Abstract: Learning semantics from unstructured 3D point clouds with fewer labels, although a realistic problem, has been under-explored. While existing weakly supervised methods can effectively learn semantics with only a small fraction of point-level annotations, we find that the vanilla bounding box-level annotation is also informative for precise semantic segmentation of large-scale 3D point clouds. In this paper, we introduce a neural architecture, called Box2Seg, to infer the semantics of 3D point clouds with bounding box-level supervision. The key to our approach is to generate accurate pseudo labels of the whole point clouds by exploring the geometric structure inside and outside each bounding box. In particular, two groups of pseudo labels are first estimated through the 3DGrabCut and Point Class Activation Mapping (PCAM) techniques, the network is further trained and refined with the combined pseudo labels. Experiments on three large-scale benchmarks including S3DIS, ScanNet, and SemanticKITTI demonstrate the competitive performance of the proposed method. In particular, the proposed network can be trained with cheap, or even off-the-shelf bounding box-level annotations and scene-level tags.

Keywords: Semantic Segmentation, 3D Point Clouds, Box-Level Supervision

1 Introduction

Semantic understanding of the 3D environment is a key enabler for robots to perceive and interact with the physical world. A number of real applications such as robotic grasping [1], autonomous navigation [2], and human-machine interaction [3] require the machine to precisely recognize (i.e., point-wise dense segmentation) its 3D surroundings. However, this remains challenging due to the complex geometrical structures of environments and limited capacity of existing 3D understanding models.

As 3D data acquisition and annotation become increasingly affordable, remarkable progress has been achieved in the task of 3D point cloud analysis in recent years [4]. This is not only reflected in the emergence of several milestone network architectures [5, 6, 7, 8, 9, 10, 11, 12, 13], but also a number of high-quality densely annotated public datasets [14, 15, 16, 17, 18, 19, 20, 21, 22]. However, as suggested in [16], it is still an open question that if existing 3D segmentation pipelines can be generalized/scaled to extremely large-scale real 3D environments with complex geometrical structures. One main issue is that most existing pipelines still heavily rely on the strong supervision signals (i.e., point-wise dense semantic labels), which are not always available and easy to collect in practice. Furthermore, the requirement of strong supervision also prevents existing approaches from taking full advantage of different forms of rich annotations in existing datasets. For example, there are more than 12 million bounding boxes in the Waymo dataset [20], while not fully exploited for segmentation tasks.

A handful of recent works have started to explore the weakly supervised semantic segmentation of 3D point clouds [23, 24, 25, 26, 27]. That is, learning semantics from a small fraction of labeled points or other indirect formats of supervision. MPRM [28] and SegGroup [26] are proposed to generate pseudo point-level labels from subcloud-level labels or seg-level labels. SQN [23], [24] and [25] are introduced to learn semantic segmentation from a small fraction of randomly annotated or actively

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.
labeled points. This is achieved by leveraging the semantic similarity within a neighborhood or gradient approximation, and contrastive pretraining followed with finetuning. Despite the promising results, it is still non-trivial to deploy these weak supervision schemes in practice as a variety of customized tools are unavailable yet (e.g., It remains challenging for existing tools to annotate sparsely distributed points or sub-cloud labels on-the-fly). Different from the aforementioned scheme, we instead turn our eyes to the commonly-used bounding boxes in 3D object detection. Is it possible to learn point-level semantics from the bounding box annotations?

In this paper, we propose a weakly supervised semantic segmentation framework called Box2Seg by leveraging the commonly used, low-cost bounding box annotations and scene-level tags as the supervision signal. Overall, the key idea is to fully exploit the geometric priors and structural cues from the limited supervision signals, further estimate the pseudo labels of the foreground points and background points separately in a divide-and-conquer way. In particular, we extend the idea of GraphCut [29, 30] from 2D images to unstructured 3D point clouds, so as to predict the point-wise semantic mask for the foreground points. Moreover, we also train a dedicated point classifier to mine the localization cues by investigating the generated point class activation maps, hence further estimate the pseudo labels of background points. Finally, the generated point-wise pseudo labels allow us to train a segmentation network in a fully supervised way, although the supervision signal is not perfectly correct. Experiments on several large-scale open benchmarks demonstrate the feasibility of our solution. The proposed Box2Seg framework can even achieve comparable performance with several early supervised baseline networks on existing datasets, by learning from limited supervision tags.

Overall, the contributions of our paper can be summarized as follows:

- To the best of our knowledge, the proposed Box2Seg framework is the first attempt to achieve semantic segmentation of large-scale 3D point clouds, with the usage of only bounding box-level annotations and scene-level tags.
- An effective approach is proposed to generate pseudo labels for both foreground and background points with 3D Grabcut and Point Class Activation Maps (PCAMs) to achieve weakly supervised learning.
- Extensive experiments conducted on three large-scale point cloud datasets and various backbones demonstrate the effectiveness and versatility of Box2Seg. In particular, the proposed framework allows leveraging the ready-to-use bounding box labels in existing large-scale 3D object detection datasets.

2 Related Works

2.1 Semantic Segmentation of 3D Point Clouds with Full Supervision

Recently, pioneering works such as PointNet [5] and SparseConvNet [11] have greatly facilitated the development of deep learning in semantic segmentation of 3D point clouds. A number of sophisticated approaches [6, 8, 7, 9, 31, 12, 13, 32] have been further proposed to substantially improve the runtime
performance from various aspects. Following [4], existing approaches can be roughly divided into four categories according to the representation of 3D point clouds used in their frameworks: 1) Point-based methods [5, 6, 7, 33, 8, 31, 9] 2) Voxel-based methods [11, 34, 35, 12, 36, 32], 3) Projection-based methods [10, 37, 38, 39], and 4) Hybrid approaches [13, 40, 41].

Although remarkable progress has been achieved, these methods usually require point-level semantic labels to provide dense supervision signal for network training, which is usually extremely expensive and not always available in practice. This issue is particularly serious for customized practical applications since not everyone has a dedicated annotation team. Motivated by this, we propose Box2Seg, a weakly-supervised framework to learn from the ready-to-use bounding box annotations and cheap scene-level tags.

2.2 Semantic Segmentation of 3D Point Clouds with Weak Supervision

To reduce the annotation cost, a handful of recent methods [24, 23, 25, 26, 28, 27] have been proposed to investigate weakly supervised semantic segmentation of 3D point clouds. These existing approaches can be divided into two categories according to their weak supervision settings: methods with limited point supervision, methods with inexact supervision.

Methods with Limited Point Supervision. These methods aim at learning the semantics of 3D point clouds using a small fraction of labeled points (e.g., 10%, 1‰) as their supervision. In general, the labeled points can be determined randomly [23, 24] or through active learning techniques [25]. Xu et al. [24] achieve weakly supervised point cloud segmentation with 10× fewer labels through gradient approximation and spatial color consistency. Hu et al. [23] empirically identify the redundancy in dense 3D annotations, and propose a network (namely, SQN) to implicitly augment the supervision signal by leveraging the semantic similarity within neighboring points. Inspired by the unsupervised PointContrast [42] framework, Hou et al. [25] further investigate label-efficient learning on 3D point clouds and active labeling techniques.

Methods with Inexact Supervision. Instead of using point-level supervisions, these methods attempt to learn from various inexact supervisions, including sub-cloud level labels [28], scene-level tags [27], seg-level labels [26], and 2D pixel-wise labels [43]. Wei et al. [44] first train a classifier using sub-cloud level labels, and then introduce a multi-path region mining module to extract the object localization region cues. They finally generate pseudo labels for the training of their segmentation network. Ren et al. [27] introduce WyPR to achieve semantic segmentation, 3D proposal generation, and 3D object detection jointly, with only scene-level tags being provided. Tao et al. [26] propose SegGroup to achieve 3D semantic segmentation and instance segmentation, by taking oversegmentation as the pre-processor and seg-level labels as the supervision signals. Wang et al. [43] propose a Graph-based Pyramid Feature Network (GPFN) to learn semantics from 3D point clouds with solely 2D supervision.

Analogous to these methods with inexact supervision, we introduce Box2Seg in this paper to achieve semantic segmentation of 3D point clouds, by leveraging annotated instance bounding boxes and scene level tags as its supervision. This is motivated by the large quantities of ready-to-use bounding box annotations (e.g., 12 million bounding boxes in the Waymo dataset [20]) in existing 3D object detection datasets [21, 20]. We aim at building a weakly supervised framework to learn semantics from cheap bounding box labels and scene level tags. Consequently, the gap between the tasks of 3D object detection and dense point cloud segmentation can be bridged.

3 The Proposed Method

3.1 Overview

Different from existing fully-supervised segmentation pipelines [31, 9, 5, 6, 7, 8], where their point-wise semantic labels are always available by default, we consider learning semantics from low-cost bounding box annotations and scene-level tags in this work. Note that, it is usually much easier to obtain bounding box annotations and scene-level tags than dense point-wise annotations. For example, the average number of instances in a scene in the ScanNet dataset [17] is around 20, and the annotation cost is around 2 minutes1 using off-the-shelf labeling tools [45, 21]. In contrast, it takes

---

1 It took around 7s to annotate an instance with bounding boxes [21].
Figure 2: Overall framework of the proposed Box2Seg. Unlike conventional fully-supervised point-cloud semantic segmentation methods that use point-level annotation for model training, our Box2Seg takes the bounding-box annotations and scene-level tags as main supervision signals. The whole pipeline consists of three steps, including foreground, background pseudo label generation, and the re-training of segmentation models. 3D Grabcut and point class activation map techniques are adopted.

22.3 minutes to conduct point-wise annotation [17]. Moreover, bounding box annotations are even ready-to-use in several 3D object detection datasets [20, 21].

As shown in Fig. 2, the key to our approach is to generate point-wise pseudo labels inside and outside each bounding box in a divide-and-conquer manner, by fully exploring the local geometric priors and structures in point clouds. In particular, we first approximate the foreground pseudo labels by generalizing the GrabCut [30] algorithm to unstructured 3D point clouds. Then, we leveraging the class activation maps to further estimate the background pseudo labels. Finally, these per-point pseudo labels are further combined to train an existing segmentation network.

3.2 Learning Foreground Pseudo Labels from Box-Level Annotations

Given a 3D point cloud \( P \) and its associated bounding box annotations (i.e., 3D coordinates of the bounding box and corresponding semantics), the first objective is to generate a foreground semantic mask for each individual instance. Inspired by the classical GraphCut methods [29, 30] in computer vision, we develop our 3D GrabCut in this work to separate foreground and background points within each bounding box.

The pipeline of our 3D GrabCut is shown in Fig. 3. Considering the unstructured and orderless nature of 3D point clouds, the raw point cloud within each bounding box is converted into a regular volumetric representation before feeding into the network. Then, the voxels are first oversegmented into geometrically homogeneous partitions. Next, foreground voxels are segmented through iterative energy minimization [30], by taking the bounding box priors and the generated soft label matrix (i.e., partitions) as the input. These voxels identified as foreground are assigned with the semantic label of the bounding box, while background voxels still remain unlabeled. Finally, semantic labels of these voxels are back-projected to the raw point clouds. By iteratively traversing all bounding boxes, we can obtain the point-level semantic mask for foreground points in each instance.

Note that, each voxel is only assigned to a single bounding box to avoid the semantic ambiguity of foreground points. If a voxel belongs to the foreground of two overlapping bounding boxes, we simply assign the voxel to the smaller bounding box. With the proposed 3D GrabCut method, the pseudo labels (i.e., semantic mask) of foreground points can be produced. Then, the geometrical information in background points is further explored.

3.3 Learning Background Pseudo Labels from Scene-Level Tags

Apart from these foreground points, the raw point clouds also contain a large number of background points (e.g., road, floor, ceiling, etc.). In particular, these points usually have relatively homogeneous
geometric patterns. Motivated by this, we propose to learn from additional scene-level class tags $y_s \in \{0, 1\}^C$, which indicates whether a specific class appears in the scene or not. Inspired by the success of Class Activation Mapping (CAM) techniques [46] in weakly supervised semantic segmentation of 2D images, we first train a point classifier under the scene-level supervision, and then further bring out the localization cues of specific classes from the weighted combination of intermediate latent feature maps.

As shown in Fig. 4, we first hierarchically encode the geometrical patterns of point cloud $\mathcal{P}$ as compact latent vectors. Considering the efficiency and the capacity of processing large-scale point clouds, we adopt RandLA-Net [31] as the backbone of our framework. Then, instead of the symmetry decoder module used in the U-Net architecture [47], we append several $1 \times 1$ convolution layers to further predict the classification probabilities. Note that, Global Average Pooling (GAP) is used at the final stage, and the sigmoid cross-entropy loss is used to train the classifier.

Once the training of the point classifier is completed, the class activation map of point $p$ with respect to semantic class $c$ is given as:

$$M_c(p) = w_{c}^T \cdot f_{cam}(p) \cdot y_c,$$  \hspace{1cm} (1)

where $M_c(p)$ is the class activation map for class $c$ at point $p$, $f_{cam}(p)$ is the last latent feature fed into the global average pooling layer. $y_c$ is an one-hot vector, which indicates the presence or absence of semantic category $c$. Therefore, we can further determine the pseudo label of each point by assigning the class with the largest activation value:

$$\tilde{y}_p = \arg\max_c (M(p)).$$  \hspace{1cm} (2)

Note that, the pseudo labels generated at this stage are only used to determine the semantics of background points (i.e., points remaining unlabeled), while the semantics of foreground points remain unchanged.

### 3.4 Learning Semantic Segmentation from Pseudo Labels

Once the point-wise pseudo segmentation of the raw point clouds are obtained, we can retrain the network with the generated pseudo labels in a fully supervised manner. Here, we take all points for network training and calculate the vanilla cross-entropy loss between the predicted logits and the “pseudo ground truth labels”:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=0}^{C} p_{hc} \log \hat{p}_{hc},$$  \hspace{1cm} (3)

where $N$ is the number of the training points, $C$ is the number of the semantic categories, $\hat{p}_{hc}$ and $p_{hc}$ represent the predicted and pseudo labels for the $i$-th point, respectively. Finally, the trained model is used to infer the semantic labels of the raw point clouds. Note that, we are aware that the generated pseudo labels are not completely correct, but similar to [28], we found that the final segmentation results are totally acceptable, even training with these flawed labels.
4 Experiments

To demonstrate the effectiveness of the proposed Box2Seg framework, we conduct extensive experiments on three representative public benchmarks in this section, including indoor S3DIS [18], ScanNet [17], and outdoor SemanticKITTI [15].

4.1 Implementation Details

The proposed Box2Seg takes RandLA-Net [31], an efficient framework for large-scale point clouds, as the backbone to learn point-wise semantic labels. However, our Box2Seg framework is quite flexible and allows to use any point-based method as its backbone. Note that, we append one 1×1 convolution layer after the encoder of RandLA-Net to build the classifier when generating pseudo labels for background points. Once we obtain all the pseudo labels, we train our Box2Seg with up to 100 epochs, and Adam is used as the optimizer. The learning rate is set to 0.01 and decreases by 5% after each epoch. Mean Intersection over Union (mIoU) and Overall Accuracy (OA) are used as the main evaluation metrics. All experiments are conducted on a PC with NVIDIA GTX 1080Ti GPU with 11G memory for a fair comparison. In all experiments, we only use spatial information for the generation of pseudo labels and the training of our segmentation model.

4.2 Evaluation on the S3DIS Dataset

The S3DIS [18] dataset is composed of 272 indoor scenarios collected from 6 areas, with 273 million points being divided into 13 semantic categories, including ceiling, floor, tables, chairs, etc. Following [9, 48], Area-5 is selected as the validation set to evaluate the final segmentation performance of our Box2Seg. Note that, the box-level labels in this dataset are not provided. Instead, we generate box-level labels based on the instance labels provided in the dataset. Three categories, including ceiling, floor, and wall, are considered as the background categories, while the rest are considered as foreground classes.

The quantitative comparison between our Box2Seg and other existing methods are shown in Table 1. It can be seen that the proposed Box2Seg can achieve better performance than existing weakly supervised approaches [24], especially in light of the bounding box annotations and scene-level tags are much easier and cheaper than sparse point-level annotations (e.g., annotating 10% or 1% points). Moreover, there is no need to develop any customized labeling tools when adopting our weak supervision scheme, several off-the-shelf tools can be seamlessly integrated into our pipeline. We also noticed a certain performance gap between the proposed method and its fully-supervised counterpart (i.e., RandLA-Net [31]), primarily because of the relatively poor performance on the background categories (10% decrease in ceiling, floor, and wall in average). To improve the segmentation performance of background points, we will attempt to incorporate more advanced CAM techniques into our framework in our future works.

We also show the visualization results on the S3DIS dataset in Fig. 5. Note that, our Box2Seg only takes the box-level supervision and scene-level tags as its supervision signals. Although there are few misclassified segments, the proposed Box2Seg framework is able to capture semantic information effectively using 3D bounding boxes and scene-level tags as its supervision signal.
### 4.3 Evaluation on the ScanNet Dataset

The ScanNet [17] dataset consists of 1613 indoor 3D scenes, with 1201 scenes, 312 scenes, and 100 scenes being used for training, validation, and online test, respectively. Similar to the S3DIS dataset, the bounding box annotations are generated from instance labels. In this dataset, only floor and wall are considered as background classes.

It can be seen that our Box2Seg outperforms existing weak supervision frameworks MPRM [28] and WyPR [27], which use subcloud labels and scene-level tags, respectively. In addition, our method also achieves better performance than several early fully supervised approaches such as PointNet++ [6] and PointCNN [7], while our framework uses weak supervision signals only. It is noticed that there is a certain performance gap between our Box2Seg and the state-of-the-art fully supervised approaches. This is primarily due to the gap in the learning capacity of the backbone network. In general, sparse convolution-based methods [11, 12, 36] are more likely to achieve better results on this dataset. In future works, we will explore the incorporation of sparse convolution operation in our framework.
Can Box2Seg use different backbones? To evaluate the versatility of our framework, we further develop a Box2Seg variant by adopting PointNet++ [6] as its backbone. The results achieved by ablated networks on the S3DIS dataset [18] are shown in Table 3. It can be seen that Box2Seg (PointNet++) can also achieve a reasonable segmentation performance, with an insignificant performance decrease compared with its fully supervised counterpart. This clearly demonstrates that our framework is flexible and generic, and allows the use of different backbones for weakly supervised semantic segmentation of 3D point clouds. Due to page limited, more experimental results are provided in supplementary materials.

### 4.5 Ablation Study

**Settings**

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full supervision</td>
<td></td>
</tr>
<tr>
<td>PointNet++ [6]</td>
<td>32.3</td>
</tr>
<tr>
<td>RandLA-Net [31]</td>
<td>63.0</td>
</tr>
<tr>
<td>Weak supervision</td>
<td></td>
</tr>
<tr>
<td>Box2Seg (PointNet++)</td>
<td>44.5</td>
</tr>
<tr>
<td>Box2Seg (RandLA-Net)</td>
<td>57.7</td>
</tr>
</tbody>
</table>

Table 3: Quantitative results of different ablated networks achieved on Area-5 of the S3DIS [18] dataset.

### 5 Conclusion

In this paper, we propose a weakly supervised framework for semantic segmentation of unstructured 3D point clouds. The key of our approach is to learn pseudo labels inside and outside bounding boxes by exploring geometrical structures through a divide-and-conquer manner. Our experiments show that satisfactory semantic segmentation results can be achieved with only bounding box and scene-level tags. It would also be interesting to further generalize our framework to 3D instance segmentation and panoptic segmentation. Furthermore, we would like to consider the problem of learning from inaccurate point-level pseudo labels.
References


