SSPC-Net: Semi-supervised Semantic 3D Point Cloud Segmentation Network

Mingmei Cheng, Le Hui, Jin Xie*, Jian Yang

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education
Jiangsu Key Lab of Image and Video Understanding for Social Security
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{chengmm, le.hui, csjxie, csjyang}@njust.edu.cn

Abstract

Point cloud semantic segmentation is a crucial task in 3D scene understanding. Existing methods mainly focus on employing a large number of annotated labels for supervised semantic segmentation. Nonetheless, manually labeling such large point clouds for the supervised segmentation task is time-consuming. In order to reduce the number of annotated labels, we propose a semi-supervised semantic point cloud segmentation network, named SSPC-Net, where we train the semantic segmentation network by inferring the labels of unlabeled points from the few annotated 3D points. In our method, we first partition the whole point cloud into superpoints and build superpoint graphs to mine the long-range dependencies in point clouds. Based on the constructed superpoint graph, we then develop a dynamic label propagation method to generate the pseudo labels for the unsupervised superpoints. Particularly, we adopt a superpoint dropout strategy to dynamically select the generated pseudo labels. In order to fully exploit the generated pseudo labels of the unsupervised superpoints, we furthermore propose a coupled attention mechanism for superpoint feature embedding. Finally, we employ the cross-entropy loss to train the semantic segmentation network with the labels of the supervised superpoints and the pseudo labels of the unsupervised superpoints. Experiments on various datasets demonstrate that our semi-supervised segmentation method can achieve better performance than the current semi-supervised segmentation method with fewer annotated 3D points.

Introduction

Due to the increasing growth of 3D point cloud data, point cloud semantic segmentation has been receiving more and more attention in the 3D computer vision community. Most of these segmentation methods focus on fully supervised segmentation with manually annotated points (Hu et al. 2020; Thomas et al. 2019; Lei, Akhtar, and Mian 2020; Zhao et al. 2019; Wang et al. 2019a). However, annotating large-scale 3D point clouds is a cumbersome process, which is costly in labor and time. Particularly, the number of point clouds in some real scenes such as the indoor scene can often reach the order of magnitude to millions. Therefore, it is difficult to obtain the accurate labels of these million points for full-supervised segmentation.

Different from full-supervised point cloud segmentation, semi-supervised segmentation aims to learn a good label prediction for point clouds with partially annotated points. Recent works have been dedicated to the semi-supervised point cloud segmentation task. Guinard *et al.* (Guinard and Landrieu 2017) propose a weakly supervised conditional random field classifier for 3D LiDAR point cloud segmentation. However, it converts the segmentation task into an optimization problem, and the contextual information in point clouds is ignored. Mei *et al.* propose a semi-supervised 3D LiDAR point cloud segmentation method (Mei et al. 2019), where the 3D data is projected to range images for feature embedding, and the inter-frame constraints are combined with some labeled samples to encourage feature consistency. Nonetheless, the constraints along the LiDAR sequential frames are not available in general 3D segmentation datasets. Lately, (Xu and Lee 2020) proposes a semi-supervised point cloud segmentation method, which employs three constraints to enhance the feature learning of unlabeled points, including block-level label penalization, data augmentation with rotation and flipping for prediction consistency, and a spatial and color smoothness constraint in local regions. Although it can obtain effective segmentation results, the long-range relations are ignored in this method.

Although some efforts have been made on semi-supervised point cloud segmentation, how to accurately predict the labels of unannotated points for segmentation is still a challenging problem. Particularly, since point clouds are irregular, it is difficult to exploit the geometry structures of point clouds to accurately infer pseudo labels of unannotated points for label propagation. In addition, the uncertainty of inferred pseudo labels of unannotated points hinders the network from learning discriminative features of point clouds, leading to inaccurate label prediction.

Aiming at the aforementioned two problems, in this paper, we propose a novel semi-supervised semantic point cloud segmentation network, named SSPC-Net. We first divide the point clouds into superpoints and build the superpoint graph, where the superpoint is a set of points with isotropically geometric features. Thus, we can convert the point-level label prediction problem in the point cloud segmentation task into the superpoint-level label prediction problem. Following the

*Corresponding author.

method in (Landrieu and Simonovsky 2018), we employ the gated graph neural network (GNN) (Li et al. 2015) for superpoint feature embedding. In order to fully exploit the local geometry structure of the constructed superpoint graph, we then develop a dynamic label propagation method to accurately infer pseudo labels for unsupervised superpoints. Specifically, the labels of supervised superpoints are gradually extended to the adjacent superpoints with high semantic similarity along the edges of the superpoint graph. We also adopt a superpoint dropout strategy to obtain the high-quality pseudo labels during the label propagation process, where the extended superpoints with low confidences are dynamically pruned. Furthermore, we propose a coupled attention mechanism to learn the discriminative context features of superpoints. We alternatively perform attention on the supervised and extended superpoints so that the discrimination of the features of the supervised and extended superpoints can be boosted each other, alleviating the uncertainty of the inferred pseudo labels of the unsupervised superpoints. Finally, we employ a combined cross-entropy loss to train the segmentation network. Extensive results on various indoor and outdoor datasets demonstrate that our method can yield good performance with only few point-level annotations.

The main contributions of this paper are summarized as: **(1)** We develop a dynamic superpoint label propagation method to accurately infer the pseudo labels of unsupervised superpoints. We also present a superpoint dropout strategy to select the high-quality pseudo labels. **(2)** We propose a coupled attention mechanism on the supervised and extended superpoints to learn the discriminative features of the superpoints. **(3)** Our proposed method can yield better performance than the current semi-supervised point cloud semantic segmentation method with fewer labels.

Related Work

Deep learning on 3D point clouds. Recently, many deep learning methods are proposed to tackle point cloud classification and segmentation. Some methods (Wu et al. 2015; Maturana and Scherer 2015; Sedaghat et al. 2016; Qi et al. 2016) voxelize point clouds and employ 3D CNNs for feature embedding. However, the voxel-based methods suffer from the large memory cost due to the high-resolution voxels. By projecting point clouds into 2D views, (Su et al. 2015; Boulch, Le Saux, and Audebert 2017; Tatarchenko et al. 2018) use classic CNNs to extract features from point clouds. However, the view-based methods are sensitive to the density of 3D data. To reduce memory cost and additional preprocessing, Qi *et al.* propose PointNet, which directly processes the unordered point clouds and uses multi-layer perceptrons (MLPs) and the maxpooling function for feature embedding. Following PointNet, many efforts (Qi et al. 2017b; Klokov and Lempitsky 2017; Wang et al. 2019a; Hua, Tran, and Yeung 2018; Li et al. 2018; Zhao et al. 2019; Wang et al. 2019b; Thomas et al. 2019; Wu et al. 2019; Liu et al. 2019; Han et al. 2020; Zhao and Tao 2020; Feng et al. 2018; Ma et al. 2018) are proposed for point cloud processing. Although these methods have achieved decent performance, their models depend on fully annotated 3D point clouds for training. However, in this paper, we focus on the semi-supervised point cloud semantic segmentation.

Semi-/Weakly supervised deep learning on 3D point clouds. Many efforts (Mei et al. 2019; Wei et al. 2020; Xu and Lee 2020) have been proposed to tackle semi-/weakly supervised point cloud semantic segmentation. In (Mei et al. 2019), Mei *et al.* introduce a semi-supervised 3D LiDAR data segmentation method. It first converts the 3D data to depth maps and then applies CNNs for feature embedding. In addition to a small part of supervised data, it also leverages the temporal constraints along the LiDAR scans sequence to boost feature consistency. Therefore, it is not practicable for general point cloud segmentation cases. Inspired by CAM (Zhou et al. 2016), Wei *et al.* propose MPRM (Wei et al. 2020) with scene-level and subcloud-level labels for weakly supervised segmentation. Specifically, it leverages a point class activation map (PCAM) to obtain the localization of each class and then generates point-wise pseudo labels with a multi-path region mining module. In this way, the segmentation network can be trained in a fully supervised manner. However, in practice, generating the subcloud-level annotation is still time-consuming. Lately, in (Xu and Lee 2020), Xu *et al.* propose a semi-supervised algorithm, which uses three constraints on the unlabeled points, $i.e.$, the block level labels for penalizing the negative categories in point clouds, data augmentation with random in-plane rotation and flipping for feature consistency and a spatial and color smoothness constraint in point clouds.

Our Method

In this section, we present our semi-supervised point cloud segmentation network and the outline of our framework is shown in Fig. 1. We first introduce the superpoint graph embedding module. Then we propose a dynamic label propagation approach combined with a superpoint dropout strategy. Next, we propose a coupled attention mechanism to learn discriminative contextual features of superpoints. Finally, we depict the framework of our method.

Superpoint Graph Embedding

To obtain the superpoints and learn the superpoint features, following (Landrieu and Simonovsky 2018), we perform an unsupervised superpoints partition approach to generate superpoints and then build superpoint graphs combined with graph neural network (GNN) for superpoints feature embedding. Denote $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as the superpoint graph built upon superpoints, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. Edge $(i, j) \in \mathcal{E}$ links node $i \in \mathcal{V}$ with $j \in \mathcal{V}$. We first perform a lightweight PointNet-like structure on the superpoints to obtain superpoints features. After that, we learn the superpoint embedding with the gated GNN used in (Li et al. 2015). Given the superpoint embeddings and the semi-supervision, we can penalize the model with incomplete supervision. For a point cloud consists of N superpoints, we define $a_i \in \{0, 1\}^N$ to indicate whether the i-th superpoint has supervision. Then the segmentation loss $\mathcal{L}_s$ on the superpoint graph embedding module can be formulated as:

$$\mathcal{L}_s = \frac{1}{A} \sum\nolimits_{i=1}^{N} a_i \cdot \mathcal{F}_{loss} \left(z_i, \boldsymbol{y}_i \right) \tag{1}$$

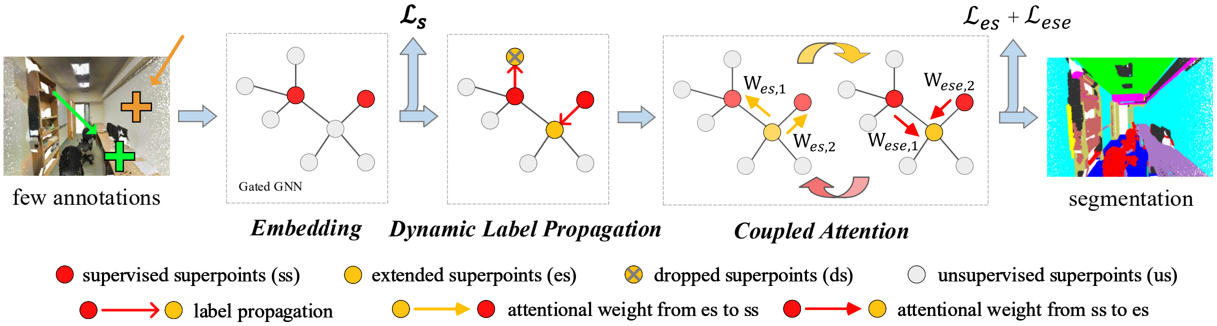
$\mathcal{L}_s$
$\mathcal{L}_{es} + \mathcal{L}_{ese}$
few annotations
Gated GNN
$W_{es,1}$
$W_{es,2}$
$W_{ese,2}$
$W_{ese,1}$
segmentation
Embedding
Dynamic Label Propagation
Coupled Attention
supervised superpoints (ss)
extended superpoints (es)
dropped superpoints (ds)
unsupervised superpoints (us)
label propagation
attentional weight from es to ss
attentional weight from ss to es

Figure 1: Overview of the proposed semi-supervised semantic point cloud segmentation network (SSPC-Net). We first leverage the gated GNN to extract superpoints features. Then based on the superpoint graph, we conduct the dynamic label propagation strategy to generate pseudo labels. Next, based on the supervised superpoints and the extended superpoints, we perform a coupled attention mechanism to further boost the extraction of discriminative contextual features in the point cloud.

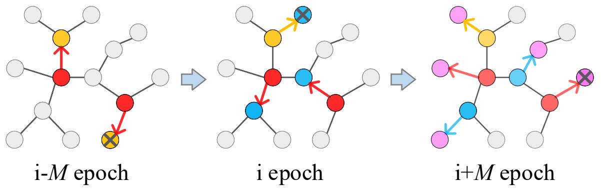

Figure 2: The procedure of our dynamic label propagation. We progressively propagate the superpoint-level label and discard the extended superpoint with low confidence.

where $\mathcal{F}_{loss}$ is the loss function and we choose the cross-entropy loss in experiments, $A = \sum_{i=1}^{N} a_i$ is adopted for normalization, z_i represents the superpoint-level label of i-superpoint and $\boldsymbol{y}_i$ is the prediction logit.

The reason why we choose the superpoint graph as the representation of point cloud is at two points. On the one hand, the superpoint is geometrically isotropic and therefore we can directly extend the point-level label to the superpoint-level label, which alleviates the lack of supervision. On the other hand, since the superpoint graph is rooted in the geometric structure of the point cloud, where the linking edges between the superpoints greatly facilitate the feature propagation. Thus we can obtain more discriminative contextual features of superpoints.

Dynamic Label Propagation

To propagate superpoint labels, we propose a dynamic label propagation strategy to generate pseudo labels. Suppose we have constructed three sets: the supervised superpoints set S, unsupervised superpoints set U, and extended superpoints set E. Note that at the beginning we set $E = \varnothing$. Besides, elements in each set indicate the index of superpoints.

For $\forall\ i \in T, T = S \cup E$, we use the adjacent superpoints to construct candidate set $\mathcal{N}_i$, where we consider propagating labels in it. Suppose z_i is the label of i-th superpoint. Note that $\forall\ j \in \mathcal{N}_i$ must satisfy two constraints: $j \in U$ and the predicted category of the j-th superpoint should be the same as that of the i-th superpoint, that is, z_i. Compared with other unsupervised superpoints, elements in $\mathcal{N}_i$ are with higher possibilities to be assigned with pseudo labels, due to the close geometric relations and the close distances to the i-th superpoint. To generate high-quality pseudo labels, we assess the confidence scores of the superpoints in $\mathcal{N}_i$ and denote the scores as $\boldsymbol{m}_i \in \mathbb{R}^{|\mathcal{N}_i|}$. Then, we enumerate all the superpoints in $\mathcal{N}_i$ and select the superpoint with the highest confidence score. The operation can be formulated as:

$$j^* = \underset{j=1,2,\ldots,|\mathcal{N}_i|}{\arg\max}\ (m_{i,j}) \qquad (2)$$

where j^* represents the index of the superpoint with the highest confidence score in $\mathcal{N}_i$. To further ensure the high quality of pseudo labels, we set the threshold τ to filter the selected superpoints with dissatisfactory confidence values. When the confidence score $m_{i,j^*} \geqslant \tau$, the j^*-th superpoint is selected and assigned with pseudo label z_i. Then j^* will be removed from the unsupervised superpoints set U and added to the extended superpoints set E. On the contrary, if there is no superpoint satisfying the constraint, no superpoint will be extended from $\mathcal{N}_i$. In the experiments, τ is empirically set to 0.9. Note that for each extension procedure, we merge the supervised superpoints set S and the extended superpoints set E to the new set $T = S \cup E$ for further extension. Because the extended superpoints with pseudo labels can also be treated as the superpoints with supervision for further label propagation. In this way, we can progressively propagate the labels of the supervised superpoints and generate more high-quality pseudo labels for unsupervised superpoints in U. What's more, Algorithm 1 shows the details of the graph-based supervision extension procedure.

Since our extension strategy is performed progressively, we consider removing the low-confidence superpoints in the extended superpoints set E. Hence, we propose a superpoint dropout strategy assessing the reliability of the extended superpoints in the embedding space. In the superpoints set $T = S \cup E$, we cluster the superpoints into c classes according to the superpoints labels or pseudo labels, where c is the number of categories. Suppose $\mathcal{C}_i$ is the i-th cluster set that contains the index of the superpoints belonging to the

Algorithm 1: Graph-based supervision extension

```
Input: Supervised superpoints set $S$, unsupervised
       superpoints set $U$, extended superpoints set
       $E$, threshold $\tau$
Output: Updated sets $U$ and $E$
$T = S \cup E$
for $i \in T$ do
    Denote $z_i$ as the label of $i$-th superpoint
    Construct the candidate supeproints set $\mathcal{N}_i$
    if $\mathcal{N}_i \neq \varnothing$ then
        Generate the confidence scores $\boldsymbol{m}_i$
        $j^* = \underset{j=1,2,\dots,|\mathcal{N}_i|}{\arg\max}\ (m_{i,j})$
        if $m_i^{j^*} \geqslant \tau$ then
            Assign pseudo label $z_i$ to the $j^*$-th
              superpoint
            $U := U \setminus \{j^*\} \quad E := E \cup \{j^*\}$
```

Algorithm 2: Superpoint dropout strategy

```
Input: Number of classes $c$, supervised superpoints
       set $S$, unsupervised superpoints set $U$,
       extended superpoints set $E$
Output: Updated sets $U$ and $E$
$T = S \cup E$
Cluster on $T$ and obtain $c$ cluster sets: $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_c$
for $i = 1 : c$ do
    Compute the feature $\boldsymbol{v}_i$ of the cluster center of $\mathcal{C}_i$
    for each $j \in E \cap \mathcal{C}_i$ do
        Generate the feature $\boldsymbol{f}_j$ of the $j$-th superpoint
        Compute the distance $d_i^j = \|\boldsymbol{f}_j - \boldsymbol{v}_i\|_2$
    Find the farthest 5% superpoints (set as $\mathcal{C}_{drop}$) in
      $E \cap \mathcal{C}_i$ from the cluster center according to the
      distance $\boldsymbol{d}_i$
    $E := E \setminus \mathcal{C}_{drop} \quad U := U \cup \mathcal{C}_{drop}$
```

i-th category. In addition, we denote $\boldsymbol{v}_i$ as the feature of the cluster center of $\mathcal{C}_i$, which is computed by averaging the features of all the superpoints in $\mathcal{C}_i$. We assess the confidence of the extended superpoints by considering its distance to the corresponding cluster center in the feature space. For $\forall\ j \in E \cap \mathcal{C}_i$, its Euclidean distance to the cluster center in the feature space is formulated as:

$$d_i^j = \|\boldsymbol{f}_j - \boldsymbol{v}_i\|_2 \tag{3}$$

where $\boldsymbol{f}_j \in \mathbb{R}^D$ is the feature of j-th superpoint, and $\boldsymbol{v}_i \in \mathbb{R}^D$ is the feature of cluster center. Smaller distance indicates the higher reliability of extended superpoints, whereas the larger distance means the higher uncertainty. Therefore, in each cluster, we discard k extended superpoints that are furthest from the cluster center, where k is set to $0.05 * |E \cap \mathcal{C}_i|$. In other words, we retain the most reliable 95% superpoints and drop the 5% unreliable superpoints in the set $E \cap \mathcal{C}_i$. Our superpoint dropout strategy is explained in Algorithm 2.

Concretely, as shown in Fig. 2, we perform our graph-based dynamic label propagation strategy every M epochs, therefore, the extended superpoints are gradually "growing" on the graph from the supervised superpoints. The reason why we conduct the extension operation in a multi-stage manner instead of every epoch is that our extension strategy is a cumulative one, which means that too much extension operations will cause redundant extended superpoints and aggravate the memory cost. Meanwhile, the model is not stable at the beginning, which is not conducive to generating extended superpoints.

Coupled Attention for Feature Enhancement

Aiming to learn more discriminative contextual features in point clouds, we propose a coupled attention mechanism. For $\forall\ i \in S$, we denote the corresponding embedding as $\boldsymbol{h}_i \in \mathbb{R}^D$. Similarly, for $\forall\ j \in E$, we denote the corresponding embedding as $\boldsymbol{h}_j$. By weighing all the extended superpoints, we extract the novel contextual feature of i-th superpoint with attention mechanism:

$$\boldsymbol{x}_i = \sum\nolimits_{j \in E} g\left(\phi(\boldsymbol{h}_i, \boldsymbol{h}_j)\right) \odot \alpha(\boldsymbol{h}_j) \tag{4}$$

where $\phi(\boldsymbol{h}_i, \boldsymbol{h}_j) = MLP(\boldsymbol{h}_i - \boldsymbol{h}_j)$ embeds the channel-wise relations between superpoints, $\alpha(\boldsymbol{h}_j) = MLP(\boldsymbol{h}_j)$ is a unary function for individual superpoint embedding, $\phi(\cdot,\cdot) : \mathbb{R}^D \to \mathbb{R}^D$ and $\alpha : \mathbb{R}^D \to \mathbb{R}^D$, $\odot$ is the Hadamard product. g is a normalization function and is defined as:

$$g\left(\phi_l(\boldsymbol{h}_i, \boldsymbol{h}_j)\right) = \frac{\exp(\phi_l(\boldsymbol{h}_i, \boldsymbol{h}_j))}{\sum_{r \in E} \exp(\phi_l(\boldsymbol{h}_i, \boldsymbol{h}_r))} \tag{5}$$

where $l = 1, 2, \dots, D$, represents l-th element of embedding $\phi_l(\cdot,\cdot)$. Consequently, the matrix representation of the attention operation on the supervised superpoints in S can be formulated as:

$$\boldsymbol{X}_s = \sum\nolimits_{j \in E} \boldsymbol{W}_{es,j} \odot \boldsymbol{H}_{e,j} \tag{6}$$

where $\boldsymbol{X}_s \in \mathbb{R}^{|S| \times D}$, $\boldsymbol{W}_{es,j} \in \mathbb{R}^{|S| \times D}$, $\boldsymbol{H}_{e,j} \in \mathbb{R}^{|S| \times D}$ and j enumerates the extended superpoints in E. Note that $\boldsymbol{W}_{es} \in \mathbb{R}^{|S| \times |E| \times D}$ represents the channel-wise weights from the extended superpoints to the supervised superpoints.

Once we obtain the attention embedding $\boldsymbol{X}_s \in \mathbb{R}^{|S| \times D}$, we can derive new segmentation logits of supervised superpoints and formulate the loss as:

$$\mathcal{L}_{es} = \frac{1}{|S|} \sum\nolimits_{i \in S} \mathcal{F}_{loss}\left(z_i, FC\left(\boldsymbol{X}_{s,i}\right)\right) \tag{7}$$

where z_i is the superpoint-level label, $\boldsymbol{X}_{s,i}$ is the attention feature of the corresponding supervised superpoint, and $\mathcal{F}_{loss}$ is the cross-entropy loss adopted in experiments. Note that FC is the fully connected layer, which maps $\boldsymbol{X}_{s,i} \in \mathbb{R}^{|S| \times D}$ from the D-dim to the dimension of the categories.

Similarly, to promote the feature characterization of the extended superpoints, we then perform attention on the extended superpoints in reverse. By weighting the new features enhanced by the attention operation of the supervised superpoints, we boost the context feature propagation and thus

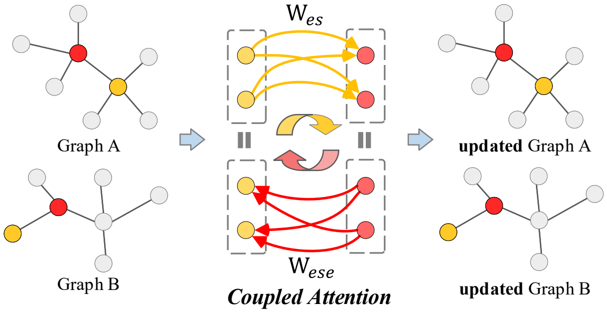

Figure 3: The coupled attention for feature enhancement.

enhance the robustness of the features of the extended superpoints. Thus, for $\forall j \in E$, the new embedding of the corresponding superpoint can be calculated as:

$$\boldsymbol{y}_j = \sum\nolimits_{i \in S} g\left(\psi(\boldsymbol{h}_j, \boldsymbol{x}_i)\right) \odot \beta(\boldsymbol{x}_i) \qquad (8)$$

where $\psi(\boldsymbol{h}_j, \boldsymbol{x}_i) = MLP(\boldsymbol{h}_j - \boldsymbol{x}_i)$ characterizes the dependencies of the extended superpoints on the attention embeddings of the supervised superpoints. $\beta(\boldsymbol{x}_i) = MLP(\boldsymbol{x}_i)$ is a unary function similar to α, $\psi(\cdot,\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^D$ and $\beta : \mathbb{R}^D \rightarrow \mathbb{R}^D$. g is a normalization function defined as:

$$g\left(\psi_l(\boldsymbol{h}_j, \boldsymbol{x}_i)\right) = \frac{\exp(\psi_l(\boldsymbol{h}_j, \boldsymbol{x}_i))}{\sum_{r \in S} \exp(\psi_l(\boldsymbol{h}_j, \boldsymbol{x}_r))} \qquad (9)$$

where $l = 1, 2, \ldots, D$, denotes l-th element of embedding $\psi(\cdot,\cdot)$. Then the matrix representation of the attention operation on the extended superpoints in E can be defined as:

$$\boldsymbol{Y}_e = \sum\nolimits_{i \in S} \boldsymbol{W}_{ese,i} \odot \boldsymbol{\mathcal{X}}_{s,i} \qquad (10)$$

where $\boldsymbol{Y}_e \in \mathbb{R}^{|E| \times D}$, $\boldsymbol{W}_{ese,i} \in \mathbb{R}^{|E| \times D}$, $\boldsymbol{\mathcal{X}}_{s,i} \in \mathbb{R}^{|E| \times D}$ and i enumerates superpoints in S. Note that $\boldsymbol{\mathcal{X}}_s$ is the feature after employing function $\beta(\cdot)$ on attention feature $\boldsymbol{X}_s$. In this way, we develop the coupled attention, *i.e.*, $\boldsymbol{W}_{ese} \in \mathbb{R}^{|S| \times |E| \times D}$ denotes the channel-wise weights from the attentional supervised superpoints to extended superpoints.

Then the loss $\mathcal{L}_{ese}$ on the extended superpoints with enhanced attention features can be formulated as:

$$\mathcal{L}_{ese} = \frac{1}{|E|} \sum\nolimits_{j \in E} \mathcal{F}_{loss}\left(z_j^p, FC\left(\boldsymbol{Y}_{e,j}\right)\right) \qquad (11)$$

where z_j^p is the pseudo label and $\mathcal{F}_{loss}$ is the cross-entropy loss as well. FC maps the feature to the category space.

Specifically, as shown in Fig. 3, our coupled attention considers the intra- and inter-relations concurrently. To encourage the feature consistency in different point clouds, we integrate the supervised superpoints and extended superpoints in various point clouds into sets S and E, respectively. The connections between S and E are constructed within and across various point cloud samples, and superpoints with the same labels are encouraged to have more similar semantic embeddings compared to those with diverse classes. As a result, by alternatively performing attention on the supervised and extended superpoints, more long-range dependencies between superpoints are built. Hence, the model learns more discriminative and robust contextual features of the supervised and unsupervised superpoints.

Framework

The framework of our model is illustrated in Fig. 1. In our framework, the superpoint graph embedding module is the basis of our point cloud feature embedding. Based on this module, the dynamic label propagation method assesses the semantic similarity between the superpoints and propagates the superpoint-level supervision along the edges of the superpoint graph. Then, with the extended superpoints searched by the dynamic label propagation module, we propose a coupled attention mechanism to boost the contextual feature learning of the point cloud.

The final objective function is a combination of the three objectives $\mathcal{L}_{final} = \mathcal{L}_s + \lambda_1 \cdot \mathcal{L}_{es} + \lambda_2 \cdot \mathcal{L}_{ese}$ and we empirically set λ_1, λ_2 to 1. As shown in Fig. 1, the dynamic label propagation module and coupled attention module are only conducted in the training stage. For testing, we obtain the inferred prediction directly from the superpoint graph embedding module.

Experiments

Implementation Details

To train our model, we adopt Adam optimizer with a base learning rate of 0.01. For the S3DIS (Armeni et al. 2016), ScanNet (Dai et al. 2017) and vKITTI (Gaidon et al. 2016) dataset, we employ the mini-batch size of 4, 8, 8, respectively. We empirically implement the dynamic label propagation module every $M = 40$ epochs.

Semi-supervision generation. To produce the semi-supervision of point clouds, we randomly select a part of the points with annotations in each class. For example, given a point cloud containing n points with c classes, suppose the supervision rate be r, then we evenly distribute the supervision budget $r \cdot n$ and randomly sample $(r \cdot n)/c$ points in each category as the supervised points. The label of superpoint is the category with the most annotated points. If there is no supervised point contained, then the superpoint will be unsupervised. Note that compared with the sampling strategy of random sampling annotated points directly in point clouds, our labeling mechanism is more in coincident with the human annotation behavior, since the random sampling strategy will result in that most of the supervised points will be occupied by the areas with simple geometric structure but more points, *e.g.*, walls, roads, etc. For evaluation, all the quantitative results are computed at the point level.

Semi-supervised Semantic Segmentation

S3DIS. S3DIS (Armeni et al. 2016) dataset is an indoor 3D dataset including 6 areas and 13 categories. Three metrics are adopted for quantitative evaluation: mean IoU (mIoU), mean class accuracy (mAcc), and overall accuracy (OA).

The quantitative and visual results are shown in Tab. 1 and Fig. 4, respectively. For a fair comparison, we test our framework with the "1pt" labeling strategy adopted in (Xu and Lee 2020) (dubbed "Semi-Seg" in Tab. 1) as well, which samples one point in each category of each block as the supervised point. It can be seen that our SSPC-Net achieves a significant gain of 9.3% in terms of mIoU with the "1pt" labeling strategy. In (Xu and Lee 2020), Xu *et al.* split the point cloud

	Method	Rate	mIoU	mAcc	OA
		6-fold cross validation			
Full	PointNet	100%	47.6	66.2	78.5
	SPGraph	100%	62.1	73.0	85.5
	PointCNN	100%	**65.3**	**75.6**	**88.1**
	RSNet	100%	56.4	66.4	-
	G+RCU2	100%	49.7	66.4	81.1
	3P-RNN	100%	56.3	73.6	86.9
Semi-	Baseline	0.002%	45.1	63.7	73.9
	SSPC-Net	0.002%	48.5	68.3	79.1
	SSPC-Net	0.01%	**54.5**	**70.8**	**80.4**
		Fold 5			
Full	PointNet	100%	41.1	49.0	-
	PointNet++	100%	47.8	-	-
	SPGraph	100%	**58.0**	**66.5**	**86.3**
	SegCloud	100%	48.9	57.3	-
	PointCNN	100%	57.2	63.8	85.9
Semi-	Semi-Seg	1pt	44.5	-	-
	Semi-Seg	10%	48.0	-	-
	Baseline	0.002%	39.6	52.1	72.4
	SSPC-Net	0.002%	43.0	56.4	76.2
	SSPC-Net	0.01%	51.5	63.8	82.0
	SSPC-Net	1pt	**53.8**	**63.9**	**83.8**

Table 1: Evaluation on the S3DIS dataset.

	Method	Rate	ScanNet		vKITTI		
			mIoU	OA	mIoU	mAcc	OA
Full	PointNet	100%	-	73.9	34.4	47.0	79.7
	PointNet++	100%	-	**84.5**	-	-	-
	SSP + SPG	100%	-	-	**52.0**	**67.3**	84.3
	G+RCU	100%	-	-	35.6	57.6	79.7
	RSNet	100%	**39.3**	79.2	-	-	-
	3P-RNN	100%	-	-	41.6	54.1	**87.8**
	3DCNN	100%	-	73.0	-	-	-
Semi-	Baseline	0.01%	24.1	38.2	35.7	53.4	79.2
	SSPC-Net	0.01%	27.1	66.6	41.0	55.7	81.2
	SSPC-Net	0.05%	**39.3**	**77.1**	**50.6**	**64.8**	**85.4**

Table 2: Evaluation on the ScanNet and vKITTI datasets.

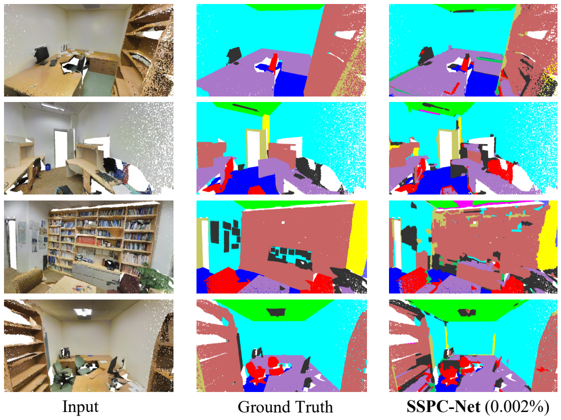

Figure 4: The visual results on the S3DIS dataset with supervision rate of 0.002%.

into blocks and then train and test their model on each block separately. Nonetheless, our model learns the embeddings of superpoints in the whole point cloud, therefore we can obtain more discriminative contextual features and yield better performance. Note that in Tab. 1, "Baseline" represents our method without the label propagation strategy and coupled attention mechanism. One can see that our SSPC-Net improves the performance from 39.6% to 43.0% in terms of mIoU with the supervision rate of 0.002% on Area 5 of the S3DIS dataset, benefiting from the pseudo labels generated from the label propagation and the discriminative contextual features extracted by the coupled attention mechanism.

ScanNet. ScanNet (Dai et al. 2017) is an indoor scene dataset containing 1513 point clouds with 20 categories. We split the dataset into a training set with 1201 scenes and a testing set with 312 scenes following (Qi et al. 2017b). We adopt overall semantic voxel labeling accuracy (OA) and mean IoU (mIoU) for evaluation.

We list the quantitative results on the testing set in Tab. 2. Similar to S3DIS, ScanNet is also an indoor dataset, but the point cloud of ScanNet is much sparser than that of S3DIS. This brings greater challenges to the propagation of supervised labels. However, the proposed model can still achieve good segmentation results and even outperform some fully supervised methods like PointNet (Qi et al. 2017a) with semi-supervision. Furthermore, the performance of the proposed model is much better than the baseline method, which further validates the effectiveness of our method.

vKITTI. vKITTI (Gaidon et al. 2016) dataset mimics the real-world KITTI dataset and contains the synthetic outdoor scenes with 13 classes (including road, tree, terrain, car, etc.). For evaluation, we split the dataset into 6 non-overlapping sub-sequences and employ 6-fold cross validation following (Ye et al. 2018). Mean IoU (mIoU), mean class accuracy (mAcc) and overall accuracy (OA) are employed for evaluation.

The quantitative results are presented in Tab. 2. With the 0.01% point-level annotations, compared with the baseline method, our model achieves better segmentation results due to the dynamic label propagation strategy and the discriminative contextual features generated from the coupled attention module. In addition, our model can achieve better or comparable performance than some fully supervised methods with only 0.01% and 0.05% of the supervised points.

Ablation Study

Contribution of individual components. In this section, we investigate the contribution of the proposed components to model performance. The evaluation results on Area5 of the S3DIS dataset of different components with the supervision ratio of 0.002% and 0.01% are shown in Tab. 3, where the components are the graph embedding (Graph Emb.), dynamic label propagation (Label Prop.), coupled attention for feature enhancement (Coup. Attn.). It can be observed that there is an obvious promotion on the performance with the addition of dynamic label propagation and coupled atten-

Components			Rate=0.002%			Rate=0.01%		
Graph Emb.	Label Prop.	Coup. Attn.	mIoU	mAcc	OA	mIoU	mAcc	OA
✓			39.6	52.1	72.4	48.5	61.2	80.3
✓	✓		40.9	55.8	73.6	50.0	60.6	80.8
✓	✓	✓	**43.0**	**56.4**	**76.2**	**51.5**	**63.8**	**82.0**

Table 3: The contribution of different components on Area5 of the S3DIS dataset with different annotation rates.

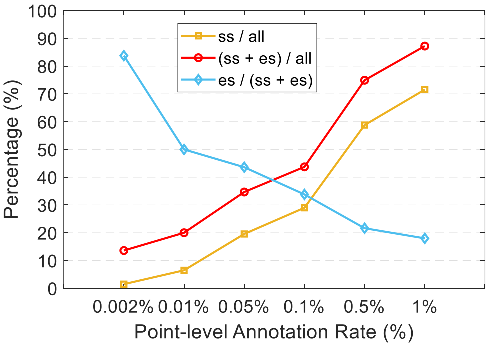

Figure 5: The percentage of supervised superpoints (ss) and extended superpoints (es) during training. Note that "all" means the overall superpoints.

tion module, which further demonstrates the effectiveness of these strategies for the semi-supervision.

Supervision rate. The number of supervised points plays an important role in the segmentation performance. The more labeled points, the smaller gap of data distribution between the semi-supervision and full supervision. To discuss the effect of various labeling rates on model performance, we test our method on Area5 of the S3DIS dataset. The results are shown in Tab. 4. Combined with Tab. 1, it can be observed that with only few labeled points, our model has already achieved effective segmentation results. With the growth of supervision, the performance of our model further increases. It is worth noting that we pay more attention to the cases of extremely few supervision signals, which is more challenging for the point cloud segmentation task.

Number of the extended superpoints. The dynamic label propagation strategy plays an important role in our model. As shown in Fig. 5, we show the proportion of the supervised superpoints and extended superpoints in the training set when testing on Area 5 of the S3DIS dataset. With the increase of the annotated points, the proportion of the supervised superpoints increases rapidly. Because the probability of a superpoint containing a supervised point is getting higher as well. However, when there are fewer supervised points, the percentage of extended superpoints is obviously larger. This demonstrates the importance of pseudo labels facing extremely few point annotations.

Quality of the extended superpoints. To analyze the quality of the extended superpoints, we evaluate the overall accuracy of the extended superpoints (OA of es) in Tab. 4. Noted that, similar to the aforementioned metrics, the quantitative results of the extended superpoints are conducted at the point level as well. From Tab. 4, one can see that the overall accuracy of extended superpoints is around 90%, which demonstrates the high quality of extended superpoints. This further proves the effectiveness of our label propagation strategy which generates high-quality pseudo labels. In addition, the high quality of pseudo labels of the extended superpoints further reveals the reason for the improved performance based on the label propagation module.

Rate	mIoU	mAcc	OA	OA of es
0.002%	43.0	56.4	76.2	87.3
0.01%	51.5	63.8	82.0	90.9
0.1%	56.2	66.1	84.6	91.0
1.0%	58.3	66.5	85.7	90.1

Table 4: Comparison of various supervision rates on Area5 of the S3DIS dataset, where "es" represents the extended superpoints.

Interval M	mIoU	mAcc	OA
20	50.2	61.1	81.2
30	50.8	63.3	81.5
40	**51.5**	**63.8**	**82.0**
50	49.6	61.5	80.7
60	49.9	62.2	81.0

Table 5: Comparison of segmentation results with various interval M of the dynamic label propagation method in the case of the supervision rate of 0.01%.

Epoch interval in dynamic label propagation. During the training, we perform the dynamic label propagation method every M epochs. For comparison, we train our model with various interval M while keeping other parameters unchanged with the supervision rate of 0.01%. The evaluation results on Area 5 of the S3DIS dataset are shown in Tab. 5. It can be observed that when $M = 40$, our model achieves the best performance.

Conclusion

In this paper, we proposed a semi-supervised point cloud segmentation network. We first partitioned the point cloud into superpoints and built superpoint graphs to explore the long-range relations in the point cloud. Then based on superpoint graphs, we proposed a dynamic label propagation method combined with a superpoint dropout strategy to generate high-quality pseudo labels for the unsupervised superpoints. Next, we proposed a coupled attention module to learn discriminative contextual features of superpoints and fully exploit the generated pseudo labels. Our method can achieve better performance than the current semi-supervised point cloud segmentation methods with fewer labels.

Acknowledgments

This work was supported by the National Science Fund of China (Grant Nos. U1713208, 61876084), Program for Changjiang Scholars.

References

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 *CVPR*.

Boulch, A.; Le Saux, B.; and Audebert, N. 2017. Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks. *3DOR* .

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*.

Feng, Y.; Zhang, Z.; Zhao, X.; Ji, R.; and Gao, Y. 2018. Gvcnn: Group-view convolutional neural networks for 3D shape recognition. In *CVPR*.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*.

Guinard, S.; and Landrieu, L. 2017. Weakly supervised segmentation-aided classification of urban scenes from 3D LiDAR point clouds. In *ISPRS Workshop*.

Han, W.; Wen, C.; Wang, C.; Li, X.; and Li, Q. 2020. Point2Node: Correlation Learning of Dynamic-Node for Point Cloud Feature Modeling. In *AAAI*.

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 *CVPR*.

Hua, B.-S.; Tran, M.-K.; and Yeung, S.-K. 2018. Pointwise convolutional neural networks. In *CVPR*.

Klokov, R.; and Lempitsky, V. 2017. Escape from cells: Deep kd-networks for the recognition of 3D point cloud models. In *ICCV*.

Landrieu, L.; and Simonovsky, M. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *CVPR*.

Lei, H.; Akhtar, N.; and Mian, A. 2020. SegGCN: Efficient 3D Point Cloud Segmentation With Fuzzy Spherical Kernel. In *CVPR*.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. Pointcnn: Convolution on x-transformed points. In *NeurIPS*.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* .

Liu, X.; Han, Z.; Liu, Y.-S.; and Zwicker, M. 2019. Point2sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network. In *AAAI*.

Ma, C.; Guo, Y.; Yang, J.; and An, W. 2018. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. *IEEE Transactions on Multimedia* 21(5): 1169–1182.

Maturana, D.; and Scherer, S. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *IROS*.

Mei, J.; Gao, B.; Xu, D.; Yao, W.; Zhao, X.; and Zhao, H. 2019. Semantic segmentation of 3D lidar data in dynamic scene using semi-supervised learning. *IEEE Transactions on Intelligent Transportation Systems* 21(6): 2496–2509.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*.

Qi, C. R.; Su, H.; Niebner, M.; Dai, A.; Yan, M.; and Guibas, L. J. 2016. Volumetric and Multi-view CNNs for Object Classification on 3D Data. In *CVPR*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*.

Sedaghat, N.; Zolfaghari, M. R.; Amiri, E.; and Brox, T. 2016. Orientation-boosted Voxel Nets for 3D Object Recognition. In *CVPR*.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. G. 2015. Multi-view convolutional neural networks for 3D shape recognition. In *ICCV*.

Tatarchenko, M.; Park, J.; Koltun, V.; and Zhou, Q.-Y. 2018. Tangent convolutions for dense prediction in 3D. In *CVPR*.

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 *ICCV*.

Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; and Shan, J. 2019a. Graph attention convolution for point cloud semantic segmentation. In *CVPR*.

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. In *CVPR*.

Wu, P.; Chen, C.; Yi, J.; and Metaxas, D. 2019. Point cloud processing via recurrent set encoding. In *AAAI*.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D shapenets: A deep representation for volumetric shapes. In *CVPR*.

Xu, X.; and Lee, G. H. 2020. Weakly Supervised Semantic Point Cloud Segmentation: Towards 10x Fewer Labels. In *CVPR*.

Ye, X.; Li, J.; Huang, H.; Du, L.; and Zhang, X. 2018. 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In *ECCV*.

Zhao, H.; Jiang, L.; Fu, C.-W.; and Jia, J. 2019. PointWeb: Enhancing local neighborhood features for point cloud processing. In *CVPR*.

Zhao, L.; and Tao, W. 2020. JSNet: Joint Instance and Semantic Segmentation of 3D Point Clouds. In *AAAI*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.