Weakly Supervised Semantic Point Cloud Segmentation: Towards 10× Fewer Labels

Xun Xu* Gim Hee Lee
Department of Computer Science, National University of Singapore
alex.xun.xu@gmail.com, gimhee.lee@nus.edu.sg

Abstract

Point cloud analysis has received much attention recently; and segmentation is one of the most important tasks. The success of existing approaches is attributed to deep network design and large amount of labelled training data, where the latter is assumed to be always available. However, obtaining 3d point cloud segmentation labels is often very costly in practice. In this work, we propose a weakly supervised point cloud segmentation approach which requires only a tiny fraction of points to be labelled in the training stage. This is made possible by learning gradient approximation and exploitation of additional spatial and color smoothness constraints. Experiments are done on three public datasets with different degrees of weak supervision. In particular, our proposed method can produce results that are close to and sometimes even better than its fully supervised counterpart with 10× fewer labels. Our code is available at the project website[1].

1. Introduction

Recent developments in point cloud data research have witnessed the emergence of many supervised approaches [19, 20, 33, 12, 29]. Most efforts of current research are dedicated into two tasks: point cloud shape classification (a.k.a. shape recognition) and point cloud segmentation (a.k.a. semantic segmentation). For both tasks, the success of the state-of-the-art methods is attributed mostly to the deep learning architecture [19] and the availability of large amount of labelled 3d point cloud data [16, 1]. Although the community is still focused on pushing forward in the former direction, we believe the latter issue, i.e. data annotation, is an overlooked bottleneck. In particular, it is assumed that all points for the point cloud segmentation task are provided with ground-truth labels, which is often in the range of 1k to 10k points for a 3d shape [34, 16]. The order of magnitude increases drastically to millions of points for a real indoor scene [11]. As a result, very accurate labels for billions of points are needed in a dataset to train good segmentation models. Despite the developments of modern annotation toolkits [16, 1] to facilitate large-scale annotation, exhaustive labelling is still prohibitively expensive for ever growing new datasets.

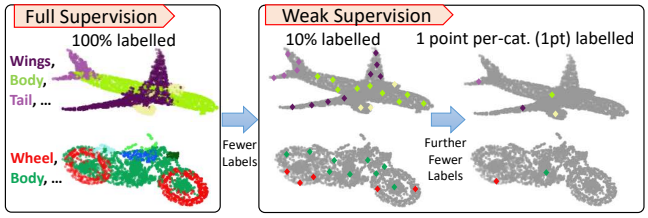


Figure 1: Illustration of the weak supervision concept in this work. Our approach achieves segmentation with only a fraction of labelled points. .

In this work, we raise the question on whether it is possible to learn a point cloud segmentation model with only partially labelled points. And, if so, how many is enough for good segmentation. This problem is often referred to as weakly supervised learning in the literature [37] as illustrated in Fig. 1. To the best of our knowledge, there are only a handful of works which tried to address related problems [6, 14]. In [6], a non-parametric conditional random field classifier (CRF) is proposed to capture the geometric structure for weakly supervised segmentation. However, it casts the task into a pure structured optimization problem, and thus fail to capture the context, e.g. spatial and color cues. A method for semi-supervised 3D LiDAR data segmentation is proposed in [14]. It converts 3D points into a depth map with CNNs applied for feature learning, and the semi-supervised constraint is generated from the temporal consistency of the LiDAR scans. Consequently, it is not applicable to general 3D point cloud segmentation.

To enable the weakly supervised segmentation with both strong contextual modelling ability and handling generic 3D point cloud data, we choose to build upon the state-of-the-art deep neural networks for learning point cloud feature embedding [19, 33]. Given partially labelled point cloud

*now with A-STAR, Singapore.
[1] https://github.com/alex-xun-xu/WeakSupPointCloudSeg

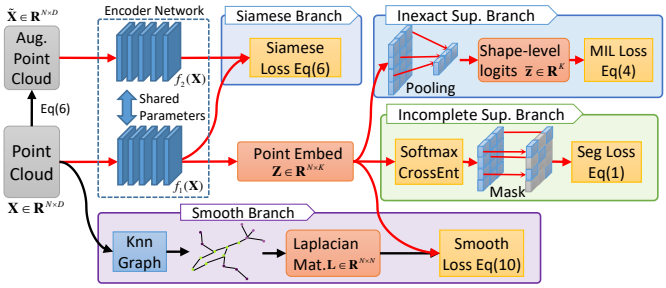


Figure 2: Our network architecture for weakly supervised point cloud segmentation. Red lines indicate back propagation flow.

data, we employ an incomplete supervision branch with softmax cross-entropy loss that penalizes only on labelled points. We observe that such simple strategy can succeed even at $10\times$ fewer labels, i.e. only 10% of the points are labelled. This is because the learning gradient of the incomplete supervision can be considered as a sampling approximation of the full supervision. In Sect. 3.2, we show our analysis that the approximated gradient converges to the true gradient in distribution, and the gap is subjected to a normal distribution with variance inversely proportional to the number of sampled points. As a result, the approximated gradient is close to the true gradient given enough labelled points. The analysis also gives an insight into choosing the best annotation strategy under fixed budget. We conclude that it is always better to extensively annotate more samples with fewer labelled points in each sample than to intensively label fewer samples with more (or fully) labelled points.

As the above method imposes constraints only on the labelled points, we propose additional constraints to the unlabelled points in three orthogonal directions. First, we introduce an additional inexact supervision branch which defines a point cloud sample level cross entropy loss in a similar way to multi-instance learning[35, 7]. It aims to suppress the activation of any point with respect to the negative categories. Second, we introduce a Siamese self-supervision branch by augmenting the training sample with a random in-plane rotation and flipping, and then encourage the original and augmented point-wise predictions to be consistent. Finally, we make the observation that semantic parts/objects are often continuous in the spatial and color spaces. To this end, we propose a spatial and color smoothness constraint to encourage spatially adjacent points with similar color to have the same prediction. Such constraint can be further applied at inference stage by solving a soft constrained optimization that resembles label propagation on a graph [38]. Our proposed network is illustrated in Fig. 2.

Our contributions are fourfold. i) To the best of our knowledge, this is the first work to investigate weakly supervised point cloud segmentation within a deep learning context. ii) We give an explanation to the success of weak supervision and provide insight into annotation strategy under a fixed labelling budget. iii) We adopt three additional losses based on inexact supervision, self-supervision and spatial and color smoothness to further constrain unlabelled data. iv) Experiments are carried out on three public dataset which serve as benchmarks to encourage future research.

2. Related Work

Weakly supervised learning aims to use weaker annotations, often in the form of partially labelled dataset or samples. In this work, we follow the definition of weak supervision made by [37]. More specifically, we are concern with two types of weak supervision: incomplete and inexact supervision.

Incomplete Supervision. This is also referred to as semi-supervised learning in the literature [38, 3, 17, 2, 10, 27, 8]. We interchangeably use semi-supervised, weakly supervised and weak supervision in this paper to refer to this type of supervision. It is assumed that only partial instances are labelled, e.g. only a few images are labelled for the recognition task [38, 36, 8], a few bounding boxes or pixels are labelled for the image segmentation task [17, 2] or a few nodes are labelled for graph inference [27]. The success is often attributed to the exploitation of problem specific assumptions including graph manifold [38, 3, 27], spatial and color continuity [17, 2], etc. Another line of works are based on ensemble learning by introducing additional constraints such as consistency between original and altered data, e.g. the addition of noise [22], rotation [10] or adversarial training [15]. This has further inspired ensemble approaches [25, 21] akin to data distillation. Up till now, most of these works emphasize on large-scale image data, while very limited works have addressed point cloud data. [14] proposes a semi-supervised framework for point cloud segmentation. However, it does not directly learn from point cloud data and the required annotation is quite large. [6] proposes to exploit the geometric homogeneity and formulated a CRF-like inference framework. Nonetheless, it is purely optimization-based, and thus fails to capture the spatial relation between semantic labels. In this work, we make use of the state-of-the-art deep neural networks, and incorporate additional spatial constraints to further regularize the model. Thus we take advantage of both spatial correlation provided by deep models and geometric priors.

Inexact Supervision. It is also referred as weakly annotation in the image segmentation community [9, 24]. They aim to infer the per-pixel prediction from a per-image level annotation [9, 24] for image segmentation tasks. The class activation map (CAM) [35] is proposed to highlight the attention of of CNN based on discriminative supervision. It is proven to be a good prior model for weakly supervised segmentation [9, 32]. Inexact supervision is often complementary to incomplete supervision, and therefore, it is also used

to improve semi-supervised image segmentation[2]. In this work, we introduce inexact supervision as a complement to incomplete supervision for the task of point cloud segmentation.

Point Cloud Analysis. It is applied on 3D shapes and has received much attention in recent years. The PointNet [19] is initially proposed to learn 3D point cloud feature through cascaded multi-layer perceptrons (mlps) for point cloud classification and segmentation. The following works [20, 33, 12, 30, 11] are subsequently proposed to exploit local geometry through local pooling or graph convolution. Among all tasks of point cloud analysis, semantic segmentation is of high importance due to its potential application in robotics and the existing works rely on learning classifiers at point-level [19]. However, this paradigm requires exhaustive point-level labelling and does not scale well. To resolve this issue, we propose a weakly supervised approach that requires only a fraction of points to be labelled. We also note that [26] proposes to add spatial smoothness regularization to the training objective. [5] proposes to refine prediction via CRF. Nevertheless, both works require full supervision, while our work is based on a more challenging weak supervision setting.

3. Methodology

3.1. Point Cloud Encoder Network

We formally denote the input point cloud data as $\{\mathbf{X}_b\}_{b=1\cdots B}$ with B individual shapes (e.g. shape segmentation) or room blocks (e.g. indoor point cloud segmentation). Each sample $\mathbf{X}_b \in \mathcal{R}^{N\times F}$ consists of N 3d points with the xyz coordinates and possibly additional features, e.g. RGB values. Each sample is further accompanied with per-point segmentation label $\mathbf{y}_b \in \{1,\cdots K\}^N$, e.g. fuselage, wing and engine of a plane. For clarity , we denote the one-hot encoded label as $\hat{\mathbf{Y}} \in \{0,1\}^{B\times N\times K}$. A point cloud encoder network $f(\mathbf{X};\Theta)$ parameterized by Θ is employed to obtain the embedded point cloud features $\mathbf{Z}_b \in \mathcal{R}^{N\times K}$. We note that the dimension of the embedding is the same as the number of segmentation categories. The recent development on point cloud deep learning [19, 20, 12] provides many candidate encoder networks, which are evaluated in the experiment section.

3.2. Incomplete Supervision Branch

We assume that only a few points in the point cloud samples $\{\mathbf{X}_b\}$ are labelled with the ground-truth. Specifically, we denote a binary mask as $\mathbf{M} \in \{0,1\}^{B\times N}$, which is 1 for a labelled point and 0 otherwise. Furthermore, we define a softmax cross-entropy loss on the labelled point as

$$l_{seg} = -\frac{1}{C}\sum_b\sum_i m_{bi}\sum_k \hat{y}_{bik}\log\frac{\exp(z_{bik})}{\sum_k\exp(z_{bik})}, \quad (1)$$

where $C = \sum_{b,i} m_{bi} = ||\mathbf{M}||_1$ is the normalization variable.

Discussion: According to the experiments, we found that our method yields competitive results with as few as 10% labelled points, i.e. $||\mathbf{M}||_1/(B\cdot N) = 0.1$. The rationale is detailed in the following. We first assume that two networks with similar weights – one trained with full supervision and the other with weak supervision should produce similar results. Assuming that both networks start with an identical initialization, the higher similarity of the gradients at each step means a higher chance for the two networks to converge to similar results. Now, we write the gradients with full supervision $\nabla_\Theta l_f$ and weak supervision $\nabla_\Theta l_w$ as

$$\begin{aligned}\nabla_\Theta l_f &= \frac{1}{B\cdot N}\sum_b\sum_i\sum_k \nabla_\Theta l_{bik}, \quad \text{and}\\ \nabla_\Theta l_w &= \frac{1}{C}\sum_b\sum_i m_{bi}\sum_k \nabla_\Theta l_{bik}, \quad (2)\\ \text{where} \quad l_{bik} &= -\hat{y}_{bik}\log\frac{\exp(z_{bik})}{\sum_k\exp(z_{bik})}.\end{aligned}$$

This relation is also illustrated in Fig. 3.

At each training step, the direction of the learning gradient is the mean of the gradients calculated with respect to each individual point. Suppose that $\nabla_\Theta l_{bik}$ is i.i.d. with expectation $E[\nabla_\Theta l_{bik}] = \mu$ and variance $Var[\nabla_\Theta l_{bik}] = \sigma^2$, and sampled mean (n samples) $S_n = mean(\nabla_\Theta l_{bik})$. We can easily verify that $E[\nabla_\Theta l_{bik}] = \nabla_\Theta l_f$ and $S_n = \nabla_\Theta l_w$ with $n = C = ||\mathbf{M}||_1$. According to the Central Limit Theorem, we have the following convergence in distribution:

$$\begin{aligned}&\sqrt{n}(S_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2),\\ \Rightarrow&\sqrt{||\mathbf{M}||_1}(\nabla_\Theta l_w - \nabla_\Theta l_f) \xrightarrow{d} \mathcal{N}(0,\sigma^2), \quad (3)\\ \Rightarrow&(\nabla_\Theta l_w - \nabla_\Theta l_f) \xrightarrow{d} \mathcal{N}(0,\sigma^2/||\mathbf{M}||_1).\end{aligned}$$

This basically indicates that the difference between the gradient of full supervision and weak supervision is subjected to a normal distribution with variance $\sigma^2/||\mathbf{M}||_1$. Consequently, a sufficient number of labelled points, i.e. sufficiently large $||\mathbf{M}||_1$, is able to approximate $\nabla_\Theta l_f$ well

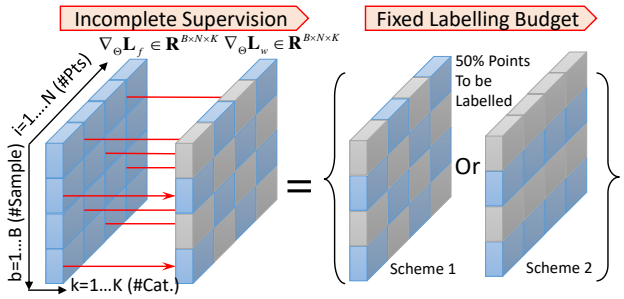


Figure 3: Illustration of incomplete supervision and labeling strategies with fixed budget.

with $\nabla_\Theta l_w$. Although the value of σ is hard to estimate in advance, we empirically found that our method yields results comparable to full supervision with $10\times$ fewer labelled points.

The analysis also provides additional insight into data annotation under a fixed budget. For example, with 50% of the total points to be labelled as illustrated in Fig. 3 (right): should we label 50% of the points in each sample (Scheme 1) or label all the points in only 50% of the samples (Scheme 2)? From the above analysis, it is apparent that Scheme 1 is better than Scheme 2 since it is closer to the i.i.d. assumption. This is further backed up by experiments in Sect. 4.4.

3.3. Inexact Supervision Branch

In addition to the Incomplete Supervision Branch, a so-called inexact supervision accompanies the annotation. Assuming each part has at least one labelled point, every training sample $\mathbf{X}_b$ is accompanied with an inexact label $\bar{\mathbf{y}}_b = \max_i \hat{\mathbf{y}}_{bi}$ simply by doing maxpooling over all points. Consequently, the inexact supervision branch is constructed in a similar fashion as multi-instance learning [18, 7]. The feature embedding $\mathbf{Z}_b$ is first globally maxpooled, i.e. $\bar{\mathbf{z}}_b = \max_i \mathbf{z}_{bi}$. We then introduce a loss for the inexact supervision branch. Since $\bar{\mathbf{z}}_b$ defines the logits on each category, the sigmoid cross entropy can be adopted as

$$l_{mil} = -\frac{1}{B \cdot K} \sum_b \sum_k \bar{y}_{bk} \log \frac{1}{1+\exp(-\bar{z}_{bk})} + (1-\bar{y}_{bk})(\log(\frac{\exp(-\bar{z}_{bk})}{1+\exp(-\bar{z}_{bk})})). \quad (4)$$

The rationale is that for those part categories that are absent from the sample, no points should be predicted with high logits. The incomplete supervision branch is only supervised on a tiny fraction of label points while the inexact supervision branch is supervised on the sample level with all points involved, so they are complementary to each other.

3.4. Siamese Self-Supervision

Despite the above two losses, majority of the unlabelled points are still not trained with any constraints. We believe additional constraints on those points can potentially further improve the results. To this end, we first introduce a Siamese self-supervision structure. We make the assumption that the prediction for any point is rotation and mirror flipping invariant. This assumption in particular holds true for 3D CAD shapes and indoor scenes with rotation in the XoY plane, e.g. the semantic label should not change with different view angle in a room. With this in mind, we design a Siamese network structure with two shared-parameter encoders $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$. Then given a training sample $\mathbf{X}$, we apply a random transformation that consists of a random mirroring along the X and/or Y axes and an XoY plane rotation, i.e.

$$\mathbf{R} = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} (2a-1)c & (2b-1)(1-c) & 0 \\ (2a-1)(1-c) & (2b-1)c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (5)$$

where $\theta \sim \mathcal{U}(0, 2\pi)$ (uniform distribution) and $a, b, c \sim \mathcal{B}(1, 0.5)$ (Bernoulli distribution). Specifically, the first matrix controls the degree of rotation and the second matrix controls mirroring and X,Y swapping. With the augmented sample denoted as $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{R}^\top$, the rotation invariant constraint is turned into minimizing the divergence between the probabilistic predictions of $g(f_1(\mathbf{X}))$ and $g(f_2(\tilde{\mathbf{X}}))$, where $g(\cdot)$ is the softmax function. We use L2 distance to measure the divergence:

$$l_{sia} = \frac{1}{B \cdot N \cdot K} \sum_b ||g(f_1(\mathbf{X}_b)) - g(f_2(\tilde{\mathbf{X}}_b))||_F^2, \quad (6)$$

and empirically found it to be better than KL-Divergence.

3.5. Spatial & Color Smoothness Constraint

Semantic labels for 3D shape or scenes are usually smooth in both spatial and color spaces. Although they can be included by the state-of-the-art convolution networks [29], explicit constraints are more beneficial in our context of weak supervision when the embedding of large amount of unlabelled points are not well constrained by the segmentation loss. Consequently, we introduce additional constraints at both training and inference stages.

Spatial & Color Manifold. A manifold can be defined on the point cloud to account for the local geometry and color by a graph. We denote the 3D coordinate channels and RGB channels, if any, as $\mathbf{X}^{\mathbf{xyz}}$ and $\mathbf{X}^{\mathbf{rgb}}$, respectively. To construct a graph for the manifold, we first compute the pairwise distance $\mathbf{P}_c \in \mathcal{R}^{N\times N}$ for channel c (xyz or rgb) as $p_{ij}^c = ||\mathbf{x}_i^c - \mathbf{x}_j^c||_2$, $\forall i, j \in \{1, \cdots N\}$. A k-nn graph can be then constructed by searching for the k nearest neighbors $NN_k(\mathbf{x})$ of each point, and the corresponding weight matrix $\mathbf{W}^c \in \mathcal{R}^{N\times N}$ is written as

$$w_{ij}^c = \begin{cases} \exp(-p_{ij}^c/\eta), & j \in NN_k(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases}, \forall i, j \in \{1, \cdots N\}. \quad (7)$$

We take the sum of both weight matrices as $w_{ij} = w_{ij}^{xyz} + w_{ij}^{rgb}$ $\forall i, j$ to produce a more reliable manifold when both xyz and rgb channels are available. This is reasonable since the xyz channel blurs the boundary and the rgb channel links faraway points, respectively. In case the manifold constructed on spatial distance and color contradicts the labelled ground-truth, we add additional must-link and must-not-link constraints [31] to $\mathbf{W}$ to strengthen the compliance to known annotations, i.e.

$$w_{ij} = \begin{cases} 1, & m_i, m_j = 1, y_i = y_j \\ -1, & m_i, m_j = 1, y_i \neq y_j \end{cases}. \quad (8)$$

We further write the Laplacian matrix [3] as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with the degree matrix denoted as $\mathbf{D} = diag(\mathbf{d})$ [28] and $d_i = \sum_j w_{ij}, \forall i \in \{1 \cdots N\}$.

Training Stage. We introduce a manifold regularizer [3] to encourage the feature embedding of each point to comply with the manifold obtained previously. More specifically, the prediction $f(\mathbf{x}_i)$ should stay close to $f(\mathbf{x}_j)$ if w_{ij} indicates high and stay unconstrained otherwise. Thus the regularizer is given by

$$\begin{aligned} l_{smo} &= \frac{1}{||\mathbf{W}||_0} \sum_i \sum_j w_{ij} ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2^2 \\ &= \frac{2}{||\mathbf{W}||_0} (\sum_i d_i f(\mathbf{x}_i)^\top f(\mathbf{x}_i) - \sum_i \sum_j w_{ij} f(\mathbf{x}_i)^\top f(\mathbf{x}_j)) \\ &= \frac{2}{||\mathbf{W}||_0} (tr(\mathbf{Z}^\top \mathbf{D} \mathbf{Z}) - tr(\mathbf{Z}^\top \mathbf{W} \mathbf{Z})) = \frac{2}{||\mathbf{W}||_0} tr(\mathbf{Z}^\top \mathbf{L} \mathbf{Z}), \end{aligned} \tag{9}$$

where $\mathbf{Z}$ is the prediction of all points.

Inference Stage. It is well known in image segmentation that the predictions of a CNN do not consider the boundaries well [4, 9] and CRF is often employed to refine the raw predictions. In weakly supervised point cloud segmentation, this issue exacerbates due to limited labels. To mitigate this problem, we introduce a semi-supervised label propagation procedure[38] to refine the predictions. Specifically, the refined predictions $\tilde{\mathbf{Z}}$ should comply with the spatial and color manifold defined by the Laplacian $\mathbf{L}$, and at the same time should not deviate too much from the network predictions $\mathbf{Z}$. We write the objective as

$$\begin{aligned} &\min_{\{\tilde{\mathbf{z}}\}} \sum_i \sum_j w_{ij} ||\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j||_2^2 + \gamma \sum_i ||\tilde{\mathbf{z}}_i - \mathbf{z}_i||_2^2, \\ \Longrightarrow &\min_{\tilde{\mathbf{Z}}} tr(\tilde{\mathbf{Z}}^\top \mathbf{L} \tilde{\mathbf{Z}}) + \gamma ||\tilde{\mathbf{Z}} - \mathbf{Z}||_F^2. \end{aligned} \tag{10}$$

A closed-form solution exists for the above optimization [38] and the final prediction for each point is simply obtained via

$$\begin{aligned} &\tilde{y}_i = \arg\max_k \tilde{z}_{ik}, \ \forall i \in \{1, \cdots N\}, \quad \text{where} \\ &\tilde{\mathbf{Z}} = \gamma(\gamma \mathbf{I} + \mathbf{L})^{-1} \mathbf{Z}. \end{aligned} \tag{11}$$

3.6. Training

The final training objective is the combination of all the above objectives, i.e. $l_{total} = l_{seg} + \lambda_1 l_{mil} + \lambda_2 l_{sia} + \lambda_3 l_{smo}$. We empirically set $\lambda_1, \lambda_2, \lambda_3 = 1$. The k-nn graph is selected as $k = 10$, $\eta = 1e3$, and γ in Eq. (11) is chosen to be 1. For efficient training, we first train the network with segmentation loss l_{seg} only for 100 epochs. Then the total loss l_{total} is trained for another 100 epochs. The default learning rate decay and batchnorm decay are preserved during the trainings of different encoder networks. The initial learning rate is fixed at $1e-3$ for all experiments and the batchsize varies from 5 to 32 for different dataset bounded by the GPU memory size. Our algorithm is summarized in Algo. 1.

4. Experiment

4.1. Dataset

We conduct experiments of our weakly supervised segmentation model on three benchmark point cloud datasets. **ShapeNet** [34] is a CAD model dataset with 16,881 shapes from 16 categories, each annotated with 50 parts. It is widely used as the benchmark for classification and segmentation evaluation. We propose a weakly supervised setting. For each training sample we randomly select a subset of points from each part to be labelled. We use the default evaluation protocol for comparison. **PartNet** [16] is proposed for more fine-grained point cloud learning. It consists of 24 unique shape categories with a total of 26,671 shapes. For the semantic segmentation task, it involves three levels of fine-grained annotation and we choose to evaluate at level 1. The incomplete weakly supervised setting is created in a similar way to ShapeNet, and we follow the original evaluation protocol. **S3DIS** [1] is proposed for indoor scene understanding. It consists of 6 areas each covering several rooms. Each room is scanned with RGBD sensors and is represented by point cloud with xyz coordinate and RGB value. For weakly supervised setting, we assume a subset of points are uniformly labelled within each room. The evaluation protocol on Area 5 as holdout is adopted.

4.2. Weakly Supervised Segmentation

Two weakly supervision settings are studied. i) 1 point label (1pt), we assume there is only 1 point within each category labelled with ground-truth. Less than 0.8% of total

Algorithm 1: Weakly Supervised Point Cloud Segmentation

input : Point Cloud $\{\mathbf{X}_b \in \mathcal{R}^{N \times D}\}$, Labels $\{\mathbf{y}_b \in \mathcal{Z}^N\}$
output: Segmentation Predictions $\{\tilde{\mathbf{y}}_b \in \mathcal{Z}^N\}$

```
/* Training Stage:                                   */
```
for $epoch \leftarrow 1$ **to** 100 **do**
  Train One Epoch: $\Theta = \Theta - \alpha \nabla_\Theta l_{seg}|_{\{\mathbf{X}_b\},\{\mathbf{y}_b\}}$;
for $epoch \leftarrow 1$ **to** 100 **do**
  `// Siamese Network`
  Sample $\phi \sim \mathcal{U}(0, 2\pi)$ and $a, b, c \sim \mathcal{B}(1, 0.5)$;
  Calculate $\mathbf{R}$ according to Eq. (5);
  Generate augmented sample $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{R}^\top$;
  `// Manifold Regularization`
  Construct Laplacian $\mathbf{L}$ according to Sect. 3.5;
  Train one epoch:
  $\Theta = \Theta - \alpha \nabla_\Theta l_{total}|_{\{\mathbf{X}_b\},\{\tilde{\mathbf{X}}_b\},\{\mathbf{y}_b\}}$;
```
/* Inference Stage:                                  */
```
Forward pass $\mathbf{Z}_b = f(\tilde{\mathbf{X}}_b; \Theta)$;
Obtain predictions $\{\tilde{\mathbf{y}}_b\}$ via Eq. (11);

points are labelled for ShapeNet under the 1pt scheme. For S3DIS, the total labelled points is less than 0.2%. ii) 10 percentage label (10%), we uniformly label 10% of all points for each training sample.

Encoder Network. We choose DGCNN [33] with default parameters as our encoder network due to its superior performance on benchmark shape segmentation and high training efficiency. However, as we point out in Sect. 5, the proposed weakly supervised methods are compatible with alternative encoder networks.

Comparisons. We compare against 3 sub-categories of methods. i) Fully supervised approaches (Ful.Sup.), including the state-of-the-art networks for point cloud segmentation. These methods serve as the upper bound of weakly supervised approaches. ii) Weakly supervised approaches (Weak Sup.), we implemented several generic weakly supervised methods and adapt them to point cloud segmentation tasks. In particular, the following methods are compared. The Π model [10] proposed to supervise on original input and the augmented input, but without the incomplete supervision on the augmented input. The mean teacher (MT) [25] model employs a temporal ensemble for semi-supervised learning. The baseline method is implemented with only the segmentation loss l_{seg} and DGCNN as encoder. Our final approach (Ours) is trained with the multi-task total loss l_{total} with label propagation in the inference stage. iii) Unsupervised approaches, these methods do not rely on any annotations but instead directly infer clusters from the spatial and color affinities. Specifically, we experiment with Kmeans and normalized cut spectral clustering[23] (Ncut). Both methods are provided with ground-truth number of parts.

Evaluation. For all datasets, we calculate the mean Intersect over Union (mIoU) for each test sample, and report the average mIoU over all samples (SampAvg) and all categories (CatAvg). For unsupervised methods, we find a best permutation between the prediction and ground-truth and then calculate the same mIoU metrics.

ShapeNet. We present the results in Tab. 1, where we make the following observations. Firstly, the weak supervision model produces very competitive results with only 1 labelled point per part category. The gap between full supervision and 1 point weak supervision is less than 12%. Secondly, we observe consistent improvement in the performance of segmentation with more labelled point from 1pt to 10%. Interestingly, the weak supervision model is comparable to full supervision even with 10% labelled points. Lastly, our proposed method that combines multiple losses and label propagation improves upon the baseline consistently, and outperforms alternative generic semi-supervised learning approaches and unsupervised clustering methods.

S3DIS. The results are presented in Tab. 2. We make observations in a similar way to ShapeNet. First, the 1pt weak supervision provides strong results. The results of our proposed multi-task model is only 1% lower than the fully supervised counterpart. Furthermore, the results of our method with only 10% labelled points is even slightly superior than the fully supervision. Finally, the results of our method consistently outperform both unsupervised and alternative weakly supervised methods.

PartNet. For the PartNet dataset, we report the average mIoU in Tab. 2. Details for each category is included in the supplementary. We also observe the same patterns from the results. The 1pt setting yields particularly strong results and our own variant outperforms all unsupervised and alternative weak supervision methods.

4.3. Qualitative Examples

We show qualitative examples of point cloud segmentation on all datasets and compare the segmentation quality. Firstly, we present the segmentation results on selected rooms from the S3DIS dataset in Fig. 4. From left to right we sequentially visualize the RGB view, ground-truth, fully supervised segmentation, weakly supervised baseline method and our final approach results. For weakly supervised methods, 10% training points are assumed to be labelled. We observe accurate segmentation of majority and continuous objects, e.g. wall, floor, table, chair and window. In particular, our proposed method is able to improve the baseline results substantially by smoothing out the noisy areas. Nonetheless, we observe some mistakes of our method at the boundaries between different objects. The segmentation results on ShapeNet are shown in Fig. 5. These examples again demonstrate the highly competitive performance by the weakly supervised approach. For both the plane and car categories, the results of the weak supervision are very close to the fully supervised ones.

4.4. Label More Points Or More Samples

Given a fixed annotation budget, e.g. the total number of labelled points, there are different combinations of labelling strategies to balance the amount of labelled samples and the amount of labelled points within each sample. In this experiment, we control these two variables and validate on ShapeNet segmentation with the PointNet encoder for efficient evaluation. We first restrict the fixed budget to be 10% of all training points. The labelling strategy is described by $x\%$ samples (Samp) each with $y\%$ labelled points (Pts) and $xy = 1000$ to satisfy the restriction. We evaluate 5 combinations and present the results in Tab. 3. The consistent improvement of mIoU with $x\%$ from 10% to 100% suggests that, given fixed total annotation budget, it is better to extensively label more samples each with fewer labelled points than intensively labelling a fraction of the dataset.

Table 1: mIoU (%) evaluation on ShapeNet dataset. The fully supervision (Ful. Sup.) methods are trained on 100% labelled points. Three levels of weak supervisions (1pt, 1% and 100%) are compared. Ours method consists of DGCNN as encoder net, MIL branch, Siamese branch, Smooth branch and Inference label propagation.

Setting		Model	CatAvg	SampAvg	Air.	Bag	Cap	Car	Chair	Ear.	Guitar	Knife	Lamp	Lap.	Motor.	Mug	Pistol	Rocket	Skate.	Table
Ful.Sup.		PointNet[19]	80.4	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
		PointNet++[20]	81.9	85.1	82.4	79.0	**87.7**	**77.3**	90.8	71.8	91.0	85.9	**83.7**	95.3	**71.6**	**94.1**	81.3	58.7	**76.4**	**82.6**
		DGCNN[33]	**82.3**	**85.1**	**84.2**	**83.7**	84.4	77.1	**90.9**	**78.5**	**91.5**	**87.3**	82.9	**96.0**	67.8	93.3	**82.6**	**59.7**	75.5	82.0
Unsup.		Kmeans	39.4	39.6	36.3	34.0	49.7	18.0	48.0	37.5	47.3	75.6	42.0	69.7	16.6	30.3	43.3	33.1	17.4	31.7
		Ncut[23]	43.5	43.2	41.0	38.0	53.4	20.0	52.1	41.1	52.1	83.5	46.1	77.5	18.0	33.5	48.0	36.5	19.6	35.0
Weak Sup.	1pt	Π Model[10]	72.7	73.2	71.1	**77.0**	76.1	59.7	85.3	**68.0**	88.9	84.3	76.5	**94.9**	44.6	88.7	74.2	45.1	67.4	60.9
		MT[25]	68.6	72.2	71.6	60.0	**79.3**	57.1	86.6	48.4	87.9	80.0	73.7	94.0	43.3	79.8	74.0	45.9	56.9	59.8
		Baseline	72.2	72.6	74.3	75.9	79.0	**64.2**	84.1	58.8	88.8	83.2	72.3	94.7	48.7	84.8	75.8	**50.6**	60.3	59.5
		Ours	**74.4**	**75.5**	**75.6**	74.4	79.2	66.3	**87.3**	63.3	**89.4**	**84.4**	**78.7**	94.5	**49.7**	**90.3**	**76.7**	47.8	**71.0**	**62.6**
	10%	Π Model[10]	79.2	83.8	80.0	82.3	78.7	74.9	89.8	**76.8**	90.6	87.4	**83.1**	95.8	50.7	87.8	77.9	55.2	74.3	82.7
		MT[25]	76.8	81.7	78.0	76.3	78.1	64.4	87.6	67.2	88.7	85.5	79.0	94.3	63.3	90.8	78.2	50.7	67.5	78.5
		Baseline	81.5	84.5	82.5	80.6	**85.7**	76.4	90.0	76.6	89.7	87.1	82.6	95.6	63.3	93.6	79.7	**63.2**	74.4	82.6
		Ours	**81.7**	**85.0**	**83.1**	**82.6**	80.8	**77.7**	**90.4**	77.3	**90.9**	**87.6**	82.9	**95.8**	**64.7**	**93.9**	**79.8**	61.9	**74.9**	**82.9**

Table 2: mIoU (%) evaluations on S3DIS (Area 5) and PartNet datasets. We compared against fully supervised (Ful.Sup.), unsupervised (Unsup.) and alternative weakly supervised (Weak. Sup.) approaches.

			S3DIS													PartNet	
Setting		Model	CatAvg	ceil.	floor	wall	beam	col.	win.	door	chair	table	book.	sofa	board	clutter	CatAvg	SampAvg
Ful.Sup.		PointNet	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	52.6	58.9	**40.3**	5.9	26.4	33.2	57.9	58.3
		PointNet++	**47.8**	90.3	95.6	69.3	0.1	**13.8**	26.7	**44.1**	64.3	**70.0**	27.8	**47.8**	30.8	38.1	65.5	67.1
		DGCNN	47.0	**92.4**	**97.6**	**74.5**	**0.5**	13.3	**48.0**	23.7	**65.4**	67.0	10.7	44.0	**34.2**	**40.0**	**65.6**	**67.2**
Unsup.		Kmeans	38.4	59.8	63.3	34.9	21.5	**24.6**	34.2	29.3	35.7	33.1	45.0	45.6	41.7	30.4	34.6	35.2
		Ncut	**40.0**	**63.5**	**63.8**	**37.2**	**23.4**	**24.6**	**35.5**	**29.9**	**38.9**	**34.3**	**47.1**	**46.3**	**44.1**	**31.5**	**38.6**	**40.1**
Weak Sup.	1pt	Π Model	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	51.4	52.6
		MT	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	52.9	53.6
		Baseline	44.0	89.8	96.7	71.5	0.0	3.0	43.2	**32.8**	60.8	58.7	15.0	41.2	22.5	36.8	50.2	51.4
		Ours	**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**	**54.6**	**55.7**
	10%	Π Model	46.3	91.8	97.1	73.8	0.0	5.1	42.0	19.6	66.7	67.2	19.1	47.9	30.6	41.3	64.1	64.7
		MT	47.9	92.2	96.8	74.1	0.0	10.4	46.2	17.7	67.0	70.7	**24.4**	50.2	**30.7**	42.2	63.8	64.5
		Baseline	45.7	92.3	**97.4**	**75.4**	0.0	**11.7**	47.2	22.9	65.3	66.7	11.7	43.6	17.8	41.5	63.1	63.9
		Ours	**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**	**64.5**	**64.9**

Table 3: Comparisons of different labelling strategies on ShapeNet segmentation. All numbers are in %.

Label Strat.	CatAvg	SampAvg
Samp=10% Pts=100%	70.37	77.71
Samp=20% Pts=50%	72.19	78.45
Samp=50% Pts=20%	74.29	79.65
Samp=80% Pts=12.5%	76.15	80.18
Samp=100% Pts=10%	77.71	80.94

5. Ablation Study

Importance of Individual Components. We analyze the importance of the proposed additional losses and inference label propagation. Different combinations of the losses are evaluated on all datasets with the 1pt annotation scheme. The results are presented in Tab. 4. We observe that the Siamese self-supervision introduces the most advantage for S3DIS. This is because S3DIS is a real dataset, where the orientations and layouts of objects are diverse, and the augmentation and consistency constraints increase the robustness of model. In contrast, the pose of test shapes are always fixed for the other two datasets, and thus they benefit less from Siamese augmentation. We also compare against the use of only data augmentation (last row), and the results suggest it is better to have the consistency constraints on unlabelled points. The results are also further improved with the multi-instance loss for inexact branch. Finally, the smooth constraint at both training (Smo.) and inference (TeLP) stages consistently bring additional advantage to the whole architecture.

Compatibility with Encoder Network. We further examine the compatibility of the proposed losses with different encoder networks. In particular, we investigate the performance with PointNet and DGCNN as the encoder network. The results are shown in Tab. 4 and it is clear that both networks exhibit same patterns.

Table 4: Ablation study on the impact of individual losses and inference label propagation and the compatibility with alternative encoder networks.

Components				PointNet			DGCNN		
MIL	Siam.	Smo.	TeLP	ShapeNet	PartNet	S3DIS	ShapeNet	PartNet	S3DIS
				65.2	49.7	36.8	72.2	50.2	44.0
	✓			66.0	50.3	41.9	73.1	51.5	44.3
✓	✓			69.0	52.1	42.2	73.4	52.9	44.4
✓	✓	✓		69.6	52.5	43.0	73.8	53.6	44.2
✓	✓	✓	✓	**70.2**	**52.8**	**43.1**	**74.4**	**54.6**	**44.5**
Data Augmentation				65.3	49.9	38.9	73.0	52.7	43.2

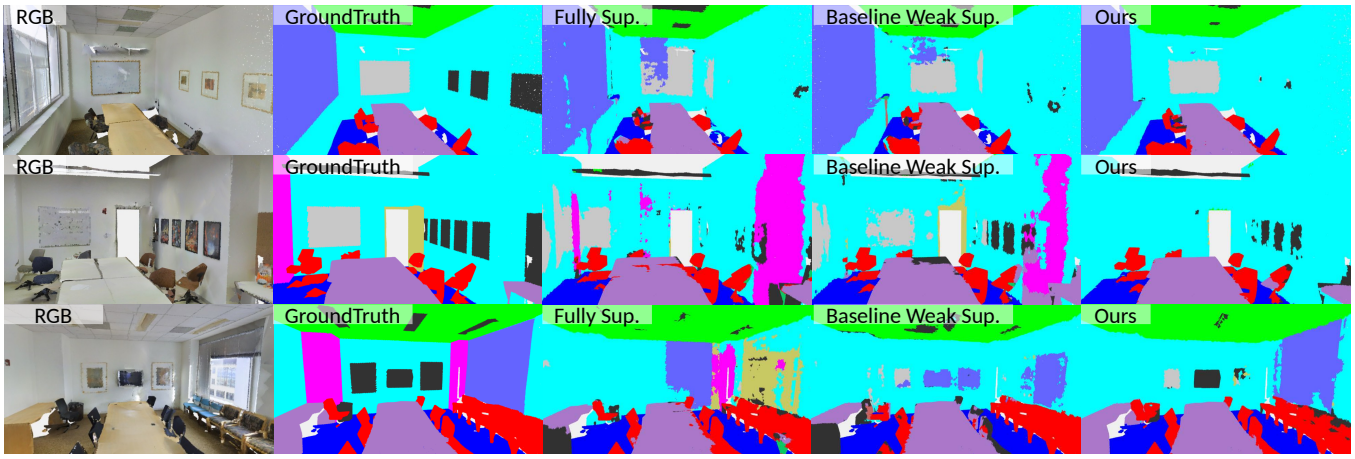


Figure 4: Qualtitative examples for S3DIS dataset test area 5. 10% labelled points are used to train the weak supervision models.

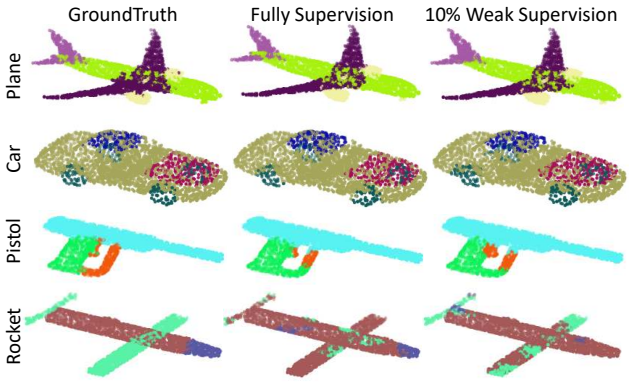
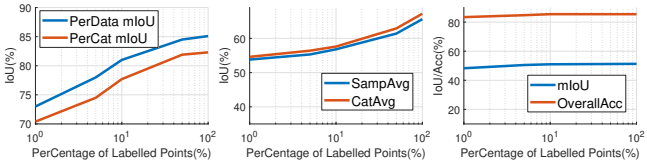


Figure 5: Qualitative examples for ShapeNet shape segmentation.



(a) ShapeNet (b) PartNet (c) S3DIS

Figure 6: The impact of amount of labelled points for all three datasets.

Amount of Labelled Data. As suggested by previous study, the amount of labelled data has an significant impact on the point cloud segmentation performance. In this section, we investigate this relation by varying the amount of labelled points. In particular, we control the percentage of labelled points to be from 1% to 100% (full supervision) with the baseline weak supervision method. The results are presented in Fig. 6. We observe that the performance on all datasets approaches the full supervision after 10% labelled points.

Point Feature Embedding. We visualize the point cloud feature embedding to further understand why weak supervision leads to competitive performance. We first project the feature before the last layer into 2D space via T-SNE [13] for both full supervision and 10% weak supervision. The projected point embeddings are visualized in Fig. 7. We observe similar feature embedding patterns. This again demonstrates a few labelled points can yield very competitive performance.

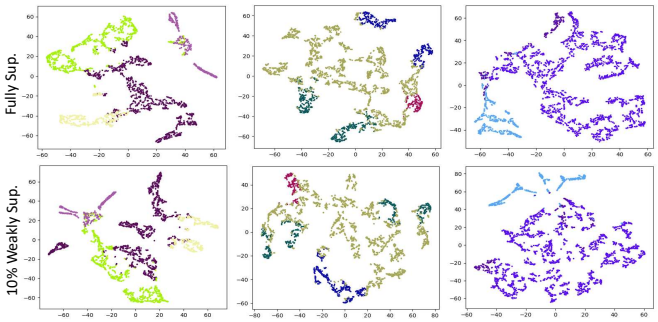


Figure 7: T-SNE visualization of point embeddings in 2D space.

6. Conclusion

In this paper, we made a discovery that only a few labelled points is needed for existing point cloud encoder networks to produce very competitive performance for the point cloud segmentation task. We provide analysis from a statistical point of view and gave insights into the annotation strategy under fixed labelling budget. Furthermore, we proposed three additional training losses, i.e. inexact supervision, Siamese self-supervision and spatial and color smoothness to further regularize the model. Experiments are carried out on three public datasets to validate the efficacy of our proposed methods. In particular, the results are comparable with full supervision with 10 $\times$ fewer labelled points.

Acknowledgement. This work was partially supported by the Singapore MOE Tier 1 grant R-252-000-A65-114.

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*, 2016.

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[3] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 2006.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPRn*, 2019.

[6] Stéphane Guinard and Loïc Landrieu. Weakly supervised segmentation-aided classification of urban scenes from 3d lidar point clouds. In *ISPRS Workshop*, 2017.

[7] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

[8] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, 2019.

[9] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[10] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

[11] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*, 2018.

[12] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NIPS*, 2018.

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[14] Jilin Mei, Biao Gao, Donghao Xu, Wen Yao, Xijun Zhao, and Huijing Zhao. Semantic segmentation of 3d lidar data in dynamic scene using semi-supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[15] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019.

[17] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[18] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[19] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[20] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017.

[21] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018.

[22] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.

[23] Jianbo SHI. Normalized cuts and image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2000.

[24] Zhiyuan Shi, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Weakly-supervised image annotation and segmentation with objects and attributes. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017.

[26] Gusi Te, Wei Hu, Amin Zheng, and Zongming Guo. Rgcnn: Regularized graph cnn for point cloud segmentation. In *ACM MM*, 2018.

[27] Max Welling Thomas N. Kipf. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[28] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

[29] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, 2019.

[30] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *CVPR*, 2018.

[31] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 2014.

[32] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. In *CVPR*, 2018.

[33] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019.

[34] Li Yi, Vladimir G Kim, Duygu Ceylan, I Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, Leonidas Guibas, et al. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics*, 2016.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2015.

[36] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.

[37] Zhihua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 2017.

[38] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.