A Streamlined Framework for Bev-Based 3d Object Detection with Prior Masking

[1]yazdani Hasan, [2]PriyaGupta

1. yazhassid@gmail.com,Noida International University **2**. priya.gupta@niu.edu.in,Noida International University

Abstract: Bird's-Eye-View (BEV) representation has become a dominant paradigm for 3D object detection in autonomous driving, offering a unified space for multi-view camera feature fusion. However, existing methods often suffer from computational redundancy by processing all spatial locations in the BEV plane with equal priority, including large, uninformative areas like the sky and distant, irrelevant background. This paper introduces a novel and streamlined framework that leverages Prior Masking to focus computational resources on semantically meaningful regions. Our approach generates a sparse set of candidate BEV queries conditioned on a priori knowledge of likely object locations, derived from classagnostic instance segmentation masks projected into the BEV space. This "mask-guided sampling" strategy eliminates redundant computations in empty areas, leading to a more efficient and focused detection pipeline. Experiments on the nuScenes dataset demonstrate that our method achieves a significant reduction in computational cost (FLOPs) and inference time while maintaining competitive accuracy with state-of-the-art BEV detectors. The framework provides a principled approach to integrating geometric and semantic priors, paving the way for more efficient and deployable autonomous perception systems.

Keywords: 3D Object Detection, Bird's-Eye-View, Autonomous Driving, Prior Masking, Computational Efficiency, Sparse Attention.

1. Introduction

The ability to accurately perceive 3D surroundings from camera images is a cornerstone of autonomous driving systems. While LiDAR-based methods have historically dominated 3D detection, vision-based approaches have gained significant traction due to the lower cost of cameras and their rich semantic information [1]. Among these, Bird's-Eye-View (BEV) perception has emerged as a powerful framework, transforming features from multiple perspective views into a unified top-down representation where 3D detection can be naturally formulated [2, 3].

Core to modern BEV detectors like BEVDet [4] and BEVFormer [5] is the concept of "BEV queries"—a grid of learnable parameters that interact with image features via cross-attention to populate the BEV space with scene information. A fundamental inefficiency in this paradigm is the dense and uniform nature of this query grid. The model expends significant computational effort "reasoning" about every cell in the BEV plane, including vast areas that are either empty (e.g., sky, upper parts of buildings) or contain no objects of interest (e.g., distant vegetation, sidewalks beyond a relevant range). This is computationally wasteful and can introduce noise during training.

In this work, we posit that a significant portion of the BEV space can be a priori identified as noninformative for the task of detecting dynamic agents like vehicles and pedestrians. We propose a Streamlined Framework with Prior Masking that introduces a lightweight, class-agnostic masking stage to guide the allocation of BEV queries. Instead of a dense grid, our model generates a sparse set of queries only at locations likely to contain objects, as determined by projecting 2D instance segmentation masks into the BEV space using estimated depth. This prior mask acts as a spatial filter, focusing the model's representational capacity and computational budget on critical regions.

Our contributions are threefold:

1. We introduce the concept of Prior Mask for BEV-based 3D detection, a simple yet effective mechanism to sparsify the BEV query space.
2. We design a streamlined framework that integrates a fast, class-agnostic mask generator with a standard BEV encoder, enabling efficient mask-guided query sampling.
3. We demonstrate on the nuScenes dataset that our method reduces computational overhead by over 30% while preserving detection accuracy, offering a superior trade-off for real-world deployment.

2. Related Work

2.1 BEV-based 3D Object Detection

The shift from perspective view to BEV representation has been a key advancement. LSS [6] pioneered this by lifting image features to 3D using predicted depth distributions. BEVDet [4] formalized a four-stage pipeline (view transform, BEV encoder, head, post-processing) that many subsequent works follow. BEVFormer [5] introduced a temporal fusion mechanism using deformable attention, setting a new state-of-the-art. These methods, however, rely on a dense BEV grid, which is the core inefficiency our work addresses.

2.2 Sparse and Query-Based Detection

DETR [7] revolutionized 2D detection by replacing non-maximum suppression with a set of object queries. This idea was extended to 3D in works like DETR3D [8], which uses 3D object queries to reference image features. While these methods are inherently sparse in their output, their query initialization is often still random or learned, without leveraging strong spatial priors. Our method can be seen as an intelligent, data-dependent initialization of sparse queries for a BEV-centric paradigm.

2.3 Efficiency in Vision Models

Numerous techniques exist for improving model efficiency, including network pruning, knowledge distillation, and efficient architecture design [9]. Our approach is most closely related to spatial sparsification methods, which skip computations in less important regions of the feature space. Prior Masking provides a task-specific and semantically meaningful way to achieve this sparsification in the BEV domain.

3. Methodology

Our framework, illustrated in Figure 1, consists of three main components: (1) a Prior Mask Generator , (2) a Mask-Guided BEV Query Sampler , and (3) a Standard BEV Detection Head .

3.1 Prior Mask Generation

The goal of this stage is to quickly identify regions in the 2D images that are likely to correspond to objects in 3D space. We prioritize speed and generality over detailed classification.

Input: Surround-view images \(\{I_i\}_{i=1}^{N} \), where \(N \) is the number of cameras (e.g., 6 for nuScenes).

Process: We employ a lightweight, class-agnostic instance segmentation model (e.g., a modified CondInst [10]) to generate a set of binary masks \(\{M_i\}_{i=1}^{N} \). Each mask \(M_i \) labels pixels belonging to any foreground object instance, without distinguishing between cars, pedestrians, etc.

Output: For each mask \(M_i \), we obtain a set of connected components, each representing a candidate object.

3.2 Mask-Guided BEV Query Sampling

This is the core of our streamlining process. We convert the 2D masks into a sparse set of 3D BEV queries.

Lifting 2D Masks to 3D: For each segmented instance in a 2D mask, we sample points within its bounding box. For each sampled pixel \(p=(u, v) \), we use a pre-trained monocular depth estimator (e.g., a lightweight version of [11]) to get a depth distribution $D(p)$. We then unproject the pixel to a 3D point $P_{cam} = (X, Y, Z)$ in the camera coordinate system.

BEV Projection and Query Placement: The 3D point P_{cam} is transformed into the ego vehicle's BEV coordinate system: $P_{bev} = (x, y)$. Instead of placing a single query at the continuous coordinate (x, y), we assign it to the corresponding discrete cell in the BEV grid. This process creates a heatmap of "potential object locations" in the BEV plane.

Sparsification: We apply non-maximum suppression (NMS) in the BEV space to the accumulated points to avoid duplicate queries for the same object viewed by multiple cameras. The final output is a sparse set of K BEV query locations $Q_{sparse} = \{q_1, q_2, ..., q_K\}$, where $K \ll H \times W$ of the original dense grid.

3.3 Sparse BEV Feature Encoding and Detection Head

With the sparsified query set Q_{sparse}, we proceed with the standard BEV detection pipeline but with significantly reduced computation.

Feature Interaction: The sparse BEV queries interact with the multi-view image features via cross-attention, as in [5]. However, since the number of queries K is much smaller, the computational cost of this step is drastically reduced.

Detection Head: The updated, content-aware BEV queries are fed directly into a standard 3D detection head (e.g., a CenterPoint [12] style head) to predict the final bounding boxes, including their center, size, orientation, and velocity.

The loss function is identical to that of the baseline dense BEV detector, ensuring a fair comparison.

4. Experiments

4.1 Experimental Setup

Dataset: We evaluate our method on the large-scale nuScenes dataset [13], using the standard validation split.

Metrics: We report the standard nuScenes Detection Score (NDS) and mean Average Precision (mAP), along with key efficiency metrics: GFLOPs, inference time (ms), and the number of BEV queries.

Baselines: We use BEVDet [4] and BEVFormer [5] as our primary baselines. Our framework is implemented by modifying their dense BEV query mechanism.

4.2 Results and Analysis

Table 1: Performance and Efficiency on nuScenes val set

Method	Backbone	Image Size	BEV Queries	GFLOPs	mAP ↑	NDS ↑	Inference Time (ms) ↓
BEVDet [4]	ResNet-50	256x704	16,384 (128x128)	437	0.298	0.379	320
BEVDet (Ours)	ResNet-50	256x704	~2,500	298	0.292	0.374	220
BEVFormer [5]	ResNet-101	900x1600	40,000 (200x200)	1,175	0.416	0.517	620
BEVFormer (Ours)	ResNet-101	900x1600	~5,000	802	0.408	0.512	430

The results demonstrate the effectiveness of our Prior Masking framework. Our method achieves a reduction of over 30% in GFLOPs and inference time for both baselines, with only a marginal drop in accuracy (≤ 0.008 in mAP and NDS). This confirms that a large portion of computation in dense BEV models is indeed redundant.

4.3 Ablation Study

We ablate the key component of our method:

A. Dense Baseline (BEVFormer): 40k queries, 0.517 NDS, 1175 GFLOPs.

B. + Random Query Sampling: 5k random queries. Result: 0.481 NDS. This shows that naive sparsification severely harms performance.

C. + Prior Masking (Ours): ~5k guided queries. Result: 0.512 NDS. This confirms that our mask-guided sampling is crucial for maintaining performance while being efficient.

5. Conclusion and Future Work

We presented a streamlined framework for BEV-based 3D detection that addresses the computational redundancy of dense BEV grids through Prior Masking. By leveraging simple 2D segmentation and depth priors to focus attention on likely object locations, our method achieves a superior efficiency-accuracy trade-off. This makes it highly suitable for real-time autonomous driving applications where computational resources are constrained.

A limitation of our current approach is its dependence on the quality of the initial mask and depth predictions. Future work will explore end-to-end training of the mask generator within the detection pipeline and investigate the integration of temporal priors to further enhance the stability and accuracy of the query sampling process.

References

[1] Chen, Y., et al. (2022). Focal Sparse Convolutional Networks for 3D Object Detection. IEEE TPAMI .

[2] Philion, J., & Fidler, S. (2020). Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. ECCV .

[3] Wang, Y., et al. (2023). A Survey on Bird's-Eye-View Representation for Autonomous Driving. IEEE TIV .

[4] Huang, J., & Huang, G. (2022). BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View. arXiv preprint arXiv:2112.11790 .

[5] Li, Z., et al. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. ECCV .

[6] Philion, J., & Fidler, S. (2020). Lift, Splat, Shoot. ECCV .

[7] Carion, N., et al. (2020). End-to-End Object Detection with Transformers. ECCV .

[8] Wang, Y., et al. (2021). DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. CoRL .

[9] Han, K., et al. (2020). A Survey on Visual Transformer. arXiv preprint arXiv:2012.12556 .

[10] Tian, Z., et al. (2020). Conditional Convolutions for Instance Segmentation. ECCV .

[11] Godard, C., et al. (2019). Digging Into Self-Supervised Monocular Depth Estimation. ICCV .

[12] Yin, T., et al. (2021). Center-based 3D Object Detection and Tracking. CVPR .

[13] Caesar, H., et al. (2020). nuScenes: A multimodal dataset for autonomous driving. CVPR .