GaPro: Box-Supervised 3D Point Cloud Instance Segmentation Using Gaussian Processes as Pseudo Labelers

Tuan Duc Ngo Binh-Son Hua Khoi Nguyen
VinAI Research, Hanoi, Vietnam
{v.tuannd42, v.sonhb, v.khoindm}@vinai.io

Abstract

Instance segmentation on 3D point clouds (3DIS) is a longstanding challenge in computer vision, where state-of-the-art methods are mainly based on full supervision. As annotating ground truth dense instance masks is tedious and expensive, solving 3DIS with weak supervision has become more practical. In this paper, we propose GaPro, a new instance segmentation for 3D point clouds using axis-aligned 3D bounding box supervision. Our two-step approach involves generating pseudo labels from box annotations and training a 3DIS network with the resulting labels. Additionally, we employ the self-training strategy to improve the performance of our method further. We devise an effective Gaussian Process to generate pseudo instance masks from the bounding boxes and resolve ambiguities when they overlap, resulting in pseudo instance masks with their uncertainty values. Our experiments show that GaPro outperforms previous weakly supervised 3D instance segmentation methods and has competitive performance compared to state-of-the-art fully supervised ones. Furthermore, we demonstrate the robustness of our approach, where we can adapt various state-of-the-art fully supervised methods to the weak supervision task by using our pseudo labels for training. The source code and trained models are available at https://github.com/VinAIResearch/GaPro.

1. Introduction

This paper addresses the challenging problem of box-supervised 3D point cloud instance segmentation (BS-3DIS), which seeks to segment every point into instances of predefined classes using only axis-aligned 3D bounding boxes as supervision during training. This problem arises to address the huge annotating cost of fully-supervised 3D point cloud instance segmentation (3DIS) where every point in the point cloud is manually labeled. Compared to 3DIS, BS-3DIS is considered significantly harder. First, axis-aligned boxes cannot capture the shape or geometry of objects as they only represent the very coarse extent of the objects. Second, unlike instance mask where points only belong to at most one mask, points can belong to multiple boxes as visualized in Fig. 1, resulting in the ambiguous point-object assignment.

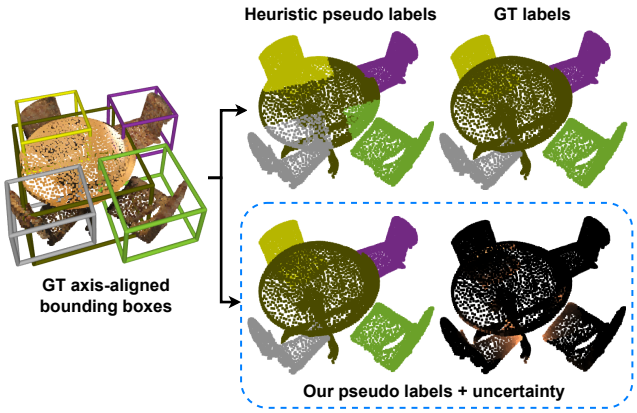

Figure 1: Weakly supervised instance segmentation relies on high-quality pseudo labels to achieve competitive performance. Given only axis-aligned bounding box annotations, pseudo labels based on heuristics [5] often have large errors in box overlapping regions and thus yield inferior performance. Our GaPro predicts the pseudo labels and their confidence using Gaussian Processes, via resolving the ambiguity in box overlapping regions.

The task of box-supervised 3D point cloud instance segmentation has received little attention, with Box2Mask [5] being the first attempt. However, due to the ambiguity in point-object assignments, the point-wise predicted boxes are unreliable for clustering. This leads to a significant performance gap compared to fully supervised methods, such as Mask3D [33], which achieves an mAP of 55.2 on ScanNetV2 [6], compared to Box2Mask's 39.1 (around 30%) using the same backbone. Furthermore, Box2Mask is not adaptable to new advances in fully supervised 3DIS, as it is designed as a standalone method.

To address these limitations, we propose a novel pseudo-labeling method that can be used as a universal plugin for any 3DIS network and offers an instant solution for any new fully supervised 3DIS approach, with a smaller performance gap between fully supervised and BS-3DIS versions, typically

around 10%. In particular, we formulate it as a learning problem with two unknowns: the network's parameters and the ground-truth object masks. Our goal is to construct pseudo object masks from box supervision and optimize the network's parameters using these pseudo labels. To achieve this, we propose using Gaussian Process (GP) on each pair of overlapping 3D bounding boxes to infer the optimal pseudo labels of object masks and their uncertainty values, which are constrained by the given 3D bounding boxes. Next, we modify a 3DIS network to predict additional uncertainty values along with the object mask to match the inferred pseudo labels obtained from the GP. GP plays a key role in our approach. First, it models the similarity relationship among regions of the point cloud, which enables effective label propagation from determined regions (belonging to a single box) to undetermined regions (belonging to multiple boxes). Second, it estimates the uncertainty of the predictions with weak labels, providing informative indications for annotators to correct uncertain regions of pseudo labels for training the 3D instance segmentation network.

We evaluate our approach on various state-of-the-art 3DIS methods, including PointGroup [19], SSTNet [26], SoftGroup [40], ISBNet [30], and SPFormer [36], using two challenging datasets: ScanNetV2 [6] and S3DIS [1]. Our box-supervised versions of these methods achieve comparable performance to their fully-supervised counterparts on both datasets, outperforming other weakly-supervised 3DIS methods significantly.

In summary, the contributions of our work are as follows:

- We propose GaPro, a weakly-supervised 3DIS method based on 3D bounding box supervision. We devise a systematic approach to generate pseudo object masks from 3D axis-aligned bounding boxes so that fully supervised 3DIS methods can be retargeted for weak supervision purposes.
- We propose an efficient Gaussian Process to resolve the ambiguity of pseudo labels in the overlapped region of two or more bounding boxes by inferring both the pseudo masks and their uncertainty values.
- Our GaPro achieves competitive performance with the SOTA fully-supervised approaches and outperforms other weakly-supervised methods by a large margin on both ScanNetV2 and S3DIS datasets.

In the following, Sec. 2 reviews prior work; Sec. 3 specifies GaPro; and Sec. 4 presents our implementation details and experimental results. Sec. 5 concludes with some remarks and discussions.

2. Related Work

This section reviews some related work on 3D point cloud instance segmentation and weakly-supervised instance segmentation in 2D and 3D, and the usage of the Gaussian Process in the 3D point cloud.

3D Point Cloud Instance Segmentation (3DIS) approaches are categorized into box-based, cluster-based, and dynamic convolution (DC)-based methods. Box-based methods [15, 45, 47] detect and segment the foreground region inside each 3D proposal box to get instance masks. Cluster-based methods cluster points into instances based on the predicted object centroid [42, 19, 2, 40, 7], or build a tree/graph then cut the subtrees/subgraphs as clusters [26, 18]. DC-based methods [13, 36, 14, 43, 33, 27] generate kernels representing different object instances to convolve with point-wise features to produce instance masks. Among these methods, DC-based approaches are preferred due to their superior performance, since they do not rely on error-prone intermediate predictions like proposal boxes or clusters. However, fully-supervised 3DIS approaches require costly point-wise instance annotation for training which hinders their application in practice. Our proposed approach only uses 3D instance boxes (represented by two points) as supervision, which is much cheaper to obtain. Our approach can be applied to all the aforementioned fully-supervised 3DIS approaches, allowing them to transform into BS-3DIS versions.

Weakly-supervised 2D image instance segmentation aims to segment images into instances of predefined classes using weaker supervision than instance masks. Different types of weak supervision include image-level classes [10, 22, 50], instance points [3, 37, 11], and instance boxes [17, 38, 48, 23, 21, 24, 4, 46, 25, 20]. Box supervision is particularly attractive because it provides a stronger signal for training with only two points per instance. Box-supervised approaches (BS-2DIS) compensate for the lack of ground-truth masks by regularizing the training of instance segmenters with priors. Various methods have been proposed for BS-2DIS, such as BoxInst [38] with tight-box prior loss and color smoothness, LevelSetBox [24] with level set evolution, Mask Auto-Labelers [20] using Conditional Random Fields, and BoxTeacher [4] employing consistency regularization of the Mean-teacher technique to generate pseudo instance masks conditioned by ground-truth boxes. Although BS-2DIS is less challenging than BS-3DIS, the structured and dense properties of 2D images that these regularization techniques imply do not hold in 3D point clouds, thus, we cannot trivially apply these methods in BS-3DIS.

Box-supervised 3D point cloud instance segmentation (BS-3DIS) aims to segment all instances of predefined classes, utilizing the supervision of axis-aligned 3D bounding boxes, which correspond to two 3D points per instance. Compared to point supervision techniques such as PointContrast [44] and CSC [16], BS-3DIS [5, 9] is considered more appropriate in 3DIS segmentation with less supervision. This is because the former provides valuable information about object extent through its only one bounding box per instance whereas the latter relies on selecting specific labeled

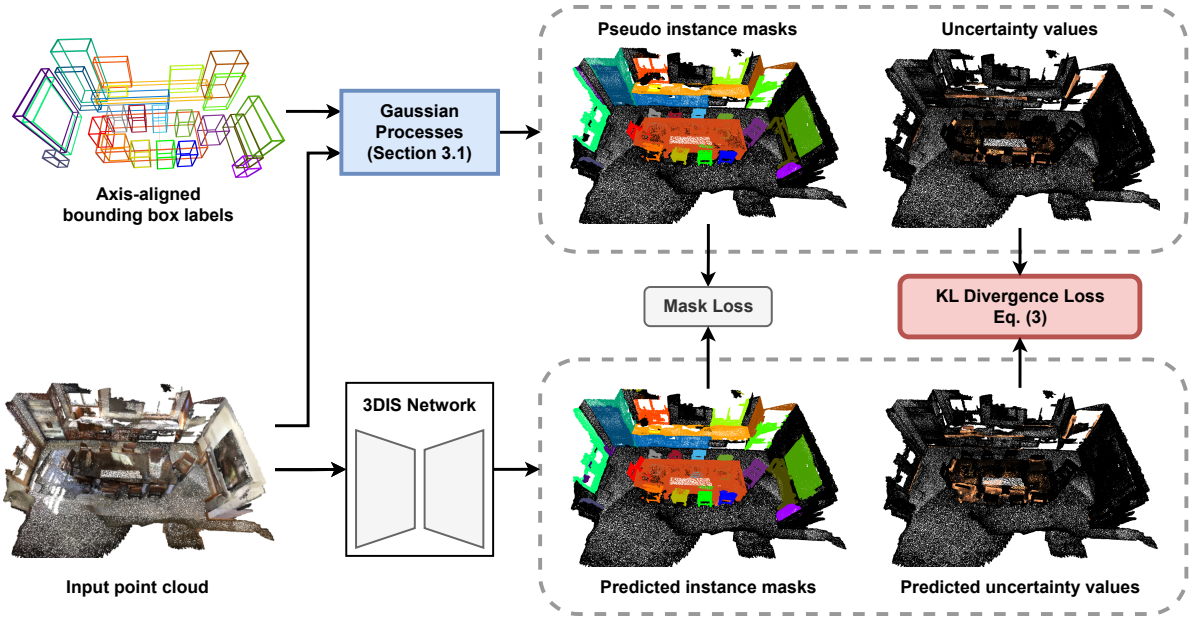

Figure 2: **Overall architecture of our approach**. GaPro is a two-step approach consisting of leveraging Gaussian Processes to generate pseudo instance masks and their uncertainty values, and training a 3DIS network to match its prediction against these pseudo labels with a new KL divergence loss along with the mask loss.

points, resulting in more sensitive results. Box2Mask [5] was the first to introduce BS-3DIS utilizing point clustering to group points based on their predicted bounding boxes. WISGP [9] employs simple heuristics to propagate labels from determined points to undetermined points and uses the pseudo labels to train a fully-supervised 3DIS model. In contrast, our proposed approach utilizes uncertainty when predicting object masks with weak labels as additional pseudo labels. Furthermore, our approach incorporates Gaussian Processes to model pairwise similarity between regions, including determined-determined, determined-undetermined, and undetermined-undetermined relationships. This results in a more effective global label propagation than the local propagation between neighboring points utilized by [9].

Gaussian Process (GP) in 3D point cloud methods including [34, 8, 39] leverage GP to model the relationship among regions to predict semantic segmentation in the fully-supervised setting. On the other hand, our approach utilizes GP in the weakly-supervised setting of 3D instance segmentation, that is, to estimate the distribution of object masks from the provided GT 3D boxes to train a 3DIS network.

3. Our Approach

Problem statement: In training, we are given a 3D point cloud $\mathbf{P} \in \mathbb{R}^{N \times 6}$ where N is the number of points, and each point is represented by a 3D position and RGB color vector. We are also provided a set of 3D axis-aligned bounding boxes $\mathbf{B} \in \mathbb{R}^{K \times 6}$ and their classes $\mathbf{L} \in \{1, \dots, C\}^{K \times 1}$, where K is the number of instances and C is the number of object classes, as the box-supervision. Each bounding box is represented by two corners with minimum and maximum XYZ coordinates. Our approach, GaPro, attempts to generate pseudo object masks of these K instances, $\mathbf{M} \in \{0, 1\}^{K \times N}$, and use them to train a 3DIS network Φ. In testing, given a new point cloud $\mathbf{P'} \in \mathbb{R}^{N' \times 6}$, Φ predicts the masks $\widehat{\mathbf{M}} \in \{0, 1\}^{K' \times N'}$ of all K' instances of the C object classes.

The overall architecture of GaPro is depicted in Fig. 2, which is a two-step approach that involves generating pseudo instance masks and their uncertainty values from box annotations with Gaussian Processes and training a 3DIS network with the resulting labels with a devised KL divergence loss along with the previous mask loss.

3.1. Gaussian Processes as Pseudo Labelers

We observed that 3D point clouds are sparse, and if a 3D point is within an axis-aligned bounding box representing an instance, it likely belongs to that instance. Using this geometric prior, we can roughly assign points to instances to generate pseudo object masks. However, as the axis-aligned bounding boxes do not accurately fit the complex shapes of objects, there are often overlapped regions among these boxes, leading to points belonging to multiple instances, as shown in Fig. 1. Consequently, assigning points to instances becomes challenging. To overcome this issue, we propose using Gaussian Process (GP) as a probabilistic assigner to resolve conflicts that arise from overlapping boxes. We choose the Gaussian process for two reasons. Firstly, GP consid-

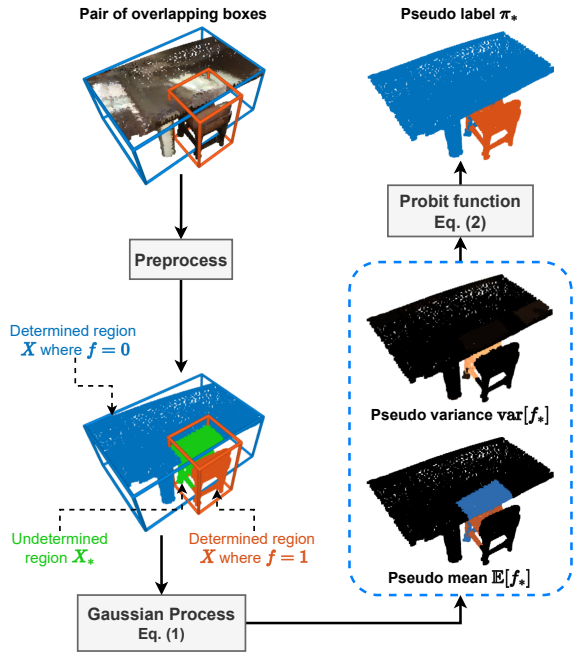

Figure 3: **Our Gaussian Process**. For each pair of overlapping boxes, the determined and undetermined regions are identified and taken as input into a Gaussian Process to produce pseudo mean and variance values. Then the Probit function is utilized to output the posterior Bernoulli distribution as pseudo labels.

ers the complete relationships among regions, allowing the similarity between determined regions and the similarity between undetermined regions to affect label propagation from determined to undetermined regions. Secondly, GP outputs a probabilistic distribution, enabling the modeling of uncertainty in the pseudo labels.

To begin, we divide the input point cloud into two non-overlapping sets: the *determined set* and the *undetermined set*. The determined set includes points that belong to at most one bounding box, and we assign these points to the corresponding label of the bounding box that encloses them. Points outside all bounding boxes are labeled as background. However, in the undetermined set, it is challenging to assign the correct labels to points that reside in the overlapped regions of bounding boxes. To solve this problem, we treat the assignment of points in the overlapped region of two boxes as a binary classification task and use the Gaussian Process as a probabilistic classifier.

While there are some regions that result from the intersections of more than two boxes, our analysis of the overlapping box labels in the ScanNetV2 [6] and S3DIS [1] 3DIS datasets shows that 95.4% of cases involve only two boxes, and the remainder involve three or four boxes. In these infrequent cases, we select the pair with the largest overlap to use for the GP. Additionally, both datasets include superpoints – clusters of points grouped together based on their RGB color and position values. We can use these superpoints as elements in the GP rather than individual points, which can help reduce processing time, as utilized in Mask3D [33] and SPFormer [36]. Therefore, we will refer to both superpoints and individual points as regions going forward.

Our devised GP is illustrated in Fig. 3. Given two overlapping bounding boxes, the training data for GP is n_1 determined regions $\mathbf{X} \in \mathbb{R}^{n_1 \times 6}$ with their noise-free labels $\mathbf{f} \in \{0,1\}^{n_1}$, or $p(\mathbf{f}) = \mathcal{N}(\mathbf{f}, \mathbf{0})$. The GP seeks to produce the outputs of n_2 testing undetermined regions $\mathbf{X}_* \in \mathbb{R}^{n_2 \times 6}$ including the underlying Gaussian distributions $p(\mathbf{f}_*) = \mathcal{N}(\mathbb{E}[\mathbf{f}_*], \text{var}[\mathbf{f}_*])$ of labels $\mathbf{f}_*$, and the pseudo labels $\boldsymbol{\pi}_*$ inferred from the distribution.

In particular, we denote the output as the concatenation of the training labels $\mathbf{f}$ and the unknown $\mathbf{f}_*$, which follows the joint multivariate Gaussian distribution:

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right), \tag{1}$$

where $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X}) \in \mathbb{R}_+^{n_1 \times n_1}, \mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*) \in \mathbb{R}_+^{n_1 \times n_2}, \mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*) \in \mathbb{R}_+^{n_2 \times n_2}$ are the covariance matrices that capture the relationship between determined regions, determined-undetermined regions, and undetermined regions, respectively. $\kappa(x, x') = s^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2\right)$ is the radial basis kernel where l and s control the length scale and output scale. We create separate a GP model for each pair of overlapping bounding boxes. The hyperparameters, i.e., length scale l and output scale s, are optimized by using the determined regions.

The pseudo labels $\boldsymbol{\pi}_*$ can be computed as posterior:

$$\begin{aligned} \boldsymbol{\pi}_* = p(\mathbf{f}_* = 1 \mid \mathbf{X}_*, \mathbf{X}, \mathbf{f}) &\approx \int \sigma(\mathbf{f}_*)\, p(\mathbf{f}_*)\, d\mathbf{f}_*, \\ &\approx \sigma\left(\frac{\mathbb{E}[\mathbf{f}_*]}{\sqrt{1 + \frac{\pi}{8}\text{var}[\mathbf{f}_*]}}\right), \end{aligned} \tag{2}$$

where the last approximation is the probit approximation, and σ is sigmoid activation.

For each object, the final binary mask $\mathbf{m} \in \{0,1\}^{1 \times N}$ is obtained by attaching the regions $\mathbf{X}_*$ whose $\boldsymbol{\pi}_* \geq 0.5$ to the foreground regions of the object. Also, the mean map $\mathbf{e} \in [0,1]^{1 \times N}$ is constructed by setting the mean of the determined regions to their labels and the mean of the undetermined regions to $\mathbb{E}[\mathbf{f}_*]$. Finally, the variance map $\mathbf{v} \in \mathbb{R}_+^{1 \times N}$ is constructed by setting the variance of the determined regions to 0 and the variance of the undetermined regions to var$[\mathbf{f}_*]$.

3.2. Training a 3DIS Network with Pseudo Labels

After getting the pseudo masks $\mathbf{M} \in \{0,1\}^{K \times N}$ from GP, we are ready to train any 3DIS network Φ. However, to

leverage the informative cues from the mean $\mathbf{E} \in [0,1]^{K\times N}$ and variance $\mathbf{V} \in \mathbb{R}_{+}^{K\times N}$ maps also inferred from GP, rather than predicting only instance masks $\widehat{\mathbf{M}}$, we can simply modify the last layer of the network to predict two additional outputs: the mean $\widehat{\mathbf{E}}$ and the variance $\widehat{\mathbf{V}}$ representing the predicted Gaussian distribution.

For training the mask prediction $\widehat{\mathbf{M}}$, we use two loss functions: dice loss [35] and BCE loss following prior 3DIS work. For training the mean $\widehat{\mathbf{E}}$ and variance $\widehat{\mathbf{V}}$ predictions, we devise a new loss function based on KL divergence for each location i as follows:

$$L_{\text{KL}}(i) = \begin{cases} \log \frac{\hat{\mathbf{v}}_i}{\mathbf{v}_i} + \frac{\mathbf{v}_i^2 + (\mathbf{e}_i - \hat{\mathbf{e}}_i)^2}{2\hat{\mathbf{v}}_i^2} - \frac{1}{2}, & \text{if } \mathbf{v}_i > 0 \\ (\mathbf{e}_i - \hat{\mathbf{e}}_i)^2 + \hat{\mathbf{v}}_i^2, & \text{if } \mathbf{v}_i = 0, \end{cases} \tag{3}$$

where $\mathbf{e}_i, \mathbf{v}_i$ are the mean and variance at location i. When the variance is positive, we want to match two Gaussian distributions using KL divergence. Otherwise, they are Dirac Delta functions, so the predicted mean is matched with the pseudo mean and the predicted variance is matched with the pseudo variance using the MSE loss. As will be shown in the experiments, using the L_{KL} helps boost performance compared to only using mask loss.

Self-training: The feature for each point/superpoint can either be the input features (RGB color and position) or the pointwise deep feature extracted from a pretrained 3DIS network. Thus, after training the 3DIS network with the pseudo labels, we can utilize its pointwise deep features as $\mathbf{X}$ and $\mathbf{X}_*$, and then rerun the GP to obtain better pseudo labels. This strategy is referred to as *self-training*.

4. Experiments

Datasets. We conduct experiments on two datasets: ScanNetV2 [6] and S3DIS [1]. *ScanNetV2* consists of 1201, 312, and 100 scans with 18 object classes for training, validation, and testing, respectively. We report the evaluation results on the validation and test sets of ScanNetV2. The *S3DIS* dataset contains 271 scenes from 6 areas with 13 categories. We use Area 1, 2, 3, 4, 6 for training and Area 5 for evaluation.

Evaluation metrics. The average precision (AP) metrics commonly used in object detection and instance segmentation are adopted, including AP_{50} and AP_{25} are the scores with IoU thresholds of 50% and 25%, AP is the averaged score with IoU thresholds from 50% to 95% with a step size of 5%, and Box AP means the AP of the 3D axis-aligned bounding box prediction. Additionally, the S3DIS is also evaluated using mean coverage (mCov), mean weighed coverage (mWCov), mean precision (mPrec_{50}), and mean recall (mRec_{50}) with IoU threshold of 50%.

Implementation details. We implement our devised Gaussian Process by using GPytorch [12] to estimate l, s and compute $\boldsymbol{\pi}_*, \mathbb{E}[\mathbf{f}_*], \text{var}[\mathbf{f}_*]$ efficiently. We leverage the Adam optimizer with a learning rate of 0.1. For reference, it takes approximately 5 hours to generate pseudo labels for the entire ScanNetV2 training set (1201 scenes) on a single V100. We leverage our pseudo labels to train 5 different 3DIS methods, including PointGroup [19], SSTNet [26], SoftGroup [40], ISBNet [30], and SPFormer [36] based on their publicly released implementations. For methods that do not provide the code on S3DIS, we reproduce them based on the implementation details in their papers. All the models are trained from scratch and the hyper-parameters and the training details are kept the same as the original methods.

4.1. Comparison to Prior Work

Our direct comparison includes Box2Mask [25] and WISGP [9]. Their details are specified in Sec. 2.

Quantitative results. For ScanNetV2, we present the instance segmentation results for both the validation set and hidden test set in Tab. 1. It is obviously seen that our GaPro's versions of 3DIS methods outperform other box-supervised 3DIS methods by a significant margin on both sets, even with a smaller backbone (SPConv compared to Minkowski). Notably, our results are consistently comparable to SOTA fully supervised methods in AP, achieving about 90%. These findings demonstrate the effectiveness of our approach and the potential of our pseudo labels for improving standard 3DIS methods. For S3DIS, Tab. 2 presents the results on Area 5 of the S3DIS dataset. Our proposed GaPro achieves superior performance compared to Box2Mask, with large margins in both AP and AP_{50} when applied to SoftGroup and ISBNet. Additionally, when applied to PointGroup and SSTNet, our approach outperforms the WISGP's versions by a significant margin, demonstrating the robustness and effectiveness of our proposed pseudo labels.

Qualitative results. We visualize the qualitative results of pseudo labels of Box2Mask [5] and our method on ScanNetV2 training set in Fig. 4. Our approach generates more precise pseudo instance masks than Box2Mask. Additionally, our method performs well even in challenging scenarios where objects are densely packed or share edges (2nd and 3rd row respectively), our method is able to accurately label points in overlapped regions.

4.2. Ablation Study

We conduct ablation studies to justify the design choices of our proposed method. All these ablation experiments are conducted on ISBNet [30] on the validation set of the ScanNetV2 dataset unless otherwise stated.

Handling undetermined regions. We first explore different techniques for handling undetermined regions (i.e., regions belonging to multiple boxes) in our proposed method. Tab. 3 summarizes the results of our experiments. In setting A,

Method	Sup.	Backbone	Test set				Val set			
			AP	% full	AP_{50}	AP_{25}	AP	% full	AP_{50}	AP_{25}
Mask3D [33]	Mask	Minkowski	56.6	-	78.0	87.0	55.2	-	73.7	83.5
PointGroup [19]		SPConv	40.7	-	63.6	77.8	34.8	-	51.7	71.3
SSTNet [26]		SPConv	50.6	-	69.8	78.9	49.4	-	64.3	74.0
SoftGroup [40]		SPConv	50.4	-	76.1	86.5	46.0	-	67.6	78.9
ISBNet [30]		SPConv	55.9	-	76.3	84.5	54.5	-	73.1	82.5
SPFormer [36]		SPConv	54.9	-	77.0	85.1	56.3	-	73.9	82.9
CSC [16]	Point	Minkowski	29.3	51.8%	59.2	70.2	15.9	28.8%	28.9	49.6
PointContrast [44]		Minkowski	27.8	49.1%	47.1	64.5	27.8	50.4%	47.1	64.5
Box2Mask [5] (stand-alone)	Box	Minkowski	43.3	-	67.7	80.3	39.1	-	59.7	71.8
WISGP [9] + PointGroup [19]		SPConv	-	-	-	-	31.3	89.9%	50.2	64.9
WISGP [9] + SSTNet [26]		SPConv	-	-	-	-	35.2	71.2%	56.9	70.2
GaPro + PointGroup [19]	Box	SPConv	39.4	96.8%	62.3	74.5	33.4	96.0%	53.7	69.8
GaPro + SSTNet [26]		SPConv	45.8	90.5%	65.2	75.0	43.9	88.9%	60.1	70.8
GaPro + SoftGroup [40]		SPConv	42.1	83.5%	62.9	79.4	41.3	89.8%	62.7	77.3
GaPro + ISBNet [30]		SPConv	49.3	88.2%	69.8	81.0	50.6	92.8%	69.1	79.3
GaPro + SPFormer [36]		SPConv	48.2	87.7%	69.2	82.4	51.1	90.8%	70.4	79.9

Table 1: **3D instance segmentation results on ScanNetV2 hidden test set and validation set in AP metrics.** For reference purposes, we show the results of methods that use other types of supervision, such as Mask or Point in gray. The main metric for comparison is **AP**. The column **% full** indicates the percentage of the current method's performance compared to its corresponding fully supervised counterpart in the **AP** column. For the backbone, Minkowski is much heavier than SPConv. For Point supervision, we used 200 points per scene (or 10-20 points per instance).

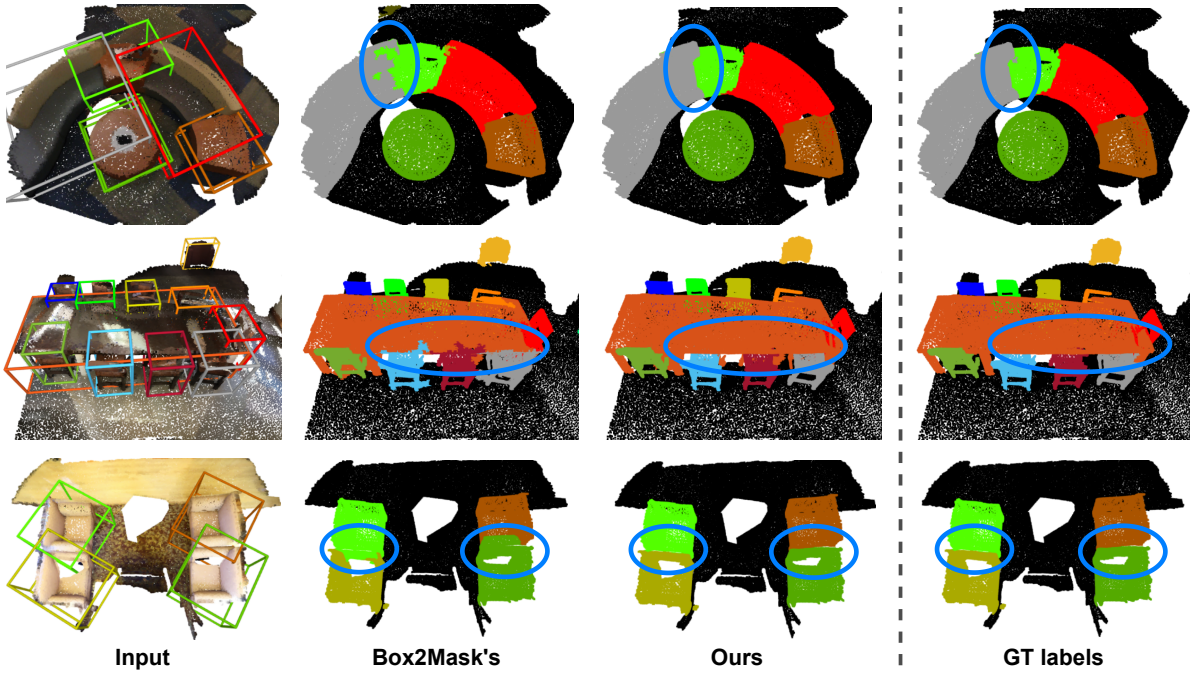

Figure 4: **Representative examples on ScanNetV2 training set**. Each row shows an example with the input and axis-aligned bounding box labels, Box2Mask [5]'s pseudo labels, our pseudo labels, and GT labels, respectively. Our approach produces highly accurate instance masks, particularly in regions with overlapping GT bounding boxes (blue circles).

we evaluate the approach of ignoring undetermined regions during training and only using the determined regions as pseudo labels. Next, inspired by the heuristics proposed by [5], we assign undetermined points to the smaller box. This

Method	Sup.	AP	AP_{50}	mPrec	mRec
Mask3D [33]	Mask	56.6	68.4	68.7	66.3
PointGroup [19]		-	57.8	61.9	62.1
SSTNet [26]		42.7	59.3	65.6	64.2
SoftGroup [40]		51.6	66.1	73.6	66.6
ISBNet [30]		54.0	65.8	74.2	72.7
Box2Mask	Box	-	-	66.7	65.5
Box2Mask*		43.6	54.6	64.4	67.4
WISGP + PointGroup		33.5	48.6	50.0	52.8
WISGP + SSTNet		37.2	51.0	44.3	56.7
GaPro + PointGroup	Box	42.5	56.8	59.3	61.3
GaPro + SSTNet		44.7	57.4	54.3	62.7
GaPro + SoftGroup		47.0	62.1	64.8	67.0
GaPro + ISBNet		50.5	61.2	66.7	72.4

Table 2: **3DIS results on S3DIS on Area 5**. The methods that use mask supervision are displayed in gray and are solely for reference purposes. The primary metric for comparison is the **AP**. A * symbol indicates that we reproduced Box2Mask on the S3DIS dataset based on their public code. For the backbone of each method, please refer to Tab. 1.

approach, setting B, results in a +3.7 improvement in AP compared to ignoring undetermined regions. By replacing the previous heuristic rule with a simple linear classifier, setting C, we achieve 44.2 in AP. In setting D1, we apply GP classification at the point level rather than the superpoint level. This approach significantly outperforms the heuristics-based approach from row 2 by a margin of +4 in AP. Finally, in settings D2 and D3, we explore two variations of GP applied at the superpoint level. The D2 approach performs GP regression which predicts the mask value as a continuous value between 0 and 1, while the D3 approach performs GP classification directly on the superpoints. The latter achieves the highest results, with a +1 improvement in AP over the regression-based approach.

Furthermore, we evaluate the quality of pseudo masks by comparing them to GT labels in the *training* set of ScanNetV2 using AP and AP_{90} metrics. Tab. 4 shows that our GP-generated pseudo labels outperform setting A, B, and C. In E, we replace the labels of D3 predicted with high uncertainty by the GT labels so as to quantify the usefulness GP's uncertainty. This replacement leads to a notable improvement, 88.0 in AP while applying the same strategy for points with low uncertainty results in a lower AP of 86.3.

Impact analysis of each component is summarized in Tab. 5. In rows 1 and 2, we compare the performance with and without our GP-based pseudo labels. The results show a significant improvement in AP of up to +10 when our pseudo labels are used. In row 3, we add a KL divergence loss during training with no additional cost to encourage the distribution of predicted masks to match the distribution of pseudo labels. This brings a further improvement of +0.3 in AP. In row 4, we incorporate self-training to refine the quality of our pseudo labels, resulting in a higher quality of training data and a performance boost of +0.8 in AP. Finally, in row 5, we combine all the components to produce our proposed approach, which achieves the best performance.

	Handling of undetermined points	AP	AP_{50}
	A: No pseudo labels in overlapped regions	38.1	59.1
	B: Box2Mask: assign points to smaller boxes	41.8	64.8
	C: Linear Classifier with points	44.2	64.5
GaPro	D1: GP Classification with points	45.7	67.2
GaPro	D2: GP Regression with superpoints	47.8	67.7
GaPro	D3: **GP Classification with superpoints**	**48.9**	**68.4**

Table 3: Handling the undetermined regions to produce pseudo labels.

Handling of undetermined points	AP	AP_{90}
A: No pseudo labels in overlapped regions	53.6	22.5
B: Box2Mask: assign points to smaller box	64.4	27.6
C: Linear classifier with points	69.4	34.1
D3: **GaPro (ours)**	**85.9**	**63.1**
E: Ours w/ uncertainty-guided GT replacement	88.0	67.2

Table 4: Quality of pseudo labels. We compute APs on the GT labels in the training set of ScanNetV2.

Our pseudo labels	KL loss	Self-train.	AP	AP_{50}	AP_{25}
			38.1	59.1	72.7
✓			48.9	68.4	79.0
✓	✓		49.2	68.1	78.5
✓		✓	50.0	68.3	79.0
✓	✓	✓	**50.6**	**69.1**	**79.3**

Table 5: Impact of our GaPro's components. **Our Pseudo Labels**: the proposed pseudo labels in Sec. 3.1, **KL Loss**: KL divergence loss, **Self-train.**: Self-training.

GP parameters	Superpoint	AP	AP_{50}
Fixed		46.3	66.3
Fixed	✓	48.0	67.2
Learnable		48.5	67.7
Learnable	✓	**50.6**	**69.1**

Table 6: Different configurations of GP. For fixed parameters, we set the length $l = 0.5$ and output scales $s = 1$.

Study on the configuration of GP is represented in Tab. 6. We found that allowing the GP parameters, i.e., length scale l and output scale s, to be learned resulted in a performance gain of more than 2 in AP. Furthermore, running GP on the superpoint level led to an additional improvement of 2 in AP compared to the version with point level.

Study on the features of GP is shown in Tab. 7. The first

Feature type	AP	AP_{50}
Pos.	48.5	67.9
Pos. + Norm.	49.0	68.1
Deep	**50.6**	**69.1**

Table 7: Impact of different features to GP.

Loss type	AP	AP_{50}
None	50.0	68.3
MSE	49.9	68.5
KL Loss	**50.6**	**69.1**

Table 8: Different losses to use with uncertainty values.

Method	Venue	Box AP_{50}	Box AP_{25}
VoteNet [31]	ICCV 19	33.5	58.6
3DETR [29]	ICCV 21	47.0	65.0
GroupFree [28]	ICCV 21	52.8	69.1
RGBNet [41]	CVPR 22	55.2	70.6
HyperDet3D [49]	CVPR 22	57.2	70.9
FCAF3D [32]	ECCV 22	57.3	71.5
GaPro + PointGroup	-	52.6	66.0
GaPro + SSTNet	-	57.8	67.8
GaPro + SoftGroup	-	60.2	73.4
GaPro + SPFormer	-	65.9	**78.9**
GaPro + ISBNet	-	**67.0**	77.1

Table 9: 3D object detection results on ScanNetV2 val set.

two rows present the results when we use only the position and normal of the point cloud as input to GP. When using *deep* features obtained from a 3DIS network pretrained on our pseudo labels, the performance improved by +1.6 in AP.

Study on different losses to use with uncertainty values is reported in Tab. 8. In row 2, simply using MSE loss for all points brings no difference to the overall performance. Our KL divergence loss helps improve the AP by 0.6 in row 3.

3D Object Detection Results. Our approach infers axis-aligned 3D bounding boxes, i.e., by taking the min and max coordinates of each dimension of the predicted instance masks, and we compare our results with other 3D object detection methods in Tab. 9. Notably, our findings demonstrate that when trained with the same level of annotations, the GaPro versions of 3DIS methods can outperform SOTA 3D object detection methods by a significant margin, achieving a Box AP_{50} increase of +8.6.

5. Discussion

Limitations: Although our approach assumes accurately annotated bounding boxes for all considered objects to generate pseudo labels, this assumption is no longer valid when the bounding boxes are noisy or incomplete. To simulate such scenarios, we conducted two experiments: (1) adding Gaussian noise to the coordinates of two defining corners of GT boxes to create noisy bounding boxes, and (2) randomly dropping accurate GT boxes to create incomplete GT bounding boxes (Tabs. 10a and 10b, respectively). As shown in our experiments, the quality of the bounding boxes

Cor. noise	AP	AP_{50}
2cm	48.3	67.4
5cm	45.0	65.7
10cm	43.0	64.2
10% dim	34.3	58.6
20% dim	21.0	43.5

(a) GT boxes with corner noises.

Drop rate	AP	AP_{50}
5%	49.6	68.2
10%	49.1	68.1
20%	48.2	66.7
50%	41.6	61.2
80%	30.6	48.6

(b) Dropping GT boxes.

Table 10: Results drop with noisy and incomplete boxes.

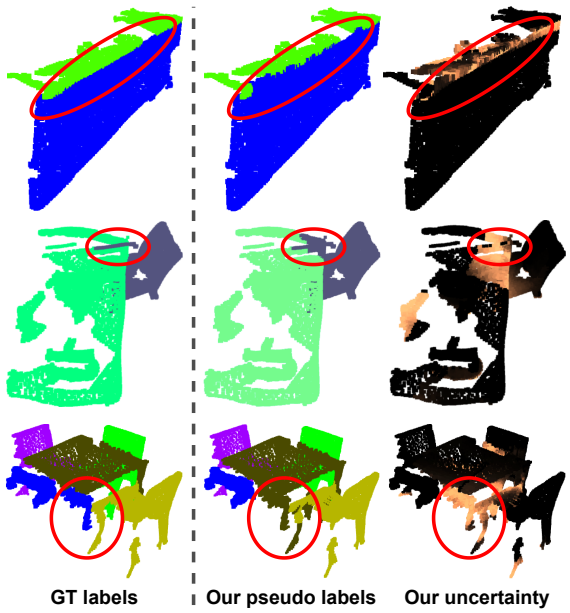

Figure 5: Examples of our imperfect GP pseudo labels with their informative uncertainty values for annotators to correct.

can significantly affect the accuracy of our pseudo labels. Moreover, even with accurate GT boxes, our pseudo labels may not be perfect in cases where there are overlapping boxes between adjacent objects or connecting objects with ambiguous shapes, as exemplified in Fig. 5. In such cases, our uncertainty values can provide useful indications for annotators to correct the pseudo labels.

Conclusion: In this work, we have introduced GaPro, a novel approach for instance segmentation on 3D point clouds using axis-aligned 3D bounding box supervision. Our approach generates high-quality pseudo instance masks along with associated uncertainty values, leading to superior performance compared to previous weakly supervised methods and competitive performance with SOTA fully supervised methods, achieving an accuracy of approximately 90%. Additionally, our method's robustness has allowed for the easy adaptation of various fully supervised to weakly supervised versions using our pseudo labels, showing its potential for applications where obtaining fine-grain labels is costly.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 4, 5

[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*, pages 15467–15476, 2021. 2

[3] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022. 2

[4] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, and Wenyu Liu. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. *arXiv preprint arXiv:2210.05174*, 2022. 2

[5] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. 1, 2, 3, 5, 6

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2, 4, 5

[7] Shichao Dong, Guosheng Lin, and Tzu-Yi Hung. Learning regional purity for instance segmentation on 3d point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 56–72. Springer, 2022. 2

[8] B. Douillard, J. Underwood, N. Kuntz, V. Vlaskine, A. Quadros, P. Morton, and A. Frenkel. On the segmentation of 3d lidar point clouds. In *2011 IEEE International Conference on Robotics and Automation*, pages 2798–2805, 2011. 3

[9] Heming Du, Xin Yu, Farookh Hussain, Mohammad Ali Armin, Lars Petersson, and Weihao Li. Weakly-supervised point cloud instance segmentation with geometric priors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4271–4280, 2023. 2, 3, 5, 6

[10] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5957–5966, 2017. 2

[11] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 319–336. Springer, 2022. 2

[12] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018. 5

[13] Tong He, Chunhua Shen, and Anton van den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 354–363, 2021. 2

[14] Tong He, Wei Yin, Chunhua Shen, and Anton van den Hengel. Pointinst3d: Segmenting 3d instances by points. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 286–302. Springer, 2022. 2

[15] 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. 2

[16] 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. 2, 6

[17] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[18] Le Hui, Linghua Tang, Yaqi Shen, Jin Xie, and Jian Yang. Learning superpoint graph cut for 3d instance segmentation. In *Advances in Neural Information Processing Systems*, 2022. 2

[19] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2, 5, 6, 7

[20] Shiyi Lan, Xitong Yang, Zhiding Yu, Zuxuan Wu, Jose M Alvarez, and Anima Anandkumar. Vision transformers are good mask auto-labelers. *arXiv preprint arXiv:2301.03992*, 2023. 2

[21] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3406–3416, 2021. 2

[22] Issam H Laradji, David Vazquez, and Mark Schmidt. Where are the masks: Instance segmentation with image-level supervision. *arXiv preprint arXiv:1907.01430*, 2019. 2

[23] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021. 2

[24] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 1–18. Springer, 2022. 2

[25] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Risheng Yu, Xiansheng Hua, and Lei Zhang. Box2mask: Box-

supervised instance segmentation via level-set evolution. *arXiv preprint arXiv:2212.01579*, 2022. 2, 5

[26] 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. 2, 5, 6, 7

[27] Jiaheng Liu, Tong He, Honghui Yang, Rui Su, Jiayi Tian, Junran Wu, Hongcheng Guo, Ke Xu, and Wanli Ouyang. 3d-queryis: A query-based framework for 3d instance segmentation. *arXiv preprint arXiv:2211.09375*, 2022. 2

[28] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 8

[29] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 8

[30] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. *arXiv preprint arXiv:2303.00246*, 2023. 2, 5, 6, 7

[31] 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/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 8

[32] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: fully convolutional anchor-free 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 477–493. Springer, 2022. 8

[33] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 4, 6, 7

[34] Myung-Ok Shin, Gyu-Min Oh, Seong-Woo Kim, and Seung-Woo Seo. Real-time and accurate segmentation of 3-d point clouds based on gaussian process regression. *IEEE Transactions on Intelligent Transportation Systems*, 18(12):3363–3377, 2017. 3

[35] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017. 5

[36] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 2, 4, 5, 6

[37] Chufeng Tang, Lingxi Xie, Gang Zhang, Xiaopeng Zhang, Qi Tian, and Xiaolin Hu. Active pointly-supervised instance segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 606–623. Springer, 2022. 2

[38] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[39] Shrihari Vasudevan, Fabio Ramos, Eric Nettleton, Hugh Durrant-Whyte, and Allan Blair. Gaussian process modeling of large scale terrain. In *2009 IEEE International Conference on Robotics and Automation*, pages 1047–1053, 2009. 3

[40] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022. 2, 5, 6, 7

[41] Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng Li, Bernt Schiele, and Liwei Wang. Rbgnet: Ray-based grouping for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1110–1119, 2022. 8

[42] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 2

[43] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 235–252. Springer, 2022. 2

[44] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 2, 6

[45] 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. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019. 2

[46] Siwei Yang, Longlong Jing, Junfei Xiao, Hang Zhao, Alan Yuille, and Yingwei Li. Asyinst: Asymmetric affinity with depthgrad and color for box-supervised instance segmentation. *arXiv preprint arXiv:2212.03517*, 2022. 2

[47] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 2

[48] Bingfeng Zhang, Jimin Xiao, Jianbo Jiao, Yunchao Wei, and Yao Zhao. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8082–8096, 2021. 2

[49] Yu Zheng, Yueqi Duan, Jiwen Lu, Jie Zhou, and Qi Tian. Hyperdet3d: Learning a scene-conditioned 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5585–5594, 2022. 8

[50] 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. 2