The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Weakly Supervised 3D Segmentation via Receptive-Driven Pseudo Label Consistency and Structural Consistency

Yuxiang Lan[1*], Yachao Zhang[1*], Yanyun Qu [1†], Cong Wang[2] Chengyang Li [3], Jia Cai [3]
Yuan Xie [3†], Zongze Wu[4]
[1]School of Informatics, Xiamen University, Fujian, China
[2]Huawei Technologies, Shanghai, China
[3]School of Computer Science and Technology, East China Normal University, Shanghai, China
[4] School of Mechatronics and Control Engineering, Shenzhen University, Guangdong, China
{lanyuxiang,yachaozhang}@stu.xmu.edu.cn, yxie@cs.ecnu.edu.cn, yyqu@xmu.edu.cn

Abstract

As manual point-wise label is time and labor-intensive for fully supervised large-scale point cloud semantic segmentation, weakly supervised method is increasingly active. However, existing methods fail to generate high-quality pseudo labels effectively, leading to unsatisfactory results. In this paper, we propose a weakly supervised point cloud semantic segmentation framework via receptive-driven pseudo label consistency and structural consistency to mine potential knowledge. Specifically, we propose three consistency contrains: pseudo label consistency among different scales, semantic structure consistency between intra-class features and class-level relation structure consistency between pair-wise categories. Three consistency constraints are jointly used to effectively prepares and utilizes pseudo labels simultaneously for stable training. Finally, extensive experimental results on three challenging datasets demonstrate that our method significantly outperforms state-of-the-art weakly supervised methods and even achieves comparable performance to the fully supervised methods.

Introduction

Point cloud semantic segmentation attracts more and more attention in the field of 3D computer vision due to its wide applications in many scenarios, including remote sensing, AR/VR, robotics, and automatic driving. Fully supervised deep learning methods (Wu, Qi, and Fuxin 2019; Hu et al. 2020; Han et al. 2020; Yan et al. 2020; Ma et al. 2020) are prevalently studied, which depends on densely annotated datasets. However, the full annotation for large-scale datasets with amount of millions of points is labor-intensive and time-consuming. Taking ScanNet-v2 (Dai et al. 2017) as an example, it takes 22.3 minutes to annotate one scene on average, and the entire dataset contains 1513 scenes. Due to the lack of supervision information, the performance will be severely degraded when directly extending the fully supervised method to weakly supervised learning (Hu et al. 2022). Therefore, recent works (Wei et al. 2020; Xu and Lee 2020; Cheng et al. 2021; Zhang et al. 2021a; Hu et al. 2022; Liu, Qi, and Fu 2021) have been attracting more attention to the weakly supervised point cloud segmentation task.

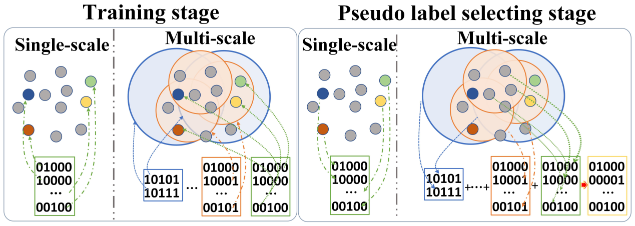

Figure 1: Comparison of single-scale inference and multi-scale inference in training stage and pseudo label selection stage. $\boxed{10101}$ indicates the multiple scale label in *Training stage* and prediction in *Pseudo label selection stage*

In 3D weakly supervised learning, current popular methods can be roughly divided into two categories: pseudo labeling based methods (Zhang et al. 2021a; Cheng et al. 2021; Liu, Qi, and Fu 2021) and consistency based methods (Xu and Lee 2020; Zhang et al. 2021b). Current pseudo labeling based methods (Zhang et al. 2021a; Cheng et al. 2021; Liu, Qi, and Fu 2021) generate pseudo labels at a single scale and use a pre-defined fixed threshold for pseudo label selection. Firstly, single-scale inference only supervises the output prediction, in the case of sparse supervision information, only a few valid labels can be used, which seriously affects the performance of the network, resulting in unsatisfactory results. So, it is not an ideal solution. In contrast, as shown in Figure 1, receptive-driven multi-scale inference generates multiple scale labels by enriching the supervision information within the receptive field in the process of down-sampling. Then, using multiple scale labels to supervise the features of multiple scales, which will more fully mine and exploit extremely limited supervised information. Secondly, in the pseudo label selection stage, only using single-scale inference to select pseudo labels will generate a large number of false pseudo labels, which will bring too much noise for training. Additionally, these works (Zhang et al. 2021a; Cheng et al. 2021; Liu, Qi, and Fu

[*]These authors contributed equally.
[†]Corresponding authors.

2021) pay more attention to how to generate pseudo labels, but ignore how to make full use of the auxiliary supervision information introduced by pseudo labels.

The consistency based methods (Xu and Lee 2020; Zhang et al. 2021b) adopt the self-supervision method based on consistency. These methods focus on meticulously designing data augmentation methods and constraining the consistency between original data predictions and their augmented data predictions. These methods pay attention on the perturbation invariance, but the connection between labeled and unlabeled points is still not well established. Moreover, double network branches introduce additional computation for the training networks and limit the practicality of the method to a certain extent.

To remedy these problems, in this paper, we propose a weakly supervised point cloud semantic segmentation framework via **R**eceptive-driven **P**seudo label consistency and **S**tructural **C**onsistency (RPSC), which contains three consistency: Receptive-driven pseudo label consistency, Semantic structural consistency and Relation structural consistency. Concretely, we generate corresponding multi-hot labels for each scale during down-sampling in training stage. In this way, as the receptive field expands and the unlabeled points cooperate with labeled points to share one multi-hot label, enriching more supervision information and effectively improving the availability of multi-hot labels in the weakly supervised setting. In the decoder, we generate score predictions for each scale, which are supervised by multi-hot labels for the corresponding scales. By supervising the features of each hidden layer, the supervision information can be fully excavated, and the discriminability of the final segmentation result can be improved. Furthermore, to alleviate noisy pseudo labels due to incorrect predictions, we use the consistency between single-scale pseudo labels and multi-scale pseudo labels to select reliable pseudo labels.

In semantic structural consistency and relation structural consistency, which are proposed to constrain the network training based on a key analysis that in a common space, the same category between labeled points and unlabeled points should have consistent feature distribution, and different categories should also have consistent category-level relation. Specifically, we first build a prototype memory bank and update it smoothly with the features of each batch of labeled points to obtain a more robust prototype. Then, in the feature space, keep the unlabeled points and the corresponding labeled point prototypes with the same category more compact, so the features of the labeled and unlabeled points have a consistent distribution, which is semantic structure consistency. Finally, we use the prototypes to calculate the category-level relation between labeled and unlabeled points, respectively. We ensure the relation structure consistency between labeled and unlabeled points, which provides additional supervision information beyond the point level, *i.e.*, category-level relation supervision.

Our contributions are summarized as follows:

• We propose a weakly supervised point cloud semantic segmentation framework via receptive-driven pseudo label consistency and structural consistency, which proposes three consistency constraints to effectively prepares and utilizes pseudo labels simultaneously for stable training.

• We introduce a receptive-driven pseudo label consistency method, which uses receptive-driven multi-scale scoring and pseudo label consistency to select high-quality pseudo labels. And structural consistency is established to constrain the consistency between labeled and unlabeled points with semantic features and category-level relations.

• Extensive experimental results demonstrate that RPSC significantly outperforms state-of-the-art weakly supervised competitors and even obtains comparable performance to fully competitors on three challenging datasets.

Related Work

Fully Supervised Point Cloud Semantic Segmentation. With the large-scale fully-annotated point cloud datasets (Armeni et al. 2016; Dai et al. 2017; Hackel et al. 2017; Behley et al. 2019), fully-supervised point cloud semantic segmentation has made great progress in recent years. 3D semantic segmentation methods can be roughly divided into voxel-based methods and point-based methods. Voxel-based methods (Graham, Engelcke, and Van Der Maaten 2018; Meng et al. 2019; Yan et al. 2021) voxelize point clouds into regular 3D grids and process them using dense 3D CNNs or sparse convolutions. However, information loss is inevitable during the voxelization process.

To avoid structuring, point-based methods (Qi et al. 2017a,b; Li et al. 2018; Wang et al. 2019; Wu, Qi, and Fuxin 2019; Lei, Akhtar, and Mian 2020a; Zhang et al. 2020; Yan et al. 2020) are proposed to directly process the raw unordered point cloud data. PointNet(Qi et al. 2017a) and PointNet++ (Qi et al. 2017a) are the pioneering ones. RandLA-Net (Hu et al. 2020) utilizes a random downsampling strategy to build an efficient and lightweight neural architecture, which makes it possible to process large-scale point cloud datasets efficiently. However, the methods mentioned above are fully supervised that require a large number of abeled samples. Thus, in this paper, we focus on weakly supervised point cloud semantic segmentation.

Weakly/Semi Supervised Point Cloud Semantic Segmentation. In weakly supervised learning, only a small amount of data is annotated, which reduces the annotating cost. Due to the few supervision, it is challenging to enforce the constraints for point cloud semantic segmentation. There are three weakly labeling methods: sub-cloud (Wei et al. 2020) which labels semantic categories in various subsets of each point cloud, labeling a tiny fraction of points (Xu and Lee 2020; Zhang et al. 2021a; Cheng et al. 2021; Zhang et al. 2021b), and a fraction of the point clouds with full labels (Jiang et al. 2021). In this paper, we adopt the second weakly labeling method of labeling a tiny fraction of points.

Weakly supervised methods can be roughly divided into two categories: consistency based methods and pseudo labeling based methods. The former (Tarvainen and Valpola 2017; Zhang et al. 2018) originates from the semi-supervised classification of 2D images. They depend on the fact that the predictions for the perturbed samples and the predictions for the original samples should be consistent. In PSD (Zhang et al. 2021b), a perturbed self-distillation framework is introduced by constructing perturbed samples

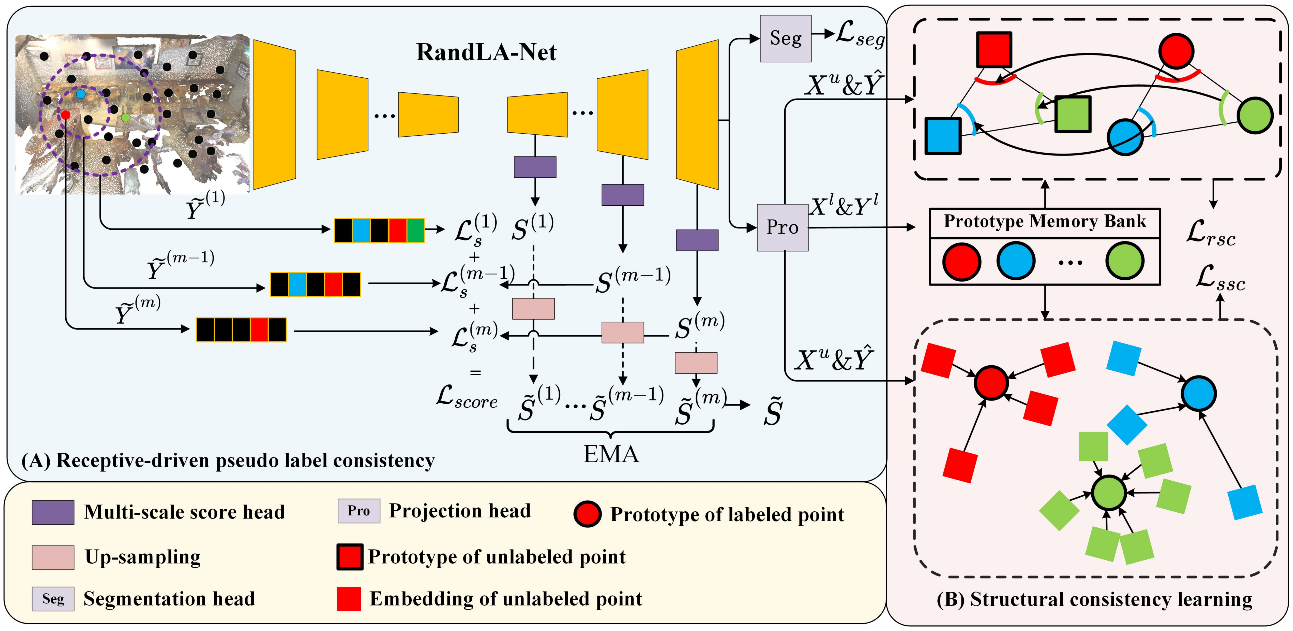

Figure 2: The framework of RPSC. (A) is used to generate the multi-scale labels for adding auxiliary supervision in training stage. In pseudo label selection stage, the reliability of pseudo labels are selected rely on the consistency between single-scale pseudo labels and multi-scale pseudo labels generated from (A). As we get the pseudo labels, we train the network by proposed structural consistency learning shown in (B), which contains relation structural consistency and semantic structural consistency

to ensure the predictive consistency among perturbed samples and original samples. However, consistency methods in both 2D images and 3D point clouds rely on well-designed and domain-specific data augmentation. Instead of perturbing consistency that requires task-specific redesign (Zhang et al. 2022), we focus on the pseudo-labels, which is a more flexible way.

Pseudo labeling based methods (Lee 2013; Rizve et al. 2021; Wang and Wu 2020; Iscen et al. 2019; Yarowsky 1995; Hu et al. 2021b) assign pseudo labels depending on the prediction confidence for unlabeled data assuming that high confidence corresponds to good accuracy. For the 3D point cloud segmentation, Zhang *et al.* (Zhang et al. 2021a) introduced a sparse label propagation method to generate pseudo labels to regularize network learning. In SPCC-Net (Cheng et al. 2021), a dynamic label propagation scheme is proposed based on the built superpoint graphs, involving label noisy training inevitably. We design three consistency regularization methods, semantic structural consistency and relation structural consistency, to make full use of pseudo labels.

Method

Problem Definition

Let a point cloud be denoted as $P = \{P^l, P^u\}$, where $P^l = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_L} = \{X^l, Y^l\}$ and $P^u = \{(\boldsymbol{x}_i, \varnothing)\}_{i=1}^{N_U} = \{X^u, \varnothing\}$ are labeled and unlabeled sets with N_L labeled points and N_U unlabeled points ($N_L \ll N_U$), respectively. $y_i \in \{1, 2, \cdots, C\}$ is the corresponding label of the labeled point $\boldsymbol{x}_i$, where C is the number of category. $N = N_L + N_U$ denotes the total number of points in one point cloud. We use clicked annotation setting in our paper, which means only one point or several points are labeled with the ground-truth for every category, and the annotated points are chosen randomly to alleviate the cost of annotation. Weakly supervised semantic segmentation involves learning a parameterized model $f_{\boldsymbol{\theta}}$ on P to predict the semantic category of each point in $X^l \cup X^u$.

Overview

The framework of our proposed receptive-driven pseudo label consistency and structural consistency (RPSC) weakly supervised point cloud semantic segmentation method is shown in Figure 2. RPSC jointly trains the backbone $f(\cdot)$ including the segmentation head, the projection head $g(\cdot)$ and the multi-scale score head $h^{(i)}(\cdot)$.

Specifically, we first perform a pre-training, *i.e.,* initialize the network parameters (backbone and multi-scale score head) by the available labeled data to construct cross entropy as segmentation loss and multi-scale score loss. After that, the single-scale pseudo labels generated from segmentation head for each unlabeled point is represented as:

$$\hat{Y} = \{\hat{y}_i\}_{i=1}^{N_U} = \{argmax f(\boldsymbol{x}_i)\}_{i=1}^{N_U}, \tag{1}$$

where $\hat{Y}$ is the set of all single-scale pseudo labels, $\hat{y}_i \in \{1, 2, \cdots, C\}$ denotes the single-scale pseudo label of the i^{th} unlabeled point, and $f(\boldsymbol{x}_i)$ is i^{th} point prediction distribution. Then we use the receptive-driven pseudo label consistency constraint to select high-quality pseudo labels with the help of the receptive-driven multi-scale scoring and pseudo label consistency selection. It explores a method for pseudo label selection under consistency constraint between single-scale pseudo labels and multi-scale pseudo labels. Finally, we project all points to a common feature space and enforce the constraints of the structural consistency to update network parameters, as well as to improve the quality of pseudo labels. We alternately perform pseudo label selection and network parameters updating.

Receptive-Driven Pseudo Label Consistency

Current pseudo labeling methods get poor performance by using the single-scale inference to select pseudo labels. Conversely, we propose a receptive-driven pseudo label consistency method, which contains two key components: receptive-driven multi-scale scoring in the training stage to gain the auxiliary supervision and pseudo label consistency selection in the pseudo label selection stage to select high-quality pseudo labels.

Receptive-Driven Multi-Scale Scoring. In order to take advantage of multi-scale inference, Inspired by RFCC (Gong et al. 2021; Hu et al. 2021a), we use original labels to generate multi-hot labels for each scale during down-sampling in the encoder, and the process is shown in Figure 3. As the down-sampling proceeds, the multi-hot labels are generated by performing the logical “OR” operation on the label vectors contained in the local neighborhood for each point of the current scale. Then each score head includes a linear layer and sigmoid activation function is attached for each decoder layer, generating a C-dimensional score prediction $\boldsymbol{S^{(i)}} \in \mathbb{R}^{N^{(i)} \times C}$ for i^{th} decoder layer, where $N^{(i)}$ is the number of points in the i^{th} decoder layer. We use the generated multi-hot labels to supervise the score predictions by BCE loss at each scale before up-sampling, and accumulate the losses at each scale to get the final loss as:

$$\mathcal{L}_{score} = \sum_{i=1}^{m} \mathcal{L}_s^{(i)} = \sum_{i=1}^{m} BCE(\widetilde{Y}^{(i)}, \boldsymbol{S^{(i)}}), \tag{2}$$

where $\widetilde{Y}^{(i)}$ are multi-hot labels of the scale corresponding to the i^{th} decoder layer, m represents the number of scales, and $BCE(\cdot,\cdot)$ denotes the binary cross entropy.

Pseudo Label Consistency Selection. In order to better fuse multi-scale score prediction, according to the previous down-sampling process, an inverse up-sampling operation acts on the score prediction $\boldsymbol{S^{(i)}}$ of each scale. We apply the nearest neighbor interpolation method to gradually restore the $\boldsymbol{S^{(i)}}$ to the same size as the original input, obtaining $\widetilde{\boldsymbol{S}}^{\boldsymbol{(i)}} \in \mathbb{R}^{N \times C}$. At last, we fuse the score prediction of multiple scales together through the Exponential Moving Average (EMA) to obtain the final receptive-driven multi-scale score prediction $\widetilde{\boldsymbol{S}} \in \mathbb{R}^{N \times C}$:

$$\widetilde{\boldsymbol{S}} \triangleq \delta \cdot \widetilde{\boldsymbol{S}} + (1 - \delta) \cdot \widetilde{\boldsymbol{S}}^{\boldsymbol{(i)}},\ i = \{2, \ldots, m\}. \tag{3}$$

When $i = 1$, $\widetilde{\boldsymbol{S}}$ is initialized as $\widetilde{\boldsymbol{S}}^{\boldsymbol{(1)}}$. And the fused score label $\hat{s}_i$ of i^{th} point generated as $\hat{S} = \{\hat{s}_i\}_{i=1}^{N_U} = \{argmax\ \widetilde{\boldsymbol{S}}_i\}_{i=1}^{N_U}$.

As shown in Figure 3, each label aggregates all the annotation information of a region. It is obvious that the proportion of valid labels increases as down-sampling proceeds.

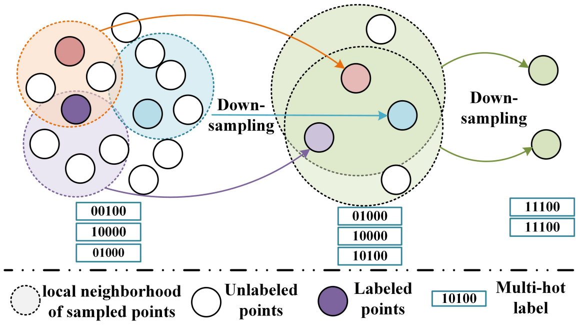

Figure 3: The generation process of multi-hot labels in encoder. 0 or 1 at each position indicates whether the corresponding category exists

To take advantage of the supervision information enriched by multi-hot labels, we use the generated multi-hot labels to supervise the score predictions before up-sampling instead of after up-sampling. Similarly, based on the advantage of multi-scale inference, we use the loss at each scale to accumulate the final loss, instead of only using the final fused multi-scale score prediction to calculate the loss. More importantly, using multi-scale inference to mine auxiliary supervision information at multiple scales and supervising the hidden layer features in the network during the training stage will help to improve the overall performance of the network.

At last, we use the prediction consistency of single-scale pseudo labels and multi-scale pseudo labels to select high-quality pseudo labels. We believe that if the pseudo labels under the two different scales can be consistent, it has a high credibility. So, the selected pseudo labels can be identified by a binary indicator ω_i for each unlabeled point $\boldsymbol{x}_i$:

$$\omega_i = \left\{ \begin{array}{ll} 1, & if\ max(f(\boldsymbol{x}_i)) > \mathcal{T}_{th}\ and\ \hat{y}_i = \hat{s}_i \\ 0, & otherwise \end{array} \right., \tag{4}$$

where $\mathcal{T}_{th}$ is the confidence threshold, which is set to 0.7 by experience. When $\omega_i = 1$, it means that the pseudo label is selected, otherwise it is discarded.

Structural Consistency Learning

In order to make better use of the extremely limited supervised information, we further mine the prior knowledge to enforce the constraints for training network and improving the quality of pseudo labels. And we make a reasonable assumption that between the labeled points and unlabeled points, the distribution in the same class and the relation of the pair-wise classes are both similar. Thus, we conduct two consistency constraint named semantic structural consistency and relation structural consistency.

Semantic Structural Consistency. It serves to constrain unlabeled points and labeled points to have consistent semantic representations. We hold that the features learned by labeled points are more accurate than those learned by unlabeled points. So, we use the features of labeled points to supervise unlabeled points. To alleviate the instability of single batch samples, we introduce a prototype memory bank to obtain a more robust feature representation. Firstly, for each point

$\boldsymbol{x}_i$, through projection head with L_2 normalization, it can be projected to a common feature space, getting feature representation $g(\boldsymbol{x}_i)$. Then a prototype memory bank $\{\boldsymbol{t}_{\boldsymbol{i}}^{\boldsymbol{mb}}\}_{i=1}^C$ is built to store prototypes of each category, which updated by calculating the prototype of labeled points in each batch during training. Given a batch train samples, *i.e.,* B point clouds, the category prototype $\boldsymbol{t}_c^l$ is defined as:

$$\boldsymbol{t}_c^l = \frac{1}{|\Omega_c|} \sum_{\boldsymbol{x}_i \in \Omega_c} g(\boldsymbol{x}_i), \tag{5}$$

where $\Omega_c = \{\boldsymbol{x}_i | y_i = c, i = \{1, \dots, BN_L\}\}$. In the first training batch, the prototype memory bank is initialized with the computed prototype. And then it is updated with the prototypes computed from the current batch by EMA:

$$\boldsymbol{t}_{\boldsymbol{c}}^{\boldsymbol{mb}} \triangleq \epsilon \cdot \boldsymbol{t}_{\boldsymbol{c}}^{\boldsymbol{mb}} + (1 - \epsilon) \cdot \boldsymbol{t}_c^l,\ c = \{1, \dots, C\}. \tag{6}$$

The prototype memory bank and unlabeled points feature $\{g(\boldsymbol{x}_i)\}_{i=1}^{N_U}$ are used to construct semantic structural consistency matrix M^s with the size of $C \times N_U$:

$$M_{ij}^s = \begin{cases} \|\boldsymbol{t}_{\boldsymbol{i}}^{\boldsymbol{mb}} - g(\boldsymbol{x}_j)\|_2, & if\ \hat{y}_j = i \\ 0, & otherwise \end{cases}. \tag{7}$$

where M_{ij}^s is the distance between i^{th} prototype and j^{th} unlabeled point in the feature space. We make the unlabeled points and their corresponding class prototypes compact and formulate the semantic structural consistency loss as:

$$\mathcal{L}_{ssc} = \frac{1}{\sum_{j=1}^{N_U} \omega_j} \sum_{i=1}^{C} \sum_{j=1}^{N_U} \omega_j \cdot M_{ij}^s, \tag{8}$$

where ω_j is pseudo label indicator defined in the Eq.(4). Introducing the prototype memory bank and updating it with the EMA can store more supervision information and reduce the interference of noise on network training.

Relation Structural Consistency. Both ground truth and pseudo labels only introduce independent point-level supervision. We further take the additional category-level relation supervision into account, and introduce the constraint of relation structural consistency, i.e., the similarity between any two categories of labeled and unlabeled points should be consistent.

Specifically, we design relation structural consistency based on the category prototypes instead of raw features, as a way to reduce the adverse effects of noisy pseudo labels. The category prototype of unlabeled points $\{\boldsymbol{t}_c^u\}_{c=1}^C$is defined as:

$$\boldsymbol{t}_c^u = \frac{1}{|\Omega_c^u|} \sum_{\boldsymbol{x}_i \in \Omega_c^u} g(\boldsymbol{x}_i), \tag{9}$$

where $\Omega_c^u = \{\boldsymbol{x}_i | \hat{y}_i = c, i = \{1, \dots, N_u\}\}$. Retrieving the labeled category prototypes from prototype memory bank, we quantify category-level relation both in labeled and unlabeled points by cosine similarity:

$$e_{ij}^l = {\boldsymbol{t}^{\boldsymbol{mb}}}_i^T \cdot {\boldsymbol{t}^{\boldsymbol{mb}}}_j, e_{ij}^u = {\boldsymbol{t}^{\boldsymbol{u}}}_i^T \cdot \boldsymbol{t}_j^u,\ i,j = \{1, 2, \cdots, C\} \tag{10}$$

The relation values is further normalized by:

$$\hat{e}_{ij}^l = e_{ij}^l / \sum_{j=1}^{C} e_{ij}^l,\ \hat{e}_{ij}^u = e_{ij}^u / \sum_{j=1}^{C} e_{ij}^l. \tag{11}$$

We have relation matrix of labeled point and unlabeled point as $E^l = \{\hat{e}_{ij}^l\} \in \mathbb{R}^{C \times C}$ and $E^u = \{\hat{e}_{ij}^u\} \in \mathbb{R}^{C \times C}$. Obviously, neither relation matrix is symmetric. Because in one scene with multiple categories, the mutual relation between two categories is not equal. To maintain the category-level relation structural consistency between labeled points and unlabeled points, we adopt the Kullback-Leibler divergence as the learning objective and formulate relation structural consistency loss as:

$$\begin{aligned} \mathcal{L}_{rsc} =& \frac{1}{C} \sum_{i=1}^{C} D_{KL}(E_{i\cdot}^l || E_{i\cdot}^u) \\ &+ \alpha \cdot \sum_{i=1}^{N_L} \sum_{j=1}^{N_L} \|g(\boldsymbol{x}_i)^T \cdot g(\boldsymbol{x}_j) - \boldsymbol{y}_i^T \cdot \boldsymbol{y}_j\|, \end{aligned} \tag{12}$$

where $D_{KL}(E_{i\cdot}^l || E_{i\cdot}^u) = \sum_{j=1}^{C} \hat{e}_{ij}^l log \frac{\hat{e}_{ij}^l}{\hat{e}_{ij}^u}$ in the first term, $E_{i\cdot}^l$ and $E_{i\cdot}^u$ denote the i^{th} row of the matrix. The second term is the regularization term, which using the ground truth to supervise labeled points such that each category has a distinct boundary in the feature space.

Objective Function

The total loss of RPSC contains four loss terms: (1) a segmentation loss $\mathcal{L}_{seg}$, (2) a receptive-driven multi-scale score loss $\mathcal{L}_{score}$; (3) a semantic structural consistency loss $\mathcal{L}_{ssc}$ and (4) a relation structural consistency loss $\mathcal{L}_{rsc}$. Specifically, $\mathcal{L}_{seg}$ calculates the cross-entropy between the model's prediction and the ground truth for labeled points, and that between the model's prediction and the pseudo label for unlabeled point, which is formulated as:

$$\begin{aligned} \mathcal{L}_{seg} =& \frac{1}{N_L} \sum_{i=1}^{N_L} CE(one\_hot(y_i), f(\boldsymbol{x}_i)) \\ &+ \frac{1}{N_U} \sum_{i=1}^{N_U} \omega_i \cdot CE(one\_hot(\hat{y}_i), f(\boldsymbol{x}_i)). \end{aligned} \tag{13}$$

where $CE(y, p)$ denotes the cross entropy between two distributions y and p, $one\_hot(\cdot) : \mathbb{R} \to \mathbb{R}^C$ is an function which converts the label to an one-hot vector. Our overall training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{score} \cdot \mathcal{L}_{score} + \lambda_{ssc} \cdot \mathcal{L}_{ssc} + \lambda_{rsc} \cdot \mathcal{L}_{rsc}, \tag{14}$$

where λ_{score}, λ_{ssc} and λ_{rsc} are scalar hyperparameters for trade-off these losses.

Experiments

Experimental Details

Datasets. To thoroughly evaluate our RPSC, we adopt three challenging large-scale point cloud benchmarks: S3DIS (Armeni et al. 2016), ScanNet-v2 (Dai et al. 2017) and SemanticKITTI (Behley et al. 2019). The first two are indoor

	Methods	Setting	S3DIS	ScanNet	SemKITTI
Fully	PointConv ('19) (Wu, Qi, and Fuxin 2019)	100%	50.3	55.6	-
	KPConv ('19) (Thomas et al. 2019)		**67.1**	**68.4**	-
	PointGCR ('20) (Ma et al. 2020)		54.4	60.8	-
	RandLA-Net ('20) (Hu et al. 2020)		63.0	64.5	53.9
	SPH3D ('20) (Lei, Akhtar, and Mian 2020b)		59.5	61.0	-
	PointASNL ('20) (Yan et al. 2020)		-	63.0	46.8
	Point2Node ('20) (Han et al. 2020)		63.0	-	-
	JSNet ('20) (Zhao and Tao 2020)		54.5	-	-
	SegGCN ('20) (Lei, Akhtar, and Mian 2020a)		63.6	58.9	**54.3**
Weakly	MPRM ('20) (Wei et al. 2020)	sub-cloud	-	41.1	-
	Xu and Lee ('20) (Xu and Lee 2020)	$1pt$-b ($\sim$ 2‰)	44.5	-	-
	Xu and Lee ('20) (Xu and Lee 2020)	10%	48.0	-	-
	SSPC-Net ('21) (Cheng et al. 2021)	0.1‰	51.5	27.1	41.0
	SSPC-Net ('21) (Cheng et al. 2021)	$1pt$-b ($\sim$ 2‰)	53.8	-	-
	Zhang *et al.* ('21) (Zhang et al. 2021a)	$1pt$-c	45.8	-	-
	Zhang *et al.* ('21) (Zhang et al. 2021a)	1%	61.8	51.1	-
	PSD ('21) (Zhang et al. 2021b)	1%	63.5	54.7	-
	SQN ('22) (Hu et al. 2022)	1‰	61.4	56.9	50.8
	Baseline	1‰	59.9	54.4	48.8
	Ours	1‰	63.1	57.5	50.9
	Baseline	$1pt$-c	41.7	40.5	42.5
	Ours	$1pt$-c	56.3	48.6	44.8
	Baseline	$50pt$-c	58.7	54.6	49.2
	Ours	$50pt$-c	**64.0**	**58.7**	**51.0**

Table 1: Quantitative results (mIoU, %) on Area 5 of S3DIS (Armeni et al. 2016), ScanNet-v2 (Dai et al. 2017) and SemanticKITTI (SemKITTI) (Behley et al. 2019). *Per class mIoU results are shown in the supplementary materials*

scene datasets and the last one is outdoor scene dataset. Each scene of three datasets contains $10^6 \sim 10^8$ points. We use all points of the original test set for the fair comparison.

Following the weakly supervised setting of previous works (Zhang et al. 2021a,b), we randomly select x points annotated with the ground truth of each category in each point cloud, which denotes as xpt-c, and $x = \{1, 50\}$. It is different from (Xu and Lee 2020; Cheng et al. 2021), which annotate the points for each category in one block ($1m \times 1m$). For a clearer representation, we denote their labeling method as xpt-b. Usually, a point cloud will be divided into several blocks. Therefore, xpt-b has more labels than our xpt-c. The annotated proportions of the three datasets at the $1pt$-c and $50pt$-c settings are shown in Table 2. In SemanticKITTI, the official dataset divides a street scene into several small pieces. We use the official division as an input scene like RandLA-Net (Hu et al. 2020).

Implementation Details. Our framework uses RandLA-Net (Hu et al. 2020) as backbone and it is trained on a single NVIDIA Tesla T4 with Tensorflow 1.14. The Adam Optimizer is adopted for training with an initial learning rate of 0.01 and momentum of 0.9. We first pre-train our network for 100 epoches using labeled points. Then we perform 10 iterations of training. In each training iteration, we train our network for 30 epoches by $\mathcal{L}_{total}$ in Eq. (14) with only loading the parameters of the encoder part of the previous model. In all experiments, we set the hyperparameters $\delta = 0.8$, $\epsilon = 0.9$ and $\alpha = 0.1$ empirically, the scalar hyperparameters $\lambda_{score} = 1.5$, $\lambda_{ssc} = 0.75$ and $\lambda_{rsc} = 0.1$ are selected through experiments (experiment results and analysis are shown in supplementary materials), while the batch size is kept fixed to 8 in all dataset experiments.

Experiment Results

Results on Area-5 of S3DIS. In the Table 1, we give the quantitative comparison results of state-of-the-art approaches including fully supervised and weakly supervised methods on the Area 5 test set. It shows that compared with other weakly supervised methods, RPSC achieves the best performance with the same amount of annotations. At the $1pt$-c setting, RPSC gains 10.5% mIoU against Zhang *et al.* (Zhang et al. 2021a) with the same number of labeled points. Interestingly, RPSC only need $50pt$-c or 1‰ labeled points to exceed the performance of the backbone network (RandLA-Net) with full supervision. Under the same weakly supervised setting of 1‰, RPSC gains 1.7% in mIoU against SQN, and even exceeds the setting of other methods by 1%.

Results on ScanNet-v2. The comparison results on ScanNet are shown in the fifth column of Table 1. At the $1pt$-c setting, RPSC gains 7.5% mIoU against MPRM (Wei et al. 2020) with fewer annotated points. At the 1‰ and $50pt$-c settings, RPSC can achieve better performance than most weakly supervised methods without using more annotation information than them. *As analyzed by SQN (Hu et al. 2022), 1T1C (Liu, Qi, and Fu 2021) has two serious problems with ground truth label leakage and misleading (over-exaggerated) labeling ratios in the experiment, so it can be regarded as almost full supervision methods on ScanNet. For fairness, we did not compare with them directly.*

Results on SemanticKITTI. The sixth column of Table 1 lists the test results on the online test set of SemanticKITTI dataset. Our $1pt$-c setting has a similar amount of annotation

Settings	S3DIS	ScanNet	SemKITTI
$1pt\text{-}s$	0.009‰	0.054‰	0.090‰
$50pt\text{-}s$	0.44‰	2.17‰	4.02‰

Table 2: Annotation proportion of $1pt\text{-}c$ and $50pt\text{-}c$ settings on all datasets

	Base.w.PLF	PLC	SSC	RSC	mIoU
#1	✓				51.4
#2	✓	✓			54.5
#3	✓		✓		52.7
#4	✓			✓	53.8
#5	✓	✓	✓		55.1
#6	✓	✓		✓	55.8
#7	✓		✓	✓	54.6
#8	✓	✓	✓	✓	**56.3**

Table 3: Results (mIoU, %) of ablation study at the $1pt\text{-}c$ setting on Area 5 of S3DIS. Base.w.PLF means using Baseline with pseudo labeling framework

as the 0.1‰ setting, but our RPSC can be improved by 3.8% mIoU over SSPC-Net. Similarly, at the 0.1‰ setting, RPSC also gains 0.1% compared to SQN.

Ablation Study

To further study the effectiveness of three main components: receptive-driven pseudo label consistency, semantic structural consistency and relation structural consistency, we show the ablation study results in Table 3.

Effectiveness of Pseudo Label Consistency (PLC). From #1 → #2, #3 → #5, #4 → #6, #7 → #8 in Table 3, the performance achieves 3.1%, 2.4%, 2% and 1.7% improvement after using pseudo label consistency under the four contrasting settings. The above results fully demonstrate the effectiveness of our proposed pseudo label consistency.

Effectiveness of Semantic Structural Consistency (SSC). From #1 → #3, #2 → #5, #4 → #7, #6 → #8 in Table 3, we can find that when the unlabeled points are supervised by the prototype of the labeled points, that is, after the constraint of semantic structure consistency is established, the performance of the network can be improved.

Effectiveness of Relation Structural Consistency (RSC). Comparing #1 → #4, #2 → #6, #3 → #7, #5 → #8, the performance achieves 2.4%, 1.3%, 1.9% and 1.2% improvement, because of the usage of relation structural consistency.

Method	# Para.	T_{test}	mIoU
Baseline	**1.02**	**210**	41.7
RPSC	1.23	213	**56.3**

Table 4: The total test time (in seconds), network parameters (# Para., in millions) and mIoU (%) on Area 5 of S3DIS

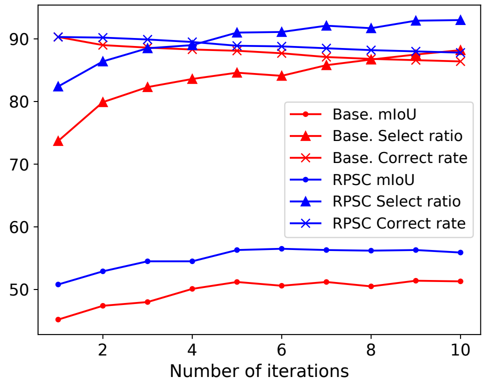

Figure 4: The analysis of pseudo label on S3DIS. The figure shows the trend of mIoU, the pseudo labels selection ratio, and the correct rate of selected pseudo labels with iteration.

Analysis

Generalizability of the Framework. We conduct further experiments to demonstrate our method is backbone independent and generalized. When we choose PointNet++ (Qi et al. 2017b) as the backbone, RPSC still achieves the 49.9% mIoU at the $50pt\text{-}c$ setting. which is even higher than the results of Xu and Lee (48.0%) with 10% labeled points, whose backbone is also PointNet++. Therefore, RPSC is a generalized framework which can be instantiated with other deep segmentation models for point clouds.

Efficiency Analysis. We conduct the efficiency analysis at the $1pt\text{-}c$ setting on Area 5 of S3DIS, whose results are shown in Table 4. In the training stage, we introduce an additional projection head and multiple score head, which make the parameter increase by 0.21M compared to Baseline. In the test stage, without the projection head and the multi-scale score head, no extra computation is introduced against Baseline, so RPSC have almost the same test time as Baseline, yet improve the performance by 14.6% mIoU.

Pseudo Label Consistency Efficiently. As shown in Figure 4, both the number and correct rate of pseudo labels selected by using our pseudo label consistency selection method are higher than those selected by single-scale inference, which in turn leads to better network performance (mIoU).

Conclusions

We propose a weakly supervised point cloud semantic segmentation framework to efficiently prepare and exploit pseudo labels for mining auxiliary supervision information from sparse labels, which has two benefits: obtaining reliable pseudo labels and training stable. Such advantages are based on three consistency contrains: the pseudo label consistency, the semantic structure consistency and relation structure consistency. The first one uses consistency between single-scale pseudo labels and multi-scale pseudo labels to select high-quality pseudo labels. The last two provide additional supervision information besides the direct supervision of pseudo labels. Extensive experimental results on three benchmarks demonstrate the effectiveness of RPSC with extremely sparse labels.

Acknowledgments

This work was supported by the National Key Research and Development Program of China No. 2020AAA0108301, the National Natural Science Foundation of China No. 6217224, 61876161, 62222602, 62176092, Shanghai Science and Technology Commission No.21511100700, Natural Science Foundation of Shanghai (20ZR1417700), CAAI-Huawei MindSpore Open Fund, the Fundamental Research Funds for the Central Universities.

References

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 *CVPR*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 9297–9307.

Bouville, M. 2008. Crime and punishment in scientific research. arXiv:0803.4058.

Cheng, M.; Hui, L.; Xie, J.; and Yang, J. 2021. SSPC-Net: Semi-supervised Semantic 3D Point Cloud Segmentation Network. In *AAAI*, volume 35, 1140–1147.

Clancey, W. J. 1979. *Transfer of Rule-Based Expertise through a Tutorial Dialogue*. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Clancey, W. J. 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, 556–560. Menlo Park, Calif: IJCAI Organization.

Clancey, W. J. 1984. Classification Problem Solving. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, 45–54. Menlo Park, Calif.: AAAI Press.

Clancey, W. J. 2021. The Engineering of Qualitative Models. Forthcoming.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 5828–5839.

Engelmore, R.; and Morgan, A., eds. 1986. *Blackboard Systems*. Reading, Mass.: Addison-Wesley.

Gong, J.; Xu, J.; Tan, X.; Song, H.; Qu, Y.; Xie, Y.; and Ma, L. 2021. Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11673–11682.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 9224–9232.

Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J. D.; Schindler, K.; and Pollefeys, M. 2017. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 91–98.

Han, W.; Wen, C.; Wang, C.; Li, X.; and Li, Q. 2020. Point2Node: Correlation Learning of Dynamic-Node for Point Cloud Feature Modeling. In *CVPR*, 10925–10932.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20(1): 3–19.

Hu, Q.; Yang, B.; Fang, G.; Guo, Y.; Leonardis, A.; Trigoni, N.; and Markham, A. 2022. SQN: Weakly-Supervised Semantic Segmentation of Large-Scale 3D Point Clouds. In *ECCV*.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, 11108–11117.

Hu, R.; Liu, Y.; Gu, K.; Min, X.; and Zhai, G. 2021a. Toward a No-Reference Quality Metric for Camera-Captured Images. *IEEE Transactions on Cybernetics*.

Hu, R.; Liu, Y.; Wang, Z.; and Li, X. 2021b. Blind quality assessment of night-time image. *Displays*, 69: 102045.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label propagation for deep semi-supervised learning. In *CVPR*, 5070–5079.

Jiang, L.; Shi, S.; Tian, Z.; Lai, X.; Liu, S.; Fu, C.-W.; and Jia, J. 2021. Guided Point Contrastive Learning for Semi-supervised Point Cloud Semantic Segmentation. In *ICCV*, 6423–6432.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML,Workshop*.

Lei, H.; Akhtar, N.; and Mian, A. 2020a. SegGCN: Efficient 3D Point Cloud Segmentation With Fuzzy Spherical Kernel. In *CVPR*, 11611–11620.

Lei, H.; Akhtar, N.; and Mian, A. 2020b. Spherical kernel for efficient graph convolution on 3d point clouds. *IEEE TPAMI*.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, 820–830.

Liu, Z.; Qi, X.; and Fu, C.-W. 2021. One Thing One Click: A Self-Training Approach for Weakly Supervised 3D Semantic Segmentation. In *CVPR*, 1726–1736.

Ma, Y.; Guo, Y.; Liu, H.; Lei, Y.; and Wen, G. 2020. Global Context Reasoning for Semantic Segmentation of 3D Point Clouds. In *CVPR*, 2931–2940.

Meng, H.-Y.; Gao, L.; Lai, Y.-K.; and Manocha, D. 2019. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *ICCV*, 8500–8508.

NASA. 2015. Pluto: The 'Other' Red Planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 5099–5108.

Rice, J. 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*.

Robinson, A. L. 1980. New Ways to Make Microcircuits Smaller. *Science*, 208(4447): 1019–1022.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 1195–1204.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 6411–6420.

Wang, G.-H.; and Wu, J. 2020. Repetitive reprediction deep decipher for semi-supervised learning. In *AAAI*, 6170–6177.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics*, 38(5): 1–12.

Wei, J.; Lin, G.; Yap, K.-H.; Hung, T.-Y.; and Xie, L. 2020. Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds. In *CVPR*, 4384–4393.

Wu, W.; Qi, Z.; and Fuxin, L. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 9621–9630.

Xu, X.; and Lee, G. H. 2020. Weakly Supervised Semantic Point Cloud Segmentation: Towards 10x Fewer Labels. In *CVPR*, 13706–13715.

Yan, X.; Gao, J.; Li, J.; Zhang, R.; Li, Z.; Huang, R.; and Cui, S. 2021. Sparse Single Sweep LiDAR Point Cloud Segmentation via Learning Contextual Shape Priors from Scene Completion. In *AAAI*, volume 35, 3101–3109.

Yan, X.; Zheng, C.; Li, Z.; Wang, S.; and Cui, S. 2020. PointASNL: Robust Point Clouds Processing using Nonlocal Neural Networks with Adaptive Sampling. In *CVPR*, 5589–5598.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of The Association for Computational Linguistics*, 189–196.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. In *ICLR*.

Zhang, Y.; Li, M.; Xie, Y.; Li, C.; Wang, C.; Zhang, Z.; and Qu, Y. 2022. Self-supervised Exclusive Learning for 3D Segmentation with Cross-Modal Unsupervised Domain Adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3338–3346.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021a. Weakly Supervised Semantic Segmentation for Large-Scale Point Cloud. In *AAAI*, volume 35, 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 *ICCV*, 15520–15528.

Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; and Foroosh, H. 2020. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*, 9601–9610.

Zhao, L.; and Tao, W. 2020. Jsnet: Joint instance and semantic segmentation of 3d point clouds. In *AAAI*, volume 34, 12951–12958.