

SELECT: A Submodular Approach for Active LiDAR Semantic Segmentation

Ruiyu Mao, Sarthak Kumar Maharana, Xulong Tang, Yunhui Guo, *Member, IEEE*

Abstract—LiDAR-based semantic segmentation plays a vital role in autonomous driving by enabling detailed understating of 3D environments. However, annotating LiDAR point clouds is extremely costly and, requires assigning semantic labels to millions of points with complex geometry structure. Active Learning (AL) has emerged as a promising approach to reduce labeling costs by querying only the most informative samples. Yet, existing AL methods face critical challenges when applied to large-scale 3D data: outdoor scenes contain an *overwhelming number of points* and suffer from severe class imbalance, where rare classes have far fewer points than dominant classes.

To address these issues, we propose SELECT, a voxel-centric submodular approach tailored for active LiDAR semantic segmentation. Our method targets both scalability problems and class imbalance through three coordinated stages. First, we perform *Voxel-Level Submodular Subset Selection*, which efficiently identifies representative voxels without pairwise comparisons, ensuring scalability. Second, we estimate *Voxel-Level Model Uncertainty* using Monte Carlo dropout, aggregating point-wise uncertainties to identify informative voxels. Finally, we introduce *Submodular Maximization for Point-Level Class Balancing*, which selects a subset of points that enhances label diversity, explicitly mitigating class imbalance. Experiments on SemanticPOSS, SemanticKITTI, and nuScenes benchmarks demonstrate that SELECT achieves superior performance compared to prior active learning approaches for 3D semantic segmentation.

Index Terms—LiDAR Semantic Segmentation, Active Learning, Submodular Optimization

I. INTRODUCTION

LIDAR-based semantic segmentation [85], [86] plays a crucial role in real-world applications such as autonomous driving, robotics, urban mapping, and augmented reality [21]–[23], as it provides a detailed understanding of the surrounding environment by providing a semantic label for each point in a LiDAR scan. Compared to image-based semantic segmentation, LiDAR-based scene understanding offers unique advantages, including robustness to lighting variations and precise 3D spatial information.

While LiDAR sensors can capture millions of points per second, this high data density presents a major bottleneck: the cost of obtaining point-level semantic annotations. In contrast to 2D images—where annotators can assign labels via bounding boxes or polygons [29], [65] with relative ease—3D point cloud annotation is inherently more complex. Annotators must interpret sparse and irregular point distributions from multiple viewpoints, often without clear object boundaries

The authors are with the Department of Computer Science, the University of Texas at Dallas, Richardson, Texas, 75080, USA. Yunhui Guo is the corresponding author. E-mail: {ruiyu.mao, sarthak.maharana, xulong.tang, yunhui.guo}@utdallas.edu.

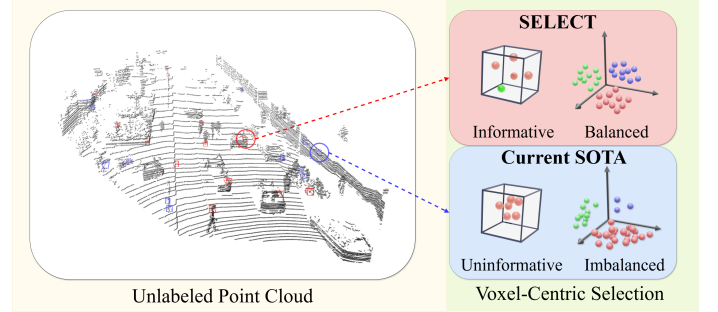


Fig. 1: The proposed SELECT employs a unified submodular approach that ensures the selection of points that are both informative and balanced in label distribution. Here, informativeness refers to points that either belong to rare classes, where the model exhibits low confidence, or are situated along object boundaries, where semantic ambiguity is high. This stands in contrast to state-of-the-art (SOTA) methods [84], which often struggle to jointly capture these critical aspects for active LiDAR semantic segmentation.

[15], [53]. For instance, annotating the SemanticKITTI dataset required over 1,700 hours for just 43,000 scans [4]. This makes exhaustive manual annotation not only labor-intensive but also economically unsustainable, particularly for large-scale autonomous driving datasets.

To address the high annotation cost, active learning (AL) [2], [31] has become a widely adopted solution. By starting with a small labeled set and iteratively selecting the most informative samples for annotation, AL strategies aim to maximize model performance within a fixed labeling budget. The informativeness of a chosen sample can be assessed in several ways, such as by calculating sampling uncertainty using maximum Shannon entropy [58], [82] or by estimating model changes. Other strategies focuses on selecting the most *representative* samples to avoid redundancy, employing methods like greedy coresot algorithms [60], [61], clustering-based techniques [26], [27], [62], [74], and spatial structure based algorithms [49], [97].

However, directly applying traditional active learning techniques to LiDAR semantic segmentation faces three key challenges: **1)** Each scene contains hundreds of thousands of points, making point-wise informativeness estimation computationally expensive. **2)** Accurately identifying informative points is difficult, especially when the model’s predictions are unreliable in sparse or ambiguous regions [84]. **3)** Outdoor LiDAR scenes are naturally imbalanced—frequent classes like *car* may dominate, while rare classes like *pedestrian* or *bicycle* are severely underrepresented [19], [28]. This imbalance leads

to biased model training, where rare classes are prone to misclassifications.

A recent voxel-centric active learning method, Annotator [84], was introduced to address the scalability issue in LiDAR-based semantic segmentation. Instead of selecting individual points, Annotator partitions each LiDAR scan into voxels and evaluates each voxel’s informativeness using a metric called the Voxel Confusion Degree (VCD). This metric quantifies the diversity of predicted labels within a voxel, under the assumption that higher class confusion indicates greater informativeness. Voxels with the highest VCD scores are selected for annotation. By operating at the voxel level, Annotator is more computationally efficient than traditional point-based method [88] or regional-based method [49], and better suited to the sparse nature of outdoor LiDAR scenes.

However, while Annotator improves scalability and informativeness, it still suffers from two major limitations. First, it lacks an effective mechanism to estimate model uncertainty, which is crucial for selecting truly uncertain and informative regions. Second, it does not account for class imbalance, leading to over-selection of frequent classes and neglect of rare but critical ones. These gaps reduce the method’s effectiveness in complex outdoor environments where both uncertainty and class diversity are central challenges.

In this paper, we introduce *SELECT*, a unified submodular framework for voxel-centric active LiDAR semantic segmentation (see Fig. 1). *SELECT* is specifically designed to address the aforementioned limitations through three tightly integrated stages: **1) Voxel-Level Submodular Subset Selection:** To reduce computational cost, *SELECT* employs a submodular function to efficiently select a subset of semantically informative voxels, without the need for costly pairwise comparisons. This design ensures scalability to large-scale and sparsely distributed LiDAR datasets. **2) Voxel-Level Model Uncertainty Estimation:** To enhance the selection of uncertain regions, *SELECT* uses Monte Carlo dropout [66] to measure uncertainty across point-wise label distributions within each voxel. This approach enables more reliable identification of ambiguous regions. **3) Submodular Maximization for Point-Level Class Balancing:** To address class imbalance, *SELECT* incorporates a submodular diversity objective to prioritize the selection of points that improve label distribution balance across rare and frequent classes.

Together, these components form a robust and scalable framework that jointly optimizes for informativeness, uncertainty, and label diversity. Compared to the state-of-the-art active learning methods for LiDAR semantic segmentation, *SELECT* achieves consistent performance gains across multiple benchmarks: +8.17% mIoU on SemanticPOSS [20], +5.06% on SemanticKITTI [4], and +5.05% on nuScenes [1].

Our main contributions are summarized as follows:

- We identify the core limitations of current active learning approaches for LiDAR semantic segmentation—namely, poor scalability, weak uncertainty modeling, and an inability to address class imbalance.
- We propose *SELECT*, a scalable and uncertainty-aware submodular approach that ensures selected samples are both informative and balanced in class distribution.

- We conduct extensive experiments on three public outdoor LiDAR semantic segmentation benchmarks, demonstrating that *SELECT* significantly outperforms state-of-the-art methods in terms of segmentation accuracy under limited annotation budgets.

II. RELATED WORK

Despite the rapid progress in fully supervised approaches in recent years [5], [12], [14], [14], [32], [51], [52], label-efficient 3D semantic segmentation remains a relatively nascent field [56]. Researchers’ efforts in this area generally fall into three main categories: weakly supervised learning, unsupervised and self-supervised learning, and active learning.

Weakly Supervised Learning. In contrast to fully supervised learning, which requires exhaustive per-point annotations, weakly supervised learning aims to leverage limited annotations—such as sub-scene-level labels [34], [57], sparse point-level annotations, or bounding-box-level annotations [48]—to generate segmentation maps while significantly reducing annotation costs [3], [48]. While these methods are able to reduce the number of labeled points needed during training, some still require substantial annotations to achieve results comparable to fully supervised approaches. Additionally, the sub-scene-level annotation process remains challenging [15] due to the complex and unclear surface structures inherent in point cloud scans [53].

Unsupervised and Self-supervised Learning. In 3D point cloud semantic segmentation, unsupervised and self-supervised learning methods aim to reduce reliance on extensive manual annotations by automatically learning meaningful representations from unlabeled data [59]. These approaches can be broadly categorized based on their use of auxiliary modalities: image- or text-based methods [9], [11] and point-cloud-based methods [6]–[8], [24]. Although these methods have advanced the field by reducing dependence on labeled data, challenges remain. This cross-modal dependency often introduces additional issues, such as complex alignment problems between the 2D and 3D domains and increased computational complexity due to the fusion of multi-modal features [10].

Active Learning. With an abundance of unlabeled outdoor point cloud data, active learning has emerged as the de facto strategy for reducing annotation costs [2], [31], [33], by selecting samples for annotation by an oracle based on their uncertainty or representativeness, while maintaining model performance. There is a vast body of literature on active learning, encompassing methods based on CoreSet selection [60], [61], adversarial learning [30], and clustering techniques [26], [27]. Data sampling strategies from the unlabeled pool often rely on *uncertainty* measures [35]–[39], selecting samples based on the maximum Shannon entropy of the posterior probability, decision tree heuristics [40], variance of the training error [55], or the largest gradient magnitudes [54].

Active Learning for 3D Semantic Segmentation. Active learning has been extensively studied and applied to tasks such as image classification [89]–[91], 2D/3D object detection [50], [64], [72], [99], [100], and 2D semantic segmentation

[101], [102]. Recently, with growing interest in applying active learning to 3D semantic segmentation, uncertainty [98] and diversity [92], [97] have emerged as common strategies for point-cloud selection. In [98], the authors estimate inter-frame inconsistency to capture the inherent uncertainty present in LiDAR sequences. [88] divides the point cloud space into multiple components and sample a few points from each component, for labeling. Annotator [84] employs a computational geometry approach and suggests selecting informative point clouds from a voxel-centric perspective, turning out to be highly effective in learning with a limited budget.

Submodular Functions. Submodular functions and their optimization have been found to be widely used in data subset selection [73], [77], [78], [103], active learning [75], [76], [105], continual learning [79], and video summarization [67], [104]. The appealing properties of submodular functions aid in modeling both diversity and relevance when selecting classes or data subsets, while also preserving key features within each set. Submodularity ensures that as elements are added to a subset, the marginal gain diminishes, inherently encouraging a diverse and balanced selection. This, in turn, facilitates the maximization of overall relevance.

III. BACKGROUND

3D LiDAR Semantic Segmentation. LiDAR-based 3D semantic segmentation [85], [86] assigns semantic labels to individual points in a point cloud, delineating distinct objects and structures within the scene. We denote $\mathcal{P}_i = \left\{ \mathcal{V}_j = \{(P_k, c_k)\}_{k=1}^{N_j} \right\}_{j=1}^{N_i}$ as the i -th LiDAR scan which is also known as a point cloud in training dataset. N_i is the number of voxels in a given point cloud \mathcal{P}_i and N_j is the number of points in the voxel \mathcal{V}_j belonging to point cloud \mathcal{P}_i . Let P_k represent a single point in a particular point cloud as $P_k = (x_k, y_k, z_k, r_k)$, where x_k, y_k, z_k are the 3D Cartesian coordinates of the point and r_k denotes the reflectance rate generated by the LiDAR sensors. c_k is the corresponding ground truth semantic label of the point P_k . In 3D LiDAR semantic segmentation, a segmentation model S_ϕ , parameterized by ϕ , processes the raw point cloud data \mathcal{P}_i through a series of sparse convolutional blocks to extract and refine features f_k , for each point. The final classifier layer then classifies each point's corresponding features f_k into the semantic label c_k , where c_k belongs to the set $\{1, 2, 3, \dots, C\}$, with C representing the total number of semantic classes.

Voxel-centric Active Learning for LiDAR Semantic Segmentation. Following [84], in the initial stage, for each point cloud \mathcal{P}_i , a small number of voxels are randomly selected for labeling. The segmentation model S_ϕ undergoes warm-up training on the labeled voxels. Let the labeled voxels in the point cloud \mathcal{P}_i be denoted as D_L^i , and the remaining unlabeled voxels as D_U^i . During active learning, for each query iteration $q \in \{1, 2, \dots, Q\}$, the given active learning method selects N_q unlabeled voxels $D_q^i \subset D_U^i$ from each point cloud. The points within these selected voxels are then sent to human

annotators for labeling. The newly labeled voxels are merged with the existing labeled set:

$$D_L^i = D_L^i \cup D_q^i. \quad (1)$$

The active learning process continues until either the Q -th query round is reached or the total number of labeled points across all selected voxels exceeds the predefined annotation budget N_{budget} .

Using voxels as the selection unit not only improves selection efficiency but also reduces annotation difficulty. Instead of requiring annotators to manually distinguish between overlapping objects with fuzzy edges or complex surfaces across large portions of the full point cloud scene, they only need to label localized regions corresponding to the selected voxels, greatly simplifying the task.

Voxel-Level Selection vs. Point-Level Budget. While voxel-centric active learning improves selection efficiency, the number of points per voxel can vary considerably due to the sparse and irregular nature of LiDAR data. As a result, selecting the same number of voxels with different methods may result in drastically different labeling costs. To ensure fair comparison and consistent annotation effort, we define the annotation budget N_{budget} based on the total number of labeled points, not voxels.

Voxelization. Similar to [84], the voxelization process initiates with a raw point cloud \mathcal{P}_i as input. For k -th point's coordinates \mathcal{G}_k in a given voxel, the voxelized coordinates \mathcal{G}'_k are determined by dividing the raw coordinates (x_k, y_k, z_k) by the voxel size λ and rounding down the results to the nearest integers using $\mathcal{G}'_k = \text{int} \left(\left\lfloor \frac{(x_k, y_k, z_k)}{\lambda} \right\rfloor \right)$. Subsequently, each point is hashed based on its voxelized coordinates to filter out redundant points with closely situated coordinates. To better capture a comprehensive view while minimizing computational complexity and mitigating the effects of noise and sparsity, we opt for a voxel size λ of 0.25 during the selection phase and λ of 0.05 during training. The influence of voxel size on our proposed method will be further explored in the ablation study section.

Submodular Functions. A submodular function f is a scalar function that is defined over a ground set of elements E , forming a discrete space, that has decreasing marginal returns i.e., the following property is satisfied $\forall S \subseteq T \subseteq E$,

$$f(T \cup \{x\}) - f(T) \leq f(S \cup \{x\}) - f(S) \quad (2)$$

$\forall x \in E \setminus T$. If $f(S) < f(T)$ for $S \subseteq T$, f is considered to be strictly monotone. [71] provides detailed proof demonstrating that Shannon entropy [82], a goto criterion of sample selection in active learning, is submodular. To select *representative* data subsets, f is maximized as $\max_{A \in \mathcal{K}} f(S)$, where \mathcal{K} is a constrained set on E as $\mathcal{K} \subseteq 2^E$. With submodular functions naturally modeling notions of diversity and representativeness, they are optimized efficiently by simple solutions involving greedy algorithms [68].

IV. PROPOSED METHOD

To address the challenges of active 3D LiDAR semantic segmentation, our proposed SELECT employs a three-stage

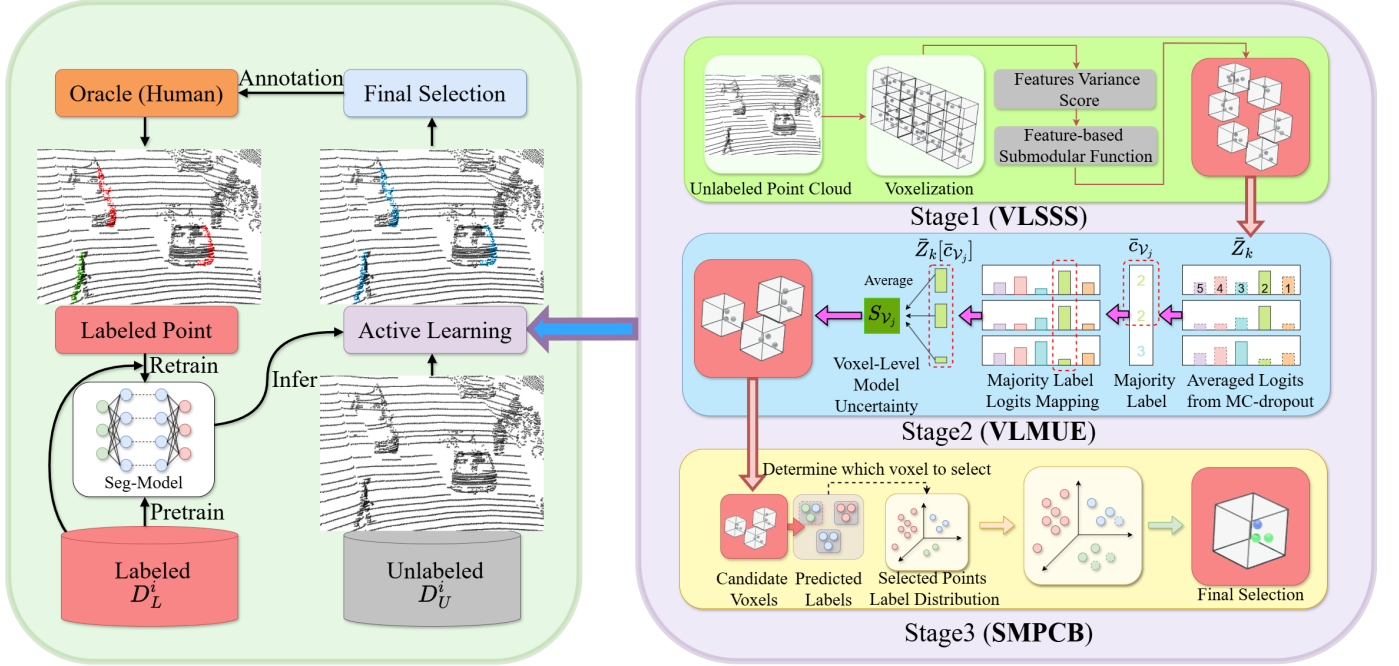


Fig. 2: **Left:** The active learning pipeline for LiDAR semantic segmentation. **Right:** The proposed SELECT, which consists of three key stages — efficiently selecting points that are both informative and well-balanced in label distribution for training the LiDAR semantic segmentation model.

hierarchical framework that incrementally filters out uninformative voxels (Figure 2). In each active learning round, given a point cloud, SELECT follows these three key stages:

- **Stage 1: Voxel-Level Submodular Subset Selection (VLSSS)** selects a representative subset of voxels to minimizing redundancy and ensure scalability;
- **Stage 2: Voxel-Level Model Uncertainty Estimation (VLMUE)** identifies uncertain voxels based on voxel-level model predictions, targeting regions where additional labels are likely to reduce model error;
- **Stage 3: Submodular Maximization for Point-Level Class Balancing (SMPCB)** promotes label diversity by selecting voxels that contribute to a balanced class representation, addressing the class imbalance issue prevalent in outdoor scenes.

Stage 1: Voxel-Level Submodular Subset Selection. Despite their sparsity, outdoor LiDAR point clouds contain a massive number of points distributed across a large spatial range, typically resulting in over 10,000 voxel grid cells per scan. However, only a small subset of these voxels—such as those along object boundaries or covering rare semantic regions—contribute meaningfully to model improvement. The vast majority are either uninformative (e.g., background, ground) or redundant (e.g., repeated structure).

Traditional clustering-based selection techniques like K-means or Gaussian Mixture Models (GMM) are computationally infeasible in this context. They require either constructing and storing pairwise distance matrices (quadratic in the number of voxels) or iteratively optimizing probabilistic assignments via the EM algorithm, which becomes prohibitively

expensive for large-scale outdoor scenes. More details are provided in the Ablation Study section.

To address this scalability bottleneck, we propose a feature-based submodular subset selection strategy that efficiently selects a set of voxels with high diversity and semantic representativeness—without any pairwise computation. Submodular functions exhibit a natural diminishing returns property, enabling fast greedy selection while preserving coverage and diversity of the selected voxels.

Specifically, for each unlabeled point cloud \mathcal{P}_i , we extract a point-level feature vector f_k for each point using the backbone network, just before the final classifier layer. We then compute a voxel-level representation $f_{\mathcal{V}_j}$ by averaging the features of all points within voxel \mathcal{V}_j :

$$f_{\mathcal{V}_j} = \frac{1}{N_j} \sum_{k=1}^{N_j} f_k \quad (3)$$

To measure the informativeness of each voxel, we define a variance-based submodular score function $\sigma(f_{\mathcal{V}_j})$, which captures the dissimilarity between the point-level features and favors voxels with distinct semantic patterns:

$$\sigma(f_{\mathcal{V}_j}) = \frac{1}{D} \sum_{d=1}^D (f_{\mathcal{V}_j,d} - \mu_j)^2, \quad \mu_j = \frac{1}{D} \sum_{d=1}^D f_{\mathcal{V}_j,d} \quad (4)$$

We denote D as the feature dimension. The score $\sigma(f_{\mathcal{V}_j})$ is then plugged into a concave gain function $g(x) = \log(1 + x)$ to define our final sub-modular objective in the following equation:

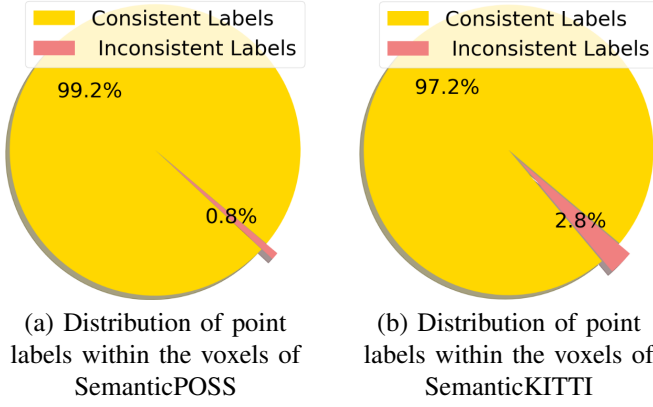


Fig. 3: We summarize the statistics of commonly used datasets for the LiDAR semantic segmentation task, including SemanticPOSS [20] and SemanticKITTI [4]. We observe that most points within the same voxel share the same semantic label. Therefore, a reasonable assumption for the voxel-level label is to assign it based on the majority predicted label among the points contained within the voxel.

$$\max_{D_S^i \subset D_U^i, |D_S^i|=\Lambda_1} \sum_{V_j \in D_S^i} g(\sigma(f_{V_j})) \quad (5)$$

where D_U^i denotes the remaining unlabeled voxels, D_S^i represents the voxels selected in stage 1, and Λ_1 indicates their total number. This formulation encourages selection of voxels that contribute non-redundant, semantically rich features. The greedy selection procedure avoids computing similarity matrices or clustering centroids and instead iteratively adds the voxel with the highest marginal gain, which allows the method to scale linearly with the number of candidate voxels. Compared to scene-level or region-based methods that often require auxiliary networks or per-point uncertainty evaluations, our feature-based submodular selection offers a simple yet scalable mechanism to select a subset of representative voxels.

Stage 2: Voxel-Level Model Uncertainty Estimation. After reducing the voxel candidate pool to Λ_1 informative candidates as in the Stage 1, the next challenge is to prioritize those voxels where the model remains *most uncertain*. This allows SELECT to focus its annotation budget on samples where the model either lacks confidence or struggles to disambiguate different semantic classes, both of which are critical for improving model performance.

One remaining challenge is that semantic segmentation models operate at the point level, not at the voxel level. Therefore, we need to first devise a method to aggregate point-wise predictions into a voxel-level uncertainty score. To this end, we leverage a key observation about voxelized LiDAR data: In outdoor LiDAR scenes, points within the same voxel often share the same semantic label (Fig. 3). This is due to the sparsity of LiDAR data and the physical coherence of structures (e.g., all points in a small voxel along a vehicle surface typically belong to the “car” class).

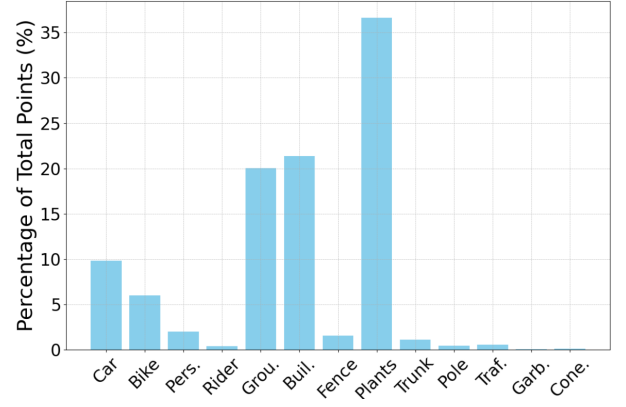


Fig. 4: Label distribution of LiDAR data points is highly imbalanced, as shown in the SemanticPOSS [20] dataset.

Given this, we hypothesize that the model’s uncertainty at the voxel level can be accurately estimated by evaluating how consistently it predicts the same label across the points inside that voxel. If the predicted labels vary widely across points in a voxel, or if the predicted probability for the dominant class is low, it signals either: (1) The model is *uncertain*, or (2) The voxel likely contains points from *multiple semantic classes*, such as object boundaries or rare regions — both are useful for model learning.

To estimate uncertainty, we use Monte Carlo dropout (MC-dropout) during inference to generate multiple predictions per point. For each point P_k , we collect the averaged logit \bar{Z}_k and use it to determine the hypothetical or predicted label $\bar{c}_k = \arg \max(\bar{Z}_k)$. For a voxel V_j , we assign a hypothetical voxel label \bar{c}_{V_j} by taking the majority label of the hypothetical point labels within the voxel:

$$\bar{c}_{V_j} = \operatorname{argmax}_{c \in C} \left(\sum_{k=1}^{N_j} \mathbb{1}(\bar{c}_k = c) \right) \quad (6)$$

We compute the voxel-level uncertainty score S_{V_j} as the average logit of the majority class across all points within the voxel:

$$S_{V_j} = \frac{1}{N_j} \sum_{k=1}^{N_j} \bar{Z}_k[\bar{c}_{V_j}] \quad (7)$$

A lower value of S_{V_j} implies that the model has low confidence in its majority-class prediction, indicating either intra-voxel label ambiguity or high epistemic uncertainty. We therefore rank all Λ_1 voxels using S_{V_j} and select the Λ_2 voxels with the lowest scores for further consideration in the third stage.

This uncertainty-aware selection ensures that SELECT does not waste annotations on regions where the model is already confident. Instead, it targets hard regions (ambiguous structures, underrepresented classes, or object boundaries) where supervision is likely to be more impactful.

Stage 3: Submodular Maximization for Point-Level Class Balancing. After identifying uncertain voxels in Stage 2, a critical challenge remains: ensuring that the final set of selected voxels does not disproportionately represent frequent

classes. Outdoor LiDAR datasets exhibit significant class imbalance, with common classes (e.g., road, car) dominating the point distribution, while rare classes (e.g., pedestrian, bicyclist) are heavily underrepresented (Fig. 4). As observed in our experiments, existing active learning methods, such as Annotator [84], often query more samples from frequent classes, resulting in an imbalanced label distribution that can degrade performance on rare classes. To address the issue of class imbalance in 3D scenes, where certain classes have significantly more points than others, we aim to select a subset of Λ_3 voxels from the Λ_2 voxels identified in Stage 2, ensuring a more balanced representation of all classes.

We begin by calculating the current class distribution among the selected voxels. Let N_{D_L} represent the total number of points in the selected voxels, and $N_{D_L}^c$ denote the number of points belonging to class c . For a given voxel \mathcal{V}_j , $N_{\mathcal{V}_j}$ is the total number of points it contains, and $N_{\mathcal{V}_j}^c$ is the number of points in \mathcal{V}_j predicted to be of class c . To evaluate the impact of adding voxel \mathcal{V}_j on class balance, we compute the relative proportion $\Lambda_{j,c}$ for each class c :

$$\Lambda_{j,c} = \frac{N_{D_L}^c + N_{\mathcal{V}_j}^c}{N_{D_L} + N_{\mathcal{V}_j}} \quad (8)$$

Next, we apply the softmax function to normalize these proportions, yielding $\mathcal{A}_{j,c}$, which represents the normalized probability of points belonging to each class c if voxel \mathcal{V}_j is selected:

$$\mathcal{A}_{j,c} = \frac{e^{\Lambda_{j,c}}}{\sum_{c=1}^C e^{\Lambda_{j,c}}} \quad (9)$$

We then calculate the entropy $H(\mathcal{V}_j)$ of this distribution:

$$H(\mathcal{V}_j) = - \sum_{c=1}^C \mathcal{A}_{j,c} \log \mathcal{A}_{j,c} \quad (10)$$

Higher entropy values indicate a more balanced class distribution. Using a greedy algorithm, we iteratively select voxels from Λ_2 that maximize $H(\mathcal{V}_j)$, until Λ_3 voxels are chosen. This process ensures that the selected subset contributes to a balanced representation of all classes, enhancing the model's performance across both frequent and rare classes.

By employing this three-stage approach as proposed in SELECT, we achieve effective, informative, and balanced subset selection from the set of unlabeled voxels in each point cloud, which is crucial for various real-world applications of active 3D LiDAR semantic segmentation.

V. EXPERIMENTS

A. Experiments setup

Datasets. For the experiments, we use SemanticPOSS [20], a commonly used dataset for the LiDAR semantic segmentation task. This dataset contains over 2,988 LiDAR scans for training and 500 for validation. We also include SemanticKITTI [4] in our experiments, which has more than 29,130 training scans and 6,019 validation scans. The training scans contain more than 1.6 billion points and 19 semantic classes. Finally, we employ the nuScenes [1] dataset, a large-scale dataset mainly

for autonomous driving, containing more than 1,000 urban driving scenes captured in Singapore and Boston, US.

Baselines. We benchmark our proposed method against several active learning (AL) strategies to validate its effectiveness. This includes both generic AL methods and previous state-of-the-art AL methods for LiDAR semantic segmentation:

- **Random** [31]: This baseline method involves randomly selecting samples from the unlabeled pool for each point cloud.
- **Entropy** [82]: This baseline method leverages model predictions to calculate an entropy score for each sample. Samples with higher entropy, indicating greater uncertainty, are selected.
- **BGADL** [42]: It is a method that combines Bayesian active learning with generative models to efficiently select and generate informative training samples, enhancing the performance of deep learning classifiers.
- **BADGE** [63]: An active learning method designed for deep neural networks which select a batch of samples with diverse gradient magnitudes.
- **CORESET** [60]: This method identifies a compact, representative subset of training data for annotation.
- **ReDAL** [49]: A region-based and diversity-aware active learning framework for point cloud semantic segmentation. ReDAL selects informative sub-scene regions based on softmax entropy, color discontinuity, and structural complexity, and employs a diversity-aware selection algorithm to avoid redundant annotations, thereby reducing labeling costs while maintaining high segmentation performance.
- **SQN** [41]: A weakly supervised semantic segmentation approach that leverages the local semantic homogeneity of point clouds to select samples.
- **BaSAL** [45]: A size-balanced warm start active learning method for LiDAR semantic segmentation. BaSAL addresses class imbalance and the cold start problem by sampling object clusters based on their characteristic sizes, enabling effective model initialization without requiring a pretrained model.
- **LiDAL** [98]: An active learning framework that exploits inter-frame uncertainty in LiDAR sequences. LiDAL measures inconsistencies in model predictions across frames to identify uncertain regions, enhancing sample selection and incorporating pseudo-labels to improve segmentation performance.
- **MiLAN** [46]: A method that combines self-supervised representations with minimal annotations for LiDAR semantic segmentation. MiLAN selects informative scans using self-supervised features and clusters points within these scans, allowing annotators to label entire clusters with a single click, significantly reducing annotation effort.
- **SSDR** [47]: A superpoint-based active learning approach that considers both spatial and structural diversity. SSDR-AL groups point clouds into superpoints and employs a graph reasoning network to select the most informative and diverse samples, effectively reducing annotation costs

Method	Car	Bike	Pers.	Rider	Grou.	Buil.	Fence	Plants	Trunk	Pole	Traf.	Garb.	Cone.	mIoU
Random [31]	21.43	48.14	36.34	16.79	69.47	40.64	27.42	70.45	13.01	0.00	3.45	0.00	0.00	26.70
Entropy [82]	31.23	43.13	39.43	18.63	69.58	69.20	43.09	71.16	25.84	16.45	11.83	5.63	7.89	34.85
BGADL [42]	29.13	37.87	44.42	9.98	71.01	55.21	45.11	74.46	25.84	7.31	5.56	4.13	6.67	32.05
BADGE [63]	26.35	47.39	47.55	5.13	77.42	74.28	40.41	74.57	21.20	11.09	5.61	11.41	10.13	34.78
CORESET [60]	25.88	45.90	43.59	22.11	66.76	64.02	46.29	68.55	30.24	23.79	15.48	9.89	11.50	36.38
ReDAL [49]	28.18	48.04	45.21	28.21	65.43	72.21	40.55	75.17	15.53	11.42	10.02	3.43	2.21	34.28
SQN [41]	29.72	41.23	42.98	26.21	75.34	72.02	33.58	70.19	17.34	25.45	5.54	1.33	2.46	34.11
BaSAL [45]	33.94	40.35	46.63	32.59	66.55	64.23	41.92	67.89	15.29	4.56	5.03	1.22	8.76	32.99
LiDAL [98]	36.47	48.55	34.07	15.55	70.32	73.56	33.45	67.97	33.21	16.56	18.88	8.66	5.81	35.62
MILAN [46]	28.20	37.56	42.01	23.61	65.66	64.58	47.71	72.22	30.57	14.21	6.65	15.63	2.22	34.68
SSDR [47]	35.35	50.19	48.32	20.63	77.63	72.48	40.55	72.56	27.21	18.41	9.89	10.59	12.33	38.19
Margin [13]	25.44	47.79	38.34	30.45	78.42	60.35	43.68	70.45	15.51	22.32	25.35	2.21	0.00	37.33
Annotator [84]	29.64	48.04	39.70	27.85	77.20	70.44	50.47	69.70	20.23	22.06	20.77	4.44	12.44	35.40
SELECT	41.00	51.02	57.63	50.90	79.52	75.00	52.90	76.25	36.66	27.60	31.06	18.57	11.56	46.90

TABLE I: The proposed SELECT achieves the highest mIoU on the SemanticPOSS dataset with SPVCNN [14] as the segmentation model.

Method	Ba.ier	Bi.cle	Bus	Car	Con-Veh.	Mo.cle	Ped.	Fr.aff.	Trailer	Truck	Dri_Surf.	Oth_flat	Side.	Terr.	Man.	Veget.	mIoU
Random [31]	37.88	2.12	0.00	60.57	8.48	0.00	0.00	0.19	1.27	44.32	84.81	0.00	39.72	55.01	77.64	67.58	29.97
Entropy [82]	29.31	0.00	1.01	66.49	0.00	1.20	0.00	11.09	0.00	41.45	88.36	0.00	34.97	59.83	68.62	68.39	29.40
BGADL [42]	27.61	0.53	2.12	58.34	0.00	0.00	0.00	2.13	0.00	43.98	80.70	0.00	36.12	46.18	68.99	66.33	27.06
BADGE [63]	40.59	0.77	0.00	60.79	0.00	0.00	0.00	15.23	0.00	48.89	85.31	5.95	44.34	59.05	73.66	74.37	31.87
CORESET [60]	25.11	0.00	3.76	71.75	0.00	0.88	0.00	9.76	0.00	47.69	87.88	0.00	40.12	62.01	70.01	69.16	30.51
ReDAL [49]	36.71	0.00	0.70	70.65	0.00	0.00	0.00	2.89	0.00	45.55	80.06	0.00	37.04	61.34	73.12	66.00	29.63
Sqn [41]	28.09	0.49	1.65	63.01	0.00	0.55	1.04	4.37	0.09	35.78	86.94	0.00	33.33	57.81	68.20	67.03	28.02
BaSAL [45]	25.12	0.00	2.33	59.88	0.00	0.72	5.91	4.76	0.00	34.68	79.22	0.00	30.41	48.67	61.93	60.22	25.93
LiDAL [98]	33.33	0.00	0.00	58.00	0.00	0.00	0.00	2.13	0.00	44.38	83.00	3.78	37.01	52.28	70.90	68.98	28.36
MILAN [46]	33.59	0.00	2.22	67.21	0.00	0.63	0.00	2.44	0.00	43.19	90.12	0.00	40.36	61.88	73.12	71.33	30.38
SSDR [47]	26.61	0.00	0.00	63.23	0.00	0.76	0.00	5.87	0.00	44.18	87.01	0.00	49.90	62.12	77.01	75.24	30.81
Margin [13]	28.78	0.00	7.78	66.01	0.00	0.00	0.00	1.20	0.00	50.24	91.24	0.00	41.17	64.89	59.44	70.21	30.06
Annotator [84]	32.40	0.00	1.46	68.79	0.00	0.00	9.59	8.28	0.51	46.03	90.24	0.17	45.06	62.34	76.47	75.75	32.31
SELECT	43.74	0.00	3.01	74.33	0.00	0.00	27.87	0.00	15.40	50.88	91.16	18.30	51.17	64.55	78.40	78.99	37.36

TABLE II: The proposed SELECT achieves the highest mIoU on the nuScenes (16-class) dataset with MinkNet [12] as the segmentation model.

while maintaining segmentation accuracy.

- **Margin [13]:** This method calculates the margin as the difference between the highest and second-highest logit values for each point. The voxel-level margin score is determined by finding the maximum difference among these pairs within the voxel. The voxel with the lowest margin score, suggesting minimal confidence in prediction accuracy, is selected.
- **Annotator [84]:** This is a recently proposed active learning method for LiDAR semantic segmentation, where the Voxel Confusion Degree (VCD) is calculated to assess the class diversity within each voxel, based on the predicted labels from the segmentation model. Voxels exhibiting the highest VCD, indicating significant class diversity, are prioritized for selection.

Implementation Details. We use MinkNet [12] and SPVCNN [14] as the primary backbone segmentation models. All the models are trained on four NVIDIA RTX A5000 GPUs. Across all datasets and backbones, we employ a batch size of 4. We utilize the SGD optimizer and implement the linear warmup with a cosine decay scheduler, setting the learning rate at 0.01.

Active Learning Setups. To initiate the active learning process, we start by randomly selecting one voxel \mathcal{V}_j from unlabeled voxel D_U^i in a given point cloud \mathcal{P}_i . A human annotator then labels each point P_k within that voxel. These labeled points are used to train the segmentation model. We train the segmentation model for three epochs. To ensure fairness in our comparisons, all active learning methods selected one voxel per point cloud in each active learning round, resulting in $N_q = 1$. In our method, the number of voxels selected at different stages, denoted by Λ , was configured as $\Lambda_1 = 200$, $\Lambda_2 = 5$, and $\Lambda_3 = 1$ across all datasets. We set our budget N_{budget} to 6000 points for all baselines. The initial epoch consists of the initialization process we previously discussed, and every three epochs, we perform one query of active learning, with a total of $Q = 5$ rounds of queries. After the Q -th query or once the budget N_{budget} is exhausted, we stop the active learning process and continue training the model until epoch 60.

B. Results

Results on the SemanticPOSS dataset. We evaluate the performance of SELECT against various baselines, as detailed

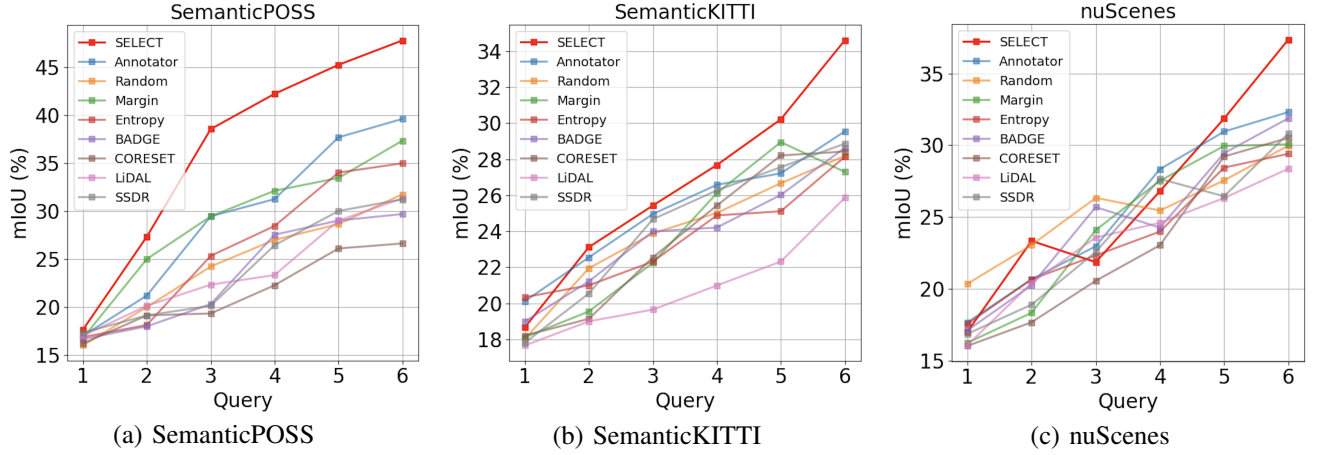


Fig. 5: The mIoU results of AL baselines and SELECT across each active learning round for the SemanticPOSS, SemanticKITTI, and NuScenes datasets.

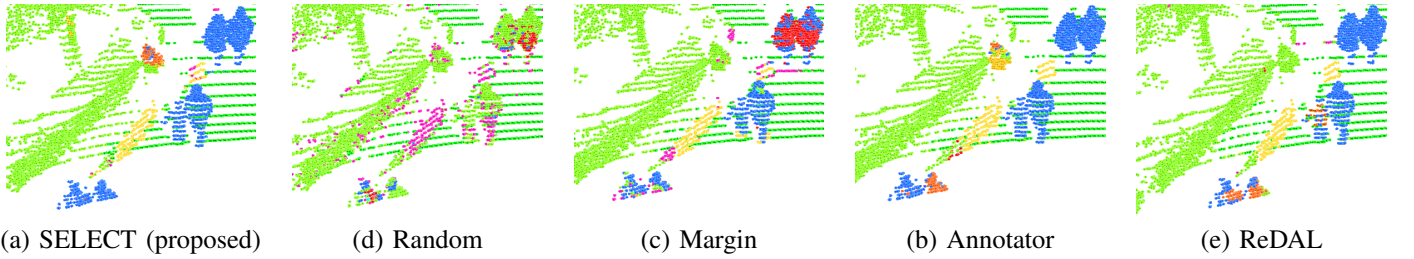


Fig. 6: Visualization of inference results on the SemanticPOSS dataset using MinkNet. Our method, SELECT, accurately recognizes small and rare classes such as pedestrians (blue points) and produces clearer boundaries between objects from different classes.

in Table I, employing SPVCNN [14] as the backbone segmentation models. The results clearly demonstrate that SELECT surpasses all previous active learning methods using different segmentation backbones. Specifically, SELECT achieves up to 15% improvements in mIoU scores for the *Car*, *Person*, and *Rider* classes compared to Annotator [84], which is the SOTA method for active LiDAR semantic segmentation when using the SPVCNN backbone. Another critical observation is that most baselines, including Annotator, perform unsatisfactorily, often approaching zero mIoU on the rare classes such as *Cone*, indicating their inability to select informative points with balanced label distribution from the unlabeled pool D_U^i for each point cloud. In contrast, SELECT achieves notable mIoU scores of 8.92% and 15.31% on rare classes like *Garbage-can* (Garb.) and *Cone* when using MinkNet [12] as backbone model, with detailed per-class table results provided in the Supplementary Materials. Fig. 5 (a) further illustrates that our approach consistently outperforms other methods throughout the active learning process, confirming the effectiveness of SELECT. Due to space limitations, visualizations of the selected points in each point cloud are provided in the Supplementary.

Results on SemanticKITTI and nuScenes datasets. To evaluate the generality and robustness of SELECT, we further test our approach on the SemanticKITTI [4] and nuScenes [1] datasets. Comprehensive per-class results for SemanticKITTI dataset can be found in the Supplementary Materials, due to a

space limitation. Using MinkNet as the backbone segmentation model on SemanticKITTI, SELECT demonstrates an overall performance improvement of 5.06% over Annotator, with up to a 10% increase in mIoU for rare classes such as *Parking*, *Trunk*, and *Pole*. On the nuScenes dataset as shown in Table II, our approach achieves a 5.05% overall performance gain compared to Annotator, with notable improvements in classes like *Other Flat Surfaces* and *Pedestrian*, where mIoU increased by 18.13% and 18.28%, respectively. For detailed results per query, please refer to Fig. 5 (b) for SemanticKITTI, Fig. 5 (c) for the nuScenes and Supplementary Materials.

Analysis of Selected Voxels by SELECT. With the outstanding segmentation results achieved across different datasets, we further investigate the informativeness of the voxels selected by SELECT. As previously described, informativeness is defined as voxels that either contain points from rare classes, where the model exhibits low confidence, or are located along the boundaries between objects. From Fig. 7, we observe that although less than 1% of the voxels in the entire dataset contain points from more than one class, our method successfully identifies approximately 9% of selected voxels containing points from multiple classes, compared to only 5% achieved by the previous state-of-the-art method.

Visualization of SELECT. To more intuitively demonstrate the segmentation results of the model trained with our active learning approach, we visualize the predictions on the Sema-

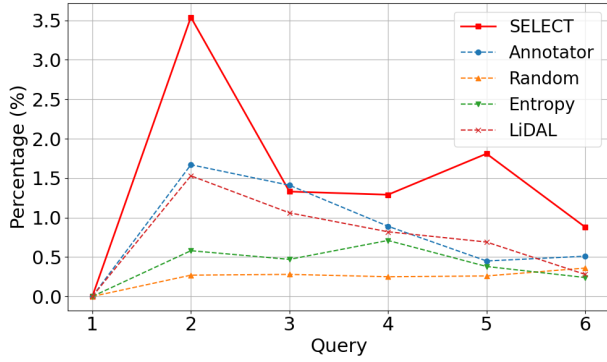


Fig. 7: The plot shows that the proposed SELECT selects more voxels containing points from multiple classes, suggesting it focuses on object boundaries, which are more informative than voxels fully within one class. Results on the SemanticPOSS [20] dataset using MinkNet [12].

ticPOSS dataset using MinkNet. As shown in Fig. 6, even under a limited annotation budget, our method achieves high segmentation accuracy for certain rare-class points, such as pedestrians. Although this comes at the cost of slightly reduced accuracy on background classes like buildings, fences, and plants compared to other baselines, the overall results highlight the effectiveness of our AL selection strategy in prioritizing rare and informative samples.

Furthermore, regarding the ability to discover points from rare classes, such as bicycles, motorcycles, and pedestrians in the SemanticKITTI dataset, the previous state-of-the-art method struggles to select points from these categories, while our method consistently identifies a considerable number of points belonging to these rare classes. More detailed results are provided in the Supplemental Materials.

C. Ablation Study

To analyze the proposed method, we conducted ablation studies on the impact of different stages in SELECT, voxel size, and various selection methods in the VLSSS stage. Due to space limitations, additional ablation studies are available in the Supplementary.

The effect of different stages in SELECT. To validate the effectiveness of our proposed three-stage method, we present a performance comparison, in Table III, involving six variants of our approach against the *Random* selection method, which serves as our baseline. We utilize MinkNet [12] as the backbone and evaluate our method on the SemanticPOSS [20] dataset. Our results show that individual stages of our method impact the mIoU score differently. Specifically, using only the Voxel-level Submodular Subset Selection (VLSSS) stage results in a near 5% improvement over the baseline, while the Voxel-level Model Uncertainty Estimation (VLMUE) stage alone increases mIoU by nearly 14%. Conversely, the Submodular Maximization for Point-Wise Class Balancing (SMPCB) stage decreases mIoU by 3% relative to the baseline. This indicates that the VLMUE stage is most effective at identifying informative samples. Additionally, implementing

only Stage 2 or Stages 1 and 2 results in zero mIoU for the “garbage-can” and “cone/stone” classes, suggesting a bias towards non-informative samples for rare classes. Solely using Stage 3 underperforms compared to the baseline but when combined with other stages, the overall performance will increase, highlighting the importance of integrating all stages to avoid performance drops.

TABLE III: Performance of different active learning stage criteria on SemanticPOSS (mIoU %)

VLSSS	VLMUE	SMPCB	mIoU (%)
–	–	–	31.75
✓	–	–	36.52
–	✓	–	45.21
–	–	✓	28.64
✓	✓	–	43.28
✓	–	✓	37.53
✓	✓	✓	47.77

Effect of voxel size λ . In our study, we investigate the effect of voxel size λ on the field-of-view using Minknet [12] and evaluate on SemanticPOSS [20] dataset. A larger λ results in voxels \mathcal{V}_j containing more points, which consequently decreases the number of voxels into which each point cloud \mathcal{P}_i is divided. In contrast, a smaller λ allows for a finer division of point clouds, resulting in a larger number of voxels, each containing fewer points. According to Table IV, we observe that training time decreases with an increase in λ , as fewer voxels need to be processed during the active learning stage. However, the mIoU scores tend to decline as the voxel size increases. This is because larger voxels encompass more points, leading to a higher likelihood of points being filtered out due to the hashing based on voxelized coordinates, resulting in a loss of semantic and structural information. Using a smaller voxel size improves the mIoU score, but it significantly increases computational time. For example, reducing λ to 0.05 nearly triples the training time without a substantial improvement in performance. Therefore, to strike a balance between voxel granularity and training efficiency, our chosen voxel size of 0.25 offers an optimal compromise between performance and training time.

TABLE IV: Effect of voxel size (λ) on performance and training time

Voxel size (λ)	mIoU (%)	Training time (hours)
0.05	46.61	17.0
0.25	47.77	6.3
0.5	43.56	6.0
0.75	37.70	5.2

TABLE V: The use of a feature-based submodular function in the first stage reduces training time and leads to a higher mIoU compared to clustering methods.

Clustering method	mIoU (%)	Training time (hours)
Random [31]	34.2	5.4
K-Means++ [18]	43.76	40.1
K-Means [17]	45.19	43.4
GMM [96]	39.20	47.9
Ours	47.77	6.3

Effect of different selection methods in VLSSS stage. We assess the effectiveness of our feature-based submodular

function by considering both computation time and model performance. Using MinkNet [12] as the model, we conduct evaluations on the SemanticPOSS [20] dataset. We compared our submodular approach against Gaussian mixture models (GMM) [96], K-means [17], and K-means++ [18], and all methods select $\Lambda_1 = 200$ voxels per active query for each point cloud. Detailed comparisons of performance and computation times are provided in Table V. Our method significantly reduces computational time compared to the clustering-based baselines, which on average require almost six times the computational resources of our approach. Additionally, our method outperforms these baselines, demonstrating its superior ability to identify informative samples efficiently.

VI. CONCLUSION

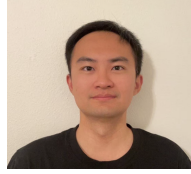
We propose SELECT, a unified submodular approach for voxel-centric active 3D LiDAR semantic segmentation, designed to overcome the limitations of existing active learning methods. Compared to current active learning approaches for LiDAR semantic segmentation, SELECT efficiently performs active learning while ensuring that the selected points are both informative and well-balanced in label distribution. Extensive experiments on SemanticPOSS, SemanticKITTI, and nuScenes demonstrate that the proposed SELECT enables the LiDAR semantic segmentation model to be trained on a limited number of points while achieving high segmentation accuracy.

REFERENCES

- [1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 11621–11631.
- [2] S. Dasgupta, “Two faces of active learning,” *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1767–1781, 2011.
- [3] Z. Liu, X. Qi, and C.-W. Fu, “One thing one click: A self-training approach for weakly supervised 3D semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2021, pp. 1726–1736.
- [4] 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 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019.
- [5] H. Thomas, C. R. Qi, J.-E. Deschaut, B. Marcotegui, F. Goulette, and L. J. Guibas, “KPConv: Flexible and deformable convolution for point clouds,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 6411–6420.
- [6] K. Hassani and M. Haley, “Unsupervised multi-task feature learning on point clouds,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 8160–8171.
- [7] A. Thabet, H. Alwassel, and B. Ghanem, “Self-supervised learning of local features in 3D point clouds,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 938–939.
- [8] G. Mei, C. Saltori, E. Ricci, N. Sebe, Q. Wu, J. Zhang, and F. Poiesi, “Unsupervised point cloud representation learning by clustering and neural rendering,” *Int. J. Comput. Vis.*, vol. 132, no. 8, pp. 3251–3269, 2024.
- [9] C. Xu, S. Yang, B. Zhai, B. Wu, X. Yue, W. Zhan, et al., “Image2point: 3D point-cloud understanding with pretrained 2D convnets,” *arXiv preprint arXiv:2101.02691*, 2021.
- [10] Y. Wu, M. Xing, Y. Zhang, Y. Xie, K. Peng, and Y. Qu, “Fusion-then-distillation: Toward cross-modal positive distillation for domain adaptive 3D semantic segmentation,” *IEEE Trans. Circuits Syst. Video Technol.*, 2025.
- [11] B. Sun, Y. Liu, X. Wang, B. Tian, L. Chen, and F.-Y. Wang, “3D unsupervised learning by distilling 2D open-vocabulary segmentation models for autonomous driving,” *arXiv preprint arXiv:2405.15286*, 2024.
- [12] C. B. Choy, J. Gwak, and S. Savarese, “4D spatio-temporal ConvNets: Minkowski convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 3075–3084.
- [13] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, “Multi-class active learning for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Miami, FL, USA, 2009, pp. 2372–2379.
- [14] 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 Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K., 2020, pp. 685–702.
- [15] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “A dataset for semantic scene understanding of LiDAR sequences,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 9297–9307.
- [16] D. A. Reynolds, “Gaussian Mixture Models,” in **Encyclopedia of Biometrics**, S. Z. Li and A. Jain, Eds. Boston, MA, USA: Springer, 2009, pp. 659–663.
- [17] S. P. Lloyd, “Least squares quantization in PCM,” **IEEE Trans. Inf. Theory**, vol. 28, no. 2, pp. 129–137, Mar. 1982.
- [18] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in **Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)**, New Orleans, LA, USA, Jan. 2007, pp. 1027–1035.
- [19] M. Li, S. Lin, Z. Wang, Y. Shen, B. Zhang, and L. Ma, “Class-imbalanced semi-supervised learning for large-scale point cloud semantic segmentation via decoupling optimization,” *Pattern Recognit.*, vol. 156, 2024, Art. no. 110701.
- [20] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, and H. Zhao, “SemanticPOSS: A point cloud dataset with large quantity of dynamic instances,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.09147>
- [21] W. Zhang, H. Wang, X. Li, and C. Xu, “Deep learning-based LiDAR point cloud semantic segmentation for robotics: A survey,” *IEEE Trans. Robot.*, vol. 39, no. 2, pp. 150–170, 2023.
- [22] J. Yan, Z. Chen, and B. Liu, “LiDAR-based urban scene understanding for smart city applications: A review,” *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17088–17104, 2022.
- [23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5099–5108.
- [24] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, “Spatio-temporal self-supervised representation learning for 3D point clouds,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 6535–6545.
- [25] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, “Batch mode active learning and its application to medical image classification,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), Pittsburgh, PA, USA, 2006, pp. 417–424.
- [26] H. T. Nguyen and A. Smeulders, “Active learning using pre-clustering,” in Proc. 21st Int. Conf. Mach. Learn. (ICML), Banff, AB, Canada, 2004, p. 79.
- [27] Z. Bodó, Z. Minier, and L. Csató, “Active learning with clustering,” in Proc. Active Learn. Exp. Design Workshop, Int. Conf. Artif. Intell. Statist. (AISTATS), Ft. Lauderdale, FL, USA, 2011, pp. 127–139.
- [28] J. Han, K. Liu, W. Li, and G. Chen, “Subspace prototype guidance for mitigating class imbalance in point cloud semantic segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Milan, Italy, 2024, pp. 255–272.
- [29] J. Dai, K. He, and J. Sun, “BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1635–1643.
- [30] D. Gissin and S. Shalev-Shwartz, “Discriminative active learning,” *arXiv preprint arXiv:1907.06347*, 2019.
- [31] B. Settles, “Active learning literature survey,” Univ. Wisconsin–Madison, Dept. Comput. Sci., Madison, WI, USA, Tech. Rep. 1648, 2009.
- [32] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2021, pp. 9939–9948.
- [33] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A survey of deep active learning,” *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021.
- [34] Z. Ren, I. Misra, A. G. Schwing, and R. Girdhar, “3D spatial recognition without spatially labeled 3D,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2021, pp. 13204–13213.
- [35] D. D. Lewis, “A sequential algorithm for training text classifiers: Corrigendum and additional data,” *ACM SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

- [36] D. Wang and Y. Shang, "A new active labeling method for deep learning," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Beijing, China, 2014, pp. 112–119.
- [37] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in Proc. Eur. Conf. Mach. Learn. (ECML), Berlin, Germany, 2006, pp. 413–424.
- [38] Y.-Y. Kim, K. Song, J. Jang, and I.-C. Moon, "LADA: Look-ahead data acquisition via augmentation for deep active learning," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 22919–22930, 2021.
- [39] A. Parvaneh, E. Abbasnejad, D. Teney, G. R. Haffari, A. Van Den Hengel, and J. Q. Shi, "Active learning by feature mixing," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 12237–12246.
- [40] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in Proc. 11th Int. Conf. Mach. Learn. (ICML), New Brunswick, NJ, USA, 1994, pp. 148–156.
- [41] Z. Hu, J. Shang, X. Bai, C.-L. Tai, and H. Fu, "SQN: Weakly-supervised semantic segmentation of large-scale 3D point clouds," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 14699–14708.
- [42] T. Tran, T.-T. Do, I. Reid, and G. Carneiro, "Bayesian generative active deep learning," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6295–6304.
- [43] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," *arXiv preprint arXiv:1906.03671*, 2019.
- [44] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.
- [45] B. Wei, B. Gao, C. Li, and H. Zhao, "BaSAL: Size-balanced active learning for LiDAR semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 4812–4821.
- [46] B. Samet, P. Dvornik, E. Belilovsky, and I. Laptev, "MiLAN: Minimal annotation for LiDAR semantic segmentation via self-supervised learning," *arXiv preprint arXiv:2404.10986*, 2024.
- [47] F. Shao, Y. Luo, P. Liu, J. Chen, Y. Yang, Y. Lu, and J. Xiao, "Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning," in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 2575–2585.
- [48] J. Chibane, F. Engelmann, T. A. Tran, and G. Pons-Moll, "Box2Mask: Weakly supervised 3D semantic instance segmentation using bounding boxes," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, 2022, pp. 681–699.
- [49] 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 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 15510–15519.
- [50] Y. Luo, Z. Chen, Z. Wang, X. Yu, Z. Huang, and M. Baktashmotlagh, "Exploring active 3D object detection from a generalization perspective," in Proc. Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, 2023.
- [51] Z. Hu, X. Bai, J. Shang, R. Zhang, J. Dong, X. Wang, G. Sun, H. Fu, and C.-L. Tai, "VMNet: Voxel-mesh network for geodesic-aware 3D semantic segmentation," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada, 2021, pp. 15488–15498.
- [52] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-L. Tai, "JSENet: Joint semantic segmentation and edge detection network for 3D point clouds," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K., 2020, pp. 222–239.
- [53] T. Hackel, N. Savinov, Ľ. Ladický, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3D.net: A new large-scale point cloud classification benchmark," *arXiv preprint arXiv:1704.03847*, 2017.
- [54] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," *Adv. Neural Inf. Process. Syst.*, vol. 20, 2007.
- [55] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996.
- [56] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, 2019, pp. 4332–4341.
- [57] J. Wei, G. Lin, K.-H. Yap, T.-Y. Hung, and L. Xie, "Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 4384–4393.
- [58] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, 1991.
- [59] C. Wang, L. Jiang, X. Wu, Z. Tian, B. Peng, H. Zhao, and J. Jia, "GroupContrast: Semantic-aware self-supervised representation learning for 3D understanding," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2024, pp. 4917–4928.
- [60] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.
- [61] C. Guo, B. Zhao, and Y. Bai, "DeepCore: A comprehensive library for coreset selection in deep learning," in Proc. Int. Conf. Database Expert Syst. Appl. (DEXA), Vienna, Austria, 2022, pp. 181–195.
- [62] M. Wang, F. Min, Z.-H. Zhang, and Y.-X. Wu, "Active learning through density clustering," *Expert Syst. Appl.*, vol. 85, pp. 305–317, 2017.
- [63] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," *arXiv preprint arXiv:1906.03671*, 2019.
- [64] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "Localization-aware active learning for object detection," in Proc. Asian Conf. Comput. Vis. (ACCV), Perth, WA, Australia, 2019, pp. 506–522.
- [65] L. Castrejón, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a Polygon-RNN," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 5230–5238.
- [66] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2016.
- [67] V. Kaushal, R. Iyer, K. Doctor, A. Sahoo, P. Dubal, S. Kothawade, R. Mahadev, K. Dargan, and G. Ramakrishnan, "Demystifying multi-faceted video summarization: Tradeoff between diversity, representation, coverage and importance," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa Village, HI, USA, 2019, pp. 452–461.
- [68] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Program.*, vol. 14, no. 1, pp. 265–294, 1978.
- [69] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.
- [70] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecy, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, and J. M. Alvarez, "Scalable active learning for object detection," in Proc. IEEE Intell. Veh. Symp. (IV), Las Vegas, NV, USA, 2020, pp. 1430–1435.
- [71] J. Bilmes, "Submodularity in machine learning and artificial intelligence," *arXiv preprint arXiv:2202.00132*, 2022.
- [72] Y. Luo, Z. Chen, Z. Fang, Z. Zhang, M. Baktashmotlagh, and Z. Huang, "Kecor: Kernel coding rate maximization for active 3D object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Paris, France, 2023, pp. 18279–18290.
- [73] S. Kothawade, V. Kaushal, G. Ramakrishnan, J. Bilmes, and R. Iyer, "Prism: A rich class of parameterized submodular information measures for guided data subset selection," in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 9, 2022, pp. 10238–10246.
- [74] R. Mao, X. Ouyang, and Y. Guo, "Inconsistency-based data-centric active open-set annotation," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 5, 2024, pp. 4180–4188.
- [75] S. Kothawade, N. Beck, K. Killamsetty, and R. Iyer, "Similar: Submodular information measures based active learning in realistic scenarios," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18685–18697.
- [76] S. Kothawade, S. Ghosh, S. Shekhar, Y. Xiang, and R. Iyer, "Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information," in Proc. Eur. Conf. Comput. Vis. (ECCV), Tel Aviv, Israel, 2022, pp. 1–16.
- [77] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Florence, Italy, 2014, pp. 3311–3315.
- [78] K. Killamsetty, G. S. Abhishek, A. Lnu, G. Ramakrishnan, A. Evfimievski, L. Popa, and R. Iyer, "Automata: Gradient based data subset selection for compute-efficient hyper-parameter tuning," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 28721–28733.
- [79] R. Tiwari, K. Killamsetty, R. Iyer, and P. Shenoy, "GCR: Gradient coreset based replay buffer selection for continual learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 99–108.
- [80] C. Yin, B. Qian, S. Cao, X. Li, J. Wei, Q. Zheng, and I. Davidson, "Deep similarity-based batch mode active learning with exploration-exploitation," in Proc. IEEE Int. Conf. Data Min. (ICDM), New Orleans, LA, USA, 2017, pp. 575–584.
- [81] E. Bıyık, K. Wang, N. Anari, and D. Sadigh, "Batch active learning using determinantal point processes," *arXiv preprint arXiv:1906.07975*, 2019.
- [82] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

- [83] S. Song, S. P. Lichtenberg, and J. Xiao, “SUN RGB-D: A RGB-D scene understanding benchmark suite,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 567–576.
- [84] B. Xie, S. Li, Q. Guo, C. Liu, and X. Cheng, “Annotator: A generic active learning baseline for LiDAR semantic segmentation,” *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.
- [85] B. Gao, Y. Pan, C. Li, S. Geng, and H. Zhao, “Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6063–6081, 2021.
- [86] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss, “Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset,” *Int. J. Robot. Res.*, vol. 40, no. 8–9, pp. 959–967, 2021.
- [87] Y. Yan, Y. Mao, and B. Li, “SECOND: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [88] M. Liu, Y. Zhou, C. R. Qi, B. Gong, H. Su, and D. Anguelov, “LESS: Label-efficient semantic segmentation for LiDAR point clouds,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, 2022, pp. 70–89.
- [89] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, “Multi-class active learning for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 2372–2379.
- [90] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, “Cost-effective active learning for deep image classification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, 2016.
- [91] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 9368–9377.
- [92] S. Ye, Z. Yin, Y. Fu, H. Lin, and Z. Pan, “A multi-granularity semi-supervised active learning for point cloud semantic segmentation,” *Neural Comput. Appl.*, vol. 35, no. 21, pp. 15629–15645, 2023.
- [93] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. L. Yuille, and Z. Zhou, “Making your first choice: To address cold start problem in vision active learning,” *arXiv preprint arXiv:2210.02442*, 2022.
- [94] Y. Zhu, J. Lin, S. He, B. Wang, Z. Guan, H. Liu, and D. Cai, “Addressing the item cold-start problem by attribute-driven active learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 631–644, 2019.
- [95] N. Houlsby, J. M. Hernández-Lobato, and Z. Ghahramani, “Cold-start active learning with robust ordinal matrix factorization,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, 2014, pp. 766–774.
- [96] D. A. Reynolds, “Gaussian mixture models,” in *Encyclopedia of Biometrics*, vol. 741, A. Jain and P. Flynn, Eds., Springer, 2009, pp. 659–663.
- [97] F. Shao, Y. Luo, P. Liu, J. Chen, Y. Yang, Y. Lu, and J. Xiao, “Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning,” in *Proc. 30th ACM Int. Conf. Multimedia (ACM MM)*, Lisbon, Portugal, 2022, pp. 2575–2585.
- [98] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, 2022, pp. 248–265.
- [99] H. H. Aghdam, A. Gonzalez-Garcia, J. van de Weijer, and A. M. López, “Active learning for deep detection neural networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 3672–3680.
- [100] R. Mao, S. K. Maharana, R. K. Iyer, and Y. Guo, “STONE: A submodular optimization framework for active 3D object detection,” *arXiv preprint arXiv:2410.03918*, 2024.
- [101] Y. Siddiqui, J. Valentin, and M. Nießner, “ViewAL: Active learning with viewpoint entropy for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 9433–9443.
- [102] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, Quebec City, QC, Canada, 2017, pp. 399–407.
- [103] E. Jain, T. Nandy, G. Aggarwal, A. Tendulkar, R. Iyer, and A. De, “Efficient data subset selection to generalize training across models: Transductive and inductive networks,” *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2024.
- [104] V. Kaushal, S. Kothawade, A. Tomar, R. Iyer, and G. Ramakrishnan, “How good is a video summary? A new benchmarking dataset and evaluation framework towards realistic video summarization,” *arXiv preprint arXiv:2101.10514*, 2021.
- [105] N. Beck, D. Sivasubramanian, A. Dani, G. Ramakrishnan, and R. Iyer, “Effective evaluation of deep active learning on image classification tasks,” *arXiv preprint arXiv:2106.15324*, 2021.



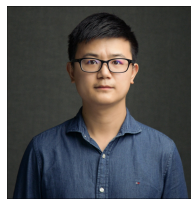
Ruiyu Mao is a Ph.D. candidate in Computer Science at the University of Texas at Dallas, advised by Professor Yunhui Guo. His primary research interests include computer vision, autonomous driving, and active learning. He is dedicated to developing intelligent agents capable of actively and continuously learning in dynamic and evolving environments, with a particular focus on autonomous driving applications leveraging both vision-based and LiDAR-based technologies.



Sarthak Kumar Maharana is a Ph.D. candidate in Computer Science at the University of Texas at Dallas, where he is advised by Prof. Yunhui Guo. He holds a Master's degree in Electrical Engineering from the University of Southern California, Los Angeles, and a Bachelor's degree in Electrical Engineering from IIIT-Bhubaneswar, India. His research focuses on computer vision and multimodal learning, with an emphasis on developing efficient methods to adapt foundational models to distributional shifts, aiming to enhance robustness and generalizability.



Xulong Tang is a Ph.D. student in the Department of Computer Science at The University of Texas at Dallas. His research interests include extended reality (XR), computer vision, multimedia retrieval, and multimodal learning. His primary research focuses on immersive choreography system, where he explores multimodal generation techniques by combining music, text, and motion to create customized dances that are precisely aligned with accompanying music.



Yunhui Guo is an assistant professor in the Department of Computer Science at the University of Texas at Dallas. He was a postdoctoral researcher at UC Berkeley. He earned his PhD in Computer Science from the University of California, San Diego. His research interests include machine learning and computer vision, with a focus on developing intelligent agents that can continuously learn, dynamically adapt to evolving environments without forgetting previously acquired knowledge, and repurpose existing knowledge to handle novel scenarios.