A Appendix

A.1 Performance on the Camera-based method

Although we design our motion-guided temporal modeling (MTM) module based on the LiDAR domain, we also explore the performance of MTM on camera-based methods. Thus, we integrate the MTM into the advanced camera-based detector CAPE [8] with two frames as input for temporal fusion on the nuScenes [2] validation set. As shown in Table 1, our MTM can also boost the performance of the camera-based method, which effectively demonstrates the generality of our method.

Table 1: Performance of camera-based method with MTM. The C represents camera. * denotes our reproduced results. All models are trained by four NVIDIA RTX 4090 GPUs with 24 epochs and without CBGS [13]. The batch size is set to 4.

Method	Year	Modality	Frames	Resolution	Backbone	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
CAPE* [8]	CVPR 2023	C	1	704×256	R50	27.5	35.9	0.794	0.286	0.642	0.847	0.215
+MTM	-	C	2	704×256	R50	**31.6**	**43.8**	**0.752**	**0.277**	**0.558**	**0.438**	**0.182**
CAPE* [8]	CVPR 2023	C	1	800×320	V2-99	39.7	46.3	0.693	0.270	0.438	0.747	0.206
+MTM	-	C	2	800×320	V2-99	**43.9**	**53.6**	**0.656**	**0.266**	**0.380**	**0.350**	**0.183**

A.2 Performance breakdown for each category

We report the detailed performance of QTNet for each category on the nuScenes [2] testing benchmark, as shown in Table 2. Compared with our LiDAR-only baseline TransFusion-L [1], QTNet brings consistent improvements on most categories, especially on the construction vehicle (+7.0% AP), motorcycle (+6.6% AP), and bicycle (+6.3% AP).

Table 2: Comparison with state-of-the-art methods on the nuScenes testing set for each category. The L and C represent LiDAR and camera, respectively. C.V., Ped., M.C., B.C., T.C., and B.R. represent construction vehicle, pedestrian, motorcycle, bicycle, traffic cone, and barrier, respectively. The column of Frames denotes the number of key frame. † denotes future information is used.

Method	Modality	Frames	mAP	NDS	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	B.C.	T.C.	B.R.
CenterPoint [12]	L	1	60.3	67.3	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7	78.4	71.1
TransFusion-L [1]	L	1	65.5	70.2	86.2	56.7	66.3	58.8	28.2	86.1	68.3	44.2	82.0	**78.2**
VISTA [5]	L	1	63.7	70.4	84.7	54.2	64.0	55.0	29.1	83.6	71.0	45.2	78.6	71.8
LidarMultiNet [10]	L	1	67.0	71.6	86.9	57.4	64.7	61.0	31.5	87.2	**75.3**	47.6	**85.1**	73.5
VoxelNeXt [4]	L	1	64.5	70.0	84.6	53.0	64.7	55.8	28.7	85.8	73.2	45.7	79.0	74.6
LargeKernel3D [3]	L	1	65.3	70.5	85.9	55.3	66.2	60.2	26.8	85.6	72.5	46.6	80.0	74.3
LinK [7]	L	1	66.3	71.0	86.1	55.7	65.7	62.1	30.9	85.8	73.5	47.5	80.4	75.1
3DVID† [11]	L	3	65.4	71.4	87.5	56.9	63.5	60.2	32.1	82.1	74.6	45.9	78.8	69.3
MGTANet† [6]	L	3	65.4	71.2	**87.7**	56.9	64.6	59.0	28.5	86.4	72.7	47.9	83.8	65.9
QTNet	L	3	68.2	72.0	86.5	57.2	**68.3**	**63.0**	34.3	88.1	74.9	49.7	82.7	77.0
QTNet	L	4	**68.4**	**72.2**	86.6	**57.7**	**68.3**	62.9	**35.2**	**88.2**	74.9	**50.5**	82.8	77.3

Besides, we report the detailed performance of QTNet for each category on the nuScenes [2] validation benchmark, as shown in Table 3.

Table 3: Comparison with different baselines on the nuScenes validation set for each category. * denotes our reproduced results.

Method	Modality	Frames	mAP	NDS	Car	Truck	Bus	Trailer	C.V.	Ped.	M.C.	B.C.	T.C.	B.R.
TransFusion* [1]	LC	1	67.1	70.7	87.7	61.6	75.9	42.4	26.5	88.0	75.2	63.8	77.3	72.2
+QTNet	LC	4	68.5	71.6	87.8	63.0	76.6	43.1	27.7	89.3	77.5	68.8	78.3	72.5
DeepInteraction* [9]	LC	1	69.9	72.6	**88.5**	64.4	**79.2**	44.5	**30.1**	88.9	79.0	67.8	**80.0**	**76.4**
+QTNet	LC	4	**70.3**	**73.1**	88.4	**64.7**	79.0	**44.8**	29.4	**89.4**	**80.5**	**70.6**	79.7	76.1
BEVFusion* [**?**]	LC	1	69.9	72.6	**88.5**	64.4	**79.2**	44.5	**30.1**	88.9	79.0	67.8	**80.0**	**76.4**
+QTNet	LC	4	**70.3**	**73.1**	88.4	**64.7**	79.0	**44.8**	29.4	**89.4**	**80.5**	**70.6**	79.7	76.1
TransFusion-L* [1]	L	1	65.0	70.0	86.7	60.4	75.3	41.6	24.6	86.8	71.8	56.5	74.4	**71.8**
+QTNet	L	3	66.3	70.8	87.1	61.1	75.5	**43.0**	**25.7**	**87.8**	75.2	61.2	**75.6**	71.4
+QTNet	L	4	**66.5**	**70.9**	**87.2**	**61.5**	**75.8**	**43.0**	**25.7**	**87.8**	**75.5**	**61.5**	75.4	71.4

A.3 Visualization

To illustrate the superiority of our QTNet, we visualize the results of TransFusion-L [1] on the nuScenes [2] validation set for comparison. As shown in Figure 1, QTNet can detect the hard-detected objects for TransFusion-L and boost the detection performance thanks to our proposed temporal fusion module MTM. As shown in Figure 2, QTNet successfully correct the angle error of objects for TransFusion-L thanks to our proposed temporal fusion module MTM. Besides, as shown in Figure 3, we compare TransFusion-L and QTNet along the temporal dimension for better presentation. It can be seen that the object on the lower left, which is moving away from the ego vehicle, is not detected in t frame by TransFusion-L. However, QTNet can still capture the object in t frame, benefiting from our effective temporal fusion.

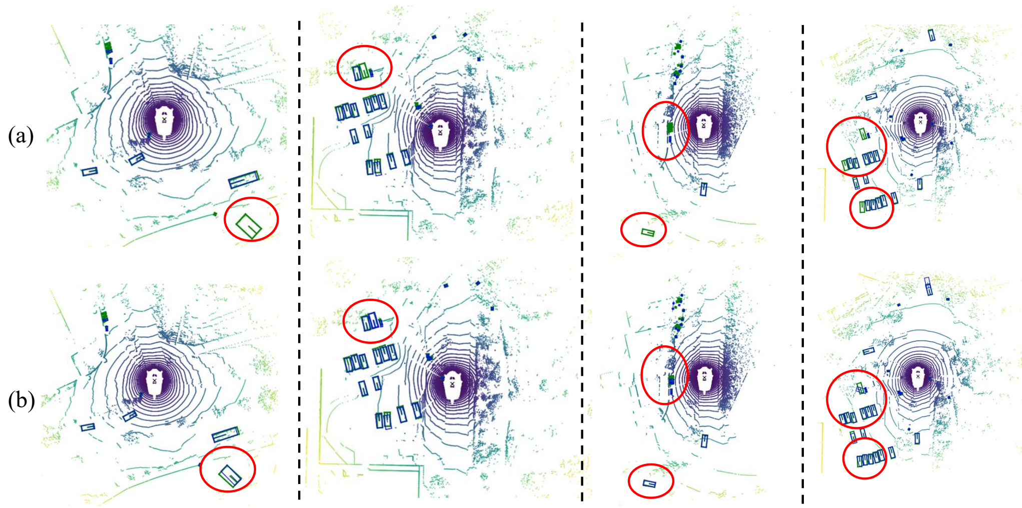

Figure 1: Comparison of LiDAR-only baseline TransFusion-L (a) and QTNet (b) on the nuScenes validation set. Blue and green boxes are the prediction and ground truth boxes. It can be seen that TransFusion-L fails to detect the hard-detected objects. However, thanks to the temporal information, QTNet detects these objects successfully.

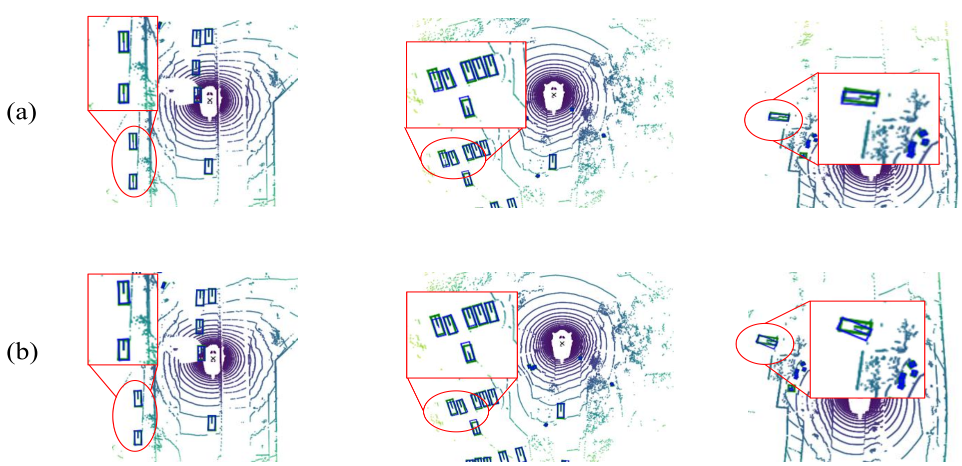

Figure 2: Comparison of LiDAR-only baseline TransFusion-L (a) and QTNet (b) about orientation of objects. Thanks to the temporal information, QTNet successfully corrected the orientation error.

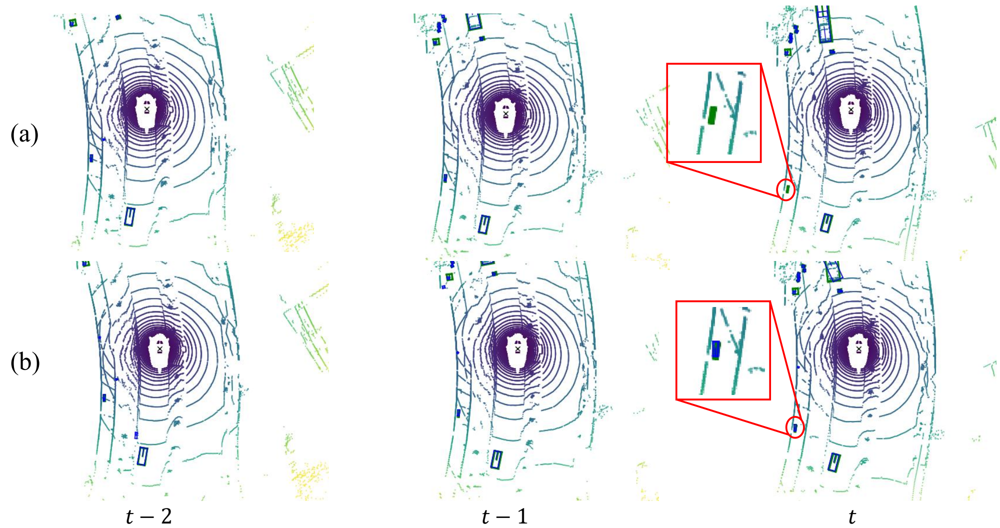

Figure 3: Comparison of LiDAR-only baseline TransFusion-L (a) and QTNet (b) along the temporal dimension. The ego vehicle is moving from bottom to top.

A.4 Discussions of potential societal impacts

Effectively utilizing temporal information is vital for autonomous driving. QTNet improves 3D detection performance with negligible computation cost and latency by a lightweight temporal fusion module MTM, which can utilize temporal information to improve the safety of autonomous driving in the real world. However, temporal fusion usually requires sensor synchronization in time, which puts forward higher requirements for the hardware of autonomous driving.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[2] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[3] Yukang Chen, Jianhui Liu, Xiaojuan Qi, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Scaling up kernels in 3d cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[4] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[5] Shengheng Deng, Zhihao Liang, Lin Sun, and Kui Jia. Vista: Boosting 3d object detection via dual cross-view spatial attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Junho Koh, Junhyung Lee, Youngwoo Lee, Jaekyum Kim, and Jun Won Choi. Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection. In *AAAI Conference on Artificial Intelligence*, 2023.

[7] Tao Lu, Xiang Ding, Haisong Liu, Gangshan Wu, and Limin Wang. Link: Linear kernel for lidar-based 3d perception. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[8] Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, and Xiang Bai. Cape: Camera view position embedding for multi-view 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[9] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. In *Advances in Neural Information Processing Systems*, 2022.

[10] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. In *AAAI Conference on Artificial Intelligence*, 2023.

[11] Junbo Yin, Jianbing Shen, Xin Gao, David Crandall, and Ruigang Yang. Graph neural network and spatiotemporal transformer attention for 3d video object detection from point clouds. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[13] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. In *arXiv preprint arXiv:1908.09492*, 2019.