FAC: 3D Representation Learning via Foreground Aware Feature Contrast

Kangcheng Liu[1,†], Aoran Xiao[1], Xiaoqin Zhang[2], Shijian Lu[1,†], Ling Shao[3]
[1]Nanyang Technological University [2]Wenzhou University [3]UCAS-Terminus AI Lab, UCAS

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 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. 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 and object detection tasks. All codes, data, and models are available.

1. Introduction

3D scene understanding is crucial to many tasks such as robot grasping and autonomous navigation [12, 21, 30]. However, most existing work is fully supervised which relies heavily on large-scale annotated 3D data that is often laborious to collect. Self-supervised learning (SSL), which allows learning rich and meaningful representations from large-scale unannotated data, has recently demonstrated great potential to mitigate the annotation constraint [1, 5]. 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 2D recognition tasks [6, 29].

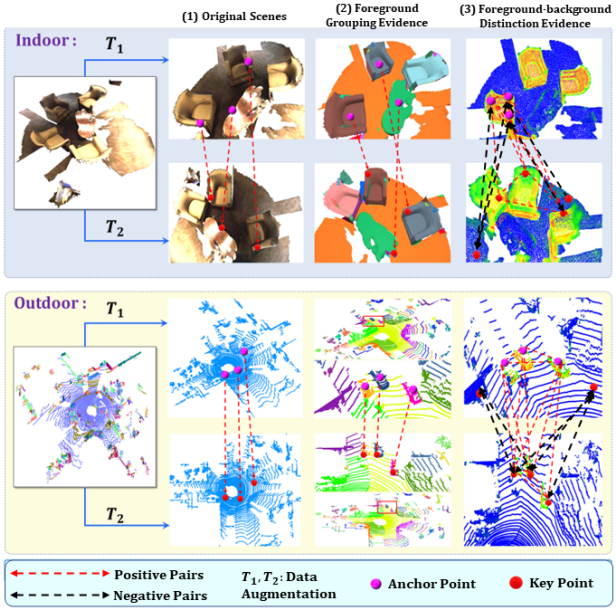

Figure 1. Constructing informative contrast pairs matters in contrastive learning: Conventional contrast requires strict point-level correspondence. The proposed method FAC takes both foreground grouping and foreground-background distinction cues into account, thus forming better contrast pairs to learn more informative and discriminative 3D feature representations.

Contrastive learning has also been explored for point cloud representation learning in various downstream tasks such as semantic segmentation [7, 18, 22, 42], instance segmentation [19, 20], and object detection [26, 44]. However, many successful 2D contrastive learning methods [6, 14, 46] 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, [22] employ max-pooled features of two augmented scenes to form the contrast, but they tend to overemphasize holistic information and overlook informative

† Corresponding Authors.

features about foreground objects. [19, 26, 42] 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 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* 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 foreground-enhanced contrast guides the learning toward foreground regions while foreground-background contrast enhances distinctions among foreground and background features effectively in a complementary manner, the two collaborating to learn more informative and discriminative representation as illustrated in Fig. 1.

The contributions of this work can be summarized in three aspects. *First*, we propose FAC, a foreground-aware feature contrast framework for large-scale 3D pre-training. *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. *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 [15] and 3D detection backbone networks including PV-RCNN, PointPillars [25], and PointRCNN [36]. It is also applicable to both indoor dense RGB-D and outdoor sparse LiDAR point clouds.

2. Related Work

2.1. 3D Scene Understanding

3D scene understanding aims for the understanding of 3D depth or point-cloud data and it involves several downstream tasks such as 3D semantic segmentation [9, 31], 3D object detection [11], etc. It has recently achieved very impressive progress as driven by the advance in 3D deep learning strategy and the increasing large-scale 3D datasets. Different approaches have been proposed to address various challenges in 3D scene understanding. For example, the point-based approach can learn point clouds well but is often stuck by high computational costs while facing large-scale point-cloud datasets. Voxel-based approach [10, 32, 47] is computation and memory efficient but often suffers from information loss from the voxel quantification. In addition, voxel-based SparseConv network [15, 39] has shown very promising performance in indoor scene segmentation, while combining point and voxel often has a clear advantage in outdoor LiDAR-based detection [25, 47]. 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

Contrastive Pre-training. Recent years have witnessed notable success in contrastive learning for learning unsupervised representations. For example, contrast scene context (CSC) [19, 42] explores contrastive pre-training with scene context descriptors. However, it focuses too much on optimizing low-level registered point features but overlooks the regional homogeneous semantic patterns and high-level feature correlations. Some work employs max-pooled scene-level information for contrast [22, 26], 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 [22] and synthetic CAD models [8, 34, 41, 43]. STRL [22] introduces a mechanism of learning from dynamic 3D scenes synthetic 3D by regarding 3D scenes are RGB-D video sequences. Randomrooms [34] synthesizes man-made 3D scenes by randomly putting synthetic CAD models into regular synthetic 3D scenes. 4DContrast [8] leverages spatio-temporal motion priors of synthetic 3D shapes to learn a better 3D representation.

Masked Generation-based Pre-training. Masked image modeling has demonstrated its effectiveness in various image understanding tasks with the success of vision transformers [3]. Recently, mask-based pre-training [17, 27, 33, 35, 40] has also been explored for the understanding of

small-scale 3D shapes [4,37,41]. However, mask-based designs usually involve a transformer backbone [27,28,33,45] that has a high demand for both computation and memory while handling large-scale 3D scenes.

3. Method

As illustrated in Fig. 2, our proposed FAC framework is composed of four components: data augmentation, backbone network feature extraction, feature matching, and foreground-background aware feature contrastive optimizations with matched contrast pairs. 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 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) [42] or scene level as in DepthContrast (DCon). 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 loss can be applied for contrast:

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

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 [42] directly adopts registered point-level feature pairs while DCon uses the max-pooled scene-level feature pairs for 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 very useful in considering contexts for downstream tasks such as semantic segmentation and detection. In our proposed geometric region-level foreground-aware contrast, we obtain regions by leveraging the off-the-shelf point cloud over-segmentation techniques [16]. The adoption of over-segmentation is motivated by its merits in three major aspects. First, it can work in a completely unsupervised 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 framework FAC works effectively with mainstream over-segmentation approaches without fine-tuning.

Regional Sampling for Balanced Learning. We design a simple but effective region sampling technique to obtain meaningful foreground from the geometrically homogeneous regions as derived via over-segmentation. Specifically, we first count the number of points in each region and rank regions according to the number of points it contains. 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 the foreground regions which leads to more informative and discriminative representations.

Contrast for Local Regional Consistency. Different from the above-mentioned PCon [42] and DCon, we directly exploit region homogeneity to obtain contrast pairs. Specifi-

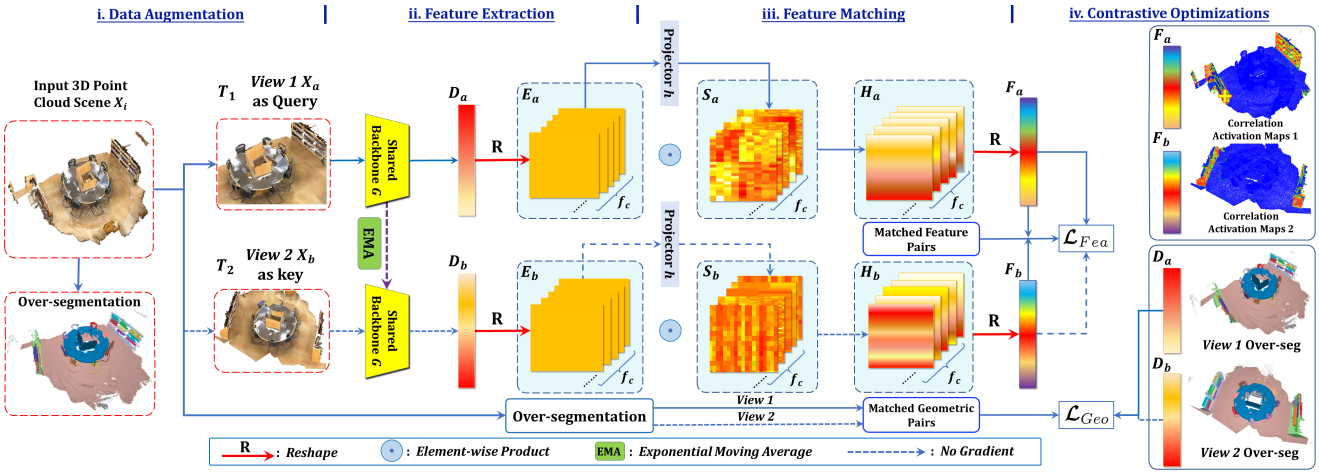

Figure 2. **The framework of our proposed FAC.** FAC takes two augmented 3D point cloud views as input which first extracts the backbone features $\boldsymbol{D}_a$ and $\boldsymbol{D}_b$ for foreground aware contrast with $\boldsymbol{\mathcal{L}}_{Geo}$. The backbone features are then reshaped to regularized 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 $\boldsymbol{\mathcal{L}}_{Fea}$. Hence, FAC exploits complementary foreground awareness and foreground-background distinction within and across views for more informative representation learning.

cally, 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 $\boldsymbol{\mathcal{D}}_m$ within a region as the anchor in contrast to enhance robustness:

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

Regarding $\boldsymbol{\mathcal{D}}_m$ as the anchor, we propose a foreground aware geometry contrast loss $\boldsymbol{\mathcal{L}}_{Geo}$ 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:

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

Here, $\boldsymbol{d}_j^+$ and $\boldsymbol{d}_j^-$ denote the positive and negative samples with $\boldsymbol{\mathcal{D}}_m$, respectively. 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 [42] 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 [19] in data efficiency in empirical experimental results.

3.3. Foreground-Background 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.

Siamese Correspondence Network for Adaptive Correlation Mining. Given the input 3D scene $\boldsymbol{X}_i$ with N points, FAC 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 [19, 44]. 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 min-

ing. The global feature-level discriminative representation learning is enhanced by the proposed SCN, enabling subsequent contrast with the matched feature.

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.

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

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. FAC is verified as effective qualitatively in point activation maps and quantitatively in downstream transfer learning and data efficiency.

3.4. Joint Optimization of FAC

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

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

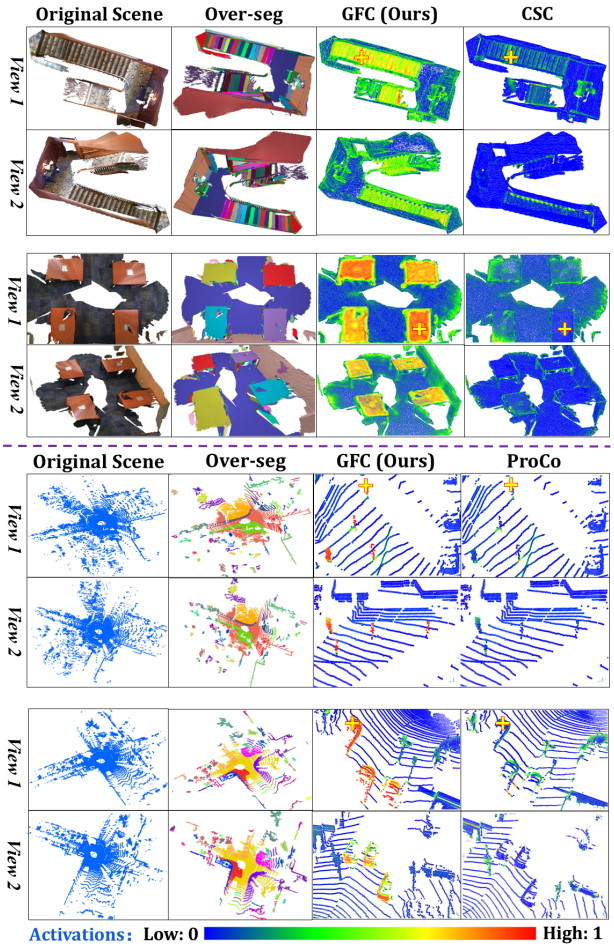

Figure 3. Visualizations of projected point correlation maps over the indoor ScanNet (1st-4th rows) and the outdoor KITTI [13] (5th-8th rows) with respect to the query points highlighted by yellow crosses. The *View 1* and *View 2* in each sample show the intra-view and cross-view correlations, respectively. We compare FAC with the state-of-the-art CSC [19] on segmentation (rows 1-4) and ProCo [44] on detection (rows 5-8). FAC clearly captures better feature correlations within and across views (columns 3-4).

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 learnt unsupervised representations [19]. 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

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 [36]	*From Scratch*	63.51	88.64	75.23	72.47	55.49	48.90	42.23	85.41	66.39	61.74
		ProCo [44]	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**
	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 [44]	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**
100% (about 3.71k frames)	PointRCNN [36]	*From Scratch*	69.45	90.02	80.56	78.02	62.59	55.66	48.69	89.87	72.12	67.52
		DCon	70.26	89.38	80.32	77.92	65.55	57.62	50.98	90.52	72.84	68.22
		ProCo [44]	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**
	PV-RCNN	*From Scratch*	70.57	-	84.50	-	-	57.06	-	-	70.14	-
		GCC-3D [26]	71.26	-		-	-	-	-	-	-	-
		STRL [22]	71.46	-	84.70	-	-	57.80	-	-	71.88	-
		PCon [42]	71.55	91.40	84.18	82.25	65.73	57.74	52.46	91.47	72.72	67.95
		ProCo [44]	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**

Table 1. Data-efficient ***3D object detection*** on KITTI [13]. We pre-train the backbone network of PointRCNN [36] and PV-RCNN on Waymo and transfer to KITTI with 20% and 100% annotation ratios in fine-tuning. FAC outperforms the state-of-the-art ProCo [44] consistently for two settings. *'From Scratch'* denotes the model trained from scratch. All experimental results are averaged over three runs.

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 [25]	-	*From Scratch*	23.05	18.08	27.15	26.17	30.31	18.79	11.28	9.28
		ProCo [44]	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**
	VoxelNet [47]	-	*From Scratch*	20.88	17.83	21.95	21.45	27.98	20.52	12.70	11.53
		ProCo [44]	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**
10% (around 8k frames)	PointPillars [25]	-	*From Scratch*	51.75	46.58	54.94	54.32	54.01	41.53	46.31	43.88
		ProCo [44]	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**
	VoxelNet [47]	-	*From Scratch*	54.04	51.24	54.37	53.74	51.45	45.05	56.30	54.93
		ProCo [44]	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**

Table 2. Data-efficient ***3D object detection*** experimental results on Waymo with 1% and 10% labeled training data. Similar experimental results are obtained for KITTI in Table 1 for FAC as compared with the state-of-the-art ProCo [44].

Label Ratio	10%	20%	40%	80%	100%
From Scratch	0.39	4.67	22.09	33.75	35.48
CSC [19]	8.68	20.96	29.27	36.75	39.32
ProCo [44]	12.64	21.87	31.95	37.83	40.56
FAC (Ours)	**20.96**	**27.35**	**35.93**	**39.91**	**42.83**

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

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 [19]	29.32	49.93	59.45	64.63	68.96
	FAC (Ours)	**35.25**	**51.95**	**61.28**	**65.84**	**69.52**
S3DIS Sem. Seg.	*From Scratch*	35.75	44.38	51.86	58.72	61.83
	CSC [19]	36.48	45.07	52.95	59.93	62.65
	FAC (Ours)	**43.73**	**49.28**	**54.76**	**61.05**	**63.22**
SemanticKITTI [2] Sem. Seg.	*From Scratch*	28.36	33.58	46.37	50.15	54.56
	CSC [19]	32.78	37.55	49.62	55.67	58.89
	FAC (Ours)	**39.92**	**41.75**	**52.37**	**57.65**	**60.17**

Table 4. Data efficient ***3D semantic segmentation*** (mIoU%) results with the limited scene reconstructions [19] on ScanNet, S3DIS, and SemanticKITTI (SK) [2] with diverse label ratios.

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

3D Object Detection. The object detection experiments involve two backbones including VoxelNet [47] and PointPillars [25]. Following ProCo [44], we pre-train the model on Waymo and fine-tune it on KITTI [13] and Waymo. Following ProCo [44] and CSC [19], 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 [44], 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 [13], we pre-train the network with Adam [23] optimizer and follow ProCo [44] for epoch and batch size setting for fair comparisons with existing works [26, 44]. In indoor object

detection on ScanNet, we follow CSC [19] to adopt SparseConv [15] as the backbone network and VoteNet as the 3D detector, and follow its training settings with the limited number of scene reconstructions [19].

3D Semantic Segmentation. For 3D segmentation, we strictly follow CSC [19] in the limited reconstruction setting. Specifically, we pre-train on ScanNet and fine-tune the pre-trained model on indoor S3DIS, ScanNet and outdoor SemanticKITTI (SK) [2]. 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 [19] and PCon [42]. In addition, we test the ScanNet pre-trained model on 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.

4.2. Data-efficient Transfer Learning

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 [13] as shown in Table 1 and Fig. 4. 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. 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.

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 2 shows, FAC outperforms ProCo [44] clearly and consistently, demonstrating its potential in reducing data annotations. In addition, we conducted experiments for indoor detection on ScanNet. As Table 3 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 holistic object-level representations.

3D Semantic 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 [19]. This shows that FAC can learn superb representations that capture similar features while suppressing distinct ones.

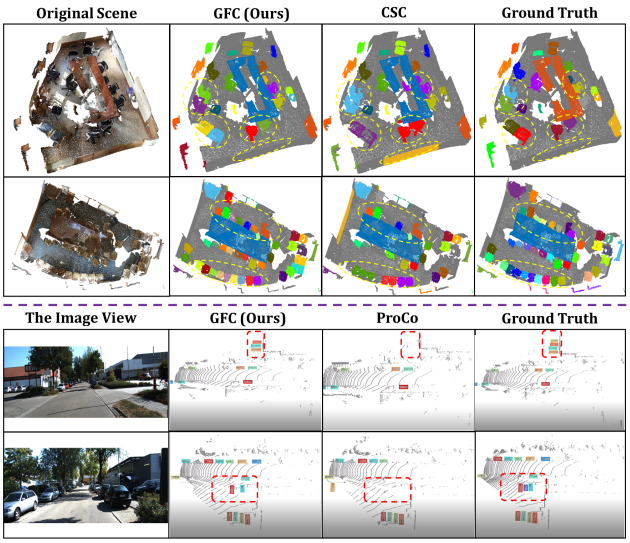

Figure 4. Visualizations of indoor 3D segmentation over ScanNet compared with CSC [19] as fine-tuned with 10% labeled training data and outdoor object detection over KITTI [13] with 20% labeled training data compared with ProCo [44]. Different segmented instances and detected objects are highlighted with different colors. Differences in prediction are highlighted with yellow ellipses and red boxes.

We also conduct quantitative experiments as shown in the Table 4, where we adopt limited labels (*e.g.*, {1%, 5%, 10%, 20%}) in training. We can 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 [19] significantly when only 1% labels are used, demonstrating its great capacity in learning informative representations with limited labels. Note FAC achieves more improvements while working with less labeled data. For semantic segmentation over dataset SK [2], FAC achieves consistent improvement and similar trends with decreasing labeled data.

4.3. Ablation Study and Analysis

We perform extensive ablation studies over key designs in FAC. Specifically, we examine the effectiveness of the proposed regional sampling, feature matching network, and the two losses. 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 [36] as the 3D detector.

Regional Sampling and Feature Matching. Regional

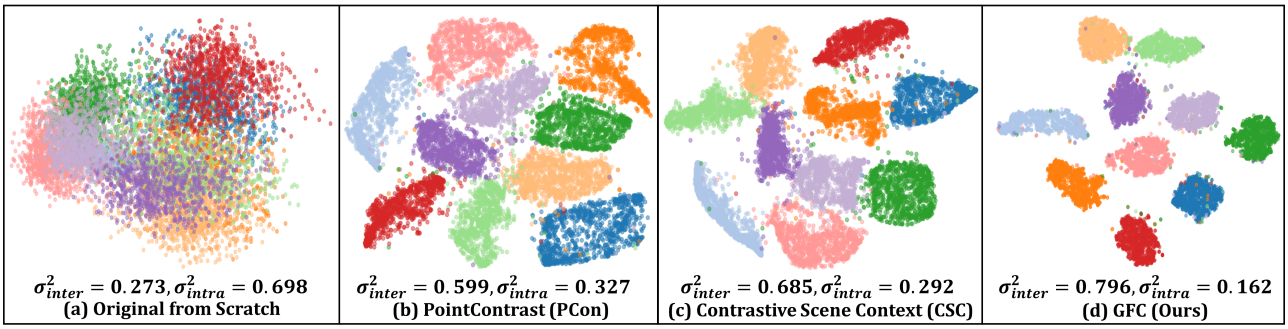

Figure 5. t-SNE [38] visualization of feature embeddings for SemanticKITTI semantic segmentation fine-tuned with 5% percent label (ScanNet Pre-trained). Ten classes with the least number of points are shown, where σ^2_{intra}, σ^2_{inter} denote intra- and inter-class variance. FAC learns a more compact feature space with the smallest intra-class variance and largest inter-class variance as compared with state-of-the-art methods PCon [42], CSC [19].

Case	*Sampling*	$\mathcal{L}_{Geo}$	$\mathcal{L}_{Fea}$	Sem. mIoU% (Sc)	Sem. mIoU% (SK)	Det. AP@50% (Sc)	Det. mAP% (K)
Baseline				47.06	33.58	0.39	63.51
Case 1	✓			48.61	36.21	8.39	64.78
Case 2		✓		50.62	39.87	16.46	66.36
Case 3			✓	50.93	40.32	17.98	66.29
Case 4	✓		✓	51.34	40.56	18.57	66.72
Case 5		✓	✓	51.48	40.68	18.68	67.28
Case 6	✓	✓		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 5. Ablation study of diverse modules of FAC on ScanNet (Sc) and SemanticKITTI (SK) [2], & KITTI (K) [13].

sampling samples points in the foreground regions as anchors. Table 5 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 [24] (i.e., **H-FAC**) as shown in Table 5. 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.

FAC Losses. FAC employs a foreground-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 5 cases 4 and 6, including either loss clearly outperforms the *Baseline* as well as the state-of-the-art CSC [19] in segmentation and ProCo [44] 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 1 and Table 3. At last, the **full FAC** in Table 5 including both losses learn better representations with the best performance in various downstream tasks.

Feature Visualization with t-SNE [38]. We employ t-SNE to visualize the feature representations that are learnt for SemanticKITTI [2] semantic segmentation task as illustrated in Fig. 5. Compared with other contrastive learning approaches such as PCon [42] and CSC [19], FAC learns a more compact and discriminative feature space that can clearly separate features of different semantic classes. As Fig. 5 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. Conclusion

We propose a *foreground-aware* contrast framework (FAC) for 3D unsupervised pre-training. 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 SCN to find the well-correlated feature contrast pairs within and across views for both the foreground and background segments. Extensive experiments demonstrate the superiority of FAC in both knowledge transfer and data efficiency.

Acknowledgement

This study is supported in part by Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), and in part by the National Natural Science Foundation of China [grant no. U2033210], and in part by the Zhejiang Provincial Natural Science Foundation [grant no. LDT23F02024F02].

References

[1] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14605–14615, 2022. 1

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 6, 7, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[7] Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8382–8391, 2021. 1

[8] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *European Conference on Computer Vision*. Springer, 2022. 2

[9] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *European Conference on Computer Vision*, pages 681–699. Springer, 2022. 2

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[11] Emeç Erçelik, Ekim Yurtsever, Mingyu Liu, Zhijie Yang, Hanzhen Zhang, Pınar Topçam, Maximilian Listl, Yılmaz Kaan Çaylı, and Alois Knoll. 3d object detection with a self-supervised lidar scene flow backbone. In *European Conference on Computer Vision*, 2022. 2

[12] Rinon Gal, Amit Bermano, Hao Zhang, and Daniel Cohen-Or. Mrgan: Multi-rooted 3d shape representation learning with unsupervised part disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2039–2048, 2021. 1

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5, 6, 7, 8

[14] Akash Gokul, Konstantinos Kallidromitis, Shufan Li, Yusuke Kato, Kazuki Kozuka, Trevor Darrell, and Colorado J Reed. Refine and represent: Region-to-object representation learning. *arXiv preprint arXiv:2208.11821*, 2022. 1

[15] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9224–9232, 2018. 2, 7

[16] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2270–2287, 2014. 3

[17] Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8376–8384, 2019. 2

[18] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae: Unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 10441–10450. IEEE, 2019. 1

[19] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 1, 2, 4, 5, 6, 7, 8

[20] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021. 1

[21] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision*, pages 187–203, 2018. 1

[22] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 1, 2, 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 6

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2005. 8

[25] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 2, 6

[26] Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, and Luc Van Gool. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3293–3302, 2021. 1, 2, 6

[27] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *European Conference on Computer Vision*, 2022. 2, 3

[28] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*. Springer, 2022. 3

[29] Kangcheng Liu, Yuzhi Zhao, Qiang Nie, Zhi Gao, and Ben M Chen. Weakly supervised 3d scene segmentation with region-level boundary awareness and instance discrimination. In *European Conference on Computer Vision 2022 (ECCV 2022)*, pages 37–55. Springer, Cham, 2022. 1

[30] Minghua Liu, Yin Zhou, Charles R Qi, Boqing Gong, Hao Su, and Dragomir Anguelov. Less: Label-efficient semantic segmentation for lidar point clouds. In *European Conference on Computer Vision*, pages 70–89. Springer, 2022. 1

[31] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 754–763, 2021. 2

[32] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3164–3173, 2021. 2

[33] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*. Springer, 2022. 2, 3

[34] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3283–3292, 2021. 2

[35] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5376–5385, 2020. 2

[36] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 2, 6, 7

[37] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 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*, pages 1588–1597, 2019. 3

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[39] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 2

[40] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9782–9792, 2021. 2

[41] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. 2, 3

[42] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 1, 2, 3, 4, 6, 7, 8

[43] Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, and Tetsuya Ogata. Point cloud pre-training with natural 3d structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21283–21293, 2022. 2

[44] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 1, 4, 5, 6, 7, 8

[45] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. 3

[46] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. *arXiv preprint arXiv:2207.05027*, 2022. 1

[47] Yin Zhou and Oncel Tuzel. 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*, pages 4490–4499, 2018. 2, 6