Less Is More: Label Recommendation for Weakly Supervised Point Cloud Semantic Segmentation

Zhiyi Pan[1,3], **Nan Zhang**[1], **Wei Gao**[1*], **Shan Liu**[2], **Ge Li**[1]
[1]SECE, Shenzhen Graduate School, Peking University
[2]Media Laboratory, Tencent
[3]Peng Cheng Laboratory
{panzhiyi, zhangnan}@stu.pku.edu.cn, gaowei262@pku.edu.cn, shanl@tencent.com, geli@ece.pku.edu.cn

Abstract

Weak supervision has proven to be an effective strategy for reducing the burden of annotating semantic segmentation tasks in 3D space. However, unconstrained or heuristic weakly supervised annotation forms may lead to suboptimal label efficiency. To address this issue, we propose a novel label recommendation framework for weakly supervised point cloud semantic segmentation. Distinct from pre-training and active learning, the label recommendation framework consists of three stages: inductive bias learning, recommendations for points to be labeled, and weakly supervised point cloud semantic segmentation learning. In practice, we first introduce the point cloud upsampling task to induct inductive bias from structural information. During the recommendation stage, we present a cross-scene clustering strategy to generate centers of clustering as recommended points. Then we introduce a recommended point positions attention module LabelAttention to model the long-range dependency under sparse annotations. Additionally, we employ position encoding to enhance the spatial awareness of the segmentation network. Throughout the framework, the useful information obtained from inductive bias learning is propagated to subsequent semantic segmentation networks in the form of label positions. Experimental results demonstrate that our framework outperforms weakly supervised point cloud semantic segmentation methods and other methods for labeling efficiency on S3DIS and ScanNetV2, even at an extremely low label rate.

Introduction

Point cloud semantic segmentation is one of the fundamental tasks in 3D scene understanding and holds significant importance in applications such as autonomous driving (Li et al. 2020), AR/VR (Blanc et al. 2020), and robotics (Shan et al. 2020). With the advent of weakly supervised learning, the requirement for fine-grained annotation of massive point cloud datasets for semantic segmentation has been alleviated, garnering considerable attention from researchers.

However, existing works mostly focus on designing network architectures and loss functions to leverage limited sparse annotations, but impose few constraints on the annotation forms. Some works (Zhang et al. 2021a; Hu et al. 2022a) require sparse annotations to be as random as possible, while others (Liu, Qi, and Fu 2021; Wu et al. 2022b) additionally mandate a one-thing-one-click format. Such weakly supervised annotation forms are suboptimal. The constraints of randomness and one-thing-one-click do not take into account the varying importance of points during network learning, leading to supervisory information deficiency. Even worse, annotators often exhibit preferences in practice, leading to an inhomogeneous labeling distribution (Pan et al. 2023). Consequently, the sparse annotations may not sufficiently support effective network learning.

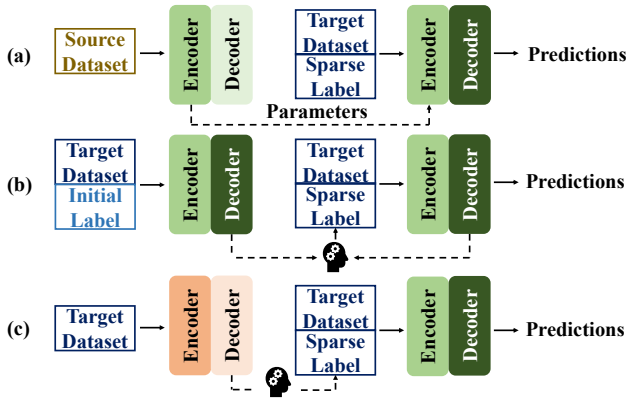


Figure 1: Process comparison of (a) pre-training framework, (b) active learning framework, and (c) label recommendation framework.

Therefore, we propose a novel label recommendation framework to enhance the annotation efficiency for weakly supervised point cloud semantic segmentation. As illustrated in Figure 1 (c), the framework consists of three stages: inductive bias learning, recommendations for points to be labeled, and point cloud semantic segmentation learning based on the recommended annotations. While it shares similarities with pre-training and active learning frameworks, it has essential differences as shown in Figure 1. The inductive bias learning stage extracts facilitative inductive bias for label recommendation in an unsupervised manner. In the annotation recommendation generation stage, annotations are generated based on the point embedding, and then annotators label these recommendation points to produce sparse

*Wei Gao is the corresponding author.

annotations. Finally, in the point cloud semantic segmentation learning stage, the segmentation network is trained based on the recommended sparse annotations.

Regarding the proposed annotation recommendation framework, this work addresses and attempts to solve three key challenges: (1) How to select appropriate pretext tasks for inductive bias learning in the label recommendation framework? (2) How to efficiently choose points to be labeled based on point embedding? (3) How to fully leverage the recommended annotations for weakly supervised point cloud semantic segmentation learning?

For the first challenge, We resort to unsupervised tasks for inductive bias learning. Surprisingly, we discovered that inductive bias learning for label recommendation requires explicit restrictions at the point level rather than inter-point contrastive supervision. Creating the generative unsupervised tasks necessitates altering the original data distribution, leading to domain adaptation issues. To address this, we adopted $2\times$ point cloud upsampling based on the principle of minimizing divergence as the pretext task. In response to the second challenge, we employed clustering to use the cluster centers of features as recommended sparse annotations. Furthermore, we proposed a cross-scene clustering strategy that can perceive feature information from multiple scenes and reduce the complexity introduced by cross-scene scenarios. Regarding the third issue, we proposed an attention module based on the positions of sparse annotations named LabelAttention, allowing gradients to be directly propagated from annotated points to unannotated points during backpropagation, thereby alleviating long-range dependency issues caused by sparse annotations. Additionally, we introduce position encoding to enhance the spatial perception of models.

Our experiments demonstrated that, at a 0.01% label rate, the proposed framework achieved over 90% performance of fully supervised learning on both S3DIS (Armeni et al. 2016) and ScanNetV2 (Dai et al. 2017), surpassing other competitors. Extensive ablation studies further confirmed the effectiveness of the proposed module in our framework.

Related Work

Weakly Supervised Point Cloud Semantic Segmentation. Although weakly supervised point cloud semantic segmentation is still in its infancy, a large number of well-established works have emerged. Xu and Lee is the first to introduce the weakly supervised point cloud semantic segmentation task. They propose a dual-branch framework trained with consistency loss, inexact loss, and color space prior constraints. Subsequent works either improve the dual-branch input paradigm (Unal, Dai, and Van Gool 2022; Wu et al. 2022b), introduce information interaction between the two branches (Zhang et al. 2021b; Cheng et al. 2021), or impose new constraints between the results from branches (Li et al. 2022a). Some methods (Liu, Qi, and Fu 2021, 2023; Zhang et al. 2021a) propose single-branch frameworks with label propagation. These methods usually require designing a relation metric module to predict the similarity relationship matrix, which helps propagate labels from annotated points to unannotated points step by step. Additionally, SQN (Hu et al. 2022a) and GaIA (Lee, Yang, and Han 2023) provide solutions from the perspectives of feature queries and information gain, respectively. Recently, AAD (Pan et al. 2023) analyzes weakly supervised point cloud semantic segmentation under non-uniform distribution annotations.

Indeed, these works are designed for frameworks based on given sparse annotations but do not explore how to label sparse annotations that harness the segmentation potential of subsequent networks.

Pre-training for Point Cloud Understanding. Due to the ability of unsupervised pretext tasks to learn information from any point cloud scene without annotations, pre-trained networks can significantly enhance point cloud understanding after fine-tuning the target dataset. Pre-training can be divided into contrastive learning-based pre-training and generative-based pre-training. PointContrast (Xie et al. 2020) is an impactful work, which proposes a contrastive learning-based pre-training framework for point cloud understanding by view transformations to construct positive and negative pairs. DepthContrast (Zhang et al. 2021c) and crosspoint (Afham et al. 2022) utilize geometric transformation and cross-modal transformations to create positive and negative pairs, to adapt to single-view point clouds. SegContrast (Nunes et al. 2022) implements contrastive learning using pre-extracted segments as the basic unit.

Differently, generative pre-training methods construct a pretext task by applying operations such as masking (Yu et al. 2022; Wang et al. 2021) and mixing (Sun et al. 2022) to the original point cloud scene. Subsequently, the corresponding decoder and loss function are designed for a complete pre-training.

Label-efficient Learning for Point Cloud. Currently, learning for label efficiency on point clouds typically relies on pre-training or active learning frameworks. Pre-training-based label-efficient learning investigates how to obtain beneficial point embedding from limited annotations. ACD (Gadelha et al. 2020) achieves label-efficient learning by combining Approximate Convex Decomposition with contrastive loss as self-supervised signals. Hou et al. proposes the Contrast Scene Contexts to integrate spatial contexts for pre-training, and efficient annotations are selected based on the pre-trained features. In contrast, active learning focuses on how to design the metric function to select the regions to be labeled online. Shi et al. constructs manual labeling metrics at point-level, superpoint-level, and shape-level. ReDAL (Wu et al. 2021) calculates annotation metrics based on feature entropy, color, and geometric structure with region density selection strategy. Shao et al. measures spatial and structural diversities via graph reasoning network. LiDAL (Hu et al. 2022b) considers both inter-frame divergency and entropy to solve region selection for LiDAR point cloud scenes. Rong, Cui, and Shen chooses 2D rendered images rather than points or superpoints in 3D space .

The closely related work to ours is LESS (Liu et al. 2022), which introduces a heuristic pre-segmentation for LiDAR outdoor datasets. It automatically expands sparse annotations into pseudo-labels at the segment level. Their segmentation network incorporates prototype learning and multi-

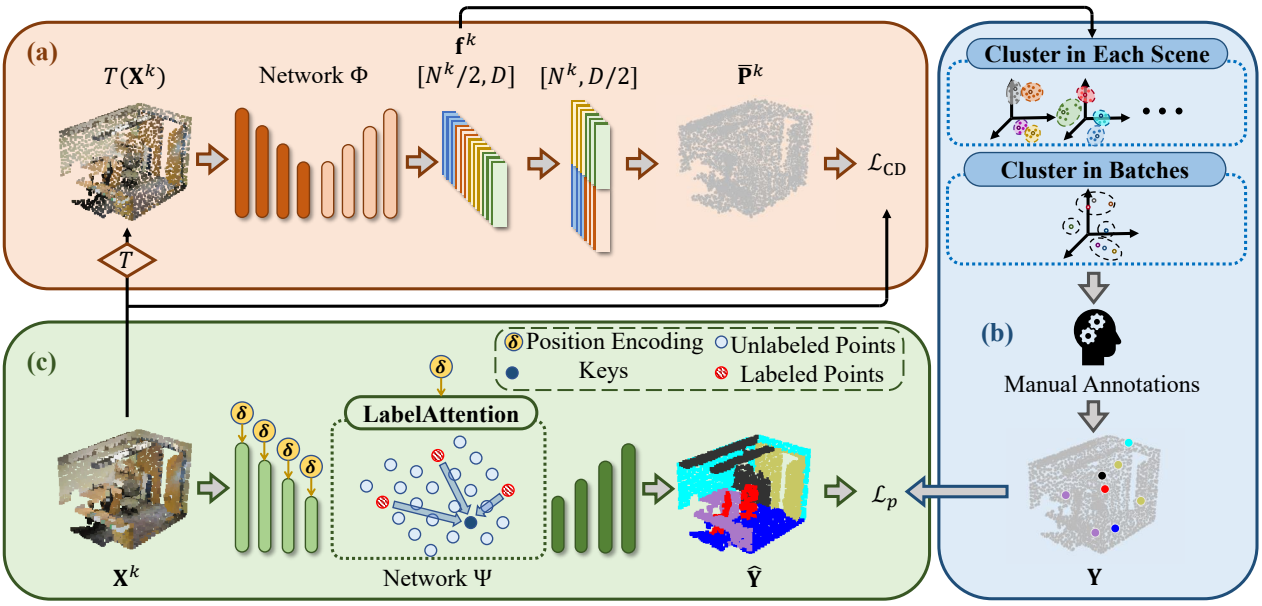


Figure 2: The proposed label recommendation framework contains three sequential stages: (a) inductive bias learning, (b) recommendation, and (c) weakly supervised point cloud semantic segmentation learning. In (a), the inductive bias is cultivated via $2\times$ point cloud upsampling. The subsequent stage (b) leverages dual-tier clustering for cross-scene clustering and annotators to derive the recommendation $\mathbf{Y}$. In (c), the network integrates LabelAttention and position encoding to harness $\mathbf{Y}$ for comprehensive semantic segmentation learning, culminating in a refined understanding of the point cloud data.

scan distillation to enhance the accuracy of single-scan segmentation. Conversely, our framework adopts an unsupervised task for learning-based label recommendation while refraining from pseudo-label extension to mitigate noises.

Method

Overview

In this paper, we propose a label recommendation framework to explore label efficiency for weakly supervised point cloud semantic segmentation. The point cloud dataset on this framework is defined as $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^K\}$, where K denotes the number of point cloud scenes and $\mathbf{X}^k = [\mathbf{P}^k, \mathbf{F}^k] \in \mathbb{R}^{N_k \times (3+\mathcal{F})}$. $\mathbf{P}^k \in \mathbb{R}^{N_k \times 3}$ and $\mathbf{F}^k \in \mathbb{R}^{N_k \times \mathcal{F}}$ represent the 3D spatial coordinate information and other $\mathcal{F}$-dimensional information (e.g. color, normal vector, etc.) respectively. N_k denotes the number of points in $\mathbf{X}^k$.

As shown in Figure 2, the whole framework contains three sequential stages, including inductive bias learning, recommendation, and point cloud semantic segmentation learning. In the inductive bias learning stage, an unsupervised task is chosen as a pretext task to obtain high-level feature representations $\mathbf{f}^k$ corresponding to the input $\mathbf{X}^k$, which can be written as $\mathbf{f}^k = \Phi[T(\mathbf{X}^k)]$, where Φ denotes the network trained on the unsupervised task and T can be a mask operation in the spatial or feature domain for generative unsupervised tasks. Accordingly, the loss function depends on the choice of pretext task.

In the recommendation stage, we perform clustering based on $\mathbf{f}^k$ obtained from the previous stage, to take the clustering centroids as recommendation points to be labeled by $\mathcal{C}^k = \text{cluster}(\mathbf{f}^k, \alpha N_k)$, where α and αN_k denote the proportion of sparse annotation and the number of clusters, respectively. In particular, we propose a dual-tier cross-scene clustering strategy that adapts to the distribution of objects in multiple scenes. The human annotators sparsely annotate the point cloud dataset according to the recommended points, and the sparse annotation $Y^k = \{\mathbf{Y}_i^k \mid i \in \mathcal{C}^k\}$.

The point cloud semantic segmentation learning stage performs weakly supervision based on $\mathbf{X}^k$ and $\mathbf{Y}^k$, with a partial cross-entropy loss:

$$\mathcal{L}_p = \frac{1}{|\mathcal{C}^k|} \sum_{i \in \mathcal{C}^k} \left(\mathbf{Y}_i^k \log(\hat{\mathbf{Y}}_i^k) + \hat{\mathbf{Y}}_i^k \log(\mathbf{Y}_i^k) \right), \quad (1)$$

where $\hat{\mathbf{Y}}^k = \Psi(\mathbf{X}^k)$ denotes the prediction result of semantic segmentation network Ψ with $\mathbf{X}^k$ as input.

Differences from pretraining. (1) Different Objectives: The purpose of pre-training is to achieve general parameters for the encoder, whereas label recommendation aims to recommend annotation regions for point cloud understanding. (2) Different Datasets: Pre-training typically utilizes large-scale point cloud datasets, entailing numerous training epochs. In contrast, label recommendation is performed on a given segmentation dataset with fewer training iterations. (3) Different Network Requirements: Pre-training necessitates maintaining the same encoder structure before and after the pre-training. Conversely, label recommendation transmits information through the points to be labeled, allowing different structures.

Differences from active learning. (1) Different Annotation Recommendation Stage: In active learning, the recommendation stage occurs during task training. In contrast, the recommendation stage predates task training in label recommendation learning, preventing disruption to the process. (2) Different Frequency of Annotation Recommendations: Active learning requires multiple and continuous provisions of new annotations during the training process, while label recommendation requires only one overall recommendation step. (3) Different Task: Active learning operates within the context of the given task training, whereas the recommendation stage in label recommendation is before network training, involving the selection of an unsupervised pretext task.

Inductive Bias Learning Stage

We investigate several mainstream unsupervised learning tasks including generative and contrastive learning unsupervised methods. Generative unsupervised methods first utilize a sampling function T to mask information from the point cloud scene as input. Subsequently, the original point cloud scene serves as the supervisory signal to guide the generation of concealed information. Contrastive learning as another solution transforms the point cloud scene using the bijection function T and then constructs positive and negative pairs to provide constraints. Based on the theoretical analysis and experimental results, we choose the $2\times$ point cloud upsampling as the pretext task, then T denotes the uniform $2\times$ downsampling function, and the Chamfer Distance function used for supervision is defined as:

$$\mathcal{L}_{\text{CD}} = \frac{1}{N_k}\sum_i \min_j ||\mathbf{P}_i^k - \bar{\mathbf{P}}_j^k||_2^2 + \frac{1}{N_k}\sum_i \min_j ||\bar{\mathbf{P}}_i^k - \mathbf{P}_j^k||_2^2, \tag{2}$$

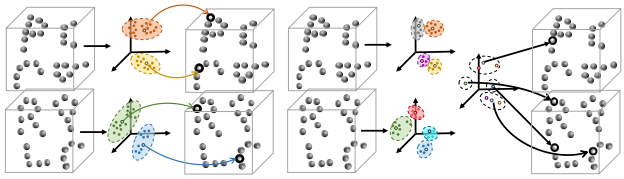
Inspired by the single point cloud object upsampling methods (Yu et al. 2018; Li et al. 2022b), we define the predictions $\bar{\mathbf{P}}^k$ on point cloud scene upsampling as:

$$\bar{\mathbf{P}}^k = \text{MLPs}\big(\text{reshape}(\mathbf{f}^k, [N_k/2, D] \to [N_k, D/2])\big), \tag{3}$$

where D denotes the channel dimension of $\mathbf{f}^k$.

Discussion of pretext tasks. Table 3 showcases our evaluation of diverse methods, including two contrastive learning methods (Hardest-Contrastive (Choy, Park, and Koltun 2019) and PointInfoNCE (Xie et al. 2020)) and two generative methods (point cloud upsampling and point cloud colorize). Despite the commendable performance of contrastive learning within the pre-training domain, its efficacy within the label recommendation framework appeared relatively diminished compared to generative methods. We attribute this observation to the label recommendation framework focus on point-level feature acquisition, demanding distinct supervision for individual points. In contrast, contrastive learning draws upon features from multiple points and lacks explicit point-level guidance.

In generative unsupervised methods (Yu et al. 2022; Cheng et al. 2023), the sampled point cloud as input introduces a distribution discrepancy between the training data $T(\mathbf{P}^k)$ and the target data $\mathbf{P}^k$. The mitigation of distribution disparities constitutes a foundational pursuit in the realm of domain adaptation. Drawing inspiration from divergence minimization-based domain adaptation (Zhang, Li, and Ogunbona 2017), we elect the $2\times$ point cloud upsampling as our pretext task. This selection stems from its relatively subdued influence on distribution when juxtaposed with alternative generative tasks. While the variational autoencoder (Doersch 2016) theoretically exerts the least distribution impact, it grapples with training stability issues due to posterior collapse (He et al. 2019). Therefore, we ultimately settle upon the $2\times$ point cloud upsampling, guided by its favorable equilibrium between distributional influence and training stability.



(a) Single-scene Clustering **(b) Cross-scene Clustering**

Figure 3: Schematic comparison of single-scene clustering and cross-scene clustering.

Recommendation Stage

An effective recommendation should furnish more holistic guidance for subsequent semantic segmentation learning, encompassing a wider spectrum of the scene's intricacies. In light of this perspective, we regard clustering centers as the foremost candidates for representative point recommendations. Nevertheless, employing clustering solely within a singular scene may yield insufficient annotations for intricate scenes, juxtaposed with redundant annotations for simpler scenes, as depicted in Figure 3 (a). By adopting a cross-scene clustering strategy, we facilitate the automated allocation of annotation quantities, taking into account the intricacies inherent to different scenes. Moreover, harnessing features across diverse scenes augments the precision of annotation region recommendations and the breadth of supervisory information coverage.

However, cross-scene clustering introduces heightened algorithmic complexity. To address this, we propose a dual-tier clustering algorithm for the recommendation stage. Firstly, we execute clustering within a single scene with the label rate α and the expansion coefficient β. Following this, the second clustering is executed across the aggregation of clustering centers extracted from numerous point cloud scenes. This process can be represented as follows:

$$\{\mathcal{C}^k\}_{k\in B} = \text{cluster}\Big(\big\{\text{cluster}(\mathbf{f}^k, \alpha\beta N_k)\big\}, \alpha\sum_{k\in B} N_k\Big), \tag{4}$$

where B denotes the index set of point cloud scenes in one batch and β denotes the expansion coefficient of the first clustering. In practice, we choose Kmeans as the clustering algorithm. Finally, the human annotators provide sparse annotations $\mathbf{Y}$ based on the location of the clustering centers.

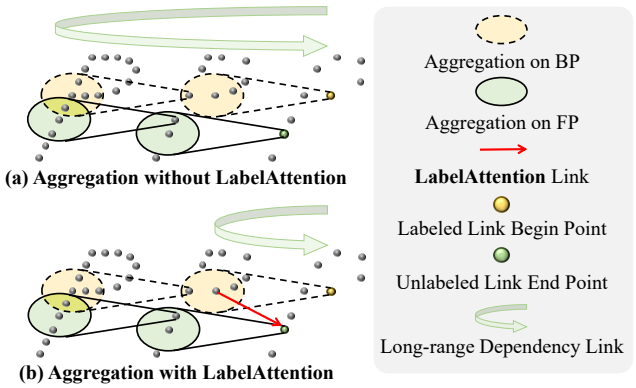


Figure 4: LabelAttention effectively shortens the information interaction link between labeled and unlabeled points. BP and FP denote backward propagation and forward propagation, respectively.

Algorithm complexity analysis. For the sake of comparison, we assume that each scene has the same number of points, i.e. $\sum_{k\in B} N_k = |B|N_k$. Since the maximum number of iterations of Kmeans and the feature dimension are constants, the conventional cross-scene clustering has a time complexity of $O(\alpha|B|^2N_k^2)$, while our proposed dual-tier cross-scene clustering has a time complexity of $O(\alpha\beta|B|N_k^2 + \alpha^2\beta|B|N_k^2)$. Since the label rate $\alpha \ll 1$, the time complexity can be simplified to $O(\alpha\beta|B|N_k^2)$. When the expansion coefficient $\beta < |B|$, the dual-tier cross-scene clustering can effectively reduce the time complexity of the algorithm. Correspondingly, the space complexity can be reduced from $O(|B|N_k)$ to $O(N_k)$.

Weakly Supervised Learning Stage

Effectively harnessing the limited supervision inherent in sparse annotations constitutes a pivotal factor in enhancing segmentation performance. The paucity of annotated points intensifies the complexity of modeling long-range dependencies. In this context, unlabeled points solely glean supervision from labeled points through multiple feature extractions, devoid of direct interaction with unlabeled points. The applicability of local and global attention mechanisms, integral for capturing long-range dependencies within fully supervised networks, is constrained within the context of weak supervision. Local attention predominantly concentrates on neighborhood regions, commonly comprising a substantial pool of unlabeled points, while global attention diffuses supervision efficacy attributed to annotated points, given the significant prevalence of unlabeled counterparts.

To address this issue, we propose the LabelAttention module to consciously impose attention at labeled locations. The features $\mathbf{f}_i'$ after LabelAttention can be defined as: (For a clear presentation, we simplify $\mathbf{f}_i$ to $\mathbf{f}_i^k$)

$$\mathbf{f}_i' = \mathbf{f}_i + \sum_{j\in\mathcal{C}} \rho(\varphi(\mathbf{f}_i) - \psi(\mathbf{f}_j)) \odot \phi(\mathbf{f}_j + \delta), \qquad (5)$$

where φ, ψ, ϕ are point-level linear transformations, ρ consists of a softmax and a mapping function, δ is a positional encoding, and $\odot$ donates the Hadamard product. We use the attention module to aggregate the features from $\{\mathbf{f}_j\}_{j\in\mathcal{C}}$, which allows the sparsely supervised gradient at the labeled points to propagate quickly to all positions. According to the analysis of self-attention in AFGCN (Zhang et al. 2023), our utilization of LabelAttention is only concentrated within the last encoder block. In the inference phase, we similarly use the positions of the labeled points in the test set to construct LabelAttention.

In set abstract and downsampling layers, we further utilize position encoding to enhance the perception of the network to the geometric space, which is formulated as:

$$\delta = \mathrm{MLPs}(p_j - p_i). \qquad (6)$$

Relative position-based encoding facilitates the spatial positional alignment between unlabeled and labeled points within the feature aggregation process, and local topological capture during feature downsampling, resulting in more robust representations.

Experiments and Analysis

Experiment Settings

Dataset. S3DIS (Armeni et al. 2016) covers six large-scale indoor areas, totaling about 271 rooms, with varying room layouts, furniture arrangements, and object placements. Area 5 is used for validation and the remaining areas are allocated for network training. ScanNetV2 (Dai et al. 2017) encompasses 1,513 scanned scenes originating from 707 diverse indoor environments. Our study adheres to the official ScanNetV2 partition, employing 1,201 scenes for training and allocating 312 scenes for validation.

Method	Supervision	mIoU (%)
PointNet++ (Qi et al.)	100%	33.9
PointCNN (Li et al.)	100%	45.8
PointNeXt (Qian et al.)	100%	71.2
PTV2 (Wu et al.)	100%	75.2
Zhang et al. 2021a	1%	51.1
PSD (Zhang et al.)	1%	54.7
HybirdCR (Li et al.)	1%	56.8
GaIA (Lee, Yang, and Han)	1%	65.2
AADNet (Pan et al.)	1%	66.8
Ours	1%	**67.4**
SQN (Hu et al.)	0.01%	35.9
PointMatch† (Wu et al.)	0.01%	57.1
OTOC++† (Liu, Qi, and Fu)	0.01%	60.6
Hou et al. 2021	20 pts/scene	55.5
MILTrans (Yang et al.)	20 pts/scene	54.4
OTOC† (Liu, Qi, and Fu)	20 pts/scene	59.4
PointMatch† (Wu et al.)	20 pts/scene	62.4
AADNet (Pan et al.)	20 pts/scene	62.5
Ours†	20 pts/scene	**63.7**

Table 1: Quantitative results of semantic segmentation on the test set of ScanNetV2. † denotes the super-voxel setting.

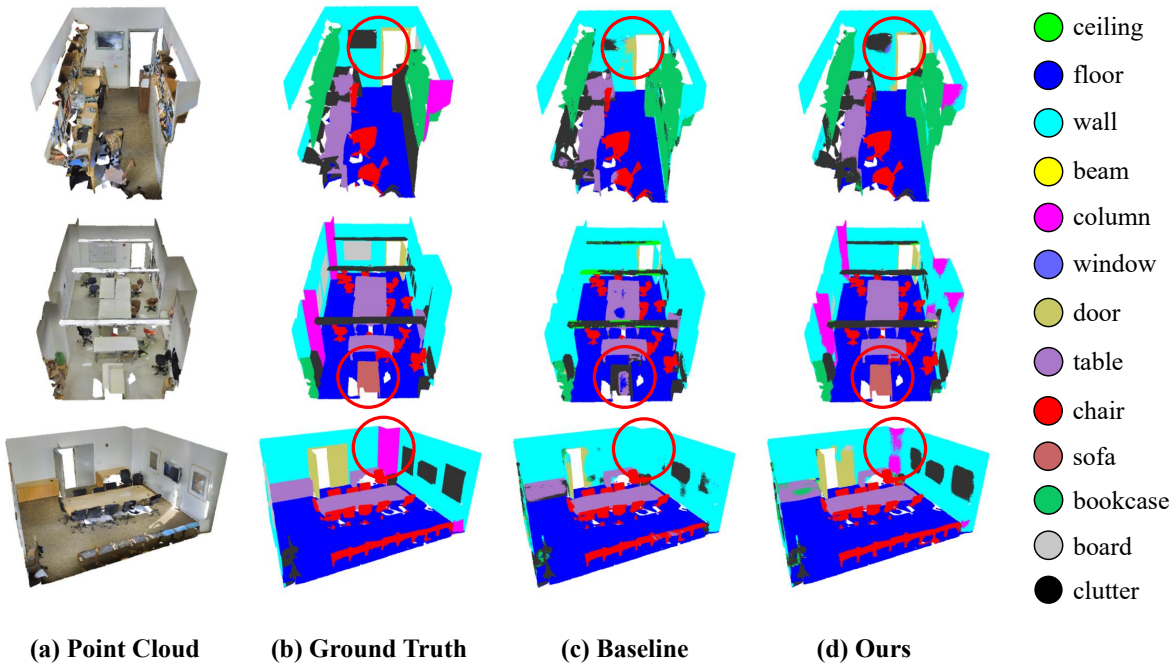


Figure 5: Visual comparison on S3DIS Area 5 at 0.01% label rate. Red circle for highlighting.

Method	Supervision	mIoU (%)
PointNet++ (Qi et al.)	100%	53.5
PointCNN (Li et al.)	100%	57.3
PointNeXt (Qian et al.)	100%	70.5
PTV2 (Wu et al.)	100%	71.6
Π Model (Laine and Aila)	0.2%	44.3
MT (Tarvainen and Valpola)	0.2%	44.4
Xu and Lee 2020	0.2%	44.5
SQN (Hu et al.)	0.1%	61.4
PointMatch† (Wu et al.)	0.1%	63.4
AADNet (Pan et al.)	0.1%	67.2
Ours	0.1%	**68.9**
Zhang et al. 2021a	0.03%	45.8
PSD (Zhang et al.)	0.03%	48.2
HybirdCR (Li et al.)	0.03%	51.5
OTOC† (Liu, Qi, and Fu)	0.02%⋆	50.1
MILTrans (Yang et al.)	0.02%⋆	51.4
GaIA (Lee, Yang, and Han)	0.02%⋆	53.7
DAT (Wu et al.)	0.02%⋆	56.5
OTOC++† (Liu, Qi, and Fu)	0.02%⋆	56.6
SQN (Hu et al.)	0.01%	45.3
PointMatch† (Wu et al.)	0.01%	59.9
AADNet (Pan et al.)	0.01%	60.8
Ours	0.01%	**62.9**

Table 2: Quantitative results of semantic segmentation on S3DIS Area 5. 0.02%⋆ denotes the one-thing-one-click annotation. † denotes the super-voxel setting.

Implementation. In inductive bias learning stage, we use pointnet++ (Qi et al. 2017) as the backbone with an initial learning rate of 10^{-3}, adamw optimizer (weight decay is 10^{-4}) for 30 epochs. For S3DIS, we first utilize the entire voxel-downsampled strategy to sample 24,000 points from the original point cloud as the ground truth and downsample ground truth to 12,000 points as inputs to the inductive bias learning network. For ScannetV2, the input resolution of the upsampling network is 32,000. Inductive bias learning is trained with one NVIDIA V100 GPU on S3DIS and four NVIDIA V100 GPUs on ScanNetV2, respectively. For the recommendation stage, we use Kmeans as the clustering algorithm, where cross-scene clustering is computed separately for the training and test sets. The maximum number of iterations is 100. The expansion coefficient $\beta = \frac{1}{2}B$ with $B = 8$. In the weakly-supervised point cloud semantic segmentation learning stage, we use PointNeXt-L (Qian et al. 2022) as the backbone and follow the official settings of PointNeXt. Following AAD (Pan et al. 2023), the label-aware downsampling (LaDS) is imposed to fully utilize the sparse annotation. mIoU is used as the evaluation metric.

Comparsion Results

Figures 1 and Figures 2 illustrate that our framework achieves SOTA on both ScanNetV2 and S3DIS compared to label-efficient learning and weakly-supervised point cloud semantic segmentation methods. Our label recommendation framework achieves more than 90% of fully-supervised segmentation performance at the 0.01% label rate. Compared to the baseline using randomly sampled sparse annotations, our method improves 4.5% and 1.8% mIoU at 0.01% and 0.1% label rates on S3DIS, respectively.

Method		mIoU (%)
Hardest-Contrastive Loss	Contrastive	43.7
PointInfoNCE Loss		42.6
Point Colorize	Generative	45.1
2× Point Upsampling		**49.6**
4× Point Upsampling		42.9
8× Point Upsampling		43.2
2× Point Completion		48.2

Table 3: Ablation study for pretext task in the inductive bias learning stage.

Figure 5 demonstrates that the label commendation framework has a significant accuracy improvement in detailed categories (e.g., doorframes with slender shapes) and confusing categories (e.g., columns and walls, sofas and chairs).

Ablation Study

We perform ablation experiments without LaDS on S3DIS Area 5. The baseline denotes label recommendation based on features obtained in the inductive bias learning stage with single-scene Kmeans, followed by PointNeXt-L for point cloud semantic segmentation learning.

The choice of pretext task. Table 3 provides a comprehensive comparison of the two unsupervised learning paradigms, revealing the notable superiority of the generation task over the contrastive learning task. Furthermore, it elucidates that the point-generation task significantly outperforms the color-generation task. Our exploration extends to investigating the impact of generation distribution uniformity (upsampling versus completion) and generation complexity (upsampling rates) on the point cloud label recommendation framework. Our findings underscore that the 2× upsampling, which minimizes perturbations to the original distribution, emerges as the most efficacious pretext task.

Scene Batch Size	1	4	8	16	32
mIoU (%)	49.6	50.3	**51.2**	48.6	47.8

Table 4: Ablation study for scene batch size in the recommendation stage.

The cross-scene clustering. Table 4 provides an analysis, evaluating the influence of cross-scene clustering across varying scene batch sizes. Our observations illustrate that the performance shows a trend of increasing and then decreasing with increased scene size. A small scene batch size leads to insufficient cross-scene information, while a large scene batch size leads to complex clustering features. In addition, we visualize the recommendation results after single-scene and cross-scene clustering in Figure 6, and cross-scene clustering focuses more on long-tail categories with variable

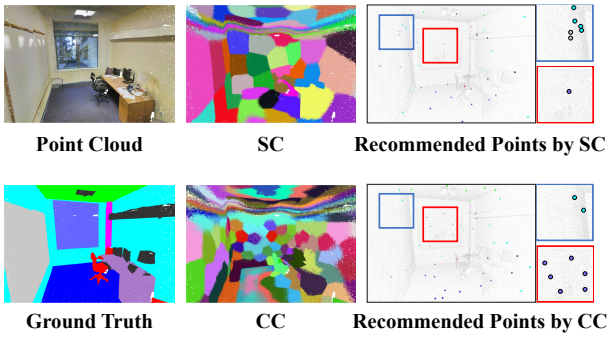


Figure 6: Cluster results and recommended points by single-scene clustering (SC) and cross-scene clustering (CC).

shapes (e.g., windows with variable perspectives) and less on sample head categories (e.g., walls).

Method	mIoU (%)	P. (M)	GFLOPs
PointNeXt-L	51.2	**7.13**	24.4
+δ	55.0	7.54	**33.0**
+δ +RandomAttention	54.4	7.42	30.0
+δ +LocalAttention	55.7	7.42	31.6
+δ +GlobalAttention	55.8	7.42	28.7
+δ +LabelAttention	**56.1**	7.42	30.0

Table 5: Ablation Study for LabelAttention and position encoding in the weakly supervised learning stage. P. is an abbreviation for network parameters.

The LabelAttention and position encoding. In Table 5, we perform an ablation study on position encoding and various attention modules. It can be found that position encoding is significantly helpful within the realm of weakly supervised learning. In addition, LabelAttention mitigates the long-range dependence issue by explicitly aggregating feature information with labeled points, resulting in optimal performance. While GlobalAttention similarly addresses the long-range dependence, it simultaneously increases the computational demands as well as introduces susceptibility to non-robust features beyond labeled points. However, our method can effectively improve performance without significantly increasing the computational cost.

Conclusion

In this paper, we design a novel labeled recommendation framework to address weakly-supervised point cloud semantic segmentation and explore each stage under this framework. Empirical experiments demonstrate that our framework outperforms conventional weakly-supervised semantic segmentation methods and other generalized label-efficient learning frameworks. We expect the label recommendation framework to be a compelling alternative in the field of weakly-supervised point cloud semantic segmentation and provide a new perspective for the point cloud community.

Acknowledgements

This work was supported in part by Natural Science Foundation of China under Grants 62172021, in part by Shenzhen Fundamental Research Program under Grant GXWD20201231165807007-20200806163656003, and in part by National Key R&D Program of China (No.2020AAA0103501).

References

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. 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*, 9902–9912.

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Blanc, T.; El Beheiry, M.; Caporal, C.; Masson, J.-B.; and Hajj, B. 2020. Genuage: visualize and analyze multidimensional single-molecule point cloud data in virtual reality. *Nature Methods*, 17(11): 1100–1102.

Cheng, M.; Hui, L.; Xie, J.; and Yang, J. 2021. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1140–1147.

Cheng, X.; Zhang, N.; Yu, J.; Wang, Y.; Li, G.; and Zhang, J. 2023. Null-Space Diffusion Sampling for Zero-Shot Point Cloud Completion. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Choy, C.; Park, J.; and Koltun, V. 2019. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8958–8966.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Doersch, C. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Gadelha, M.; RoyChowdhury, A.; Sharma, G.; Kalogerakis, E.; Cao, L.; Learned-Miller, E.; Wang, R.; and Maji, S. 2020. Label-efficient learning on point clouds using approximate convex decompositions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 473–491. Springer.

He, J.; Spokoyny, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Hou, J.; Graham, B.; Nießner, M.; and Xie, S. 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15587–15597.

Hu, Q.; Yang, B.; Fang, G.; Guo, Y.; Leonardis, A.; Trigoni, N.; and Markham, A. 2022a. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, 600–619. Springer.

Hu, Z.; Bai, X.; Zhang, R.; Wang, X.; Sun, G.; Fu, H.; and Tai, C.-L. 2022b. Lidal: Inter-frame uncertainty based active learning for 3d lidar semantic segmentation. In *European Conference on Computer Vision*, 248–265. Springer.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Lee, M. S.; Yang, S. W.; and Han, S. W. 2023. GaIA: Graphical Information Gain based Attention Network for Weakly Supervised Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 582–591.

Li, M.; Xie, Y.; Shen, Y.; Ke, B.; Qiao, R.; Ren, B.; Lin, S.; and Ma, L. 2022a. 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*, 14930–14939.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31.

Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M. A.; Cao, D.; and Li, J. 2020. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8): 3412–3432.

Li, Z.; Li, G.; Li, T. H.; Liu, S.; and Gao, W. 2022b. Semantic point cloud upsampling. *IEEE Transactions on Multimedia*.

Liu, M.; Zhou, Y.; Qi, C. R.; Gong, B.; Su, H.; and Anguelov, D. 2022. Less: Label-efficient semantic segmentation for lidar point clouds. In *European Conference on Computer Vision*, 70–89. Springer.

Liu, Z.; Qi, X.; and Fu, C.-W. 2021. 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*, 1726–1736.

Liu, Z.; Qi, X.; and Fu, C.-W. 2023. One Thing One Click++: Self-Training for Weakly Supervised 3D Scene Understanding. *arXiv preprint arXiv:2303.14727*.

Nunes, L.; Marcuzzi, R.; Chen, X.; Behley, J.; and Stachniss, C. 2022. SegContrast: 3D point cloud feature representation learning through self-supervised segment discrimination. *IEEE Robotics and Automation Letters*, 7(2): 2116–2123.

Pan, Z.; Zhang, N.; Gao, W.; Liu, S.; and Li, G. 2023. Adaptive Annotation Distribution for Weakly Supervised Point Cloud Semantic Segmentation. *arXiv preprint arXiv:2312.06259*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; and Ghanem, B. 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35: 23192–23204.

Rong, M.; Cui, H.; and Shen, S. 2023. Efficient 3D Scene Semantic Segmentation via Active Learning on Rendered 2D Images. *IEEE Transactions on Image Processing*.

Shan, T.; Englot, B.; Meyers, D.; Wang, W.; Ratti, C.; and Rus, D. 2020. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 5135–5142. IEEE.

Shao, F.; Luo, Y.; Liu, P.; Chen, J.; Yang, Y.; Lu, Y.; and Xiao, J. 2022. Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2575–2585.

Shi, X.; Xu, X.; Chen, K.; Cai, L.; Foo, C. S.; and Jia, K. 2021. Label-efficient point cloud semantic segmentation: An active learning approach. *arXiv preprint arXiv:2101.06931*.

Sun, C.; Zheng, Z.; Wang, X.; Xu, M.; and Yang, Y. 2022. Self-supervised point cloud representation learning via separating mixed shapes. *IEEE Transactions on Multimedia*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Unal, O.; Dai, D.; and Van Gool, L. 2022. Scribble-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2697–2707.

Wang, H.; Liu, Q.; Yue, X.; Lasenby, J.; and Kusner, M. J. 2021. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9782–9792.

Wu, T.-H.; Liu, Y.-C.; Huang, Y.-K.; Lee, H.-Y.; Su, H.-T.; Huang, P.-C.; and Hsu, W. H. 2021. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15510–15519.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022a. Point transformer v2: Grouped vector attention and partition-based pooling. *arXiv preprint arXiv:2210.05666*.

Wu, Y.; Yan, Z.; Cai, S.; Li, G.; Yu, Y.; Han, X.; and Cui, S. 2022b. Pointmatch: a consistency training framework for weakly supervisedsemantic segmentation of 3d point clouds. *arXiv preprint arXiv:2202.10705*.

Wu, Z.; Wu, Y.; Lin, G.; Cai, J.; and Qian, C. 2022c. Dual Adaptive Transformations for Weakly Supervised Point Cloud Segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, 78–96. Springer.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 574–591. Springer.

Xu, X.; and Lee, G. H. 2020. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13706–13715.

Yang, C.-K.; Wu, J.-J.; Chen, K.-S.; Chuang, Y.-Y.; and Lin, Y.-Y. 2022. An mil-derived transformer for weakly supervised point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11830–11839.

Yu, L.; Li, X.; Fu, C.-W.; Cohen-Or, D.; and Heng, P.-A. 2018. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2790–2799.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19313–19322.

Zhang, J.; Li, W.; and Ogunbona, P. 2017. Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1859–1867.

Zhang, N.; Pan, Z.; Li, T. H.; Gao, W.; and Li, G. 2023. Improving Graph Representation for Point Cloud Segmentation via Attentive Filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1244–1254.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021a. Weakly supervised semantic segmentation for large-scale point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3421–3429.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021b. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15520–15528.

Zhang, Z.; Girdhar, R.; Joulin, A.; and Misra, I. 2021c. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10252–10263.