

General Point Model Pretraining with Autoencoding and Autoregressive

Zhe Li^{1,†}, Zhangyang Gao^{3,†}, Cheng Tan^{3,†}, Bocheng Ren¹, Laurence T. Yang^{1,2,*}, Stan Z. Li^{3,*}

¹ Huazhong University of Science and Technology

² Zhengzhou University

³ AI Lab, Research Center for Industries of the Future, Westlake University

keycharon0122@gmail.com, {gaozhangyang,tancheng,stan.zq.li}@westlake.edu.cn, ltyang@ieee.org

Abstract

The pre-training architectures of large language models encompass various types, including autoencoding models, autoregressive models, and encoder-decoder models. We posit that any modality can potentially benefit from a large language model, as long as it undergoes vector quantization to become discrete tokens. Inspired by the General Language Model, we propose a General Point Model (GPM) that seamlessly integrates autoencoding and autoregressive tasks in a point cloud transformer. This model is versatile, allowing fine-tuning for downstream point cloud representation tasks, as well as unconditional and conditional generation tasks. GPM enhances masked prediction in autoencoding through various forms of mask padding tasks, leading to improved performance in point cloud understanding. Additionally, GPM demonstrates highly competitive results in unconditional point cloud generation tasks, even exhibiting the potential for conditional generation tasks by modifying the input’s conditional information. Compared to models like Point-BERT, MaskPoint, and PointMAE, our GPM achieves superior performance in point cloud understanding tasks. Furthermore, the integration of autoregressive and autoencoding within the same transformer underscores its versatility across different downstream tasks. Codes are available at <https://github.com/gentlefress/GPM>

1. Introduction

In recent years, the natural language processing (NLP) [4, 19, 20, 29, 40] and computer vision (CV) [3, 7, 9, 49, 57, 66] realms have witnessed a proliferation of transformer-based pretrained models. Their primary advantage lies in the inclusion of massive parameters and data in the training process, overcoming inductive biases introduced by traditional Convolutional Neural Networks [21]. Point clouds serve as fundamental data structures in fields like autonomous driving and robotics, thus emphasizing the escalating significance of tasks related to point cloud representation learn-

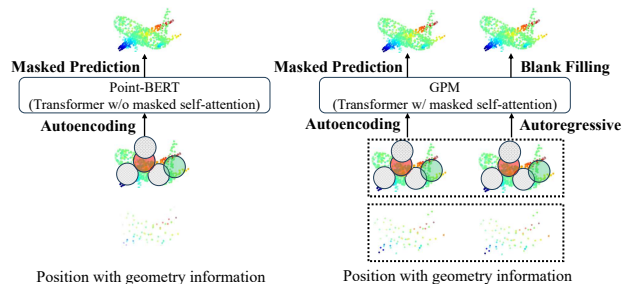


Figure 1. Comparison of pre-training frameworks between GPM and Point-BERT [63]. Point-BERT (left) conducts the masked prediction task in an autoencoding manner, while GPM (right) combines both autoencoding and autoregressive tasks, performing tasks for both masked prediction and masked region blank filling.

ing and generation. However, the realm of pretrained point cloud models based on transformers remains relatively scarce. Existing transformer-based point cloud models [13, 65] encounter inevitable inductive biases due to local feature aggregation [13], neighbor embedding [65], and the scarcity of annotated data [9]. These biases deviate from the mainstream of standard Transformers. Self-supervised models [12, 15, 20, 25, 26, 39, 40, 61] have emerged as the dominant methodology. They excel in learning high-quality representations across modalities without extensive labeled data, models like GPT [39] even dominate in text generation. In this end, there is a compelling need to design a transformer-based point cloud model that minimizes inductive biases, learns superior point cloud representations from limited data, and simultaneously possesses the capacities for conditional and unconditional point cloud generation. The transformer-based BERT [20] has marked a significant milestone in the field of natural language processing (NLP). Its pre-training phase, involving masked prediction tasks, grants the model the capacity to learn language representations, showcasing the prowess of autoencoding language models. Given BERT’s proficiency in learning language representations, the question arises: can we extend BERT’s

capabilities to the realm of point clouds? However, since point clouds lack a direct analog to the concept of 'words' in language, constructing a discrete vocabulary for point clouds, as done in language models, is infeasible. Point clouds are composed of individual points, and considering each point as a token would lead to a quadratic increase in computational cost as the number of tokens grows. Furthermore, each point in a point cloud contains limited semantic information, necessitating the maximization of geometric information utilization. Inspired by Point-BERT [63], we embarked on a journey to vector quantization, transforming point clouds into discrete tokens. Through the application of the farthest point sampling (FPS) [36] algorithm, we segment the point cloud into multiple patches, each patch encapsulates plentiful geometric information, akin to a unit. By constructing a vocabulary specific to the point cloud domain, we obtain labels for each unit, discretizing the entire point cloud. This paves the way for the completion of the autoencoding model's masked prediction tasks in the point cloud domain.

Transformer-based GPT [39] has dominated the field of text generation, with its core being an autoregressive language model, predicting the next token based on preceding tokens. In recent years, researchers have progressively extended this paradigm into the realm of images [7, 41, 42], ushering in a new era of image generation. In this endeavor, we bring this paradigm to the domain of point clouds, aiming to empower it for autoregressive conditional and unconditional point cloud generation tasks. After discretization, each patch of the point cloud possesses distinct geometric information, enabling it to excel in autoregressive tasks.

While these pre-training frameworks can be adapted to the point cloud domain, they lack the flexibility to satisfy all point cloud tasks. Therefore, we aim to integrate these two pre-training frameworks from the NLP domain and apply them to the point cloud domain, depicted in Figure 1.

In this work, we introduce a novel approach termed **General Point Model (GPM)** with autoencoding and autoregressive, which integrates point cloud representation learning and point cloud generation (both conditional and unconditional) within the same transformer framework. Similar to Point-BERT, we employ a point cloud tokenizer designed through dVAE-based point cloud reconstruction, where the point cloud can be discretely labeled based on the learned vocabulary. Additionally, we utilize the Masked Point Modeling (MPM) task. This involves partitioning the point cloud into distinct patches and randomly masking some adjacent patches. This aims to predict the masked portion, with the goal of learning geometric information between neighboring blocks and capturing meaningful geometric features for understanding the point cloud.

Our primary contributions can be summarized as follows:

- We introduce **autoencoding + autoregressive** as a novel point cloud transformer paradigm, unifying tasks in point cloud understanding and generation.
- We partition the input into two segments, the first part predicts masked points (autoencoding), the second part generates masked points (autoregressive), enhancing comprehension and generation.
- Our approach demonstrates notable competitiveness across various point cloud understanding tasks and unconditional generation tasks. Moreover, it exhibits the potential for conditional generation, such as text- or image-conditioned point clouds generation.

2. Related Work

2.1. Self-supervised Learning

Self-supervised learning (SSL) has garnered significant attention owing to the substantial labor required for acquiring annotated data in large quantities. SSL enables models to learn feature representations from unlabeled data, adapting to downstream tasks. The core of SSL lies in the design of proxy tasks to replace traditional classification tasks, thereby learning feature representations. ELMo [44] employs bidirectional LSTMs [16] and generates subsequent words from left to right given the representation of preceding content. ViP [25] employs a dynamically updated momentum encoder for contrastive learning and designs a text swapping task to enhance the model's sentence representation capability. In the computer vision domain, Image-GPT [7] trains a sequence transformer to autoregressively predict pixels, showing promising representation learning capabilities without incorporating specific knowledge about the 2D input structure after pretraining. Furthermore, SSL in the field of point clouds has garnered significant attention. ACT [8] employs cross-modal autoencoders as teacher models to acquire knowledge from other modalities. The generative approaches [1, 22, 33, 45, 63, 64] are the most pertinent research to our work. Point-MAE [34] extends MAE by randomly masking point patches and reconstructing the masked regions. Point-M2AE [64] further utilizes a hierarchical transformer architecture and designs corresponding masking strategies. However, mask-based point modeling methods still face the issue of shape leakage, limiting their effective generalization to downstream tasks. In this paper, we leverage autoregressive pretraining on point clouds and address the unique challenges associated with point cloud attributes. Our concise design avoids position information leakage, thereby enhancing generalization capabilities.

2.2. General Language Model

In the field of natural language processing, GLM [10] integrates both autoregressive and autoencoding methods. This enables it to perform tasks related to sentence representa-

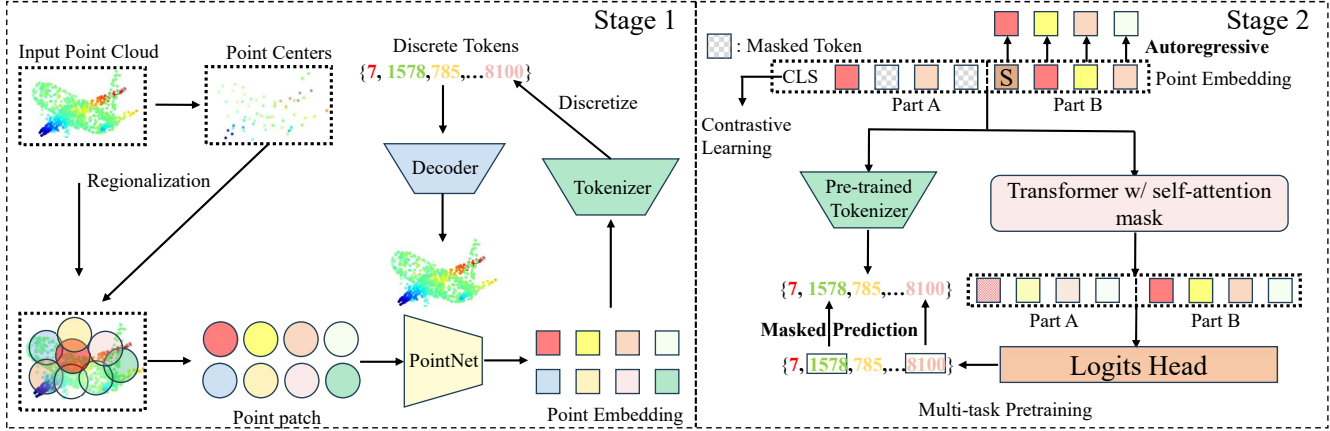


Figure 2. The framework of our GPM. We divide the entire pipeline into two stages: 1) The input point cloud is divided into several sub-clouds. Then, we utilize Mini-PointNet [35] to obtain a sequence of point embeddings. These embeddings are transformed into a sequence of discrete point tokens using a pre-trained tokenizer; 2) During pre-training, for PartA, certain portions of the masked embeddings are masked and replaced with [M]. They are predicted in an autoencoding manner. For PartB, masks for PartA are generated in an autoregressive manner.

tion as well as conditioned and unconditioned generation through fine-tuning. To the best of our knowledge, we are the first to integrate autoregressive and autoencoding techniques on a single transformer in the domain of point clouds. In downstream tasks, we have the potential to fine-tune point cloud representations and perform unconditioned generation even in zero-shot scenarios, with the possibility of fine-tuning towards conditioned generation.

3. Methods

In this study, we endeavor to seamlessly integrate BERT-style and GPT-style pre-trained strategies, extending their applicability to point cloud transformers. Following the Point-BERT[63], our training process unfolds in two distinctive stages. Firstly, we embark on training a specialized Tokenizer, a critical step towards acquiring discrete point labels for each input point cloud. Concurrently, a dedicated discrete VAE (dVAE) is employed to masterfully reconstruct the discrete point cloud. The subsequent stage involves the splicing of two sequences of point cloud embeddings, partA for autoencoding and partB for autoregressive. The overall idea of our approach is illustrated in Figure 2.

3.1. Stage 1: Discrete Variational Autoencoder Pre-training

Point Cloud Partitioning. We posit that discrete tokens derived from points encapsulate crucial geometric information, allowing for a discrete representation of any point cloud. However, a naive strategy of assigning one token to each point in a point-wise reconstruction task poses a formidable computational challenge. This arises from the quadratic complexity of self-attention within transformers.

Building upon the foundations laid by Point-BERT [63] and ViT[9], we adopt a strategy of partitioning each point cloud into distinct patches, each serving as a single token. Specifically, for an input point cloud $p \in \mathbb{R}^{\mathcal{N} \times 3}$, we initiate the process by extracting m center points $\mathcal{C}_u \in \mathbb{R}^{m \times 3}$ from the point cloud p through farthest point sampling (FPS). Subsequently, we employ the k -nearest neighbor (kNN) algorithm to identify the k nearest neighboring points for every center point. This forms m localized patches or sub-clouds denoted as $\mathcal{P}_u \in \mathbb{R}^{m \times k \times 3}$. To ensure these patches are free of bias, we normalize them by subtracting their respective center coordinates. This operation effectively disentangles the structural patterns from the spatial coordinates inherent to each local patch.

Point Cloud Embedding. Following Point-BERT [63], we employ a mini-PointNet [35] to embed the point patches as a sequence of point embedding $\{h_i\}_{i=1}^m$. For the further vector quantization, we pre-define a learnable codebook $\mathcal{V} := \{(s, z_s)\}_{s \in \mathcal{S}}^{d_z}$, where d_z is the dimension of codes, \mathcal{S} is the size of codebook and s is the index of embedding in \mathcal{V} . We adopt DGCNN [53] as tokenizer $\mathcal{Q} : h_i \mapsto z_i$, which maps $\{h_i\}_{i=1}^m$ into $\{z_i\}_{i=1}^m$ in \mathcal{V} . However, given the non-differentiable nature of the discrete tokens, applying reparameterization gradients for dVAE training becomes unfeasible. As suggested in [41], we resort to the Gumbel-softmax relaxation technique [18], coupled with a uniform prior, as a workaround during the dVAE training process.

Point Cloud Reconstruction. Given $\{z_i\}_{i=1}^m$ as input, the decoder $\mathcal{D}(\cdot)$ is tasked with the reconstruction of the entire point cloud. Inspired by [63], to effectively capture inter-

point relationships and bolster the representational capacity of discrete point tokens across various local structures, we adopt the DGCNN [53]-FoldingNet [60] architecture to reconstruct the whole point cloud.

The overarching reconstruction objective can be denoted as $\mathbb{E}_{z \sim \mathcal{Q}(z|h)} \log D(\hat{p}|z)$. Inspired by Point-BERT [63], the entire reconstruction process can be conceptualized as the maximization of the Evidence Lower Bound (ELB) for the log-likelihood, denoted as $\mathcal{D}(\hat{p}|p)$:

$$\sum_{i=1}^{\mathcal{N}} \log D(\hat{p}_i|p_i) \geq \sum_{i=1}^{\mathcal{N}} (\mathbb{E}_{z_i \sim \mathcal{Q}(z_i|p_i)}) [\log D(p_i|z_i)] - \text{KL}[\mathcal{Q}(z|p_i), \mathcal{D}(z|\hat{p}_i)]. \quad (1)$$

Simultaneously, we account for an intuitive reconstruction loss, leveraging the $l1$ Chamfer Distance [11] to supervise prediction point cloud with the ground-truth point cloud:

$$\mathcal{L}_{\text{CD}}(\mathcal{P}, \mathcal{G}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \min_{g \in \mathcal{G}} \|p - g\| + \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \min_{p \in \mathcal{P}} \|g - p\|, \quad (2)$$

where \mathcal{P} denotes the predicted point clouds and \mathcal{G} is the ground truth point clouds. Additional, we follow [41], optimize the KL-divergence \mathcal{L}_{KL} between the generated point cloud distribution and a uniform distribution:

$$\mathcal{L}_{\text{dVAE}} = \mathcal{L}_{\text{CD}} + \mathcal{L}_{\text{KL}}. \quad (3)$$

3.2. Stage 2: GPM Pre-training

To ensure GPM excel in both point cloud representation learning and point cloud generation task, we formulate a training framework that amalgamates autoencoding and autoregressive techniques motivated by [10].

Autoencoding Masked Sequence Generation. We employ the standard Transformers [51], which encompass multi-headed self-attention layers and FFN blocks. Initially, we partition each input point cloud into m local patches with designated center points. Subsequently, these local patches undergo discretization into a code sequence $\{f_i\}_{i=1}^m$ via our pre-trained dVAE in stage 1. Formally, the input embeddings $\{e_i\}_{i=1}^m$ are constructed as a combination of point embeddings $\{f_i\}_{i=1}^m$ and positional embeddings $\{pos_i\}_{i=1}^m$. Adhering to language model [20], a class token [CLS] is appended to the input sequences, rendering the transformer input sequence as $I = \{[\text{CLS}], e_1^0, e_2^0, \dots, e_i^0\}$.

Drawing inspiration from [63], we initiate by selecting a central point \mathcal{C}_i alongside its corresponding sub-cloud \mathcal{P}_i , integrate it with m neighboring sub-clouds as a coherent local region. Within this region, we apply a masking operation to obscure all local patches, generating what we refer to as the "masked point cloud". Specifically, we substitute all the masked point embeddings with a universally learnable

pre-defined mask embeddings [M]. We mark the masked position as $\mathcal{M} = \{1, \dots, b\}$, and the final input embeddings $E^{\mathcal{M}} = \{e_i, i \notin \mathcal{M}\} \cup \{[\text{M}] + pos_i, i \in \mathcal{M}\}$ are fed into the transformer.

The objective of our Masked Point Modeling (MPM) task is to deduce the geometric arrangement of absent portions and reconstruct point tokens aligned with the masked positions using the available information. Geometric information refers to the relative positions and morphological attributes of points within a localized region. These crucial geometric details provide vital cues for understanding and processing the shape and structure of the local region, playing a pivotal role in delving deep into the intricacies of point cloud data. Formally, the pre-training objective can be expressed as maximizing the log-likelihood of the point tokens e_i conditioned on the masked input embeddings $E^{\mathcal{M}}$:

$$\max_{E \in W} \sum \mathbb{E} \left[\sum_{i \in \mathcal{M}} \log P(e_i|E^{\mathcal{M}}) \right], \quad (4)$$

where W is the set of all input embeddings. Similar to [63], we also adopt Point Patch Mixing and contrastive learning to help the model to better learn high-level semantics. Point Patch Mixing involves mixing sub-clouds without complex alignment techniques. However, training solely with masked precision has limitations in understanding high-level semantics. To overcome this, we incorporate contrastive learning to enhance semantic understanding. Assuming the masking ratio is β , the loss function is defined as:

$$-\beta \log \frac{\exp(q\hat{k}_1/\tau)}{\sum_{i=0}^K \exp(qk_i)} - (1-\beta) \log \frac{\exp(q\hat{k}_2/\tau)}{\sum_{i=0}^K \exp(qk_i)}, \quad (5)$$

where q is the feature of a mixed sample that comes from two other samples, whose features are \hat{k}_1 and \hat{k}_2 . k_i are extracted by the momentum feature encoder, τ is the temperature and K is the size of memory bank.

Autoregressive Sequence Generation. While Point-BERT models the relationship between unmasked and masked regions using the Masked Point Modeling (MPM) task, it does not adequately model the interactions within masked regions. Furthermore, our aim extends beyond acquiring high-quality point cloud representations; we also aspire to proficiently execute point cloud generation tasks. Hence, for the latter segment, PartB, we engage in autoregressive mask generation tasks to enhance the MPM task.

In the case of PartB, the mask operation is ignored. Our focus lies in acquiring the generated tokens from the masked positions of PartA. We take the initial $n - 1$ tokens from PartA, appending the start token [S] at its forefront to get the PartB $E^{\mathcal{B}} = \{[\text{S}], e_1, \dots, e_{n-1}\}$. Following the autoregressive approach, the prediction of the next token relies

on the preceding token:

$$\max_{\theta} \mathbb{E} \left[\sum_{i=1}^n \log P_{\theta}(e_i | \text{PartA}, e_{<i}, [S]) \right]. \quad (6)$$

Although no masking operation is performed on PartB, since the mask positions of PartA are multiple consecutive sub-clouds instead of a single one, in PartB we are also equivalent to autoregressive generation of a continuous mask span. We extract the generated tokens from PartB, corresponding to the masked positions in PartA, and compare them with the discrete tokens produced by the Tokenizer:

$$\mathcal{L}_{AR} = \text{CrossEntropy}(e_{j,j \in \mathcal{M}}, \text{argmax} \mathcal{Q}(h)_{j,j \in \mathcal{M}}). \quad (7)$$

In this way, it allows our model to acquire proficiency in an autoencoding bidirectional encoding scheme for PartA, and an autoregressive unidirectional encoder for PartB.

3.3. Multi-task Pretraining

The masking of 15% of tokens in BERT is tailored for downstream natural language understanding tasks, whereas Point-BERT masks 25% ~ 45% to acquire enhanced point cloud representations. Our GPM focus on both point cloud representation learning and point cloud generation, we concatenate PartA and PartB as input to the transformer, aiming to simultaneously perform both autoencoding and autoregressive tasks. Therefore, we need to establish specific attention masks to facilitate the interaction of information between these two segments.

Motivated by [10], the tokens in PartA can participate in MLM task in PartA and autoregressive generation in PartB. PartB tokens can not be observed by PartA, but those antecedents. PartA is treated as the conditioning for autoregressive generation in PartB, akin to prefix tuning [23] in NLP. The implement of GPM multi-task pretraining is depicted in Figure 3.

The tokens output by the pretrained tokenizer $\mathcal{Q}(\cdot)$ are regarded as labels. PartA conducts a masked prediction task in an autoencoding manner, and the prediction loss between the predicted masked portion and the labels is computed as follows:

$$\mathcal{L}_{AE} = \text{CrossEntropy}(e_{j,j \in \mathcal{M}}, \text{argmax} \mathcal{Q}(h)_{j,j \in \mathcal{M}}). \quad (8)$$

Although Equation 7 and 8 exhibit identical formal structures. However, one entails mask prediction through autoencoding, while the other involves mask generation in an autoregressive way, indicating a fundamental distinction between the two approaches.

4. Experiments

In this section, we commence by outlining the configurations of our pretraining scheme. Subsequently, we pro-

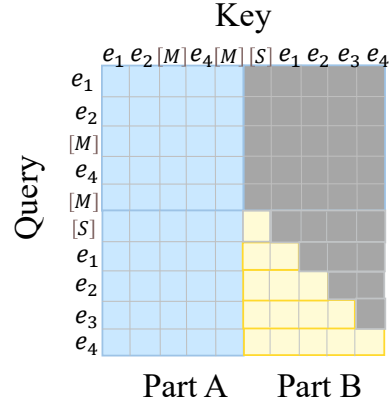


Figure 3. Self-attention mask. The regions in grey are masked. Tokens in PartA can attend to themselves, but not to PartB. Tokens in PartB can attend to PartA and their antecedents in PartB.

ceed to assess the proposed model through various downstream tasks, encompassing object classification, partial segmentation, few-shot learning, transfer learning (to validate the model’s representation learning capabilities), and point cloud generation (to validate its generative capabilities). Additionally, we conduct ablation studies on GPM.

4.1. Pre-training Setups and Implementation

Pre-training Data. Following the dataset configuration similar to Point-BERT [63], we employ ShapeNet [5] as our pretraining dataset, encompassing over 50,000 unique 3D models spanning 55 common object categories. From each 3D model, we sample 1024 points, dividing them into 64 point patches (sub-clouds) and each sub-cloud contains 32 points. Utilizing a lightweight PointNet [35] with two MLP layers, we project each sub-cloud into a 64-point embedding, serving as input for both dVAE and Transformer [51].

dVAE Pretraining Setups. The purpose of pretraining the dVAE is to acquire a high-quality tokenizer, enabling the reconstruction of discretized features back into the original point cloud to the fullest extent. Our dVAE is composed of a tokenizer and a decoder. To be specific, the tokenizer consists of a 4-layer DGCNN [53], while the decoder encompasses a 4-layer DGCNN followed by a FoldingNet [60]. Additionally, the FoldingLayer establishes a connection between a 2D grid and the input, ultimately generating a 3D point cloud.

During the training of dVAE [41], we set the number of words in the codebook N to 8192. We employ the common l_1 -style Chamfer Distance loss in the reconstruction process. Due to the small numerical value of this l_1 loss, the weight of the KL loss in Equation 1 must be smaller than that in image tasks. In the initial 10,000 steps of training,

the weight of the KL loss is kept at 0 and gradually raised to 0.1 in the subsequent 100,000 steps. Our learning rate is set to 0.0005, and follows a cosine annealing schedule spanning 60,000 steps. Consistent with prior works [41] and Point-BERT [63], we decay the temperature in the Gumble-softmax function from 1 to 0.0625 over 100,000 steps. The training of dVAE encompasses a total of 150,000 steps, with a batch size of 64. It’s worth noting that our dVAE’s network architecture closely mirrors that of Point-BERT.

GPM Pretraining Setups. In our experiments, we adhere to the standard Transformer architecture [9], which consists of a stack of Transformer blocks [51], each composed of multi-head self-attention layers and feed-forward networks. LayerNorm is applied in both layers. We set the depth of our Transformer to 12, feature dimension to 384, and the number of heads to 6. A random depth of 0.1 [17] is applied in our Transformer encoder. During GPM pretraining, the weights of the Tokenizer learned by dVAE are kept fixed. Input point embeddings from 25% ~ 45% are randomly masked, and the model is trained to infer the expected point labels at these masked positions. Unlike Point-BERT [63], which does not have a generation task during pretraining, we incorporate both autoregressive and autoencoding tasks in our work. Thus, an attention mask is necessary, as illustrated in Figure 3. For MoCo, we set the same configuration as [63]. More Details can be found in the supplementary 6.1 and 6.2.

4.2. Downstream Tasks

4.2.1 Point Cloud Representation Evaluation

Object classification on ModelNet40. We evaluate our pre-trained model on the ModelNet40 dataset [55], comprising 12,311 clean 3D CAD models spanning 40 categories. Following the Point-BERT [63] setups, we employ a two-layer MLP with a dropout rate of 0.5 as the classification head for the task. We optimize the model using AdamW with a weight decay of 0.05 and a learning rate of 0.0005, while employing a batch size of 32 and a cosine annealing schedule. We conduct comparisons with various Transformer-based models with identical Point-BERT settings, denoting [ST] for the standard Transformer architecture and [T] for Transformer models with specific designs or inductive biases. The results in Table 1 demonstrate that our model not only outperforms Point-BERT in classification metrics on this dataset but also exhibits autoregressive generation capabilities.

Object classification on ScanObjectNN. The practical applicability of our model on real datasets serves as a crucial metric. Therefore, the pre-trained models are transferred to the ScanObjectNN dataset [50], which comprises

Methods	number of points	Acc
<i>Supervised Learning</i>		
PointNet[35]	1k	89.2
SO-Net[22]	1k	92.5
PointNet++ [36]	1k	90.5
PointCNN[24]	1k	92.2
DGCNN[53]	1k	92.9
MVTN[14]	1k	93.8
RSCNN [43]	1k	92.9
GBNet[38]	1k	93.8
PointMLP[32]	1k	94.5
DensePoint[28]	1k	92.8
PointNeXt[37]	1k	94.0
P2P-RN101[54]	1k	93.1
P2P-HorNet[54]	1k	94.0
KPConv[48]	~6.8k	92.9
<i>Self-Supervised Representation Learning</i>		
[T]PCT [65]	1k	93.2
[T]PointTransformer [13]	-	93.7
[ST]NPCT [13]	1k	91.0
[ST]Transformer[51]	1k	91.4
[ST]Transformer-OcCo[51]	1k	92.1
[ST]Point-BERT[63]	1k	93.2
[ST]MaskPoint[27]	1k	93.8
[ST]Point-MAE[34]	1k	93.8
[ST]GPM	1k	94.0
<i>Methods with Cross-modal Information and Teacher Models</i>		
ACT[8]	1k	93.7
[ST]Transformer[51]	4k	91.2
[ST]Transformer-OcCo[51]	4k	92.2
[ST]Point-BERT[63]	4k	93.4
[ST]GPM	4k	93.1
[ST]Point-BERT[63]	8k	93.8
[ST]MaskPoint[27]	8k	-
[ST]Point-MAE[34]	8k	94.0
[ST]GPM	8k	94.3

Table 1. Classification results on ModelNet40. All results are expressed as percentages accuracy.

approximately 15,000 objects extracted from real indoor scans, encompassing 2902 point clouds from 15 categories. This dataset presents a greater challenge as it involves sampling from real-world scans with backgrounds and occlusions. We conducted experiments on three main variants, namely OBJ-BG, OBJ-ONLY, and PB-T50-RS, in line with prior work. The experimental results are summarized in Table 2. It is observed that GPM exhibits an improvement of approximately 2.77%, 1.88%, and 1.73% over the regular Point-BERT across the three variants.

Few-shot learning. To demonstrate the ability to acquire knowledge for new tasks with limited training data, we evaluate our model under the setting of few-shot learning, following the methodology of previous work [47, 63]. In the typical ‘ W -way S -shot’ setup, we initially randomly select W classes and then sample $(S+20)$ objects for each class [46]. The model is trained on $W \times S$ samples (support set) and evaluated on the remaining $20W$ samples (query set). We conduct 10 independent experiments for each setting and report the average performance and standard deviation across the 10 runs. As shown in Table 3, our approach outperforms other methods in all tests, achieving absolute improvements of 0.4%, 0.4%, and 0.6% over Point-BERT.

Methods	OBJ-BG	OBJ-ONLY	PB-T50-RS
Supervised Learning			
PointNet[35]	73.3	79.2	68.0
SpiderCNN[58]	77.1	79.5	73.7
PointNet++ [36]	82.3	84.3	77.9
PointCNN[24]	86.1	85.5	78.5
DGCNN[53]	82.8	86.2	78.1
MVTN[14]	92.6	92.3	82.8
BGA-DGCNN[50]	-	-	79.7
BGA-PN++ [50]	-	-	80.2
GBNet[38]	-	-	81.0
PointMLP[32]	-	-	85.4
PointNeXt[37]	-	-	87.7
P2P-RN101[54]	-	-	87.4
P2P-HorNet[54]	-	-	89.3
Self-Supervised Representation Learning			
Transformer[51]	79.86	80.55	77.24
Transformer-OcCo[51]	84.85	85.54	78.79
Point-BERT[63]	87.43	88.12	83.07
MaskPoint[27]	89.3	88.1	84.3
Point-MAE[34]	90.0	88.2	85.2
GPM	90.2	90.0	84.8

Table 2. Classification results on ScanObjectNN. All results are expressed as percentages accuracy.

Methods	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Learning				
DGCNN-rand [53]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
OcCo [52]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
Self-Supervised Representation Learning				
Point-BERT [63]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
MaskPoint [27]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
Piont-MAE [34]	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
Piont-M2AE [64]	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
GPM	97.2 ± 2.6	98.7 ± 2.2	92.9 ± 4.2	95.0 ± 3.0

Table 3. The results of few-shot classification on the ModelNet40 dataset. For each experimental setting, we conduct 10 independent experiments and report the average accuracy (%) along with its standard deviation.

Part segmentation. We evaluate the representation learning capability of our approach on the ShapeNetPart dataset [62], aiming to predict more fine-grained class labels for each point. This dataset consists of 16 categories and comprises 16,881 objects. The point cloud is downsampled to 2048 points, and the segmentation head [34] connects features $\mathcal{F}^4, \mathcal{F}^8, \mathcal{F}^{12}$ extracted from the 4-th, 6-th, and 12-th layers of the transformer blocks. Subsequently, average pooling, max pooling, and upsampling are employed to generate features for each point, followed by label prediction using MLP. In our experiments, we evaluate the performance under the settings of "5 way 10 shot", "5 way 20 shot", "10 way 10 shot", and "10 way 20 shot". The experimental results shown in Table 4 demonstrate the superior performance of our GPM compared to all other methods.

4.2.2 Point Cloud Generation Evaluation

In the pre-training phase, we perform unconditional autoregressive generation on tokens in PartB, endowing the model with the capability of point cloud unconditional generation.

Methods	Cls.mIoU	Inst.mIoU
Supervised Learning		
PointNet[35]	80.39	83.7
PointNet++ [36]	81.35	85.1
DGCNN[53]	82.33	85.2
PointMLP [32]	84.6	86.1
Self-Supervised Representation Learning		
Transformer[51]	83.42	85.1
Transformer-OcCo[51]	83.42	85.1
PointContrast [56]	-	85.1
CrossPoint [2]	-	85.5
Point-BERT [63]	84.11	85.6
Point-MAE [34]	-	86.1
GPM	84.20	85.8

Table 4. Part segmentation results on the ShapeNetPart dataset. We report the average intersection mIoU over the union of all classes (Cls.) and instances (Inst.).

Category	Model	JSD (\downarrow)		MMD (\downarrow)		Cov (% \uparrow)		1-NNA % (% \downarrow)	
		CD	EMD	CD	EMD	CD	EMD	CD	EMD
Airplane	Point-Flow	4.92	0.217	3.24	46.91	48.40	75.68	75.06	
	Point-BERT	4.07	0.189	2.75	50.17	53.44	70.16	61.07	
	GPM	3.32	0.176	2.47	52.34	56.44	68.22	60.30	
Chair	Point-Flow	1.74	2.42	7.87	46.83	46.98	60.88	59.89	
	Point-BERT	1.23	1.82	7.11	49.73	50.07	57.30	56.61	
	GPM	0.98	1.67	6.82	51.08	52.14	54.09	54.27	
Car	Point-Flow	0.87	0.91	5.22	44.03	46.59	60.65	62.36	
	Point-BERT	0.73	0.82	4.96	47.27	52.13	59.09	60.48	
	GPM	0.66	0.75	4.38	49.67	54.17	58.88	58.16	

Table 5. Generation results on ShapeNet dataset. \uparrow : the higher the better, \downarrow : the lower the better. MMD-CD is multiplied by 10^3 ; MMD-EMD and JSD are multiplied by 10^2 .

Therefore, in downstream tasks, without fine-tuning, we conduct an autoregressive point cloud generation task on the ShapeNet dataset. Existing models do not possess both point cloud generation and point cloud classification tasks. Hence, our model is the first known point cloud transformer that integrates these two tasks into one. The specific metrics are shown in Table 5. We compare our results with generation model Point-Flow [59] and Point-BERT. More details about downstream tasks and conditional point cloud generation results can be found in the supplementary 6.3.

4.3. Ablation Study

To validate the fundamental design of the GPM model, we conducted a comprehensive set of ablation studies. In order to assess the impact of autoregressive mask generation on the autoencoding masked prediction task, we present results comparing fine-tuning scenarios on the ModelNet40 dataset: one with *autoencoding* only, and the other with a combination of *autoencoding + autoregressive*. Furthermore, to demonstrate that the order of autoencoding and autoregressive tasks in the sequence does not affect point cloud understanding, we conduct additional experiments, the results are summarized in the Table 6. More ablation results are shown in the supplementary 6.4.

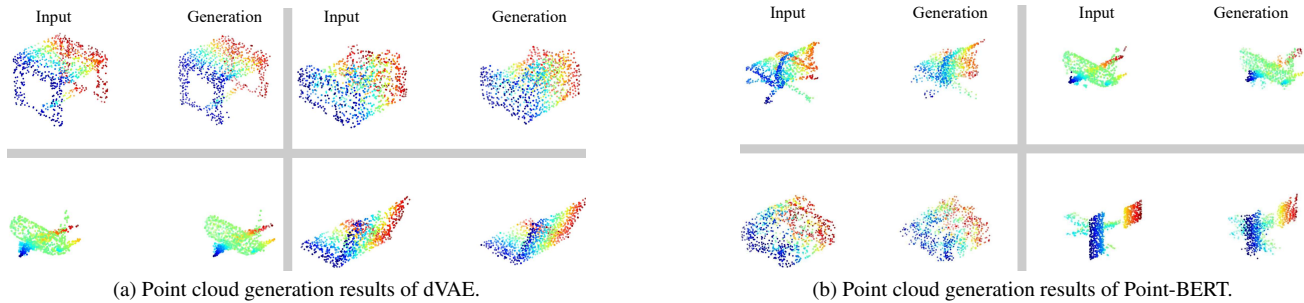


Figure 4. The reconstruction results visualization on dVAE and Point-BERT.

Tasks	npoints	Acc
autoencoding only	1k	92.9
autoregressive \Rightarrow autoencoding	1k	93.6
autoencoding \Rightarrow autoregressive	1k	93.8
autoencoding only	4k	93.4
autoregressive \Rightarrow autoencoding	4k	93.9
autoencoding \Rightarrow autoregressive	4k	93.9

Table 6. Results on autoencoding, autoencoding \Rightarrow autoregressive and autoregressive \Rightarrow autoencoding on the ModelNet dataset. It is clear that autoregressive can enhance the completion accuracy of autoencoding, the order of autoencoding and autoregressive does not have a significant impact on point cloud understanding tasks.

4.4. Analysis

Compare with Point-BERT. Motivated by BERT, Point-BERT leverages MPM as a pretext task for training. Due to the assumption of independence between MLM and MPM, Point-BERT fails to capture interdependencies among masked tokens. In our approach, autoregressive generation explicitly models dependencies between masked regions, unlike autoencoding which relies solely on unmasked tokens to predict the masked ones. Autoregressive masked prediction allows previously predicted tokens to influence subsequent predictions, enhancing information exchange between them. This leads to improved performance in downstream tasks in the realm of representation learning. Simultaneously, tokens in PartA can attend to task in PartB, making PartA a generation condition for PartB, resulting in enhanced generation performance.

Compare with PointGPT. PointGPT [6] adopts a two-stage transformer architecture for learning representations of point clouds, comprising a feature extractor and a generator. It first extracts features from the point cloud us-

ing the feature extractor and then directly generates with the obtained features in the generator. During training, it directly employs Chamfer Distance [11] between generated point clouds and ground truth points as the optimization objective, rather than operating on tokens. This is because it doesn’t discretize points embeddings, thus failing to achieve genuine autoregressive generation tasks. Moreover, it doesn’t integrate the two tasks for joint training in a single transformer; instead, it trains them separately on two different transformers, which increases the complexity.

4.5. Visualization

We visualize the generation results of dVAE and Point-BERT, as shown in the Figure 4a and Figure 4b. Although Point-BERT is capable of masked tokens generation through autoencoding, our approach allows for autoregressive **conditional** generation tasks, such as text-conditioned point clouds and image-conditioned point clouds. As observed in Figure 4b, the generation results of Point-BERT exhibit some scattered points at the edges of the overall structure, appearing relatively sparse and disorganized. The generated outcome is not as satisfactory as that of GPM. More visualizations of generation results can refer to *Unconditional and Conditional Point Generation* in the supplementary 6.3.

5. Conclusion

Our study integrates both autoencoding and autoregressive tasks within a unified transformer to enhance the model’s performance in representation learning and generation at the same time. By this framework, we address the limitation of current pretraining frameworks and make the proxy tasks non-trivial, effectively enhance the information interaction among masked tokens. Moreover, our framework endows the model with the potential for conditional generation.

Acknowledgement

This work was supported by National Science and Technology Major Project (No.2022ZD0115101), National Natural Science Foundation of China Project (No.U21A20427).

References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2
- [2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. 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*, pages 9902–9912, 2022. 7
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 1
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [6] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *arXiv preprint arXiv:2305.11487*, 2023. 8
- [7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 1, 2
- [8] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *The Eleventh International Conference on Learning Representations*, 2022. 2, 6
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 3, 6
- [10] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 2, 4, 5
- [11] Haoqiang Fan, Hao Su, and Leonidas J Guibas. 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*, pages 605–613, 2017. 4, 8
- [12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL), 2021. 1
- [13] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 1, 6
- [14] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2021. 6, 7
- [15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2
- [17] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 6
- [18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2016. 3
- [19] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77, 2020. 1
- [20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 4
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1
- [22] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018. 2, 6
- [23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 5
- [24] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 6, 7
- [25] Zhe Li, T. Yang Laurence, Xin Nie, BoCheng Ren, and Xianjun Deng. Enhancing sentence representation with visually-supervised multimodal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM’23)*, 2023. 1, 2

- [26] Zhe Li, Laurence T Yang, Bocheng Ren, Xin Nie, Zhangyang Gao, Cheng Tan, and Stan Z Li. Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. *arXiv preprint arXiv:2402.02045*, 2024. [1](#)
- [27] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. [6](#), [7](#)
- [28] Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, and Chunhong Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5239–5248, 2019. [6](#)
- [29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. [1](#)
- [30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. [1](#)
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [32] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *International Conference on Learning Representations*, 2021. [6](#), [7](#)
- [33] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. [2](#)
- [34] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022. [2](#), [6](#), [7](#)
- [35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. [3](#), [5](#), [6](#), [7](#)
- [36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. [2](#), [6](#), [7](#)
- [37] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022. [6](#), [7](#)
- [38] Shi Qiu, Saeed Anwar, and Nick Barnes. Geometric back-projection network for point cloud classification. *IEEE Transactions on Multimedia*, 24:1943–1955, 2021. [6](#), [7](#)
- [39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. [1](#), [2](#)
- [40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. [1](#)
- [41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. [2](#), [3](#), [4](#), [5](#), [6](#), [1](#)
- [42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. [2](#)
- [43] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5376–5385, 2020. [6](#)
- [44] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021. [2](#)
- [45] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019. [2](#)
- [46] Charu Sharma and Manohar Kaul. Self-supervised few-shot learning on point clouds. *Advances in Neural Information Processing Systems*, 33:7212–7221, 2020. [6](#)
- [47] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. [6](#)
- [48] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. [6](#)
- [49] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. [1](#), [2](#)
- [50] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. [6](#), [7](#), [2](#)
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [4](#), [5](#), [6](#), [7](#), [2](#)
- [52] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021. [7](#)

- [53] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. [3](#), [4](#), [5](#), [6](#), [7](#), [1](#)
- [54] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in neural information processing systems*, 35:14388–14402, 2022. [6](#), [7](#)
- [55] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. [6](#), [2](#)
- [56] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. 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*, pages 574–591. Springer, 2020. [7](#)
- [57] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021. [1](#)
- [58] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European conference on computer vision (ECCV)*, pages 87–102, 2018. [7](#)
- [59] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. [7](#)
- [60] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. [4](#), [5](#), [1](#)
- [61] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. [1](#)
- [62] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. [7](#)
- [63] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [64] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022. [2](#), [7](#)
- [65] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. [1](#), [6](#)
- [66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. [1](#)