Dual Adaptive Transformations for Weakly Supervised Point Cloud Segmentation

Zhonghua Wu[1,2], Yicheng Wu[3], Guosheng Lin[1,2(✉)], Jianfei Cai[2,3], and Chen Qian[4]

[1] S-Lab, Nanyang Technological University, Singapore, Singapore
zhonghua001@e.ntu.edu.sg, gslin@ntu.edu.sg
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
[3] Department of Data Science and AI, Monash University, Melbourne, Australia
[4] SenseTime Research, Shanghai, China

Abstract. Weakly supervised point cloud segmentation, i.e. semantically segmenting a point cloud with only a few labeled points in the whole 3D scene, is highly desirable due to the heavy burden of collecting abundant dense annotations for the model training. However, existing methods remain challenging to accurately segment 3D point clouds since limited annotated data may lead to insufficient guidance for label propagation to unlabeled data. Considering the smoothness-based methods have achieved promising progress, in this paper, we advocate applying the consistency constraint under various perturbations to effectively regularize unlabeled 3D points. Specifically, we propose a novel DAT (**D**ual **A**daptive **T**ransformations) model for weakly supervised point cloud segmentation, where the dual adaptive transformations are performed via an adversarial strategy at both point-level and region-level, aiming at enforcing the local and structural smoothness constraints on 3D point clouds. We evaluate our proposed DAT model with two popular backbones on the large-scale S3DIS and ScanNet-V2 datasets. Extensive experiments demonstrate that our model can effectively leverage the unlabeled 3D points and achieve significant performance gains on both datasets, setting new state-of-the-art performance for weakly supervised point cloud segmentation.

Keywords: Weakly supervised segmentation · Point cloud segmentation · Dual adaptive transformations

1 Introduction

Recently, the deep learning (DL)-based methods have achieved significant performance gains for the point cloud segmentation task, which is a fundamental

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19821-2_5.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13691, pp. 78–96, 2022.
https://doi.org/10.1007/978-3-031-19821-2_5

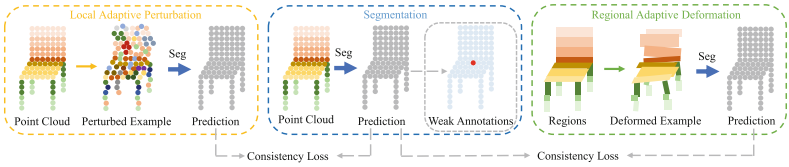

Fig. 1. Illustration of the proposed Dual Adaptive Transformation (DAT) model. We encourage DAT to produce consistent predictions under local and regional adaptive transformations. Note that, there are only few labeled points inside the whole scene to train our model. During testing, only segmentation module (blue) is used to generate the segmentation prediction. (Color figure online)

and critical step to understand realistic scenes [13] and analyze 3D geometric data [59]. However, it is extremely costly and labor-consuming to collect abundant dense annotations of 3D point clouds for model training. Thus, it is highly desirable to develop effective algorithms that can well segment point cloud data with only weak annotations of point clouds.

For semantic image segmentation tasks, there are different types of weak annotations including image-level labels [31,41,55], scribbles [22,48], or partially labeled samples [23,28,49]. For the point cloud segmentation task, following the recent work [24], we consider partially labeled samples as weak annotations for the model training, i.e., only a few sparse points inside the whole scene are labeled and all other points are unlabeled. The latest model 1T1C [24] attempts to train a segmentation model with limited labeled points and then propagate the labels to the unlabeled points as the pseudo labels for iteratively refining the model. However, such a training strategy is time-consuming and is often affected by unreliable pseudo labels, resulting in sub-optimal segmentation performance. Here, we hypothesize that the weakly supervised segmentation performance can be further improved by adding more constraints on the unlabeled 3D points.

To exploit the unlabeled data, the consistency-based learning methods have shown promising progress in natural image classification and segmentation. For example, [2,8,34] encouraged the model to produce invariant results under various strong data augmentations. However, it is non-trivial to apply these image-based strong data augmentation techniques to point cloud processing, and point cloud-specific augmentations are still under early exploration [4,19]. This motivates us to investigate an effective transformation method to leverage large amounts of unlabeled 3D points by applying sufficient smoothness constraints for weakly supervised point cloud segmentation.

Specifically, in this paper, we propose a **Dual Adaptive Transformation (DAT)** model, where we encourage consistent predictions between original and local/regional adaptively transformed point clouds data. As shown in Fig. 1, we first design a *Local Adaptive Perturbation (LAP)* module that computes the adaptive perturbations for both point coordinates and their associate features. Meanwhile, considering the feature distributions are quite different between different classes, we further embed the class-aware information into the LAP module to generate class-aware adaptive feature perturbations. Then, to capture more struc-

tural information in point clouds, we further introduce a *Regional Adaptive Deformation (RAD)* module to apply adaptive deformations on the pre-defined superpoints, which enforces the consistency constraints at the region level.

We evaluate our DAT model with two popular backbones on the large-scale S3DIS dataset [1] and ScanNet-v2 dataset [6]. Via effectively leveraging the unlabeled point clouds, our DAT model is able to segment point cloud data with very few annotations, setting new state-of-the-art (SOTA) performance for the weakly supervised point cloud segmentation task. For example, on S3DIS dataset [1], the DAT model outperforms the previous SOTA model 1T1C [24] by 6.5% under the "One Thing One Click" annotation setting. Note that our proposed strategy can be easily combined with other frameworks. For instance, based on our design, the segmentation performance of 1T1C [24] model can be further improved by 2.9%/3.0% on the ScanNet-v2 test/validation set [6], respectively.

Overall, our main contributions are three-fold:

- We propose a novel Dual Adaptive Transformation (DAT) model for weakly supervised point cloud segmentation, with the key insight that applying the consistency constraint under local and regional adaptive transformations can effectively leverage a large amount of unlabeled 3D points and facilitate a better model training.
- We introduce the Local Adaptive Perturbation (LAP) module, where we inject the adaptive perturbations to point coordinates and the associate feature inputs separately. Meanwhile, we embed the information of the class-aware point feature distribution into the generation of the local adaptive feature perturbations, which leads to better performance.
- We introduce the Regional Adaptive Deformation (RAD) module, where we generate structural adaptive deformations at the region-level, i.e. adaptive deformations such as shifting, scaling, and rotation for the superpoint regions. Such regional deformations introduce another level of the consistency constraint, which is a complement to LAP.

2 Related Work

2.1 Deep Learning on Point Clouds

DL-based methods have achieved great progress to process point cloud data. For example, PointNet model [32] used permutation-invariant operators such as pooling layers to aggregate the features from all points. Then, PointNet++ model [33] further designed a hierarchical spatial structure to extract local geometric features. Furthermore, the graph-based methods [17,18,42] built a graph for all points and applied the message passing mechanism on the graph. For instance, DGCNN [42] used a kNN graph to perform graph convolutions. To capture contextual relationships, SPG [17] constructed a graph on the sub-regions, i.e. the superpoints. DeepGCNs [18] explored the depth information in graph convolutional networks. Afterwards, [26,35,38,54] further improved the performance by directly

applying continuous convolutions on the points without any quantization. SpiderCNN [54] used polynomial functions to generate the kernel weights and the spherical convolution [35] was used to address the 3D rotation equivariance problem in Spherical CNN. KPConv [38] constructed the kernel weights based on the input coordinates and achieved good performance. Similarly, InterpCNN [26] interpolated point-wise kernel weights by utilizing the coordinate information. Different from point convolution networks, the voxel-based methods [5] firstly quantized all the points and map the points to the regular voxels and then applied 3D convolutions on the regular voxels to obtain point features.

In this paper, we adopt the point-based KPConv model [38] as our backbone, where the model is trained via encouraging the dual adaptive transformation consistency for weakly supervised point cloud segmentation. Furthermore, in Sect. 4.2, we also extend our method to the voxel-based framework MinkowiNet [5] so as to demonstrate the generalization ability of our training strategy.

2.2 Weakly Supervised Point Cloud Segmentation

There are some DL-based methods being proposed recently for the weakly supervised point cloud segmentation task [7,9,10,12,20,27,30,36,40,51,56,60]. For example, Wang et al. [39] proposed to generate point cloud segmentation labels by back-projecting 2D image annotations to 3D spaces. However, annotating large-scale image semantic segmentation datasets is extremely labor-consuming. To reduce the labeling costs, Wei et al. [44] used the Class Activation Map (CAM) [50,58] to generate pseudo segmentation masks with sub-cloud level annotations. However, its performance is limited due to the lack of localization information in labels. To address the issue, Xu et al. [53] further labeled 10% points in the whole point cloud, which is able to achieve a good performance comparable to the fully-supervised references. Then, the 1T1C method [24] under the "One Thing One Click" setting was introduced to tackle this task, which uses fewer labeled points, i.e. only labeling one point per thing in each scene.

Here, we follow the 1T1C method [24] to conduct experiments. Different from the iterative refinement mechanism used in 1T1C which brings in significant computational cost, we propose an end-to-end training strategy to train a model in the identical weakly supervised manner while without the need for any iterative refinement.

2.3 Consistency-Based Semi-supervised Learning

Our work is closely related to the consistency based semi-supervised learning (SSL) [45,46], where the basic idea is to leverage the unlabeled data based on the smoothness assumptions, i.e. deep models under various small perturbations or augmentations should output consistent results. For example, Bortsova et al. [3] enforced the model to produce invariant predictions for unlabeled images under different transformations. For semi-supervised image classification task, the VAT model [29] designed an adversarial perturbation and then encouraged the consistency between the original data and its adversarial one. Temporal

ensembling [16] and mean teacher [37] generated similar distributions for the perturbed inputs. Meanwhile, the mutual learning strategy has been studied for semi-supervised learning [47,57]. For instance, the dual-student model [14] enforced two sub-networks learn from each other via constraining the consistent predictions. FixMatch [34] further explicitly generated the pseudo labels from the data with weak augmentations and used them to guide the prediction from the strongly augmented samples.

Motivated by the consistency-based semi-supervised learning methods which encourage the model to produce consistent results under various perturbations, we propose the DAT model for the weakly supervised point cloud segmentation task with two major novel designs, i.e. the LAP and RAD modules.

3 Methods

Figure 2 gives an overview of the proposed Dual Adaptive Transformation (DAT) model, which consists of three main modules: the class-aware Local Adaptive Perturbation (LAP) module, the Regional Adaptive Deformation (RAD) module, and the original SEGmentation (SEG) module. LAP contains a novel Class-aware Perturbation Generator (CPG) to produce semantic perturbations at the point level. RAD generates structural augmented examples by applying various deformations at the region level. SEG contains a conventional point cloud segmentation backbone.

3.1 Segmentation Module

We first define a set of notations for the weakly supervised point cloud segmentation task. Specifically, consider a set of points $X = [C, F] \in \mathbb{R}^{N\times 3+D_f}$ with point coordinates $C \in \mathbb{R}^{N\times 3}$ and the corresponding features $F \in \mathbb{R}^{N\times D_f}$ as the model input, and denote $Y \in \mathbb{R}^{N\times 1}$ as the groundtruth label, which is a very sparse one with only M known entries, $M << N$. The output of SEG module is the predicted segmentation mask $\hat{Y}$. The segmentation module aims to train the backbone model with few labeled points in Y. Here, we adopt a popular segmentation framework KPConv [38] as our backbone. With the kernel parameters denoted as θ, the model prediction is given by $p(\hat{y}_i|c_i, f_i; \theta)$, $i \in \{1, ..., N\}$, where c_i and f_i are respectively the point coordinates and features of point x_i. We train the segmentation module by applying a cross-entropy loss $\mathcal{L}_{seg}$ on the few labels in Y and the corresponding predictions in $\hat{Y}$.

3.2 Local Adaptive Perturbation Module

We design a Local Adaptive Perturbation (LAP) module to generate perturbed examples X^{lap} by applying the adaptive perturbations on the point coordinates and the corresponding features. In particular, the input to LAP is the point cloud X and the output is the perturbed examples X^{lap} with the injection of the adaptive perturbations R^{ada}. Inspired by VAT [29], which is proposed for

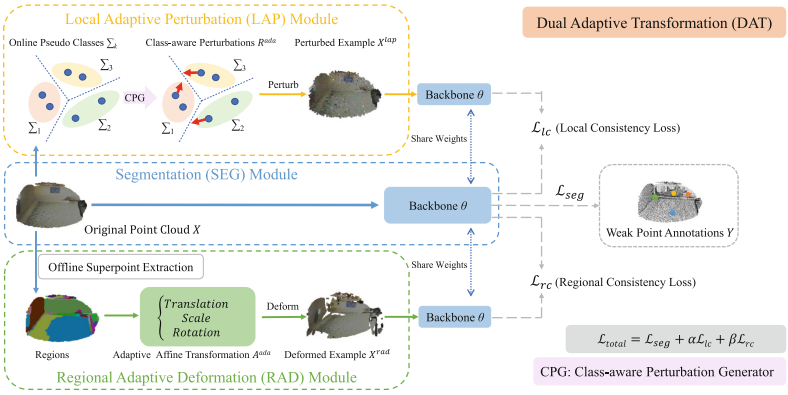

Fig. 2. Overall pipeline of our proposed Dual Adaptive Transformation (DAT) model, which consists of three main modules: the segmentation (SEG) module (blue), the Local Adaptive Perturbation (LAP) module (yellow), and the Regional Adaptive Deformation (RAD) Module (green). SEG module adopts KPConv backbone to train the model with few labeled points. LAP module is to generate class-aware perturbed examples on each point. RAD module generates structural deformed data by applying the adaptive affine transformations on each region. Note that, during testing, we only employ SEG module to process point cloud data. (Color figure online)

semi-supervised image classification, to achieve local distributional smoothness (LDS) as a smoothness constrain to regularize unlabeled data, we encourage our model to generate consistent outputs between each input point $x \in X$ and its perturbed version $x + r^{ada}$, where $r^{ada} \in R^{ada}$ is the corresponding adaptive perturbation:

$$\mathcal{LDS}(x;\theta) = D\left[p(\hat{y}|x;\theta), p(\hat{y}|x + r^{ada};\theta)\right]. \tag{1}$$

Here D is a non-negative loss function to measure the divergence between x and $x + r^{ada}$. Then, we compute r^{ada} by estimating a gradient g of $\mathcal{LDS}$ with a random input vector d as

$$\begin{aligned} g &= \nabla_R D\left[p(\hat{y}|x,\theta), p(\hat{y}|x + r,\theta)\right]\Big|_{r=\xi d} \\ r^{ada} &= \epsilon \times g/\|g\|_2, \end{aligned} \tag{2}$$

where ξ and ϵ are two hyper-parameters to control the magnitude of the perturbation, and g can be efficiently computed by applying the back-propagation on the network.

Considering the input point coordinates and features are two different types of inputs, we generate their perturbations separately. In other words, for an input point x consisting of its coordinates c and features f, we generate the adaptive perturbation data $c + r_c^{ada}$ and $f + r_f^{ada}$ with the initial random unit vectors d_c and d_f, respectively.

Class-Aware Perturbation Generator. Note that many existing perturbation based semi-supervised image classification methods [29] usually generate

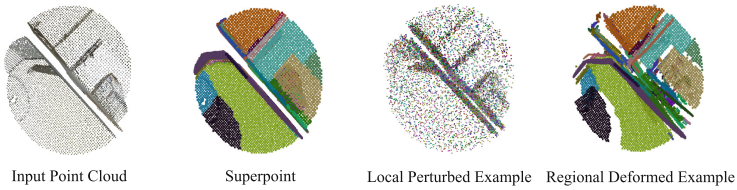

Fig. 3. Visual results for the superpoint estimation and the generated dual adaptive transformed examples during the training stage.

the initial perturbations d through sampling them from an iid Gaussian distribution. However, in the point cloud segmentation task, directly applying this to generate d_f might not be optimal. This is because, for different classes, their input point feature distributions are quite different across different dimensions. A class-agnostic iid Gaussian sampling might generate unrealistic perturbations.

Therefore, we propose a Class-aware Perturbation Generator (CPG) to obtain d_f for each point. Specifically, in each training iteration, we generate the pseudo labels $\hat{y}$ for all the points with the current model parameter $\hat{\theta}$, where $\hat{y} \in \{1, ..., K_c\}$ with K_c being the number of classes. Based on that, we establish a zero-mean multivariate normal distribution $\mathbb{N}(0, \sum_k)$. Here $\sum_k$ is the class-conditional covariance matrix estimated from all the input point features (e.g. rgbh for KPConv) that belong to the pseudo-class k. Afterward, we update the covariance matrix in an online manner [43] with the statistics of the features from each mini-batch. In this way, at each iteration, d_f is then generated by sampling from the up-to-date class-aware multivariate Gaussian distribution. For d_c, we adopt the conventional way i.e. sampling the initial input vectors from an iid Gaussian distribution, since the point clouds are unordered and the individual coordinates alone are not closely related to the class of the points. This is also observed in the PointNet model [32].

With the generated d_c and d_f, our LDS loss for point clouds now becomes:

$$\begin{aligned} \mathcal{LDS}(x;\theta) &= D\left[p(\hat{y}|c, f;\theta), p(\hat{y}|c + \xi_c d_c, f + \xi_f d_f;\theta)\right] \\ g_c &= \nabla_{\xi_c d_c} \mathcal{LDS}(x,\theta) \\ g_f &= \nabla_{\xi_f d_f} \mathcal{LDS}(x,\theta), \end{aligned} \tag{3}$$

where we use the Kullback-Leibler divergence (KL-div) for D. Finally, we obtain the r_c^{ada} and r_f^{ada} by

$$\begin{aligned} r_c^{\text{ada}} &= \epsilon_c g_c / \|g_c\|_2 \\ r_f^{\text{ada}} &= \epsilon_f g_f / \|g_f\|_2. \end{aligned} \tag{4}$$

In this way, the perturbed examples X^{lap} is obtained by point-wise adding the perturbations r_c^{ada} and r_f^{ada} on the coordinates c and the features f, respectively. One example is visualized in the third column of Fig. 3.

3.3 Regional Adaptive Deformation Module

In addition to the local adaptive perturbations, considering point clouds often contain various structural local deformations such as region shift, rotation, and scaling, we further design a regional adaptive deformation (RAD) module to generate structural local deformations. RAD module takes point cloud X as input and outputs region-level augmented examples X^{rad} by deforming each region with adaptive affine transformations A^{ada}. As shown in Fig. 2, we firstly over-segment point cloud X into a set of superpoints $S_i, i \in \{1, ..., K_s\}$ via [6,17]. For each superpoint S_i, we generate the adaptive deformed example S_i^{ada}. Combing all $S_i^{ada}, i \in \{1, ..., K_s\}$, we obtain X^{rad}.

For each superpoint S_i, we firstly generate the initial affine transformation matrices $A_{i,j}$, whose parameters are randomly sampled from an iid Gaussian distribution. Then, we deform each superpoint as

$$S_i^{int} = S_i \cdot \prod_{j=1}^{K_a} \xi_A A_{i,j}, \tag{5}$$

where $A_{i,j}, j \in \{1, ..., K_a\}$, corresponds to the j-th type of deformations. Combining all $S_i^{int}, i \in \{1, ..., K_s\}$, we obtain the initial deformed point cloud X^{int}. The $\mathcal{LDS}$ loss becomes

$$\begin{aligned} \mathcal{LDS}(X; \theta) &= D\left[p(\hat{y}|x; \theta), p(\hat{y}|x^{int}; \theta)\right] \\ g_{A_{i,j}} &= \nabla_{\xi_A A_{i,j}} \mathcal{LDS}(x; \theta). \end{aligned} \tag{6}$$

Then, we obtain the $A_{i,j}^{ada}$ by

$$A_{i,j}^{ada} = \epsilon_A g_{A_{i,j}} / \|g_{A_{i,j}}\|_2. \tag{7}$$

Finally, the regional deformed examples X^{rad} is obtained by combining all the deformed superpoints S_i^{ada}, which is computed as

$$S_i^{ada} = S_i * \prod_{j=1}^{K_a} A_{i,j}^{ada}. \tag{8}$$

Specifically, we use the following three types of affine transformations: translation, scale and rotation.

One RAD example is given in the fourth column of Fig. 3. Algorithm 1 summarizes the process of generating the adversarial examples under both LAP and RAD.

3.4 Training Losses

The overall training loss can be written as

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \alpha \mathcal{L}_{lc} + \beta \mathcal{L}_{rc} \tag{9}$$

Algorithm 1. Generating adaptive transformed examples (LAP/RAD)

Input: Training Point Cloud X
Output: Local perturbed examples X^{lap}/ Regional deformed examples X^{rad}

1. Generate initial R/A for initial transformation.
2. Compute the gradient of D with respective to R/A
   $g_R/g_A \leftarrow \nabla_r D\left[p(\hat{y}|x;\theta), p(\hat{y}|x \odot R;\theta)\right]\Big|_{r=\xi(R/A)}$ where $\odot$: $\oplus/\otimes$ for LAP/RAD, respectively.
3. Normalize the gradient to generate adaptive perturbations R^{ada}/A^{ada}.
   $R^{ada} \leftarrow \epsilon \cdot g_R/\|g_R\|_2$ or $A^{ada} \leftarrow \epsilon \cdot g_A/\|g_A\|_2$
4. Generate the adversarial examples X^{lap} / X^{rad} by injecting the R^{ada} / A^{ada} to Point Cloud X.

where $\mathcal{L}_{seg}$, $\mathcal{L}_{lc}$ and $\mathcal{L}_{rc}$ are *Segmentation Loss*, *Local Consistency Loss* and *Regional Consistency Loss*, respectively, and α and β are trade-off weights, both set as 2 to balance the losses. Segmentation Loss $\mathcal{L}_{seg}$ is to guide the segmentation prediction with the limited annotations in Y. Specifically, we follow the KPConv [38] by using the cross entropy loss for $\mathcal{L}_{seg}$ to train the segmentation prediction $\hat{Y}$. Local Consistency Loss $\mathcal{L}_{lc}$ encourages the consistency and penalizes the prediction difference between the original point cloud X and the local perturbed examples X^{lap}. Regional Consistency Loss $\mathcal{L}_{rc}$ ensures the consistency between X and its regional deformed examples X^{rad}. $\mathcal{L}_{pc}$ and $\mathcal{L}_{rc}$ are defined as

$$\begin{aligned} \mathcal{L}_{pc} &= D\left[p(\hat{y}|x;\theta), p(\hat{y}|x^{lap};\theta)\right] \\ \mathcal{L}_{rc} &= D\left[p(\hat{y}|x;\theta), p(\hat{y}|x^{rad};\theta)\right] \end{aligned} \tag{10}$$

where D is the KL-div loss.

4 Experiments and Results

4.1 Implementation Details

Datasets. Following the 1T1C [24] model, we conduct experiments on two large-scale point cloud datasets - the S3DIS [1] and ScanNet-v2 [6]. The S3DIS dataset consists of 3D scans of 271 rooms with 13 categories belonging to 6 areas. For fair comparisons, we train the segmentation model on Area 1, 2, 3, 4, 6 and test on Area 5 as [24]. The ScanNet-v2 dataset contains 1201, 312, and 100 3D scans for training, validation, and testing, respectively.

Weak Annotation Scheme. For fair comparisons, on the S3DIS dataset, we label the data under the "One Thing One Click" (OTOC) setting as in 1T1C [24]. We randomly select a point in each object with the identical probability as the labeled points. Therefore, only 0.02% of points have annotations inside the whole point cloud. On the ScanNet-v2 dataset, we evaluate our DAT model on the "3D Semantic label with Limited Annotations" benchmark [6]. In this benchmark, only 20 points are labeled in each room scene.

Table 1. Comparison of our DAT with several existing methods on the S3DIS Area-5 set. Note that, we report the performance as final results based on the KPConv [38] backbone.

Method	Supervision (%)	mIoU (%)
PointNet [32]	100%	41.1
PointCNN [21]	100%	57.3
Xu et al. [53]	0.2%	44.5
Xu et al. [53]	10%	48.0
GPFN [39]	16.7% 2D	50.8
GPFN [39]	100% 2D	52.5
1T1C [24]	0.02% (OTOC)	50.1
1T1C [24]	0.06% (OTTC)	55.3
Our DAT	0.02% (OTOC)	**56.5**
Our DAT	0.06% (OTTC)	**58.5**
Our upper bound	100%	65.4

Table 2. Comparison of our DAT with its variant methods with the KPConv framework. Note that, all experiments are conducted under the OTOC setting on the S3DIS dataset

Method	Random noises		LAP	RAD	mIoU (%)
	Features	Coordinates			
Our baseline					50.1
Ours w/ Noise	✓				49.1
Ours w/ Noise		✓			52.9
Ours w/ Noise	✓	✓			52.6
Ours w/ PAP			✓		53.9
Ours w/ RAD				✓	54.8
Our DAT			✓	✓	**56.5**

Experiment Setting. If there is no special declaration, we implement our proposed DAT training method based on the KPConv *rigid* model. We use SGD to train the model with learning rate of 0.01 and batch size of 2. Following 1T1C [24], we use the geometrical partition results [17] and mesh segment results [6] as the superpoints for S3DIS and ScanNet-v2 datasets, respectively. We set the hyper-parameters $\xi_c = 10$, $\xi_f = 0.1$, $\xi_A = 0.1$, $\epsilon_c = 1$, $\epsilon_f = 0.05$, $\epsilon_A = 0.05$. During the model training, to reduce the GPU memory consumption, we employ the segmentation loss $\mathcal{L}_{seg}$ at all iterations and randomly apply local consistency loss $\mathcal{L}_{lc}$ or regional consistency loss $\mathcal{L}_{rc}$ with an equal probability of 0.5 to train our model. All of our experiments are conducted on a single NVIDIA RTX 3090 GPU with PyTorch 1.7.0 and CUDA 11.0.

Table 3. Ablation studies of our DAT about the Class-aware Perturbation Generator (CPG) used in our LAP module under the OTOC setting on the S3DIS dataset.

Method	LAP			RAD	mIoU (%)
	Feat. w/o CPG	Feat. w/ CPG	Coordinates		
Ours w/o RAD	✓				51.3
Ours w/o RAD		✓			**51.7**
Ours w/o RAD	✓		✓		53.3
Ours w/o RAD		✓	✓		**53.9**
Our DAT	✓		✓	✓	55.1
Our DAT		✓	✓	✓	**56.5**

4.2 Evaluations on S3DIS Dataset

Comparing with State-of-the-Art Methods. Table 1 shows the results of our DAT and several SOTA methods on the S3DIS Area 5 dataset. Via effectively exploiting the unlabeled data, the DAT model with few labeled points training achieves comparable results with the upper bound (i.e. the fully-supervised KPConv model with 100% labeled data training). Furthermore, under the "OTOC" setting, the DAT model significantly outperforms the second-best 1T1C method by 6.4% mIOU gains on the S3DIS dataset. In addition, we further perform the "One Thing Three clicks" (OTTC) setting, where we annotate three points for each target. Our model outperforms the corresponding second-best method 1T1C [24] by 3.2%.

Ablation Studies Comparisons with Baselines. We perform the ablation studies on the S3DIS dataset, to show the effectiveness of our proposed DAT. The first baseline is that we only use the segmentation loss $\mathcal{L}_{seg}$ on a few labeled points to train the segmentation model, which is denoted as "Our Baseline" in Table 2. Our proposed DAT outperforms "Our baseline" by 6.4%. Another baseline is that we apply random noises to all the points to generate perturbed examples. Then we use KL-div loss to encourage the prediction consistency between the original point cloud and the perturbed examples. Specifically, similar to our designed LAP, we are able to apply random noises to point coordinates, point features, or both, which is denoted as "Ours w/ Noise". As Table 2 shows, the DAT significantly outperforms two baseline methods, which suggests that our adaptive perturbations achieve better regularization to the unlabeled data compared to the random noises.

Effects of LAP and RAD. To demonstrate the effects of two novel modules, as shown in Table 2, with separately applying consistency loss on the transformed examples generated by LAP (Ours w/ LAP) or RAD (Ours w/ RAD), we are able to significantly improve mIoU results compared with the "Our Baseline". This suggests that enforcing the consistency between the prediction of transformed

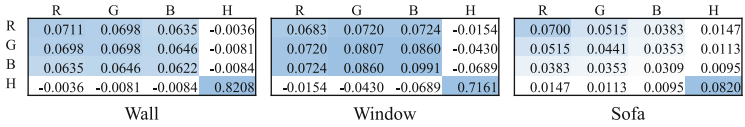

Fig. 4. Three covariance matrices estimated via our designed CPG module under the OTOC setting on the S3DIS datasets.

Table 4. Ablation studies of our DAT on different affine transformations used in RAD module under the OTOC setting on the S3DIS dataset.

Method	LAP	RAD			mIoU (%)
		Translation	Scale	Rotation	
Ours w/ RAD		✓			54.1
Ours w/ RAD			✓		54.6
Ours w/ RAD				✓	53.9
Ours w/ RAD		✓	✓		**54.8**
Ours w/ RAD		✓		✓	54.6
Ours w/ RAD		✓	✓	✓	54.5
Our DAT	✓	✓			55.5
Our DAT	✓		✓		55.9
Our DAT	✓			✓	55.2
Our DAT	✓	✓	✓		56.0
Our DAT	✓	✓		✓	55.2
Our DAT	✓	✓	✓	✓	**56.5**

examples and the original point clouds can predict better segmentation masks. "Our DAT" denotes that we apply the consistency loss on both LAP and RAD. Table 2 shows combining both modules can further improve mIoU by 2.6% and 1.7% compared with only using LAP or RAD, respectively.

Effects of CPG. We further verify the effectiveness of our designed CPG used in the LAP module. "Feat. w/o CPG" denotes that we generate the initial perturbation d_f from the iid Gaussian distribution, instead of the class-aware multivariate Gaussian distribution. Table 3 shows that our class-aware perturbation generator is able to boost segmentation performance under all settings, which suggests that the class-aware information is critical in the point cloud segmentation task.

Besides, Fig. 4 gives three examples of the computed covariance matrices in the CPG, where we randomly select them from all 13 covariance matrices. We can observe that different classes have different covariance matrices.

Different Affine Transformations in RAD. Table 4 shows the mIoU results for our DAT with different affine transformations. "Ours w/ RAD" indicates that we only apply the consistency loss on the deformed examples generated by RAD,

and "Our DAT" indicates that we make use of all the transformed examples generated by LAP and RAD to train the model. As Table 4 shows, "Our DAT" achieves the best performance by using all three affine transformation methods (i.e. translation, scale and rotation).

Generalization Ability. To verify the generalization ability, we further use our training strategy to train a voxel-based segmentation framework (i.e. MinkowskiNet [5]). Unlike the point-based methods, the voxel-based methods firstly project the point cloud into regular voxels and then apply 3D sparse convolution on it. Since the projecting operation is non-differentiable and cannot back-propagate the gradients to point coordinates, we only employ the LAP

Table 5. To show the generalization ability, we further show the results with MinkowskiNet32 [5] backbone on the S3DIS Area-5 set. "Our DAT*" denotes we only use our LAP module to train the backbone.

Method	Supervision (%)	mIoU (%)
Our Baseline	0.02% (OTOC)	48.7
Our Baseline	0.06% (OTTC)	55.0
Our DAT*	0.02% (OTOC)	**54.6**
Our DAT*	0.06% (OTTC)	**58.2**
Our Upper bound	100%	65.4

Table 6. Comparison of our DAT model with several existing methods on the ScanNet-v2 test set. "Our DAT†" denotes that our DAT is built upon the 1T1C [24] model.

Method	Supervision	mIoU (%)
Pointnet++ [33]	100%	33.9
PointCNN [21]	100%	45.8
MinkowskiNet [5]	100%	73.6
Virtual MVFusion [15]	100%+2D	74.6
MPRM [44]	Scene-level	24.4
MPRM [44]	Subcloud-level	41.1
MPRM+CRF [44]	Subcloud-level	43.2
CSC_LA_SEM [11]	20 points	53.1
Viewpoint_BN_LA_AIR [25]	20 points	54.8
PointContrast_LA_SEM [52]	20 points	55.0
1T1C [24]	20 points	59.4
Our Baseline	20 points	51.6
Our DAT	20 points	55.2
Our DAT†	20 points	**62.3**
Our Upper Bound	100%	68.4

module to add adaptive perturbations on the input features with the CPG module (labeled as "Our DAT*" in Table 5). Table 5 shows, under the OTOC/OTTC setting, our model improves the mIoU results by 5.9%/3.2% compared to their respective "Our Baseline", which demonstrates that such novel training strategy is general and effective, and can be easily applied to various point cloud frameworks.

4.3 Evaluations on ScanNet-v2 Dataset

Tables 6 and 7 respectively give the results on the test and validation set of ScanNet-v2 dataset in the "3D Semantic label with Limited Annotations" benchmark. We use the officially given 20 points annotations as the sparse labels to train the model. Compared with "Our Baseline", our DAT (denoted as "Our DAT") with the KPConv backbone can achieve impressive performance gains of 3.2% and 3.9% mIoU on ScanNet-v2 test and validation sets, respectively.

Meanwhile, such a training strategy can be easily combined with existing models for point cloud segmentation. For example, on the ScanNet-v2 dataset, we build our DAT upon the 1T1C model, which is used to generate the pseudo labels for all training data. Then we use the pseudo labels to train our DAT. Based on the 1T1C model (denoted as "Our DAT †" in Tables 6 and 7), our DAT can further improve the mIoU results by 2.9% and 3.0% on the ScanNet-v2 test and validation set compared with 1T1C, respectively. This suggests that our training strategy can further improve the performance of other SOTA models.

Table 7. Comparison of our DAT model with several existing methods on the ScanNet-v2 validation set. "Our DAT†" denotes that our DAT is built upon the 1T1C [24] model.

Method	Supervision	mIoU (%)
1T1C [24]	20 points	61.4
Our Baseline	20 points	54.6
Our DAT	20 points	58.9
Our DAT†	20 points	**64.4**
Our Upper Bound	100%	68.5

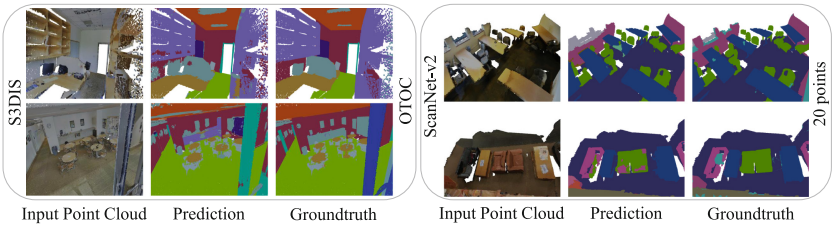

Fig. 5. Two results of our DAT on the S3DIS (first two rows, under the "OTOC" setting) and ScanNet-v2 datasets (last two rows, under the "20 points" setting).

4.4 Qualitative Results

Figure 5 shows the segmentation results obtained by our proposed DAT model on the S3DIS and ScanNet-v2 dataset. It reveals that the DAT model can successfully preserve most of the object structures and segment the 3D point clouds accurately, only with the weak annotation training.

5 Conclusion

In this paper, we have presented a Dual Adaptive Transformations (DAT) model for the weakly supervised point cloud segmentation task, with two novel designs, i.e. the LAP and RAD module. First, the LAP module generates point-wise adaptive coordinate perturbations and class-aware adaptive feature perturbations based on the online estimated class distribution. Second, we propose the RAD module to generate regional adaptive deformations by applying a set of adaptive affine transformations on the superpoint regions. Extensive experimental results under multiple weakly supervised settings have demonstrated that our proposed DAT model achieves new SOTA segmentation performance on the S3DIS and ScanNet-v2 datasets.

Acknowledgments. This study is supported under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This research is partly supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20220-0007) and Tier 1 (RG95/20). This research is also partially supported by Monash FIT Start-up Grant and SenseTime Gift Fund.

References

1. Armeni, I., et al.: 3D semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534–1543 (2016)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems 32 (2019)
3. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M.: Semi-supervised medical image segmentation via learning consistency under transformations. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 810–818. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_90
4. Chen, Y., et al.: PointMixup: augmentation for point clouds. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 330–345. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_20
5. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3075–3084 (2019)

6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839 (2017)
7. Deng, S., Dong, Q., Liu, B., Hu, Z.: Superpoint-guided semi-supervised semantic segmentation of 3D point clouds. arXiv preprint arXiv:2107.03601 (2021)
8. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv preprint arXiv:1906.01916 (2019)
9. Gao, B., Pan, Y., Li, C., Geng, S., Zhao, H.: Are we hungry for 3D LiDAR data for semantic segmentation? arXiv preprint arXiv:2006.04307 3, 20 (2020)
10. Hamdi, A., Rojas, S., Thabet, A., Ghanem, B.: AdvPC: transferable adversarial perturbations on 3D point clouds. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 241–257. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_15
11. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3D scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15587–15597 (2021)
12. Hu, Q., et al.: SQN: weakly-supervised semantic segmentation of large-scale 3D point clouds with 1000x fewer labels. arXiv preprint arXiv:2104.04891 (2021)
13. Jaritz, M., Gu, J., Su, H.: Multi-view PointNet for 3D scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
14. Ke, Z., Wang, D., Yan, Q., Ren, J., Lau, R.W.: Dual student: breaking the limits of the teacher in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6728–6736 (2019)
15. Kundu, A., et al.: Virtual multi-view fusion for 3D semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 518–535. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_31
16. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
17. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4558–4567 (2018)
18. Li, G., Muller, M., Thabet, A., Ghanem, B.: DeepGCNs: can GCNs go as deep as CNNs? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9267–9276 (2019)
19. Li, R., Li, X., Heng, P.A., Fu, C.W.: PointAugment: an auto-augmentation framework for point cloud classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6378–6387 (2020)
20. Li, X.: SnapshotNet: self-supervised feature learning for point cloud data segmentation using minimal labeled data. Ph.D. thesis, City University of New York (2021)
21. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: PointCNN: convolution on X-transformed points. In: Advances in Neural Information Processing Systems 31, pp. 820–830 (2018)
22. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3159–3167 (2016)
23. Liu, W., Wu, Z., Ding, H., Liu, F., Lin, J., Lin, G.: Few-shot segmentation with global and local contrastive learning. arXiv preprint arXiv:2108.05293 (2021)

24. Liu, Z., Qi, X., Fu, C.W.: One thing one click: a self-training approach for weakly supervised 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1726–1736 (2021)
25. Luo, L., Tian, B., Zhao, H., Zhou, G.: Pointly-supervised 3D scene parsing with viewpoint bottleneck. arXiv preprint arXiv:2109.08553 (2021)
26. Mao, J., Wang, X., Li, H.: Interpolated convolutional networks for 3D point cloud understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1578–1587 (2019)
27. Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., Van Gool, L.: Towards a weakly supervised framework for 3D point cloud object detection and annotation. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 4454–4468 (2021)
28. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high- and low-level consistency. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1369–1379 (2021). https://doi.org/10.1109/TPAMI.2019.2960224
29. Miyato, T., Maeda, S., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2018)
30. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3D: out-of-context data augmentation for 3D scenes. In: 2021 International Conference on 3D Vision (3DV), pp. 116–125. IEEE (2021)
31. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5038–5047. IEEE (2017)
32. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413 (2017)
34. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
35. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360 imagery. In: Advances in Neural Information Processing Systems 30, pp. 529–539 (2017)
36. Tao, A., Duan, Y., Wei, Y., Lu, J., Zhou, J.: SegGroup: seg-level supervision for 3D instance and semantic segmentation. arXiv preprint arXiv:2012.10217 (2020)
37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
38. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: KPConv: flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6411–6420 (2019)
39. Wang, H., Rong, X., Yang, L., Feng, J., Xiao, J., Tian, Y.: Weakly supervised semantic segmentation in 3D graph-structured point clouds of wild scenes. arXiv preprint arXiv:2004.12498 (2020)
40. Wang, P., Yao, W.: A new weakly supervised approach for ALS point cloud semantic segmentation. arXiv preprint arXiv:2110.01462 (2021)
41. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354–1362 (2018)

42. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph. (TOG) **38**(5), 1–12 (2019)
43. Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., Wu, C.: Regularizing deep networks with semantic data augmentation. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3733–3748 (2021)
44. Wei, J., Lin, G., Yap, K.H., Hung, T.Y., Xie, L.: Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4384–4393 (2020)
45. Wu, Y., et al.: Mutual consistency learning for semi-supervised medical image segmentation. Med. Image Anal. **81**, 102530 (2022)
46. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. arXiv preprint arXiv:2203.01324 (2022)
47. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 297–306. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_28
48. Wu, Z., Lin, G., Cai, J.: Keypoint based weakly supervised human parsing. Image Vis. Comput. **91**, 103801 (2019)
49. Wu, Z., Shi, X., Lin, G., Cai, J.: Learning meta-class memory for few-shot semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 517–526 (2021)
50. Wu, Z., Tao, Q., Lin, G., Cai, J.: Exploring bottom-up and top-down cues with attentive learning for webly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12936–12945 (2020)
51. Xiang, C., Qi, C.R., Li, B.: Generating 3D adversarial point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9136–9144 (2019)
52. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: PointContrast: unsupervised pre-training for 3D point cloud understanding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 574–591. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_34
53. Xu, X., Lee, G.H.: Weakly supervised semantic point cloud segmentation: towards 10x fewer labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13706–13715 (2020)
54. Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Yu.: SpiderCNN: deep learning on point sets with parameterized convolutional filters. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 90–105. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_6
55. Zhang, T., Lin, G., Liu, W., Cai, J., Kot, A.: Splitting vs. merging: mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12367, pp. 663–679. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58542-6_40
56. Zhang, Y., Qu, Y., Xie, Y., Li, Z., Zheng, S., Li, C.: Perturbed self-distillation: weakly supervised large-scale point cloud semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15520–15528 (2021)

57. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
58. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
59. Zhou, Y., Tuzel, O.: VoxelNet: end-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4490–4499 (2018)
60. Zhu, X., et al.: Weakly supervised 3D semantic segmentation using cross-image consensus and inter-voxel affinity relations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2834–2844 (2021)