This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Bringing Masked Autoencoders Explicit Contrastive Properties for Point Cloud Self-Supervised Learning

Bin Ren[1,2,3], Guofeng Mei[4], Danda Pani Paudel[3], Weijie Wang[2,4], Yawei Li[5], Mengyuan Liu[6⋆], Rita Cucchiara[7], Luc Van Gool[3,5], and Nicu Sebe[2]

[1] University of Pisa, Italy
[2] University of Trento, Italy
[3] INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria
[4] Fondazione Bruno Kessler, Italy
[5] ETH Zürich, Switzerland
[6] Peking University, Shenzhen Graduate School, China
[7] University of Modena and Reggio Emilia, Italy
bin.ren@unitn.it

Abstract. Contrastive learning (CL) for Vision Transformers (ViTs) in image domains has achieved performance comparable to CL for traditional convolutional backbones. However, in 3D point cloud pretraining with ViTs, masked autoencoder (MAE) modeling remains dominant. This raises the question: Can we take the best of both worlds? To answer this question, we first empirically validate that integrating MAE-based point cloud pre-training with the standard contrastive learning paradigm, even with meticulous design, can lead to a decrease in performance. To address this limitation, we reintroduce CL into the MAE-based point cloud pre-training paradigm by leveraging the inherent contrastive properties of MAE. Specifically, rather than relying on extensive data augmentation as commonly used in the image domain, we randomly mask the input tokens twice to generate contrastive input pairs. Subsequently, a weight-sharing encoder and two identically structured decoders are utilized to perform masked token reconstruction. Additionally, we propose that for an input token masked by both masks simultaneously, the reconstructed features should be as similar as possible. This naturally establishes an explicit contrastive constraint within the generative MAE-based pre-training paradigm, resulting in our proposed method, Point-CMAE. Consequently, Point-CMAE effectively enhances the representation quality and transfer performance compared to its MAE counterpart. Experimental evaluations across various downstream applications, including classification, part segmentation, and few-shot learning, demonstrate the efficacy of our framework in surpassing state-of-the-art techniques under standard ViTs and single-modal settings. The source code and trained models are available at https://github.com/Amazingren/Point-CMAE.

Keywords: Point Cloud Pre-training · Self-Supervised Learning · Vision Transformer

⋆ Corresponding author. Email:nkliuyifang@gmail.com

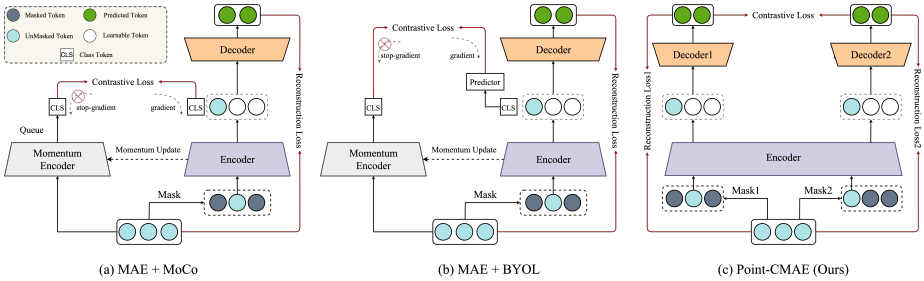

Fig. 1: Illustration of: (a) MAE with MoCo-style contrastive learning [22], requires an additional queue to store negative samples throughout the pre-training process. (b) MAE with BYOL-style contrastive learning [17], requires an asymmetric structure with a fully connected predictor layer to exclude negative samples. (c) The proposed Point-CMAE employs two identically structured decoders, updated differently, to introduce explicit contrastive properties within the MAE-based generative self-supervised pre-training.

1 Introduction

Understanding 3D scenes is critical for diverse applications, from autonomous vehicles navigating urban environments to robotic manipulation tasks [2, 5]. Point clouds, representing objects and surroundings precisely, offer an advantage due to their easy acquisition and accurate geometry capture [38]. However, annotating 3D point cloud data is even more resource-intensive compared to image data, as each point requires labeling. Self-supervised learning (SSL) has emerged as a key solution, the effectiveness of which has already been validated in natural language processing [4, 11, 49], computer vision [8, 9, 22, 74], and multimodal learning [27, 48, 72], enabling informative feature representations from unlabeled input data. As a result, it is widely utilized in pre-training for 3D point clouds, demonstrating comparative or even better performance compared to its supervised counterpart in downstream tasks such as classification, segmentation, and detection [30, 31, 36, 45, 52, 69].

Specifically, contrastive (single/cross-modal) and generative-based (reconstruct/predict) pre-training strategies are widely employed. CL is renowned for capturing global information to improve the model's discriminative ability, and this has been demonstrated to be effective in various domains such as 2D images and point cloud pre-training methods based on convolutional neural networks (CNNs) [40]. However, deploying single-modal CL to point clouds with ViT often results in inferior performance compared to MAE-based pre-training. On the other hand, using MAE with a Chamfer constraint for point cloud reconstruction can lead to sub-optimal solutions [30], as ground truth points are sampled from the original point clouds, causing inconsistencies when forcing predicted points to match them. In addition, the primary concept behind MAE typically emphasizes local relationships. These challenges collectively hinder the overall performance of ViTs-based point cloud pre-training. Consequently, a natural question arises: *Can we leverage the strengths of both paradigms*?

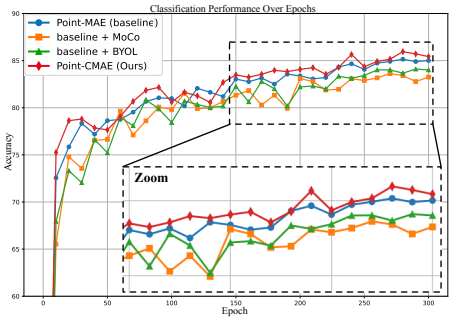

Fig. 2: The classification comparison of different contrastive learning pipelines, integrated with the baseline method Point-MAE [42], is conducted on the ScanObjectNN [55] datasets.

In other words, can we harness the power of contrastive learning to learn consistent and abstract feature representations, while also utilizing MAE to enhance the model's ability to capture local data structures? To explore this question, we propose to enrich the MAE paradigm with contrastive learning ability, where the masked and the unmasked point cloud tokens are treated as two augmentations. Based on this setting, we deploy MoCo [22], and BYOL [17] style CL with Point-MAE [42] to point cloud pre-training with ViTs structure. The illustration of these two kinds of MAE-based CL is shown in Fig. 1 (a) and Fig. 1 (b), respectively. Our experimental results, as shown in Fig. 2, confirm that this approach frequently results in the occurrence of a performance drop. This is primarily because making CL effective for ViTs-based point cloud pre-training necessitates a well-designed framework and a tailored approach to conducting separate data augmentation for point cloud data. These requirements significantly increase the complexity of the entire problem, highlighting the non-trivial nature of even bringing contrastive learning to an MAE-based paradigm.

To overcome the above limitations, we propose Point-CMAE, a simple yet effective method that explicitly integrates contrastive properties into the MAE-based point cloud pre-training paradigm using a ViTs architecture. The main idea of the proposed Point-CMAE is simplified in Fig. 1 (c). Specifically, for a given point cloud token, we mask it twice randomly to construct the contrastive input pairs instead of applying heavy data augmentation. We then perform MAE for each of the masked input tokens using a weight-sharing encoder and two identically structured decoders. In this design, each decoder independently recovers the masked point cloud input token, which naturally forces the encoder to learn more common and representative features that satisfy the requirements of both decoders. Though this largely improves the semantic understanding of the encoder, the reconstruction constraint, the Chamfer distance loss, is usually used to minimize the distance between the predicted masked points and the ground truth points regarding the fact the ground truth points are also just one sampling of the original point cloud, which naturally leads to a sub-optimal problem. Since there are two different masked input tokens, there is an almost certain probability that some tokens will be simultaneously double-masked. As a remedy, we propose that for a token that is masked in both instances, the recovered feature should be as close as possible in both cases. This naturally introduces an explicit constraint at the feature level, ensuring that the encoder outputs more informative features for downstream tasks. As a result, Fig. 2 shows that the proposed Point-CMAE

can perform a decent convergence with obvious improvement at the beginning of the fine-tuning compared to the contrastive counterpart.

In summary, our main contributions are recapped as follows:

1. We experimentally validate that integrating contrastive learning into the MAE paradigm for point cloud pre-training with ViTs architecture is non-trivial and often leads to a severe performance drop.
2. To address the aforementioned issue, we propose Point-CMAE, a method that enhances the MAE paradigm by integrating the advantages of contrastive learning through a straightforward yet effective design. This significantly enhances the informativeness and representativeness of the encoder.
3. Extensive experimental results on various point cloud downstream tasks such as object classification, part segmentation, and few-shot classification demonstrate that our proposed Point-CMAE achieves new state-of-the-art performance under the standard ViTs setting for single-modal data.

2 Related Work: Self-supervised Learning for Point Cloud

Self-supervised learning (SSL), as one type of unsupervised learning where the supervision signals can be acquired from the data itself, has attracted more and more attention to computer vision [17, 21]. As a result, a lot of methods have been proposed to advance this technique from different perspectives [3, 8, 17, 21, 41, 51, 54, 70]. Especially, recent 3D point cloud understanding also embraces a promising development owing to the SSL. Similarly to the image domain, these approaches can be mainly divided into two mainstream, *i.e.*, the contrastive pre-training and the generative pre-training approaches.

Contrastive Pre-training based approaches [9, 25, 50, 64] aim to learn instance discriminative representations to distinguish one sample from the others. Especially, PointContrast [64], as the pioneering approach that constructs two point clouds from different perspectives and compares point feature similarities for point cloud pre-training. Info3D [53] aims to maximize the mutual information between the 3D shape and a geometrically transformed version of the same shape with a MoCo [22]-like memory bank for caching the negative examples. Proposal-Contrast [68] enhances proposal representations by analyzing the geometric point relationships within each proposal, achieving this by optimizing for inter-cluster and inter-proposal separation to better adapt to 3D detection properties. FAC [32] forms advantageous point pairs from the same foreground segment with similar semantics and captures feature correlations within and across different point cloud views using adaptive learning. Note that the above methods that explored the contrastive training for point cloud are all based on CNN backbones. With the overwhelming development of ViTs, Point-BERT [69] firstly includes the classic MoCo-style contrastive learning into ViTs-based pre-training pipeline, but as a side exploration compared to its generative learning property. MaskPoint [30] converts the point cloud into discrete occupancy values, using binary classification to distinguish masked object points and sampled noise points. SoftClu [39] and CluRender [40] use clustering and rendering for point-level supervision, extracting

discriminative features without data augmentation. More recently, ReCon [45] integrated contrastive learning into pre-training to enhance performance within its generative pipeline. However, their approach employs supervised contrastive learning with a pre-constructed label set to mitigate overfitting, a prevalent concern in ViTs-based methods. Nonetheless, developing a self-supervised contrastive learning strategy tailored for ViTs-based point cloud pre-training remains challenging, particularly in effectively addressing the overfitting issue, which poses a significant obstacle in this domain.

Generative Pre-training based approaches are proposed inspired by the success of *mask and reconstruct* strategy used in BERT from NLP [11] to MAE [21] in vision with Transformers [13, 52, 56]. It prioritizes the encoder's pre-training by reconstructing masked information or its 2D projections. Notable examples include Point-BERT [69] and Point-MAE [30], which are proposed for point cloud pre-training with the masked reconstruction strategy. Point-M2AE [71] develops a hierarchical network that effectively models geometric and feature information progressively. The up-following works like TAP [60] and Ponder [24] focus on generating 2D projections of the point cloud as part of their pre-training strategies. Joint-MAE [19] addresses the correlation between 2D images and 3D point clouds, introducing hierarchical modules for cross-modal interaction to reconstruct masked information across both modalities. PointGPT [7] extends the concept of GPT [49] to point clouds pre-training with post-pre-training with larger datasets. PointDif [75] concentrates on refining the training approach with diffusion models [23]. Despite MAE models exhibiting favorable optimization properties [61] and delivering promising performance, their focus is on learning relationships among the tokens within the same input image, rather than modeling the relation among different samples as in contrastive learning, which results in less discriminative learned representations [26] or data filling issues [45, 65]. Moreover, in the context of point cloud reconstruction, the commonly used chamfer loss compels the model to precisely match the ground truth set. However, the ground truth itself represents just one sample from the true underlying distribution, posing a challenging optimization problem that often results in suboptimal performance [30]. To tackle this issue, we propose injecting feature-level contrastive properties into MAE pre-training.

3 Methodology

The overview of the proposed method Point-CMAE is illustrated in Fig. 3. Before diving into the detailed introduction of the proposed Point-CMAE, we first provide the point cloud embedding and the masked autoencoder for point cloud pre-training by ViTs in Sec. 3.1. Then the explanation of why the proposed method is designed via dual-masking augmentation is introduced in Sec. 3.2. Building upon the proposed dual-masking framework, we explicitly introduce contrastive properties into the MAE paradigm, as detailed in Sec 3.3.

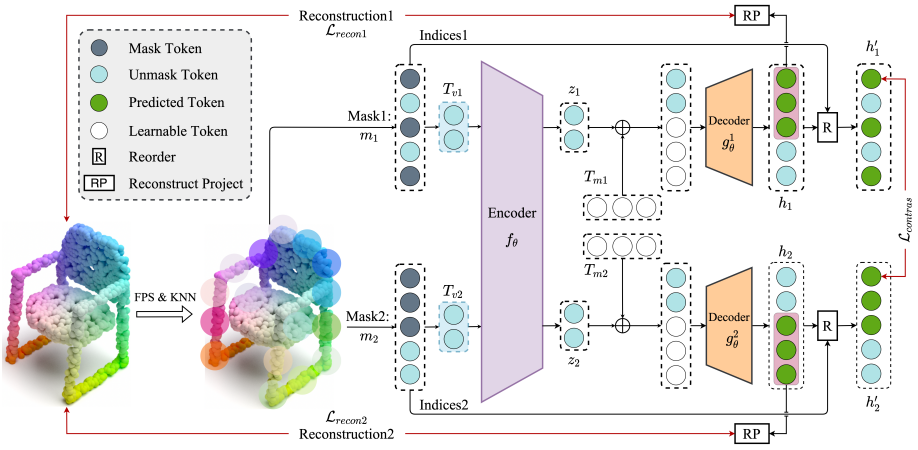

Fig. 3: The framework of the proposed Point-CMAE. The symbols $\oplus$ denote the token dimension concatenation. The point patch embedding is denoted as "FPS&KNN". The symbol $\oplus$ denotes the token dimension concatenation.

3.1 Preliminaries

Point Patch Embedding. Different from images that lie on regular grids which can be naturally divided into patches, point clouds are known to be irregular and less structured, based on this property, we follow Point-BERT [69] first to divide the input point cloud ($X^i \in \mathbb{R}^{p\times 3}$, p denotes the number of the points) into irregular point patches via Farthest Point Sampling (*i.e.*, FPS($\cdot$)) and K-Nearest Neighborhood (*i.e.*, KNN($\cdot$)) algorithm and output n center points c ($c = \text{FPS}(X^i), c \in \mathbb{R}^{n\times 3}$) and the corresponding neighbor points P ($P = \text{KNN}(X^i, c), P \in \mathbb{R}^{n\times k\times 3}$) of each center point. Finally, a lightweight PointNet [43] (*i.e.*, PointNet($\cdot$)) which mainly consists of MLPs is applied to the point patches (usually only for the visible point patches) to achieve the embedded tokens T ($T = \text{PointNet}(P), T \in \mathbb{R}^{n\times C}$, C denotes the embedding dimension).

Masked Autoencoder for point cloud with ViTs. For the embedded tokens T, to deploy the MAE strategy requires a mask m applied to T and outputs the visible tokens $T_v \in \mathbb{R}^{(1-r)n\times C}$ and the masked tokens $T_m \in \mathbb{R}^{rn\times C}$, where r denotes the mask ratio. Then the MAE of the point cloud [42, 69] can be summarized as point tokens which are masked with random mask m are fed into the encoder $f_\theta(\cdot)$, and then the decoder $g_\phi(\cdot)$ predicts the original masked points X_m with distribution $\mathcal{D}$:

$$\min_{\theta,\phi} \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \left[\mathcal{M}\left(g_\phi(z \oplus T_m), X_m\right)\right], \quad z = f_\theta(T_v). \tag{1}$$

Here $z \in \mathbb{R}^{rn\times C}$ denotes the latent feature of visible tokens T_v, $\mathcal{M}$ denotes the similarity measurement, and it was usually set as the Chamfer-Distance (*i.e.*, CD($\cdot$)) [15]. The symbol $\oplus$ denotes the token dimension concatenation. θ and ϕ are the trainable parameters of the encoder and the decoder, respectively.

3.2 Better Initialization for ViTs Encoder with Dual Masking

Inspired by the conclusion drawn in [28] that *the output feature is robust only if the most significant input part is not masked out*. Thereby, we propose to increase the invariance property of ViTs pre-training by increasing the probability of the most important input point patch that will not be masked. Unlike the heavy augmentation strategies (*e.g.*, color jittering) adopted in the image, point clouds contain only the position information. Moreover, point cloud pre-training is very sensitive to the geometry data augmentation (*i.e.*, a simple rotation usually brings a large performance increase) [12, 45], which will lead to ambiguity in knowing whether the contribution is made by the geometry data augmentation or the method itself.

To this end, we propose using two masks, m_1 and m_2, that share the same mask ratio r but incorporate different sources of randomness (*i.e.*, $m_1 \neq m_2$) as the augmentation operation. The same encoder f_θ is then used to process the visible tokens T_{v_1} and T_{v_2}, outputting the encoded features z_1 and z_2. Next, two separate decoders, g_θ^1 and g_θ^2, which share the same architecture but have weights that are updated differently based on their own inputs, are used. The main idea of the proposed method is illustrated in Fig. 3. Finally, the output of each decoder is projected back to the point cloud space for two separate reconstruction chamfer losses. Based on Eq. 1, the loss function can now be rewritten as follows:

$$\begin{aligned} \mathcal{L}_{re} &= \mathcal{L}_{recon1} + \mathcal{L}_{recon2} \\ &= \mathrm{CD}(\mathrm{RP}(g_\phi^1(z_1)), X_{m_1}) + \mathrm{CD}(\mathrm{RP}(g_\phi^2(z_2)), X_{m_2}), \end{aligned} \tag{2}$$

here $\mathrm{RP}(\cdot)$ is a fully connected layers-based projection head aiming to reconstruct masked point patches via projecting only the masked features back to the point coordinate cloud space. This simple yet effective design brings three advantages for point cloud pre-training: i) It decreases the possibility that a significant input part will be masked out (*e.g.*, though this token will be masked via m_1, now we have the possibility that it will not be masked out in m_2). ii) Two separate decoders force the encoder to learn better representations since two decoders for two different masks m_1 and m_2 require the encoder's output to be more informative to meet the requirements of both g_ϕ^1 and g_ϕ^2. iii) It naturally enriches the MAE paradigm with contrastive properties that largely improve the classification performance. Both i) and ii) are supported by the experimental results in Sec. 4 while the analysis of iii) is provided in Sec. 3.3.

3.3 Explicit Feature Level Contrastive Constraint

Given the ground truth points for the point cloud that are also one of the samplings of the original point cloud, directing using the Chamfer loss to minimize the difference between the ground truth points and the predicted points usually leads to a sub-optimize issue [30], especially for point cloud data that contains the position of each point that is largely different from the MAE in image domain where the pixel-level contraction is naturally more informative to reconstruct.

With the same masking ratio r for masks m_1 and m_2 in the dual-masking pipeline, there is a substantial probability that a token in the embedded tokens $T \in \mathbb{R}^{n \times C}$ could be masked simultaneously by both masks. The probability of this occurring can be calculated as:

$$p = 1 - (1 - r^2)^n. \tag{3}$$

In particular, given an example that $r = 0.6$, $n = 64$, $p \approx 0.945 >> 0$. Based on this observation, we propose to let the features h_1 and h_2 from both the decoders g_ϕ^1 and g_ϕ^2 of a certain point token that can be as close as possible to the feature level. Specifically, because in point-cloud MAE pretraining, the visible token after the encoder f_θ is directly concatenated with its corresponding masked token before passing through the corresponding decoder, and as the results, the output of the decoder still follows the same [visible, masked] order. This makes it non-trivial to find a token that was masked by both m_1 and m_2. As a remedy, we first recorded the output features from both decoders based on the indices when conducting the masking operations before f_θ:

$$h_1' = \mathrm{R}(Indices1, h_1), h_2' = \mathrm{R}(Indices2, h_2), \tag{4}$$

here $\mathrm{R}(\cdot)$ indicates reorder operation. Then the contrastive constraint can be written as:

$$\mathcal{L}_{contras} = \frac{m_1 \cap m_2}{|m_1 \cap m_2|} \sum_{i}^{n} (1 - \mathcal{M}(h_{1i}', h_{2i}')), \tag{5}$$

here $|m_1 \cap m_2|$ denotes the number of the co-masked point tokens by both m_1 and m_2. $\mathcal{M}$ is the cosine similarity measurement. Then the total optimization objective can be written as follows:

$$\mathcal{L} = \mathcal{L}_{re} + \lambda \mathcal{L}_{contras}, \tag{6}$$

$\mathcal{L}_{contras}$ here serve as a regularization term of $\mathcal{L}_{re}$, and λ is the regularization weight. As a result, the feature-level regularization, $\mathcal{L}_{contras}$, brings MAE explicit contrastive properties to ease the sub-optimize issue that was inherited from the MAE-based point cloud pre-training induced by the Chamfer loss [30].

4 Experiments

4.1 Self-supervised Pre-training Setups

Pre-training. We pre-train the proposed method ShapeNet [6]. ShapeNet is a synthetic 3D dataset that contains 52,470 3D shapes across 55 object categories. We pre-train our model only on the training set, which contains 41,952 shapes. For each 3D shape, we sample 1024 points to serve as the input for the mode, We set n as 64, which means each point cloud is divided into 64 patches. Furthermore, the KNN algorithm selects the $k = 32$ nearest point as a point patch. Following [30, 42], the proposed method is pre-trained for 300 epochs using an AdamW

Table 1: Classification results on ScanObjectNN. `DA`: rotation data augmentation is used during fine-tuning. The overall accuracy, *i.e.*, OA (%) is reported.

Method	#Params(M)	DA	OBJ_BG	OBJ_ONLY	PB_T50_RS
Supervised Learning Only					
PointNet [43]	3.5	-	73.3	79.2	68.0
SpiderCNN [66]	-	-	77.1	79.5	73.7
PointNet++ [44]	1.5	-	82.3	84.3	77.9
DGCNN [58]	1.8	-	82.8	86.2	78.1
PointCNN [29]	0.6	-	86.1	85.5	78.5
BGA-DGCNN [55]	1.8	-	-	-	79.7
BGA-PN++ [55]	1.5	-	-	-	80.2
DRNet [46]	-	-	-	-	80.3
GBNet [47]	8.8	-	-	-	80.5
SimpleView [16]	-	-	-	-	80.5±0.3
PRANet [10]	2.3	-	-	-	81.0
MVTN [20]	-	-	-	-	82.8
PointMLP [37]	13.2	-	-	-	85.4±0.3
with Standard ViTs and Single-Modal Self-Supervised Learning (FULL)					
Transformer [56]	22.1	×	79.86	80.55	77.24
OcCo [57]	22.1	×	84.85	85.54	78.79
Point-BERT [69]	22.1	×	87.43	88.12	83.07
MaskPoint [30]	22.1	×	89.30	88.10	84.30
Point-MAE [42]	22.1	×	**90.02**	88.29	85.18
Point-CMAE (Ours)	22.1	×	**90.02**	**88.64**	**85.95**
Point-CMAE (Ours)	22.1	✓	**93.46**	**91.05**	**88.75**
with Hierarchical ViTs / Multi-Modal/Post-Process Self-Supervised Learning (FULL)					
Point-M2AE [71]	15.3	×	91.22	88.81	86.43
Joint-MAE [19]	-	×	90.94	88.86	86.07
ACT [12]	22.1	✓	93.29	91.91	88.21
PointGPT-S [7]	29.2	✓	93.39	92.43	89.17
with Standard ViTs and Single-Modal Self-Supervised Learning (MLP-LINEAR)					
Point-MAE [42]	22.1	×	82.58 ± 0.58	**83.52** ± 0.41	73.08 ± 0.30
Point-CMAE (Ours)	22.1	×	**83.48 ± 0.31**	83.45 ± **0.35**	**73.15 ± 0.11**
with Standard ViTs and Single-Modal Self-Supervised Learning (MLP-3)					
Point-MAE [42]	22.1	×	84.29 ± 0.55	85.24 ± 0.41	77.34 ± **0.12**
Point-CMAE (Ours)	22.1	×	**85.88 ± 0.53**	**85.60 ± 0.35**	**77.47** ± 0.13

optimizer [35]. In the autoencoder's backbone, the encoder has 12 Transformer blocks while the decoder has 4 ViTs encoder blocks. Each Transformer block has 384 hidden dimensions and 6 heads. MLP ratio in Transformer blocks is set to 4. The batch size was set to 128 during the entire pre-training. The initial learning rate was set to 0.0005 with cosine learning rate decay (the decay weight was 0.05) employed. More details regarding our experimental configuration and implementation are provided in our supplementary materials (*i.e.*, *Supp. Mat.*)

Transfer Protocol. Similar to [12, 45], we adopt three variants of transfer learning protocols for classification tasks during fine-tuning. *i.e.*, (a) **Full**: Fine-tuning pre-trained models by updating all backbone and classification heads. (b)

Table 2: Classification results on the ModelNet40 dataset. The overall accuracy, *i.e.*, OA (%) is reported. `[ST]`: standard Transformer architecture. *: The reproduced results.

Method	[ST]	#Point	OA (%)
Supervised Learning Only			
PointNet [43]	-	1k P	89.2
PointNet++ [44]	-	1k P	90.7
PointCNN [29]	-	1k P	92.5
DGCNN [58]	-	1k P	92.9
DensePoint [33]	-	1k P	93.2
PointASNL [62]	-	1k P	92.9
DRNet [46]	-	1k P	93.1
Point Trans. [14]	×	1k P	92.8
PCT [18]	×	1k P	93.2
PointTransformer [73]	×	1k P	93.7
NPCT [18]	✓	1k P	91.0
Transformer [56]	✓	1k P	91.4
OcCo [57]	✓	1k P	92.1
Point-BERT [69]	✓	1k P	93.2
Point-MAE [42]	✓	1k P	93.8
Point-MAE* [42]	✓	1k P	93.5
Point-CMAE (Ours)	✓	1k P	**93.6**
with Self-Supervised Representation Learning (MLP-LINEAR)			
Point-MAE [42]	✓	1k P	91.22±0.26
Point-CMAE (Ours)	✓	1k P	**92.30±0.32**
with Self-Supervised Representation Learning (MLP-3)			
Point-MAE [42]	✓	1k P	92.33±0.09
Point-CMAE (Ours)	✓	1k P	**92.60±0.19**

MLP-Linear: The classification head is a single-layer linear MLP, and we only update these head parameters during fine-tuning. (c) **MLP-3**: The classification head is a three-layer non-linear MLP (*i.e.*, the same as the one used in FULL), and we only update these head parameters during fine-tuning.

4.2 Transfer Learning on Downstream Tasks

To assess the efficacy of the pre-trained model, we gauged its performance on various fine-tuned tasks using numerous real-world datasets.

3D Real-Word Object Classification. We use the scanned ScanObjectNN [55] dataset to evaluate the shape recognition ability of the pre-trained model of our method. The ScanObjectNN [55] dataset covers around 15K real-world objects from 15 categories, and it is divided into three subsets: OBJ-BG (objects and background), OBJ-ONLY (only objects), and PB-T50-RS (objects, background, and artificially added perturbations). For a fair comparison, we report the results without voting strategy [34]. The results in Tab. 1 demonstrate that: (i) Without increasing parameters, the proposed Point-CMAE significantly improves accuracy using standard ViT architecture under the single-modal full-tuning protocol, even surpassing supervised methods. Notably, it achieves a **0.77%** improvement under the challenging PB_T50_RS setting. (ii) When fine-tuning with rotation data augmentation, as used by ACT and PointGPT, Point-CMAE still outperforms ACT on OBJ_BG (by **0.17%**) and PB_T50_RS (by **0.54%**), despite both ACT and PointGPT using multi-modal information. (iii) The performance of Point-CMAE on MLP-Linear and MLP-3 further highlights the enhanced generalization of the pre-trained model, even when only fine-tuning the classification head.

3D Synthetic Object Classification. We also evaluate 3D shape classification on the synthetic ModelNet40 dataset [63], which contains 12,311 clean 3D CAD models across 40 object categories. Following the standard protocols of [42,69], we split ModelNet40 into 9,843 training instances and 2,468 testing instances. Data augmentation during training includes random scaling and random translation. The results, presented in Tab.2, use the voting strategy for fair comparison

Table 3: Few-shot classification on **ModelNet40**, overall accuracy (%) is reported.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Representation Learning				
PointNet [43]	52.0 ± 3.8	57.8 ± 4.9	46.6 ± 4.3	35.2 ± 4.8
PointNet-OcCo [57]	89.7 ± 1.9	92.4 ± 1.6	83.9 ± 1.8	89.7 ± 1.5
PointNet-CrossPoint [1]	90.9 ± 4.8	93.5 ± 4.4	84.6 ± 4.7	90.2 ± 2.2
DGCNN [59]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
DGCNN-CrossPoint [1]	92.5 ± 3.0	94.9 ± 2.1	83.6 ± 5.3	87.9 ± 4.2
with Self-Supervised Representation Learning (Full)				
Transformer [56]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
OcCo [57]	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
Point-BERT [69]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
MaskPoint [30]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± **4.0**	93.4 ± 3.5
Point-MAE [42]	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± **3.0**
Point-CMAE (Ours)	**96.7** ± **2.2**	**98.0** ± **0.9**	**92.7** ± 4.4	**95.3** ± 3.3
with Hierarchical ViTs/Multi-Modal/Post-Process Self-Supervised Learning (Full)				
Point-M2AE [71]	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
Joint-MAE [19]	96.7 ± 2.2	97.9 ± 1.8	92.6 ± 3.7	95.1 ± 2.6
Point-GPT [7]	96.8 ± 2.0	98.6 ± 1.1	92.6 ± 4.6	95.2 ± 3.4
ACT [12]	96.8 ± 2.3	98.0 ± 1.4	93.3 ± 4.0	95.6 ± 2.8
with Self-Supervised Representation Learning (Mlp-Linear)				
Point-MAE	**91.1** ± 5.6	91.7 ± 4.0	83.5 ± 6.1	89.7 ± **4.1**
Point-CMAE (Ours)	90.4 ± **4.2**	**94.1** ± **3.9**	**89.2** ± **5.5**	**92.3** ± 4.5
with Self-Supervised Representation Learning (Mlp-3)				
Point-MAE	95.0 ± **2.8**	96.7 ± 2.4	90.6 ± 4.7	93.8 ± 5.0
Point-CMAE (Ours)	**95.9** ± 3.1	**97.5** ± **2.0**	**91.3** ± **4.6**	**94.4** ± **3.7**

when reproducing Point-MAE [42] and Point-CMAE under the Full evaluation protocol. Key findings include that Point-CMAE demonstrates effectiveness on the synthetic dataset, with slight improvements over baseline methods. This effectiveness is further confirmed under the MLP-Linear and MLP-3 protocols, where the pre-trained backbone is frozen.

Few-shot Classification. We follow previous works [30, 42, 45, 69] to conduct few-shot learning experiments on ModelNet40 [63], adopting num_cls-way, num_sample-shot setting, where num_cls is the number of classes that randomly selected from the dataset and num_sample is the number of objects randomly sampled for each class. We use the above-mentioned $num\_cls \times num\_sample$ objects for training. During testing, we randomly sample 20 unseen objects from each of num_cls classes for evaluation. The results with the setting of num_cls in 5, 10 and num_sample in 10, 20 are presented in Tab. 3. Following all three protocols, we conduct 10 independent experiments for each

Table 4: Part segmentation on **ShapeNetPart**. The class mIoU (mIoU_C) and the instance mIoU (mIoU_I) are reported, with model parameters #P and FLOPs #F.

Method	mIoU_C (%) ↑	mIoU_I (%) ↑	#P (M) ↓	#F (G) ↓
Supervised Representation Learning				
PointNet [43]	80.4	83.7	3.6	0.5
PointNet++ [44]	81.9	85.1	1.0	1.7
DGCNN [58]	82.3	85.2	1.3	2.4
Transformer [56]	83.4	85.1	22.1	4.8
PointMLP [37]	84.6	86.1	13.2	31.4
with Self-Supervised Representation Learning				
Transformer [56]	83.4	85.1	22.1	4.8
OcCo [57]	83.4	84.7	22.1	-
PointContrast [64]	-	85.1	37.9	-
CrossPoint [1]	-	85.5	-	-
Point-BERT [69]	84.1	85.6	22.1	4.8
MaskPoint [30]	84.6	86.0	22.1	4.8
Point-MAE [42]	84.2	**86.1**	22.1	4.8
Point-CMAE (Ours)	**84.9**	86.0	22.1	4.8
with Hierarchical ViTs / Multi-Modal/Post-Process Self-Supervised Learning				
Point-M2AE [71]	84.8	86.5	12.8	3.6
PointGPT-L [7]	84.8	86.6	29.2	6.4
ACT [12]	84.7	86.1	22.1	4.8
Recon [45]	84.8	86.4	43.6	5.3

setting and report mean accuracy with standard deviation. Tab. 3 shows that: (i) Besides the truth that our self-supervised Point-CMAE outperforms the supervised solutions by a large margin, Point-CMAE brings significant improvements of +8.8%, +4.7%, +8.1%, +5.9% respectively for the four settings over from scratch FULL transferring baseline [56]. (ii) Our Point-CMAE not only consistently achieves the best performance compared to our baseline [42], but it is also worth pointing out that with the standard ViTs architecture pre-trained with only point cloud data, Point-CMAE achieves competitive or even better performance compared to other state-of-the-art methods which were proposed either with complex hierarchical ViTs structure (*i.e.*, Point-M2AE) or trained with multi-modal information (*i.e.*, Joint-MAE, Point-GPT, and ACT). (iii) For the MLP-Linear and MLP-3 transfer protocol, we observed that the Point-CMAE achieves obvious improvements (*e.g.*, 2.4%, 5.7%, and 2.6% for the 5way20shot, 10way10shot, and the 10way20shot under the MLP-Linear protocol) compared to Point-MAE with smaller deviations.

3D Part Segmentation. We evaluate the segmentation performance of the proposed Point-CMAE on ShapeNetPart [67] dataset. It contains 16,881 objects of 2,048 points from 16 categories with 50 parts in total. The segmentation head in our method is the same as in Point-MAE [42], which is relatively simple and does not use any propagating operation or DGCNN [58]. We use the learned features from the 4th, 8th, and 12th layers of the Transformer block, and concatenate the three levels of features. Then an average pooling and a max pooling were applied separately to obtain two global features. For Transformer, we used per-category

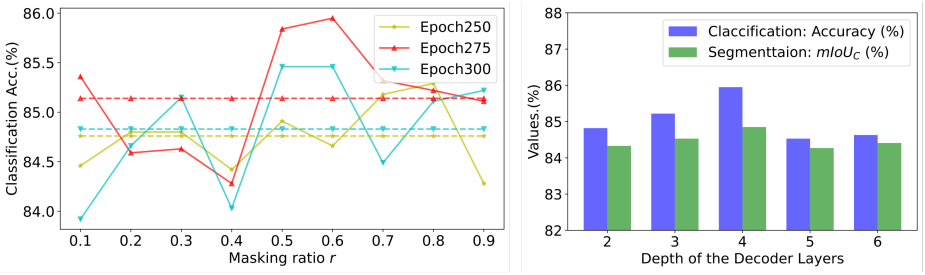

Fig. 4: (a) The classification results on the ScanobjectNN [55] (PB_T50_RS) dataset cross different masking ratios. The corresponding average results across the entire masking ratios are depicted with the dashed lines. (b) Both the classification on the ScanobjectNN [55] (PB_T50_RS) and the part segmentation results on the ShapeNet-Part [67] are provided to study how the depth of the decoder affects the pre-traing.

mean IoU (mIoU$_C$) and mean IoU averaged over all test instances (mIoU$_I$) to assess performance. The part segmentation results provided in Tab. 4 show that the proposed Point-CMAE archives the best per-category mean IoU and the second-best mean IoU averaged over all test instances. Especially, a significant improvement 0.7% compared to Point-MAE [42] on mIoU$_C$. When compared to other state-of-the-art methods that adopt hierarchical ViTs architecture (*i.e.*, Point-M2AE), utilize post-training on extra-large datasets (*i.e.*, PointGPT-L), or incorporate multi-modal information (*i.e.*, ACT and Recon), our Point-CMAE still slightly outperforms these methods in terms of the evaluation metric mIoU$_C$.

4.3 Ablation Study

The sweet spot of the mask ratio. The mask ratio has been validated significantly for MAE-based self-supervised learning in both the image and point cloud domains [21, 42, 69]. To determine a suitable masking ratio for our method, we varied the mask ratio from 0.1 to 0.9. The corresponding results are depicted in Fig. 4 (a). Specifically, using a pre-trained Point-CMAE model, we followed Recon [45] and evaluated the model checkpoints from epochs 250, 275, and 300. We found that the checkpoint from epoch 275 performs well for our method, as indicated by the average results across mask ratios shown with dashed lines in Fig. 4 (a). A mask ratio of 0.6 consistently produced the best results compared to other ratios. Therefore, we adopted a mask ratio of 0.6 throughout this paper, which is also consistent with findings in related works [42, 45].

The effect of the depth of the ViTs decoder. Fig. 4 (b) shows the fine-tuning performance for both the classification and the part segmentation tasks with different numbers of the ViTs layers within the decoder. It can be seen that the performance increase when the depth is in increased from 2 to 4, while it decrease when the depth is further increased from 4 to 6. We set the decoder 4 ViTs transformer layers throughout our work based on this observation. This setting is also consistent with our baseline method Point-MAE [42].

Table 5: The effect of each component on ScanObjectNN [55] and ShapeNetPart [67] datasets for classification (OA) and part segmentation (mIoU_C).

Methods	Dual-masking	Same Encoder	Same decoder	Contrastive	OA(%) ↑	mIoU_C (%) ↑
Baseline [42]	×	-	-	-	85.18	84.20
(a)	✓	×	×	×	85.51	84.32
(b)	✓	×	×	✓	85.60	84.45
(c)	✓	×	✓	×	85.23	84.33
(d)	✓	×	✓	✓	85.32	84.51
(e)	✓	✓	×	×	85.70	84.65
(f)	✓	✓	×	✓	85.95	84.85
(g)	✓	✓	✓	×	85.43	84.50
(h)	✓	✓	✓	✓	85.37	84.43

The effect of each component. The ablation studies regarding how each component affects both the classification and the part segmentation performance of the proposed Point-CMAE are shown in Tab. 5. It mainly uncovers that: (i) When just using the proposed dual-masking strategy (Sec. 3.2) from (a), an obvious improvement can be achieved for both the classification and the segmentation. (ii) Using one same encoder during pre-training is better than using two separate encoders. (iii) Using the same decoder to reconstruct the point cloud from two masks degenerates the overall performance for both the classification and the segmentation tasks. (iv) The contrastive learning strategy proposed in our paper brings obvious improvements. In particular, built upon the same encoder and two separate decoders during pre-training, (f) archives the best results compared to the rest, and we set (f) as our full model.

5 Conclusion

We propose Point-CMAE, a self-supervised method that integrates the MAE pre-training paradigm with explicit contrastive properties for point clouds. Specifically, we experimentally demonstrate that directly combining classic contrastive learning with generative MAE degrades the SSL performance. To address this, we propose a simple dual masking strategy that effectively introduces explicit contrastive properties. The feature-level contrastive constraint enables Point-CMAE to achieve notable improvements over the baseline method, even surpassing some state-of-the-art methods that use complex hierarchical architectures, post-training techniques, or multi-modal information. Additionally, our findings indicate that masking plays a significant role in MAE-based SSL, highlighting the importance of careful attention to masking strategies, especially for point clouds.

Acknowledgement. This work was partially supported by Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064), the National Natural Science Foundation of China (No. 62203476), the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU, the PRIN project CREATIVE (Prot. 2020ZSL9F9), the EU Horizon project ELIAS (No. 101120237), and the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9902–9912 (2022) 11, 12
2. An, Z., Sun, G., Liu, Y., Liu, F., Wu, Z., Wang, D., Van Gool, L., Belongie, S.: Rethinking few-shot 3d point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3996–4006 (2024) 2
3. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (2021) 4
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2
5. Cai, J., Huang, W., You, Y., Chen, Z., Ren, B., Liu, H.: Spsd: Semantics and deep reinforcement learning based motion planning for supermarket robot. IEICE TRANSACTIONS on Information and Systems **106**(5), 765–772 (2023) 2
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 8
7. Chen, G., Wang, M., Yang, Y., Yu, K., Yuan, L., Yue, Y.: Pointgpt: Auto-regressively generative pre-training from point clouds. Advances in Neural Information Processing Systems **36** (2024) 5, 9, 11, 12
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 2, 4
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021) 2, 4
10. Cheng, S., Chen, X., He, X., Liu, Z., Bai, X.: Pra-net: Point relation-aware network for 3d point cloud analysis. IEEE Transactions on Image Processing **30**, 4436–4448 (2021) 9
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2, 5
12. Dong, R., Qi, Z., Zhang, L., Zhang, J., Sun, J., Ge, Z., Yi, L., Ma, K.: Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In: The Eleventh International Conference on Learning Representations (2022) 7, 9, 11, 12
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2020) 5
14. Engel, N., Belagiannis, V., Dietmayer, K.: Point transformer. IEEE access **9**, 134826–134840 (2021) 10
15. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017) 6

16. Goyal, A., Law, H., Liu, B., Newell, A., Deng, J.: Revisiting point cloud shape classification with a simple and effective baseline. In: International Conference on Machine Learning. pp. 3809–3820. PMLR (2021) 9
17. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020) 2, 3, 4
18. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**, 187–199 (2021) 10
19. Guo, Z., Zhang, R., Qiu, L., Li, X., Heng, P.A.: Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 791–799 (2023) 5, 9, 11
20. Hamdi, A., Giancola, S., Ghanem, B.: Mvtn: Multi-view transformation network for 3d shape recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2021) 9
21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 4, 5, 13
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 2, 3, 4
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 5
24. Huang, D., Peng, S., He, T., Yang, H., Zhou, X., Ouyang, W.: Ponder: Point cloud pre-training via neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16089–16098 (2023) 5
25. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021) 4
26. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 5
27. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 2
28. Kong, X., Zhang, X.: Understanding masked image modeling via learning occlusion invariant feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6241–6251 (2023) 7
29. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems **31** (2018) 9, 10
30. Liu, H., Cai, M., Lee, Y.J.: Masked discrimination for self-supervised learning on point clouds. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II. pp. 657–675. Springer (2022) 2, 4, 5, 7, 8, 9, 11, 12
31. Liu, H., Ren, B., Liu, M., Ding, R.: Grouped temporal enhancement module for human action recognition. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 1801–1805. IEEE (2020) 2

32. Liu, K., Xiao, A., Zhang, X., Lu, S., Shao, L.: Fac: 3d representation learning via foreground aware feature contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9476–9485 (2023) 4
33. Liu, Y., Fan, B., Meng, G., Lu, J., Xiang, S., Pan, C.: Densepoint: Learning densely contextual representation for efficient point cloud processing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5239–5248 (2019) 10
34. Liu, Y., Fan, B., Xiang, S., Pan, C.: Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8895–8904 (2019) 10
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018) 9
36. Ma, Q., Li, Y., Ren, B., Sebe, N., Konukoglu, E., Gevers, T., Van Gool, L., Paudel, D.P.: Shapesplat: A large-scale dataset of gaussian splats and their self-supervised pretraining. arXiv preprint arXiv:2408.10906 (2024) 2
37. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In: International Conference on Learning Representations (2021) 9, 12
38. Mei, G., Huang, X., Zhang, J., Wu, Q.: Overlap-guided coarse-to-fine correspondence prediction for point cloud registration. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022) 2
39. Mei, G., Saltori, C., Poiesi, F., Zhang, J., Ricci, E., Sebe, N., Qiang, W.: Data augmentation-free unsupervised learning for 3d point cloud understanding. In: The 33rd British Machine Vision Conference Proceedings (2022) 4
40. Mei, G., Saltori, C., Ricci, E., Sebe, N., Wu, Q., Zhang, J., Poiesi, F.: Unsupervised point cloud representation learning by clustering and neural rendering. International Journal of Computer Vision pp. 1–19 (2024) 2, 4
41. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4
42. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: European conference on computer vision. pp. 604–621. Springer (2022) 3, 6, 8, 9, 10, 11, 12, 13, 14
43. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 6, 9, 10, 11, 12
44. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017) 9, 10, 12
45. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In: International Conference on Machine Learning. pp. 28223–28243. PMLR (2023) 2, 5, 7, 9, 11, 12, 13
46. Qiu, S., Anwar, S., Barnes, N.: Dense-resolution network for point cloud classification and segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3813–3822 (2021) 9, 10
47. Qiu, S., Anwar, S., Barnes, N.: Geometric back-projection network for point cloud classification. IEEE Transactions on Multimedia **24**, 1943–1955 (2021) 9
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2

49. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018) 2, 5
50. Rao, Y., Lu, J., Zhou, J.: 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. pp. 5376–5385 (2020) 4
51. Ren, B., Liu, M., Ding, R., Liu, H.: A survey on 3d skeleton-based action recognition using learning method. Cyborg and Bionic Systems **5**, 0100 (2024) 4
52. Ren, B., Liu, Y., Song, Y., Bi, W., Cucchiara, R., Sebe, N., Wang, W.: Masked jigsaw puzzle: A versatile position embedding for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20382–20391 (2023) 2, 5
53. Sanghi, A.: Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 626–642. Springer (2020) 4
54. Tang, H., Ding, L., Wu, S., Ren, B., Sebe, N., Rota, P.: Deep unsupervised key frame extraction for efficient video classification. ACM Transactions on Multimedia Computing, Communications and Applications **19**(3), 1–17 (2023) 4
55. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019) 3, 9, 10, 13, 14
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 5, 9, 10, 11, 12
57. Wang, H., Liu, Q., Yue, X., Lasenby, J., Kusner, M.J.: Unsupervised point cloud pre-training via occlusion completion. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9782–9792 (2021) 9, 10, 11, 12
58. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) **38**(5), 1–12 (2019) 9, 10, 12
59. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019) 11
60. Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J.: Take-a-photo: 3d-to-2d generative pre-training of point cloud models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5640–5650 (2023) 5
61. Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., Guo, B.: Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint arXiv:2205.14141 (2022) 5
62. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 9621–9630 (2019) 10
63. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015) 10, 11
64. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 574–591. Springer (2020) 4, 12

65. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., Hu, H.: On data scaling in masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10365–10374 (2023) 5
66. Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Y.: Spidercnn: Deep learning on point sets with parameterized convolutional filters. In: Proceedings of the European conference on computer vision (ECCV). pp. 87–102 (2018) 9
67. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016) 12, 13, 14
68. Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C.Z., Shen, J., Wang, W.: Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In: European conference on computer vision. pp. 17–33. Springer (2022) 4
69. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: 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. pp. 19313–19322 (2022) 2, 4, 5, 6, 9, 10, 11, 12, 13
70. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International conference on machine learning. pp. 12310–12320. PMLR (2021) 4
71. Zhang, R., Guo, Z., Gao, P., Fang, R., Zhao, B., Wang, D., Qiao, Y., Li, H.: Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. Advances in neural information processing systems **35**, 27061–27074 (2022) 5, 9, 11, 12
72. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8552–8562 (2022) 2
73. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021) 10
74. Zhao, M., Liu, M., Ren, B., Dai, S., Sebe, N.: Denoising diffusion probabilistic models for action-conditioned 3d motion generation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4225–4229. IEEE (2024) 2
75. Zheng, X., Huang, X., Mei, G., Hou, Y., Lyu, Z., Dai, B., Ouyang, W., Gong, Y.: Point cloud pre-training with diffusion models. arXiv preprint arXiv:2311.14960 (2023) 5