Multi-Modality Affinity Inference for Weakly Supervised 3D Semantic Segmentation

Xiawei Li[1*], Qingyuan Xu[1*], Jing Zhang[1†], Tianyi Zhang[2], Qian Yu[1], Lu Sheng[1], Dong Xu[3]

[1]School of Software, Beihang University
[2]College of Computer Science and Technology, Zhejiang University
[3]Department of Computer Science, The University of Hong Kong
{ZY2121108,ZY2121121,zhang_jing,qianyu,lsheng}@buaa.edu.cn, tianyizhang0213@zju.edu.cn, dongxu@hku.hk

Abstract

3D point cloud semantic segmentation has a wide range of applications. Recently, weakly supervised point cloud segmentation methods have been proposed, aiming to alleviate the expensive and laborious manual annotation process by leveraging scene-level labels. However, these methods have not effectively exploited the rich geometric information (such as shape and scale) and appearance information (such as color and texture) present in RGB-D scans. Furthermore, current approaches fail to fully leverage the point affinity that can be inferred from the feature extraction network, which is crucial for learning from weak scene-level labels. Additionally, previous work overlooks the detrimental effects of the long-tailed distribution of point cloud data in weakly supervised 3D semantic segmentation. To this end, this paper proposes a simple yet effective scene-level weakly supervised point cloud segmentation method with a newly introduced multi-modality point affinity inference module. The point affinity proposed in this paper is characterized by features from multiple modalities (e.g., point cloud and RGB), and is further refined by normalizing the classifier weights to alleviate the detrimental effects of long-tailed distribution without the need of the prior of category distribution. Extensive experiments on the ScanNet and S3DIS benchmarks verify the effectiveness of our proposed method, which outperforms the state-of-the-art by $\sim 4\%$ to $\sim 6\%$ mIoU. Codes are released at https://github.com/Sunny599/AAAI24-3DWSSG-MMA.

Introduction

Point cloud data capture rich object and scene geometric and appearance information, which serves as an essential data representation for various applications, such as autonomous driving, augmented reality, and robotic. Point cloud semantic segmentation plays a key role in 3D scene understanding, and has been extensively explored (Qi et al. 2017a,b; Wang et al. 2019b,a; Thomas et al. 2019; Shi, Wang, and Li 2019; Wu, Qi, and Fuxin 2019; Hu et al. 2020; Zhao et al. 2021; Xu et al. 2021; Lai et al. 2022). However, the success of most of the methods is based on learning in a fully supervised manner, requiring extensive point-level annotations.

---

*These authors contributed equally.

†Corresponding author

To reduce the annotation costs, some recent works have delved into the realm of weakly supervised semantic segmentation (WSSS) methods (Bearman et al. 2016; Xu and Lee 2020; Zhang et al. 2021b; Liu, Qi, and Fu 2021; Hou et al. 2021; Zhang et al. 2021a; Yang et al. 2022; Wei et al. 2020; Ren et al. 2021; Yang et al. 2022). These methods can be categorized based on different levels of supervision, including partially labeled points, sub-cloud level annotations, and scene-level annotations. Among these, segmentation with scene-level labels is the most challenging scenario as point-wise annotations are completely unavailable, which is the focus of our paper.

Most of the current methods with scene-level annotations (Wei et al. 2020; Ren et al. 2021; Yang et al. 2022) are proposed based on pseudo labels which can be obtained by Class Activation Map (CAM) (Zhou et al. 2016) or Multiple Instance Learning (MIL) (Maron and Lozano-Pérez 1997). The key to successful weakly supervised semantic segmentation is how to expand the semantic regions to achieve completeness and preciseness of the localized objects. The 3D point clouds of the RGB-D scans provide accurate object shape and scale information without occlusion and distortions, while the corresponding RGB data provide additional color and texture information. However, the current weakly supervised point cloud segmentation based on RGB-D scans fails to fully take advantage of all the data modalities. Moreover, current methods fail to fully exploit the point affinity that can be readily inferred from the feature extraction network. We argue that point affinity that characterizes the similarity between points is essential for learning from scene-level weak labels.

In addition, the long-tailed distribution of point cloud data due to the extremely imbalanced point data have detrimental effects on both point-wise classification performance and point-wise feature learning, which has largely overlooked by previous research on weakly-supervised point cloud semantic segmentation. For example, when the point cloud data have long-tailed distribution, the point-wise classifier for semantic segmentation will be biased towards the head classes with more number of samples (Wang et al. 2022; Kang et al. 2019; Wang et al. 2021; Cui et al. 2019; Jamal et al. 2020). Moreover, the point features of the tail classes with less number of samples will also be learned similar to the point features of the head classes since the feature ex-

tractor may also be dominated by the head classes. These detrimental effects will further affect the affinity learning that relies on feature similarity. However, the lack of per-point category information in scene-level weak supervision makes it challenging to obtain category distribution. There-fore, how to take full advantage of both RGB and geome-try information in the point cloud data while addressing the long-tailed distribution issue for affinity learning in the con-text of weakly-supervised semantic segmentation remain a challenging task.

To this end, we propose a novel multi-modality affinity (MMA) enhanced weakly supervised semantic segmenta-tion (WSSS) method by fully exploiting the point feature affinities from multiple data modalities and eliminating the influence of long-tail data distribution. The geometric infor-mation derived from point clouds and the color and texture information captured in RGB data offer distinct perspectives for characterizing feature affinity. By leveraging these com-plementary data modalities, we propose to generate multi-modality affinities based on both pure geometric data as well as color-appended RGB-D data. For simplicity, we mask out the RGB features from the input point cloud data to model the geometric affinity and use the original point cloud data to model the RGB-enriched data affinity. We also enhance the affinity by normalizing the weights of the point-wise classi-fier. This normalization process assists in mitigating the net-work's tendency to misclassify data points into the dominant head categories. Consequently, it enhances the point-wise features of the tail categories, allowing for improved affinity inference. The obtained multi-modality affinity matrices are used to refine the WSSS-related objective functions. Exten-sive experiments are conducted on ScanNet (Dai et al. 2017) and S3DIS (Armeni et al. 2016) benchmarks, and the results demonstrate that our method significantly outperforms the state-of-the-art scene-level weakly supervised point cloud semantic segmentation methods.

Related Work

Weakly-supervised Point Cloud Segmentation with Sparse Labels

The main idea of learning from sparse annotations focuses on propagating information from labeled points to the unla-beled points. For example, Liu *et al.* (Liu, Qi, and Fu 2021) generates a super-voxel graph between labeled points and unlabeled to guide the iterative training. Yang *et al.*(Yang et al. 2022) design transformer model derived by multiple instance learning (MIL), where the two clouds with shared category yield a positive bag while with different classes produce a negative bag. Recently, PSD (Zhang et al. 2021b) proposes to learn 3D point affinity based on sparsely labeled points. Though the affinity can be precisely learned based on the supervised learning loss with sparsely labeled points, this method is not directly applicable to our setting, where only the scene-level category supervision is provided without any ground-truth point-level supervision. Therefore, these meth-ods with sparse labels mainly focus on relationship between labeled and unlabeled points, which is not applicable in our setting when point-level supervision is not available.

Weakly-supervised Point Cloud Segmentation with Scene-level Labels

Compared to weakly supervised point cloud segmentation based on sparse point-level annotations, the methods by scene-level annotations are less exploited. The state-of-the-art methods generally generate pseudo labels in the first step, and then refine the segmentation results via self-training. Two types of strategies are commonly used for generating pseudo labels by previous methods: class activation maps (CAM)-based and multi-instance learning (MIL)-based. Wei *et al.*(Wei et al. 2020) propose a Multi-Path Region Min-ing model involving spatial, channel, and point-wise paths to generate CAM of sub-cloud as pseudo. It is common to design two branches for a network and use consistency loss for self-supervised learning between the original point cloud and the augmented(Ren et al. 2021) or perturbed (Zhang et al. 2021a) point clouds. Similarly, Ren *et al.*(Ren et al. 2021) propose to jointly learn semantic segmentation, 3D proposal generation, and 3D object detection in a two-branch framework. However, current methods do not effec-tively leverage the complementary RGB and geometric in-formation present in point cloud data. Additionally, these methods often overlook the long-tail distribution of differ-ent categories, resulting in suboptimal performance.

Methodology

Given a set of M point clouds with scene-level annotations: $D = \{P_m, \boldsymbol{y}_m\}_{m=1}^M$, where $P_m \in \mathbb{R}^{N \times (3+K)}$ denotes the mth point cloud and $\boldsymbol{y}_m \in \{0,1\}^C$ is a C-dimensional bi-nary vector indicating which categories are present in point cloud P_m, we aim to derive a segmentation model, which classifies each point into one of the C categories. Each $P_m \in \mathbb{R}^{N \times (3+K)}$ represents a whole 3D scene with N 3D coordinates together with K-dimensional auxiliary features, such RGB, object normals, and height.

Overview. As shown in Figure 1, our framework consists of three main modules: a *feature extraction module* and a *segmentation module*, and a *multi-modality affinity infer-ence module* to enhance the segmentation results. Specifi-cally, our **feature extraction module** hierarchically extract multiple scales of point set features by gradually grouping point features in local regions. In the **multi-modality affin-ity inference module**, to differentiate between pure geomet-ric data and RGB-enriched data, we employ a masking tech-nique to exclude the RGB features from the input data, re-sulting in the pure geometric data. Simultaneously, we re-tain the original point cloud data as the RGB-enriched data. The two types of input data are fed into the shared backbone network to obtain the respective point affinity of the cor-responding modality. The **segmentation module** then uti-lizes the learned point affinity to refine the MIL (Multi-ple Instance Learning) objective, as well as the point-level pseudo-labels. These refinements are instrumental in guid-ing the self-training process of semantic segmentation.

MIL-based 3D Semantic Segmentation

We use a MIL-based weakly-supervised 3D semantic seg-mentation model as our baseline model, which consists of a
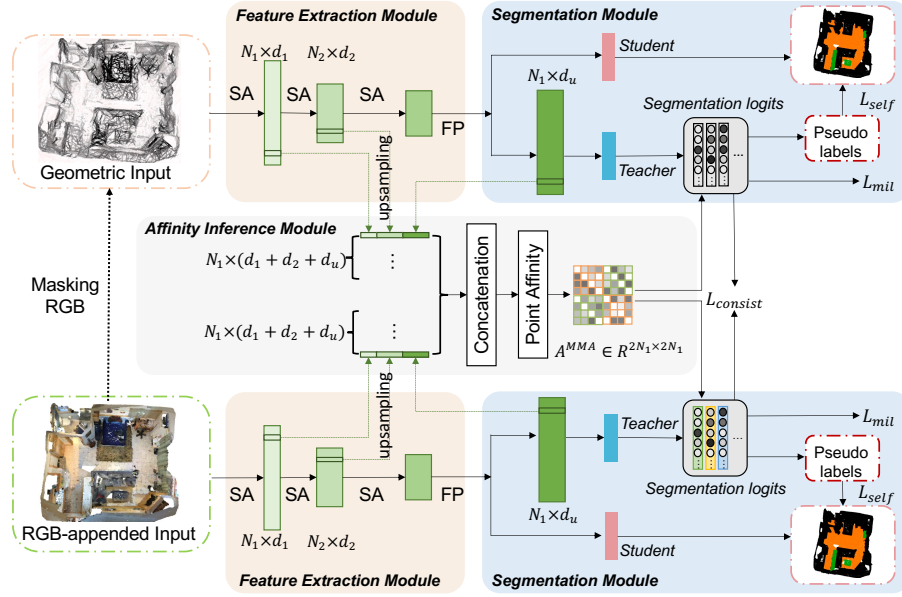
Figure 1. Pipeline of the proposed method. A two-stream architecture with shared parameters is adopted, where the two streams take pure geometric point clouds and RGB-appended point clouds as inputs, respectively. Three main modules are involved: a *feature extraction module*, a *segmentation module*, and a *multi-modality affinity inference module*.

feature extraction network and a segmentation head.

Feature Extraction Module. In this work, we choose our backbone network based on PointNet++ (Qi et al. 2017b) to hierarchically extract multiple scales of point set features by gradually grouping point features in local regions. The four set abstraction (SA) layers down-sample the point cloud from N points to $N_1 = 2048$, $N_2 = 1024$, $N_3 = 512$, and $N_4 = 256$ points, respectively by each layer. Two feature propagation (FP) layers then up-sample the points from N_4 to N_2 points. Note that other backbone networks that extract multiple scales of point features can also be selected as the feature extraction network.

Segmentation Module. Inspired by (Ren et al. 2021), we use two segmentation heads to process extracted features from the backbone networks. Each head predicts a segmentation logits matrix over C classes for all points, denoted as $U_{seg} \in \mathbb{R}^{N_1 \times C}$ and $S_{seg} \in \mathbb{R}^{N_1 \times C}$, respectively. The first head is to find the semantically discriminative points through a MIL-loss and produce pseudo labels by the points with highly confident predictions, which is denoted as the teacher head. The pseudo labels are generated by following (Ren et al. 2021). The second head takes the pseudo labels produced by the first head for self-training, which is denoted as the student head. However, we use a slightly different structure for the teacher and student segmentation heads. Specifically, the teacher head contain one FP layer to up-sample the points from $N_2 = 1024$ to $N_1 = 2048$ and one fully-connected (FC) layer to predict the per-point segmentation logits. The student head contains one FP layer and is followed by a one-layer Transformer Encoder (Vaswani et al. 2017). The motivation for adding a Transformer layer to the student head is two-fold. The first is to borrow the self-

attention module in Transformer for capturing long-range dependencies to ensure object completeness. The second is to explicitly enforce the teacher head and the student head to learn different features such that the pseudo labels produced by the teacher head could benefit more on the student head, which is inspired by co-training (Blum and Mitchell 1998).

Loss Function. The loss function of our weakly supervised semantic segmentation module is defined as,

$$\mathcal{L}_{wsss} = \mathcal{L}_{mil} + \mathcal{L}_{self}, \tag{1}$$

where $\mathcal{L}_{mil}$ is the MIL loss and $\mathcal{L}_{self}$ is the self-training loss. We define the scene-level MIL loss as,

$$\mathcal{L}_{mil} = -\sum_{c=1}^{C} (\boldsymbol{y}[c] \log \boldsymbol{\sigma}[c] - (1 - \boldsymbol{y}[c]) log(1 - \boldsymbol{\sigma}[c])), \tag{2}$$

where $\boldsymbol{\sigma}[c] = sigmoid(\frac{1}{N_1} \sum_{i=1}^{N_1} U_{seg}[i,c])$ converts the per-point logits U_{seg} into a scene-level prediction $\boldsymbol{\sigma}$ via average pooling and a sigmoid activation. The self-training loss for each scene is formulated by

$$\mathcal{L}_{self} = -\frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{c=1}^{C} \hat{Y}[i,c] \log \psi[i,c], \tag{3}$$

where $\psi[i,c] = softmax(S_{seg}[i,c])$ denotes the probability of the ith point being predicted as class c, N_1 denotes the number of points in U_{seg}, and $\hat{Y}[i,c] \in \{0,1\}$ is the point-wise pseudo label generated by the teacher head (Ren et al. 2021) by selecting the high confident points within the scene categories.

Long-tail-aware Multi-modality Point Affinity

We leverage point features from various data modalities to enhance their similarity, while eliminating the influence of long-tail distribution.

Multi-modality Point Affinity. We argue that geometric information from point cloud and color information encoded in RGB data can characterize the feature affinity from different perspectives by taking advantages of different data modalities. Moreover, to make the most of the multi-scale features extracted by the backbone, we incorporate features from multiple layers of feature extractor. We obtain the RGB-appended point affinity by concatenating multiple scales of point features $F \in \mathbb{R}^{N_1 \times D} = concate(F_1, F_2, F_u)$, where $F_1 \in \mathbb{R}^{N_1 \times d_1}, F_2 \in \mathbb{R}^{N_1 \times d_2}, F_u \in \mathbb{R}^{N_1 \times d_u}$. F_u is obtained by teacher head, is a high-level abstract points representation. F_1 and F_2 indicate the multiple scales of features produced by different SA layers from the backbone network. Note that the N_1 points of F_2 is obtained by up-sampling via linear interpolation. We model the geometric affinity by masking out the RGB values from the input point cloud for simplicity. Similarly, we use $\tilde{F} \in \mathbb{R}^{N_1 \times D} = concate(\tilde{F}_1, \tilde{F}_2, \tilde{F}_u)$ to denote the multi-scale geometric features by aggregating multiple scales of features from RGB-masked point clouds. We define the multi-scale multi-modality features by concatenating the multi-scale geometric features and RGB-appended features as $F^M = [F \in \mathbb{R}^{N_1 \times D}; \tilde{F} \in \mathbb{R}^{N_1 \times D}]$ and $F^M \in \mathbb{R}^{2N_1 \times D}$. Thus, our final multi-scale multi-modality affinity matrix is defined as,

$$A^{MMA}[i,j] = \max\left(\theta, \frac{\langle F^M[i,\cdot], F^M[j,\cdot] \rangle}{\|F^M[i,\cdot]\|\|F^M[j,\cdot]\|}\right), \quad (4)$$

where $i, j = 1, 2, ..., 2N_1$, and θ is the threshold for filtering out less confident affinities. Since the multi-scale multi-modality affinity matrix $A^{MMA} \in \mathbb{R}^{2N_1 \times 2N_1}$ is defined based on features from multiple modalities, the similarity of point features across-modality is also considered in our affinity matrix.

Long-tail Aware Affinity Enhancement. In point cloud-based semantic segmentation, the long-tailed distribution of data among classes happens at both the category level and point level. For example, in indoor scenarios, the category "wall" and "floor" generally not only appear in almost all of the scenes but also contain a larger amount of points in each scene compared to other objects, which are termed head classes in long-tailed distribution. The long-tailed distribution leads to the classifier weights of the head classes are learned to have larger norms, resulting in greater logits for head classes in each sample. This can be attributed to that large classifier weights norms cause the gradient leans towards head classes during back propagation. This biases the network's learned features towards head categories. Since the affinity matrix is calculated based on feature similarity, the head classes data points will contribute more to the affinity values than the tail classes, which contradict to the fact that the within-class point features should have larger affinity values. Therefore, the long-tail issue not only affects the

final segmentation results but also is detrimental to affinity inference. Unfortunately, previous research has overlooked this issue, and utilizing affinity affected by the long-tail distribution may lead to error accumulation.

In WSSS, the point-level category distribution is not accessible with scene-level annotations. Thus, inspired by Decoupling Representation and Classifier (Kang et al. 2019), we enhance affinity inference by dealing with the long-tail issue via a simple method through normalizing classifier weight(NCW), which alleviates the long-tail issue without the need of the prior of category distribution. Formally, let $W = \{\boldsymbol{w_i}\} \in \mathbb{R}^{d \times C}$, where $\boldsymbol{w_i} \in \mathbb{R}^d$ are the classifier weights corresponding to class i of the teacher and student segmentation head. We normalize W to obtain $\widehat{W} = \{\widehat{\boldsymbol{w_i}}\}$ via $\widehat{\boldsymbol{w_i}} = \frac{\boldsymbol{w_i}}{\|\boldsymbol{w_i}\|}$, where $\|\cdot\|$ denotes the l_2 norm.

Objective Functions

With the learned multi-modality affinity matrix A^{MMA}, we obtain the refined segmentation logits matrix $U_{seg}^{refined}$ and $\tilde{U}_{seg}^{refined}$ of the teacher segmentation head by multiplying the original logits matrix by the affinity matrix:

$$\left[U_{seg}^{refined}; \tilde{U}_{seg}^{refined}\right] = A^{MMA}\left[U_{seg}; \tilde{U}_{seg}\right], \quad (5)$$

where $\tilde{U}_{seg} \in \mathbb{R}^{N_1 \times C}$ denotes the predicted segmentation logits matrix over C classes of the RGB-masked point cloud data $\tilde{P}$. The refined segmentation logits of a point aggregate information from the points with similar features both within and across modalities, which are considered to be able to improve the pseudo labels for self-training loss in Eq.(3).

To achieve message passing between multi-modality affinities and further improve the segmentation performance, we explicitly impose the prediction consistency constraint between the original point cloud data P and the RGB-masked point cloud data $\tilde{P}$ with horizontal transformation, which is inspired from previous WSSS methods that introduce different contrastive or consistency learning strategies by augmenting the original point cloud data (Yang et al. 2022; Ren et al. 2021; Ahn and Kwak 2018; Zhang et al. 2021b). The consistency loss is defined as,

$$\mathcal{L}_{consist} = \frac{1}{N_1}\sum_{i=1}^{N_1}\sum_{c=1}^{C}\left|U_{seg}^{refined}[i,c] - \tilde{U}_{seg}^{refined}[i,c]\right|. \quad (6)$$

Moreover, the refined self-training loss $\mathcal{L}_{self}^{refined}$ is defined:

$$\mathcal{L}_{self}^{refined} = -\frac{1}{N_1}\sum_{i=1}^{N_1}\sum_{c=1}^{C}\hat{Y}^{refined}[i,c]\log\psi[i,c], \quad (7)$$

where $\hat{Y}^{refined}[i,c] \in \{0,1\}$ is the point-wise pseudo label generated by $U_{seg}^{refined}$.

The final objective function of our method is

$$\mathcal{L} = \mathcal{L}_{mil} + \mathcal{L}_{self}^{refined} + \mathcal{L}_{consist}. \quad (8)$$

Experiments

Experimental Setting

Datasets and evaluation metrics. We evaluate the proposed approach MMA on two benchmarks, ScanNet (Dai

et al. 2017) and S3DIS (Armeni et al. 2017) datasets. Scan-Net is a commonly-used indoor 3D point cloud dataset for semantic segmentation. It contains 1513 training scenes (1201 scenes for training, 312 scenes for validation) and 100 test scenes, annotated with 20 classes. S3DIS is also an indoor 3D point cloud dataset, which contains 6 indoor areas and has 13 classes. By following the previous work, we use area 5 as the test data. Mean intersection over union (mIoU) is used as the evaluation metric of the segmentation results.

Implementation details. The RGB-appended input point clouds are a set of 10-dimensional vectors, including coordinates (x,y,z), color (R,G,B), surface normal, and height, while the pure geometric input is produced by masking out the RGB values with 0. PointNet++ (Qi et al. 2017a) is adopted as the backbone feature extraction module to extract point cloud features. The segmentation module has two segmentation heads: a teacher head for providing pseudo labels with MIL loss and a student head for self-training. The teacher head is a multi-label classification model, which contains one FP layer to upsample the points to 2048 and one fully-connected (FC) layer to predict per-point logit and then through average to get per-class logit. The student head contains one FP layer to upsample points and follows one Transformer Encoder layer to capture the long-range dependencies of points, then an FC layer to predict per-point logit. Multi-scale module use features obtained from the first and second SA layers and U_{seg} FP layer. The details of the architecture of teacher head and student head can be found in the supplementary material. The model is trained on 3090 GPU with batch size 8 for 300 epochs. We use AdamW optimizer with an initial learning rate of 0.0014 and decay to half at 160 epochs and 180 epochs. All hyper-parameters are tuned based on the validation set.

Comparison with State-of-the-arts

We mainly compare our approach to other 3D weakly supervised segmentation methods utilizing scene labels, including MPRM (Wei et al. 2020), WyPR (Ren et al. 2021), and MIL-Derived (Yang et al. 2022). This type of supervision is challenging for large-scale point cloud datasets. MPRM uses various attention modules to mine local and global context information. WyPR joint learning of segmentation and detection to get a better feature representation, and gain high performance. MIL-Derived proposes a transformer model to explore pair-wise cloud-level supervision, where two clouds of the same category yield a positive bag while two of different classes produce a negative bag.

Results on ScanNet. Table 1 reports the mIoU results of the proposed method and the state-of-the-art baseline methods . It can be seen that our proposed method performs better than existing methods (MPRM (Wei et al. 2020), WyPR (Ren et al. 2021), MIL-Derived (Yang et al. 2022)) by large margins(+15.8%, +8.1%, +11.5%) on the ScanNet validation set. And our method outperforms WyPR by 6.6% in terms of the test mIoU. In Table 3, we report the per-class IoU on ScanNet, including PCAM (Wei et al. 2020), MPRM (Wei et al. 2020), WyPR (Ren et al.

Method	Supervision	Val.	Test
PointNet++ (Qi et al. 2017b)	Full	-	33.9
PointCNN (Li et al. 2018)	Full	-	45.8
KPConv (Thomas et al. 2019)	Full	-	68.4
MinkNet (Choy, Gwak, and Savarese 2019)	Full	-	73.6
MPRM (Wei et al. 2020)	Scene	21.9	-
WyPR (Ren et al. 2021)	Scene	29.6	24
WyPR+prior (Ren et al. 2021)	Scene	31.1	-
MIL-Derived (Yang et al. 2022)	Scene	26.2	-
Ours (MMA)	Scene	**37.7**	**30.6**

Table 1. 3D semantic segmentation on ScanNet.

2021). Obviously, the proposed MMA achieves the highest mIoU, and significantly improves the performance in "floor", "chair", "sofa", "table", "shelf", "desk", "toilet" and "bathtub" against WyPR. These categories are often co-occurred and easily mis-classified. In addition, our "MMA" as shown in Table 3 improves the performance of objects with either discriminative geometric or color information, such as "floor", "chair", "table", "curtain", "shower curtain", and "bathtub". The Multi-modality affinity in MMA takes advantage of both geometric and color information and thus improves the performance to a large margin.

Method	Supervision	Test
PointNet++ (Qi et al. 2017b)	Full	53.5
PointCNN (Li et al. 2018)	Full	57.3
KPConv (Thomas et al. 2019)	Full	70.6
MinkNet (Choy, Gwak, and Savarese 2019)	Full	65.4
MPRM (Wei et al. 2020)	Scene	10.3
WyPR (Ren et al. 2021)	Scene	22.3
MIL-Derived (Yang et al. 2022)	Scene	12.9
Ours (MMA)	Scene	**26.3**

Table 2. 3D semantic segmentation on S3DIS.

Results on S3DIS. Table 2 shows the S3DIS results of the proposed method and the baseline methods. It can be seen that our proposed method achieves much higher mIoU scores, and outperforms MPRM, MIL-Derived and WyPR with gains of 16.0%, 13.4% and 4.0%.

Discussion. When compared with the state-of-the-art methods, our method owns different advantages. Firstly, MPRM (Wei et al. 2020) proposes various attention modules to refine the point features by local and global context information. The attention modules are only applied to the output features of the backbone network, which might fail to capture multi-scale attentions. Moreover, the multiple modalities of RGB-D data are not explicitly exploited. By contrast, our MMA affinity takes both multi-scale and multi-modality similarities into account and the obtained MMA affinity matrix is directly applied to the segmentation logits the refine the segmentation results, which directly improve the pseudo labels for self-training. Secondly, WyPR (Ren et al. 2021) achieves better results than MPRM. However, the good results are achieved by jointly training with 3D object detection, which relies on a costly selective search step. Differently, our method with a single segmentation

Method	eval.	mIoU	wall	floor	cabinet	bed	chair	sofa	table	door	window	shelf	picture	counter	desk	curtain	fridge	sc	toilet	sink	bathtub	other
PCAM	train	22.1	54.9	48.3	**14.1**	34.7	32.9	45.3	26.1	0.6	3.3	46.5	0.6	6.0	7.4	26.9	0	6.1	22.3	8.2	52.0	6.1
MPRM	train	24.4	47.3	41.1	10.4	43.2	25.2	43.1	21.5	9.8	12.3	45.0	9.0	13.9	21.1	40.9	1.8	29.4	14.3	9.2	39.9	10.0
WyPR	train	30.7	**59.3**	31.5	6.4	58.3	31.6	47.5	18.3	17.9	**36.7**	34.1	6.2	**36.1**	24.3	**67.2**	**8.7**	38	17.9	**28.9**	35.9	8.2
Ours(MMA)	train	**41.2**	52	**84.2**	13.2	**63.2**	**53.1**	**62.4**	**41.1**	**18.0**	31.1	**46.6**	**10.2**	24.6	**43**	65.8	6.6	**59.8**	**43**	24.3	**64.3**	**17.6**
WyPR	val	29.6	**58.1**	33.9	5.6	56.6	29.1	45.5	19.3	15.2	**34.2**	33.7	6.8	**33.3**	22.1	65.6	6.6	36.3	18.6	24.5	39.8	6.6
WyPR+prior	val	31.1	52.0	77.1	6.6	54.3	35.2	40.9	29.6	9.3	28.7	33.3	4.8	26.6	27.9	**69.4**	**8.1**	27.9	24.1	25.4	32.3	8.7
Ours(MMA)	val	**37.7**	49.8	**84.3**	**11.1**	**56.8**	**54.3**	**51.6**	**45.2**	**16.0**	22.2	**42.0**	**6.8**	28.6	**37.2**	59.9	4.2	**43.0**	**41.6**	**27.4**	**59.7**	**12.3**

Table 3. 3D semantic segmentation result for 20 classes on ScanNet. Here sc refers to the 'shower curtain' class

task outperforms WyPR to a large margin. Lastly, the MIL-derived Transformer (Yang et al. 2022) modeling the similarities across scenes to improve the weak labels. However, the cross-scene similarity might be hard to model due to the large cross-scene variations. Moreover, our method can be a complementary to MIL-derived Transformer by exploiting similarities from different perspectives.

Ablation Study and Further Analysis

We report the results of ablation study to show the effectiveness of each components of our method. We also conduct further analysis by qualitative results. The experiments in this section are conducted on ScanNet validation set.

Contributions of Components. We report the results of ablation study to demonstrate the contribution of different components of our method. The results are shown in Table 4 with 8 experiments denoted by "A.#". The baseline method in "A.1" is our MIL-based segmentation model as described in Section , which consists of a MIL-loss, a self-training loss, and a cross transformation consistency loss between the original point cloud and the geometrically augmented point cloud. The baseline method achieves 23.8% mIoU. Then we add the multi-modality affinity to the baseline model in "A.3", which achieves obtains 28.4% with 4.6% performance gain compared to the baseline. To alleviate the imbalance issue that may affect both the affinity matrix and the segmentation results, the classifier normalization is added and the performance achieves 36.2% as shown in "A.4", which significantly improve the results. "A.8" is our final results by using the proposed MMA-refined model, which further improve "A.4" by using additional multi-scale information and achieves the best results. The results also show that though simply introducing NCW (baseline + NCW in "A.2") can improve the baseline result (in "A.1") by 3%, "A.8" (resp., "A.4") can further improve "A.6" (resp., "A.3") by nearly 8%, which verifies that NCW not only balances the classification results but also enhances the point affinity. In other words, NCW only works significantly well when jointly working with the proposed MMA module. Even without NCW, we add multi-modality affinity and multi-scale affinity to baseline in "A.6" can achieve 30.1% with 6.3% performance gain compared to the baseline. Adding multi-scale affinity to the baseline in "A.2" can improve baseline result by 1.9%. To validate the effectiveness of the additional light-weight Transformer block in our student segmentation head, we conduct experiments by re-

moving the Transformer block as shown in "A.7", the performance drops from 37.7% to 33.9%. To sum up, all the components in our method contribute to the final results.

	Baseline	Multi-modality Aff.	Multi-scale Aff.	$\widehat{w}$	results
A.1	✓				23.8
A.2	✓			✓	26.9
A.3	✓	✓			28.4
A.4	✓	✓		✓	36.2
A.5	✓		✓		25.7
A.6	✓	✓	✓		30.1
A.7	w/o Trans.	✓	✓	✓	33.9
A.8	✓	✓	✓	✓	**37.7**

Table 4. Ablation study on ScanNet validation set.

Qualitative Results. We qualitatively illustrate more segmentation results in Figure 2, where the columns indicate:(a) Point cloud data, (b) Ground-Truth segmentation results,(c) Baseline results, (d) Results of "Baseline+Multi-modality affinity", (e) Results of our final method. From the results, we can find that our final method achieves the best results. For the baseline method (Figure 2(c)), different objects are not clearly distinguished with clear boundary, and the semantic categories for many objects are mis-classified.

In our final model with the additional multi-modality affinity module (Figure 2 (e)), the discriminative information from both geometric modality and color modality can both benefit the final results. Although the original RGB-appended point cloud data are more informative, the RGB data and the geometric data are entangled, and thus the complementary multi-modality information is hard to be fully exploited. For example, in the RGB-appended point cloud data, the rich geometric data with shape and scale information might be overwhelmed by the RGB information that suffers from lighting conditions, shadows, and reflections in many cases. As shown in Figure 2 #2 and #3, "floor" is wrongly segmented in our single-stream baseline variant due to the serious reflections in the RGB data. By contrast, our multi-modality affinity (Figure 2 (d)) successfully corrects the results thanks to the enhanced geometric information. In addition, the "floor" is geometrically located with smaller "z" value and smaller height, which is distinguished from the objects placed on it with higher height (such as "table" and "chair") as shown in most of the scenes such as #1, #2, and #3. For the "door" and "wall" classes in scene #1, we observe that by taking advantages from the discriminative color information, our method can better distinguish "door"
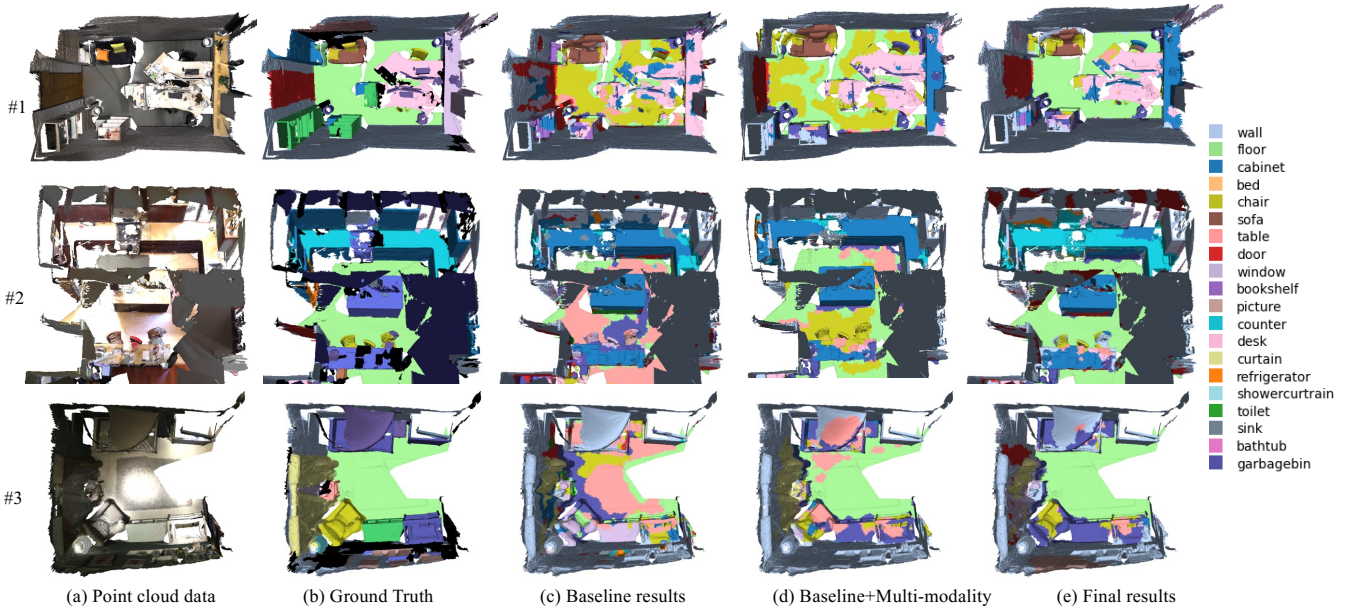
Figure 2. Qualitative segmentation results: (a) Point cloud data, (b) Ground-truth segmentation results, (c) Baseline results, (d) Results of "Baseline+Multi-modality affinity w/o $\hat{W}$" variant of our method, (e) Results of our final method ("Baseline+Multi-modality affinity").

from "wall" when compared with the baselines.

Analysis of Multi-scale Affinity. In Table 5, we further evaluate the effectiveness of multi-modality multi-scale affinity. Specifically, the "F_u^M only" variant is the multi-modality affinity-only baseline. The "$concate(F_u^M, F_1^M)$" method concatenates the F_u^M features with F_1^M (i.e., features from first SA layer). The "$concate(F_u^M, F_1^M, F_2^M)$" method concatenates the F_u^M features with both the first SA layer features F_1^M and the second SA layer feature F_2^M. Based on multi-modality, multi-scale can further enhance the model's performance. Therefore, it is necessary for multi-modality and multi-scale to work in conjunction with each other.

Method	mIoU(%)
F_u^M	36.2
$concate(F_u^M, F_1^M)$	36.6
$concate(F_u^M, F_1^M, F_2^M)$	**37.7**

Table 5. Evaluation of the effectiveness of multi-scale affinity on ScanNet validation set. $concate(\cdot, \cdot)$ denotes concatenation.

Analysis of Normalizing Classifier Weights for Affinity. To demonstrate the effectiveness of the Normalizing Classifier Weights (NCW) module on improving affinity, we conducted a comparative analysis between the MMA w/o $\widehat{W}$ affinity, MMA affinity. We split the ScanNet classes into three groups by the number of points in each category in training samples: head classes each contains over 8 million points, medium classes each has between 1.2 million

and 8 million points, and tail classes with under 0.3 million points. We demonstrate in supplementary materials the categories included in the head, medium, and tail of ScanNet, respectively. We evaluate mAP and mIoU for each subset, the results are shown in Table 6, where Δ represents the relative performance difference between MMA and MMA w/o $\widehat{W}$. mAP reflects the performance of multi-label classification, which impact the quality of generated pseudo-labels. We observed that MMA w/o $\widehat{W}$ affinity can significantly enhance the performance of medium and tail classes while maintaining the effectiveness of the head set. Notably, there was a 7.2% performance gain in the tail set. Due to the improved classification performance of the medium and tail categories, the points previously wrongly categorized as head class have been correctly classified to the medium and tail categories. Thus, the mIoU for all splits are improved.

method	mAP			mIoU		
	Head	Medium	Tail	Head	Medium	Tail
MMA w/o $\widehat{W}$	100.0	86.4	65.7	55.3	18.8	30.6
MMA	100.0	88.3	72.9	68.6	29.6	36.0
Δ	**0**	**1.9**	**7.2**	**13.3**	**10.8**	**5.4**

Table 6. mAP, mIoU of MMA and MMA w/o $\widehat{W}$ on ScanNet

Visualization of Point Class Relationship Maps Enhanced by Point Affinity. We further visualize the binary map of the class relationship enhanced by point affinity in Figure 3. If two points are predicted as the same class, the value is 1, otherwise the value is 0. The columns indicate the affinity matrices produced based on: (a) Ground-
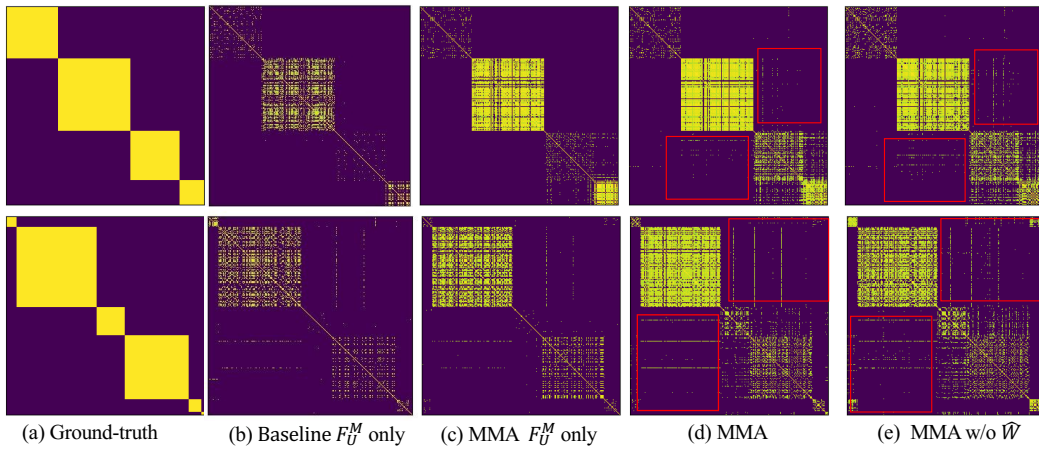
| (a) Ground-truth | (b) Baseline F_u^M only | (c) MMA F_u^M only | (d) MMA | (e) MMA w/o $\widehat{W}$ |

Figure 3. Class relationship maps enhanced by point affinity.

truth, (b) Baseline F_u^M only, (c) MMA F_u^M only, (d) MMA (i.e., $concate(F_u^M, F_1^M, F_2^M)$), (e) MMA w/o $\widehat{W}$. In Figure 3(b), different categories get confused because the affinities between points of the same class are not higher than those between points of different classes. In contrast, Figure 3(c) shows much better results than (b). This suggests that our MMA approach enhances the intra-class similarity by fully exploiting multiple data modalities and considering the long-tail issues. Figure 3(c) and (d) demonstrate that our multi-scale features can further improve the affinity of objects from different scales. By comparing Figure 3(d) and (e), the affinities of points from different classes in (d) with the proposed long-tail aware normalization is much smaller than that in (e), especially in the highlighted red boxes, which is the key to learn discriminative features.

Conclusion

This paper proposes a novel multi-modality point affinity (MMA) enhanced weakly supervised 3D semantic segmentation method. The proposed MMA considers the point feature similarities by taking advantage of complementary information from different data modalities. The point affinity is also enhanced by normalizing the classifier weights to alleviate the detrimental effects of long-tailed data distribution to the affinity matrix. Extensive experiments demonstrate the effectiveness of the proposed method on two commonly used indoor scene understanding benchmark datasets.

Acknowledgments

References

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4981–4990.

Armeni, I.; Sax, S.; Zamir, A. R.; and Savarese, S. 2017. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*.

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

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, 549–565. Springer.

Blum, A.; and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, 92–100.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3075–3084.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

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

Hou, J.; Graham, B.; Nießner, M.; and Xie, S. 2021. Exploring data-efficient 3d scene understanding with contrastive

scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15587–15597.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11108–11117.

Jamal, M. A.; Brown, M.; Yang, M.-H.; Wang, L.; and Gong, B. 2020. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7610–7619.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; and Jia, J. 2022. Stratified Transformer for 3D Point Cloud Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8500–8509.

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

Liu, Z.; Qi, X.; and Fu, C.-W. 2021. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1726–1736.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

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

Ren, Z.; Misra, I.; Schwing, A. G.; and Girdhar, R. 2021. 3d spatial recognition without spatially labeled 3d. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13204–13213.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019a. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10296–10305.

Wang, T.; Zhu, Y.; Zhao, C.; Zeng, W.; Wang, J.; and Tang, M. 2021. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3103–3112.

Wang, W.; Zhao, Z.; Wang, P.; Su, F.; and Meng, H. 2022. Attentive Feature Augmentation for Long-Tailed Visual Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 5803–5816.

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

Wei, J.; Lin, G.; Yap, K.-H.; Hung, T.-Y.; and Xie, L. 2020. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4384–4393.

Wu, W.; Qi, Z.; and Fuxin, L. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9621–9630.

Xu, M.; Ding, R.; Zhao, H.; and Qi, X. 2021. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3173–3182.

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

Yang, C.-K.; Wu, J.-J.; Chen, K.-S.; Chuang, Y.-Y.; and Lin, Y.-Y. 2022. An MIL-Derived Transformer for Weakly Supervised Point Cloud Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11830–11839.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021a. Weakly supervised semantic segmentation for large-scale point cloud. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4): 3421–3429.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021b. Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15500–15508.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.