Neural Computing and Applications (2023) 35:15629–15645
https://doi.org/10.1007/s00521-023-08455-7

ORIGINAL ARTICLE

A multi-granularity semisupervised active learning for point cloud semantic segmentation

Shanding Ye[1] · Zhe Yin[1] · Yongjian Fu[1] · Hu Lin[1] · Zhijie Pan[1]

Received: 10 June 2022 / Accepted: 3 March 2023 / Published online: 17 April 2023

Abstract

Recent successes in point cloud semantic segmentation heavily rely on a large amount of annotated data. Furthermore, three-dimensional point cloud data are generally sparse and unorganized, and a frame of point cloud usually includes more than 100,000 points, which increases the difficulty of point cloud annotation. To reduce the annotation efforts, we propose a multi-granularity semisupervised active learning pipeline which aims to select representative, uncertain and diverse data to annotate. To better exploit annotating budget, we first leverage the conventional point cloud registration algorithm to develop a matching score function which is used to select a representative subset. And then we change the annotating units from a point cloud scan to segmented regions through two semisupervised methods. Subsequently, in each active selection step, segmented region information is calculated with two terms: softmax entropy and point cloud intensity, and the latter serves to encourage region diversity. Finally, to further reduce annotation effort, semisupervised learning is introduced to our pipeline to automatically select a portion of unlabeled segmented regions with high confidence and assign pseudolabels to them. Extensive experiments show that our approach greatly outperforms previous active learning methods, and we obtain the mean class intersection-over-union performance of 95% fully supervised learning with merely 3% of labeled data on SemanticKITTI dataset.

Keywords Active learning · Semisupervised learning · Convolutional neural network · Supervoxel · Semantic segmentation

1 Introduction

In recent years, with the aid of deep learning, autonomous driving achieves significant breakthroughs in multiple tasks, like object detection, motion forecasting, and semantic segmentation. As an emerging field among them, point cloud semantic segmentation (PCSS) is usually used to understand the driving-scene and draw more and more attention. Especially in the past 5 years, numerous novel PCCS methods [18, 35, 47] based on deep learning frameworks have been proposed. And several public datasets of PCSS have also been released, such as Semantic3D [15], ScanNet [9], SemanticKITTI [3].

To achieve superior performance of the model, deep learning generally relies on a large amount of annotated data to strengthen the large-scale model. However, the performance of the model is still not saturated with respect to the size of annotated data [54]. Moreover, it costs lots of human labor and time to annotate a large amount of data, and sometimes only relevant professionals can annotate data [4]. More importantly, 3D point cloud data are generally sparse and unorganized, and a point cloud often includes more than 100,000 points [3], which results in difficulties of point cloud annotation. Active learning (AL) is an effective method to solve this problem. The purpose of AL is to select the most informative and representative

✉ Zhijie Pan
zhijie_pan@zju.edu.cn

Shanding Ye
ysd@zju.edu.cn

Zhe Yin
pidandan@zju.edu.cn

Yongjian Fu
yjfu@zju.edu.cn

Hu Lin
11621080@zju.edu.cn

[1] College of Computer Science and Technology, Zhejiang University, 38 Zheda Road, Hangzhou 310027, Zhejiang, China

samples from the unlabeled data to annotate, which greatly reduces the cost of annotation.

Existing AL methods are mostly at the sample level and focus less on dense prediction tasks. Most of the works [36] are proposed for image processing and natural language processing tasks. However, since point cloud is an unorganized and irregular structure, these methods for image cannot be directly applied to it. In addition, compared with images, point cloud typically contains rich geometric information [33] and intensity information. Besides, it is often collected in sequence, which contains temporal information [3]. This information, which is mostly not involved in recent works [26, 43, 52], has the potential to improve the AL model performance.

In this paper, we focus on these characteristics of the point cloud and propose a novel sample selection and annotation pipeline. Specifically, our proposed method takes representativeness, uncertainty, and diversity into consideration and conducts multi-granularity sample selection: inter-frame and intra-frame. For inter-frame selection, we consider the sample representativeness within the sequence, so as to single out a subset which could represent for the entire sequence distribution. In other words, the coverage area of adjacent frames usually overlaps with different sizes, so it is uneconomical to label all point clouds, which will produce a lot of redundancy. Inspired by the point cloud registration algorithm [5], we develop a novel matching score function which is used to evaluate the similarity of two frames within the sequence. According to whether the matching score is smaller or bigger than a similarity threshold, we determine which one of the two frames is a member of the representative subset. As shown in Fig. 3, a representative subset selected from a sequence can cover the whole coverage of the point cloud sequence with fewer samples, reduce the occurrence of overlapping areas, and thus lower the annotation costs.

As for intra-frame selection, not all annotated points within the frame contribute to the model's improvement [52], that is, redundancy also exists in the intra-frame annotation. Besides, due to the particularity of the dense prediction task, it is laborious to annotate every point in PCSS task. To make the point cloud annotation more efficient and encourage maximizing the segmentation performance, we argue that the unit of point cloud annotation can be changed from the frame to a small portion of segmented regions [52]. Therefore, we make a tradeoff between annotating labor and efficiency to alleviate the expensive point-by-point labeling [43]. Specifically, we propose a novel method to reduce the redundancy of the intra-frame granularity under the guidance of uncertainty estimation and point cloud intensity. In detail, we first segment a point cloud into regions as the fundamental labeled units using two unsupervised algorithms [33, 46]. Next, uncertainty estimation is carried out on such segmented regions. Furthermore, to avoid selecting some typically uncertain segmented regions which exist in several point clouds, we introduce exclusive intensity information of point cloud [19] to complement segmented region information estimation. Finally, the segmented regions with uncertainty and diversity are selected to annotate.

AL aims at minimizing the training size, while exactly matching the natural demand of semisupervised learning [27]. Semisupervised learning utilizes both labeled and unlabeled data to train models and is well suited to solve *the lack of data* in real-world tasks. Pseudolabeling is one of the application methods in semisupervised learning. Its goal is to leverage the model trained by partially labeled data to predict unlabeled data for generating pseudolabels [51]. Then, data with high confidence in model prediction will be assigned pseudolabels. Therefore, the integration of semisupervised learning and AL has attracted research interest in recent years [45, 50]. However, this integration method used to PCCS is almost not involved in recent literature. In this paper, to further reduce the human annotating labor, we propose to automatically select and pseudolabel a portion of the confident unlabeled data. The proposed method aims at searching for the most certain and informative unlabeled data with the guidance of a high-confidence threshold. Specifically, we first leverage the trained model to predict unlabeled data for getting the prediction confidence. And further, the data with high prediction confidence are selected and added to the labeled data pool. Then, the labeled data and pseudolabeled data are exploited to fine-tune the model.

Experimental results show that our method significantly outperforms existing deep active learning approaches on the SemanticKITTI dataset and achieves state-of-the-art performance on the S3DIS dataset. Our proposed method could achieve the performance of 90% fully supervised learning, while less than 15% and 3% annotations are required on S3DIS and SemanticKITTI datasets, respectively. The ablation studies also verify the effectiveness of each component proposed in our method.

In summary, the major contributions of this paper are as follows:

- We propose a new multi-granularity sample selection and annotation AL pipeline for point cloud semantic segmentation.

- We introduce semisupervised learning to automatically select and annotate the high prediction confidence data for effectively reducing annotation costs.
- Experiments on challenging SemanticKITTI dataset show that our approach outperforms existing deep active learning methods in classification accuracy and could highly reduce human annotation labor and computational costs.

2 Related works

2.1 3D semantic segmentation

Recently, 3D PCSS has achieved great progress with the aid of deep learning. The purpose of 3D PCSS is to divide a point cloud into several objects according to the predicted semantic meanings of points. According to the representation of the point cloud data, 3D semantic segmentation methods can be classified into three categories: point-based [18, 35], projection-based [34], voxel-based [7, 29]. Point-based methods directly process unstructured point clouds, which suffer from efficiency bottlenecks. In order to employ the two-dimensional (2D) convolutional neural networks (CNN) architectures, projection-based methods focus on converting the 3D point cloud to 2D pseudo-images, yet resulting in information loss. Voxel-based methods convert a point cloud into 3D voxels processed by 3D volumetric convolutions. Although retains the 3D geometric information, it requires very high resolution in order not to lose much information. Overall, these methods heavily rely on fully annotated datasets, which require densely annotated point clouds that are laborious and time-consuming. To this end, we focus on how to train a model with less annotated data to achieve similar performance compared to fully supervised training.

2.2 Deep active learning

As a machine learning method, AL has been of research interest for a couple of decades for increasing label efficiency and reducing annotated costs. AL selects the most informative and representative samples from the unlabeled dataset into the labeled pool through the query strategy and then iteratively trains the model until the annotated budget is exhausted or the pre-defined termination conditions are reached. Therefore, the query strategy is becoming extremely important. The main query strategies include the uncertainty-based approach [4, 20, 23, 30], distribution-based approach [2, 14, 31] and expected model change approach [21, 41]. Various methods were proposed to measure the uncertainty of the unlabeled samples through the posterior probability of a predicted class [23], the difference between the first prediction and the second one [20], or the entropy of class posterior probabilities [30]. Some earlier studies [8, 42] also estimated the sample uncertainty referring to a committee of classifiers. The distribution-based approach queries samples by considering the selection of core subsets and chooses the samples which represent the whole dataset, like clustering algorithm [31], Gaussian process [14] and context-aware methods [2]. The expected model change approach primarily chooses the unlabeled samples that can make the largest change on the current model through estimating expected gradient length [41], expected future errors [38], or expected output changes [21].

Deep learning (DL) has achieved unparalleled breakthroughs in various fields, while DL is often very greedy for large amounts of labeled data [16]. Therefore, many researchers have high expectations for the results of combining DL and AL, referred to as deep active learning (DAL) [36], for AL's capacity to effectively reduce labeling costs. Gal et al. [12] proposed a significant AL framework for high-dimensional data based on Bayesian deep learning, to estimate uncertainty through Monte Carlo(MC) Dropout integration. However, Sener and Savarese [40] pointed out that this method is unsuitable for large datasets because of batch sampling. And then, they proposed a Core-set approach from the perspective of distribution to construct a core set which is representative of the entire original dataset. They considered minimizing the core-set loss is equivalent to the k-Center problem which can be tackled by an efficient approximate solution. William et al. [4] proposed an ensemble-based AL for deriving well-behaved uncertainty estimates for unlabeled data. Meanwhile, they compared it against the Bayesian deep learning approach [12] and the density-based approach [40], and the results show ensemble-based AL can effectively counteract the class-imbalanced problem during acquisition and lead to more calibrated predictive uncertainties. Yoo and Kweon [54] introduced a novel active learning method with a loss prediction module which is learned to predict the target loss of the unlabeled dataset. By considering the difference between a pair of loss predictions, the loss prediction module could discard the scale of the real loss changes. Inspired by semisupervised learning, some researchers [13, 17, 45, 50, 55] have assigned pseudo-labels to high-confidence samples in order to further improve the accuracy and keep the stability of the DAL model because of the majority and consistency. In addition, some researchers combined generative adversarial networks (GAN) [48], reinforcement learning [28], and transfer learning [10] with AL to achieve various purposes, respectively.

2.3 AL for semantic segmentation

Semantic segmentation has important applications in various fields, like autonomous driving [24], image processing [1], and high-resolution remote sensing [32]. Combining AL with semantic segmentation is also conducive to alleviating the annotation cost. Although many AL approaches for semantic segmentation have been proposed, most of them focus on 2D image segmentation [6, 22, 44, 53]. Recently, a few researchers are applying AL to 3D point cloud segmentation. Lin et al. [26] first combined AL with DL for semantic segmentation of large-scale airborne laser scanning (ALS) point clouds. They proposed a segment-based query function, considering interactions among points within segments, to assess the informativeness of samples. Based on the previous training framework, they introduced incremental learning to save the training time and added mutual information metric to estimate model-dependent uncertainty [25]. Shi et al. [43] proposed a super-point-based [11] AL strategy which could better exploit the limited annotation cost. And they further designed shape-level diversity and local spatial consistency constraint. Observing that only a small portion of annotated regions are sufficient for 3D scene understanding, Wu et al. [52] proposed a region-based and diversity-aware AL. In this paper, from the perspective of uncertainty, representativeness, and diversity, we propose a multi-granularity sample selection and annotation pipeline which combines the unique 3D geometric information of the point cloud and the sequential relationship between frames.

3 Methodology

In this section, we describe our multi-granularity and semisupervised AL pipeline in detail. We first introduce the architecture of our pipeline. Then, the proposed inter-frame selection approach is presented. And then, we introduce the segmented region-based inner frame selection strategy in detail. Furthermore, we illustrate how to compute the confidence of segmented regions to further apply pseudolabels for semisupervised learning task. Next, the details of the network adopted in our work are explained. Finally, we introduce how we leverage the query strategy to select the segmented regions with uncertainty and diversity for annotation and pick out segmented regions with high confidence probability for pseudolabeling.

3.1 Architecture of the proposed pipeline

The purpose of PCSS is to train a model by leveraging the dataset, and then, the model assigns a predicted label to each point, which is a dense prediction task. Therefore, the labor and time cost of sample annotation required in the training of PCSS model are very high. In order to improve the efficiency of manual annotation, we first achieve a representative subset D_{NDT} from the original point cloud dataset D_{orig} through the normal distributions transfer (NDT) algorithm. Next over-segments 3D point cloud scans from D_{NDT} into supervoxels using the voxel cloud connectivity segmentation (VCCS) [33] algorithm. Subsequently, the locally convex connected patches (LCCP) [46] algorithm is used to obtain the segmented regions from the generated supervoxels. Each segmented region contains several points, so it is convenient and time-saving to annotate such regions. So, we have a segmented 3D point cloud dataset D now, which can be divided into two subsets. One is a little labeled subset D_{L} containing randomly selected point cloud scans, and the other is a large unlabeled subset D_{U}.

Our multi-granularity and semisupervised active learning can be divided into 5 steps:

1. Achieving a representative subset D_{NDT} from the original point cloud dataset D_{orig} through the NDT algorithm.
2. Generating a segmented 3D point cloud dataset D through VCCS [33] and LCCP [46] algorithms.
3. Training a network on the current labeled subset D_{L} for assigning a label to each point.
4. Calculating the information score of segmented regions with two items: softmax entropy and intensity of point cloud as shown in Fig. 1a. And computing the softmax confidence of segmented regions as shown in Fig. 1c
5. Selecting $Top-K$ segmented regions for annotators to annotate exclusive labels, and moving them from the unlabeled subset D_{U} into the current labeled subset D_{L} as shown in Fig. 1b. Meanwhile picking out $Top-M$ segmented regions with pseudolabels from D_{U} and also feeding into D_{L} as shown in Fig. 1d.

3.2 Registration-based inter-frame selection

Generally speaking, a point cloud dataset contains multiple sequences, each of which contains multiple frames. Continuous frames in the same sequence have overlapping areas and include a large number of repeated categories, so we employ a point cloud matching approach to screen out a

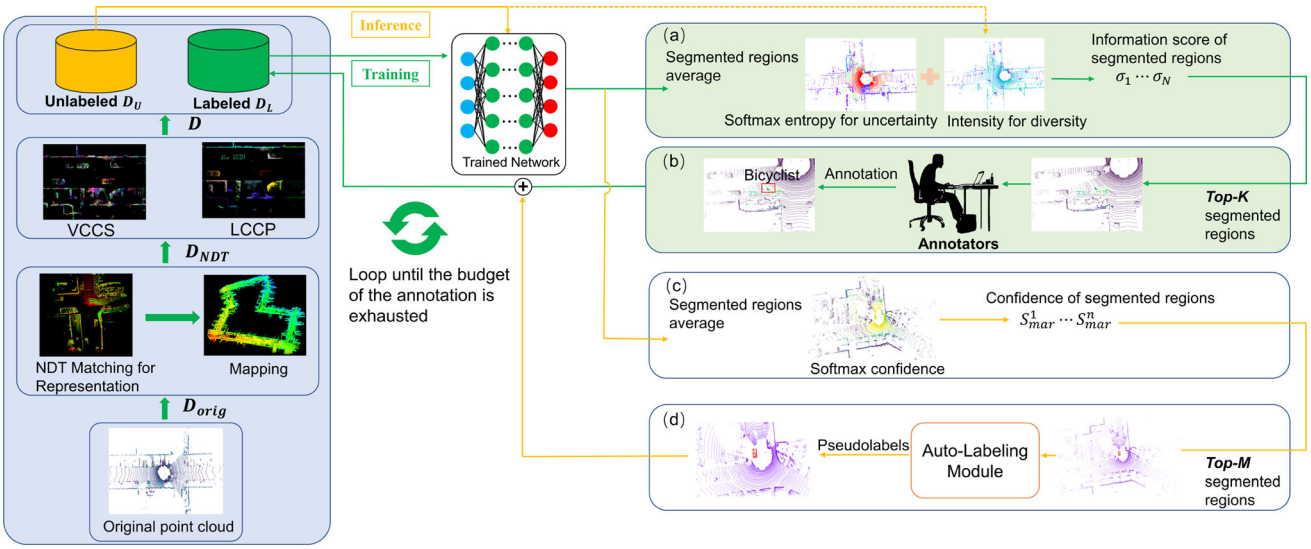

Fig. 1 Multi-granularity and semi-semisupervised active learning pipeline. In our proposed architecture, the network is first trained in supervision with labeled subset D_{L}. The network then produces softmax entropy and intensity of all segmented regions in unlabeled subset D_{U}. **a** Combining segmented region entropy with point cloud intensity to form the selection indicators. **b** The $Top-K$ segmented regions are selected for the annotator to label and moved to the labeled subset D_{L} for the next round. **c** Calculating the classification score for each segmented region. **d** Assigning pseudolabels to $Top-M$ segmented regions and moving them to the labeled subset D_{L}

subset which could represent the sequence from the perspective of building-map.

Considering robustness and efficiency, we choose the NDT algorithm [5] as the point cloud registration method. This is because NDT does not need to establish explicit correspondences between points or features, and all derivatives could be calculated analytically. The NDT transforms the discrete set of 2D points reconstructed from a single point cloud scan into a piecewise continuous and differentiable probability density, which consists of a set of normal distributions and can be used to match another scan through Newton's algorithm [5]. During the registration of the two point cloud scans through the NDT algorithm, if the registration process converges or reaches the maximum number of iterations, a registration score $\mathrm{score}_{\mathrm{match}}$ will be obtained, which is used to construct the matching score function for screening representative point clouds.

$$\mathrm{score}_{\mathrm{match}} = 1 - \sum_i \exp\left(\frac{-(x_i' - q_i)^t \sum_i^{-1} (x_i' - q_i)}{2}\right), \tag{1}$$

where x_i', $\sum_i^{-1}$ and q_i denotes the following notation:

- x_i' denotes the point x_i mapped into the coordinate frame of the target scan according to the parameters P of rotation and displacement. x_i is the reconstructed 2D point of laser scan sample i of the input scan in the coordinate frame of the input scan.
- $\sum_i$ and q_i represent the covariance matrix and the mean of the corresponding normal distribution to point x_i'.

In our work, when the registration score $\mathrm{score}_{\mathrm{match}}$ of two point cloud scans is less than a threshold δ_{match}, we consider that the overlapping area of two point cloud scans is large, and then discard the input frame and retain the target frame. On the contrary, when it is greater than δ_{match}, we take the current input frame as the target frame for the next matching. The outline of the proposed inter-frame selection approach, given a point cloud sequence $\mathbf{S} = \{s_1, s_2, \ldots, s_n\}$ of n scans and a initial representative subset $\mathbf{S}' = \{s_1\}$, is as follows:

1. Take the scan s_1 as the target frame and scan s_2 as the input frame, and then calculate their matching score $\mathrm{score}^{1-2}_{\mathrm{match}}$ through the NDT algorithm. If score $\mathrm{score}^{1-2}_{\mathrm{match}}$ is less than the threshold δ_{match}, there is no need to update subset $\mathbf{S}'$.
2. Next take the scan s_3 as the input frame, and perform the registration between scan s_3 and scan s_1. If their matching score $\mathrm{score}^{1-3}_{\mathrm{match}}$ is larger than the threshold δ_{match}, scan s_3 will be added to the subset $\mathbf{S}'$ and taken as the target frame at the same time.
3. Repeat the above steps until the point cloud registration of each frame in the sequence is completed.

And then, we can achieve a representative subset $\mathbf{S}' = \{s_1', s_2', \ldots, s_m'\}$ that represents the whole sequence. The process of inter-frame selection is illustrated in detail in Algorithm 1.

Algorithm 1 Registration Based Inter-Frame Selection

Input: Given a point cloud sequence $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}$ with n scans

Output: A representative subset $\mathbf{S}' = \{s'_1, s'_2, \cdots, s'_m\}$ with m scans

initialize $\mathbf{S}' = \{s_1\}$
$SourcePointCloud = s_1$
for $(i = 2;\ i <= n;\ i = i + 1)$ **do**
  $score_i = score_{match}(SourcePointCloud, s_i)$
  if $score_i > \delta_{threshold}$ **then**
    $SourcePointCloud = s_i$
    $\mathbf{S}'.append(s_i)$
  end if
end for
return $\mathbf{S}'$

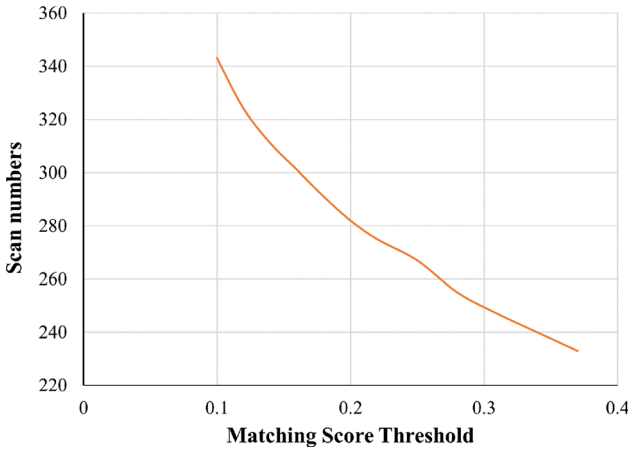

Fig. 2 The number of point clouds selected from the sequence 07 (with 1101 point cloud scans) in SemanticKITTI dataset by setting different thresholds

It is obvious that the number of point clouds selected from the same sequence will be different with different thresholds. Taking the sequence *07* (*with 1101 point cloud scans*) in SemanticKITTI dataset as an example, the number of point clouds selected by setting different thresholds is shown in Fig. 2. For example, when the threshold is $\delta_{\text{match}} = 0.2$, a representative subset $\mathbf{S}'$ (*with 330 point cloud scans*) is selected from the sequence *07*. Then, the selected point cloud scans are used to build the map, as shown in Fig. 3. The results show that the subset selected by the NDT matching algorithm can represent all the elements in the scene completely.

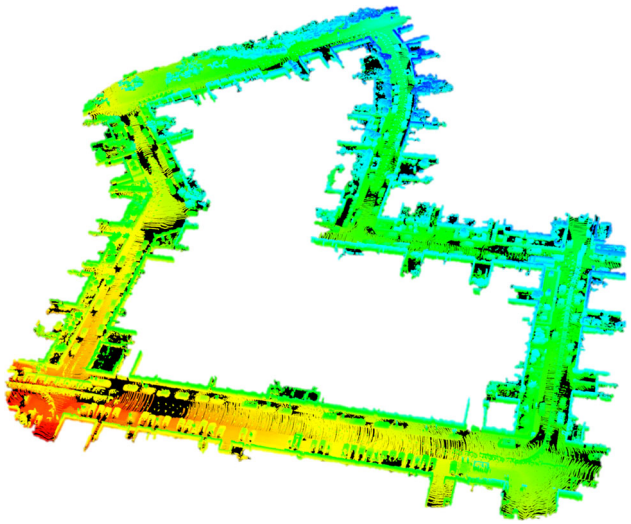

Fig. 3 Leveraging 330 representative point cloud scans selected from the sequence 07 in SemanticKITTI dataset with threshold $\delta_{\text{match}} = 0.2$ to build the map

3.3 Segmented region-based inner frame selection

The labeling cost varies greatly depending on target tasks. In the annotation process, it is relatively cheap to select closed polygons to form a semantic annotation for a 2D image, but 3D point-wise data require expensive point-by-point labeling [43, 54]. However, not all annotated points within the frame contribute to the model's improvement [52]. Besides, when annotating the same number of points, if the selected points are scattered in the whole frame, although the model performance may be very good, the difficulty and time consumption of annotation will be greatly increased, and it is hard to exploit the limited budget.

To alleviate the time and labor of manual point-by-point labeling, we first leverage VCCS [33] and LCCP [46] algorithms to segment a point cloud scan into segmented regions which can be taken as the fundamental label querying units. Then, in each active selection step, we calculate segmented regions information with softmax entropy and point cloud intensity.

3.3.1 Segmented regions generation

Geometrically constrained supervoxels All points in a point cloud scan are required to be annotated in the supervised task or conventional AL, which is labor-intensive. If we can divide a point cloud scan into connective segmented regions as the basic unit of annotation, it will greatly improve the efficiency of annotation. So, we first employ VCCS [33] algorithm to deal with the original point cloud scan for generating geometrically constrained supervoxels. The VCCS algorithm is composed of 4 parts: (1) construct

the adjacency graph for the voxel-cloud to ensure these supervoxels connection in space; (2) select a number of seed points to initialize the supervoxels; (3) calculate the normalized distance d_{norm} with three distances: spatial distance d_{s}, color distance d_{c} and distance d_{f} in fast point feature histograms (FPFH) space [39]; and (4) use a flow-constrained local iterative clustering for generating geometrically constrained supervoxels as shown in Fig. 4.

Point cloud partitioning These geometrically constrained supervoxels gained in the last step are not isolated; they can be further merged into larger segmented regions. So next we leverage LCCP [46] algorithm to segment the supervoxel adjacency graph by classifying whether the connection relation between two supervoxels is convex or concave through two criteria: extended convexity criterion (CC) and sanity criterion (SC). Finally, these small supervoxels can merge into larger segmented regions as shown in Fig. 5 through a region-growing process according to the discriminant results.

3.3.2 Calculating segmented regions information

In each AL selection step, the trained network predicts the probability $p(y_i = j|x_i)$ of each point x_i belonging to the j_{th} category. Then, we calculate the information of a segmented region from two aspects: (1) softmax entropy based on the probability; (2) point cloud intensity, which is introduced in detail as follows.

Segmented region entropy As a widely concerned aspect in AL, uncertainty sampling aims to select the most uncertain samples to annotate from unlabeled subset D_{U}. In this paper, we use softmax entropy to measure the uncertainty of a segmented region. We first obtain the softmax probability $p(y_i = j|x_i)$ of each point x_i belonging to the j_{th} category in the unlabeled subset D_{U}. Then, we calculate the region entropy E_n for the n_{th} segmented region R_n through

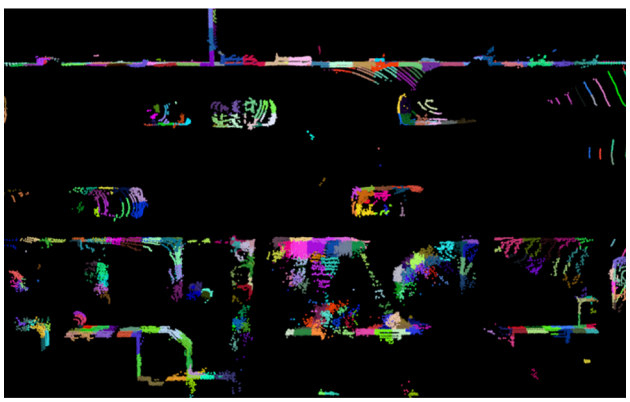

Fig. 4 Visualization of over-segmenting an original 3D point cloud scan into supervoxels using the VCCS [33] algorithm. Points of the same color belong to the same supervoxel

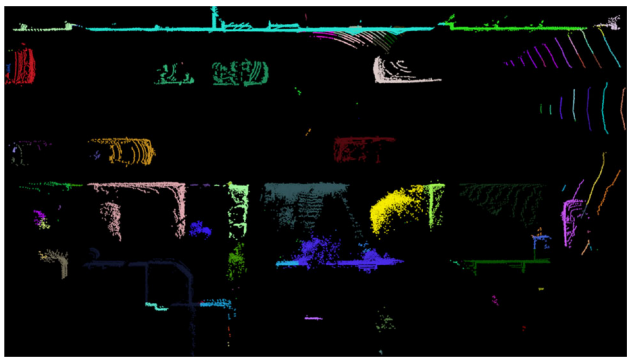

Fig. 5 Visualization of obtaining the segmented regions after using LCCP [46] algorithm to work on previous supervoxels. Points of the same color belong to the same segmented region

averaging the entropy of points within unlabeled region R_n as shown in Eq. 2,

$$E_n = \frac{1}{R_n}\sum_{i=1}^{R_n} -P(y_i = j|x_i;\Theta)\log P(y_i = j|x_i;\Theta), \tag{2}$$

where R_n contains N points, Θ denotes the network parameters. If the trained network is quite confident about a predicted category, it will assign a probability to that category greater than other categories. In this case, the entropy E_n is much lower than other categories. On the contrary, a higher entropy value is obtained when the trained network is not confident about a category in the prediction.

Point cloud intensity After obtaining the entropy E_n of each segmented region, the most obvious way is to select the top-ranked regions for annotation. However, these segmented regions with higher entropy E_n may result in redundant annotation effort if appearing in the same querying step. To increase diverse information for the network, we can leverage the intensity of each point in a point cloud scan. The reason for this is that intensity is different from material to material. The intensities of reflection on the same material are similar, while pulsed on different materials are different [19]. Based on this theory, we pick the intensity as a diversity-aware selection criterion to select diverse segmented regions for the network. We compute the region intensity score I_n for the n_{th} segmented region R_n by averaging intensity of points within unlabeled region R_n as shown in Eq. 3,

$$I_n = \frac{1}{R_n}\sum_{i=1}^{R_n} \rho_i, \tag{3}$$

where ρ_i is intensity of a point.

After calculating the softmax entropy E_n and intensity I_n of each segmented region, we can combine them linearly to form the information score σ_n of the n_{th} segmented region R_n as shown in Eq. 3.

$$\sigma_n = \alpha E_n + \beta I_n. \tag{4}$$

Finally, we can obtain a sorted information list σ,

$$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n). \tag{5}$$

3.4 Segmented region confidence estimation

In our work, at each AL iterative process, the most informative unlabeled segmented regions are selected for annotating, and the network is retrained with added labeled dataset. In this way, the redundant annotation of noninformative regions is avoided, greatly reducing human annotation labor. Actually, the subset D_{U} also contains an adequate amount of ignored unlabeled data with high confidence. After the network is trained with the initial labeled subset D_{L}, we can use its predictive capability to generate relatively accurate pseudolabels for unlabeled segmented regions in subset D_{U}.

We select the segmented regions with high confidence from subset D_{U}, when the predicted probability difference S_{mar} between the two most likely class labels is smaller than a threshold δ_H. The pseudolabel y_c^{pseudo} is defined as:

$$y_c^{\mathrm{pseudo}} = \begin{cases} \underset{j}{\operatorname{argmax}}\ p(y_i = j | x_i; \Theta), & \text{if } S_{\mathrm{mar}} > \delta_H \\ \text{None}, & \text{otherwise}, \end{cases} \tag{6}$$

where the threshold δ_H is set to a large value to achieve high confident pseudolabels. The S_{mar} is formulated as follows:

$$S_{\mathrm{mar}} = S_{\mathrm{conf}}^{c_1} - S_{\mathrm{conf}}^{c_2}, \tag{7}$$

where $S_{\mathrm{conf}}^{c_1}$, $S_{\mathrm{conf}}^{c_2}$ represent the classification scores of the highest and second highest predicted class labels for a segmented region, respectively. As shown in Eq. 8, given a segmented region R with N points, we calculate the confidence of the predicted class label for all points and achieve the classification scores $S_{\mathrm{conf}}^{c_1}$ and $S_{\mathrm{conf}}^{c_2}$ for a segmented region by averaging the predicted probabilities of all points in the segmented region.

$$\begin{aligned} S_{\mathrm{conf}}^{c_1} &= \frac{1}{N}\sum_{n=1}^{N} P\left(y_n^{c1} | R; \Theta\right), \\ S_{\mathrm{conf}}^{c_2} &= \frac{1}{N}\sum_{n=1}^{N} P\left(y_n^{c2} | R; \Theta\right). \end{aligned} \tag{8}$$

Through the probability difference S_{mar}, we can avoid selecting noisy segmented regions to assign pseudolabels.

For the segmented regions which meet the pseudolabeling condition, we can arrange each segmented region in descending order according to its probability difference S_{mar} to obtain a descending list φ_S,

$$\varphi_S = \left(S_{\mathrm{mar}}^1, S_{\mathrm{mar}}^2, \ldots, S_{\mathrm{mar}}^n\right). \tag{9}$$

3.5 PCSS network

The PCSS network is a crucial component in our pipeline for 3D deep learning. Currently, many point-based [35] and voxel-based [37] networks are proposed to process 3D data. However, most of these methods suffer from high memory consumption and computational costs. To better demonstrate the effectiveness of the proposed AL pipeline, we pick MinkowskiNet [7] based on sparse convolution and SPVCNN [29] based on point-voxel CNN as the PCSS networks in this paper.

MinkowskiNet is proposed for spatio-temporal perception which can directly process 3D point cloud scans using high-dimensional convolutions. To achieve this, it adopts sparse tensors and convolutions for three reasons:

1. The sparse tensor can better express and generalize high-dimensional spaces.
2. The sparse convolution is similar to the standard convolution which can leverage all architectural innovations such as residual connections and batch normalization.
3. The sparse convolution is efficient and fast according to only computing outputs for predefined coordinates and saving them into a compact sparse tensor.

To implement efficient and generalized sparse convolution, it proposes an open-source library which includes sparse tensor quantization, generalized sparse convolution, max pooling, and so on. Furthermore, *MinkowskiNet leverages* a hybrid kernel (cross-shaped kernel and cubic kernel) to resolve the problem of computational cost and the number of parameters in a network caused by increasing dimensions.

SPVCNN is composed of a fine-grained point-based branch that keeps the 3D data in high resolution without large memory footprint, and a coarse-grained voxel-based branch which aggregates the neighboring features without random memory accesses [29]. And for large outdoor scenes [3], it further proposes sparse point-voxel convolution (SPVConv) that enhances PVConv with the sparse convolution to enable higher resolutions in the voxel-based branch.

3.6 Annotating labels for segmented regions

On the one hand, according to the final decreasing order σ, we can select $Top - K$ segmented regions for annotators to assign labels. For the experiment, we actually regard the

ground truth of the segmented region as the labeled data instead of labeling by human annotators. Then, these labeled segmented regions D_{label} are moved from unlabeled subset D_{U} to labeled subset D_{L}. Note that only a small portion of a point cloud scan in each active selection is added to the subset D_{L} as shown in Fig. 1b, because we take the segmented region as the basic labeling unit.

On the other hand, after getting the final descending list φ_S, we select $Top - M$ segmented regions to assign pseudolabels. Then, these pseudolabeled regions D_{pseudo} are fed into the labeled subset D_{L} from unlabeled subset D_{U}. Accomplishing the segmented region information estimation, label annotation, region confidence estimation and pseudolabeling, we repeat the AL loop to fine-tune the PCSS network on the updated subset D_{L} until the annotated budget is exhausted or the iterations are reached. Note that after each fine-tuning step, we put the high-confidence samples D_{pseudo} back to D_{U} and erase their pseudolabels.

4 Experiments

In this section, we first introduce our experimental settings, including two datasets, the initial portion of all labeled point cloud scans, maximum iteration, and annotation budget. Then, we compare our approach with other existing methods to demonstrate the effectiveness of our method. Next, to verify the contribution of each individual strategy, we conduct ablation experiments. Finally, based on the experimental results, we present the limitations of our method and the directions for future work.

4.1 Experimental settings

4.1.1 Datasets

We evaluate the performance of our approach and compare it with the other AL methods on two large-scale challenging datasets, S3DIS and SemanticKITTI, respectively. S3DIS is a commonly used indoor semantic segmentation dataset which can be divided into 6 large areas, with a total of 271 rooms. We take Area5 as the validation set and perform active learning training on the remaining datasets. As for SemanticKITTI [3], it is a representative outdoor dataset which is released in 2019 for autonomous driving. SemanticKITTI consists of 22 sequences with total of *43,552 point cloud scans*, splitting sequences *00* to *10* as a training set where sequence *08* is used as the validation set and the rest sequences as the test set. And the total number of training points is $\text{total}_{\text{number}} = 2{,}349{,}559{,}532$.

4.1.2 Segmented region generation

We employ the VCCS [33] algorithm to over-segment a 3D point cloud scan into supervoxels with given voxel resolution R_{voxel} and seed resolution R_{seed}. Considering the density difference between indoor and outdoor point cloud, we set R_{voxel}, R_{seed} to a small value ($R_{\text{voxel}} = 0.05$, $R_{\text{seed}} = 0.5$) for S3DIS dataset, and a large value ($R_{\text{voxel}} = 0.15$, $R_{\text{seed}} = 3.5$) for SemanticKITTI dataset. The R_{voxel} represents the voxel resolution which will be used for the segmentation, R_{seed} denotes the distance between supervoxels. After that, flow-constrained local iterative clustering is used to generate geometrically constrained supervoxels based on spatial connection. Next, we utilize the LCCP algorithm to cluster these supervoxels into larger segmented regions through CC criterion with $\beta_{\text{Tresh}} = 10°$, and SC criterion with $\alpha_{\text{smooth}} = 0.1$. The β_{Tresh} denotes the concavity tolerance angle, and α_{smooth} is utilized to calculate the smoothness constraint.

4.1.3 Annotation budget

In each active label acquisition step, *because the number of points in different segmented regions varies*, we set the annotation budget as a fixed portion of total training points instead of a fixed number of segmented regions for the fair comparison with other methods. The number of pseudolabel acquisitions is also set as a fixed portion of the total points.

4.1.4 Active learning settings

At the beginning of each experiment, we first randomly select a small portion $x_{\text{init}}\%$ of fully labeled point clouds as the initially labeled subset D_{L} and treat the rest as the unlabeled subset D_{U}. Then, we perform K rounds as following steps: (1) Training the PCSS network on subset D_{L}; (2) Selecting a portion $x_{\text{label}}\%$ of total training points from subset D_{U} for annotation according to different AL querying methods; (3) If pseudolabels are adopted, select a portion $x_{\text{pseudo}}\%$ of total training points for assigning pseudolabels at $\delta_H = 0.9$; (4) Moving the newly annotated points into subset D_{L} and fine-tune the network. In order to ensure the reliability of the experimental results, each experiment is conducted three times and results are averaged.

Specifically, we set $x_{\text{init}} = 3\%$, $K = 7$ and $x_{\text{label}} = 2\%$ for S3DIS dataset, and $x_{\text{init}} = 1\%$, $K = 5$ and $x_{\text{label}} = 1\%$ for SemanticKITTI dataset [52].

4.1.5 Network training

For both S3DIS and SemanticKITTI datasets, the networks are trained with Adam optimizer (initial learning rate = 0.001) and cross-entropy loss [52]. And the voxel resolution of both datasets is set to 5 cm.

On the S3DIS dataset, we train the networks on 3 TITAN RTX GPUs with a batch size of nine. In the training, we first train both networks for 200 epochs on 3% of the fully labeled point cloud scans and then fine-tune the two networks for 150 epochs after adding 2% active annotated data into subset D_L each time. Since the point clouds in the S3DIS dataset do not include intensity information, we set $\alpha = 1, \beta = 0$ in Eq. 4 for the dataset.

On the SemanticKITTI dataset, we train both networks on 4 GTX 1080Ti GPUs and set the batch size to 8. In the training, we initially train both networks for 100 epochs on 1% of the fully labeled point cloud scans and then fine-tune the two networks for 30 epochs after adding 1% active annotated data into subset D_L each time. Referring to [52], the weight of softmax entropy in Eq. 4 is set as $\alpha = 1$. Based on the experimental results, we set $\beta = 0.05$.

4.2 Comparison with other methods

We compare our approach with 7 other AL methods, including random point cloud scans selection (RAND), uncertainty-based methods, such as softmax confidence (CONF [50]), softmax margin (MARG [50]), softmax entropy (ENT [50]) and segmented-entropy(SEG-ENT [13]), and diversity-based methods, such as core-set approach (CoSET [40]) and ReDAL [52], which is a recent region-based and diversity-aware AL approach.

4.2.1 Inter-frame selection

The inter-frame selection algorithm proposed in this research cannot be employed to reduce the inter-frame redundancy of the S3DIS dataset since the point clouds in the S3DIS dataset are not collected in chronological sequence. As a result, we only conduct inter-frame selection comparison experiments on the SemanticKITTI dataset. To fairly verify the effectiveness of the inter-frame selection method based on the NDT registration algorithm, we adopt the random selection method as the active query method. The experimental results are shown in Table 1. RAND and RAND_{NDT} indicate that point cloud scans are randomly selected from the original unlabeled dataset D_{orig} and the unlabeled dataset D_{NDT} for annotation, respectively. Note that the dataset D_{orig} *contains 19,130 raining point cloud scans*, after the NDT matching with the threshold $\delta_{\text{match}} = 0.1$, the dataset D_{NDT} *contains 9335 point cloud scans*. For the SPVCNN network, our inter-frame selection method can achieve 90% performance of the result of fully supervised methods ($\text{mIoU}^{\text{SPVCNN}}_{\text{supvis}} = 63.52\%$) with merely 5% of annotated data. With the MinkowskiNet network, our method is also better than RAND. Although the training data for active queries is reduced, our method makes the model be trained on more diverse and informative labeled data.

Table 1 Results of mIoU performance (%) on SemanticKITTI with SPVCNN and MinkowskiNet in frame annotation

% Labeled data	SPVCNN		MinkowskiNet	
	RAND	RAND_{NDT}	RAND	RAND_{NDT}
Init	41.84	45.99	37.74	39.56
2	45.41	49.66	42.74	43.79
3	52.19	53.84	48.82	48.47
4	54.76	56.41	52.51	52.81
5	56.89	57.78	54.67	55.69

4.2.2 Intra-frame selection

The visualization of SemanticKITTI on sequence 08 validation subset with SPVCNN network is shown in Fig. 6. And the experimental comparison results on the SemanticKITTI dataset are shown in Figs. 7 and 8 where the x-axis represents the percentage of annotated points, and the y-axis means the mIoU obtained by the network. Under both networks, our proposed multi-granularity and semisupervised AL pipeline consistently outperforms the previous methods over the PCSS task. We find that our method outperforms any other AL methods on two experiments with initial $x_{\text{init}} = 1\%$ labeled data. It verifies the effectiveness of the inter-frame selection method based on the NDT registration algorithm again.

As for the SPVCNN, in Table 2, we observe that our AL method can achieve 90% performance of the result of fully supervised methods with merely 3% of annotated data, and it reaches 97.95% fully supervised performance with 5% of annotated points. Particularly, it, respectively, outperforms the recent state-of-the-art (SOTA) method ReDAL [52] by 6.6%, 7.4%, 8.4%, and 6.9% when using 2%, 3%, 4%, and 5% labeled points. With the network of MinkowskiNet, in Table 3, our AL method can achieve 90% performance of the result of fully supervised methods ($\text{mIoU}^{\text{MinkuNet}}_{\text{supvis}} = 61.4\%$) with merely 2% of annotated data, and it can even reach 99.48% fully supervised performance with only 4% of annotated points.

On the S3DIS dataset, as shown in Figs. 9 and 10, our method highly outperforms any other AL methods except

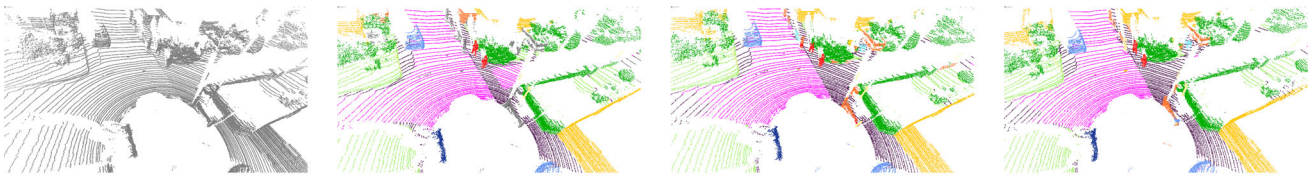

Fig. 6 Visualization of SemanticKITTI on sequence 08 validation subset with SPVCNN network. With our AL approach, the model can correctly identify persons on the sidewalk with merely 5% annotated points

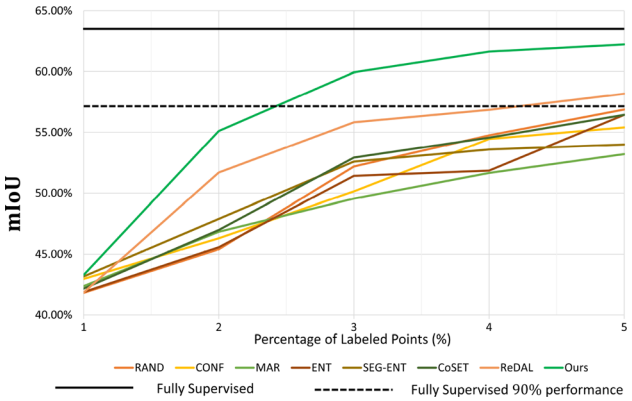

Fig. 7 Experimental results of different AL methods on SemanticKITTI with SPVCNN. We compare our multi-granularity and semisupervised AL method with other approaches. It is obvious that our method highly outperforms previous AL approaches

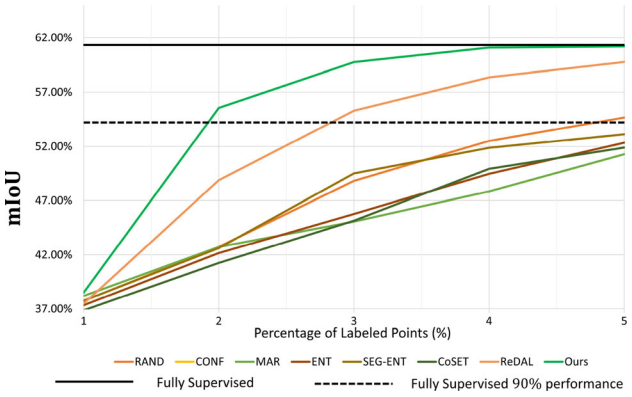

Fig. 8 Experimental results of different AL methods on SemanticKITTI with MinkowskiNet. We compare our multi-granularity and semisupervised AL method with other approaches. It is obvious that our method highly outperforms previous AL approaches

for ReDAL. As shown in Tables 4 and 5, the performance of mIoU we obtained is very close to those obtained by ReDAL. The main reason for this is that the point clouds in the S3DIS dataset do not include diverse intensity information. Therefore, we cannot leverage the intensity information of the point cloud to reduce its intra-frame redundancy which results in both networks being trained on the redundant annotated dataset. Furthermore, this result also demonstrates that our method achieves SOTA performance by leveraging segmented region entropy and pseudolabels.

4.3 Ablation studies

We verify the effectiveness of segmented region, point cloud intensity, pseudolabels and NDT in our proposed pipeline on SemanticKITTI dataset with 5% of annotated points for fair comparison.

The results are shown in Table 6 and Fig. 11 where ENT and ENT_{reg} represents querying the annotated points by calculating the softmax entropy of a point cloud scan and the segmented region entropy, respectively. *Inten*, *Pseu* and *NDT*, respectively, denote selecting the segmented regions using point cloud intensity, training the network with pseudolabels, and selecting segmented regions from a representative dataset screened out by the NDT algorithm.

In Table 6, we can observe that changing the annotating units from a point cloud scan to segmented regions contributes most to the improvement with about 6.15% mIoU. Furthermore, with the aid of *Inten*, *Pseu* and *NDT*, the mIoU performance of segmented region entropy yields an improvement of 1.90%, 2.84% and 2.49%, respectively.

Table 2 Results of mIoU performance (%) on SemanticKITTI with SPVCNN

% Labeled data	RAND	CONF	MAR	ENT	SEG-ENT	CoSET	ReDAL	Ours
Init	41.84	42.98	42.39	41.90	43.18	42.19	41.87	**43.30**
2	45.41	46.31	46.84	45.57	47.89	46.98	51.70	**55.12**
3	52.19	50.15	49.55	51.42	52.60	52.93	55.83	**59.94**
4	54.76	54.46	51.66	51.85	53.60	54.57	56.86	**61.63**
5	56.89	55.41	53.21	56.45	54.00	56.45	58.18	**62.22**

The bold text is used to highlight the best performance

Table 3 Results of mIoU performance (%) on SemanticKITTI with MinkowskiNet

% Labeled data	RAND	CONF	MAR	ENT	SEG-ENT	CoSET	ReDAL	Ours
Init	37.74	37.32	38.20	37.33	37.75	36.86	37.48	**38.50**
2	42.74	42.01	42.73	42.16	42.62	41.25	48.88	**55.56**
3	48.82	47.37	45.07	45.77	49.51	45.15	55.30	**59.75**
4	52.51	49.54	47.84	49.46	51.87	49.93	58.35	**61.08**
5	54.67	53.49	51.27	52.34	53.12	51.89	59.76	**61.18**

The bold text is used to highlight the best performance

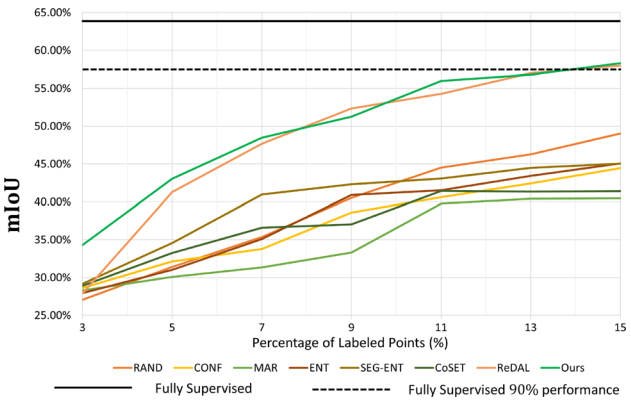

Fig. 9 Experimental results of different AL methods on S3DIS with SPVCNN. We compare our multi-granularity and semisupervised AL method with other approaches. Except for ReDAL approach, our method highly outperforms previous AL approaches

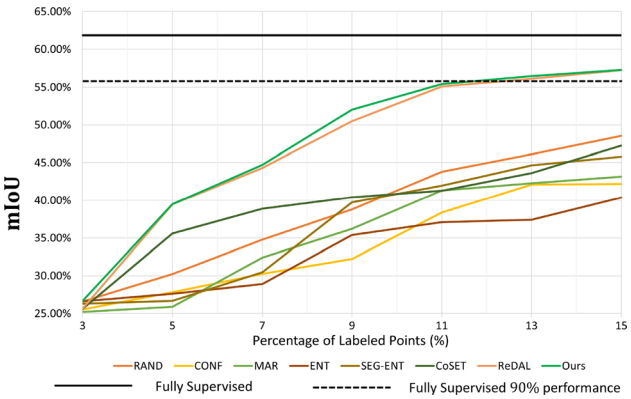

Fig. 10 Experimental results of different AL methods on S3DIS with MinkowskiNet. We compare our multi-granularity and semisupervised AL method with other approaches. Except for ReDAL approach, our method highly outperforms previous AL approaches

From the comparison of combination (ENT_{reg} + Inten) and combination (ENT_{reg} + Inten + Pseu), we find that pseudolabels play a key role in the performance of the trained network.

From Fig. 11, we observe that the performance of "ENT_{reg}" is similar to "ENT_{reg} + Inten." The reason is that without the diverse intensity information, the selected segmented regions still contain redundant regions. The result also validates the feasibility of selecting point cloud intensity information as the diversity indicator. Although the final performance of group (ENT_{reg} + Inten + Pseu) and group (ENT_{reg} + Inten + Pseu + NDT) is very close, the training data for the latter are reduced from *19,130 scans to 9335 scans after inter-frame selection*. This result shows that the inter-frame selection method effectively reduces inter-frame redundancy, and it enables the model to be trained on a more representative dataset. Despite the fact that the quantity of point clouds available for model training is reduced by 51.20%, the model performance is not compromised by the reduction in the training dataset. Besides less training data mean less training time consumption and storage consumption. The result also validates the importance of our inner selection strategy.

The group (ENT_{reg} + Pseu) outperforms the group (ENT_{reg} + NDT) by only 0.34%, and the performance of group (ENT_{reg} + Inten+NDT) is weaker than that of the

Table 4 Results of mIoU performance (%) on S3DIS with SPVCNN

% Labeled data	RAND	CONF	MAR	ENT	SEG-ENT	CoSET	ReDAL	Ours
Init	27.05	28.60	28.29	27.92	29.16	28.89	27.86	**34.29**
5	31.39	32.14	30.07	31.02	34.55	33.24	41.27	**43.04**
7	35.37	33.76	31.34	35.10	40.97	36.59	47.68	**48.48**
9	40.51	38.57	33.30	40.90	42.30	37.02	**52.34**	51.25
11	44.50	40.60	39.75	41.51	43.07	41.42	54.28	**55.98**
13	46.28	42.43	40.41	43.42	44.48	41.34	**57.01**	56.81
15	49.02	44.44	40.45	45.06	45.04	41.40	57.97	**58.30**

The bold text is used to highlight the best performance

Table 5 Results of mIoU performance (%) on S3DIS with MinkowskiNet

% Labeled data	RAND	CONF	MAR	ENT	SEG-ENT	CoSET	ReDAL	Ours
Init	26.59	25.52	25.20	26.60	26.30	25.60	25.63	**26.67**
5	30.22	27.81	25.87	27.60	26.66	35.58	39.45	**39.48**
7	34.76	30.25	32.40	28.91	30.45	38.88	44.29	**44.72**
9	38.79	32.23	36.20	35.40	39.72	40.41	50.50	**52.01**
11	43.80	38.39	41.31	37.10	41.95	41.28	55.11	**55.42**
13	46.13	42.10	42.28	37.42	44.66	43.63	56.14	**56.49**
15	48.57	42.18	43.15	40.37	45.79	47.26	57.26	**57.30**

The bold text is used to highlight the best performance

Table 6 Ablation study with 5% of annotated data on SemanticKITTI with SPVCNN network

ENT	ENT_{reg}	Inten	Pseu	NDT	mIoU (%)
√	–	–	–	–	56.45
–	√	–	–	–	59.93
–	√	√	–	–	61.07
–	√	–	√	–	61.63
–	√	–	–	√	61.42
–	√	√	√	–	62.18
-	√	√	–	√	61.58
–	√	√	√	√	62.22

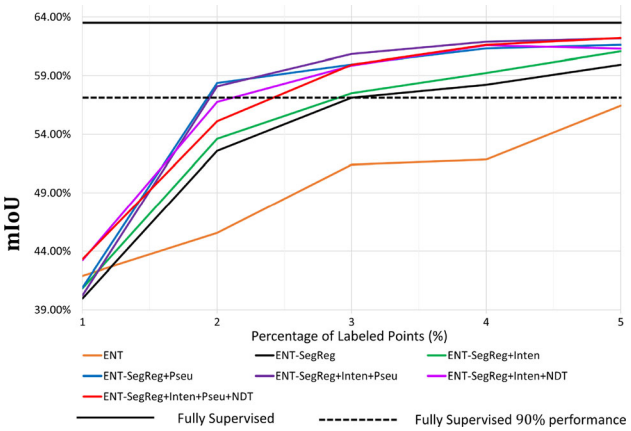

Fig. 11 Ablation study. Segmented region entropy, point cloud intensity, pseudolabels and NDT all yield improvements to mIoU

group (ENT_{reg} + Inten + Psu). It can be seen that the *Pseu* approach actually feeds the model with supplementary pseudolabeled training data, which can improve the model performance. The *NDT* method, on the other hand, enables the model to be trained on less redundant and more informative data. *Although it can improve the model performance, the NDT method is a coarse-grained selection method which filters out redundant information by the unit of frame. This way may result in the removal of data that is necessary for enhancing the model performance*. In summary, there are two ways to improve model performance, either by feeding the model with a large amount of trainable data, including pseudolabeled data, or by providing data that is diverse and representative.

4.4 Discussion

4.4.1 Per-class IoU results

A comparison of the performance of our method with fully supervised one is shown in Table 7. For the SPVCNN network, our method is on par with full supervision (Full) on most categories, and even better than that on the category of building. Although the performance on the three categories of other-vehicle, parking, and terrain is weaker than full supervised one, our method achieves 91%, 86%, and 93% of full supervised result, respectively. The main reason for that is that the inter-frame filtering method is a coarse-grained method, which may result in filtering out some useful information. Another possible reason is the imbalanced class distributions in the SemanticKITTI dataset. As for the MinkowskiNet network, our method outperforms the fully supervised result for some small objects, such as motorcycle, person and bicyclist.

4.4.2 Performance change

To investigate the relationship between segmentation performance and the proportion of annotated data, we expand the annotated data proportion to 10% and conduct experiments on the SemanticKITTI dataset, as shown in Fig. 12. The results show that our method achieves 99.15% performance of full supervision result with 10% of the annotated data. It can be seen that the model performance slowly improves from 62.11 to 62.98% as the annotated data increase from 5 to 10%. The main reason for the slow performance improvement is that as the proportion of annotated data increases, the proportion of the new annotated data that is valid for the model decreases. Another possible reason is that the diversity filtering criteria proposed in this paper, when designing the active query

Table 7 Per-class results of IoU performance(%) with 5% of annotated data on SemanticKITTI with two networks

Method	mIoU	Car	Bicycle	Motorcycle	Truck	Other-vehicle	Person	Bicyclist	Motorcyclist	Road
SPVCNN										
Full	63.5	96.7	41.4	66.0	79.9	60.1	67.1	84.9	0.0	93.6
Ours	62.2	95.9	38.6	64.9	80.4	54.8	66.6	84.0	0.0	92.8
MinkowskiNet										
Full	61.4	95.9	20.4	63.9	70.3	45.5	65.0	78.5	0.4	93.5
Ours	61.2	95.8	32.5	67.3	61.1	49.5	71.3	83.6	0.0	92.8

Method	Parking	Sidewalk	Other-ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic-sign
SPVCNN										
Full	50.6	81.1	0.0	82.8	59.5	88.0	65.3	75.2	64.4	50.5
Ours	43.6	79.6	0.0	89.3	57.5	86.2	65.0	69.9	63.2	49.8
MinkowskiNet										
Full	50.6	82.0	0.2	91.2	63.8	87.2	68.5	74.3	64.4	50.1
Ours	44.7	79.8	0.6	89.8	57.2	87.0	65.0	73.2	62.6	48.5

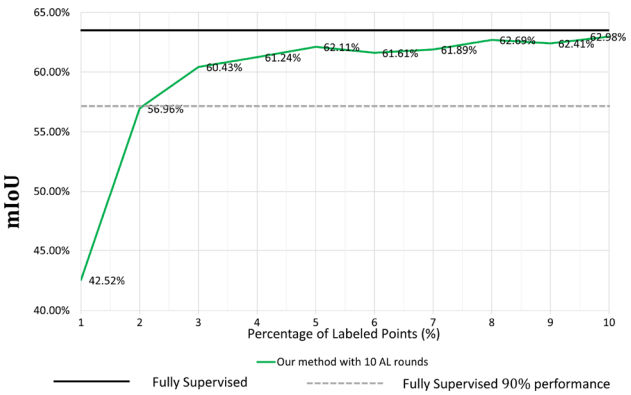

Fig. 12 Experimental results of our method on SemanticKITTI with more annotated data. Segmented region entropy, point cloud intensity, pseudolabels and NDT all yield improvements to mIoU

Table 8 Computational time (min) with 5% of annotated data on SemanticKITTI with SPVCNN network

% Labeled data	Time	ENT_{reg}	ENT_{int}	ENT_{pse}	ENT_{ndt}
Init.	T_{train}	0.83	0.78	0.78	0.83
	T_{cal}	25.93	24.00	28.30	10.88
2	T_{train}	71.98	71.50	69.15	35.63
	T_{cal}	18.85	18.63	23.88	8.97
3	T_{train}	71.92	71.02	70.13	35.93
	T_{cal}	17.90	18.67	21.87	8.22
4	T_{train}	71.93	70.28	69.96	35.93
	T_{cal}	17.15	17.28	20.68	7.95
5	T_{train}	72.47	70.08	69.90	35.40
	T_{cal}	17.10	17.18	20.62	7.90

function, only utilizes one piece of information, the point cloud intensity, which makes it difficult to feed the model with more diverse data.

4.4.3 Computational costs

We report the computational time (in minutes) of four methods presented in ablation study in Table 8 where ENT_{int}, ENT_{pse} and ENT_{ndt}, respectively, denote "ENT_{reg} + Inten," "ENT_{reg} + Pseu" and "ENT_{reg} + NDT." And T_{train} and T_{cal} denote the average time per epoch in an AL loop and the calculating time for active querying, respectively.

Because the amount of annotated data is the same for the initial training, the training time T_{train} is approximately the same for each method. It can be seen that as the proportion of annotated data increases, the calculation time T_{cal} for active querying tends to decrease. It is because the amount of data in the unlabeled dataset D_{U} gradually decreases, resulting in less computation on querying. Using point cloud intensity data for calculating segmented regions information has no impact on calculation time T_{cal}. The addition of pseudolabeling, on the other hand, increases active querying time by 19.0%, with a mean value of 23.07 min. It can be seen that compared to the ENT_{reg} method, the mean calculating time T_{cal} and training time T_{train} of the ENT_{ndt} method are reduced by 54.6% and 50.5%, respectively. The reason for computational costs reduction is that after NDT-based inter-frame selection, and the number of point clouds in the training set decreased from *19,130 to 9335*, resulting in a considerable reduction

Table 9 The mIoU performance (%) of hyper-parameters with 5% of annotated data on SemanticKITTI with SPVCNN network

% Labeled data	β			x_{pseudo} (%)			δ_{match}		
	0.025	0.05	0.1	0.25	0.5	1	0.1	0.2	0.4
Init	43.24	43.30	43.96	41.98	43.30	44.18	43.30	43.97	45.78
2	55.01	55.12	56.44	55.78	55.12	56.99	55.12	55.95	56.20
3	59.42	59.94	60.11	60.58	59.94	60.39	59.94	59.12	59.14
4	61.48	61.63	60.66	61.06	61.63	62.04	61.63	61.50	59.57
5	62.01	62.22	61.43	61.07	62.22	62.11	62.22	61.06	60.28

in computation in the unlabeled dataset. This result also validates the importance of our registration-based inter-frame selection.

4.4.4 Hyper-parameters analysis

We conduct a parametric study of three important parameters proposed in our method, which are the registration threshold (δ_{match}), the weight of point cloud intensity (β) in Eq. 4, and the proportion of pseudolabeled data (x_{pseudo}). During the experiment, we keep the other settings unchanged and then evaluate how mIoU performance varies with the set parameters, the results are shown in Table 9. It can be seen that the mIou performance decays when the point cloud intensity threshold is gradually increasing. It is because samples with uncertainty are more important for improving model performance than samples with diversity [52]. As the x_{pseudo} increases, the model performance tends to flatten out. The reason may be that the amount of mislabeled data in the D_{pseudo} dataset also increases as the x_{pseudo} increases, which has a negative effect on the model performance. Although the NDT-based inter-frame selection can effectively reduce computational costs, it can degrade model performance when the δ_{match} is set too large. This is because NDT-based inter-frame selection is a coarse-grained selection method that may result in the removal of data that is necessary for enhancing the model performance.

4.5 Limitations and future work

Although our proposed method proved to be effective in reducing human annotation labor and computational costs, there are still two pivotal limitations. The first is that the diversity filtering criteria proposed in this paper, when designing the active query function, only utilize the point cloud intensity. In fact, there is additional information that can be used, such as the color properties contained in the S3DIS dataset's point clouds. It is because regions with substantial color variances are more likely to suggest semantic diversity.

The other limitation is that the imbalance of categories in the dataset during the acquisition is not considered. Deep learning is usually trained and evaluated with the assumption that the dataset is balanced or nearly so. In reality, datasets in real-world scenarios are frequently unevenly distributed between categories, such as S3DIS and SemanticKITTI. The model trained on a skewed dataset is likely to be overwhelmed by samples coming from the majority categories. To summarize, we argue that active learning should not only select informative and diverse samples to decrease annotating costs, but should also be able to alleviate the imbalance in the labeled subset for improving the model's accuracy and robustness. In addition, scribble-annotation is a popular and effective method that retains as much information as possible to allow relatively high performance when compared to fully supervised training [49]. In future work, active learning can be integrated with scribble-annotations, i.e., only scribbling the uncertain and diverse data, to further minimize annotation labor.

5 Conclusion

In this paper, we propose a multi-granularity and semisupervised active learning pipeline for point cloud semantic segmentation. We first propose the novel inter-frame selection module based on the NDT registration algorithm to select a representative subset. Then, two key components, the segmented region entropy and point cloud intensity, are designed to select the most informative and diverse regions to annotate rather than a traditional point cloud scan. Next, through the efficient pseudolabeling method, our method further achieves high-cost efficiency. Finally, we conduct extensive experiments and ablation studies with two networks on SemanticKITTI dataset, where our method substantially achieves SOTA cost efficiency and greatly outperforms all existing works.

Acknowledgements The work was supported in part by the National Key Research and Development Program of China under Grant

No. 2021YFB2501300 and in part by the National Important Science & Technology Specific Projects under Grant No. 2017ZX01038201.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author (Z. Pan) and S. Ye on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abdel-Salam R, Mostafa R, Abdel-Gawad AH (2022) RIECNN: real-time image enhanced CNN for traffic sign recognition. Neural Comput Appl 34:6085–6096. https://doi.org/10.1007/s00521-021-06762-5
2. Aodha OM, Campbell ND, Kautz J et al (2014) Hierarchical subquery evaluation for active learning on a graph. In: 2014 IEEE conference on computer vision and pattern recognition, pp 564–571. https://doi.org/10.1109/CVPR.2014.79
3. Behley J, Garbade M, Milioto A et al (2019) SemanticKITTI: a dataset for semantic scene understanding of lidar sequences. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 9296–9306. https://doi.org/10.1109/ICCV.2019.00939
4. Beluch WH, Genewein T, Nurnberger A et al (2018) The power of ensembles for active learning in image classification. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 9368–9377. https://doi.org/10.1109/CVPR.2018.00976
5. Biber P, Straßer W (2003) The normal distributions transform: a new approach to laser scan matching. In: 2003 IEEE/RSJ international conference on intelligent robots and systems, Las Vegas, Nevada, USA, October 27–November 1, 2003. IEEE, pp 2743–2748. https://doi.org/10.1109/IROS.2003.1249285
6. Casanova A, Pinheiro PO, Rostamzadeh N et al (2020) Reinforced active learning for image segmentation. In: 8th International conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net
7. Choy C, Gwak J, Savarese S (2019) 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3070–3079. https://doi.org/10.1109/CVPR.2019.00319
8. Dagan I, Engelson SP (1995) Committee-based sampling for training probabilistic classifiers. In: Machine learning, proceedings of the twelfth international conference on machine learning, Tahoe City, California, USA, July 9–12, 1995. Morgan Kaufmann, pp 150–157. https://doi.org/10.1016/b978-1-55860-377-6.50027-x
9. Dai A, Chang AX, Savva M et al (2017) Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 2432–2443. https://doi.org/10.1109/CVPR.2017.261
10. Deng C, Xue Y, Liu X et al (2019) Active transfer learning network: a unified deep joint spectral-spatial feature learning model for hyperspectral image classification. IEEE Trans Geosci Remote Sens 57(3):1741–1754. https://doi.org/10.1109/TGRS.2018.2868851
11. Deng S, Dong Q, Liu B, Hu Z (2022) Superpoint-guided semi-supervised semantic segmentation of 3D point clouds. In: 2022 International conference on robotics and automation (ICRA), pp 9214–9220. https://doi.org/10.1109/ICRA46639.2022.9811904
12. Gal Y, Islam R, Ghahramani Z (2017) Deep Bayesian active learning with image data. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, proceedings of machine learning research, vol 70. PMLR, pp 1183–1192
13. Gu B, Zhai Z, Deng C et al (2021) Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data. IEEE Trans Neural Netw Learn Syst 32(9):4111–4122. https://doi.org/10.1109/TNNLS.2020.3016928
14. Guo Y (2010) Active instance sampling via matrix partition. In: Advances in neural information processing systems, vol 23. Curran Associates Inc., pp 802–810
15. Hackel T, Savinov N, Ladicky L et al (2017) Semantic3d.net: a new large-scale point cloud classification benchmark. In: ISPRS Annals of the photogrammetry, remote sensing and spatial information sciences, pp 91–98
16. Hinton GE, Srivastava N, Krizhevsky A et al (2012) Improving neural networks by preventing co-adaptation of feature detectors. CoRR arXiv:1207.0580
17. Hossain HMS, Roy N (2019) Active deep learning for activity recognition with context aware annotator selection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019. ACM, pp 1862–1870. https://doi.org/10.1145/3292500.3330688
18. Hu Q, Yang B, Xie L et al (2020) Randla-net: efficient semantic segmentation of large-scale point clouds. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 11105–11114. https://doi.org/10.1109/CVPR42600.2020.01112
19. Hui L, Di L, Xianfeng H et al (2008) Laser intensity used in classification of lidar point cloud data. In: IGARSS 2008—2008 IEEE international geoscience and remote sensing symposium, pp II-1140–II-1143. https://doi.org/10.1109/IGARSS.2008.4779201
20. Joshi AJ, Porikli F, Papanikolopoulos N (2009) Multi-class active learning for image classification. In: 2009 IEEE conference on computer vision and pattern recognition, pp 2372–2379. https://doi.org/10.1109/CVPR.2009.5206627
21. Käding C, Rodner E, Freytag A et al (2016) Active and continuous exploration with deep neural networks and expected model output changes. CoRR arXiv:1612.06129
22. Konyushkova K, Sznitman R, Fua P (2015) Introducing geometry in active learning for image segmentation. In: 2015 IEEE international conference on computer vision (ICCV), pp 2974–2982. https://doi.org/10.1109/ICCV.2015.340
23. Lewis DD, Catlett J (1994) Heterogeneous uncertainty sampling for supervised learning. In: Machine learning, proceedings of the eleventh international conference, Rutgers University, New Brunswick, NJ, USA, July 10–13, 1994. Morgan Kaufmann, pp 148–156. https://doi.org/10.1016/b978-1-55860-335-6.50026-x
24. Li J, Jiang F, Yang J et al (2021) Lane-deeplab: lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing 465:15–25. https://doi.org/10.1016/j.neucom.2021.08.105
25. Lin Y, Vosselman G, Cao Y et al (2020) Active and incremental learning for semantic ALS point cloud segmentation. ISPRS J Photogramm Remote Sens 169:73–92. https://doi.org/10.1016/j.isprsjprs.2020.09.003
26. Lin Y, Vosselman G, Cao Y et al (2020) Efficient training of semantic point cloud segmentation via active learning. In: ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, pp 243–250. https://doi.org/10.5194/isprs-annals-V-2-2020-243-2020

27. Liu C, Li J, He L (2019) Superpixel-based semisupervised active learning for hyperspectral image classification. IEEE J Sel Top Appl Earth Observ Remote Sens 12(1):357–370. https://doi.org/10.1109/JSTARS.2018.2880562
28. Liu Z, Wang J, Gong S et al (2019) Deep reinforcement active learning for human-in-the-loop person re-identification. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 6121–6130. https://doi.org/10.1109/ICCV.2019.00622
29. Liu Z, Tang H, Zhao S et al (2021) Pvnas: 3d neural architecture search with point-voxel convolution. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3109025
30. Luo W, Schwing AG, Urtasun R (2013) Latent structured active learning. In: Advances in neural information processing systems 26: 27th annual conference on neural information processing systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States, pp 728–736
31. Nguyen HT, Smeulders AWM (2004) Active learning using pre-clustering. In: Machine learning, proceedings of the twenty-first international conference ICML 2004, Banff, Alberta, Canada, July 4–8, 2004, ACM international conference proceeding series, vol 69. ACM. https://doi.org/10.1145/1015330.1015349
32. Pan Y, Pi D, Chen J et al (2021) FDPPGAN: remote sensing image fusion based on deep perceptual patchGAN. Neural Comput Appl 33:9589–9605. https://doi.org/10.1007/s00521-021-05724-1
33. Papon J, Abramov A, Schoeler M et al (2013) Voxel cloud connectivity segmentation - supervoxels for point clouds. In: 2013 IEEE conference on computer vision and pattern recognition, pp 2027–2034. https://doi.org/10.1109/CVPR.2013.264
34. Peng K, Fei J, Yang K et al (2022) MASS: multi-attentional semantic segmentation of LiDAR data for dense top-view understanding. IEEE Trans Intell Transp Syst 23(9):15824–15840. https://doi.org/10.1109/TITS.2022.3145588
35. Qi CR, Yi L, Su H et al (2017) Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 5099–5108
36. Ren P, Xiao Y, Chang X et al (2022) A survey of deep active learning. ACM Comput Surv 54(9):180:1-180:40. https://doi.org/10.1145/3472291
37. Riegler G, Ulusoy AO, Geiger A (2017) Octnet: learning deep 3d representations at high resolutions. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6620–6629. https://doi.org/10.1109/CVPR.2017.701
38. Roy N, Mccallum A (2001) Toward optimal active learning through Monte Carlo estimation of error reduction. In: Proceedings of the international conference on machine learning, pp 441–448
39. Rusu RB, Blodow N, Beetz M (2009) Fast point feature histograms (FPFH) for 3D registration. In: 2009 IEEE international conference on robotics and automation, pp 3212–3217. https://doi.org/10.1109/ROBOT.2009.5152473
40. Sener O, Savarese S (2018) Active learning for convolutional neural networks: a core-set approach. In: 6th International conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings. OpenReview.net
41. Settles B, Craven M, Ray S (2007) Multiple-instance active learning. In: Advances in neural information processing systems 20, proceedings of the twenty-first annual conference on neural information processing systems, Vancouver, British Columbia, Canada, December 3–6, 2007. Curran Associates Inc., pp 1289–1296
42. Seung HS, Opper M, Sompolinsky H (1992) Query by committee. In: Proceedings of the fifth annual ACM conference on computational learning theory, COLT 1992, Pittsburgh, PA, USA, July 27–29, 1992. ACM, pp 287–294. https://doi.org/10.1145/130385.130417
43. Shi X, Xu X, Chen K et al (2021) Label-efficient point cloud semantic segmentation: an active learning approach. CoRR arXiv:2101.06931
44. Siddiqui Y, Valentin J, Niessner M (2020) Viewal: active learning with viewpoint entropy for semantic segmentation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9430–9440. https://doi.org/10.1109/CVPR42600.2020.00945
45. Siméoni O, Budnik M, Avrithis Y et al (2021) Rethinking deep active learning: using unlabeled data at model training. In: 2020 25th International conference on pattern recognition (ICPR), pp 1220–1227. https://doi.org/10.1109/ICPR48806.2021.9412716
46. Stein SC, Schoeler M, Papon J et al (2014) Object partitioning using local convexity. In: 2014 IEEE conference on computer vision and pattern recognition, pp 304–311. https://doi.org/10.1109/CVPR.2014.46
47. Tatarchenko M, Park J, Koltun V et al (2018) Tangent convolutions for dense prediction in 3D. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 3887–3896. https://doi.org/10.1109/CVPR.2018.00409
48. Tran T, Do T, Reid ID et al (2019) Bayesian generative active deep learning. In: Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, proceedings of machine learning research, vol 97. PMLR, pp 6295–6304
49. Unal O, Dai D, Gool L van, Zurich E (2022) Scribble-supervised LiDAR semantic segmentation. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2697–2707
50. Wang K, Zhang D, Li Y et al (2017) Cost-effective active learning for deep image classification. IEEE Trans Circuits Syst Video Technol 27(12):2591–2600. https://doi.org/10.1109/TCSVT.2016.2589879
51. Wang J-X, Chen S-B, Ding CHQ, Tang J, Luo B (2022) RanPaste: paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens 60:1–16. https://doi.org/10.1109/TGRS.2021.3102026
52. Wu TH, Liu YC, Huang YK et al (2021) Redal: region-based and diversity-aware active learning for point cloud semantic segmentation. In: 2021 IEEE/CVF international conference on computer vision (ICCV), pp 15490–15499. https://doi.org/10.1109/ICCV48922.2021.01522
53. Xie B, Yuan L, Li S, Liu CH, Cheng X (2022) Towards fewer annotations: active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8058–8068. https://doi.org/10.1109/CVPR52688.2022.00790
54. Yoo D, Kweon IS (2019) Learning loss for active learning. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 93–102. https://doi.org/10.1109/CVPR.2019.00018
55. Yuan T, Wan F, Fu M et al (2021) Multiple instance active learning for object detection. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5326–5335. https://doi.org/10.1109/CVPR46437.2021.00529