Weakly Supervised 3D Scene Segmentation with Region-Level Boundary Awareness and Instance Discrimination

Kangcheng Liu*[1], Yuzhi Zhao[2], Qiang Nie[3], Zhi Gao[4], and Ben M. Chen[1]

[1] The Chinese University of Hong Kong, China
[2] City University of Hong Kong, China
[3] Tencent Youtu Lab, China
[4] Wuhan University, China
kcliu@mae.cuhk.edu.hk, yzzhao2-c@my.cityu.edu.hk, qnie.cuhk@gmail.com, bmchen@mae.cuhk.edu.hk, gaozhinus@gmail.com

Abstract. Current state-of-the-art 3D scene understanding methods are merely designed in a full-supervised way. However, in the limited reconstruction cases, only limited 3D scenes can be reconstructed and annotated. We are in need of a framework that can concurrently be applied to 3D point cloud semantic segmentation and instance segmentation, particularly in circumstances where labels are rather scarce. The paper introduces an effective approach to tackle the 3D scene understanding problem when labeled scenes are limited. To leverage the boundary information, we propose a novel energy-based loss with boundary awareness benefiting from the region-level boundary labels predicted by the boundary prediction network. To encourage latent instance discrimination and guarantee efficiency, we propose the first unsupervised region-level semantic contrastive learning scheme for point clouds, which uses confident predictions of the network to discriminate the intermediate feature embeddings in multiple stages. In the limited reconstruction case, our proposed approach, termed WS3D, has pioneer performance on the large-scale ScanNet on semantic segmentation and instance segmentation. Also, our proposed WS3D achieves state-of-the-art performance on the other indoor and outdoor datasets S3DIS and SemanticKITTI.

Keywords: 3D Scene Understanding, Weakly-Supervised/Semi-Supervised Learning, Region-Level Contrast, Energy Function, Segmentation

1 Introduction

The 3D scene segmentation problem, which typically consists of two important downstream tasks: point cloud semantic segmentation and instance segmentation, becomes increasingly important recently with the wide deployment of 3D sensors, such as LiDAR and RGB-D cameras [3]. Point clouds are the raw sensor data obtained by 3D sensors and the most common 3D data representation for

[*] Kangcheng Liu is the corresponding author. Y. Zhao, Q. Nie are co-second authors.

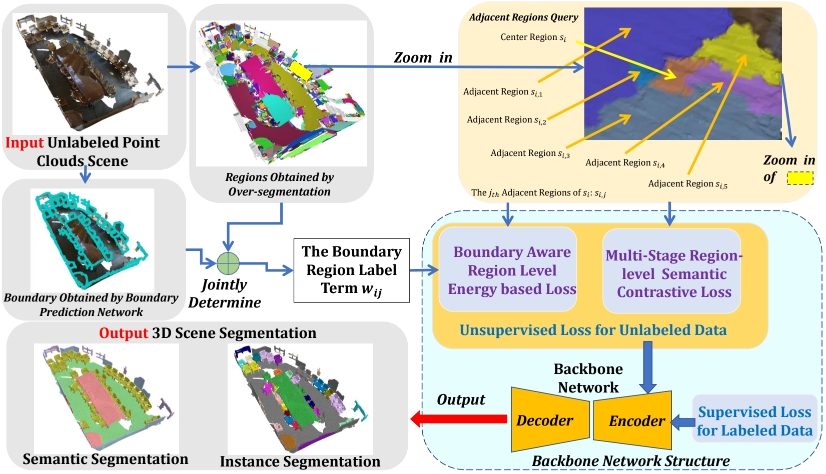
Input Unlabeled Point Clouds Scene
Regions Obtained by Over-segmentation
Zoom in
Adjacent Regions Query
Center Region s_i
Adjacent Region $s_{i,1}$
Adjacent Region $s_{i,2}$
Adjacent Region $s_{i,3}$
Adjacent Region $s_{i,4}$
Adjacent Region $s_{i,5}$
Zoom in of
The j_{th} Adjacent Regions of s_i: $s_{i,j}$
Boundary Obtained by Boundary Prediction Network
Jointly Determine
The Boundary Region Label Term w_{ij}
Boundary Aware Region Level Energy based Loss
Multi-Stage Region-level Semantic Contrastive Loss
Unsupervised Loss for Unlabeled Data
Backbone Network
Output
Decoder
Encoder
Supervised Loss for Labeled Data
Backbone Network Structure
Output 3D Scene Segmentation
Semantic Segmentation
Instance Segmentation

Fig. 1. The illustration of the **overall framework** of our proposed **WS3D**.

3D scene understanding in robotics and autonomous driving. However, the majority of point cloud understanding methods rely on heavy annotations [4, 7, 15]. Annotations of 3D point cloud requires a large amount of time and huge labours. For instance, it requires approximately half an hour per scene for ScanNet [10] or S3DIS [2] with even thousands of scenes. Though existing point cloud understanding methods [4, 7, 15] have achieved good results on these datasets, it is difficult to directly extend them to new scenes when high-quality labels are scarce. And limited number of scenes can be reconstructed in reality [18]. Therefore, developing methods that can be trained with limited 3D labels in complex scenes, termed as weakly supervised learning (WSL)-based or semi-supervised learning (SSL)-based 3D point cloud understanding, becomes in high demand. Recently, motivated by the success of WSL in images [54], many works start to tackle WSL with fewer labels in 3D, but great challenges remain. The challenges involve the meaningful information loss when 3D scenes are transformed to image [50], reliance on fully supervised image segmentation [50], sophisticated pre-processing and pre-training [55], customized labeling strategy and reliance on scene class labels for sub-clouds [52], lack of relationship mining both in low-level geometry and high-level semantics [18]. Hence, developing an effective 3D WSL framework to effectively exploit the information in limited 3D labeled data for the scene segmentation task becomes extremely important.

Weakly supervised image semantic and instance segmentation is a vehement research focus in recent years. Some simple but effective methods have been proposed for WSL-based semantic understanding such as contrastive learning [19,55] and conditional random field (CRF) [6, 43]. However, there still exist four main challenging unsolved issues. **Firstly**, the widely adopted energy-function-based conditional random field segmentation [6] relies on handcrafted feature similarities, and does not consider the boundary information. It attaches equal importance to pixels on the semantic boundary and within the same semantic object, which can cause vague and inaccurate predictions in pixel-level segmentation at the object boundaries. And how to leverage boundary information has been explored in 2D but rarely explored in 3D WSL. **Secondly**, the computation costs are both very high when applying pixel-level contrastive learning or pixel-level energy-based segmentation in a high-resolution image for every pixel pair. Furthermore, the large-scale point cloud scenes even contain billions of points, making the point-level contrastive learning intractable. **Thirdly**, the existing unsupervised contrastive-learning-based pre-training for point clouds [18, 55] only regards the geometrically registered point cloud pairs as positive samples, while does not take their important correlated semantics into consideration. **Finally**, although existing state-of-the-art detection and segmentation methods [31, 63] succeed in using multi-level feature representations in 2D, it remains challenging to design efficient 3D multi-stage contrastive learning strategies to establish more distinctive feature representations at each stage of the feature pyramid.

As depicted in Fig. 1, we propose a unified **WS3D** framework which simultaneously solves the 3D semantic segmentation and instance segmentation. We firstly use the oversegmentation [46] to obtain regions, and use a boundary prediction network as an intermediate tool to obtain boundary region labels. Then, high-confidence boundary region labels serve as guidance for our proposed unsupervised region-level energy-based loss. Meanwhile, we propose an unsupervised multi-stage region-level confidence-guided contrastive loss to enhance instance discrimination. Combined with supervised loss, complete 3D scene segmentation is achieved. Specifically, our **WS3D** includes two novel designs to address the challenging issues mentioned above and to enhance the performance. Firstly, to encourage latent instance discrimination and to guarantee efficiency, an efficient region-level contrastive learning strategy is proposed to guide network training at multiple stages, which realizes unsupervised instance discrimination. Also, to leverage boundary information as labels for semantic divisions, an energy-based loss with guidance from the semantic boundary regions is proposed to make the maximum utilization of the unlabeled data in network training.

The main contributions of our work are highlighted as follows:

1. We propose an unsupervised region-level energy-based loss to achieve region-level boundary awareness, which utilizes boundaries as additional information to assist the 3D scene segmentation.
2. We propose the first unsupervised region-level semantic contrastive learning strategy for multi-stage feature discrimination. The energy-based loss and the contrastive loss are jointly optimized for the segmentation network in a complementary manner to make full use of the unlabeled data.
3. We propose the first weakly supervised framework that can be simultaneously applied for 3D semantic segmentation and instance segmentation. We conduct a lot of experiments on ScanNet and other indoor/outdoor benchmarks such as S3DIS and SemanticKITTI with different annotation ratios. It is demonstrated State-of-the-art performance has been attained.

2 Related Work

Machine Learning for 3D Scene Understanding. Point cloud processing has become increasingly important in robotic control and scene understanding applications [33–38]. Deep-learning-based approaches are commonly selected for the downstream high-level tasks of 3D scene understanding. The deep-learning-based point cloud processing approaches can be roughly divided into voxelization-based approaches [8, 42, 45, 47, 59], transformation-based approaches [13, 14, 27, 30,53,56], and point-based approaches [1,12,23,28,29,35,36,39,41,58,60,62]. The typical point-based method is the superpoint-graph [28] proposing graph-based deep metric learning for point clouds oversegmentation, which has inspired our work. Different from them, we use the oversegmentation result as the intermediate tool to obtain the boundary region labels. Typical voxel-based method is the Sparseconv [16]. We use it as the backbone network in the task of semantic segmentation because of its high performance in inferring 3D semantics.

Pre-training for the Point Clouds Understanding. Many recent works proposed to pre-train networks on source datasets with auxiliary tasks such as the low-level point cloud geometric registration [55], the local structure prediction [48], the completion of the occluded point clouds [51], and the high-level supervised point cloud semantic segmentation [11], with effective learning strategies such as contrastive learning [55] and generative models [11]. Then, they finetuned the weights of the trained networks for the target 3D understanding tasks to boost performance on the target dataset. However, two major challenges still exist. Firstly, all the mentioned approaches depended on high-quality full annotations, which are hard to obtain for large-scale 3D scenes. Secondly, the large-scale pre-training requires a huge number of computational resources even for image understanding tasks [65]. Thus, the pre-training for large-scale point clouds understanding is hard to put into practice. The unsupervised pre-training [18] showed great capacity in unleashing the potential of a large amount of training data to serve for complicated tasks, e.g., instance segmentation. But merely utilizing unsupervised pre-training cannot explicitly make the utilization of unlabeled scenes, which results in unsatisfactory performance. Unlike previous methods benefiting a lot from pre-training, our proposed approach is trained in an end-to-end manner without pre-training.

WSL for 3D Semantic/Instance Segmentation. A large number of recent works focused on the task of 3D semantic and instance segmentation [21, 25, 30, 44] with full labels. However, applying current State-of-the-art methods (SOTAs) in a direct way for training often results in a great decrease in performance [20] for WSL, if the percentage of labeled data drops to a certain value, e.g., less than 30%. Recently, many works started to focus more on point cloud semantic segmentation with partially labeled data. Wang et al. [50] chose to transform 3D point clouds representations to 2D images, but pixel-level semantic segmentation labels are still in need during network training. Sub-cloud-level labels [52] needed additional large amount of labor to divide the whole scene into point cloud sub-scenes and to annotate the divided 3D scans into diverse categories. The iterative self-training approaches [40] took advantage of designing a sequential learning

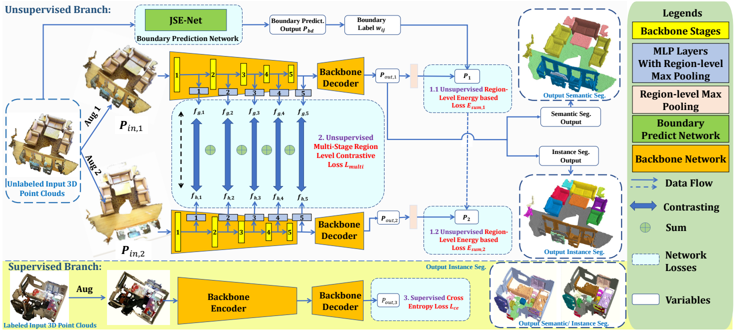

Fig. 2. **WS3D** architecture overview. **WS3D** consists of three modules: 1. **Unsupervised** region-level energy-based optimization guided by boundary labels; 2. **Unsupervised** multi-stage region-level contrastive learning with high confidence; 3. **Supervised** region-level semantic contrastive learning with labeled data. The backbone network adopts the encoder-decoder structures. The weights of the backbone network are shared in the supervised and unsupervised branch. Our framework can be integrated seamlessly to any off-the-shelf point-based or voxel-based backbones.

strategy, which was made up of two steps to provide pseudo supervisions from limited annotations. However, the OTOC merely works for the task of semantic segmentation. And it can not be easily generalized to more complicated tasks such as instance segmentation. Xu et al. [57] designed a learning framework that merely relies on a small portion of points to be annotated during training. It was designed to approximate the gradient during the learning process, where the auxiliary 3D spatial constraints and color-level evenness were also considered in the network optimizations. However, the approach was restricted to the object part segmentation, and it is difficult to annotate points in a well-proportioned and homogeneous as required. Like Xu et al. [57], we interchangeably use the terms weakly-supervised and semi-supervised for the limited reconstruction cases in this work. In summary, weakly/semi-supervised 3D semantic and instance segmentation are far from mature. More effective methods should be proposed to extract meaningful information from the unlabeled data when limited 3D scenes can be labeled.

3 Proposed Methodology

We propose a general **WS3D** framework to tackle weakly supervised 3D understanding with limited labels, as shown in Fig. 2. We choose different backbone networks for semantic and instance segmentation tasks. For semantic segmentation, we choose the effective backbone Sparseconv [16]. For instance segmentation, our backbone and point clustering procedure follow widely-used PointGroup [25]. **WS3D** consists of three modules for the network optimization: **1**. Unsupervised energy-based loss guided by boundary awareness and highly confident predictions for unlabeled data, which is discussed in Subsection 3.1; **2**.

Unsupervised multi-stage region-level contrastive learning with highly confident predictions for unlabeled data, which is discussed in Subsection 3.2. **3**. Supervised semantic contrastive learning for labeled data, which is discussed in Subsection 3.3. The three modules are integrated jointly into the optimization function for network training to accomplish the semantic or instance segmentation task.

3.1 Unsupervised Region-level Boundary Awareness

Energy-function-based conditional random field segmentation was proposed in [6] and has been widely applied. However, it works in a fully supervised manner and does not consider the semantic boundary information, which is a great indication of semantic partitions in point clouds scenes. In this Subsection, we develop a boundary-aware energy-based loss for unsupervised learning. As shown in Fig. 1, to obtain robust boundaries for unlabeled 3D points, we first perform point cloud oversegmentation [46], and also extract boundary points using an off-the shelf semantic boundary prediction network, which are both subsequently used as the conditions to define boundary regions for 3D points. Then, we propose a region-level energy-based loss based on obtained boundary region labels.

Point Clouds Oversegmentation. To obtain boundary regions, and facilitate subsequent region-level affinity computation and region-level contrastive learning, we first perform a region-level coarse clustering based on point cloud oversegmentation. The previous method depended on the region growing [46] to do oversegmentation, which relied heavily on the accurate normal estimation and were easily influenced by noises. In our work, we choose to use normal, curvature to provide the initial oversegmentation. And the oversegmentation results are shown in Fig. 1 and 4. Denote original point clouds as P_{in}. After oversegmentation, they are partitioned into Q subregions $S = \{s_1, s_2, ..., s_q\}$, where $s_i \cap s_k = \varnothing$ for any $s_i \neq s_k$ as shown in Fig. 1 and 4.

Boundary Points Extraction. As shown in Fig. 1, in addition to the oversegmentation results, we extract the semantic boundary points to further identify boundary regions. The semantic boundary often indicates the distinguishment between various semantic classes. We extract semantic boundary points by JSENet [22], as shown in Fig. 2. As for training, we first define semantic boundary points from the limited labeled scenes as ground truth. With definition of the ground truth boundary points, we design the loss following JSENet except substituting the binary cross entropy loss L_{bce} with the focal loss L_{foc} [32] to tackle the large class imbalance between the boundary points and non-boundary points. L_{foc} is as follows:

$$L_{foc} = -\frac{1}{N_i}\sum_{i=1}^{N_i}(1-b_i)^{\alpha} b_i^{gt} log(b_i) + (b_i)^{\alpha}(1-b_i^{gt}) log(1-b_i), \qquad (1)$$

where b_i denotes the binary predicted boundary map and b_i^{gt} denotes the ground truth boundary map. N_i is the total number of input points for training. We select α=2 based on the original design [32]. After its convergence, we apply the trained network to the remaining unlabeled scenes to obtain their boundary points. Examples of predicted boundary points of ScanNet [10] are shown in Fig. 4, which clearly reveal distinctions between diverse semantic classes.

Boundary Labels. After extracting semantic boundary points, we utilize them as labels of discrimination between diverse semantic categories. As shown in Fig. 1, denote the j_{th} adjacent regions of the center region s_i as $s_{i,j}$. The adjacent region query is realized by fast Octree-based K-nearest neighbour search [5]. Then, we determine the two adjacent regions as boundary regions if both s_i and $s_{i,j}$ contain boundary points. The label for boundary region $w_{i,j}$ is designed as:

$$w_{i,j} = \begin{cases} 1 & \textit{if } s_i,\ s_{i,j} \textit{ both contain boundary points;} \\ 0 & \textit{otherwise.} \end{cases} \tag{2}$$

The label $w_{i,j}$ denotes semantic boundaries of adjacent regions, which is then used to guide the optimization of the energy function for segmentation.

Energy Loss Guided by Boundary Labels. As shown in Fig. 2, we first perform data augmentation (detailed in the Appendix) for the input point clouds $\mathbf{P}_{in}$ to obtain two transformed point clouds $\mathbf{P}_{in,1}$ and $\mathbf{P}_{in,2} \in \mathbb{R}^{N \times C_{in}}$, where N is the numbers of points. C_{in} and C_{out} are the numbers of input and output feature channels, respectively. Utilizing the backbone network Sparseconv [16], we can obtain point cloud predictions $\mathbf{P}_{out,1}$ and $\mathbf{P}_{out,2}$. Then, applying region-level max pooling on the same subregions, we obtain the predicted classes $\mathbf{P}_1$, and $\mathbf{P}_2 \in \mathbb{R}^{M \times C_{out}}$ of the specific subregions. $\mathbf{P} = \{p(s_1), p(s_2), ...p(s_i), ..., p(s_R)\}$, where R is the total number of regions obtained by oversegmentation. Denote the prediction of the j_{th} neighbouring region of center region s_i as $p(s_{i,j})$. Taking the unary network prediction and pairwise affinity between the neighbouring region into account, inspired by conditional random field (CRF) in DeepLab [6], we formulate the optimization energy function E_{sum} as follows:

$$E_{sum} = \sum_i E_i(s_i) + \sum_{i<j}^{adjacent} E_{i,j}(s_i, s_{i,j}). \tag{3}$$

The first unary network prediction item $E_i(s_i)$ is the entropy regularization term. It encourages region-level prediction with high confidence, which also facilitates the contrastive learning introduced subsequently. It is formulated as:

$$E_i(s_i) = -log\, p(s_i). \tag{4}$$

We propose the pairwise affinity term of E_{sum} as:

$$\begin{aligned} E_{i,j}(s_i, s_{i,j}) = {} & H_{i,j} w_{i,j} [\epsilon - \|p(s_i) - p(s_{i,j})\|]_+^2 \\ & + H_{i,j}(1 - w_{i,j}) \|p(s_i) - p(s_{i,j})\|^2. \end{aligned} \tag{5}$$

The $H_{i,j}$ is the confidence indicator. $H_{i,j} = 1$, if the probabilities which produce $p(s_i)$ and $p(s_{i,j})$ are both larger than a threshold γ. Otherwise, it equals 0. ϵ can be any value in the range of $(0, 1)$. $p(s_i)$ and $p(s_{i,j})$ are the semantic predictions for the i_{th} center region and the j_{th} neighbouring region, respectively. And

$[x]_+$ is the maximum function $max(0, x)$. For adjacent boundary regions, we encourage their confident semantic or instance predictions to be different (i.e., larger $\|p(s_i) - p(s_{i,j})\|$); while for non-boundary adjacent regions, we force their semantic predictions to be the same (i.e., $\|p(s_i) - p(s_{i,j})\| = 0$). Different from the traditional energy function in DeepLab [6], which used handcrafted features to compare similarities, we propose to use the learned boundary region labels to guide the network's confident max-pooled region-level predictions. Therefore, the proposed boundary-aware energy function better encourages semantic separations at boundaries. Furthermore, we only consider pairwise affinity between adjacent regions instead of all pixel pairs, which greatly reduces computation costs, and avoids noises induced by distant unrelated pairs in the meanwhile.

3.2 Unsupervised Region-level Instance Discrimination

After applying entropy regularization term $E_i(s_i)$, we can obtain region-level predictions with high confidence. Note that confident region-level predictions further improve the latent feature discrimination capacity of the network, which makes contrastive learning in the latent space feasible. Therefore, we further propose **a multi-stage region-level contrastive learning** for unlabeled data. Compared with previous work only using contrastive learning with low-level geometric registrations [55], our work unleashes potentials of contrastive learning with instance discrimination to enhance distinct feature learning in latent space. The key of semantic/instance segmentation is to maintain discriminative feature representations at different stages of the backbone network [63]. Inspired by the feature pyramid network [31, 63], we propose a simple but effective multi-stage contrastive learning approach for point clouds in an unsupervised setting. As shown in Fig. 2, given the input augmented point clouds $P_{in,1}$ and $P_{in,2}$, we feed them into the backbone encoder. We add five additional MLP heads with region-level max-pooling to obtain region-level segmentation predictions at the m_{th} backbone stage, denoted as $\mathbf{f}_{g,m}$ and $\mathbf{f}_{h,m}$, respectively (five stages in our case, denote M as the total network stages. i.e., $M = 5$). After we apply the MLP heads to the extracted features at different stages, we can obtain the hierarchical feature embeddings. Unlike existing pixel-level [9] (or point-level for point clouds) contrastive learning, our proposed contrastive learning performs on the region level. The region-level semantic contrastive loss is formulated as:

$$L^m_{contrast} = -\frac{1}{\mathbf{S}_p} \sum_{(a,b)\in S_p} log \frac{H_{i,j} exp(\mathbf{f}^a_{g,m} \cdot \mathbf{f}^+_{h,m}/\tau)}{\sum_{(\cdot,c)\in S_p} H_{i,j} exp(\mathbf{f}^a_{g,m} \cdot \mathbf{f}^-_{h,m}/\tau))}, \tag{6}$$

where $(a, b) \in \mathbf{S}_p$ are latent confident predicted positive region pairs, and $(a, c) \in \mathbf{S}_p$ are latent negative region pairs. As mentioned before, $H_{i,j}$ is designed for eliminating contrastive learning candidates with low confidence degrees. Reliable region-level contrastive learning is only applied to confident predictions by the network. Note that although a recent work **GPC** [24] proposed methods to perform contrastive learning on the point clouds in a SSL manner, our work is different from their method in two aspects. Firstly, our contrastive learning is conducted at region level while **GPC** conducted contrastive learning at point level. Secondly, **GPC** focused on the selection of the positive and negative point-set samples to perform contrastive learning in a pseudo-label supervised manner on two different 3D scene samples, while we focus on unsupervised contrastive learning which disentangles different feature representations in latent spaces on the same augmented 3D scene sample, guided by confident network predictions. The final proposed multi-stage contrastive learning loss is formulated as the sum of losses at every network stage:

$$L_{multi} = \sum_{m=1}^{M} L_{contrast}^{m}. \tag{7}$$

After applying the multi-stage contrastive loss L_{multi}, the output at each stage of the network will provide more distinctive representations to attain a better performance. From our ablation experiments, the performance can be boosted by applying multi-stage contrastive loss. Combining proposed loss items $E_{sum,1}$, $E_{sum,2}$ (see Eq. 3) for the two augmented scenes $P_{in,1}$, $P_{in,2}$, and L_{multi} (see Eq. 7), we formulate the overall loss $L_{unlabeled}$ for the **WS3D** training with unlabeled data: $L_{unlabeled} = E_{sum,1} + E_{sum,2} + L_{multi}$.

3.3 Supervised Learning for Labeled Data

We also guide the network optimization by using supervision from the labeled data. As shown in Fig. 2, we use the cross-entropy loss L_{ce} to guide the supervised learning on the labeled data in the supervised branch. The loss term for the **WS3D** training with the labeled data is $L_{labeled} = L_{ce}$.

3.4 The Overall Optimization Loss Function

Leveraging our proposed region-level energy-based loss and region-level constrastive learning, the network can make use of the unlabeled data for better feature learning to boost performance. As shown in Fig. 2, for semantic segmentation and instance segmentation, we train the network in an end-to-end manner for both supervised and unsupervised branches to make full use of labeled and unlabeled data. The overall optimization function L_{total} is formulated as follows:

$$L_{total} = L_{labeled} + L_{unlabeled}. \tag{8}$$

4 Experiments

4.1 Experimental Settings

Datasets. To demonstrate the effectiveness of our proposed **WS3D** for WSL under the limited reconstruction labeling scheme, we have tested it on various of benchmarks, including S3DIS [2], ScanNet [17], and SemanticKITTI [4] for semantic segmentation, and ScanNet [17] for instance segmentation, respectively.

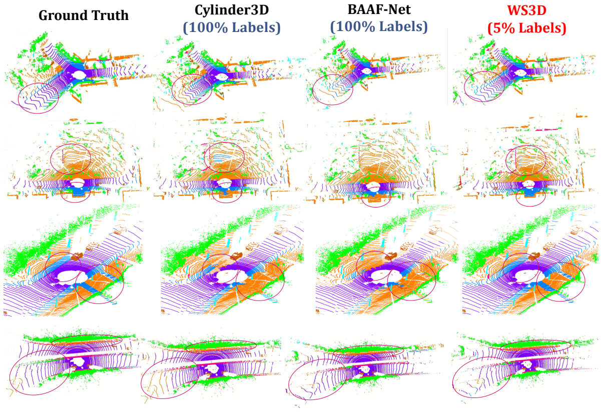

Fig. 3. Qualitative **semantic segmentation** results of proposed **WS3D** on SemanticKITTI Val. Set with 5% labeling percentage, compared with fully supervised arts Cylinder3D [64], and BAAF-Net [44] with semantics indicated by different colors. The red circles highlight the performance difference between diverse methods.

The detailed information of each dataset is put in the Appendix.

Training Set Partition. Following the typical setting in data-efficient learning in the limited reconstruction case [18, 24], we partition the training set of all tested datasets into labeled data and unlabeled data with various of labeling points proportion, e.g., {1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 100%}. For the limited reconstruction case, noted that to partition the labeled points into a specific labeling ratio, we probably need to split a maximum of one scene into two sub-scenes. One of the sub-scenes belongs to the labeled data and the other sub-scenes belong to the unlabeled data.

Implementation Details. For the task of semantic segmentation, we train the network for 500 epochs on a single NVIDIA 1080Ti GPU with a batch size of 16 during training. The initial learning rate is 1×10^{-3} and is multiplied with 0.2 every 50 epochs. We implement it by *PyTorch* and optimize it with Adam optimizer [26]. We set the hyperparameter γ as 0.8 to ensure merely highly confident prediction can be used for the network optimization. ϵ is set to 0.5. For the instance segmentation, we train the network for 580 epochs on a single NVIDIA 1080Ti GPU with a batch size of 8 during training. The other settings are the same as the semantic segmentation task.

4.2 WSL-based 3D Semantic Segmentation

Overall Experimental Results. For semantic segmentation, we tested **WS3D** on various indoor and outdoor benchmarks, including ScanNet [10], S3DIS [2], and SemanticKITTI [4]. We have done experiments with limited labeled data, e.g., only {1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 100%} data in the training set are used as labeled data. As mentioned, we have used the voxel-based method

Table 1. Comparison of **semantic segmentation** results with different labeling percentages on ScanNet val. set, and S3DIS val. set (Area 5), and SemanticKITTI val. set. 'Sup-only-GPC' denotes **GPC** model trained with only labeled data. '**WS3D**' denotes model trained with our proposed methods. We have shown the performance increase in the last row for each dataset, compared to merely trained models with labeled data (the left value) and to the SOTAs **GPC** [24] (the right value).

Datasets	Network Model	Semantic Segmentation mIOU (%) on the Validation Set According to Supervision Level (%)								
		1%	5%	10%	15%	20%	25%	30%	40%	100%
ScanNet	Sup-only-GPC	40.9	48.1	57.2	61.3	64.0	65.3	67.1	68.8	72.9
	GPC [24]	46.6	54.8	60.5	63.3	66.7	67.5	68.9	71.3	74.0
	WS3D	49.9	56.2	62.2	65.8	68.5	69.4	70.3	73.4	76.9
	↑	**+9.0/+3.3**	**+8.1/+1.4**	**+5.0/+1.7**	**+4.5/+2.5**	**+4.5/+1.8**	**+4.1/+1.9**	**+3.2/+1.4**	**+4.6/+2.1**	**+4.0/+2.9**
S3DIS	Sup-only-GPC	36.3	45.0	52.9	55.3	59.9	60.3	61.2	62.6	66.4
	GPC [24]	38.2	53.0	57.7	60.2	63.5	63.9	64.9	65.0	68.8
	WS3D	45.3	54.6	59.3	62.3	65.7	66.5	67.2	69.5	72.9
	↑	**+9.0/+7.1**	**+9.6/+1.6**	**+6.4/+1.6**	**+7.0/+2.1**	**+5.8/+2.2**	**+6.2/+2.6**	**+6.0/+2.3**	**+6.9/+4.5**	**+6.5/+4.1**
SemanticKITTI	Sup-only-GPC	28.6	34.8	43.9	47.9	53.8	55.1	55.4	57.4	65.0
	GPC [24]	34.7	41.8	49.9	53.1	58.8	59.1	59.4	59.9	65.8
	WS3D	38.9	43.7	52.3	55.5	61.4	61.7	62.1	63.2	66.9
	↑	**+10.3/+4.2**	**+8.9/+1.9**	**+8.4/+2.4**	**+7.6/+2.4**	**+7.6/+2.6**	**+6.6/+2.6**	**+6.7/+2.7**	**+5.8/+3.3**	**+1.9/+1.1**

Sparseconv [16] as the backbone. The qualitative results are shown in Fig. 3. And the quantitative semantic segmentation performance in terms of mIOU is summarized in Table 1. Our WSL model significantly surpasses the supervised-only model in **GPC** that is merely trained with labeled data, showing that our WSL can make use of the unlabeled data to enhance the feature discrimination capacity of the model. Also, it can be observed that compared with Sup-only-GPC models, the increment of performance is more obvious when the unlabeled data percentage is larger. For example, the performance increase on SemanticKITTI is 10.3% for the 1% labeling percentage, 5.8% for the 40% labeling percentage, and 1.9% for the 100% labeling percentage. This can be possibly explained by the fact that for more unlabeled data, our proposed **WS3D** can extract more meaningful semantic information from the unlabeled data based on our boundary-guided energy-based loss and confidence-guided region-level contrastive learning design. In addition, compared with current SOTAs **GPC**, our proposed **WS3D** also achieves consistently better results in semantic segmentation performance, especially when faced with very limited label circumstances (e.g., 1% labeling points). In that case, **WS3D** outperforms **GPC** by 3.3%, 7.1%, and 4.2% for ScanNet, S3DIS, and SemanticKITTI, respectively. Fig. 3 shows that we can provide comparable performance compared with fully supervised SOTAs BAAF-Net [44] and Cylinder3D [64] on SemanticKITTI with 5% labels.

Comparisons with SOTAs in the fully supervised mode. To demonstrate the feature learning capacity of our proposed **WS3D**, we have also experimented with the 100% labeling percentage for a fair comparison with fully supervised SOTAs. The results are shown in Table 2 and the last column of Table 1. We have fed the whole training set into the supervised branch and unsupervised branch in Fig. 2 simultaneously. Therefore, our proposed energy-based loss and region-level contrastive learning strategy operate as additional optimization guidance for the network training. Table 2 demonstrates that we can realize at least comparable or even better results, compared with fully supervised SOTAs.

Transductive learning. Similar to the experimental setting of [24], we have

Table 2. Comparison of SOTAs methods in the semantic segmentation performance on ScanNet validation set, S3DIS validation set (Area 5), and on SemanticKITTI validation set and test set. All results are based on the 100% label ratio. Top-two results are highlighted.

Approaches	Venue	ScanNet	S3DIS	SemanticKITTI	
		Val.	Val. (Area 5)	Val.	Test
Minkow-Network [8]	CVPR19	72.2	65.4	61.1	63.1
PointASNL [58]	CVPR20	66.4	62.6	-	58.8
KPConv [49]	ICCV19	69.2	67.1	-	46.8
SPV-NAS [47]	ECCV20	-	-	64.7	66.4
Fusion-Net [61]	ECCV20	-	67.2	63.7	61.3
MV-Fusion [27]	ECCV20	76.4	65.4	-	-
Cylinder3D [64]	CVPR21	-	-	**65.9**	**67.8**
BAAF-Net [44]	CVPR21	-	**72.1**	61.2	59.9
Sup-only-GPC	-	72.9	66.4	65.0	65.4
GPC	ICCV21	**74.0**	68.8	65.8	67.7
WS3D	-	**76.9**	**72.9**	**66.9**	**69.0**

Table 3. Comparison of experimental results on 20% and fully labeled case for the task of inductive and transductive learning, respectively. In transductive learning, the test set is also utilized for network training. We test on the task of semantic segmentation on ScanNet, S3DIS, and SemanticKITTI with the evaluation metric of mIOU(%).

Datasets	20% label			100% label		
	Base	Induct.	Transduct.	Base	Induct.	Transduct.
ScanNet Val.	64.0	68.5	71.4	72.9	76.9	77.6
S3DIS Area5 Val.	59.9	65.7	66.6	66.4	72.9	73.5
Semantic KITTI Val.	53.8	61.4	64.5	65.0	66.9	68.2
Semantic KITTI Test.	55.7	62.5	63.6	65.4	68.1	71.3

also conducted experiments evaluating the performance of **WS3D** in transductive learning. Different from inductive learning we tested above that requires the trained model to be generalized to an unseen test set, transductive learning can exploit the testing set when training. Compared with inductive learning, we add the test set as part of the unlabeled data in transductive learning. As is demonstrated in Table 3, the sem. seg. mIOU becomes higher if the network is learned in a transductive way, both for the fully labeled case with 100% labels and the weakly labeled case with 20% labels. It demonstrates that our proposed WSL approaches, including the energy-based boundary-aware loss and the region-level contrastive learning, can leverage the unlabeled data for feature learning effectively in an implicit way to enhance the final segmentation performance.

4.3 WSL-based 3D Instance Segmentation

As our method can be integrated seamlessly into various network backbones and applied to different highly-level understanding tasks, we have also integrated our method with Point-Group [25] for the instance segmentation on ScanNet with results shown in Table 4. Noticed that the performance increase is 21.7% when merely 1% data is labeled compared with the sup-only case. It further demonstrates that our proposed approaches for the unsupervised branch have effectively exploited the unlabeled data to improve the feature learning capacity of the model. As shown in Fig. 4, our proposed approach can provide explicit boundary guidance for separating diverse semantic classes, and the instance segmentation performance is comparable to those fully supervised counterparts.

4.4 Ablation Study

Ablations: In this Subsection, to analyze the significance and demonstrate the effectiveness of various components in **WS3D**, we have done comprehensive

Table 4. Comparison of the performance of **instance segmentation**, under various levels of supervision on ScanNet validation set. 'Sup-only-GPC' denotes the model trained with only labeled data. '**WS3D**' denotes the model trained with our proposed methods. We have shown the performance increase of **WS3D** in the last row.

Tested Dataset	Network Model	Ins. Seg. Results with the metric of AP@50%								
		1%	5%	10%	15%	20%	30%	35%	40%	100%
ScanNet	Sup-only-GPC	10.8	33.6	42.8	45.3	48.2	49.0	49.5	50.2	56.8
	WS3D	32.5	45.6	49.2	51.1	51.3	51.9	52.5	53.0	58.7
	↑	**+21.7**	**+12.0**	**+6.4**	**+5.8**	**+3.1**	**+2.9**	**+3.0**	**+2.8**	**+1.9**

Table 5. WS3D ablation studies on ScanNet (Left Value) and S3DIS (Right Value) Val. Set, for semantic segmentation (Metric: mIOU%) and on ScanNet Val. set for instance segmentation (Metric: AP@50%), both tested with the 5% labeled case.

Cases	Base	$w_{i,j}$	$H_{i,j}$ in EF	$H_{i,j}$ in UCSL	UCSL	MS-UCSL	SCE	mIOU%	AP@50%
No. 1	✓	✓	✓	✓		✓	✓	**56.2 / 54.6**	**45.6**
No. 2	✓		✓	✓		✓	✓	51.0 / 49.3	39.9
No. 3	✓			✓		✓	✓	49.9 / 47.2	37.0
No. 4	✓	✓		✓		✓	✓	51.6 / 52.1	40.1
No. 5	✓	✓	✓			✓	✓	51.1 / 51.4	40.7
No. 6	✓	✓	✓	✓	✓		✓	52.5 / 50.9	42.2
No. 7	✓	✓	✓				✓	49.3 / 48.0	38.1
No. 8	✓	✓	✓	✓		✓		54.3 / 52.8	42.9
No. 9	✓						✓	48.1 / 45.0	34.8

ablations on ScanNet and S3DIS datasets for different network modules on both semantic segmentation and instance segmentation tasks. The final results are summarized in Table 5. We have ablated network modules in all combinations of settings as follows. Take the ScanNet instance segmentation at AP@50% as examples: **Case 1**: The full **WS3D**. **Case 2**: Removing the boundary prediction network, and not using the guidance of $w_{i,j}$. The framework still consists of the supervised branch, unsupervised guidance of the energy function based on the predicted confident pseudo label, and contrastive learning. This setting leads to a significant drop of 5.7% on AP. **Case 3**: Removing the pairwise term in the energy-based optimization function E_{sum}, the AP drops largely by 8.6%. **Case 4**: Removing $H_{i,j}$ in the energy function, the performance drops by 5.5%. **Case 5**: Removing $H_{i,j}$ in the unsupervised contrastive learning, the performance drops by 4.9%. **Case 6**: Conducting contrastive learning only with the region-level feature $\mathbf{f}_{g,5}$ and $\mathbf{f}_{h,5}$ at the fifth network stage, rather than at multiple stages. The performance drops by 3.4%. **Case 7**: Removing the unsupervised region-level contrastive learning branch, the performance drops largely by 7.5%. **Case 8**: Removing the supervised learning branch with the cross-entropy loss, the performance drops by 2.7%. **Case 9**: Only using the supervised branch, the ins. seg. performance drops significantly by 10.8%.

Analyses: From the above ablations, some important findings are summarized: **Firstly**, not using our designed modules results in a significant performance drop (Cases No. 3, No. 7, and No. 9), which demonstrates the effectiveness of the proposed unsupervised branch and learning strategies to leverage the unlabeled data. **Secondly**, our proposed learning strategies with boundary label $w_{i,j}$ (Case

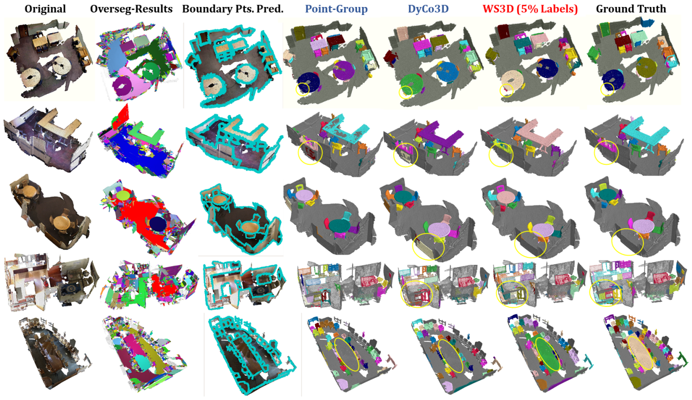

Fig. 4. Qualitative **instance segmentation** results of proposed **WS3D** on ScanNet with the 5% labeling ratio, compared with fully supervised arts, with instances indicated by different colors. And the intermediate oversegmentation results for obtaining regions and boundary predictions, with boundary points (pts.) indicated by blue.

No. 2), energy function design (No. 3), high-confidence prediction based energy function design (No. 4), high-confidence based region-level contrastive learning strategy (No. 5), multi-stage contrastive learning network design (No. 6), all have a boost on the overall semantic/instance segmentation performance. The results demonstrate that the proposed energy loss is significant for semantic/instance seg. performance, because semantic boundary labels are crucial for identifying diverse objects. **Thirdly**, removing the supervision (Case No. 8), our method still maintains the performance with a slight drop of performance by 2.7%. It further validates the robustness and feature learning capacity of our approach.

5 Conclusion

In summary, we propose a general **WS3D** framework for WSL-based 3D scene segmentation with SOTAs performance. We propose an unsupervised boundary-aware energy-based loss and a novel region-level multi-stage semantic contrastive learning strategy, which are complementary to each other to make the network learn more meaningful and discriminative features from the unlabeled data. The effectiveness of our approach is verified across three diverse large-scale 3D scene understanding benchmarks under various experiment circumstances.

Acknowledgments

This work is mainly supported by the Hong Kong PhD Fellowship Scheme awarded to Dr. Kangcheng Liu. This work is partially supported by the Hubei Province Natural Science Foundation (Grant No. 2021CFA088), and Wuhan University–Huawei Geoinformatics Innovation Laboratory.

References

1. Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y.: Spinnet: Learning a general surface descriptor for 3d point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11753–11762 (2021) 4
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1534–1543 (2016) 2, 9, 10
3. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Gall, J., Stachniss, C.: Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. The International Journal of Robotics Research (IJRR) p. 02783649211006735 (2021) 1
4. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 9297–9307 (2019) 2, 9, 10
5. Behley, J., Steinhage, V., Cremers, A.B.: Efficient radius neighbor search in three-dimensional point clouds. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 3625–3630. IEEE (2015) 7
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017) 2, 6, 7, 8
7. Cheng, R., Razani, R., Taghavi, E., Li, E., Liu, B.: 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12547–12556 (2021) 2
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3075–3084 (2019) 4, 12
9. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 715–724 (2021) 8
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5828–5839 (2017) 2, 7, 10
11. Eckart, B., Yuan, W., Liu, C., Kautz, J.: Self-supervised learning on 3d point clouds by learning discrete generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8248–8257 (2021) 4
12. Fan, H., Yu, X., Ding, Y., Yang, Y., Kankanhalli, M.: Pstnet: Point spatio-temporal convolution on point cloud sequences. In: International Conference on Learning Representations (ICLR) (2020) 4
13. Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y.: Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 264–272 (2018) 4

14. Gojcic, Z., Zhou, C., Wegner, J.D., Guibas, L.J., Birdal, T.: Learning multiview 3d point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1759–1769 (2020) 4
15. Gong, J., Xu, J., Tan, X., Song, H., Qu, Y., Xie, Y., Ma, L.: Omni-supervised point cloud segmentation via gradual receptive field component reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11673–11682 (2021) 2
16. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9224–9232 (2018) 4, 5, 7, 11
17. Hou, J., Dai, A., Niessner, M.: 3d-sis: 3d semantic instance segmentation of rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 9
18. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15587–15597 (2021) 2, 3, 4, 10
19. Hu, H., Cui, J., Wang, L.: Region-aware contrastive learning for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16291–16301 (2021) 2
20. Hu, Q., Yang, B., Fang, G., Guo, Y., Leonardis, A., Trigoni, N., Markham, A.: Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds with 1000x fewer labels. arXiv preprint arXiv:2104.04891 (2021) 4
21. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11108–11117 (2020) 4
22. Hu, Z., Zhen, M., Bai, X., Fu, H., Tai, C.l.: Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds. In: European Conference on Computer Vision (ECCV). pp. 222–239. Springer Nature (2020) 6
23. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4267–4276 (2021) 4
24. Jiang, L., Shi, S., Tian, Z., Lai, X., Liu, S., Fu, C.W., Jia, J.: Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6423–6432 (2021) 8, 10, 11
25. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4867–4876 (2020) 4, 5, 12
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
27. Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C.: Virtual multi-view fusion for 3d semantic segmentation. In: European Conference on Computer Vision. pp. 518–535. Springer (2020) 4, 12
28. Landrieu, L., Boussaha, M.: Point cloud oversegmentation with graph-structured deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7440–7449 (2019) 4

29. Lei, H., Akhtar, N., Mian, A.: Spherical kernel for efficient graph convolution on 3d point clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 4
30. Li, L., Zhu, S., Fu, H., Tan, P., Tai, C.L.: End-to-end learning local multi-view descriptors for 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1919–1928 (2020) 4
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017) 3, 8
32. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 6
33. Liu, K.: Robust industrial uav/ugv-based unsupervised domain adaptive crack recognitions with depth and edge awareness: From system and database constructions to real-site inspections. In: 30th ACM International Conference on Multimedia (ACM MM) (2022) 4
34. Liu, K., Gao, Z., Lin, F., Chen, B.M.: Fg-net: Fast large-scale lidar point cloudsunderstanding network leveraging correlatedfeature mining and geometric-aware modelling. arXiv preprint arXiv:2012.09439 (2020) 4
35. Liu, K., Gao, Z., Lin, F., Chen, B.M.: Fg-conv: Large-scale lidar point clouds understanding leveraging feature correlation mining and geometric-aware modeling. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 12896–12902. IEEE (2021) 4
36. Liu, K., Gao, Z., Lin, F., Chen, B.M.: Fg-net: A fast and accurate framework for large-scale lidar point cloud understanding. IEEE Transactions on Cybernetics (2022) 4
37. Liu, K., Han, X., Chen, B.M.: Deep learning based automatic crack detection and segmentation for unmanned aerial vehicle inspections. In: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 381–387. IEEE (2019) 4
38. Liu, K., Qu, Y., Kim, H.M., Song, H.: Avoiding frequency second dip in power unreserved control during wind power rotational speed recovery. IEEE transactions on power systems **33**(3), 3097–3106 (2017) 4
39. Liu, Y., Fan, B., Meng, G., Lu, J., Xiang, S., Pan, C.: Densepoint: Learning densely contextual representation for efficient point cloud processing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5239–5248 (2019) 4
40. Liu, Z., Qi, X., Fu, C.W.: One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1726–1736 (2021) 4
41. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. In: Advances in Neural Information Processing Systems (NIPS). pp. 965–975 (2019) 4
42. Noh, J., Lee, S., Ham, B.: Hvpr: Hybrid voxel-point representation for single-stage 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14605–14614 (2021) 4
43. Obukhov, A., Georgoulis, S., Dai, D., Van Gool, L.: Gated crf loss for weakly supervised semantic image segmentation. arXiv preprint arXiv:1906.04651 **6** (2019) 2

44. Qiu, S., Anwar, S., Barnes, N.: Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1757–1767 (2021) 4, 10, 11, 12
45. Que, Z., Lu, G., Xu, D.: Voxelcontext-net: An octree based framework for point cloud compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6042–6051 (2021) 4
46. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: 2011 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–4. IEEE (2011) 3, 6
47. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. arXiv preprint arXiv:2007.16100 (2020) 4, 12
48. Thabet, A., Alwassel, H., Ghanem, B.: Self-supervised learning of local features in 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 938–939 (2020) 4
49. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 6411–6420 (2019) 12
50. Wang, H., Rong, X., Yang, L., Feng, J., Xiao, J., Tian, Y.: Weakly supervised semantic segmentation in 3d graph-structured point clouds of wild scenes. arXiv preprint arXiv:2004.12498 (2020) 2, 4
51. Wang, H., Liu, Q., Yue, X., Lasenby, J., Kusner, M.J.: Unsupervised point cloud pre-training via occlusion completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9782–9792 (2021) 4
52. Wei, J., Lin, G., Yap, K.H., Hung, T.Y., Xie, L.: Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4384–4393 (2020) 2, 4
53. Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K.: Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4376–4382. IEEE (2019) 4
54. Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16765–16774 (2021) 2
55. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: European Conference on Computer Vision (ECCV). pp. 574–591. Springer (2020) 2, 3, 4, 8
56. Xu, C., Wu, B., Wang, Z., Zhan, W., Vajda, P., Keutzer, K., Tomizuka, M.: Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. arXiv preprint arXiv:2004.01803 (2020) 4
57. Xu, X., Lee, G.H.: Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13706–13715 (2020) 5
58. Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S.: Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5589–5598 (2020) 4, 12

59. Ye, M., Xu, S., Cao, T.: Hvnet: Hybrid voxel network for lidar based 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1631–1640 (2020) 4
60. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11784–11793 (2021) 4
61. Zhang, F., Fang, J., Wah, B.W., Torr, P.H.: Deep fusionnet for point cloud semantic segmentation. In: European Conference on Computer Vision (ECCV). pp. 644–663 (2020) 12
62. Zhang, Z., Hua, B.S., Yeung, S.K.: Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1607–1616 (2019) 4
63. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 2881–2890 (2017) 3, 8
64. Zhu, X., Zhou, H., Wang, T., Hong, F., Li, W., Ma, Y., Li, H., Yang, R., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 10, 11, 12
65. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020) 4