Power of Cooperative Supervision: Multiple Teachers Framework for Advanced 3D Semi-Supervised Object Detection

Supplementary Material

1. Overview

In this supplementary material, we provide additional analyzes and experiments to comprehensively validate the effectiveness of MultipleTeachers. In Sec. 2 we elaborate on the unique strengths of our proposed LiO dataset. In Sec. 3, we provide a detailed explanation of the implementation details for MultipleTeachers. In Sec. 4 we conduct extensive ablation studies to analyze the impact of hyperparameters and evaluate the compatibility of our framework with various baseline detectors, highlighting its adaptability and robustness. Finally, Sec. 5 includes a comparative visualization of pseudo-labels generated by MultipleTeachers and the strong baseline method HSSDA [4]. These visualizations emphasize the superior quality and accuracy of our pseudo-labels, further demonstrating the advantages of our framework in enhancing 3D semi-supervised object detection.

2. Unique Strengths of LiO Dataset

We summarize the unique strengths of our LiO. First, it captured diverse and complex urban environments, spanning large, medium, and small cities across various conditions (day/night and weather). It provides finely differentiated annotations for a wide range of object categories, including multiple types like kick scooters, fire trucks, and ambulance. We ensured its high quality by conducting at least three rounds of iterative feedback between annotators and expert reviewers. Second, the data was collected by a 128-channel RS Ruby LiDAR, and out of 585 raw sequences recorded. We carefully selected 105 sequences with the highest object density to maximize object diversity. These frames were sampled at 2 Hz and annotated across seven major object classes. Despite fewer total time of data, our LiO achieves scenario diversity comparable to public datasets, while also offering a more diverse set of object classes to facilitate detailed evaluation (see Tab. 1).

3. Implementation Details

As shown in Tab. 2, MultipleTeachers follows the training configurations of HSSDA. For the KITTI dataset [3], the detection ranges are defined as [0, -40, 70, 40] along the X and Y axes, and [-3, 1] along the Z axis. For the Waymo Open Dataset (WOD) [9] and LiO datasets, the detection ranges are set to [-75, -75, 75, 75] on the X and Y axes, while the Z axis is defined as [-4, 2] for WOD and [-5, 5] for LiO. For nuScenes dataset [1] , the detection ranges are defined as [-52, -52, 52, 52] along the X and Y axes, and [-5, 3] along the Z axis. In addition, Tab. 3 presents the data ratios according to the settings for each dataset, which were used to conduct various experiments on the four datasets.

To achieve state-of-the-art (SOTA) performance, MultipleTeachers leverages advanced techniques, including shuffle data augmentation and dual-dynamic thresholding [4]. In addition, the PointGen module is carefully configured by setting its angular parameter, deg to 45-degree and the parameters γ and σ to 0.4 and 0.6, respectively.

To maintain a consistent update of pseudo-labels, MultipleTeachers applies dual-dynamic threshold updates every 5 epochs. Similarly, the PointGen memory bank, which is responsible for generating pseudo-points, is refreshed at the same interval. This periodic update ensures that the pseudo-point samples (P_{obj}) remain representative and contribute effectively to performance gains.

4. Additional Experimental Results

Analysis of Hyperparameters in PointGen. We analyze the impact of key hyperparameters (γ, σ, deg) on the performance of our detector. Specifically, γ and σ are weight parameters that balance the confidence score S_{conf} and density score $S_{density}$ in PointGen module. To evaluate their effects, we conduct experiments with different combinations of γ and σ, as detailed in Tab. 4 and 5. The results show that the model achieves the best performance when γ is set to 0.4 and σ to 0.6.

Furthermore, we investigate the influence of deg which determines the angular size of the pie-shaped regions in PointGen. In KITTI, experiments are carried out with deg values of 15, 30 and 45. In LiO, experiments are performed with various deg settings of 15, 30, 45, 60, and 120. As shown in Tab. 6 and 7, 45-degree yields the highest mAP, achieving 72.2 mAP on the KITTI moderate level and 50.8 mAP on the LiO dataset. However, performance remains stable across the range of angles, indicating that PointGen is relatively robust to changes in deg.

Effectiveness of PointGen. To demonstrate the effectiveness of the PointGen module, we present the results of its application across different network configurations in Tab. 8. Notably, the highest detection accuracy is observed when PointGen is applied solely to the teacher network prior to pseudo-label generation. This suggests that PointGen contributes to generating more accurate pseudo-labels in the specialized teacher network, ultimately leading to enhanced detection performance in the student network.

Dataset	Class Distribution (%)				3D Box Counts (K)						
	Veh.	L.V.	Ped.	Cyc.	Car	Ped.	Bic.	Bus	Tru.	O.V.	Mot.
KITTI [3]	82	-	13	5	29	4	1	-	-	-	-
WOD [9]	68	-	31	1	6023	2771	66	-	-	-	-
nuScenes [1]	56	17	25	3	493	222	11	16	88	15	13
H3D [7]	46	12	41	1	515	458	12	117	95	30	2
ONCE [5]	60	5	16	19	146	37	47	4	7	-	-
LiO	62	11	22	4	471	168	10	17	61	6	22

Table 1. Distributions of class categories and 3D box counts between the LiO dataset and public datasets. Note that, 'Veh.', 'L.V.', 'Ped.', 'Cyc.', 'Bic.', 'Tru.', 'O.V.', and 'Mot.' indicate vehicle, large vehicle, pedestrian, cyclist, bicycle, truck, other vehicle, and motorcycle, respectively.

Configuration	Dataset			
	KITTI	WOD	LiO	nuScenes
Detector	PV-RCNN [8]			CenterPoint [11]
Batch size	2			2
Weight decay	0.001			0.01
Learning rate	0.01			0.001
Optimizer	Adam			Adam
Epoch	80	30	30	20
Det.R. -x axis	70	150	150	108
Det.R. -y axis	80	150	150	108
Det.R. -z axis	4	6	10	8
V.S. -x axis	0.05	0.1	0.1	0.075
V.S. -y axis	0.05	0.1	0.1	0.075
V.S. -z axis	0.1	0.2	0.25	0.2

Table 2. Detailed training configuration of MultipleTeachers. Note that 'Det.R.' denotes detection range (m) and 'V.S.' is voxel size (m).

Dataset	Setting	Ratio (%)	
		L.Split	U.Split
KITTI	All (100%)	1	99
		2	98
		15	85
		20	80
WOD	Small (5%)	1	4
	Medium (20%)	1	19
	Large (100%)	1	99
nuScenes	Medium (20%)	5	15
LiO	Small (16K)	1	99
		2	98
		15	85
	Large (112K)	15	85

Table 3. The experiment data ratios according to the settings for each dataset. Note that 'L.Split' denotes labeled split and 'U.Split' is unlabeled split.

Density	Confidence	mAP		
		Easy	Mod.	Hard
0.2	0.8	83.1	70.7	65.4
0.4	0.6	82.4	70.4	65.3
0.6	**0.4**	**83.9**	**72.2**	**67.0**
0.8	0.2	83.3	70.9	65.8

Table 4. Ablation study on various combinations of confidence and density scores (γ and σ) in the Pseudo-points generator (PointGen) on 2% KITTI labeled split. The evaluation metrics are divided into difficulty levels: Easy, Moderate (Mod.), and Hard.

Density	Confidence	mAP	Car	Ped.	Mot.	Bic.	...
0.2	0.8	50.6	78.4	45.7	50.9	20.8	...
0.4	0.6	50.1	78.6	45.7	49.5	19.4	...
0.6	**0.4**	**50.8**	**78.8**	**46.4**	**51.8**	20.5	...
0.8	0.2	50.1	77.9	45.6	50.0	19.3	...

Table 5. Ablation study on various combinations of confidence and density scores in the PointGen on 2% LiO labeled split.

Degree	mAP		
	Easy	Moderate	Hard
15	83.6	71.3	65.9
30	82.7	70.6	65.3
45	**83.9**	**72.2**	**67.0**

Table 6. Ablation study on the degree-wise PointGen on 2% KITTI labeled split.

Comparison with SOTA on KITTI. Due to the HSSDA github repository does not provide a KITTI 20% split, we contacted the authors of DetMatch [6] and obtained their official 20% split. On this split, MultipleTeachers achieves the same performance as the SOTA baseline (see Tab. 9).

Component-wise Analysis on WOD. To further validate the superiority of our proposed modules, we individually analyze the effectiveness of each component by integrating the MGen and PointGen modules into the baseline model (see Tab. 10). Specifically, replacing the general SGen module with our proposed MGen module substantially improves detection performance, resulting in a notable gain of 4.4 mAP (Level 1). Furthermore, combining both the MGen and PointGen modules leads to an additional gain of 1.9 mAP (Level 1). These incremental improvements clearly demonstrate the effectiveness and advantages of our proposed modules.

Degree	mAP	Car	Ped.	Mot.	Bic.	...
15	50.3	78.0	45.4	49.7	19.9	...
30	50.3	78.1	45.7	49.1	20.4	...
45	**50.8**	**78.8**	**46.4**	**51.8**	**20.5**	...
60	50.7	78.2	45.7	49.9	20.4	...
120	50.3	78.2	45.8	49.1	19.9	...

Table 7. Ablation study on the degree-wise PointGen using 2% LiO labeled split.

Method	PointGen		mAP
	Teacher	Student	
MutipleTeachers †			70.2
MutipleTeachers		✓	70.3
MutipleTeachers	✓		**72.2**

Table 8. Ablation study on the impact of PointGen in various teacher-student network on 2% KITTI labeled split. Note that, † denotes the baseline method that does not use PointGen.

Dataset	Method	mAP	Car	Ped.	Cyc.
KITTI	DetMatch [6]	68.7	78.7	57.6	69.6
	HSSDA [4]	**71.6**	82.5	59.1	73.2
	MultipleTeachers	**71.6**	83.1	57.2	74.4

Table 9. Performance comparison of MultipleTeachers with SOTA models on 20% KITTI labeled split.

Method	SGen	MGen	PointGen	mAP	
				L1	L2
MultipleTeachers	✓			38.4	33.4
		✓		42.8	37.8
		✓	✓	**44.7**	**39.4**

Table 10. Ablation study on the effect of each component in MultipleTeachers. All experiments are performed on ’Small’ setting of WOD. The last row presents the results achieved by utilizing the proposed MGen and PointGen modules, respectively.

Trade-off between Performance and Cost. We conduct an extensive ablation in which we varied the number of teachers and measure accuracy, parameter count, training time, and GPU memory on 8 RTX 3090 GPUs. Across both WOD and nuScenes, we observe the same encouraging pattern: once the teacher pool grows to three or more, training time rises only modestly and peak memory barely increases, yet accuracy leaps dramatically. In other words, every extra teacher costs little but pays off handsomely—delivering a remarkably favorable cost-to-performance ratio (see Tab. 11). Especially, we add teachers for traffic-cones and barriers on nuScenes. Most importantly, inference always runs on a single student network, so deployment introduces zero additional latency or memory overhead. In short, MultipleTeachers demands just a modest, one-time training investment yet unlocks substantial accuracy gains while preserving real-time performance.

Dataset	#Tea.	mAP	#Param.	Time (h.)		Mem. (GB)	
			(M)	Train	Test	Train	Test
WOD	1	38.4	26	13.9	0.27	10.2	4.2
	3	42.8	52	25.9	0.27	13.9	4.2
nuScenes	1	38.4	17	4.2	0.04	7.6	2.6
	4	41.7	40	5.7	0.04	7.7	2.6
	5	42.0	48	6.1	0.04	7.9	2.6

Table 11. Comparison of performance and cost on various number of teachers. ‘Tea.’ indicates the number of teachers, ‘Param.’ denotes the number of model weight parameters, and ‘Mem.’ is GPU memory usage. PV-RCNN is adopted as the baseline detector on ’Small’ setting of WOD, while CenterPoint [11] is used as the baseline on ’Medium’ setting of nuScenes. Note that the PointGen module is not used in these experiments.

Dataset	Detector	SSL Method	mAP
LiO (Small 2%)	PV-RCNN [8]	Baseline ∗	30.4
		HSSDA ∗[4]	40.4
		MultipleTeachers	**49.1**
	Voxel-RCNN [2]	Baseline ∗	31.3
		HSSDA ∗[4]	42.4
		MultipleTeachers	**50.8**
	SECOND [10]	Baseline ∗	18.5
		HSSDA ∗[4]	26.4
		MultipleTeachers	**37.4**
LiO (Large)	PV-RCNN [8]	Baseline ∗	59.4
		HSSDA ∗[4]	60.8
		MultipleTeachers	**61.4**
	Voxel-RCNN [2]	Baseline ∗	61.1
		HSSDA ∗[4]	62.5
		MultipleTeachers	**62.7**

Table 12. Experimental results for different detectors on ‘Small’ and ‘Large’ setting of LiO. For a fair comparison, all models incorporate SSL components from MultipleTeachers, including MGen and PointGen. Note that ‘∗’ is re-implemented by us.

Adaptability across Diverse Detectors. Additional experiments on the LiO dataset further demonstrate the robust generalization capability of MultipleTeachers. As presented in Tab. 12, our proposed method consistently outperforms the recent HSSDA model, on various detectors. These competitive results further confirm that our framework can be effectively applied across various detectors.

5. Pseudo-Labels Visualization

In this subsection, we analyze the quality of pseudo-labels, a critical factor in determining the detection accuracy of the student network. As depicted in Fig. 1, we visualize and compare pseudo-labels generated at training epochs 20, 60, and 80 using the KITTI dataset with 2% labeled split. In the ground truth (GT) boxes of the KITTI dataset, we define that ‘Van’ class is included within Car category for consistency. The regions enclosed by red circle lines indicate areas

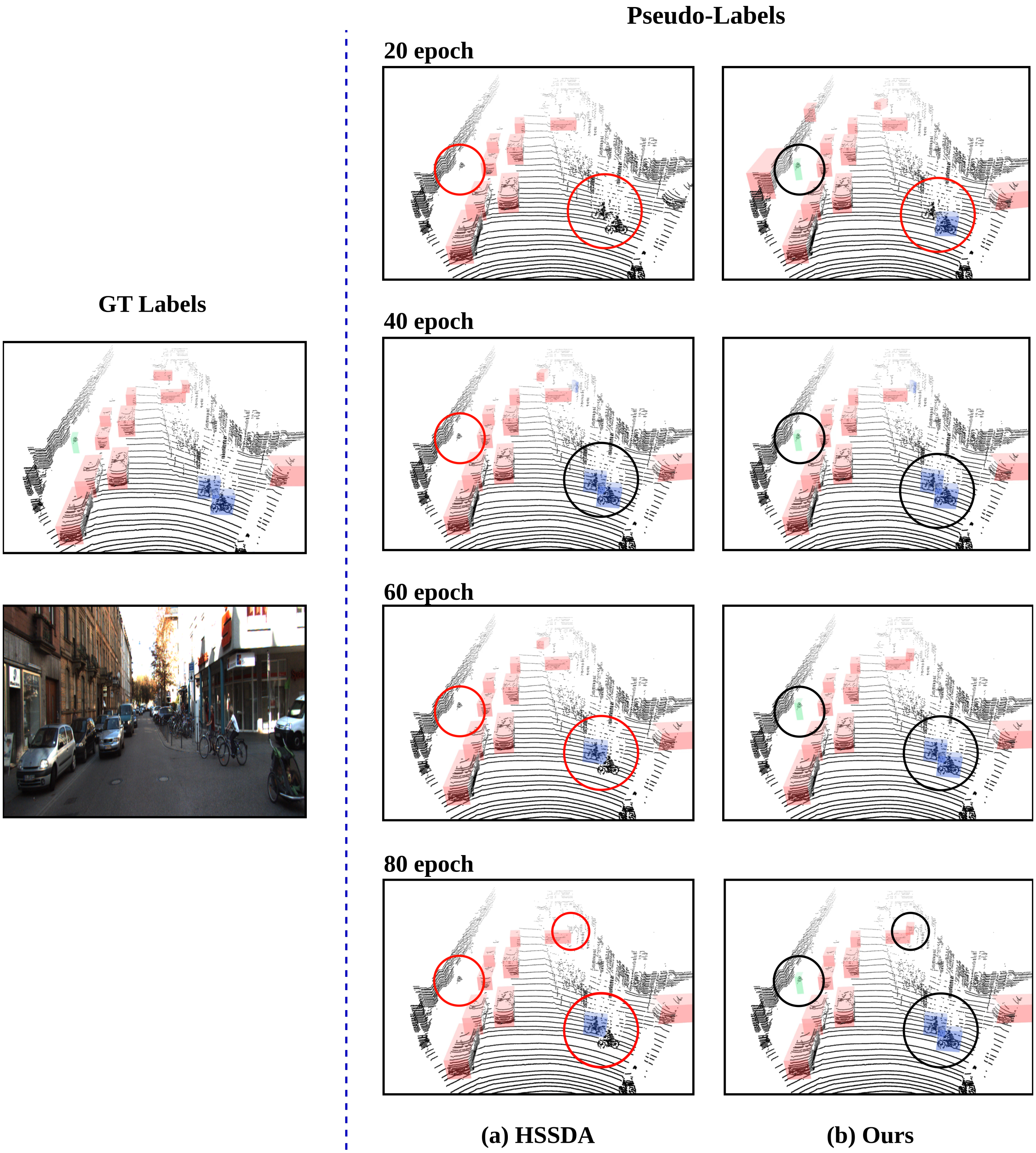

Figure 1. **Qualitative comparison of pseudo-labels between MultipleTeachers and HSSDA on the KITTI dataset with 2% labeled data split.** This figure compares pseudo-labels generated by our MultipleTeachers framework and the previous approach, showcasing the progressive accuracy improvements achieved by MultipleTeachers. Pseudo-labels are visualized every 20 epochs by the teacher network, highlighting the enhanced accuracy and reliability of our method in capturing diverse object characteristics.

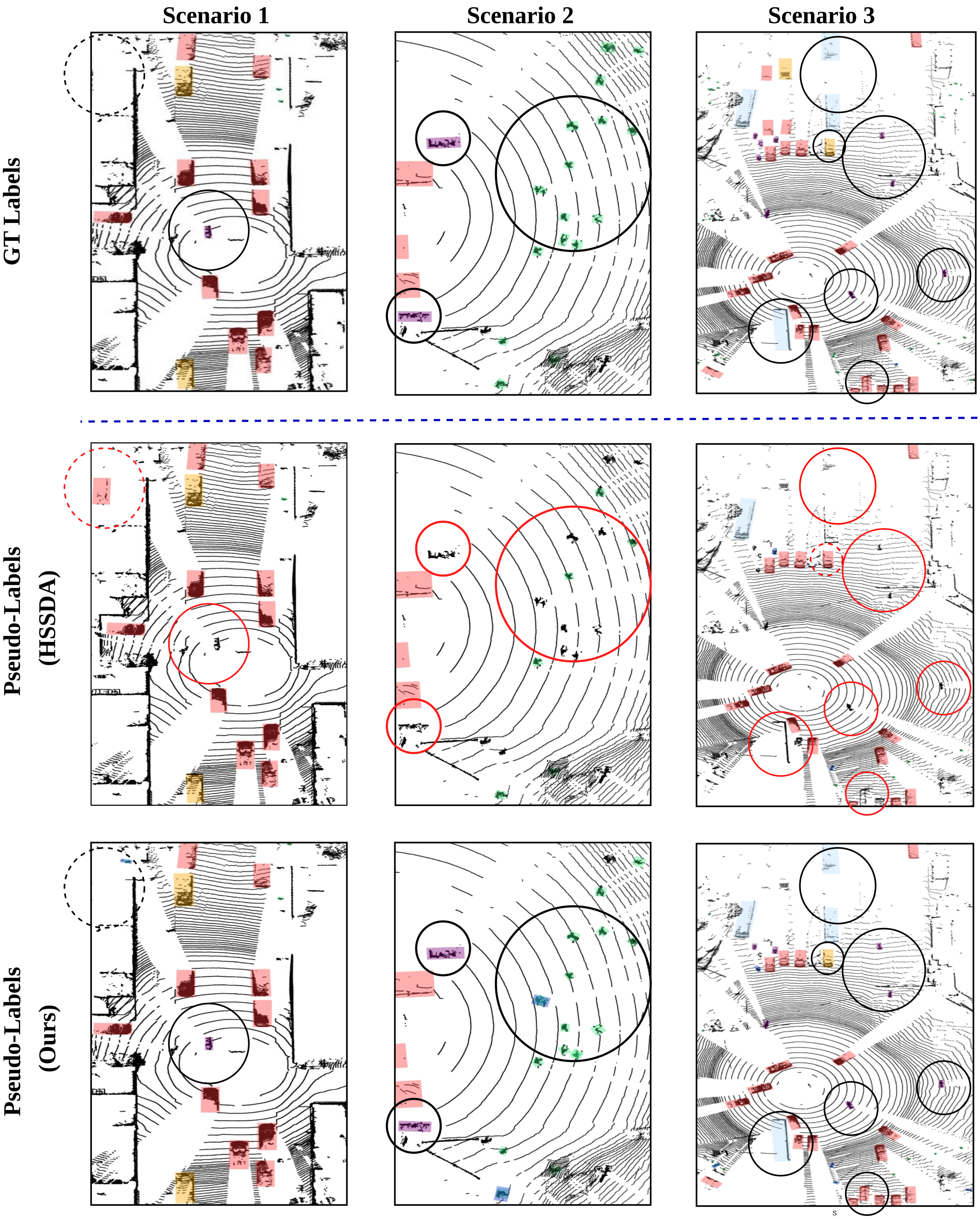

Figure 2. **Qualitative comparison of pseudo-labels on the LiO dataset using 2% labeled data split.** The top row displays the ground truth annotations, the middle row presents pseudo-labels generated by HSSDA, and the bottom row shows pseudo-labels predicted by our MultipleTeachers framework. This comparison highlights the improved precision and reliability of pseudo-labels produced by our approach, demonstrating its ability to capture finer object details and reduce false negatives effectively.

with false negatives, while black circle lines highlight true positives.

The HSSDA model exhibits limitations, struggling to produce accurate pseudo-labels for the cyclist class, even after several epochs, and failing to detect the pedestrian class at all until 80 epochs. In contrast, our MultipleTeachers framework demonstrates consistent and robust performance, generating high-quality pseudo-labels for both cyclist and pedestrian throughout the training progress.

This improvement underscores the significance of progressively accurate weight updates in the teacher network, enabled by the C-EMA module. This mechanism effectively enhances pseudo-label quality, which translates into superior detection performance. This analysis further validates the efficacy of MultipleTeachers in addressing challenges related to pseudo-labeling, particularly in sparse and complex environments.

Fig. 2 presents the experimental results on the LiO dataset, comparing the pseudo-label quality generated by the HSSDA and MultipleTeachers frameworks across various scenarios. The first row illustrates the GT boxes, the second row shows predictions from the HSSDA, and the last row displays predictions from our MultipleTeachers method.

Our approach demonstrates a marked improvement in pseudo-label accuracy, as evidenced in the regions highlighted by black circle lines indicating true positives. MultipleTeachers effectively captures both small objects such as pedestrians, motorcycles, and bicycles, and larger ones like trucks and buses, surpassing the performance of HSSDA. In contrast, the regions enclosed by red circle lines and dotted red circle lines highlight areas where HSSDA exhibits false negatives and false positives, respectively.

These results emphasize the advantages of our category-specialized teacher networks in generating precise pseudo-labels. By producing reliable labels, MultipleTeachers framework enables the student network to achieve significantly enhanced 3D object detection performance, particularly in challenging and diverse urban scenarios. This evaluation further validates the robustness of our method for semi-supervised learning tasks in 3D object detection.

References

[1] 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 *CVPR*, pages 11618–11628, 2020. 1, 2

[2] 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. 3

[3] Andreas. Geiger, Phili. Lenz, and Raquel. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 1, 2

[4] Chuandong. Liu, Chenqiang. Gao, Fangcen. Liu, Pengcheng. Li, Deyu. Meng, and Xinbo. Gao. Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection. In *CVPR*, pages 23819–23828, 2023. 1, 3

[5] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, and Chunjing Xu. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2

[6] J. Park, C. Xu, Y. Zhou, M. Tomizuka, and W. Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. In *ECCV*, pages 370–389, 2022. 2, 3

[7] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. pages 9552–9557, 2019. 2

[8] 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

[9] Pei. Sun, Henrik. Kretzschmar, Xerxes. Dotiwalla, Aurelien. Chouard, Vijaysai. Patnaik, Paul. Tsui, James. Guo, Yin. Zhou, Yuning. Chai, and Benjamin. Caine. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 1, 2

[10] Yan. Yan, Yuxing. Mao., and Bo. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[11] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *CVPR*, pages 11779–11788, 2021. 2, 3