Supplementary Material: HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Contrastive Regularization

Mengtian Li[1], Yuan Xie[1], Yunhang Shen[2], Bo Ke[2], Ruizhi Qiao[2], Bo Ren[2], Shaohui Lin[1,†], Lizhuang Ma[1,†]
[1]School of Computer Science and Technology
East China Normal University, Shanghai, China
[2]Tencent Youtu Lab
mtli@stu.ecnu.edu.cn, {yxie,shlin,lzma}@cs.ecnu.edu.cn
{odysseyshen,boke,ruizhiqiao,timren}@tencent.com

1. Overview

In the supplementary material, we start with more details of the training setup in Sec. 1.1 and model complexity in Sec. 1.2. Further, we give the per-class scores of ScanNet-V2 [4], Semantic3D [6] and SemanticKITTI [2] in Sec. 1.3, and provide more visual results in Sec. 1.4. Finally, we compare the dynamic augmentor and the fixed one in Sec.1.5.

1.1. Training Setup

Weakly Setting. We create the weakly-supervised dataset by randomly annotating a tiny fraction of points in a class for each point cloud sample. Specifically, we set up two weakly-supervised training methods: 1pt and 1%. At 1pt setting, we annotate one point for each class for each point cloud sample. At 1% setting, we select 1% of the points that are labeled for each class randomly, and these labeled points will not change during the training. Thus, at the semantic level, we only annotate some points for each semantic class as this is a form of weak-supervision (incomplete supervision) defined by Zhou *et al*. [20]. In addition, Xu *et al*. [17] and Zhang *et al*. [18] also define incomplete supervision as a weakly-supervised task. Therefore, we follow the definition in this paper.

Training configuration. Here we have supplemented the experimental details of the main paper. Our network training is conducted on the RTX Titan GPU with 24 GB memory. We use a grid size of $4cm$ for indoor dataset and $6cm$ for outdoor dataset to down-sample the raw point clouds, let the barycenter of each small grid be the selected point. Then the network takes input point clouds of size 40960 points for all datasets during training.

† Corresponding authors.

Method	Training time	Network parameters	Total reference time
PSD(1%) [18]	302	1.10	263
HybridCR(1%)	387	1.51	279

Table 1. The training time of per-epoch (in seconds), the network parameters (in millions) and total test time (in seconds) on S3DIS.

1.2. Model Complexity

We list the training time of per-epoch, the network parameters, and the total test time in Tab. 1 compared with PSD [18]. Since the parameters of the Siamese network are shared, only the parameters of dynamic point cloud augmentor are added compared to the PSD, so that the parameters of HybridCR are more by 0.41M than PSD. Since the augmentation operation and pseudo label selection are only introduced in the training phase, the training time of HybridCR is 85s per epoch longer than PSD. In comparison, the total reference time of HybridCR is relatively similar with PSD. Considering the significant improvements on quantitative results provided by HybridCR, it is still an efficient method.

1.3. Detailed Quantitative Results

Evaluation on ScanNet-V2. We present the segmentation performance of per-class on the ScanNet-V2 and choose the weakly-supervised setting of 1% for comparison. From Tab. 2. It can be observed that our HybridCR achieves 56.8% mIoU and 2.1% improvements against PSD. Moreover, in the aspect of specific classes, our method gains 11.1%, 8.8%, 8.4%, improvements in “door”, “other-furniture”, “curtain” against PSD, respectively, and achieve the best performance on “picture”. While HybridCR can significantly improve the performance of these classes and demonstrate that our method can learn more discrimina-

Set.	Methods	**mIoU(%)**	bath-tub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	floor	other-furniture	picture	refrigerator	shower-curtain	sink	sofa	table	toilet	wall	window
Fully	PointNet [12]	33.9	58.4	47.8	45.8	25.6	36.0	25.0	24.7	27.8	26.1	67.7	18.3	11.7	21.2	14.5	36.4	34.6	23.2	54.8	52.3	25.2
	PCNN [1]	49.8	55.9	64.4	56.0	42.0	71.1	22.9	41.4	43.6	35.2	94.1	32.4	15.5	23.8	38.7	49.3	52.9	50.9	81.3	75.1	50.4
	SegGCN [8]	58.9	83.3	73.1	53.9	51.4	78.9	44.8	46.7	57.3	48.4	93.6	39.6	6.1	50.1	50.7	59.4	70.0	56.3	87.4	77.1	49.3
	PointConv [16]	66.6	78.1	75.9	69.9	64.4	82.2	47.5	77.9	56.4	50.4	95.3	42.8	20.3	58.6	75.4	66.1	75.3	58.8	90.2	81.3	64.2
	KPConv [14]	68.4	84.7	75.8	78.4	64.7	81.4	47.3	77.2	60.5	59.4	93.5	45.0	18.1	58.7	80.5	69.0	78.5	61.4	88.2	81.9	63.2
	RFCR [5]	70.2	88.9	74.5	81.3	67.2	81.8	49.3	81.5	62.3	61.0	94.7	47.0	24.9	59.4	84.8	70.5	77.9	64.6	89.2	82.3	61.1
	HybridCR	59.9	87.2	70.7	68.3	56.1	78.4	46.3	61.6	46.5	45.6	93.6	42.7	20.7	46.4	56.7	53.1	69.5	48.0	71.3	76.9	58.4
weakly	PSD(1%) [18]	54.7	57.1	67.8	65.9	46.5	77.8	38.8	52.8	49.2	30.4	93.3	38.7	30.7	43.1	38.2	52.6	66.9	57.2	71.6	60.9	50.6
	HybridCR(1%)	56.8	58.9	65.8	66.8	42.3	80.2	36.7	61.2	58.1	45.5	90.1	47.5	33.4	41.0	37.5	51.1	70.5	60.8	71.0	60.1	57.9

Table 2. Quantitative results of per class on ScanNet-V2 [4]. (mIoU %)

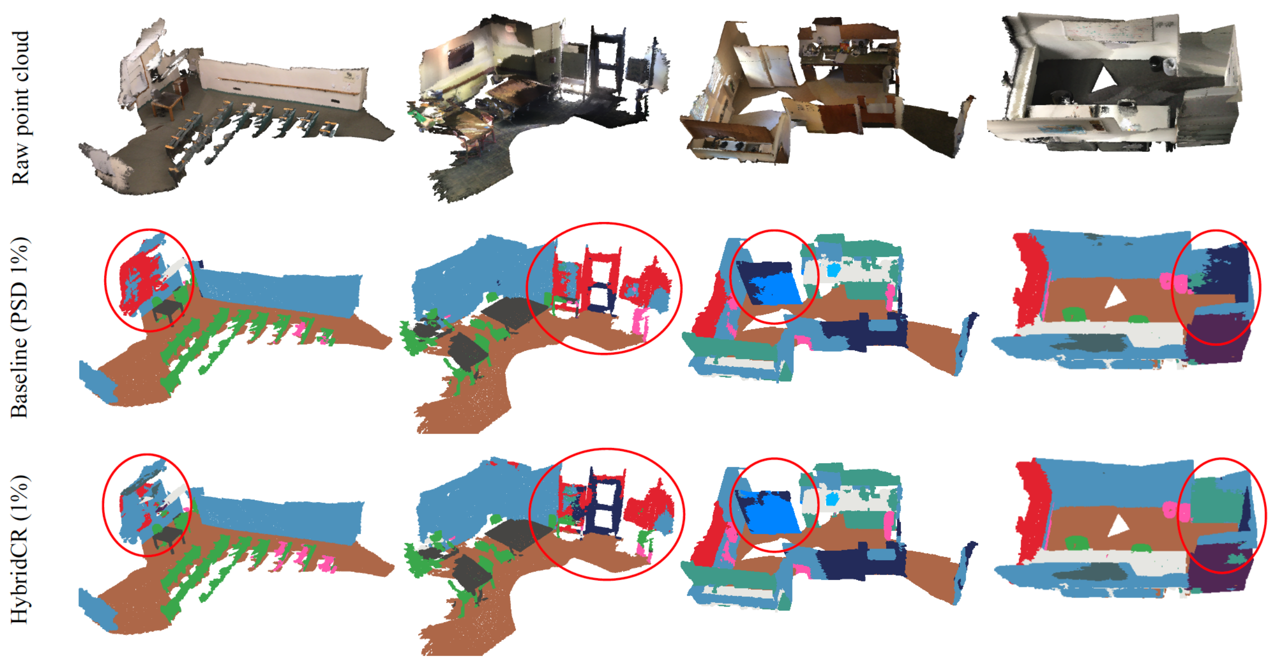

Figure 1. Visualization results on the test set of ScanNet-V2. Raw point cloud, results of the baseline and ours are presented separately from top to bottom.

tive features. Besides, We achieve comparable performance close to the fully-supervised SegGCN [8], which shows that our method is effective for weakly-supervised point cloud segmentation.

Evaluation on Semantic3D. We conduct the quantitative evaluations on Semantic3D (reduced-8) and list the per-class scores in Tab. 3. Mean Intersection-over-Union (mIoU) and Overall Accuracy (OA) of all classes are used as the standard metrics. We compared some full supervised methods published in recent years such as SnapNet [3], SEGCloud [13], ShellNet [19], KPConv [14], RandLA-Net [7], and PointGCR [9], RFCR [5]. At 1% setting, HybridCR achieves 76.8% and 94.9% in terms of both mIoU and OA, comparable to the fully-supervised methods. Compared with the fully supervised RandLA-Net, HybridCR is 0.6% lower than RandLA-Net in mIoU while 0.1% higher in OA, respectively. But, we achieve the best performance in the classes of "man-made" and "nature". Therefore,, the results show that HybridCR can generate to the sparse outdoor dataset.

Evaluation on SemantucKITTI. We conduct the quantitative evaluations on SemanticKITTI and list the per-class scores in Tab. 4. We compared some full supervised methods published in recent years, including PointNet [11],SqueezeSegV2 [15], DarkNet53Seg [2], RangeNet53++ [10] and RandLA-Net [7]. It can be found that HybridCR achieves the best performance among the fully-supervised setting comparison. At 1% setting, HybridCR reports 52.3% in mIoU, which are close to the performance of the fully-supervised methods. Compared with

Set.	Methods	mIoU(%)	OA	man-made.	natural.	high-veg.	low-veg.	buildings	hard-scape	scanning-art.	cars
Fully	SnapNet [3]	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
	SEGCloud [13]	61.3	88.1	83.9	66.0	86.0	40.5	91.1	30.9	27.5	64.3
	ShellNet [19]	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
	KPConv [14]	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	77.3	79.7
	RandLA-Net [7]	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
	PointGCR [9]	69.5	92.1	93.8	80.0	64.4	66.4	93.2	39.2	34.3	85.3
	RFCR [5]	77.8	95.0	94.2	89.1	85.7	54.4	95.0	43.8	76.2	83.7
	HybridCR	77.4	95.0	97.3	84.1	87.7	58.2	95.2	48.2	67.5	81.0
weakly	PSD(1%) [18]	75.8	94.3	97.1	91.0	86.7	48.1	95.1	46.5	63.2	79.0
	HybridCR(1%)	76.8	94.9	97.8	94.0	86.6	52.9	95.3	47.1	64.9	75.5

Table 3. Quantitative results of per class on Semantic3D (reduced-8) [6]. (mIoU %, OA %)

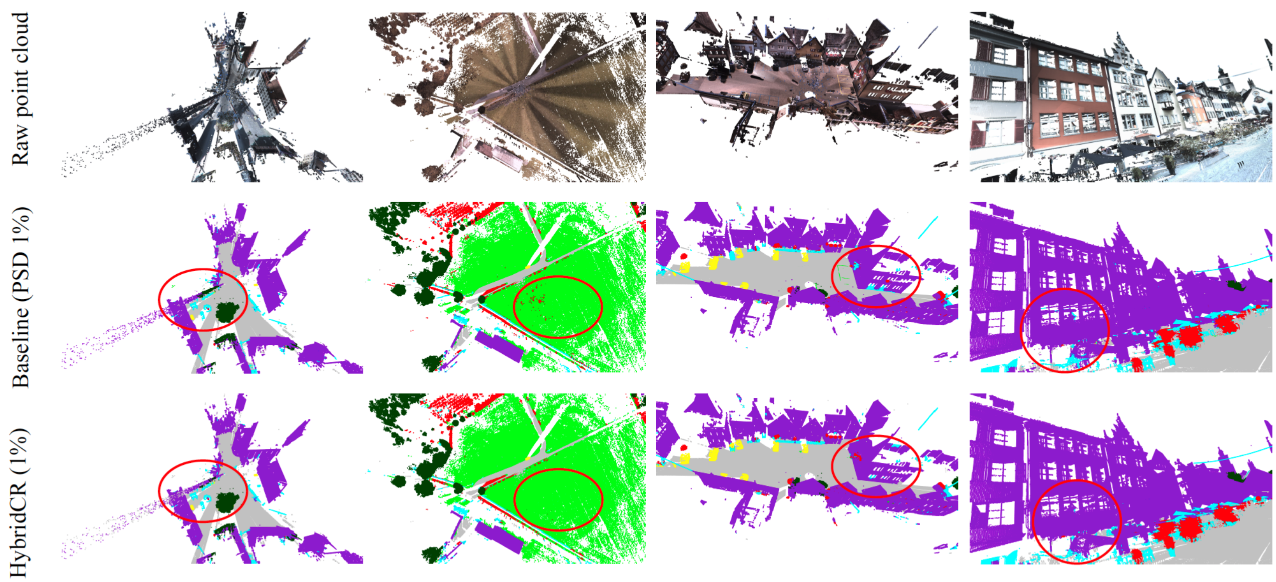

Figure 2. Visualization results on the test set of Semantic3D. Raw point cloud, results of the baseline and ours are presented separately from top to bottom.

the fully supervised DarkNet53Seg and RangeNet53++, our HybridCR is 2.4% 0.1% higher in mIoU, respectively. Besides, we achieve the best performance in the "vegetation", "trunk" and "bycicle" classes. Therefore, the results demonstrate that HybridCR has reliable performance on the outdoor dataset.

1.4. Quantitative Results

Visualization on ScanNet-V2. In Fig. 1, we show visualization results on the test set of ScanNet-V2. Since there is no public ground truth, we show the raw point clouds at the top row and our segmentation results at the bottom row. It can be observed that HybridCR can achieve good segmentation results for most classes. At the 1% setting, the segmentation precision of small corners and boundaries *e.g.*, " wall" and "door" area compared to PSD, is further improved.

Visualization on Semantic3D. Fig. 2 shows the visualization results on the test set of Semantic3D. Since there is no public ground truth, we show the raw point cloud at the top row and our segmentation results at the bottom row. In general, it can be seen that HybridCR achieves good qualitative segmentation results at 1% setting. Our method can also make more accurate predictions for some categories (e.g., "low-veg.", "buildings" and "man-made") with a small number of points.

Visualization on SemanticKITTI. Fig 3 shows more qualitative results of HybridCR on the validation split. It can be seen that our method achieves consistent segmentation results to ground-truth, especially in "road" and "car", which are difficult to distinguish while critical on sparse outdoor scenes in the auto-driving application.

1.5. Dynamic *vs*. fixed augmentor of multiple runs.

In Tab. 5, we report 1pt and 1% results with mean and std.dev. (5 runs) on S3DIS Area-5, as well as dynamic and

Set.	Methods	**mIoU(%)**	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
Fully	PointNet [11]	14.6	61.6	35.7	15.8	1.4	41.4	46.3	0.1	1.3	0.3	0.8	31.0	4.6	17.6	0.2	0.2	0.0	12.9	2.4	3.7
	SqueezeSegV2 [15]	39.7	88.6	67.6	45.8	17.7	73.7	81.8	13.4	18.5	17.9	14.0	71.8	35.8	60.2	20.1	25.1	3.9	41.1	20.2	36.3
	DarkNet53Seg [2]	49.9	91.8	74.6	64.8	27.9	84.1	86.4	25.5	24.5	32.7	22.6	78.3	50.1	64.0	36.2	33.6	4.7	55.0	38.9	52.2
	RangeNet53++ [10]	52.2	91.8	75.2	65.0	27.8	87.4	91.4	25.7	25.7	34.4	23.0	80.5	55.1	64.6	38.3	38.8	4.8	58.6	47.9	55.9
	RandLA-Net [7]	53.9	90.7	73.7	60.3	20.4	86.9	94.2	40.1	26.0	25.8	38.9	81.4	61.3	66.8	49.2	48.2	7.2	56.3	49.2	47.7
	HybridCR	54.0	90.5	73.9	59.1	21.2	88.3	93.9	42.7	22.8	31.6	36.8	81.7	61.7	66.1	50.2	45.5	4.5	57.4	49.5	49.0
weakly	HybridCR(1%)	52.3	89.4	72.9	61.5	20.6	85.8	92.7	30.2	27.3	27.7	23.6	83.2	64.5	69.3	50.1	45.8	3.9	55.2	41.8	48.2

Table 4. Quantitative results of per class on SemanticKITTI [2]. (mIoU %)

Method	#1	#2	#3	#4	#5	#6	#7	#8
1pt	48.2±(0.3)	50.7±(0.3)	49.8±(0.5)	50.2±(0.2)	51.1±(0.2)	50.8±(0.1)	51.0±(0.3)	51.5±(0.2)
Dynamic(1pt)	-	50.7±(0.3)	-	-	-	50.8±(0.1)	51.1±(0.3)	51.5±(0.2)
Fix(1pt)	-	47.2±(0.4)	-	-	-	47.7±(0.3)	48.0±(0.2)	48.3±(0.3)
1%	63.5±(0.1)	64.5±(0.3)	63.9±(0.4)	64.0±(0.2)	65.0±(0.3)	64.7±(0.4)	65.1±(0.2)	65.3±(0.3)
Dynamic(1%)	-	64.5±(0.3)	-	-	-	64.7±(0.4)	65.1±(0.2)	65.3±(0.3)
Fix(1%)	-	59.8±(0.1)	-	-	-	61.8±(0.2)	61.4±(0.3)	62.6±(0.1)

Table 5. Dynamic *vs.* fixed augmentor on S3DIS Area-5 in 1pt and 1%. #1-#8 are the ablation settings in Tab. 3 of the main paper.

fixed augmentors. Note that #1-#8 are the ablation settings in Tab. 3 of the main paper. For 1pt and 1% setting, it can be found that the dynamic augmentor outperform the fixed one by 3.2% and 2.7% mIoU at the ablation setting #8, respectively.

References

[1] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *TOG*, 37(4):1–12, 2018. 2

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 1, 2, 4

[3] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. *3DOR@ Eurographics*, 3, 2017. 2, 3

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 2

[5] Jingyu Gong, Jiachen Xu, Xin Tan, Haichuan Song, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *CVPR*, pages 11673–11682, 2021. 2, 3

[6] T Hackel, N Savinov, L Ladicky, JD Wegner, K Schindler, and M Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:91, 2017. 1, 3

[7] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, pages 11108–11117, 2020. 2, 3, 4

[8] Huan Lei, Naveed Akhtar, and Ajmal Mian. Seggcn: Efficient 3d point cloud segmentation with fuzzy spherical kernel. In *CVPR*, pages 11611–11620, 2020. 2

[9] Yanni Ma, Yulan Guo, Hao Liu, Yinjie Lei, and Gongjian Wen. Global context reasoning for semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2931–2940, 2020. 2, 3

[10] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019. 2, 4

[11] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 2, 4

[12] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 2

[13] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *3DV*, pages 537–547. IEEE, 2017. 2, 3

[14] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, pages 6411–6420, 2019. 2, 3

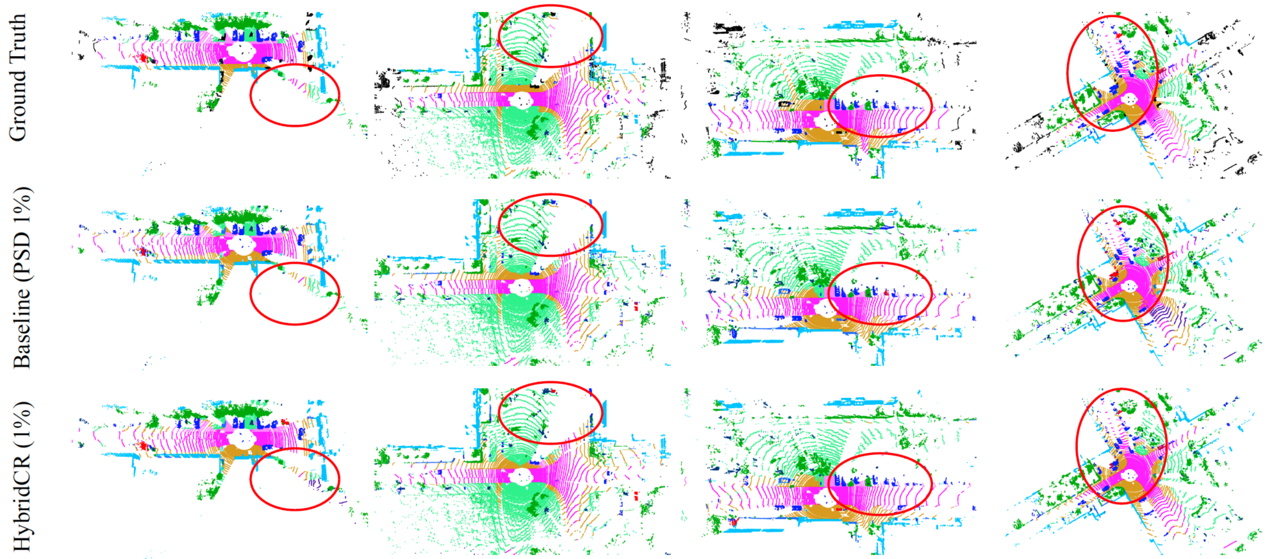

Figure 3. Visualization results on the validation set of SemanticKITTI. Ground truth, results of the baseline and ours are presented separately from top to bottom.

[15] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, pages 4376–4382. IEEE, 2019. 2, 4

[16] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, pages 9621–9630, 2019. 2

[17] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *CVPR*, pages 13706–13715, 2020. 1

[18] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *ICCV*, pages 15520–15528, 2021. 1, 2, 3

[19] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In *ICCV*, pages 1607–1616, 2019. 2, 3

[20] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. 1