MSMDFusion – Supplementary Material

1. Effect of Model Size

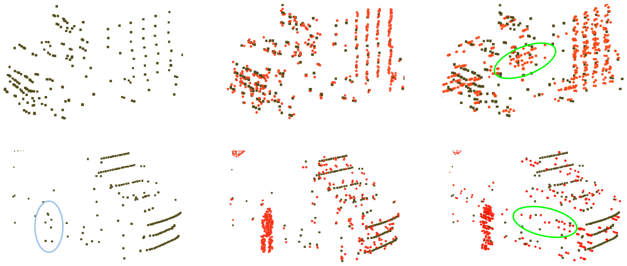
We vary the number of parameters of our main multi-modal interaction components (GMA-Conv) and test their performances to verify that the improvements are not simply because of the increased model capacity. Table 1 shows the performances of three models of different sizes. Specifically, compared to our final model ("medium"), the "small" model with about half of its parameters can achieve similar performances. And further increasing the model capacity to "large" does not bring extra improvements.

Model Size	mAP	NDS	Params	FPS
small	66.86	68.78	9.6M	**3.2**
medium	**66.93**	**68.93**	17.0M	2.1
large	66.67	68.83	27.9M	0.8

Table 1. Effects of using different numbers of parameters for multi-modal interaction (GMA-Conv). "medium" is our final model.

This demonstrates that parameter size hardly affects our model's performance, and the sufficient multi-modal interaction brought by our MDU and GMA-Conv, as well as the multi-scale fusion framework are indeed the main contributing factors.

To verify the number of model parameters' influence on the multi-modal interaction, we modify the model size and inspect the performances. Since the only parameters for multi-modal interaction are included in the GMA-Conv, we enlarge and reduce the parameters in each GMA-Conv shown as "small" and "large" in the Table 1, in respectively. We only calculate the total parameters within all GMA-Conv blocks for simplicity. The results show that the parameter numbers of GMA-Conv only have marginal effects on the model's performances, which demonstrates that MSMDFusion [2] mainly benefits from the multi-scale LiDAR-camera interaction and MDU and GMA-Conv within each scale, rather than merely enlarging the model capacity.



(a) LiDAR points

(b) LiDAR points + virtual points (1NN)

(c) LiDAR points + virtual points (6NN)

Figure 1. Comparison of (a) the original LiDAR points, (b) LiDAR points (black) and virtual points (red) generated by retrieving 1 nearest neighbor of the seeds, and (c) LiDAR points and virtual points generated by retrieving 6 nearest neighbors of the seeds. For fair comparison, we keep the total number of virtual points in (b) and (c) the same by varying the number of seeds sampled per instance.

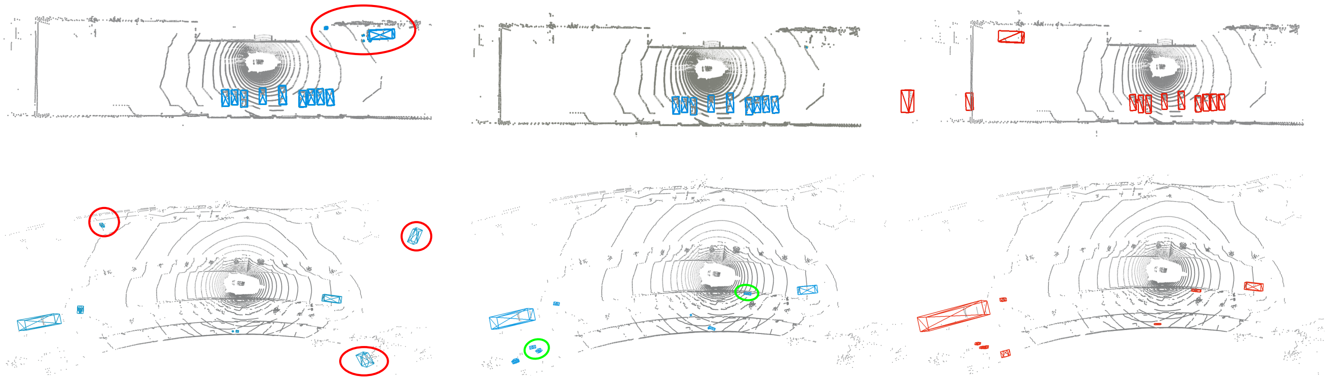
2. Qualitative Results

2.1. Visualization of virtual points

To intuitively illustrate the benefits of our Multi-Depth Unprojection strategy (MDU), we visualize the virtual points generated with different numbers of Nearest Neighbors (NN) in the raw point space as shown in Fig.1, where each row contains a different scene, and the original LiDAR points and the generated virtual points are shown in black and red, respectively. From these cases, we have the following observations. (i) The pedestrian in the bottom row of Fig.1(a) is only composed of a handful of LiDAR points (highlighted with blue circle), which can be easily confused with the nearby car. With the generated virtual points supplemented ((b) and (c) in the bottom row), the pedestrian can now be easily distinguished from the car. (ii) The virtual points generated with the 1-NN version of MDU tend to cluster around the real LiDAR points, while the 6-NN version of MDU can generate virtual points that better capture the objects' surface in 3D (in green circles). This demonstrates the benefits of our multi-depth unprojection strategy.

2.2. Visualization of detection results

We further qualitatively compare the 3D detection results predicted by a strong LiDAR-only baseline (i.e.,



(a) LiDAR-only baseline prediction (b) Our MSMDFusion prediction (c) Ground-truth bounding boxes

Figure 2. Comparison of detection results generated by (a) the LiDAR-only baseline and (b) our MSMDFusion model, and (c) the ground-truth bounding boxes for reference. We use red and green circles to mark the differences on the figure.

TransFusion-L [1]) and our MSMDFusion on the nuScenes validation set. The visualized results are shown in Fig.2, where each row contains a different scene. It can be seen that our MSMDFusion has significant advantages over the LiDAR-only baseline in two aspects. (i) Since the LiDAR point clouds lack semantic information, noisy points reflected from cluttered backgrounds may mislead the model's predictions as indicated by the red circle in Fig.2(a). By progressively introducing abundant semantics from images into the detector, those noisy parts can be correctly recognized as background by our model as shown in Fig.2(b). (ii) Small or faraway objects in the point cloud can contain only a limited number of LiDAR points, therefore, the LiDAR-only model may fail to detect them. By utilizing the 2D instance priors provided in camera images to generate virtual points and performing LiDAR-camera fusion at multiple scales to incorporate multi-granularity information, our model can effectively capture a part of the small or faraway objects (green circles in the bottom row of Fig.2(b)).

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 2

[2] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Xiaolin Wei, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. *arXiv preprint arXiv:2209.03102*, 2022. 1