BSNet: Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation

Jiahao Lu[1], Jiacheng Deng[1], Tianzhu Zhang[1,2,*]
[1]University of Science and Technology of China, [2]Deep Space Exploration Lab
{lujiahao, dengjc}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

Abstract

3D instance segmentation (3DIS) is a crucial task, but point-level annotations are tedious in fully supervised settings. Thus, using bounding boxes (bboxes) as annotations has shown great potential. The current mainstream approach is a two-step process, involving the generation of pseudo-labels from box annotations and the training of a 3DIS network with the pseudo-labels. However, due to the presence of intersections among bboxes, not every point has a determined instance label, especially in overlapping areas. To generate higher quality pseudo-labels and achieve more precise weakly supervised 3DIS results, we propose the Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation (BSNet), which devises a novel pseudo-labeler called Simulation-assisted Transformer. The labeler consists of two main components. The first is Simulation-assisted Mean Teacher, which introduces Mean Teacher for the first time in this task and constructs simulated samples to assist the labeler in acquiring prior knowledge about overlapping areas. To better model local-global structure, we also propose Local-Global Aware Attention as the decoder for teacher and student labelers. Extensive experiments conducted on the ScanNetV2 and S3DIS datasets verify the superiority of our designs. Code is available at https://github.com/peoplelu/BSNet.

1. Introduction

3D instance segmentation is a fundamental task in 3D scene understanding, primarily focused on predicting masks and categories for every foreground object within a scene. Current instance segmentation methods are mainly in fully supervised settings [25, 30, 34, 36, 37] and achieve commendable results. However, the time-consuming nature of point-level annotations poses a significant challenge. In contrast, annotating instances with 3D bboxes (object-level) is notably easier, requiring only the annotations for center points and dimensions (length, width, height). Nevertheless, a notable limitation stems from the use of bboxes, which cannot capture the detailed shape or geometry of objects. Consequently, bridging the gap between object-level and point-level annotations remains a challenge.

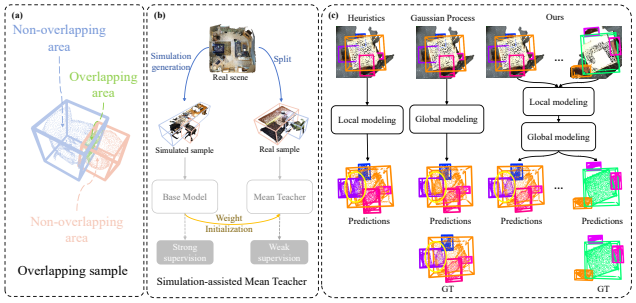

Figure 1. (a) The visualization of an overlapping sample. (b) The proposed Simulation-assisted Mean Teacher helps the labeler acquire prior knowledge from simulated samples. (c) Our method improves local-global structure modeling of overlapping samples to generate better pseudo-labels (especially in yellow circles).

To solve the above challenge, existing methods [8, 11, 35] conduct several explorations. Box2Mask [8] parameterizes bboxes and utilizes them as labels. However, due to bbox overlap, some point clouds may exist within multiple bboxes, introducing ambiguity in point-object assignments. As illustrated in Figure 1(a), instance labels for point clouds in non-overlapping areas are determined as they only belong to one bbox. In contrast, overlapping areas are governed by two different bboxes, resulting in indeterminate instance labels. Consequently, the point-wise predicted bboxes cannot be reliably used for clustering. To better address the ambiguity in overlapping areas, WISGP [11] employs straightforward heuristics based on local structure modeling. Concretely, for each indeterminate point, WISGP selects the most common label from its neighboring points as its label. On the other hand, Gapro [29] uses Gaussian Process (GP) [35] to train individual overlapping samples, modeling global structure by fitting the similarity relationships between all points into a Gaussian distribution. Then Gapro computes posterior probability to achieve binary classification for overlapping areas.

*Corresponding Author

Based on the preceding discussion, we identify two key issues in bbox-supervised 3DIS: 1) *How to generate labels for overlapping areas?* Using Mean Teacher [3, 10, 20, 48] to generate and continuously optimize pseudo-labels for overlapping areas is an effective way. This paradigm allows for the online update of pseudo-labels during the training process, continuously transferring knowledge from a teacher network to a student network. Besides, the teacher network employs Exponential Moving Average (EMA) to integrate information from historical students, providing more stable learning targets for the student network. However, given that non-overlapping areas contain a single object with a clear structure, while overlapping areas involve two intertwined objects, there is a significant disparity in complexity between them. Hence, it is difficult to infer accurate pseudo-labels for overlapping areas solely according to non-overlapping area labels. To tackle this issue, given the abundance of non-overlapping bboxes in the dataset with definite labels, we can construct simulated overlapping samples using these bboxes. As illustrated in Figure 1(b), we can train a base model on these simulated samples, and transfer this model to real datasets. By this way, the information loss resulting from the absence of labels in overlapping areas can be compensated and higher quality pseudo-labels can be predicted. 2) *How to better model structure of overlapping samples?* As illustrated in Figure 1(c), current methods [11, 29] either tend to focus on local structure modeling, bringing dedicated local structure representations like WISGP [11], or emphasize global relationship modeling, results in a more effective connection between overlapping areas and non-overlapping areas like Gapro [29]. Both types of modeling are crucial, but there is currently no approach effectively integrating these two aspects. Consequently, it is essential to devise a universal network proficient in extracting local structure features efficiently while fostering interactions between overlapping and non-overlapping areas, yielding more precise pseudo-labels.

To achieve the above goals, we introduce a novel pseudo-labeler called Simulation-assisted Transformer (SAFormer), which is trained based on an innovative training strategy called Simulation-assisted Mean Teacher (SMT) and incorporates a special decoder called Local-Global Aware Attention (LGA). In order to solve the **first problem** in the previous paragraph, we introduce the SMT. Concretely, student network is directly supervised with definite instance labels for non-overlapping areas, and for overlapping areas, pseudo-labels generated by teacher network are used as supervision. As to teacher network, we use EMA to updates its parameters. This approach yields more accurate predictions for overlapping areas compared to classical statistical methods like GP [35]. Furthermore, to address the challenge of suboptimal pseudo-label quality, we generate simulated overlapping samples using non-overlapping bboxes. And these simulated samples are used to train a base model, producing weights that serve as the initialization for teacher and student networks. This fundamentally equips the network with the ability to distinguish overlapping areas, and the higher quality pseudo-labels generated by the teacher network aid in the rapid training of the Mean Teacher. Additionally, when applied to multiple datasets, only a brief finetuning is required instead of retraining the pre-trained weights. Taking S3DIS [2] as an example, the model's training time decreases from 42 hours to just 1.7 hours. As to the **second problem**, we introduce the LGA. Concretely, we first initialize two learnable queries, with each representing one of the two foreground instances. We then employ local-structure attention to effectively model the local structure of each instance and aggregate structural relationships within each instance into a holistic representation through queries. Subsequently, by employing global-context attention, we facilitate the aggregation of global information, especially interactions between the two foreground instances and interactions between overlapping areas and non-overlapping areas. Through this design, we can effectively model category, structure, and contextual information adaptively. Additionally, we can leverage the response values of the overlapping area points to the queries to remove background points from the overlapping areas.

In summary, the main contributions of this work are as follows: (i) We propose a weakly supervised 3D instance segmentation method called BSNet, which uses bboxes as annotations and devises a novel pseudo-labeler. (ii) We design a pioneering pseudo-labeler called SAFormer, which for the first time incorporates the deep neural network and the Mean Teacher paradigm, and innovatively constructs simulated samples to facilitate training. Besides, with the help of LGA, SAFormer can accurately predict pseudo-labels for overlapping areas, thus achieving precise weakly supervised 3DIS results. (iii) Extensive experimental results on two standard benchmarks, ScanNetV2 [9] and S3DIS [2], verify the superiority of our designs.

2. Related Work

In this section, we briefly overview related works on 3D instance segmentation, weakly supervised 3D instance segmentation and the Mean Teacher paradigm.

3D Instance Segmentation. 3D instance segmentation is a fundamental task for 3D scene understanding, which can be categorized into proposal-based, grouping-based and transformer-based methods. Proposal-based methods [12, 23, 45, 46] extract 3D bboxes and utilize a mask learning branch to predict the object mask inside each box. Grouping-based methods [5, 19, 21, 30, 37, 42, 50] predict semantic categories and geometric offsets for each point, and then employ clustering algorithms to group the points

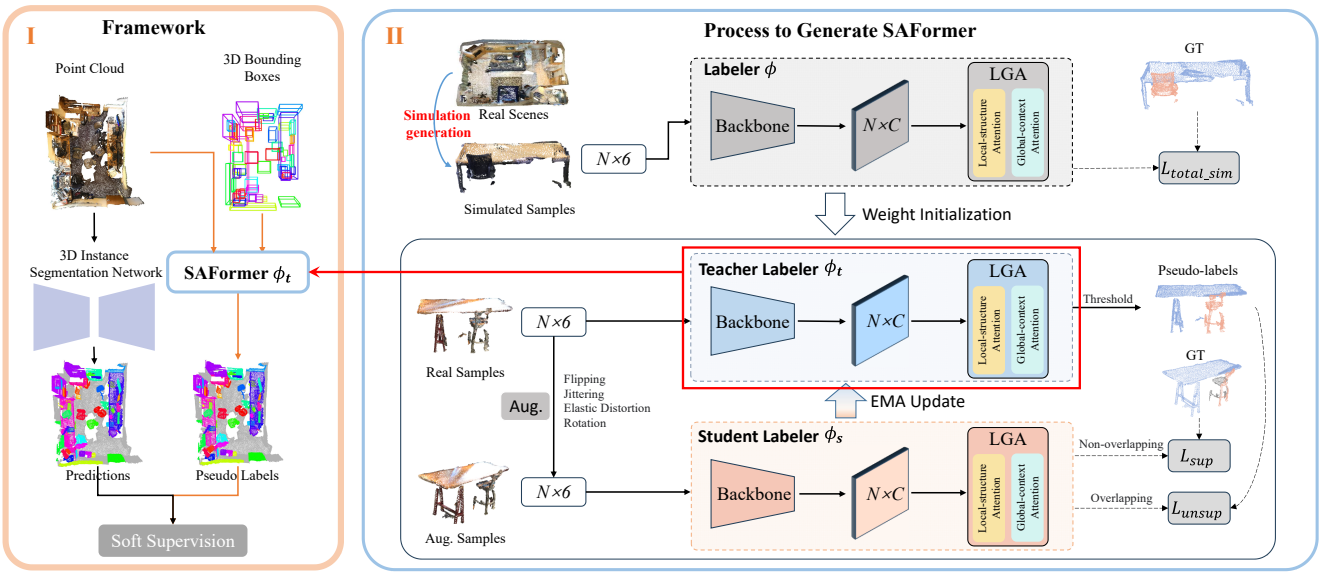

Figure 2. **(I) The overall framework of our method BSNet. (II) The total process to generate an outstanding pseudo-labeler SAFormer.** BSNet is a novel two-step method consisting of generating pseudo instance labels by SPFormer and using the pseudo instance labels to train a 3D instance segmentation network.

into instances. Transformer-based methods [25, 34, 36] are the state-of-the-art paradigm, where queries are used to represent instances and global information is aggregated into queries through a transformer decoder [4, 6, 7]. Although these fully supervised methods have achieved superior performance, they still have significant limitations in practice due to the time-consuming point-wise instance annotation. Therefore, some weakly supervised instance segmentation methods have been proposed to alleviate this problem.

Weakly Supervised 3D Instance Segmentation. The current weakly supervised 3D instance segmentation methods [8, 11, 17, 29, 43] can be divided into two categories: sparse points as annotations and 3D bboxes as annotations. Sparse-point annotation methods [17, 43] primarily utilize sparsely labeled point clouds as supervision to train the network. In comparison to methods using sparse points as annotations, 3D bboxes provide richer instance information such as category and shape size, enabling the network to better handle instance segmentation tasks. Box2mask [8] uses bboxes as supervision, allowing the network to predict bbox for each individual point. WISGP [11] leverages 3D local geometric information to generate point-level labels from bbox annotations. Gapro [29] employs GP [35] to model the global similarity relationships between overlapping and non-overlapping regions. However, these methods have relatively simple modeling of structure in overlapping samples and do not effectively incorporate category, structure, and contextual information. In contrast, our proposed Local-Global Aware Attention enhances the capacity to model both local structures and global relationships.

Mean Teacher Paradigm. The Mean Teacher paradigm has been widely researched in various tasks, such as UDA for semantic segmentation [1, 18, 47], semi-supervised object detection [26, 40, 44], weakly supervised object detection [39, 41], UDA for object detection [3, 20], and UDA for person ReID [13, 16, 48]. This paradigm helps to avoid the interative self-training complicated multi-stage training process. SoftTeacher [44] introduces the first end-to-end pseudo labeling framework in semi-supervised object detection, gradually improving the quality of pseudo labels during a curriculum. To mitigate the issue of low-quality pseudo-labels, CMT [3] identifies the alignment and synergy between Mean Teacher and contrastive learning. UNRN [48] proposes the estimation and exploitation of the credibility of assigned pseudo-labels for each sample, reducing the impact of noisy pseudo-labels generated by the teacher network. Based on the above research, we introduce a Simulation-assisted Mean Teacher approach, which employs the Mean Teacher paradigm to generate stable pseudo-labels in real-time and constructs simulated samples to assist the network in acquiring prior knowledge about overlapping areas.

3. Method

3.1. Overview

As illustrated in Figure 2(I), the framework of our method begins by generating pseudo object masks for instances in the training set based on bbox annotations. Subsequently, these pseudo object masks are employed to train a 3DIS network. Throughout the entire process, the most critical step is to generate an outstanding pseudo-labeler to predict pseudo-labels for overlapping areas, as shown in

Figure 2(II). In the generation process, two distinct designs stand out. The first one is the adoption of a unique training strategy called Simulation-assisted Mean Teacher (SMT), which can be divided into two steps: Simulated Sample Generation in Section 3.2.1 and Mean Teacher Approach in Section 3.2.3. The second one is a novel decoder named Local-Global Aware Attention (LGA) in Section 3.2.2. First, we generate simulated samples using non-overlapping bboxes from real datasets. These simulated samples are then used to train a labeler ϕ. Subsequently, we utilize the weights of ϕ to initialize the teacher labeler ϕ_t and the student labeler ϕ_s. Finally, we finetune the labelers ϕ_s, ϕ_t using the Mean Teacher approach to generate pseudo-labels in overlapping areas. The resulting labeler ϕ_t is denoted as SAFormer. After obtaining the pseudo-labels, we employ them for soft supervision of the 3DIS network.

3.2. Process to Generate SAFormer

We develop a novel pseudo-labeler called SAFormer, which accurately predicts labels for overlapping areas, leading to precise results in bbox-supervised 3DIS. Next, we will sequentially introduce the generation process.

3.2.1 Simulated Sample Generation

The abundant non-overlapping bboxes in ScanNetV2 [9] with definite instance labels allow us to generate simulated overlapping samples. As illustrated in Figure 3, we begin by extracting real overlapping samples O and non-overlapping objects P from real scenes. Subsequently, we conduct an analysis of the class distribution and spatial distribution within these real overlapping samples. To be more specific, we determine which class pairs make up overlapping samples, counting the number n of samples for each class pair and calculate the mean μ and variance σ of the distances between the center points of each class pair. After obtaining the statistical data, we commence the simulation of the distribution. Firstly, we perform sampling of class pairs based on the distribution of n. Assuming the sampled class pair is denoted as (a, b), we then uniformly sample one object point cloud for each of the classes a and b from the set P. After obtaining these two point clouds, we perform gaussian sampling based on the corresponding μ and σ to obtain a distance d, representing the distance between the two point clouds. Finally, for the sake of simplicity, we directly translate one of the point clouds along the X or Y-axis by the distance d. It is worth noting that, before performing the distance translation, we align the center points of the point cloud pairs.

To better maintain physical plausibility, we make scene adjustments based on the following two principles [32]: 1) gravity: objects should not float in the air; 2) collision: these two objects should not exhibit any collision. Specific details are covered in the supplementary materials. Another purpose of designing the collision constraint is to verify whether two objects can be matched when constructing simulated samples. Concretely, after multiple (with an upper limit of M) distance samplings and collision constraint corrections, some object pairs may still fail to generate overlapping regions. Such pairs are subsequently excluded from use. Finally, considering that real overlapping samples contain background noise points, we add an appropriate number of $floor$ points to represent the presence of background noise points.

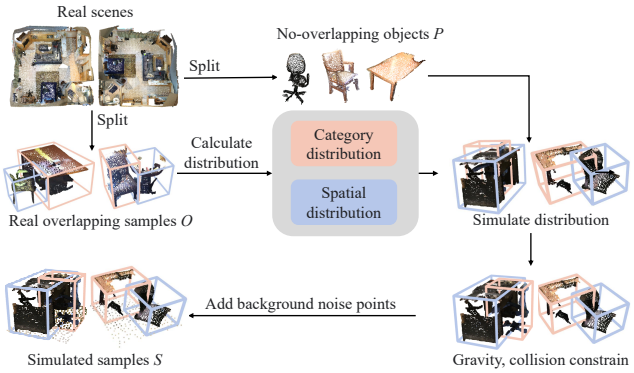

Figure 3. **The process of generating simulated samples.** There are numerous non-overlapping objects (P) with definite instance labels in real scenes. We can generate simulated samples (S) based on the distribution of real overlapping samples (O) and the physical plausibility.

During the aforementioned process, we have obtained simulated overlapping samples S. Currently, we utilize these samples to train a labeler ϕ. The labeler ϕ primarily consists of two components: a lightweight 3D-UNet based on sparse convolution [14, 15, 22] and LGA. In next section, we provide a detailed introduction of LGA.

3.2.2 Local-Global Aware Attention

As shown in Figure 4, LGA mainly contains local-structure attention and global-context attention. Assuming that the input point cloud consists of N points, with each point containing position coordinates (x, y, z) and color information (r, g, b). First, we input the point cloud into a lightweight 3D-UNet to obtain point-level features F. Subsequently, following SPFormer [36], we aggregate the point-level features F into superpoint-level features F_{sup} using average pooling. Next, we initialize two learnable queries Q_1, Q_2, representing two foreground instances respectively. To better model local structure, we separately employ the self-attention layer and the feed-forward layer within the non-overlapping areas of different instances. This approach ensures that each local region interacts with similar regions belonging to the same instance, significantly enhancing the discriminative and representational capabilities of local structures. Specifically, we concatenate $F_{sup,1}$ with Q_1 and $F_{sup,2}$ with Q_2 to form $F_{v,1}, F_{v,2}$, and then input them sep-

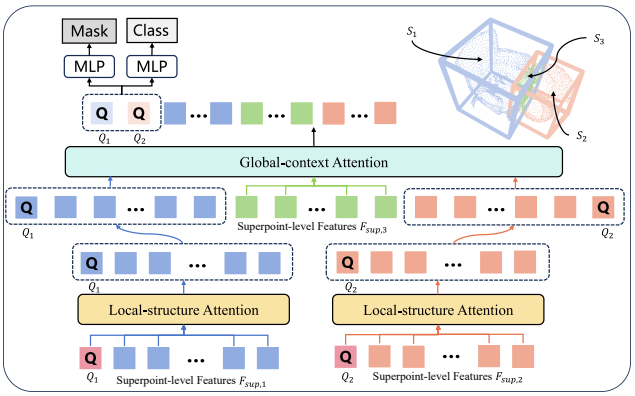

Figure 4. **The Local-Global Aware Attention.** Two foreground queries are input into local-structure attention and global-context attention to generate corresponding masks. S_1, S_2 represent non-overlapping areas. S_3 represents overlapping areas.

arately into the self-attention layer, just as follows:

$$F'_{v,i} = \text{Softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{\mathcal{C}}})\mathcal{V}, i = 1, 2, \tag{1}$$

where $\mathcal{Q} = F_{v,i}W_q$, $\mathcal{K} = F_{v,i}W_k$, $\mathcal{V} = F_{v,i}W_v$, and W_q, W_k, W_v denote the linear transform matrices for queries, keys, and values, respectively. Finally we input F'_v into the feed-forward layer to obtain F''_v.

After modeling the local structures within each non-overlapping area S_1, S_2, and aggregating these structures into a holistic representation through foreground queries, we need to incorporate global information. The specific approach involves concatenating the features $F''_{v,1}, F''_{v,2}$, and $F_{sup,3}$. Then, we input this concatenated features into the self-attention layer and the feed-forward layer. Through this method, we can model relationships between non-overlapping and overlapping areas, between the two foreground instances, and aggregate these global relationships into Q_1 and Q_2, respectively. Finally, to classify the overlapping areas, we obtain the masks $M_{ins,1}, M_{ins,2}$ for the two objects by calculating the dot product between $F_{sup,3}$ and Q_1, Q_2. The final mask is obtained through the Sigmoid function followed by a threshold of 0.5:

$$M_i = \text{Sigmoid}(M_{ins,i}) > 0.5, i = 1, 2. \tag{2}$$

Since M_1 and M_2 represent two different foreground object masks, for areas where both M_1 and M_2 are not activated, we classify them as background areas. This approach naturally helps the labeler filter out background points, which is an improvement over Gapro [29], as Gapro overlooks the presence of background points. Furthermore, to better assist the labeler in learning unified knowledge for the same class, we add a class prediction head.

For training on the simulated overlapping samples S, since the instance labels are complete, we directly use the shared losses from SPFormer [36] and Mask3D [34]:

$$L_{total\_sim} = \lambda_1 L_{cls} + \lambda_2 L_{bce} + \lambda_3 L_{dice}, \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, L_{cls} is the cross-entropy loss, L_{bce} is the binary cross-entropy loss, L_{dice} is the dice loss [27].

3.2.3 Mean Teacher Approach

During the above process, the labeler ϕ has learned prior knowledge about overlapping scenes through training on the simulated samples. Subsequently, we used the pre-trained weights as the initial weights for both the teacher labeler ϕ_t and the student labeler ϕ_s. Then, we apply data augmentation to the real overlapping samples, including flipping, jittering, elastic distortion, and so on. The original samples are fed into the teacher labeler, while the augmented samples are input into the student labeler. Since the labels for non-overlapping areas in real samples are known, we can directly supervise these areas. As for the overlapping areas, to better leverage the predictions of the teacher labeler, we select high-confidence pseudo-labels for the overlapping areas based on a fixed threshold τ. The teacher labeler updates its parameters using the EMA technique. With this design, the teacher labeler can continuously update pseudo-labels online and transfer knowledge to the student. Simultaneously, the student labeler can employ EMA to transmit the acquired knowledge back to the teacher. Going a step further, with the initialization weights obtained through simulation, the teacher labeler gains the ability to distinguish overlapping areas. It can generate higher quality pseudo-labels, accelerating the Mean Teacher's training speed. Finally, the well-trained teacher labeler ϕ_t is referred to as SAFormer, which is used to generate final pseudo-labels for overlapping regions.

For finetuning on the real samples, we adopt a weakly supervised approach. The specific approach can be divided into two parts. First, for non-overlapping areas, where the labels are known but only partial labels of the complete objects, we only supervise the non-overlapping areas.:

$$M' = Q_s F^T_{s,sup,1\cup 2}, \tag{4}$$

where Q_s represents the instance queries of the student labeler, $1 \cup 2$ represents the union of S_1 and S_2,

$$L_{sup} = \lambda_2 L_{bce}(M', M'_{gt}) + \lambda_3 L_{dice}(M', M'_{gt}). \tag{5}$$

Next, for overlapping areas, we obtain high-confidence pseudo-labels based on a threshold τ from the teacher labeler and solely supervise the overlapping areas that have corresponding pseudo-labels:

$$M''_{ps} = Q_t F^T_{t,sup,3} \odot (Q_t F^T_{t,sup,3} > \tau), \tag{6}$$

$$M'' = Q_s F^T_{s,sup,3} \odot (Q_t F^T_{t,sup,3} > \tau), \tag{7}$$

where $\odot$ represents hadamard product, Q_t represents the instance queries of the teacher labeler, M''_{ps} represents the pseudo-labels of overlapping areas,

$$L_{unsup} = \lambda_2 L_{bce}(M'', M''_{ps}) + \lambda_3 L_{dice}(M'', M''_{ps}). \tag{8}$$

The total loss for real samples is:

$$L_{total\_real} = \lambda_1 L_{cls} + L_{sup} + L_{unsup}. \quad (9)$$

3.3. Training a 3DIS Network

Due to the fact that the instance labels for points within non-overlapping bboxes and non-overlapping areas of overlapping bboxes are definite, we can combine these determined instance labels with the pseudo-labels obtained through SAFormer. Then the combined labels are used to train a 3DIS network. It's worth noting that since the predicted pseudo-label values $\in [0, 1]$, which reflect confidence, employing a soft supervision is a better choice. Assuming there are K instances, the pseudo masks $M \in [0, 1]^{K \times N}$,

$$L'_{bce} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} L_{bce}(M_{pred,i,j}, M_{i,j}) * M_{i,j}}{\sum_{i=1}^{K} \sum_{j=1}^{N} M_{i,j}}, \quad (10)$$

where M_{pred} represents the results predicted by the 3DIS network. The total soft loss is:

$$L_{total\_soft} = \widehat{\lambda_1} L_{cls} + \widehat{\lambda_2} L'_{bce} + \widehat{\lambda_3} L_{dice} + L_{net}, \quad (11)$$

where $\widehat{\lambda_1}, \widehat{\lambda_2}, \widehat{\lambda_3}$ are hyperparameters specific to different 3DIS networks, and L_{net} are loss functions unique to different 3DIS networks.

4. Experiments

4.1. Experimental Setup

Datasets and metrics. We conduct our experiments on ScanNetV2 [9] and S3DIS [2] datasets. ScanNetV2 includes 1,613 scenes with 18 instance categories. Among them, 1,201 scenes are used for training, 312 scenes are used for validation, and 100 scenes are used for test. S3DIS is a large-scale indoor dataset collected from six different areas, which contains 272 scenes with 13 instance categories. Following previous works [29, 37], we train on Area 1, 2, 3, 4, 6 and evaluate on Area 5. AP@25 and AP@50 represent the average precision scores with IoU thresholds 25% and 50%, and mAP represents the average of all the APs with IoU thresholds ranging from 50% to 95% with a step size of 5%. On ScanNetV2, we report mAP, AP@50 and AP@25. Moreover, we also report the Box AP@50 and AP@25 results following Gapro [29]. On S3DIS, we report mAP and AP@50.

Implementation details. The whole method BSNet is trained on a single RTX3090. As to the training setting of the pseudo-labeler SAFormer, first we train 100 epochs on simulated samples with a batch size of 64, which takes about 6 hours. Next, we fintune 5 epochs on real samples of ScannetV2 training set with a batch size of 64, which takes about 90 minutes. During inference, it takes approximately 10 minutes to generate pseudo-labels for the entire training set. As to S3DIS, it takes about 100 minutes for fintuning with 5 epochs and 10 minutes to generate pseudo-labels. Given that our pseudo-labeler only needs to be trained once on the simulated samples when applied to multiple datasets, so the more datasets we apply, the more efficient the method is. As to the backbone of SAFormer, we use a lightweight 3D-UNet based on sparse convolution [14, 15, 22] with 3 blocks and 32 media channels. At last, we tune the hyperparameters $\mathbf{M}, \tau, \lambda_1, \lambda_2, \lambda_3$ as 8, 0.9, 2, 5, 2.

Table 1. **Comparison on ScanNetV2 validation set.** %full indicates the percentage of the current method's performance compared to its corresponding fully supervised method. ISBNet† refers that we use the pseudo-labels generated by "Box2Mask [8]: assign points to smaller box" to supervise ISBNet [30].

Method	Sup.	mAP	%full	AP@50	%full	AP@25
Mask3D [34]	Mask	55.2	-	73.7	-	83.5
PointGroup [19]		34.8	-	51.7	-	71.3
SSTNet [21]		49.4	-	64.3	-	74.0
ISBNet [30]		54.5	-	73.1	-	82.5
SPFormer [36]		56.3	-	73.9	-	82.9
CSC [17]	Point	15.9	28.8%	28.9	39.2%	49.6
PointContrast [43]		27.8	50.4%	47.1	63.9%	64.5
Box2Mask(stand-alone) [8]	Box	39.1	-	59.7	-	71.8
ISBNet†		41.8	76.7%	64.8	88.6%	-
WISGP [11] + PointGroup		31.3	89.9%	50.2	97.1%	64.9
WISGP + SSTNet		35.2	71.3%	56.9	88.5%	70.2
GaPro [29] + ISBNet		50.6	92.8%	69.1	94.5%	79.3
GaPro + SPFormer		51.1	90.8%	70.4	95.3%	79.9
Ours + ISBNet		52.8	**96.9%**	71.6	97.9%	82.6
Ours + SPFormer		**53.3**	94.7%	**72.7**	**98.4%**	**83.4**

4.2. Comparison with state-of-the-art methods

ScanNetV2. As shown in Table 1, we compare our approach with existing state-of-the-art methods on the ScanNetV2 validation set. Attributed to the innovative construction of simulated samples by SMT and the capability of LGA to model local and global information, our proposed SAFormer can generate higher-quality pseudo-labels to supervise the 3DIS network. Consequently, our box-supervised 3DIS method outperforms other methods by a significant margin in terms of mAP, AP@50 and AP@25. It is worth emphasizing that our results can achieve 95% in terms of mAP when compared to the corresponding fully supervised methods. This signifies a substantial improvement over previous approaches, which typically achieves only about 90% performance. To vividly illustrate the differences between our method and others, we visualize the qualitative results of pseudo-labels in Figure 5. From the regions highlighted in yellow circles, it is evident that our method can generate more accurate pseudo-labels for overlapping areas.

S3DIS. We evaluate our method on S3DIS using Area 5 in Table 2. Our proposed method achieves superior performance compared to previous methods, with large margins in both mAP and AP@50, demonstrating the effectiveness

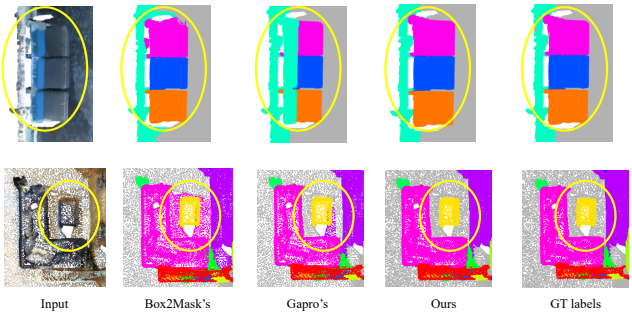

Figure 5. **Qualitative results on ScanNetV2 training set.** Our approach produces highly accurate pseudo instance masks, particularly in overlapping areas (yellow circles).

Table 2. **Comparison on S3DIS on Area 5.** Box2Mask* represents the results of Box2Mask [8] reproduced by Gapro [29] on the S3DIS dataset based on their public code.

Method	Sup.	mAP	%full	AP@50	%full
Mask3D [34]		56.6	-	68.4	-
PointGroup [19]		-	-	57.8	-
SSTNet [21]	Mask	42.7	-	59.3	-
SoftGroup [21]		51.6	-	66.1	-
ISBNet [30]		54.0	-	65.8	-
Box2Mask* [8]		43.6	-	54.6	-
WISGP [11] + PointGroup		33.5	-	46.8	81.0%
WISGP + SSTNet		37.2	87.1%	51.0	86.0%
GaPro [29] + SoftGroup	Box	47.0	91.1%	62.1	93.9%
GaPro + ISBNet		50.5	93.5%	61.2	93.0%
Ours + SoftGroup		51.4	**99.6%**	62.8	95.0%
Ours + ISBNet		**53.0**	98.1%	**64.3**	**97.7%**

and generalization of our method.

4.3. Ablation Study

The following experiments in Table 4 are conducted on ISBNet on the validation set of ScanNetV2, while the others are performed on the training set of ScanNetV2.

Comparison on pseudo-labels. Firstly, we use the metric mAcc to evaluate the quality of pseudo-labels in overlapping areas. Assuming that the predicted pseudo-labels of overlapping areas are P, the GT of overlapping areas are P^{gt}, there are N overlapping areas, there are M_i points in overlapping area i, and $\mathbb{I}$ is the indicator function,

$$\text{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{M_i}\mathbb{I}(P_{i,j} == P_{i,j}^{gt})}{M_i}. \tag{12}$$

The higher mAcc represents the better quality of pseudo-labels. With the help of mAcc, we can explore the impact of different techniques for handling overlapping areas.

As depicted in Table 3, setting B represents the current state-of-the-art technique for handling overlapping areas, and its performance is significantly higher than setting A. We compare our method SAFormer with these methods and conduct an ablation study of each component in setting C. In setting C0, attributed to the neural network's strong fitting capability, even without utilizing our proposed SMT and LGA, our base performance still surpasses the current state-of-the-art method Gapro. In setting C1, we directly train the labeler with a backbone and LA on the real scenes. The results show an improvement of 10.0 in mAcc compared to Gapro, indicating that deep neural networks can accurately predict overlapping area labels through dedicated local structure modeling and the accumulation of multiple samples. In setting C2, we replace LA with GA, resulting in a 5.4 increase in mAcc compared to Gapro. The results suggest the importance of global information, particularly the interaction between the two foreground instances and between overlapping areas and non-overlapping areas. In setting C3, we add the design of GA based on C1, resulting in a 4.4 improvement in mAcc. From the results of C1, C2, and C3, we can conclude that local structure modeling and global relationship modeling complement each other. Good local structures form the basis for modeling global relationships, and modeling global relationships can better unleash the potential of good local structures. In setting C4, to provide stable pseudo-labels for overlapping areas and facilitate information transfer between teacher and student labelers, we add MT. The improved performance in mAcc proves its effectiveness. Finally, to help the labeler gain the ability to distinguish overlapping areas, we add SSG in C5. This enables the labeler to predict higher quality pseudo-labels and achieve faster training speed, as shown in Table 7.

Table 3. **Quality of pseudo-labels in overlapping areas.** Base refers to utilizing a 3D-UNet and a mask and classification head. LA, GA, MT, SSG represent Local-structure Attention, Global-context Attention, Mean Teacher, Simulated Sample Generation.

Handling of overlapping areas	mAcc
A: Box2Mask: assign points to smaller box	24.1
B: Gapro: GP classification with superpoints	38.1
C0: Base	41.5
C1: Ours (LA)	48.1
C2: Ours (GA)	43.5
C3: Ours (LA + GA)	52.5
C4: Ours (LA + GA + MT)	55.3
C5: Ours (LA + GA + MT + SSG)	**59.6**

Table 4. **Effect of our method's components.** Our pseudo-labels: the pseudo-labels generated by our proposed pseudo-labeler SAFormer. Soft loss: the soft loss proposed in Section 3.3.

Our pseudo-labels	Soft loss	mAP	AP@50	AP@25
✗	✗	38.1	59.1	72.7
✓	✗	52.3	71.2	82.1
✓	✓	**52.8**	**71.6**	**82.6**

Effect of our method's components. Table 4 shows 3DIS results with different components. In the first row, we evaluate the approach of ignoring overlapping areas during training and only using the determined regions as pseudo-labels. The second row showcases the efficacy of the pseudo-labels produced by our proposed labeler SAFormer, resulting in a 14.2 improvement in mAP. In the last row, to validate the impact of the soft loss, we conduct a corresponding ablation experiment and achieve a performance

Table 5. **Effect of different pseudo-label utilization methods.** Base refers that pseudo-labels are directly used to train the 3DIS network. Iterative self-training refers that updating pseudo-labels offline after each training round and then using the updated pseudo-labels to further optimize the labeler. After multiple iterations, the latest pseudo-labels are used to train the 3DIS network.

Method	mAcc
Base	52.5
Iterative self-training	52.6
Mean Teacher	**55.3**

Table 6. **Effect of different steps in SSG.** SD, GCC, ABP represent simulating distribution, gravity-collision constrain, adding background points respectively.

SD	GCC	ABP	mAcc
✓	✗	✗	58.5
✓	✓	✗	59.3
✓	✓	✓	**59.6**

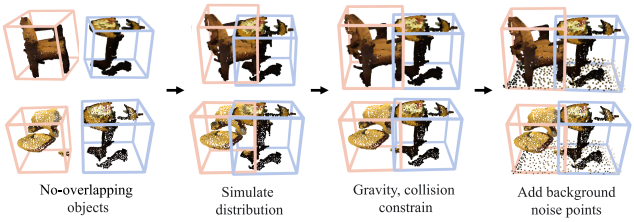

Figure 6. **Qualitative visualization results of our SSG.**

boost of 0.5 in mAP.

Effect of different pseudo-label utilization methods. As shown in Table 5, we observe that iterative self-training contributes minimally to performance improvement, whereas Mean Teacher results in a 2.8 increase in mAcc. The findings highlight that Mean Teacher can generate higher quality pseudo-labels by facilitating information transfer between the student and teacher labeler.

Effect of different steps in SSG. Table 6 illustrates that as the simulated overlapping samples become more realistic, the quality of pseudo-labels is getting better. It's worth noting that adding background points results in a 0.3 increase in mAcc. This is partly because it makes the samples more realistic. On the other hand, it is because our designed mask activation using sigmoid function can naturally filter out background points. In order to illustrate the generation process more vividly, we visualize the qualitative results in Figure 6. It is shown that the generated simulated samples successfully combine the individual 3D shapes in a meaningful way.

Effect of SSG. As shown in Table 7, with the assistance of SSG, the labeler can predict higher quality pseudo-labels. Moreover, owing to the labeler's initialization with simulated samples, the teacher labeler can furnish more stable and accurate pseudo-labels in the early stages of training, thereby expediting the overall training process.

Effect of a class head. Based on Table 8, it can be deduced that the incorporation of a class head helps the labeler acquire unified representations for the same class, resulting in more precise pseudo-labels.

Table 7. **Effect of SSG to MT.**

Setting	Training time(h)	mAcc
w/o SSG	40	55.3
w SSG	**1.5**	**59.6**

Table 8. **Effect of a class head.**

Setting	mAcc
w/o class head	59.2
w class head	**59.6**

Table 9. **Comparison of parameters and training time.** T represents the total training time, which includes the time to generate pseudo-labels and the time to train 3DIS network with the pseudo-labels. $\widehat{\mathrm{P}}$ represents the pseudo-labeler parameters, P represents the corresponding 3DIS network parameters, and %full denotes the proportion of $\widehat{\mathrm{P}}$ to P.

Method	T(h)	$\widehat{\mathrm{P}}$(M)	P(M)	%full
Gapro + ISBNet	150	-	30.7	-
Gapro + SPFormer	80	-	17.6	-
Ours + ISBNet	72	2.4	30.7	7.8%
Ours + SPFormer	37	2.4	17.6	13.6%

4.4. Parameters and Training Time Analysis

Table 9 reports the parameters and the training time on ScanNetV2 training set. For a fair comparison, the reported training time is measured on the same device. Our pseudo-labeler utilizes only about 10% of the corresponding 3DIS network parameters, making it very lightweight. And in terms of time, it is less than half of Gapro's. This can be attributed to different self-training ways and different objects to which self-training is applied. Gapro performs iterative self-training on pseudo-labeler and 3DIS network, while our method performs Mean Teacher self-training only on pseudo-labeler. Therefore, our method not only eliminates the high time cost caused by repeated training of the 3DIS network, but also greatly alleviates the training time of Mean Teacher through the design of SMT.

5. Conclusion

In this paper, we propose the Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation, which devises a novel pseudo-labeler called SAFormer. To the best of our knowledge, SAFormer is the first labeler incorporating the deep neural network and Mean Teacher in this task, and innovatively constructs simulated samples to facilitate training. Furthermore, the well-designed transformer decoder LGA effectively models local structures and global relationships of point clouds. Extensive experiments conducted on two widely used box-supervised 3D instance segmentation benchmarks demonstrate the superior performance of our method.

6. Acknowledgements

This work was partially supported by the Youth Innovation Promotion Association CAS 2018166, Anhui Provincial Natural Science Foundation (Grant 2308085QF222) and National Defense Basic Scientific Research program (JCKY2021601B013).

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. 3

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2, 6

[3] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23839–23848, 2023. 2, 3

[4] 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

[5] 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

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3

[8] 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*, pages 681–699. Springer, 2022. 1, 3, 6, 7

[9] 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. 2, 4, 6

[10] Jiacheng Deng, Chuxin Wang, Jiahao Lu, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Zhe Zhang. Se-ornet: Self-ensembling orientation-aware network for unsupervised point cloud shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5364–5373, 2023. 2

[11] 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, 2, 3, 6, 7

[12] 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. 2

[13] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 3

[14] Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 4, 6

[15] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 4, 6

[16] Jian Han, Ya-Li Li, and Shengjin Wang. Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 790–798, 2022. 3

[17] 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

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 3

[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, 6, 7

[20] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 2, 3

[21] 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, 6, 7

[22] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 806–814, 2015. 4, 6

[23] 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. 2

[24] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, pages 2949–2958, 2021.

[25] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18516–18526, 2023. 1, 3

[26] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14482–14491, 2022. 3

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[28] 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.

[29] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Gapro: Box-supervised 3d point cloud instance segmentation using gaussian processes as pseudo labelers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17794–17803, 2023. 1, 2, 3, 5, 6, 7

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 2, 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.

[32] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3283–3292, 2021. 4

[33] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022.

[34] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 1, 3, 5, 6, 7

[35] 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. 1, 2, 3

[36] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2393–2401, 2023. 1, 3, 4, 5, 6

[37] 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

[38] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35: 29975–29988, 2022.

[39] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9367–9376, 2022. 3

[40] Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3240–3249, 2023. 3

[41] Yuting Wang, Velibor Ilic, Jiatong Li, Branislav Kisačanin, and Vladimir Pavlovic. Alwod: Active learning for weakly-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6459–6469, 2023. 3

[42] 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

[43] 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. 3, 6

[44] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 3

[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. *Advances in neural information processing systems*, 32, 2019. 2

[46] 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

[47] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 3

[48] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zheng-Jun Zha. Exploiting sample uncertainty for domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3538–3546, 2021. 2, 3

[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.

[50] Min Zhong, Xinghao Chen, Xiaokang Chen, Gang Zeng, and Yunhe Wang. Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2