Supplementary Material for Bayesian Self-Training for Semi-Supervised 3D Segmentation

Ozan Unal[1,2], Christos Sakaridis[1], and Luc Van Gool[1,3,4]

[1]ETH Zurich, [2]Huawei Technologies, [3]KU Leuven, [4]INSAIT
{ozan.unal, csakarid, vangool}@vision.ee.ethz.ch

1 Implementation Details and Dataset

1.1 3D Semantic Segmentation

Implementation details: For 3D semantic segmentation, we use Cylinder3D [14] in our experiments without modification for a fair comparison to existing work. We use a labeling percentage of $p_\tau = 0.75$ and only require a single iteration of self-training. We use K=9 passes.
Dataset: We run our experiments on the SemanticKITTI [2] which is the most popular dataset for LiDAR semantic segmentation. SemanticKITTI consists of 11 sequences, with sequence 8 reserved for validation. We further show results on ScribbleKITTI [11], a realistically weakly labeled dataset built on top of SemanticKITTI consisting of only 8% labeled points. Following precedent we uniformly sample the data at varying thresholds [8] and report the mIoU on the *val*-set.

1.2 3D Instance Segmentation

Implementation details: For 3D instance segmentation, we use the same backbone Sparse U-Net [7, 13] in our experiments. We use a labeling percentage of $p_\tau = 0.75$ for instance masks and again only require a single iteration of self-training. We use K=9 passes.
Dataset: We evaluate our method on the indoor 3D datasets ScanNet [6] and S3DIS [1]. ScanNet consists of 1613 scenes, of which 312 are reserved for validation. For the training splits, we follow the limited reconstruction setting from Hou *et al.* [7] and report the mAP and AP at threshold 50% and 25% on the *val*-set. S3DIS is a much smaller dataset consisting of only 271 scenes collected from 6 different areas, of which we reserve Area 5 as the validation set following precedent [5]. Due to its small size, we tackle the semi-supervised setting for labeling splits ranging from 5% up to 50%.

1.3 Dense 3D Visual Grounding

Implementation details: For 3D visual grounding, we use the current state-of-the-art ConcreteNet [12] in our experiments. We use $p_\tau = 75\%$ for the grounding

Table 1: Classwise results for semi-supervised 3D semantic segmentation on SemanticKITTI [2].

	Method	mIoU	car	bicycle	m.cycle	truck	o.vehicle	person	bicyclist	m.cyclist	road	parking	sidewalk	o.ground	building	fence	vegetation	trunk	terrain	pole	t.sign
1%	*Sup-only* [14]	45.4	90.9	24.5	2.8	35.1	20.4	31.7	49.5	0.0	85.5	23.4	67.5	**1.3**	85.0	46.0	84.1	49.1	70.3	55.0	40.6
	DiAL [10]	45.4	91.2	13.2	5.4	47.3	14.5	29.0	37.3	0.0	86.8	22.6	70.3	1.2	86.7	45.4	84.7	59.4	70.9	55.8	40.8
	CBST [15]	48.8	92.4	16.3	6.4	**61.9**	**27.0**	35.7	49.4	0.0	88.9	29.4	73.2	0.7	**89.1**	49.5	83.9	51.4	68.1	59.8	44.0
	CBST [4]	46.7	92.0	13.5	7.1	37.8	12.7	33.0	**54.5**	0.0	89.8	25.0	73.8	0.0	88.8	**50.1**	83.6	57.4	67.8	58.2	42.1
	LaserMix [8]	**50.6**	91.8	**35.7**	19.8	37.5	25.6	**53.6**	45.7	**2.5**	87.8	**33.5**	71.3	0.7	87.3	43.8	84.6	**62.7**	69.3	59.8	**47.6**
	Ours	49.8	**93.1**	21.3	**24.8**	44.7	20.8	48.3	38.3	0.0	**90.3**	29.5	**73.9**	0.0	85.0	42.5	**88.2**	59.5	**75.1**	**64.6**	46.5
10%	*Sup-only* [14]	56.1	93.4	38.4	47.7	65.7	31.0	61.9	64.9	0.0	90.7	37.7	75.3	0.9	89.2	50.5	86.4	56.0	73.9	56.2	46.0
	DiAL [10]	57.1	94.1	40.5	**58.4**	56.0	38.0	**66.5**	75.6	0.0	88.4	22.7	72.0	1.5	87.9	49.3	86.7	66.1	74.2	58.0	49.2
	CBST [15]	58.3	93.6	40.3	43.5	80.4	33.8	57.6	78.1	0.0	91.6	36.3	76.6	**5.1**	89.2	51.1	86.3	61.9	71.2	61.3	49.7
	CBST [4]	58.7	94.0	38.7	51.0	60.3	39.8	65.7	80.0	0.0	91.4	33.2	76.4	2.9	89.8	53.8	87.2	65.7	**74.6**	61.5	50.0
	LaserMix [8]	60.0	93.8	44.9	**58.4**	65.6	39.4	65.8	80.9	**0.2**	92.0	**44.2**	77.1	3.9	89.1	49.0	86.2	**66.8**	72.3	58.4	51.2
	Ours	**61.7**	**95.4**	**47.3**	53.8	**82.3**	**43.2**	63.6	**81.2**	0.0	**93.8**	43.2	**79.8**	1.4	**89.9**	**54.6**	**87.3**	66.1	73.3	**63.9**	**51.8**
20%	*Sup-only* [14]	57.8	94.0	31.6	47.3	**89.5**	38.3	57.9	79.1	0.0	91.6	29.6	76.1	0.9	87.8	43.6	86.6	63.7	72.5	61.8	47.5
	DiAL [10]	59.2	94.4	38.7	52.5	81.2	45.8	64.2	78.0	0.0	90.9	35.2	75.7	1.8	89.2	49.8	86.3	65.6	72.6	56.0	47.6
	CBST [15]	59.4	94.2	41.8	51.4	77.7	39.8	65.4	79.8	0.0	91.7	29.8	76.3	3.5	89.2	49.7	87.1	66.1	**74.2**	60.1	51.3
	CBST [4]	59.6	94.2	41.8	52.9	78.2	39.6	66.1	80.6	0.0	91.9	30.2	76.4	**3.7**	89.2	50.0	87.0	66.6	73.7	60.0	51.1
	LaserMix [8]	61.9	94.4	46.0	**68.0**	74.3	47.6	68.1	83.7	0.2	92.6	**42.7**	78.0	1.9	89.7	52.9	86.0	**69.3**	70.6	59.2	51.7
	Ours	**63.7**	**96.4**	**50.7**	61.6	79.2	**55.4**	**73.3**	**85.3**	**1.1**	**93.8**	40.6	**80.1**	2.1	**89.9**	**58.0**	**87.5**	66.4	72.5	**63.4**	**52.9**
50%	*Sup-only* [14]	58.7	93.9	40.4	48.0	81.4	33.7	65.7	79.7	0.0	91.9	32.6	76.7	1.3	89.0	51.8	87.2	61.4	72.5	58.7	48.7
	DiAL [10]	60.0	94.1	41.3	57.7	64.6	39.5	65.3	86.8	0.0	91.3	32.8	75.2	3.5	89.7	48.6	85.4	65.9	70.6	58.7	49.1
	CBST [15]	59.7	94.9	40.9	54.4	75.3	43.8	67.3	86.8	0.0	91.5	33.3	75.7	2.6	89.3	50.7	86.7	63.9	72.4	56.4	48.8
	CBST [4]	60.5	94.6	43.3	55.3	80.5	42.5	67.9	84.6	0.0	92.0	34.3	76.9	2.2	89.8	52.3	86.0	67.4	71.1	59.5	49.4
	LaserMix [8]	62.3	94.7	48.4	**64.7**	65.2	44.5	71.0	**88.3**	**2.1**	92.7	**43.0**	78.4	2.0	**90.3**	**54.9**	**88.1**	**68.1**	**75.3**	**66.6**	51.7
	Ours	**64.1**	**96.4**	**51.1**	64.3	**84.6**	**57.5**	**70.8**	83.9	0.0	**93.4**	41.5	**79.6**	**6.3**	88.8	53.8	87.9	67.8	73.8	64.3	**52.7**

pseudo-labels when including verbal cues in the unlabeled set and only require a single iteration of self-training. We use K=9 passes.

Dataset: We evaluate our method on the ScanRefer [3] dataset using axis-aligned bounding boxes that are fitted onto the predicted instance masks. ScanRefer builds on top of ScanNet but only consists of a single furniture arrangement per room. We construct three semi-supervised splits of 5%, 10%, and 20% where we ensure the data split is valid for both the number of 3D scenes and the number of available utterances.

1.4 Classwise 3D Semantic Segmentation Results

In reference to the Tab. 1 on the main manuscript, we provide Tab. 1 and Tab. 2 that show the classwise IoUs for the same semi-supervised 3D semantic segmentation experiments on SemanticKITTI and ScribbleKITTI respectively (apart from [9]). Here we observe two key feats of our Bayesian self-training method: (i) Our method outperforms existing work in head classes that are more prone to the softmax overconfidence issue; (ii) our method shows great improvements over the baseline on classes that have similar 3D shapes (e.g. car, truck, other vehicle). The inclusion of uncertainty-based filtering allows the network to reduce the number of false positives, allowing better separation between geometrically similar objects.

Table 2: Classwise results for semi-supervised 3D semantic segmentation on ScribbleKITTI [11].

	Method	mIoU	car	bicycle	m.cycle	truck	o.vehicle	person	bicyclist	m.cyclist	road	parking	sidewalk	o.ground	building	fence	vegetation	trunk	terrain	pole	t.sign
1%	*Sup-only* [14]	39.2	83.2	13.8	3.4	26.3	11.8	28.0	25.2	0.0	72.5	13.0	59.5	0.2	86.6	33.7	78.7	55.7	58.4	54.0	40.3
	DiAL [10]	41.0	82.3	15.8	7.1	**32.0**	15.4	23.7	36.3	0.0	75.0	12.6	61.4	**0.9**	85.3	30.0	80.1	57.0	67.0	56.1	41.3
	CBST [15]	41.5	83.7	22.1	5.9	28.3	13.4	27.1	34.7	0.0	74.0	14.4	61.7	0.2	**88.1**	36.6	80.3	58.7	60.4	57.1	41.4
	CBST [4]	41.4	82.8	18.2	11.4	20.9	15.1	22.5	35.5	0.0	74.7	**15.7**	61.6	0.4	86.0	34.2	82.2	58.4	69.9	56.7	40.0
	LaserMix [8]	44.2	82.6	**25.5**	18.8	29.0	**19.8**	41.1	**47.2**	**0.6**	71.5	10.5	64.2	2.2	85.1	33.5	82.0	59.9	65.8	54.5	45.2
	Ours	**43.9**	**85.7**	13.8	**23.6**	7.9	7.6	**48.8**	30.2	0.0	**81.3**	**15.7**	**68.8**	0.2	84.2	**40.0**	**85.7**	**60.3**	**71.2**	**63.9**	**45.3**
10%	*Sup-only* [14]	48.0	85.7	25.6	21.3	52.8	29.9	46.5	47.2	0.1	79.5	15.4	63.8	0.3	85.4	39.6	84.8	59.7	71.5	57.7	45.8
	DiAL [10]	50.1	83.7	32.6	45.1	41.0	34.7	56.0	59.2	0.0	75.9	14.0	64.0	0.7	85.6	37.9	83.3	62.6	68.2	59.7	47.0
	CBST [15]	50.6	85.8	31.4	30.5	58.5	24.4	55.1	58.8	0.0	82.6	15.3	67.8	0.5	87.7	40.0	82.8	62.5	65.0	62.0	50.8
	CBST [4]	51.8	84.6	34.9	47.1	37.5	29.5	60.1	69.1	0.0	79.8	16.5	67.3	**2.7**	88.0	39.2	84.5	64.5	71.0	60.4	47.9
	LaserMix [8]	53.7	85.8	34.7	45.6	**54.9**	35.8	63.2	73.6	**1.3**	79.8	25.0	68.2	1.8	87.7	35.4	84.0	65.8	70.8	59.4	48.2
	Ours	**58.9**	**89.9**	**41.4**	**60.6**	51.0	**50.4**	**66.7**	**76.8**	0.0	**86.7**	**29.4**	**74.7**	1.6	**89.4**	**53.7**	**88.3**	**68.5**	**75.2**	**63.3**	**51.2**
20%	*Sup-only* [14]	52.1	86.9	38.0	39.5	67.3	29.7	56.5	69.9	0.0	79.0	16.0	66.0	0.3	87.0	38.6	84.3	60.6	66.2	58.8	45.2
	DiAL [10]	52.8	85.9	27.9	41.5	55.5	33.0	64.1	72.0	**1.2**	81.0	22.5	67.8	1.2	89.1	39.9	82.9	63.7	66.9	60.5	46.7
	CBST [15]	53.3	86.6	36.8	40.9	72.9	28.3	58.0	69.5	0.0	81.1	18.3	68.2	0.7	88.7	44.3	83.6	63.3	64.4	60.3	47.5
	CBST [4]	53.9	85.4	37.2	44.7	58.9	32.9	63.5	71.0	0.0	81.6	23.1	69.2	1.9	88.4	38.2	83.8	65.7	69.2	60.2	48.9
	LaserMix [8]	55.1	88.0	38.8	51.3	54.8	36.6	60.2	73.9	0.0	78.8	22.7	71.9	1.5	**90.3**	43.3	85.3	66.5	**70.9**	60.3	**51.6**
	Ours	**60.1**	**90.7**	**45.1**	**56.7**	**76.1**	**44.2**	**71.1**	**80.9**	0.0	**88.7**	**30.5**	**76.6**	**3.8**	89.5	**53.7**	**84.6**	**69.1**	65.7	**65.0**	**51.6**
50%	*Sup-only* [14]	53.8	87.5	37.2	41.3	**71.4**	29.6	58.8	80.4	0.0	81.1	16.7	67.5	0.4	88.4	39.4	83.1	64.4	65.5	61.8	47.5
	DiAL [10]	53.9	86.9	33.6	46.2	48.9	33.2	62.8	77.7	0.0	82.7	22.8	68.6	3.2	89.2	38.6	83.8	66.4	68.0	62.3	48.5
	CBST [15]	54.5	87.6	39.5	36.7	65.9	35.7	62.8	78.1	0.0	82.4	20.4	69.6	0.1	88.8	42.3	84.2	64.0	67.4	60.1	50.1
	CBST [4]	54.8	85.1	35.2	45.2	68.6	32.0	65.7	77.9	0.2	81.2	21.7	69.0	1.6	89.2	40.2	84.5	65.1	70.1	60.9	48.5
	LaserMix [8]	56.8	88.0	40.8	51.6	63.1	38.4	61.7	79.9	**2.0**	83.1	26.1	71.2	2.8	**90.1**	41.7	85.9	**69.5**	70.5	63.0	51.6
	Ours	**61.2**	**95.1**	**45.2**	**57.1**	62.7	**45.5**	**71.2**	**82.1**	0.0	**93.1**	**42.3**	**78.4**	**4.3**	89.6	**52.4**	**88.1**	63.3	**74.3**	**64.7**	**53.7**

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019)
3. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: ECCV (2020)
4. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
5. Chu, R., Ye, X., Liu, Z., Tan, X., Qi, X., Fu, C.W., Jia, J.: TWIST: Two-way inter-label self-training for semi-supervised 3D instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: ICCV (2017)
7. Hou, J., Graham, B., Niessner, M., Xie, S.: Exploring data-efficient 3D scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
8. Kong, L., Ren, J., Pan, L., Liu, Z.: LaserMix for semi-supervised LiDAR semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
9. Unal, O., Dai, D., Hoyer, L., Can, Y.B., Van Gool, L.: 2d feature distillation for weakly- and semi-supervised 3d semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7336–7345 (January 2024)
10. Unal, O., Dai, D., Unal, A.T., Van Gool, L.: Discwise active learning for LiDAR semantic segmentation. IEEE Robotics and Automation Letters (2023)
11. Unal, O., Dai, D., Van Gool, L.: Scribble-supervised LiDAR semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
12. Unal, O., Sakaridis, C., Saha, S., Gool, L.V.: Four ways to improve verbo-visual fusion for dense 3d visual grounding (2024), `https://arxiv.org/abs/2309.04561`
13. Wu, Y., Shi, M., Du, S., Lu, H., Cao, Z., Zhong, W.: 3d instances as 1d kernels. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX. pp. 235–252. Springer (2022)
14. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
15. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: The European Conference on Computer Vision (ECCV) (2018)