Supplementary Material of Homogeneous Multi-modal Feature Fusion and Interaction for 3D Object Detection

Xin Li[1], Botian Shi[2], Yuenan Hou[2], Xingjiao Wu[1,3], Tianlong Ma[1,3], Yikang Li[2†], and Liang He[1†]

[1]East China Normal University [2]Shanghai AI Lab [3]Fudan University
{sankin0528, wuxingjiao2885}@gmail.com
{shibotian, houyuenan, liyikang}@pjlab.org.cn {tlma, lhe}@cs.ecnu.edu.cn
[†]Corresponding author

1 Elaborated Implementation Details

Depth Bins Discretization. As shown in Figure 2 of the main content, we obtain feature maps F from the last layer of the backbone. The feature maps F are used to generate the image-voxel features. Following [3, 2], we utilize discrete representations of depth. We categorize the continuous depth map into depth bin intervals since the estimation of long-range regions inherently yields large errors and thus needs to be relatively suppressed. In the designed depth range $[d\_min, d\_max]$, and each of depth bin interval is set as β. Then, the length of the next bin is always β longer than the previous bin. The calculation of the interval length β is provided in Equation 1:

$$\beta = \frac{2 \times (d\_max - d\_min)}{T \times (T + 1)} \tag{1}$$

The T categorizes it into 80 intervals while the background depth is set to non-category.

Inference Settings. At the inference stage, we first perform non-maximum suppression (NMS) in the RPN with IoU threshold 0.7 and keep the top 100 region proposals as the input of detect head. Then, after refinement, NMS is applied again with IoU threshold 0.1 and score threshold 0.65 to remove the redundant predictions.

1.1 BEV Visualization

To verify the effectiveness of the proposed modules for better cross-modal fusion, we visualize the original point cloud features and the fused features through different feature fusion methods in BEV, as shown in Fig. 1. Apparently, the proposed IVLM and QFM can better leverage richer image information to enhance the point cloud features. Adopting the VFIM is conducive to producing sharper and more accurate features that are useful for 3D detection.

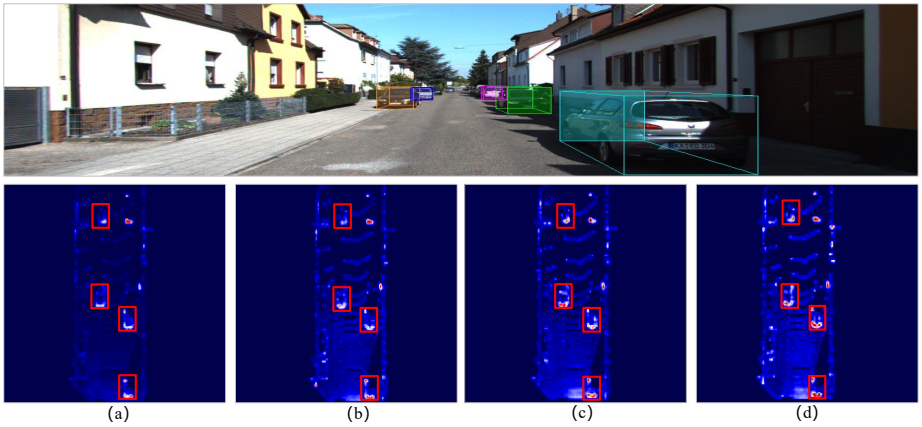

Fig. 1: The image with 3D ground-truth boxes is shown on the top row. (a) Original point cloud features without fusing images. (b) Fused features without lifting images to 3D voxel space. (c) Fused features by lifting images to 3D voxel space. (d) Fused features with our HMFI. Regions with large differences are highlighted using red rectangles.

2 More Ablation Study

Effect of the location of IVLM. As shown in Table 4 of the main content, we can find that the performance of the fusion model without IVLM is worse than achieving feature fusion on image voxel features generated by the IVLM. In this section, we also explore the impact of the IVLM's location on feature interactions. We consider two positions of IVLM: before QFM and after QFM. The former is our proposed HMFI pipeline. While the latter is that the QFM is applied to the 2D image representations directly, then introducing the IVLM lifts the 2D image representations into image voxel features. Finally, we adopt the VFIM on these two pipelines to achieve the cross-modal feature interaction. From Table 1 (a), our HMFI can achieve better performances than the scheme of introducing the IVLM after feature fusion by QFM, which also indicates that the homogeneous structure is a preferred way to build the cross-modal feature fusion and interaction.

Effect of the hyperparameter γ. In the proposed HMFI method, VFIM can bring significant performance gains and the feature interaction plays a crucial role in cross-modal feature fusion. Hyperparameter γ is set to optimize VFIM together with the whole network. Hence, we explore the effect of the hyperparameter γ. From the Table 1 (b), it shows that setting γ as 0.1 can achieve the optimal performance. Therefore, we choose $\gamma = 0.1$ for joint feature interaction.

Table 1: AP_{Easy}, $AP_{Mod.}$, and AP_{Hard} are the mAP performance of easy, moderate, and hard levels respectively.

(a) The location of IVLM.

Position of IVLM	AP_{Easy}	$AP_{Mod.}$	AP_{Hard}
Baseline [1]	81.34	71.76	67.09
Before QFM (HMFI)	**83.36**	**73.89**	**68.98**
After QFM	82.21	72.26	67.43

(b) Hyperparameter γ

γ	AP_{Easy}	$AP_{Mod.}$	AP_{Hard}
0.05	82.96	73.21	68.61
0.1	**83.36**	**73.89**	**68.98**
0.5	82.76	73.13	68.51

3 The Performance of HMFI on Single Stage Detector

To further verify the effectiveness of our HMFI, we also adopt the HMFI including the IVLM, QFM and VFIM on the commonly-used single stage detector [4]. We still use the anchor-based assignment following [4]. Other modules and configurations are kept the same to ensure fair comparison. It suggests that our HMFI can bring a significant performance gain of over 1.8 AP on all difficulty levels of the KITTI val set. In particular, HMFI achieves a remarkable gain of +2.48 AP on the hard level, which strongly demonstrates the effectiveness and generalization of our method.

Table 2: The Effectiveness of HMFI on Single Stage Detector [4].

Method	AP_{Easy}	$AP_{Mod.}$	AP_{Hard}
Baseline [4]	75.71	65.71	62.59
HMFI (Ours)	77.69	67.72	65.07
Improvements	+1.98	+2.01	+2.48

References

1. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In: AAAI. pp. 1201–1209 (2021)
2. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: CVPR. pp. 8555–8564 (2021)
3. Tang, Y., Dorn, S., Savani, C.: Center3d: Center-based monocular 3d object detection with joint depth understanding. In: DAGM German Conference on Pattern Recognition. pp. 289–302. Springer (2020)
4. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)