International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02340-z

Generalized Robot Vision-Language Model via Linguistic Foreground-Aware Contrast

Kangcheng Liu[1,2] · **Chaoqun Wang**[3] · **Xiaodong Han**[4] · **Yong-Jin Liu**[5] · **Baoquan Chen**[6]

Received: 30 November 2023 / Accepted: 24 December 2024

Abstract

Contrastive learning has recently demonstrated great potential for unsupervised pre-training in 3D scene understanding tasks. However, most existing work randomly selects point features as anchors while building contrast, leading to a clear bias toward background points that often dominate in 3D scenes. Also, object awareness and foreground-to-background discrimination are neglected, making contrastive learning less effective. To tackle these issues, we propose a general foreground-aware feature contrast FAC++ framework to learn more effective point cloud representations in pre-training. FAC++ consists of two novel contrast designs to construct more effective and informative contrast pairs. The first is building positive pairs within the same foreground segment where points tend to have the same semantics. The second is that we prevent over-discrimination between 3D segments/objects and encourage grouped foreground-to-background distinctions at the segment level with adaptive feature learning in a Siamese correspondence network, which adaptively learns feature correlations within and across point cloud views effectively. Our proposed approach enhances both the local coherence as well as the overall feature discrimination. Moreover, we have designed the linguistic foreground-aware regional point sampling to enhance more balanced foreground-aware learning, which is termed FAC++. Visualization with point activation maps shows that our contrast pairs capture clear correspondences among foreground regions during pre-training. Quantitative experiments also show that FAC++ achieves superior knowledge transfer and data efficiency in various downstream 3D semantic segmentation, instance segmentation as well as object detection tasks. All codes, data, and models are available at: (https://github.com/KangchengLiu/FAC_Foreground_Aware_Contrast).

Keywords Self-supervised learning · Vision-language models · Representation learning · Data-efficient learning · 3D vision

1 Introduction

Understanding 3D scenes is crucial for many tasks such as grasping robots, embodied robotic control and planning, smart manufacturing, virtual reality/augmented reality, and autonomous navigation (Huang et al. 2018; Liu et al. 2017; Liu et al. 2022g; Liu et al. 2022h; Liu et al. 2022c; Liu 2022d). Moreover, the data-efficient vision language model plays an important role in embodied intelligence, promoting label-efficient scene parsing (Liu and Cao 2023, Liu 2023d, Liu 2022d, Liu 2022e). However, most existing work is fully supervised, which relies heavily on large-scale annotated 3D data that is often very laborious to collect. Self-supervised learning (SSL), which allows learning rich and meaningful representations from large-scale unannotated data, has

Communicated by Minsu Cho.

The preliminary results of this work are published at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR 2023).

✉ Kangcheng Liu
kcliu@mae.cuhk.edu.hk

[1] The College of Electrical and Information Engineering, Hunan University (HNU), Changsha, China

[2] The Division of Engineering and Applied Science, California Institute of Technology (Caltech), Pasadena, USA

[3] The School of Control Science and Engineering, Shandong University, Jinan, China

[4] The School of Control Engineering, Minjiang University, Fuzhou, China

[5] The Department of Computer Science and Technology, Tsinghua University, Beijing, China

[6] The School of Artificial Intelligence, Peking University, Beijing, China

Published online: 14 January 2025

recently demonstrated great potential to mitigate the annotation constraint. It learns with auxiliary supervision signals derived from unannotated data, which are usually much easier to collect. In particular, contrastive learning as one prevalent SSL approach has achieved great success in various visual downstream 2D recognition tasks.

Contrastive learning has also been explored for robot point cloud-based representation learning in various downstream tasks such as semantic segmentation (Xie et al., 2020a), instance segmentation (Hou et al., 2021), and object detection (Yin et al., 2022). However, many successful 2D contrastive learning methods do not work well for 3D point clouds, largely because point clouds often capture wide-view scenes which consist of complex points of many irregularly distributed foreground objects as well as a large number of background points. Several studies attempt to design specific contrast to cater to the geometry and distribution of point clouds. For example, Huang et al. (2021) employs max-pooled features of two augmented scenes to form the contrast, but they tend to over-emphasize holistic information and overlook informative features about foreground objects. Xie et al. (2020a), Hou et al. (2021), Liang et al. (2021) directly use registered point/voxel features as positive pairs and treat all non-registered as negative pairs, causing many false contrast pairs in semantics.

We propose exploiting scene foreground evidence and foreground-background distinction to construct more *foreground grouping aware* and *foreground-background distinction aware* contrast for learning discriminative 3D representations. For *foreground grouping aware* contrast, we first obtain regional correspondences with over-segmentation (Papon et al., 2013) and then build positive pairs with points of the same region across views, leading to semantic coherent representations. In addition, we design a sampling strategy to sample more foreground point features while building contrast, because the background point features are often less-informative and have repetitive or homogeneous patterns. For *foreground-background distinction aware* contrast, we first enhance foreground-background point feature distinction and then design a Siamese correspondence network that selects correlated features by adaptively learning affinities among feature pairs within and across views in both foreground and background to avoid over-discrimination between parts/objects. Visualizations show that, in a complementary manner, foreground-enhanced contrast effectively guides the learning toward foreground regions while foreground-background contrast enhances distinctions among foreground and background features. The two designs collaborates to learn more informative and discriminative representation as illustrated in Fig. 1.

This work is a significant extension of the preliminary version of the published conference work (Liu et al., 2023), where basic ideas of forming contrast between

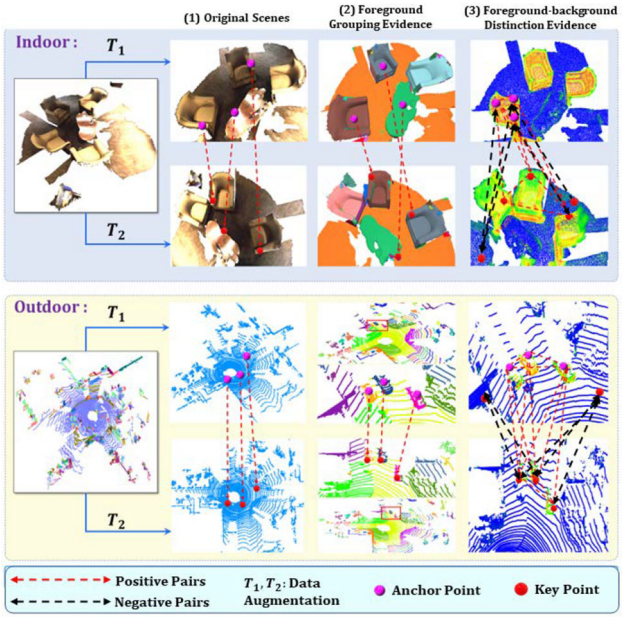

Fig. 1 FAC++ takes both the foreground grouping and the foreground-background distinction guidance into account, thus forming better contrast pairs to learn more informative and discriminative 3D feature representations. FAC++ provides more language-aware information with the 3D visual-linguistic aligned prompts

grouped foreground and background are introduced to tackle the final data-efficient 3D scene understanding during the fine-tuning stage to enhance the foreground grouping and foreground-background distinction. The core ideas are infuse 3D foreground linguistic information through 2D vision-language aligned prompts. As demonstrated by our extensive experiments, our proposed framework can provide linguistic foreground-aware feature contrast, which provides valuable guidance from the vision-language models. In summary, we extensively enriched previous works in the following aspects:

First, we propose to enhance the foreground-background discrimination directly with explicit foreground points queries of FAC/FAC++ from the 3D vision-language models, which achieved more generalized pre-training for representation learning, and enhance the final data-efficient learning as well as the open-vocabulary recognition performance.

Second, we added the experiments in instance segmentation, where we have achieved superior performance for that the superior foreground-aware instance discrimination in feature representation learning is successfully achieved.

Third, we added comprehensive experimental about the open-world recognition, which demonstrates apart from state-of-the-art performance in data-efficient learning, our proposed approach is also compatible with the current state-of-the-art 3D open vocabulary recognition approaches, and also has superior effectiveness in terms of recognizing novel categories and novel semantic classes.

Fourth, we added very comprehensive experimental comparative analysis about the efficiency as well as the training/inference time analysis of our proposed FAC++, which demonstrates our proposed framework has relatively low computational complexity thanks to our proposed regional contrastive designs rather than conducting contrast point-level. The proposed method can be seamlessly integrated to the current state-of-the-art vision-language models. In the meanwhile, our proposed can work harmoniously with the prevailing State-of-the-art backbone network approaches, including SparseConv as well as PV-RCNN. It demonstrates superior data-efficiency, training efficiency, as well as compatibility of our proposed framework.

The contributions of this work can be summarized in the following aspects. *First,* we propose FAC/FAC++, a foreground-aware feature contrast framework for large-scale 3D pre-training. FAC samples median-sized regions as foreground regions, while FAC++ leverages foreground prompts to enhance the foreground-aware feature representations. *Second,* we construct region-level contrast to enhance the local coherence and better foreground awareness in the learned representations. *Third,* on top of that, we design a Siamese correspondence framework that can locate well-matched keys to adaptively enhance the intra- and inter-view feature correlations, as well as enhance the foreground-background distinction. *Fourthly,* we propose leveraging current prevailing vision-language models to extend the model's generalization capacity while encountered with novel categories, and demonstrate the open-world recognition capacity of the model by extensive experiments. *Lastly,* extensive experiments over multiple public benchmarks show that FAC++ achieves superior self-supervised learning when compared with the state-of-the-art. FAC++ is compatible with the prevalent 3D segmentation backbone network SparseConv and 3D detection backbone networks including PV-RCNN, PointPillars (Lang et al., 2019), and Point-RCNN (Shi et al., 2019). It is also applicable to both indoor dense RGB-D and outdoor sparse LiDAR point clouds. Therefore, the proposed framework has been demonstrated very effective in constructing a generalized robotic vision-language learning model leveraging linguistic foreground aware contrast.

2 Related Work

2.1 3D Scene Understanding

The intelligence of the robot embodied is highly dependent on the powerful parsing and perception capacity (Liu 2023a, 2023c, 2023b; Liu and Ou 2022a; Liu and Ou 2022b; Liu 2022a; Liu et al. 2022f; Liu 2022c). Understanding the 3D scene aims at understanding the depth of the 3D or point cloud data and involves several downstream tasks such as 3D semantic segmentation (Chibane et al., 2022; Yang et al., 2022), 3D object detection (Erçelik et al., 2022), etc. It has recently achieved very impressive progress as driven by the advance in 3D deep learning strategy and the increasing large-scale 3D benchmark datasets (Liu and Chen 2022b; Liu et al. 2024; Liu et al. 2022a; Liu et al. 2022b; Liu et al. 2022d; Liu and Chen 2022a). Different approaches have been proposed to address various challenges in 3D scene understanding. For example, the point-based approach (Liu et al., 2022e, 2019) can learn point features well but is often stuck by high computational costs while facing large-scale point-cloud input stream. Voxel-based approach (Mao et al., 2021; Zhou & Tuzel, 2018; Liu, 2022b; Liu et al., 2022d, 2021b) is computation and memory efficient but often suffers from information loss from the voxel quantification. In addition, voxel-based SparseConv network (Vu et al., 2022) has shown very promising performance in indoor 3D scene parsing, while combining point and voxel often has a clear advantage in outdoor LiDAR-based detection (Lang et al., 2019; Zhou & Tuzel, 2018; Shi et al., 2020). Our proposed SSL framework shows consistent superiority in indoor/outdoor 3D perception tasks, and it is also backbone-agnostic.

2.2 Self-Supervised Pre-training on Point Clouds

2.2.1 Contrastive Pre-training

Recent years have witnessed notable success in contrastive learning for learning unsupervised representations (Li & Heizmann, 2022; Wang et al., 2021b; Zhang et al., 2022; Sanghi, 2020). For example, contrast scene context (CSC) (Hou et al., 2021; Xie et al., 2020a) explores contrastive pre-training with scene context descriptors. However, it focuses too much on optimizing low-level registered point features but neglects the regional homogeneous semantic patterns and high-level feature correlations. Some work employs max-pooled scene-level information for contrast (Huang et al., 2021; Liang et al., 2021), but it tends to sacrifice local geometry details and object-level semantic correlations, leading to sub-optimal representations for dense prediction tasks such as semantic segmentation. Differently, we explicitly consider regional foreground awareness as well as feature correlation and distinction among foreground and background regions which lead to more informative and discriminative representations in 3D downstream tasks.

Further, many approaches incorporate auxiliary temporal or spatial 3D information for self-supervised contrast with augmented unlabeled datasets (Huang et al., 2021) and synthetic CAD models (Chen et al., 2022). STRL (Huang et al., 2021) introduces a mechanism of learning from dynamic 3D scenes synthetic 3D by regarding 3D scenes are RGB-D video sequences. Randomrooms (Rao et al., 2021) synthe-

sizes man-made 3D scenes by randomly putting synthetic CAD models into regular synthetic 3D scenes. 4DContrast (Chen et al., 2022) leverages spatio-temporal motion priors of synthetic 3D shapes to learn a better 3D representation. However, most of these prior studies rely on extra supervision from auxiliary spatio-temporal information. Differently, we perform self-supervised learning over original 3D scans without additional synthetic 3D models.

2.2.2 Masked Generation-Based Pre-training

Masked image modeling has demonstrated its effectiveness in various image understanding tasks (Xie et al., 2022; He et al., 2022) with the success of vision transformers (Liu et al., 2021a; Carion et al., 2020). Recently, mask-based pre-training (Liang et al., 2022; Pang et al., 2022; Wang et al., 2021b) has also been explored for the understanding of small-scale 3D shapes (Uy et al., 2019). Confidence-based discrimination has also been shown to be effective in semi-supervised learning and confidence-level-based discriminative pre-training (Liu, 2022f). However, mask-based designs usually involve a transformer backbone (Liang et al., 2022; Pang et al., 2022) that has a high demand for both computation and memory while handling large-scale 3D scenes. We focus on pre-training with contrastive learning, which is compatible with both point-based and voxel-based backbones.

3 Method

As illustrated in Fig. 2, our proposed FAC/FAC++ frameworks are composed of four components: data augmentation, backbone network feature extraction, feature matching, and foreground-background aware feature contrastive optimizations with matched contrast pairs. The differences of them merely lie in that FAC samples median-sized regions as foreground regions, while FAC++ leverages foreground prompts to enhance the foreground-aware feature representations. In the following, we first revisit typical contrastive learning approaches for 3D point clouds and discuss their limitations that could lead to less informative representations. We then elaborate our proposed FAC from three major aspects: (1) Regional grouping contrast that exploits local geometry homogeneity from over-segmentation to encourage semantic coherence of local regions; (2) A correspondence framework that consists of a Siamese network and a feature contrast loss for capturing the correlations among the learned feature representations; (3) Optimization losses that take advantage of the better contrast pairs for more discriminative robot self-supervised learning.

3.1 Point- and Scene-Level Contrast Revisited

The key in contrastive learning-based 3D SSL is to construct meaningful contrast pairs between the two augmented views. Positive pairs have been constructed at either point level as in PointContrast (PCon) (Xie et al., 2020a) or scene level as in DepthContrast (DCon) (Zhang et al., 2022). Concretely, given the augmented views of 3D partial point/depth scans, the contrastive loss is applied to maximize the similarity of the positive pairs and the distinction between negative pairs. In most cases, InfoNCE (Oord et al., 2018) loss can be applied for contrast:

$$\mathcal{L}_{ctra} = -\frac{1}{\|\mathbf{B}_p\|} \sum_{(a,b)\in\mathbf{B}_p} \log \frac{\exp(\boldsymbol{f}^a_{g1} \cdot \boldsymbol{f}^b_{g2}/\tau)}{\sum_{(\cdot,c)\in\mathbf{B}_p} \exp(\boldsymbol{f}^a_{g1} \cdot \boldsymbol{f}^c_{g2}/\tau))}. \tag{1}$$

Here $\boldsymbol{f}_{g1}$ and $\boldsymbol{f}_{g2}$ are the feature vectors of two augmented views for contrast. $\mathbf{B}_p$ is the index set of matched positive pairs. $(a, b) \in \mathbf{B}_p$ is a positive pair whose feature embeddings are forced to be similar, while $\{(a, c)|(\cdot, c) \in \mathbf{B}_p, c \neq b\}$ are negative pairs whose feature embeddings are encouraged to be different. PCon (Xie et al., 2020a) directly adopts registered point-level pairs while DCon (Zhang et al., 2022) uses the max-pooled scene-level feature pairs for conducting contrast.

Despite their decent performance in 3D downstream tasks, the constructed contrast pairs in prior studies tend to be sub-optimal. As illustrated in Fig. 1, point-level contrast tends to overemphasize the fine-grained low-level details and overlook the region-level geometric coherence which often provides object-level information. Scene-level contrast aggregates the feature of the whole scene for contrast, which can lose the object-level spatial contexts and distinctive features, leading to less informative representations for downstream tasks. We thus conjecture that region-level correspondences are more suitable to form the contrast, and this has been experimentally verified as illustrated in Fig. 1, more details to be elaborated in ensuing Subsections.

3.2 Foreground-Aware Contrast

Region-wise feature representations have been shown to be very useful in considering contexts for downstream tasks such as semantic segmentation and detection (He et al., 2017; Zhang et al., 2020; Bai et al., 2022). In our proposed geometric region-level foreground-aware contrast, we obtain regions by leveraging the off-the-shelf point cloud over-segmentation techniques (Papon et al., 2013; Guo et al., 2014). The adoption of over-segmentation is motivated by its merits in three major aspects. First, it can work in a completely unsupervised

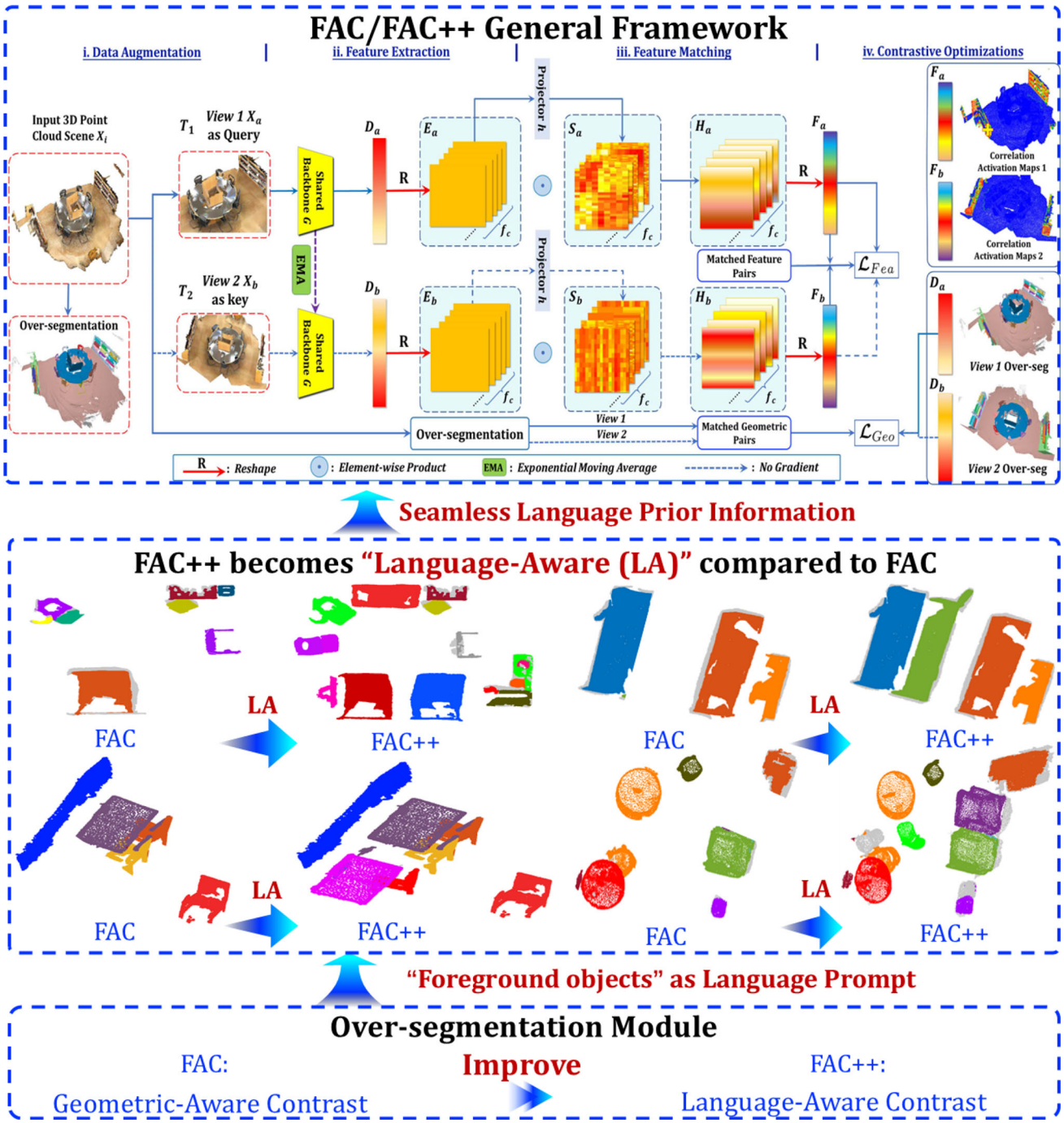

Fig. 2 The framework of our proposed FAC/FAC++. FAC samples median-sized regions as foreground regions, while FAC++ leverages foreground prompts to enhance the foreground-aware feature representations. They both take two augmented 3D point cloud views as input which first extracts the backbone features $\boldsymbol{D}_a$ and $\boldsymbol{D}_b$ for foreground aware contrast with $\mathcal{L}_{Geo}$. The backbone features are then reshaped to regularized representation $\boldsymbol{E}_a$ and $\boldsymbol{E}_b$ to find correspondences across two views for feature matching. Specifically, we adopt the projector $\boldsymbol{h}$ to transfer $\boldsymbol{E}_a$ and $\boldsymbol{E}_b$ to feature maps $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ to adaptively learn their correlations and produce enhanced representations $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$. Finally, $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$ are reshaped back to $\boldsymbol{F}_a$ and $\boldsymbol{F}_b$ where matched feature pairs are enhanced with feature contrast loss $\mathcal{L}_{Fea}$. Hence, both our proposed FAC and FAC++ exploits complementary foreground awareness and foreground-background distinction within and across views for more informative and meaningful representation learning. We have compared the performance of proposed FAC++ compared with FAC the in a detailed manner as demonstrated in this Figure below. It can be shown qualitatively and clearly that by leveraging "foreground objects" as the language prompts, our proposed FAC++ becomes language-aware compared with FAC. The proposed simple but effective module can provide seamless language prior information in our generalized FAC++ framework. As demonstrated from our experiments, the foreground awareness will also provide a large performance boost compared with the previous approaches

manner without requiring any annotated data. Second, our proposed regional sampling (to be described later) allows us to filter out background regions such as ceilings, walls, and ground in an unsupervised manner, where the background regions are often represented by geometrically homogeneous patterns with a large number of points. Regions with a very limited number of points can also be filtered out, which are noisy in both geometry and semantics. Third, over-segmentation provides geometrically coherent regions with high semantic similarity, while diverse distant regions tend to be semantically distinct after sampling, which effectively facilitates discriminative feature learning. Specifically, over-segmentation divides the original point clouds scene into I class-agnostic regions $S = \{s_1, s_2, ..., s_i\}$, and $s_i \cap s_k = \emptyset$ for any $s_i \neq s_k$. Our empirical experiments show that our proposed framework works effectively with mainstream over-segmentation approaches without fine-tuning.

3.2.1 Foreground Prompted Regional Sampling for Balanced Learning

In our preliminary conference version, we designed a simple but effective region sampling technique to obtain meaningful foreground from the geometrically homogeneous regions as derived via over-segmentation as introduced above. Specifically, we first count the number of points in each region and rank regions according to the number of points they contain. We then identify the region having the median number of points as s_{med}. Next, we select H regions having the closest number of points with s_{med} to form contrast pairs. Extensive experiments show that this sampling strategy is effective in the downstream task. We conjecture that the massive points in background regions encourage biased learning towards repetitive and redundant information, while regions with very limited points are noisy in both geometry and semantics. Our sampling strategy can encourage balanced learning towards foreground regions which leads to more informative and discriminative representations.

3.2.2 Foreground Prompts

Note that for FAC, merely selecting the median-sized regions can not always guarantee that all points lie within this region are foreground points. Therefore, we designed an approach to better guide the contrastive optimizations to more effectively enhance the foreground-background distinctive representations. Specifically, we designed an effective approach for the language-queried foreground sampling based on the current prevailing vision-language models (Bai et al., 2023). We directly utilize the model from the OpenMask3D (Takmaz et al., 2023) to obtain the aligned 3D and language co-embeddings. Then, we utilize the pre-trained models with the corresponding language textual query termed "foreground" to obtain the final filtered regions with foreground points. It tuns out that this simple but effective design filters out many non-foreground points and provide more clearly separate foreground regions, which will be of significance to the final downstream 3D scene understanding tasks.

3.2.3 Contrast with Local Regional Consistency

Different from the above-mentioned PCon (Xie et al., 2020a) and DCon (Zhang et al., 2022), we directly exploit region homogeneity to obtain contrast pairs. Specifically, taking the average point feature within a region as the anchor, we regard selected features within the same region as positive keys and in different regions as negative keys. Benefiting from the region sampling strategy, we can focus on the foreground for better representation learning. Denote the number of points within a region as $\mathcal{N}(s_i)$ and the backbone feature as $\boldsymbol{D}$, we aggregated their point feature $\boldsymbol{d}_j \in \boldsymbol{D}$ to produce an average regional feature $\mathcal{D}_m$ within a region as the anchor in contrast to enhance the robustness:

$$\mathcal{D}_m = \frac{1}{\|\mathcal{N}(s_i)\|} \sum_{j \in \mathcal{N}(s_i)} \boldsymbol{d}_j. \tag{2}$$

Regarding $\mathcal{D}_m$ as the anchor, we propose a foreground aware geometry contrast loss $\mathcal{L}_{Geo}$ pulling the point feature to its corresponding positive features in the local geometric region, and pushing it apart from negative point features of different separated regions:

$$\mathcal{L}_{Geo} = -\frac{1}{\|\mathbf{B}_p\|} \sum_{(a,b)\in\mathbf{B}_p} \log \frac{\exp(\mathcal{D}_m^a \cdot \boldsymbol{d}_j^{b,+}/\tau)}{\sum_{(\cdot,c)\in\mathbf{B}_p} \exp(\mathcal{D}_m^a \cdot \boldsymbol{d}_j^{b,-}/\tau))}. \tag{3}$$

Here, $\boldsymbol{d}_j^+$ and $\boldsymbol{d}_j^-$ denote the positive and negative samples, respectively with $\mathcal{D}_m$. We set the number of positive and negative point feature pairs for each regional anchor as k equally. Note our proposed foreground contrast is a generalized version of PCon (Xie et al., 2020a) with foreground enhanced and it returns to PCon if all regions shrink to a single point. Benefiting from the regional geometric consistency and balanced foreground sampling, the foreground aware contrast alone outperforms the state-of-the-art CSC (Hou et al., 2021) in data efficiency in our empirical experimental results.

3.3 Language-Guided Foreground-Background Distinction Aware Contrast

As illustrated in Fig. 2, we propose a Siamese correspondence network (SCN) to explicitly identify feature correspondences within and across views and introduce a feature contrast loss to adaptively enhance their correlations. The SCN is merely

used during the pre-training stage for improving the representation quality. After pre-training, only the backbone network is fine-tuned for downstream tasks.

3.3.1 Siamese Correspondence Network for Adaptive Correlation Mining

Given the input 3D scene $\boldsymbol{X}_i$ with N points, our proposed FAC/FAC++ framework first transforms it to two augmented views $\boldsymbol{X}_a$ and $\boldsymbol{X}_b \in \mathbb{R}^{N\times f_{in}}$, and obtain backbone feature $\boldsymbol{D}_a$ and $\boldsymbol{D}_b \in \mathbb{R}^{N\times f_c}$ by feeding the two views into the backbone network $\mathcal{G}$ and its momentum update (via exponential moving average), respectively (f_c is the number of feature channels). For fair comparisons, we adopt the same augmentation scheme with existing work (Yin et al., 2022; Hou et al., 2021). In addition, We reshape the backbone point-level features to feature maps $\boldsymbol{E}_a$ and $\boldsymbol{E}_b \in \mathbb{R}^{m\times\frac{N}{m}\times f_c}$ to obtain regularized point cloud representations and reduce computational costs. We then apply the projector $\boldsymbol{h}$ to $\boldsymbol{E}_a$ and $\boldsymbol{E}_b$ respectively to obtain feature maps $\boldsymbol{S}_a$ and $\boldsymbol{S}_b \in \mathbb{R}^{m\times\frac{N}{m}\times f_c}$ of the same dimension as $\boldsymbol{E}_a$ and $\boldsymbol{E}_b$. We adopt two simple point-MLPs with a ReLU layer in between to form the projector $\boldsymbol{h}$. The feature maps $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ work as learnable scores which adaptively enhance the significant and correlated features within and across two views. Finally, we conduct element-wise product between $\boldsymbol{E}$ and $\boldsymbol{S}$ to obtain the enhanced feature $\boldsymbol{H}_a$ and $\boldsymbol{H}_b \in \mathbb{R}^{m\times\frac{N}{m}\times f_c}$ and further transform them back to point-wise features $\boldsymbol{F}_a$ and $\boldsymbol{F}_b \in \mathbb{R}^{N\times f_c}$ for correspondence mining. The global feature-level discriminative representation learning is enhanced by the proposed SCN, enabling subsequent contrast with the matched feature. It should also be noted that the Siamese Correspondence Network will not exert any auxiliary extra computational burden, for the fact the weights of the Siamese Correspondence Network are shared in essence.

3.3.2 Contrast with the Matched Feature and Foreground-Background Distinction

With the obtained sampled foreground-background pairs labeled as negative, we conduct feature matching to select the most correlated positive contrastive pairs. As illustrated in Fig. 2, we evaluate the similarity between $\boldsymbol{F}_a$ and $\boldsymbol{F}_b$ and select the most correlated pairs for contrast. The regional anchors are selected in the same manner as in Subsection 3.2. Concretely, we first introduce an average feature $\boldsymbol{\mathcal{F}}_m^a$ for point feature within a region as the anchor when forming contrast, given as $\boldsymbol{\mathcal{F}}_m^a = \frac{1}{\|\mathcal{N}(s_i)\|}\sum_{j\in\mathcal{N}(s_i)}\boldsymbol{f}_j^a$, based on the observation that points in the same local region tends to have the same semantic. For *j-th* point-level feature $\boldsymbol{f}_j^b \in \mathbb{R}^c$ in $\boldsymbol{F}_b$, we calculate its similarity $S_{p,j}$ with regional feature $\boldsymbol{\mathcal{F}}_m^a \in \mathbb{R}^c$:

$$S_{p,j} = \mathcal{F}_{\mathcal{S}}(\boldsymbol{\mathcal{F}}_m^a, \boldsymbol{f}_j^b). \tag{4}$$

Here $\mathcal{F}_{\mathcal{S}}(\boldsymbol{x},\boldsymbol{y})$ denotes the cosine similarity between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We sample the top-k elements from $S_{p,j}$ as positive keys with the regional feature $\boldsymbol{\mathcal{F}}_m^a$ from both foreground and background point features. The top-k operation is easily made differentiable by reformulating it as an optimal transport problem. Besides, we equally select other k foreground-background pairs as negative.

$$\boldsymbol{\mathcal{L}}_{Fea} = -\frac{1}{\|\mathbf{B}_p\|}\sum_{(a,b)\in\mathbf{B}_p}\log\frac{\exp(\boldsymbol{\mathcal{F}}_m^a\cdot\boldsymbol{f}_j^{b,+}/\tau)}{\sum_{(\cdot,c)\in\mathbf{B}_p}\exp(\boldsymbol{\mathcal{F}}_m^a\cdot\boldsymbol{f}_j^{c,-}/\tau))}. \tag{5}$$

Here, $\boldsymbol{f}_j^{b,+}$ denotes the positive keys of the identified k most similar elements with $\boldsymbol{\mathcal{F}}_m^a$ from $\boldsymbol{F}_b$ in another view. $\boldsymbol{f}_j^{b,-}$ denotes the sampled other k negative point features in a batch, respectively. Therefore, the well-related cross-view point features can be adaptively enhanced with the learning of the point-level feature maps $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ of 3D scenes. Our feature contrast enhances the correlations at the feature level within and across views by explicitly finding the region-to-point most correlated keys for the foreground anchor as the query. With learned feature maps, the features of well-correlated foreground/background points are adaptively emphasized while foreground-background distinctive ones are suppressed. Our proposed framework is verified to be very effective qualitatively in point activation maps and quantitatively in downstream transfer learning and data efficiency.

We have also illustrated our framework clearly as given in Fig. 2. As demonstrated in Fig. 2, our proposed framework has incorperated simple but effective designs for enhancing the foreground-aware feature representations. Leveraging the "foreground objects" as the linguistic prompts, our proposed FAC++ becomes language-aware compared with the FAC.

The details of the proposal generation are illustrated apparantly in Fig. 2. We directly use the prompt of “foreground objects” as the linguistic prompt to the 3D vision-language model of OpenMask3D for semantic scene parsing. Utilizing “foreground objects” as the linguistic prompt, the OpenMask3D can explicitly obtain the foreground regions with the segmented foreground masks. Leveraging “foreground objects” as the linguistic prompts, we can formulate the “language-aware” contrast and provide seamless language prior information for our generalized FAC++ framework.

In our subsequent experimental results, we have examined extensively regarding the performance of our proposed framework for semantic segmentation, instance segmentation, as well as for the object detection tasks,both for the

open-world learning as well as the data-efficient learning circumstances. The visual-linguistic aligned foreground-aware feature representations will have a boost on the final semantic scene parsing performance for the fact more discriminative feature representations are learnt through regional foreground-aware contrastive feature learning.

3.4 Joint Optimization of Our Framework

Considering both local region-level foreground geometric correspondence and global foreground-background distinction within and across views, the overall objective function of FAC/FAC++ framework $\mathcal{L}_{Sum}$ is as follows:

$$\mathcal{L}_{Sum} = \alpha \mathcal{L}_{Geo} + \beta \mathcal{L}_{Fea}. \tag{6}$$

Here α, β are the weights balancing two loss terms. We empirically set $\alpha = \beta = 1$ without tuning.

4 Experiments

Data-efficient learning and knowledge transfer capacity have been widely adopted for evaluating self-supervised pre-training and the learned unsupervised representations (Hou et al., 2021). In the following experiments, we first pre-train models on large-scale unlabeled data and then fine-tune them with small amounts of labeled data of downstream tasks to test their data efficiency. We also transfer the pre-trained models to other datasets to evaluate their knowledge transfer capacity. The two aspects are evaluated over multiple downstream tasks including 3D semantic segmentation, instances segmentation, and object detection. Details of the involved datasets are provided in the Appendix.

4.1 Experimental Settings

4.1.1 3D Object Detection

The object detection experiments involve two backbones including VoxelNet (Zhou & Tuzel, 2018) and PointPillars (Lang et al., 2019). Following ProCo (Yin et al., 2022), we pre-train the model on Waymo and fine-tune it on KITTI and Waymo (Table 1).

Following ProCo (Yin et al., 2022) and CSC (Hou et al., 2021), we augment data via random rotation, scaling and flipping, and random point dropout for fair comparisons. We set hyper-parameters τ in $\mathcal{L}_{Fea}$ and $\mathcal{L}_{Geo}$ at 0.1 following ProCo (Yin et al., 2022), H=f_c=m=20, and the total number of positive/negative pairs as 4096 in all experiments including detection and segmentation without tuning. In outdoor object detection on Waymo and KITTI, we pre-train the network with Adam (Kingma & Ba, 2014) optimizer and follow ProCo (Yin et al., 2022) for epoch and batch size setting for fair comparisons with existing works (Liang et al., 2021; Yin et al., 2022). In indoor object detection on ScanNet, we follow CSC (Hou et al., 2021) to adopt SparseConv as the backbone network and VoteNet as the 3D detector, and follow its training settings with the limited number of scene reconstructions (Hou et al., 2021).

4.1.2 3D Semantic and Instance Segmentation

For 3D segmentation, we strictly follow CSC (Hou et al., 2021) in the limited reconstruction setting. Specifically, we pre-train on ScanNet and fine-tune pre-trained models on indoor S3DIS, ScanNet and outdoor SemanticKITTI (SK) (Behley et al., 2019). We use SGD in pre-training with a learning rate of 0.1 and batch size of 32 for 60K steps to ensure fair comparisons with other 3D pre-training methods including CSC (Hou et al., 2021) and PCon (Xie et al., 2020a). In addition, we test the model pre-trained upon ScanNet for SK to evaluate its learning capacity for outdoor sparse LiDAR point clouds. The only difference is that we fine-tune the model for 320 epochs for SK but 180 epochs for indoor datasets. The longer fine-tuning with SK is because transferring models trained on indoor data to outdoor data takes more time to optimize and converge (Table 2).

4.2 Data-efficient Transfer Learning

4.2.1 3D Object Detection

One major target of self-supervised pre-training is more data-efficient transfer learning with less labeled data for fine-tuning. We evaluate data-efficient transfer from Waymo to KITTI as shown in Table 3 and Fig. 6. We can see that FAC outperforms the state-of-the-art consistently. With 20% labeled data for fine-tuning, FAC achieves comparable performance as training *from scratch* by using 100% training data, demonstrating its potential in mitigating the dependence on heavy labeling efforts in 3D object detection. Also, as demonstrated in Table 3, our proposed FAC++ demonstrates superior performance in the tasks of open-world 3D scene understanding and outperforms previous state-of-the-art by a larger margin. The results reveal that our proposed foreground prompted regional sampling approach can have a significant boost on the final semantic and instance segmentation. These results also to some extent indicate more accurate foreground extraction as well as foreground object awareness can boost the final constrative representations in an effective manner. As Fig. 3 shows, FAC has clearly larger activation for inter- and intra-view objects such as vehicles and pedestrians, indicating its learned informative and discriminative representations.

Table 1 Data-efficient *3D object detection* experimental results on Waymo with 1% and 10% labeled training data

Fine-tuning with different label ratios	3D Detector	Approach	Pre-training schedule	Overall		Vehicle		Pedestrian		Cyclist	
				AP%	APH%	AP%	APH%	AP%	APH%	AP%	APH%
1% (around 0.8k frames)	PointPillars (Lang et al., 2019)	–	*From Scratch*	23.05	18.08	27.15	26.17	30.31	18.79	11.28	9.28
		ProCo (Yin et al., 2022)	Pre-trained	31.65	26.34	35.88	35.08	37.61	25.22	21.47	18.73
		FAC (Ours)	Pre-trained	**33.57**	**28.13**	**37.92**	**36.59**	**39.22**	**26.78**	**23.02**	**20.22**
		FAC++ (Ours)	Pre-trained	**34.58**	**29.22**	**38.86**	**37.62**	**39.33**	**26.89**	**23.63**	**21.36**
	VoxelNet (Zhou & Tuzel, 2018)	–	*From Scratch*	20.88	17.83	21.95	21.45	27.98	20.52	12.70	11.53
		ProCo (Yin et al., 2022)	Pre-trained	38.36	34.78	37.60	36.91	39.74	31.70	37.74	35.73
		FAC (Ours)	Pre-trained	**40.15**	**36.65**	**39.57**	**38.76**	**41.59**	**33.42**	**39.43**	**37.39**
		FAC++ (Ours)	Pre-trained	**41.69**	**37.63**	**40.76**	**39.79**	**42.53**	**34.89**	**40.93**	**41.36**
10% (around 8k frames)	PointPillars (Lang et al., 2019)	–	*From Scratch*	51.75	46.58	54.94	54.32	54.01	41.53	46.31	43.88
		ProCo (Yin et al., 2022)	Pre-trained	54.08	49.43	57.54	56.93	56.97	45.25	47.74	46.10
		FAC (Ours)	Pre-trained	**55.16**	**50.51**	**58.60**	**57.91**	**57.98**	**46.88**	**49.19**	**47.38**
		FAC++ (Ours)	Pre-trained	**56.69**	**51.67**	**59.87**	**59.12**	**59.07**	**47.89**	**50.63**	**48.56**
	VoxelNet (Zhou & Tuzel, 2018)	–	*From Scratch*	54.04	51.24	54.37	53.74	51.45	45.05	56.30	54.93
		ProCo (Yin et al., 2022)	Pre-trained	59.00	56.30	58.83	58.23	57.75	51.75	60.42	58.91
		FAC (Ours)	Pre-trained	**60.16**	**57.23**	**59.90**	**59.71**	**58.85**	**52.46**	**61.33**	**59.79**
		FAC++ (Ours)	Pre-trained	**61.57**	**58.16**	**60.99**	**60.89**	**59.98**	**53.65**	**62.82**	**61.36**

The bold highlights the results of our proposed approaches
Similar comparably better experimental results are obtained for KITTI in Table 3 for FAC and FAC++ as compared with the state-of-the-art ProCo (Yin et al., 2022)

Table 2 Data-efficient *3D instance segmentation* results with the limited number of scene reconstructions on ScanNet (Dai et al., 2017) and S3DIS Area-5 validation set with SparseConv and PointGroup (Jiang et al., 2020) as the backbone network

Dataset & Task	Approach	1%		5%		10%		20%		100%	
		mAP	AP@50%	mAP	AP@50%	mAP	AP@50%	mAP	AP@50%	mAP	AP@50%
ScanNet (Dai et al., 2017) Ins. Seg.	*From Scratch*	5.17	9.95	18.38	31.92	26.75	42.76	29.39	48.12	34.53	56.97
	PCon (Xie et al., 2020a)	7.23	12.55	19.46	35.42	27.09	43.97	30.28	49.56	37.29	58.06
	CSC (Hou et al., 2021)	7.13	13.26	20.93	36.75	27.37	44.93	30.62	50.61	38.69	59.43
	FAC (Ours)	**13.25**	**21.95**	**27.61**	**44.88**	**30.22**	**48.23**	**34.57**	**53.86**	**40.56**	**60.59**
	FAC++ (Ours)	**14.63**	**22.78**	**28.73**	**45.96**	**36.76**	**52.67**	**37.68**	**57.93**	**41.68**	**62.67**
S3DIS Ins. Seg.	*From Scratch*	9.55	13.66	20.33	30.59	25.86	36.75	28.77	40.68	40.69	59.32
	PCon (Xie et al., 2020a)	13.42	15.96	22.93	33.65	27.17	38.73	31.29	43.18	43.15	60.56
	CSC (Hou et al., 2021)	14.66	16.70	24.91	34.23	29.77	41.08	33.59	44.77	46.25	63.48
	FAC (Ours)	**19.79**	**24.62**	**28.50**	**39.25**	**35.36**	**45.31**	**35.86**	**45.89**	**47.76**	**65.77**
	FAC++ (Ours)	**21.87**	**26.78**	**31.74**	**41.56**	**38.89**	**46.57**	**36.92**	**48.25**	**48.98**	**66.87**
ScanNet200 (Rozenberszki et al., 2022) Ins. Seg.	*From Scratch*	2.55	9.87	12.38	20.67	23.97	30.69	27.79	39.58	37.69	56.32
	PCon (Xie et al., 2020a)	8.67	15.89	19.93	25.68	25.69	36.89	30.76	43.58	43.15	46.68
	CSC (Hou et al., 2021)	14.66	17.26	21.91	30.23	28.97	39.59	33.63	45.87	46.25	53.48
	ProCo (Yin et al., 2022)	15.31	17.87	23.78	34.67	29.78	40.93	35.29	46.36	47.27	55.18
	FAC (Ours)	**21.16**	**25.32**	**28.50**	**36.56**	**35.36**	**42.31**	**36.39**	**46.97**	**47.79**	**63.86**
	FAC++ (Ours)	**27.76**	**31.61**	**32.97**	**40.67**	**39.88**	**47.69**	**39.26**	**49.39**	**53.67**	**68.67**

FAC clearly outperforms CSC (Hou et al., 2021) and PCon (Xie et al., 2020a) in all label ratios. The increment is more notable with extremely limited labeled data. It can also be demonstrated that our proposed approach can provide more discriminative feature embedding space for the open-vocabulary instance scene parsing in ScanNet200 benchmark

Table 3 The data-efficient *3D object detection* on KITTI

Fine-tuning with different label ratios	3D Detector	Pre-training schedule	mAP (Mod).	Car			Pedestrian			Cyclist		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
20% (about 0.74k frames)	PointRCNN (Shi et al., 2019)	*From Scratch*	63.51	88.64	75.23	72.47	55.49	48.90	42.23	85.41	66.39	61.74
		ProCo (Yin et al., 2022)	66.20	88.52	77.02	72.56	58.66	51.90	44.98	90.27	69.67	65.05
		FAC (Ours)	**68.11**	**89.95**	**78.75**	**73.98**	**59.93**	**53.98**	**46.36**	**91.56**	**72.30**	**67.88**
		FAC++ (Ours)	**69.89**	**91.26**	**80.59**	**75.36**	**61.39**	**55.57**	**47.87**	**92.89**	**73.68**	**69.96**
	PV-RCNN	*From Scratch*	66.71	91.81	82.52	80.11	58.78	53.33	47.61	86.74	64.28	59.53
		ProCo (Yin et al., 2022)	68.13	91.96	82.65	80.15	62.58	55.05	50.06	88.58	66.68	62.32
		FAC (Ours)	**69.73**	**92.87**	**83.68**	**82.32**	**64.15**	**56.78**	**51.29**	**89.65**	**68.65**	**65.63**
		FAC++ (Ours)	**71.27**	**94.79**	**85.92**	**84.99**	**66.87**	**57.98**	**52.92**	**91.39**	**70.56**	**67.77**
100% (about 3.71k frames)	PointRCNN (Shi et al., 2019)	*From Scratch*	69.45	90.02	80.56	78.02	62.59	55.66	48.69	89.87	72.12	67.52
		DCon (Zhang et al., 2022)	70.26	89.38	80.32	77.92	65.55	57.62	50.98	90.52	72.84	68.22
		ProCo (Yin et al., 2022)	70.71	89.51	80.23	77.96	66.15	58.82	52.00	91.28	73.08	68.45
		FAC (Ours)	**71.83**	**90.53**	**81.29**	**78.92**	**67.23**	**59.97**	**53.10**	**92.23**	**74.59**	**69.87**
		FAC++ (Ours)	**73.37**	**92.59**	**82.97**	**80.59**	**69.76**	**61.99**	**55.88**	**93.89**	**76.38**	**71.96**
	PV-RCNN	*From Scratch*	70.57	–	84.50	–	–	57.06	–	–	70.14	–
		GCC-3D (Liang et al., 2021)	71.26	–		–	–	–	–	–	–	–
		STRL (Huang et al., 2021)	71.46	–	84.70	–	–	57.80	–	–	71.88	–
		PCon (Xie et al., 2020a)	71.55	91.40	84.18	82.25	65.73	57.74	52.46	91.47	72.72	67.95
		ProCo (Yin et al., 2022)	72.92	92.45	84.72	82.47	68.43	60.36	55.01	92.77	73.69	69.51
		FAC (Ours)	**73.95**	**92.98**	**86.33**	**83.82**	**69.39**	**61.27**	**56.36**	**93.75**	**74.85**	**71.23**
		FAC++ (Ours)	**75.77**	**94.49**	**88.53**	**85.97**	**71.92**	**64.21**	**59.23**	**95.57**	**76.39**	**73.52**

We pre-train the backbone network of PointRCNN (Shi et al., 2019) and PV-RCNN on Waymo and transfer to KITTI with 20% and 100% annotation ratios in fine-tuning. FAC as well as FAC++ outperforms the state-of-the-art ProCo (Yin et al., 2022) consistently for two settings. *'From Scratch'* denotes the model trained from scratch. All experimental results are averaged over three runs

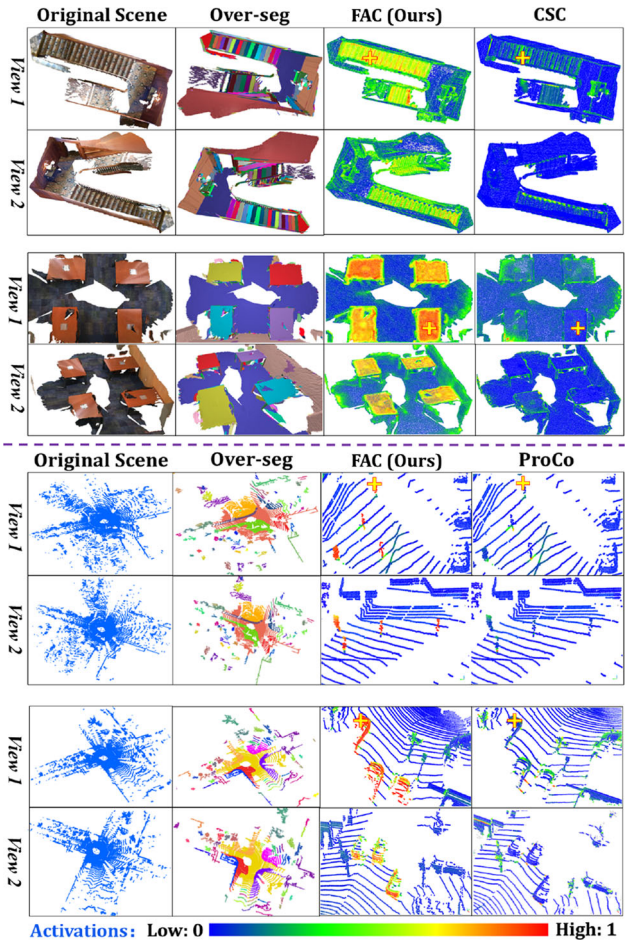

Fig. 3 Visualizations of projected point correlation maps over the indoor ScanNet (1st-4th rows) and the outdoor KITTI (5th-8th rows) with respect to the query points highlighted by yellow crosses. The *View 1* and *View 2* in each sample show the intra-view and cross-view correlations, respectively. We compare FAC with the state-of-the-art CSC (Hou et al., 2021) on segmentation (rows 1–4) and ProCo (Yin et al., 2022) on detection (rows 5–8). FAC clearly captures better feature correlations within and across views (columns 3–4) (Color figure online)

We also study data-efficient learning while performing intra-domain transfer to the Waymo validation set in an extremely label-scarce circumstance with 1% labels. As Table 1 shows, FAC outperforms ProCo (Yin et al., 2022) clearly and consistently, demonstrating its potential in reducing data annotations. Moreover, it can also be demonstrated that our proposed FAC++ provides more superior performance gain while trained with less labeled data, demonstrating its label-efficient learning capacity. It can be attributed to that superior foreground-background distinctive representations are learned during the pre-training stage, which boost the final 3D object detection performance. In addition, we conducted experiments for indoor detection on ScanNet. As Table 4 shows, FAC achieves excellent transfer and improves AP significantly by 20.57% with 10% labels compared with *From Scratch*. Also, the improvement is larger when less annotated data is applied. The superior object detection performance is largely attributed to our proposed foreground-grouping aware contrast that leverages informative foreground regions to form the contrast pairs and the adaptive feature contrast that enhances holistic object-level representations.

Table 4 Data-efficient *3D object detection* average precision (AP%) with the limited number of scene reconstructions on ScanNet with VoteNet as the backbone network

Label ratio	10%	20%	40%	80%	100%
From Scratch	0.39	4.67	22.09	33.75	35.48
CSC (Hou et al., 2021)	8.68	20.96	29.27	36.75	39.32
ProCo (Yin et al., 2022)	12.64	21.87	31.95	37.83	40.56
FAC (Ours)	**20.96**	**27.35**	**35.93**	**39.91**	**42.83**
FAC++ (Ours)	**22.89**	**28.87**	**37.12**	**41.23**	**44.18**

4.2.2 3D Semantic and Instance Segmentation

We first conduct qualitative analysis with point activation maps over the dataset ScanNet. As Fig. 3 shows, FAC can find more semantic relationships within and across 3D scenes as compared with the state-of-the-art CSC (Hou et al., 2021) (Fig. 4). This shows that FAC can learn discriminative representations that capture similar features while suppressing distinct ones. As Fig. 5 shows, FAC also produces clearly better instance segmentation as compared with CSC (Hou et al., 2021). Specifically, CSC tends to fail to distinguish adjacent instances such as chairs while FAC can handle such challenging cases successfully.

We also conduct extensive quantitative experiments as shown in the Table 5, where we adopt limited labels (e.g., {1%, 5%, 10%, 20%}) in training. We can apparantly see that FAC outperforms the *baseline From Scratch* by large margins consistently for both semantic segmentation tasks under different labeling percentages. In addition, FAC outperforms the state-of-the-art CSC (Hou et al., 2021) significantly when only 1% labels are used, demonstrating its capacity in learning informative representations with limited labels. Note FAC achieves more improvements while working with less labeled data. It can also be demonstrated in Table 5, our proposed foreground-prompted regional sampling can also have very beneficial results in the task of semantic segmentation. For example, it improves the performance from 35.25 to 37.71 for the task of semantic segmentation under the 1% labeled setting. It to some extent validates the better foreground-aware representation can boost the performance of the final data-efficient semantic segmentation. For semantic segmentation over the dataset SK (Behley et al., 2019), FAC achieves consistent improvement and similar trends with decreasing labeled data (Tables 6, 7).

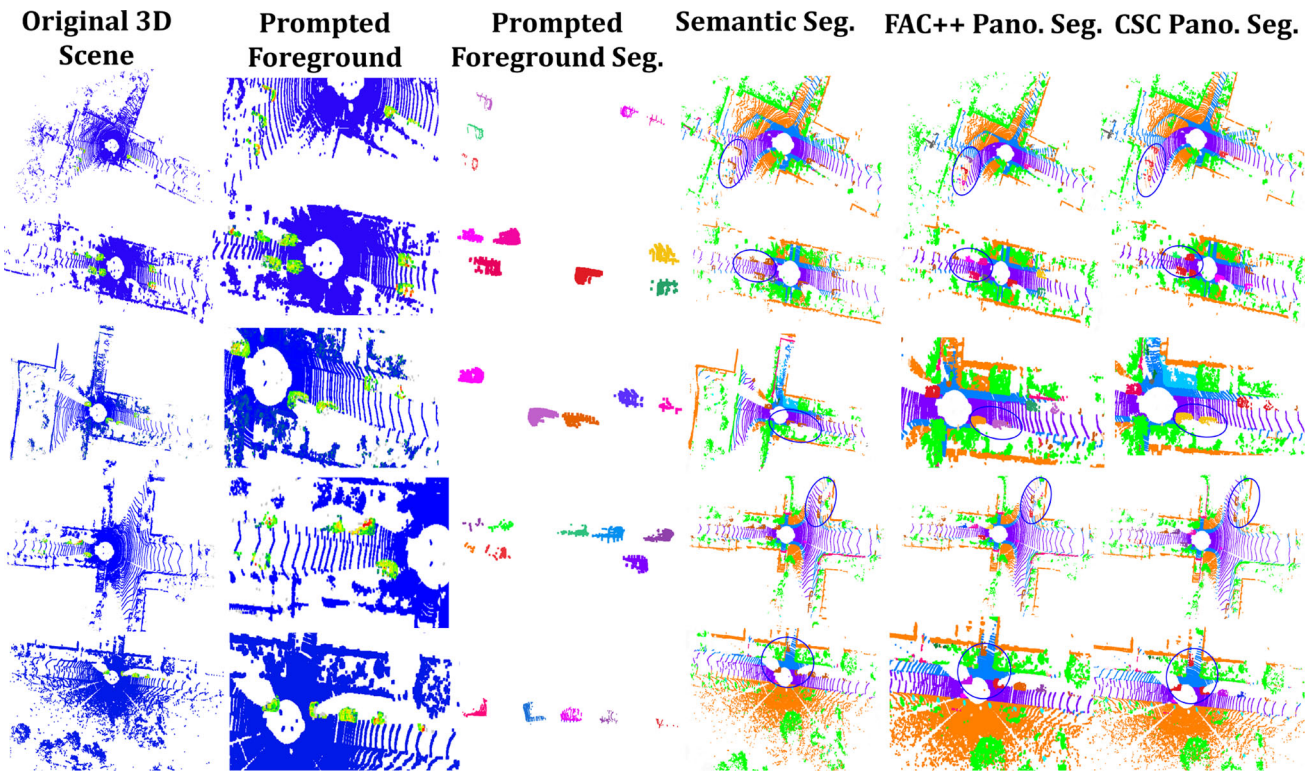

Fig. 4 Visualizations of the outdoor panoptic segmentation in SemanticKITTI. It is demonstrated that our proposed language queries can provide explicit foreground regional information, and the final panoptic segmentation performance is qualitatively superior, which successfully separates between diverse foreground objects very explicitly

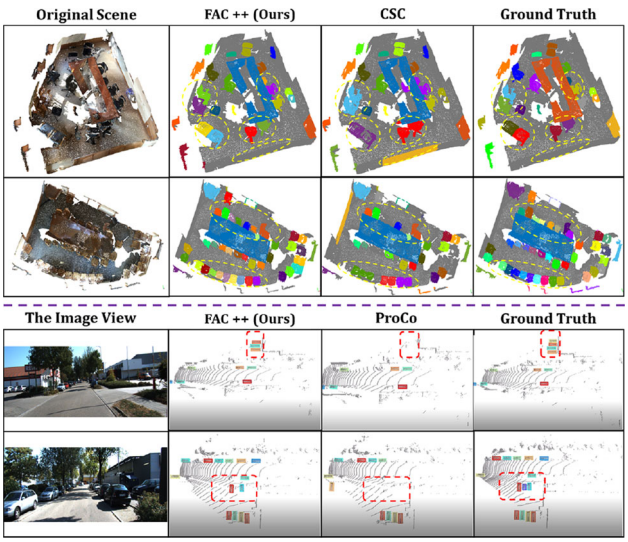

Fig. 5 Visualizations of indoor 3D segmentation over ScanNet compared with CSC (Hou et al., 2021) as fine-tuned with 10% labeled training data and outdoor object detection over KITTI with 20% labeled training data compared with ProCo (Yin et al., 2022). Different segmented instances and detected objects are highlighted in different colors. Differences in prediction are highlighted with the yellow ellipses as well as the red boxes

We also study data-efficient learning while performing intra-domain transfer to the Waymo validation set in an extremely label-scarce circumstance with 1% labels. As Table 1 shows, FAC outperforms ProCo (Yin et al., 2022) clearly and consistently, demonstrating its potential in reducing data annotations. In addition, we conducted experiments for indoor detection on ScanNet. As Table 4 shows, FAC achieves excellent transfer and improves AP significantly by 20.57% with 10% labels compared with *From Scratch*. Also, the improvement is larger when less annotated data is applied. The superior object detection performance is largely attributed to our foreground-aware contrast that leverages informative foreground regions to form the contrast, and the adaptive feature contrast that enhances the holistic object-level representations.

The data-efficient open-world recognition results are shown in Figs. 5 and 6. It can be demonstrated that better foreground object awareness can be effectively capture by our proposed *FAC++* compared with the state-of-the-art (Ding et al., 2023). In Figs. 5 and 6, it can be also demonstrated that the foreground-aware representations can be apparently captured and well maintained (Table 8).

In the meanwhile, we have examined extensively about the domain transfer learning performance as demonstrated

Table 5 Data-efficient *3D semantic segmentation* (mIoU%) results with the limited scene reconstructions (Hou et al., 2021) on ScanNet, S3DIS, and SemanticKITTI (SK) (Behley et al., 2019) with diverse labelling ratios

Dataset & Task	Label ratio	1%	5%	10%	20%	40%
ScanNet Sem. Seg.	*From Scratch*	25.65	47.06	56.72	60.93	63.72
	CSC (Hou et al., 2021)	29.32	49.93	59.45	64.63	68.96
	FAC (Ours)	**35.25**	**51.95**	**61.28**	**65.84**	**69.52**
	FAC++ (Ours)	**37.71**	**53.58**	**62.32**	**66.92**	**70.63**
S3DIS Sem. Seg.	*From Scratch*	35.75	44.38	51.86	58.72	61.83
	CSC (Hou et al., 2021)	36.48	45.07	52.95	59.93	62.65
	FAC (Ours)	**43.73**	**49.28**	**54.76**	**61.05**	**63.22**
	FAC++ (Ours)	**44.68**	**50.29**	**55.89**	**58.92**	**64.36**
SemanticKITTI (Behley et al., 2019) Sem. Seg.	*From Scratch*	28.36	33.58	46.37	50.15	54.56
	CSC (Hou et al., 2021)	32.78	37.55	49.62	55.67	58.89
	FAC (Ours)	**39.92**	**41.75**	**52.37**	**57.65**	**60.17**
	FAC++ (Ours)	**41.39**	**43.13**	**53.26**	**59.87**	**62.38**
ScanNet200 (Rozenberszki et al., 2022) Sem. Seg. (Behley et al., 2019)	*From Scratch*	7.86	13.87	19.37	21.89	24.36
	CSC (Hou et al., 2021)	13.97	17.96	25.57	29.86	36.98
	ProCo (Yin et al., 2022)	**22.69**	**27.68**	**33.68**	**38.65**	**39.17**
	FAC (Ours)	**29.78**	**38.96**	**39.37**	**46.65**	**48.17**
	FAC++ (Ours)	**32.69**	**41.13**	**43.69**	**51.87**	**53.96**

Also, we have evaluated extensively about the results of the proposed approach on ScanNet200 (Rozenberszki et al., 2022) benchmark. It can be demonstrated explicitly that our proposed FAC++ has relatively significant boost on the final semantic 3D scene parsing performance. In the meanwhile, it can be demonstrated clearly that our proposed approach can clearly boost the ultimate open-vocabulary semantic parsing performance with the visual-linguistic aligned foreground-aware feature representations

Table 6 Ablation study of diverse modules of FAC for downstream tasks on ScanNet (Sc) and SemanticKITTI (SK) (Behley et al., 2019), & KITTI (K)

Case	*Sampling*	$\mathcal{L}_{Geo}$	$\mathcal{L}_{Fea}$	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Det. AP@50% (Sc)	Det. mAP% (K)
Baseline				47.17	33.58	0.39	63.51
Case 1 (Merely with Sampling)	✓			48.61	36.21	8.39	64.78
Case 2 (Merely with $\mathcal{L}_{Geo}$)		✓		50.62	39.87	16.46	66.36
Case 3 (Merely with $\mathcal{L}_{Fea}$)			✓	50.93	40.32	17.98	66.29
Case 4 (Merely w/ $\mathcal{L}_{Geo}$)	✓		✓	51.34	40.56	18.57	66.72
Case 5 (Merely w/ Sampling)		✓	✓	51.48	40.68	18.68	67.28
Case 6 (Merely w/ $\mathcal{L}_{Fea}$)	✓	✓		51.46	40.97	18.79	67.22
FAC (Full)	✓	✓	✓	**51.95**	**41.75**	**20.96**	**68.11**
H-FAC	✓	✓	✓	51.66	41.58	20.67	67.65

Table 7 Ablation study of diverse modules of FAC++ for downstream scene understanding tasks on ScanNet (Sc) and SemanticKITTI (SK) (Behley et al., 2019), & KITTI (K)

Case	*Sampling*	$\mathcal{L}_{Geo}$	$\mathcal{L}_{Fea}$	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Det. AP@50% (Sc)	Det. mAP% (K)
Baseline				47.87	35.08	4.78	64.68
Case 1 (Merely with Sampling)	✓			48.93	37.47	16.21	65.52
Case 2 (Merely with $\mathcal{L}_{Geo}$)		✓		51.38	39.98	17.89	66.77
Case 3 (Merely with $\mathcal{L}_{Fea}$)			✓	51.67	40.75	18.86	67.19
Case 4 (Merely w/o $\mathcal{L}_{Geo}$)	✓		✓	52.28	41.87	19.19	66.82
Case 5 (Merely w/o Sampling)		✓	✓	52.55	42.33	19.76	68.12
Case 6 (Merely w/o $\mathcal{L}_{Fea}$)	✓	✓		52.89	42.56	20.87	68.61
FAC++ (Full)	✓	✓	✓	**53.58**	**43.13**	**22.89**	**69.96**
H-FAC++	✓	✓	✓	53.42	42.97	21.76	69.22

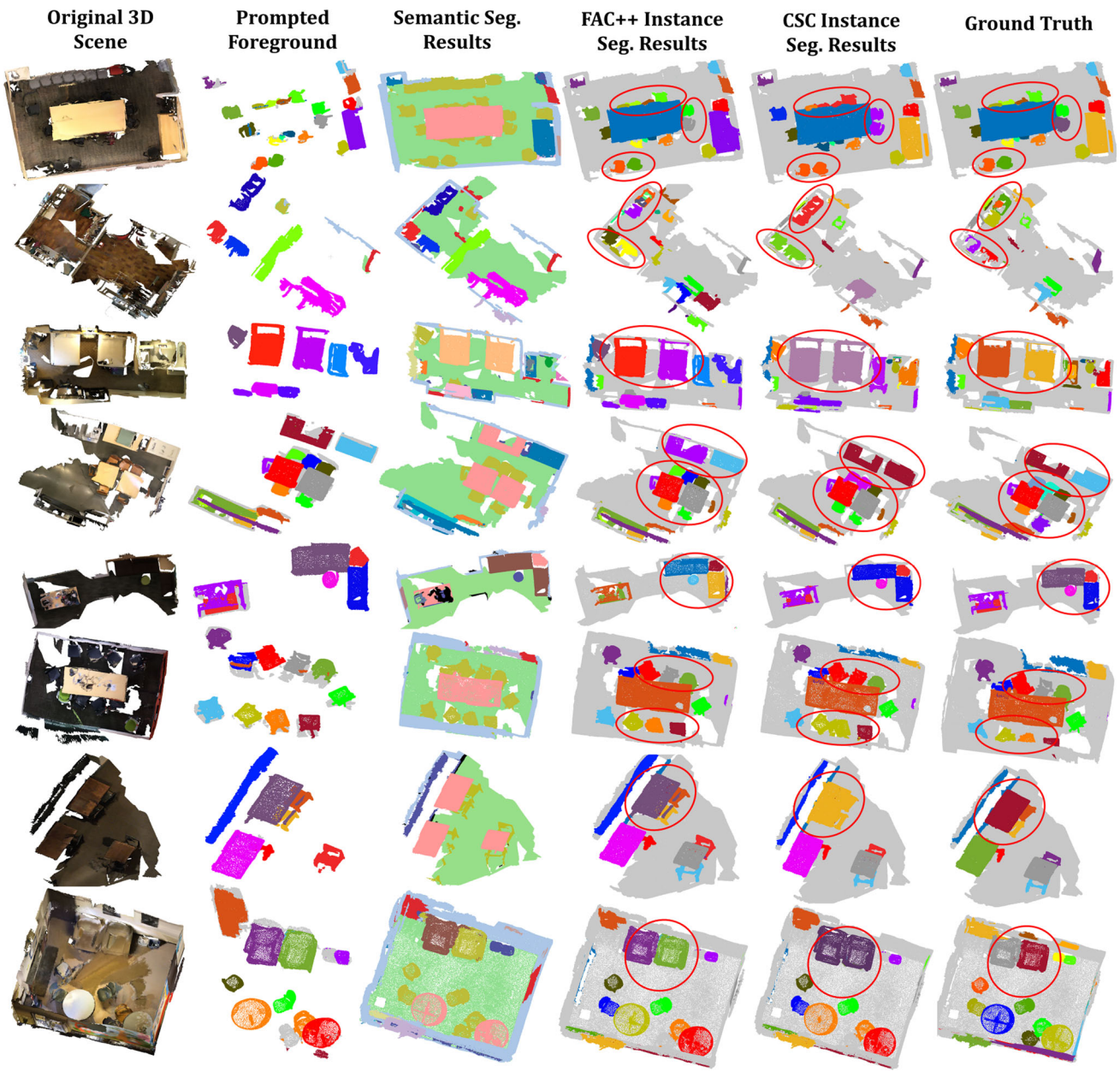

Fig. 6 Visualizations of indoor 3D segmentation over ScanNet compared with CSC (Hou et al., 2021) as fine-tuned with 10% labeled training data and outdoor object detection over KITTI with 20% labeled training data compared with ProCo (Yin et al., 2022). It can also be demonstrated that our proposed language query provides explicit as well as clearly separated foreground regional information. Different segmented instances and detected objects are highlighted with different colors. Differences in prediction are highlighted with yellow ellipses and red boxes

in Table 9. First, it can be shown that our proposed FAC is less influenced by the domain gap compared with the previous SOTAs ProCo. Second, it can be found that our proposed FAC++ can generalize well between indoor and outdoor benchmarks for KITTI and ScanNet as demonstrated in Table 9. It can be inferred from Table 9 that the vision-language aligned representations will improve the domain generalization capacity, and make the model easy to generalize well between indoor and outdoor circumstances.

4.3 Open-world 3D scene Understanding

We have also extensively evaluated the open-world 3D scene understanding performance of our proposed approaches for the final open-world 3D scene understanding tasks. Specif-

Table 8 Comparisons of the open vocabulary learning performance

Datasets	Models	Few-shot settings		
		B15/N4	B12/N7	B10/N9
ScanNet (Dai et al., 2017)	3DGenZ (Cheraghian et al., 2020)	20.6/56.0/12.6	19.8/35.5/13.3	12.0/63.6/6.6
	3DTZSL (Michele et al., 2021)	10.5/36.7/6.1	3.8/36.6/2.0	7.8/55.5/4.2
	LSeg3D (Wang et al., 2021a)	0.0/64.4/0.0	0.9/55.7/0.1	1.8/68.4/0.9
	PLA without caption (Ding et al., 2023)	39.7/68.3/28.0	24.5/70.0/14.8	25.7/75.6/15.5
	PLA (Ding et al., 2023)	65.3/68.3/62.4	55.3/69.5/45.9	53.1/76.2/40.8
	FAC & PLA (Ours)	68.9/70.6/66.8	60.3/71.7/58.8	58.7/77.5/51.6
	FAC & **PLA (Ours)**	70.8/71.9/68.3	61.9/72.9/60.5	59.6/78.8/52.9
	Fully-supervised	74.5/68.4/79.1	73.6/72.0/72.8	69.9/75.8/64.9
Datasets	Models	Few-shot settings		
		B12/N3	B10/N5	B6/N9
NuScenes (Caesar et al., 2020)	3DGenZ (Cheraghian et al., 2020)	01.6/53.3/00.8	01.9/44.6/01.0	01.1/52.6/00.5
	3DTZSL (Michele et al., 2021)	01.2/21.0/0.6	06.4/17.1/03.9	2.61/18.52/03.15
	LSeg-3D (Ding et al., 2023)	0.6/74.4/0.3	0.0/72.5/0.0	2.66/69.72/0.21
	PLA without caption (Ding et al., 2023)	25.5/75.8/15.4	10.7/76.0/05.7	8.95/65.83/6.32
	PLA (Ding et al., 2023)	47.7/73.4/35.4	24.3/73.1/14.5	15.63/60.32/12.38
	FAC & **PLA (Ours)**	52.8/77.3/46.9	51.6/79.5/26.8	42.5/53.6/58.3
	FAC++ & **PLA (Ours)**	67.7/79.6/51.8	65.9/85.8/47.3	59.6/77.5/68.9
	Fully-supervised	73.7/76.6/71.1	74.8/76.8/72.8	74.6/75.9/72.3

It can be demonstrated that our proposed approach provides very superior open-world recognition performance compared with the diverse SOTAs

Table 9 The FAC and FAC++ transfer learning performance for the task of the semantic segmentation (Metric: mIOU%) as well as object detection (Metric: Average Precision (AP)), respectively

Case No.	Case No.	Diverse 3D scene understanding tasks	
		Sem Seg (mIoU%)	Obj Det (mAP%)
ProCo (Yin et al., 2022)	KITTI → ScNet	59.8 (↓ −17.9)	52.8 (↓ −16.5)
	Waymo → ScNet	61.0 (↓ −16.7)	53.5 (↓ −15.8)
	ScNet → KITTI	60.1 (↓ −11.5)	45.5 (↓ −15.3)
	ScNet → Waymo	53.1 (↓ −17.6)	49.0 (↓ −14.1)
FAC	KITTI → ScNet	71.8 (↓ −5.9)	61.8 (↓ −7.5)
	Waymo → ScNet	72.1 (↓ −5.6)	59.5 (↓ −9.8)
	ScNet → KITTI	67.5 (↓ −4.1)	55.5 (↓ −5.3)
	ScNet → Waymo	63.8 (↓ −6.9)	55.6 (↓ −7.5)
FAC++	KITTI → ScNet	78.9 (↑ +1.2)	72.1 (↑ +2.8)
	Waymo → ScNet	80.3 (↑ +2.6)	70.6 (↑ +1.3)
	ScNet→ KITTI	73.9 (↑ +2.3)	62.3 (↑ +1.5)
	ScNet→ Waymo	75.2 (↑ +4.5)	65.4 (↑ +2.3)

It can be demonstrated apparently that our proposed FAC++ has superior transfer learning capacity compared with the previous state-of the art approach (Yin et al., 2022). The phonomenon can be attributed to our regional contrastive designs, which largely benefit the ultimate domain tranfer learning capacity. It should also be noted that our proposed vision-language aligned representation will also have a significant boost on the final domain transfer learning performance. It can also be concluded that our proposed vision-language aligned representations work well for both within and across the indoor and outdoor circumstances

ically, for the open-world 3D scene understanding, we train the model with our proposed FAC for pre-training before we apply the subsequent open-world instance-level 3D scene understanding of PLA (Ding et al., 2023). It is demonstrated that our proposed approach has superior performance in terms of open-world 3D scene understanding. As demonstrated in Table 8, our proposed approach achieves superior open-world 3D scene understanding performance. It can be demonstrated that our proposed FAC provides superior performance while combined with the previous state-of-the-art PLA (Ding et al., 2023), which demonstrates that the foreground-background distinctive representation is also very fundamental to the final open-world scene understanding performance.

In this Subsection, we further evaluate the performance of the open-world recognition capacity of our proposed approach FAC. The results of open-world recognition are demonstrated in Table 8. We have compared our combined FAC & PLA pre-training with merely adopting the PLA (Ding et al., 2023) pre-training for establishing the accurate point-language associations. It can be demonstrated that our proposed approach has shown superior performance in terms of open-world recognition. For example, in the setting of B15/N4 our proposed FAC++ has outperformed merely using PLA, which is the previous state-of-the-art by 3.6/2.3/4.4 respectively. The superior performance can be ascribed to that our proposed FAC has demonstrated remarkable performance in establishing foreground-aware feature contrast. We directly use the settings in the PLA (Ding et al., 2023) and split the categories on ScanNet (Dai et al., 2017) and Nuscene (Caesar et al., 2020) into base and novel categories. It can be demonstrated that our proposed method has superior performance in terms of the open-vocabulary few-shot learning for diverse partitioning of original and novel classes. It can also be demonstrated in Table 8 that under diverse spliting of base and novel categories during the data-efficient learning, our porposed FAC++ provides consistent superior performance while conducting 3D scene understanding, demonstrating both its superiority and robustness in terms of open-world 3D recognition while encountered with diverse novel semantic categories and classes.

5 Ablation Study and Analysis

We perform extensive ablation studies over several key technical designs in FAC. Specifically, we examine the effectiveness of the proposed regional sampling, feature matching network, and the two proposed losses. At the same time, we evaluate the performance of data-efficient learning of our proposed FAC++ as compared with FAC for diverse tasks Lastly, we provide t-SNE visualizations to compare the FAC-learnt feature space with the state-of-the-art. In the ablation studies, we adopt 5% labels in semantic segmentation experiments, 10% labels in indoor detection experiments on ScanNet, and 20% labels in outdoor object detection experiments on KITTI with PointRCNN (Shi et al., 2019) as the 3D detector.

5.1 Regional Sampling and Feature Matching

Regional sampling samples points in the foreground regions as anchors. The ablation experiment without sampling means that we do not use the foreground sampling and use the random sampled point features to acquire the contrast pairs. Table 6 shows related ablation studies as denoted by *Sampling*. We can see that both segmentation and detection deteriorate without *Sampling*, indicating that the foreground regions in over-segmentation may provide important object information while forming contrast. It validates that the proposed regional sampling not only suppresses noises but also mitigates the learning bias towards the background, leading to more informative representations in downstream tasks. In addition, we replace the proposed Siamese correspondence network with Hungarian bipartite matching (Kuhn, 2005) (i.e., **H-FAC**) as shown in Table 6. We can observe consistent performance drops, indicating that our Siamese correspondence framework can achieve better feature matching and provides well-correlated feature contrast pairs for downstream tasks. More comparisons of matching strategies are reported in the Appendix.

5.2 FAC Losses

FAC employs a foreground grouping-aware geometric loss $\mathcal{L}_{Geo}$ and a feature loss $\mathcal{L}_{Fea}$ that are critical to its learned representations in various downstream tasks. The geometric loss guides foreground-aware contrast to capture local consistency while the feature loss guides foreground-background distinction. They are complementary and collaborate to learn discriminative representations for downstream tasks. As shown in Table 6 cases 4 and 6, including either loss clearly outperforms the *Baseline* as well as the state-of-the-art CSC (Hou et al., 2021) in segmentation and ProCo (Yin et al., 2022) in detection. For example, only including $\mathcal{L}_{Geo}$ (Case 6) achieves 67.22% and 18.79% average precision in object detection on KITTI and ScanNet, outperforming ProCo (66.20% and 12.64%) by 1.02% and 6.15%, respectively as shown in Table 3 and Table 4. At last, the **full FAC** in Table 6 including both losses learn better representations with the best performance in various downstream tasks.

5.3 FAC++ Ablations

The ablation study results of FAC++ are demonstrated in Table 7. It is demonstrated that our proposed FAC++ has generally slightly better performance as compared with FAC.

The foreground grouping-aware geometric loss $\mathcal{L}_{Geo}$ and the feature loss $\mathcal{L}_{Fea}$ are both very significant for the final scene understanding performance, and dropping either of them will result in significant information loss. As shown in Table 7 cases 4 and 6, including either loss clearly outperforms the *Baseline* as well as the state-of-the-art CSC (Hou et al., 2021) in segmentation and ProCo (Yin et al., 2022) in detection. At the same time, the foreground prompted regional sampling in FAC++ is also very significant for the final scene understanding performance. As validated in Table 7 case 2, 3, 5, the foreground prompted sampling is of significance to the final downstream performance and removing the foreground sampling results in the performance drop. For example, the performance drops 3.13% while comparing case 5 with the full **FAC**. It validates the effectiveness of our proposed foreground prompted sampling in sampling very meaningful and effective foreground-aware feature representations.

5.4 Feature Visualization with t-SNE (Van der Maaten & Hinton, 2008)

We employ t-SNE to visualize the feature representations that are learnt for SemanticKITTI (Behley et al., 2019) semantic segmentation task as illustrated in Fig. 7. Compared with other contrastive learning approaches such as PCon (Xie et al., 2020a) and CSC (Hou et al., 2021), FAC learns a more compact and discriminative feature space that can clearly separate features of different semantic classes. As Fig. 7 shows, the FAC-learnt features have the smallest intra-class variance and largest inter-class variance, demonstrating that the FAC-learnt representations help learn more discriminative features in the downstream task.

5.5 Detailed Supplementary Experimental Results about our proposed 3D Vision-language Model

In this supplementary material for experimental details, additional experimental results and details that are not included in the main paper due to space limits are provided. We include further parameter analyses and ablation studies testing the robustness of our proposed FAC that are not included due to the space limits. More qualitative and quantitative illustrations are also provided:

- The details of our further experimental settings in pre-training including data augmentation and hardware settings (see Sect. 6).
- Details of the experimental datasets involved during the pre-training and testing of our proposed FAC (see Sect. 7).
- Additional quantitative experimental results and analyses are provided (see Sect. 8).
- Additional qualitative experimental results and analyses are provided (see Sect. 9.1).
- Details of the further parameter analysis testing the robustness of the proposed FAC are provided. (see Sect. 11).
- Future directions of this work (see Sect. 11.1).

6 Further Pre-training Experimental Settings

6.1 Data Augmentation Details

We utilize four common types of data augmentation to generate augmented two different views in pre-training, including random rotation ($[-180°, 180°]$) along an arbitrary axis (applied independently for both two views), random scaling ([0.8, 1.2]), random flipping along X-axis or Y-axis, and random point dropout. We follow ProCo (Yin et al., 2022) in random point dropout and sample 100k points from the original point cloud for each of the two augmented views. 20k points are chosen from the same indexes to ensure a 20% overlap for the two augmented views, while the other 80k points are randomly sampled from the remaining point clouds. Our data augmentation strictly follows previous work ProCo (Yin et al., 2022) and CSC (Hou et al., 2021) for fair comparisons with them. Concretely, we follow ProCo (Yin et al., 2022) for outdoor 3D object detection on KITTI and Waymo (Sun et al., 2020) and follow CSC (Hou et al., 2021) for other experimental cases for data augmentation.

6.2 Hardware Settings

We next report the hardware used in our experiments. The PCon (Xie et al., 2020a), ProCo (Yin et al., 2022) and CSC (Hou et al., 2021) use data parallel on eight NVIDIA Tesla V100 GPUs with at least 16 GB GPU memory per card as reported in their papers. Limited by computational resources, we use data parallel on four NVIDIA 2080 Ti GPUs with 11 GB GPU memory per card in all experiments. For experiments in outdoor 3D object detection, we directly report the results of ProCo (Yin et al., 2022) in Tables 1 and 2 of our main paper according to its original paper. It can be seen that FAC still outperforms the state-of-the-art approach ProCo (Yin et al., 2022) consistently even if much fewer computational resources are used. For all other experiments, we reimplement the CSC (Hou et al., 2021), ProCo (Yin et al., 2022), PCon (Xie et al., 2020a) and use the same hardware and experimental settings as our proposed FAC in experiments for a fair comparison in Tables 3, 4, and 5 of our main paper. Specifically, we use data parallel on four NVIDIA 2080 Ti GPUs with 11 GB GPU memory per card.

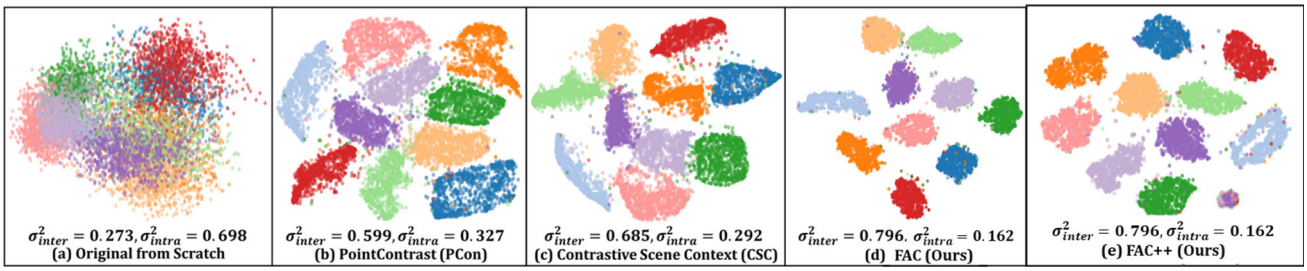

Fig. 7 t-SNE (Van der Maaten & Hinton, 2008) visualization of feature embeddings for SemanticKITTI semantic segmentation fine-tuned with 5% percent label (ScanNet Pre-trained). Ten classes with the least number of points are shown, where σ^2_{intra}, σ^2_{inter} denote intra- and inter-class variance. FAC learns a more compact feature space with the smallest intra-class variance and largest inter-class variance as compared with state-of-the-art methods PCon (Xie et al., 2020a), CSC (Hou et al., 2021)

7 Dataset Details

7.1 S3DIS

S3DIS is a large indoor point cloud scene understanding dataset across six large-scale indoor areas. The total number of scenes is 271. Area 5 is utilized for testing and other areas are used as the training set. Benefiting from Sparse convolution of Minkowski engine, we do not partition the 3D scene into small rooms. The S3DIS dataset has more than 215 million points with thirteen semantic classes. It is used to test the effectiveness of the proposed FAC for both indoor semantic segmentation and instance segmentation.

7.2 ScanNet-v2 (Sc) (Dai et al., 2017)

ScanNet-v2 is a large-scale and comprehensive 3D indoor scene understanding dataset consisting of 1,513 3D scans. The dataset has been adopted for tasks of semantic segmentation, instance segmentation, and object detection. The dataset is divided into 1,201 scans as the training set and 312 scans as the validation set. The number of the semantic category is 21 for semantic segmentation. The ScanNet-v2 (Dai et al., 2017) benchmark is used to test the effectiveness of the proposed FAC for indoor semantic segmentation, instance segmentation as well as indoor object detection. Also, it is used as the pre-training dataset for indoor scene understanding tasks and the outdoor semantic segmentation task on SemanticKITTI (Behley et al., 2019).

7.3 KITTI (K)

KITTI is a large-scale driving-scene dataset that covers sequential outdoor LiDAR point clouds. The KITTI 3D point cloud object detection dataset consists of 7481 labeled samples. The labeled 3D LiDAR scans are split into the training set with 3,712 scans and the validation set with 3,769 scans. The mean average precision (mAP) with 40 recall positions is typically utilized to evaluate the 3D object detection performance. The 3D IoU (Intersection over Union) thresholds are set as 0.7 for cars and 0.5 for cyclists and pedestrians. The KITTI is used to test the effectiveness of the proposed FAC for outdoor 3D object detection.

7.4 SemanticKITTI (SK) (Behley et al., 2019)

SemanticKITTI is derived from the above-mentioned KITTI dataset and annotated with point-level semantics. It is made up of more than 43 thousand (43,552) LiDAR scans. It is annotated with nineteen semantic classes. We follow the official split and use sequences 00-10 for training except sequence 08 for validation. The SemanticKITTI (Behley et al., 2019) is used to test the effectiveness of the proposed FAC for outdoor semantic segmentation.

7.5 Waymo (Sun et al., 2020)

Waymo (Sun et al., 2020) is a large-scale driving-scene dataset that encompasses 158,361 LiDAR scans from 798 scenes for training and 40,077 LiDAR scans for validation. It is approximately twenty times larger than KITTI. The whole training set (without label) is utilized for pre-training different 3D detection backbone networks. The training set of the Waymo (Sun et al., 2020) benchmark is used as the pre-training dataset for outdoor 3D object detection. Its validation set is also utilized to test the effectiveness of the proposed FAC for downstream fine-tuning in outdoor 3D object detection.

8 More Quantitative Experiment Results

In this Section, we include further quantitative experiments that are not included in the main paper due to space limits for the following three experimental cases:

Table 10 Data-efficient *3D object detection* on KITTI

Fine-tuning with diverse label ratios	3D detector	Pre-train. Schedule	mAP (Mod).	Car			Pedestrian			Cyclist		
				Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
20% (about 0.74k frames)	PointRCNN (Shi et al., 2019)	*From Scratch*	63.51	88.64	75.23	72.47	55.49	48.90	42.23	85.41	66.39	61.74
		ProCo (Yin et al., 2022)	66.20	88.52	77.02	72.56	58.66	51.90	44.98	90.27	69.67	65.05
		FAC (Ours)	**68.11**	**89.95**	**78.75**	**73.98**	**59.93**	**53.98**	**46.36**	**91.56**	**72.30**	**67.88**
		FAC++ (Ours)	**69.89**	**91.26**	**80.59**	**75.36**	**61.39**	**55.57**	**47.87**	**92.89**	**73.68**	**69.96**
	PV-RCNN	*From Scratch*	66.71	91.81	82.52	80.11	58.78	53.33	47.61	86.74	64.28	59.53
		ProCo (Yin et al., 2022)	68.13	91.96	82.65	80.15	62.58	55.05	50.06	88.58	66.68	62.32
		FAC (Ours)	**69.73**	**92.87**	**83.68**	**82.32**	**64.15**	**56.78**	**51.29**	**89.65**	**68.65**	**65.63**
		FAC++ (Ours)	**71.27**	**94.79**	**85.92**	**84.99**	**66.87**	**57.98**	**52.92**	**91.39**	**70.56**	**67.77**
50% (about 1.85k frames)	PointRCNN (Shi et al., 2019)	*From Scratch*	66.73	89.12	77.85	75.36	61.82	54.58	47.90	86.30	67.76	63.26
		ProCo (Yin et al., 2022)	69.23	89.32	79.97	77.39	62.19	54.47	46.49	91.26	73.25	68.51
		FAC (Ours)	**70.78**	**90.97**	**81.52**	**78.96**	**63.78**	**56.01**	**48.23**	**92.52**	**74.57**	**69.65**
		FAC++ (Ours)	**71.89**	**91.92**	**82.68**	**79.99**	**64.87**	**57.16**	**49.38**	**93.68**	**75.68**	**70.69**
	PV-RCNN	*From Scratch*	69.63	91.77	82.68	81.90	63.70	57.10	52.77	89.77	69.12	64.61
		ProCo (Yin et al., 2022)	71.76	92.29	82.92	82.09	65.82	59.92	55.06	91.87	72.45	67.53
		FAC (Ours)	**73.25**	**93.35**	**84.39**	**83.69**	**67.25**	**61.45**	**56.87**	**92.17**	**73.96**	**68.98**
		FAC++ (Ours)	**74.53**	**94.78**	**85.53**	**85.23**	**68.39**	**62.68**	**57.98**	**93.53**	**74.97**	**71.28**
100% (about 3.71k frames)	PointRCNN (Shi et al., 2019)	*From Scratch*	69.45	90.02	80.56	78.02	62.59	55.66	48.69	89.87	72.12	67.52
		DCon (Zhang et al., 2022)	70.26	89.38	80.32	77.92	65.55	57.62	50.98	90.52	72.84	68.22
		ProCo (Yin et al., 2022)	70.71	89.51	80.23	77.96	66.15	58.82	52.00	91.28	73.08	68.45
		FAC (Ours)	**71.83**	**90.53**	**81.29**	**78.92**	**67.23**	**59.97**	**53.10**	**92.23**	**74.59**	**69.87**
		FAC++ (Ours)	**73.37**	**92.59**	**82.97**	**80.59**	**69.76**	**61.99**	**55.88**	**93.89**	**76.38**	**71.96**
	PV-RCNN	*From Scratch*	70.57	–	84.50	–	–	57.06	–	–	70.14	–
		GCC-3D (Liang et al., 2021)	71.26	–		–	–	–	–	–	–	–
		STRL (Huang et al., 2021)	71.46	–	84.70	–	–	57.80	–	–	71.88	–
		PCon (Xie et al., 2020a)	71.55	91.40	84.18	82.25	65.73	57.74	52.46	91.47	72.72	67.95
		ProCo (Yin et al., 2022)	72.92	92.45	84.72	82.47	68.43	60.36	55.01	92.77	73.69	69.51
		FAC (Ours)	**73.95**	**92.98**	**86.33**	**83.82**	**69.39**	**61.27**	**56.36**	**93.75**	**74.85**	**71.23**
		FAC++ (Ours)	**75.77**	**94.49**	**88.53**	**85.97**	**71.92**	**64.21**	**59.23**	**95.57**	**76.39**	**73.52**

We pre-train the backbone network of PointRCNN (Shi et al., 2019) and PV-RCNN on Waymo (Sun et al., 2020) and transfer to KITTI with 20%, 50%, and 100% annotation ratios in fine-tuning. FAC and FAC++ outperforms the state-of-the-art ProCo (Yin et al., 2022) consistently across different settings. *'From Scratch'* denotes the model trained from scratch

Table 11 Data-efficient *3D object detection* on Waymo (Sun et al., 2020) with two state-of-the-art 3D object detection backbone networks including PV-RCNN and CenterPoint fine-tuned with 20% training labels

3D Detector	Pre-training Schedule	Overall AP%/APH%	Vehicle		Pedestrian		Cyclist	
			AP%	APH%	AP%	APH%	APH%	AP%
PV-RCNN	*From Scratch*	59.84 / 56.23	64.99	64.38	53.80	45.14	60.61	61.35
GCC-3D (Liang et al., 2021)	Pre-trained	61.30 / 58.18	65.65	65.10	55.54	48.02	62.72	61.43
ProCo (Yin et al., 2022)	Pre-trained	62.62 / 59.28	66.04	65.47	57.58	49.51	64.23	62.86
FAC (Ours)	Pre-trained	**64.57 / 61.75**	**68.27**	**67.76**	**59.96**	**51.27**	**66.97**	**64.87**
FAC++ (Ours)	Pre-trained	**65.76 / 62.89**	**69.53**	**68.95**	**60.97**	**52.38**	**67.88**	**65.92**
CenterPoint	*From Scratch*	63.46 / 60.95	61.81	61.30	63.62	57.79	64.96	63.77
GCC-3D (Liang et al., 2021)	Pre-trained	65.29 / 62.79	63.97	63.47	64.23	58.47	67.68	66.44
ProCo (Yin et al., 2022)	Pre-trained	66.42 / 63.85	64.94	64.42	66.13	60.11	68.19	67.01
FAC (Ours)	Pre-trained	**68.07 / 65.33**	**65.67**	**65.89**	**68.90**	**62.70**	**69.06**	**69.27**
FAC++ (Ours)	Pre-trained	**69.28 / 66.86**	**66.95**	**66.92**	**69.97**	**63.89**	**70.69**	**70.53**
CenterPoint-2-Stages (CP2)	*From Scratch*	65.29 / 62.47	64.70	64.11	63.26	58.46	65.93	64.85
GCC-3D (CP2) (Liang et al., 2021)	Pre-trained	67.29 / 64.95	66.45	65.93	66.82	61.47	68.61	67.46
ProCo (CP2) (Yin et al., 2022)	Pre-trained	68.08 / 65.69	66.98	66.48	68.15	62.61	69.04	67.97
FAC (Ours)	Pre-trained	**69.95 / 67.68**	**68.37**	**68.56**	**69.77**	**65.01**	**70.55**	**69.38**
FAC++ (Ours)	Pre-trained	**71.16 / 68.93**	**69.58**	**69.76**	**70.29**	**66.38**	**71.83**	**71.72**

We implement and configure diverse backbone networks with the codebase OpenPCDet (Team, 2020). Compared with GCC-3D (Liang et al., 2021) and ProCo (Yin et al., 2022), our FAC and FAC++ obtains clear performance gains consistently with different backbone networks. The detectors are trained with 20% training samples in the training set and evaluated on the validation set. The SOTA transfer learning performance can be achieved with our proposed FAC++

Table 12 Comparison of the FAC and FAC++ training time compared with the other state-of-the-art 3D pre-training approaches PCon (Xie et al., 2020a) and CSC (Hou et al., 2021) on different pre-training datasets including ScanNet (Dai et al., 2017) and Waymo (Sun et al., 2020)

Pre-training dataset	Method	Epoch training time	Total training time
ScanNet (Dai et al., 2017)	PCon (Xie et al., 2020a)	15.25	3965.3
	CSC (Hou et al., 2021)	16.13	4193.8
	ProCo (Yin et al., 2022)	21.65	5629.6
	FAC (Ours)	14.59	1750.8 (↓ **126.49%**)
	FAC++ (Ours)	14.78	1773.6 (↓ **123.57%**)
Waymo (Sun et al., 2020)	PCon (Xie et al., 2020a)	20.63	5363.8
	CSC (Hou et al., 2021)	21.22	5517.2
	ProCo (Yin et al., 2022)	25.39	6601.5
	FAC (Ours)	17.69	2299.7 (↓ **133.24%**)
	FAC++ (Ours)	18.21	2367.3 (↓ **126.58%**)

The unit of the training time is (Miniutes per Epoch). Our proposed framework converges at 130 epoch, while the previous approaches converge at 260 epoch approximately. It can be demonstrated that our proposed approach has comparatively less training time, which reveals the training efficiency of our proposed approaches. Compared with the previous most efficient approaches in 3D pre-training, our proposed approach has relatively much less training time. The training time reduces by 123.57% and 126.58% for ScanNet and Waymo, respectively. It further demonstrates the efficiency of our proposed regional contrastive design compared with the point-level contrastive approaches

8.1 KITTI 3D Object Detection

We enrich the experiments of Table 1 in the main paper as shown in Table 10 in this supplementary material. We add the fine-tuning results of data-efficient 3D object detection on KITTI with 50% labeled training data. From Table 10, we can see that although the increments are not as significant as the case when fine-tuned with 20% labeled training data, FAC can still have a notable boost on data-efficient learning performance when fine-tuned with 50% labeled training data. It can also be observed that the improvement is generally more significant as compared with the fine-tuned results with full supervision (100% labeled training data).

8.2 Waymo 3D Object Detection

We enrich the experimental results of Table 2 of the main paper as shown in Table 1 in this supplementary material. We have added the fine-tuning results of data-efficient 3D object detection on Waymo (Sun et al., 2020) with more labeled training data including cases with 50% and 100% labeled training data (compared with 1% and 10% cases). It can be seen that FAC consistently improves the performance with more labeled training data, which further demonstrates the effectiveness of FAC when fine-tuned with the abundant labeled training data.

8.3 Performance of FAC with Other State-of-the-art 3D Object Detection Backbone Networks

We also test the performance of FAC with two state-of-the-art 3D object detection backbone networks including PV-RCNN and Centerpoint as shown in Table 11. We add fine-tuning results of data-efficient 3D object detection on Waymo (Sun et al., 2020) with 20% labeled training data. We implement and configure diverse backbone networks with the codebase OpenPCDet. It can be seen that apart from the 3D backbone network that has been tested in Table 1, our proposed FAC also has consistent improvement when pre-trained and fine-tuned with different backbone networks including the state-of-the-art PV-RCNN and CenterPoint, demonstrating the compatibility of FAC while integrated with different 3D object detection backbone networks (Table 12).

9 More Qualitative Experiment Results

In this Section, we provide more qualitative experiment results. *First*, we provide more visualizations of the point activation maps to test the learnt representation of our proposed FAC. Concretely, like Fig. 3 in the main paper, visualizations of projected point correlation maps over the indoor ScanNet (Dai et al., 2017) and the outdoor KITTI with respect to the query points are provided in Fig. 8.

Second, we visualize qualitative data-efficient experimental results on various 3D scene understanding tasks with diverse labeling percentages when fine-tuned on downstream tasks including 3D instance segmentation on ScanNet-v2 (Dai et al., 2017) as illustrated in Figs. 9 and 10, 3D semantic segmentation on SemanticKITTI (Behley et al., 2019) as illustrated in Figs. 11 and 12, and 3D object detection on KITTI as illustrated in Figs. 13 and 14, respectively.

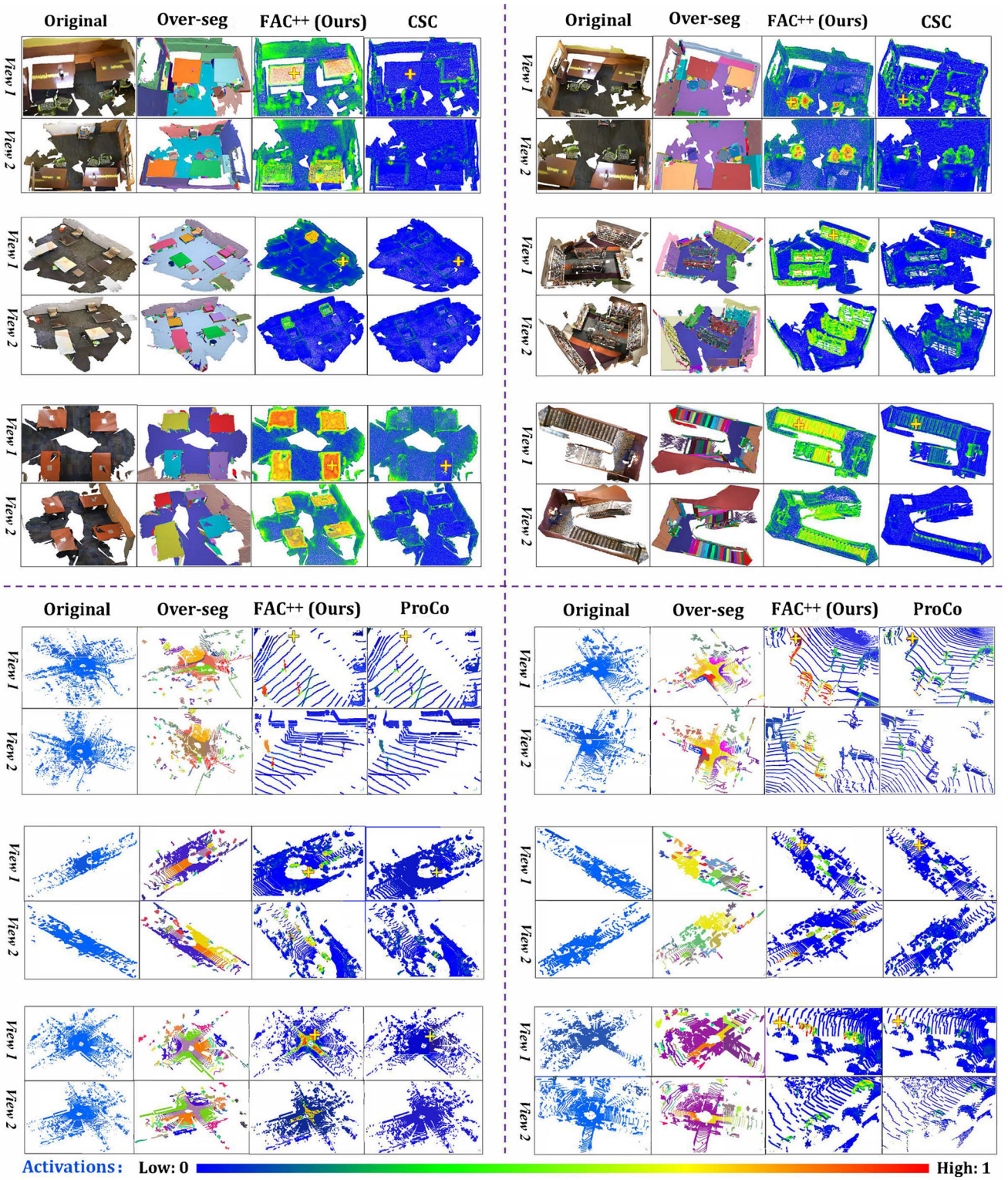

Fig. 8 Visualizations of *projected point activation maps* over the indoor ScanNet (Dai et al., 2017) (Above the purple dash line) and the outdoor KITTI (Below the purple dash line) with respect to the query points highlighted by yellow crosses. The *View 1* and *View 2* in each sample show the intra-view and cross-view correlations, respectively. We compare our proposed FAC++ with the state-of-the-art CSC (Hou et al., 2021) on instance segmentation (Above the purple dash line) and ProCo (Yin et al., 2022) on detection (Below the purple dash line). FAC++ clearly captures better feature correlations within and across views as shown in columns 3–4 and columns 7-8 compared with the state-of-the-art approaches CSC (Hou et al., 2021) and ProCo (Yin et al., 2022), respectively

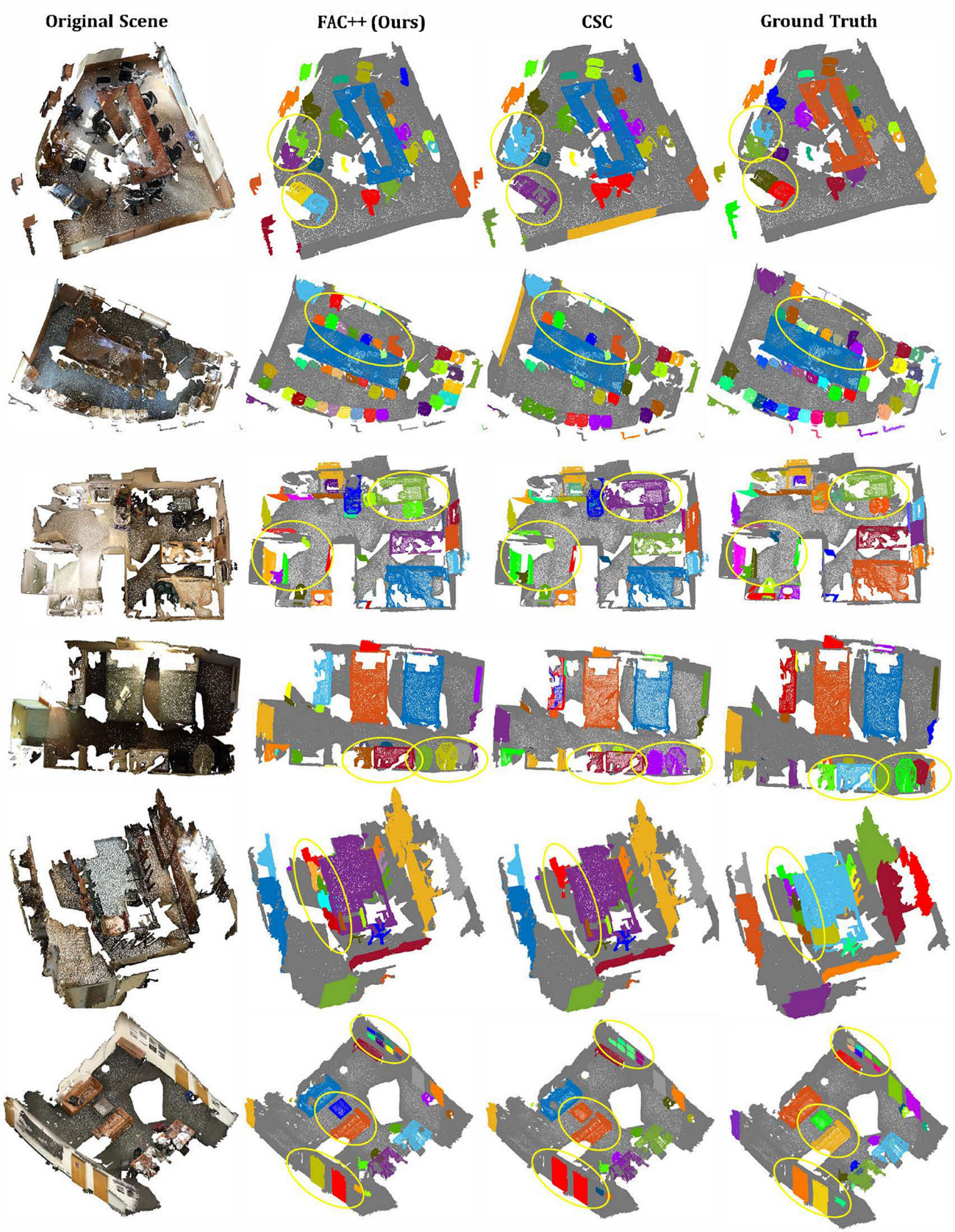

Fig. 9 Visualizations of indoor *3D instance segmentation* over ScanNet (Dai et al., 2017) as fine-tuned with *10% labeled training data*. Different segmented instances are indicated by different colours. Differences in prediction are highlighted by yellow ellipses

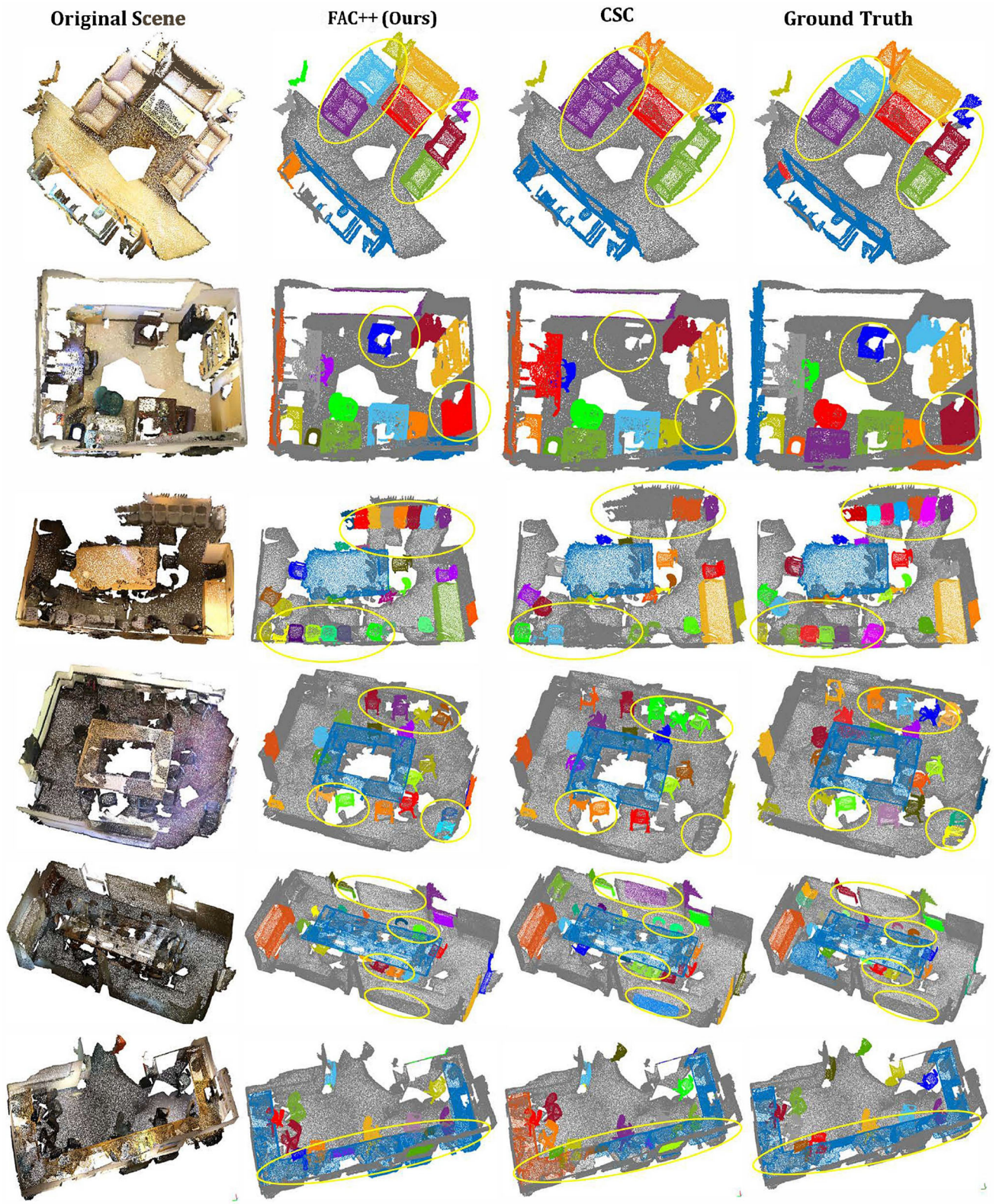

Fig. 10 Visualizations of indoor *3D instance segmentation* over ScanNet (Dai et al., 2017) as fine-tuned with *20% labeled training data*. Different segmented instances are indicated by different colours. Differences in prediction are highlighted by yellow ellipses. It can be demonstrated that our proposed FAC++ has gained superior performance increment compared with the previous SOTA approaches

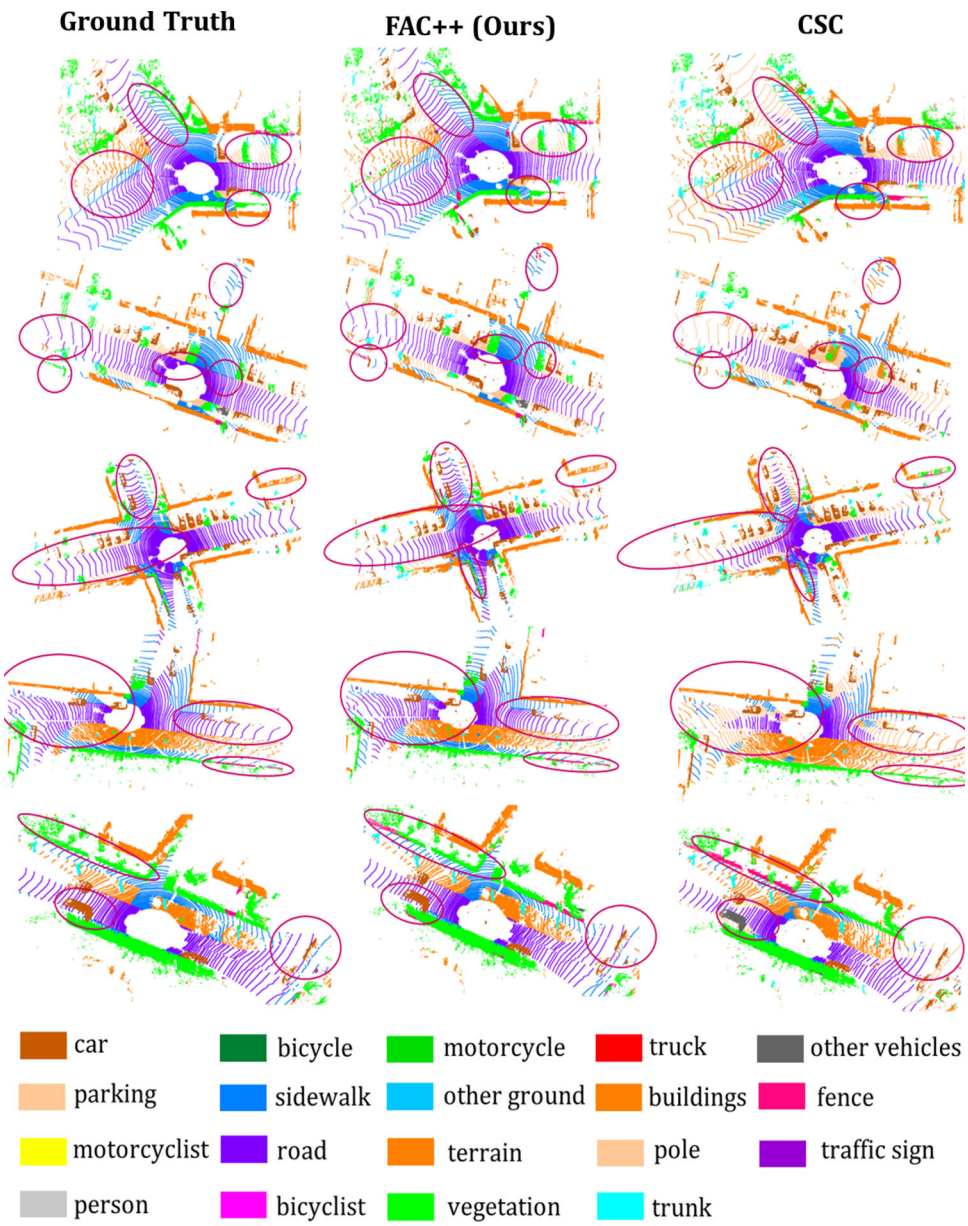

Fig. 11 Comparisons of outdoor *3D Semantic Segmentation* Results on SemanticKITTI (Behley et al., 2019) benchmark fine-tuned with *10% labeled training data* (ScanNet Dai et al. 2017 pre-trained). Note that the SemanticKITTI (Behley et al., 2019) has no color channel as input for the task of semantic segmentation. Therefore, we visualize the ground truth without visualizing the original scene (no color channel). Differences in prediction are highlighted by the red ellipses

9.1 Point Correlation Maps Visualization

As illustrated in Fig. 3, it is clear that our proposed FAC can effectively find both intra- and inter-view feature correlations of the same semantics compared with the state-of-the-art CSC (Hou et al., 2021) and ProCo (Yin et al., 2022). For example, as illustrated in Fig. 3, FAC has clearly larger activation for the inter- and intra-view objects of the same semantics as the query point, such as the vehicle, pedestrian, and road. It further demonstrates that FAC learns informative and discriminative representations which capture similar features while suppressing distinct ones (Table 13).

9.2 Data-Efficient Instance Segmentation

The qualitative experimental results of instance segmentation when fine-tuned with 10% and 20% labeled training data are shown in Figs. 9 and 10. It can be seen that in both the above two cases, the state-of-the-art CSC tends to fail to distinguish adjacent instances such as chairs, desks, and sofas, while FAC can handle these challenging cases successfully.

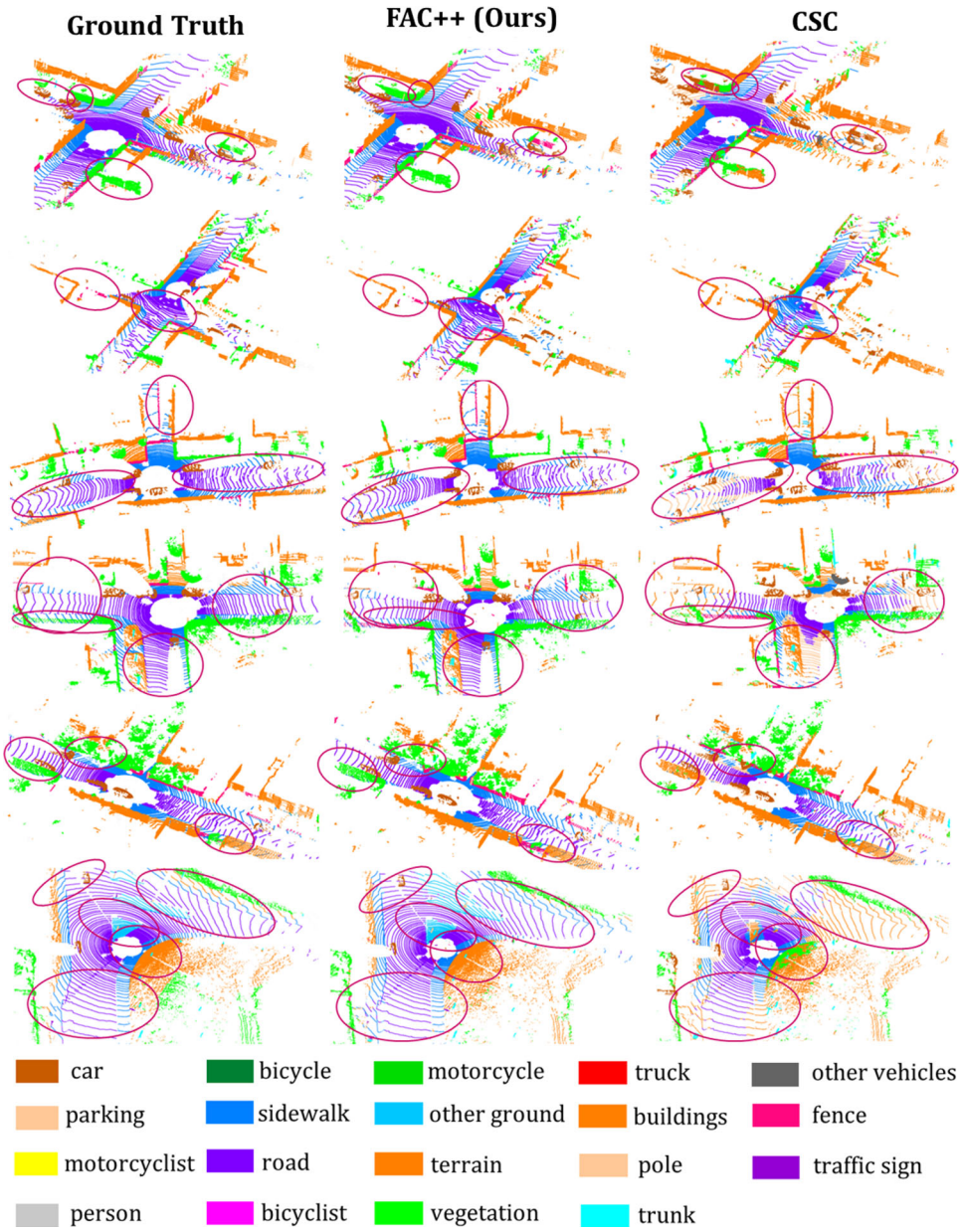

Fig. 12 Comparisons of outdoor *3D Semantic Segmentation* Results on SemanticKITTI (Behley et al., 2019) benchmark fine-tuned with *20% labeled training data* (ScanNet Dai et al. 2017 pre-trained). Note that the SemanticKITTI (Behley et al., 2019) has no color channel as input for the task of semantic segmentation. Therefore, we visualize the ground truth without visualizing the original scene (no color channel). Differences in prediction are highlighted by red ellipses

9.3 Data-Efficient Semantic Segmentation

The qualitative experimental results of semantic segmentation when fine-tuned with 10% and 20% labeled training data are shown in Figs. 11 and 12, respectively. It can be seen that CSC produces many false predictions, while FAC can provide more accurate semantic predictions as compared with the ground truth. It indicates more informative representation is learnt with FAC, which ultimately benefits the downstream semantic segmentation tasks. Also, it demonstrates that the model obtained from indoor pre-training on indoor ScanNet (Dai et al., 2017) can successfully generalize to outdoor SemanticKITTI (Behley et al., 2019) with data-efficient fine-tuning, which also manifests the generalization capacity of the learnt representation by FAC.

9.4 Data-Efficient Object Detection

The qualitative experimental results of object detection when fine-tuned with 20% and 50% labeled training data are shown in Figs. 13 and 14, respectively. It can be observed that compared with ProCo (Yin et al., 2022), our proposed FAC has clear more accurate predictions in detecting vehicles for outdoor sparse LiDAR point clouds in both 20% and 50% labeled training data cases. It further verifies that FAC learns generalized representations that can be applied for both segmentation and detection.

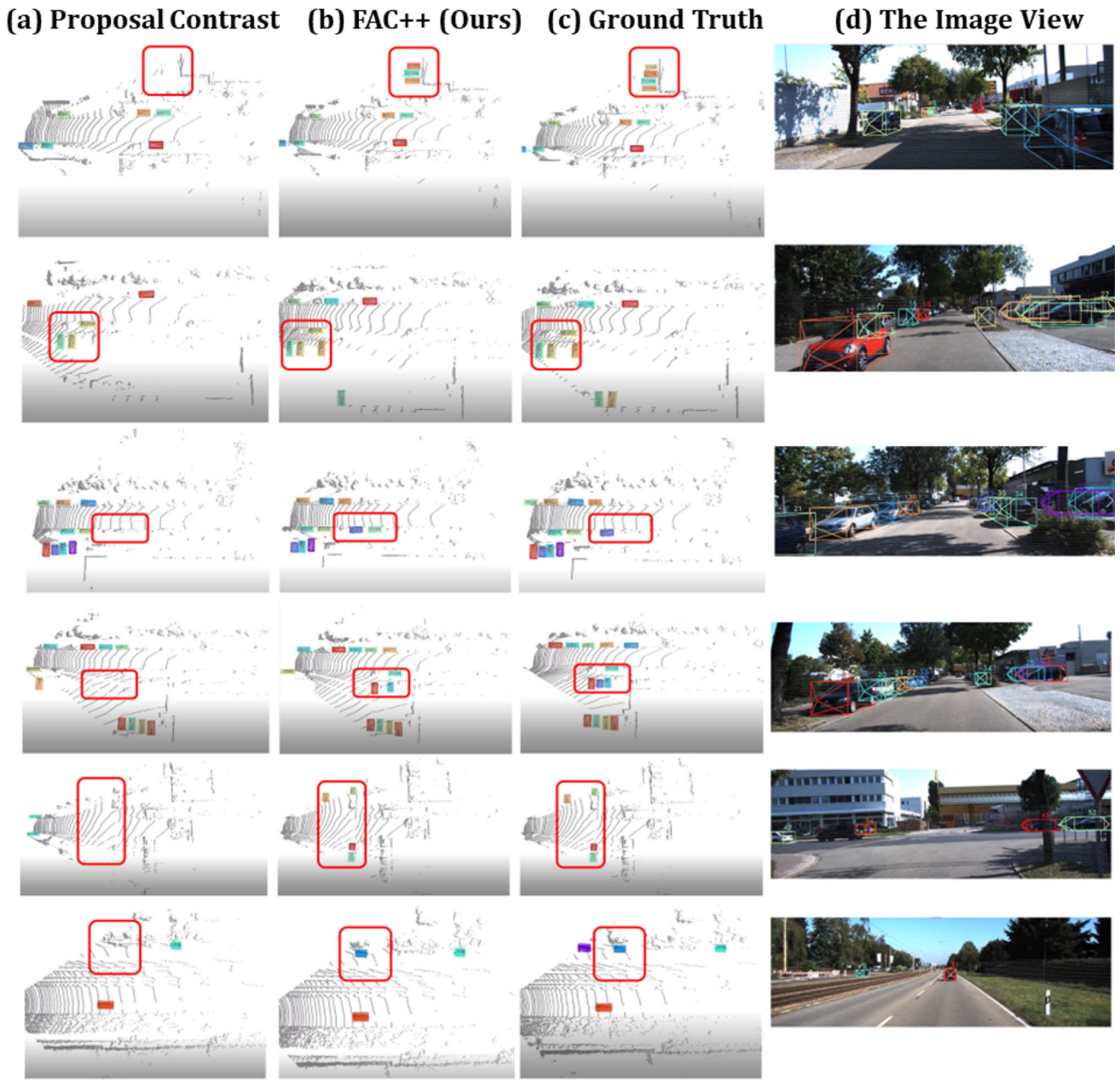

Fig. 13 Comparison of *3D Object detection* fine-tuned with *20% labeled training data* on KITTI benchmark (pre-trained on Waymo Sun et al. 2020) compared with the state-of-the-art approach ProCo (Yin et al., 2022). It can be seen that we can provide more accurate detection results as compared with the state-of-the-art approach ProCo (Yin et al., 2022). Different detected objects are indicated by different colors. Differences in prediction are also highlighted by red rectangles

10 Parameter Analysis of the Proposed GFC

First, as shown in Table 14, we examine the influence about the percentage of the foreground ratio on the downstream 3D scene understanding fine-tuning performance. It can be seen that too few and too large number of selected foreground region will both results in a performance drop for downstream tasks. It can be explained that selecting too few regions will overlook some smaller objects within the scene, and selecting a large number of selected regions will result in less foreground and more background regions being sampled, both impairing learning informative and meaningful representations. According to our parameter analysis, selecting within an appropriate range for the number of regions in each scene can all achieve satisfactory overall results for different downstream tasks. We obtain the range by the rig-

Fig. 14 Comparison of *3D Object detection* fine-tuned with *50% labeled training data* on KITTI benchmark (pre-trained on Waymo Sun et al. 2020) compared with the state-of-the-art approach ProCo (Yin et al., 2022). It can be seen that we can provide more accurate detection results as compared with the state-of-the-art approach ProCo (Yin et al., 2022). Different detected objects are indicated by different colors. Differences in prediction are also highlighted by red rectangles

Table 13 Comparison of the FAC and FAC++ training time with the other state-of-the-art 3D pre-training approaches PCon (Xie et al., 2020a) and CSC (Hou et al., 2021) on different pre-training datasets including ScanNet (Dai et al., 2017) and Waymo (Sun et al., 2020)

Pre-training dataset	Method	Inference time per 10^3 Million cloud		
		Sem	Ins	Obj
ScanNet (Dai et al., 2017)	PCon (Xie et al., 2020a)	1.987 s	2.126 s	1.238 s
	CSC (Hou et al., 2021)	2.178 s	2.356 s	1.387 s
	ProCo (Yin et al., 2022)	2.328 s	2.521 s	1.418 s
	FAC (Ours)	0.312 s	0.386 s	0.298 s (↓ **315.44**%)
	FAC++ (Ours)	0.365 s	0.416 s	0.322 s (↓ **294.89**%)
Waymo (Sun et al., 2020)	PCon (Xie et al., 2020a)	2.897s	3.126 s	2.626 s
	CSC (Hou et al., 2021)	3.218 s	3.529 s	2.729 s
	ProCo (Yin et al., 2022)	3.238 s	3.529 s	2.839 s
	FAC (Ours)	0.729 s	0.989 s	0.665 s (↓ **294.89**%)
	FAC++ (Ours)	0.767 s	0.728 s	0.605 s (↓ **334.05**%)

We have examined the tasks extensively for semantic segmentation, instance segmentation, as well as object detection, respectively. The efficiency boost mainly comes from the It can be demonstrated that our proposed approach has comparatively superior efficiency as well as effectiveness compared with the previous approaches. It can be proved clearly that due to our proposed regional contrastive designs which substitute the point-level contrast, our proposed approach has superior increment in its efficiency

orous parameter analysis as demonstrated in Table 14. It can be seen clearly that the performance can be well guaranteed if the foreground ratio is kept within the range of 20–80%.

11 Efficiency Analysis of the Proposed FAC

To test the efficiency of FAC, we reported the training time of diverse 3D pre-training approaches in Table 12. Specifically, we compare with state-of-the-art 3D pre-training approach PCon (Xie et al., 2020a) and CSC (Hou et al., 2021). It can be seen that compared with CSC (Hou et al., 2021), FAC introduces less than 1% training overhead on SemanticKITTI (Behley et al., 2019), and the fine-tuning time merely increases by approximately 9 s every epoch. The computational overhead mainly comes from the Siamese Correspondence network and the top-k operation. The SCN is light-weighted while the top-k operation has also been implemented with optimal transport in an efficient manner (Xie et al., 2020b) as illustrated in the main paper. In summary, the efficiency analysis with training time validates that our proposed FAC merely adds subtle extra computational overhead in pre-training.

11.1 Future Direction

In the future, we believe two directions deserve to be further explored to better unleash the potential of 3D unsupervised representation learning. The *first* is constructing large-scale 3D datasets with motion and spatio-temporal statistics for pre-training. The *second* is designing more advanced self-supervised learning techniques leveraging both geometry-aware and semantics-correlated features considering motion and spatio-temporal statistical cues (Tables 15, 16).

12 Conclusion

We propose a *foreground-aware* feature contrast framework (FAC) for unsupervised 3D pre-training in robot 3D vision-based scene parsing. The proposed framework has been proven very effective in constructing a generalized robotic vision-language learning model leveraging linguistic foreground aware contrast. FAC builds better contrastive pairs to produce more geometrically informative and semantically meaningful 3D representations. Specifically, we design a regional sampling technique to promote balanced learning of over-segmented foreground regions and eliminate noisy ones, which facilitates building foreground-aware contrast pairs based on regional correspondence. Moreover, we enhance foreground-background distinction and propose a plug-in-play Siamese correspondence network to find the well-correlated feature contrast pairs within and across views for both the foreground and background segments. Extensive experiments demonstrate the effectiveness as well as the superiority of FAC in terms of both the knowledge transfer and the data efficiency.

Table 14 Ablation studies on using different percentage of over-segmented foreground regions provided by the VCCS (Papon et al., 2013)

Selected foreground point ratios in over-seg %	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Ins. AP@50% (Sc)	Det. AP@50% (Sc)	Det. mAP% (K)
5	50.25	39.89	42.89	19.67	67.85
10	51.37	40.87	42.79	18.98	67.96
15	51.78	41.56	44.97	20.76	68.43
20	51.95	41.75	44.88	**20.96**	68.11
30	**52.71**	**41.93**	**45.20**	20.95	68.42
40	51.92	40.86	44.92	20.85	**68.71**
60	50.98	40.85	44.53	20.89	67.85
75	51.87	39.89	43.83	20.45	67.56
80	52.69	40.36	43.98	20.87	67.97
90	51.88	39.93	44.85	20.69	67.76
100	52.31	40.69	44.21	20.22	67.39
The instance segmentation ground truth	**53.69**	**41.97**	**45.38**	**21.89**	**69.05**

The tasks examined for the pre-training encompass semantic segmentation (mIoU%), instance segmentation (AP@50%), as well as object detection (mAP%) on ScanNet (Sc) (Dai et al., 2017), SemanticKITTI (SK) (Behley et al., 2019) and KITTI (K), respectively. In the meanwhile, it can also be demonstrated that directly utilizing the instance ground truth will help FAC++ obtain the best results. However, the increment is slight over simply utilizing off-the-shelf vanilla instance segmentation approaches. Utilizing the instance ground truth will exert huge burden for the labeling efforts, which will make it hard for the real-world robotic applications. Therefore, we directly use the oversegmentation approach of the VCCS for generating the foreground regions rather than utilize the ground truth

Table 15 Ablation study of the rational equivalent data augmentation in FAC++ for downstream scene understanding tasks on ScanNet (Sc) ((Dai et al., 2017) and SemanticKITTI (SK) (Behley et al., 2019), & KITTI (K) for the tasks of semantic segmentation, instance segmentation, as well as the object detection in the case of z / SO(3) as well as SO(3) / SO(3)

Experimental results for the z / SO(3) as well as SO(3) / SO(3)	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Det. AP@50% (Sc)	Det. mAP% (K)
Basic data augmentation	48.87 / 50.28	42.21 / 45.98	36.76 / 39.71	64.98 / 68.59
SO(3) Equivalent data augmentation (Esteves et al., 2018)	49.32 / 51.58	43.76 / 47.78	37.61 / 40.19	66.98 / 68.96
SPConv-ShellNet	51.76 / 53.69	45.78 / 49.26	38.59 / 40.56	67.31 / 69.32
Rot-Equiv-FG (Liu et al., 2022, 2020, 2021)	50.16 / 53.46	45.76 / 48.67	39.21 / 39.67	67.89 / 69.56
SPConv-PointCNN	50.86 / 53.98	45.87 / 49.28	39.87 / 40.26	67.95 / 69.78
SE(3) Transformer	52.55 / 53.98	45.78 / 49.86	40.21 / 40.67	68.27 / 69.86
Vector neurons (Deng et al., 2021)	52.89 / 53.97	45.82 / 49.69	40.27 / 41.78	69.61 / 69.98
Equivalent GNN (Passaro & Zitnick, 2023)	53.36 / 54.87	45.73 / 49.82	40.66 / 42.76	70.16 / 70.68
Equivact (Yang et al., 2024)	53.86 / 55.86	45.86 / 49.97	40.95 / 42.81	71.23 / 71.97
Equal-motion (Xu et al., 2023)	**54.53 / 56.95**	**45.98 / 50.56**	**40.96 / 42.97**	**71.62 / 72.31**

It can be demonstrated explicitly that the rotational equivariant frameworks also have boost on the final scene parsing performance. It can be demonstrated the rotational equivalent framework will have a marginal boost on the ultimate semantic scene parsing performance. Therefore, the data augmentation approaches also have a large impact on the final scene parsing performance, although the training speed might be compromised with approximately 6% according to our experimental results

Table 16 Ablation study of different over-segmentation approaches on the final scene parsing performance

Experimental results for the z / SO(3) as well as SO(3) / SO(3)	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Det. AP@50% (Sc)	Det. mAP% (K)
Visual-similarity (Chen et al., 2003)	51.96 / 52.46	44.51 / 46.75	38.76 / 39.89	65.97 / 68.69
HKS (Sun et al., 2009)	52.89 / 53.28	45.16 / 47.96	38.21 / 39.87	65.52 / 68.96
SIKS (Bronstein & Kokkinos, 2010)	50.65 / 52.65	46.98 / 48.98	40.56 / 40.87	66.77 / 68.51
F-PFH (Rusu et al., 2009)	49.26 / 50.38	45.87 / 47.76	39.61 / 40.67	67.89 / 69.18
WKS (Aubry et al., 2011)	50.47 / 52.98	45.97 / 47.68	39.87 / 40.89	67.82 / 69.98
SHOT (Salti et al., 2014)	50.55 / 53.68	46.28 / 48.97	40.61 / 41.87	68.12 / 70.38
VCCS (Papon et al., 2013)	53.59 / 55.26	45.98 / 49.29	40.96 / 42.63	69.61 / 70.89
Superpoint-graph (Landrieu & Simonovsky, 2018)	54.26 / 55.97	43.13 / 45.61	41.36 / 42.12	70.16 / 70.78
Original PFH (Rusu et al., 2008)	54.29 / 55.86	45.96 / 47.98	41.95 / 42.21	71.23 / 71.97
Adapted PFH (Liu, 2023d)	54.36 / 55.75	45.98 / 48.56	41.96 / 43.27	71.57 / 72.61
Point-SIFT (Jiang et al., 2018)	54.53 / 56.85	46.95 / 49.78	42.67 / 43.98	71.62 / 72.31
S-SPG (Landrieu & Boussaha, 2019)	54.53 / 56.92	46.98 / 50.86	43.56 / 44.57	71.76 / 73.87
Predator (Huang et al., 2021)	**55.67 / 56.95**	**46.88 / 51.26**	**43.98 / 44.69**	**71.67 / 73.98**

The tested appraoches are for ScanNet (Sc) (Dai et al., 2017) and SemanticKITTI (SK) (Behley et al., 2019), & KITTI (K) for the tasks of semantic segmentation, instance segmentation, as well as the object detection in the case of z / SO(3) as well as SO(3) / SO(3). It can be demonstrated the approaches utilized for the over-segmentation merely have a slight impact on the final semantic scene parsing performance. The reason behind can be attribute to the regional representations are comparatively more robust compared with the point-level ones. Meanwhile, combined with our proposed foreground prompted visual-linguistic aligned feature contrastive designs, our proposed approach can attain superior performance robustness regarding the over-segmentation approaches selected

Acknowledgements This work is supported by National Natural Science Foundation of China Fund (No. 62403400) leading by Prof. Kangcheng Liu. Prof. Kangcheng Liu is the sole corresponding author of this work.

References

Aubry, M., Schlickewei, U., & Cremers, D. (2011). The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 1626–1633). IEEE.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966

Bai, Y., Chen, X., Kirillov, A., Yuille, A.,& Berg, A. C. (2022). Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16,061–16,070).

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., & Gall, J. (2019). SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9297–9307).

Bronstein, M. M., & Kokkinos, I. (2010). Scale-invariant heat kernel signatures for non-rigid shape recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 1704–1711). IEEE.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11,621–11,631).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer

Chen, D. Y., Tian, X. P., Shen, Y. T., & Ouhyoung, M. (2003). On visual similarity based 3D model retrieval. In *Computer graphics forum* (vol. 22, pp. 223–232). Wiley Online Library.

Chen, Y., Nießner, M., & Dai, A. (2022). 4DContrast: Contrastive learning with dynamic correspondences for 3D scene understanding. In *European Conference on Computer Vision*. Springer.

Cheraghian, A., Rahman, S., Campbell, D., & Petersson, L. (2020). Transductive zero-shot learning for 3D point cloud classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 923–933).

Chibane, J., Engelmann, F., Anh Tran, T., & Pons-Moll, G. (2022). Box2Mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In: European conference on computer vision (pp. 681–699). Springer.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828–5839).

Deng, C., Litany, O., Duan, Y., Poulenard, A., Tagliasacchi, A., & Guibas, L. J. (2021). Vector neurons: A general framework for SO (3)-equivariant networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12,200–12,209).

Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., & Qi, X. (2023). PLA: Language-driven open-vocabulary 3D scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7010–7019).

Erçelik, E., Yurtsever, E., Liu, M., Yang, Z., Zhang, H., Topçam, P., Listl, M., Çaylı, Y. K., & Knoll, A. (2022). 3D object detection with a self-supervised lidar scene flow backbone. In *European Conference on Computer Vision*

Esteves, C., Allen-Blanchette, C., Makadia, A., & Daniilidis, K. (2018). Learning SO (3) equivariant representations with spherical CNNs. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 52–68).

Guo, Y., Bennamoun, M., Sohel, F., Lu, M., & Wan, J. (2014). 3D object recognition in cluttered scenes with local surface features: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(11), 2270–2287.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16,000–16,009).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

Hou, J., Graham, B., Nießner, M., Xie, S. (2021). Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15,587–15,597).

Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., & Schindler, K. (2021) PREDATOR: Registration of 3D point clouds with low overlap. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4267–4276).

Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S. C. (2018). Holistic 3D scene parsing and reconstruction from a single RGB image. In *Proceedings of the European conference on computer vision* (pp. 187–203).

Huang, S., Xie, Y., Zhu, S. C., & Zhu, Y. (2021). Spatio-temporal self-supervised representation learning for 3D point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6535–6545).

Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C. W., & Jia, J. (2020). PointGroup: Dual-set point grouping for 3D instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4867–4876).

Jiang, M., Wu, Y., Zhao, T., Zhao, Z., & Lu, C. (2018). PointSIFT: A sift-like network module for 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kuhn, H. W. (2005). The method for the assignment problem. *Naval Research Logistics (NRL), 52*(1), 7–21.

Landrieu, L., & Boussaha, M. (2019). Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7440–7449).

Landrieu, L., & Simonovsky, M. (2018). Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4558–4567)

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12,697–12,705)

Li, L., & Heizmann, M. (2022). A closer look at invariances in self-supervised pre-training for 3d vision. In *European Conference on Computer Vision*. Springer.

Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., Liang, X., Zhang, W., Li, Z., & Van Gool, L. (2021). Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3293–3302).

Liang, Y., Zhao, S., Yu, B., Zhang, J., & He, F. (2022). MeshMAE: Masked autoencoders for 3d mesh data analysis. In *European conference on computer vision*

Liu, K. (2023a). Learning-based defect recognitions for autonomous UAV inspections. arxiv preprint arxiv:2302.06093

Liu, K. (2022a). An enhanced lidar-inertial slam system for robotics localization and mapping. arXiv preprint arXiv:2212.14209

Liu, K. (2022b). An integrated lidar-slam system for complex environment with noisy point clouds. arXiv preprint arXiv:2212.05705

Liu, K. (2022c). An integrated visual system for unmanned aerial vehicles tracking and landing on the ground vehicles. arXiv preprint arXiv:2301.00198

Liu, K. (2022d). A robust and efficient lidar-inertial-visual fused simultaneous localization and mapping system with loop closure. In *2022 12th International conference on cyber technology in automation, control, and intelligent systems (CYBER)* (pp. 1182–1187). IEEE.

Liu, K. (2022e). Robust industrial UAV/UGV-based unsupervised domain adaptive crack recognitions with depth and edge awareness: from system and database constructions to real-site inspections. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 5361–5370)

Liu, K. (2022f). Semi-supervised confidence-level-based contrastive discrimination for class-imbalanced semantic segmentation. In *2022 12th International conference on CYBER technology in automation, control, and intelligent systems (CYBER)* (pp. 1230–1235). IEEE.

Liu, K. (2023b). Learning-based defect recognitions for autonomous UAV inspections. arXiv preprint arXiv:2302.06093

Liu, K. (2023c). A lidar-inertial-visual slam system with loop detection. arXiv preprint arXiv:2301.05604

Liu, K. (2023d). RM3D: Robust data-efficient 3D scene parsing via traditional and learnt 3D descriptors-based semantic region merging. *International Journal of Computer Vision, 131*(4), 938–967.

Liu, K., & Cao, M. (2023). DLC-SLAM: A robust LiDAR-slam system with learning-based denoising and loop closure. *IEEE/ASME Transactions on Mechatronics, 28*(5), 2876–2884.

Liu, K., & Chen, B. M. (2022a). Industrial UAV-based unsupervised domain adaptive crack recognitions: From database towards real-site infrastructural inspections. *IEEE Transactions on Industrial Electronics, 70*(9), 9410–9420.

Liu, K., & Chen, B. M. (2022b). Industrial UAV-based unsupervised domain adaptive crack recognitions: From system setups to real-site infrastructural inspections. *IEEE Transactions on Industrial Electronics, 70*, 9410–9420.

Liu, K., Gao, Z., Lin, F., & Chen, B.M.(2020). FG-Net: Fast large-scale lidar point cloudsunderstanding network leveraging correlated-feature mining and geometric-aware modelling. arXiv preprint arXiv:2012.09439

Liu, K., Gao, Z., Lin, F., & Chen, B. M.(2021). FG-Conv: Large-scale LiDAR point clouds understanding leveraging feature correlation mining and geometric-aware modeling. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 12,896–12,902). IEEE.

Liu, K., Gao, Z., Lin, F., & Chen, B. M. (2022). FG-Net: A fast and accurate framework for large-scale lidar point cloud understanding. *IEEE Transactions on Cybernetics, 53*(1), 553–564.

Liu, K., Han, X., & Chen, B. M. (2019). Deep learning based automatic crack detection and segmentation for unmanned aerial vehicle inspections. In *2019 IEEE international conference on robotics and biomimetics (ROBIO)* (pp. 381–387). IEEE.

Liu, K.,& Ou, H. (2022a). A light-weight lidar-inertial slam system with high efficiency and loop closure detection capacity. In *2022 International conference on advanced robotics and mechatronics (ICARM)* (pp. 284–289). IEEE.

Liu, K.,& Ou, H. (2022b). A light-weight lidar-inertial slam system with loop closing. arXiv preprint arXiv:2212.05743

Liu, K., Qu, Y., Kim, H. M., & Song, H. (2017). Avoiding frequency second dip in power unreserved control during wind power rotational speed recovery. *IEEE Transactions on Power Systems, 33*(3), 3097–3106.

Liu, K., Xiao, A., Huang, J., Cui, K., Xing, Y., & Lu, S. (2022a). D-LC-Nets: Robust denoising and loop closing networks for lidar slam in complicated circumstances with noisy point clouds. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3097–3106).

Liu, K., Xiao, A., Huang, J., Cui, K., Xing, Y., & Lu, S. (2022b). D-LC-Nets: Robust denoising and loop closing networks for lidar slam in complicated circumstances with noisy point clouds. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 12,212–12,218). IEEE.

Liu, K., Xiao, A., Zhang, X., Lu, S., & Shao, L. (2023). FAC: 3D representation learning via foreground aware feature contrast. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9476–9485).

Liu, K., Yang, G., Zhang, J., Zhao, Z., Chen, X., & Chen, B. M. (2022c). Datasets and methods for boosting infrastructure inspection: A survey on defect segmentation and detection. In *2022 IEEE 17th international conference on control & automation (ICCA)* (pp. 23–30). IEEE.

Liu, K., Zhao, Y., Gao, Z., & Chen, B. M. (2022d). WeakLabel3D-Net: A complete framework for real-scene lidar point clouds weakly supervised multi-tasks understanding. In *2022 international conference on robotics and automation (ICRA)* (pp. 5108–5115). IEEE.

Liu, K., Zhao, Y., Nie, Q., Gao, Z., & Chen, B. M. (2022e). Weakly supervised 3d scene segmentation with region-level boundary awareness and instance discrimination. In *European Conference on Computer Vision 2022 (ECCV 2022)* (pp. 37–55). Springer, Cham.

Liu, K., Zheng, X., Wang, C., Wang, H., Liu, M., & Tang, K. (2024). Online robot navigation and and manipulation with distilled vision-language models. arXiv preprint arXiv:2401.17083

Liu, K., Zhou, X., & Chen, B. M. (2022f). An enhanced lidar inertial localization and mapping system for unmanned ground vehicles. In *2022 IEEE 17th international conference on control & automation (ICCA)* (pp. 587–592). IEEE.

Liu, K., Zhou, X., Zhao, B., Ou, H., & Chen, B. M. (2022g). An integrated visual system for unmanned aerial vehicles following ground vehicles: Simulations and experiments. In *2022 IEEE 17th international conference on control & automation (ICCA)* (pp. 593–598). IEEE.

Liu, M., Zhou, Y., Qi, C. R., Gong, B., Su, H., & Anguelov, D. (2022h). Less: Label-efficient semantic segmentation for lidar point clouds. In *European conference on computer vision* (pp. 70–89). Springer.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021a). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10,012–10,022).

Liu, Z., Qi, X., & Fu, C. W. (2021b). One thing one click: A self-training approach for weakly supervised 3D semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1726–1736).

Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., & Xu, C. (2021). Voxel transformer for 3D object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3164–3173).

Michele, B., Boulch, A., Puy, G., Bucher, M., & Marlet, R. (2021). Generative zero-shot learning for semantic segmentation of 3D point clouds. In *2021 International Conference on 3D vision (3DV)* (pp. 992–1002). IEEE.

Oord, A.v.d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

Pang, Y., Wang, W., Tay, F. E., Liu, W., Tian, Y., & Yuan, L. (2022). Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*. Springer.

Papon, J., Abramov, A., Schoeler, M., & Worgotter, F. (2013). Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2027–2034).

Passaro, S., & Zitnick, C. L. (2023). Reducing SO (3) convolutions to SO (2) for efficient equivariant GNNs. In *International conference on machine learning* (pp. 27,420–27,438). PMLR.

Rao, Y., Liu, B., Wei, Y., Lu, J., Hsieh, C. J., & Zhou, J. (2021). RandomRooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3D object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3283–3292).

Rozenberszki, D., Litany, O., & Dai, A. (2022). Language-grounded indoor 3D semantic segmentation in the wild. In *European conference on computer vision* (pp. 125–141). Springer.

Rusu, R. B., Blodow, N., & Beetz, M. (2009). Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation (ICRA)* (pp. 3212–3217). IEEE.

Rusu, R. B., Blodow, N., Marton, Z. C., & Beetz, M. (2008). Aligning point cloud views using persistent feature histograms. In *2008 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 3384–3391). IEEE.

Salti, S., Tombari, F., & Di Stefano, L. (2014). Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding, 125*, 251–264.

Sanghi, A. (2020). Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning. In *European conference on computer vision* (pp. 626–642). Springer

Shi, S., Wang, X., & Li, H.(2019). PointRCNN: 3D object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–779).

Shi, S., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(8), 2647–2664.

Sun, J., Ovsjanikov, M., & Guibas, L. (2009). A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum* (vol. 28, pp. 1383–1392). Wiley Online Library.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., & Vasudevan V. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2446–2454).

Takmaz, A., Fedele, E., Sumner, R. W., Pollefeys, M., Tombari, F., & Engelmann, F. (2023). OpenMask3D: Open-vocabulary 3D instance segmentation. arXiv preprint arXiv:2306.13631

Team, O. (2020). OpenPCDet: An open-source toolbox for 3d object detection from point clouds. OD Team

Uy, M. A., Pham, Q. H., Hua, B. S., Nguyen, T., & Yeung, S. K. (2019). Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1588–1597).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11), 2579–2605.

Vu, T., Kim, K., Luu, T. M., Nguyen, T., Yoo, & C. D. (2022). SoftGroup for 3D instance segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2708–2717).

Wang, H., Cong, Y., Litany, O., Gao, Y., & Guibas, L. J. (2021a). 3DIoUMatch: Leveraging IoU prediction for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14,615–14,624).

Wang, H., Liu, Q., Yue, X., Lasenby, J., & Kusner, M. J. (2021b). Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9782–9792).

Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L., & Litany, O. (2020a). PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *European Conference on Computer Vision* (pp. 574–591). Springer.

Xie, Y., Dai, H., Chen, M., Dai, B., Zhao, T., Zha, H., Wei, W., & Pfister, T. (2020b). Differentiable top-k with optimal transport. In *Advances in neural information processing systems* (vol. 33, pp. 20520–20531).

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9653–9663).

Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, Y.G., Wang, X., & Wang, Y. (2023). EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1410–1420).

Yang, G., Liu, K., Zhao, Z., Zhang, J., Chen, X., & Chen, B. M. (2022). Datasets and methods for boosting infrastructure inspection: A survey on defect classification. In *2022 IEEE 17th international conference on control & automation (ICCA)* (pp. 15–22). IEEE.

Yang, J., Deng, C., Wu, J., Antonova, R., Guibas, L., & Bohg, J. (2024). EquivAct: SIM(3)-equivariant visuomotor policies beyond rigid object manipulation. In *2024 IEEE international conference on robotics and automation (ICRA)* (pp. 9249–9255). IEEE.

Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C. Z., Shen, J., & Wang, W. (2022). ProposalContrast: Unsupervised pre-training for LiDAR-based 3D object detection. In *European conference on computer vision* (pp. 574–591). Springer.

Zhang, Z., Bai, M., & Li, E. (2022). Self-supervised pretraining for large-scale point clouds. In *Advances in neural information processing systems*.

Zhang, Z., Sun, B., Yang, H., & Huang, Q. (2020). H3dnet: 3D object detection using hybrid geometric primitives. In *European conference on computer vision* (pp. 311–329). Springer.

Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4490–4499).