Predict to Detect: Prediction-guided 3D Object Detection using Sequential Images
- Supplementary Materials -

Sanmin Kim Youngseok Kim In-Jae Lee Dongsuk Kum
KAIST
{sanmin.kim, youngseok.kim, oliver0922, dskum}@kaist.ac.kr

In this supplementary material, we provide additional information for the proposed method and supplementary experiments. Specifically, we present implementation details, additional ablation studies, experiments on KITTI dataset, results on 3D object tracking, and additional qualitative results.

A. Implementation Details

A.1. Baselines

We use BEVDepth [10] and BEVstereo [9] as two baselines of our work. We employ the official codes.[1]

BEVDepth To enhance depth estimation from images, BEVDepth introduces explicit depth supervision using LiDAR point clouds. A camera-awareness depth estimation module and a depth refinement module are proposed to facilitate depth prediction. Furthermore, BEVDepth leverages multi-frame by concatenating all temporal features after warping.

BEVStereo BEVStereo extends BEVDepth pipeline to leverage multiple frames effectively. BEVStereo adopts a dynamic temporal stereo which can save memory cost by reducing cost volume. Furthermore, a parameter evolution algorithm is proposed for noisy features. BEVStereo inspired by MaGNet [1].

A.2. Training settings

Table A1 presents the training recipes and hyperparameters used in P2D.

A.3. Prediction query-based cross attention

We provide detailed information on the prediction query-based cross attention (PQCA). Given the object queries $Q \in \mathbb{R}^{K \times C_q}$ that are the output of the prediction head in P2D, PQCA aggregates temporal BEV features $F_{BEV}^{1:T}$ using a deformable attention [16].

[1] https://github.com/Megvii-BaseDetection/BEVDepth

backbone	ResNet50	ResNet101
image size	256×704	512×1408
batch size	16	
epoch	24	
optimizer	AdamW	
base lr	2e-4	
backbone lr	2e-4	
lr scheduler	0.1 at [19, 23]	
weight decay	0.01	

Table A1. Training settings of P2D with different backbones.

$$PQCA(Q_q, p_q, \{F_{BEV}^{1:T}\}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{t=1}^{T} \sum_{n=1}^{N} A_{mtqn} \cdot \boldsymbol{W}'_m F_{BEV}^{t}(p_q + \Delta p_{mtqn}) \right], \tag{A1}$$

where Q_q denotes the prediction-guided object query at $p_q = (i, j)$. m, t and n index the attention head, timestep, and sampling point, respectively. $\boldsymbol{W}_m \in \mathbb{R}^{C \times C_v}$ and $\boldsymbol{W}'_m \in \mathbb{R}^{C_v \times C}$ are learnable weights, A_{mtqn} denotes the attention weight of n^{th} sampling point in the feature F_{BEV}^{t} and the m^{th} attention head. A_{mtqn} is normalized as $\sum_{t=1}^{T} \sum_{n=1}^{N} A_{mtpn} = 1$.

We set $M = 8$, $T = 3$ with 1-second interval, and $N = 9$. We stack 6 layers of PQCA to generate a spatio-temporal feature.

B. Additional Quantitative Results

B.1. Per-class results

Table B2 presents the per-class evaluation results obtained by comparing P2D with each baseline under the same experimental settings for a fair comparison. We utilized a

Methods	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor.	Bicyle	T.C.	Barrier	mAP
BEVDepth	50.4	27.4	**38.2**	14.6	8.2	29.2	34.8	**34.2**	46.7	50.3	33.4
P2D (BEVDepth)	**52.3**	**30.6**	36.8	**18.3**	**8.9**	**32.5**	**36.9**	34.0	**52.5**	**56.6**	**36.0**
BEVStereo	50.9	28.7	38.8	**16.7**	8.7	35.1	36.8	33.8	49.2	50.4	34.9
P2D (Stereo)	**54.1**	**31.8**	**39.7**	16.5	**10.6**	**38.0**	**40.3**	**36.8**	**52.2**	**54.0**	**37.4**

Table B2. Per-class AP results on nuScenes *val* set. 'C.V', 'Ped.', 'Motor.', 'T.C.' stands for construction vehicle, pedestrian, motorcycle, and traffic cone, respectively.

Methods	mATE	mASE	mAOE	mAVE	mAAE
BEVDepth	0.662	0.282	0.663	**0.135**	0.260
+ P2D	**0.649**	**0.276**	**0.574**	0.147	**0.251**
BEVStereo	0.647	0.284	0.601	0.131	0.259
+ P2D	**0.644**	**0.279**	**0.595**	**0.123**	**0.248**

Table B3. Results on the static objects. Only objects with a velocity lower than 1m/s are evaluated.

K	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE	FPS
1024	0.359	0.474	**0.642**	0.271	0.515	0.413	0.218	**10.8**
2048	**0.360**	**0.474**	0.643	**0.271**	**0.512**	**0.412**	**0.217**	**10.8**
4096	0.359	0.473	0.642	0.272	0.515	0.415	0.218	10.7
8192	0.358	0.472	0.640	0.273	0.517	0.423	0.219	10.4

Table B4. Ablation on the number of queries in the prediction-guided cross attention.

ResNet50 backbone and an image size of 256×704. The results reveal that P2D consistently outperforms most classes.

B.2. Static objects

Table B3 demonstrates the evaluation results for static objects with a velocity lower than 1 m/s. While P2D with the BEVDepth baseline shows a marginal increase in mAVE, the small absolute value (0.012 m/s) renders it insignificant. In the case of BEVStereo, there is a marginal improvement in all true positive errors. These outcomes confirm that P2D's primary performance enhancement is in estimating moving objects, which hold greater significance in the driving environment.

B.3. Number of queries

In the prediction-guided object query-based cross attention, we select object queries to effectively aggregate temporal BEV features. To investigate the impact of the number of queries on the model's performance, we conduct an ablation study and report the results in Table B4. The results indicate that both detection performance and FPS decrease as the number of queries increases. Although the performance difference is marginal, We believe that choosing an appropriate number of object queries can help the model to focus on the foreground area and prevent the generation of false positives, thereby improving the overall detection performance.

Methods	AMOTA ↑	AMOTP ↓	MOTA ↑	IDS ↓	Frag ↓	MT ↑
DEFT [2]	0.177	1.564	0.156	6901	3420	1951
Time3D [8]	0.214	1.360	0.173	N/A	N/A	N/A
QD3DT [7]	0.217	1.550	0.198	6856	3001	1893
TripletTrack [11]	0.268	1.504	0.245	**1044**	3978	2085
MUTR3D [14]	0.270	1.494	0.245	6018	2749	2221
PolarDETR [3]	0.273	1.185	0.238	2170	1924	2266
P2D	**0.377**	**1.122**	**0.337**	1212	**858**	**2713**

Table B5. 3D object tracking results on nuScenes *test* set. We attached a greedy assignment algorithm on P2D for the tracking task.

B.4. Tracking

The ability of P2D to accurately estimate the location and velocity of moving objects makes P2D promising to apply for the object tracking task. To evaluate the performance of P2D for 3D object tracking, we conduct experiments on the nuScenes tracking dataset. We trained P2D for tracking using a ResNet101 backbone and an image size of 512×1408, and adopted a simple greedy assignment algorithm similar to CenterTrack [15]. The results are reported in Table B5. The results demonstrate that P2D outperforms several other state-of-the-art models. Notably, we observed significant improvements in both Frag (number of track fragmentations) and MT (number of mostly tracked trajectories), indicating that the prediction strategy employed in P2D is also effective in object tracking.

C. Experiments on KITTI Dataset

To prove the effectiveness of the prediction scheme even in a different dataset and model architecture, we conduct experiments on KITTI dataset [4] with DD3D [12] baseline.

C.1. Experiments settings

DD3D [12] is a monocular 3D object detection model that benefits from depth pre-training with another dataset

Methods	TTA	Pedestrian						Cyclist					
		AP_{3D}			AP_{BEV}			AP_{3D}			AP_{BEV}		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
DD3D		10.90	8.22	6.53	13.73	10.38	8.37	3.92	2.20	2.08	4.73	2.66	2.40
P2D		**12.86**	**9.21**	**7.63**	**15.07**	**11.22**	**8.76**	**6.56**	**3.52**	**3.40**	**7.36**	**3.87**	**3.77**
DD3D	✓	11.19	8.80	7.04	13.17	10.45	8.46	4.40	2.56	2.48	5.27	2.95	2.82
P2D	✓	**13.60**	**10.08**	**8.17**	**16.18**	**12.01**	**9.87**	**7.67**	**3.90**	**3.82**	**10.62**	**5.26**	**5.42**

Table B6. KITTI-3D *val* set evaluation on *Pedestrian* and *Cyclist*. TTA denotes Test Time Augmentation.

Methods	TTA	Car					
		AP_{3D}			AP_{BEV}		
		Easy	Moderate	Hard	Easy	Moderate	Hard
DD3D		20.06	16.08	14.04	27.93	22.53	19.93
P2D		**22.60**	**16.74**	**14.38**	**31.59**	**24.39**	**20.95**
DD3D	✓	22.32	16.92	15.11	32.20	24.77	22.08
P2D	✓	**28.08**	**19.69**	**17.09**	**39.04**	**28.34**	**24.57**

Table B7. KITTI-3D *val* set evaluation on *Car*. TTA denotes Test Time Augmentation.

[6]. The architecture of DD3D is based on FCOS3D and achieves state-of-the-art performance on KITTI 3D object detection dataset [5]. We employ DD3D as our baseline and add a prediction scheme to verify the effectiveness of the prediction strategy on KITTI dataset.

C.2. Results on KITTI dataset

Different from P2D with BEVDepth, we employ the prediction as an auxiliary task, and the results are reported in Table B6 and Table B7. We extend our model from the official code[2] with the DLA34 [13] backbone, keeping all other training settings remain the same as DD3D. The results in Table B6 and Table B7 demonstrate that the prediction scheme significantly improves detection performance in all three classes. suggesting that the effectiveness of our prediction scheme is not limited to a specific model or dataset but is generally applicable.

D. Additional Qualitative Results

We visualize sample cases in Figure D1 and Figure D2 to compare the performance of P2D against the baseline method. These figures depict the sequence of detection results obtained from both methods with ground truth bounding boxes, covering 5 frames with 0.5 second intervals.

In the first scenario of Figure D1, there are moving objects ahead and behind the ego vehicle as indicated by the dotted blue boxes. During this sequence, the baseline method detects the front vehicle in only two frames ($t-3$ and $t-2$), and it fails to detect the vehicle in the rear. In contrast, P2D detects both vehicles in front and rear throughout all timesteps. This case demonstrates the ability of P2D to effectively detect moving objects.

In the second scenario depicted in Figure D2, vehicles on the right side are visible at the first timestep ($t-4$), but they become occluded by a truck as the ego vehicle moves forward. The baseline method starts to lose these occluded objects from $t-2$ and fails to detect all four occluded vehicles at time t. On the other hand, P2D successfully keeps track of the occluded objects and does not miss any of them throughout all timesteps.

Detecting occluded objects using motion features is critical as pedestrians can suddenly appear from such blind areas. Moreover, since these vehicles are not annotated as ground truth due to the absence of Lidar points, these detection results are counted as false positives. Thus, we argue that motion cues can enhance 3D object detection beyond just improving performance metrics. The presented qualitative analysis highlights the importance of incorporating motion cues in 3D object detection and demonstrates the superior ability of P2D in detecting occluded objects.

[2] https://https://github.com/TRI-ML/dd3d

Figure D1. Visualization of a sequence containing moving objects. The blue dotted rectangles in the BEV view indicate the vehicles moving in the same lane as the ego vehicle. Despite an error in localization, P2D (right) successfully detects these moving objects, while the baseline (left) fails to fully detect them.

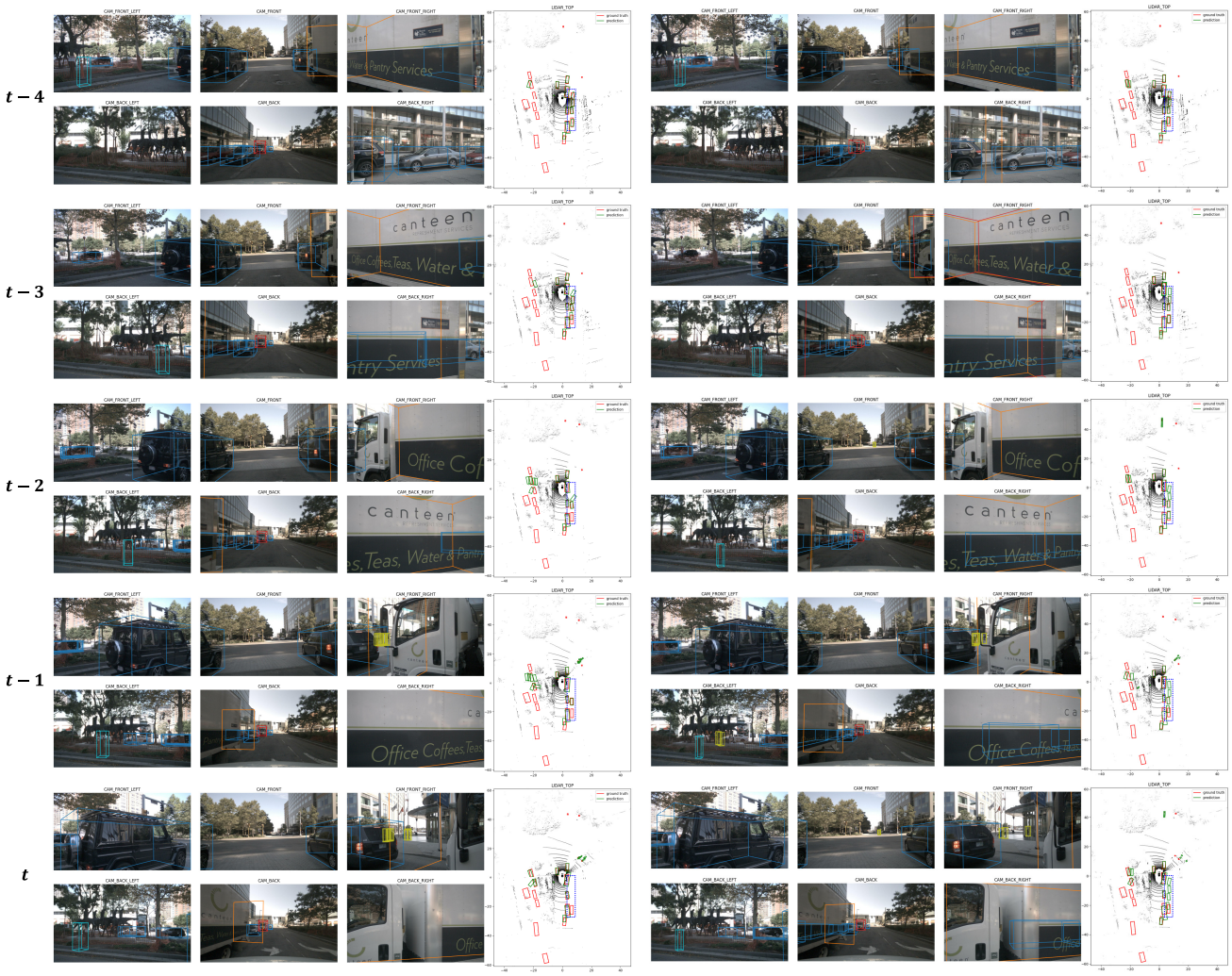

Figure D2. Visualization of a sequence with occluded objects. The blue dotted rectangles in the BEV view indicate the objects which are highly occluded in the image view. P2D (right) leverages temporal information to detect such an occluded object that appears in previous frames, while the baseline (left) fails to detect it due to occlusion.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[2] Mohamed Chaabane, Peter Zhang, Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021. 2

[3] Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Chang Huang, and Wenyu Liu. Polar parametrization for vision-based surround-view 3d detection. *arXiv preprint arXiv:2206.10965*, 2022. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE conference on computer vision and pattern recognition*, 2012. 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE conference on computer vision and pattern recognition*, 2012. 3

[6] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 3

[7] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1992–2008, 2022. 2

[8] Peixuan Li and Jieyu Jin. Time3d: End-to-end joint monocu-

lar 3d object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[9] 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*, 2023. 1

[10] 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*, 2023. 1

[11] Nicola Marinello, Marc Proesmans, and Luc Van Gool. Triplettrack: 3d object tracking using triplet embeddings and lstm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[12] 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*, 2021. 2

[13] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[14] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[15] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, 2020. 2

[16] 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. 1