

QE-BEV: Query Evolution for Bird's Eye View Object Detection in Varied Contexts – Supplementary Materials –

Anonymous Authors

1 ADDITIONAL RESULTS

1.1 Quantitative Performance on nuScenes [1] (test set)

Table 1: Performance comparison on nuScenes test split.

Method	Backbone	Epochs	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
DETR3D [15]	V2-99	24	47.9	41.2	0.641	0.255	0.394	0.845	0.133
PETR [10]	V2-99	24	50.4	44.1	0.593	0.249	0.383	0.808	0.132
UVTR [5]	V2-99	24	55.1	47.2	0.577	0.253	0.391	0.508	0.123
BEVFormer [7]	V2-99	24	56.9	48.1	0.582	0.256	0.375	0.378	0.126
BEVDet4D [2]	Swin-B [12]	90 [‡]	56.9	45.1	0.511	0.241	0.386	0.301	0.121
PolarFormer [3]		24	57.2	49.3	0.556	0.256	0.364	0.440	0.127
PETrv2 [11]	V2-99	24	59.1	50.8	0.543	0.241	0.360	0.367	0.118
Sparse4D [8]	V2-99	48	59.5	51.1	0.533	0.263	0.369	0.317	0.124
BEVDepth [6]	V2-99	90 [‡]	60.0	50.3	0.445	0.245	0.378	0.320	0.126
BEVStereo [4]	V2-99	90 [‡]	61.0	52.5	0.431	0.246	0.358	0.357	0.138
SOLOFusion [14]	ConvNeXt-B [13]	90 [‡]	61.9	54.0	0.453	0.257	0.376	0.276	0.148
SparseBEV [9]		24	62.7	54.3	0.502	0.244	0.324	0.251	0.126
QE-BEV	V2-99	24	65.7	59.2	0.326	0.261	0.293	0.252	0.183

The performance of QE-BEV compared with other state-of-the-art methods on the test set of nuScenes is in Table 1. We still outperform all the baselines by an obvious gap of +3.0 in NDS and +4.9 in mAP. It also reveals considerable advantages.

1.2 Parameter Sensitivity in Dynamic Adaptation and Temporal Fusion

The optimal values for key parameters are discussed with respect to their impact on model performance. As shown in Figure 1a, the optimal value for β is around 0.6, providing the best blend of initial and dynamically aggregated features. Deviating too much from this value results in suboptimal performance. Similarly, Figure 1b shows that the value of $\alpha = 0.4$ yields the highest NDS and mAP, suggesting that balancing the current and previous dynamic queries effectively captures temporal information.

2 CORRECTION OF TYPOS

Due to an oversight, the data depicted in Figure 6 of the main paper was inadvertently taken from a previous experiment, and not from the most recent results that were intended for this submission. The revised Figure 6 is provided to ensure accurate evaluation of our work, as shown in Figure 2. Additionally, the '60.5' in the original manuscript's statement 'QE-BEV outshines its competitors with an NDS of 60.5' is a typo error; it should be 61.1.

We have also submitted the log files from both experiments as supplementary materials to verify the authenticity of the research data. I deeply regret this mistake and sincerely apologize for any inconvenience it may have caused. We value the opportunity to contribute to the ACM MM conference and appreciate your understanding and support.

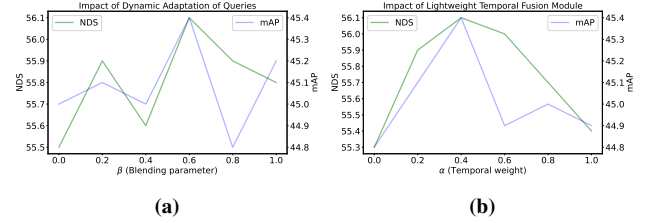


Figure 1: Sensitivity analysis of parameters β and α in the model.

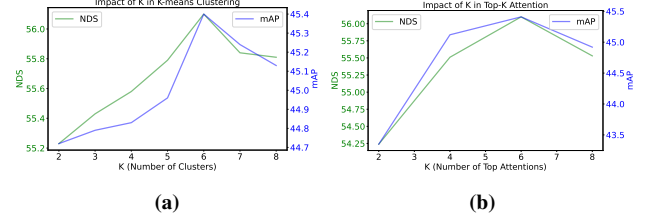


Figure 2: Sensitivity analysis of parameters β and α in the model.

REFERENCES

- [1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.
- [2] Junjie Huang and Guan Huang. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054* (2022).
- [3] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. 2022. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398* (2022).
- [4] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. 2022. BEVStereo: Enhancing Depth Estimation in Multi-view 3D Object Detection with Dynamic Temporal Stereo. *arXiv preprint arXiv:2209.10248* (2022).
- [5] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. 2022. Unifying Voxel-based Representation with Transformer for 3D Object Detection. In *Advances in Neural Information Processing Systems*.
- [6] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. 2022. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092* (2022).
- [7] Zhiqi Li, Wenhao Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. 2022. BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In *European Conference on Computer Vision (ECCV)*. Springer, 1–18.
- [8] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. 2022. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581* (2022).
- [9] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. 2023. SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos. *arXiv preprint arXiv:2308.09244* (2023).
- [10] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. 2022. PETR: Position Embedding Transformation for Multi-view 3D Object Detection. In *European Conference on Computer Vision (ECCV)*. Springer, 531–548.
- [11] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. 2022. PETrv2: A Unified Framework for 3D Perception from Multi-Camera Images. *arXiv preprint arXiv:2206.01256* (2022).

- [12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
- [13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

- [14] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443* (2022).
- [15] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*. PMLR, 180–191.

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232