Weakly-supervised Point Cloud Instance Segmentation with Geometric Priors

Heming Du[1,2,3], Xin Yu[2], Farookh Hussain[2], Mohammad Ali Armin[3], Lars Petersson[3], Weihao Li[3]
[1]Australian National University, [2]University of Technology Sydney, [3]Data61, CSIRO

Abstract

This paper investigates how to leverage more readily acquired annotations, i.e.*, 3D bounding boxes instead of dense point-wise labels, for instance segmentation. We propose a Weakly-supervised point cloud Instance Segmentation framework with Geometric Priors (WISGP) that allows segmentation models to be trained with 3D bounding boxes of instances. Considering intersections among bounding boxes in a scene would result in ambiguous labels, we first group points into two sets,* i.e.*, univocal and equivocal sets, indicating the certainty of a 3D point belonging to an instance, respectively. Specifically, 3D points with clear labels belong to the univocal set while the rest are grouped into the equivocal set. To assign reliable labels to points in the equivocal set, we design a Geometry-guided Label Propagation (GLP) scheme that progressively propagates labels to linked points based on geometric structure,* e.g.*, polygon meshes and superpoints. Afterwards, we train an instance segmentation model with the univocal points and equivocal points labeled by GLP, and then employ it to assign pseudo labels for the remainder of the unlabeled points. Lastly, we retrain the model with all the labeled points to achieve better instance segmentation performance. Experiments on large-scale datasets ScanNet-v2 and S3DIS demonstrate that WISGP is superior to competing weakly-supervised algorithms and even on par with a few fully-supervised ones.*

1. Introduction

Point cloud instance segmentation aims to classify 3D points into multiple objects of interest. Current approaches [9, 4, 25, 16, 2] commonly require point-level instance labels for training, where semantic and instance labels are manually assigned to each point. It is often time-consuming and laborious to label millions of points in each scene [21]. In comparison, it takes much less effort to annotate instances with 3D instance bounding boxes. While learning with point-level annotations has been widely studied, there lack solutions to leveraging 3D bounding boxes for instance segmentation.

While the community has seen effective solutions to fully supervised point cloud instance segmentation [27, 20, 4, 9], there lacks a solution to the weakly-supervised problem. On the one hand, directly applying existing fully-supervised methods to accommodate 3D bounding boxes would incur significant accuracy degradation, because points exhibited in multiple bounding boxes would introduce severe ambiguity to the network training. On the other hand, although several existing works [10, 28, 29] have investigated leveraging scribbles, image tags, and bounding boxes for 2D instance segmentation, it is non-trivial to adapt them to point clouds due to the distinctive natures of 2D pixels and 3D points.

In light of the above considerations, we propose a weakly-supervised framework for point cloud instance segmentation, which, based on 3D bounding box annotations, effectively incorporates local geometric priors of point clouds into the learning procedure. A 3D point generally belongs to one of the following cases: (i) a point is not within any bounding boxes, thus viewed as background, or (ii) a point lies in one or more than one bounding boxes, which is the focus of our research[1]. Based on the label reliability, we group points with clear labels into a *univocal* set. In general, those points only reside in a single bounding box, so their labels are trustworthy.

Unlike point-wise instance masks, 3D bounding boxes might intersect or even contain one another. As a result, it is challenging to estimate labels of points within the intersection areas. We categorize these points into an *equivocal* set. Since equivocal points do not have clear labels, we propose to explore geometric relationships between the univocal and equivocal points for reliable label assignment. Specifically, we adopt two generic and elementary structural representations of point clouds, *i.e.*, polygon meshes and superpoints, to capture the local geometry of scenes. Here, polygon meshes imply geometric connectivity among different points while superpoints indicate the appearance and spatial similarity within a local region. Then, we intro-

[1]Please note that erroneous 3D points have been removed during annotation.

duce a Geometry-guided Label Propagation (GLP) scheme to progressively propagate the labels of univocal points to the equivocal ones based on the geometric priors. In this manner, we can generate robust labels for points in the equivocal set.

After GLP, some equivocal points still remain unlabeled due to the incomplete or missing geometric connections among 3D points. Inspired by pseudo labeling [14], we propagate labels to unlabeled points by exploring their high-level semantic similarity to the labeled points. Thereby, we train an instance segmentation network with the univocal set and the equivocal points labeled by geometry-induced priors. Once the network is trained, we use it to assign pseudo labels to the unlabeled equivocal points. After obtaining pseudo-labels of the equivocal points, we will retrain the instance segmentation network to pursue better performance.

Experiments on two popular widely-used 3D indoor datasets ScanNet-V2 [3] and S3DIS [1] demonstrate the effectiveness of our method. In particular, our method significantly outperforms the baselines and achieves comparable performance with respect to the fully-supervised method PointGroup [9]. Moreover, our method is backbone-agnostic and can be conveniently incorporated into existing 3D point cloud instance segmentation networks, *i.e*., PointGroup [9] and SSTNet [16].

2. Related Works

Fully-supervised point cloud instance segmentation. Point cloud instance segmentation methods [27, 27, 7] group 3D points into different objects and predict their categories. They can be categorized into two groups: top-down and bottom-up. Top-down methods adopt a paradigm of detection followed by segmentation. For instance, 3D-BoNet [27] directly regresses bounding boxes for all instances and then predicts instance masks. Hou *et al*. [7] predict bounding boxes and estimate instance masks by fusing both geometric and color cues. Liu *et al*. [20] propose a Gaussian instance center network to predict instance center heatmaps. Bottom-up methods first obtain point-wise semantic labels and then group points into instances. Liu *et al*. [19] leverage sparse convolution to process point clouds and then predict point affinity. Wang *et al*. [25] propose to segment instances and semantics concurrently. Jiang *et al*. [9] estimate point offsets to object centers for clustering 3D instances. Engelmann *et al*. [4] introduce a graph convolutional network to refine proposal features. Furthermore, Liang *et al*. [16] and Chen *et al*. [2] improve segmentation performance by adopting hierarchical aggregation schemes. The above mentioned methods require point-level labels in training, and they will suffer dramatic performance degradation when only weak annotations are available.

Weakly-supervised 2D instance segmentation. Instance segmentation predicts a semantic label and an instance number for each pixel. Since obtaining pixel-wise annotation is time-consuming, weakly-supervised learning is an alternative way to bypass expensive annotations. Previous works mainly tackle the task by generating pseudo masks first and then retraining the segmentation model. Khoreva *et al*. [10] propose weakly-supervised semantic labeling to generate instance-level pseudo labels from 2D bounding boxes. Zhou *et al*. [29] take advantage of class activation maps [28] to obtain instance-level representations. Li *et al*. [15] introduce variation smoothing to produce high-quality pseudo masks. However, due to the different natures of 2D pixels and 3D points, these methods are not suitable to tackle point cloud instance segmentation.

Semi-/weakly-supervised 3D semantic segmentation. Semi-supervised 3D semantic segmentation methods leverage only a small portion of annotated 3D points as supervision to learn semantic segmentation. Xu *et al*. [26] approximate gradients of unlabeled points from labeled points to optimize their semantic segmentation network. Hou *et al*. [8] label 0.1% of points and train a 3D semantic segmentation network by encoding spatial information through contrastive learning. Furthermore, Liu *et al*. [21] introduce a self-training semantic segmentation approach to generate semantic pseudo labels from one point per object. Note that Liu *et al*.'s method focuses on semantic segmentation rather than instance segmentation. Compared with these methods, our work focuses on instance segmentation rather than semantic segmentation. Moreover, different from aforementioned sparse point supervision, 3D bounding box annotations will inevitably introduce noisy supervision to network training, thus posing great challenges to 3D instance segmentation. Liao *et al*. [17] propose a semi-supervised point cloud object detection and instance segmentation framework (SPIB) with parts of bounding boxes as supervision. Unlike the specifically designed architecture of SPIB, our method is designed to be a generic model-agnostic framework with weak supervision.

3. Proposed Method

In this work, we design a Weakly-supervised point cloud Instance Segmentation framework with Geometric Priors (WISGP) to segment instances from 3D bounding box annotations. We first group points that can obtain explicit labels from 3D bounding boxes into a univocal set P_u. Then, we group the rest of the points into an equivocal set P_e and propose Geometry-guided Label Propagation (GLP) to assign highly-confident labels to these points. In particular, we introduce geometric priors, *i.e*., polygon meshes and superpoints, to establish local geometric connections among 3D points and then propagate reliable labels to equivocal points iteratively. After GLP, unlabeled points may still exist. To finish the last piece of the puzzle, we predict highly-confident pseudo labels and then assign them to the remain-

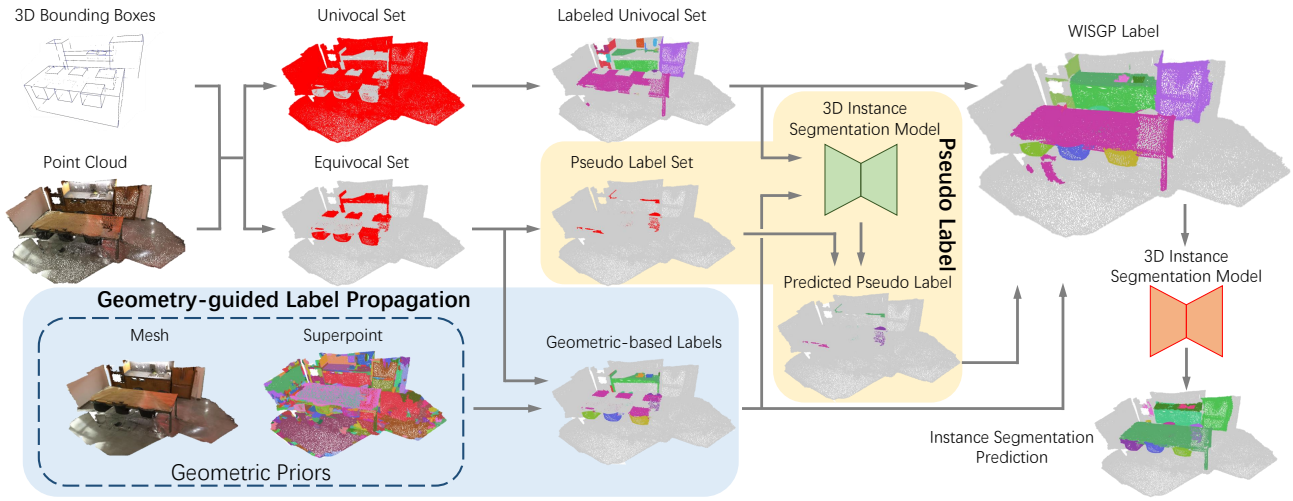

Figure 1: Pipeline of WISGP. Point clouds are split into two complementary sets: *univocal* and *equivocal* sets. We assign labels to univocal points and propagate labels to equivocal points with the help of geometric priors. Then, we train an instance segmentation model with labels of univocal points and geometry-induced labels to generate pseudo labels. Last, we retrain the point cloud instance segmentation model with the acquired labels.

ing unlabeled points. Finally, we train an instance segmentation model with high-quality point-level labels, as illustrated in Figure 1.

3.1. Univocal and Equivocal Sets

We observe that the spatial relationship between points and their occupying bounding boxes can roughly determine the certainty of a 3D point belonging to an instance. Based on the certainty of points, we split points into two complementary sets: a univocal set and an equivocal set. Then, we propose to propagate labels with high confidence (*i.e.*, univocal points) to uncertain points (*i.e.*, equivocal points). In this manner, we can obtain more highly-confident point labels for instance segmentation.

Univocal set. We categorize points that can achieve confident labels from 3D bounding boxes into a univocal set P_u. A point enclosed by only a single bounding box is assigned to the label of the bounding box and is regarded as a univocal point. Meanwhile, we find that some erroneous points caused by inaccurate 3D registration might appear in object bounding boxes. For those points, we manually label them by bounding boxes and then remove them without increasing labeling efforts, as shown in Figure 2a. In addition, for points that are outside all bounding boxes and do not belong to any object of interest, we consider them as background points.

Equivocal set. For the points residing in the intersection regions of bounding boxes, it is challenging to assign labels to them directly based on the annotated 3D bounding boxes. As a point only comes from one particular object, assigning multiple labels to a point would introduce ambiguity and thus misleads the instance segmentation network during training. For instance, as seen in Figure 2b, the gray points lie in the intersection of two 3D bounding boxes. Either misusing them as chair points or ignoring them would let a network misunderstand the 3D structure of a chair and the other object, *i.e.*, table in this case. Therefore, it is important to distinguish points located in the intersection areas of multiple 3D bounding boxes in order to infer correct object structure in instance segmentation. To this end, we group points located in the intersection area of multiple 3D bounding boxes into an equivocal set P_e, and then explore local geometric priors of point clouds and high-level semantic similarity to provide reliable labels to equivocal points, which is one of our key contributions.

3.2. Univocal Point Label Assignment

We first determine labels L_u for points in the univocal set by assigning semantic class c_i and instance identity id_i of the i-th 3D bounding box B_i to the univocal points residing in the bounding box B_i. Besides, we notice that some points that do not belong to any classes of interest or belong to backgrounds are occasionally included in a bounding box. For them, we did not manually remove them since they are physical points but do not belong to any categories. In fact, we hope a network can learn statistical structure of objects in training and thus ignore the side effects of incorporating those points to the univocal set. Last, points outside all 3D bounding boxes are regarded as background points.

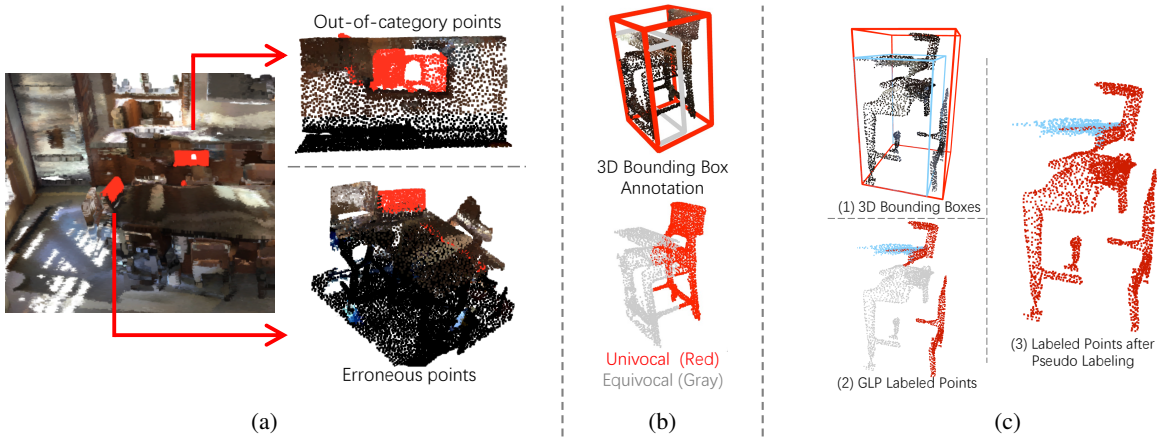

Figure 2: Different strategies to address different types of points. (a) Illustration of out-of-category and erroneous points. Microwave does not belong to annotated categories in ScanNet-v2, and red points in the top right of (a) belong to a microwave. In the bottom right of (a), red points are erroneous points caused by inaccurate the 3D registration or reconstruction process. (b) Demonstration of an intersection region of two 3D bounding boxes. Red and gray bounding boxes indicate bounding boxes of a chair and a table. Red points belong to the univocal set while equivocal points located in the intersection are highlighted in gray. (c) Pseudo labeling unlabeled equivocal points. In the bottom left of (c), GLP labels the chair and table by red and blue points, respectively. However, due to missing geometric connections among points, gray equivocal points are still unlabeled after GLP. In the right side of (c), we propagate labels for these unlabeled equivocal points by pseudo labeling.

3.3. Equivocal Point Label Assignment

To fully exploit bounding box annotations in instance segmentation, we aim to further mine the information from the equivocal set. Our motivation is to propagate labels of univocal points L_u to the equivocal points which are geometrically linked to the univocal points. To be specific, since equivocal points are located in the intersection volumes, we first measure the relative spatial relationships among 3D bounding boxes to decide the necessity of label prorogation. Then, we introduce geometric prior knowledge, including polygon meshes and superpoints, to deduce relationships among points. The geometric priors provide strong clues for us to propagate labels with high-confidence.

3.3.1 Bounding box spatial relationship inference

The spatial relationships among overlapping 3D bounding boxes typically fall into two cases: (i) inclusion relationship: a bounding box lies in another one, or (ii) overlapping relationship: bounding boxes intersect with others. In order to deduce the spatial relationship of bounding boxes, we calculate the intersection score $S_{i|j} = \frac{|P_i \cap P_j|}{|P_i|}$ on the point level, where P_i represents the set of points located in the i-th bounding box. In our experiments, we set the intersection score threshold to 0.9. The point set P_i is the subset of P_j, if S_i is over the threshold. In other words, the ith 3D bounding box is considered as being contained in the j-th 3D bounding box. Otherwise, P_i and P_j are regarded as overlapping sets, and we remark the relationship between the i-th and j-th 3D bounding boxes as an overlapping relationship.

3.3.2 Point label propagation via geometric priors.

In accordance with the different spatial relationships among 3D bounding boxes, we present a geometry-guided label propagation (GLP) scheme to estimate the semantic labels of equivocal points in different scenarios as follows.

Inclusion relationship. If the intersection score $S_{i|j}$ indicates the existence of an inclusion relationship between two bounding boxes, we presume that all the points in the intersection area mainly represent an object enclosed by the included bounding box. In other words, let B_i represent a 3D bounding box that is included in another bounding box B_j, and equivocal points located in the intersection area belong to the instance indicated by B_i. Therefore, we assign the semantic labels c_i of the enclosed bounding box B_i to these equivocal points. Note that if a bounding-box has been enclosed by another one but still has partial overlapping with others, the overlapping points in the bounding box should not be labeled by the label of the bounding box. Instead, we will switch to the overlapping scenario to label the equivocal points of the bounding box.

Overlapping relationship. When multiple bounding boxes overlap, we aim to designate one specific semantic category for each equivocal point. To achieve this goal, we estimate semantic labels of these points based on the as-

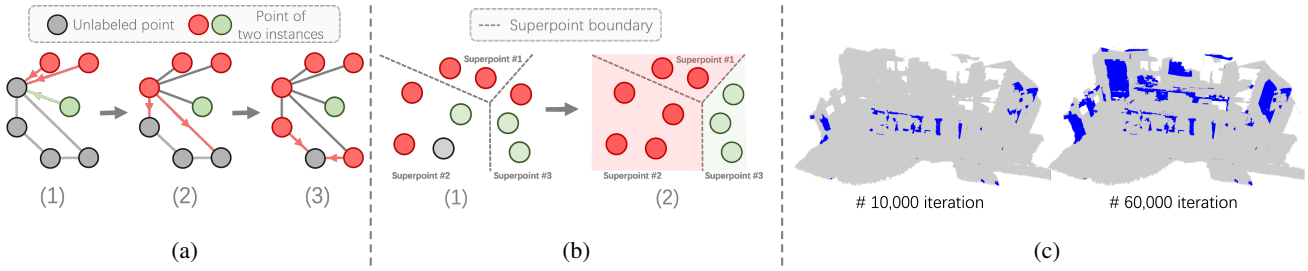

Figure 3: Illustration of geometry-guided label propagation. (a) Mesh based label propagation. (b) Superpoint based label propagation. Unlabeled equivocal points are highlighted in gray, while red and green points represent points belonging to two different instances, respectively. Given superpoints, we not only propagate labels to the unlabeled points but also smooth the noise labels in each superpoint. (c) Geometry-guided label propagation as the iterations proceed. Labeled points are highlighted in blue. Visualization of intermediate iteration results will be provided in the supplementary material.

signed labels of surrounding points. Specifically, we select the most common semantic class from n neighbors with assigned semantic labels $C = \{c_1, \ldots, c_n\}$, which starts from univocal points, as illustrated in Figure 3a. Considering the point cloud is sparse and irregular, retrieving neighboring points with Euclidean distance might be unsuitable and would neglect geometric structure of objects and scene layouts. To tackle this issue, we introduce polygon meshes to establish the geometric relationship among points and then measure the relationship between two points. Specifically, points that are enclosed by the same polygon mesh are considered as neighbors, and meshes are constructed by Marching Cubes algorithm [22] from 3D point clouds.

In addition, we introduce superpoints [16] to smooth semantic labels in local regions, as visible in Figure 3b. In our experiments, superpoints are the result of a hand-crafted graph-based segmentation method [5]. Given point coordinates and colors, the graph-based segmentation method groups mesh-connected vertices into a superpoint according to their appearance and spatial location similarities. Although the results of the graph-based segmentator are still coarse, superpoints can provide complementary appearance and geometric hints to polygon meshes, indicating that points within a local area should share the same label. Thanks to the superpoints, we can more reliably propagate semantic labels. To spread the labels to a large extent, we run GLP iteratively, as shown in Figure 3c (results from the intermediate iterations will be provided in the supplementary material.). When no more points are incorporated, our GLP terminates.

Instance label propagation. Instance labels are also essential for 3D instance segmentation model training. After assigning semantic labels to equivocal points, we also need to determine the instance labels for those points. Since an object holds unique semantic and instance labels, points within an object instance should share the same semantic label. Following this, we take advantage of semantic labels as constraints in assigning instance labels to equivocal points. To be specific, for each equivocal point, we first search the most common instance label from its neighboring points that have the same semantic label as the equivocal point. Here, we only consider neighboring points with assigned instance labels for label propagation. Similar to semantic label estimation, we employ polygon meshes rather than Euclidean distance to search neighbors of equivocal points. Moreover, we also propagate the instance labels to larger regions by iteratively repeating the same procedure, as illustrated in Figure 3c (the intermediate results of the instance label propagation will be shown in the supplementary material).

Pseudo label on unlabeled equivocal points. After GLP, some equivocal points are still unlabeled because of incomplete or missing geometric connections among 3D points. As demonstrated in Figure 2c, unlabeled equivocal points are usually located in isolated regions without connecting to labeled points geometrically. Inspired from [14], we aim to predict pseudo labels for unlabeled equivocal points, denoted as a pseudo label set P_{pl}. The pseudo label set is a subset of the equivocal set P_e. In this manner, we are able to achieve more labeled points for final instance segmentation learning. Here, we propagate label information via high-level semantic similarity and instance closeness learned by a neural network.

Considering labels for the univocal set and the equivocal set labeled by GLP are reliable, we only generate pseudo labels for points in P_{pl} by an instance segmentation network. To be specific, we first train an instance segmentation model with labels of univocal points and GLP-labeled equivocal points. Then, we assign pseudo labels predicted by the trained model to the unlabeled equivocal points. As a result, we obtain point-level labels from 3D bounding boxes for instance segmentation. Then, we can train an instance segmentation network with the generated point-level labels. Note that, our label generation procedure is generic and ag-

Table 1: Comparison of different supervisions trained with PointGroup [9] and SSTNet [16] on ScanNet-v2. The upper part demonstrates the results with PointGroup and the lower one shows the results with SSTNet. Specifically, the top row of each part shows the results of PointGroup and SSTNet with full supervision, respectively. The bottom part shows the results of instance segmentation models with weak supervision. $*$ stands for the model with full supervision. † represents using PointGroup as the segmentation model, while ‡ stands for adopting SSTNet in training.

Method	mAP	bathtub otherfur.	bed picture	booksh. refriger.	cabinet shower.	chair sink	counter sofa	curtain table	desk toilet	door window
PointGroup$*$ [9]	0.348	0.597 0.339	0.376 0.208	0.267 0.246	0.253 0.416	0.712 0.298	0.069 0.434	0.266 0.385	0.140 0.758	0.229 0.275
Baseline†	0.251	0.313 0.248	0.243 0.219	0.232 0.235	0.197 0.261	0.572 0.121	0.055 0.334	**0.265** 0.209	0.050 0.631	0.131 0.197
WISGP †	**0.313**	**0.402** **0.309**	**0.347** **0.262**	**0.262** **0.307**	**0.272** **0.331**	**0.691** **0.238**	**0.059** **0.339**	0.199 **0.391**	**0.087** **0.737**	**0.182** **0.224**
SSTNet$*$ [16]	0.494	0.777 0.520	0.566 0.403	0.258 0.438	0.406 0.489	0.818 0.549	0.225 0.526	0.384 0.557	0.281 0.929	0.429 0.343
Baseline ‡	0.293	0.290 0.312	**0.351** **0.339**	**0.248** **0.322**	0.186 0.253	0.661 0.171	**0.092** **0.427**	0.208 0.343	0.093 0.665	0.233 0.257
WISGP ‡	**0.352**	**0.455** **0.330**	0.328 0.284	0.238 0.314	**0.304** **0.321**	**0.753** **0.329**	0.088 **0.427**	**0.239** **0.394**	**0.176** **0.834**	**0.278** **0.259**

Table 2: Results on S3DIS. $*$, †, ‡ indicate fully-supervised models, PointGroup backbone, and SSTNet backbone, respectively.

Method	mAP	AP@50	mPrec	mRec
PointGroup$*$ [9]	-	0.578	0.619	0.642
Baseline †	0.232	0.352	0.407	0.415
WISGP †	**0.335**	**0.486**	**0.500**	**0.528**
SSTNet$*$ [16]	0.427	0.593	0.655	0.642
Baseline ‡	0.282	0.412	0.377	0.479
WISGP ‡	**0.372**	**0.510**	**0.443**	**0.567**

nostic against different point cloud instance segmentation networks.

4. Experiments

To validate the effectiveness of proposed WISGP, we conduct extensive experiments on challenging real-world scenes, *i.e.*, ScanNet-V2 [3] dataset and S3DIS [1]. Furthermore, to demonstrate the superiority of our proposed method when only bounding box annotations are available, we compare with two state-of-the-art instance segmentation architectures *i.e.*, PointGroup [9] and SSTNet [16].

PointGroup [9] adopts a U-Net architecture with Submanifold Sparse Convolution (SSC) and Sparse Convolution (SC) [6] and predicts a semantic score and offset vector per point. PointGroup clusters points into instances twice with original coordinates and shifted coordinates according to the semantic and affinity predictions. SSTNet [16] employs a Sparse Convolution based U-Net to simultaneously predict semantic and affinity. Furthermore, SSTNet proposes a Semantic Superpoint Tree Network to cluster points and a CliqueNet to prune errors during grouping instances. For clarification, when we compare with different methods, we adopt the same network architecture as the competing methods.

Table 3: Comparison with SPIB [17] on the ScanNet-v2 validation set. †, ‡ indicate PointGroup backbone, and SSTNet backbone, respectively. Note that SPIB uses all the training annotations.

Method	mAP	AP@50	AP@25
SPIB [17]	-	-	0.614
WISGP †	0.313	0.502	0.649
WISGP ‡	**0.352**	**0.569**	**0.702**

4.1. Dataset and Evaluation

ScanNet-v2 [3] has 1,613 indoor scenes with 18 instance classes. The dataset is split into training, validation, and test, containing 1,201, 312, and 100 scenes, respectively. We acquire 3D bounding boxes by following the procedure in VoteNet [24]. As bounding box annotations do not label the floor and wall categories, we treat these two categories as a background class in ScanNet-V2.

S3DIS [1], known as Stanford 3D Indoor Scene Dataset (S3DIS) dataset, contains 6 large-scale indoor areas with 271 rooms. Each point in a scene point cloud is annotated by one of the 13 semantic categories. Following the standard training and testing splits [1, 9], we train methods on

Table 4: Impacts of different point sets.

Univocal Set	Equivocal Set				mAP	AP@50	AP@25
	GLP w/ inclusion	GLP w/ mesh	GLP w/ superpoint	Pseudo labeling			
✓					0.251	0.482	0.643
✓	✓		✓		0.271	0.476	0.634
✓	✓	✓			0.262	0.482	0.662
✓	✓	✓	✓		0.289	0.515	0.676
✓	✓	✓	✓	✓	**0.313**	**0.529**	**0.693**

Table 5: Impacts of pseudo labels.

Pseudo Label	mAP	AP@50	AP@25
Both univocal and equivocal sets	0.282	0.513	0.674
Equivocal set alone	0.292	0.513	0.672
WISGP	**0.313**	**0.529**	**0.693**

Area 1, 2, 3, 4, 6 and then evaluate them on Area 5. Additionally, we adopt 3D bounding boxes provided in Stanford 2D-3D-Semantics (2D-3D-S) dataset [1].

We train our model on the training dataset and evaluate on the validation set for ScanNet-V2 and on the testing set for S3DIS. In order to ensure fairness, we report the performance on the model trained with the same training epochs as the compared methods. For ablation studies, we employ PointGroup as the instance segmentation network to demonstrate the contributions of our proposed components.

Following the work [3], we use widely-adopted evaluation metrics: mean average precision [18] at overlap 0.25 (mAP@25), overlap at 0.5 (mAP@50) and overlap in a range $[0.5 : 0.05 : 0.95]$ (mAP). Meanwhile, similar to the methods [9, 16], we adopt the mean precision (mPrec) and mean recall (mRec) with IoU threshold 0.5 to evaluate methods on the S3DIS dataset.

4.2. Implementation Details

We employ the Adam optimizer [11] to train PointGroup [9] with the batch size of 12 and a learning rate of 10^{-3} on ScanNet-v2 dataset. We train models with 384 epochs on 4 Nvidia P100 for 50 hours. In addition, following the publicly released training configuration of SSTNet [16] on ScanNet-v2, we train SSTNet with the AdamW optimizer [23] for 512 epochs. For training on S3DIS, we adopt similar configurations as in ScanNet-v2.

Furthermore, surface meshes are provided in ScanNet-v2 and S3DIS. To be specific, surface meshes are acquired using the Marching Cubes algorithm [22] on the implicit TSDF. The superpoints of ScanNet-v2 are obtained by applying a 3D adapted graph-based segmentation algorithm [5, 3]. For superpoints of S3DIS, we adopt Supervized SuperPoint (SSP) [12] and SuperPoint Graph (SPG) [13] to generate superpoints over point clouds, following the procedure in SSTNet [16]. Note that our geometric priors, *i.e*. surface mesh and superpoints, are produced during data pre-processing. Thus, there is no extra time consumption on surface mesh and superpoint generation in training. We will release our code and data, to facilitate reproducibility and future work.

4.3. Main Results

We present the performance of instance segmentation models with WISGP on the validation set of ScanNet-v2 and the test set of S3DIS in the Table 1 and Table 2. In order to demonstrate the improvement of our method, we treat models trained on the univocal set as our baselines.

As demonstrated in Table 1, WISGP outperforms corresponding baselines by a large margin on ScanNet-v2. Compared with the baseline, our result is 24% higher on PointGroup and 20.1% higher on SSTNet. Meanwhile, our weakly-supervised method achieves 89.9% and 71.2% mAP of fully supervised PointGroup and SSTNet, respectively. With PointGroup, WISGP achieves higher performance than the baseline on all classes except curtain. With SSTNet, WISGP achieves higher performance than the baseline on 13 classes out of 18 classes. We notice that points on object boundaries that have been contained in other 3D bounding boxes, such as pictures and refrigerators, could influence network training. Therefore, models trained with our framework significantly outperform baselines. Moreover, since Liao *et al*. [17] did not release code and pertinent data for instance segmentation, it is difficult to provide comparable results on the other evaluation metrics. Thus, we compare SPIB on mAP@25, and WISGP performs 8.8% better than SPIB on mAP@25, as demonstrated in Table 3.

In general, fully supervised methods with point-level labeling can be considered as an upper bound for our weakly-supervised method with bounding box annotations. Surprisingly, we observe that WISGP outperforms its full-supervised counterpart on 4 classes (*e.g*. cabinet and picture) with PointGroup. Furthermore, WISGP also achieves a remarkable improvement over the baseline of PointGroup and SSTNet across all the evaluation metrics on S3DIS, as shown in Table 2. Our method is 44.4% higher than the baseline of PointGroup and outperforms the baseline of SSTNet by 31.9% on mAP. Meanwhile, WISGP achieves nearly 82% performance of fully supervised versions with either PointGroup or SSTNet backbones. Compared with ScanNet-v2, the mesh of a room in S3DIS is rather coarse, which means there would be multiple points on the same mesh face. Therefore, we further establish connections

among points between neighboring mesh faces. As expected, our WISGP model gains significant improvements over baselines on S3DIS. All in all, the superior performance of WISGP on both ScanNet-v2 and S3DIS implies that WISGP achieves a promising generalization ability thanks to the generic local geometric priors. The visual results of our methods on both datasets are provided in the supplementary material due to the page limit.

4.4. Ablation Study

Impacts of different point sets. In order to analyze the impacts of model training with different sets, we demonstrate the comparisons of adopting (i) the univocal set (Univocal Set), (ii) both the univocal and labeled equivocal sets (Univocal + full GLP), and (iii) all points in the 3D scene (WISGP), as indicated in Table 4. Compared to the model trained on the univocal set only, GLP improves performance for instance segmentators by propagating labels to equivocal points according to geometric priors. This comparison suggests that GLP components provide some equivocal points with reliable labels. With these reliable labels, segmentation models increase the instance segmentation accuracy. Furthermore, after adopting predicted pseudo labels on the unlabeled equivocal points, we observe further performance improvement. It implies the effectiveness of our pseudo labeling.

Label propagation based on different geometric priors. To analyze the impacts of geometric priors, we present the ablation on using different geometric priors in label propagation for the equivocal points, as shown in Table 4. We separately remove polygon meshes and superpoints from GLP and notice that both geometric priors improve the performance of segmentation models. Furthermore, compared to GLP without superpoints, we observe that smoothing labels in local regions based on superpoints lead to a significant improvement on ScanNet-v2.

Impact of pseudo labeling. Table 5 shows the comparison of generating pseudo labels on (i) both the univocal and labeled equivocal sets, (ii) the labeled equivocal set, and (iii) the unlabeled set. As indicated in Table 5, assigning pseudo labels to either the univocal set or the labeled equivocal set degrades the performance of instance segmentation. This implies that both labels of univocal points and equivocal points obtained from 3D bounding boxes and geometric priors are more reliable. On the other hand, some points in the equivocal set are not annotated after adopting GLP and are ignored during training. Furthermore, after pseudo labeling unlabeled equivocal points, we observe the performance improvement. This indicates that this fashion of employing the pseudo labels is more suitable and thus our method can better exploit the information of 3D points.

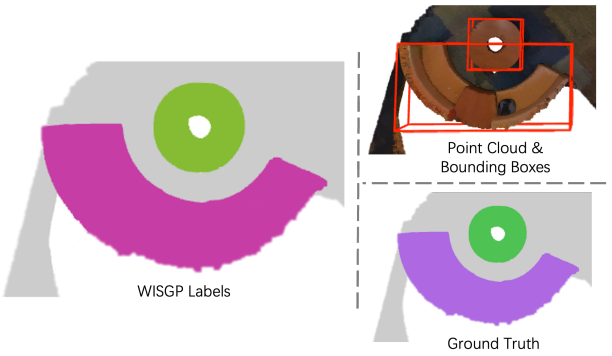

Figure 4: Demonstration of WISGP labels on objects with irregular shapes.

5. Discussion and Limitation

Different from point-level annotations, bounding-boxes might be not effective to annotate highly irregular objects. This would affect the final instance segmentation performance. However, highly irregular objects rarely appear in our experiments. In Figure 4, our method can produce reliable labels on objects with irregular shapes, *i.e.*, half ring sofa, by introducing the geometric priors. This indicates that our WISGP achieves good generalization ability. In addition, when labeling bounding-boxes, some object points which are outside the categories of interest may appear in a univocal set. These points could degrade the final segmentation performance as a network may recognize them as one of the classes of interest. To ameliorate this issue, we can actually ask annotators to remove those objects similar to the erroneous points during labeling without increasing too many manual efforts.

6. Conclusion

In this paper, we proposed a weakly-supervised point cloud Instance segmentation framework by fully exploiting local geometric priors of point clouds, namely WISGP. Benefiting from the introduction of local geometric priors represented by polygon meshes and superpoints, our framework effectively propagates reliable point-level labels to the neighboring points within multiple bounding boxes. We further leverage pseudo labeling to propagate labels to unlabeled points that share the high-level semantic similarity with the labeled ones. By fully exploring the geometric and semantic similarities of 3D scenes, we obtain high-quality point-level annotations, leading to promising instance segmentation performance. More importantly, our framework is model-agnostic. With our WISGP, fully-supervised methods can be easily accommodated with 3D bounding box annotations for instance segmentation.

References

[1] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, Feb. 2017.

[2] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15467–15476, October 2021.

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[4] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020.

[5] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[6] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[7] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019.

[8] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021.

[9] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Loic Landrieu and Mohamed Boussaha. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7440–7449, 2019.

[13] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.

[14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[15] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6964–6973, 2021.

[16] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2783–2792, 2021.

[17] Yongbin Liao, Hongyuan Zhu, Yanggang Zhang, Chuangguan Ye, Tao Chen, and Jiayuan Fan. Point cloud instance segmentation with semi-supervised bounding-box mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*, 2019.

[20] Shih-Hung Liu, Shang-Yi Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *arXiv preprint arXiv:2007.09860*, 2020.

[21] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1726–1736, 2021.

[22] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[25] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2019.

[26] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13706–13715, 2020.

[27] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *arXiv preprint arXiv:1906.01140*, 2019.

[28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[29] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.