SEMI-SUPERVISED LIDAR SEMANTIC SEGMENTATION WITH SPATIAL CONSISTENCY TRAINING

Lingdong Kong[¶], **Jiawei Ren**[¶], **Liang Pan & Ziwei Liu**[✉]
S-Lab, Nanyang Technological University, Singapore
{lingdong001,jiawei011}@e.ntu.edu.sg {liang.pan,ziwei.liu}@ntu.edu.sg

ABSTRACT

We study the underexplored semi-supervised learning (SSL) in LiDAR semantic segmentation, as annotating LiDAR point clouds is expensive and hinders the scalability of fully-supervised methods. Our core idea is to leverage the strong spatial cues of LiDAR point clouds to better exploit unlabeled data. We propose LaserMix to mix laser beams from different LiDAR scans and encourage the model to make consistent and confident predictions before and after mixing. Our framework has three appealing properties. 1) Generic: LaserMix is agnostic to LiDAR representations hence our SSL framework can be universally applied. 2) Statistically grounded: We provide a detailed analysis to theoretically explain the applicability of the proposed framework. 3) Effective: Comprehensive experiments on popular LiDAR segmentation datasets demonstrate our effectiveness and superiority. Notably, we achieve competitive results over fully-supervised counterparts with $2\times$ to $5\times$ fewer labels and improve the supervised-only baseline significantly by relatively 10.8%. We hope this concise yet high-performing framework could facilitate future research in semi-supervised LiDAR segmentation.

1 INTRODUCTION

LiDAR segmentation, crucial for autonomous vehicle perception, enables semantic perception of the 3D environment Roriz et al. (2021). However, densely annotating LiDAR point clouds is costly and hinders the scalability of fully-supervised methods Unal et al. (2022); Hu et al. (2021); Kong et al. (2021). Semi-supervised learning (SSL) that leverages unlabeled data is a promising solution for scalable LiDAR segmentation Gao et al. (2021); Triess et al. (2021).

Semi-supervised LiDAR segmentation is underexplored, with modern SSL frameworks designed for 2D image recognition Berthelot et al. (2019); Sohn et al. (2020) and semantic segmentation Ouali et al. (2020); Ke et al. (2020) tasks only delivering sub-par performance on 3D data. A recent study Jiang et al. (2021) proposed a point contrastive learning framework, but it overlooks important properties specific to LiDAR point clouds by not differentiating indoor and outdoor scenes.

This work investigates using spatial prior for semi-supervised LiDAR segmentation, where the significance of spatial cues is especially pronounced. LiDAR point clouds accurately reflect real-world distributions, heavily reliant on spatial areas in LiDAR-centered 3D coordinates. As illustrated in Fig. 1 (left), top laser beams cover long distances and primarily detect vegetation, while middle and bottom beams detect cars and roads at medium and close distances, respectively.

To effectively leverage this strong spatial prior, we propose **LaserMix** to mix laser beams from different LiDAR scans, and then encourage the LiDAR segmentation model to make consistent and confident predictions before and after mixing. Our SSL framework is statistically grounded, which consists of the following components:

1) Partitioning the LiDAR scan into low-variation areas. We observe a strong distribution pattern on laser beams as shown in Fig. 1 (left) and thus propose the laser partition.

2) Efficiently mixing every area in the scan with foreign data and obtaining model predictions. We propose LaserMix to manipulate the laser-grouped areas from two LiDAR scans in an intertwining way as depicted in Fig. 1 (middle) and serves as an efficient LiDAR mixing strategy for SSL.

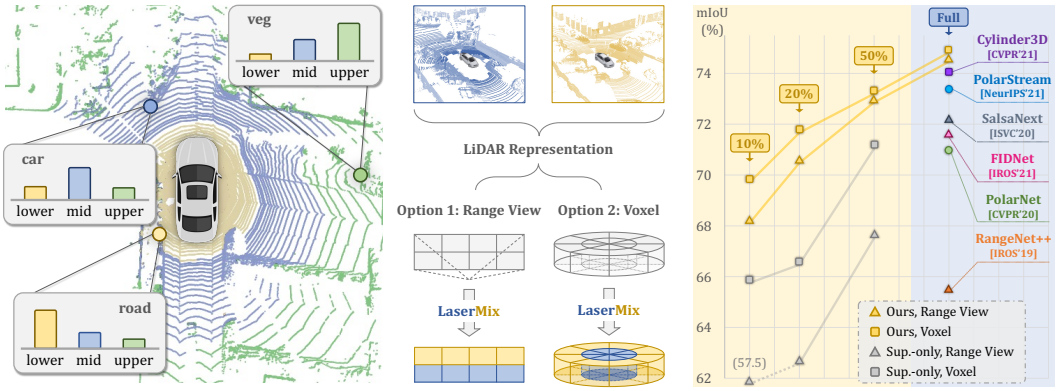


Figure 1: **Left:** The LiDAR point cloud contains strong spatial prior. Objects and backgrounds around the ego-vehicle have a patterned distribution on different (lower, middle, upper) laser beams. **Middle:** Following the scene structure, the proposed LaserMix blends beams from different LiDAR scans, which is compatible with various popular LiDAR representations. **Right:** We achieve superior results over SoTA methods in both low-data (10%, 20%, and 50% labels) and high-data (full labels) regimes on nuScenes Fong et al. (2022).

3) Encouraging models to make confident and consistent predictions on the same area in different mixing. We hence propose a mixing-based teacher-student training pipeline.

Despite the simplicity of our overall pipeline, it achieves competitive results over the fully supervised counterpart using $2\times$ to $5\times$ fewer labels as shown in Fig. 1 (right) and significantly outperforms all prevailing semi-supervised segmentation methods on nuScenes Fong et al. (2022) (up to $+5.7\%$ mIoU) and SemanticKITTI Behley et al. (2019) (up to $+3.5\%$ mIoU). Moreover, LaserMix directly operates on point clouds so as to be agnostic to different LiDAR representations, *e.g.*, range view Milioto et al. (2019) and voxel Zhu et al. (2021). Therefore, our pipeline is highly compatible with existing state-of-the-art (SoTA) LiDAR segmentation methods under various representations Zhang et al. (2020); Thomas et al. (2019); Zhao et al. (2021). Spatial prior is proven to play a pivotal role in the success of our framework through comprehensive empirical analysis.

2 Approach

This section presents our SSL framework that leverages priors in LiDAR data by encouraging spatial consistency in predictions. Due to space limits, we place the statistical derivation in Appendix.

2.1 LaserMix

Partition. LiDAR sensors have a fixed number of laser beams which are emitted isotropically around the ego-vehicle with predefined inclination angles as shown in Fig. 2. To obtain a proper set of spatial areas A, we propose to partition the LiDAR point cloud based on laser beams. Specifically, points captured by the same laser beam have a unified inclination angle to the sensor plane. For point i, its inclination is defined as: $\phi_i = \arctan(\frac{p_i^z}{\sqrt{(p_i^x)^2+(p_i^y)^2}})$, where (p^x, p^y, p^z) is the Cartesian coordinates of the LiDAR points. Given two LiDAR scans x_1 and x_2, we first group all points from each scan by their inclination angles. Concretely, to form m non-overlapping areas, a set of $m+1$ inclination angles $\Phi = \{\phi_0, \phi_1, \phi_2, ..., \phi_m\}$ will be evenly sampled within the range of the minimum and maximum inclination angles in the dataset (defined by sensor configurations), and the area set $A = \{a_1, a_2, ..., a_m\}$ can be formed by bounding area a_i in the inclination range $[\phi_{i-1}, \phi_i)$.

Role in our framework: Laser partition effectively "excites" a strong spatial prior in the LiDAR point cloud. As shown in Fig. 1 (left), we find an overt pattern in semantic classes detected by each laser beam. More concrete evidence on this aspect has been included in the Appendix. Despite being an empirical choice, we will show in later sections that laser partition significantly outperforms other partition choices, including random points (*MixUp*-like partition Zhang et al. (2018)), random areas (*CutMix*-like partition Yun et al. (2019)), and other heuristics like azimuth α (sensor horizontal direction) or radius r (sensor range direction) partitions.

Mixing. LaserMix mixes the aforementioned laser partitioned areas A from two scans in an intertwining way, *i.e.*, one takes from odd-indexed areas $A_1 = \{a_1, a_3, ...\}$ and the other takes from even-indexed areas $A_2 = \{a_2, a_4, ...\}$, so that each area's neighbor will be from the other scan:

$$\tilde{x}_1, \tilde{x}_2 = \text{LaserMix}(x_1, x_2)\,, \quad \tilde{x}_1 = x_1^{a_1} \cup x_2^{a_2} \cup x_1^{a_3} \cup \cdots\,, \quad \tilde{x}_2 = x_2^{a_1} \cup x_1^{a_2} \cup x_2^{a_3} \cup \cdots\,, \tag{1}$$

where $x_i^{a_j}$ is the data crop of x_i in the area a_j. The semantic labels are mixed in the same way. LaserMix is directly applied to the point clouds and is thus agnostic to the various LiDAR representations Hu et al. (2020); Milioto et al. (2019); Zhang et al. (2020); Zhu et al. (2021). We show LaserMix's instantiations with the *range view* and *voxel* representations as in Fig. 1 (middle), since they are currently the most efficient and the best-performing options, respectively.

Role in our framework: The cost for directly computing the marginal probability in Eq. 5 in Appendix on real-world LiDAR data is prohibitive; we need to iterate through all areas in A and all outside data in X_{out}, which requires $|A| \cdot |X_{\text{out}}|$ predictions in total. To reduce the training overhead, we take advantage of the fact that a prediction in an area will be largely affected by its neighboring areas and let X_{out} fill only the neighbors instead of all the remaining areas. By mixing two scans through *"intertwining"* the areas, the neighbors of each area are filled with data from the other scan, reducing the cost from $|A|$ to 1 on average. The scan before and after mixing counts as two data fillings, therefore $|X_{\text{out}}| = 2$. Overall, the training overhead is reduced from $|A| \cdot |X_{\text{out}}|$ to 2: only one prediction on original data and one additional prediction on mixed data are required for each LiDAR scan.

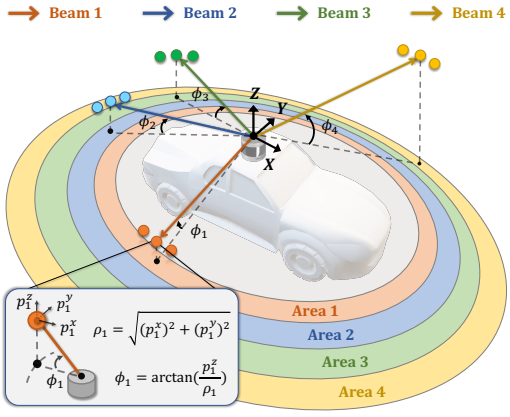


Figure 2: **Laser partition example**. We group LiDAR points (p_i^x, p_i^y, p_i^z) whose inclinations ϕ_i are within the same inclination range into the same area, as depicted in the color regions.

2.2 Overall Pipeline

There are two branches in our pipeline, one Student net $\mathcal{G}_\theta^s$ and one Teacher net $\mathcal{G}_\theta^t$. During training, a batch is composed of half labeled data and half unlabeled data. We collect the predictions from both $\mathcal{G}_\theta^s$ and $\mathcal{G}_\theta^t$, and produce pseudo-labels from Teacher net's prediction with a predefined confidence threshold T. For labeled data, we compute the cross-entropy loss between the Student net's prediction and the ground-truth as $\mathcal{L}_{\text{sup}}$. For unlabeled data, LaserMix blends every scan with a random labeled scan, together with their pseudo-label or ground-truth. Then, we let $\mathcal{G}_\theta^s$ predict on the mixed data and compute the cross-entropy loss $\mathcal{L}_{\text{mix}}$ (*w/* mixed labels). Moreover, we adopt the mean teacher idea in Tarvainen & Valpola (2017) and use Exponential Moving Average (EMA) to update the weights of $\mathcal{G}_\theta^t$ from $\mathcal{G}_\theta^s$, and compute the L2 loss between their predictions as: $\mathcal{L}_{\text{mt}} = ||\mathcal{G}_\theta^s(x) - \mathcal{G}_\theta^t(x)||_2^2$, where $||\cdot||_2^2$ is the L2 norm. The overall loss function is: $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{mix}}\mathcal{L}_{\text{mix}} + \lambda_{\text{mt}}\mathcal{L}_{\text{mt}}$, where λ_{mix} and λ_{mt} are loss weights. We use the Teacher net during inference as it empirically gives more stable results. There will be no extra inference overhead.

Role in our framework: Our overall pipeline minimizes the marginal entropy. Since the objective for minimizing the entropy has a hard optimization landscape, pseudo-labeling is a common resort in practice Lee (2013). Unlike conventional pseudo-label optimization in SSL which only aims to encourage the predictions to be confident, minimizing the marginal entropy requires all predictions to be both confident and consistent. Hence, we use the ground-truth and pseudo-label as an anchor and encourage the model's predictions to be confident and consistent with these supervision signals.

3 Experiments

Data. We build three SSL benchmarks upon nuScenes Fong et al. (2022), SemanticKITTI Behley et al. (2019), and ScribbleKITTI Unal et al. (2022). For all three sets, we uniformly sample 1%, 10%, 20%, and 50% labeled training scans and assume the remaining ones as unlabeled. This is in

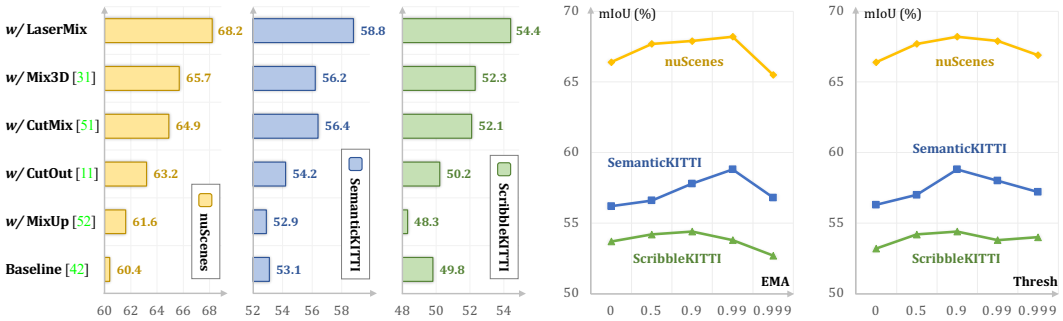


Figure 3: Ablation studies on: **Left:** Different mixing-based techniques used in point partition & mixing; **Middle:** Different EMA decay rates between the Teacher net and the Student net; **Right:** Different confidence thresholds T used in the pseudo-label generation.

line with the conventional settings from the semi-supervised image segmentation community Ouali et al. (2020); Ke et al. (2020); Chen et al. (2021).

Implementation Details. We adopt FIDNet Zhao et al. (2021) and Cylinder3D Zhu et al. (2021) as the segmentation backbones for the *range view* and the *voxel* options, respectively. The input resolution of range images is set as 64×2048 for SemanticKITTI and ScribbleKITTI, and 32×1920 for nuScenes. The voxel resolution is fixed as $[240, 180, 20]$ for all three sets. The number of spatial areas m in LaserMix is uniformly sampled from 2 to 6 areas. We denote the supervised-only baseline as *sup.-only*. Due to the lack of LiDAR SSL works, we also compare SoTA consistency regularization Tarvainen & Valpola (2017); Chen et al. (2021); French et al. (2020) and entropy minimization Zou et al. (2018) methods from semi-supervised image segmentation.

Table 1: Benchmarking results among different SSL methods with the LiDAR *range view* (top) and *voxel* (bottom) representations.

Repr.	Method	nuScenes 1%	nuScenes 10%	nuScenes 20%	nuScenes 50%	SemanticKITTI 1%	SemanticKITTI 10%	SemanticKITTI 20%	SemanticKITTI 50%	ScribbleKITTI 1%	ScribbleKITTI 10%	ScribbleKITTI 20%	ScribbleKITTI 50%
Range View	*Sup.-only*	38.3	57.5	62.7	67.6	36.2	52.2	55.9	57.2	33.1	47.7	49.9	52.5
	MeanTeacher	42.1	60.4	65.4	69.4	37.5	53.1	56.1	57.4	34.2	49.8	51.6	53.3
	CBST	40.9	60.5	64.3	69.3	39.9	53.4	56.1	56.9	35.7	50.7	52.7	54.6
	CutMix-Seg	43.8	63.9	64.8	69.8	37.4	54.3	56.6	57.6	36.7	50.7	52.9	54.3
	CPS	40.7	60.8	64.9	68.0	36.5	52.3	56.3	57.4	33.7	50.0	52.8	54.6
	LaserMix	**49.5**	**68.2**	**70.6**	**73.0**	**43.4**	**58.8**	**59.4**	**61.4**	**38.3**	**54.4**	**55.6**	**58.7**
Voxel	*Sup.-only*	50.9	65.9	66.6	71.2	45.4	56.1	57.8	58.7	39.2	48.0	52.1	53.8
	MeanTeacher	51.6	66.0	67.1	71.7	45.4	57.1	59.2	60.0	41.0	50.1	52.8	53.9
	CBST	53.0	66.5	69.6	71.6	48.8	58.3	59.4	59.7	41.5	50.6	53.3	54.5
	CPS	52.9	66.3	70.0	72.5	46.7	58.7	59.6	60.5	41.4	51.8	53.9	54.8
	LaserMix	**55.3**	**69.9**	**71.8**	**73.2**	**50.6**	**60.0**	**61.9**	**62.3**	**44.2**	**53.7**	**55.1**	**56.8**

Comparative Studies. Tab. 1 benchmarks results for various SSL methods. For all three sets under different data splits, we observe significant improvements in our approach over the *sup.-only* baseline. Such gains are especially evident in *range view*, which reach up to 11.2% mIoU. We also observe constant improvements for the *voxel* option, which provide on average 4.1% mIoU gains over all splits across all sets. The results verify the effectiveness of our framework and further highlight the importance of leveraging unlabeled data in LiDAR semantic segmentation.

Ablation Studies. Fig. 3 (left) compares LaserMix with other mixing methods. MixUp and CutMix can be considered as setting A to random points and random areas, respectively. We observe that MixUp has no improvements over the baseline on average since there is no distribution pattern in random points. CutMix has a considerable improvement over the baseline, as there is always a structure prior in scene segmentation, *i.e.*, the same semantic class points tend to cluster, which reduces the entropy in any continuous area. This prior is often used in image semantic segmentation SSL French et al. (2020). However, our spatial prior is much stronger, where not only the area structure but also the area's spatial position has been considered.

4 Conclusion

We propose a novel SSL approach that utilizes the unique spatial prior in LiDAR point clouds. Our statistically-grounded SSL pipeline includes a novel LiDAR mixing technique, LaserMix, that intertwines laser beams from different LiDAR scans. Our approach is demonstrated to be effective and superior through empirical analysis on three popular LiDAR semantic segmentation datasets, and its simplicity sheds light on the scalable deployment of the LiDAR semantic mapping system.

REFERENCES

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9297–9307, 2019.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622, 2021.

Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, pp. 3795–3802, 2022.

Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *British Machine Vision Conference (BMVC)*, 2020.

Biao Gao, Yancheng Pan, Chengkun Li, Sibo Geng, and Huijing Zhao. Are we hungry for 3d lidar data for semantic segmentation? a survey of datasets and methods. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2004.

Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Niki Trigoni Zhihua Wang, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11108–11117, 2020.

Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Ales Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds with 1000x fewer labels. *arXiv preprint arXiv:2104.04891*, 2021.

Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6423–6432, 2021.

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, pp. 429–445, 2020.

Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. *arXiv preprint arXiv:2111.15242*, 2021.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning Workshops (ICMLW)*, volume 3, 2013.

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

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12674–12684, 2020.

Ricardo Roriz, Jorge Cabral, and Tiago Gomes. Automotive lidar technology: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin D Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6411–6420, 2019.

Larissa T Triess, Mariella Dreissig, Christoph B Rist, and J Marius Zöllner. A survey on deep domain adaptation for lidar perception. In *IEEE Intelligent Vehicles Symposium Workshops (IVW)*, pp. 350–357, 2021.

Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9601–9610, 2020.

Yiming Zhao, Lin Bai, and Xinming Huang. Fidnet: Lidar point cloud semantic segmentation with fully interpolation decoding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4453–4458, 2021.

Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9939–9948, 2021.

Yang Zou, Zhiding Yu, B V K Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision (ECCV)*, pp. 289–305, 2018.

A APPENDIX

A.1 LEVERAGING THE SPATIAL PRIOR FOR SSL

Spatial Prior Formulation. The distribution of real-world objects/backgrounds has a strong correlation to their spatial positions in the LiDAR scan. Objects/backgrounds inside a specified spatial area of a LiDAR point cloud follow similar patterns, *e.g.*, the close-range area is most likely *road* while the long-range area consists of *building*, *vegetation*, etc. In another word, there exists a spatial area $a \in A$ where LiDAR points and semantic labels inside the area (denoted as X_{in} and Y_{in}, respectively) will have relatively low variations. Formally, the conditional entropy $H(X_{\text{in}}, Y_{\text{in}}|A)$ is smaller. Therefore, when estimating the parameter θ of the segmentation network $\mathcal{G}_\theta$, we would expect:

$$\mathbb{E}_\theta[H(X_{\text{in}}, Y_{\text{in}}|A)] = c\,, \tag{2}$$

where c is a small constant. Similar to the classic entropy minimization Grandvalet & Bengio (2004), the constraint in Eq. 2 can be converted to a prior on the model parameter θ using the principle of entropy maximization:

$$P(\theta) \propto \exp(-\lambda H(X_{\text{in}}, Y_{\text{in}}|A)) \propto \exp(-\lambda H(Y_{\text{in}}|X_{\text{in}}, A))\,, \tag{3}$$

where $\lambda > 0$ is the Lagrange multiplier corresponding to constant c; $H(X_{\text{in}}|A)$ has been ignored for being independent of the model parameter θ. We consider Eq. 3 as the formal formulation of the spatial prior and discuss how to empirically compute it in the following sections.

Marginalization. To utilize the spatial prior defined in Eq. 3, we empirically compute the entropy $H(Y_{\text{in}}|X_{\text{in}}, A)$ of the LiDAR points *inside* area A as follows:

$$\hat{H}(Y_{\text{in}}|X_{\text{in}}, A) = \hat{\mathbb{E}}_{X_{\text{in}}, Y_{\text{in}}, A}[P(Y_{\text{in}}|X_{\text{in}}, A) \log P(Y_{\text{in}}|X_{\text{in}}, A)]\,, \tag{4}$$

where $\hat{\cdot}$ denotes the empirical estimation. The end-to-end LiDAR segmentation model $\mathcal{G}_\theta$ usually takes full-sized data as inputs during inference. Therefore, to compute $P(Y_{\text{in}}|X_{\text{in}}, A)$ in Eq. 4, we first pad the data *outside* the area to obtain the full-sized data. Here we denote the data *outside* the area as X_{out}; we then let the model infer $P(Y_{\text{in}}|X_{\text{in}}, X_{\text{out}}, A)$, and finally marginalize X_{out} as:

$$P(Y_{\text{in}}|X_{\text{in}}, A) = \hat{\mathbb{E}}_{X_{\text{out}}}[P(Y_{\text{in}}|X_{\text{in}}, X_{\text{out}}, A)]\,. \tag{5}$$

The generative distribution of the padding $P(X_{\text{out}})$ can be directly obtained from the dataset.

Training. Finally, we train the segmentation model $\mathcal{G}_\theta$ using the standard maximum-a-posteriori (MAP) estimation. We maximize the posterior that can be computed by Eq. 3, Eq. 4 and Eq. 5, which is formulated as follows:

$$\begin{aligned} C(\theta) &= L(\theta) - \lambda \hat{H}(Y_{\text{in}}|X_{\text{in}}, A) = L(\theta) \\ &- \lambda \hat{\mathbb{E}}_{X_{\text{in}}, Y_{\text{in}}, A}[P(Y_{\text{in}}|X_{\text{in}}, A) \log P(Y_{\text{in}}|X_{\text{in}}, A)]. \end{aligned} \tag{6}$$

Here, $L(\theta)$ is the likelihood function computed using labeled data, *i.e.*, the conventional supervised learning. Minimizing $\hat{H}(Y_{\text{in}}|X_{\text{in}}, A)$ requires the marginal probability $P(Y_{\text{in}}|X_{\text{in}}, A)$ to be confident, which further requires $P(Y_{\text{in}}|X_{\text{in}}, X_{\text{out}}, A)$ to be both confident and consistent with respect to different outside data X_{out}.

In summary, our proposed SSL framework in Eq. 6 encourages the segmentation model to make confident and consistent predictions at a predefined area, regardless of the data outside the area. The predefined area set A determines the "strength" of the prior. When setting A to the full area (*i.e.*, the whole point cloud), our framework degrades to the classic entropy minimization framework Grandvalet & Bengio (2004).

Implementation. There are three key steps for implementing our framework:

- *Step 1)*: Select a proper partition set A which maintains strong spatial prior;
- *Step 2)*: Efficiently compute the marginal probability, *i.e.*, $P(Y_{\text{in}}|X_{\text{in}}, A)$;
- *Step 3)*: Efficiently minimize the marginal entropy, *i.e.*, $\hat{H}(Y_{\text{in}}|X_{\text{in}}, A)$.

We propose a simple yet effective implementation following these steps in Sec. 2 of the main paper.

A.2 Case Study: Spatial Prior in LiDAR Data

A.2.1 Laser Partition

As mentioned in the main body of this paper, the LiDAR point clouds collected by the LiDAR sensor on top of the autonomous vehicle contain inherent spatial cues, which lead to strong patterns in laser beam partition. In this section, we conduct a case study on SemanticKITTI Behley et al. (2019) to verify our findings. The LiDAR scans in SemanticKITTI are collected by the Velodyne-HDLE64 sensor, which contains 64 laser beams emitted isotropically around the ego-vehicle with predefined inclination angles. In this study, we split each LiDAR point cloud into eight non-overlapping areas, *i.e.*, $A = \{a_1, a_2, ..., a_8\}$. Each area a_i contains points captured from the consecutive 8 laser beams.

A.2.2 Spatial Prior

As can be seen from the fourth column in Tab. 2, different semantic classes have their own behaviors in these predefined areas. Specifically, the *road* class occupies mostly the first four areas (close to the ego-vehicle) while hardly appearing in the last two areas (far from the ego-vehicle). The *vegetation* class and the *building* class behavior conversely to *road* and appear at the long-distance areas (*e.g.*, a_6, a_7, a_8). The dynamic classes, including *car*, *bicyclist*, *motorcyclist*, and *person*, tend to appear in the middle-distance areas (*e.g.*, a_4, a_5, a_6). Similarly, from the heatmaps shown in the fifth column in Tab. 2, we can see that these semantic classes tend to appear (lighter colors) in only certain areas. For example, the *traffic-sign* class has a high likelihood to appear in the long-distance regions from the ego-vehicle (upper areas in the corresponding heatmap).

These unique distributions reflect the spatial layout of street scenes in the real world. In this work, we propose to leverage these strong spatial cues to construct our SSL framework. The experimental results verify that the spatial prior can better encourage consistency regularization in LiDAR segmentation under annotation scarcity.

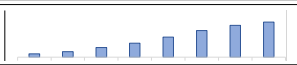
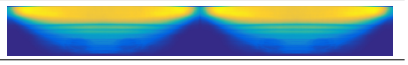
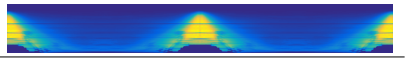
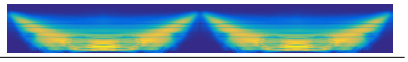
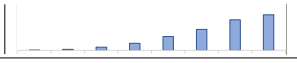
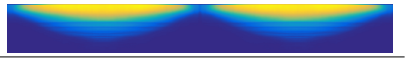
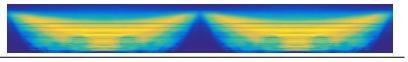
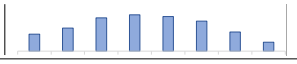
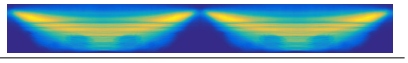
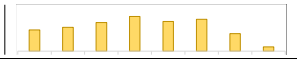
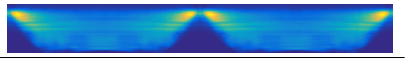
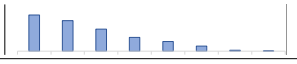
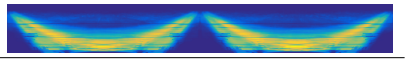
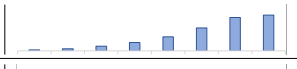
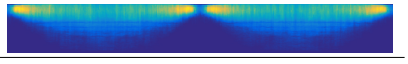
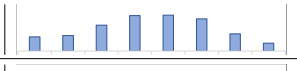
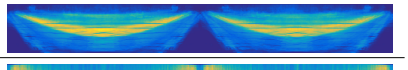
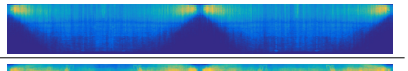
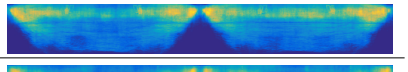
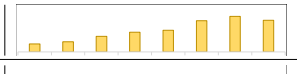
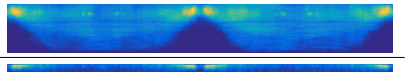
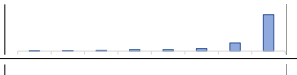
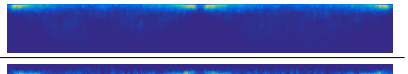
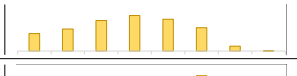
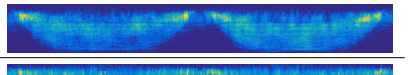
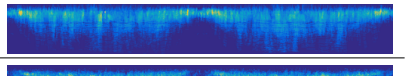
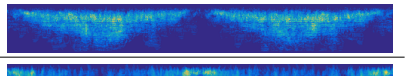
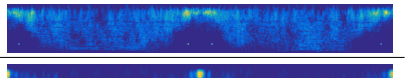
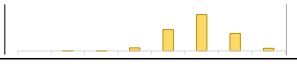
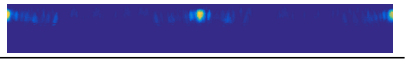
A.3 Public Resources Used

We acknowledge the use of the following public resources, during the course of this work:

- nuScenes[1] ..... CC BY-NC-SA 4.0
- nuScenes-devkit[2] ..... Apache License 2.0
- SemanticKITTI[3] ..... CC BY-NC-SA 4.0
- SemanticKITTI-API[4] ..... MIT License
- ScribbleKITTI[5] ..... Unknown
- FIDNet[6] ..... Unknown
- Cylinder3D[7] ..... Apache License 2.0
- TorchSemiSeg[8] ..... MIT License
- Mix3D[9] ..... Unknown
- MixUp[10] ..... Attribution-NonCommercial 4.0
- CutMix[11] ..... MIT License

[1] https://www.nuscenes.org/nuscenes.
[2] https://github.com/nutonomy/nuscenes-devkit.
[3] http://semantic-kitti.org.
[4] https://github.com/PRBonn/semantic-kitti-api.
[5] https://github.com/ouenal/scribblekitti.
[6] https://github.com/placeforyiming/IROS21-FIDNet-SemanticKITTI.
[7] https://github.com/xinge008/Cylinder3D.
[8] https://github.com/charlesCXK/TorchSemiSeg.
[9] https://github.com/kumuji/mix3d.
[10] https://github.com/facebookresearch/mixup-cifar10.
[11] https://github.com/clovaai/CutMix-PyTorch.

Table 2: A case study on the **strong spatial prior** in the LiDAR data (statistics calculated from the SemanticKITTI Behley et al. (2019) dataset in this example). For each semantic class, we show its type (static or dynamic), occupation (valid # of points in percentage), distribution among eight areas ($A = \{a_1, a_2, ..., a_8\}$, *i.e.*, eight laser beam groups), and the heatmap in range view (lighter colors correspond to areas that have a higher likelihood to appear and vice versa).

Class	Type	Proportion	Distribution	Heatmap
vegetation	static	24.825%		
road	static	22.545%		
sidewalk	static	16.353%		
building	static	12.118%		
terrain	static	8.122%		
fence	static	7.827%		
car	dynamic	4.657%		
parking	static	1.681%		
trunk	static	0.580%		
other-ground	static	0.396%		
pole	static	0.296%		
other-vehicle	dynamic	0.229%		
truck	dynamic	0.193%		
traffic-sign	static	0.061%		
motorcycle	dynamic	0.045%		
person	dynamic	0.036%		
bicycle	dynamic	0.018%		
bicyclist	dynamic	0.014%		
motorcyclist	dynamic	0.004%		

- CutMix-Seg[12] ..... MIT License
- CBST[13] ..... Attribution-NonCommercial 4.0
- MeanTeacher[14] ..... Attribution-NonCommercial 4.0

[12] https://github.com/Britefury/cutmix-semisup-seg.
[13] https://github.com/yzou2/CBST.
[14] https://github.com/CuriousAI/mean-teacher.