ECO-3D: Equivariant Contrastive Learning for Pre-training on Perturbed 3D Point Cloud

Ruibin Wang, Xianghua Ying[*], Bowei Xing, Jinfa Yang
Key Laboratory of Machine Perception (MOE)
School of Intelligence Science and Technology, Peking University
{robin_wang, xhying, xingbowei, jinfayang}@pku.edu.cn

Abstract

In this work, we investigate contrastive learning on perturbed point clouds and find that the contrasting process may widen the domain gap caused by random perturbations, making the pre-trained network fail to generalize on testing data. To this end, we propose the Equivariant COntrastive framework which closes the domain gap before contrasting, further introduces the equivariance property, and enables pre-training networks under more perturbation types to obtain meaningful features. Specifically, to close the domain gap, a pre-trained VAE is adopted to convert perturbed point clouds into less perturbed point embedding of similar domains and separated perturbation embedding. The contrastive pairs can then be generated by mixing the point embedding with different perturbation embedding. Moreover, to pursue the equivariance property, a Vector Quantizer is adopted during VAE training, discretizing the perturbation embedding into one-hot tokens which indicate the perturbation labels. By correctly predicting the perturbation labels from the perturbed point cloud, the property of equivariance can be encouraged in the learned features. Experiments on synthesized and real-world perturbed datasets show that ECO-3D outperforms most existing pre-training strategies under various downstream tasks, achieving SOTA performance for lots of perturbations.

Introduction

Contrastive learning has been widely accepted as a self-supervised technique to learn representations in a label-free manner and achieved promising results in image domain (Wu et al. 2018; Van den Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020; He et al. 2020; Chen et al. 2020b). Recently, the 2D contrastive paradigm has been successfully applied in 3D domain (Xie et al. 2020; Sanghi 2020; Alliegro et al. 2021; Yu et al. 2021; Afham et al. 2022), learning meaningful representations of point clouds. These methods usually perform augmented transformations on each point cloud to generate the positive pairs and regard different point clouds as negative pairs. Despite learning useful features under the pre-aligned clean datasets such as ModelNet40 (Wu et al. 2015) and ShapeNet (Yi et al. 2016), we find that existing contrastive frameworks may fail to extract generalized features under point clouds having random perturbations such as noise, occlusion, etc. Since the perturbations normally exist in real-world 3D scans and require huge human effort to eliminate, enabling contrastive learning on perturbed data has significant practical meanings.

To investigate the contrastive learning on perturbed point clouds, we pre-train PointNet (Qi et al. 2017) following the most employed contrastive framework (Chen et al. 2020b) on both clean and perturbed ModelNet40 (Wu et al. 2015). The pre-trained encoders are then fine-tuned to perform the classification task and compared with the baseline which is trained in fully-supervised settings. The experimental results are illustrated in Fig. 1 (a). For the clean dataset, fine-tuning the pre-trained networks generally converges to better classification accuracy than the fully-supervised baseline. However, the results on the perturbed dataset are completely different. Fine-tuning performance is worse than fully-supervised training. The performance decrease is observed under more than one perturbation. It seems that the contrastive framework not only fails to facilitate useful feature extraction from the perturbed point clouds but learns harmful features, causing lower accuracy than the baseline.

To find out why contrastive learning fails on the perturbed point clouds, we visualize the features of training and testing data learned via the supervised and contrastive framework. As in Fig. 1(b), under the supervised framework, the features of testing and training data are more dispersedly distributed under the ② perturbed dataset than the ① clean dataset. This can be easily explained since the random perturbations lead to the testing data having perturbations unseen in the training data, thus increasing the training and testing domain gap. More surprisingly, we observe that such a domain gap is further enlarged under the contrastive framework③. We think it may be because the contrastive enhanced features can hardly generalize to testing data due to the domain gap thus leading to even worse domain gaps. Therefore, we try to enable contrastive learning on the perturbed point clouds by first closing the domain gap. A simple method is to manually delete the perturbations. However, this requires huge efforts and can hardly be generalized to different perturbations.

In this work, we design a novel Variational AutoEncoder (VAE) architecture to close the domain gap in the embedded space. The proposed VAE is composed of a hierarchical encoder to progressively map the perturbed data into low

[*] Xianghua Ying is the corresponding author.

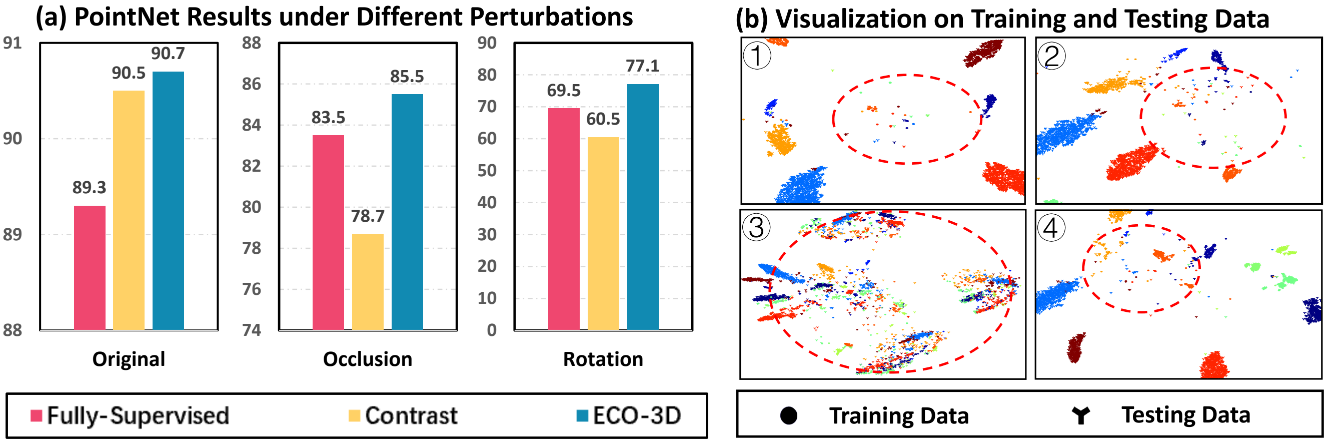

Figure 1: (a) Results of training PointNet (Qi et al. 2017) under one original (Wu et al. 2015) and two perturbed datasets (Taghanaki et al. 2020) with different training strategies. (b) t-SNE (Van der Maaten and Hinton 2008) visualizations on the encoded features of trained PointNet: ① Supervised training on the original dataset ② Supervised training on the noise dataset ③ Contrastive pre-training and fine-tuning on the noise dataset. ④ ECO-3D pre-training and fine-tuning on the noise dataset.

and high-level embedding and a decoder to let these embedding respectively contain information of less perturbed point clouds and correspondingly separated perturbations, via distinct reconstruction. Specifically, for the less perturbed point embedding, we expect its decoded point clouds to have fewer perturbations. In contrast, decoding on the combination of point and perturbation embedding precisely reconstructs the perturbed shapes. The distinct reconstruction strategy obtains less perturbed point embedding distributed in a similar domain. Then, existing contrastive frameworks can be employed based on these embeddings. To transform the embedding for generating contrastive pairs, we simply mix up the less perturbed point embedding with different perturbation embedding. Compared with hand-craft pre-processing, our method automatically narrows down the domain gap and is compatible with different perturbation types. The feature visualization of the less perturbed point embedding generated by VAE is shown in Fig. 1 (b) ④, where the training and testing data points are less dispersed, corresponding to narrowed domain gap.

Based on the new contrastive framework, we attempt to pursue the equivariance property to further improve the pre-training results under various perturbations. Essentially, our contrastive paradigm encourages the learned features to be invariant to the perturbations. Yet, the invariance property is a trivial instance of a broader class called equivariance. Formally, given a point cloud $X \in \mathbb{R}^{N\times 3}$ and a set of transformations $T_A : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^{N\times 3}$, a neural network $\phi : \mathbb{R}^{N\times 3} \rightarrow \mathcal{F}$ is called equivariant if for each A there exists an equivariant transformation $S_A : \mathcal{F} \rightarrow \mathcal{F}$ so that: $S_A[\phi(X)] = \phi(T_A[X])$. Therefore, the equivariance can be intuitively understood as the property that representations transform according to the way the inputs transform and invariance is a special case of equivariance when S_A is an identity mapping. Equivariant features naturally reflect the input transformations. Recently, (Dangovski et al. 2021) encourage the equivariance simply by predicting the input transformations from the learned features, which improves image pre-training. However, in point cloud pre-training, similar equivariant properties have less been investigated.

In sight of this, we generalize the equivariant framework established in (Dangovski et al. 2021) to our contrastive learning paradigm. To achieve that, a Vector Quantizer (Van Den Oord et al. 2017) is introduced in VAE, which discretizes the perturbation embedding into one-hot perturbation labels. By predicting the perturbation token labels from the learned representations, the equivariance properties are encouraged. Besides, a learnable weight is used to balance the invariance and the equivariance during pre-training.

Our final 3D pre-training framework named Equivariant COntrastive Learning (ECO-3D) achieves promising performance for various kinds of perturbations. We test ECO-3D on several point networks under both synthesized and real-world perturbations. Experiments show that ECO-3D achieves better pre-training performance under most perturbation types than existing pre-training strategies, obtaining SOTA performance for various downstream tasks. Ablations confirm the effectiveness of the designs in our framework. Our contributions can be summarized as:

- For the first time, we observe the degeneration of point cloud contrastive learning under perturbations and propose ECO-3D pre-training which successfully learns generalized features from perturbed point clouds and alleviates the degeneration under various perturbations.
- In ECO-3D, we invent a hierarchical VAE to close the perturbation domain gap based on which a novel contrastive framework is designed to overcome degeneration. A Vector Quantizer is further equipped to help pursue feature equivariance. We also propose a learnable weight to adaptively adjust the equivariance scale.
- Experiments suggest that ECO-3D outperforms most existing self-supervised pre-training frameworks on synthesized and real-world perturbed datasets, achieving SOTA performance under various downstream tasks.

Related Work

3D Variational AutoEncoders. Variational AutoEncoders (VAEs) (Kingma and Welling 2013; Higgins et al. 2017) are major deep generative models for learning the content distribution of various kinds of data including images (Bepler et al. 2019; Liu, Breuel, and Kautz 2017), texts (Hu et al. 2017; Yang et al. 2017) and sound (Tjandra et al. 2019; Eloff et al. 2019). Recently, VAE has shown remarkable performance in representing 3D point clouds. For example, (Achlioptas et al. 2018; Han et al. 2019) encodes the input into continuous latent variables for precise shape reconstruction and generation. (Eckart et al. 2021; Yu et al. 2021) models the discrete distribution of inputs to tokenize the point clouds into discrete geometric partitions, serving for their designed pretext tasks. These 3D VAEs mainly learn one distribution for the input point clouds.

Our work differs from these researches in that we try to simultaneously model two complementary distributions from perturbed point clouds, *i.e.* the less perturbed point distribution and the perturbation distribution. To achieve that, we develop a hierarchical VAE to extract the embedding of different levels and learn the two independent distributions via distinct reconstructions. Moreover, a Vector Quantizer (Van Den Oord et al. 2017; Razavi et al. 2019) is adopted for discretizing the perturbation embedding into tokens which serve as the perturbation labels for the equivariant learning.

Self-Supervised Pre-Training on 3D Point Clouds. Self-Supervised pre-training is a type of unsupervised learning where the supervision signals can be generated from the data itself (Jing and Tian 2020). In images, multiple pretext tasks have been designed for generating supervision signals including jigsaw puzzles (Noroozi and Favaro 2016), contrastive learning (Van den Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020; He et al. 2020; Chen et al. 2020b), masked encoding (Chen et al. 2020a; He et al. 2021), etc. Recently, self-supervised pre-training has been successfully applied in 3D point clouds via carefully-designed 3D pretext tasks, such as the orientation estimation (Poursaeed et al. 2020), 3D jigsaw (Sauder et al. 2019), occlusion completion (Wang et al. 2021), local-global reasoning (Rao, Lu, and Zhou 2020; Sharma and Kaul 2020; Thabet, Alwassel, and Ghanem 2020). Meanwhile, pretext tasks originally designed in images also be transferred into point clouds such as contrastive learning (Alliegro et al. 2021; Sanghi 2020; Afham et al. 2022) and masked modeling (Yu et al. 2021).

However, these pre-training methods are commonly performed under the ideal 3D dataset without perturbations, which can be less practical for the 3D data collected from realistic scenarios. For the first time, we investigate the pre-training under perturbed point clouds. Our framework builds on the most employed pretext task, contrastive learning, and further introduces VAE to close the domain gaps in perturbed point clouds. We also design a novel pretext task that encourages equivariance property in contrastive learning.

Invariance vs. Equivariance. The contrastive loss (Sohn 2016; Wu et al. 2018) essentially enables the learned features to be invariant to the augmented transformations. By doing so, the learned features are insensitive to the transformations and focus more on other useful features (Xiao et al. 2020). However, many recent studies (Dangovski et al. 2021; Metzger et al. 2020; Gidaris, Singh, and Komodakis 2018; Doersch, Gupta, and Efros 2015; Gidaris et al. 2019) in image domain have shown that making the learned features sensitive to some augmentations by predicting them generates better representations. In sight of this, (Dangovski et al. 2021) propose to pursue the sensitivity by extending the invariance property of existing contrastive learning to its broader class *i.e.*, equivariance. They encourage the non-trivial equivariance in the learned features by predicting the transformation categories from the augmented images.

Inspired by their success, we introduce the non-trivial equivariance in our contrastive learning framework on perturbed point clouds. Different from (Dangovski et al. 2021), our work more conveniently pursues the equivariance under various perturbations by predicting the perturbation labels which are automatically generated from the Vector Quantizer in VAE. Moreover, a learnable weight is introduced for balancing the invariance and equivariance properties.

Method

This section introduces detailed instructions on developing the ECO-3D framework, which is mainly composed of two parts. The first part (Sect.) pre-trains a VAE to extract less perturbed point embedding and perturbation embedding from perturbed point clouds. The perturbation embedding will be discretized via the Vector Quantizer into one-hot tokens, serving as the perturbation labels. Based on the extracted embedding and labels, in the second part (Sect. -), both the pretext tasks of contrastive and equivariant learning are designed for pre-training the target networks.

VAE for Extracting Embedding and Tokens

The designed VAE for extracting embedding and tokens from perturbed point clouds consists of three modules: a hierarchical encoder to progressively learn the less perturbed point embedding and the perturbation embedding, a vector quantizer to discretize the continuous embeddings into one-hot tokens, and a decoder to regularize the encoder via distinct reconstruction. Detailed structures are shown in Fig. 2.

Hierarchical Encoder. The encoder includes a low-level branch to learn the point embedding and a high-level branch to extract the perturbation embedding. Since the learned embedding will be further quantized into discrete tokens, we follow (Yu et al. 2021) to encode in a discrete manner. Concretely, given an N-point clouds $P \in R^{N\times 3}$, we firstly divide them into G discrete patches as $\boldsymbol{p} = \{p_i\}_{i=1}^{G}$. Then, the output embedding of the low and high-level encoders are $\underline{\boldsymbol{f}} = \{\underline{f_i}\}_{i=1}^{G}, \underline{f_i} \in R^{C_1}$ and $\overline{\boldsymbol{f}} = \{\overline{f_i}\}_{i=1}^{G}, \overline{f_i} \in R^{C_2}$.

Vector Quantization. The quantization module learns the distribution $\mathcal{Q}_q(\boldsymbol{z}|\boldsymbol{f})$ of discrete latent variables $\boldsymbol{z}$ from embedding $\boldsymbol{f}$. It is mainly composed of two parts: a probabilistic modeling network learning the latent distribution from embedding $\boldsymbol{f}$ and a quantizer discretizing the variables $\boldsymbol{z}$ sampled from the latent distribution. We adopt DGCNN (Wang et al. 2019) for probabilistic modeling and follow (Ramesh et al. 2021) to discrete the sampled latent variables via jointly optimizing a codebook $\boldsymbol{e}$. The quantizer

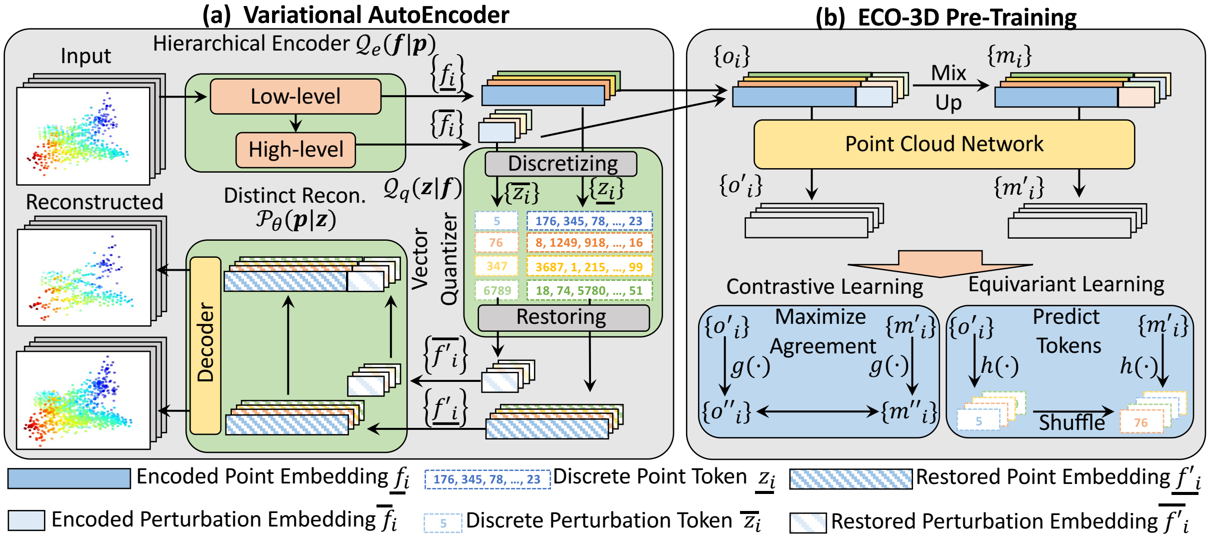

Figure 2: Illustration of ECO-3D framework. (a) The pre-trained VAE consists of a hierarchical encoder to extract the point and perturbation embedding $\{\underline{f_i}, \overline{f_i}\}$ from the perturbed inputs P, a vector quantizer converts the embedding into discrete tokens $\{\underline{z_i}, \overline{z_i}\}$, and a decoder respectively reconstruct the original and down-sampled point clouds from restored embedding $\{\underline{f'_i}, \overline{f'_i}\}$. (b) In the ECO-3D pre-training, the target network learns the point cloud representations $\{o'_i, m'_i\}$ respectively from original and mixed embedding combinations $\{o_i, m_i\}$. These representations are used for contrastive and equivariant pretext tasks.

discretizes both the low and high-level embedding in parallel, respectively generating the corresponding tokens $\underline{\boldsymbol{z}} = \{\underline{z_i}\}_{i=1}^G, \underline{z_i} \in R^{M_1}$ and $\overline{\boldsymbol{z}} = \{\overline{z_i}\}_{i=1}^G, \overline{z_i} \in R^{M_2}$.

Distinct Reconstruction. Finally, the decoder outputs the distribution $\mathcal{P}_\theta(\boldsymbol{p}|\boldsymbol{z})$ of input shapes from the latent variables $\boldsymbol{z}$ via two steps: restores the continuous embedding via $\boldsymbol{f}' = \boldsymbol{z} \cdot \boldsymbol{e}$ and then decodes the input distribution based on $\boldsymbol{f}'$ via reconstruction. We adopt FoldingNet (Yang et al. 2018) for the reconstruction process. Since the embeddings from different branches are expected to contain different information *i.e.* the low-level branch contains less perturbed point information and the high-level branch contains perturbation information, we propose the distinct reconstruction.

Specifically, we constraint the decoder to reconstruct the down-sampled coarse shapes of perturbed point clouds from the low-level embedding $\underline{\boldsymbol{f}'}$ and reconstruct the fine shape of the perturbed point cloud from the combination of the low-level and high-level embedding $\boldsymbol{f}' = \textbf{Concat}(\underline{\boldsymbol{f}'}, \overline{\boldsymbol{f}'})$. We use the down-sampled point clouds as the reconstruction target for the less perturbed point embedding, which avoids the effort of obtaining the clean point clouds. Note that, for the rotation perturbation that can not be eliminated via down-sampling, we align the rotated inputs with the reconstructed shapes to eliminate the rotation.

Optimization Objective. Following the objective derived in VAE (Kingma and Welling 2013), we maximize the evidence lower bound (ELB) of the log-likelihood as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{p}) = & E_{\boldsymbol{z}\sim\mathcal{Q}_\phi(\boldsymbol{z}|p)}\left[\log\mathcal{P}_\theta(\boldsymbol{p}|\boldsymbol{z})\right] \\ & - D_{KL}\left[\mathcal{Q}_\phi(\boldsymbol{z}|\boldsymbol{p})||\mathcal{P}_\theta(\boldsymbol{z})\right],\end{aligned} \tag{1}$$

where ϕ, θ are parameters respectively from the encoder $\mathcal{Q}$ and the decoder $\mathcal{P}$. Similar to (Yu et al. 2021), the first term refers to the reconstruction error between the input and the decoded point clouds. The second term regularizes the distribution of latent variables to be close to a pre-defined prior by the KL-Divergence. Differently, in our framework, the reconstruction error is modified to meet the distinct reconstruction. We employ ℓ_1 Chamfer Distance for calculating the reconstruction error between different point clouds. For the regularization of KL-Divergence, we follow (Ramesh et al. 2021) to initialize the prior $\mathcal{P}_\theta(\boldsymbol{z})$ via the uniform categorical distribution over the M codebook vectors.

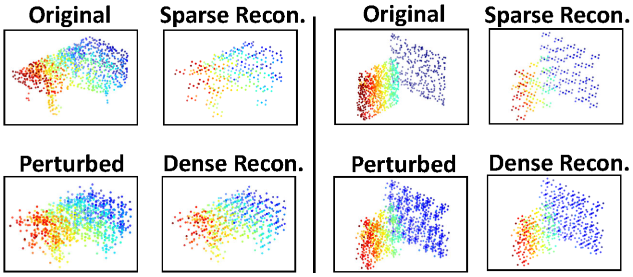

Figure 3: Visualizations of the distinct reconstructions on objects (piano, chair) via pre-trained VAE. The sparse reconstruction removes the perturbed points. Meanwhile, the dense reconstruction precisely restores the perturbed shapes.

In Fig. 3, we visualize the distinct reconstructions from our pre-trained VAE. As expected, the perturbations are significantly reduced in the reconstructed sparse shapes. The pre-training setting of VAE and more visualization results are provided in our Supp. Material.

Contrastive Learning by Mixing Embedding

The pre-trained VAE converts the randomly perturbed point clouds into less perturbed point embedding and correspondingly separated perturbation embedding. We propose to mix the information in the point and perturbation embedding to generate positive/negative pairs for contrast. We use batch notations to describe the mixing process.

Embedding Mix Up. Given perturbed N-point clouds $\boldsymbol{p} \in R^{B\times N\times 3}$ of batch size B , the pre-trained encoder converts them into the point embedding $\{\underline{f_i}\}_{i=1}^{B}, \underline{f_i} \in R^{GC_1}$ and perturbation embedding $\{\overline{f_i}\}_{i=1}^{B}, \overline{f_i} \in R^{GC_2}$. The contrastive pairs can be generated by mixing up the original combination of point and perturbation embedding. Formally, the original combinations $\{o_i\}_{i=1}^{B}$ and the mixed combinations $\{m_i\}_{i=1}^{B}$ are generated as:

$$o_i = \text{Concat}(\overline{f_i}, \underline{f_i}); \qquad m_i = \text{Concat}(\text{Shuffle}[\ \overline{\boldsymbol{f}}\ ]_i, \underline{f_i}). \tag{2}$$

Both the original and mixed point cloud embedding combinations will be input into the point cloud network learning the original and augmented representations $\{o'_i\}_{i=1}^{B}$ and $\{m'_i\}_{i=1}^{B}$ for the contrastive pre-training.

Contrastive Loss. NT-Xent (Chen et al. 2020b) is used to contrast learned representations at an instance level:

$$\mathcal{L}_{con} = -\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(g(o'_i), g(m'_i))/\tau)}{\sum_{k=1}^{B} \mathbf{1}_{[k\neq i]} \exp(\text{sim}(g(o'_i), g(m'_k))/\tau)}, \tag{3}$$

where $\mathbf{1}_{[k\neq i]}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter. $g(\cdot)$ refers to the non-linear projector (Chen et al. 2020b) used to improve the representation quality of the layer before it. Eq. (3) encourages the learned representations from the same point embedding to be invariant under different perturbation embedding.

Equivariant Learning by Predicting Tokens

We follow (Dangovski et al. 2021) to encourage the equivariance property by predicting the perturbation categories from the learned representations. Instead of manually labeling the perturbation, our quantizer proposed in Sect. conveniently generates the perturbation labels $\overline{\boldsymbol{z}}$. We use the notation in Sect. to describe the equivariant learning task.

Predicting Tokens. Given the representations of i-th point cloud m'_i and o'_i which contain original and shuffled perturbation embedding, we encourage the equivariance property by correctly predicting the token of the perturbation embedding from the representations. Concretely, a non-linear network $h(\cdot)$ is adopted to transform the representation into categorical logits. The equivariant loss can then be denoted by the CrossEntropy (CE) between the predicted logits and ground truth perturbation tokens $\overline{z_i}$ as:

$$\mathcal{L}_{equ} = \sum_{i=1}^{B} \text{CE}(h(o'_i), \overline{z_i}) + \sum_{i=1}^{B} \text{CE}(h(m'_i), \text{Shuffle}[\ \overline{\boldsymbol{z}}\ ]_i). \tag{4}$$

Note that, the token $\overline{\boldsymbol{z}}$ in the second term is shuffled as the $\overline{\boldsymbol{f}}$ shuffled in Eq. (2) to match the order.

Learnable Weight. For each perturbation, (Dangovski et al. 2021) conducts a pre-test to investigate the influence of encouraging invariance or equivariance on the pre-trained features. Based on the pre-test results, a fixed weight is used to control the influence during pre-training. Unlike (Dangovski et al. 2021), we propose using a learnable weight ρ to automatically balance the invariance and equivariance during the pre-training. The pre-trained loss function can then be defined as $\rho\mathcal{L}_{con} + (1-\rho)\mathcal{L}_{equ}$. The learnable weight reduces the effort of pre-tests. The learned weight directly implies the ideal trade-off between Equivariance and Invariance properties. The effectiveness of the learnable weight can be demonstrated in Tab. 4, where the learned ρ for different perturbations are also illustrated.

Experiments

Experiment Setup

Datasets. We experiment on three 3D point cloud datasets with the synthesized or real-world perturbations, including RobustPointSet(Taghanaki et al. 2020), ScanObejectNN (Uy et al. 2019), and ShapNetC, to verify the ECO-3D framework. Specifically, RobustPointSet (Taghanaki et al. 2020) is generated by performing six synthesized perturbations on the original ModelNet40 (Wu et al. 2015). We select five perturbations (Noise, Rotation, Occlusion, Translate, and Missing Parts) for experiments. ScanObejectNN (Uy et al. 2019) contains 3D scans with five real-world perturbations. We adopt three perturbations with increasing difficulty (OBJ-BG, PB-T25-R, PB-T50-RS). The detailed meaning of each perturbation is provided in our Supp. Material. In addition to the classification task, we generate ShapeNetPart-C based on ShapeNetPart (Yi et al. 2016) for testing our method on the part segmentation task. Three synthesized corruptions (Sun et al. 2022) (Shear, Cutout, Background Noise) are adopted. We adopt the overall accuracy (OA) and mean intersection-over-union (mIoU) to evaluate the classification and segmentation, respectively.

Pre-Trained Architectures. We test ECO-3D under two representative networks, PointNet (Qi et al. 2017) and DGCNN (Wang et al. 2019).

Training Overview. ECO-3D consists of three main training steps. First, the proposed VAE is trained under the perturbed dataset to produce tokens. Then, the selected networks are pre-trained via VAE tokens under the same perturbed dataset. Finally, the pre-trained networks are leveraged under various downstream tasks. Detailed implementations of these steps can be found in our Supp. Material.

Compared Approaches. We respectively compare ECO-3D with the fully-supervised baseline, 3D contrastive learning (Sanghi 2020) and other 3D self-supervised pre-training methods including 3D Jigsaw (Sauder et al. 2019) which predicts the original order of shuffled voxels and OcCo (Wang et al. 2021) which learns to complete occlusions.

Fine-Tune Performance

To compare the learned representations of different pre-training approaches, we fine-tune the pre-trained networks

RobustPointSet	PointNet (Qi et al. 2017)					DGCNN (Wang et al. 2019)				
	Baseline	Contrast	Jigsaw	OcCo	Ours	Baseline	Contrast	Jigsaw	OcCo	Ours
Noise	85.69	82.12	83.56	86.57	**87.01**	88.01	87.15	88.16	88.24	**88.98**
Rotation	69.52	60.54	63.24	74.12	**77.05**	78.11	76.11	81.35	82.01	**83.59**
Occlusion	83.47	78.69	80.15	**85.45**	85.10	87.21	86.04	87.35	**89.87**	89.08
Translation	88.74	86.64	87.01	89.25	**89.69**	90.50	89.89	90.01	90.84	**91.06**
MissingPart	85.12	82.01	84.39	86.10	**86.35**	88.33	87.95	88.56	89.01	**89.81**
ScanObjectNN	PointNet (Qi et al. 2017)					DGCNN (Wang et al. 2019)				
	Baseline	Contrast	Jigsaw	OcCo	Ours	Baseline	Contrast	Jigsaw	OcCo	Ours
OBJ-BG	66.54	63.15	66.58	67.98	**69.94**	82.63	81.82	82.58	84.36	**85.21**
PB-T25-R	59.31	57.45	59.35	61.20	**62.75**	79.16	78.05	80.35	80.98	**82.30**
PB-T50-RS	69.28	68.57	69.38	70.35	**72.35**	80.89	78.94	81.00	82.56	**85.94**

Table 1: Fine-tuning results (OA, %) of two architectures under the RobustPointSet and the real-world ScanObjectNN.

ShapeNetPart-C	PointNet (Qi et al. 2017)					DGCNN (Wang et al. 2019)				
	Baseline	Contrast	Jigsaw	OcCo	Ours	Baseline	Contrast	Jigsaw	OcCo	Ours
Shear	65.08	64.25	68.15	71.32	**74.44**	78.70	77.25	81.36	81.68	**83.34**
Cutout	64.12	62.99	65.35	66.14	**68.67**	72.69	71.07	77.19	78.10	**80.64**
BG Noise	60.21	59.08	62.24	63.18	**64.65**	77.18	76.24	81.01	81.15	**82.67**

Table 2: Fine-tuning results (mIoU, %) of two architectures under the perturbed ShapeNetPart-C.

under the perturbed datasets same as pre-training to perform the classification and segmentation task.

Shape Classification. Two perturbed point cloud datasets RobustPointSet and ScanObjectNN are used. The comparison results of two architectures under different training approaches are illustrated in Tab. 1. ECO-3D successfully improves the performance of existing contrastive learning on perturbed point clouds and outperforms the fully-supervised baseline for both datasets. Our method is slightly inferior to OcCo under occlusion perturbation since OcCo enables the network to see more occlusion types.

Part Segmentation. The part segmentation tasks are conducted under ShapeNetPart-C. The comparison results are illustrated in Tab. 2. Similar to the classification tasks, ECO-3D achieves the best mIoU for all three perturbations.

Transfer Performance

We show the ECO-3D learns more generalized features during pre-training by conducting the transfer learning task. Specifically, we pre-trained the networks under the RobustPointSet with three synthesized perturbations (Rotation, Noise, Missing Part) and fine-tune the pre-trained networks on the ScanObjectNN. The fine-tuned results are compared with other training approaches in Tab. 3. As can be seen, ECO-3D achieves SOTA transfer learning performance on all three real-world perturbations. Moreover, it highly improves the fully-supervised baseline.

We further verify the transferability of self-supervised features by pre-training networks under noise point clouds and then fine-tuning them under clean data. Fine-tuning the noise pre-trained PointNet and DGCNN get 91.0% and 93.2% under the clean ModelNet, surpassing their supervised baselines 89.2% and 92.2%. It suggests that ECO-3D extracts clean shape features from perturbations.

Ablation Study

We conduct ablation studies to verify designs in ECO-3D. Experiments are conducted under the PointNet architecture.

VAE Designs. We ablate the designs used in the encoder, quantizer and decoder. The experiments are mainly conducted under rotation perturbations. We compare different designs with respect to the pre-training loss of VAE (loss1), the point network (loss2) and the final performance (acc.) in Tab. 4. Our design achieves better performance.

ECO-3D Pretext. To show that it is the pretext tasks instead of the VAE embedding that help to improve the results, we gradually remove the pretext task during pre-training. For the None version, the network will be directly fine-tuned with embedding input. The results are compared in Tab. 5. Each pretext task improves the final performance.

Balanced Weight. In Tab. 6, we report the results under balanced weight and show the learned weight after pre-training. Compared with results in Tab. 5, the balanced weight benefits the pre-training results.

Computation Cost. It takes about 1.5h to train VAE for preparing the tokens. In contrast, Jigsaw and OcCo use about 1.3h and 2h to prepare the jigsaw label and occlusion data. The GPU mem. and total time of pre-training are: Jigsaw (1420M,3.33h), OcCo (1344M,2.83h), ECO-3D (1899M,1.78h). ECO-3D reaches the pre-training convergence faster. The results are recorded using a single 2080Ti.

Loss Visualization

To verify that the network effectively optimizes the pretext objective, we visualize the evolution of the two proposed losses during the pre-training. As illustrated in Fig. 4, both the $\mathcal{L}_{con}$ and $\mathcal{L}_{equ}$ are effectively optimized under different perturbations. Moreover, DGCNN converges to a lower

ScanObjectNN	PointNet (Qi et al. 2017)					DGCNN (Wang et al. 2019)				
	Baseline	Contrast	Jigsaw	OcCo	Ours	Baseline	Contrast	Jigsaw	OcCo	Ours
OBJ-BG	66.54	62.08	65.14	67.10	**68.41**	82.63	80.45	82.46	82.96	**84.22**
PB-T25-R	59.31	57.10	59.03	60.21	**61.28**	79.16	77.58	80.10	80.53	**81.03**
PB-T50-RS	69.28	67.12	69.01	69.15	**71.10**	80.89	77.12	79.92	80.81	**84.18**

Table 3: Transfer results (OA, %) of two architectures pre-trained under RobustPointSet and fine-tune under ScanObjectNN.

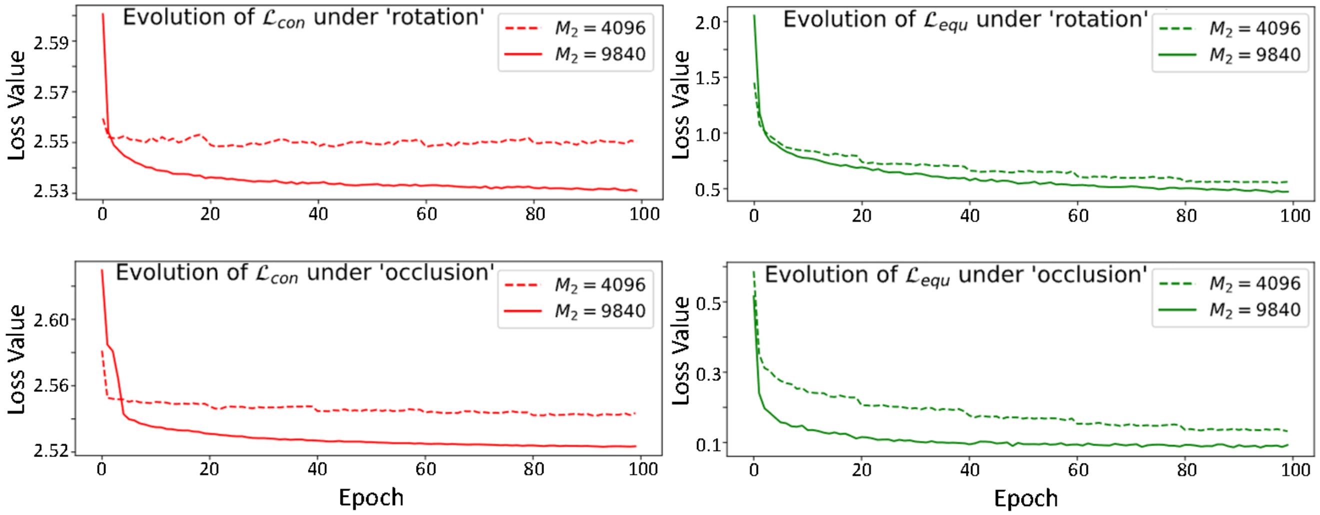

Figure 4: Evolution of the contrastive and equivariant losses for pre-training two networks under different perturbations.

Robust	Loss1		Loss2		OA
PointSet	$\mathcal{L}_{rec}$	$\mathcal{L}_{kl}$	$\mathcal{L}_{con}$	$\mathcal{L}_{equ}$	(%)
Encoder Architecture					
w/o Hierar.	0.088	0.857	3.787	1.587	74.15
w Hierar.	**0.040**	**0.009**	**2.550**	**0.562**	**77.05**
Quantizer Number					
One	**0.039**	0.215	**2.535**	0.687	76.95
Two	0.040	**0.009**	2.550	**0.562**	**77.05**
Reconstruction Strategy					
Original	**0.021**	**0.001**	3.216	1.762	74.19
Distinct	0.040	0.009	**2.550**	**0.562**	**77.05**

Table 4: Verification on the designs of VAE.

RobustPointSet	ECO-3D Pretext			
	None	$\mathcal{L}_{con}$	$\mathcal{L}_{equ}$	Both
Noise	84.38	86.05	86.18	86.96
Rotation	69.75	76.14	71.05	77.01
Occlusion	83.01	83.18	84.97	85.01

Table 5: Verification on the ECO-3D pretext tasks.

RobustPointSet	Weighted: $\rho\mathcal{L}_{con} + (1-\rho)\mathcal{L}_{equ}$
Noise	**87.01** $(\rho = 0.415)$
Rotation	**77.05** $(\rho = 0.895)$
Occlusion	**85.10** $(\rho = 0.244)$

Table 6: Verification on the balanced weight.

loss value compared with PointNet. This may suggest that the ECO-3D pretext tasks can be better optimized under more advanced architectures, obtaining discriminative pre-training features to improve downstream tasks.

Conclusion

We investigate contrastive learning under perturbed point clouds and find the domain gap caused by perturbations will result in the degraded performance of existing contrastive frameworks. To this end, we propose ECO-3D framework which closes the domain gap via a pre-trained VAE before contrasting and introduces the equivariance property during contrasting. Extensive experiments show that ECO-3D significantly outperforms existing self-supervised pre-training frameworks for various downstream tasks. Specifically, ECO-3D achieves SOTA performance on five synthesized and three real-world perturbation types under classification. For segmentation, ECO-3D reaches the best performance on three synthesized perturbations. More importantly, the pre-trained features have a better generalization for transfer learning between synthesized and real-world perturbations. In addition to outperforming existing pre-training approaches, our method also significantly beats the fully-supervised baseline, which provides a new technique route for improving model performance under perturbations.

Acknowledgements

This work was supported in part by the National Key R&D Program of China Grant No. 2020YFB1708002 and NNSFC Grant No. 61971008.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *ICML*, 40–49. PMLR.

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. *arXiv preprint arXiv:2203.00680*.

Alliegro, A.; Valsesia, D.; Fracastoro, G.; Magli, E.; and Tommasi, T. 2021. Denoise and contrast for category agnostic shape completion. In *CVPR*, 4629–4638.

Bepler, T.; Zhong, E.; Kelley, K.; Brignole, E.; and Berger, B. 2019. Explicitly disentangling image content from translation and rotation with spatial-VAE. *NeurIPS*, volume 32.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020a. Generative pretraining from pixels. In *ICML*, 1691–1703. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Dangovski, R.; Jing, L.; Loh, C.; Han, S.; Srivastava, A.; Cheung, B.; Agrawal, P.; and Soljačić, M. 2021. Equivariant Contrastive Learning. *arXiv preprint arXiv:2111.00899*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*, 1422–1430.

Eckart, B.; Yuan, W.; Liu, C.; and Kautz, J. 2021. Self-supervised learning on 3d point clouds by learning discrete generative models. In *CVPR*, 8248–8257.

Eloff, R.; Nortje, A.; van Niekerk, B.; Govender, A.; Nortje, L.; Pretorius, A.; Van Biljon, E.; van der Westhuizen, E.; van Staden, L.; and Kamper, H. 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. *arXiv preprint arXiv:1904.07556*.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2019. Boosting few-shot visual learning with self-supervision. In *ICCV*, 8059–8068.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Han, Z.; Wang, X.; Liu, Y.-S.; and Zwicker, M. 2019. 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 *ICCV*, 10441–10450.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *ICML*, 1587–1596. PMLR.

Jing, L.; and Tian, Y. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 4037–4058.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *NeurIPS*, volume 30.

Metzger, S.; Srinivas, A.; Darrell, T.; and Keutzer, K. 2020. Evaluating Self-Supervised Pretraining Without Using Labels. *arXiv preprint arXiv:2009.07724*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 69–84.

Poursaeed, O.; Jiang, T.; Qiao, H.; Xu, N.; and Kim, V. G. 2020. Self-supervised learning of point clouds via orientation estimation. In *3DV*, 1018–1028.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*, 8821–8831. PMLR.

Rao, Y.; Lu, J.; and Zhou, J. 2020. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *CVPR*, 5376–5385.

Razavi, A.; Van den Oord, A.; Vinyals, O.; et al. 2019. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, volume 32.

Sanghi, A. 2020. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *ECCV*, 626–642.

Sauder, J.; Sievers, B.; Sievers, B.; and Sievers, B. 2019. Self-supervised deep learning on point clouds by reconstructing space. *NeurIPS*, volume 32.

Sharma, C.; and Kaul, M. 2020. Self-supervised few-shot learning on point clouds. *NeurIPS*, volume 33.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *NeurIPS*, volume 29.

Sun, J.; Zhang, Q.; Kailkhura, B.; Yu, Z.; Xiao, C.; and Mao, Z. M. 2022. Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions. *arXiv preprint arXiv:2201.12296*.

Taghanaki, S. A.; Luo, J.; Zhang, R.; Wang, Y.; Jayaraman, P. K.; and Jatavallabhula, K. M. 2020. Robustpointset: A dataset for benchmarking robustness of point cloud classifiers. *arXiv preprint arXiv:2011.11572*.

Thabet, A.; Alwassel, H.; and Ghanem, B. 2020. Self-supervised learning of local features in 3d point clouds. In *CVPRW*, 938–939.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *ECCV*, 776–794.

Tjandra, A.; Sisman, B.; Zhang, M.; Sakti, S.; Li, H.; and Nakamura, S. 2019. VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019. *arXiv preprint arXiv:1905.11449*.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 1588–1597.

Van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, arXiv–1807.

Van Den Oord, A.; Vinyals, O.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *NeurIPS*, volume 30.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Liu, Q.; Yue, X.; Lasenby, J.; and Kusner, M. J. 2021. Unsupervised point cloud pre-training via occlusion completion. In *ICCV*, 9782–9792.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics*, 38(5): 1–12.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 1912–1920.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 3733–3742.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2020. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 574–591.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 206–215.

Yang, Z.; Hu, Z.; Salakhutdinov, R.; and Berg-Kirkpatrick, T. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *ICML*, 3881–3890. PMLR.

Yi, L.; Kim, V. G.; Ceylan, D.; Shen, I.-C.; Yan, M.; Su, H.; Lu, C.; Huang, Q.; Sheffer, A.; and Guibas, L. 2016. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics*, 35(6): 1–12.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2021. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling. *arXiv preprint arXiv:2111.14819*.