2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

End-to-End Semi-Supervised 3D Instance Segmentation with PCTeacher

Linfeng Li[1] and Na Zhao[2†]

Abstract— **3D instance segmentation is a fundamental and critical task for enabling robots to operate effectively in unstructured 3D environments. In order to address the challenges posed by the high demand for large-scale annotated data and the limited availability of such data in the context of 3D instance segmentation, we study semi-supervised 3D instance segmentation problem and propose a novel end-to-end framework based on the mean teacher paradigm, named PCTeacher. Our PCTeacher generates both point-level and cluster-level pseudo labels to harness knowledge from unlabeled data. It notably enhances the training stability through end-to-end training and improves pseudo-label quality. Specifically, for point-level pseudo labels, PCTeacher employs a multi-view fusion strategy to achieve higher precision and recall. Regarding cluster-level pseudo labels, it introduces a hybrid grouping strategy to generate more potential proposals and utilizes a point-cluster agreement-based thresholding (PCAT) mechanism to fully exploit cluster-level pseudo labels. By combining and strengthening both point-level and cluster-level pseudo labels, our PCTeacher achieves state-of-the-art performance on two benchmark datasets across multiple labeled data ratios with a more compact network compared to the existing method.**

I. INTRODUCTION

3D instance segmentation is a vital task in which individual instances within a 3D scene are identified and segmented. It holds significant value for robotic systems as it enables them to navigate and interact with objects in physical 3D environments [1], [2]. While fully supervised methods [3], [4], [5] have achieved impressive performance, these approaches heavily depend on the availability of large-scale and well-annotated datasets. However, the annotation process for 3D instance segmentation can be incredibly challenging and time-consuming, particularly for 3D indoor datasets, where the amount and diversity of data make annotation exceptionally difficult. This limitation poses a significant challenge to the practical implementation of these methods in real-world applications involving robot-object interaction.

Semi-supervised learning (SSL) is an effective solution to address the aforementioned data annotation burden, as it allows the model to achieve comparable performance to fully-supervised models by learning from a smaller set of labeled data while utilizing the vast amount of available unlabeled data. To the best of our knowledge, TWIST [6] is the first and only work studying this demanding yet under-explored semi-supervised 3D instance segmentation task. TWIST is a multi-stage framework that alternates between generating pseudo-labels and training the network with both labeled and pseudo-labeled data. In the pseudo label generation stage, TWIST incorporates a re-correction module to refine the pseudo labels (*i.e.* semantic and offset predictions) for relatively confident clusters. Although TWIST has made pioneering strides in this area, it exhibits several limitations: 1) The iterative training process employed by TWIST could introduce instability and training difficulties, potentially affecting the convergence and overall training efficiency. 2) TWIST's exclusive reliance on cluster-level pseudo labels neglects valuable information from points that are not included in the predicted clusters, limiting its knowledge extracted from missing foreground points and background points like wall, floor, and ceiling. 3) The filtering of cluster-level pseudo labels (*i.e.* selecting relatively confident clusters) in TWIST is based on a predefined fixed threshold, which poses challenges in finding a good trade-off between precision and recall of pseudo labels.

To tackle these limitations, we propose a novel end-to-end semi-supervised instance segmentation framework, which is capable of generating both accurate **P**oint-level and **C**luster-level pseudo labels, named PCTeacher. PCTeacher adopts the mean teacher paradigm [7], where the teacher model is updated as an exponential moving average (EMA) of the student model, accumulating historical knowledge from the student model. Our PCTeacher utilizes this teacher model to generate pseudo-labels on-the-fly, enabling end-to-end training. This enhances training stability compared to the self-training method used in TWIST. Notably, our PCTeacher generates both point-level and cluster-level pseudo labels, allowing for comprehensive knowledge extraction from unlabeled data. These pseudo labels are further enhanced in quality through our specialized designs.

For point-level pseudo labels, PCTeacher employs a multi-view fusion strategy, which combines features from different view augmentations of each point cloud to eliminate potential false point-level pseudo labels and discover missed point-level pseudo labels, enhancing precision and recall for both foreground and background points. Regarding cluster-level pseudo labels, our PCTeacher incorporates a hybrid grouping technique that increases the number of potential proposals, thereby improving the recall of cluster-level pseudo labels. Moreover, PCTeacher introduces a Point-Cluster Agreement-based Thresholding (PCAT) mechanism, which leverages the agreement between point-level and cluster-level predictions to produce high-certain (*i.e.* high-precision) and mid-certain cluster-level pseudo labels. These pseudo labels can provide varying degrees of supervision and can be used with different cost functions to fully exploit valuable information. Additionally, we choose points from clusters with extreme certainty to reduce noise in point-level pseudo labels, result-

[1] School of Computer Science and Engineering, Nanyang Technological University
[2] Information Systems Technology and Design, Singapore University of Technology and Design, na_zhao@sutd.edu.sg
† corresponding author

ing in more precise pseudo offset supervision.

Our contributions are as follows: 1) We propose the PCTeacher, an end-to-end 3D instance segmentation framework, which is able to produce both high-quality point-level and cluster-level pseudo labels online. 2) PCTeacher provides reliable point-level pseudo labels with semantic information via multi-view fusion for all points, including those within and outside potential clusters, even background points. 3) PCTeacher improves the recall and reliability of the cluster-level pseudo labels through hybrid grouping and point-cluster agreement-based thresholding. 4) PCTeacher achieves state-of-the-art performance on ScanNet and S3DIS under various labeled ratios, validating its effectiveness in semi-supervised 3D instance segmentation.

II. Related Work

3D Instance Segmentation. 3D instance segmentation involves assigning semantic labels to each instance for a given point cloud. Based on their prediction pipelines, the existing methods for accomplishing 3D instance segmentation can be categorized into two main approaches: proposal-based methods [8], [9], [10], [11] and clustering-based approaches [12], [13], [14], [4], [15], [3], [5]. Proposal-based methods adopt a top-down approach that involves generating object proposals and subsequently predicting the instance masks within each proposed region. Similar to the 2D domain, the efficacy of these methods is highly dependent on the quality of the proposals generated in the initial stage. In contrast, clustering-based methods consider a bottom-up policy that directly produces point-level labels and groups points into clusters based on their similarities. SGPN [12] employs PointNet-like networks to extract features and cluster points by their features. MTML [15] devises a multi-task learning strategy to aggregate points belonging to the same instance while separating points from different clusters. Mask3D[5] represents instance masks as queries and utilizes a transformer decoder to predict instance masks. Pointgroup [13] performs point-level offset and semantic predictions and clusters the original and offset-shifted point sets. Building upon the pipeline of Pointgroup, HAIS [14] enhances performance by hierarchical aggregation. SoftGroup [4] introduces a soft group strategy, allowing each point to have associations with multiple classes. Given the outstanding performance and flexibility of SoftGroup, we choose it as the backbone segmentation model for our approach.

Semi-supervised Learning. Semi-supervised learning has been extensively studied in the 2D domain as a means to reduce the dependence of deep learning on large labeled datasets. Consistency-based and pseudo-labeling methods have emerged as two represetative techniques for semi-supervised computer vision tasks. Consistency-based methods are based on the assumption that predictions of the same image should remain stable under various perturbations such as data augmentation [16] and model regularization [7], [17]. Pseudo-labeling methods aim to generate high-quality pseudo-labels for unlabeled data using proper criteria. Fixmatch [18] first proposes a teacher-student model where the teacher generates pseudo-labels for weakly-augmented images, and the student is trained to predict these labels when processing strongly-augmented versions of the same images. Flexmatch [19] and Freematch [20] extend this pipeline with adaptive classification thresholds. Pseudo-labeling methods have demonstrated state-of-the-art performance. One of our contributions is to effectively adapt these semi-supervised learning techniques for 3D instance segmentation task.

Semi-supervised 3D Instance Segmentation. In the context of 3D scene understanding tasks, most existing semi-supervised learning approaches are proposed for 3D semantic segmentation [21], [22] or 3D object detection [23], [24]. To the best of our knowledge, only one work [6] has been specifically designed for semi-supervised 3D instance segmentation. TWIST [6] introduces a multi-stage framework that generates cluster-level pseudo labels, offering semantic and offset supervision to the points within instances. However, relying solely on high-quality cluster-level pseudo labels can lead to the omission of semantically correct foreground points in low-quality objects and background points that can also provide valuable information to improve classification capability. Moreover, cluster-level pseudo labels filtered with a fixed threshold can be unreliable. To address these issues, we propose an end-to-end framework that generates precise point-level and cluster-level pseudo labels.

III. Methodology

A. Framework Overview

Fig. 1 illustrates the overall framework of our end-to-end framework PCTeacher, which is composed of a student model and a teacher model. PCTeacher leverages both labeled and unlabeled data to train the student model. The teacher model generates both point- and cluster-level pseudo labels for unlabeled data and is updated by the student model. These pseudo labels offer valuable guidance for the point-level semantic and offset prediction, as well as the cluster-level classification branches of the student model.

We utilize SoftGroup [4] as the backbone for our 3D instance segmentation model, which includes a point-wise prediction network and a cluster refinement network. The former network consists of a semantic branch that predicts semantic logits $\hat{S} \in \mathbb{R}^{N\times(\mathbb{C}_f+\mathbb{C}_b)}$ and an offset branch that outputs $\hat{O} \in \mathbb{R}^{N\times 3}$, where N, $\mathbb{C}_f$, and $\mathbb{C}_b$ represent the total number of points, foreground classes, and background classes, respectively. $\hat{O}$ estimates the distance from each point to the geometric center of the instance the point belongs. Subsequently, a soft group strategy is performed to form instances (*i.e.* clusters). Finally, each cluster yields its classification prediction $\hat{C} \in \mathbb{R}^{\mathbb{C}_f}$, class-aware IoU prediction $\hat{E} \in \mathbb{R}^{\mathbb{C}_f}$, and mask prediction $\hat{M} \in \mathbb{R}^{m\times\mathbb{C}_f}$ through the cluster refinement network.

B. PCTeacher Training

At the start of PCTeacher's training, both the teacher and student models are initialized with a pre-trained model trained exclusively on labeled data. In each training iteration, we form the training data batch by sampling from both the

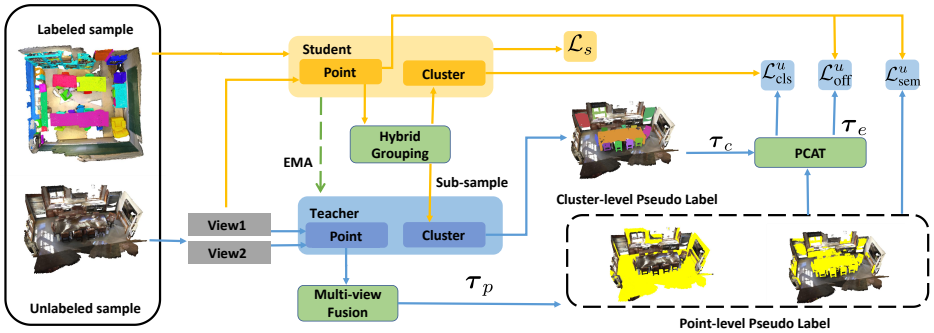


Fig. 1. **Overall framework of our PCTeacher**. The student accepts a view of the unlabeled data and utilizes the hybrid grouping to generate proposals. Meanwhile, the teacher performs pseudo-labeling on two views of the unlabeled data and the proposals from the student to generate point-level and cluster-level pseudo labels, respectively. The student model is updated with a supervised and unsupervised loss while the teacher model is updated by EMA. The variables τ_p, τ_c, and τ_e represent the thresholds for point-level pseudo label, cluster-level pseudo label and IoU thresholds, respectively.

unlabeled set D_u and labeled set D_l according to a fixed sampling ratio γ. The teacher model generates pseudo labels by processing the weakly augmented version of the unlabeled data, while the student model is trained to predict labels for the strongly augmented version of the same data. The student model is also supervised by the ground truths of the labeled data. Thus, the overall loss can be expressed as $\mathscr{L}_{\text{overall}} = \mathscr{L}_s + \mathscr{L}_u$. Following SoftGroup [4], the supervised loss $\mathscr{L}_s$ comprises five components, *i.e.* semantic, offset, classification, IoU and mask losses. In this paper, we compute the unsupervised loss $\mathscr{L}_u$ as:

$$\mathscr{L}_u = \mathscr{L}^u_{\text{sem}} + \mathscr{L}^u_{\text{off}} + \mathscr{L}^u_{\text{cls}}, \tag{1}$$

where $\mathscr{L}^u_{\text{sem}}$, $\mathscr{L}^u_{\text{off}}$ represent unsupervised point-level semantic and offset loss, while $\mathscr{L}^u_{\text{cls}}$ represents unsupervised cluster-level classification loss. We will detail them in Section III-C and III-D. The teacher model is updated with the student model via EMA mechanism:

$$\theta_T = \alpha\theta_T + (1-\alpha)\theta_S, \tag{2}$$

where θ_T and θ_S are the parameters of the teacher and student models, and α is the hyper-parameter that balances the historical parameters of the teacher and the current parameters of the student. Thanks to the temporal integration of historical student models, the teacher model maintains stable and superior performance throughout the entire training process, as shown in Fig. 2. Consequently, it produces more accurate pseudo-labels compared to the self-training method.

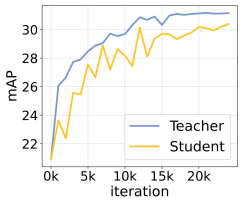


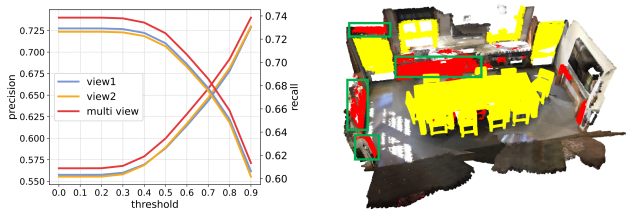
Fig. 2. Performance comparison between the teacher and student models on ScanNet val set under 5% labeled setting.

Data augmentation is crucial in teacher-student based semi-supervised learning. In our approach, the weak augmentation consists of random cropping, flipping, rotating, and scaling operations. The strong augmentation includes all the operations in the weak augmentation, as well as an additional sub-sampling within the input to the cluster refinement of the student model. This sub-sampling strategy disrupts local geometric relationships while preserving global object relationships. It prevents the student from memorizing point coordinates but encourages learning of underlying object structures and global context.

C. Point-Level Pseudo Label Generation

The reliance on solely high-quality cluster-level pseudo labels often results in a low recall of pseudo labels. This is evident in the yellow regions depicted in Fig. 3(b), where the high-quality cluster-level pseudo labels cover only a fraction of points. Nonetheless, TWIST discards point-level pseudo labels due to their low quality or reliability. On the contrary, the teacher model integrated with the multi-view fusion strategy in our PCTeacher allows us to generate more accurate pseudo labels throughout the entire training process and therefore circumvents the limitations of self-training method in TWIST.

Considering that a good model should exhibit consistent predictions across different augmentations of the same points, we feed the teacher model with the same scene after two random augmentations to obtain the semantic predictions $\hat{S}^1_T$ and $\hat{S}^2_T$. The final semantic probability is calculated by aggregating the predictions as: $\hat{S}_T = \text{softmax}((\hat{S}^1_T + \hat{S}^2_T)/2)$. We employ a point-level pseudo-label filtering technique based on a threshold τ_p. Balancing the precision and recall of pseudo labels is a crucial challenge in SSL. Using multi-view fusion, we effectively eliminate the false point-level pseudo labels that the model may be overconfident in one view, and also recall the samples that may have been overlooked in one view. Consequently, this approach offers an opportunity to achieve high recall while maintaining a high level of accuracy compared to the single-view scenario shown in Fig. 3(a). The point-level pseudo labels consist of both foreground and background point-level pseudo labels. As shown in Fig. 3(b) green boxes, the foreground point-level pseudo labels could cover numerous easily classified points in the low-quality clusters. Interestingly, we find that the background point-level pseudo labels can also provide useful contextual information, which beneficially contributes to the performance improvement of the model (*c.f.* Tab. III).



(a) (b)

Fig. 3. **Effect of Multi-view Fusion.** (a) Precision and recall of point-level pseudo labels on 20 unlabeled scenes at various thresholds under ScanNet 10% labeled setting. (b) Point-level pseudo labels complement cluster-level pseudo labels. Cluster-level pseudo labels are highlighted in yellow and red parts emphasized with green boxes are foreground point-level pseudo labels that are not included in cluster-level pseudo labels.

Overall, all these point-level pseudo-labels are leveraged to provide guidance for the semantic branch of the student model, and the loss can be formulated as:

$$\mathscr{L}^{u}_{\text{sem}} = \frac{1}{\sum m_i} \sum_{i=1}^{N_p} \text{CE}(\hat{S}_{S|i}, \text{argmax}(\hat{S}_{T|i})) \cdot m_i, \tag{3}$$

where N_p refers to the number of points, $\hat{S}_{S|i}$ denotes the semantic predictions of the student model on point i from view 1, and i indexes each individual point in the point cloud. m_i represents a binary mask, where $m_i = \mathbb{1}[\max(\hat{S}_{T|i}) > \tau_p]$.

Similarly, we can formulate unsupervised offset loss as:

$$\mathscr{L}^{u*}_{\text{off}} = \frac{1}{\sum m_i} \sum_{i=1}^{N_p} \text{L1}(\hat{O}_{S|i}, \hat{O}_{T|i})) \cdot m_i, \tag{4}$$

where $\hat{O}_{S|i}$ and $\hat{O}_{T|i}$ denotes the offset predictions of the student model and the teacher model, respectively. However, directly supervising point-level offsets in this manner results in a performance decrease, as indicated in Tab. III. This decline might be attributed to inaccurate predictions and the susceptibility of offset supervision. Consequently, we introduce an improved version based on cluster-level predictions, detailed in Sec. III-D.

D. Cluster-level Pseudo Label Generation

TWIST [6] generates instance proposals by clustering points with class labels derived from their maximum semantic scores in the shifted coordinate space. However, in limited data scenarios, inaccurate offset predictions cause a scarcity of generated instance proposals, thus leading to low recall of cluster-level pseudo labels. Moreover, TWIST applies a fixed thresholding to filter pseudo labels, which would struggle with the trade-off between precision and recall of pseudo labels. To address these issues, we propose hybrid grouping to enhance the recall of cluster-level pseudo labels and further introduce point-cluster agreement-based thresholding to obtain reliable cluster-level pseudo labels.

1) Hybrid Grouping: To generate more potential instance proposals and improve the recall of cluster-level pseudo labels, we introduce hybrid grouping by performing cluster grouping in both shifted and original coordinate spaces. In the shifted coordinate space, we utilize a soft grouping mechanism [4]. Unlike TWIST that associates each point with the most probable class, the soft mechanism allows each point to be associated with multiple classes, thereby mitigating challenges arising from errors in semantic prediction. Furthermore, to counteract the adverse effects of inaccurate offset predictions, we cluster points in the original coordinate space based on their class labels derived from the maximum semantic score. Notably, the presence of inaccurate offset predictions and the inherent randomness in the grouping process result in numerous mismatches between the proposals generated by the teacher and student models. To achieve a one-to-one matching, we directly provide the proposals generated by the student to the teacher model.

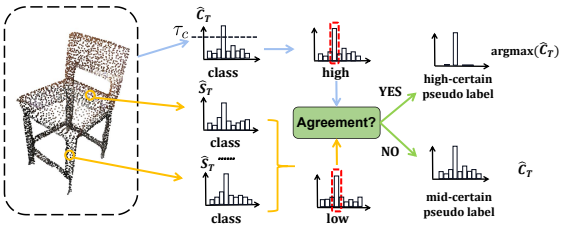


Fig. 4. Effect of point-cluster agreement-based thresholding.

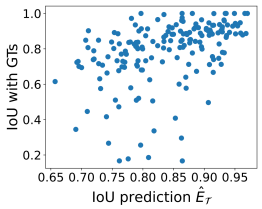


Fig. 5. **Correlation between the IoU predictions and actual IoU.** IoU prediction score $\hat{E}_T$ on 20 unlabeled scenes and actual IoU with ground truths under ScanNet 5% labeled setting.

2) Point-Cluster Agreement-based Thresholding: Similar to TWIST, we initially utilize a threshold τ_c to filter cluster-level pseudo labels based on the classification predictions of the teacher $\hat{C}_T$. However, we observed a discrepancy between the classification prediction of the cluster and the semantic prediction of its constituent points. This suggests that we can incorporate the agreement between the low-level (point-level) classification prediction and high-level (cluster-level) classification prediction into the criteria for filtering cluster-level pseudo labels. As shown in Fig. 4, for a given cluster j, we define $\text{high}_j = \text{argmax}(\hat{C}_{T|j})$ as the high-level prediction and $\text{low}_j = \text{argmax}(\frac{1}{\sum n_i} \sum \hat{S}_{T|i} \cdot n_i)$ as the low-level prediction, where n_i is a binary mask indicating whether a point i belongs to cluster j. By conducting the agreement check between high_j and low_j for each cluster j, we can categorize the cluster-level pseudo labels into *high-certain* and *mid-certain* cluster-level pseudo labels. In the case of high-certain cluster-level pseudo labels, where the classification is reliable, we apply the cross-entropy loss by converting the pseudo labels into one-hot vectors. For the mid-certain cluster-level pseudo labels, the distribution of class scores entails valuable implicit information despite the mismatch between low-level and high-level prediction. Therefore, we adopt the soft label (*i.e.* teacher predictions) as the target and employ the KL-divergence loss. For each proposal j,

TABLE I

COMPARISON WITH TWIST [6] ON **SCANNET V2** AND **S3DIS** DATASET WITH 1%, 5%, 10% AND 20% LABELED DATA SETTINGS.

Dataset	Method	1%			5%			10%			20%		
		mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}
ScanNet v2	Sup-only	5.1	9.8	17.6	18.2	32.0	47.0	26.7	42.8	58.9	29.3	47.9	63.0
	TWIST	9.6 (+4.5)	17.1 (+7.3)	26.2(+8.6)	27.0(+8.8)	44.1(+12.1)	56.2(+9.2)	30.6(+3.9)	49.7(+6.9)	63.0(+4.1)	32.8(+3.5)	52.9(+5.0)	66.8(+3.8)
	Softgroup	6.7	13.1	21.9	20.9	37.2	52.1	26.1	45.3	59.7	34.0	53.8	67.9
	Ours	**11.7**(+5.0)	**22.5**(+9.4)	**33.1**(+11.2)	**31.1**(+10.2)	**51.3**(+14.1)	**66.9**(+14.8)	**33.2**(+7.1)	**52.8**((+7.5)	**67.8**(+8.1)	**39.0**(+5.0)	**60.6**(+6.8)	**74.1**(+6.2)
S3DIS	Sup-only	9.0	12.7	20.7	21.5	30.4	42.8	25.2	36.8	48.3	29.9	41.2	54.5
	TWIST	17.9 (+8.9)	22.5 (+9.8)	27.1(+6.4)	27.1(+5.6)	37.1(+6.7)	48.6(+5.8)	33.6(+8.4)	45.6(+8.8)	55.8(+7.5)	36.7(+6.8)	48.4(+7.2)	59.7(+5.2)
	Softgroup	10.2	16.1	26.4	21.2	32.3	44.3	28.4	36.4	48.1	35.3	47.8	59.6
	Ours	**19.9**(+9.7)	**26.7**(+10.6)	**38.9**(+12.5)	**30.2**(+9.0)	**40.5**(+8.2)	**51.7**(+7.4)	**35.7**(+7.3)	**48.4**((+12.0)	**57.3**(+9.2)	**39.4**(+4.1)	**53.3**(+5.5)	**63.4**(+3.8)

the cluster-level classification loss can be defined as:

$$\mathscr{L}^{u}_{\text{cls}|j} = \begin{cases} \text{CE}(\hat{C}_{S|j}, \text{argmax}(\hat{C}_{T|j})) \cdot k_j, & \text{high}_j = \text{low}_j; \\ \text{KL}[\hat{C}_{S|j} \parallel \hat{C}_{T|j}] \cdot k_j, & \text{high}_j \neq \text{low}_j. \end{cases} \tag{5}$$

Here $k_j = \mathbb{1}[\max(\hat{C}_{T|j}) > \tau_c]$. PCAT ensures high-certain pseudo labels while simultaneously extracting useful knowledge from the mid-certain pseudo labels.

Motivated by the observation of a significant positive correlation between the IoU predictions $\hat{E}_T$ of high-certain cluster-level pseudo labels and the actual IoU values shown in Fig. 5, we further select several extreme-certain clusters based on an IoU threshold τ_e. These extreme-certain clusters possess both reliable classification information and high-quality masks. As a result, we first compute pseudo centroid for each extreme-certain cluster $\mathbf{c}_j$ as $a_j = \frac{1}{|\mathbf{c}_j|}\sum_{p_i \in \mathbf{c}_j}(p_i + \hat{O}_{T|i})$, where $\hat{O}_{T|i}$ denotes the teacher model offset predictions on point i from view 1 and p_i refers to original point coordinate. Subsequently, we calculate the offset loss for all points within the selected extreme-certain clusters, and revise Eq. 4 as:

$$\mathscr{L}^{u}_{\text{off}} = \frac{1}{N_{\mathbf{c}}} \sum \frac{1}{|\mathbf{c}_j|} \sum_{p_i \in \mathbf{c}_j} \text{L1}(\hat{O}_{S|i}, (a_j - p_i)), \tag{6}$$

where $N_{\mathbf{c}}$ refers to the number of extreme-certain clusters.

IV. EXPERIMENTS

A. Experiment Settings

Datasets. We validate the effectiveness of our method on two benchmark datasets: ScanNet v2 [25] and S3DIS [26]. ScanNet v2 comprises diverse indoor scenes, with 1,201 scans for training and 312 for validation. S3DIS consists of 272 scans acquired from 6 large areas, and we perform validation on Area 5 and leverage the remaining areas for training. Following the previous work [6], we randomly sample 1%, 5%, 10% and 20% scans from the training set as labeled training data and form the rest of the training data into unlabeled training data on both datasets.

Implementation Details. For each labeled ratio, we first pre-train our backbone SoftGroup[4] on the available labeled data with batch size 8 with a learning rate set at 0.004 until convergence. After initializing the student and teacher models with the pre-trained backbone, we train the entire framework for 24,000 iterations on 4 Tesla v100 GPUs using Adam optimizer with a learning rate of 0.002. The batch size is set to 16, and the sampling ratio γ is 0.5.

TABLE II

COMPARISON WITH POINTCONTRAST [27], CSC [28] ON **SCANNET V2** AND **S3DIS** DATASET WITH 1%, 5%, 10% AND 20% LABELED DATA SETTINGS. MAP IS USED AS THE EVALUATION METRIC.

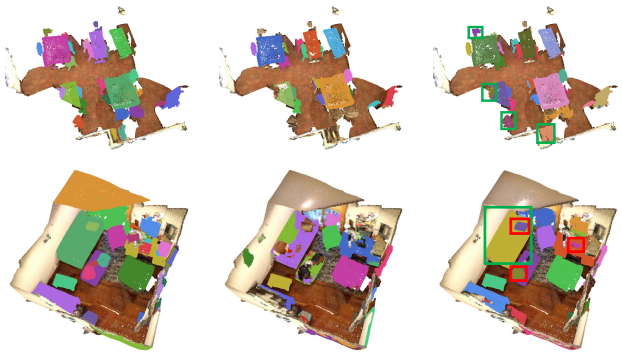
dataset	Method	1%	5%	10%	20%
ScanNet v2	PointContrast	7.2	19.4	27.0	30.2
	CSC	7.1	20.9	27.3	30.6
	Ours	**11.7**	**31.1**	**33.2**	**39.0**
S3DIS	PointContrast	13.4	22.9	27.1	31.2
	CSC	14.6	24.9	29.7	33.5
	Ours	**19.9**	**30.2**	**35.7**	**39.4**

We set the EMA update parameter α to 0.999. For the point-level pseudo labels, we set the confidence threshold $\tau_p = 0.8$ for both background and foreground pseudo labels. For the cluster-level pseudo labels, we set the classification confidence threshold $\tau_c = 0.7$. and the IoU threshold $\tau_e = 0.9$. Regarding data augmentation, we utilize random cropping, flipping, rotating, and scaling as our weak augmentation and perform a random sub-sampling of 90% on the input of the student cluster component. During inference, we report the performance of the teacher model.

B. Main Results

Tab. I compares our method with the previous state-of-the-art method, TWIST, across various label ratios. Additionally, as we employed a different backbone from TWIST, we assess the impact of the backbone by comparing with the fully-supervised counterpart. Specifically, 'Sup-only' denotes the fully-supervised performance of TWIST's backbone model, while 'SoftGroup' represents the fully-supervised performance of our backbone model, both are trained only on labeled data. Note that we report the performance of TWIST without any self-supervised pre-training for a fair comparison. As shown in the table, our method achieves superior performance and exhibits significant improvements across different evaluation protocols (mAP, AP50, and AP25) across all label ratio settings on both datasets.

Furthermore, Tab. II compares our method with two representative self-supervised approaches, PointContrast [27] and CSC [28], as self-supervised pre-training is also a potential solution to tackle the problem of insufficient labeled data. The two baseline methods initially pre-train on the unlabeled ScanNet dataset and subsequently fine-tune using the available labeled data. This table clearly demonstrates that our method outperforms these self-supervised approaches, underscoring our effective utilization of unlabeled data.



(a) (b) (c)

Fig. 6. **Qualitative Results on ScanNet v2 val set under 5% labeled setting**. (a) Ground truth. (b) SoftGroup. (c) Our PCTeahcer.

TABLE III

ABLATION STUDIES ON THE UNSUPERVISED LOSS TERMS.

No.	$\mathcal{L}^u_{\text{sem}_f}$	$\mathcal{L}^u_{\text{sem}_b}$	$\mathcal{L}^u_{\text{cls}}$	$\mathcal{L}^{u*}_{\text{off}}$	$\mathcal{L}^u_{\text{off}}$	mAP
1	✓					26.7
2	✓	✓				28.0
3	✓	✓	✓			30.4
4	✓	✓	✓	✓		28.3
5			✓		✓	28.5
6	✓	✓	✓		✓	**31.1**

In Fig. 6, we provide a qualitative comparison with SoftGroup on the ScanNet v2 dataset under the 5% labeled setting. In the first row, depicting a relatively spacious scene, both SoftGroup and PCTeacher perform well. However, PCTeacher excels in recognizing masks for challenging objects, as indicated by the green boxes. In the second row, showcasing a challenging cluttered scene where instances are frequently stacked or positioned adjacent to one another, SoftGroup's performance is subpar. In contrast, our PCTeacher generates higher-quality masks, highlighted by the green boxes, and also identifies masks in red boxes, which are missed positive masks by SoftGroup.

C. Ablation study

We conduct experiments on 5% labeled ScanNet v2 to validate the effectiveness of our key design choices.

Ablation on Different Loss Terms. In Tab. III, we incrementally incorporate individual unsupervised loss term into the baseline (row 1) and evaluate its performance. Comparing the 2^{nd} row to 1^{st} row, the 1.3% improvement demonstrates the effectiveness of incorporating semantic information from background point-level pseudo labels. When we introduce classification supervision at cluster level (row 3), the performance improves by an additional 2.4%. As demonstrated in row 4, incorporating naive unsupervised point-level offset supervision (*c.f.* Eq. 4) leads to a performance drop of -2.1%. In contrast, our revised offset supervision (*c.f.* Eq. 6) based on extreme-certain clusters overcomes this drop and achieves the best performance, as shown in row 6.

Moreover, when comparing the 6^{th} row with the 5^{th} and 2^{nd} rows, we can observe a clear improvement attributed to both point-level pseudo labels and cluster-level pseudo labels, highlighting the complementary nature of these two types of labels introduced in our PCTeacher.

TABLE IV

EFFECT OF INDIVIDUAL MODULE. MVF DENOTES MULTI-VIEW FUSION AND HG DENOTES HYBRID GROUPING.

No.	MVF	HG	PCAT	mAP
1		✓	✓	29.3
2	✓		✓	30.1
3	✓	✓		28.7
4	✓	✓	✓	31.1

TABLE V

EFFECT OF POINT-CLUSTER AGREEMENT-BASED THRESHOLDING.

No.	Method	Threshold τ_c	mAP
1	One-threshold	0.7	28.7
2	Multi-threshold	0.7 & 0.9	30.7
3	PCAT	0.7	31.1

Ablation on Multi-View Fusion and Hybrid Grouping. Tab. IV assesses the impact of the multi-view fusion strategy and hybrid grouping by disabling them individually. In the 1^{st} row, we produce point-level pseudo labels from only one view, and in the 2^{nd} row, we disable the hybrid grouping by restricting the operation within the shifted coordinate. Comparing the 4^{th} row to the 1^{st} and 2^{nd} rows, we observe a performance increase of 1.8% and 1%, indicating the clear contributions of the multi-view fusion strategy and hybrid grouping in enhancing the quality of point-level pseudo labels and increasing the potential clusters, respectively.

Ablation on Point-Cluster Agreement-based Thresholding. In the 3^{rd} of Tab. IV, we replace PCAT with a fixed threshold to filter cluster-level pseudo labels. Comparing it with the final model in the last row demonstrates the superiority of our point-cluster agreement-based thresholding, which is adaptive and significantly outperforms the fixed threshold approach. In Tab. V, we further compare PCAT with a stronger baseline, *i.e.*, a multi-threshold strategy in row 2, which involves utilizing both a high threshold and a low threshold. Notably, even when compared to this strong baseline, our PCAT still outperforms it, meanwhile eliminating the need for additional threshold.

V. CONCLUSION

We have proposed PCTeacher, a novel end-to-end semi-supervised 3D instance segmentation framework that generates point-level and cluster-level pseudo labels online. We design multi-view fusion to enhance the accuracy and recall of point-level pseudo labels. Moreover, our hybrid grouping improves recall of cluster-level pseudo labels, and point-cluster agreement-based thresholding obtains reliable cluster-level pseudo labels. PCTeacher complements the strengths of both types of pseudo labels and outperforms the previous state-of-art methods under different label ratios on two benchmark datasets.

Acknowledgment. This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

REFERENCES

[1] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019.

[2] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.

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

[4] 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.

[5] 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.

[6] Ruihang Chu, Xiaoqing Ye, Zhengzhe Liu, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, and Jiaya Jia. Twist: Two-way inter-label self-training for semi-supervised 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1100–1109, 2022.

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] Jean Lahoud, Bernard Ghanem, Marc Pollefeys, and Martin R Oswald. 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2019.

[16] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[18] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[19] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.

[20] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[21] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6423–6432, 2021.

[22] Shuang Deng, Qiulei Dong, and Bo Liu. Scss-net: Superpoint constrained semi-supervised segmentation network for 3d indoor scenes. *arXiv preprint*, 2021.

[23] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020.

[24] Yucheng Han, Na Zhao, Weiling Chen, Keng Teck Ma, and Hanwang Zhang. Dual-perspective knowledge enrichment for semi-supervised 3d object detection. *AAAI*, 2024.

[25] 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.

[26] 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.

[27] 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.

[28] 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.