Supplementary Material of LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion

Xin Li[1] Tao Ma[2] Yuenan Hou[3] Botian Shi[3] Yuchen Yang[4] Youquan Liu[5]
Xingjiao Wu[4] Qin Chen[1] Yikang Li[3*] Yu Qiao[3] Liang He[1,6*]
[1]East China Normal University [2]The Chinese University of Hong Kong
[3]Shanghai AI Laboratory [4]Fudan University
[5]Hochschule Bremerhaven [6]Shanghai Key Laboratory of Multidimensional Information Processing
{sankin0528, wuxingjiao2885}@gmail.com {qchen, lhe}@cs.ecnu.edu.cn
{matao, shibotian, houyuenan, youquanliu, liyikang, qiaoyu}@pjlab.org.cn

1. More Implementation Details

In this section, we provide more detailed experimental settings on Waymo Open Dataset (WOD) [19] and KITTI [5] datasets. For WOD [19], we adopt the two-stage training recipe. We take CenterPoint [27] as our RPN. We use Adam optimizer with one-cycle learning rate policy, with max learning rate 3×10^{-3}, weight decay 0.01 and momentum 0.85 to 0.95. We also follow [27] to use the common data augmentations including global rotation, global scaling, translation along the z-axis and gt-sampling [24] to train our RPN for 20 epochs. Batch size is set as 64 and we use 8 NVIDIA A100 GPUs. When the model is trained in the last 5 epochs, we follow the same fading strategy proposed in [21] to remove gt-sampling augmentation. While in the two-stage refinement, we do not use gt-sampling to train the local-to-global fusion in our LoGoNet for 6 epochs. Batch size, the number of NVIDIA A100 GPUs and the learning rate settings are the same as the first stage. As for applying multi-frame cross-modal fusion in LoGoNet, besides the general classification and regression loss functions, we also add the IoU loss function [30, 32] to better account for the center-based object detection. More specifically, an IoU score is added in the prediction head which is supervised with the highest IoU between the prediction and all ground truth in a smooth L1 loss, and we use it to update the predicted confidence score during inference.

For KITTI [5], our LoGoNet is trained following the same training configuration as Voxel-RCNN [3]. We train the whole model end-to-end for 80 epochs where we use 8 NVIDIA A100 GPUs and batch size is set as 2 per GPU. We adopt the one-cycle learning rate policy, with maximum learning rate being 1×10^{-2}, weight decay being 0.01 and momentum being selected from 0.85 to 0.95. Due to the extremely imbalanced object distribution in the KITTI dataset, we follow [2, 12] to adopt the multi-modal gt-sampling during training.

Both in WOD and KITTI, for the image branch, we do not perform any data augmentations on images. We take Swin-Tiny [14] as the image backbone, initialize it from the public detection model and fix its weights during training.

*Corresponding author

Table A. Effect on the type of position information for LoF. XYZ indicates spatial grid locations, D and R indicate the number and the centroids of all points in each grid respectively.

Type	3D APH L2		
	VEH	PED	CYC
XYZ+D	68.14	66.69	69.07
XYZ+D+R	68.26	66.97	69.23

2. More Quantitative Results

2.1. Type of Position Information for LoF.

Table A shows the effect of position information composition in local fusion. This information in each grid is encoded by the MLP to generate grid features and fuse local image features. We find that richer grid information brings performance gain of 0.12%, 0.28%, and 0.15% on APH (L2) on the vehicle, pedestrian, and cyclist, respectively.

2.2. Detailed Comparison on the KITTI Test Set

We show the detailed comparison between LoGoNet and other state-of-the-art detectors on the KITTI *test* set in Table B. It shows that LoGoNet surpasses all published methods on the three classes simultaneously with 69.35 mAP. Notably, for the first time, LoGoNet outperforms existing

Table B. Comparison with state-of-the-art approaches for all three classes on the KITTI *test* set with AP being calculated at 40 recall positions. The mAPs are averaged over the APs of easy, moderate and hard levels. Best in bold.

Method	Modality	Car				Pedestrian				Cyclist				mAP
		Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP	
SECOND [24]	L	83.34	72.55	65.82	73.90	48.73	40.57	37.77	42.36	71.33	52.08	45.83	56.41	57.56
PointPillars [10]	L	82.58	74.31	68.99	75.29	51.45	41.92	38.89	44.09	77.10	58.65	51.92	62.56	60.65
STD [26]	L	87.95	79.71	75.09	80.92	53.29	42.47	38.35	44.70	78.69	61.59	55.30	65.19	63.60
SE-SSD [31]	L	91.49	82.54	77.15	83.73	-	-	-	-	-	-	-	-	-
PV-RCNN [17]	L	90.25	81.43	76.82	82.83	52.17	43.29	40.29	45.25	78.60	63.71	57.65	66.65	64.91
PDV [6]	L	90.43	81.86	77.36	83.21	47.80	40.56	38.46	42.27	83.04	67.81	60.46	70.44	65.30
F-PointNet [16]	L+I	82.19	69.79	60.59	70.86	50.53	42.15	38.08	43.59	72.27	56.12	49.01	59.13	57.86
AVOD-FPN [9]	L+I	83.07	71.76	65.73	73.52	50.46	42.27	39.04	43.92	63.76	50.55	44.93	53.08	56.84
PointPainting [20]	L+I	82.11	71.70	67.08	73.63	50.32	40.97	37.84	43.05	77.63	63.78	55.89	65.77	60.82
EPNet [8]	L+I	89.81	79.28	74.59	81.23	52.79	44.38	41.29	46.15	-	-	-	-	-
3D-CVF [28]	L+I	89.20	80.05	73.11	80.79	-	-	-	-	-	-	-	-	-
SFD [22]	L+I	91.73	84.76	77.92	84.80	-	-	-	-	-	-	-	-	-
Graph-VoI [25]	L+I	**91.89**	83.27	77.78	84.31	-	-	-	-	-	-	-	-	-
VFF [12]	L+I	89.50	82.09	79.29	83.62	-	-	-	-	-	-	-	-	-
FocalsConv [2]	L+I	90.55	82.28	77.59	83.47	-	-	-	-	-	-	-	-	-
HMFI [11]	L+I	88.90	81.93	77.30	82.71	50.88	42.65	39.78	44.44	84.02	70.37	62.57	72.32	66.49
CAT-Det [29]	L+I	89.87	81.32	76.68	82.62	**54.26**	45.44	41.94	47.21	83.68	68.81	61.45	71.31	67.05
LoGoNet (Ours)	L+I	91.80	**85.06**	**80.74**	**85.87**	53.07	**47.43**	**45.22**	**48.57**	**84.47**	**71.70**	**64.67**	**73.61**	**69.35**

all published methods by a large margin, surpasses the recent multi-modal method CAT-Det [29] method by 2.30% mAP and the LiDAR-only detector [6] by 4.05% mAP.

2.3. Evaluation Regarding Distance.

In Table C, we also report the comparison between our LoGoNet and other state-of-the-art methods on the WOD test leaderboard* based on performance regarding different distances for the vehicle class. It is evident that our method outperforms all previous methods by remarkable margins on all distance ranges in both LEVEL 1 and LEVEL 2. In particular, LoGoNet outperforms all previous methods at detecting distant objects by a large margin and surpasses the state-of-the-art method CenterFormer [32] by 2.53% APH (L2). It strongly demonstrates the effectiveness of the proposed local-to-global cross-modal fusion.

2.4. Inference Time Analysis

The inference time of multimodal 3D object detection is a vital factor considering its practicality in autonomous driving. We report the inference time of LoGoNet on both WOD and KITTI benchmarks. LoGoNet is evaluated using one NVIDIA A100 GPU and the batch size is set as 1. Table D shows the comparison between LoGoNet and previous competitive methods. LoGoNet achieves the best tradeoff between the accuracy and efficiency among all methods.

2.5. Model Ensembling Settings.

We follow [4, 7, 13] to use different test time augmentations, including point cloud global rotation, global scaling

*https://waymo.com/open/challenges/ 020/3d-detection/

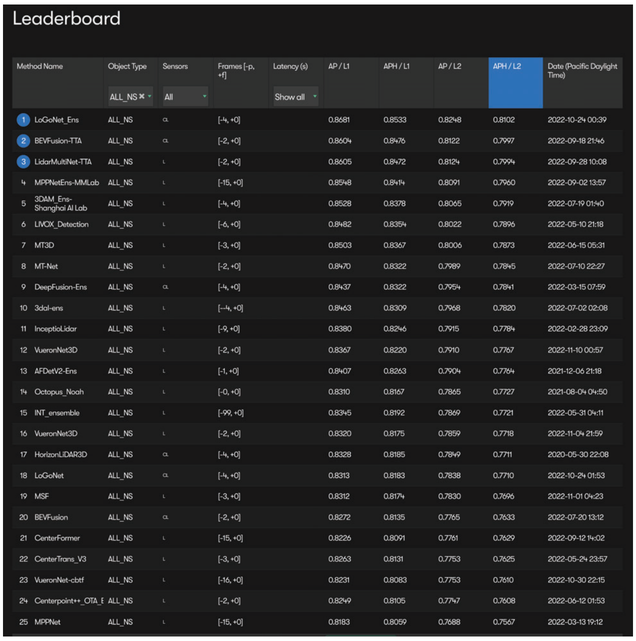
Table C. Performance comparisons with the state-of-the-art methods on the WOD test set for vehicle detection. † means multimodal methods.

Difficulty	Method	Vehicle APH			
		Overall	0-30m	30-50m	50m-Inf
LEVEL 1	PV-RCNN [17]	80.57	92.98	79.57	60.47
	CenterPoint++ [27]	82.33	92.42	81.61	64.13
	AFDetV2 [7]	81.22	92.12	79.29	61.75
	INT [23]	84.29	93.37	84.07	67.64
	DeepFusion† [13]	82.82	93.23	81.38	63.79
	MPPNet [1]	83.88	93.23	83.33	67.70
	CenterFormer [32]	84.94	94.17	84.21	67.96
	BEVFusion† [15]	84.55	94.04	83.67	67.25
	LoGoNet† (Ours)	**86.10**	**94.38**	**85.45**	**70.85**
LEVEL 2	PV-RCNN [17]	73.23	92.03	73.52	48.62
	CenterPoint++ [27]	75.05	91.17	75.89	52.02
	AFDetV2 [7]	73.89	90.85	73.50	50.03
	INT [23]	77.62	92.32	79.01	55.97
	DeepFusion† [13]	75.69	92.01	75.90	52.07
	MPPNet [1]	76.91	92.04	77.94	55.76
	CenterFormer [32]	78.28	93.12	79.06	56.32
	BEVFusion† [15]	77.48	92.89	78.14	55.08
	LoGoNet† (Ours)	**79.30**	**93.26**	**80.16**	**58.75**

and translation along z-axis, which is similar to the data augmentation in the training process. To be more specific, we use [0°,±22.5°, ±45°, ±135°, ±157.5°, 180°] for yaw rotation, [0.95, 1.05] for global scaling, and [-0.2m, 0m, 0.2m] for translation along the z-axis. For ensembling, we adopt the model ensemble by the 3D version of weighted box fusion (WBF) [18] to ensemble different models with the above test time augmentations. We obtain different

Table D. Inference time and performance comparisons on the WOD and KITTI *val* sets with competitive methods. We average the 3D mAPH (L2) on WOD *val* set. The mAP is averaged over the APs of moderate level across three classes on the KITTI *val* set. ‡ denotes the results are reported in [6].

Method	Modality	Waymo		KITTI	
		FPS	mAPH (L2)	FPS	mAP
PV-RCNN [17]	L	2.53	58.14	7.04	70.99
Voxel-RCNN‡ [3]	L	**10.98**	57.47	**13.51**	72.97
PDV [6]	L	2.94	60.56	7.41	73.44
EPNet [8]	L+I	-	-	9.10	67.85
VFF [12]	L+I	-	-	5.00	74.58
DeepFusion [13]	L+I	3.13	67.00	-	-
LoGoNet (Ours)	L+I	3.88	**71.38**	10.69	**74.70**


Leaderboard

	Method Name	Object Type	Sensors	Frames [-p, +f]	Latency (s)	AP / L1	APH / L1	AP / L2	APH / L2	Date (Pacific Daylight Time)
		ALL_NS	All		Show all					
1	LoGoNet_Ens	ALL_NS	CL	[-4, +0]		0.8681	0.8533	0.8248	0.8102	2022-10-24 00:39
2	BEVFusion-TTA	ALL_NS	CL	[-2, +0]		0.8604	0.8476	0.8122	0.7997	2022-09-18 21:46
3	LidarMultiNet-TTA	ALL_NS	L	[-2, +0]		0.8605	0.8472	0.8124	0.7994	2022-09-28 10:08
4	MPPNetEns-MMLab	ALL_NS	L	[-15, +0]		0.8548	0.8414	0.8091	0.7960	2022-09-02 13:57
5	3DAM_Ens-Shanghai AI Lab	ALL_NS	L	[-4, +0]		0.8528	0.8378	0.8065	0.7919	2022-07-19 01:40
6	LIVOX_Detection	ALL_NS	L	[-6, +0]		0.8482	0.8354	0.8022	0.7896	2022-05-10 21:18
7	MT3D	ALL_NS	L	[-3, +0]		0.8503	0.8367	0.8006	0.7873	2022-06-15 05:31
8	MT-Net	ALL_NS	L	[-2, +0]		0.8470	0.8322	0.7989	0.7845	2022-07-10 22:27
9	DeepFusion-Ens	ALL_NS	CL	[-4, +0]		0.8437	0.8322	0.7954	0.7841	2022-03-15 07:59
10	3dal-ens	ALL_NS	L	[--4, +0]		0.8463	0.8309	0.7968	0.7820	2022-07-02 02:08
11	InceptioLidar	ALL_NS	L	[-9, +0]		0.8380	0.8246	0.7915	0.7784	2022-02-28 23:09
12	VueronNet3D	ALL_NS	L	[-2, +0]		0.8367	0.8220	0.7910	0.7767	2022-11-10 00:57
13	AFDetV2-Ens	ALL_NS	L	[-1, +0]		0.8407	0.8263	0.7904	0.7764	2021-12-06 21:18
14	Octopus_Noah	ALL_NS	L	[-0, +0]		0.8310	0.8167	0.7865	0.7727	2021-08-04 04:50
15	INT_ensemble	ALL_NS	L	[-99, +0]		0.8345	0.8192	0.7869	0.7721	2022-05-31 04:11
16	VueronNet3D	ALL_NS	L	[-2, +0]		0.8320	0.8175	0.7859	0.7718	2022-11-04 21:59
17	HorizonLiDAR3D	ALL_NS	CL	[-4, +0]		0.8328	0.8185	0.7849	0.7711	2020-05-30 22:08
18	LoGoNet	ALL_NS	CL	[-4, +0]		0.8313	0.8183	0.7838	0.7710	2022-10-24 01:53
19	MSF	ALL_NS	L	[-3, +0]		0.8312	0.8174	0.7830	0.7696	2022-11-01 04:23
20	BEVFusion	ALL_NS	CL	[-2, +0]		0.8272	0.8135	0.7765	0.7633	2022-07-20 13:12
21	CenterFormer	ALL_NS	L	[-15, +0]		0.8226	0.8091	0.7761	0.7629	2022-09-12 14:02
22	CenterTrans_V3	ALL_NS	L	[-3, +0]		0.8263	0.8131	0.7753	0.7625	2022-05-24 23:57
23	VueronNet-cbtf	ALL_NS	L	[-16, +0]		0.8231	0.8083	0.7753	0.7610	2022-10-30 22:15
24	Centerpoint++_OTA_E	ALL_NS	L	[-2, +0]		0.8249	0.8105	0.7747	0.7608	2022-06-12 01:53
25	MPPNet	ALL_NS	L	[-15, +0]		0.8183	0.8059	0.7688	0.7567	2022-03-13 19:12

Figure A. Screenshot of the Waymo 3D detection leaderboard on the date of CVPR deadline, *i.e.*, Nov 12, 2022.

types of models with 5-frames and 3-frames with different gird sizes of [0.075m, 0.075m, 0.15m] and [0.1m, 0.1m, 0.15m]. The resulting model is named as LoGoNet_Ens in Table 1 of the main text.

2.6. Screenshot of Waymo 3D Detection Leaderboard

We submit detection results of LoGoNet to Waymo 3D detection leaderboard. As shown in Fig. A, our LoGoNet ranks **1**st on the detection leaderboard at the time of submission.

References

[1] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *ECCV*, 2022. 2

[2] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *CVPR*, pages 5428–5437, 2022. 1, 2

[3] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, pages 1201–1209, 2021. 1, 3

[4] Zhuangzhuang Ding, Yihan Hu, Runzhou Ge, Li Huang, Sijia Chen, Yu Wang, and Jie Liao. 1st place solution for waymo open dataset challenge–3d detection and domain adaptation. *arXiv preprint arXiv:2006.15505*, 2020. 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1

[6] Jordan SK Hu, Tianshu Kuai, and Steven L Waslander. Point density-aware voxels for lidar 3d object detection. In *CVPR*, pages 8469–8478, 2022. 2, 3

[7] Yihan Hu, Zhuangzhuang Ding, Runzhou Ge, Wenxin Shao, Li Huang, Kun Li, and Qiang Liu. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *AAAI*, pages 969–979, 2022. 2

[8] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *ECCV*, pages 35–52, 2020. 2, 3

[9] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IROS*, pages 1–8, 2018. 2

[10] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 2

[11] Xin Li, Botian Shi, Yuenan Hou, Xingjiao Wu, Tianlong Ma, Yikang Li, and Liang He. Homogeneous multi-modal feature fusion and interaction for 3d object detection. In *ECCV*, pages 691–707. Springer, 2022. 2

[12] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, pages 1120–1129, 2022. 1, 2, 3

[13] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, pages 17182–17191, 2022. 2, 3

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[15] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 2

[16] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, pages 918–927, 2018. 2

[17] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10529–10538, 2020. 2, 3

[18] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 2

[19] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 1

[20] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, pages 4604–4612, 2020. 2

[21] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pages 11794–11803, 2021. 1

[22] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *CVPR*, pages 5418–5427, 2022. 2

[23] Jianyun Xu, Zhenwei Miao, Da Zhang, Hongyu Pan, Kaixuan Liu, Peihan Hao, Jun Zhu, Zhengyang Sun, Hongmin Li, and Xin Zhan. Int: Towards infinite-frames 3d detection with an efficient framework. In *ECCV*, pages 193–209. Springer, 2022. 2

[24] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2

[25] Honghui Yang, Zili Liu, Xiaopei Wu, Wenxiao Wang, Wei Qian, Xiaofei He, and Deng Cai. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In *ECCV*, 2022. 2

[26] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, pages 1951–1960, 2019. 2

[27] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 1, 2

[28] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *ECCV*, pages 720–736, 2020. 2

[29] Yanan Zhang, Jiaxin Chen, and Di Huang. Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In *CVPR*, pages 908–917, 2022. 2

[30] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3555–3562, 2021. 1

[31] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*, pages 14494–14503, 2021. 2

[32] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *ECCV*, 2022. 1, 2