 and Fig. 11. We observe that our MvACon performs better in dense scenes.

Supplementary visualization for learned cluster heatmap We provide the detailed visualization results of learned cluster contexts with raw images in Fig. 12. This uses the same scene shown in Fig. 3 of our main paper. The only difference is that we include raw images in supplementary materials. We observe that the learned cluster heatmap has high response on foreground contexts.

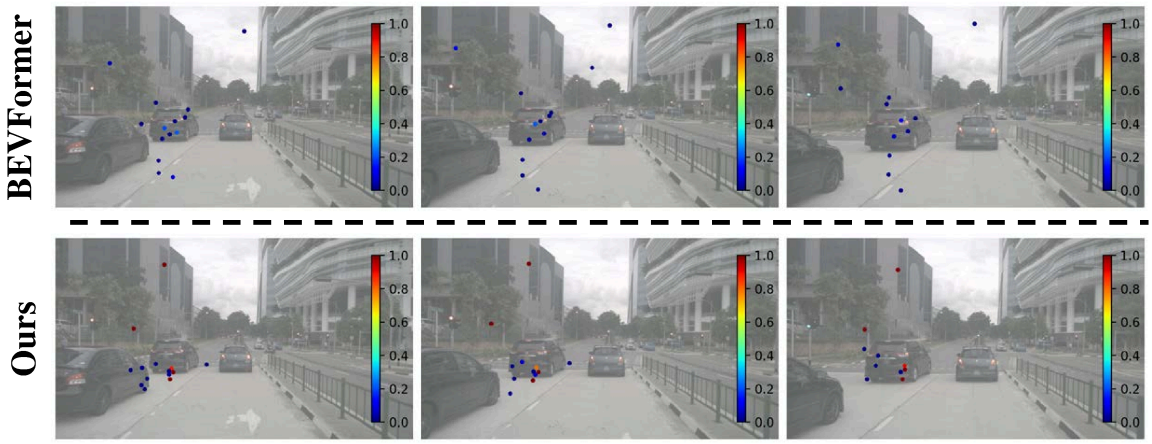


Figure 6. Visualization results of the deformable points originating from a 2D reference point across 3 consecutive frames on NuScenes validation set. This 2D reference point is projected from a 3D BEV (Bird's Eye View) anchor point in the BEVFormer encoder. We use the same BEV anchor point as the one presented in our main paper. From left to right, we exhibit the deformable points outputted from the encoder's final layer, arranged in chronological order (t-1, t, t+1).

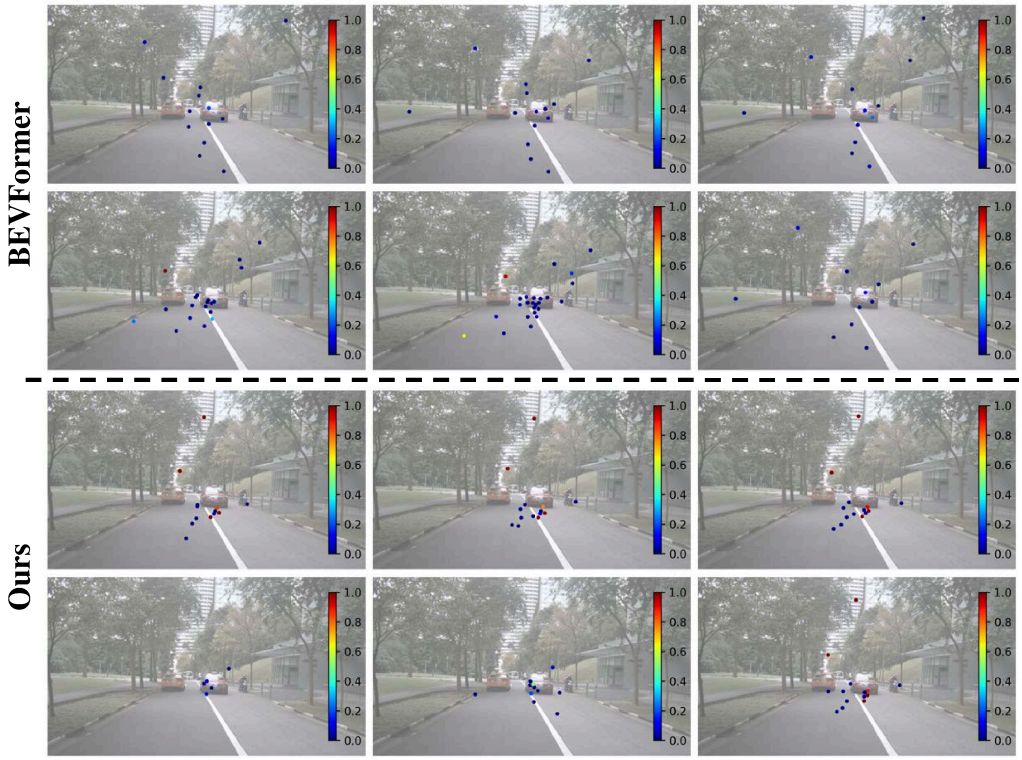


Figure 7. Visualization results of the deformable points originating from a 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder, on NuScenes validation set. We utilize the a BEV anchor point one the right car. From left to right and up to bottom, we display the deformable points output from each layer (#1-#6) in the encoder, respectively.

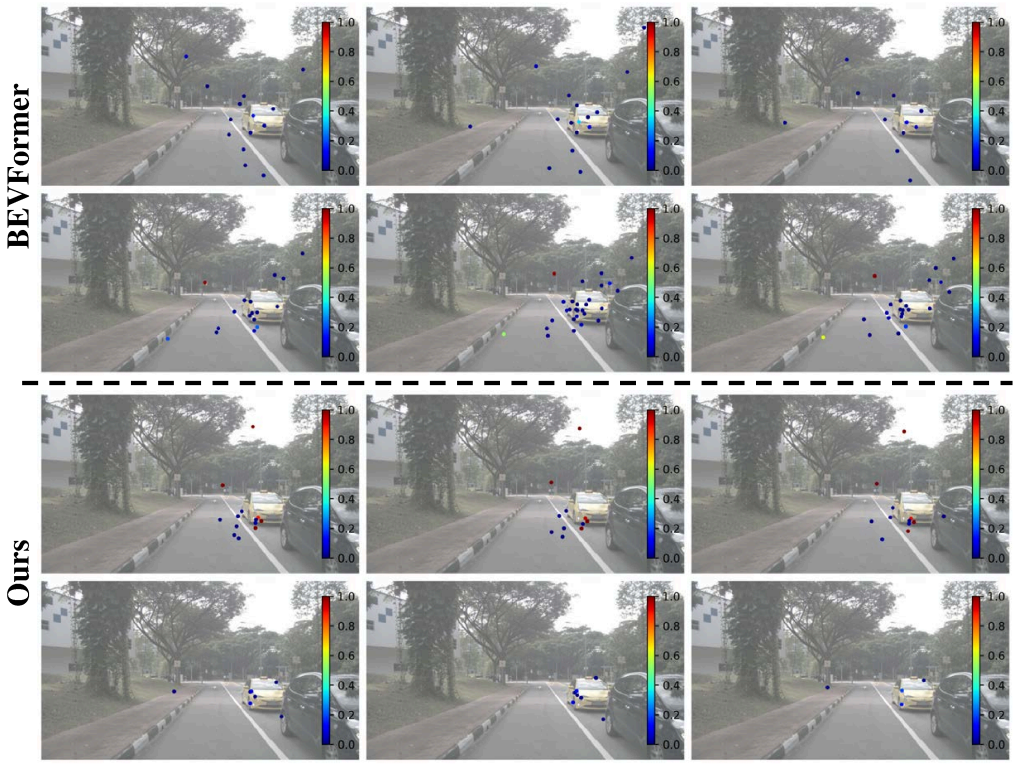


Figure 8. Visualization results of the deformable points originating from a 2D reference point, which is projected from a 3D BEV anchor point in the BEVFormer encoder, on NuScenes validation set. We utilize the a BEV anchor point one the yellow car. From left to right and up to bottom, we display the deformable points output from each layer (#1-#6) in the encoder, respectively.

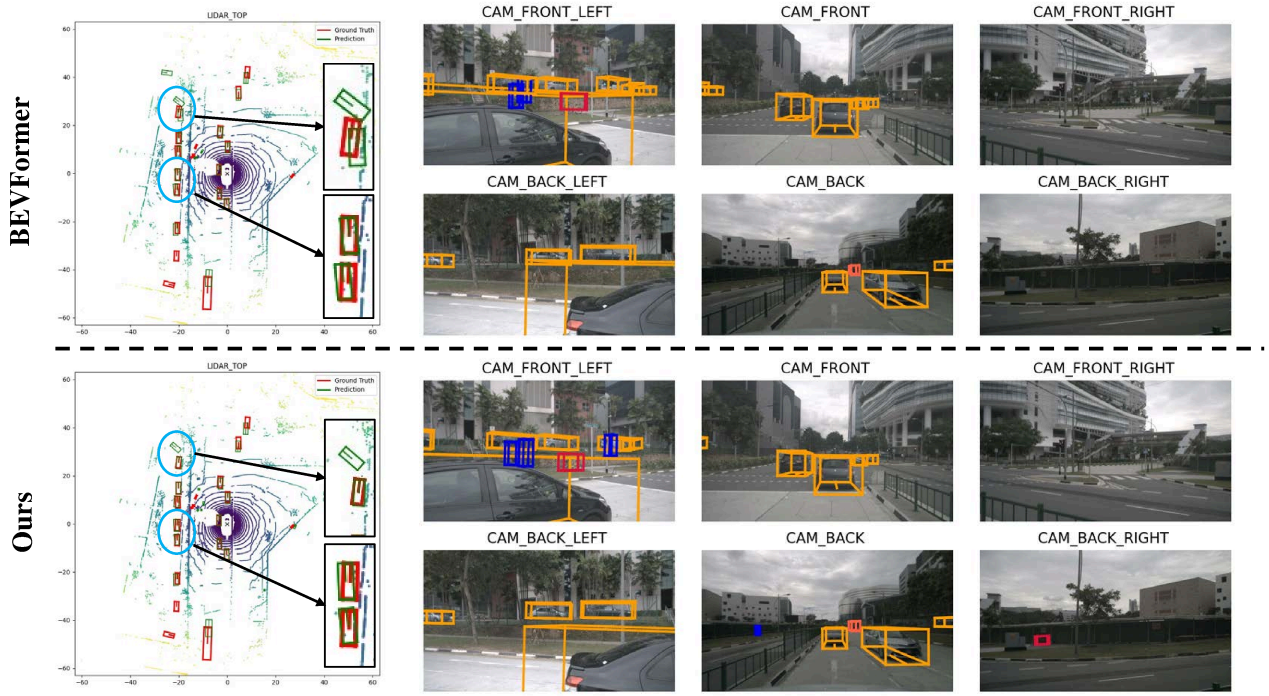


Figure 9. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

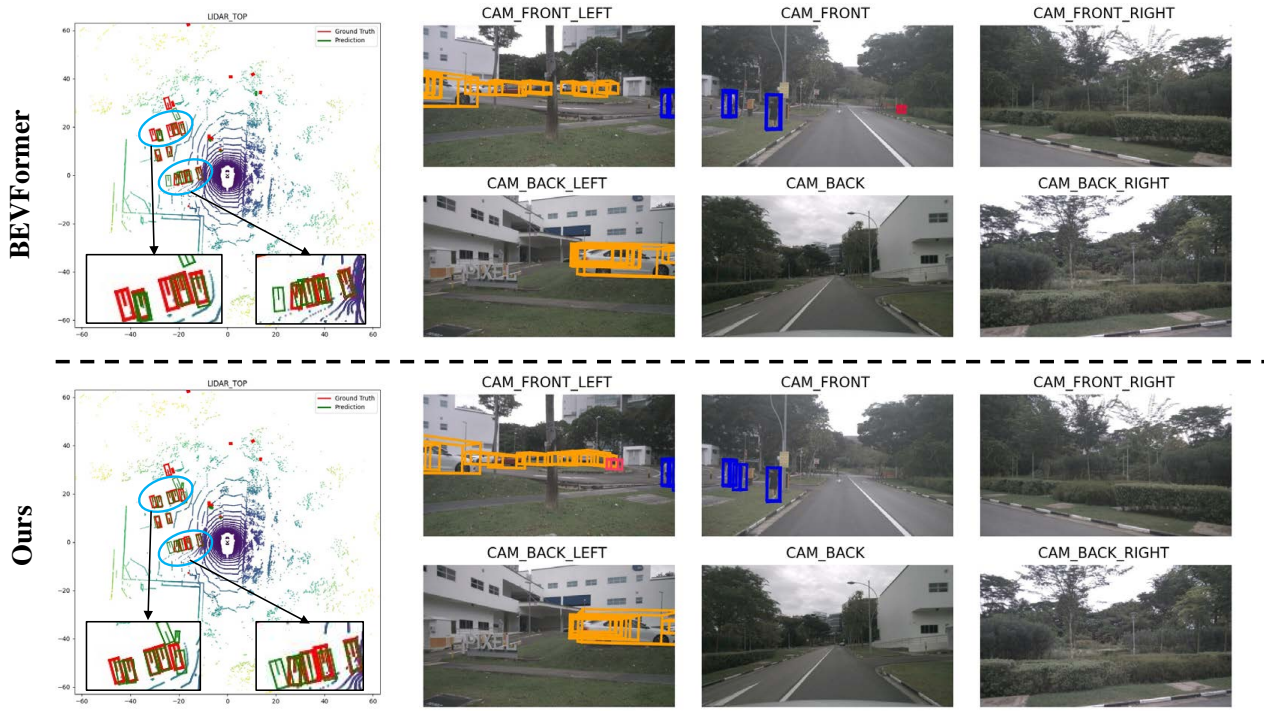


Figure 10. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

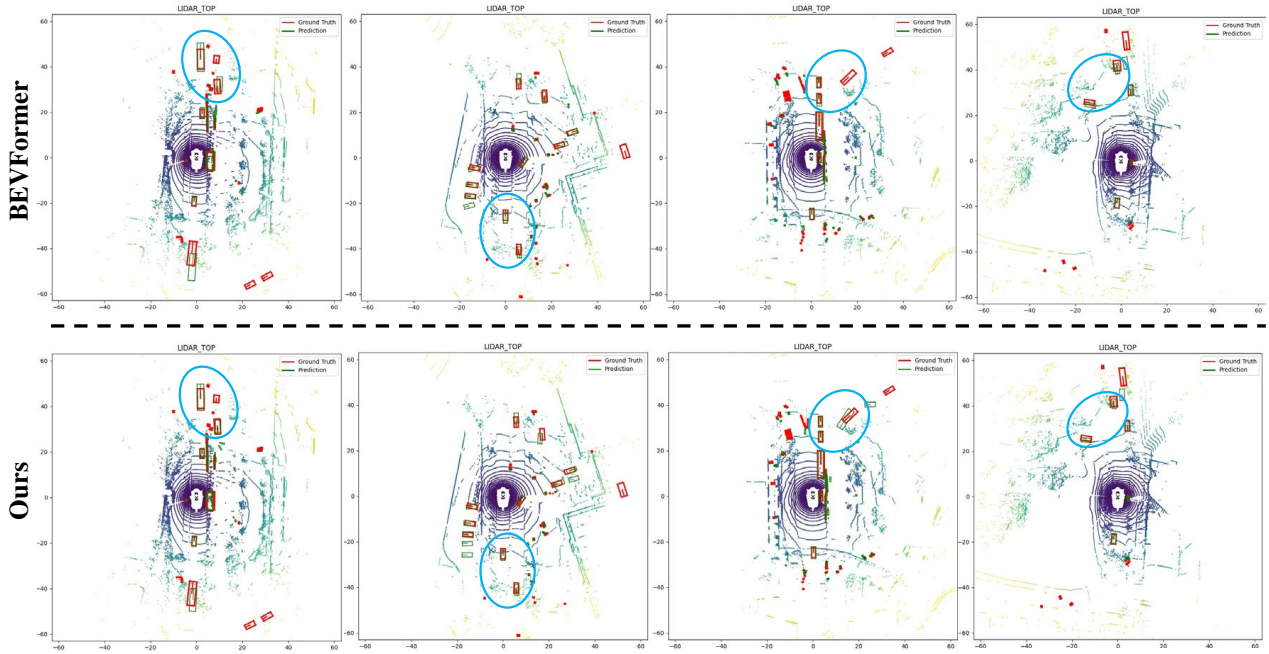


Figure 11. Qualitative comparisons between BEVFormer and our MvACon method on NuScenes validation set.

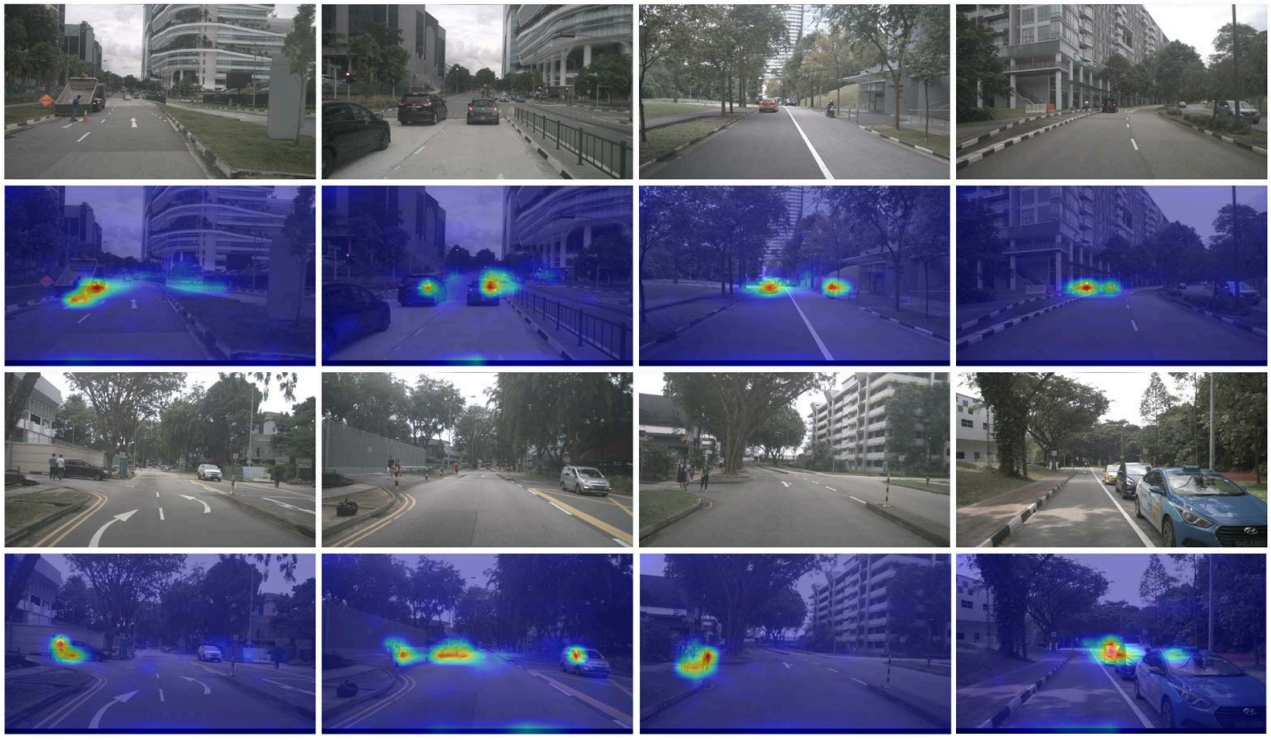


Figure 12. Visualization results of learned cluster contexts with raw images in our proposed attentive contextualization module on NuScenes validation set. We sum all the learned clusters along the channel and upsample it to the original image resolution through bilinear interpolation. We observed that the learned cluster context encodes abundant context information in the scene.