When 3D Bounding-Box Meets SAM: Point Cloud Instance Segmentation with Weak-and-Noisy Supervision

Qingtao Yu[2], Heming Du[2], Chen Liu[1], Xin Yu[1] *
[1] School of Information and Electrical Engineering, The University of Queensland
[2]College of Engineering Computting and Cybernetics, Australian National University
{Terry.Yu, u6341996}@anu.edu.au {xin.yu, uqcliu32}@uq.edu.au

Abstract

Learning from bounding-boxes annotations has shown great potential in weakly-supervised 3D point cloud instance segmentation. However, we observed that existing methods would suffer severe performance degradation with perturbed bounding box annotations. To tackle this issue, we propose a complementary image prompt-induced weakly-supervised point cloud instance segmentation (CIP-WPIS) method. CIP-WPIS leverages pretrained knowledge embedded in the 2D foundation model SAM and 3D geometric prior to achieve accurate point-wise instance labels from the bounding box annotations. Specifically, CIP-WPIS first selects image views in which 3D candidate points of an instance are fully visible. Then, we generate complementary background and foreground prompts from projections to obtain SAM 2D instance mask predictions. According to these, we assign the confidence values to points indicating the likelihood of points belonging to the instance. Furthermore, we utilize 3D geometric homogeneity provided by superpoints to decide the final instance label assignments. In this fashion, we achieve high-quality 3D point-wise instance labels. Extensive experiments on both Scannet-v2 and S3DIS benchmarks proves that our method not only achieves state-of-the-art performance for bounding-boxes supervised point cloud instance segmentation, but also exhibits robustness against noisy 3D bounding-box annotations.

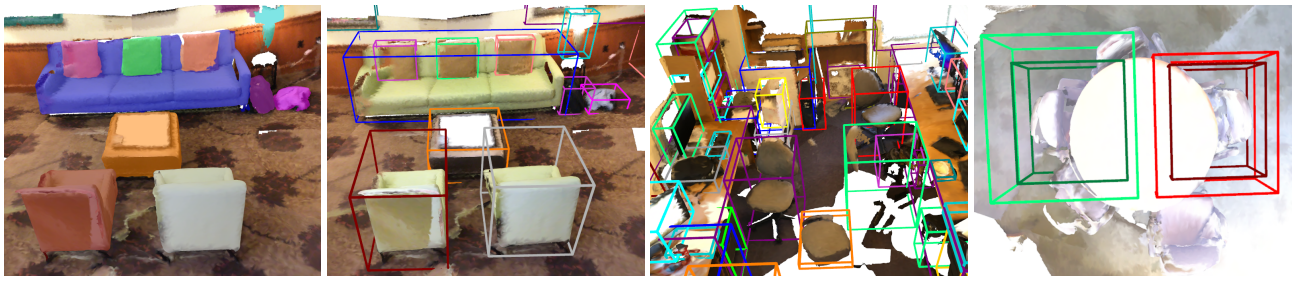
1. Introduction

Indoor point cloud instance segmentation is one of the fundamental tasks in 3D scene understanding [9, 14, 16, 19, 23, 27, 28, 30, 31]. The goal is to predict instance masks of 3D points and corresponding semantic labels. Current 3D indoor instance segmentation methods are mainly designed on fully-supervised annotations, *i.e.*, point-wise annotation. However, such annotation procedures are often time-consuming and laborious due to the vast quantity of points in each scene. Hence, there has been a growing interest in investigating weakly-supervised alternatives. Among different types of weak supervisions for point clouds, 3D instance bounding boxes stand out as a prospective direction. First, annotating bounding boxes is considerably efficient, as it only involves drawing a single box around each object. More importantly, each bounding box is a naturally richer instance representation, thus making it more capable of handling instance-level segmentation.

While several 3D bounding-boxes-based instance segmentation methods have been proposed [5, 8], they all utilize the minimum axis-aligned instance bounding-boxes as annotations. In other words, these bounding boxes are the tightest ones enclosing instance point clouds along the 3D world coordination. However, in practice, manual annotations inherently contain errors or noise. As a result, when bounding-box annotations have minor perturbations, these methods would experience significant degradation in performance. Therefore, it is crucial to consider the existence of noise and develop a counter-algorithm to mitigate its adverse effects. In this work, we proposed a complementary image prompt-induced weakly-supervised point cloud instance segmentation method under noisy bounding-box annotations. Our method merely requires each bounding box to cover the entire instance without any strict constraints on tightness. In other words, annotators can draw relatively looser bounding boxes as annotations, even if they are slightly larger than the objects themselves and may subsequently introduce a higher amount of noise.

As the first attempt to tackle this issue, we aim to leverage the recent advance of the 2D foundation model, *i.e.*, Segment Anything Model (SAM) [15]. SAM performs promptable instance segmentation on the 2D domain. In other words, SAM cuts the objects in the image, and which object gets cut out depends on the given prompt. Hence, we intend to generate 2D image prompts from 3D weak supervision signals and achieve accurate point-wise instance la-

*Corresponding author



(a) Point-wise annotation. (b) Bounding-box annotation. (c) Indoor complex point cloud scene. (d) Noisy bounding boxes.

Figure 1. (a) and (b) compare point-wise annotation and bounding-box annotation in the point cloud instance segmentation task. Bounding boxes annotation notably streamline the labeling process. (c) shows the complex indoor point clouds scene, exhibiting extensive instance overlapping and highly irregular point distribution. This suggests that manual annotated bounding boxes are unavoidable suffer from inaccuracy. (d) plots the examples of tight bounding-box annotations (□ and □) and the relaxed (*i.e.* noisy) bounding boxes (□ and □). Our work aims at alleviating the negative effects brought by noisy bounding-box annotations.

bels based on SAM predictions. In this case, any advanced fully supervised segmentation network can be used as the following procedure.

To transfer the powerful performance of SAM on 2D domains for 3D data, we treat the points that potentially belong to the instance as candidate points and project them into multiple image planes. After that, a greedy view selection algorithm is employed to choose the suitable views for projection. Based on the projected locations of candidate points, foreground 2D instance bounding boxes and sampled background pixel coordinates can be obtained as complimentary prompts. With these prompts, we use the produced SAM predictions to effectively assign confidence values to each projected point, indicating its likelihood of belonging to the instance. Finally, we apply a voting scheme to uniquely assign each point to instances according to the rank of confidence. Moreover, to mitigate the side effects of potential noisy projections and inaccurate SAM predictions, we exploit the 3D geometric structure of point clouds to facilitate our label refinement process. Our method does not require additional training or fine-tuning and can be adopted into any fully-supervised network.

With our CIP-WSIS, we can accurately assign point labels even in the presence of noisy bounding-box annotations. Extensive experiments on both widely-used ScanNet-v2 and S3DIS benchmarks demonstrate that our method is robust against various levels of annotation noise for 3D bounding boxes. In particular, our CIP-WSIS outperforms the state-of-the-art Box2Mask [5] by a large margin of 10% on AP on Scannet-v2 validation set with noise-free bounding-box annotations. There is only a 2% decrease in performance even with an increased noise rate. Overall, our contributions are three-fold:

- To the best of our knowledge, we are the first to explore the noisy weakly supervision problem on 3D instance segmentation tasks. To tackle the problem, we propose a complementary image prompt-induced weakly-supervised point cloud instance segmentation (CIP-WPIS) method that mitigates the model performance degradation problem caused by perturbed bounding box annotations.
- We introduce a 3D confidence ensemble module to mine the instance knowledge ensembles in the large 2D foundation model. With the 2D instance knowledge and the 3D geometric constraints, we can obtain accurate labels for each point.
- Our proposed is a flexible plug-in module that can be easily integrated into any fully supervised 3D instance segmentation methods, avoiding re-designing specific weak-supervised network structures.

2. Related Work

2.1. Fully-supervised 3D Instance Segmentation

Early methods can be grouped into two classes: proposal-based and grouping-based paradigms. Proposal-based methods employ a top-down strategy to generate region proposals and then segment the instance within each proposal. For instance, [36] and [11] first regress 3D bounding boxes for all instances and then leverage the point features to produce instance masks. Grouping-based methods implement a bottom-up pipeline that produces point-wise predictions and then cluster points into different instances. For example, MASC [17] leverages the mesh graph to cluster the instances after extracting the semantic features of points. To group instances more robustly, Jiang *et al.* [14] proposes to estimate point offsets with respect to object centers. Following such design, Liang *et al.* [16] and Chen [3] improve the performance by adopting hierarchical aggregation schemes. Furthermore, [9] introduces an additional graph convolutional network to refine the grouping outputs. Recently, SoftGroup [30, 31] leverages the advantages of both strategies. They proposed an architecture with bottom-up soft grouping and a subsequent top-down refinement.

State-of-the-art methods further improve 3D instance segmentation by utilizing advancements in transformers. The latest works, SPFormer [28] and Mask3D [27], both implement a transformer decoder following the design of cutting-edge 2D segmentation methods [2, 4]. Instead of clustering points, such methods learn a set of instance queries and compute instance masks directly based on the similarities between point features and query vectors. Such a strategy can better model the relationship between objects and points while accelerating the inference process at the same time. Moreover, an attention mask mechanism is incorporated to enhance training efficiency. However, all these fully-supervised methods require clean point-level annotations, and they would suffer severe performance drops when noisy annotations are provided.

2.2. Weakly-supervised 3D Instance Segmentation

Sparse-point weak supervision approaches only use a small portion of labeled point clouds to train the network. For example, Xu *et al*. [35] propose to propagate gradients of labeled points to those of labeled points in optimizing their semantic segmentation network. Hou *et al*. [12] annotate 0.1% of points and train a 3D semantic segmentation network by contrastive learning. Liu *et al*. [21] generate semantic pseudo labels from one point per object by implementing a contrastive learning strategy. In addition, Wu *et al*. [33] introduced a dual transformer model to effectively regularize unlabeled 3D points through an adversarial strategy at both the point level and region level. Note that most of these works focus on semantic segmentation instead of instance segmentation.

Compared to sparse point supervision, 3D bounding box annotations provide rough instance information such as object center and size, thus making it more capable of handling instance-level segmentation tasks. However, such a method has received little attention. Chibane *et al*. [5] adopts bounding-boxes as supervision. They estimate the instance bounding boxes for each superpoint and then follow a clustering technique to decide which group of superpoints belongs to the same instance. Du *et al*. [8] leverages 3D local geometric information to generate point-level labels from bounding-box annotations. However, all these methods highly rely on the accuracy of bounding boxes.

2.3. Foundation Models

The emergence of foundation models has received lots of attention, such as [7, 24–26]. These models are trained on vast amounts of data and hence demonstrate superior performance. Very recently, the Meta Research team has released the "Segment Anything Model" (SAM) [15]. It is trained on over 1 billion masks on 11 million images. With efficient prompting, it can create high-quality, generalized masks for image instance segmentation. Due to the excellent generalization performance of SAM, it has shown extensive use for other downstream tasks as an off-the-shelf tool [13, 18, 20, 22, 32, 37]. In our work, we aim to leverage SAM to provide 2D prior knowledge in our 3D instance segmentation.

3. Proposed Method

The overall framework of our proposed CIP-WPIS is illustrated in Figure 2. The description of our method is detailed in the following sections.

3.1. Candidate Points Initialization

We aim to initialize points inside the bounding boxes that potentially belong to the corresponding instance as candidates. Instead of considering all the contained points as candidate points, we invoke 3D superpoints to filter out some unlikely points for the efficiency of the following procedures. Superpoints are small clusters of points symbolizing local geometric continuity, formed via a normal-based graph cluster technique [10]. Following the assumption of previous works [5, 16, 28], all the points within a superpoint belong to the same instance. Given that bounding boxes only include additional points, we identify points as candidates only if the associated superpoints are entirely within the box. In other words, if any point in the superpoints lies outside the box, we can confidently exclude the entire superpoint. Following this process, we can obtain the candidate points for each instance. Note that some points in the overlapping area of boxes might be candidate points for multiple instances, and some background points might be incorrectly identified. The false-positive candidates will be corrected through the following processes with 2D pretrain knowledge from the fundamental model SAM.

3.2. View Selection

Indoor 3D point clouds are reconstructed from a sequence of RGBD images. Hence, each point of the scene is presented in at least one of the images, and this motivates us to leverage the prior knowledge from 2D for 3D tasks. We intend to select 2D image views for instances to project all the corresponding 3D candidate points. Due to the extensive overlapping and obstructions, it is usually difficult to locate a single instance view so that all the candidate points can be fully observed. Therefore, we designed a greedy view selection algorithm to progressively select a subset of 2D image views for each instance. First, we initialized an empty view set and labeled all the candidate points as unprojected. Then the view with the maximum number of visible points is added to the view set. These points, once visible, are subsequently marked as observed. We repeat the above procedure until all the candidate points are observed. The specifics of this algorithm are presented in Algorithm 1. In the projection phase, we mapped the 3D point location

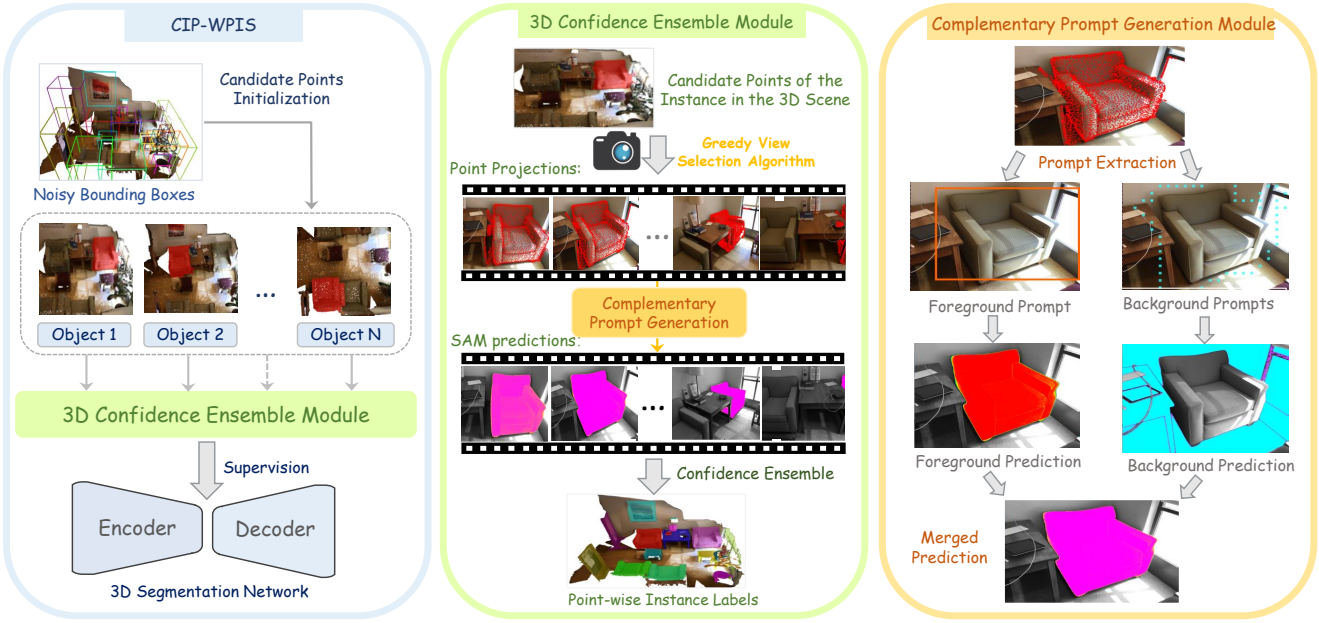


Figure 2. Workflow of our proposed CIP-WPIS. The left part depicts the whole pipeline for obtaining the point-wise instance labels from noisy bounding boxes. Specifically, we first assign the candidate points for each instance given the noisy bounding boxes. Then we devise the 3D confidence ensemble module to correct the mislabeled point of each instance as shown in the middle part plots. We first design a greedy selection algorithm to select multiple 2D views in which an instance is fully visible. Based on projected object points in each 2D view, we introduce a complementary prompt generation module to obtain the SAM predictions from various views. After that, we integrate these predictions to indicate whether the point belongs to the instance. The right part details the complementary prompt generation module. Complementary background and foreground prompts are introduced to obtain the object mask for each instance.

onto the 2D image plane utilizing the given camera view information. The projected 2D coordinates of each point are computed as below following the pinhole camera model:

$$\begin{bmatrix} u \\ v \\ z \end{bmatrix} = K \cdot P \cdot \begin{bmatrix} X \\ 1 \end{bmatrix} \qquad \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{z} \begin{bmatrix} u \\ v \end{bmatrix},$$

where K and P represent the camera intrinsic and extrinsic parameter matrix. X is the input point location vector under 3D world coordinates and x, y is the projected 2D pixel coordinates. Subsequently, we capture the points in the image view by clipping according to the dimensions of the image. We can determine the visibility of points based on the alignment between the depth of the RGBD image and the z-axis value from the camera coordinates.

3.3. Complementary Prompt Generation for SAM

Once comprehensive image views for all instances within the scene are selected, we derive 2D prompts based on the projected coordinates for these views. In this context, we adopt a complementary prompt strategy by computing a 2D bounding box of projected pixels as the foreground prompt and sampling pixels around the projected area as the background prompt. The foreground prompt can segment the target instance within the bounding box range, and

Algorithm 1 Greedy View Selection

Input: Number of instances $\mathbb{N}$
Camera views set $\mathbb{V}$
Candidate points sets $\mathbb{P}$ ($|\mathbb{P}| = \mathbb{N}$)
Output: Sets of selected instance views V
Procedure:
$V \leftarrow \{V_i | V_i = \emptyset, i \in \{1..\mathbb{N}\}\}$
$S \leftarrow \{S_j | S_j = project(\mathbb{P}, \mathbb{V}_j), \mathbb{V}_j \in \mathbb{V}\}$
for $i = 0$ **to** $\mathbb{N}$ **do**
  $P \leftarrow \mathbb{P}_i$
  while $P \neq \emptyset$:
    $j = \arg\max\{|S_j| | S_j \in S\}$
    $V_i \leftarrow V_i \cup \{\mathbb{V}_j\}$
    $P \leftarrow P \setminus S_j$
return V

the background prompt aids in identifying irrelevant parts of the instance in the image plane. Such dual types of prompts collectively facilitate each other to achieve the optimal 2D instance segmentation results using SAM.

The exclusive reliance on a single type of prompt would decrease the precision of SAM predictions, primarily due to two key factors. First, the major issue with foreground

prompts is their lack of accuracy. Since our input candidate points are highly noisy, the bounding box formed from the projected points generally exceeds the optimal size. When there are multiple instances in the box area, SAM may over-include additional parts other than the desired instance. At some viewing angles, when the image only includes a small fraction of the true positives but a larger number of false positives, such a problem can be enlarged. On the other hand, background prompts present their own set of challenges. Compared to bounding boxes, point prompts provide weaker constraints. Given a point as a prompt, SAM will return the most likely corresponding instance based on its pretrained knowledge. Since SAM is a generalized model trained with millions of diverse 2D object masks, it may have a different semantic understanding from the specific dataset. For instance, when there is a sink within the cabinet and the sink is the target instance, SAM sometimes interprets the cabinet and sink as a singular instance if the prompt is on the cabinet. Without a foreground prompt, the whole sink will be treated as part of the cabinet, which adversely impacts the prediction results for both classes.

With these two types of prompts, we can leverage SAM to compute instance masks on the 2D image plane. Note that each mask generated by SAM is a float-type score matrix with the size of the input image. Then, we merge the SAM mask predictions as the following formulation and output a single mask in the view m of instance k:

$$H_k^m = H_k^{m,f} - \beta \max\{H_k^{m,b_1}, ..., H_k^{m,b_n}\},$$

where $H_k^{m,f}$ and H_k^{m,b_i} represent the mask predictions of view from the forground bounding box prompt (f) and background point prompts (b). In the equation, $\max$ donates the element-wise maximum operation, and n is the number of background prompts. Since each H_k^{m,b_1} only represents one instance of an irrelevant image area, we implement such a strategy to merge the background predictions as a single background mask. In addition, β is a hyperparameter. According to our ablation study, $\beta = 0.5$ achieves the best performance. In this scenario, we derive the 2D instance scores for each pixel in the selected views. A higher score demonstrates a greater possibility that the pixel belongs to the prompt-induced instance.

3.4. 3D Confidence Ensemble from Selected Views

Based on the obtained 2D instance prior from SAM, we designed the following function to assign the confidence values for each candidate point:

$$C_{p,k} = \frac{\sum_m \Phi(p, V_{k,m}) \cdot H_k^m[i,j]}{\sum_m \Phi(p, V_{k,m})}.$$

The confidence result C has two inputs: the 3D point p and the instance ID k, which signifies the likelihood of the point having the corresponding instance label. It calculates the average of the 2D mask scores of the projected pixel in the views of the instance ($V_{k,m}$). In the equation, $[i, j]$ represents the 2D projected pixel coordinate of p. Implicit function Φ stands for the visibility of point p under instance view $V_{k,m}$, which outputs a value of 1 if the point is visible in the view and 0 if otherwise. Note that each point might appear a different number of times across the views. Hence, a normalization factor is added below, denoting the count of visible views.

After obtaining the confidence value of each point from 2D prior, we finalize the instance label for the candidate points guided by 3D geometric homogeneity provided by superpoints. We first compute the confidence of each superpoint as the mean confidence of its included points and then assign the instance labels to superpoints based on that. A higher confidence value indicates a stronger correspondence between the superpoint and the instance. With such a 3D cluster-level label correction approach, we can handle potential projection noise and 2D prediction errors.

We designed the following two steps to determine the instance labels of superpoints. Firstly, we set a confidence threshold at 0 to filter out the highly likely irrelevant superpoints from the candidate superpoints of the instance. Despite this, there are still massive superpoints that have positive confidence values associated with multiple instances. Therefore, we assign such superpoints to the instance with maximum confidence. This voting strategy aims to assign the points with the most correlated instance label. Following this design, additional included background points can be eliminated with negative confidence values and ambiguous points in the overlapping bounding box area can be uniquely and accurately allocated. In this fashion, we extensively leverage 2D prior knowledge from SAM and 3D geometric information to achieve correct point-wise instance labels, which can be seamlessly integrated with any fully supervised 3D instance segmentation network.

4. Experiments

To validate the effectiveness of our proposed CIP-WSIS, we conduct experiments on two challenging datasets, *i.e.*, ScanNet-V2 [6] and S3DIS [1].

4.1. Datasets and Evaluation Metrics

ScanNet-V2 [6] dataset contains 1, 613 scans with 3D semantic and instance annotations. The dataset is split into training, validation, and testing sets, with 1, 201, 312, and 100 scans. It contains 18 object semantic categories. We train on the training set and report results on the validation set for comparison with other methods. An efficient normal-based graph cut image segmentation method [10] is utilized for superpoint generation. Mean Average Precision (mAP) serves as the common evaluation metric for instance segmentation on the Scannet-V2 dataset. It calculates the av-

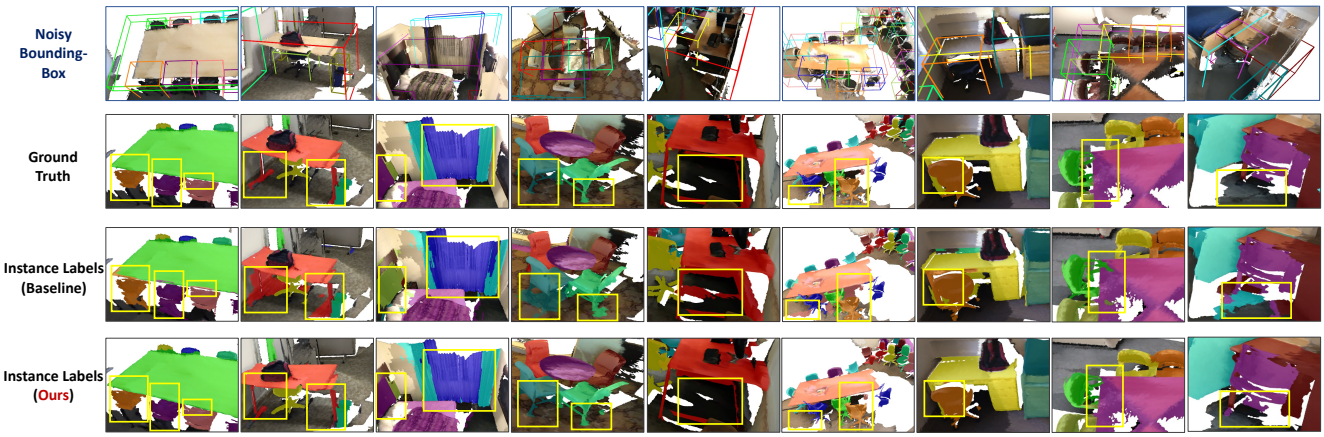


Figure 3. Visualizations of instance labels generated by the baseline model and our proposed method. To show the effectiveness of our method, we select the bounding boxes with the highest noisy rate ($\lambda = 0.3$) as the input. Regions for comparison are highlighted by □. Even the input bounding boxes containing large unrelated regions, the instance labels generated by our method have minor differences from the ground-truth labels.

Sup.	Method	AP	AP_{50}	AP_{25}
Mask	PointGroup [14]	0.348	0.517	0.713
	SSTNet [16]	0.494	0.643	0.740
	SoftGroup^{++} [30]	0.458	0.674	0.791
	ISBNet [23]	0.545	0.731	0.825
	Mask3D [27]	0.552	0.737	0.835
	SPFormer [28]	0.563	0.739	0.829
Point	PointContrast [34]	0.348	0.517	0.713
	CSC [12]	0.494	0.643	0.740
Box	Box2Mask	0.391	0.597	0.718
	WISGP [8]	0.352	0.569	0.702
	Ours&Softgroup^{++} [30]	0.395	0.629	0.773
	Ours&SPFormer [28]	**0.475**	**0.693**	**0.786**

Table 1. The results of our method under noisy-free bounding boxes on the Scannet-V2 validation set. For reference purposes, we also include the results of methods using other types of supervision (Sup.), such as masks or sparse points (200 points per scene). Our method can be used as a plugin to leverage the power of a fully supervised network, and it outperforms the existing weakly supervised design.

erage scores of all foreground classes among different Intersection over Union (IoU) thresholds, ranging from 50% to 95%, with increments of 5%. In addition, AP_{50} and AP_{25} represent the scores corresponding to IoU thresholds of 50% and 25%, respectively. We present the mAP, AP_{50}, and AP_{25} results on the ScanNetv2 validation dataset.

Stanford 3D Indoor Scene Dataset (S3DIS): S3DIS [1] is a large-scale indoor dataset displaying six distinctive areas from three separate campus buildings. It contains 272 scans and is annotated with instance masks over 13 semantic classes. Following the common splits, our method is trained in Area 1, 2, 3, 4, and 6 and evaluated in Area 5. To compare with the previous works [5, 8], we adopt the mean precision (mPrec) and mean recall (mRec) at overlap 0.5 in S3DIS evaluation.

4.2. Experiment Settings

Noisy Bounding-boxes: We input noisy bounding boxes to each instance for point-wise 3D label generation. Each axis-aligned bounding box can be represented by a 6-dimension vector, including the xyz coordinates of minimum corners and maximum corners. To simulate noisy annotations, we choose different hyper-parameter λ values to enlarge the minimum bounding boxes. Each noisy bounding box can be expressed as $[C_{min} - 0.5X, C_{max} + 0.5X]$, where C stands for bounding box corners and $X = \lambda(C_{max} - C_{min})$. To mimic human labeling in real-world settings, we add a minor Gaussian permutation with a standard deviation of $0.5\lambda X$. A larger λ will lead to a higher noise rate. To exhibit robustness, we select different λ from 0 to 0.3.

Network Selection: We selected three different seeds to generate noises and report their average performance. To evaluate the adaptive capacity of our method under different types of backbones, two representative fully supervised methods are selected (SPF and SoftGroup $^{++}$) for our experiment. Softgroup^{++} stands as a representation of the traditional methods [9, 14, 16, 31]. It involves a bottom-up grouping and a top-down refinement after feature extrac-

Method	Noise parameter λ											
	$\lambda = 0$			$\lambda = 0.1$			$\lambda = 0.2$			$\lambda = 0.3$		
	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}
Box2Mask [5]	0.398	0.592	0.712	0.366	0.563	0.690	0.359	0.539	0.670	0.301	0.515	0.669
WISGP * [8]	0.352	0.569	0.702	0.333	0.524	0.645	0.294	0.492	0.599	0.261	0.458	0.562
Base & Softgroup^{++} [30]	0.393	0.623	0.765	0.368	0.598	0.759	0.362	0.591	0.752	0.354	0.592	0.757
Base & SPF [28]	0.473	**0.689**	**0.787**	0.437	0.672	0.780	0.396	0.625	0.761	0.364	0.620	0.771
Ours & Softgroup^{++}	0.395	0.629	0.773	0.371	0.584	0.741	0.366	0.582	0.745	0.364	0.588	0.749
Ours & SPF	**0.475**	0.693	0.786	**0.465**	**0.691**	**0.777**	**0.452**	**0.672**	**0.768**	**0.446**	**0.668**	**0.761**

Table 2. The results of our method under noisy bounding boxes of different noisy rates on the ScanNet-V2 validation set. The main metric for comparison is AP. Symbol $*$ indicates the method is reproduced by ourselves to test on the noisy bounding-box setting.

Noise	Wrong Points	Wrong Superpoints
$\lambda = 0$	5.2m / 3.6m	64k / 43k
$\lambda = 0.1$	10m / 5.9m	114k / 67k
$\lambda = 0.2$	13m / 7.1m	141k / 81k
$\lambda = 0.3$	16 m / 10m	174k / 107k

Table 3. Quantitative evaluation of inaccurately assigned points and superpoints throughout the entire Scannet-V2 training set, compared between the baseline (left) and our technique (right). Due to the significant improvement in label accuracy, our method delivers a better final prediction.

Sup.	Method	mPrec	mRec
Mask	PointGroup [14]	55.3	42.4
	SSTNet [16]	65.5	64.2
	SoftGroup^{++} [30]	73.6	66.6
	Mask3D [27]	68.7	66.3
	SPFormer [28]	72.8	67.1
Box	Box2Mask	66.7	**65.5**
	WISGP&PointGroup [8]	50.0	52.8
	WISGP&SSTNet [8]	44.3	56.7
	Ours&SPFormer	**69.1**	64.2

Table 4. The results of our method under noisy-free bounding boxes on S3DIS folder-5.

tion. SPF [28], similar to Mask3D [27], implemented a different type of strategy using transformers [29] and introduces learnable queries as instance vectors. To further demonstrate the effectiveness of the method, we report the improved label accuracy in Table 3.

4.3. Implementation Details

Training Strategy: The fully-supervised technique usually imports a pre-train checkpoint on the same dataset with a less strong fully-supervised network backbone. To adapt to the weakly supervised scenario, we manually removed such pertaining. In this case, we make some minor adjustments to the previous training configuration, *i.e.*, increasing the number of training epochs and reducing the learning rates. To compare with the second type of baseline, we adopt the same configuration settings for training. Moreover, *ceiling* and *floor* classes are considered to have the same semantic label as *background* for supervision.

Prompt Generation: Our method employs complementary prompts to guide SAM predictions. The foreground prompt is the 2D bounding box of projected pixels of candidate points. Background prompts are the sampled pixels around the projected area. Specifically, we divide the image into multiple 32 $\times$ 32 windows. If one window contains a projected point, it will be marked as 'True' and vice versa. Thus, we can obtain a boolean matrix, and the number of columns and rows are image height and width divided by 32. With a simple kernel multiplication, we can achieve the windows that are close to the projected area while not including any projections. A smaller window size leads to a slightly better performance, but it will increase the number of prompts, which would cause higher computation time.

4.4. Main Result

The majority of the experiment is conducted on the ScanNet-V2. The quantitative results are shown in Table 1, 2, and qualitative results can be visualized in Figure 3. As the first attempt to tackle the problem, we tried our best to adapt previous works and create the following two baselines for comparison. The first one is the existing box-supervised approach with the same noisy bounding boxes. Another one is state-of-the-art fully-supervised methods with point-wise labels generated from each bounding box with only the can-

didate point initialization step. To ensure each point has a unique instance label, we implemented a simple heuristic following the design of previous works [5], which is choosing the instance with the smallest bounding box if the point is a candidate point of multiple instances. This is because smaller objects are often fully contained in bounding boxes of larger objects. The purpose of establishing this baseline is to ensure that the improved performance isn't just because of the change in network structure. As a result, we achieve state-of-the-art 3D instance segmentation performance under noise-free and noisy bounding box annotations. Especially for noisy bounding boxes, our method only has around 2% performance degradation as the noise rate increases to the next level.

We carry out extra experiments under the S3DIS dataset with SPFormer as the network structure. We achieved state-of-the-art performance in terms of mPrec under noisy-free bounding box supervision, as shown in Table 4. For noisy bounding boxes supervision, our method suffers nearly 5% performance drop for mPrec and 8% for mRec on average with a 0.1 increase of λ. In contrast, the baseline labels with the same network structure drop 9% and 13% on average, respectively. Therefore, we prove that our CIP-WPIS is robust against noise.

4.4.1 Ablation Study

Prompt Using Strategy: We examine another way of acquiring 2D mask scores, which is the most straightforward design. SAM allows users to feed an arbitrary number of prompts with multiple types to generate a single mask prediction. Hence, we input foreground and background prompts together to predict a single heatmap and use such heatmap to do the confidence assignment. The box foreground prompt is naturally a positive signal. Background prompts are manually set as negative signals for the predictor. However, we observed that when the number of prompts gets more, SAM tends to give unstable results and leads to a drop in performance.

Hyperparameter Selection: In addition, we evaluated different hyperparameters β for merging the SAM predictions of each prompt. A smaller β would be less effective in identifying the excessively included background points. And a bigger β over-crop the true positive candidate points. The two ablation study results are combined in Table 5.

β	Single	Merged	mAP	AP_{50}	AP_{25}
0.5	✓		0.411	0.628	0.754
0.2		✓	0.442	0.681	0.767
0.8		✓	0.409	0.616	0.732
0.5		✓	**0.465**	**0.691**	**0.777**

Table 5. Performance comparison of different ways of leveraging prompts and different selections of hyperparameter β under noise rate $\lambda = 0.1$ with SPFormer network structure under Scannet Dataset.

4.5. Further Discussion

Robustness: We observed that our method outperforms the baseline with a bigger gap as the noise level rises. One contributing factor is the strong ability of the 2D foundation model that gives robust predictions against noisy prompts. Another important factor is that the number of candidate points increases along with the noise level, leading to a rise in the number of selected views. Such a consequence can be beneficial for generating a more reliable confidence value due to our averaging process. Therefore, in the most ideal case, if every RGBD image is taken into consideration, the labeling accuracy can be further improved. However, the computational burden will surge dramatically since each 3D scene may overall contain thousands of high-resolution image frames. Therefore, the greedy view selection approach is introduced as a trade-off design between final performance and computation cost.

Limitations and Further Work: Even though our introduced method notably enhances label accuracy, it cannot match the precision of human annotation. Since this work focuses on demonstrating the effectiveness of our auto-labeling module, our proposed method doesn't involve additional innovation on the fully-supervised network structure under noisy masks. While our method provides a confidence value for each point associated with each instance, we anticipate further studies to improve noisy bounding box supervised segmentation from a soft labeling perspective.

5. Conclusion

In this work, we propose an annotation noise-aware weakly supervised point cloud instance segmentation method taking advantage of the image-domain information provided by the foundation model SAM and the geometric local consistency of point clouds. In particular, we generate prompts on the image plane based on given weak supervision, and leverage the foundation model to mine the image-domain instance mask predictions. We then rectify erroneously assigned 3D point labels according to the 3D geometric consistency. As a result, we achieved high-quality 3D point instance labels. Extensive experiments demonstrate that our method outperforms the state-of-the-art, especially in the presence of noisy bounding-box annotations.

Acknowledgements: This research is funded in part by ARC-Discovery grant (DP220200800 to XY) and ARC-DECRA grant (DE230100477 to XY). We thank all anonymous reviewers and ACs for their constructive suggestions.

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. 5, 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[3] 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. 2

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 3

[5] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *ECCV*, pages 681–699, 2022. 1, 2, 3, 6, 7, 8

[6] 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. 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] 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. 1, 3, 6, 7

[9] 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. 1, 2, 6

[10] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 3, 5

[11] 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

[12] 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. 3, 6

[13] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023. 3

[14] 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. 1, 2, 6, 7

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 3

[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. 1, 2, 3, 6, 7

[17] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*, 2019. 2

[18] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 3

[19] 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. 1

[20] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023. 3

[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. 3

[22] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 3

[23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 6

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 3

[27] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 1, 3, 6, 7

[28] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 1, 3, 6, 7

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[30] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, Junyeong Kim, and Chang D Yoo. Softgroup++: Scalable 3d instance segmentation with octree pyramid grouping. *arXiv preprint arXiv:2209.08263*, 2022. 1, 2, 6, 7

[31] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 1, 2, 6

[32] Yonghui Wang, Wengang Zhou, Yunyao Mao, and Houqiang Li. Detect any shadow: Segment anything for video shadow detection. *arXiv preprint arXiv:2305.16698*, 2023. 3

[33] Zhonghua Wu, Yicheng Wu, Guosheng Lin, Jianfei Cai, and Chen Qian. Dual adaptive transformations for weakly supervised point cloud segmentation. In *European Conference on Computer Vision*, pages 78–96. Springer, 2022. 3

[34] 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. 6

[35] 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. 3

[36] 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. 2

[37] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3