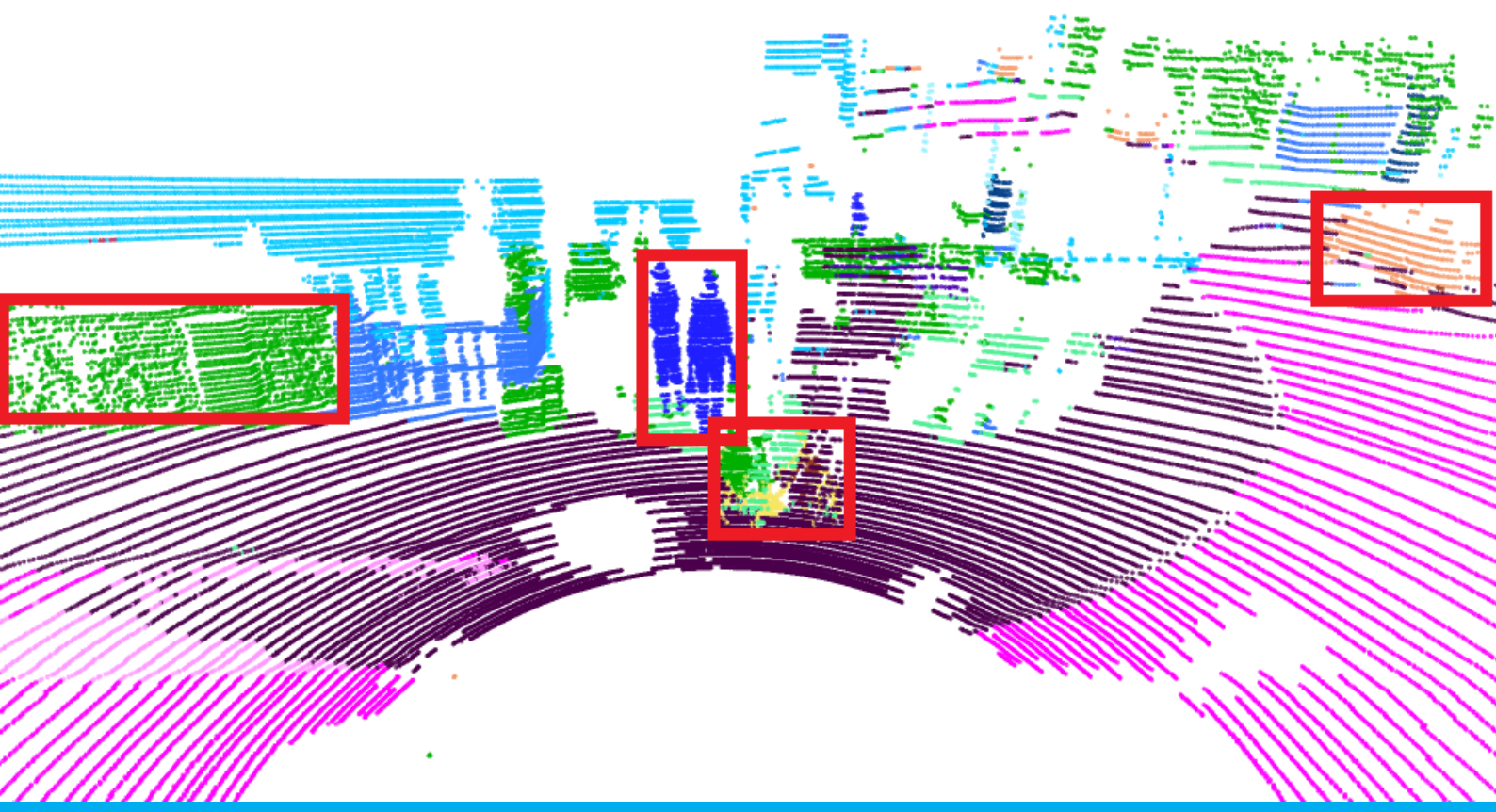
BaSAL: Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation

MSc Thesis

Jiarong Wei



TU Delft

BaSAL: Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation

MSc Thesis

by

Jiarong Wei

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on May 25, 2023

Student number:	5471524	
Project duration:	December, 1, 2022 – May 25, 2023	
Thesis Committee:	Dr. Holger Caesar,	TU Delft, daily supervisor
	Dr. J.F.P. Kooij,	TU Delft

This thesis is confidential and cannot be made public until June 31, 2023.

TU Delft

Acknowledgement

I would like to express my sincere gratitude to all the people that helped and accompanied me during these two years.

First and foremost, I extend my deepest appreciation to my supervisor, Holger. His consistent guidance has been invaluable throughout the process of my master thesis. His patience, expertise, and constructive feedback have played a pivotal role in my professional development.

Also, I would like to acknowledge the professors, and staff at TU Delft. Their commitment to education and their support in creating a conducive learning atmosphere have been instrumental in my growth during these two years.

Special thanks to the members of TU Delft intelligent vehicles group. Their collaborative spirit and open exchange of ideas have greatly enhanced my understanding of the field. Engaging in discussions and reviewing the latest research papers within this group have broadened my view and enriched my knowledge.

I would also like to express my appreciation to my friends. Together, we have faced challenges, shared experiences, and created lasting memories. Their companionship and shared pursuit of knowledge motivate me to live a better life.

Lastly, I extend my heartfelt thanks to my parents and family for their unconditional support throughout my master journey. Their unwavering belief in my abilities, constant encouragement and sacrifices have been a source of inspiration for me to overcome the obstacles.

Jiarong Wei
Delft, May 2023

BaSAL: Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation

Jiarong Wei

Abstract

Active learning has been proposed as a solution to mitigate the expensive and time-consuming process of annotating large-scale autonomous driving datasets. The process typically involves a model initialization phase, followed by multiple iterations aiming at selecting the most informative data based on the initial model. However, we find two problems that have not explicitly been solved in this process. First, current large-scale autonomous driving datasets suffer from the class imbalance problem, yet no strategy has been specifically designed to address this issue. Second, selecting the initial data from an entirely unlabeled pool of data, commonly referred to as the cold start problem of active learning, remains challenging. In this study, we propose a Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation (BaSAL) framework to address these problems. Our framework introduces a novel size-based clustering pipeline that uses a size-based cluster for non-ground points or a grid for ground points as a basic query unit. The cluster sizes are heavily correlated with their semantic classes, allowing us to more actively control the class distribution of the selected data. We also propose a warm start strategy to alleviate the cold start problem. Different from the commonly used random point cloud scan selection for model initialization, our warm start strategy selects data from the basic query units and can improve the initial model by a large margin. Experiments show that our approach can achieve over 95% of the performance of fully supervised learning while using only 5% of data, outperforming existing active learning methods on SemanticKITTI [4] and getting on par performance with the state-of-the-art method on nuScenes [7].

1. Introduction

Autonomous driving has become increasingly popular, and a robust perception system is crucial for the safe operation of autonomous vehicles. Among the various sensors used in the perception system, LiDAR has emerged as a reliable and high-resolution tool. To cope with complex driving conditions, semantic segmentation is a crucial component of LiDAR-based perception systems as it allows for the identification of different classes in a LiDAR point cloud.

Developing an effective LiDAR semantic segmentation network that can perform well requires access to large-scale autonomous driving datasets like SemanticKITTI [4] and nuScenes [7]. The cost to label these datasets is high, and deep learning networks are data-hungry, making it challenging to reduce the amount of data required. To address this problem, researchers have proposed some methods [57] [24] [30] [42] that leverage active learning to reduce the input data needed while maintaining competitive performance compared to fully supervised methods. Active learning is an effective machine learning technique to mitigate the onerous annotation burden. It typically involves an initialization phase, where a model is trained using the initial data selected from the unlabeled set, and an iterative phase, where the most informative data is iteratively selected for label acquisition based on the model trained in the previous iteration.

Recent research on active learning [57] [24] [30] [42] mainly focuses on improving the iterative phase, specifically the data information metric. However, two problems remain unsolved. Firstly, none of the methods have explicitly designed a strategy to address the class imbalance problem. In large-scale autonomous driving datasets, some classes are often overrepresented, while others are underrepresented. This problem becomes even more intractable in the active learning use case where annotation budgets are limited. Secondly, the cold start problem of selecting the initial data from an entirely unlabeled pool of data remains a challenge. In the context of active learning applied in LiDAR semantic segmentation, current works [57] [24] apply VCCS [38], an over-segmentation algorithm that divides the point cloud according to point connectivity. While this approach is effective in dividing the point cloud into connected regions, it does not address the class imbalance problem. Also, they ignore the cold start problem and simply randomly select point cloud scans to train the initial

model. The model trained on data selected by cold start often has low performance, which in turn affects the data selection process in later iterations.

In response to the aforementioned challenges, we propose a Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation (BaSAL) framework. Our approach utilizes a size-based point cloud clustering pipeline to provide an effective solution to the class imbalance problem inherent in large-scale autonomous driving datasets. By implementing a warm start strategy to initialize a high-performing model and subsequently selecting informative data combining softmax uncertainty and feature diversity ([43]), our method achieves competitive performance relative to fully supervised approaches with only a small amount of data. Our framework consists of three primary components: the size-based point cloud clustering pipeline, the warm start strategy, and the data information measure.

Our size-based point cloud clustering pipeline aims to achieve greater point cloud granularity in a class balanced manner. Specifically, we transform the basic query unit from a point cloud scan to a non-ground size-based cluster (small, medium, and large), corresponding to different classes of objects with different sizes in the real world, or a ground grid. By controlling the distribution of these size-based clusters or grids, the class imbalance problem can be alleviated.

We also provide a solution for the cold start problem. Instead of randomly selecting several point cloud scans, we select a certain amount of data from the non-ground small, medium, and large cluster sets and the ground grid set for model initialization. This approach leads to a much stronger initial model compared to the baseline random selection method. By initializing a more accurate model, subsequent active learning iterations can more effectively select the most useful information.

For the information measure in later active learning iterations, we combine softmax entropy [53] [54] and feature diversity (CoreSet) [43]. Softmax entropy is a commonly used measure in active learning to select the data that the model is most uncertain of. Feature diversity (CoreSet) [43] is a method to select data that is diverse in the feature space. By combining these two measures, we aim to more effectively select the most informative data under a fixed annotation budget.

Experiments show that by using only 5% of points, our method can achieve over 95% of the performance of fully supervised learning, outperforming existing active learning methods on SemanticKITTI [4] and matching the state-of-the-art method LiDAL [24] on nuScenes [7] dataset. Our ablation studies also verify the effectiveness of each component in our method.

In summary, our contribution can be outlined as follows:

- We propose a novel size-based point cloud clustering pipeline that increases the granularity of the dataset and enables the alleviation of the class imbalance problem.

- We propose a warm start strategy for initializing a high-performance model, which is proven to contribute to the overall model improvement.

- Experiments show that our framework can achieve competitive results compared to the fully supervised approach with significantly reduced data input, outperforming existing active learning methods on SemanticKITTI [4] and getting on par performance with the state-of-the-art method on nuScenes [7].

2. Related work

2.1. LiDAR Semantic Segmentation

LiDAR semantic segmentation is a task to allocate a label for each point in the point cloud. Current LiDAR semantic segmentation methods can be divided into four categories: point-based, voxel-based, projection-based, and fusion-based. Mainstream point-based methods directly read from the raw point cloud to learn the point features following the design of the pioneering PointNet [40]. Later work improves the PointNet baseline by improving the efficiency [23], learning the local neighborhood features [41], or designing a new convolution kernel [51]. Typical voxel-based methods represent the point cloud by assigning each point to a corresponding voxel. Except for dividing the point cloud into regular grids, some methods propose different ways of point cloud partition and design new 3D convolution methods. PolarNet [63] applies polar partitioning replacing Cartesian partitioning, making the points more evenly distributed. Cylinder3d [64] represents the point cloud using cylinder partition and designs an asymmetrical 3d convolution to extract features from the cylinders. Projection-based methods transform point clouds to a 2d representation in either Bird-Eye-View (BEV) [1] or Range-View (RV) [11] [34] [55] [56] [58], then apply the highly optimized 2D convolution to make their network efficient. Fusion-based methods try to combine the advantages of point-based, voxel-based, and projection-based methods while avoiding their shortcomings. Although sometimes computationally expensive, these methods generally can produce competitive results [62] [19] [32] [50] [59].

2.2. Active Learning

Active learning is a subfield of machine learning that aims at minimizing the cost of obtaining data labels. Active

learning algorithms mainly vary in their query strategies. According to [44], active learning has three typical scenarios, including membership query synthesis, stream-based selective sampling, and pool-based active learning. Recent research about active learning mainly focuses on pool-based active learning, which assumes that there is a small set of labeled data and a large pool of unlabeled data available and iteratively draws queries from the unlabeled pool.

2.2.1 Query Strategies for Active Learning

Pool-based active learning typically uses uncertainty-based methods, diversity-based methods, or their combination as the query strategy. Uncertainty-based strategies characterize model uncertainty, measured by entropy measurement [22], ensemble methods [5], Gaussian process [25], Bayesian approach [17], a learned loss prediction [60], or a discriminator score [48]. Diversity-based methods claim that while uncertainty-based strategies are often effective, they can also be prone to selecting similar items to appear in the same querying batch. [43] converts batch selection into a core-set construction problem to ensure diversity in the labeled data. [26] [2] consider model uncertainty and diversity at the same time. [36] selects the most representative samples while avoiding repeatedly labeling samples in the same cluster.

Recent active learning research stands on the shoulder of previous work and has more task-specific designs. ReDAL [57] is the pioneering work to apply active learning on LiDAR semantic segmentation for large-scale autonomous driving datasets. ReDAL divides the point cloud into regions and selects those regions with high uncertainty and diversity. Following ReDAL, LiDAL [24] selects the informative point cloud by exploiting inter-frame uncertainty among LiDAR frames. [30] proposes a diversity-based active learning method that enforces spatial diversity and temporal diversity for 3D object detection. [42] combines detection entropy and prediction entropy as the selection criteria for P&P tasks. [47] considers divergence score as their information measure and proves that inconsistencies in model predictions across viewpoints can provide a reliable measure of uncertainty. All the work mentioned in various directions aims at a more efficient active learning framework design.

2.2.2 Class imbalance problem

Class imbalance is a natural problem for datasets. A common solution is resampling, typically undersampling the overrepresented classes and oversampling the underrepresented classes or other more specific resampling techniques. Undersampling is one method to address the class imbalance problem by removing examples from the majority class and is reported in [15] to be effective in learning on large imbalanced datasets. Oversampling is another method to address the class imbalance problem by resampling the small class through random sampling or duplication until it is not underrepresented. [8] combines the method of oversampling the minority class and under-sampling the majority class and achieves better classifier performance. One-sided sampling is another resampling method similar to undersampling, in which redundant and borderline training examples are identified and removed from training data. [27] reports that one-sided sampling is effective in learning with two-class large imbalanced datasets.

Except for resampling techniques, active learning also provides solutions to the class imbalance problem. [3] summarizes two kinds of active learning techniques to cope with the class imbalance problem: Density-Sensitive Active Learning (information density [45], pre-clustering [37], alternate density-sensitive heuristics [12]) and Skew-Specialized Active Learning [52]. [14] argues that their proposed Support Vector Machine (SVM) active learning strategy, which queries a small pool of data at each iterative step, can be a more efficient alternative solution to the class imbalance problem compared to traditional resampling methods. [65] also analyzes the effect of the common resampling techniques including under-sampling and over-sampling, and proposes a bootstrap-based over-sampling (BootOS) method that works better than ordinary over-sampling in active learning for Word Sense Disambiguation (WSD) task. [6] alleviates the class imbalance problem in image classification tasks by proposing a general optimization framework. CBAL [13] proposes a class balance active learning framework, which adds equal numbers of instances from both object classes (cancer and non-cancer) to solve the class imbalance problem in histopathology. From our knowledge, there is no solution explicitly proposed for the class imbalance problem in active learning use case for large-scale autonomous driving datasets.

2.2.3 Cold start problem

The problem of how to choose the initial set of data for annotation from an entirely unlabeled pool of data is known as the cold start problem for active learning. The cold start problem is first observed in recommender systems, where solutions to remedy the insufficient information due to the lack of user history are needed. This concept is then used in other fields. In natural language processing (NLP), [61] seeks a solution for the cold start problem by pre-training models using self-supervision and it attributes the cold start problem to model instability and data scarcity. [9] attributes the possible causes of the cold start problem to a biased and outlier initial query and applies contrastive learning as a solution. [28] explores the effectiveness of the K-center algorithm to select the initial queries. Similarly, [39] shows that

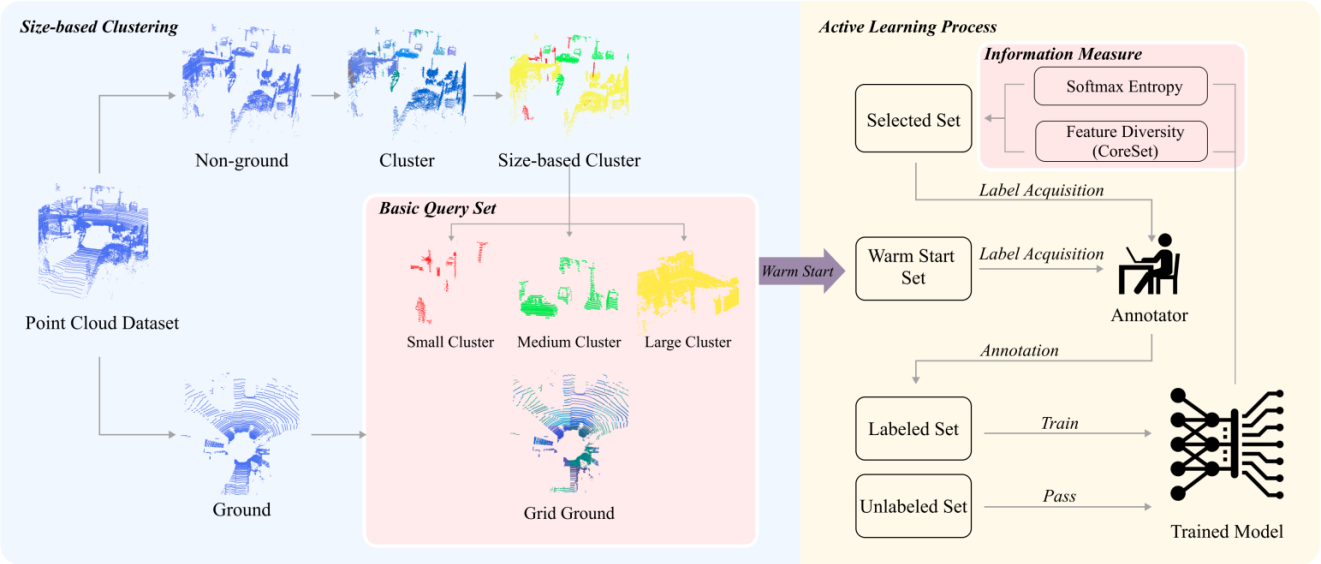


Figure 1. Overview of BaSAL. Our framework consists of a size-based clustering module and the active learning process. The size-based clustering module transforms all the point clouds in the dataset into non-ground size-based clusters (small, medium, and large clusters) and ground grids. For the active learning process, we first use our warm start strategy to initialize a model. All the unlabeled data is then passed through the model, and the most informative data is selected by our data information measure (softmax entropy, feature diversity (CoreSet) [43]). The selected data is then labeled and the labeled data is used to train a new model. The labeled data and the unlabeled data are updated accordingly.

a simple K-means clustering algorithm works fairly well at the beginning of active learning, as it is capable of covering diverse classes and selecting a similar number of data per class. Most recently, a series of studies [20] [21] [49] [35] continue to propose new strategies for selecting the initial query from the entire unlabeled data and highlight that typical data (defined in varying ways) can significantly improve the learning efficiency of active learning at a low budget. However, for active learning applied in LiDAR semantic segmentation, as far as we know, there is no study discussing how to select the initial data for label acquisition.

3. Method

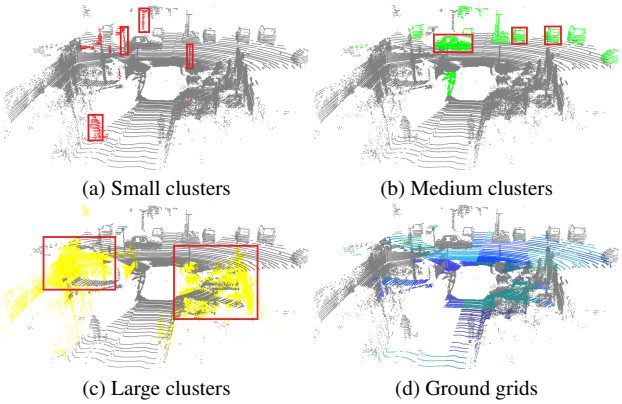
3.1. Overview

Our BaSAL framework aims to solve the class imbalance problem and the cold start problem of active learning. The overview of our BaSAL framework is shown in Figure 1. We introduce a size-based point cloud clustering pipeline to partition the dataset into non-ground small, medium, large cluster sets and a ground grid set as basic query sets. As the cluster sizes are heavily correlated with their semantic classes, by setting the amount of data selected from these sets, we can have a more precise control of class balance. For the active learning process, we first implement a warm start strategy to improve the performance of the initial model to cope with the cold start problem. For iterations after the initialization, we combine softmax entropy and feature diversity (CoreSet [43]) to measure the information of all the unlabeled data. We select the highly informative data, query their label and retrain the model. Then the labeled, unlabeled basic query sets are updated accordingly.

3.2. Size-based Point Cloud Clustering

Our size-based point cloud clustering pipeline is shown in Algorithm 1. The input of Algorithm 1 includes the raw point cloud dataset $\mathcal{D}$ and the boundary lengths L_1, L_2. The output is non-ground size-based cluster sets ($\mathcal{D}_{cs}$, $\mathcal{D}_{cm}$, $\mathcal{D}_{cl}$), and a ground grid set $\mathcal{D}_{gg}$. There are in total four procedures in our algorithm, including ground plane removal, ground partition, clustering, and size division.

The main target of our size-based point cloud clustering pipeline is to better control the class distribution in the data used for model training. Specifically, we need to develop a method to separate different classes in a point cloud based solely on their geometric features. Real-world autonomous driving environments typically consist of the road class (road, sidewalk, parking area, etc.) and the object class (tree, person, bicyclist, etc.). To first separate these two classes, we use a ground plane removal algorithm to divide all the point clouds in the dataset into ground points $\mathcal{D}_g$ consisting of road classes and non-ground points $\mathcal{D}_{ng}$ consisting of object classes. The ground points $\mathcal{D}_g$ are then



(a) Small clusters (b) Medium clusters

(c) Large clusters (d) Ground grids

Figure 2. This figure shows an example of our basic query units consisting of small, medium, large clusters, and ground grids. Small clusters normally contain small, thin, or occluded structures, e.g., pedestrians, road signs, poles, trunks, occluded cars. Medium clusters typically consist of vehicles. Large clusters mainly have buildings and vegetation. Ground grids mostly consist of road, sidewalk, terrain, and other-ground.

divided into grids $\mathcal{D}_{gg}$ for greater granularity.

For the non-ground points $\mathcal{D}_{ng}$ representing object classes, we use a clustering algorithm to divide them into clusters $\mathcal{D}_{cng}$. Given the significant variability of point cloud geometric features both within and across classes, we select size as the criterion to further separate the clusters. The size l of a cluster is defined by the sum of the length, width, and height of its bounding box. The cluster size l is then classified according to the boundary lengths L_1, L_2. Clusters with size $0 < l < L_1$, $L_1 <= l < L_2$, and $l > L_2$ are classified into small cluster set $\mathcal{D}_{cs}$, medium cluster set $\mathcal{D}_{cm}$, and large cluster set $\mathcal{D}_{cl}$ respectively.

An example of our size-based clusters is shown in Figure 2. Without knowing the labels of any point in the dataset, the minimum query unit for a dataset becomes a non-ground size-based cluster or a ground grid, corresponding to different classes of objects with different sizes in the real world. The class component of the size-based clusters is shown in Table 4 in the Supplementary Material, and it shows that our size-based clusters are heavily correlated with their semantic classes. By properly controlling the amount of different size-based clusters, we can effectively mitigate the class imbalance problem.

3.3. Warm start initialization

Recent work [57] [24] that applies active learning in LiDAR semantic segmentation randomly selects 1% total frames for model initialization and subsequently selecting informative data based on the initialized model, which is a cold start. Although this approach is straightforward, the cold start model often yields low performance and the selected informative data by this model will also be affected.

Algorithm 1 Size-based Point Cloud Clustering

Input: The raw point cloud dataset $\mathcal{D}$; The boundary length L_1, L_2 to divide the small, medium, and large size-based clusters.

procedure GROUND PLANE REMOVAL
  $\mathcal{D}_g, \mathcal{D}_{ng} \leftarrow$ *Ground plane removal on* $\mathcal{D}$
procedure GROUND PARTITION
  $\mathcal{D}_{gg} \leftarrow$ *Normal grid partition on* $\mathcal{D}_g$
procedure CLUSTERING
  $\mathcal{D}_{cng} \leftarrow$ *Clustering on* $\mathcal{D}_{ng}$
procedure SIZE DIVISION
  $N \leftarrow$ *Number of clusters in* $\mathcal{D}_{cng}$
  for $i \leftarrow 0, 1, 2, \ldots, N$ **do**
    $l \leftarrow$ *Cluster bounding box size*
    if $l > 0$ **and** $l < L_1$ **then**
      $\mathcal{D}_{cs} \leftarrow \mathcal{D}_{cng}[i]$
    if $l >= L_1$ **and** $l < L_2$ **then**
      $\mathcal{D}_{cm} \leftarrow \mathcal{D}_{cng}[i]$
    if $l >= L_2$ **then**
      $\mathcal{D}_{cl} \leftarrow \mathcal{D}_{cng}[i]$

Output: Non-ground small, medium, large size-based cluster sets $\mathcal{D}_{cs}$, $\mathcal{D}_{cm}$, $\mathcal{D}_{cl}$, and a ground grid set $\mathcal{D}_{gg}$

To improve the traditional frame-level cold start, we develop a warm start strategy that improves the initialization model by a large margin.

Our warm start strategy is straightforward: we allocate a budget $x\%$ for selecting a proportion of data from the basic query sets. As shown in Equation 1, we divide the warm start budget $x\%$ to p_{cs}, p_{cm}, p_{cl}, and p_{gg}. Then we randomly select data from $\mathcal{D}_{cs}$, $\mathcal{D}_{cm}$, $\mathcal{D}_{cl}$, and $\mathcal{D}_{gg}$ according to these budgets. Note that the budget here refers to the proportion of the number of points selected relative to the total number of points in a dataset. We denote the resulting selected data as $\mathcal{D}_{small}$, $\mathcal{D}_{medium}$, $\mathcal{D}_{large}$, and $\mathcal{D}_{ground}$. The $*$ operator here refers to the process of selecting p proportion of data from a basic query set $\mathcal{D}$.

$$\begin{aligned} x &= p_{cs}+p_{cm}+p_{cl}+p_{gg} \\ \mathcal{D}_{small} &= p_{cs} * \mathcal{D}_{cs} \\ \mathcal{D}_{medium} &= p_{cm} * \mathcal{D}_{cm} \\ \mathcal{D}_{large} &= p_{cl} * \mathcal{D}_{cl} \\ \mathcal{D}_{ground} &= p_{gg} * \mathcal{D}_{gg} \end{aligned} \tag{1}$$

We then combine these four parts of data as our warm start data $\mathcal{D}_{init}$ to train the initial model, as shown in Equation 2. The warm start data, compared to the randomly selected frames, balances the class distribution and improves the initial model performance. With a stronger initial

model, the active learning algorithm can more effectively select informative data in later iterations.

$$\mathcal{D}_{init} = \mathcal{D}_{small} \cup \mathcal{D}_{medium} \cup \mathcal{D}_{large} \cup \mathcal{D}_{ground} \quad (2)$$

3.4. Information measure

We combine softmax entropy [22] and feature diversity (Coreset) [43] to select the most informative data for label acquisition. Softmax entropy aims to select the data that the model is most uncertain of, as these points are likely to contain useful information. Feature diversity (Coreset) prioritizes unlabeled data that is far from the labeled data to ensure diversity. By combining these two methods, we aim at a balance between uncertainty and diversity.

3.4.1 Softmax entropy

To compute the softmax entropy, we pass all the unlabeled data through the model trained in the previous iteration and obtain the per-point prediction probability p. Assuming there are N_u unlabeled basic query units q_j ($j = 1, 2, 3, \ldots, N_u$), each containing n points, we compute the average entropy of a point k ($k = 1, 2, 3, \ldots, n$) in a basic unit q_j using Equation 3.

$$s_{q_j}^{(u)} = -\frac{1}{n} \sum_{k=1}^{n} p_k \log(p_k) \quad (3)$$

This yields a ranking for uncertainty, denoted as $r_q^{(u)}$. $r_q^{(u)} = rank(s_{q_j}^{(u)}) = \{r_{q_1}^{(u)}, r_{q_2}^{(u)}, \ldots, r_{q_j}^{(u)}\}$, where $r_{q_j}^{(u)}$ represents the uncertainty score ranking of q_j.

3.4.2 Feature diversity

We apply the Coreset [43] approach as the feature diversity measure. Its main idea is to condense the information of the whole dataset into a small subset so that the model trained on the subset can have on par performance with the fully supervised network. To achieve this, CoreSet selects the samples from the unlabeled dataset that are the furthest away from the labeled dataset in the feature space. In our implementation, we choose the output of the layer before the last classification layer of the encoder-decoder network as point features. Specifically, we denote $h(x_l; \theta)$, $h(x_u; \theta)$ as the prediction output of the labeled data and the unlabeled data. $h(x_l; \theta)$ has shape (N_l, K) and $h(x_u; \theta)$ has shape (N_u, K), where N_l, N_u denote the number of samples in the labeled set and the unlabeled set, and K denotes the feature dimension. The steps to calculate the feature diversity score are shown in Equation 4. We first compute the Euclidean distance between the samples in the unlabeled set and those in the labeled set D_{q_j,q_i}. Then, for each sample in the unlabeled set, its diversity score is computed by summing up the feature space distance between it and all the samples in the labeled set.

$$\begin{aligned} D_{q_j,q_i} &= ||h(x_u; \theta)_j - h(x_l; \theta)_i||_2 \\ s_{q_j}^{(d)} &= \sum_{j=1}^{N_u} D_{q_j,q_i} \end{aligned} \quad (4)$$

Then we have a ranking for feature diversity, denoted as $r_q^{(d)}$. $r_q^{(d)} = rank(s_{q_j}^{(d)}) = \{r_{q_1}^{(d)}, r_{q_2}^{(d)}, \ldots, r_{q_j}^{(d)}\}$, where $r_{q_j}^{(d)}$ represents the diversity score ranking of q_j.

3.4.3 Combination

The selected data from the unlabeled set for label acquisition is determined by the uncertainty ranking $r_q^{(u)}$ and the diversity ranking $r_q^{(d)}$. We acquire the combined rankings as shown in Equation 5.

$$\mathcal{Q}^* = argsort_q \left(\frac{1}{r_q^{(u)}} + \frac{1}{r_q^{(d)}} \right)_{1:K} \quad (5)$$

With the combined ranking of all the basic units in the unlabeled set, we select the top K basic units as the query set $\mathcal{Q}^*$ until the labeling budget is exhausted.

4. Experiments

4.1. Datasets and Evaluation Metric

SemanticKITTI [4] SemanticKITTI is a large-scale autonomous driving dataset based on the KITTI Vision Odometry Benchmark [18] recorded in Germany. It contains 22 sequences (seq 00 - 07, seq 09 - 10 for training, seq 08 for validation, and seq 11-22 for test). We perform our active learning framework on the training sequences and test the model performance on the validation sequences.

nuScenes [7] nuScenes is the first large-scale autonomous driving dataset containing the full sensor suite, which was recorded in Boston and Singapore. It contains 1000 scenes (700 scenes for training, 150 scenes for validation, and 150 scenes for testing). We perform our active learning framework on the training scenes and test the model performance on the validation scenes.

Metrics Following the typical LiDAR semantic segmentation settings, we also use the mean intersection over union (mIoU) metric to measure the model performance. 19 classes are used to evaluate SemanticKITTI and 16 classes are used to evaluate nuScenes.

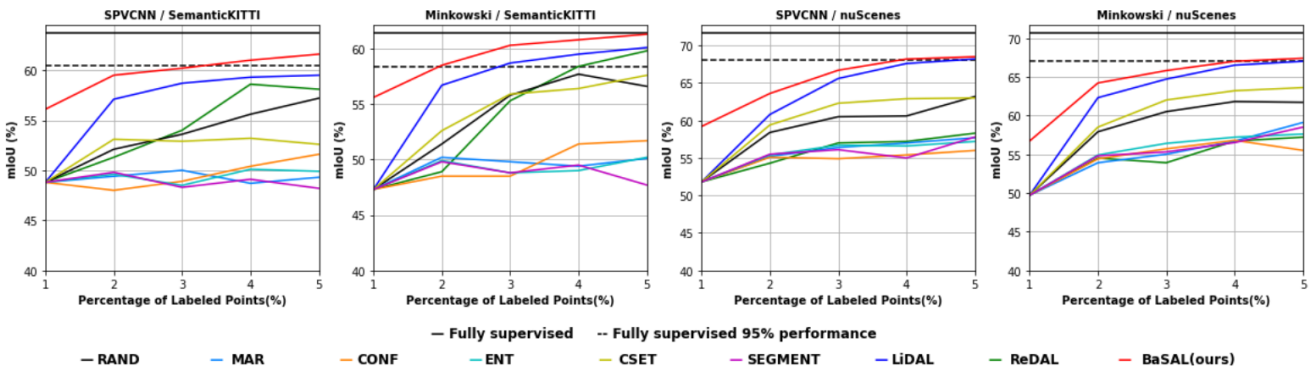


Figure 3. **Experiment results of different active learning strategies on SemanticKITTI [4], nuScenes [7] using SPVCNN [50], Minkowski [10] network.** We compare our BaSAL strategy with other existing works. Our method outperforms all existing active selection approaches on SemanticKITTI and gets on par performance with the state-of-the-art active learning method LiDAL [24] on nuScenes.

4.2. Implementation Details

4.2.1 Network Architectures

Following ReDAL [57] and LiDAL [24], for better comparison, we also use SPVCNN [50] based on point-voxel CNN, and MinkowskiNet [10] based on sparse convolution, as our backbone networks.

4.2.2 Baseline Active Learning Methods

We select nine baseline methods for comparison, including random point cloud selection (RAND), softmax confidence (CONF) [53], softmax margin (MAR) [53], softmax entropy (ENT) [53], MC-dropout (MCDR) [16], core-set selection (CSET) [43], segment-entropy (SEGENT) [31], ReDAL [57] and LiDAL [24].

4.2.3 Active Learning Protocol

Our active learning protocol follows part of the settings of ReDAL [57] and LiDAL [24]. For the model initialization step, different from their random selection, we apply our warm start strategy to select p_{cs}, p_{cm}, p_{cl}, p_{gg} of points from the small, medium, large cluster sets and the ground grid set, in total x_{init} of data. The selected data is then labeled and trained as model initialization. The active learning process after initialization consists of K rounds of the following actions: 1. Train the model on the current labeled set 2. Select x_{active} of data from the current unlabeled set for label acquisition according to our information measure. 3. Update the labeled set and the unlabeled set.

The labeling budget here is measured by the percentage of the labeled points. For both SemanticKITTI [4] and nuScenes [7], we set $x_{init} = 1\%$, $p_{cs} = 0.25\%$, $p_{cm} = 0.25\%$, $p_{cl} = 0.25\%$, $p_{gg} = 0.25\%$, $K = 4$, and $x_{active} = 1\%$. To ensure the reliability of the results, all the experiments are performed three times and the average results are reported.

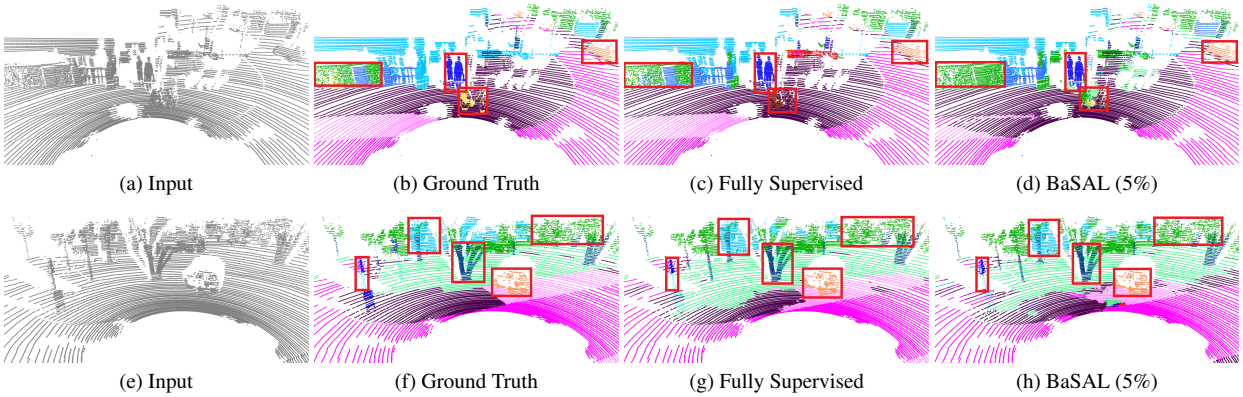
4.2.4 Size-based Point Cloud Clustering

We implement our size-based point cloud clustering pipeline following Algorithm 1 introduced in Section 3.2. The input of the Algorithm 1 includes the raw point cloud dataset $\mathcal{D}$ and boundary lengths L_1, L_2. We choose $L_1 = 5m$, $L_2 = 10m$ for both datasets. For Ground Plane Removal, we apply Patchwork++ [29], a fast and robust state-of-the-art ground segmentation algorithm on 3D point clouds. For Clustering, we apply HDBSCAN [33], a density-based clustering algorithm that is effective at discovering clusters of varying densities. We use it to divide the non-ground points into clusters, setting the 'min_cluster_size', the parameter representing the minimum number of points required to form a cluster, and 'min_sample', the parameter to control the density of clusters, to 50, 1 for SemanticKITTI [4] and 50, 10 for nuScenes [7] respectively, because nuScenes is generally sparser than SemanticKITTI. The clusters with l smaller than $L_1 = 5m$, between $L_1 = 5m$ and $L_2 = 10m$, and larger than $L_2 = 10m$ are classified to small cluster set $\mathcal{D}_s$, medium cluster set $\mathcal{D}_m$, and large cluster set $\mathcal{D}_s$ respectively. For Ground Division, the ground points are divided into grids with size $10m$ by $10m$.

4.3. Main Results

4.3.1 Comparisons among different active selection strategies

We verify our method compared with 9 other active selection strategies mentioned in Section 4.2.2. The implementation details of these methods are shown in the Supplementary Material.



(a) Input (b) Ground Truth (c) Fully Supervised (d) BaSAL (5%)

(e) Input (f) Ground Truth (g) Fully Supervised (h) BaSAL (5%)

Figure 4. Visualization for the inference results on SemanticKITTI [4] validation set with SPVNAS [50] network architecture. With our active learning strategy, the model can recognize different classes (person, bicycle, car, trunk, road sign, vegetation, sidewalk, road, etc.). For the bicycle on the sidewalk ((b), (c), (d)), our model performs even better than the fully supervised model.

The experiment results are shown in Figure 3. The data of all the baseline experiments including ReDAL and LiDAL is taken from LiDAL [24]. In each subplot, the x-axis means the percentage of the labeled points and the y-axis represents the mIoU achieved by the network. Our proposed BaSAL method outperforms all the methods on SemanticKITTI dataset and gets on par performance with LiDAL [24] on nuScenes dataset.

For SemanticKITTI, we can achieve 98% performance of fully supervised learning on SPVCNN and 100% performance of fully supervised learning on MinkowskiNet using only 5% of labels, outperforming all existing methods. For nuScenes, we can achieve 95% performance of fully supervised learning on both networks, with on par performance as the state-of-the-art method LiDAL [24].

The qualitative results are shown in Figure 4. Compared to the fully supervised model, our model can also recognize most of the object classes and ground classes. For the small classes (e.g., the bicycle on the sidewalk), our model performs even better than the fully supervised model.

4.3.2 Warm start

As shown in Figure 3, our warm start strategy far outperforms the random selection baseline when using 1% of data. We explain why our warm start is much more effective than the cold start random initialization in Table 2. The 'Class distribution' row in the table reveals that SemanticKITTI [4] has a highly imbalanced class distribution. For example, the 'road' and 'vegetation' classes account for 22.01% and 23.18% of the whole dataset, while the 'bicyclist' and 'motorcyclist' classes only have 0.013% and 0.0037%. The data selected by our warm start strategy, compared to that by the cold start random initialization, scales the amount of data of the underrepresented classes by a large margin (motorcycle, person, bicyclist, motorcyclist, etc.). While some classes may be slightly sacrificed (road), the overall performance of all the classes is considerably improved.

4.4. Ablation study

4.4.1 Component analysis

In this section, we conduct a series of controlled experiments to prove the effectiveness of our components. All our experiments are conducted on the SemanticKITTI [4] validation set evaluating the model trained on SPVCNN [50] network with 5% data. The results are shown in Table 1.

FL	SC	WS	ENT	FD	mIoU(%)
✓					57.2
	✓				60.8
	✓	✓			61.4
	✓	✓	✓		61.8
	✓	✓	✓	✓	62.1

Table 1. Ablation study. FL: Frame Level strategy; SC: Size-based Point Cloud Clustering; WS: Warm Start; ENT: Entropy-based Uncertainty; FD: Feature Diversity [43].

For FL, we randomly select 1% of total points based on a frame level and randomly increase 1% of data in each iteration (the same process as the random selection baseline). Changing from FL to SC, we implement our size-based point cloud clustering framework, but using the traditional frame-level model as 1% initialization (without warm start), and randomly select data from non-ground size-based cluster sets and ground grid sets. This brings the largest improvement (3.6%) for our model compared to the random selection baseline. For SC+WS, we use our warm start strategy as 1% model initialization, keeping other settings the same as SC, and it contributes to 0.6% improvement for

	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Class distribution(1e-2)	4.23	0.017	0.040	0.20	0.23	0.035	0.013	0.0037	19.88	1.47	14.39	0.39	13.27	7.24	26.69	0.60	7.81	0.29	0.061
1% Cold Start(1e-4)	4.23	0.017	0.040	0.20	0.23	0.035	0.013	0.0037	19.88	1.47	14.39	0.39	13.27	7.24	26.69	0.60	7.81	0.29	0.061
1% Warm Start(1e-4)	13.63	0.030	0.49	0.11	0.39	0.27	0.12	0.050	10.59	0.79	7.35	0.24	13.31	5.67	29.69	3.86	5.51	1.42	0.33
Ratio Warm / Cold (data)	3.22	1.80	**12.29**	0.55	1.66	**7.83**	**9.77**	**13.13**	0.53	0.54	0.51	0.61	1.00	0.78	1.11	**6.39**	0.71	**4.98**	**5.32**
1% Cold Start mIoU(%)	93.10	12.09	15.26	54.89	22.20	6.66	48.85	0.29	89.08	25.28	72.40	0.27	86.56	48.57	86.61	55.93	73.56	60.15	35.23
1% Warm Start mIoU(%)	94.08	29.33	64.09	43.28	22.99	68.83	85.20	2.59	85.58	19.07	69.76	1.94	85.94	42.82	83.58	66.44	63.67	57.26	43.46
Ratio Warm / Cold (mIoU)	1.01	**2.43**	**4.20**	0.79	1.04	**10.33**	1.74	**8.93**	0.96	0.75	0.96	**7.18**	0.99	0.88	0.97	1.19	0.87	0.95	1.23

Table 2. This table explains why our warm start strategy far outperforms the random selection baseline. 'Class distribution' shows the class distribution of SemanticKITTI. Note that the class distribution here is defined by the number of points of a class relative to the total number of points of this class in the dataset. '1% Cold Start' and '1% Warm Start' show the class distribution of 1% cold start and 1% warm start initialization. '1% Cold Start mIoU' and '1% Warm Start mIoU' show the per class mIoU of the model trained using the cold start strategy and the warm start strategy. 'Scaling factor (data)' and 'Scaling factor (mIoU)' compares the scaling of the class distribution and the mIoU value between 1% cold start and 1% warm start.

our model. For SC+WS+ENT, we use softmax entropy as our information measure to select data for label acquisition and improve our model by 0.4%. For SC+WS+ENT+FD, we combine softmax entropy and feature diversity (CoreSet) [43] as our information measure, and it brings 0.3% improvement to our model.

4.4.2 Warm Start with different proportion

Strategy	WS1+4IT	WS2+3IT	WS3+2IT	WS4+1IT	WS5
WS mIoU(%)	56.1	58.9	59.8	60.0	61.0
WS+IT mIoU(%)	62.1	**62.4**	60.8	61.4	61.6

Table 3. Performance of using different proportion of data as warm start with SPVCNN [50] network on SemanticKITTI [4]. WSn+(5-n)IT ((n = 1, 2, 3, 4, 5)) means applying the warm start strategy with n percent of data and continuing (5-n) active learning iterations. WS mIoU(%) means the model performance of using the warm start strategy only, without looking at later active learning iterations; WS+IT mIoU(%) means the resulted model performance of the whole process including warm start and later iterations.

With 5% of the labeling budget, a common active learning practice is to use 1% data for model initialization and perform active learning for four iterations. However, this may not be the optimal solution. With the same budget, we carry on some experiments with different proportion of initialization. We test using 1%, 2%, 3%, 4% of data for warm start initialization, then train the model for 4, 3, 2, 1 iterations, and directly using 5% of data without any active learning iteration. Our experiment result is shown in Table 3. We find that for SemanticKITTI [4] using SPVCNN [50] network, the optimal solution in our setting is using 2% of warm start and training the model for 3 iterations. This warm start strategy has the highest performance because it can reach a relatively high performance (58.9%) for model initialization, which can help to improve the efficiency to select useful data in further active learning iterations. At the same time, it also leaves 3% of the annotation budget for active learning iterations, which allows the model to iteratively select informative data that benefits the model.

5. Conclusion

In this paper, we present a Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation (BaSAL) framework. Aiming at mitigating the class imbalance problem in large-scale autonomous driving datasets, we propose a size-based clustering pipeline that divides the non-ground point cloud into clusters and classifies them according to the cluster size. As the cluster sizes are heavily correlated with their semantic classes, by controlling the amount of the size-based clusters used for model training, our active learning scheme can effectively mitigate the class imbalance problem. To solve the cold start problem in active learning, we propose a warm start strategy and explore the effect of using different proportion of data for model initialization. We also combine model uncertainty and feature diversity (CoreSet [43]) in the iterations after the warm start initialization to select informative data for label acquisition. Experiments show that our framework can outperform existing methods on SemanticKITTI and get on par performance with the state-of-the-art method LiDAL [24] on nuScenes. For future work, we believe that our handcrafted ground-plane removal and clustering steps can be improved, e.g. via a learning-based approach. Also, our ablation studies show that there should be an optimal point to switch from the warm start initialization to the active learning iterations. Relevant experiments can be done to explore this optimal point to further improve our model.

References

[1] Eren Erdal Aksoy, Saimir Baci, and Selcuk Cavdar. Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. In *2020 IEEE intelligent vehicles symposium (IV)*, pages 926–932. IEEE, 2020. 2

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 3

[3] Josh Attenberg and Şeyda Ertekin. Class imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, page 101–149, 2013. 3

[4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1, 2, 6, 7, 8, 9, 14

[5] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018. 3

[6] Javad Zolfaghari Bengar, Joost van de Weijer, Laura Lopez Fuentes, and Bogdan Raducanu. Class-balanced active learning for image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, page 1536–1545, 2022. 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 6, 7, 14, 15

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 3

[9] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. *arXiv preprint arXiv:2210.02442*, 2022. 3

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 7, 14, 15

[11] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. *arXiv preprint arXiv:2003.03653*, 2020. 2

[12] Pinar Donmez and Jaime G. Carbonell. Paired-sampling in density-sensitive active learning. 2008. 3

[13] Scott Doyle, James Monaco, Michael Feldman, John Tomaszewski, and Anant Madabhushi. A class balanced active learning scheme that accounts for minority class problems: Applications to histopathology. In *OPTIMHisE Workshop (MICCAI)*, page 19–30, 2009. 3

[14] Seyda Ertekin, Jian Huang, and C. Lee Giles. Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 823–824, 2007. 3

[15] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004. 3

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 7

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

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[19] Martin Gerdzhev, Ryan Razani, Ehsan Taghavi, and Liu Bingbing. Tornado-net: multiview total variation semantic segmentation with diamond inception module. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9543–9549. IEEE, 2021. 2

[20] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022. 4

[21] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022. 4

[22] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 3, 6

[23] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 2

[24] Zeyu Hu, Xuyang Bai, Runze Zhang, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Lidal: Inter-frame uncertainty based active learning for 3d lidar semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 248–265. Springer, 2022. 1, 2, 3, 5, 7, 8, 9, 14

[25] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 3

[26] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019. 3

[27] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer, 1997. 3

[28] Adrian Lang, Christoph Mayer, and Radu Timofte. Best practices in pool-based active learning for image classification. 2021. 3

[29] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3D point cloud. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 13276–13283, 2022. 7

[30] Zhihao Liang, Xun Xu, Shengheng Deng, Lile Cai, Tao Jiang, and Kui Jia. Exploring diversity-based active learning for 3d object detection in autonomous driving. *arXiv preprint arXiv:2205.07708*, 2022. 1, 3

[31] Y Lin, G Vosselman, Y Cao, and MY Yang. Efficient training of semantic point cloud segmentation via active learning. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2:243–250, 2020. 7, 13

[32] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020. 2

[33] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017. 7

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

[35] Vishwesh Nath, Dong Yang, Holger R Roth, and Daguang Xu. Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 297–308. Springer, 2022. 4

[36] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 3

[37] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 3

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

[39] Kossar Pourahmadi, Parsa Nooralinejad, and Hamed Pirsiavash. A simple baseline for low-budget active learning. *arXiv preprint arXiv:2110.12033*, 2021. 3

[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

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

[42] Sean Segal, Nishanth Kumar, Sergio Casas, Wenyuan Zeng, Mengye Ren, Jingkang Wang, and Raquel Urtasun. Just label what you need: fine-grained active selection for perception and prediction through partially labeled scenes. *arXiv preprint arXiv:2104.03956*, 2021. 1, 3

[43] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2, 3, 4, 6, 7, 8, 9, 13

[44] Burr Settles. Active learning literature survey. 2009. 3

[45] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, page 1070–1079, 2008. 3

[46] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 13

[47] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, page 9433–9443, 2020. 3

[48] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 3

[49] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022. 4

[50] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 2, 7, 8, 9, 14, 15

[51] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[52] Katrin Tomanek and Udo Hahn. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, page 105–112, 2009. 3

[53] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014. 2, 7, 13

[54] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2, 13

[55] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud.

In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893. IEEE, 2018. 2

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

[57] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15510–15519, 2021. 1, 3, 5, 7, 13

[58] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020. 2

[59] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021. 2

[60] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019. 3

[61] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. *arXiv preprint arXiv:2010.09535*, 2020. 3

[62] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *European Conference on Computer Vision*, pages 644–663. Springer, 2020. 2

[63] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. 2

[64] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020. 2

[65] Jingbo Zhu and Eduard Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, page 783–790, 2007. 3

Supplementary Material for BaSAL: Class Balanced Warm Start Active Learning for LiDAR Semantic Segmentation

1. Baseline methods

1.1 Random selection (RAND)

Random selection method randomly selects a portion of point cloud scans from the unlabeled dataset for label acquisition. It is commonly used as a baseline for active learning methods.

1.2 Core-set (CSET)

Core-set [43] is a diversity-based active learning strategy aiming to select a small subset so that a model trained on the selected subset has a similar performance to that trained on the whole dataset. This method first extracts the feature of each sample. Then, it selects a small number of samples from the unlabeled dataset that is the furthest away from the labeled dataset in the feature space for label acquisition. In the implementation, we choose the middle layer of the encoder-decoder network as the feature.

1.3 Least confidence sampling (CONF)

Least confidence sampling methods [54] query the sample whose prediction has the least confidence. As shown in Equation 6, they first pass all the unlabeled point cloud scans X with N through the model θ, then calculate the confidence of predicted class label ($\hat{y_n^1}$) for all points and produce the score for a point cloud scan (S_{CONF}) by averaging the value of all points in a scan. After that, the point cloud scans with the least confidence score in the unlabeled set are selected for label acquisition.

$$S_{CONF} = \frac{1}{N}\sum_{i=1}^{N} P(\hat{y_n^1}|X;\theta) \tag{6}$$

1.4 Softmax Entropy (ENT)

Entropy is an indicator to measure the information of a probability distribution in the information theory [46]. Some previous active learning approaches query samples with the highest entropy value in the predicted probability [53]. As shown in Equation 7, given a point cloud scan X with N points and a model θ, we calculate the softmax entropy value for all points and produce the score for a point cloud scan (S_{ENT}) by averaging the value of all points in a scan. After that, we select a portion of point cloud scans with the largest entropy in the unlabeled dataset for label acquisition.

$$ENT = -\sum P^p(c) log(P^p(c)) \tag{7}$$

where $P^p(c)$ is the probability of point p belonging to class c.

1.5 Softmax Margin (MAR)

Some previous active learning methods [53] query instances with the smallest model decision margin, which is the predicted probability difference between two most likely class labels. As shown in Equation, given a point cloud scan X with N points and a model θ, we calculate the difference between the two most likely class labels for all points and produce the score for a point cloud scan (S_{MAR}) by averaging the value of all points in a scan. After that, we select a portion of point cloud scans with the largest score in the unlabeled dataset for label acquisition.

$$S_{MAR} = \frac{1}{N}\sum_{i=1}^{N} P(\hat{(}y_n^1)|X;\theta) - P(\hat{(}y_n^2)|X;\theta) \tag{8}$$

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

1.6 Segment-entropy (SEGMENT)

Lin [31] proposes segment entropy to measure the point cloud information in the deep active learning pipeline. This method assumes that each geometrically related area should share similar semantic annotations. Therefore, it calculates the entropy of the distribution of predicted labels in a small area to estimate model uncertainty.

1.7 ReDAL

Region-based and diversity-aware active learning (ReDAL) [57] is the pioneer to apply active learning to LiDAR semantic segmentation. It divides a 3D scene into sub-scene regions and then estimates the region

Size-based clusters	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign
Small	1.93,	4.04,	**65.35**,	0.64,	3.69,	39.61,	**50.86**,	**67.68**,	0.03,	0.04,	0.12,	0.52,	1.31,	0.61,	1.65,	27.86,	0.39,	23.68,	29.17
Medium	**79.44**,	13.66,	14.29,	13.97,	37.01,	22.89,	26.92,	23.51,	0.1,	0.24,	0.18,	1.45,	4.93,	2.34,	7.51,	39.04,	1.58,	16.57,	27.55
Large	15.85,	**69.5**,	15.36,	**81.57**,	**57.34**,	33.0,	20.39,	3.96,	0.22,	0.19,	1.1,	8.34,	**88.53**,	**82.55**,	**79.53**,	29.87,	7.91,	**55.43**,	43.11
Ground	2.79,	12.81,	5.0,	3.82,	1.95,	4.5,	1.83,	4.86,	**99.65**,	**99.53**,	**98.6**,	**89.7**,	5.23,	14.51,	11.31,	3.24,	**90.12**,	4.33,	0.18

Table 4. This table shows the component of our non-ground small, medium, large cluster sets, and the ground grid set. Note that the component here refers to the proportion of the number of points of a class relative to the total number of points of this class in the dataset.

information utilizing three metrics: softmax entropy, color discontinuity, and structural complexity. With the estimated region information scores, this method further designs a diversity-aware selection algorithm to avoid visually similar regions appearing in a querying batch for labeling. The results in our work about ReDAL are all taken from their official paper.

1.8 LiDAL

Following the footsteps of ReDAL, LiDAL [24] is the recent state-of-the-art work that applies active learning to LiDAR semantic segmentation. Their core idea is that a well-trained model should generate robust results irrespective of viewpoints for scene scanning and thus the inconsistencies in model predictions across frames provide a very reliable measure of uncertainty for active sample selection. The results in our work about LiDAL are all taken from their official paper.

2. More results

2.1 The class distribution of the size-based clusters

Our size-based point cloud clustering pipeline divides the whole dataset into non-ground size-based cluster sets and a ground grid set. The components of these sets are shown in Table 4. The small cluster set contains most of the motorcycle, bicyclist, and motorcyclist class. The medium cluster set contains most of the car class. The large cluster set contains most of the bicycle, truck, other-vehicle, building, fence, vegetation, and pole class. The ground grid set contains most of the road, parking, sidewalk, other-ground, and terrain class. The bicycle class here is mostly classified into large cluster sets mainly because in the real world, the still bicycles are often parked together in a group or laid around other structures. Another bicycle class 'bicyclist' that is typically small in the real world is mostly classified into small cluster sets.

2.2 Training results

This section shows the detailed training results. For SemanticKITTI [4], we set the train batch size to 10. We first train the warm start data for 100 epochs, then finetune the model for 30 epochs for each iteration. For nuScenes [7], we set the train batch size to 30. The warm start data is first trained for 200 epochs and then finetuned for 150 epochs for each iteration.

Methods	Init (1%)	2%	3%	4%	5%
RAND	48.8	52.1	53.6	55.6	57.2
MAR	48.8	49.4	50.0	48.7	49.3
CONF	48.8	48.0	48.9	50.4	51.6
ENT	48.8	49.6	48.5	50.1	49.9
CSET	48.8	53.1	52.9	53.2	52.6
SEGMENT	48.8	49.8	48.3	49.1	48.2
ReDAL	48.8	51.3	54.0	58.6	58.1
LiDAL	48.8	57.1	58.7	59.3	59.5
BaSAL	56.1	58.6	59.5	61.6	62.1

Table 5. Mean intersection over union scores on SemanticKITTI [4] validation set with SPVCNN [50].

Methods	Init (1%)	2%	3%	4%	5%
RAND	47.3	51.4	55.8	57.7	56.6
MAR	47.3	50.2	49.8	49.4	50.1
CONF	47.3	48.5	48.5	51.4	51.7
ENT	47.3	49.9	48.8	49.0	50.2
CSET	47.3	52.6	55.9	56.4	57.6
SEGMENT	47.3	49.8	48.8	49.5	47.7
ReDAL	47.3	56.7	58.7	59.5	60.1
LiDAL	47.3	51.4	55.8	57.7	56.6
BaSAL	55.6	58.5	60.3	60.8	61.3

Table 6. Mean intersection over union scores on SemanticKITTI [4] validation set with MinkowskiNet [10].

Methods	Init (1%)	2%	3%	4%	5%
RAND	51.8	58.4	60.5	60.6	63.2
MAR	51.8	55.2	56.4	57.0	57.7
CONF	51.8	55.1	54.9	55.4	56.0
ENT	51.8	55.4	56.7	56.6	57.2
CSET	51.8	59.4	62.3	62.9	63.0
SEGMENT	51.8	55.5	56.1	55.0	57.8
ReDAL	51.8	54.3	57.0	57.2	58.3
LiDAL	51.8	60.8	65.6	67.6	68.2
BaSAL	59.2	63.6	66.7	68.2	68.5

Table 7. Mean intersection over union scores on nuScenes [7] validation set with SPVCNN [50].

Methods	Init (1%)	2%	3%	4%	5%
RAND	49.7	57.9	60.5	61.8	61.7
MAR	49.7	53.9	55.0	56.7	59.1
CONF	49.7	54.4	55.7	56.8	55.5
ENT	49.7	54.9	56.4	57.2	57.6
CSET	49.7	58.5	62.0	63.2	63.6
SEGMENT	49.7	54.8	55.3	56.5	58.5
ReDAL	49.7	54.5	53.9	56.7	57.2
LiDAL	49.7	62.3	64.7	66.5	67.0
BaSAL	56.7	64.2	65.8	67.0	67.4

Table 8. Mean intersection over union scores on nuScenes [7] validation set with MinkowskiNet [10].