Supplementary Material for "ReDAL: Region-based and Diversity-aware Active Learning for Point Cloud Semantic Segmentation"

Tsung-Han Wu[1] Yueh-Cheng Liu[1†] Yu-Kai Huang[1†]
Hsin-Ying Lee[1] Hung-Ting Su[1] Ping-Chia Huang[1] Winston H. Hsu[1,2]

[1]National Taiwan University [2]Mobile Drive Technology

Abstract

The supplementary material is organized as follows: Section 1 describes the implementation details. Section 2 explains the baseline active learning methods. Section 3 shows the original data of line charts or tables in the main paper.

1. Implementation Details

As explained in the main paper, the pipeline of our ReDAL contains four steps: (1) Train the deep learning model in supervision with labeled dataset D_L. (2) Calculate region information score using softmax entropy, color discontinuity, and structure complexity. (3) Diversity-aware selection by penalizing visually similar regions appearing in the same querying batch. (4) The top-ranked regions are labeled by annotators and added to the labeled dataset D_L. This section explains the implementation details of the first three steps, and the fourth step has been explained clearly in the main paper. Note that the following symbols are the same as those in Section 3 of the main paper.

1.1. Network Training

For both S3DIS [1] and SemanticKITTI [2] datasets, the networks are trained with Adam optimizer (initial learning rate = 0.001) and cross-entropy loss. We train the network on 8 V100 GPUs and set the batch size to 16. We set voxel resolution to 5cm for both datasets.

On the S3DIS dataset, the deep learning model was trained for 200 epochs on 3% of the initial fully labeled point cloud scan and then fine-tuned for 150 epochs after adding 2% labeled data each time for both network architecture backbones. On the SemanticKITTI dataset, the deep learning model was trained for 100 epochs on 1% of the initial fully labeled point cloud scan and then fine-tuned for 30 epochs after adding 1% labeled data each time for both network architectures.

[†]Co-second authors contribute equally.

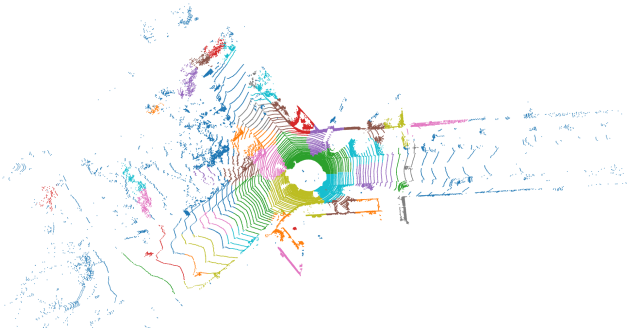

Figure 1. **Visualization of divided sub-scene regions in SemanticKITTI dataset.** Points of the same color in neighboring places belong to the same region.

1.2. Region Information Estimation

We utilize the VCCS algorithm [7] to divide a 3D scene into multiple sub-scene regions. In the algorithm, the whole 3D space is initially divided into multiple regions with two hyper-parameters R_{seed}, R_{voxel}, where R_{seed} indicates the initial distance between regions and R_{voxel} represents the minimal region resolution. After that, the clustering procedure adjusts the region boundary based on spatial or color connectivity iteratively. For the S3DIS dataset, we set R_{seed}, R_{voxel} to a small value ($R_{seed} = 1.0, R_{voxel} = 0.1$) since objects in an indoor scene are small. For the SemanticKITTI dataset, we set R_{seed}, R_{voxel} to a large value ($R_{seed} = 10, R_{voxel} = 0.5$). The reason is that the point cloud is sparse in outdoor 3D space, and choosing larger parameters (R_{seed}, R_{voxel}) can avoid creating small, unrepresentative regions. An example of divided sub-scene regions of the SemanticKITTI dataset is shown in Figure 1.

As mentioned in Section 3.2 of the main paper, we linearly combine softmax entropy, color discontinuity, and structural complexity as region information score. For color discontinuity and structural complexity, we calculate

color differences and surface variation for each point and its k-nearest neighbors ($k = 50$ in both datasets). As for the weight of the linear combination of these three terms, which is described in Eq. 4 of the main paper, we set $\alpha = 1, \beta = 0.1, \gamma = 0.05$ for S3DIS dataset and $\alpha = 1, \beta = 0, \gamma = 0.05$ for SemanticKITTI dataset. Note that the value $\alpha = 1.0, \beta = 0.1, \gamma = 0.05$ is empirically decided for we found that model uncertainty is much more important than the color discontinuity and structural complexity terms. In addition, since the SemanticKITTI dataset does not have point-by-point color information, we set $\beta = 0$ for the dataset.

1.3. Diversity-aware Selection

As explained in Section 3.3 of the main paper, we measure the similarity of these regions by clustering their corresponding region features. We set the number of clusters of all regions $M = 400, 150$ for the S3DIS and SemanticKITTI datasets, respectively. For both datasets, we set the decay rate $\eta = 0.95$. Note that our diversity-aware selection algorithm does not create too much computational burden. On the SemanticKITTI dataset, our diversity-aware selection algorithm only takes only 0.58 ms per region on average.

Note that we empirically found that k in k-nn (mentioned in the previous sub-section), decay rate η and the number of clusters M is not sensitive to the experimental results, where all values are determined via grid search.

2. Baseline Active Learning Methods

In this section, we describe the implementation of the baseline active learning methods used in our experiments.

Random selection (RAND) Randomly select a portion of point cloud scans in the unlabeled dataset for label acquisition. The strategy is commonly used as the baseline for active learning methods [11, 5, 8, 6].

Margin sampling (MAR) Some previous active learning methods query instances with the smallest model decision margin, which is the predicted probability difference between the two most likely class labels [11]. As shown in Eq. 1, given a point cloud scan X with N points and fixed model parameter θ, we calculate the difference between the two most likely class labels for all points and produce the score for a point cloud scan (S_{MAR}) by averaging the value of all points in a scan. After that, we select a portion of point cloud scans with the largest score in the unlabeled dataset for label acquisition.

$$S_{MAR} = \frac{1}{N}\sum_{n=1}^{N} P(\hat{y_n^1}|X;\theta) - P(\hat{y_n^2}|X;\theta), \quad (1)$$

where $\hat{y_n^1}$ is the first most probable label class and $\hat{y_n^2}$ is the second most probable label class.

Least confidence sampling (CONF) Many previous active learning methods query the sample whose prediction has the least confidence [11, 12]. As can be observed in Eq. 2, given a point cloud scan X with N points and fixed model parameter θ, we calculate the confidence of predicted class label ($\hat{y_n^1}$) for all points and produce the score for a point cloud scan (S_{CONF}) by averaging the value of all points in a scan. After that, we select a portion of point cloud scans with the least confidence score in the unlabeled dataset for label acquisition.

$$S_{CONF} = \frac{1}{N}\sum_{n=1}^{N} P(\hat{y_n^1}|X;\theta) \quad (2)$$

Softmax entropy (ENT) Entropy is an indicator to measure the information of a probability distribution in the information theory [9]. Some previous active learning approaches query samples with the highest entropy value in the predicted probability [11]. As shown in Eq. 3, given a point cloud scan X with N points and fixed model parameter θ, we calculate the softmax entropy value for all points and produce the score for a point cloud scan (S_{ENT}) by averaging the value of all points in a scan. After that, we select a portion of point cloud scans with the largest entropy in the unlabeled dataset for label acquisition.

$$S_{ENT} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{c} P(y_n^i|X;\theta)\log P(y_n^i|X;\theta), \quad (3)$$

where c represents the total number of labels, and $P(y_n^i|X;\theta)$ represents the probability that the model predicts point n as class i.

Core-Set (CSET) Sener *et al.* [8] proposed a purely diversity-based deep active selection strategy named Core-Set. The strategy aims to select a small subset so that a model trained on the selected subset has a similar performance to that trained on the whole dataset. The method first extracts the feature of each sample. Then, it selects a small number of samples from the unlabeled dataset that is the furthest away from the labeled dataset in the feature space for label acquisition. In the implementation, we choose the middle layer of the encoder-decoder network as the feature.

Segment entropy (SEGENT) Lin *et al.* [6] proposed *segment entropy* to measure the point cloud information in the deep active learning pipeline. This method assumes that

each geometrically related area should share similar semantic annotations. Therefore, it calculates the entropy of the distribution of predicted labels in a small area to estimate model uncertainty.

MC-Dropout (MCDR) [4, 5] combined Bayesian active learning with deep learning, which estimated model uncertainty by Monte Carlo Dropout. In the implementation, we set the dropout rate to 0.3 and perform 10 dropout predictions. Note that since there is no dropout layer in MinkowskiNet [3], we did not compare with this baseline when using MinkowskiNet.

3. Experimental Result

Due to space limitations, we show the original experimental results here, which are shown in the line charts of the main paper. Table 1, 2, 3, 4 shows the original data of Figure 5 in the main paper. Table 5, 6 present the original data of Table 1, 2 in the main paper.

% Labeled Data	RAND	MAR	CONF	ENT	CSET	SEGENT	MCDR	ReDAL (Ours)
init.	27.05	28.29	28.60	27.92	28.89	29.16	28.33	27.86
5	31.39	30.07	32.14	31.02	33.24	34.55	29.30	**41.27**
7	35.37	31.34	33.76	35.10	36.59	40.97	33.68	**47.68**
9	40.51	33.30	38.57	40.90	37.02	42.30	40.00	**52.34**
11	44.50	39.75	40.60	41.51	41.42	43.07	41.65	**54.28**
13	46.28	40.41	42.43	43.42	41.34	44.48	44.04	**57.01**
15	49.02	40.45	44.44	45.06	41.40	45.04	45.06	**57.97**

Table 1. Results of IoU performance (%) on S3DIS [1] with SPVCNN [10].

% Labeled Data	RAND	MAR	CONF	ENT	CSET	SEGENT	ReDAL (Ours)
init.	26.59	25.20	25.52	26.60	25.60	26.30	25.63
5	30.22	25.87	27.81	27.60	35.58	26.66	**39.45**
7	34.76	32.40	30.25	28.91	38.88	30.45	**44.29**
9	38.79	36.20	32.23	35.40	40.41	39.72	**50.50**
11	43.80	41.31	38.39	37.10	41.28	41.95	**55.11**
13	46.13	42.28	42.10	37.42	43.63	44.66	**56.14**
15	48.57	43.15	42.18	40.37	47.26	45.79	**57.26**

Table 2. Results of IoU performance (%) on S3DIS [1] with MinkowskiNet [3].

% Labeled Data	RAND	MAR	CONF	ENT	CSET	SEGENT	MCDR	ReDAL (Ours)
init.	41.84	42.39	42.98	41.90	42.19	43.18	42.92	41.87
2	45.41	46.84	46.31	45.57	46.98	47.89	47.57	**51.70**
3	52.19	49.55	50.15	51.42	52.93	52.60	50.08	**55.83**
4	54.76	51.66	54.46	51.85	54.57	53.60	53.56	**56.86**
5	56.89	53.21	55.41	56.45	56.45	54.00	54.40	**58.18**

Table 3. Results of IoU performance (%) on SemanticKITTI [2] with SPVCNN [10].

% Labeled Data	RAND	MAR	CONF	ENT	CSET	SEGENT	ReDAL (Ours)
init.	37.74	38.20	37.32	37.33	36.86	37.75	37.48
2	42.74	42.73	42.01	42.16	41.25	42.62	**48.88**
3	48.82	45.07	47.37	45.77	45.15	49.51	**55.30**
4	52.51	47.84	49.54	49.46	49.93	51.87	**58.35**
5	54.67	51.27	53.49	52.34	51.89	53.12	**59.76**

Table 4. Results of IoU performance (%) on SemanticKITTI [2] with MinkowskiNet [3].

method	mIoU	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Full	61.4	95.9	20.4	63.9	70.3	45.5	65.0	78.5	0.4	93.5	50.6	82.0	0.2	91.2	63.8	87.2	68.5	74.3	64.4	50.1
RAND	54.7	94.7	9.5	45.0	**66.8**	38.6	52.0	47.8	0.0	90.2	38.5	76.1	**1.8**	88.3	**55.5**	**87.9**	**64.0**	**76.5**	60.2	45.6
ReDAL	**59.8**	**95.4**	**29.6**	**58.6**	63.4	**49.8**	**63.4**	**84.1**	**0.5**	**91.5**	**39.3**	**78.4**	1.2	**89.3**	54.4	87.4	62.0	74.1	**63.5**	**49.7**

Table 5. **Results of IoU performance (%) with only 5% labeled points.** The table shows that our ReDAL achieve better results on most classes compared with baseline random selection. For some classes of small items and objects with complex boundaries, our ReDAL greatly surpass the random selection baseline and even outperform fully supervised result, such as bicycle and bicyclist.

method	total	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Full	10^3	43.68	0.17	0.41	2.02	2.40	0.36	0.13	0.04	205.22	15.19	148.59	4.03	137.00	74.69	275.57	6.23	80.67	2.95	0.63
RAND	10^3	43.89	0.14	0.34	3.51	2.12	0.42	0.11	0.05	206.86	14.07	147.32	4.02	137.63	74.47	274.47	6.21	80.54	3.02	0.73
ReDAL	10^3	33.71	0.25	0.51	8.01	11.36	1.27	0.21	0.07	168.16	20.15	145.77	16.92	132.22	78.68	252.65	9.25	114.45	4.48	1.87

Table 6. **Labeled Class Distribution Ratio (‰).** With limited annotation budgets, our active method ReDAL queries more labels on small objects like person and bicycle but less on large uniform areas like road and vegetation. The selection strategy can mitigate the label imbalance problem and improve the performance on more complicated object scenes without hurting much on large areas as shown in Table 5.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 1, 4

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1, 4

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 3, 4

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 3

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192, 2017. 2, 3

[6] Y Lin, G Vosselman, Y Cao, and MY Yang. Efficient training of semantic point cloud segmentation via active learning. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2:243–250, 2020. 2

[7] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2013. 1

[8] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 2

[9] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

[10] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702. Springer, 2020. 4

[11] D. Wang and Y. Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, 2014. 2

[12] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2