Modified BEVFormer Architecture with Multiscale Cross-Attention

Thorsten Herd*† , Philipp Heidenreich* und Christoph Stiller†

Abstract: In this paper, we enhance the BEVFormer for 3D object detection by using a new multiscale cross-attention approach. We show that our proposed architecture provides an improved performance and requires less computation. We introduce a layer-wise upscaling of the BEV grid features and design them to align with the image features of matching spatial resolution. Moreover, we reduce the number of parameters of the initial BEV grid to prevent overfitting. The proposed enhancements are vital for making automated driving systems more efficient and reliable.

Keywords: Automated driving, Deep learning, Multi-camera 3D object detection, Spatiotemporal transformers

1 Introduction

3D visual perception is crucial for automated driving systems. To tackle multi-camera 3D object detection, especially methods with feature representation in bird's-eye view (BEV) have recently attracted considerable research interest. Here, the BEV feature space can be obtained by camera-to-BEV view transformations [6] or spatiotemporal transformers [10]. In this paper, we focus on the latter BEVFormer architecture, which uses spatial cross-attention to encode image features into a BEV grid.

In practical applications, a fast deployment is important, so we have a closer look at the computationally costly components of the BEVFormer architecture, which are the image backbone and the BEV encoder. Note that the computational cost of the image backbone is mainly influenced by the input image size and the backbone dimension, whereas the computational cost of the BEV encoder is mainly influenced by the BEV grid size, the number of BEV grid layers and the number of used image feature maps.

In this paper, we propose to reduce the architecture overhead of the BEV encoder with the following contributions:

- We propose an architecture modification of the BEV encoder using the multiscale cross-attention principle, as presented in Figure 1. We show that the proposed architecture requires less computations compared to the original BEV encoder while maintaining a similar performance. The modifications include a layer-wise upscaling of the BEV grid features and designing the BEV grid features to attend to image features with the matching spatial resolution. In addition, we present detailed ablation studies to justify the proposed design choices.

*Stellantis, Opel Automobile GmbH, Rüsselsheim am Main. (thorsten.herd@external.stellantis.com)
†Institute of Measurement and Control Systems, Karlsruhe Institute of Technology, Karlsruhe.

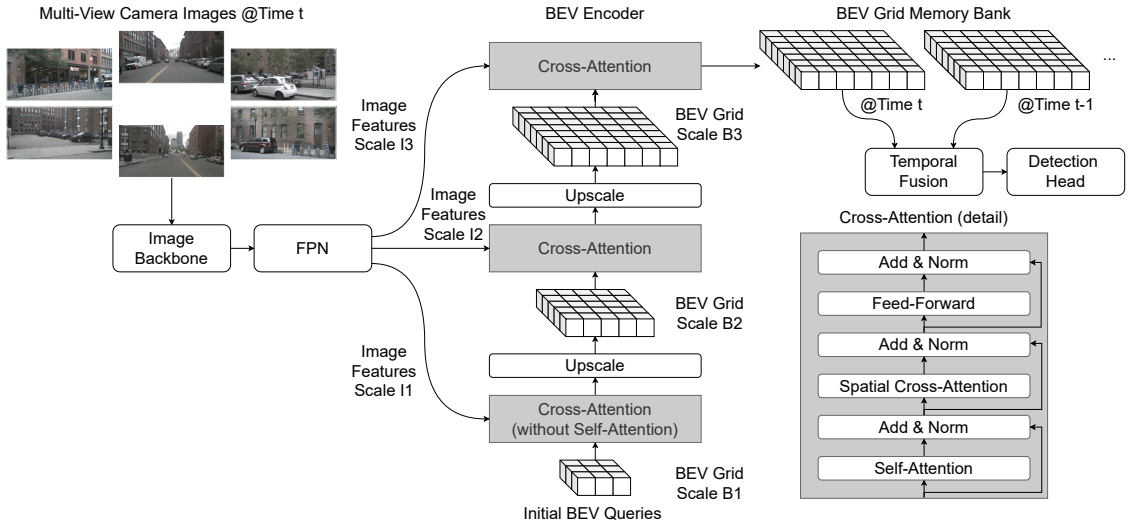

Figure 1: Multi-camera 3D object detection using a modified BEVFormer architecture with multiscale cross-attention.

- We reduce the number of learnable BEV queries to one. This is deemed to avoid that the initial BEV queries learn a position label statistic. We show that this improves the performance for our proposed architecture.

The remainder of the paper is structured as follows. The related work on multi-camera 3D object detection with BEV features is described in Section 2. In Section 3, the overall network and the proposed BEV encoder is presented. Subsequently, the results and the ablation studies are presented in sections 4 and 5, respectively. Lastly, the conclusion is drawn in Section 6.

2 Related Work

Recently, research on camera-based 3D perception with machine learning focused on approaches utilizing a birds-eye-view (BEV) feature representation. These networks usually implement a BEV encoder, which transforms the image features from the perspective view into a top-down BEV representation. Since a correct transformation requires pixel-wise depth information, which is not provided by RGB cameras, different methods have been developed. The methods can be distinguished into two main categories. The first category follows the Lift-Splat-Shoot paradigm (LSS) [11], which explicitly lifts the image features into 3D space by implementing a neural network for pixel-wise depth estimation. Examples for this category are BEVDet [6], which improves the LSS with augmentation strategies and modified non-maxima suppression, and BEVDepth [8], which adds a LiDAR supervision to the depth estimation during training. Similar to BEVDepth, BEVStereo [9] implements stereo depth estimation to improve the view transform.

In the second category, transformer architectures are utilized to encode image features into the BEV grid. Each BEV feature is obtained by sampling image features using the attention mechanism [12]. The developed methods mostly differ by the key query combinations. One example is the BEVFormer [10] architecture. It introduces grid-shaped

learnable queries, which only attend to their corresponding image regions by implementing deformable attention [15]. In subsequent work, BEVFormerV2 [13] further adds a perspective 3D head as an additional task, to encourage the image backbone to learn more 3D relevant features. Another network of this category is PolarFormer [7], which uses a BEV grid in polar coordinates and additional modifications. To tackle unconstrained object scale variation, they further introduce multiscale polar representations, which are computed in parallel and interact through an additional cross-attention block.

Most BEV-based detection networks are computationally intensive, and only a few works have focused on an optimization for a fast deployment. MatrixVT [14] modifies LSS by replacing the computationally intensive view transformation with efficient matrix operations in addition to a compression of image features to reduce memory footprint. Likewise, in [5], a deploy-friendly implementation of BEVDet is provided, which uses an efficient preprocessing of computationally expensive operations.

3 Proposed Architecture

Figure 1 shows the overall architecture, which is based on BEVFormer, extended by the proposed multiscale cross-attention approach. For structural simplicity, only the temporal fusion module is taken from the BEVFormerV2 architecture, so that overall comparability is ensured. Image features are extracted from multi-view images by an image backbone and fed into a feature pyramid network (FPN) to obtain multiscale image features of scales I_1, I_2, and I_3. These are encoded by the BEV encoder into a unified BEV grid using multiscale cross-attention. To further accomplish temporal scene understanding, the BEV grid features of the last N time steps are stored in a memory bank. Together with the current features, they are fused by concatenation and subsequent convolutions [13]. Similar to BEVFormer, a DETR detection head [2] is used to decode the BEV grid features into 3D detections.

The core of this work is the improved BEV encoder, as shown in the center column of Figure 1. Similar to BEVFormer, the image features are encoded to a BEV grid via cross-attention. To this end, learnable BEV queries are initialized and attend to the multiscale image features in three consecutive layers. Each layer consists of a cross-attention block, which contains self-attention, spatial cross-attention, and a feed forward network, respectively, with skip connections between every module. The self-attention and spatial cross-attention models are implemented as deformable attention [10]. To reduce the number of attention operations while still maintaining the final size of the BEV grid, an upsampling of the intermediate BEV grid between each layer is introduced.

In the following the three key components of this work are described.

3.1 Multiscale Cross-Attention

The number of computations is heavily dependent on the number of query-key pairs in the attention blocks. Due to the deployment of the deformable attention method, the computational complexity is reduced to a linear dependency on the number of queries. Despite this reduction, the complexity still scales quadratically with the BEV query grid dimensions. In this work a sequential upscaling of the BEV grid is introduced, to further reduce the computational complexity. The upscaling policy is thereby a simple nearest

neighbor method to avoid additional overhead. To assess the computational complexity, the number of attention operations is considered:

$$N = \sum_{l=1}^{L} q_l \cdot (k_{\text{self}} + f \cdot k_{\text{cross}}) \tag{1}$$

where L is the number of BEV feature layers, q_l is the number of BEV queries in layer l, k_{self} and k_{cross} are the number of keys per query for the self- and cross-attentions, respectively, and f is the number of image feature maps used per BEV grid layer. In our experiments, we use $L = 3$ and $k_{\text{self}} = k_{\text{cross}} = 64$.

In addition to the complexity reduction, the approach also enables to explicitly encode global and local features. The BEV grid cells of lower layers correspond to a larger area in 3D space and thus are able to extract more global information from the image features. As the number of grid cells increases with every layer, these corresponding areas become smaller and hence, more locally detailed image features can be extracted.

3.2 Matching Multiscale Image Features

Using multiple image feature maps of different scales has proven to increase the detection performance of the BEVFormer network [10]. To efficiently incorporate multiscale image features into the multiscale cross-attention approach, they are fed separately into each layer of the BEV encoder. This procedure is illustrated in Figure 1. The order of the scales is chosen from low to high spatial dimension, to further leverage the explicit encoding of global and local features, discussed in Subsection 3.1. With every encoder layer, the resolution of the BEV grid increases and simultaneously queries more specific and detailed image features.

3.3 Query Initialization and Positional Encoding

As the first modification, the positional encodings for the BEV grid are omitted and the number of learnable BEV queries is reduced to one. To still maintain a grid size larger than 1×1, the learnable query is duplicated and arranged as needed. We hypothesize that the reduction leads to a better generalization since it is independent from the position distribution of the ground truth labels. Having different learnable BEV queries at distinct positions could potentially introduce a bias towards labels that occur more frequently in one location. By placing the same query in every location, this bias is prevented and the query learns more global information.

4 Results

We compare our proposed network to the BEVFormer-small and BEVFormer-base configuration. Additionally, we include two smaller networks with a ResNet50 backbone for a potential deployment, also based on our multiscale cross-attention concept. To achieve a fair comparison, the training policy and schedule from the original BEVFormer is adapted [10]. The models are trained on the nuScenes [1] train split without additional data augmentation and evaluated on the val split. The performance is measured through

the nuScenes native metrics mAP and NDS, as well as our indicator in Eq. (1) for the computational complexity of the BEV encoder. The results are shown in Table 1.

Table 1: Comparison of the BEVFormer architecture with the proposed modified BEV encoder. The performance is evaluated on the nuScenes val split. N indicates the computational complexity of the BEV encoder. L is the number of BEV grid layers and f is the number of used image feature maps. The BEVFormer results are taken from [10].

Method	Image size	Image backbone	BEV grid size $B \times B$			L	f	BEV ops $N/10^6$	mAP	NDS
BEVFormer-base	1600×900	ResNet101	200			6	4	76.80	41.6	51.7
BEVFormer-small	1280×720	ResNet101	150			3	1	8.64	37.5	47.9
			B_1	B_2	B_3					
Proposed	1280×720	ResNet101	50	100	200	3	1	6.62	40.1	51.4
Proposed-light 1	800×450	ResNet50	32	64	128	3	1	2.75	31.5	42.6
Proposed-light 2	800×450	ResNet50	16	32	64	3	1	0.69	28.2	39.7

The proposed network outperforms BEVFormer-small with fewer computations in the BEV encoder. This shows the effectiveness of the proposed multiscale cross-attention with matching image features. Although the BEVFormer-base configuration still exceeds the proposed network, we are able to produce competitive results by requiring significantly less computations in the BEV encoder. In addition, the BEVFormer-base configuration has a bigger image input size, more encoder layers, and uses $f = 4$ different scales of image features, whereas our approach only uses 3 in total, $f = 1$ for each layer.

The two proposed light networks use a smaller backbone, smaller final BEV grid sizes, and a smaller image input size. This makes them a good option for a potential deployment, showing the scalability of the proposed method. The mAP results are thereby comparable to other deployment intended camera-only networks such as BEVDet [6].

5 Ablation Studies

For each of the proposed modifications, extensive experiments are conducted and presented in the following subsections. Here, we use a ResNet50 backbone [4] with an image input resolution of 800×450 pixel. We adapt the same training and evaluation schedules and parameters as in Section 4.

5.1 Multiscale Cross-Attention

The first ablation study aims to show the influence of multiscale cross-attention and the matching of multiscale image feature maps individually. For every experiment, we take the nuScenes metrics and our complexity measure 1 as performance metrics. Table 2 displays the conducted experiments and results. We use the two proposed light networks from Table 1 as reference configurations, i.e., with three layers and final BEV grid sizes 128×128 and 64×64, respectively. For each final BEV grid size, we compare the original cross-attention method without upsampling against our proposed upsampling by factor of two. Additionally, we test three different image feature input combinations:

- All: In each encoder layer, the BEV queries attend to all three image feature maps of scales I_1, I_2, and I_3.
- Single: In each encoder layer, the BEV queries attend to a single image feature map with the smallest spatial dimension, I_1.
- Multiscale: In each encoder layer, the BEV queries attend to a different image feature map with increasing spatial resolution, as shown in Figure 1.

Table 2: Ablation studies of the proposed multiscale cross-attention method for different BEV grid final sizes, without upsampling or with an upscaling by factor of two, and using different image feature input combinations. A ResNet50 backbone with an input image resolution of 800×450 pixel is used. The performance is evaluated on the nuScenes val split. N indicates the computational complexity of the BEV encoder.

Ablations	BEV grid final size	BEV grid upscaling	Image feature input combinations	BEV ops $N/10^6$	mAP	NDS
	128×128	✗	All	12.58	30.3	42.5
	128×128	✗	Single	6.29	29.2	40.8
	128×128	✗	Multiscale	6.29	30.9	42.6
	128×128	✓	All	5.51	30.4	41.9
	128×128	✓	Single	2.75	29.2	41.0
Proposed-light 1	128×128	✓	Multiscale	2.75	31.5	42.6
	64×64	✗	All	3.15	28.7	30.4
	64×64	✗	Single	1.57	27.3	39.4
	64×64	✗	Multiscale	1.57	28.8	41.1
	64×64	✓	All	1.38	28.7	40.3
	64×64	✓	Single	0.69	27.5	39.1
Proposed-light 2	64×64	✓	Multiscale	0.69	28.2	39.8

In general, all results with a larger BEV grid size exhibit an overall improved performance. Besides that, the general behavior among the two BEV grid sizes are comparable. For both multiscale and single configuration, the mAP and NDS metrics for our multiscale and original cross-attention method show similar performance, although the number of operations of the BEV encoder decreases over 50% for our method. The ablations show an overall increase of performance when using image feature maps at different scales. Exchanging the original multiscale with our matching multiscale approach does not result in a performance decrease but has the benefit of reducing the computation by a factor of three. Furthermore, an increase in mAP is experienced when combining the multiscale cross-attention with matching image features, when using a final BEV grid size of 128×128. However, this configuration does not lead to an enhancement for a final BEV grid size of 64×64, suggesting that the scales of BEV grids and image feature maps must exhibit a certain ratio to be beneficial.

5.2 Query initialization

To verify our assumption about the impact of the query initialization in Subsection 3.3, we conducted a small ablation study. The corresponding results are shown in Table 3. The network architecture for this ablation is the same as Proposed-light 1 in Table 1. In

total, we compare three initialization policies. The first two have 32×32 unique learnable queries with sinusoidal [12] and learned positional encoding [3], respectively. The third is our initialization approach with only one learned query and zero positional encoding. As shown in the results, our initialization approach outperforms the original ones in both mAP and NDS metrics. This supports our hypothesis, that reducing the number of learned parameters leads to a better generalization across the training dataset.

Table 3: Influence of number of learnable BEV queries and positional encodings during initialization. The general network architecture is Proposed-light 1, as shown in Table 1.

Learnable BEV queries	Positional encodings	mAP	NDS
32×32	Sinusoidal	30.8	42.4
32×32	Learned	30.6	42.0
1×1	Zero	31.5	42.6

6 Conclusion

This paper presents a modified BEVFormer architecture to reduce the computational complexity of the BEV encoder and obtain a better feature alignment. To this end, we introduce a multiscale cross-attention method, which upscales the query grid after each attention layer to produce a final output with high spatial dimensions. Although we show that the computations of the BEV encoder can be largely reduced by our method, the image backbone remains a bottleneck when it comes to fast deployment. Hence, we also considered two light networks with a ResNet50 backbone. We show that a clever matching of multiscale image feature maps to the multiscale BEV grids leads to an increase in performance with less operations in the BEV encoder. An extensive ablation study is conducted, in which we show the effectiveness of our BEV encoder and the influence of each modification.

Acknowledgement

This work is partly funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) and partly financed by the European Union in the frame of NextGenerationEU within the project STADT:up (FKZ 19A22006P).

References

[1] Holger Caesar et al. *nuScenes: A multimodal dataset for autonomous driving*. Tech. rep. arXiv:1903.11027 [cs, stat] type: article. arXiv, May 2020. DOI: 10.48550/arXiv.1903.11027.

[2] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. Tech. rep. arXiv:2005.12872 [cs] version: 3 type: article. arXiv, May 2020. DOI: 10.48550/arXiv.2005.12872.

[3] Jonas Gehring et al. “Convolutional sequence to sequence learning”. In: *International conference on machine learning*. PMLR. 2017, pp. 1243–1252.

[4] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[5] Junjie Huang and Guan Huang. “Bevpoolv2: A cutting-edge implementation of bevdet toward deployment”. In: *arXiv preprint arXiv:2211.17111* (2022).

[6] Junjie Huang et al. *BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View*. Tech. rep. arXiv:2112.11790 [cs] type: article. arXiv, June 2022. DOI: 10.48550/arXiv.2112.11790.

[7] Yanqin Jiang et al. “Polarformer: Multi-camera 3d object detection with polar transformer”. In: *Proceedings of the AAAI conference on Artificial Intelligence*. Vol. 37. 1. 2023, pp. 1042–1050.

[8] Yinhao Li et al. *BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection*. Tech. rep. arXiv:2206.10092 [cs] type: article. arXiv, Nov. 2022.

[9] Yinhao Li et al. “Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2. 2023, pp. 1486–1494.

[10] Zhiqi Li et al. *BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers*. Tech. rep. arXiv:2203.17270 [cs] type: article. arXiv, July 2022.

[11] Jonah Philion and Sanja Fidler. *Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D*. Tech. rep. arXiv:2008.05711 [cs] type: article. arXiv, Aug. 2020. DOI: 10.48550/arXiv.2008.05711.

[12] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. (Visited on 09/15/2023).

[13] Chenyu Yang et al. *BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision*. Tech. rep. arXiv:2211.10439 [cs] version: 1 type: article. arXiv, Nov. 2022.

[14] Hongyu Zhou et al. “Matrixvt: Efficient multi-camera to bev transformation for 3d perception”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 8548–8557.

[15] Xizhou Zhu et al. *Deformable DETR: Deformable Transformers for End-to-End Object Detection*. Tech. rep. arXiv:2010.04159 [cs] type: article. arXiv, Mar. 2021.