# General 3D Vision-Language Model With Fast Rendering and Pre-Training Vision-Language Alignment

Kangcheng Liu, *Member, IEEE*, Yong-Jin Liu, *Senior Member, IEEE*, and Baoquan Chen, *Fellow, IEEE*

***Abstract*—Current prevailing vision-language models have achieved remarkable progress in 3D scene understanding while trained in the closed-set setting and with full labels. The major bottleneck for the current robot 3D scene recognition approach for robotic applications is that these models do not have the capacity to recognize any unseen novel classes beyond the training categories in diverse real-world robot applications such as robot manipulation as well as robot navigation. In the meantime, current state-of-the-art 3D scene understanding approaches primarily require a large number of high-quality labels to train neural networks, which merely perform well in a fully supervised manner. Therefore, we are in urgent need of a framework that can simultaneously be applicable to both 3D point cloud segmentation and detection, particularly in the circumstances where the labels are rather scarce. This work presents a generalized and straightforward framework for dealing with 3D scene understanding when the labeled scenes are quite limited. To extract knowledge for novel categories from the pre-trained vision-language models, we propose a hierarchical feature-aligned pre-training and knowledge distillation strategy to extract and distill meaningful information from large-scale vision-language models, which helps benefit the open-vocabulary scene understanding tasks. To leverage the boundary information, we propose a novel energy-based loss with boundary awareness benefiting from the region-level boundary predictions. To encourage latent instance discrimination and to guarantee efficiency, we propose the unsupervised region-level semantic contrastive learning scheme for point clouds, using confident predictions of the neural network to discriminate the intermediate feature embeddings at multiple stages. In the limited reconstruction case, our proposed approach, termed *WS3D++*, ranks 1st on the large-scale ScanNet benchmark on both the task of semantic segmentation and instance segmentation. Also, our proposed *WS3D++* achieves state-of-the-art data-efficient learning performance on the other large-scale real-scene indoor and outdoor datasets S3DIS and SemanticKITTI. Extensive experiments with both indoor and outdoor scenes demonstrated the effectiveness of our approach in both data-efficient learning and open-world few-shot learning.**

***Index Terms*—3D scene understanding, data-efficient learning, region-level contrast, energy function, 3D vision-language model.**

## I. Introduction

THE typical 3D scene parsing problem, which usually encompasses several important downstream tasks: point cloud semantic segmentation, instance segmentation, and object detection, becomes increasingly important with the wide deployment of 3D sensors, such as LiDAR and RGB-D cameras [1], [2], [3], [4], [5], [6]. Point clouds are raw sensor data obtained from 3D sensors and the most simple and common 3D data representation for understanding 3D scenes of robot navigation, robot grasping, and manipulation tasks. Despite significant success in deep neural networks applied to 3D visual perception, two major challenges hinder the construction of more scalable visual perception systems in 3D worlds. One is the **closed-set assumption**, which means the model only performs well while recognizing the categories that appear in the training set and struggles in recognizing the novel unseen object categories or concepts. Another is the heavy **reliance on large amounts of high-quality labeled data**. Large-scale 3D scenes are very laborious to label, which also makes it very hard for deep network models to perform well with very limited annotations.

*Close-set assumption:* One of the major bottlenecks in scaling up visual perception systems is the poor generalization capacity while encountered with diverse novel semantic classes or severe domain shifts. To endow the model with the capacity for adapting the learned representation and make it conform to different data distributions as well as recognize diverse novel categories, pioneer researches such as CLIP [7], Flamingo [8], and Otter [9] have demonstrated the great potentials in learning well-aligned visual linguistic representation from large-scale image-text pairs on the Internet for improving the model generalization capacity. To this end, subsequent approaches have been proposed in establishing abundant vision-language associations for different visual recognition tasks including detection and segmentation using the large-scale vision-language model (VLM) [10], [11], [12]. The paired visual-linguistic feature representation can enable the recognition of a large number of novel objects or concepts with natural language supervision because the visual and the lexical language features are well-matched

Received 19 December 2023; revised 11 February 2025; accepted 15 April 2025. Date of publication 2 May 2025; date of current version 6 August 2025. This work was supported in part by the Natural Science Foundation of China under Grant 62461160309. The work of Kangcheng Liu was supported in part by the Natural Science Foundation of China under Grant 62403400 and in part by the National Oversea Excellent Young Scientist Fund. Recommended for acceptance by M. Sun. *(Corresponding author: Kangcheng Liu.)*

Kangcheng Liu was with the Division of Engineering and Applied Science, California Institute of Technology (Caltech), Pasadena, CA 91125 USA. He is now with the Department of Electrical and Information Engineering, Hunan University, Hunan 410012, China (e-mail: kcliu@hnu.edu.cn, kcliu@caltech.edu).

Yong-Jin Liu is with the BNRist, MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing 100190, China.

Baoquan Chen is with the National Key Lab of Genaral AI, Peking University, Beijing 100871, China, and also with the School of Artificial Intelligence, Peking University, Beijing 100871, China.

The code is at: WS3D++ Code link.

Digital Object Identifier 10.1109/TPAMI.2025.3566593

in their shared semantic feature space. Despite the remarkable performance achieved in developing diverse vision-language foundation models such as SAM [13] and SEEM [14] for image-based scene understanding, it remains very difficult for CLIP [7] to benefit downstream 3D scene understanding because it is difficult to raise the feature dimension to 3D and establish explicit correlations or find clear alignments between large-scale scene/object-level 3D point clouds as well as linguistic semantic concepts. Moreover, it is even harder to transfer the informative knowledge to various downstream 3D scene understanding tasks. These limitations severely restrict the scalability of VLM to handle diverse unseen 3D scenes containing diverse novel 3D object categories.

*Reliance on large-scale labeled data:* A critical prerequisite for fully exploiting the capacity of the fully supervised deep learning approaches is the accessibility to large-scale well-annotated high-quality training data. Most point cloud understanding methods rely on heavy annotations [15], [16], [17]. However, the annotation of large-scale 3D point cloud scenes is rather time-consuming and labor-intensive. For instance, it requires around thirty minutes to label a single scene for ScanNet [18] or S3DIS [19] with thousands of scenes. Though existing point cloud understanding methods [15], [16], [17] have achieved decent results on these datasets, it is difficult to directly extend them to novel scenes when the high-quality labeled data is scarce. In the meanwhile, it is often the case that a limited number of scenes can be reconstructed in real applications [20]. Therefore, developing methods that can be trained with very limited labeled scenes, termed data-efficient 3D scene understanding with limited scene-level annotation, becomes in high demand. Data-efficient semantic and instance segmentation [21], [22], [23] is a vehement research topic for image-level scene understanding. Some simple but successful methods have been proposed, such as contrastive learning [24], [25] which learns a meaningful and discriminative representation, and conditional random field (CRF) [26], [27] for pseudo label propagation. However, there still exist four main challenging unsolved issues while scaling up these approaches to 3D scene understanding. *First*, the widely adopted energy function-based conditional random field segmentation [27] relies on handcrafted feature similarities and does not consider explicit boundary information. It attaches equal importance to pixels on semantic boundaries and within same semantic objects, which can cause vague and inaccurate predictions in pixel-level segmentation at object boundaries. And how to leverage boundary information has been explored in 2D but rarely explored in 3D data efficient learning [28]. *Second*, the computation costs are both very high when applying point-level contrastive learning or point-level energy-based segmentation in a dense point cloud scene for every point pair [29], [30]. Furthermore, large-scale point cloud scenes even contain billions of points, making point-level contrastive learning intractable in computational costs. *Third*, the existing unsupervised contrastive learning-based pre-training for point clouds [20], [25], [31], [32] only considers geometrically registered point/voxel pairs as the positive samples, while it does not explicitly consider explicit regional information, let alone the hierarchical alignments.

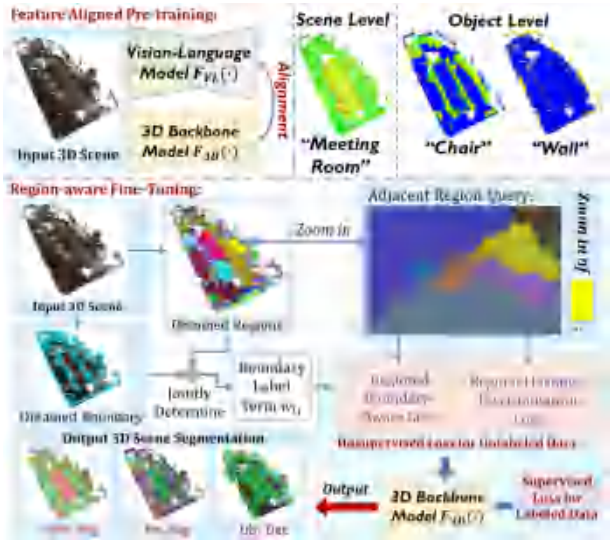


Fig. 1. The final overall illustrative diagram of our proposed *WS3D++*. We integrate language-3D feature associated pre-training and data-efficient fine-tuning as a general scene parsing vision-language model to achieve effective *data-efficient* as well as *open-vocabulary* 3D scene understanding for 3D scenes.

Driven by the above motivations in terms of both generalization capacity and data efficiency, we propose an effective two-stage framework, involving **unsupervised hierarchical vision-language pre-training** and **label-efficient fine-tuning** to boost the label/data-efficiency in 3D scene understanding. As shown in Fig. 1, *in the pre-training stage*, we leverage the rendering techniques to construct well-aligned 2D views for large-scale 3D scenes to establish more accurate coarse-to-fine vision-language associations. Then, we leverage the off-the-shelf object detector and the pre-trained large-scale vision-language model CLIP [7] to construct the *hierarchical feature representations* from both the global scene level to the local object level. We also propose an effective knowledge distillation strategy that distills the informative visual-language-aligned representation of the image encoder in CLIP [7] to the 3D backbone network. As is demonstrated by our extensive experiments, the open-vocabulary scene parsing performance can be also enhanced.

*During the fine-tuning stage*, as shown in Fig. 1, we propose a unified *WS3D++* framework that simultaneously solves the 3D scene understanding problem under the data-efficient setting. We first use the over-segmentation [33] to obtain regions and use a boundary prediction network (BPN) as an intermediate tool to obtain boundary region labels. Then, high-confidence boundary region labels serve as the guidance for our proposed region-level energy-based loss. Meanwhile, we propose a region-level confidence-guided contrastive loss to enhance instance discrimination. Specifically, our *WS3D++* includes two innovative designs to address the very challenging label scarcity issues and to enhance performance. Firstly, to encourage latent instance discrimination and to guarantee efficiency, an efficient region-level feature contrastive learning strategy is proposed to guide network training at multiple stages, which realizes the unsupervised instance discrimination. Also, to leverage boundary information as labels for the final semantic divisions, an

energy-based loss with guidance from the semantic boundary regions is proposed to take the maximized advantage of the unlabeled data in network training. Combined with supervised loss, the labeled data can also be leveraged to boost the final downstream 3D scene understanding performance.

*WS3D++* is a significant extension of the preliminary version of the conference work *WS3D* [4], where basic ideas of boundary awareness and contrastive instance discrimination are introduced to tackle the data-efficient 3D scene understanding during fine-tuning stage. In summary, we extensively enriched previous works in the following aspects:

*First*, we propose a generalized pre-training approach for data-efficient learning, which establishes accurate alignments between language and 3D point cloud in terms of both object-level and scene-level semantics in a hierarchical manner. *Second*, we propose to leverage the rendering technique that makes explicit associations between image and point cloud to facilitate 2D-to-3D matching and subsequent language-to-3D matching. *Third*, we visualized the language-queried activation maps directly on the 3D scenes, which demonstrates that the proposed approach learns better visual-linguistic alignment between the language descriptions and the visual object-level information. *Finally*, we evaluate our proposed approach comprehensively in diverse data-efficient and open-world learning settings for both the 3D semantic segmentation and 3D instance segmentation tasks.

The contributions of our work are highlighted as follows:

1) During the *pre-training stage*, we first propose an effective design which distills rich knowledge from the large-scale vision-language model into the 3D point cloud modality. Specifically, we propose leveraging rendering to obtain explicit scene-level and object-level 2D-3D feature associations, establishing a more accurate vision-language association hierarchically than the original CLIP encoder. We have demonstrated by extensive experiments that our proposed approach can realize superior compatibility with prevailing weakly supervised approaches.
2) During the *pre-training stage*, we first propose a global scene-to-sentence matching and then propose a local object-to-word matching approach, respectively, to establish the well-aligned vision-language feature representations at both the scene level and the object level, which largely facilitates the subsequent effective contrastive learning with the mostly matched visual-language contrastive pairs. The proposed designs have both enhanced the data-efficient learning and the knowledge-transfer capacity of the model, as demonstrated by our extensive experiments on both the 3D object detection and the 3D semantic/instance segmentation tasks.
3) During the *fine-tuning stage*, we propose a region-aware energy-based optimization approach to achieve the region-level boundary awareness, which utilizes the boundary as additional information to help assist the 3D scene segmentation and understanding. Furthermore, we propose the unsupervised region-level semantic contrastive learning strategy for multi-stage feature discriminations. The energy-based loss and the contrastive loss are jointly optimized for pre-training the backbone network in a complementary manner, which take full advantage of the unlabeled data.
4) Integrating the above two stages as a whole, we propose a unified framework termed *WS3D++*. State-of-the-art performance has been achieved by it with extensive experiments conducted on ScanNet [18] and other indoor/outdoor benchmarks such as S3DIS [19], SemanticKITTI [16] and NuScenes [36] in diverse experimental settings without bells and whistles. Finally, our proposed approach achieves pioneer performance on the very large-scale ScanNet [18] dataset in diverse downstream tasks of 3D scene understanding, including tasks among 3D semantic segmentation, 3D instance segmentation, and 3D object detection.[1]

To the best of our knowledge, this is the pioneer work which comprehensively evaluates across diverse 3D label-efficient scene understanding downstream tasks with our proposed 3D open-vocabulary recognition approach termed *WS3D++*. Our endeavor is orthogonal to the 3D backbone network designs and thus can be seamlessly integrated with the prevailing 3D point cloud detection or segmentation models. Our comprehensive results provide solid baselines for future researches in the data-efficient 3D scene understanding.

## II. Related Work

*Learning-based Point Cloud Understanding:* Deep-network-based approaches are widely adopted for point cloud understanding and the learning-based approaches have wide industrial applications. Fully supervised approaches can be roughly categorized into voxel-based [37], [38], [39], [40], [41], projection-based [42], [43], [44], [45], [46], [47], and point-based approaches [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59]. The voxel-based approaches [60], [61], [62] which are built upon SparseConv [60] and voxelize the point cloud for efficient processing have achieved remarkable performance in 3D scene parsing. Therefore, we use SparseConv [60] as our backbone architecture for downstream semantic understanding tasks because of its high performance in inferring 3D semantics.

*Pre-training for 3D Representation Learning:* Many recent works propose to pre-train networks on source datasets with auxiliary tasks such as low-level point cloud geometric registration [25], 3D local structural prediction [63], the completion of the occluded point clouds [64], and the foreground-background feature discrimination [30], with effective learning strategies such as contrastive learning [25] and masked generative modelling [65], [66]. Then the finetune the weights of the trained networks for the downstream target tasks to boost the robot scene parsing performances. However, several major challenges still exist. *First*, the above pre-training approaches all rely on the closed-set assumption, which means that the model can barely be transferred to recognize novel categories that do not appear within the training data. *Second*, the above methods require accessibility to the well-registered augmented point cloud [20], [25], [31], [32] to construct the pre-training contrast views,

[1] http://kaldir.vc.in.tum.de/scannet_benchmark/data_efficient/

which are very hard to obtain for large-scale 3D scenes. *Third*, a large number of computational power is required in the pre-training stage. Therefore, the designed pre-training approaches need to be very simple and lightweight, thus making it easier to directly transfer the model to large-scale point clouds.

Recently, with the development of large-scale vision-language models such as CLIP [7] and Flamingo [8], we can largely benefit the recognition capacity of 3D scene understanding models by distilling informative knowledge from large-scale vision-language models. For example, some pioneer works such as PointCLIP [67], [68] and ULIP [69] successfully transfer the knowledge from vision-language models to boost the downstream 3D shape classification tasks. The 3D CAD shapes are transferred to multi-view depth maps, thus they can be fed into the CLIP visual encoder and used the image representations paired with the corresponding point cloud as a bridge to obtain the correlations between the 3D and textual features. Moving beyond object-level recognition tasks, pioneer works also explore how to establish alignment among images, language, and 3D point cloud scenes for the task of open-vocabulary 3D scene understanding [70], [71], [72], which target localizing, detecting, and segmenting novel object categories that do not exist in the annotation. Compared with them, our proposed simple but effective framework can be both applicable to data-efficient learning and open-vocabulary scene understanding.

*Label/Data-Efficient Learning for 3D:* Recent studies have produced many elaborately designed backbone networks for 3D semantic/instance segmentation [35], [46], [73], [74], [75], as well as for 3D object detection [76], [77], [78]. However, they rely on full supervision. Directly applying current SOTAs (State-of-the-art) methods for training will result in a great decrease in performance [79] for WSL, if the percentage of labeled data drops to a certain value, e.g., less than 30%. Recently, many works have started to focus on point cloud semantic segmentation with partially labeled data. Wang [80] et al. choose to transform point clouds to images, but pixel-level semantic segmentation labels are required for network training. Sub-cloud annotations [81] require extra labor to separate the sub-clouds and to label points within the sub-clouds. Liu [82] proposed a robust data-efficient 3D scene parsing framework. It leverages the complementary merits of the superior generalization capacity of the traditional 3D descriptors and the strong feature description capacity learned 3D descriptors to learn very robust local features. Then using the descriptor guided learned region merging [3], superior performance can be achieved on downstream tasks. Liu et al. [83], [84] proposed self-training techniques to tackle scene understanding in weak supervision, which design a two-stage training scheme to produce iteratively optimized pseudo labels from weak labels during training. Despite satisfactory results, these approaches still have not learned generalized representation applicable to diverse tasks. Xu et al. [85] adopt a weakly-supervised training strategy, which combines training with coarse-grained information and partial points using 10% labels. However, their tested cases are limited to object part segmentation, and it is difficult to uniformly choose points to label. The convex decomposition [86] is conducted in an approximate manner to perform 3D scene parsing on the object parts. More approaches [87] have been proposed recently, which utilize class prototypes and masked point cloud modeling [66], [88], [89] to learn informative representations for downstream 3D scene understanding. Conceptfusion [90] has also been proposed to alleviate the labelling burden via rendering with the proposed CLIP-driven queryable 3D point cloud maps. To sum up, although approaches have been proposed to alleviate the data efficiency problem, the models for weakly supervised learning lack the capacity to recognize novel categories beyond the labeled training set. Our framework tackles open-set and data-efficient learning problems and is widely applicable to diverse 3D scene understanding downstream tasks.

## III. Proposed Methodology

We propose a general *WS3D++* framework to tackle weakly supervised 3D understanding with limited labels, as demonstrated in Fig. 1. Our framework consists of both the vision-language pre-training stage shown in Fig. 2 and fine-tuning stage illustrated in Fig. 4. *During the pre-training stage*, we first propose the hierarchical contrastive learning strategy with the help from rendering for more accurate vision-language alignments at both the scene level and object level. Then we also design a distillation strategy to distill point-language-aligned representations from 2D image network to 3D point-cloud network to endow 3D networks with the open-vocabulary recognition capacity. Finally, *during the fine-tuning stage*, we directly propose a weakly supervised approach, we perform fine-tuning with regional boundary awareness and region-aware instance discrimination, which significantly improve model discrimination capacities when the labeled data are rather scarce.

### A. Hierarchical Vision-Language Knowledge Associations and Distillations for Pre-Training

We propose a hierarchical alignment strategy in pre-training, which employs the rendering approach as a bridge to effectively align 3D vision and language embeddings, thus capturing coarse-to-fine associations for visual-linguistic synergized representations from the global scene level to local object level. It enables extracting more accurate 3D-language associations in a hierarchical manner.

*Multi-view rendering:* To obtain paired 2D-3D representations, we propose leveraging multi-view rendering to obtain paired 2D views from 3D point cloud scenes. The pairing process consists of two steps, the first is to convert point cloud scenes into meshes and the second is to render 2D images based on the different views of the 3D meshes. In terms of point-to-mesh transformation, we utilize the Delaunay triangulation approach [94], [95] to convert the point cloud into meshes, which is demonstrated as a very effective method for surface reconstruction. It connects the points in point cloud scenes by forming triangles that satisfy the Delaunay criterion which guarantees no point lies inside the circumcircle of any triangle. This method generates a triangle mesh that approximates the surface of the point cloud [96]. In terms of mesh-to-image transformation, we leverage the rendering pipeline including the vertex transformation, projection, and rasterization, as well as

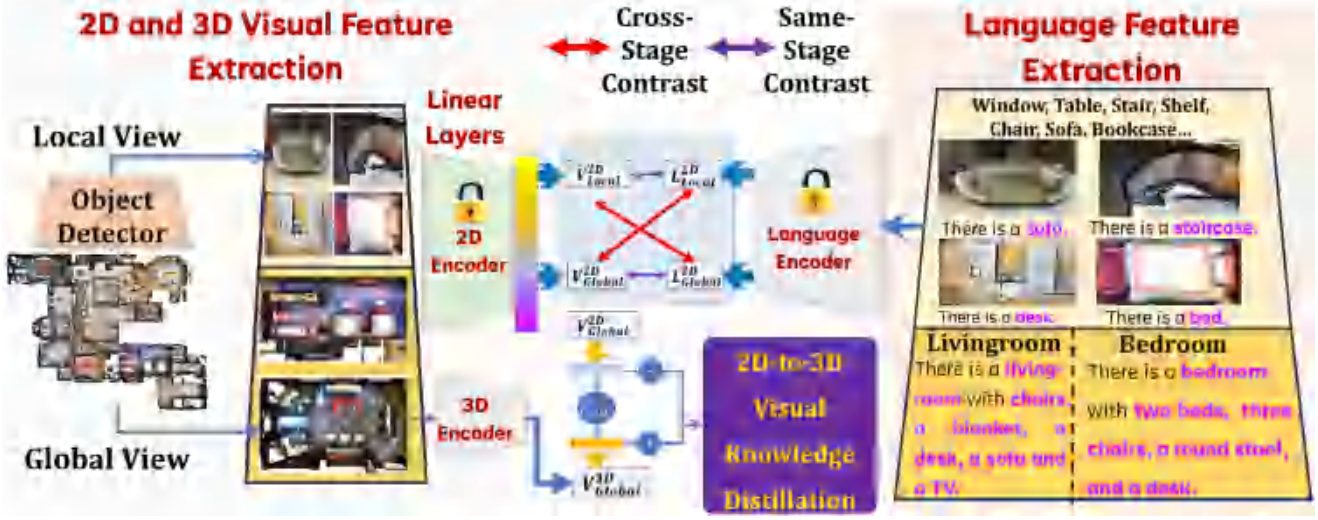


Fig. 2. The pre-training paradigm of our proposed *WS3D++*. We propose the hierarchical global to local feature alignments to establish the hierarchical vision-language aligned feature representations during the pre-training. This proposed paradigm helps to learn more powerful visual-linguistic aligned feature representation during the pre-training stage. We have further shown the final visualizations and comparisons with CLIP text presentations ranging from both the global view level to the local object category level. The results have further demonstrate the vision-language aligned feature representation for 3D scene parsing.

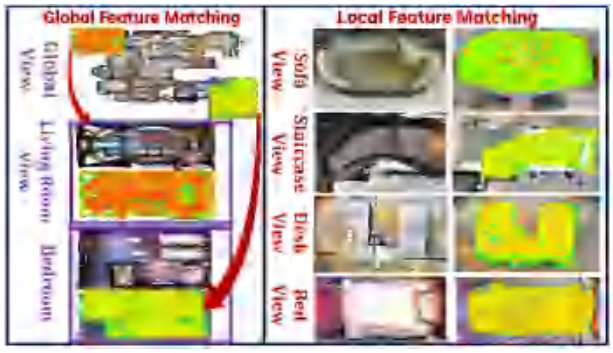


Fig. 3. The feature matching visualization of our proposed *WS3D++*. We propose hierarchical global to local feature alignments to establish hierarchical vision-language aligned feature representations during pre-training from both the global view level to the local object level. This kind of paradigm helps to learn more powerful visual-linguistic matched representations ranging from both the global view-level to the local object category-level. In the above figure, we have shown the matching at the global view on the left and the matching at local object level on the right. It can be demonstrated that our proposed approach can establish matched feature representation at both the global room feature level and the local object feature level.

shading. We directly use the rendering library OpenGL [97] to render images from meshes. The process involves projecting the 3D vertices onto a 2D image plane based on camera parameters and applying shading and lighting calculations of 3D meshes to determine the specific color of each pixel. By our simple rendering design, the world-to-camera extrinsic transformation matrix $T_e$ containing both rotation and translation information between the 2D pixels and 3D points can be easily obtained.

*2D to 3D Alignment:* After multi-view rendering, the strict 2D-3D alignment can be easily established if the camera's intrinsic $T_i$ is obtained from the standard calibration [98] and the extrinsic $T_e$ is obtained from the rendering. To be more specific, given the 3D point $\mathsf{p}_{3D} \in \mathsf{R}^3$ as well as its 2D corresponding pixel coordinate $\mathsf{p}_{2D} = (u, v)$, if we consider the pin-hole camera model, the transformation can be represented as $\hat{\mathsf{p}}_{2D} = T_i \cdot T_e \cdot \hat{\mathsf{p}}_{3D}$. The $\hat{\mathsf{p}}_{2D}$ and $\hat{\mathsf{p}}_{3D}$ are represented within the homogeneous coordinates, and they are strictly paired. Therefore, we can strictly determine the correspondence between $\hat{\mathsf{p}}_{2D}$ and $\hat{\mathsf{p}}_{3D}$. Moreover, we can find an explicit association between each element of the textual feature $\mathsf{F}_T$ and the 3D feature $\mathsf{F}_{3D}$ while passing through the backbone network.

*United 2D and 3D Proposal Generation:* As shown in Fig. 6, according to our experiments, we found that it is difficult to directly find the object-level information based on the 2D proposals merely due to the large information loss while rendering, the proposals provided by the 2D region proposal network (RPN) [100] can not effectively capture the 3D object information within the holistic scene. On the other hand, merely relying on the 3D proposals provided by the 3D RPN [101] still can not guarantee accurate proposal generation for the fact that some objects are too adjacent in their geometry. To this end, we propose to leverage the union of 2D and 3D RPN to capture intact and all-inclusive holistic proposals within the 3D scenes. Denote the region proposals as $R_{2D}$ and $R_{3D}$ respectively, the final holistic proposal generation $R_H$ is formulated as $R_H = R_{2D} \cup R_{3D}$. According to our experiments, this simple design can considerably boost the performance for the fact that the objects are more clustered and closely distributed within the indoor 3D scenes. Taking the union of 2D and 3D object proposals into consideration also guarantees that the optimization merely considers the regions where the object really exists and prevents the models from taking the pure background into consideration during the optimizations. According to our previous experiments, although the average precision and recall can merely achieve 38.7/45.9% and 47.6/56.8% for 2D and 3D object proposal generation, respectively, the 3D scene parsing performance can still be well-maintained. By combining the 2D proposal with the 3D using the union operation while omitting the duplicated 2D proposals, the object proposal generation

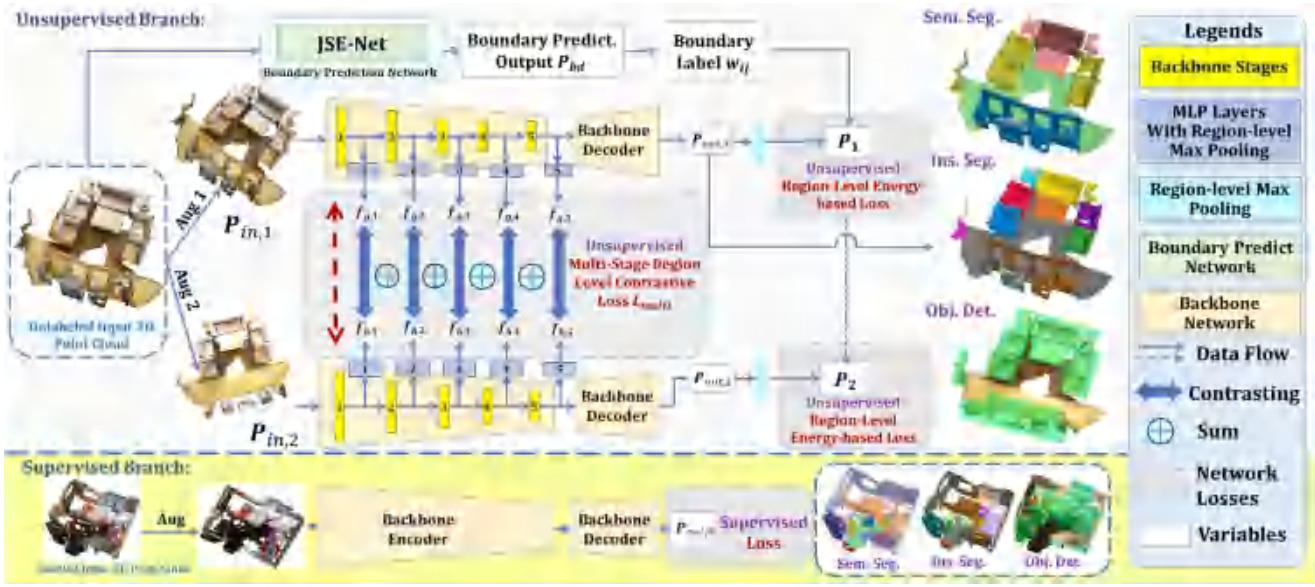


Fig. 4. The **fine-tuning paradigm** of our proposed *WS3D++*. *WS3D++* [4] consists of three proposed modules: 1. The **unsupervised** region-level energy-based optimization guided by boundary labels; 2. The **unsupervised** multi-stage region-level contrastive learning with high confidence; 3. The **supervised** region-level semantic contrastive learning with labeled data. The backbone network adopts encoder-decoder structures. The weights of the backbone network are shared in the supervised and unsupervised branches. Integrated with the proposed pre-training paradigm illustrated in Fig. 2, by our proposed hierarchical feature aligned pre-training and regional fine-tuning, more effective label-efficient learning as well as open-vocabulary learning is realized.

precision and recall can achieve a very high score of 65.2% and 76.8%, respectively, which demonstrates the superior effectiveness of them in generating region proposals.

In the next stage, we perform vision-language matched contrastive candidate selection at both the object level and the scene level. The object-level vision-language matching makes alignment between the individual object and the descriptive word. While the scene-level vision-language matching make alignment between the whole scene with the descriptive sentence. **In this way, the vision-language matched candidate can be selected for effective constrastive optimizations taking both global and local feature representation into account.**

Although the accuracy of proposal generation is not that high, our proposed approach can guarantee the final scene parsing performance as reported and demonstrated in Table I. The reason behind the phenomenon can be explained by that as long as the region proposal can be found, our framework can leverage the well-aligned representation in the vision language model (CLIP) to boost the final 3D semantic scene parsing performance, both in the open-vocabulary and closed-set scenarios. As validated by the previous experimental results and our qualitative validation in Fig. 6, the quality of object proposal can be guaranteed, if the domain gap is not that large.

*Global 2D Scene-to-Sentence Matching:* In the first place, we perform the global scene-to-sentence matching. For the rendered 2D image of the 3D scene, we utilize the GPT-4 to obtain the direct sentence-level description of the holistic scene. The sentence is given by the most similar description generated by GPT-4. Benefiting from the previous effective multi-view rendering designs, our proposed approach can have a holistic multi-view abstraction of the 3D environment. Denote the extract global scene-level visual feature as $V_{global}$ and denote the global text-level feature as $L_{global}$, then we can evaluate the alignment between visual and linguistic feature utilizing the similarity between $V_{global}$ and $L_{global}$.

*Local 2D Object-to-Word Matching:* At the next step, we perform local object-to-word matching. In contrast to the global scene to scentence matching, here we conduct local object-to-word matching, which establishes and enhances vision-language association at more fine-grained object level and word level. The association is established at the object level for visual part and at the word level for language part. Specifically, we use the CLIP image encoder to generate a set of local visual feature embeddings $V_{l}ocal$ and utilize the text encoder to output a set of local word-level feature embeddings $L_{l}ocal$. In the next step, we calculate aggregated similarity using the cosine similarity $\mathcal{G}_{sim}$ between the 3D visually encoded features and the textual features:

$$\mathsf{S}(\mathsf{V},\mathsf{L}) = \mathcal{G}_{sim}\left(V^{2D}_{gobal}, L^{2D}_{gobal}\right) + \mathcal{G}_{sim}\left(V^{2D}_{local}, L^{2D}_{local}\right) \quad (1)$$

The operation can be interpreted as a bi-directional operation, which means that for each region proposal image patch, we find the textual concept that fits best with the semantics of the region. And for each textual concept, we find the region that has best correspondence with it among multiple regions. We model it as an **optimal transport** problem, which finds the most similar visual feature by formulating it as the differentiable Top-$k$ with respect to the related anchor textual description [102] both globally and regionally. The region-to-word pairs with Top-$k$ maximum activations $\mathsf{S}(\mathsf{V},\mathsf{L})$ are finally regarded as the positive pairs in contrast. It can ultimately ensure learning highly discriminative representations at both the global scene-to-sentence level and the local object-to-word level. Finally, it can be demonstrated by our experiments that the alignments at the local object-to-word

TABLE I
COMPARISONS OF THE OPEN-VOCABULARY LEARNING PERFORMANCE ON SCANNET

| Datasets | Models | 80% | | | 100% | | |
|---|---|---|---|---|---|---|---|
| | | B15/N4 | B12/N7 | B10/N9 | B15/N4 | B12/N7 | B10/N9 |
| ScanNet [18] | Baseline | 39.7 / 48.8 / 38.9 | 37.9 / 44.9 / 42.6 | 31.9 / 38.8 / 33.9 | 43.6 / 50.7 / 39.6 | 38.8 / 48.9 / 45.8 | 33.9 / 42.9 / 38.8 |
| | Merely the Pre-training Stage | 55.9 / 52.9 / 48.8 | 46.9 / 48.8 / 46.9 | 43.9 / 46.7 / 43.8 | 69.6 / 67.8 / 59.8 | 55.8 / 50.9 / 48.8 | 53.6 / 47.9 / 46.8 |
| | 3DGenZ [91] | 32.6 / 76.8 / 26.6 | 25.9 / 46.5 / 26.3 | 28.3 / 63.6 / 18.6 | 42.9 / 76.9 / 39.6 | 32.9 / 56.8 / 33.9 | 27.1 / 69.8 / 32.8 |
| | 3DGenZ [91] (Pr) | 42.8 / 78.8 / 42.9 | 39.6 / 68.3 / 51.8 | 27.6 / 79.5 / 18.9 | 46.8 / 88.6 / 48.6 | 43.8 / 78.8 / 49.9 | 37.9 / 83.7 / 38.9 |
| | 3DTZSL [92] | 17.6 / 45.8 / 6.1 | 23.8 / 36.6 / 12.8 | 13.8 / 67.5 / 15.7 | 42.9 / 49.6 / 16.7 | 27.8 / 38.9 / 15.9 | 16.9 / 69.8 / 18.8 |
| | 3DTZSL [92] (Pr) | 49.6 / 58.7 / 48.9 | 38.8 / 50.6 / 38.6 | 33.9 / 69.8 / 33.9 | 59.6 / 69.8 / 59.6 | 48.9 / 59.6 / 48.7 | 38.6 / 73.9 / 39.6 |
| | LSeg3D [93] | 0.0 / 64.4 / 0.0 | 26.9 / 55.7 / 22.8 | 11.8 / 77.6 / 23.9 | 33.6 / 76.8 / 33.8 | 28.8 / 72.6 / 29.2 | 26.6 / 68.5 / 26.9 |
| | LSeg3D [93] (Pr) | 38.9 / 78.8 / 29.9 | 32.8 / 73.6 / 26.8 | 28.6 / 66.8 / 22.8 | 49.6 / 82.7 / 46.9 | 46.9 / 76.8 / 43.2 | 38.9 / 70.3 / 39.6 |
| | PLA without caption [70] | 39.7 / 73.3 / 28.0 | 31.5 / 72.3 / 25.6 | 28.7 / 68.6 / 23.5 | 45.7 / 72.9 / 36.9 | 31.9 / 77.8 / 26.8 | 29.8 / 79.8 / 28.3 |
| | PLA without caption [70] (Pr) | 53.9 / 82.6 / 45.8 | 48.6 / 80.6 / 40.8 | 42.6 / 75.5 / 37.8 | 56.7 / 84.3 / 52.9 | 53.9 / 83.8 / 45.3 | 52.1 / 80.8 / 45.1 |
| | PLA [70] | 69.2 / 70.6 / 66.9 | 60.9 / 68.9 / 59.9 | 58.2 / 78.6 / 73.8 | 69.3 / 73.3 / 72.2 | 62.7 / 73.6 / 61.3 | 65.6 / 81.8 / 76.8 |
| | PLA (Pr) [70] | 79.9 / 82.9 / 81.9 | 78.5 / 80.6 / 68.9 | 65.9 / 77.8 / 57.8 | 82.5 / 88.2 / 84.9 | 80.8 / 85.7 / 73.9 | 76.8 / 85.3 / 63.9 |
| | *WS3D++* (Ours) | 72.3 / 69.1 / 73.3 | 68.6 / 65.6 / 68.6 | 73.2 / 71.9 / 71.6 | 77.9 / 76.6 / 78.6 | 73.8 / 72.8 / 75.8 | 71.9 / 72.8 / 73.9 |
| | *WS3D++* (Ours) (Pr) | 75.8 / 73.8 / 77.5 | 78.6 / 75.3 / 78.9 | 71.8 / 85.6 / 79.5 | 86.7 / 83.7 / 85.8 | 82.7 / 80.9 / 83.6 | 78.9 / 88.9 / 83.8 |
| Datasets | Models | 10% | | | 50% | | |
| | | B15/N4 | B12/N7 | B10/N9 | B15/N4 | B12/N7 | B10/N9 |
| ScanNet [18] | Baseline | 27.9 / 18.2 / 21.9 | 23.0 / 33.6 / 31.2 | 19.6 / 28.6 / 27.6 | 35.1 / 20.6 / 36.6 | 31.7 / 41.5 / 38.2 | 28.5 / 32.3 / 28.2 |
| | Merely the Pre-training Stage | 35.8 / 58.3 / 36.8 | 33.8 / 37.9 / 31.8 | 30.8 / 70.3 / 25.6 | 37.2 / 63.2 / 39.7 | 48.9 / 53.8 / 49.9 | 49.6 / 86.8 / 49.8 |
| | 3DGenZ [91] | 17.3 / 52.3 / 9.7 | 15.3 / 28.8 / 12.5 | 11.2 / 62.5 / 5.2 | 18.6 / 52.8 / 11.6 | 17.8 / 33.6 / 12.9 | 11.8 / 72.3 / 6.3 |
| | 3DGenZ [91] (Pr) | 28.5 / 72.9 / 28.2 | 23.6 / 38.9 / 23.6 | 15.6 / 68.9 / 12.9 | 29.7 / 75.6 / 29.7 | 29.5 / 45.3 / 45.7 | 23.3 / 43.9 / 11.9 |
| | 3DTZSL [92] | 8.9 / 33.9 / 6.1 | 3.3 / 35.8 / 1.8 | 7.3 / 47.3 / 2.8 | 16.2 / 32.5 / 23.2 | 21.9 / 31.3 / 6.8 | 9.9 / 63.5 / 8.8 |
| | 3DTZSL [92] (Pr) | 26.8 / 43.8 / 28.7 | 18.6 / 52.6 / 23.2 | 16.8 / 49.5 / 19.2 | 31.3 / 49.3 / 28.7 | 28.6 / 43.9 / 27.3 | 15.3 / 66.9 / 21.7 |
| | LSeg3D [93] | 0.0 / 55.6 / 0.0 | 0.6 / 43.6 / 0.3 | 1.2 / 56.1 / 0.8 | 19.9 / 65.3 / 13.6 | 12.9 / 56.8 / 8.7 | 9.9 / 58.5 / 8.7 |
| | LSeg3D [93] (Pr) | 28.8 / 65.3 / 23.8 | 26.7 / 62.8 / 16.7 | 23.7 / 62.8 / 22.9 | 32.8 / 76.7 / 28.7 | 28.9 / 73.9 / 26.9 | 26.8 / 73.2 / 25.6 |
| | PLA without caption [70] | 35.3 / 53.2 / 18.2 | 26.7 / 52.1 / 9.9 | 19.6 / 56.7 / 9.7 | 36.5 / 55.9 / 22.8 | 28.2 / 58.7 / 18.8 | 25.8 / 62.1 / 15.2 |
| | PLA without caption [ ] (Pr) | 39.3 / 73.3 / 35.7 | 28.6 / 67.8 / 29.6 | 24.5 / 62.6 / 26.8 | 42.9 / 75.6 / 41.8 | 39.3 / 73.6 / 36.9 | 35.6 / 72.3 / 33.8 |
| | PLA [70] | 51.3 / 58.8 / 48.6 | 49.3 / 55.9 / 44.9 | 44.7 / 52.2 / 40.8 | 65.7 / 72.8 / 65.7 | 58.8 / 66.5 / 49.8 | 53.6 / 63.5 / 46.8 |
| | PLA (Pr) [70] | 61.8 / 67.6 / 73.9 | 58.7 / 59.3 / 56.8 | 68.8 / 82.9 / 47.6 | 72.5 / 82.8 / 79.9 | 71.8 / 81.9 / 61.8 | 70.3 / 83.8 / 49.8 |
| | *WS3D++* (Ours) | 65.9 / 75.2 / 68.3 | 61.3 / 65.6 / 62.7 | 58.6 / 60.9 / 56.8 | 71.8 / 76.8 / 81.6 | 65.8 / 69.5 / 68.9 | 62.8 / 67.8 / 68.2 |
| | *WS3D++* (Ours) (Pr) | 69.9 / 68.8 / 67.8 | 67.9 / 67.7 / 66.9 | 63.6 / 66.2 / 65.6 | 70.9 / 70.8 / 71.9 | 72.8 / 71.8 / 69.8 | 69.8 / 75.6 / 72.1 |

It can be demonstrated that our proposed approach provides very superior open-world recognition performance compared with the diverse SOTAs. The results are given as hIoU / $\text{mIoU}_{\mathcal{B}}$ / $\text{mIoU}_{\mathcal{N}}$, respectively. We have compared intensively with diverse labeled ratios. We apply panoptic quality mean intersection over union (mIoU), which can be divided into segmentation quality and recognition quality as metrics for the instance segmentation task. These evaluation metrics are computed on base (B) and novel (N) categories, with the superscripts of B and N (e.g. $\text{mIoU}_{\mathcal{B}}$/$\text{mIoU}_{\mathcal{N}}$), respectively. Furthermore, we have also utilized the harmonic metric such as harmonic IoU (hIoU) as the major indicators for open-world tasks following popular zero-shot learning work to consider the partition between base and novel classes.

level both have a considerable boost on the final scene parsing performance, both qualitatively and quantitatively.

*2D-3D Visual Feature Distillation:* The main purpose is to conduct 2D-3D Visual Feature Distillation to *obtain the aligned 2D-3D visual feature in the 2D-3D-language co-embedded feature space.* We use the global 3D backbone to extract the 3D visual feature representation $V^{3D}_{global}$, and then utilize the $\mathcal{KL}$ divergence as the distillation loss to further distill the vision-language aligned informative knowledge from 2D feature space to 3D. This process can also be interpreted as the 2D-3D explicit feature alignment/distillation process. Compared with the mean square error loss such as the $\mathcal{L}_1$ or $\mathcal{L}_2$ losses, the $\mathcal{KL}$ divergence has improved regression capacity and ensured smoother gradient, which to some extent overcomes overfitting problems while distilling knowledge. The $\mathcal{KL}$ distillation loss $\mathcal{L}^{Dist}_{\mathcal{KL}}$ finally operates on the final two normalized 2D/3D vector for feature alignment:

$$\mathcal{L}^{Dist}_{\mathcal{KL}} = \mathsf{Div}_{\mathcal{KL}}\left(\frac{\mathsf{V}^{2D}_{global}}{||\mathsf{V}^{2D}_{global}||}, \frac{\mathsf{V}^{3D}_{global}}{||\mathsf{V}^{3D}_{global}||}\right) \tag{2}$$

*Contrastive Language-Vision Optimizations:* Note that when conducting contrastive learning, we regard textual features as anchors because textual descriptions are highly semantic and contain rich information, whereas the images contain too much low-level information and pixel-level details. Denote the $\mathsf{V}^{3D,+}_{global}$ and the $\mathsf{L}^{3D,-}_{global}$ as the positive and the negative feature with respect to the anchor textual feature $\boldsymbol{F}^{a}_{T}$, respectively, the designed contrastive discrimination loss is formulated as follows:

$$\mathcal{L}_{Ctr} = -\frac{1}{\|\mathbf{H}\|}\sum_{(a,b)\in\mathbf{H}} \times \log\frac{\exp(\mathsf{L}^{3D}_{global}\cdot\mathsf{V}^{3D,+}_{global}/\tau)}{\sum_{(\cdot,c)\in\mathbf{B}}\exp(\mathsf{L}^{3D}_{global}\cdot\mathsf{V}^{3D,-}_{global}/\tau))}. \tag{3}$$

The final pre-training optimization loss is the joint consideration of contrastive language-vision optimization and 2D-3D visual feature distillation with the balancing $\lambda_{\mathcal{KL}}$ set to 0.5 empirically. Note that $\lambda_{\mathcal{KL}}$ set to range of 0.3-3.0 will not influence the performance too much according to our evaluation.

$$\mathcal{L}_{Pretrain} = \mathcal{L}_{Ctr} + \lambda_{\mathcal{KL}}\mathcal{L}^{Dist}_{\mathcal{KL}} \tag{4}$$

According to our extensive experiments, our simple pre-training approach provides the well-aligned vision-language-3D aggregated co-embedding. It considerably facilitates vision-language-associated knowledge transfers from 2D to 3D, boosting both the

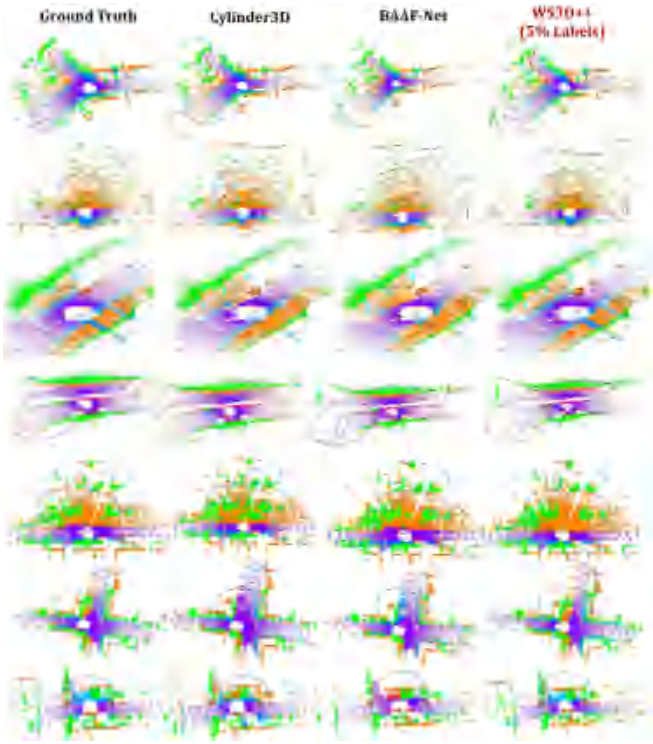


Fig. 5. Qualitative **semantic segmentation** results of the proposed *WS3D++* on SemanticKITTI validation set the with the 5% labeling percentage, compared with the fully supervised state-of-the-art Cylinder3D [34], and BAAF-Net [35] with the diverse semantics indicated by different colors. The red circles highlight the final performance difference between diverse comparative approaches.

final label efficiency and the final recognition capacity of unseen novel categories.

### B. Region-Aware Fine-Tuning

During the fine-tuning stage, our proposed framework consists of three subparts for the network optimization: **1**. Unsupervised energy-based loss guided by boundary awareness and highly confident network predictions for unlabeled data, which is discussed in our original ECCV conference work [4]; **2**. Unsupervised multi-stage region-level contrastive learning with highly confident predictions for unlabeled data. **3**. Supervised semantic contrastive learning for labeled data. The three important modules above are integrated jointly into the optimization function for network training to accomplish the final downstream detection or segmentation tasks with a very limited labeled data, with all remaining data unlabeled.

## IV. Experiments

### A. Pre-Training Experimental Settings

For the indoor scene understanding tasks, we pre-train the network on ScanNet [18]. And for the outdoor scene parsing tasks, we pre-train the network on NuScenes [36] dataset. For the dataset partition, we follow the official partition of ScanNet-V2 [18] using 1,201 scans as the pre-training dataset. The NuScenes [36] is an outdoor autonomous driving dataset that


Fig. 6. The captured 2D and 3D region proposals. It is demonstrated qualitatively clearly that more precise object proposals are captured by proposed united 2D/3D proposal generation approach. It can be demonstrated that clear superior regional proposal generation performance can be well guaranteed.

contains 7000 training scenes, the dataset provides the camera's intrinsic and extrinsic parameters, thus we can obtain the 2D to 3D transformations and alignments very easily from designed rendering approaches. For the indoor and outdoor pre-training, we pre-train the network for 500 epochs and then we fine-tune the network on diverse downstream tasks. The hyper-parameter $k$ in Top-$k$ is set to 3. The initial learning rate is set to $5 \times 10^{-4}$ and is multiplied with 0.2 every 50 epochs.

### B. Finetuning Experimental Settings

*Datasets:* During the fine-tuning, to demonstrate the effectiveness for both data-efficient learning and open-world recognition of the proposed *WS3D* and *WS3D++* under the limited scene reconstruction labeling scheme, we have tested it on various benchmarks, including S3DIS [19], ScanNet [103],

TABLE II
BOUNDARY PREDICTION AVERAGE PRECISION (AP) DIVERSE LABEL RATIOS FOR OUTDOOR SEMANTICKITTI (SKITTI) BENCHMARK AND INDOOR SCANNET AND S3DIS BENCHMARK

| Labeling Percentage | The AP% of Boundary Prediction (BP) | | | WS3D++ Sem. Seg. mIOU% | | |
|---|---|---|---|---|---|---|
| | SKITTI (Outdoor) | ScanNet (Indoor) | S3DIS (Indoor) | SKITTI (Outdoor) | ScanNet (Indoor) | S3DIS (Indoor) |
| 1% | **56.7** / 61.6 | **55.8** / 59.1 | **56.2** / 61.7 | 51.6 | 58.8 | 55.8 |
| 5% | **53.9** / 60.8 | **58.9** / 63.9 | **59.6** / 69.2 | 59.3 | 67.9 | 64.7 |
| 10% | **55.7** / 61.5 | **62.2** / 65.7 | **62.6** / 66.9 | 68.7 | 76.7 | 68.2 |
| 15% | **57.6** / 67.9 | **65.2** / 68.2 | **62.9** / 67.8 | 75.2 | 76.2 | 68.8 |
| 20% | 59.7 / 66.5 | **66.3** / 69.8 | 65.2 / 68.7 | 72.9 | 79.8 | 75.8 |
| 25% | **62.6** / 65.6 | **67.8** / 70.3 | **65.8** / 67.8 | 74.7 | 77.1 | 75.5 |
| 30% | 64.6 / 70.2 | **71.6** / 73.8 | **68.9** / 72.8 | 78.7 | 76.3 | 75.9 |
| 40% | **68.6** / 72.8 | **73.2** / 75.3 | **71.2** / 73.5 | 75.6 | 81.6 | 77.8 |
| 100% | **83.7** / 88.7 | **76.9** / 82.6 | **76.8** / 83.7 | 83.9 | 88.8 | 82.9 |

Left of '/': The Average Precision (AP) of Semantic Boundary Prediction with focal Loss $L_{fl}$. Right of '/': The average precision with the cross entropy loss $L_{ce}$.

and SemanticKITTI [16] for semantic segmentation, and ScanNet [103] for instance segmentation, respectively. Detailed information on each dataset and training details are put into the Appendix. The pre-training is conducted on ScanNet training set for indoor benchmark and on Waymo for outdoor benchmark.

*Training Set Partition:* Following the typical setting in data-efficient learning in the limited reconstruction case [20] [104], we partition the training set of all tested datasets into labeled data and unlabeled data with various labeling points percentage, 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-scene belongs to the unlabeled data.

*Implementation Details:* For the task of semantic segmentation, we fine-tune the network for 500 epochs on a single NVIDIA 1080Ti GPU with a batch size of 16 during training. The initial learning rate is set to $1\times10^{-3}$ and is multiplied with 0.2 every 50 epochs. We implement it in *PyTorch* and optimize it with *Adam* optimizer [110]. We set the hyperparameter $\gamma$ as 0.8 to ensure that merely highly confident prediction can be used for network optimization. $\epsilon$ is set to 0.5. We empirically choose $\alpha = \beta = 1$, while $\lambda_u = 0.1$. For 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.

### C. Data-Efficient 3D Semantic Segmentation

*Overall Experimental Results:* For the semantic segmentation, we have tested *WS3D++* on versatile indoor and outdoor benchmarks, including ScanNet [18], S3DIS [19], and SemanticKITTI [16]. We have done extensive experiments with limited labeled data, e.g., only {1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 100%} data in training set are available as labeled data. The qualitative results are shown in Fig. 5. In the meanwhile, the quantitative semantic segmentation performance is summarized in Table III. As mentioned, we have used the voxel-based method SparseConv [60] as the backbone. Our WSL model significantly surpasses the supervised-only model in **GPC** that is merely trained with labeled data, showing that our WSL can effectively make use of the unlabeled data to enhance the feature discrimination capacity of the model. Also, it can be observed 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 SOTA **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. 5 shows that we can provide comparable performance compared with fully supervised SOTAs BAAF-Net [35] and Cylinder3D [34] on SemanticKITTI with 5% labels. As shown in Table III, the performance of our enhanced approach *WS3D++* has remarkably increased performance compared with *WS3D++* and previous SOTAs, demonstrating the effectiveness of our proposed vision-language knowledge-associated pre-training.

TABLE III
COMPARISON OF **SEMANTIC SEGMENTATION** RESULTS WITH DIFFERENT LABELING PERCENTAGES ON SCANNET VALIDATION SET, S3DIS VALIDATION SET (AREA 5), AND SEMANTICKITTI VALIDATION SET (SEQUENCE 08)

| Datasets | Approaches | Semantic Segmentation mIoU (%) on the Validation Set According to Supervision Level (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 15% | 20% | 25% | 30% | 40% | 100% |
| ScanNet | Sup-only-GPC–Baseline | 40.9 | 48.1 | 57.2 | 61.3 | 64.0 | 65.3 | 67.1 | 68.8 | 72.9 |
| | Sup-only-GPC (Pr) | 44.8 | 56.9 | 59.8 | 64.9 | 66.7 | 69.9 | 66.8 | 63.8 | 70.8 |
| | **GPC** [104] | 46.6 | 54.8 | 60.5 | 63.3 | 66.7 | 67.5 | 68.9 | 71.3 | 74.0 |
| | **GPC** (Pr) [104] | 52.5 | 59.3 | 67.8 | 68.9 | 69.7 | 69.8 | 75.3 | 76.5 | 76.6 |
| | Mix3D [105] | 47.1 | 55.2 | 61.1 | 63.5 | 66.8 | 67.9 | 68.7 | 71.5 | 73.3 |
| | Mix3D [105] (Pr) | 52.7 | 58.9 | 65.1 | 67.9 | 70.2 | 70.6 | 71.3 | 73.8 | 78.6 |
| | PointMixup [106] | 44.5 | 55.2 | 69.6 | 63.7 | 66.3 | 66.9 | 68.8 | 70.2 | 73.8 |
| | PointMixup [106] (Pr) | 46.8 | 59.8 | 75.5 | 68.1 | 69.8 | 73.7 | 72.8 | 75.3 | 78.6 |
| | RDPL [107] | 47.6 | 56.3 | 61.1 | 63.8 | 66.8 | 67.8 | 69.7 | 71.8 | 74.6 |
| | RDPL [107] (Pr) | 50.7 | 59.6 | 66.9 | 68.3 | 71.6 | 71.5 | 72.8 | 73.9 | 76.8 |
| | Active-ST [108] | 48.7 | 56.7 | 62.5 | 65.6 | 67.8 | 68.7 | 71.2 | 72.5 | 75.8 |
| | Active-ST [108] (Pr) | 49.9 | 59.5 | 67.9 | 67.9 | 75.3 | 70.6 | 73.7 | 75.2 | 78.1 |
| | *WS3D++* | 52.6 +11.7 | 59.1 +11.0 | 65.2 +8.0 | 66.9 +3.6 | 71.8 +6.6 | 72.0 +4.7 | 73.9 +5.3 | 75.3 +4.6 | 81.7 +8.8 |
| | *WS3D++* (Pr) | 58.8 +17.9 | 67.9 +19.8 | 76.5 +18.5 | 78.6 +14.9 | 79.8 +15.8 | 80.1 +13.8 | 80.8 +11.6 | 83.8 +12.8 | 88.8 +15.9 |
| | Merely Pre-training Stage | 43.9 | 52.3 | 56.3 | 61.6 | 65.3 | 67.6 | 68.5 | 69.8 | 71.9 |
| S3DIS | Sup-only-GPC–Baseline | 35.3 | 44.5 | 52.9 | 53.8 | 59.9 | 60.3 | 61.2 | 62.6 | 66.4 |
| | Sup-only-GPC (Pr) | 48.6 | 51.7 | 60.8 | 61.9 | 70.8 | 72.8 | 69.7 | 68.8 | 73.6 |
| | **GPC** [104] | 38.2 | 53.0 | 57.7 | 60.2 | 63.5 | 63.9 | 64.9 | 65.0 | 68.8 |
| | **GPC** [104] (Pr) | 47.8 | 58.7 | 63.8 | 66.8 | 68.5 | 69.7 | 68.8 | 68.9 | 71.2 |
| | **Mix3D** [105] | 39.2 | 50.3 | 58.3 | 61.1 | 64.1 | 64.8 | 65.9 | 66.9 | 69.6 |
| | **Mix3D** [105] (Pr) | 47.2 | 57.9 | 63.2 | 68.2 | 69.5 | 70.6 | 69.8 | 69.8 | 73.9 |
| | *WS3D++* | 48.6 +12.3 | 57.7 +12.7 | 61.2 +8.3 | 66.9 +11.6 | 70.6 +10.7 | 71.1 +10.8 | 72.6 +11.4 | 75.3 +12.7 | 82.0 +15.6 |
| | *WS3D++* (Pr) | 55.8 +19.5 | 64.7 +19.7 | 68.2 +15.3 | 68.8 +13.0 | 75.8 +15.9 | 75.5 +15.2 | 75.9 +14.7 | 77.8 +15.2 | 82.9 +16.5 |
| | Merely Pre-training Stage | 45.2 | 52.8 | 58.7 | 63.8 | 68.6 | 69.8 | 71.2 | 72.6 | 73.9 |
| SemanticKITTI | Sup-only-GPC–Baseline | 28.6 | 34.8 | 43.9 | 47.9 | 53.8 | 55.1 | 55.4 | 57.4 | 65.0 |
| | Sup-only-GPC (Pr) | 36.3 | 41.8 | 50.3 | 55.6 | 57.6 | 63.8 | 60.8 | 59.8 | 67.8 |
| | **GPC** [104] | 34.7 | 41.8 | 49.9 | 53.1 | 58.8 | 59.1 | 59.4 | 59.9 | 65.8 |
| | **GPC** [104] (Pr) | 36.9 | 45.6 | 53.8 | 55.8 | 59.9 | 63.9 | 65.6 | 64.8 | 69.5 |
| | LESS [109] | 37.1 | 42.5 | 50.5 | 53.9 | 59.5 | 59.6 | 60.5 | 63.5 | 66.9 |
| | LESS (Pr) [109] | 39.6 | 46.7 | 53.9 | 56.8 | 63.9 | 65.7 | 65.3 | 68.7 | 68.8 |
| | Mix3D [105] | 37.7 | 42.9 | 50.8 | 54.1 | 59.9 | 60.9 | 61.8 | 61.3 | 68.8 |
| | Mix3D [105] (Pr) | 41.8 | 46.6 | 53.9 | 58.6 | 64.7 | 65.6 | 66.6 | 65.9 | 73.7 |
| | *WS3D++* | 46.8 +18.2 | 48.6 +13.8 | 55.2 +11.3 | 62.8 +14.9 | 65.9 +12.1 | 67.9 +12.8 | 68.6 +13.2 | 70.9 +13.5 | 76.8 +11.8 |
| | *WS3D++* (Pr) | 51.6 +33.0 | 59.3 +24.5 | 68.7 +24.8 | 69.8 +21.9 | 72.9 +19.1 | 74.7 +19.6 | 76.3 +20.9 | 76.6 +19.2 | 83.9 +18.9 |
| | Merely Pre-training Stage | 42.9 | 46.6 | 51.7 | 61.7 | 63.8 | 67.8 | 69.6 | 71.8 | 75.8 |

'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 (left value) and to the SOTAs GPC [104] (right value). The results in brackets (Pr) mean that the results after using our proposed pre-training approach to increase the final performance. It should be noted that the pre-training models for all weakly supervised approaches are the same with our proposed *WS3D++* pre-training. It can be demonstrated that our proposed pre-training approach has a significant boost on the final scene parsing performance. Note that our subscript highlights the performance increment compared with the "Sup-only-GPC".

TABLE IV
COMPARISON OF EXPERIMENTAL RESULTS ON 20% AND FULLY LABELED CASE FOR THE TASK OF INDUCTIVE AND TRANSDUCTIVE LEARNING FOR OUR PROPOSED *WS3D++*, RESPECTIVELY

| Datasets | 20% label | | | 100% label | | |
|---|---|---|---|---|---|---|
| | Base | Induct. | Transduct. | Base | Induct. | Transduct. |
| ScanNetv2-*WS3D++* | 69.9 | 71.8 | 85.8 | 78.9 | 81.7 | 87.8 |
| ScanNetv2-*WS3D++* (Pr) | 75.3 | 79.8 | 87.8 | 85.6 | 88.8 | 91.8 |
| S3DIS *WS3D++* | 62.8 | 70.6 | 75.8 | 69.7 | 82.0 | 86.8 |
| S3DIS *WS3D++* (Pr) | 68.7 | 75.8 | 78.7 | 73.8 | 82.9 | 87.3 |
| Semantic KITTI Val. *WS3D++* | 58.9 | 65.9 | 79.3 | 74.3 | 76.8 | 88.8 |
| Semantic KITTI Val. *WS3D++* (Pr) | 62.8 | 72.9 | 69.1 | 68.6 | 83.9 | 85.5 |
| Semantic KITTI Test. *WS3D++* | 66.7 | 71.8 | 72.6 | 71.6 | 77.9 | 79.6 |
| Semantic KITTI Test. *WS3D++* (Pr) | 70.8 | 75.9 | 78.8 | 79.1 | 84.3 | 86.5 |

In transductive learning, the test set is also utilized for network training. We test the task of semantic segmentation on ScanNet, S3DIS, and SemanticKITTI with the evaluation metric of mIoU(%).

TABLE V
COMPARISON OF CURRENT STATE-OF-THE-ART (SOTA) APPROACHES IN THE LIMITED RECONSTRUCTION CASE FOR THE **3D OBJECT DETECTION** TASKS WITH DIFFERENT RATIOS OF LABELED DATA

| Datasets | Models | 5% | | 10% | | 20% | | 100% | |
|---|---|---|---|---|---|---|---|---|---|
| | | Induct. | Transduct. | Induct. | Transduct. | Induct. | Transduct. | Induct. | Transduct. |
| SUN RGB-D [59], [111] | Baseline | 29.9 ± 1.5 | 33.5 ± 0.8 | 34.4 ± 1.1 | 40.7 ± 0.9 | 41.1 ± 0.3 | 47.5 ± 0.5 | 51.7 ± 0.9 | 53.8 ± 0.7 |
| | Merely the Pre-training Stage | 35.8 ± 1.1 | 38.2 ± 1.1 | 39.3 ± 0.7 | 45.1 ± 0.7 | 46.1 ± 0.9 | 56.9 ± 0.7 | 51.9 ± 1.1 | 65.8 ± 1.5 |
| | SESS [112] | 34.2 ± 2.0 | 38.1 ± 0.7 | 42.9 ± 0.8 | 45.3± 0.9 | 47.9 ± 0.4 | 51.6 ± 0.2 | 50.7 ± 0.5 | 53.7 ± 0.5 |
| | SESS (Pr) [112] | 42.8 ± 3.2 | 42.6 ± 0.7 | 48.2 ± 0.8 | 53.8 ± 0.9 | 50.5 ± 0.6 | 56.3 ± 0.7 | 55.2 ± 0.3 | 63.9 ± 0.6 |
| | 3D-IOUMatch [93] | 39.1 ± 1.9 | 46.3 ± 0.7 | 45.5 ± 1.5 | 53.5 ± 0.3 | 51.6 ± 0.5 | 55.6 ± 0.8 | 55.6 ± 0.3 | 62.3 ± 0.7 |
| | 3D-IOUMatch (Pr) [93] | 43.6 ± 3.2 | 48.5 ± 0.9 | 48.5 ± 1.8 | 56.9 ± 0.5 | 59.8 ± 0.7 | 56.8 ± 0.8 | 65.7 ± 0.5 | 69.1 ± 0.7 |
| | SPD [113] | 38.5 ± 0.7 | 44.3 ± 0.8 | 46.0 ± 1.0 | 51.7 ± 0.9 | 60.9 ± 0.5 | 56.9 ± 0.9 | 67.9 ± 0.7 | 73.9 ± 0.8 |
| | SPD [113] (Pr) | 43.6 ± 1.2 | 47.8 ± 1.7 | 49.6 ± 0.8 | 53.8 ± 1.1 | 58.6 ± 1.6 | 68.9 ± 2.5 | 69.2 ± 0.9 | 83.8 ± 0.5 |
| | *WS3D*++ (Ours) | 52.7 ± 0.7 | 54.5 ± 0.8 | 53.7 ± 0.9 | 63.8 ± 0.8 | 69.8 ± 0.7 | 72.9 ± 1.1 | 68.6 ± 0.9 | 82.8 ± 0.5 |
| | *WS3D*++ (Ours) (Pr) | 56.9 ± 0.8 | 59.7 ± 1.5 | 58.7 ± 1.8 | 68.9 ± 1.3 | 68.6 ± 1.8 | 73.7 ± 1.6 | 72.6 ± 1.1 | 83.9 ± 0.9 |
| ScanNet-V2 [18] | Baseline | 27.9 ± 0.5 | 30.8 ± 1.5 | 31.1 ± 0.7 | 33.2 ± 0.5 | 41.6 ± 0.9 | 44.5 ± 1.2 | 45.8 ± 0.8 | 48.9 ± 0.9 |
| | Merely the Pre-training Stage | 35.2 ± 1.5 | 32.8 ± 1.1 | 35.6 ± 0.8 | 36.6 ± 0.6 | 43.9 ± 1.1 | 47.9 ± 0.8 | 49.3 ± 0.7 | 56.8 ± 0.9 |
| | SESS [112] | 32.2 ± 0.8 | 36.8 ± 1.1 | 39.7 ± 0.9 | 44.5 ± 0.5 | 48.2 ± 0.4 | 49.6 ± 0.7 | 55.8 ± 0.1 | 56.8 ± 0.7 |
| | SESS (Pr) [112] | 35.9 ± 0.9 | 43.9 ± 1.2 | 43.5 ± 1.2 | 48.9 ± 0.5 | 49.6 ± 0.6 | 53.2 ± 0.7 | 56.9 ± 0.6 | 59.6 ± 0.9 |
| | 3D-IOUMatch [93] | 40.0 ± 0.9 | 46.3 ± 0.7 | 47.2 ± 0.4 | 49.6 ± 0.6 | 52.8 ± 1.2 | 62.9 ± 0.8 | 56.9 ± 1.2 | 63.7 ± 0.8 |
| | 3D-IOUMatch (Pr) [93] | 45.6 ± 2.0 | 52.8 ± 0.8 | 53.3 ± 1.5 | 55.6 ± 1.7 | 57.7 ± 1.2 | 59.6 ± 0.8 | 61.8 ± 1.6 | 66.9 ± 1.3 |
| | SPD [113] | 41.5 ± 0.5 | 44.3 ± 0.8 | 43.2 ± 1.2 | 46.2 ± 0.5 | 51.9 ± 0.5 | 55.6 ± 0.9 | 54.8 ± 0.5 | 58.6 ± 0.9 |
| | SPD [113] (Pr) | 45.2 ± 0.7 | 48.8 ± 1.9 | 52.8 ± 1.6 | 55.2 ± 0.7 | 53.8 ± 1.8 | 65.2 ± 2.5 | 59.9 ± 0.5 | 68.9 ± 1.6 |
| | *WS3D*++ (Ours) | 55.3 ± 1.1 | 57.6 ± 1.2 | 59.8 ± 0.8 | 66.8 ± 1.2 | 64.6 ± 0.5 | 72.8 ± 1.1 | 68.7 ± 0.5 | 76.6 ± 1.1 |
| | *WS3D*++ (Ours) (Pr) | 66.6 ± 1.5 | 68.1 ± 1.7 | 67.6 ± 1.6 | 74.5 ± 1.3 | 69.5 ± 0.9 | 76.9 ± 0.9 | 73.6 ± 0.7 | 77.8 ± 0.6 |

The mean average precision (mAP) given also with the mean ± standard deviation in three runs of diverse random splits are reported. Baseline is the circumstance which is trained with merely labelled data. It can be demonstrated our proposed pre-training approach have improved the performance of 3D object detection by a considerable margin.

We have reported the experimental results about semantic boundary prediction as demonstrated in Table I of our revised manuscript, it is demonstrated that our proposed approach can provide a very high-quality label at a very low-label regime, and the boundary prediction average precision can be kept at more than 50% (50.9% at least). The phenomenon has also been found at previous research including SQN, Lasermix, and our Weak-lab3DNet [59]. The reason behind is that: As point clouds are essentially samples of the 3D world, the distribution of points in a very close local neighborhood is comparatively homogeneous, revealing strong semantic similarity/homogeneity. Moreover, our proposed weakly supervised approach can be regarded as an amplification of those rather sparse supervision signals, which largely facilitates ultimate semantic boundary prediction. As we have demonstrated in Table I, the average precision (AP) of boundary prediction can still be maintained at a relatively high value (more than 50.9%). Our proposed approach can realize promising performance on outdoor benchmark SemanticKITTI and indoor benchmarks including S3DIS and ScanNet, which demonstrates the superior effectiveness of our proposed weakly supervised semantic boundary prediction.

### *D. Data-Efficient 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 [74] for the instance segmentation on ScanNet with results shown in Table VI. Notice that the performance

TABLE VI
COMPARISON OF THE PERFORMANCE OF **INSTANCE SEGMENTATION**, UNDER VARIOUS LEVELS OF SUPERVISION ON SCANNET VALIDATION SET

| Tested Dataset | Approaches | Ins. Seg. Results with the metric of AP@50% | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 15% | 20% | 30% | 35% | 40% | 100% |
| ScanNet | Merely the Pre-training Stage | 39.7 | 46.7 | 53.8 | 55.6 | 58.6 | 63.5 | 65.2 | 65.7 | 65.8 |
| | Sup-only-GPC–Baseline [104] | 10.8 | 33.6 | 42.8 | 45.3 | 48.2 | 49.0 | 49.5 | 50.2 | 56.8 |
| | Sup-only-GPC (Pr) [104] | 16.8 | 35.9 | 33.8 | 48.9 | 56.6 | 57.9 | 58.8 | 66.8 | 68.9 |
| | Mix3D [105] | 12.7 | 34.7 | 43.1 | 45.7 | 48.7 | 49.6 | 50.2 | 51.3 | 57.6 |
| | Mix3D (Pr) [105] | 20.6 | 39.7 | 48.9 | 50.9 | 53.9 | 58.9 | 58.7 | 55.2 | 63.9 |
| | GPC [104] | 16.9 | 38.6 | 44.9 | 47.2 | 48.6 | 49.7 | 51.2 | 52.0 | 57.7 |
| | GPC (Pr) [104] | 23.8 | 45.8 | 53.8 | 52.8 | 54.5 | 55.1 | 57.2 | 57.1 | 60.6 |
| | SPIB_Ins | 17.1 | 38.9 | 45.3 | 47.9 | 48.9 | 50.5 | 51.6 | 52.7 | 57.6 |
| | SPIB_Ins (Pr) | 23.7 | 52.8 | 52.9 | 55.3 | 54.8 | 55.8 | 56.9 | 58.8 | 64.8 |
| | *WS3D*++ (Ours) | **45.9** | **61.8** | **66.8** | **67.7** | **65.1** | **65.2** | **62.7** | **68.7** | **58.8** |
| | *WS3D*++ (Ours) (Pr) | **48.2** | **72.2** | **71.6** | **70.5** | **67.9** | **68.7** | **71.6** | **75.8** | **73.9** |

'Sup-only-GPC' denotes the model trained with only labeled data. *'WS3D'* denotes the model trained with our proposed methods. In the last row, we have shown the performance increase of *WS3D*. *WS3D*-Open abandons our feature-aligned contrastive pretraining with the contrastive loss term $L_{CL}^{total}$ and directly uses the CLIP [7] feature encoder with the knowledge distillation loss term $L_{KL}^{Dist}$. *WS3D*-Open is examined to validate the effect of hierarchical feature-aligned pretraining in improving data efficiency. The results of our proposed approach are highlighted in orange.

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. Our proposed *WS3D* and *WS3D++* both provide explicit boundary guidance for separating diverse kinds of semantic classes, and the instance segmentation performance with the very limited labeling percentage is comparable to those fully supervised counterparts.

### *E. Data-Efficient 3D Object Detection*

For the data-efficient 3D object detection, following our previous work [82], we extensively evaluate current approaches extensively on SUN RGB-D [111] and ScanNet [18] benchmarks for 3D object detection tasks with the strong 3D object

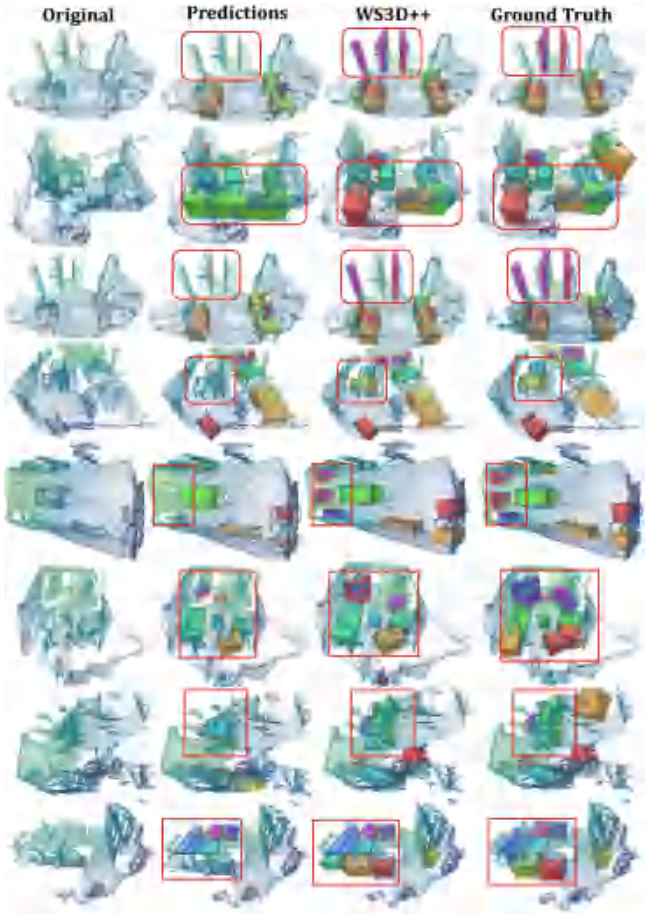


Fig. 7. The final ScanNet [18] object detection results and performance. The comparative differences in detection predictions are highlighted in the rectangles.


Fig. 8. The object detection comparisons on KITTI [99] validation set. It can be demonstrated that our proposed *WS3D++* can provide very accurate bounding box predictions qualitatively compared with the previous state-of-the-art merely using 2D bounding boxes, which demonstrates the effectiveness of using 3D regional as well as 3D object-level information in facilitating effective feature representation learning.

detection backbone VoteNet [115]. It can be demonstrated the open-vocabulary designs can to some extent boost the 3D object detection performance, which demonstrate the generalization capacity of our fundation model. As shown in Table V, we have also tested the ablated approach *WS3D-Open* which abandons our feature-aligned pre-training term $\mathcal{L}_{Ctr}^{total}$ and directly uses the CLIP [7] feature encoder for the knowledge distillation loss term with $\mathcal{L}_{\mathcal{KL}}^{Dist}$. It can be demonstrated that the performance degradation can be observed when comparing *WS3D-Open* with *WS3D++*, which demonstrate the effectiveness of our proposed hierarchical feature aligned pre-training in improving the data efficiency in downstream scene parsing.

### *F. Qualitative and Quantitative Results of the Open-World 3D Recognition Approaches*

In this Subsection, we further evaluate the performance of the open-world recognition capacity of our proposed approach. The results of open-world recognition are shown in Table I. We have also compared our work with the previous approach PLA [70] in establishing the sufficient point-language associations for the open-world robot learning. The results demonstrate that our proposed approach has superior performance in open-world recognition. We directly use the settings in the PLA [70] and split the categories on ScanNet [18] and Nuscene [36] into base and novel categories. It can also be validated that *WS3D-Open*, which abandons our feature-aligned pre-training and directly use the CLIP [7] feature encoder, provides slightly inferior performance compared with *WS3D++*, validating the effectiveness of our language-3D matching strategy designs. *WS3D++* exhibits superior performance in terms of the open-vocabulary few-shot learning for diverse partitioning of original and novel classes. The open-world recognition results are shown in Figs. 9 and 10. It can be demonstrated that better foreground object awareness can be effectively capture by our proposed *WS3D++* compared with PLA [70], with superior segmentation performance guided by the textual prompts. The superior open-world recognition performance can be achieved while conducting open-world learning in diverse spliting of based and novel classes, including B15/N4, B12/N7, B10/N9 for the ScanNet [18] as well as B12/N3 and B10/N5 for the NuScenes [36]. It demonstrates robustness of our proposed approach. Also, as demonstrated in Fig. 11, the *WS3D++* language driven-3D scene segmentation results are very precise as shown qualitatively, which is corresponding to the object queried by the language, and it demonstrates that the inference can be done based on the object,

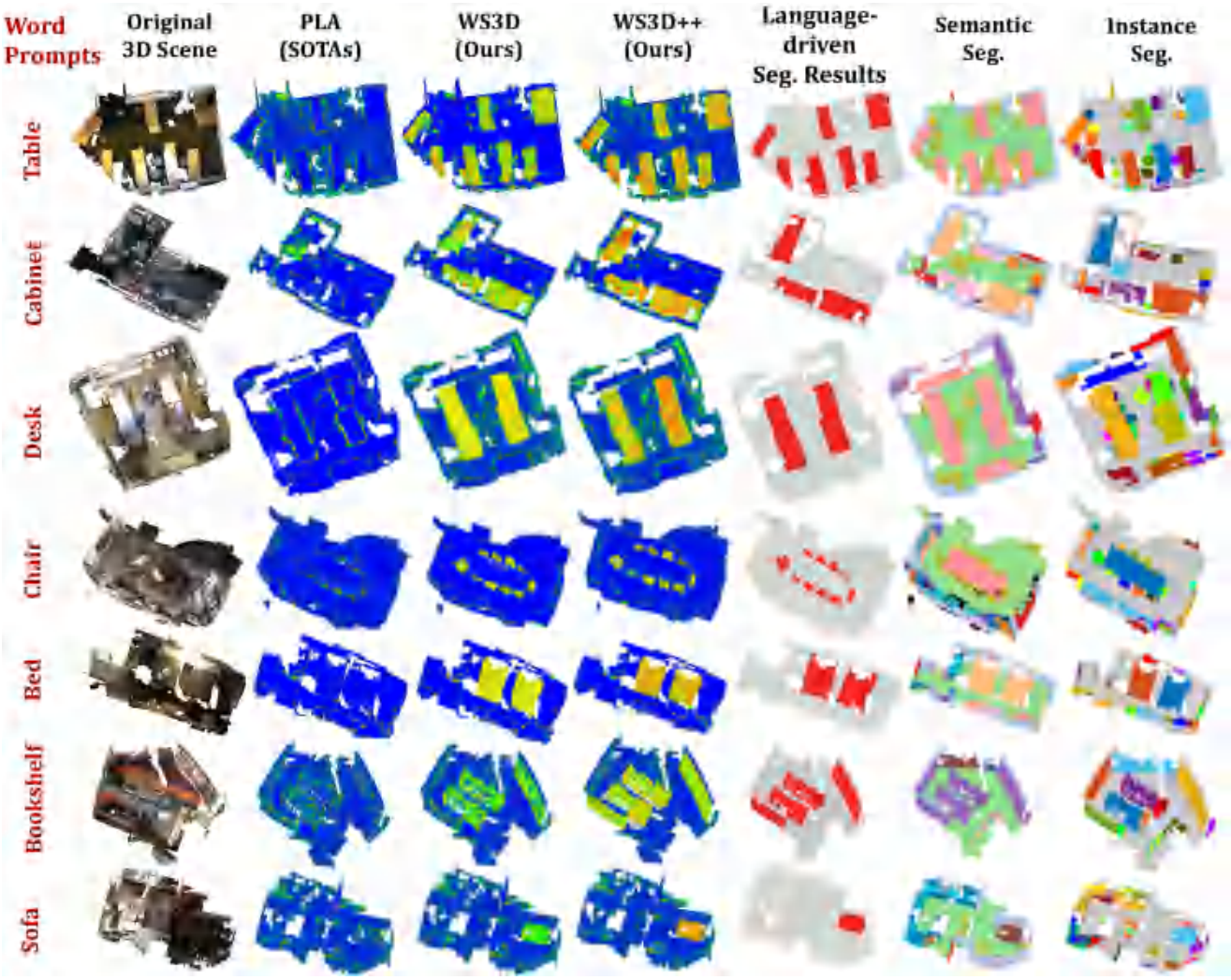


Fig. 9. Segmentation result comparisons with CLIP prompts for the indoor ScanNet benchmark. It can be seen from our results that clear object-level vision-language matched information can be captured with the designed visual prompts. It reveals the effectiveness of our designed hierarchical visual linguistic feature-aligned representation learning approaches.

material, properties, affordance, room type, etc. It demonstrates that our proposed *WS3D++* can enable the scene-level object recognition based on the semantic language queries. As further shown in Fig. 13, through our effective rendering techniques, which establish the explicit 2D-3D association, the aligned representation of 2D-3D-language co-embeddings can be learned and the object information can also be enhanced through finding the similarity among diverse views through contrastive learning approaches. Also, by combining 2D and 3D region proposals, more complete and apparent object-level information can be clearly captured both from 2D views and 3D views. It turns out that in the first place, while initializing other various weakly supervised approaches, our proposed approach can realize consistent improvement on the final performance of weakly supervised scene parsing, which demonstrates the superior generalization capacity benefiting from our designed pre-training of our proposed WS3D++ framework. In the second pace, the performance of our proposed WS3D++ is comparatively superior compared with the existing weakly supervised comparative approaches listed above.

### G. Instance Discrimination Capacity

We show t-SNE visualizations of the learned latent feature representations for various semantic classes in Fig. 12. The case study task is the semantic segmentation on the S3DIS dataset with a supervision level of 5%. It is demonstrated that more distinctive and better separated point-wise feature embeddings are provided by our proposed unsupervised region-level contrastive learning, which can be attributed to its strong instance discrimination capacity. And more separated feature space can be provided and maintained with our proposed *WS3D++* compared to *WS3D* and GPC. This strong instance discrimination capacity can be explained by more discriminative feature representations guided by 3D vision-language aligned representations, and is thus more beneficial to high-level semantic and instance segmentation performances both in terms of data efficiency and open-world recognition capacity. Also, our proposed hierarchical feature alignment also provides more separated feature space, which means that the feature alignment successfully enhances the final instance discrimination capacity.

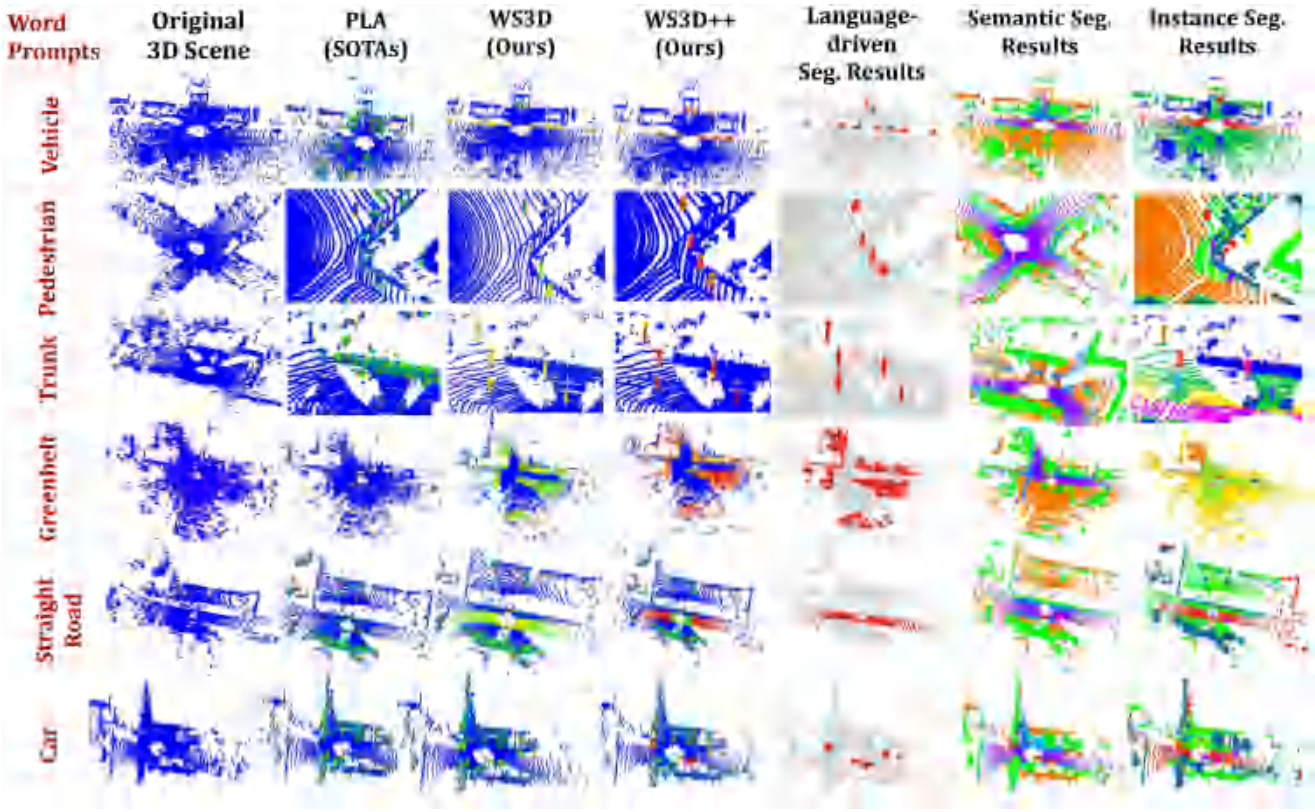


Fig. 10. The segmentation result visualizations and comparisons with CLIP text encoder prompts for the outdoor KITTI benchmark. It can be demonstrated the final foreground object awareness can be clearly captured as compared with the previous SOTAs approach PLA [70]. Meanwhile, as shown in the last three columns, we can provide clear segmentation for the corresponding visual objects based on the textual prompts. The results further demonstrate that vision-language aligned representations can be effectively and sufficiently learnt.

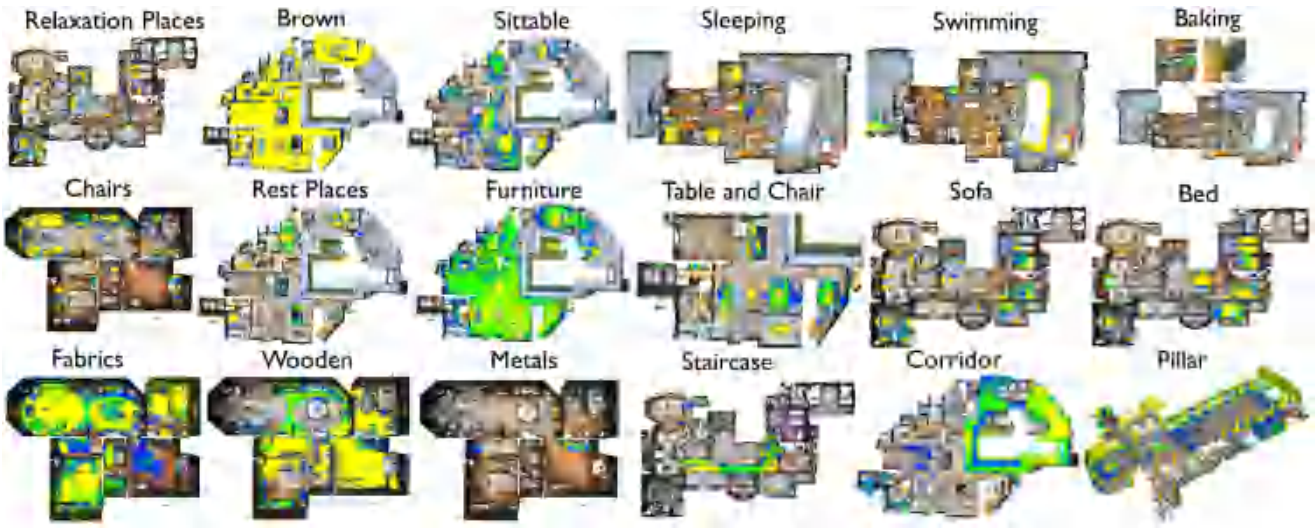


Fig. 11. The final 3D scene-level activations results based on the language query of *WS3D++*. Better zoom in for details.

*Robot Arm Grasping Example:* We have deployed our approach for the task of open-world perception in robot grasping in our extended robotic research work. We use our proposed approach for segmentation and use the ROS 2 Gazebo-based framework to implement the other components of the system, such as kinematics/dynamics modeling, motion planning, low-level control, point cloud-based pose estimation, etc. Our proposed approach has robust performance and decent accuracy in grasping, which demonstrates the potential of our proposed *WS3D++* in industrial manipulation applications in grasping and dropping novel objects beyond the training set.

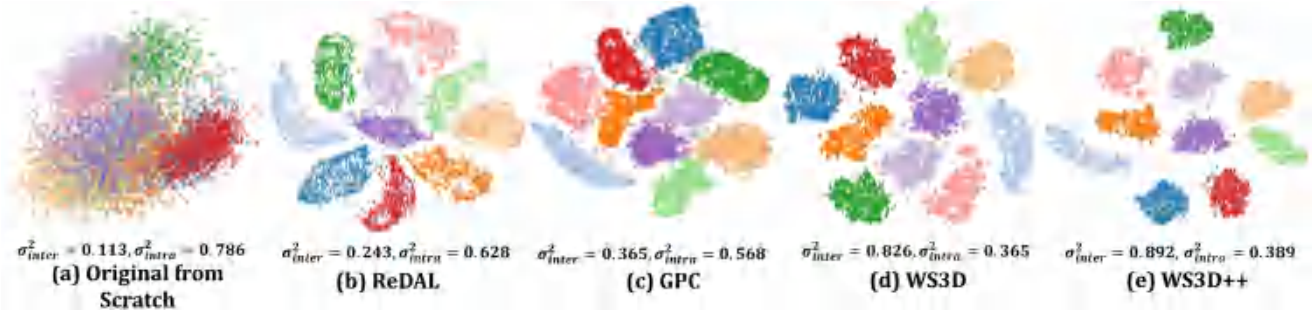


Fig. 12. t-SNE visualization in semantic segmentation of the proposed *WS3D* under the 5% labeling percentage on S3DIS validation set. The diverse feature embeddings are indicated by different colors and are normalized into $[-1, 1]$ for better visualization. It has been validated that more discriminative features can be acquired for diverse semantic classes with our proposed unsupervised region-aware fine-tuning strategy (demonstrated by *WS3D*) and our proposed hierarchical vision-language knowledge associated and distilled pre-training (demonstrated by *WS3D++*).


Fig. 13. The *WS3D++* corresponding detection and segmentation projections on six rendered 2D views through our proposed multi-view rendering approach. It can be demonstrated that our proposed rendering has established an explicit modality association between the final 2D views and 3D views. Also, by combining 2D and 3D region proposals, more complete and apparent object-level information can be clearly captured both from 2D views and 3D views. Best zoom-in for viewing.

## V. Conclusion

In this paper, we proposed a general *WS3D++* framework for *open-vocabulary* and *data-efficient* 3D scene parsing. The whole framework involves both the pre-training and the fine-tuning stages. During the pre-training stage, we propose the hierarchical feature alignment strategy to acquire accurate regional 3D-linguistic pairs, thus the performance can be enhanced to a large extent. At the same time, 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. Our approach can maximally exploit the unlabeled data to enhance the performance both for 3D point clouds semantic segmentation and instance segmentation, and object detection under various labeling percentages in the limited reconstruction case. Our proposed label-efficient learning framework, termed *WS3D++*, provides conprehensive baselines for future 3D scene parsing methods when the label is inaccessible or limited. The proposed pre-training as well as fine-tuning approach can have a significant boost on the final open-vocabulary and data-efficient semantic scene parsing in term of efficiency, effectiveness, and robustness.

## REFERENCES


[1] J. Behley et al., "Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The semanticKITTI dataset," *Int. J. Robot. Res.*, vol. 40, pp. 959–967, 2021.

[2] S. Ao, Q. Hu, H. Wang, K. Xu, and Y. Guo, "Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1255–1264.

[3] Z. Zhang, B. Yang, B. Wang, and B. Li, "GrowSP: Unsupervised semantic segmentation of 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17619–17629.

[4] K. Liu, Y. Zhao, Q. Nie, Z. Gao, and B. M. Chen, "Weakly supervised 3D scene segmentation with region-level boundary awareness and instance discrimination," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 37–55.

[5] Z. Song and B. Yang, "OGC: Unsupervised 3D object segmentation from rigid dynamics of point clouds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 30798–30812.

[6] D. Rozenberszki, O. Litany, and A. Dai, "Unscene3D: Unsupervised 3D instance segmentation for indoor scenes," 2023, *arXiv:2303.14541*.

[7] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[8] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.

[9] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," 2023, *arXiv:2305.03726*.

[10] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 23033–23044.

[11] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 696–712.

[12] K. Zheng et al., "Regularized mask tuning: Uncovering hidden knowledge in pre-trained vision-language models," 2023, *arXiv:2307.15049*.

[13] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[14] X. Zou et al., "Segment everything everywhere all at once," 2023, *arXiv:2304.06718*.

[15] J. Gong et al., "Omni-supervised point cloud segmentation via gradual receptive field component reasoning," in *Proc. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11673–11682.

[16] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9297–9307.

[17] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "2-S3Net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12547–12556.

[18] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.

[19] I. Armeni et al., "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1534–1543.

[20] J. Hou, B. Graham, M. Nießner, and S. Xie, "Exploring data-efficient 3D scene understanding with contrastive scene contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15587–15597.

[21] W. Shen et al., "A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9284–9305, Aug. 2023.

[22] P.-C. Yu, C. Sun, and M. Sun, "Data efficient 3D learner via knowledge transferred from 2D model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 182–198.

[23] P. Hu, S. Sclaroff, and K. Saenko, "Leveraging geometric structure for label-efficient semi-supervised scene segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 6320–6330, 2022.

[24] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16291–16301.

[25] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Pointcontrast: Unsupervised pre-training for 3D point cloud understanding," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 574–591.

[26] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," 2019, *arXiv: 1906.04651*.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[28] S. Rong, B. Tu, Z. Wang, and J. Li, "Boundary-enhanced co-training for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19574–19584.

[29] T. Feng, W. Wang, X. Wang, Y. Yang, and Q. Zheng, "Clustering based point cloud representation learning for 3D analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8283–8294.

[30] K. Liu, A. Xiao, X. Zhang, S. Lu, and L. Shao, "FAC: 3D representation learning via foreground aware feature contrast," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9476–9485.

[31] L. Wiesmann, L. Nunes, J. Behley, and C. Stachniss, "KPPR: Exploiting momentum contrast for point cloud-based place recognition," *IEEE Robot. Automat. Lett.*, vol. 8, no. 2, pp. 592–599, Feb. 2023.

[32] B. Pang, H. Xia, and C. Lu, "Unsupervised 3D point cloud representation learning by triangle constrained contrast for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5229–5239.

[33] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 1–4.

[34] X. Zhu et al., "Cylindrical and asymmetrical 3D convolution networks for LiDAR-based perception," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6807–6822, Oct. 2022.

[35] S. Qiu, S. Anwar, and N. Barnes, "Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1757–1767.

[36] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.

[37] Z. Que, G. Lu, and D. Xu, "Voxelcontext-net: An octree based framework for point cloud compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6042–6051.

[38] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal convnets: Minkowski convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3075–3084.

[39] J. Noh, S. Lee, and B. Ham, "HVPR: Hybrid voxel-point representation for single-stage 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14605–14614.

[40] J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, "Towards efficient 3D object detection with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 21300–21313.

[41] T. Vu, K. Kim, T. M. Luu, T. Nguyen, J. Kim, and C. D. Yoo, "SoftGroup: Scalable 3D instance segmentation with octree pyramid grouping," 2022, *arXiv:2209.08263*.

[42] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. 2019 Int. Conf. Robot. Automat.*, 2019, pp. 4376–4382.

[43] C. Xu et al., "SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation," 2020, *arXiv: 2004.01803*.

[44] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, "GVCNN: Group-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 264–272.

[45] A. Kundu et al., "Virtual multi-view fusion for 3D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 518–535.

[46] L. Li, S. Zhu, H. Fu, P. Tan, and C.-L. Tai, "End-to-end learning local multi-view descriptors for 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1919–1928.

[47] Z. Gojcic, C. Zhou, J. D. Wegner, L. J. Guibas, and T. Birdal, "Learning multiview 3D point cloud registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1759–1769.

[48] L. Landrieu and M. Boussaha, "Point cloud oversegmentation with graph-structured deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7440–7449.

[49] Z. Zhang, B.-S. Hua, and S.-K. Yeung, "ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1607–1616.

[50] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5589–5598.

[51] Y. Liu et al., "DensePoint: Learning densely contextual representation for efficient point cloud processing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5239–5248.

[52] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.

[53] S. Ao, Q. Hu, B. Yang, A. Markham, and Y. Guo, "SpinNet: Learning a general surface descriptor for 3D point cloud registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11753–11762.

[54] H. Fan, X. Yu, Y. Ding, Y. Yang, and M. Kankanhalli, "PSTNet: Point spatio-temporal convolution on point cloud sequences," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[55] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3D point clouds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3664–3680, Oct. 2021.

[56] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel CNN for efficient 3D deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 965–975.

[57] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3D point clouds with low overlap," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4267–4276.

[58] K. Liu, Z. Gao, F. Lin, and B. M. Chen, "FG-CONV: Large-scale LiDAR point clouds understanding leveraging feature correlation mining and geometric-aware modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 12896–12902.

[59] K. Liu, Y. Zhao, Z. Gao, and B. M. Chen, "Weaklabel3D-Net: A complete framework for real-scene LiDAR point clouds weakly supervised multi-tasks understanding," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 5108–5115.

[60] B. Graham, M. Engelcke, and L. Van der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9224–9232.

[61] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "LargeKernel3D: Scaling UP kernels in 3D sparse CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13488–13498.

[62] J. Liu, Y. Chen, X. Ye, Z. Tian, X. Tan, and X. Qi, "Spatial pruned sparse convolution for efficient 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 6735–6748.

[63] A. Thabet, H. Alwassel, and B. Ghanem, "Self-supervised learning of local features in 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 938–939.

[64] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner, "Unsupervised point cloud pre-training via occlusion completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9782–9792.

[65] B. Eckart, W. Yuan, C. Liu, and J. Kautz, "Self-supervised learning on 3D point clouds by learning discrete generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8248–8257.

[66] X. Wu, X. Wen, X. Liu, and H. Zhao, "Masked scene contrast: A scalable framework for unsupervised 3D representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9415–9424.

[67] R. Zhang et al., "PointCLIP: Point cloud understanding by CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8552–8562.

[68] X. Zhu et al., "PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2639–2650.

[69] L. Xue et al., "ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1179–1189.

[70] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "PLA: Language-driven open-vocabulary 3D scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7010–7019.

[71] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Lowis3D: Language-driven open-world instance-level 3D scene understanding," 2023, *arXiv:2308.00353*.

[72] J. Yang, R. Ding, Z. Wang, and X. Qi, "RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding," 2023, *arXiv:2304.00962*.

[73] Q. Hu et al., "Randla-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11108–11117.

[74] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "PointGroup: Dual-set point grouping for 3D instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4867–4876.

[75] R. Ding, J. Yang, L. Jiang, and X. Qi, "DODA: Data-oriented sim-to-real domain adaptation for 3D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 284–303.

[76] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10529–10538.

[77] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2647–2664, Aug. 2021.

[78] S. Shi et al., "PV-RCNN : Point-voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 531–551, 2023.

[79] Q. Hu et al., "SQN: Weakly-supervised semantic segmentation of large-scale 3D point clouds with 1000x fewer labels," 2021, *arXiv:2104.04891*.

[80] H. Wang, X. Rong, L. Yang, J. Feng, J. Xiao, and Y. Tian, "Weakly supervised semantic segmentation in 3D graph-structured point clouds of wild scenes," 2020, *arXiv: 2004.12498*.

[81] J. Wei, G. Lin, K.-H. Yap, T.-Y. Hung, and L. Xie, "Multi-path region mining for weakly supervised 3D semantic segmentation on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4384–4393.

[82] K. Liu, "Rm3d: Robust data-efficient 3D scene parsing via traditional and learnt 3D descriptors-based semantic region merging," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 938–967, 2022.

[83] Z. Liu, X. Qi, and C.-W. Fu, "One thing one click: A self-training approach for weakly supervised 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1726–1736.

[84] Z. Liu, X. Qi, and C.-W. Fu, "One thing one click: Self-training for weakly supervised 3D scene understanding," 2023, *arXiv:2303.14727*.

[85] X. Xu and G. H. Lee, "Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13706–13715.

[86] M. Gadelha et al., "Label-efficient learning on point clouds using approximate convex decompositions," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 473–491.

[87] R. Li, A.-Q. Cao, and R. de Charette, "Coarse3D: Class-prototypes for contrastive learning in weakly-supervised 3D point cloud segmentation," 2022, *arXiv:2210.01784*.

[88] L. Liu et al., "CPCM: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation," 2023, *arXiv:2307.10316*.

[89] M. Xu et al., "MM-3DScene: 3D scene understanding by customizing masked modeling with informative-preserved reconstruction and self-distilled consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4380–4390.

[90] K. M. Jatavallabhula et al., "Conceptfusion: Open-set multimodal 3D mapping," *Robot.: Sci. Syst.*, 2023, pp. 1–17.

[91] A. Cheraghian, S. Rahman, D. Campbell, and L. Petersson, "Transductive zero-shot learning for 3D point cloud classification," in *Proc. IEEE/CVF winter Conf. Appl. Comput. Vis.*, 2020, pp. 923–933.

[92] B. Michele, A. Boulch, G. Puy, M. Bucher, and R. Marlet, "Generative zero-shot learning for semantic segmentation of 3D point clouds," in *Proc. 2021 Int. Conf. 3D Vis.*, 2021, pp. 992–1002.

[93] H. Wang, Y. Cong, O. Litany, Y. Gao, and L. J. Guibas, "3Dioumatch: Leveraging IoU prediction for semi-supervised 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14615–14624.
[94] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *Int. J. Comput. Inf. Sci.*, vol. 9, no. 3, pp. 219–242, 1980.
[95] D.-T. Lee, "On k-nearest neighbor Voronoi diagrams in the plane," *IEEE Trans. Comput.*, vol. 100, no. 6, pp. 478–487, Jun. 1982.
[96] P. Cignoni, C. Montani, and R. Scopigno, "DeWall: A fast divide and conquer delaunay triangulation algorithm in ed," *Comput.-Aided Des.*, vol. 30, no. 5, pp. 333–341, 1998.
[97] D. Shreiner et al., *OpenGL Programming Guide: The Official Guide to Learning OpenGL*. London, England: Pearson Education, 2009.
[98] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[99] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[100] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[101] S. Shi, X. Wang, and H. Li, "PointrCNN: 3D object proposal generation and detection from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.
[102] Y. Xie et al., "Differentiable top-k with optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 20520–20531.
[103] J. Hou, A. Dai, and M. Niessner, "3D-SIS: 3D semantic instance segmentation of RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4416–4425.
[104] L. Jiang et al., "Guided point contrastive learning for semi-supervised point cloud semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6423–6432.
[105] A. Nekrasov, J. Schult, O. Litany, B. Leibe, and F. Engelmann, "Mix3D: Out-of-context data augmentation for 3D scenes," in *Proc. 2021 Int. Conf. 3D Vis.*, 2021, pp. 116–125.
[106] Y. Chen et al., "PointMixup: Augmentation for point clouds," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 330–345.
[107] Y. Lan et al., "Weakly supervised 3D segmentation via receptive-driven pseudo label consistency and structural consistency," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1222–1230.
[108] G. Liu, O. van Kaick, H. Huang, and R. Hu, "Active self-training for weakly supervised 3D scene semantic segmentation," 2022, *arXiv:2209.07069*.
[109] M. Liu, Y. Zhou, C. R. Qi, B. Gong, H. Su, and D. Anguelov, "Less: Label-efficient semantic segmentation for LiDAR point clouds," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 70–89.
[110] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[111] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567–576.
[112] N. Zhao, T.-S. Chua, and G. H. Lee, "SESS: Self-ensembling semi-supervised 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11079–11087.
[113] B. Xie et al., "SPD: Semi-supervised learning and progressive distillation for 3-D detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3503–3513, Mar. 2024.
[114] Y. Liao, H. Zhu, Y. Zhang, C. Ye, T. Chen, and J. Fan, "Point cloud instance segmentation with semi-supervised bounding-box mining," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10159–10170, Dec. 2022.
[115] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9277–9286.

**Kangcheng Liu** (Member, IEEE) received the BEng degree from the Harbin Institute of Technology and the PhD degree from the Chinese University of Hong Kong with honors. He is a full professor with Hunan University, China. Before that, he was a senior scientist and research associate professor with the California Institute of Technology, serving as the leading Principal investigator or the technical lead of several projects in China and America. His interests are swarm robot systems, swarm intelligence, advanced optics and circuit systems, and artificial intelligence. He has been nominated to serve as the program committee member, associate editor, and reviewer of international flagship journals and conferences: Nature Machine Intelligence, Nature Medicine, Artificial Intelligence Journal, International Journal of Computer Vision, International Journal of Robotics Research, Journal of Machine Learning Research, ACM Transactions on Graphics, ICRA, ICPR, ICIP, and CASE. He has published more than 50 works in journals including IJCV and T-PAMI as the first author and the corresponding author. His work has also published as special nomination and won the best paper candidate award in International Journal of Computer Vision, IEEE International Conference on Robotics and Automation, IEEE International Conference on Intelligent Robots and Systems and IEEE International Conference on Advanced Robotics and Mechatronics. He served as the reviewer of above 90 kinds of journals and conferences such as Science Robotics, Nature Machine Intelligence, IJCV, T-PAMI, T-RO, T-CYB, JFR, AIJ, IJRR, CVPR, ICCV, ECCV, TOG, and ACM SIGGRAPH.

**Yong-Jin Liu** (Senior Member, IEEE) received the BEng degree from Tianjin University, Tianjin, China, in 1998, and the MPhil and PhD degrees from the Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is currently a professor with the BNRist, MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing, China.

**Baoquan Chen** (Fellow, IEEE) received the MS degree in electronic engineering from Tsinghua University, Beijing, China, and the 2nd MS and PhD degrees in computer science from the State University of New York at Stony Brook, New York, USA. He is currently an Endowed Boya Professor with the National Key Lab of Genaral AI, Peking University, where he is the associate dean of the School of Artificial Intelligence.