GAFusion: Adaptive Fusing LiDAR and Camera with Multiple Guidance for 3D Object Detection

Supplementary Material

Methods	Modality	AP/L1	APH/L1	AP/L2	APH/L2
PV-RCNN++	L	80.26	78.71	75.00	73.52
MPPNet	L	81.83	80.59	76.88	75.67
CenterFormer	L	82.26	80.91	77.60	76.29
DeepFusion	LC	81.90	80.48	76.91	75.54
BEVFusion	LC	82.72	81.35	77.65	76.33
GAFusion(ours)	LC	**83.10**	**81.73**	**77.97**	**76.69**

Table 7. Comparison on the Waymo test set. The models in the table are without ensemble or test-time augmentation.

Class	3D obeject	2D object
Vehicle	6.1M	9.0M
Pedestrian	2.8M	2.7M
Cyclist	67k	81k
Sign	3.2M	-

Table 8. The number of 3D and 2D objects of different categories on the Waymo dataset.

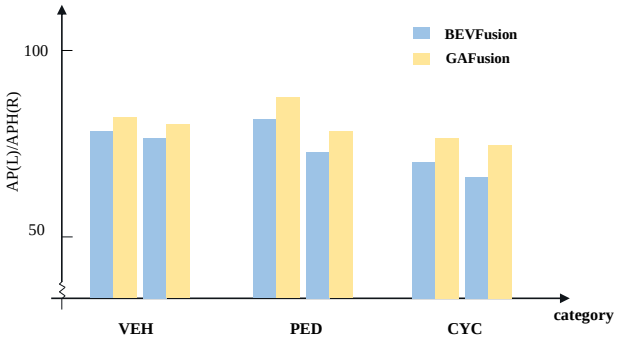

Figure 9. Comparison of BEVFusion and GAFusion on AP/APH (L2) for each category. For each category, the left is mAP and the right is mAPH. "VEH" is vehicle, "PED" is pedestrian, and "CYC" is cyclist.

6. Rationale

The supplementary document is organized as follows:

- The proposed model achieves the results on the Waymo dataset.
- The generalization study of the proposed LGAFT.
- Visualization of the model's prediction results in the night-time;

6.1. the results on the Waymo dataset.

The dataset contains 3,000 driving segments, each of which consists of 20 seconds of continuous driving footage. Each segment covers data from five high-resolution Waymo LiDARs and five front and side cameras. The entire dataset contains a total of 600,000 frames, with about 25 million 3D bounding boxes and 22 million 2D bounding boxes.

As shown in Table 7, we present the results of GAFusion on the Waymo test set and compare it with several models on the Waymo 3D detection leaderboard. We do not use test-time augmentation (TTA) or multi-model ensemble. GAFusion achieves excellent results on the Waymo 3D detection challenge with 76.69 mAPH (L2) detection performance.

In Table 8, we counted the number of 2D and 3D objects of different categories on the Waymo dataset. It can be seen that there are a large number of relatively small objects (pedestrian, cyclist and sign) on the dataset, which indicates that focusing on the detection accuracy of small objects can benefit the overall performance of the model.

Meanwhile, we compare BEVFusion and the proposed model on vehicle (VEH), pedestrian (PED), and cyclist(CYC) categories in Fig. 9. We observe that GAFusion has a significant improvement over BEVFusion, especially in small and distant objects detection.

Methods	LGAFT	mAP↑	NDS↑
BEVFusion		70.2	72.9
BEVFusion	✓	71.0$_{\uparrow 0.8}$	73.4$_{\uparrow 0.5}$
MSMDFusion		71.5	74.0
MSMDFusion	✓	72.1$_{\uparrow 0.6}$	74.3$_{\uparrow 0.3}$

Table 9. Generalization of LGAFT module on the nuScenes dataset.

6.2. Generalization of LGAFT module

To demonstrate the effectiveness and generalization of the proposed module LGAFT, we introduce this module into BEVFusion [26] and MSMDFusion [13]. As shown in Table 9, with the LGAFT module, BEVFusion achieves a sufficient improvement of 0.8% mAP and 0.5% NDS and MSMDFusion brings about 0.6% mAP and 0.3% NDS gain. It can be seen that the LGAFT module achieves a relatively significant improvement in different frameworks. This is attributed to the fact that LGAFT can adaptively fuse the BEV features of different modalities from a global perspective, which greatly enhances the interaction of semantic and geometric information and improves the intrinsic correlation between LiDAR and camera.

Figure 10. Visualization results of BEVFusion and GAFusion at night on the nuScenes validation set. The red circles and boxes show the detection ability of GAFusion for small and occluded objects.

6.3. Visualization

We show the visualization results on the nuScenes validation set of GAFusion and BEVFusion [26] at night in Fig. 10. With the help of our proposed modules, we find that GAFusion achieves excellent performance even at night. This is mainly reflected in its detection performance for distant small objects, which is attributed to better guidance and larger receptive fields.