

Discwise Active Learning for LiDAR Semantic Segmentation

Ozan Unal^{1,*}, Dengxin Dai², Ali Tamer Unal³, and Luc Van Gool^{1,4,5}

Abstract—While LiDAR data acquisition is easy, labeling for semantic segmentation remains highly time consuming and must therefore be done selectively. Active learning (AL) provides a solution that can iteratively and intelligently label a dataset while retaining high performance and a low budget. In this work we explore AL for LiDAR semantic segmentation. As a human expert is a component of the pipeline, a practical framework must consider common labeling techniques such as sequential labeling that drastically improve annotation times. We therefore propose a discwise approach (DiAL), where in each iteration, we query the region a single frame covers on global coordinates, labeling all frames simultaneously. We then tackle the two major challenges that emerge with discwise AL. Firstly we devise a new acquisition function that takes 3D point density changes into consideration which arise due to location changes or ego-vehicle motion. Next we solve a mixed-integer linear program that provides a general solution to the selection of multiple frames while taking into consideration the possibilities of disc intersections. Finally we propose a semi-supervised learning approach to utilize all frames within our dataset and improve performance.

I. INTRODUCTION

Dense prediction tasks such as LiDAR semantic segmentation require large amounts of labeled data. While new LiDAR frames are easy to acquire, labeling, especially in 3D, is not only tedious but also highly costly. This severely hinders the scalability and deployability of such tasks in any practical application such as autonomous driving. To retain an economical budget, labeling must therefore be done selectively, as exhaustively labeling all acquired frames is simply not feasible. The question then remains, amongst a pool of acquired LiDAR frames, how do we select which parts should be labeled?

Active learning (AL) provides a framework to intelligently decide what data to label and has proven to be widely successful in comparable 2D tasks where similar economical constraints apply [1], [2]. In an AL setting, a model is initially trained on a small amount of labeled data. Based on the model’s predictions of the unlabeled data (often on its uncertainty), an acquisition function determines which data to query an external human oracle for a label. The human expert labels the selected data points, which then expand the training set. The model can then be retrained and thus the loop restarts until the predetermined budget is reached. Such a framework that incorporates the human expert into the pipeline often drastically reduces labeling costs while retaining performance. In this work, our goal is to investigate AL for LiDAR semantic segmentation.

This work was funded by Toyota Motor Europe via TRACE Zurich.

¹Computer Vision Lab, ETH Zurich; ²Huawei Technologies, Zurich Research Center; ³Department of Industrial Engineering, Bosphorus University; ⁴VISICS, ESAT/PSI, KU Leuven; ⁵INSAIT

*ozan.unal@vision.ee.ethz.ch

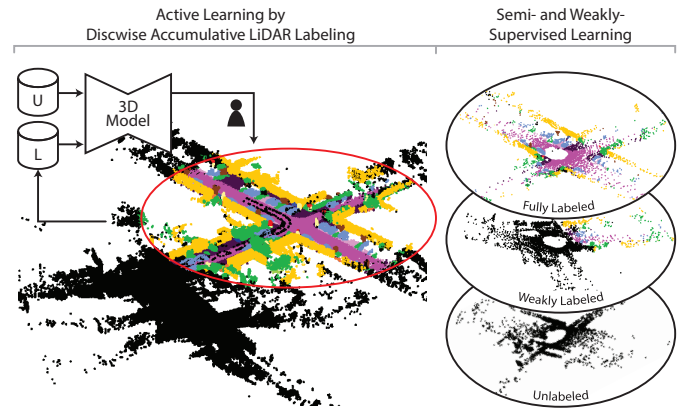


Fig. 1. Accumulation of sequential LiDAR point clouds can severely reduce labeling costs, thus an active learning framework must learn to select accumulated regions for labeling to take advantage of this commonly used technique. When discs are queried for labels, the resulting dataset consists of fully labeled, weakly labeled and unlabeled frames.

Before we move forward, as we aim to include a human labeler into our pipeline, it is important to understand the commonly used techniques for labeling LiDAR sequences in order to ground our AL approach in reality and ultimately provide a deployable solution. Unlike 2D images, LiDAR scans comprise of 3D points that provide coordinates in the real world. This means sequential point clouds can be projected onto a global coordinate system and concatenated to form a denser representation. This common trick can immensely reduce labeling costs [3], [4], [5]. Most outdoor scenes are dominated by static structures (e.g. road, building, vegetation) and static objects (e.g. parked vehicles, bicycles) thus when working on concatenated point clouds they only need to be labeled once to have them labeled on all frames. In fact, labeling the area of a single frame in global coordinates for concatenated points takes less time than labeling two individual frames based on our user study.

To exploit the sequential labeling of concatenated frames within our pipeline and save time labeling, instead of a single frame, we propose constructing the problem of AL for LiDAR semantic segmentation using a disc as the unit data point, i.e. the region a single frame covers on global coordinates. Therefore when a disc is selected to be labeled, the oracle labels all points from any frame that fall within the boundaries of the selected area. An illustration of our proposed discwise labeling can be seen in Fig. 1 - left. Exploring discwise AL yields two main challenges which can be identified when compared to the analogous 2D setting.

Firstly, in a common imagewise AL pipeline the pixel-wise epistemic uncertainty is aggregated over the image via

the sum [6], [7] or mean [8], [9] to form the acquisition function. However, unlike 2D images, outdoor LiDAR scenes have vastly varied point counts which can have unintentional consequences when naively aggregating pointwise uncertainty. To combat such effects, we propose a simple but intuitive solution where we initially pool information within globally defined voxels via a symmetric function and then aggregate voxelwise uncertainty over discs.

Secondly, following imagewise aggregation, in common 2D AL pipelines the images yielding the highest uncertainty are then queried for labels. However unlike 2D images, discs can intersect with one another in global coordinates and intersecting regions do not provide any additional information. A naive strategy of disallowing intersecting solutions is not guaranteed to yield the optimal disc selection that maximizes uncertainty. Therefore to form a general solution, we construct and solve a mixed-integer linear program that maximizes the information within the total area formed by the selected discs. Our solution allows discs to intersect, if the resulting union would yield a higher uncertainty and collapses to the commonly chosen argmax function when selecting only a single disc.

Having established the major building blocks, we construct a basic yet effective AL pipeline (DiAL) for LiDAR semantic segmentation where a model trains on discwise labeled frames, an acquisition function aggregates uncertainty over discs while respecting point count variability and a selection algorithm selects discs that maximize information.

Finally, we tackle the problem of semi-supervised training via a mean teacher framework [10], [4] to utilize all three types of frames in our actively labeled dataset (see Fig. 1).

Our contributions can be summarised as follows:

- We develop a simple active learning pipeline for LiDAR semantic segmentation named DiAL that aligns well with the established techniques for annotating. We query discs on accumulated frames to exploit the effectiveness of sequential data labeling.
- We aggregate uncertainty over discs via a two step method to respect the point count variability of outdoor LiDAR data and limit bias introduced via aggregation.
- We construct and solve a mixed-integer linear problem that outputs a number of (possibly intersecting) discs, allowing a generalized solution to querying multiple samples each iteration.
- With a simple semi-supervised learning approach that utilizes labeled, partially labeled and unlabeled frames, our method performs at 99.8% relative to fully supervised learning on dense labels, with only 50 discs ($\sim 0.5\%$ cost of framewise labeling the entire dataset).

II. RELATED WORK

Active Learning for LiDAR Segmentation: Active learning provides a framework that leverages existing labeled data points to intelligently decide how to expand the dataset, most commonly via the use of a human oracle [1], [2]. The core goal is extract as much performance as possible from a model by labeling as few samples. A commonly used strategy to

maximize information yield per labeled data is to sample data points that show high uncertainty. Common metrics include the softmax confidence [11], [12], softmax margin [13], [14], [6], softmax entropy [15], [12], or Bayesian approaches that utilize learning by disagreement (BALD) [16], [9]. Recent work have investigated active learning within the context of 3D semantic segmentation [17], [18], [19]. Specifically, ReDAL [17] proposes labeling highest information yielding regions within point cloud frames based on softmax entropy, color discontinuity, and structural complexity. However to propose a general AL framework for point cloud segmentation, the sequential nature of outdoor LiDAR scans remain to be considered. LiDAL [18] takes advantage of this inherent property and utilizes cross-frame predictions as a measure of uncertainty to again, find and label high information providing regions. While both methods shows great performance at low percentage point counts (1% – 5%), we argue that the goal of an AL pipeline should not be to reduce point counts but to reduce labeling times. It is therefore vital to consider the sequential nature of LiDAR frames during the costly labeling stage.

Semi-Supervised LiDAR Semantic Segmentation: LiDAR semantic segmentation remains to be a challenging and computationally expensive task with current research focused on understanding how to best process the unordered data structure: directly operating on points [20], [21], [22], projecting it onto 2D [23], [24] and the now prevailing strategy, voxelization and 3D sparse convolution [25], [26], [27]. Recently, data-efficient methods have risen in popularity that aim to reduce the labeling cost of the dense prediction task. There are two common strategies: (i) weakly-supervised that provides weak labels for all available frames (e.g. scribble labels with ScribbleKITTI [4]), (ii) semi-supervised that provides dense labels for only a subset of frames, while the rest remain unlabeled [28]. For semi-supervised learning, SemiSup [28] proposes using a pseudo-label guided point contrastive loss and SSPC [29] utilizes self-training to reduce the gap to their fully supervised baselines.

III. ACTIVE LEARNING FOR LiDAR SEGMENTATION

For LiDAR semantic segmentation, given the vast differences in both time and budgetary requirements for data acquisition and labeling, the latter must be done selectively. An active learning framework provides a solution to intelligently decide what data to label next. In the following sections, we explore the individual steps that typically form the loop of an active learning framework:

III-A A model, trained on the labeled dataset, is used to determine pointwise uncertainty for the unlabeled data.

III-B The uncertainty is aggregated over the chosen unit of data, which in our case is the accumulated region (disc).

III-C The data points yielding the highest uncertainty are queried for labels and thus expand the labeled dataset.

A. Pointwise Epistemic Uncertainty Estimation

The goal in LiDAR semantic segmentation is to discover the dependency of the pointwise distribution over the labels

$y \in Y$ on an input variable $x \in X$ via the model weights w . Formally, reducing this to a classification task where y can be of class $c \in C$, we define the conditional probability as:

$$p(y = c|x, X, Y) = \int p(y = c|x, \omega)p(\omega|X, Y)d\omega \quad (1)$$

As the posterior distribution $p(w|X, Y)$ is intractable, we approximate it with a variational distribution $q(\omega)$ and aim to minimize the Kullback-Leibler divergence between the two distributions $\text{KL}(q(\omega)|p(\omega|X, Y))$. Following Gal. *et al.* [9], we further approximate the variational distribution via Monte Carlo integration by employing variational inference, i.e. by applying stochastic forward passes (i.e. a model is trained with dropout and dropout is performed during inference):

$$\begin{aligned} p(y = c|x, X, Y) &\approx \int p(y = c|x, \omega)q(\omega)d\omega \\ &\approx \frac{1}{N} \sum_{n=1:N} p(y = c|x, \hat{\omega}_n) \end{aligned} \quad (2)$$

We measure the pointwise epistemic uncertainty within our training set X as the mutual information between the model parameters and the model posterior:

$$\mathbb{I}[y, \omega|x, X] = \mathbb{H}[y|x, X] - \mathbb{E}_{p(\omega|X)} [\mathbb{H}[y|x, \omega]] \quad (3)$$

with $\mathbb{H}[y|x, X] = -\sum_c p(y = c|x, X) \log p(y = c|x, X)$ defining the pointwise entropy. Inserting Eq. 2 into Eq. 3:

$$\begin{aligned} \mathbb{I}[y, \omega|x, X] &\approx -\sum_c \left(\frac{1}{N} \sum_n \hat{p}_c \right) \log \left(\frac{1}{N} \sum_n \hat{p}_c \right) \\ &\quad + \frac{1}{N} \sum_c \sum_n \hat{p}_c \log \hat{p}_c \end{aligned} \quad (4)$$

with \hat{p} denoting the softmax output of a stochastic forward pass. For the derivation, please refer to [30].

B. Discwise Uncertainty Aggregation

Having established a metric to capture pointwise uncertainty, the next step in an active learning pipeline is to aggregate this information over a chosen *unit of data*. This aggregate will then be used to guide the ensuing data unit selection step. The choice of the data unit is therefore not trivial as it will determine what will be presented to the human annotator. To make a practical choice, it is important to first understand the labeling process of LiDAR point clouds for semantic segmentation to ground our choice in reality and take advantage of commonly utilized techniques.

LiDAR labeling is notoriously difficult and requires a high level of expertise from the annotator. While scenes are represented in 3D, labeling is done on 2D projections onto a monitor screen. As the projection is view dependent, the labeling process requires constant navigation and readjustments within the 3D space. Labeling a 2D area on a screen labels every point that falls within that selected area, commonly causing unintentional regions to be labeled, thus requiring regular corrections.

While working with 3D point clouds does come with obvious disadvantages for labeling, there is one strength that drastically speeds up the process. Given sequential data (as commonly acquired in outdoor LiDAR applications), individual frames can be projected onto a global coordinate system via odometry poses to form a concatenated point cloud sequence. This allows each global area to be labeled once, in order to have it labeled on every frame. Given that most outdoor scenes are dominated by static things (e.g. parked vehicles, bicycles) and static stuff (e.g. road, building, vegetation), this technique can save thousands of hours for large-scale datasets. To quantify the impact of this method, we utilize in-house annotators to label multiple frames as well as their corresponding discs. We define a disc as the region a single frame would cover on global coordinates (see Fig. 1 - left for an illustration of a labeled disc). We observe that with an overhead of 10% – 80%¹, one can not only fully label a single frame, but also weakly label several other frames within the same sequence by utilizing concatenated point cloud sequences. Unlike previous work that focus on reducing the total number of points labeled [17], [18], our goal in this work is to minimize labeling times. Therefore following this observation, to take advantage of sequential labeling, we propose using discs as units of data within our active learning pipeline.

To aggregate pointwise uncertainty over a disc, we first consider 2D analogous aggregation methods such as the sum [6], [7], [31] or the mean [8], [9]. However we note that unlike 2D images that have constant pixel counts, LiDAR discs can have vastly different point counts which can bias such methods. In specific, the *mean* fails to consider the variation in discwise point counts due to environmental changes (e.g. highway scenes have an overall low number of points compared to in-city 4-way junctions), and the *sum* fails to consider local density changes due to the variation in ego vehicle velocity (e.g a slow moving vehicle would accumulate more frames and thus more points within the same global environment compared to a fast moving car).

To combat these effects, we propose an intuitive two step aggregation method.

Firstly, to combat the local density changes, we aggregate mutual information (MI) within local neighborhoods via a symmetric function, i.e. we pool within globally defined voxels. Formally, we define the aggregated voxelwise MI as:

$$\mathbb{I}_j = f_{x_i \in V_j} \mathbb{I}[y_i, w|x_i, X] \quad (5)$$

with the set V_j containing all points x_i within the voxel centered around v_j of length l_{voxel} , i.e. $V_j = \{x_i \mid \|x_i - v_j\| \leq l_{voxel}/2\}$ and f chosen as min or max [20], [21].

Having voxelwise uncertainty, we can then aggregate the mutual information within discs to form our acquisition metric to guide our decision on what region to label next. To combat

¹For SemanticKITTI [3], we observe that a disc takes approximately 3.5 hours to label due to their fixed range, while the corresponding frame can vary significantly based on content and environment (which is in line with the reported labeling times of accumulated tiles [3]).

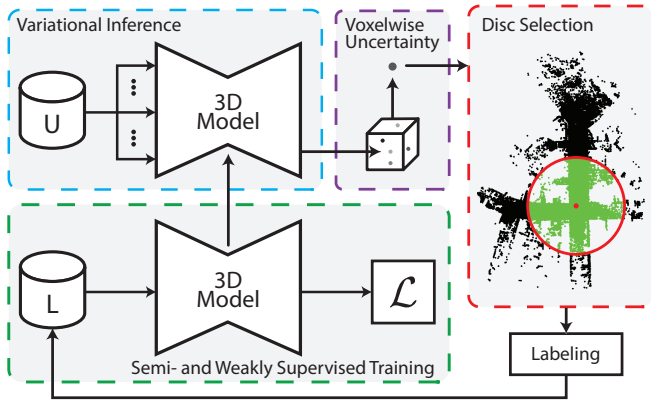


Fig. 2. Illustration of the active learning pipeline. We first estimate pointwise epistemic uncertainty, aggregate it on globally defined voxels and select discs that maximize the total information. The points within each disc is then labeled by an annotator.

discwise point count variations we refer back to the sum. Formally, we define the discwise uncertainty α_i as:

$$\alpha_i = \sum_j \mathbf{z}_{j,i} \mathbb{I}_j \quad (6)$$

with $\mathbf{z}_{j,i}$ defining a binary mask of voxels within the disc of radius R centered around x_i :

$$z_{j,i} = \begin{cases} 1, & \text{if } \|v_j - x_i\| - R \leq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

With the two step aggregation, we aim to maximize the pooled information we obtain from globally defined local neighborhoods based on the conditions set by function f .

It should be noted that, we consider a disc as the unit of data to remain comparable to current state-of-the-art framewise active learning strategies. However our method provides a general solution to the selection of any region shape within the registered point cloud sequence, as Eq. 7 can be freely designed based on annotator preferences.

C. Disc Selection for Labeling

Given discwise uncertainty, we can now query the disc \hat{i} that would yield the highest amount of new information:

$$\hat{i} = \operatorname{argmax}_i \alpha_i. \quad (8)$$

Selecting a single disc to label in each iteration of the active learning algorithm comes with additional fix costs that are associated with not only the initiation of the data labeling pipeline (e.g. set up time for the annotator) but also with the training of the model. A common solution to this issue is to query multiple data samples in each iteration [32], [1], [17].

While the selection of a single disc can be done via Eq. 8, the selection of multiple discs in each iteration cannot be trivially handled. Unlike with 2D images, discs can intersect in global coordinates and intersecting regions do not provide any additional information to the system. Yet, naively prohibiting intersections within the selection of discs would not guarantee the highest information yield.

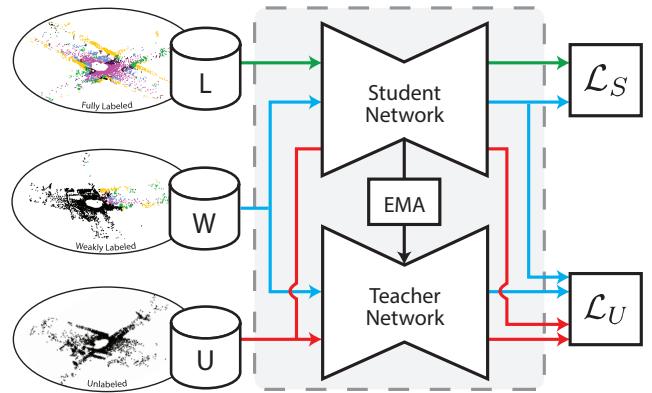


Fig. 3. Mean teacher framework for semi-supervised LiDAR semantic segmentation to take advantage of fully labeled (L), weakly labeled (W) and unlabeled (U) frames within the overall dataset. Colored paths show the distinct paths for the types of frames.

To this end, we propose a general solution by constructing a mixed-integer linear program that maximizes the information within the union of the areas formed by the discs. To select discs centered around possible points x_i of radius R , we solve the following problem:

$$\begin{aligned} \max \quad & \sum_j \mathbf{v}_j \mathbb{I}_j \\ \text{s.t.} \quad & \sum_i \mathbf{x}_i = N \\ & \mathbf{v}_j \leq \sum_i z_{j,i} \mathbf{x}_i, \quad \forall j \\ & 0 \leq \mathbf{v}_j \leq 1, \quad \forall j \\ & \mathbf{x}_i \in \{0, 1\}, \quad \forall i \end{aligned} \quad (9)$$

with \mathbf{x}_i and \mathbf{v}_j denoting binary and real variables that determine if a certain center point or voxel has been selected.

Here, the first constraint fixes the total number of selected discs to a predetermined value of N for each iteration, and the second constraint ensures that we select the union of the disc masks $z_{j,i}$ (see Eq. 7) for selected disc centers $x_i = 1$. Although \mathbf{v}_j 's are defined as $\{\mathbf{v}_j \in \mathbb{R} \mid 0 \leq \mathbf{v}_j \leq 1\}$, the \mathbf{x}_i variables being binary guarantees that \mathbf{v}_j 's would assume binary values in any optimal solution.

In ensuing steps of the AL pipeline, we set an additional constraint to fix the previously selected discs \hat{i} ($x_i = 1 \forall \hat{i}$) and adjust the first constraint accordingly ($\sum_i \mathbf{x}_i = tN$ for time step t). Our overall proposed AL framework for LiDAR semantic segmentation is illustrated in Fig. 2.

IV. SEMI-SUPERVISED LiDAR SEGMENTATION

By labeling discs, i.e. during any point in our active learning pipeline, our overall dataset consists of three types of frames: (i) fully labeled frames that span a complete disc region, (ii) weakly labeled that partially intersect with the labeled disc in global coordinates and (iii) unlabeled frames. When naively training with a supervised loss, we only make use of labeled points. To better utilize all information within our dataset, we construct a semi-supervised training approach that can utilize all forms of information to improve the quality of the final dense predictions of our model.

Points of all frame types can be divided into two subsets: labeled, and unlabeled. For labeled points we apply a supervised loss H that is typically chosen as cross entropy:

$$\mathcal{L}_S = H(\hat{y}, y) \quad (10)$$

given the predicted class distribution \hat{y} , and ground truth y .

To extend the supervision to unlabeled points we utilize a mean teacher framework (MT) [10]. The MT framework consists of a student that is typically trained using backpropagation (see Eq. 10), and a teacher network who’s weights (θ^{EMA}) are computed as the exponential moving average (EMA) of the student’s.

$$\theta_t^{\text{EMA}} = \beta\theta_{t-1}^{\text{EMA}} + (1 - \beta)\theta_t \quad (11)$$

To make use of the more robust representation capabilities of the moving average weights, we apply a consistency loss between the student’s predictions and the teacher’s via [4]:

$$\mathcal{L}_U = \text{KL}(p(\hat{y}_i), p(\hat{y}_i^{\text{EMA}})) \quad (12)$$

with \hat{y}^{EMA} denoting the teacher’s predicted class distribution. To increase the variation between the student and teacher’s outputs and therefore more effectively utilize the unsupervised loss, we apply perturbations to the student’s input in the form of rotation, Gaussian jitter, translation and scaling.

Formally, we define the total loss as:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_U \quad (13)$$

Our proposed mean-teacher based semi-supervised learning pipeline is illustrated in Fig. 3.

V. EXPERIMENTS

Implementation Details: In our experiments we use the popular SPVCNN [26] as a baseline model which already includes dropout layers. We use GUROBI [33] to solve the optimization problem. We do 10 stochastic forward passes for each sample to compute the mutual information. We set a radius of $50m$ for discs following precedent on LiDAR frame processing [25], [26] and a voxel size of $0.5m$ for the initial local aggregation. The disc size allows us to remain comparable to framewise methods. Furthermore, querying large regions allows us to better capture the underlying long tailed distribution of driving scenes.

We select $N = 5$ discs on each step, starting with an initial state of 1 arbitrarily chosen disc (sequence 0, disc 100). We set $\beta = 0.99$ following Unal *et al.* [4]. We set f as the min-operation unless stated otherwise. Furthermore, in this work we limit the possible disc centers x_i to the ego vehicle path to reduce the search space. It should be noted that while we impose this constraint for our experiments, our AL pipeline along with the mixed-integer linear program in Eq. 9 forms a general solution, allowing a free choice x_i and even $z_{j,i}$.

Dataset: We test and extensively ablate our method on SemanticKITTI [3] which is the most popular large-scale autonomous driving dataset for LiDAR semantic segmentation. It consists of 11 sequences with publicly available labels,

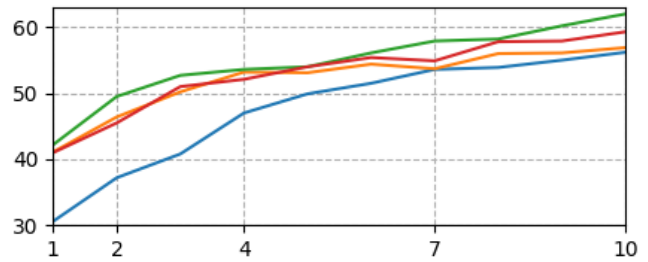


TABLE I
COMPARISON OF DIFFERENT UNCERTAINTY METRICS: (I) THE SOFTMAX CONFIDENCE (CONF), (II) THE SOFTMAX MARGIN (MAR), AND (III) THE SOFTMAX ENTROPY (ENT) TO MUTUAL INFORMATION (MI).

	Exp. Time	1%	2%	3%	4%	5%
		(1%)	2%	3%	4%	5%
Framewise	RAND _{fr}	48.8	52.1	53.6	55.6	57.2
	RAND _{re}	48.8	51.7	55.0	56.1	58.2
	SEGENT	48.8	49.8	48.3	49.1	48.2
	CSET	48.8	53.1	52.9	53.2	52.6
	ReDAL [17]	48.8	51.3	54.0	58.6	58.1
	LiDAL [18]	48.8	57.1	58.7	59.3	59.5
	DiAL (Ours)	44.4*	53.6	55.9	56.3	58.1

	Exp. Time	1%	2%	3%	4%	5%
		60%	90%	99%	99%	99%
DiAL (Discwise Ours)		61.4	63.8	63.8	63.8	63.8

TABLE II
COMPARISON OF AL METHODS UNDER FIXED BUDGET CONSTRAINTS. * INDICATES A DIFFERENT INITIAL STATE.

among which sequence 8 is reserved for validation. The training sequences consist of 19130 frames. Evaluation is carried via the mIoU metric over 19 classes on the *val*-set. We also showcase that our method works on two further datasets ScribbleKITTI [4] and nuScenes [5] and with other baseline models (MinkowskiNet [27]), with similar performance gains over random disc selection.

A. Results and Ablation Studies

In this section, we conduct extensive experiments and ablation studies to isolate the effect of each of our proposed components in DiAL, and compare them to both baseline methods and current approaches in the literature in order to show their necessity and effectiveness for discwise AL.

Uncertainty Metric: We compare the mutual information to other metrics commonly used for uncertainty estimation for classification tasks, including (i) the softmax confidence (CONF), (ii) the softmax margin (MAR), and (iii) the softmax entropy (ENT). More detailed explanation on all baseline approaches can be found in Hu *et al.* [18]. As seen in Tab. I, the mutual information provides strictly better performance within our proposed active learning pipeline for LiDAR semantic segmentation, and thus we retain this metric for the remainder of the experiments.

Disc vs. Frame: We compare our proposed strategy of discwise active learning to methods that do not take sequential labeling into consideration, including random selection of

Steps	0	1	2	4	7	10
Framework	1.1	19.2	22.6	26.1	32.6	40.0
Discwise	25.7	42.2	49.5	53.6	57.9	62.0
Δ	+24.6	+23.0	+26.9	+27.5	+25.3	+22.0

TABLE III

COMPARISON OF FRAMEWISE AND DISCWISE LABELING UNDER REALISTIC LOW BUDGET CONSTRAINTS.

Labeled	1%	3%	6%	12%
ReDAL [M39]	25.7	37.7	42.1	47.6
LiDAL [M12]	25.7	40.2	45.8	49.9
DiAL (Ours)	25.7	42.2	49.5	53.6

TABLE IV

COMPARISON OF AL WITH SEQUENTIAL LABELING.

frames ($RAND_{fr}$), random selection of framework regions ($RAND_{re}$), segment-entropy (SEGENT), core-set selection (CSET), ReDAL [17] and LiDAL [18]. A detailed explanation on baseline approaches can be found in Hu *et al.* [18].²

Firstly in Tab. II - top, we compare our approach applied with framework active learning to existing approaches. As seen, our baseline framework active learning framework performs on par with existing methods such as ReDAL [17].

Second, we showcase the effectiveness of sequential labeling in an AL setting in Tab. II - bottom, where we observe significantly better performance compared to our framework AL baseline under a fixed labeling budget.

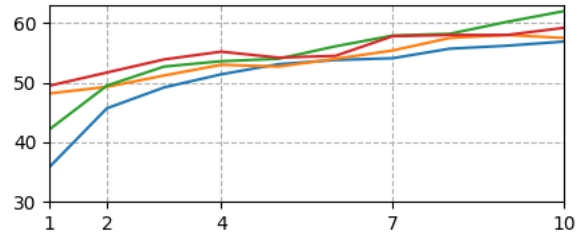
Of course these results are completely expected when we consider the percentage of points labeled. Here, with the aid of Tab. II we would like draw the readers attention to how misleading the percentage of labeled points might be when comparing it to labeling times of LiDAR point clouds. While the reported percentages may seem sparse, by the time it takes to label 1% of points on individual frames (191 frames), around 60% of the total point count can be labeled (95 discs)³. These findings once again emphasize the need to ground LiDAR based active learning research in commonly used techniques for labeling to ensure practicality.

To yet again showcase the advantages of discs over frames in low budget scenarios, we also provide an additional comparison in Tab. III where we see an overwhelming advantage of sequential labeling within AL.

Adapting SOTA to sequential AL: We strongly argue that all AL methods *should* utilize sequential labeling but not all of them are naturally extendable to incorporate sequential labeling. Specifically, ReDAL and LiDAL rely on memory

²To compare to the aforementioned framework methods under a fixed labeling budget, we need to first determine the equivalent number of discs to label, as, to emphasize once again, matching the percentage of labeling points would not match the labeling times. Assuming point percentage is directly correlated with frame percentage, to label 1% of points, we determine the appropriate number frames to label as 1% of the total frame count. To compensate for the time difference between discwise and framework labeling, we then apply our active learning strategy with discwise selection until a conservative $(1\%)F/2$ discs are queried, with F denoting the total frame count in the *train*-set. For the initial state, we randomly sample discs while minimizing intersections. We therefore ensure that each method is compared under the same labeling budget.

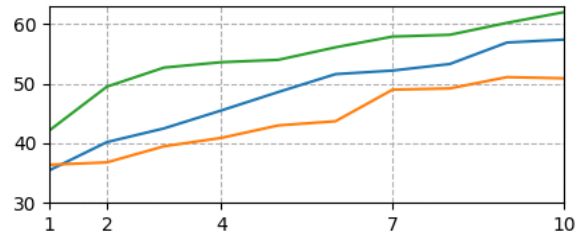
³The point count of each disc can correspond to up to 100 times the point count of the corresponding frame. We note that this ratio holds for most datasets as ego-vehicles often move at low velocities during data acquisition. Even at higher velocities, we argue that this example provides sufficient grounds that emphasizes the benefits of sequential labeling.



f	D	0	1	2	4	7	10
-	Mean	25.7	35.9	45.7	51.4	54.1	56.9
-	Sum	25.7	48.2	49.3	53.0	55.4	57.5
Max	Sum	25.7	49.5	51.7	55.2	57.8	59.2
Min	Sum	25.7	42.2	49.5	53.6	57.9	62.0

TABLE V

COMPARISON OF DISC AGGREGATION STRATEGIES. f DENOTES THE SYMMETRIC FUNCTION AND D DENOTES THE DISCWISE AGGREGATION.



Steps	0	1	2	4	7	10
RAND	25.7	35.5	40.2	45.5	52.2	57.4
HPCS	25.7	36.4	36.8	40.9	49.0	50.9
Ours	25.7	42.2	49.5	53.6	57.9	62.0

TABLE VI

COMPARISON OF DISC SELECTION STRATEGIES (I) RANDOM (RAND), (II) HIGHEST POINT COUNT SELECTION (HPCS) AND (III) OURS.

heavy region extraction methods (VCCS, SV clustering) limiting them from being used on accumulated point clouds. Still to exploit sequential labeling, we construct a new pipeline where we use ReDAL and LiDAL to select a region to label, fit a convex hull onto this region and label all points within the hull in the accumulated point cloud sequence. This allows our method, which is developed directly with sequential labeling at its core, to better compare to baseline approaches under sequential labeling. In Tab. IV we observe that our discwise AL strategy outperforms both work. This is because both LiDAL and ReDAL often pick distant sparse regions that show a higher uncertainty in the current frame, however when sequential labeling, also contain a large number of high certainty points from other frames.

Disc Aggregation: In Tab. V, we compare our disc aggregation strategy of voxelwise pooling followed by discwise sum to 2D analogous methods of discwise mean, and discwise sum. We further compare the effects of different symmetric functions within our 2-stage setup.

Firstly we observe the limitations of common aggregation strategies. The sum performs similarly, if not better at low step counts to our proposed methods. As it is biased towards frames with high point counts, the high information yield at early frames allows easier learning of head classes. This trait later becomes its own demise as the distribution of labels becomes increasingly long tailed, limiting the models capability to learn

Method	mIoU	car	bicycle	motorcycle	truck	other vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other ground	building	fence	vegetation	trunk	terrain	pole	traffic sign
SPVCNN (Dense labels) [26]	63.8	97.1	35.2	64.6	72.7	64.3	69.7	82.5	0.2	93.5	50.8	81.0	0.3	91.1	63.5	89.2	66.1	77.2	64.1	49.4
RAND	57.4	95.1	8.2	56.1	69.5	31.1	62.7	80.8	0.0	91.8	34.2	76.8	0.1	91.4	61.7	88.6	55.5	76.5	62.9	48.3
HPCS	50.9	93.9	5.3	31.7	29.6	28.3	53.9	51.0	0.0	92.3	40.0	78.5	0.0	90.2	55.2	87.2	57.5	76.1	60.6	36.0
DiAL (Supervised)	62.0	96.1	30.1	62.7	69.7	54.4	68.5	85.5	0.0	89.7	45.7	76.8	0.1	91.0	61.8	88.3	66.5	75.3	64.8	50.7
Δ (RAND)	+4.6	+1.0	+21.9	+6.6	+0.2	+23.3	+5.8	+4.7	0.0	-2.1	+11.5	0.0	0.0	-0.4	+0.1	-0.3	+11.0	-1.2	+1.9	+2.4
DiAL (Semi-supervised)	63.7	96.0	42.0	63.7	82.5	46.1	74.6	86.3	0.0	93.8	44.0	79.6	13.0	89.9	56.8	87.1	66.0	71.7	65.2	51.6
Δ (Supervised)	+1.7	-0.1	+12.1	+1.0	+12.8	-8.3	+6.1	+0.8	0.0	+4.1	-1.7	+2.8	+12.9	-1.1	-5.0	-1.2	+0.5	-3.6	+0.4	+0.9

TABLE VII

COMPARISON OF THE CURRENT BASELINES AT STEP 10 (51 DISCS) TO OUR PROPOSED ACTIVE LEARNING FRAMEWORK WITH AND WITHOUT SEMI-SUPERVISED TRAINING. ALONGSIDE THE MIOU, WE PROVIDE CLASSWISE IOU METRICS TO SHOW THE ADVANTAGES OF ACTIVE LEARNING.

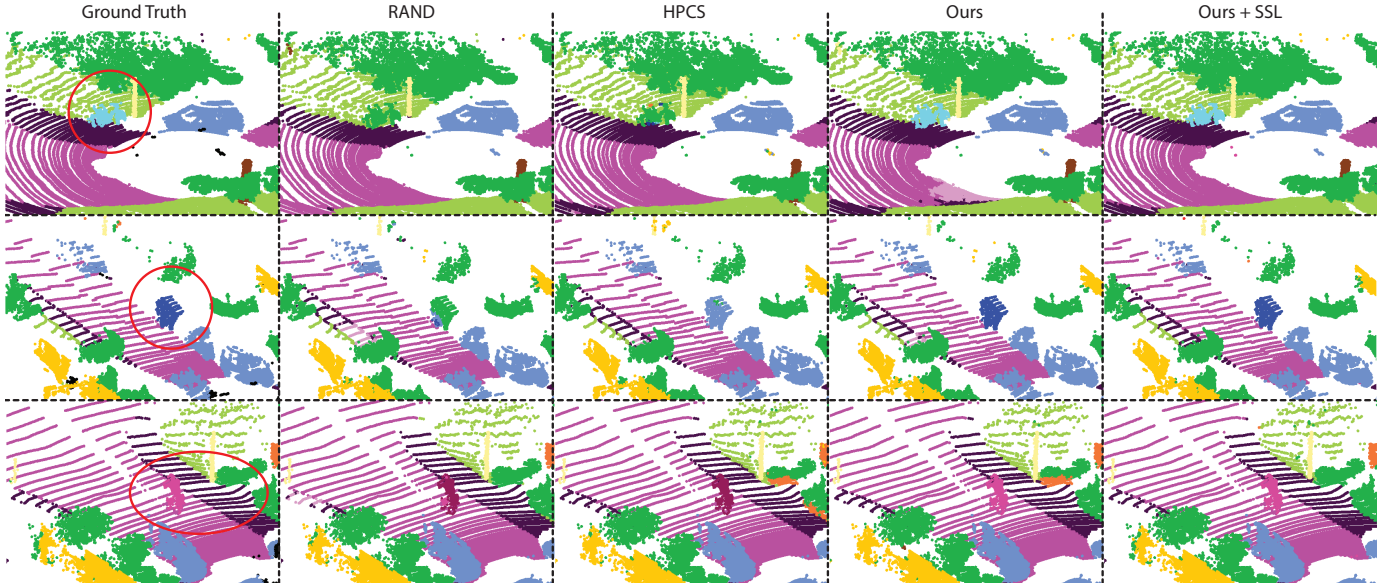


Fig. 4. Example results from the SemanticKITTI validation set comparing current baseline approaches to our proposed discwise active learning pipeline trained in a supervised and semi-supervised manner (+ SSL).

Steps	1	2	4	7	10
RAND	35.5	37.0	44.0	48.8	53.8
Ours	40.8	49.5	51.9	56.5	59.0
Δ	+5.3	+12.5	+7.9	+7.7	+5.2

TABLE VIII

COMPARISON WITH MINKNET [27]

Steps	1	2	4	7	10
RAND	33.6	36.6	44.2	50.6	52.8
Ours	37.7	39.4	47.3	52.4	54.4
Δ	+4.1	+2.8	+3.1	+1.8	+1.6

TABLE IX

COMPARISON ON SCRIBBLEKITTI [4].

Steps	1	2	4	7	10
RAND	34.6	36.2	42.3	48.0	51.1
Ours	41.7	44.1	47.4	50.8	55.0
Δ	+7.1	+7.9	+5.1	+2.8	+3.9

TABLE X

COMPARISON ON NUSCENES [5].

difficult examples.

As expected, the max operation outperforms the min during the initial iterations as high density discs often lead to multi-class voxels that show a large variation of uncertainty. While again the high information yield at early frames helps, it comes with a detriment at later stages. As the min operation aims to maximize the minimum information, it expects each voxel to contribute substantially to the learning process, i.e. it favors purely uncertain voxels within the selected discs. These result in a more even distribution thus allowing the strategy to outperform others at later steps.

Disc Selection: In Tab. VI, we compare our disc sampling strategy of mixed-integer linear optimization to maximize total MI to (i) random selection of discs, and to (ii) the heuristic approach of highest point count sampling (HPCS), where each iteration we sample N discs that are associated with the frames

that contain the highest number of points. In all baselines, we prohibit intersections between discs to prevent trivial and harming solutions of selecting neighboring discs around the argmax⁴. As seen, our AL pipeline outperforms baseline methods with considerable margins⁵.

In Tab. VII we also provide a classwise breakdown of the IoU for the final step. Here it can be seen that through intelligent disc selection, our AL approach allows substantial improvements for mainly tail classes such as bicycle, motor-

⁴Given a LiDAR sensor that operates at 10Hz and an ego-vehicle that moves in a straight line at 50km/h, each disc will be separated by only 2m. This is further reduces at lower velocities and while turning. Thus when intersections are allowed, if a disc at x_i yields the highest uncertainty, it is very likely that the neighboring discs at x_{i-1} or x_{i+1} would yield the second highest, resulting in a severe overlap of voxels.

⁵Comparing our step 10 performance (51 discs) to 1% results from Tab. II (95 discs) we see that with almost half the discs, we still maintain an advantage over random sampling.

cycle, other-vehicle, person and bicyclist. In Fig. 4 we also provide visual results that demonstrate this effect.

Semi-Supervised Training: In Tab. VII we further compare our active learning model trained on labeled points to our semi-supervised approach. As seen (in step 10) the performance is improved significantly (1.7% mIoU), with major improvements seen for classes such as bicycle, truck, person and the difficult other-ground category. Our AL method performs 99.8% relative to the baseline model trained on a fully labeled dataset, with only 51 discs labeled (0.5% annotation cost of framewise labeling the entire dataset).

Model Independence: Our proposed discwise AL pipeline is completely model independent. In Tab. VIII we provide additional results to demonstrate this by showing that our approach continues to significantly outperform baseline random sampling while using MinkowskiNet [27].

Other Datasets: Our method works well against the random baseline for both ScribbleKITTI [4] and nuScenes [5] which we showcase in Tab. IX and Tab. X respectively.

VI. CONCLUSION

In this work we tackle the problem of active learning for LiDAR semantic segmentation while considering common labeling techniques that minimize labeling times. To this end, we propose querying discs to utilize the sequential nature of LiDAR frames. With discwise AL, we label an accumulated region only once to have it labeled on all frames. We further devise a new acquisition function that tackles bias introduced due to the variability of local and discwise point counts. Later we construct a mixed-integer linear program to provide a general solution to disc selection that considers intersections. Finally we improve the performance of our model via a semi-supervised approach that utilizes a mean teacher. Through extensive studies we show that while simple, our discwise AL framework is highly effective and can boast high performance gains compared to baseline approaches.

Limitations: To enable sequential labeling objects must remain static. In cases where multiple varying class object paths collide in global coordinates, the joint object cloud must be labeled on a per frame basis. Time savings are still substantial as most of the environment is dominated by static structures. Still, in our proposed active learning framework, we do not consider the additional cost of labeling colliding moving objects. We hope future work can tackle this issue.

REFERENCES

- [1] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *CVPR*, 2012.
- [2] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *CVPR*, 2020.
- [3] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *ICCV*, 2019.
- [4] O. Unal, D. Dai, and L. Van Gool, "Scribble-supervised lidar semantic segmentation," in *CVPR*, 2022.
- [5] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, "Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking," *RA-L*, 2022.
- [6] S. Desai and D. Ghose, "Active learning for improved semi-supervised semantic segmentation in satellite images," in *WACV*, January 2022.
- [7] I. C. Saidu and L. Csató, "Active learning with bayesian unet for efficient semantic image segmentation," *Journal of Imaging*, 2021.
- [8] F. Ozdemir, Z. Peng, P. Fuerstahl, C. Tanner, and O. Goksel, "Active learning for segmentation based on bayesian sample queries," *Knowledge-Based Systems*, vol. 214, 2021.
- [9] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *ICML*, 2017.
- [10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NIPS*, 2017.
- [11] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *EMNLP*, 2008.
- [12] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, 2016.
- [13] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *ECML*. Springer, 2006.
- [14] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *CVPR*, 2009, pp. 2372–2379.
- [15] R. Hwa, "Sample selection for statistical parsing," *Computational linguistics*, vol. 30, no. 3, 2004.
- [16] A. Kirsch, J. Van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *NIPS*, vol. 32, 2019.
- [17] T.-H. Wu, Y.-C. Liu, Y.-K. Huang, H.-Y. Lee, H.-T. Su, P.-C. Huang, and W. H. Hsu, "Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation," in *ICCV*, 2021.
- [18] Z. Hu, X. Bai, R. Zhang, X. Wang, G. Sun, H. Fu, and C.-L. Tai, "Lidal: Inter-frame uncertainty based active learning for 3d lidar semantic segmentation," in *ECCV*. Springer, 2022.
- [19] M. Liu, Y. Zhou, C. R. Qi, B. Gong, H. Su, and D. Anguelov, "Less: Label-efficient semantic segmentation for lidar point clouds," in *ECCV*. Springer, 2022, pp. 70–89.
- [20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.
- [21] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *NIPS*, 2017.
- [22] O. Unal, L. Van Gool, and D. Dai, "Improving point cloud semantic segmentation by learning 3d object detection," in *WACV*, 2021.
- [23] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *ICRA*, 2018.
- [24] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in *ICRA*, 2019.
- [25] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *CVPR*, 2021.
- [26] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *ECCV*, 2020.
- [27] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *CVPR*, 2019.
- [28] L. Jiang, S. Shi, Z. Tian, X. Lai, S. Liu, C.-W. Fu, and J. Jia, "Guided point contrastive learning for semi-supervised point cloud semantic segmentation," in *ICCV*, October 2021.
- [29] M. Cheng, L. Hui, J. Xie, and J. Yang, "Sspc-net: Semi-supervised semantic 3d point cloud segmentation network," in *AAAI*, 2021.
- [30] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.
- [31] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *MICCAI*. Springer, 2017.
- [32] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "We don't need no bounding-boxes: Training object class detectors using only human verification," in *CVPR*, 2016.
- [33] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2022. [Online]. Available: <https://www.gurobi.com>