HybridCR: Weakly-Supervised 3D Point Cloud Semantic Segmentation via Hybrid Contrastive Regularization

Mengtian Li[1], Yuan Xie[1], Yunhang Shen[2], Bo Ke[2], Ruizhi Qiao[2], Bo Ren[2], Shaohui Lin[1,†], Lizhuang Ma[1,†]
[1]School of Computer Science and Technology
East China Normal University, Shanghai, China
[2]Tencent Youtu Lab
mtli@stu.ecnu.edu.cn, {yxie,shlin,lzma}@cs.ecnu.edu.cn
{odysseyshen,boke,ruizhiqiao,timren}@tencent.com

Abstract

To address the huge labeling cost in large-scale point cloud semantic segmentation, we propose a novel hybrid contrastive regularization (HybridCR) framework in weakly-supervised setting, which obtains competitive performance compared to its fully-supervised counterpart. Specifically, HybridCR is the first framework to leverage both point consistency and employ contrastive regularization with pseudo labeling in an end-to-end manner. Fundamentally, HybridCR explicitly and effectively considers the semantic similarity between local neighboring points and global characteristics of 3D classes. We further design a dynamic point cloud augmentor to generate diversity and robust sample views, whose transformation parameter is jointly optimized with model training. Through extensive experiments, HybridCR achieves significant performance improvement against the SOTA methods on both indoor and outdoor datasets, e.g., S3DIS, ScanNet-V2, Semantic3D, and SemanticKITTI.

1. Introduction

Learning the precise semantic meanings of large-scale point clouds is a fundamental perception task for intelligent machines to understand complex 3D scenes. Existing deep-learning-based methods heavily rely on the availability and quantity of labeled point cloud data for training [5, 21, 22, 29]. However, 3D point-wise labeling is time-consuming and labor-intensive. Hence, we aim to explore weakly-supervised learning to maximize the data efficiency and reduce efforts to annotate 3D point clouds.

Recently, several 3D point cloud weakly-supervised semantic segmentation methods have been emerged, which can be generally divided into three groups: (1) Consistency regularization [33, 38] employs consistency constrain from the distribution of prediction after randomly modifying the input or model function. (2) Pseudo labeling, *a.k.a.* self-training [4, 18, 37], uses the model predictions as supervision. (3) Contrastive pre-training [9, 32] focuses on model pre-trained, which is then followed by fine-tuning with fewer labels for downstream tasks.

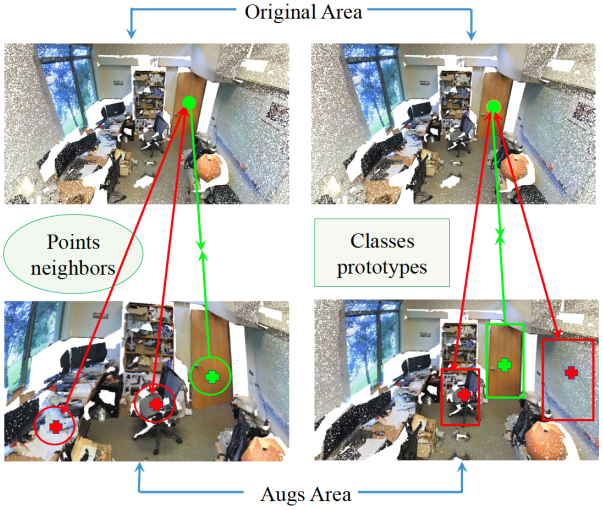


Figure 1. Hybrid contrastive regularization at local and global. **Left:** The anchor point is encouraged to be similar to the matched positive point and its neighbors (in green circle) while being dissimilar to negative points and their neighbors (in red circle). **Right:** The anchor point is encouraged to be similar to the matched positive point and other points that belong to the same class (in green box) while being dissimilar to negative points of different classes (in red box).

Although the existing methods have achieved encouraging results, some limitations remain to be addressed. Firstly, they do not adequately consider the semantic properties of neighbors and global characteristics of 3D classes for large-scale scenarios, failing to fully exploit the limited yet valu-

† Corresponding authors.

able annotations [33]. Secondly, many pipelines [33, 38] used fixed/handcrafted data augmentation to get multi-view representation resulting in sub-optimal learning, as the strength and types of the augmentation depends strongly on model and dataset size. Besides, the shape complexity of the samples is ignored in the fixed augmentation. Thirdly, the existing methods [9, 37] usually involve multiple stages pre-training and fine-tuning, which raise difficult training and deploy in practice compared to the end-to-end training scheme.

To address the above shortcomings, we explore to simultaneously leverage the consistency and contrastive property in label space and feature space, respectively. Inspired by recent 3D PSD [38] and 2D FixMatch [27], we combine the pseudo label and consistency regularization strategy in an end-to-end training scheme for large-scale point clouds. To better use contrastive information, we redesign the positive pairs and negative pairs of anchor points. A key observation is that high-level semantic scene understanding requires not only local but also global geometric features, making point cloud instances contrasting more sufficiently Besides, motivated by PointAugment [15] in the classification task, we further introduce dynamic point cloud augmentor to provide transformations for consistency and contrastive regularization with jointly optimization.

To implement the above idea, we propose a new paradigm, called hybrid contrastive regularization (HybridCR), for weakly-supervised semantic segmentation on large-scale point clouds, which consists of local and global guidance contrastive learning along with dynamic point cloud transformations. As shown in Fig. 1, local guidance contrastive regularization forces data sample of different views to be close to their neighbors and far away from other points. For global guidance contrastive regularization, each sample is imposed to to be close to the prototype of its class and far away from different classes prototypes. Fundamentally, HybridCR explicitly and effectively considers the semantic similarity among the local neighboring points and global characteristics of 3D point cloud classes. Furthermore, the proposed dynamic point cloud augmentor use multi-layer perceptrons (MLPs) and Gaussian noises to enrich the data diversity in context-wise displacement, where the parameters of augmentor can be jointly optimized with model training. Extensive experiments show that HybridCR achieves the SOTA performance for both indoor scenes, *i.e*., S3DIS [1] and ScanNet-V2 [6], and outdoor scenes, *i.e*., Semantic3D [8] and SemanticKITTI [2], demonstrating the effectiveness of our proposed framework.

To summarize, our contributions are four-fold:

- We propose the first framework HybridCR to leverage both point consistency and contrastive properties for weakly-supervised point cloud semantic segmentation in an end-to-end manner.
- We introduce the local and global guidance contrastive regularization to promote high-level 3D semantic scene understanding tasks.
- We design a novel dynamic point cloud augmentor to transform diverse and robust sample views, which is jointly optimized with the whole training process.
- HybridCR achieves significant performance over recent weakly-supervised methods and gains 2.4% and 1.0% AP improvements on average in indoor and outdoor datasets, respectively.

2. Related Work

2.1. Weakly-supervised point cloud segmentation

Weakly-supervised learning is an effective way to reduce high labor costs. Some weakly labeling methods have made preliminary attempts, such as labeling a tiny fraction of points [18, 33, 38] or semantic classes [31]. Existing methods use various means to improve the expressive ability of models. They can be roughly divided into three categories:

Consistency regularization achieves a perspective performance in weakly-supervised image classification [28, 36, 40]. Xu *et al*. [33] introduce a multi-branch supervision method for point cloud feature where two types of point cloud augmentation and consistency regularization are adopted. Zhang *et al*. [38] provide additional supervision by perturbed self-distillation for implicit information propagation. Shi *et al*. [26] investigate label-efficient learning and introduce a super-point-based active learning strategy. Despite benefiting from the consistency of different network branches, they fail to consider the contrastive property in feature space.

Pseudo labeling creates supervision from the predictions of a trained models [14, 24], assigned by neighborhood graphs [11], or self-training [19, 35]. In the weakly-supervised setting. Zhang *et al*. [37] propose a transfer learning-based method and introduced sparse pseudo labels to regularize network learning. Hu *et al*. [18] propose a self-training strategy to utilize the pseudo labels to improve the network performance. Cheng *et al*. [4] utilize a dynamic label propagation scheme to generate pseudo labels based on the built super-point graphs. However, they only use the pseudo labels to gain more supervised signals and ignore the consistency property in label space.

Contrastive pre-training first proposed by Xie *et al*. [32] and initiate the efforts through presenting a contrastive learning framework for point cloud scenes. However, it mainly focuses on downstream tasks with 100% labels. Hou *et al*. [9] leverage the inherent properties of scenes to expand the network transferability. Li *et al*. [12] propose the guided point contrastive loss and leverage pseudo-label to learn discriminative features. However,

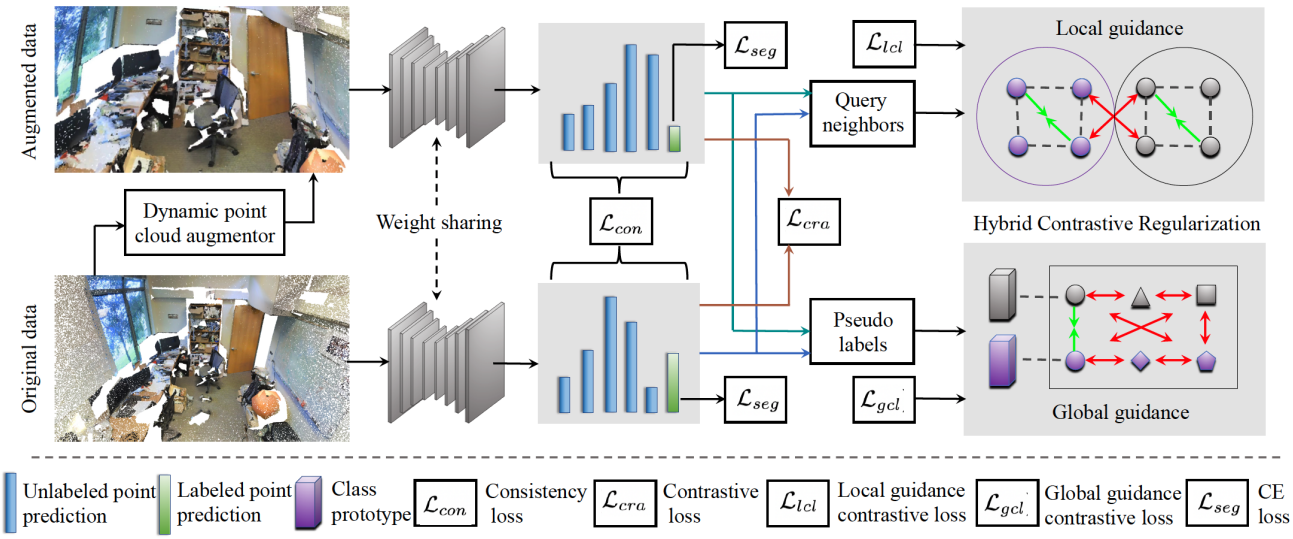


Figure 2. The original point clouds are first fed into the dynamic augmentor to generate augmented points. Then, the original points and augmented ones pass through the Siamese network to generate the model's predictions on all points, as well as pseudo labels for unlabeled points with high confidence. Point-level consistency loss L_{con} and contrast loss L_{cra} are used for the predictions of all points, while softmax cross-entropy loss L_{seg} performs on the supervision of labeled points. Meanwhile, pseudo labels are used to compute the prototypes for each class. Finally, HybridCR is conducted on both local and global perspectives to form local and global guidance contrastive losses (*i.e.*, L_{lcl} and L_{gcl}) to provide regularization for feature learning. By this way, HybridCR serves for the weakly-supervised framework in the end-to-end training scheme

they only conduct the point-level contrast in feature space while ignoring the inherent property of point clouds, *i.e.*, geometry structures and classes semantics.

HybridCR redesigns the local and global positive and negative pairs of large-scale point clouds and fully explores how to leverage and simultaneously enforce both consistency and contrastive properties in an end-to-end manner.

2.2. Point cloud augmentation

The data augmentation in existing networks [33, 38] mainly includes random rotation, scaling, and jittering, which are handcrafted/fixed throughout the training process. Li *et al*. [15] propose an auto-augmentation framework by leveraging adversarial learning strategy. Chen *et al*. [3] present this by interpolation between examples. Kim *et al*. [13] leverage local weighted transformations to produces non-rigid deformations. However, they merely focus on the object-level point clouds. Besides, it is complex to implement them in practical applications, which brings difficulties to tune the parameters during training and merely focus on the object-level point clouds. We introduce a dynamic point cloud augmentor to generate diverse transformations for large-scale point clouds during training.

3. Method

In this part, we first describe the notations and preliminaries in Sec. 3.1. Then, we present the general framework of HybridCR with the local and global guidance contrastive regularization in Sec. 3.2. Next, we introduce the dynamic point cloud augmentor in Sec. 3.3. Lastly, we present the overall objective for training in Sec. 3.4.

3.1. Preliminaries

Problem setup and notation. We let $\mathcal{D}$ be the point cloud dataset, which is defined as $\left\{\left(X^l, Y^l\right), \left(X^u, \varnothing\right)\right\} = \left\{\left(x_1^l, y_1^l\right), \ldots, \left(x_M^l, y_M^l\right), x_{M+1}^u, \ldots, x_N^u\right\}$, where N denotes the total number of points, M is the number of labeled points, X^l and X^u are the sets of the labeled and unlabeled points. For X^u, the labels are absence that is often replaced by pseudo labels Y^p generated on-the-fly. Thus, $Y = Y^l \cup Y^p$ are the whole label sets for weakly-supervised semantic segmentation. Note that Y^l is fixed, but Y^p is updated during training. Formally, given a large-scale point clouds with a tiny fraction of labels as input, weakly-supervised semantic segmentation aims to learn the function: $f_{\boldsymbol{\theta}} : X^l \cup X^u \mapsto Y$. Specifically, for 1% setting, the number of labeled points is $M = 1\% \times N$, and all the labeled points are selected randomly. The 1pt represents only one point labeled with the ground truth for each class, so the number of labeled points M equals to the number of classes C. Note that all the labeled points are selected randomly.

Point-level consistency and contrast. Point-level consistency [33,38] has been widely used in weakly-supervised point cloud semantic segmentation, which enforces a corresponding point pair with different augmentations into a

Siamese network to have same feature representation. Formally, the point-level consistency loss is formulated as:

$$\mathcal{L}_{con} = \frac{1}{2N}\sum_{i=1}^{N} JS\left(\tilde{y}_i \| \hat{y}_i\right), \tag{1}$$

where $\tilde{y}_i = f_\theta(x_i)$ and $\hat{y}_i = f_\theta(\hat{x}_i)$ are the predicted probabilities of the i-th point through the original branch and the data augment branch, respectively. JS is the Jensen-Shannon divergence.

Point-level contrast in self-supervised learning [32] is promoted by the supervised dense prediction tasks, $e.g.$, semantic segmentation, which performs dense per-point classification. Point-level contrast aims to pull the anchor (point x_i) to data-augmentation point while pushing it away from other points in the prediction space. Therefore, point-level contrastive loss is formulated as:

$$\mathcal{L}_{cra} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\tilde{y}_i \cdot \hat{y}_i/\tau\right)}{\sum_{j=1}^{N} \mathbb{1}_{[j\neq i]} \exp\left(\tilde{y}_i \cdot \hat{y}_j/\tau\right)}, \tag{2}$$

where $\mathbb{1}_{[j\neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 *iff* $j \neq i$ and τ is a temperature hyper-parameter. Note that Eq. 1 and Eq. 2 are computed across all points.

Pseudo label generation and selection. Pseudo labeling [14] uses the model's class prediction as supervision to train again, and benefits from the popular 2D Fixmatch [27]. It estimates the probability for all points by ground truth labels Y^l and generated pseudo labels Y^p. Let p_i be the probability outputs of the network with parameter $\boldsymbol{\theta}$ of point x_i, the p_{ic} represents the probability of class c being present in x_i. Using these output probabilities, the pseudo-label y_{ic}^p of x_i is generated. After generation, the pseudo-labels are selected with the high-confidence predictions by obtain a binary vector $\boldsymbol{g}_i$. Let $\boldsymbol{g}_i = [g_{i1}, \ldots, g_{iC}] \subseteq \{0,1\}^C$ be the selected pseudo-labels, which is obtained as:

$$g_{ic} = \mathbb{1}\left[p_{ic} \geq \tau_p\right], \tag{3}$$

where $g_{ic} = 1$ if y_{ic}^p is selected and $g_{ic} = 0$ otherwise. τ_p is the confidence thresholds for labels. The label is selected when the probability score is sufficiently high ($p_{ic} \geq \tau_p$).

High-level semantic scene understanding tasks require not only local but also global information, directly contrasting 3D instances merely on the point-level is insufficient [17, 32]. Therefore, this motivates us to explore more effective contrastive strategies to fully leverage the inherent properties of point clouds in both geometry structures and classes semantics.

3.2. Hybrid Contrastive Regularization

As depicted in Fig. 2, we propose a compact weakly-supervised semantic segmentation framework for large-scale point cloud that contains the novel hybrid contrastive regularization strategy (HybridCR) with the effective dynamic point cloud augmentor. The original point clouds are firstly fed into the dynamic point cloud augmentor to generate different transformations. Then, the original input points and augmented points pass through the Siamese network to generate pseudo labels using the model's predictions on unlabeled points. The model is encouraged to learn similar and robust features during training by matching 3D point pairs with different transformations. Meanwhile, the generated pseudo labels are used to compute the prototypes for each class. Finally, HybridCR is conducted on both local and global guidance perspectives to learn the feature relationship between unlabeled and labeled points, which also leverages the traditional segmentation loss for labeled points with point-level consistency and contrast losses.

3.2.1 Local guidance contrastive regularization

The local neighbor information is essential for feature learning on the objects of the point clouds. For example, occlusions and holes always exist in objects in indoor and outdoor scenes. If the model learns the local structure information (sphere, corner, *etc*.) from other complete objects, it can enhance the robustness of the model on incomplete objects during training. While the local feature of the point clouds mainly comes from the points and their neighbors, which inspires us to model the local information of the point cloud by the proposed local guidance contrastive regularization. To accomplish this, we first query the neighbor points for the anchor, and then force differently augmented views of each point to be close to their neighbors and far away from other points.

Given a 3D query point x_i with its coordinates xyz, we search its nearest K neighbor points by the point-wise Euclidean distance, and their encoded feature vectors are aggregated to generate a mean vector κ_i, which is computed by $\frac{1}{|\mathcal{N}(x_i)|}\sum_{j\in\mathcal{N}(x_i)} y_j$. Based on this, we construct the local guidance contrastive loss $\mathcal{L}_{lcl}$ following InfoNCE [20] by pulling $\tilde{y}_i$ close to κ_i, while pushing it away from the neighbor vector of other points:

$$\mathcal{L}_{lcl} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\tilde{y}_i \cdot \kappa_{ik}/\tau\right)}{\sum_{j=1}^{N} \mathbb{1}_{[j\neq i]} \exp\left(\tilde{y}_i \cdot \kappa_{jk}/\tau\right)}. \tag{4}$$

In fact, the proposed local guidance contrastive loss is more generalized to Eq. 2. Note that Eq. 4 is degenerated to Eq. 2 if K is set to 1.

3.2.2 Global guidance contrastive regularization

The global information is critical for point cloud objects and scenes recognition, where the objects from the same class should share the similar semantic features, even though they are very different in appearances. On the contrary, objects

belonging to different classes should be distinguishable in the feature space, no matter how similar they look. For example, chairs and tables are similar in appearance but belong to different classes. Thus, it is necessary for the network to obtain the critical information to avoid this trap. To end this, we leverage the semantic information from class labels by the proposed global guidance contrastive regularization. To accomplish this, we take the mean embedding of labeled points to generate its prototype ρ for each class, and ρ_{ic} is the prototype of the i-th point belonging to the c-th class. According to this, we construct the global guidance contrastive loss $\mathcal{L}_{gcl}$ by pulling $\tilde{y}_i$ close to ρ_i, while pushing it far away from the prototypes of the remaining classes:

$$\mathcal{L}_{\text{gcl}} = -\frac{1}{M_l}\sum_{i=1}^{M_l}\log\frac{\exp\left(\tilde{y}_i \cdot \rho_{ic}/\tau\right)}{\sum_{j=1}^{M_l}\mathbb{1}_{[j\neq i]}\exp\left(\tilde{y}_i \cdot \rho_{jc'}/\tau\right)}, \quad (5)$$

where $M_l = M + M_p$ and M_p is the number of the selected pseudo labels (defined in Eq. 3), and $c^{'}$ is the class different with class c. Therefore, negative samples are from the prototype of $C-1$ classes except the c-th class. Note that, if the dataset has C classes, this is essentially equivalent to a negative size of $C-1$. This is practically important when dealing with dataset with large number of classes. Thus, $\mathcal{L}_{gcl}$ can retain the feature learning property of $\mathcal{L}_{cra}$ in Eq. 2 mostly as well as resolve the memory bottleneck issue.

3.3. Dynamic point cloud augmentor

Data augmentation is an essential component in the proposed HybridCR, which generates varied anchors, positive and negative examples, and extract invariant representations by adding the particular noise in the input. Inspired by [15], we use MLPs and Gaussian noises to implement the learnable dynamic point cloud augmentor, which enrichs the data diversity in context-wise displacement and generate different transformations in the same scene.

Fig. 3 presents the proposed augmentor architecture. First, we use shared 4-layer MLPs with progressive dimensions of $[64, 128, 1024, 512]$ to extract $F \in \mathbb{R}^{N\times d}$. Then, two separate linear projection layers compute H and G. We regress the augmentation function specific to input sample $\mathcal{D}$ using two separate components in the architecture: (1) global-wises regression to produce transformation $\mathcal{M} \in \mathbb{R}^{N\times N}$, and (2) context-wise regression to produce displacement $\mathcal{S} \in \mathbb{R}^{N\times 3}$. In particular, we introduce two d dimension noise vectors based on a Gaussian distribution and concatenate them with H and G. Then, we employ MLPs to obtain $\mathcal{M}$ and $\mathcal{S}$. Note that the noise vectors enable the augmentor to explore more diverse choices in regressing the transformation matrix. Using $\mathcal{M}$ and $\mathcal{S}$, we then generate the augmented sample $\mathcal{D}' = \mathcal{D}\cdot\mathcal{M} + \mathcal{S}$. The proposed dynamic point cloud augmentor is more flexible

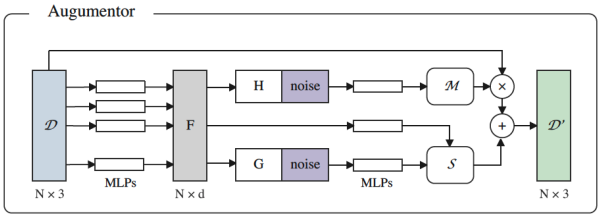


Figure 3. The architecture of dynamic point cloud augmentor.

to the traditional augmentor adopted in [33, 38] with jointly optimizing during training.

3.4. Overall objective.

As discussed above, HybridCR could serve as the effective contrastive regularization strategy for weakly-supervised point cloud semantic segmentation framework in end-to-end training scheme. The overall objective of network is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{con} + \mathcal{L}_{seg} + \lambda(\mathcal{L}_{cra} + \mathcal{L}_{lcl} + \mathcal{L}_{gcl}), \quad (6)$$

where λ is a balance parameter. $\mathcal{L}_{seg}$ is the cross-entropy based segmentation loss on the labeled points, which is formulated as:

$$\mathcal{L}_{seg} = -\frac{1}{CM}\sum_{i=1}^{M}\sum_{c=1}^{C} y_{ic}\log\frac{\exp\left(\tilde{y}_{ic}\right)}{\sum_{c=1}^{C}\exp\left(\tilde{y}_{ic}\right)}, \quad (7)$$

where y_{ic} denotes the ground truth label of point x_i. We also apply Eq. 7 into the augmentation data to learn the network parameter $\boldsymbol{\theta}$. We solve Eq. 6 by Adam optimizer. Further, HybridCR can serves as an effective auxiliary feature learning loss when expanded to fully-supervised manner.

4. Experiments

4.1. Experiment setting

Experimental datasets contains S3DIS [1], ScanNet-V2 [6], Semantic3D [8] and SemanticKITTI [2]. **S3DIS** is a commonly-used indoor 3D point cloud dataset for semantic segmentation. It has 271 point cloud scenes across 6 areas with 13 classes. **ScanNet-V2** is also an indoor 3D point cloud datasets, which contains 1,613 3D scans with the total of 20 classes. The whole data is split into a training set (1201 scans), a validation set (312 scans), and a testing set (100 scans). **Semantic3D** is an outdoor dataset that provides a large-scale labeled 3D point cloud with over 4 billion points. It covers a range of diverse urban scenes, and the raw 3D points have 8 classes with multiple information, such as 3D coordinates, RGB information, and intensity. **SemanticKITTI** is a large-scale outdoor point cloud dataset for 3D semantic segmentation in an autonomous driving scenario and has 19 classes. The dataset contains 22 sequences that are divided into a training set (10 sequences

Settings	Methods	mIoU(%)	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clutter
Fully	PointNet [21]	41.1	88.8	97.3	69.8	0.1	4.0	46.3	10.8	58.9	52.6	5.9	40.3	26.4	33.2
	KPConv [29]	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
	RandLA-Net [10]	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
	RFCR [7]	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
	PSD [38]	65.1	92.3	97.1	80.7	0.0	32.4	55.5	68.1	78.9	86.8	71.1	70.6	59.0	53.0
	HybridCR	65.8	93.6	98.1	82.3	0.0	24.4	59.5	66.9	79.6	87.9	67.1	73.0	66.8	55.7
10%	Xu *et al.* [33]	48.0	90.9	97.3	74.8	0.0	8.4	49.3	27.3	69.0	71.7	16.5	53.2	23.3	42.8
1%	Zhang *et al.* [37]	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 [38]	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	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
1pt(0.2%)	Π Model [25]	44.3	89.1	97.0	71.5	0.0	3.6	43.2	27.4	62.1	63.1	14.7	43.7	24.0	36.7
	MT [28]	44.4	88.9	96.8	70.1	0.1	3.0	44.3	28.8	63.6	63.7	15.5	43.7	23.0	35.8
	Xu *et al.* [33]	44.5	90.1	97.1	71.9	0.0	1.9	47.2	29.3	62.9	64.0	15.9	42.2	18.9	37.5
1pt(0.03%)	RandlA-Net [10]	40.7	83.7	90.7	61.2	0.0	11.9	40.8	15.2	52.0	51.7	14.9	50.5	25.3	31.8
	PSD [38]	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	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

Table 1. Quantitative results on Area-5 of S3DIS. "*" denotes the results of the method trained by us using the official code. Note that our 1pt denotes only one labeled point for each class in the entire rooms instead of small blocks (e.g., 1×1 meter) of Xu *et al.* [33]. The number of labeled points in our 1pt setting accounts for 0.03% of the total points, which is about 0.2% in Xu *et al.* [33].

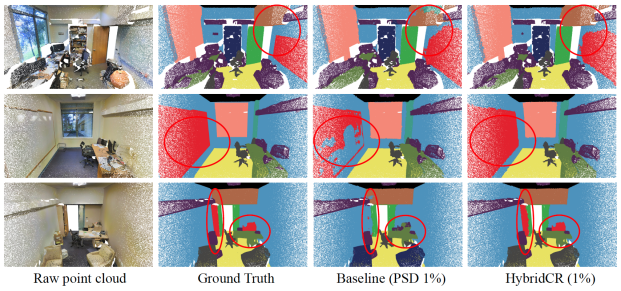


Figure 4. Visualization results on the test set of S3DIS Area-5. Raw point cloud, semantic labels, results of the baseline and ours are presented separately from left to right.

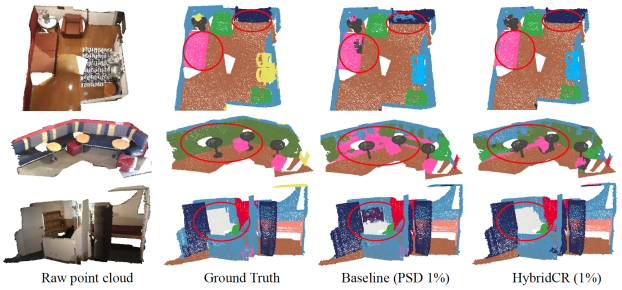


Figure 5. Visualization results on the validation set of ScanNet-V2. Raw point cloud, semantic labels, results of the baseline and ours are presented separately from left to right.

with ∼19k frames), a validation set (1 sequence with ∼4k frames), and a testing set (11 sequences with ∼20k frames).

Implementation details. We use Adam Optimizer with an initial learning rate of 0.001 and momentum of 0.9 to train 100 epochs for all datasets on an NVIDIA RTX Titan GPU. The number of neighbor points K is 16, the batch size is 6, the initial learning rate is 0.01 with the decay rate 0.98, and the iteration steps for each epoch are set to 500. Note that we choose point-based backbone PSD [38] as our baseline due to its effectiveness and efficiency.

Evaluation Protocols. We evaluate the final performance on all points of the original test set. For quantitative comparison, we use the mean Intersection-over-Union (mIoU) as the standard metrics. We experimentally study two types of weak labels: 1pt and 1% settings. Further, we extend HybridCR to the fully-supervised manner.

4.2. Comparison with the SOTA Methods

Quantitative Results on S3DIS and ScanNet-V2. First, we compare HybridCR with the SOTA methods on S3DIS Area-5, whose quantitative results are summarized in Tab. 1. Obviously, the proposed HybridCR achieves the highest mIoU in the settings of 1pt and 1%, compared to Zhang *et al.* [37], PSD [38], Π Model [25], MT [28], Xu *et al.* [33] and RandLA-Net [10]. For example, our method outperforms PSD and RandLA-Net by 3.3% and 10.8% at the setting of 1pt(0.03%), respectively. Moreover, our method also achieves 7.0% performance gains over Xu *et al.* [33], which utilizes the more labeled points about 0.2%. In the aspect of specific class at the setting of 1pt(0.03%), our method significantly improve the performance with 8.7%, 16.4% and 8.9% improvements in "chair", "table", and "sofa" against PSD, respectively.

For the setting of 1%, our method achieves 1.8% mIoU gains over the PSD baseline, and even surpasses Xu *et*

Set.	Method	S3DIS	ScanNet-V2		Sem3D.	SemKitti.	
		6-fold	val	test	test	val.	test
Fully	PointCNN [16]	65.4	-	45.8	-	-	-
	DGCNN [30]	56.1	-	-	-	-	-
	ShellNet [39]	66.8	-	-	69.4	-	-
	PointASNL [34]	68.7	66.4	63.0	-	-	46.8
	KPConv [29]	70.6	69.2	68.4	74.6	-	58.8
	RandLA-Net [10]	70.0	-	57.8*	77.4	-	53.9
	RFCR [7]	70.9*	-	70.2	77.8	-	-
	PSD [38]	70.3*	-	-	-	-	-
	HybridCR	70.7	59.5	59.9	77.4	53.2	54.0
sub.	WyPR [23]	-	31.1	24.0	-	-	-
	MPRM [31]	-	43.2	41.1	-	-	-
1%	Zhang *et al*. [37]	65.9	-	51.1	72.6	-	-
	PSD [38]	68.0	-	54.7	75.8	-	-
	HybridCR	69.2	56.9	56.8	76.8	51.9	52.3

Table 2. Quantitative results (mIoU(%) on S3DIS 6-fold, ScanNet-V2 validation set, Semantic3D (reduced-8) and SemanticKITTI validation set, with fully labeled data and 1% labeled data. Particularly, in experiments with 100% labeled data, our hybrid contrastive loss serves as an auxiliary feature learning loss. "*" denotes the results of the method trained by us using the official code.

al. [33] at the setting of 10%. To explain, our method learns diverse geometry structures from the large-scale point cloud data by adding the proposed hydrid contrastive regularization. Based on that, our method only uses the 1% points to outperform the fully supervised RandLA-Net and PSD. For a fair comparison, we also expand the comparison with other methods on S3DIS at the 6-fold setting, whose results are presented in Tab. 2. For ScanNet-V2, compared to WyPR [23] and MPRM [31] based on scene/subcloud-level annotation, HybridCR achieves the highest mIoU of 56.8% at 1% setting on test set. Meanwhile, HybridCR achieves 5.7% mIoU gains over Zhang *et al*. at the same number of label annotation. Besides, our method achieves 2.1% mIoU gains over RandLA-Net at fully supervised situation.

Qualitative Results on S3DIS and ScanNet-V2. We present the qualitative results of S3DIS and ScanNet-V2 in Fig. 4 and Fig. 5, respectively. On S3DIS, HybridCR achieves better segmentation on "board" and "chair" compared to PSD. Moreover, the segmentation results of HybridCR are very consistency to the ground-truth. On ScanNet-V2, we observe that HybridCR achieves good and truthfully segmentation results. On ScanNet-V2, HybridCR achieves good performance on " sofa" and "desk" compared to PSD. The reason could be that HybridCR can effectively leverage the diverse transformations generated by dynamic point cloud augmentor to improve the representation ability and promote segmentation performance.

Quantitative Results on Semantic3D and SemanticKITTI. We further evaluate HybridCR on outdoor large-scale point cloud datasets Semantic3D (reduced-8) and SemanticKITTI and present the results in Tab. 2,

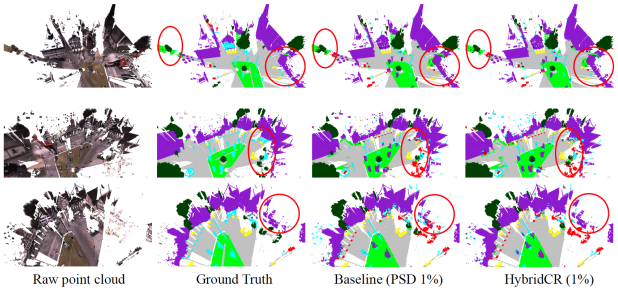


Figure 6. Visualizations on validation set of Semantic3D. Raw point cloud, semantic labels, results of the baseline and ours are presented separately from left to right.

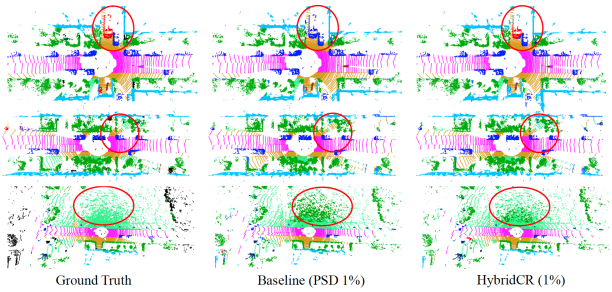
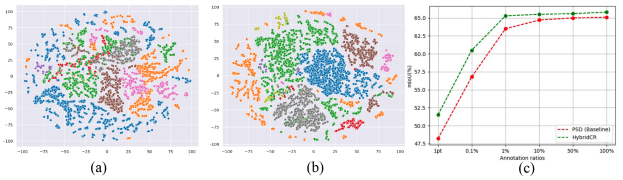


Figure 7. Visualization results on the validation set of SemanticKITTI. Semantic labels, results of the baseline and ours are presented separately from left to right.

respectively. For Semantic3D, our method also achieves a better performance with 4.2% and 1.0% mIoU improvements at the setting of 1% compared to Zhang *et al*. [37] and PSD. For SemanticKITTI, our method reports the results as 51.9% and 52.3% on validation and test dataset at the setting of 1%. It can be seen that our method surpasses other point-based approaches by a large margin with limited annotations.

Qualitative Results on Semantic3D and SemanticKITTI. We give the qualitative results of Semantic3D and SemanticKITTI in Fig. 6 and Fig. 7, respectively. On Semantic3D, our method improvement over the PSD, especially achieves precisely segmentation on "buildings". On SemanticKITTI, it can be seen that our method achieves consistency segmentation results to ground-truth, especially in "road" and "car", which are difficult to distinguish while critical on sparse outdoor scenes in the auto-driving application. The results demonstrate the effectiveness of our method on outdoor datasets.

Results on fully supervised settings. We further expand the comparison with current SOTA methods on fully-supervised setting in both indoor and outdoor datasets, whose quantitative results are summarized in Tab. 2. It can be observed that HybridCR is competitive among them. *e.g*., HybridCR surpass RandLA-Net with 0.7% and 2.1% mIoU improvements on S3DIS and ScanNet-V2, respectively, and gains 0.1% mIoU improvement on SemanticKITTI. Moreover, HybridCR outperforms KPConv by 1.8% in mIoU on Semantic3D.



(a) (b) (c)

Figure 8. Visualization of point embedding at 1% setting. (a) is the embedding of PSD, (b) is the embedding of HybridCR. The scene is randomly selected from the test set of S3DIS. (c) is the relation between the number of labeled points and performances.

4.3. Ablation Study

We further evaluate the effectiveness of the essential components for ablation study, including dynamic point cloud augmentor and local/global guidance contrastive regularization. All experiments are conducted on the S3DIS Area-5 and results are shown in Tab. 3. Note that, the #1 is reported by PSD, the #8 is reported by HybridCR, we report the results with mean and std.dev.(5 runs).

Effectiveness of dynamic data augmentor. To verify that the improvement caused by the data augmentation, we compare the *Base.* with the *Aug.*. Comparing #1 and #2 at 1pt and 1% setting, it achieves 2.5% and 1.0% gains over Base., respectively. For #5 and #8 at 1pt and 1% setting, it achieves 0.4% and 0.3% gains over HybridCR, respectively. The results indicate the HybridCR gains much benefits from the *Aug.* with the diverse transformations.

Effectiveness of local guidance contrastive loss. From the comparison between #1 and #3 at 1pt and 1% setting, it outperforms 1.6% and 0.4% in mIoU over *Base.*, respectively. For #7 and #8, it gains 0.5% and 0.2% improvements over HybridCR, respectively. These results show that the *Local.* further improves the performances because it leverages the neighboring information during model training while enhancing the feature learning.

Effectiveness of global guidance contrastive loss. Similarly, from the comparison of #1 and #4, it outperforms *Base.* by 2.0% and 0.5% at the setting of 1pt and 1% setting, respectively. For #6 and #8, it achieves 1.3% and 0.6% gains over HybridCR, respectively. The results demonstrate that the *Global.* effectively improves the performances of the weakly-supervised semantic segmentation task with classes prototypes.

4.4. Analysis

Visualization of point embedding. As shown in Fig. 8 (a) and (b), compared with PSD, the learned point embeddings of HybridCR become more compact and separate. It suggests that the segmentation network generates more discriminative features and produce promising results by enjoying the advantage of local and global guidance contrastive losses and the effective transformations generated by dynamic point cloud augmentor.

Labeled points and the performance. We further discuss the relationship between performances and label ratios {1pt, 0.1%, 1%, 10%, 50%, 100%} in Fig. 8 (c). With the increase of ratios, the performances of the two methods are improved, and the growth trend is gradually slowing down. Note that the performances decreases marginally with the ratio less than 1%, which indicates that keeping a certain amount of supervising signal is essential. Moreover, the performance at the ratio 10% is near to 100%, which shows that the dense annotations are unnecessary to obtain favorable segmentation results.

	Base.	Aug.	Local.	Global.	1pt	1%
#1	✓				48.2±(0.3)	63.5±(0.1)
#2	✓	✓			50.7±(0.3)	64.5±(0.3)
#3	✓		✓		49.8±(0.5)	63.9±(0.4)
#4	✓			✓	50.2±(0.2)	64.0±(0.2)
#5	✓		✓	✓	51.1±(0.2)	65.0±(0.3)
#6	✓	✓	✓		50.8±(0.3)	64.7±(0.4)
#7	✓	✓		✓	51.0±(0.1)	65.1±(0.2)
#8	✓	✓	✓	✓	51.5±(0.4)	65.3±(0.3)

Table 3. Ablations of different components on Area-5 of S3DIS.

5. Conclusion

In this paper, we propose a hybrid contrastive regularization framework for weakly supervised large-scale point cloud semantic segmentation. With our proposed local and global guidance contrastive regularization, the network learns more discriminative features by leveraging the neighboring points and the pseudo-labels. Meanwhile, we propose a dynamic point cloud augmentor to benefit contrastive strategy with more diverse transformations with jointly optimizing during training. Extensive experimental results on indoor and outdoor dataset demonstrate that HybridCR achieves significant gains compared with the SOTA methods. Moreover, the effectiveness of the introduced key components is verified by ablation studies. The results further demonstrate our method's effectiveness in exploiting limited labeled large-scale point clouds and improving the model generalization ability.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2019YFC1521104), National Natural Science Foundation of China (No.72192821, No.61972157, No.62102151, No.62176092), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0102), Shanghai Science and Technology Commission (No.22YF1420300, No.21511101200, No.21511100700, No.21YF1411200), Shaohui Lin is Sponsored by CAAI-Huawei MindSpore Open Fund (No.CAAIXSJLJJ-2021-031A).

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016. 2, 5

[2] 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 *ICCV*, pages 9297–9307, 2019. 2, 5

[3] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *ECCV*, pages 330–345. Springer, 2020. 3

[4] Mingmei Cheng, Le Hui, Jin Xie, and Jian Yang. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In *AAAI*, volume 35, pages 1140–1147, 2021. 1, 2

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 1

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 5

[7] 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 *CVPR*, pages 11673–11682, 2021. 6, 7

[8] T Hackel, N Savinov, L Ladicky, JD Wegner, K Schindler, and M Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:91, 2017. 2, 5

[9] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, pages 15587–15597, 2021. 1, 2

[10] 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 *CVPR*, pages 11108–11117, 2020. 6, 7

[11] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, pages 5070–5079, 2019. 2

[12] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, pages 6423–6432, 2021. 2

[13] Sihyeon Kim, Sanghyeok Lee, Dasol Hwang, Jaewon Lee, Seong Jae Hwang, and Hyunwoo J Kim. Point cloud augmentation with weighted local transformations. In *ICCV*, pages 548–557, 2021. 3

[14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2, 4

[15] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Pointaugment: an auto-augmentation framework for point cloud classification. In *CVPR*, pages 6378–6387, 2020. 2, 3, 5

[16] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *NeurIPS*, 31:820–830, 2018. 7

[17] Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas Funkhouser, and Hao Dong. P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. *arXiv preprint arXiv:2012.13089*, 2020. 4

[18] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *CVPR*, pages 1726–1736, 2021. 1, 2

[19] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, 2006. 2

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1, 6

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

[23] Zhongzheng Ren, Ishan Misra, Alexander G Schwing, and Rohit Girdhar. 3d spatial recognition without spatially labeled 3d. In *CVPR*, pages 13204–13213, 2021. 7

[24] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2020. 2

[25] Laine Samuli and Aila Timo. Temporal ensembling for semi-supervised learning. In *ICLR*, volume 4, page 6, 2017. 6

[26] Xian Shi, Xun Xu, Ke Chen, Lile Cai, Chuan Sheng Foo, and Kui Jia. Label-efficient point cloud semantic segmentation: An active learning approach. *arXiv preprint arXiv:2101.06931*, 2021. 2

[27] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33, 2020. 2, 4

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 2, 6

[29] 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 *ICCV*, pages 6411–6420, 2019. 1, 6, 7

[30] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 2019. 7

[31] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly super-

vised 3d semantic segmentation on point clouds. In *CVPR*, pages 4384–4393, 2020. 2, 7

[32] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, pages 574–591. Springer, 2020. 1, 2, 4

[33] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *CVPR*, pages 13706–13715, 2020. 1, 2, 3, 5, 6, 7

[34] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, pages 5589–5598, 2020. 7

[35] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 2

[36] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[37] Yachao Zhang, Zonghao Li, Yuan Xie, Yanyun Qu, Cuihua Li, and Tao Mei. Weakly supervised semantic segmentation for large-scale point cloud. In *AAAI*, volume 35, pages 3421–3429, 2021. 1, 2, 6, 7

[38] 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 *ICCV*, pages 15520–15528, 2021. 1, 2, 3, 5, 6, 7

[39] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In *ICCV*, pages 1607–1616, 2019. 7

[40] Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Time-consistent self-supervision for semi-supervised learning. In *ICML*, pages 11523–11533. PMLR, 2020. 2