

Learning Inter-Superpoint Affinity for Weakly Supervised 3D Instance Segmentation

Linghua Tang, Le Hui, and Jin Xie^(✉)

Nanjing University of Science and Technology, Nanjing, China
{tanglinghua, le.hui, csjxie}@njjust.edu.cn

Abstract. Due to the few annotated labels of 3D point clouds, how to learn discriminative features of point clouds to segment object instances is a challenging problem. In this paper, we propose a simple yet effective 3D instance segmentation framework that can achieve good performance by annotating only one point for each instance. Specifically, to tackle extremely few labels for instance segmentation, we first over-segment the point cloud into superpoints in an unsupervised manner and extend the point-level annotations to the superpoint level. Then, based on the superpoint graph, we propose an inter-superpoint affinity mining module that considers the semantic and spatial relations to adaptively learn inter-superpoint affinity to generate high-quality pseudo labels via semantic-aware random walk. Finally, we propose a volume-aware instance refinement module to segment high-quality instances by applying volume constraints of objects in clustering on the superpoint graph. Extensive experiments on the ScanNet-v2 and S3DIS datasets demonstrate that our method achieves state-of-the-art performance in the weakly supervised point cloud instance segmentation task, and even outperforms some fully supervised methods. Source code is available at <https://github.com/fpthink/3D-WSIS>.

1 Introduction

Point cloud instance segmentation is a classic task in 3D computer vision, and it can be applied in many fields, including indoor navigation systems, augmented reality, and robotics. The fully supervised instance segmentation methods [17,2,12] have achieved impressive results, but they rely on numerous manually labeled data. However, annotating a large number of point clouds is extremely time-consuming and expensive. Thus, it is meaningful to segment point clouds in a semi-/weakly supervised manner that requires a small number of annotations. However, how to fully exploit the limited labels to improve the performance of instance segmentation is still a challenging problem.

Few efforts have been dedicated to semi-/weakly supervised point cloud instance segmentation. As a pioneer, Liao *et al.* [18] proposed a semi-supervised point cloud instance segmentation method using bounding boxes as supervision, where a network is used to generate bounding box proposals. And instance segmentation is achieved by refining the point cloud within the bounding box

proposals. Besides, Tao *et al.* [24] proposed a two-stage seg-level supervision 3D instance and semantic segmentation method, which first leverages a segment grouping network to generate pseudo labels for the whole scenes, and then the generated pseudo point-level labels are used as the ground truth to train the network. However, these simple pseudo label generation strategies cannot effectively generate high-quality pseudo labels, resulting in poor 3D instance segmentation results.

In this paper, we propose a simple yet effective weakly supervised 3D instance segmentation framework, which can achieve impressive results with one point annotation per instance. For weakly supervised point cloud instance segmentation with few annotated labels, our intuition lies in two folds: (1) Under rare annotations, effective label propagation is essential to produce high-quality pseudo labels, especially in 3D instance segmentation. (2) Weakly supervised 3D instance segmentation is more challenging than weakly supervised 3D semantic segmentation, so we consider introducing the object volume constraint to improve the instance segmentation results. Specifically, we first use an unsupervised method [14] to oversegment the point cloud into superpoints and build the superpoint graph. In this way, point-level labels can be extended to superpoint-level labels. Then, we propose an inter-superpoint affinity mining module to generate high-quality pseudo labels based on a few annotated superpoint-level labels. Based on the superpoint graph, we leverage the semantic and spatial information of adjacent superpoints to adaptively learn inter-superpoint affinity, which can be used to propagate superpoint labels along the superpoint graph via semantic-aware random walk. Finally, we propose a volume-aware instance refinement module to improve instance segmentation performance. Based on the trained model using superpoint-level propagation, we can obtain coarse instance segmentation results through superpoint clustering and further infer the object volume information from the instance segmentation results. The object volume information contains the number of voxels and the radius of the object. The inferred object volume information is regarded as the ground truth of the corresponding instance to retrain the network. In the test phase, based on the object volume information, we utilize the predicted object volume information to introduce a volume-aware instance clustering algorithm for segmenting high-quality instances. Extensive experiments on the ScanNet-v2 [6] and S3DIS [1] datasets can demonstrate the effectiveness of our method.

The main contributions of our paper are as follows:

- We present an inter-superpoint affinity mining module that considers the semantic and spatial relation to adaptively learn inter-superpoint affinity for random-walk based label propagation.
- We present a volume-aware instance refinement module, which guides the superpoint clustering on the superpoint graph to segment instances by using the object volume information.
- Our simple yet effective framework achieves state-of-the-art weakly supervised 3D instance segmentation performance on popular datasets ScanNet-v2 and S3DIS.

2 Related Work

2.1 3D Semantic Segmentation

Fully supervised 3D semantic segmentation. Many methods have been proposed to achieve point cloud semantic segmentation. Some methods [15,25,11] project point clouds into a series of regular 2D images from different views, and then fuse features extracted through 2D convolutional neural networks (CNNs). To apply 3D CNNs on the irregular point cloud and alleviate large memory costs, many efforts [8,5] first voxelize the point cloud into voxels and then utilize the sparse convolutional neural network to extract features of the point cloud. PointNet [20] directly extracts features from points with shared multi-layer perceptrons and max-pooling layer. Inspired by PointNet, different local feature aggregation operators [21,26,32,4] are proposed to work on point cloud, which directly consume point cloud. Besides, various methods [30,10] capture intrinsic spatial and geometric features by constructing the graph on the point cloud. Various approaches exploit different local feature aggregation networks to extract discriminative point features and use multi-layer perceptrons to achieve 3D semantic segmentation.

Semi-/Weakly supervised 3D semantic segmentation. Inspired by class activation map in 2D images, Wei *et al.* [31] introduce a multi-path region mining module to generate pseudo labels, which only requires cloud-level weak labels. Xu *et al.* [33] use three additional losses to constrain on unlabeled points, achieving impressive performance with 10% labels. Cheng *et al.* [3] use a dynamic label propagation strategy to generate pseudo labels, and learn discriminative features with a coupled attention module. Zhang *et al.* [36] exploit the consistency generated by perturbation to obtain additional supervision and propagate implicit labels by constructing the graph topology of the point cloud. Liu *et al.* [19] first build a supervoxel graph on the point cloud and then conduct label propagation by learning the similarity among graph nodes. Li *et al.* [16] utilize a hybrid contrastive regularization strategy with point cloud augmentation to provide additional constraints for network training. To generate pseudo labels for outdoor point cloud scenes, Shi *et al.* [22] design a matching module to propagate pseudo labels in both temporal and spatial spaces.

2.2 3D Instance Segmentation

Fully supervised 3D instance segmentation. Compared with point cloud semantic segmentation, instance segmentation is more challenging because it not only requires predicting semantic scores but also distinguishing instances of the same class. According to the different manners of generating instances, instance segmentation methods can be mainly divided into clustering-based methods and proposal-based methods. Given point clouds as input, clustering-based methods regard instance segmentation as the post-processing task after network inference, and the result is obtained by clustering on point clouds with the predicted features. As a pioneer, Wang *et al.* [28] introduce a similarity matrix to measure

the distances between the features of all point pairs, which guides clustering points as proposals. Wang *et al.* [29] integrate semantic and instance segmentation into a parallel framework, which benefits from each other task. Lahoud *et al.* [13] design a multi-task neural network architecture, where instances are simultaneously separated in the feature vector space and direction vector space by a discriminative loss [7] and a directional loss. Jiang *et al.* [12] generate proposals by clustering points on the original and offset-shifted coordinate spaces, which benefits from both advantages. Hou *et al.* [9] jointly learn color and geometry features for instance segmentation from different modalities. Lately, Chen *et al.* [2] introduce a hierarchical aggregation method that iteratively clusters point clouds into instance proposals. Liang *et al.* [17] propose a semantic superpoint tree structure and achieved instance segmentation by tree traversal and splitting. Vu *et al.* [27] design a soft group algorithm to reduce the semantic prediction errors to significantly boost the segmentation performance.

For proposal-based methods, instance segmentation consists of two procedures, first generating rough proposals and then predicting precise instance masks. Yang *et al.* [34] propose an end-to-end trainable network which directly generates 3D bounding boxes as proposals and infers point-wise instance masks for points inside proposals. Instead of getting proposals via 3D bounding box regression, Yi *et al.* [35] introduce an approach to obtain proposals by object generation, and then predict instance masks within proposals.

Semi-/Weakly supervised 3D instance segmentation. Few efforts have been made on semi-/weakly supervised point cloud instance segmentation. Tao *et al.* [24] propose a method to generate pseudo labels for the whole training scene and the generated pseudo point-level labels are used to train existing full supervised methods for point cloud instance segmentation, where one point per instance is clicked as the weak label. Nonetheless, the quality of pseudo labels is limited due to lack of learning discriminative instance features. With bounding boxes as weak labels, Liao *et al.* [18] propose a semi-supervised point cloud instance segmentation method, where a network is leveraged to generate bounding box proposals and instance segmentation is achieved by refining points within bounding box proposals.

3 Method

The overall architecture of our method is depicted in Fig. 1. The backbone network (Sec. 3.1) first takes the point cloud and superpoint graph as input and predicts superpoint-wise semantic labels and offset vectors. Then, the inter-superpoint affinity mining module (Sec. 3.2) propagates labels on the superpoint graph via semantic-aware random walk. Finally, the volume-aware instance refinement module (Sec. 3.3) learns object volume information to improve instance segmentation performance.

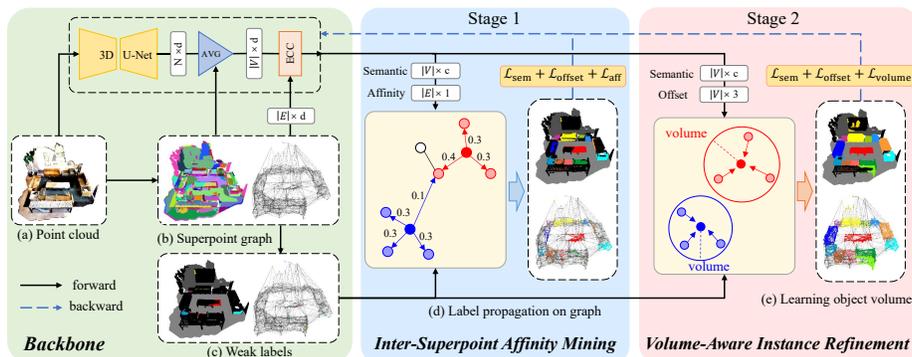


Fig. 1. Overview of our framework for weakly supervised 3D instance segmentation. We first oversegment point cloud (a) to build the superpoint graph (b) and extend weak labels (c) to the corresponding superpoint. Then, based on superpoint graph, the random walk with the predicted affinity and semantic is used for label propagation (d). Finally, combining the predicted semantic and offset, learning pseudo object volume (e) is achieved. c is the number of categories, d is the feature dimension, N is the number of points, $|V|$ is the number of superpoints, and $|E|$ is the number of edges.

3.1 Backbone Network

Superpoint graph construction. Following [14,17], we adopt an unsupervised point cloud oversegmentation method to generate the superpoints and construct the superpoint graph. The superpoint graph is a geometry representation of point clouds, defined as $G = (V, E)$. Vertice V means superpoints that are generated by aggregating points with similar geometric characteristics, and edge E indicates the prior connection relationship between adjacency superpoints which are constructed by linking k -nearest superpoints. In weakly supervised 3D instance segmentation, the benefits of the use of superpoints are two-fold. On one hand, the superpoint is a geometrically homogeneous unit, so we can extend annotated point label to the corresponding superpoint, thereby alleviating the sparsity of point-level annotations. On the other hand, the superpoint graph captures the spatial relationship of different instances so that we can utilize it to perform label propagation efficiently.

Superpoint feature extraction. Specifically, we first extract point features using 3D U-Net [8] on the point cloud and then aggregate point features into superpoint features by average pooling. After that, based on the superpoint graph, we use the edge-conditioned convolutions (ECC) [23] to extract superpoint features. Finally, we use the superpoint features to predict the semantic label of superpoints.

3.2 Inter-Superpoint Affinity Mining

To perform label propagation, we develop an inter-superpoint affinity mining module to learn the superpoint relationship in the semantic and coordinate s-

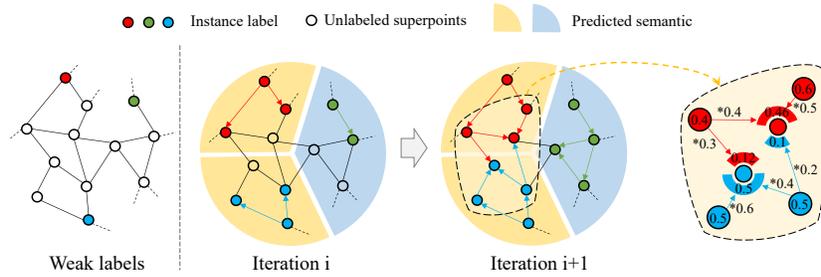


Fig. 2. The process of label propagation with predicted instance affinity and semantics.

paces. By using the learned inter-superpoint affinity, we design a simple semantic-aware random walk algorithm for label propagation on the superpoint graph.

Superpoint affinity learning. Based on the superpoint graph, we learn the relationship between two adjacent superpoints to characterize their affinity. It is desired that the learned affinity between two adjacent superpoints can guide the label propagation along the edge of the superpoint graph. Assuming the learned superpoint embedding in the backbone network is $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the number of superpoints and d is the feature dimension. Given the i -th superpoint embedding $\mathbf{X}_i \in \mathbb{R}^d$ and its first-order neighbors \mathcal{N}_i , we leverage the semantic and spatial information of the superpoints to adaptively learn inter-superpoint affinity. The affinity A_{ij} between the i -th superpoint and its j -th neighbor is formulated as:

$$A_{ij} = \frac{\exp(\sigma(\phi(\mathbf{X}_i), \psi(\mathbf{X}_j)) * \gamma(\mathbf{p}_i - \mathbf{p}_j))}{\sum_{k \in \mathcal{N}_i} \exp(\sigma(\phi(\mathbf{X}_i), \psi(\mathbf{X}_k)) * \gamma(\mathbf{p}_i - \mathbf{p}_k))} \quad (1)$$

where $\mathbf{X}_i \in \mathbb{R}^d$ and $\mathbf{X}_j \in \mathbb{R}^d$ are the superpoint embedding. $\mathbf{p}_i \in \mathbb{R}^3$ and $\mathbf{p}_j \in \mathbb{R}^3$ are the centroid coordinate of the superpoints. $\phi(\cdot)$ and $\psi(\cdot)$ are linear projections, and $\gamma(\cdot)$ is a multi-layer perceptron. $\sigma(\cdot, \cdot)$ is the dot production for learning the similarity of the i -th and g -th superpoints. In Eq. (1), semantic similarity is measured by dot production while spatial similarity is measured by subtraction. As a result, the affinity A_{ij} considers the semantic and spatial information of the superpoints. After that, we use the learned inter-superpoint affinity to update the superpoint embeddings. For the i -th superpoint, the new superpoint embedding $\tilde{\mathbf{X}}_i \in \mathbb{R}^d$ is written as:

$$\tilde{\mathbf{X}}_i = A_{ij} \cdot \rho(\mathbf{X}_j) + \mathbf{X}_i \quad (2)$$

where $\rho(\cdot)$ is linear projection. During the training, we employ a discriminative loss (dubbed \mathcal{L}_{aff}) used in [7] to draw $\tilde{\mathbf{X}}$ belonging to the same object towards each other, and make $\tilde{\mathbf{X}}$ in different objects away. It is expected that the affinity A_{ij} between the superpoints of the same instance can be enhanced.

Label propagation via semantic-aware random walk. After obtaining the inter-superpoint affinity $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$, we design a simple semantic-aware random walk to propagate labels over the superpoint graph, as shown in Fig. 2. Specifically, our semantic-aware random walk propagates labels over the superpoints graph with the same predicted semantic labels. For the c -th class, assume

that its semantic matrix is $\mathbf{S}^c \in \mathbb{R}^{|V| \times |V|}$. In \mathbf{S}^c , if the semantic class of the i -th and j -th superpoints are the same, then $S_{ij}^c = 1$, otherwise $S_{ij}^c = 0$. For label propagation, we first using the semantic matrix \mathbf{S}^c , superpoint affinity $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$ and adjacency matrix $\mathbf{M} \in \mathbb{R}^{|V| \times |V|}$ of the graph $G = (V, E)$ to compute the weight $\mathbf{P}^c \in \mathbb{R}^{|V| \times |V|}$ for the c -th class, which is formulated as:

$$\mathbf{P}^c = \mathbf{M} \odot \mathbf{S}^c \odot \mathbf{A} \quad (3)$$

where \odot is Hadamard product. Note that the weight \mathbf{P}^c considers the semantic information and superpoint affinity simultaneously. Then, we derive the transition probability matrix $\mathbf{T}^c \in \mathbb{R}^{|V| \times |V|}$ for the c -th class, and it is defined as:

$$\mathbf{T}^c = \mathbf{D}^{-1} \mathbf{P}^c, \text{ where } D_{ii} = \sum_j P_{ij}^c \quad (4)$$

The diagonal matrix \mathbf{D} is used for the row normalization of the matrix \mathbf{P}^c . Finally, the pseudo instance label I_j of the j -th superpoint is propagated by:

$$I_j = I_k, \text{ where } k = \underset{i=1, \dots, |V|}{\operatorname{argmax}} (\hat{\mathbf{T}}_{ij}^c), \hat{\mathbf{T}}^c = (\mathbf{T}^c)^t \quad (5)$$

where $\hat{\mathbf{T}}_{ij}^c$ indicates the probability of propagating the instance label of the i -th superpoint to the j -th superpoint, and t is the iteration number. In Eq. (5), the j -th superpoint selects the instance label of the k -th superpoint, which has the highest probability of propagating to the j -th superpoint. In this way, we can propagate the instance label of each annotated superpoint to unlabeled superpoints on the superpoint graph.

3.3 Volume-Aware Instance Refinement

We propose the volume-aware instance refinement module to segment instances by using object volume information. We first introduce how to predict object volume via pseudo instances. Then, we present the volume-aware instance clustering algorithm to generate instances on the superpoint graph.

Object volume prediction via pseudo instance. In order to predict object volume information, we first train the network with the pseudo labels generated by label propagation in the first stage. As shown in Fig. 1 (e), we use the pre-trained model in the first stage to generate the pseudo instances by voting the superpoints to the closest annotated point. Specifically, by adding the predicted offset vector (refer to the first stage in Sec. 3.4) to the corresponding superpoint center, we can shift each superpoint center closer to the center of the corresponding object. To generate instances, the shifted superpoints are assigned the same instance labels as the closest annotated points with the same semantic labels. Here we regard the generated instances as the pseudo instances. According to the generated pseudo instances, we compute its volume information. We consider the number of voxels inside the instance and the instance radius to measure the volume of the instance. The instance radius is defined as the distance between the instance center and the farthest point. Thus, for each pseudo instance, we can obtain its volume information after the first stage.

Volume-aware instance clustering. After predicting the volume of the object in the first stage, we additionally use the predicted volume as the supervision to retrain the network (refer to the second stage in Sec. 3.4), which

Algorithm 1 Volume-Aware Instance Clustering Algorithm.

Input: superpoint shifted coordinate $\{\tilde{\mathbf{p}}_1, \dots, \tilde{\mathbf{p}}_{|V|}\} \in \mathbb{R}^{|V| \times 3}$; superpoint semantic label $\{s_1, \dots, s_{|V|}\} \in \mathbb{R}^{|V| \times C}$; the predicted voxel number $\{u_1, \dots, u_{|V|}\} \in \mathbb{R}^{|V|}$; the predicted radius $\{r_1, \dots, r_{|V|}\} \in \mathbb{R}^{|V|}$

Output: generated instances $\mathbf{I} \in \{I_1, \dots, I_m\}$, m is the number of instances.

```

1: Initialize an array  $f$  (visited flag) of length  $|V|$  with all zeros
2: Initialize an empty proposal set  $H$ , an empty instance set  $\mathbf{I}$ 
3: // Using the predicted radius  $r$  to filter superpoints
4: for  $v = 1$  to  $|V|$  do
5:   if  $f_v == 0$  then
6:     Initialize an empty queue  $Q$ 
7:     Initialize a set  $H$ 
8:      $f_v = 1$ ;  $Q.enqueue(v)$ ; add  $v$  to  $H$ 
9:     while  $Q$  is not empty do
10:       $j = Q.dequeue()$ 
11:      for each  $k \in \{k \mid k \in \mathcal{N}_j, s_k == s_j, \|\tilde{\mathbf{p}}_k - \tilde{\mathbf{p}}_j\|_2 < \lambda r_j\}$  do
12:        if  $f_k == 0$  then
13:           $f_k = 1$ ;  $Q.enqueue(k)$ ; add  $k$  to  $H$ 
14:      add  $H$  to  $\mathbf{H}$ 
15: // Using the predicted voxel numbers  $u$  to filter proposals
16: for each  $H \in \mathbf{H}$  do
17:   compute  $\bar{w} = \text{avg}(\{u_i \mid i \in H\})$ ,  $w$  for  $H$ 
18:   if  $w > \beta \bar{w}$  then
19:     add  $H$  to  $\mathbf{I}$ 
20: for each  $H \in \mathbf{H}$  do
21:   compute  $\bar{w} = \text{avg}(\{u_i \mid i \in H\})$ ,  $w$  for  $H$ 
22:   if  $w \leq \beta \bar{w}$  then
23:      $I_{closest} = \text{findClosestInstance}(\{I \mid I \in \mathbf{I}, s_I == s_H\})$ 
24:      $I_{closest} = I_{closest} \cup H$ 
25: return  $\mathbf{I}$ 

```

is regarded as the second stage. Thus, in the second stage, we can additionally predict the instance volume information (the number of voxels and the radius) for each superpoint. For the i -th superpoint, assume that the predicted semantic is $s_i \in \mathbb{R}^{1 \times C}$, offset vector is $\mathbf{o}_i \in \mathbb{R}^3$, the number of voxels is u_i , and the radius is r_i . Note that the predicted semantic s_i is the one hot label. We first obtain the shifted coordinate of the superpoint by $\tilde{\mathbf{p}}_i = \mathbf{p}_i + \mathbf{o}_i$, which makes the superpoint close to the corresponding instance center. Based on the shifted coordinate and the graph structure, the i -th superpoint merges its neighbors $\{j \mid j \in \mathcal{N}_i, s_i = s_j, \|\tilde{\mathbf{p}}_i - \tilde{\mathbf{p}}_j\|_2 < \lambda r_i\}$ into the same cluster, where the hyperparameter λ is set to 0.25 empirically. Note that the radius r_i is used to filter the superpoints far from the object center. Here, we use the breadth-first search on the superpoint graph to group nodes in the same cluster for generating compact

instance proposals. After that, we further count the number of voxels w in the proposal to filter the fragmented proposals. The predicted number of voxels \bar{w} of the proposal is computed by averaging the predicted voxel numbers of superpoints within the proposal. If $w > \beta\bar{w}$, the corresponding proposal can be regarded as the instance. The hyperparameter β is empirically set to 0.3. Finally, the remaining proposals are aggregated to the closest instance with the same semantic label. The volume-aware instance clustering algorithm is shown in Algorithm 1.

3.4 Network Training

Our method is a two-stage framework. As shown in Fig. 1, the first stage learns the inter-superpoint affinity to propagate labels via random walk, while the second stage leverage the object volume information to refine the instance.

First stage. As shown in Fig. 1, the first stage is supervised by the semantic loss \mathcal{L}_{sem} , offset loss $\mathcal{L}_{\text{offset}}$, and affinity loss \mathcal{L}_{aff} . The semantic loss \mathcal{L}_{sem} is defined as the cross-entropy loss:

$$\mathcal{L}_{\text{sem}} = \frac{1}{\sum_{i=1}^{|V|} \mathbb{I}(v_i)} \sum_{i=1}^{|V|} \text{CE}(s_i, s_i^*) \cdot \mathbb{I}(v_i) \quad (6)$$

where s_i is the predicted label and s_i^* is the ground truth label. Note that the original annotated labels and generated pseudo labels are all regarded as the ground truth labels. If the superpoint v_i has the label or is assigned with the pseudo label, the indicator function $\mathbb{I}(v_i)$ is equal to 1, otherwise 0. In addition, we use MLP to predict the offset vector $\mathbf{o}_i \in \mathbb{R}^3$. the offset loss $\mathcal{L}_{\text{offset}}$ is used to minimize the predicted offset of the superpoint to its instance center. $\mathcal{L}_{\text{offset}}$ is defined as:

$$\mathcal{L}_{\text{offset}} = \frac{1}{\sum_{i=1}^{|V|} \mathbb{I}(v_i)} \sum_{i=1}^{|V|} \|\mathbf{o}_i - \mathbf{o}_i^*\|_1 \cdot \mathbb{I}(v_i) \quad (7)$$

where the \mathbf{o}_i is the predicted superpoint offset and the \mathbf{o}_i^* is the ground truth offset. Note the \mathbf{o}_i^* is computed by coarse pseudo instance labels. Following [7], the affinity loss \mathcal{L}_{aff} (refer to Sec. 3.2) is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{var}} &= \frac{1}{I} \sum_{i=1}^I \frac{1}{\sum_{j=1}^{|V|} \mathbb{I}(v_j, i)} \sum_{j=1}^{|V|} \left[\|\boldsymbol{\mu}_i - \widetilde{\mathbf{X}}_j\|_2 - \delta_v \right]_+^2 \cdot \mathbb{I}(v_j, i) \\ \mathcal{L}_{\text{dist}} &= \frac{1}{I(I-1)} \sum_{i_A=1}^I \sum_{\substack{i_B=1 \\ i_A \neq i_B}}^I [2\delta_d - \|\boldsymbol{\mu}_{i_A} - \boldsymbol{\mu}_{i_B}\|_2]_+^2 \\ \mathcal{L}_{\text{aff}} &= \mathcal{L}_{\text{var}} + \mathcal{L}_{\text{dist}} + \alpha \cdot \mathcal{L}_{\text{reg}}, \text{ where } \mathcal{L}_{\text{reg}} = \frac{1}{I} \sum_{i=1}^I \|\boldsymbol{\mu}_i\|_2 \end{aligned} \quad (8)$$

where I is the number of instances (equal to the number of the annotated points, *i.e.*, one point per instance). $\boldsymbol{\mu}_i$ is the mean embedding of the i -th instance and $\widetilde{\mathbf{X}}_j$ is the embedding of the j -th superpoint in Eq. (2). According to [7], the margins δ_v and δ_d are set to 0.1 and 1.5, respectively. The parameter α is set to 0.001, and $[x]_+ = \max(0, x)$ denotes the hinge. $\mathbb{I}(v_j, i)$ is the indicator function,

and $\mathbb{I}(v_j, i)$ equals to 1 if superpoint v_j is labeled as the i -th instance. Note that we only perform \mathcal{L}_{aff} on the superpoints with annotated labels or pseudo labels. The final loss function in the first stage is defined as:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{offset}} + \mathcal{L}_{\text{aff}} \quad (9)$$

Second stage. As shown in Fig. 1, the second stage is supervised by the semantic loss \mathcal{L}_{sem} , offset loss $\mathcal{L}_{\text{offset}}$, and volume loss $\mathcal{L}_{\text{volume}}$. As the affinity loss is used for label propagation, we remove it in the second stage. For the volume loss $\mathcal{L}_{\text{volume}}$, it uses the predicted object volume information as the ground truth to train the network (refer to Sec. 3.3). The $\mathcal{L}_{\text{volume}}$ is formulated as:

$$\mathcal{L}_{\text{volume}} = \frac{1}{K} \sum_{i=1}^K \sum_{j=1}^I (\|u_i - \hat{u}_j\|_1 + \|r_i - \hat{r}_j\|_1) \cdot \mathbb{I}(i, j) \quad (10)$$

where K is the number of labeled superpoints, including the original annotated labels and the generated pseudo labels. If the i -th superpoint belongs to the j -th instance, the indicator function $\mathbb{I}(i, j)$ is equal to 1, otherwise 0. \hat{u}_j and \hat{r}_j indicate the ground truth voxel numbers and radius counted from the pseudo instances, respectively. The generation of the pseudo instances refers to Sec. 3.3. The final loss function in the second stage is defined as:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{offset}} + \mathcal{L}_{\text{volume}} \quad (11)$$

4 Experiments

4.1 Experimental Settings

Datasets. ScanNet-v2 [6] and S3DIS [1] are used in our experiments to conduct 3D instance segmentation. ScanNet-v2 contains 1,613 indoor RGB-D scans with dense semantic and instance annotations. The dataset is split into 1,201 training scenes, 312 validation scenes, and 100 hidden test scenes. The instance segmentation is evaluated on 18 object categories. S3DIS contains 6 large-scale indoor areas, which has 272 rooms and 13 categories. For the ScanNet-v2 dataset, we report both validation and online test results. For the S3DIS dataset, we report both Area 5 and the 6-fold cross validation results.

Evaluation metrics. For the ScanNet-v2 dataset, the mean average precision at the overlap 0.25 (AP_{25}), 0.5 (AP_{50}) and overlaps from 0.5 to 0.95 (AP) are reported. For the S3DIS dataset, we additionally use mean coverage (mCov), mean weighted coverage (mWCov), mean precision (mPrec), and mean recall (mRec) with the IoU threshold of 0.5 as evaluation metrics.

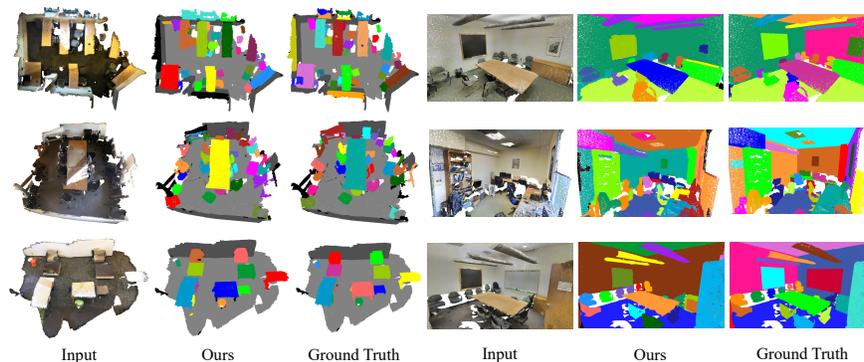
Annotation of weak labels. To generate weak labels of point clouds, we randomly click one point of each instance as the ground truth label. Note that our annotation strategy is the same as SegGroup [24]. Unlike our method and SegGroup, SPIB [18] adopts 3D box-level annotation, which annotates each instance with bounding box. Compared with time-consuming box-level annotation, clicking one point per instance is faster and more convenient.

4.2 Results

ScanNet-v2. Tab. 1 reports the quantitative results on the ScanNet-v2 validation set and hidden testing set. Compared with the existing semi-/weakly

Table 1. 3D instance segmentation results on the ScanNet-v2 validation set and online test set. “Baseline” means the model trained with the initial annotated labels only.

Method		Annotation	AP	AP ₅₀	AP ₂₅	AP	AP ₅₀	AP ₂₅
			Validation Set			Online Test Set		
Fully Sup.	SGPN [28]	100%	-	11.3	22.2	4.9	14.3	39.0
	3D-SIS [9]	100%	-	18.7	35.7	16.1	38.2	55.8
	MTML [13]	100%	20.3	40.2	55.4	28.2	54.9	73.1
	PointGroup [12]	100%	34.8	56.9	71.3	40.7	63.6	77.8
	H AIS [2]	100%	43.5	64.1	75.6	45.7	69.9	80.3
	SSTNet [17]	100%	49.4	64.3	74.0	50.6	69.8	78.9
	SoftGroup [27]	100%	46.0	67.6	78.9	50.4	76.1	86.5
Weakly Sup.	SPIB [18]	0.16%	-	38.6	61.4	-	-	63.4
	SegGroup [24]	0.02%	23.4	43.4	62.9	24.6	44.5	63.7
	Baseline	0.02%	21.2	39.0	61.3	-	-	-
	3D-WSIS (ours)	0.02%	28.1	47.2	67.5	25.1	47.0	67.8

**Fig. 3.** Visualization of the 3D instance segmentation results on the validation of ScanNet-v2 (left) and S3DIS (right). We randomly select colors for different instances.

supervised point cloud instance segmentation methods, our approach achieves state-of-the-art performance and improves the AP₂₅ from 62.9% to 67.5% with a gain of about 5% on the ScanNet-v2 validation set. Note that SPIB [18] uses the bounding box of each instance as weak labels, which is different from ours. Although we have the same number of annotated instances, clicking one point per instance provides less information than eight corners of the box. Nonetheless, our method can still achieve higher performance than SPIB.

S3DIS. Tab. 2 reports the results on Area 5 and 6-fold cross validation of the S3DIS dataset. Compared with the fully supervised 3D instance segmentation methods, our model can still achieve good results, even outperforming the fully supervised methods such as SGPN [28]. The quantitative results demonstrate the effectiveness of our method on weakly supervised 3D instance segmentation.

Table 2. 3D instance segmentation results on the S3DIS dataset. “Baseline” means the model trained with the initial annotated labels only.

Method		Annotation	AP	AP ₅₀	AP ₂₅	mCov	mWCov	mPrec	mRec
6-fold Cross Validation									
Fully Sup.	SGPN [28]	100%	-	-	-	37.9	40.8	38.2	31.2
	ASIS [29]	100%	-	-	-	51.2	55.1	63.6	47.5
	PointGroup [12]	100%	-	64.0	-	-	-	69.6	69.2
	HAIS [2]	100%	-	-	-	67.0	70.4	73.2	69.4
	SSTNet [17]	100%	54.1	67.8	-	-	-	73.5	73.4
	SoftGroup [27]	100%	54.4	68.9	-	69.3	71.7	75.3	69.8
Weakly Sup.	Baseline	0.02%	19.5	30.5	42.0	41.1	42.3	13.3	37.1
	SegGroup	0.02%	23.1	37.6	48.5	45.5	47.6	56.7	43.3
	3D-WSIS (ours)	0.02%	26.7	40.4	52.6	48.0	50.5	59.3	46.7
Area 5									
Fully Sup.	SGPN [28]	100%	-	-	-	32.7	35.5	36.0	28.7
	ASIS [29]	100%	-	-	-	44.6	47.8	55.3	42.4
	PointGroup [12]	100%	-	57.8	-	-	-	61.9	62.1
	HAIS [2]	100%	-	-	-	64.3	66.0	71.1	65.0
	SSTNet [17]	100%	42.7	59.3	-	-	-	65.5	64.2
	SoftGroup [27]	100%	51.6	66.1	-	66.1	68.0	73.6	66.6
Weakly Sup.	Baseline	0.02%	18.9	26.8	37.7	36.3	37.1	11.1	28.5
	SegGroup	0.02%	21.0	29.8	41.9	39.1	40.8	47.2	34.9
	3D-WSIS (ours)	0.02%	23.3	33.0	48.3	42.2	44.2	50.8	38.9

Visualization results. Fig. 3 shows the visualization results of instance segmentation on the ScanNet-v2 validation set and S3DIS. It can be found that although some objects of the same class are close to each other, like chairs, the instances are still segmented properly. Since we additionally predict the size of the objects in the network, our method can effectively use the size information of the objects to guide the clustering, thereby segmenting different instances that are close to each other.

4.3 Ablation Study

Effect of pseudo labels. We conduct experiments on the ScanNet-v2 validation set to verify the effectiveness of our label propagation. The quantitative results are reported in Tab. 3. “Baseline” indicates that our method trained with initial annotated labels, without pseudo labels. In the first stage, we report the mean average precision at different iterations, respectively. It can be observed that the performance gradually increases as the number of iterations increases. Furthermore, based on the stage one (dubbed “Stage 1”), it can be found that the performance is greatly improved after training in the second stage (dubbed “Stage 2”). Since the number of generated pseudo labels in the first stage is still less than that of fully annotated labels, it is difficult to effectively segment instances. In the second stage, we use the model trained in the first stage to cluster pseudo instances, so we can regard the obtained object volume as the additional supervision to train the network. The quantitative results in the second stage

Table 3. The ablation study of different components on the ScanNet-v2 validation set and S3DIS Area 5. “Baseline” means the model trained the initial annotated labels only. Note that “Stage 2” is performed based on “Stage 1”.

		ScanNet-v2 Val			S3DIS Area 5						
Settings		AP	AP ₅₀	AP ₂₅	AP	AP ₅₀	AP ₂₅	mCov	mWCov	mPrec	mRec
Baseline		21.2	39.0	61.3	18.9	26.8	37.7	36.3	37.1	11.1	28.5
Stage 1	Iter. 1	23.4	42.2	62.8	19.9	27.7	40.2	38.5	39.4	21.7	33.1
	Iter. 2	24.5	43.6	64.4	20.1	28.0	40.7	39.7	40.4	22.2	33.8
	Iter. 3	25.4	45.3	65.8	20.9	28.3	41.9	40.0	40.8	23.1	34.1
Stage 2		28.1	47.2	67.5	23.3	33.0	48.3	42.2	44.2	50.8	38.9

Table 4. The ablation study results (proportion/accuracy) of pseudo labels at different iterations on the ScanNet-v2 training set. The proportion and accuracy of pseudo labels are computed at the point level.

Stage 1	Only Random	Random+Affinity	Random+Affinity+Semantic
Iter. 1	33.7 / 39.9	29.1 / 52.7	18.2 / 81.9
Iter. 2	47.7 / 38.4	35.1 / 71.6	30.9 / 82.1
Iter. 3	48.2 / 38.3	35.2 / 73.1	31.4 / 82.5

further demonstrate that using the predicted object volume can indeed improve the performance of weakly supervised 3D instance segmentation.

Quality of pseudo labels. The pseudo labels are generated on the training set to increase supervision during training, so their quality affects network training. To further study the quality of the generated pseudo labels, we count the proportion and accuracy of pseudo labels on the ScanNet-v2 training set during training. The results are listed in Tab. 4. When only using random walk (dubbed “Only Random”), the proportion of pseudo labels is high, but the accuracy is low (39.9% at “Iter. 1”). The low-accuracy pseudo labels will affect the training of the network. If we add the extra affinity constraint (dubbed “Random+Affinity”), we can observe that the proportion of pseudo labels is lower, but the accuracy is greatly improved (52.7% at “Iter. 1”). Due to the affinity constraint, the proportion of wrong label propagation is reduced. Therefore, the quality of pseudo labels is improved and high-quality supervision is provided for network training. Furthermore, when we add the semantic constraints (dubbed “Random+Affinity+Semantic”), the accuracy of pseudo labels improves from 52.7% (“Random+Affinity”) to 81.9% (“Random+Affinity+Semantic”), which shows that the semantic constraint is useful for the weakly supervised 3D instance segmentation task. As constraints are added gradually, the proportion of the generated pseudo labels decreases, while the accuracy increases.

Label propagation times. Different label propagation times influence the quality of pseudo labels. As shown in Tab. 4, the proportion and accuracy of pseudo labels at three iterations (dubbed “Iter. 3”) is comparable to two iterations

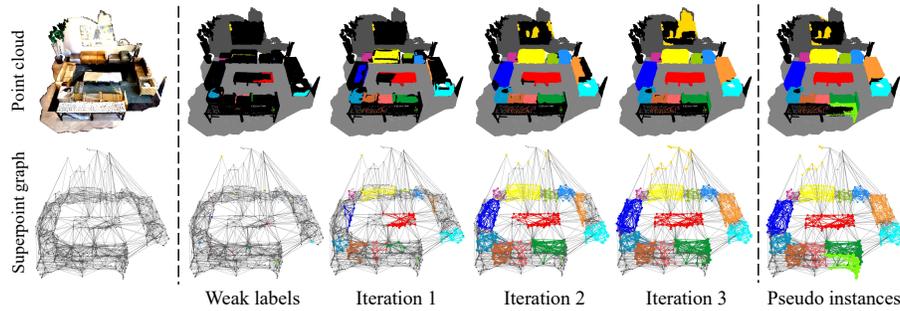


Fig. 4. Visualization results of pseudo label generation. Note that we remove the superpoints on the walls and floor for a better view.

(dubbed “Iter. 2”), and performing more iterations consumes more resources, thus we choose three iterations for label propagation.

Visualization of pseudo labels. Fig. 4 shows the pseudo labels at different iterations in the first stage on the superpoint graph. We can observe that the initial annotated labels are extremely sparse, and only a few superpoints are annotated in the superpoint graph. In the first stage, as the iteration number of label propagation increases, the labels spread to the surrounding superpoints in the graph gradually. With the constraints of the predicted affinity and semantic, the propagation of labels is restricted to the same object. In the second stage, we use the model trained in the first stage to predict pseudo instances by performing clustering on the superpoint graph. The last column shows the predicted pseudo instances. It can be observed that different instances can be effectively separated.

5 Conclusion

In this paper, we proposed a simple yet effective method for weakly supervised 3D instance segmentation with extremely few labels. To exploit few point-level annotations, we used an unsupervised point cloud oversegmentation method on the point cloud to generate superpoints and construct the superpoint graph. Based on the constructed superpoint graph, we developed an inter-superpoint affinity mining module to adaptively learn inter-superpoint affinity for label propagation via random walk. We further developed a volume-aware instance refinement module to guide the superpoint clustering on the superpoint graph by learning the object volume information. Experiments on the ScanNet-v2 and S3DIS datasets demonstrate that our method achieves state-of-the-art performance on weakly supervised 3D instance segmentation.

Acknowledgments The authors would like to thank reviewers for their detailed comments and instructive suggestions. This work was supported by the National Science Fund of China (Grant No. 61876084).

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: CVPR (2016)
2. Chen, S., Fang, J., Zhang, Q., Liu, W., Wang, X.: Hierarchical aggregation for 3D instance segmentation. In: ICCV (2021)
3. Cheng, M., Hui, L., Xie, J., Yang, J.: SSPC-Net: Semi-supervised semantic 3D point cloud segmentation network. In: AAAI (2021)
4. Cheng, M., Hui, L., Xie, J., Yang, J., Kong, H.: Cascaded non-local neural network for point cloud semantic segmentation. In: IROS (2020)
5. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR (2019)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017)
7. De Brabandere, B., Neven, D., Van Gool, L.: Semantic instance segmentation with a discriminative loss function. arXiv (2017)
8. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018)
9. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: CVPR (2019)
10. Hui, L., Yuan, J., Cheng, M., Xie, J., Zhang, X., Yang, J.: Superpoint network for point cloud oversegmentation. In: ICCV (2021)
11. Jaritz, M., Gu, J., Su, H.: Multi-view pointnet for 3D scene understanding. In: ICCVW (2019)
12. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: PointGroup: Dual-set point grouping for 3D instance segmentation. In: CVPR (2020)
13. Lahoud, J., Ghanem, B., Pollefeys, M., Oswald, M.R.: 3D instance segmentation via multi-task metric learning. In: ICCV (2019)
14. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: CVPR (2018)
15. Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M.: Deep projective 3D semantic segmentation. In: CAIP. Springer (2017)
16. Li, M., Xie, Y., Shen, Y., Ke, B., Qiao, R., Ren, B., Lin, S., Ma, L.: HybridCR: Weakly-supervised 3D point cloud semantic segmentation via hybrid contrastive regularization. In: CVPR (2022)
17. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance segmentation in 3D scenes using semantic superpoint tree networks. In: ICCV (2021)
18. Liao, Y., Zhu, H., Zhang, Y., Ye, C., Chen, T., Fan, J.: Point cloud instance segmentation with semi-supervised bounding-box mining. TPAMI (2021)
19. Liu, Z., Qi, X., Fu, C.W.: One thing one click: A self-training approach for weakly supervised 3D semantic segmentation. In: CVPR (2021)
20. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: CVPR (2017)
21. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space (2017)
22. Shi, H., Wei, J., Li, R., Liu, F., Lin, G.: Weakly supervised segmentation on outdoor 4D point clouds with temporal matching and spatial graph propagation. In: ICCV (2022)
23. Simonovsky, M., Komodakis, N.: Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: CVPR (2017)

24. Tao, A., Duan, Y., Wei, Y., Lu, J., Zhou, J.: SegGroup: Seg-level supervision for 3D instance and semantic segmentation. *TIP* (2022)
25. Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.Y.: Tangent convolutions for dense prediction in 3D. In: *CVPR* (2018)
26. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: KPConv: Flexible and deformable convolution for point clouds. In: *ICCV* (2019)
27. Vu, T., Kim, K., Luu, T.M., Nguyen, X.T., Yoo, C.D.: SoftGroup for 3D instance segmentation on 3D point clouds. In: *CVPR* (2022)
28. Wang, W., Yu, R., Huang, Q., Neumann, U.: SGPN: Similarity group proposal network for 3D point cloud instance segmentation. In: *CVPR* (2018)
29. Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively segmenting instances and semantics in point clouds. In: *CVPR* (2019)
30. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *TOG* (2019)
31. 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: *CVPR* (2020)
32. Wu, W., Qi, Z., Fuxin, L.: PointConv: Deep convolutional networks on 3D point clouds. In: *CVPR* (2019)
33. Xu, X., Lee, G.H.: Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In: *CVPR* (2020)
34. Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., Trigoni, N.: Learning object bounding boxes for 3D instance segmentation on point clouds. In: *NeurIPS* (2019)
35. Yi, L., Zhao, W., Wang, H., Sung, M., Guibas, L.J.: GSPN: Generative shape proposal network for 3D instance segmentation in point cloud. In: *CVPR* (2019)
36. Zhang, Y., Qu, Y., Xie, Y., Li, Z., Zheng, S., Li, C.: Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In: *ICCV* (2021)