Weakly Supervised Point Cloud Semantic Segmentation via Artificial Oracle

Hyeokjun Kweon*
KAIST
0327june@kaist.ac.kr

Jihun Kim*
KAIST
jihun1998@kaist.ac.kr

Kuk-Jin Yoon
KAIST
kjyoon@kaist.ac.kr

Abstract

Manual annotation of every point in a point cloud is a costly and labor-intensive process. While weakly supervised point cloud semantic segmentation (WSPCSS) with sparse annotation shows promise, the limited information from initial sparse labels can place an upper bound on performance. As a new research direction for WSPCSS, we propose a novel Region Exploration via Artificial Labeling (REAL) framework. It leverages a foundational image model as an artificial oracle within the active learning context, eliminating the need for manual annotation by a human oracle. To integrate the 2D model into the 3D domain, we first introduce a Projection-based Point-to-Segment (PP2S) module, designed to enable prompt segmentation of 3D data without additional training. The REAL framework samples query points based on model predictions and requests annotations from PP2S, dynamically refining labels and improving model training. Furthermore, to overcome several challenges of employing an artificial model as an oracle, we formulate effective query sampling and label updating strategies. Our comprehensive experiments and comparisons demonstrate that the REAL framework significantly outperforms existing methods across various benchmarks. The code is available at https://github.com/jihun1998/AO.

1. Introduction

Point cloud semantic segmentation has been extensively researched for robotics or autonomous driving applications. While fully supervised approaches [7, 10, 14, 28, 29, 35] utilizing deep learning have demonstrated remarkable progress, the densely annotated point-wise labels required for these methods pose a significant challenge due to the labor-intensive and expensive annotation process. This obstacle hinders the practical application of existing methods.

In response to this challenge, Weakly Supervised Point Cloud Semantic Segmentation (WSPCSS) has gained attention, leveraging more affordable weak labels [15, 23, 24, 26, 40, 42, 47, 48]. Among various types of weak labels, sparse annotations have become a widely adopted setting, where only a tiny subset (*e.g.*, 0.02%) of the entire point cloud is labeled. Existing WSPCSS works have proposed remedies focusing on two key aspects: 1) fully utilizing the limited information from sparse labels and 2) learning the features beneficial for segmentation from the data. The methods based on self-labeling use the regions confidently predicted by the model as pseudo-labels for training the model itself. Meanwhile, various studies have proposed self-supervised learning methods, such as consistency against data augmentation, contrastive learning, or masked modeling.

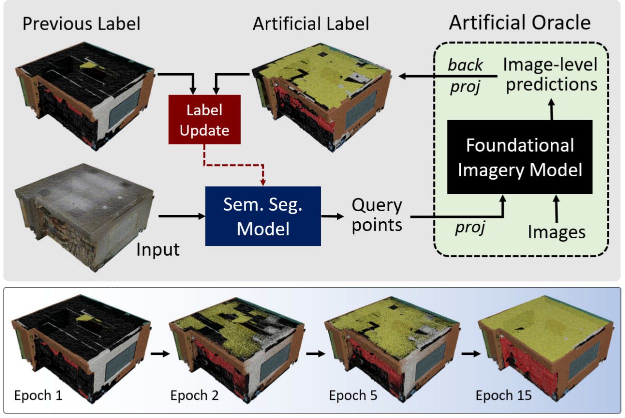

Figure 1. Visualization of the proposed REAL framework. We tackle the challenge of WSPCSS, leveraging a foundational imagery model as an artificial oracle within the context of active learning. The below examples illustrate the evolving labels as training proceeds.

While these approaches have shown substantial results, the overall amount of information provided is constrained by the initial sparse annotation, potentially placing an upper limit on performance. In this paper, instead of solely focusing on how to exploit the given limited information, we embark on a new and more challenging research direction. Specifically, we propose a novel WSPCSS method using an additional publicly available source of information—leveraging advancements in the 2D domain with the Segment Anything Model (SAM) [19].

We begin by utilizing the promptable segmentation capability of SAM to enhance initially provided sparse annotations. To bridge the gap between the 2D model and 3D point cloud data, we introduce the **Projection-based Point-to-Segment (PP2S)** module. This module comprises the projection of weakly annotated points onto images, acquiring 2D masks using SAM, and back-projecting these masks into 3D, facilitating segmentation in the point cloud. The proposed PP2S-based preprocessing effectively improves label quality, resulting in significantly enhanced semantic segmentation performance.

However, the labels preprocessed by PP2S remain incomplete due to potential noises from calibration errors or SAM ambiguity. To further enhance label quality during training, we propose the **Region Exploration via Artificial Labeling (REAL)** framework, utilizing PP2S as an artificial oracle within the context of active learning, as depicted in Fig. 1. The REAL framework samples query points according to model predictions and subsequently requests annotations from PP2S to dynamically enhance training labels. Here, unlike general active learning methods relying on a human oracle, our artificial oracle cannot directly assign the classes of the requested query. Therefore, we design specialized strategies for query sampling and label refinement to unlock the segmentation capability of the PP2S.

Note that the main novelty of our approach lies in how we effectively integrate the benefits of SAM into the realm of WSPCSS. We analyze the REAL framework through extensive experiments, including quantitative and qualitative ablation studies. Furthermore, our method achieves new state-of-the-art (SoTA) results on S3DIS [3] and ScanNetV2 [8] under all the tested settings, surpassing the existing WSPCSS methods. Remarkably, even with the 0.004% setting, our method outperforms conventional methods using the 0.02% setting, underscoring the superiority of the proposed approach.

2. Related Work

2.1. Weakly Supervised Approaches

Originating from weakly supervised learning in imagery domain [2, 21, 22, 31, 39, 45], weakly supervised approach has emerged as a cost-effective approach to mitigate the challenges associated with acquiring fully labeled data [15, 18, 23, 24, 26, 40, 42, 47, 48]. Xu and Lee [40] first propose weakly supervised configurations involving annotations of 10% of the data and a single point label for each category. PSD [48] introduces a perturbed branch and constraint self-distillation loss between the perturbed and original branches. OTOC [26] utilizes pseudo-label methods and iteratively updates pseudo-labels through a self-training strategy. MIL [42] proposes the transformer model derived from multiple instance learning and integrates adaptive global weighted pooling to their model. HybridCR [23] introduces an architecture that leverages pseudo-label methods and consistency regularization between the original and augmented branches. CPCM [24] introduces a region-wise masking strategy and contextual masked training method to integrate the benefit of the masked autoencoder. Despite the significant progress in WSPCSS, in this paper, we would like to explore a new and more challenging research direction. In this light, we suggest a novel WSPCSS method, using a publicly available additional source of information.

2.2. Point Cloud Segmentation Using 2D Models

There have been extensive efforts to achieve point cloud segmentation in an image-based manner as in [12, 16, 20, 30, 43]. With the rapid advancement in the 2D domain, multi-modal self-supervised approaches [1, 6, 13] and employing language-based model [9, 27, 34] have also been particularly prevalent. Recently, there has been a surge in attempts to address various point cloud tasks using the Segment Anything Model (SAM) [19] The zero-shot approach [44, 46], which directly applies SAM to the image corresponding to the point cloud, has demonstrated promising segmentation results. However, SAM's lack of explicit semantics limits its applicability to high-level perception tasks. Adaptation-based approaches [11] can make SAM directly learn 3D tasks; however, they necessitate dense ground truth data during the learning process, rendering them unsuitable for WSPCSS. To tackle the weakly supervised setting, this paper presents a novel WSPCSS method that effectively refines weak labels using the foundational 2D model as an artificial oracle.

2.3. Active Learning for Point Cloud Segmentation

Active Learning (AL) aims to identify a subset of instances that an oracle manually annotates during training. As in [17, 32, 36, 41], AL has also been extensively explored for WSPCSS. Given that the REAL framework actively employs SAM as an artificial oracle, it may share a conceptual scheme with conventional AL-based methods to some extent. Nevertheless, note that our weakly-supervised setting significantly differs from them. Our framework solely relies on initial sparse labels and does not involve any additional manual annotation within the training loop.

3. Method

3.1. Preliminaries

Problem setting. We denote an input point cloud as a set of N points: $X = \{x_1, \ldots, x_N\}$. In typical fully-supervised point cloud semantic segmentation, the labels for all points are provided as $Y = \{y_1, \ldots, y_N\}$ where $y_i \in \{1, \ldots, C\}$ and C is the number of classes. On the other hand, in weakly-supervised point cloud semantic segmentation, only a tiny subset of the points is labeled, and the other points are unlabelled. We formulate this setting as $(X, Y) = (X^L, Y^L) \cup (X^U, Y^U)$. Here, $X^L = \{x_1^L, \ldots, x_n^L\}$ is the points labeled with sparse annotations $Y^L = \{y_1^L, \ldots, y_n^L\}$. Conversely, $X^U = \{x_1^U, \ldots, x_{N-n}^U\}$ is not labeled, and we represent it as $Y^U = \{y_1^U, \ldots, y_{N-n}^U\}$, where $y_i^U = 0$.

Segment Anything Model. SAM is a foundational model for promptable segmentation, comprising 1) an image encoder, 2) a prompt encoder, and 3) a decoder predicting the mask using the embeddings from encoders. We denote the inference process of SAM as

$$M^{2D} = \text{SAM}(I; Z), \tag{1}$$

where I and Z are an input image and prompt, respectively. M^{2D} denotes a set of the pixels of the predicted mask. In this paper, we use a point prompt (*i.e.*, a pixel coordinate).

3.2. Projection-based Point-to-Segment

We begin by enhancing the initially given weak annotations using SAM. However, SAM is a vision foundation model designed and trained for 2D imagery. Therefore, to directly process 3D point cloud data using SAM, it is necessary to conduct additional adaptation or fine-tuning.

Instead, we propose to utilize image data that captures the same scene as the provided point cloud data. We believe that incorporating image data is not overly restrictive in practice. In fact, point cloud data is usually created from RGBD frames using Structure from Motion (SfM) [3, 8] or captured by LiDAR paired with RGBD cameras [4, 5, 33]. Hence, most point cloud datasets naturally contain RGB images, depth maps, and associated camera parameters. Under this philosophy, we introduce a **Projection-based Point-to-Segment (PP2S)** module, enabling promptable segmentation of 3D point cloud data using SAM via projection. It is noteworthy that our method utilizes image data during the training phase only and can perform semantic segmentation using input point cloud data alone during testing.

Figure 2 illustrates the visualization of PP2S. Formally, for each point cloud X, we assume that the model can access data from J cameras: images, depth maps, and projection matrices. Initially, we project the set of weakly annotated points $X^L = \{x_i^L\}$ onto the image I_j as:

$$p_{i,j} = \text{proj}_j x_i^L \quad \text{where} \quad i \in \{1, \ldots, n\}, j \in \{1, \ldots, J\} \tag{2}$$

where proj_j denotes the projection matrix, and $p_{i,j}$ represents the 2D coordinate of the ith weakly annotated points on the jth image. After the projection, we discard points that fall outside the image boundaries. We also filter out occluded points using the depth map D_j, though alternative visibility testing algorithms can be employed.

Then, SAM performs segmentation using the coordinates of projected points as individual input prompts as

$$M_{i,j}^{2D} = \text{SAM}(I_j; p_{i,j}), \tag{3}$$

where $M_{i,j}^{2D}$ is a set of the pixels of predicted 2D mask on I_j, given input point prompt $p_{i,j}$. Since $p_{i,j}$ corresponds to x_i^L and is labeled as y_i^L, the pixels in $M_{i,j}^{2D}$ are also assigned to y_i^L. As multiple cameras exist per a single point cloud, we conduct the above process for all images in parallel.

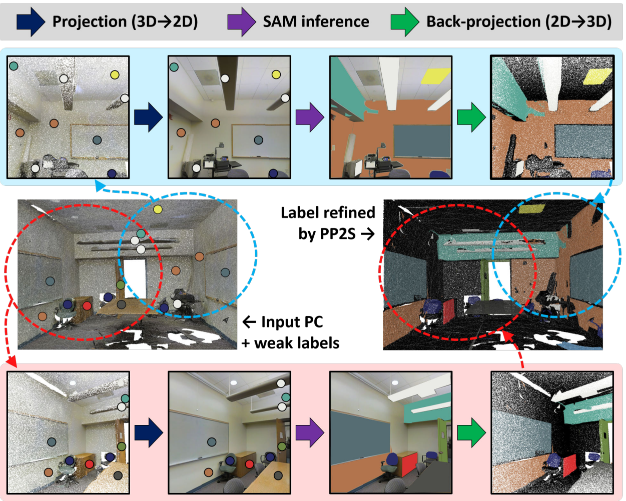

Figure 2. Visualization of the proposed **Projection-based Point-to-Segment (PP2S)** module. The process begins by projecting the weak labels onto images through projection matrices. Using these projected points as point prompts, SAM predicts 2D masks, which are subsequently back-projected onto the point cloud.

Subsequently, we back-project the acquired 2D masks to 3D coordinates. Initially, similar to Equ. 2, we project the point cloud onto the jth image as $p_{k,j} = \text{proj}_j x_k$, where x_k is the kth point in the point cloud and $p_{k,j}$ denotes the projection of x_k. If $p_{k,j}$ is an element of $M_{i,j}^{2D}$, it means that the projection result of x_k belongs to the mask of the weakly annotated point x_i^L, within the image I_j. Therefore, we define the 3D mask $M_{i,j}^{3D}$ as the set of x_k such that its projection $p_{k,j}$ is in $M_{i,j}^{2D}$.

The above process can be formulated as

$$M_{i,j}^{3D} = \{x_k | p_{k,j} \in M_{i,j}^{2D}\}, \tag{4}$$

where $M_{i,j}^{3D}$ denotes the set of points predicted by SAM to be the same segment with x_i^L from the perspective of jth image. Accordingly, $y_{i,j}^{3D}$, the class of $M_{i,j}^{3D}$, is y_i^L.

However, the 3D masks are not mutually exclusive. Given that a single 3D point could be observed from multiple images, it may be shared by more than one 3D mask. To address this, we establish a consensus among the 3D masks through a mask-wise voting system. In our system, each 3D mask $M_{i,j}^{3D}$ casts a vote for its points being set to $y_{i,j}^{3D}$. Therefore, from the perspective of each point, its class is determined by the voting of the 3D masks that include the point. The above process can be formulated as

$$v_k = \{y_{i,j}^{3D}\} \ \forall\, i, j \ \text{ such that } \ x_k \in M_{i,j}^{3D}, \tag{5}$$

where v_k is a multiset of the voted classes. Finally, we assign the class with the highest number of votes among v_k as

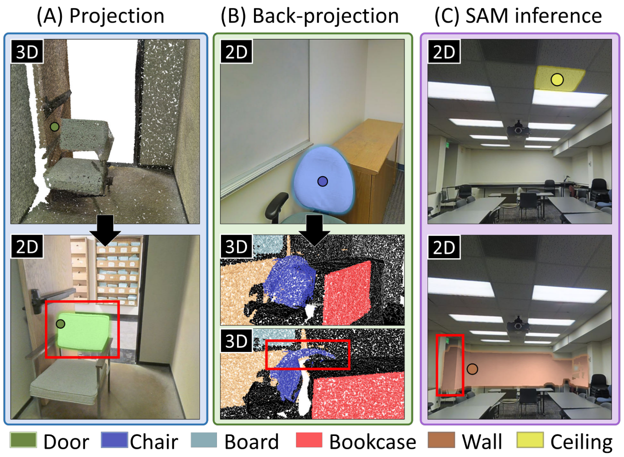

Figure 3. Possible sources of noise in PP2S. **(A):** The 3D *door* point is projected onto the *chair* in the image, resulting in an incorrect 2D mask. **(B):** Erroneous back-projection causes *wall* points in 3D to be mislabeled as *chair*. **(C) Above:** The 2D mask of the *ceiling* covers only a small portion of the entire object. **(C) Below:** The 2D mask of the *wall* encroaches into the regions of *column*.

the artificial label a_k for kth point x_k:

$$a_k = \begin{cases} 0 & \text{if } v_k = \emptyset \\ \text{Mode}(v_k) & \text{otherwise,} \end{cases} \tag{6}$$

where the Mode operator returns the most frequent element.

The proposed PP2S module effectively leverages the segmentation capability of SAM to expand the sparsely annotated initial labels. These resulting artificial labels, denoted as $A = \{a_k\}$, can directly serve as a supervised signal for training. Notably, we have observed that training with A yields significant performance improvements compared to using initial Y. The detailed experimental results are demonstrated in Section 4.

3.3. Region Exploration via Artificial Labeling

As mentioned, the proposed PP2S effectively enhances the initial sparse annotations, resulting in a substantial improvement in segmentation performance. However, we also have recognized several limitations associated with this preprocessing approach. The PP2S-enhanced labels can serve as better guidance than the initial sparse labels. Nevertheless, they are still fixed during the entire training phase and thereby impose an upper bound on performance. Notably, we cannot guide the regions not covered by the 2D masks obtained within PP2S. Furthermore, even if we assume that the initially provided sparse annotations are perfectly correct, errors stemming from factors such as noises of camera projection matrices (Fig. 3 A, B) or inherent ambiguity of SAM (Fig. 3 C) are inevitable. While the voting system can mitigate some of these, it cannot completely eliminate them.

Algorithm 1 Region Exploring via Artificial Labeling

```
Require: Segmentation model $f_\theta$, Training dataset $\mathcal{D} = \{(X, Y)\}$, Learning rate $\eta$, Number of epochs $T$
Ensure: Optimized $f_\theta$
for $(X, Y)$ in $\mathcal{D}$ do
    $Y \leftarrow \text{PP2S}(X; X^L)$
end for
Initialize the model parameter $\theta$.
for $t \leftarrow 1, 2, \dots, T$ do
    for $(X, Y)$ in $\mathcal{D}$ do
        $Y^{t-1} \leftarrow Y$
        $S \leftarrow \text{softmax}(f_\theta(X))$
        $\hat{Y} \leftarrow \text{argmax}(S)$
        $Q \leftarrow \text{QuerySampling}(X, S, Y^{t-1})$
        $A \leftarrow \text{PP2S}(X; Q)$
        $Y^t \leftarrow \text{LabelUpdate}(Y^{t-1}, A, \hat{Y})$
        $Y \leftarrow Y^t$
        $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(S, Y)$
    end for
end for
```

Then, how can we further improve label quality, especially during training? We draw inspiration from the concept of active learning. In active labeling, the model identifies a set of instances, known as a query, during training and requests human annotators (*i.e.*, oracle) to label these points. Typically, the query consists of instances that confuse the model at a given timestep, making annotating the query highly effective for the model to learn decision boundaries. However, active learning requires manual annotation within the training loop, potentially making it even more expensive than WSPCSS.

To address this challenge, we present the **Region Exploration via Artificial Labeling (REAL)** framework—a novel WSPCSS method utilizing the PP2S module as an **artificial oracle**, as depicted in Fig. 4. Specifically, the proposed PP2S module serves as an oracle in the REAL framework, responding to requests to annotate the query. Given that the artificial oracle is, in fact, a pretrained network, the REAL framework is completely relieved of the necessity for manual annotation, except for the initial sparse labels.

However, unlike a human oracle in conventional methods, our artificial oracle lacks explicit semantic knowledge. Therefore, we cannot expect the PP2S module to directly label the classes for the confusing points. Instead, we focus on leveraging what PP2S can provide to us. PP2S cannot directly annotate the given points; however, it can precisely predict which points should be grouped into a segment. To fully utilize this attribute, we define the query $Q = \{q_i\}$ to consist of the confidently identified points, instead of ambiguous ones from the perspective of the model. Trusting the prediction of the model on these points, we can self-label the query with minimal risk. We represent this pro-

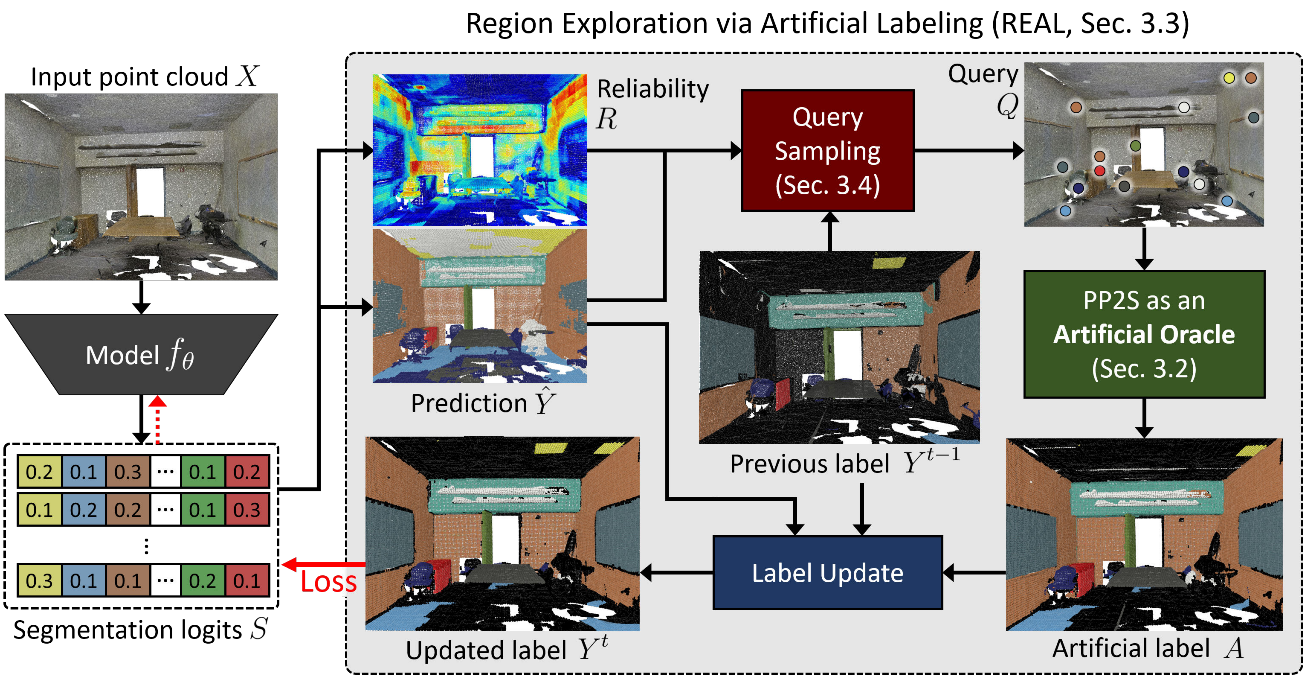

Figure 4. Visualization of the proposed Region Exploration via Artificial Labeling (REAL) framework. The objective is to dynamically enhance the quality of labels for weakly supervised point cloud semantic segmentation. In the REAL framework, the model f_θ takes the point cloud X as input and predicts semantic segmentation logits S. We then sample the most confident points from the unlabeled regions of the previous label Y^{t-1} based on the predictions $\hat{Y}$ and reliability scores R. The sampled points then serve as prompts for the proposed PP2S module, producing artificial labels A. Finally, by combining the model prediction and A, we refine Y^{t-1} into the updated labels Y^t, which serve as pseudo-ground truth for training the logits.

cess as QuerySampling in Algorithm 1, which will be thoroughly discussed in Section 3.4.

Subsequently, we request the PP2S module to act as an artificial oracle, annotating the unknown regions using the given query. For this, the self-labeled query serves as an input prompt for the PP2S module as follows:

$$A = \text{PP2S}(X; Q), \tag{7}$$

where A represents the **artificial label** obtained by PP2S from the given query Q. This approach is akin to using a set of sparsely annotated points (X^L) for PP2S in the preprocessing phase, as described in Section 3.2. The primary distinction in the REAL framework is that the input prompt is self-labeled, not manually annotated.

With the obtained artificial label A, we update the label of the previous step, Y^{t-1}, into Y^t. However, A may not be entirely reliable, as it could be more prone to noise compared to labels provided by a human oracle. This potential for noise could arise from inaccuracies in the self-labeled input query or the issues highlighted in Fig. 3.

To mitigate the risk of incorporating errors during the label-updating process, we devise a cautious approach using the prediction of the model. Specifically, if a discrepancy exists between the class assigned by the artificial label A and the model's prediction $\hat{Y}$ for a specific point, we exclude that point from the updating process. This conservative update process can be formulated as:

$$y_i^t = \begin{cases} a_i & \text{if } a_i = \hat{y}_i \\ y_i^{t-1} & \text{otherwise.} \end{cases} \tag{8}$$

Finally, the model f_θ is optimized by minimizing the cross-entropy loss between $\hat{Y}$ and Y^t.

3.4. Query Sampling Strategy for REAL

In the REAL framework, query sampling is a critical component, as the quality of the artificial label depends on the quality of the input query. We posit that an effective query must meet two key criteria: 1) it should exhibit a high level of confidence, and 2) it should aid the model in acquiring new knowledge. To address the first criterion, we design a margin-based reliability metric for evaluating the confidence of the model's prediction for each point as follows:

$$r_i = \text{Max}(s_i) - \text{Max2}(s_i), \tag{9}$$

where $s_i \in \mathbb{R}^C$ is the predicted logit of the ith point x_i, and r_i denotes its reliability. Max and Max2 are operators that return the maximum and the second maximum el-

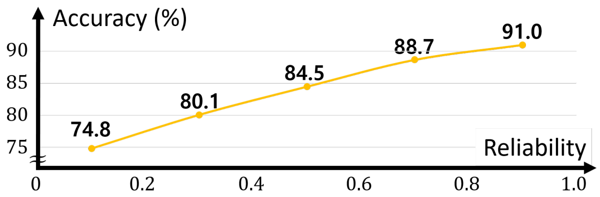

Figure 5. Verification of the reliability metric. The points presenting higher reliability show higher semantic segmentation accuracy.

ements of the input vector, respectively. We verify a robust correlation between the devised reliability metric and the accuracy of model predictions, as illustrated in Fig. 5. For instance, among points with reliability scores exceeding 0.5, the model correctly predicts the class for 84.5% of them. We set the threshold for query sampling at 0.95, corresponding to around 94% accuracy.

On the contrary, directly quantifying the utility of a query for the second criterion is inherently elusive. In this context, we propose an intuitive hypothesis: if the model has successfully leveraged the labels from the previous timestep, Y^{t-1}, the regions previously annotated in Y^{t-1} are expected to contribute less novel information during the current timestep, t. Therefore, the focus of new queries should be on effectively exploring the regions not previously or incorrectly labeled in Y^{t-1}.

To this end, we narrow the sampling range of the query to the points that are predicted by the model differently from the previous label. We formulate the above constraint as

$$Z_c = \{x_i\} \ \forall\, i \ \text{such that} \ \hat{y}_i = c \ \text{and} \ y_i^{t-1} \neq c. \tag{10}$$

Here, Z_c is the mother set for sampling q_c, the query point of the class c. Considering the first criterion, we sample q_c as the most reliable point among Z_c, and Q is the set of q_c. We formulate the above processes as

$$Q = \text{QuerySampling}(X, S, Y^{t-1}). \tag{11}$$

4. Experiments

4.1. Experimental settings

Datasets. We conduct experiments on S3DIS [3] and ScanNetV2 [8]. S3DIS comprises 6 areas with 272 rooms and 13 categories. We use the area 5 for evaluation following previous studies. ScanNetV2 includes 1613 3D scenes with 20 categories. We follow the official split (1201 training, 312 validation, and 100 test scenes). In addition to the point cloud data, the proposed REAL framework requires the data from cameras (*i.e.*, images, depth maps, and camera projection matrices) to leverage SAM. The datasets inherently include such data since they are reconstructed from the sequence of RGBD frames. During training, we utilize the data from 48 and 17 cameras per scene on average for S3DIS and ScanNetV2, respectively. Note that our framework does not require 2D data at inference time.

Table 1. A quantitative comparison between the initially sparse annotations (**Initial**) and the artificial labels from PP2S (**PP2S**). Pre, Rec, and # denote precision, recall, and the proportion of annotated points relative to the total number of points. The performance of the point cloud semantic segmentation model trained by each model is represented as **Model mIoU**. Every metric is in %.

Labels	mIoU	Pre	Rec	#	Model mIoU
Initial	0.004	100	0.004	0.004	39.7
PP2S	30.4	83.1	32.4	38.4	53.0

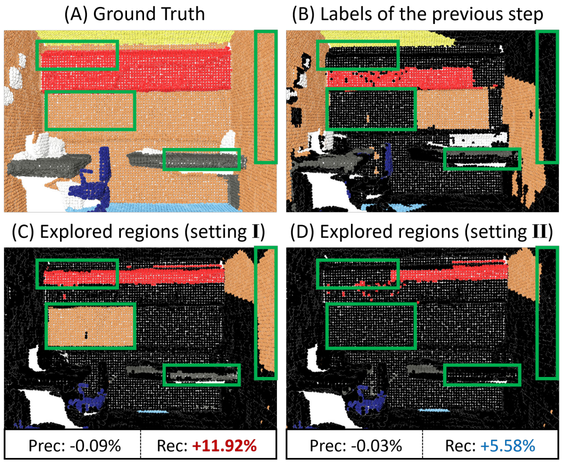
(A) Ground Truth
(B) Labels of the previous step
(C) Explored regions (setting I)
(D) Explored regions (setting II)
Prec: -0.09%
Rec: +11.92%
Prec: -0.03%
Rec: +5.58%

Figure 6. Validation of our query sampling strategy. Setting **I** samples the queries following our strategy while setting **II** samples from the entire point set. We present the improvement obtained by artificial labels in terms of precision (**Prec**) and recall (**Rec**).

Evaluation protocols. Previous studies [23, 24, 42, 48] have conducted experiments with various levels of sparsity. We believe that the most cost-efficient yet practical setting involves labeling only one point per object. Accordingly, our experiments mainly target the 1pt weak setting, which corresponds to 0.004% of points in S3DIS. For a fair comparison, we also conduct experiments under the 0.02% setting, which is the most widely used. For ScanNetV2, we follow the conventional works, using 20pt per scene. The overall performance is assessed on all points within the test set. Following the standard practice, we employ the mean Intersection over Union (mIoU) as our quantitative metric.

Implementation details. We choose PTv2 [37] network as our main backbone. To verify the effect of the backbone, we additionally test Closer [25] backbone for the 0.004% setting on S3DIS. Training details are the same as standard settings of the backbones. We employ a pretrained ViT-H model for SAM. Further details are in the *Supp.*

4.2. Analysis on REAL

PP2S-based preprocessing. We first validate the preprocessing strategy built upon the proposed PP2S. This strategy aims to enhance the initial sparse annotations into artificial labels by harnessing the promptable segmentation capability of SAM via projection. Table 1 provides various met-

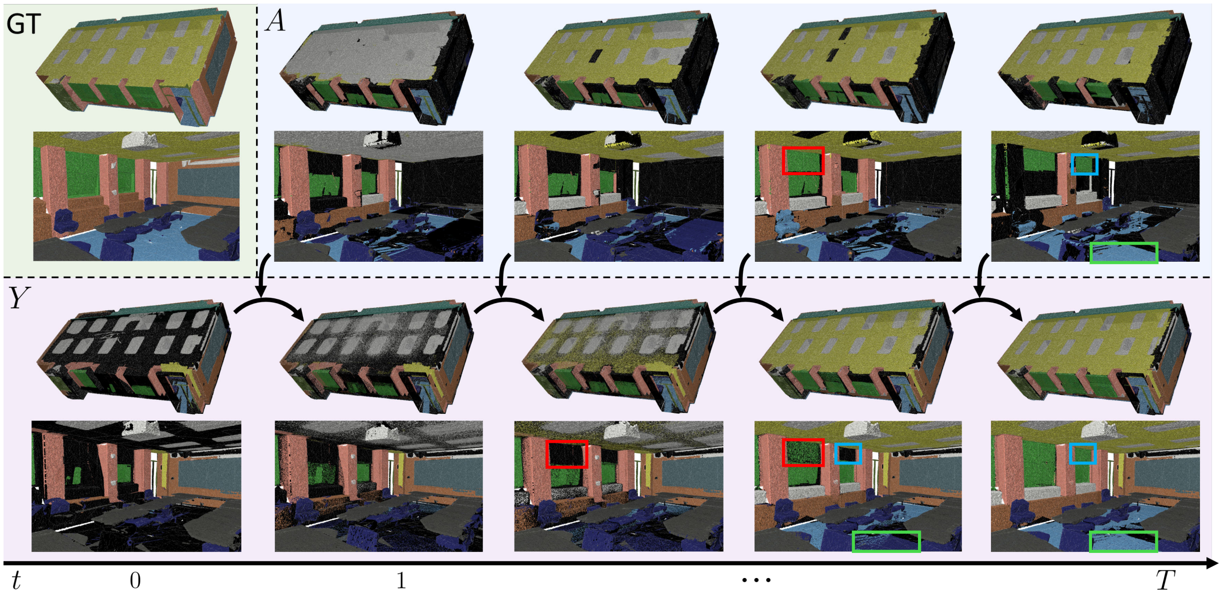

Figure 7. Progressive visualization of the behavior of the proposed REAL framework throughout training. A is the artificial label, and Y denotes the subsequent label updated at timestep t. The highlighted boxes indicate newly explored regions by artificial oracle, which were unknown (red, blue) or wrongly labeled (green) in earlier timesteps.

rics of the initial labels and those of the labels enhanced by PP2S. Furthermore, we compare the performance of the point cloud semantic segmentation model when trained with the original sparse annotations and artificial labels enhanced by PP2S. The results demonstrate that our PP2S significantly improves the quality of the label, consequently boosting the model's performance. This analysis supports the validity of PP2S as a bridge for transferring the segmentation capability of SAM into the realm of weakly supervised point cloud semantic segmentation.

Query sampling strategy. We conduct ablation studies regarding the proposed query sampling strategy introduced in Section 3.4. In this section, we compare two settings: (I) using the proposed strategy sampling from the narrowed mother set of Equ. 10 and (II) sampling the most reliable point among every point without any constraint. Figure 6 provides both qualitative and quantitative comparisons between (I) and (II), displaying the improvement achieved by the artificial labels of each setting. (I) outperforms (II) in terms of recall while maintaining precision. This distinction primarily arises from the fact that most query points of (II) are already labeled in Y^{t-1}, limiting their potential for exploring into unknown regions. On the other hand, the query of (I) is explicitly restricted to represent genuinely novel information compared with previous labels Y^{t-1}, even without compromising precision. This implies that the proposed query sampling strategy yields better labels, facilitating the model to learn segmentation effectively. Accordingly, the performance with our query sampling strategy (**62.7%**) is significantly higher than without using it (**59.8%**). Additional results are in *Supp*.

Active label enhancement. To demonstrate the behavior of the proposed REAL framework intuitively, Fig. 7 depicts progressive visualization throughout training. The labels are significantly improved as training proceeds, thanks to the artificial labels from our artificial oracle. Notably, the highlighted boxes indicate newly explored regions, which were unknown (red, blue) or wrongly labeled (green) by the labels in earlier timesteps. The REAL framework effectively integrates these regions into the current label.

4.3. Comparison with State-of-the-arts

We compare our method with SoTA methods on S3DIS, as shown in Table 2. The proposed REAL significantly enhances performance across all settings. Specifically, REAL with PTv2 [37] outperforms CPCM by 2.9% in mIoU under the 0.02% setting, which is even comparable to the SoTA results of the 1% setting. Furthermore, REAL with PTv2 surpasses the other SoTA methods under the 0.02% setting, using only 1pt for each object (0.004%). We also observe a significant performance improvement when employing the REAL in Closer [25]. These results imply that the gain of REAL is not restricted by the backbone.

Besides, Fig. 8 provides a qualitative comparison. Notably, the semantic confusion in the baseline is remarkably improved in ours. Furthermore, REAL demonstrates a meaningful enhancement in its ability to group the points of each object precisely. Consequently, REAL achieves remarkable semantic segmentation results even in complicated scenes, using only 0.004% of annotations.

Furthermore, we evaluate the performance of our method against SoTA methods on ScanNetV2 [8], as detailed in Ta-

Table 2. Comparisons of the proposed methods with SoTA methods on S3DIS test set. Every metric is in mIoU (%). **Bold** number represents the best result. † denotes the results from our reimplementation.

Settings	Methods	mIoU	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clutter
Fully	RandLA [14]	62.4	91.2	95.7	80.1	0.0	25.2	62.3	47.4	75.8	83.2	60.8	70.8	65.2	54.0
	KPConv [35]	67.1	92.8	97.3	82.4	0.0	23.9	58.0	69.0	91.0	81.5	75.3	75.4	66.7	58.9
	RFCR [10]	68.7	94.2	98.3	84.3	0.0	28.5	62.4	71.2	92.0	82.6	76.1	71.1	71.6	61.3
	Closer† [25]	66.0	93.2	98.2	81.2	0.0	24.8	49.9	69.3	90.4	80.4	72.0	74.6	68.0	55.8
	PTv2† [38]	70.9	93.8	98.5	86.1	0.0	29.5	60.5	78.0	93.2	81.8	76.6	79.2	83.0	61.7
1%	SPT [47]	61.8	91.5	96.9	80.6	0.0	18.2	58.1	47.2	75.8	85.7	65.3	68.9	65.0	50.2
	PSD [48]	63.5	92.3	97.7	80.7	0.0	27.8	56.2	62.5	78.7	84.1	63.1	70.4	58.9	53.2
	HybridCR [23]	65.3	92.5	93.9	82.6	0.0	24.2	64.4	63.2	78.3	81.7	69.0	74.4	68.2	56.5
0.03%	PSD [48]	48.2	87.9	96.0	62.1	0.0	20.6	49.3	40.9	55.1	61.9	43.9	50.7	27.3	31.1
	HybridCR [23]	51.5	85.4	91.9	65.9	0.0	18.0	51.4	34.2	63.8	78.3	52.4	59.6	29.9	39.0
0.02%	MIL [42]	51.4	86.6	93.2	75.0	0.0	29.3	45.3	46.7	60.5	62.3	56.5	47.5	33.7	32.2
	CPCM [24]	62.3	92.6	95.6	79.4	0.0	17.8	49.3	59.4	85.7	75.6	69.1	60.7	68.2	55.8
	PTv2† [38]	55.1	80.5	94.7	73.9	0.0	21.7	42.9	38.9	81.4	59.2	62.5	52.9	66.1	42.1
	+PP2S	60.0	79.0	93.1	77.5	0.1	31.5	49.7	61.4	70.0	73.7	69.9	56.9	71.0	45.6
	+REAL	**65.2**	86.9	95.8	80.2	0.0	27.0	60.3	77.8	72.3	79.9	70.8	64.7	77.4	54.1
0.004%	Closer† [25]	38.7	82.5	92.4	69.4	0.1	15.0	23.3	35.6	50.9	47.8	0.3	19.1	37.3	29.1
	+PP2S	44.9	62.9	88.8	67.9	0.0	11.7	29.2	44.7	46.4	63.2	53.5	25.7	50.7	38.6
	+REAL	**57.8**	79.0	97.7	70.8	0.1	24.4	50.4	60.1	84.5	64.6	64.8	66.8	39.9	48.7
	PTv2† [38]	39.7	76.6	87.2	65.4	0.0	9.4	30.6	22.3	61.7	50.3	41.7	14.6	22.4	34.1
	+PP2S	53.0	65.9	93.1	68.9	0.0	27.3	42.7	55.8	68.4	63.7	58.2	52.5	50.5	41.8
	+REAL	**62.7**	84.0	95.7	80.3	0.0	31.1	57.4	63.7	75.5	76.9	68.7	70.0	65.4	46.8

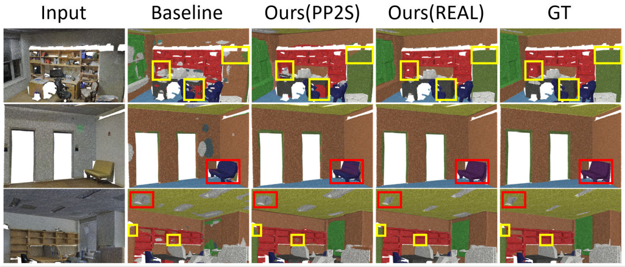

Figure 8. Qualitative comparisons between the semantic segmentation results of the baseline and ours on S3DIS under 0.004% setting. The displayed boxes indicate the regions where significant improvements are achieved by the proposed method.

Table 3. Comparisons of the proposed methods with SoTA methods on ScanNetV2 val/test set. The results are in mIoU.

Settings	Methods	Val.	Test
Fully	KPConv [35]	-	68.4
	RFCR [10]	-	70.2
1%	SPT [47]	-	51.1
	PSD [48]	-	54.7
	HybridCR [23]	56.9	56.8
20 pts	MIL [42]	57.8	54.4
	CPCM [24]	62.7	62.8
	Ours (REAL)	**66.3**	**65.5**

ble 3. Our method demonstrates superior performance on both the validation and test splits. Notably, REAL outperforms CPCM by 3.6% in the validation set and 2.7% in the test set. These results imply that our method effectively utilizes the potential of weak labels. The qualitative results for ScanNetV2 can be found in *Supp*.

5. Conclusion

Expensive and labor-intensive point-wise annotation hinders the practical application of point cloud semantic segmentation. While conventional WSPCSS methods have exhibited promising results, the fixed input sparse label potentially constrains the achievable performance. In pursuit of a novel research direction, we propose the utilization of an additional source of information, introducing the Region Exploration via Artificial Labeling (REAL) framework. This framework aims to harness the power of the image foundational model, SAM, within the context of active learning. To establish the connection between the 2D model and 3D data, we designed the Projection-based Point-to-Segment (PP2S) module. The PP2S is employed as an artificial oracle in the REAL framework, dynamically enhancing labels during training. Although using an artificial oracle eliminates the burden of manual annotation, it introduces some drawbacks due to the model's imperfections. As a remedy, we devised several strategies for query sampling and label updating. Through extensive experiments, we demonstrated the working logic of our framework in detail. Finally, the proposed REAL framework achieved new SoTA on various datasets, surpassing existing works. We believe that our approach pioneers the effective integration of the foundational image model into the realm of WSPCSS.

Acknowledgements This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636) and the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2024 (No.912768601).

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 2

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 2

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 3, 6

[4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 3

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[6] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5301, 2023. 2

[7] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 3, 6, 7

[9] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7010–7019, 2023. 2

[10] Jingyu Gong, Jiachen Xu, Xin Tan, Haichuan Song, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11673–11682, 2021. 1, 8

[11] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023. 2

[12] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021. 2

[13] Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13510–13519, 2023. 2

[14] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 1, 8

[15] Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Aleš Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *European Conference on Computer Vision*, pages 600–619. Springer, 2022. 1, 2

[16] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021. 2

[17] Zeyu Hu, Xuyang Bai, Runze Zhang, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Lidal: Inter-frame uncertainty based active learning for 3d lidar semantic segmentation. In *European Conference on Computer Vision*, pages 248–265. Springer, 2022. 2

[18] Jihun Kim, Hyeokjun Kwon, Yunseo Yang, and Kuk-Jin Yoon. Learning point cloud completion without complete point clouds: A pose-aware approach. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14157–14167. IEEE, 2023. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2

[20] Hyeokjun Kweon and Kuk-Jin Yoon. Joint learning of 2d-3d weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 35:30499–30511, 2022. 2

[21] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6994–7003, 2021. 2

[22] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023. 2

[23] Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, and Lizhuang Ma. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14930–14939, 2022. 1, 2, 6, 8

[24] Lizhao Liu, Zhuangwei Zhuang, Shangxin Huang, Xunlong Xiao, Tianhang Xiang, Cen Chen, Jingdong Wang, and Mingkui Tan. Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18413–18422, 2023. 1, 2, 6, 8

[25] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 326–342. Springer, 2020. 6, 7, 8

[26] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1726–1736, 2021. 1, 2

[27] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1190–1199, 2023. 2

[28] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[29] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1

[30] Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5575–5584, 2022. 2

[31] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023. 2

[32] Feifei Shao, Yawei Luo, Ping Liu, Jie Chen, Yi Yang, Yulei Lu, and Jun Xiao. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2575–2585, 2022. 2

[33] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[34] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 2

[35] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 1, 8

[36] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15510–15519, 2021. 2

[37] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 6, 7

[38] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 8

[39] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 2

[40] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13706–13715, 2020. 1, 2

[41] Zongyi Xu, Bo Yuan, Shanshan Zhao, Qianni Zhang, and Xinbo Gao. Hierarchical point-based active learning for semi-supervised point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18098–18108, 2023. 2

[42] Cheng-Kun Yang, Ji-Jia Wu, Kai-Syun Chen, Yung-Yu Chuang, and Yen-Yu Lin. An mil-derived transformer for weakly supervised point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11830–11839, 2022. 1, 2, 6, 8

[43] Cheng-Kun Yang, Min-Hung Chen, Yung-Yu Chuang, and Yen-Yu Lin. 2d-3d interlaced transformer for point cloud segmentation with scene-level supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 977–987, 2023. 2

[44] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2

[45] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel,*

October 23–27, 2022, Proceedings, Part XXIX, pages 326–344. Springer Nature Switzerland Cham, 2022. 2

[46] Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. Sam3d: Zero-shot 3d object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023. 2

[47] Yachao Zhang, Zonghao Li, Yuan Xie, Yanyun Qu, Cuihua Li, and Tao Mei. Weakly supervised semantic segmentation for large-scale point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3421–3429, 2021. 1, 2, 8

[48] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15520–15528, 2021. 1, 2, 6, 8