Semantic-Free Procedural 3D Shapes Are Surprisingly Good Teachers

Xuweiyi Chen Zezhou Cheng
University of Virginia
https://point-mae-zero.cs.virginia.edu/

Abstract

Self-supervised learning has emerged as a promising approach for acquiring transferable 3D representations from unlabeled 3D point clouds. Unlike 2D images, which are widely accessible, acquiring 3D assets requires specialized expertise or professional 3D scanning equipment, making it difficult to scale and raising copyright concerns. To address these challenges, we propose learning 3D representations from procedural 3D programs that automatically generate 3D shapes using simple 3D primitives and augmentations.

Remarkably, despite lacking semantic content, the 3D representations learned from the procedurally generated 3D shapes perform on par with state-of-the-art representations learned from semantically recognizable 3D models (e.g., airplanes) across various downstream 3D tasks, such as shape classification, part segmentation, masked point cloud completion, and both scene semantic and instance segmentation. We provide a detailed analysis on factors that make a good 3D procedural programs. Extensive experiments further suggest that current 3D self-supervised learning methods on point clouds do not rely on semantics of 3D shapes, shedding light on the nature of 3D representations learned.

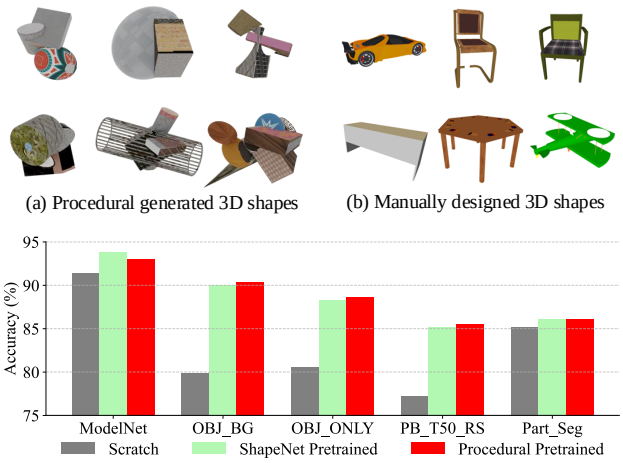


Figure 1. Self-supervised learning from **(a)** procedurally generated 3D shapes [22, 48, 52, 53] performs comparably to learning from **(b)** ShapeNet models that are semantically meaningful [8] across various downstream 3D understanding tasks. Both outperforms training from scratch significantly. In **(c)**, the x-axis represents various tasks and benchmarks: ModelNet40 [27] and three variants of ScanObjectNN [37] for shape classification, and ShapeNetPart [56] for part segmentation.

1. Introduction

Self-supervised learning (SSL) aims at learning representations from unlabeled data that can transfer effectively to various downstream tasks. Inspired by the success of SSL in language [14] and 2D images [19, 20], SSL for 3D point cloud understanding has gained considerable interest [29, 40, 57]. Recently, Point-MAE [29] and its follow-ups [44, 45, 61, 63] exploit the masked autoencoding scheme [20] for 3D point cloud representation learning, showing substantial improvements in various 3D shape understanding tasks (*e.g.*, shape classification, part segmentation, and scene instance segmentation).

However, unlike language and image data, which are abundantly available on the Internet, 3D assets are less accessible, as their creation often requires domain expertise and specialized tools such as 3D modeling software (*e.g.*, Blender) or scanning equipment (*e.g.* LiDAR sensors). This scarcity of 3D shapes, often referred to as the 3D data desert [15], has significantly hindered the scalability of representation learning methods. Recent efforts have expanded point cloud datasets at both the object level [12, 13] and the scene level [4, 16, 55], but often rely on substantial human effort. Nevertheless, challenges unique to 3D data collection—such as copyright concerns, diverse file formats, and limited scalability—remain largely unresolved.

A common belief in 3D representation learning is that strong representations require semantically meaningful 3D shapes—objects such as chairs, airplanes, or indoor scenes—which are inherently costly to curate and difficult to scale. In this work, we challenge this assumption by asking a central question: ***Do we really need semantically meaningful 3D shapes to learn strong 3D representations?***

To explore this, we investigate learning point cloud rep-

resentations from *purely synthetic data* generated by procedural 3D programs [22, 48, 52, 53], as illustrated in Fig. 1a. Our data generation pipeline begins by sampling simple 3D primitives (*e.g.*, cubes, cylinders, spheres), which are then transformed via affine operations (*e.g.*, scaling, translation, rotation) and composed into more complex geometries. We further apply augmentations such as Boolean operations to enrich topological diversity, followed by uniform surface point sampling to obtain point clouds suitable for representation learning. This approach is lightweight, efficient, and capable of generating unlimited number of 3D shapes with diverse geometric structures. Unlike standard datasets, our procedural shapes lack human-recognizable semantics.

We then pose a follow-up question: ***Are current 3D self-supervised learning (SSL) methods capable of leveraging such non-semantic, procedural shapes?*** To answer this, we generate 150K object-level and 4K scene-level synthetic 3D point clouds using our procedural programs, at a total cost of approximately 1400 CPU hours. Notably, the scale of our dataset exceeds that of widely used benchmarks such as ShapeNet (51K publicly available shapes) and ScanNet (1,513 indoor scenes). We benchmark multiple representative 3D SSL methods, including Point-MAE [29] and its recent variants [44, 61, 63], and evaluate them across a wide variety of downstream tasks such as shape classification, part segmentation, masked point cloud completion, and scene-level semantic and instance segmentation.

Our main findings are as follows:

- **Semantic-free procedural 3D shapes are surprisingly good teachers.** Despite lacking semantic content, self-supervised models trained solely on synthetic data perform on par with counterparts trained on real 3D datasets such as ShapeNet [8], Objaverse [12], and ScanNet [11]. Moreover, they significantly outperform models trained from scratch without any pretraining (Fig. 1c, Tabs. 1–5).
- **Geometric diversity plays a key role in learning effective 3D representations.** We provide detailed insights into the factors that influence the quality of procedurally generated 3D datasets. Our analysis shows that learning performance improves significantly with increased geometric diversity and larger dataset size (Tab. 4, Fig. 5).
- **Current 3D SSL methods rely more on geometric cues rather than semantic content.** Our in-depth analysis reveals strong structural similarities between representations learned from semantic-free synthetic procedural shapes and those learned from semantically meaningful 3D models (see t-SNE visualization in Fig. 7).

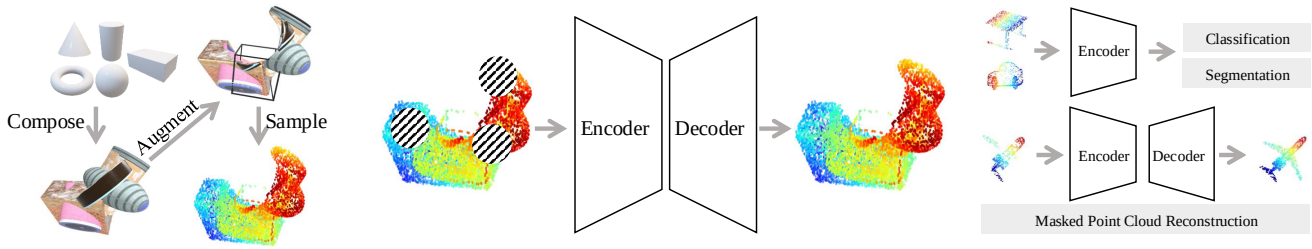
To our best knowledge, this is the first systematic large-scale study on 3D SSLs from procedural 3D shapes. Our work is inspired by recent works that successfully train large 3D reconstruction models exclusively on procedurally generated shapes [22, 48]. Our exploration is also closely related to prior efforts that learn image or video representations from procedural programs [2, 3, 58]. While recent efforts focus on scaling 3D datasets using human-designed models or 3D scans [12, 13, 55], our approach is orthogonal and complementary, leveraging procedurally generated data to bypass manual design and scanning altogether.

2. Related Work

3D Datasets. Significant efforts have been made to curate extensive 3D shape datasets [8, 10, 12, 16, 28, 33, 35, 42, 46]. For example, ShapeNet provides 3 million CAD models, with 51K high-quality shapes publicly available. More recently, Objaverse [12] and Objaverse-XL [13] expanded the 3D dataset to 10.2 million manually-created 3D shapes. However, challenges, such as format diversity and copyright and legal issues, remain unsolved. At the scene level, most commonly used datasets—ScanNet [11], Structured3D [65], Matterport3D [7], nuScenes [6], SemanticKITTI [5], and Waymo [34]—are distributed under non-commercial or research-only licenses, limiting their applicability for broader use. A more recent effort, ASE [1], introduces 100,000 synthetic indoor scenes, also under a non-commercial license. In contrast, we explore procedural 3D programs that generate shapes and scenes from simple primitives, enabling the creation of unlimited synthetic objects and scenes without licensing constraints.

Learning from Synthetic Data. Synthetic data has become popular in computer vision, especially in scenarios where ground-truth annotations are difficult to obtain or where privacy and copyright issues arise. State-of-the-art performance in mid-level or 3D vision tasks is often achieved through training on synthetic data, including tasks like optical flow [36], depth estimation [54], dense tracking [23], relighting [52], novel view synthesis [53], and material estimation [25]. Procedurally generated synthetic data has also been explored for self-supervised representation learning in images [2, 3] and videos [58], and more recently for multi-view feed-forward 3D reconstruction [48]. In this work, we explore self-supervised representation learning for point clouds, using synthetic 3D shapes generated by procedural 3D programs.

Self-supervised Learning for Point Clouds. Recent self-supervised learning (SSL) methods for point clouds generally fall into two categories: contrastive learning [9, 17, 19, 49] and masked autoencoding [15, 18, 29, 32, 40, 57, 59, 60, 62]. Contrastive approaches such as PointContrast [49] and DepthContrast [9] rely on instance discrimination [19] to learn view-invariant features. Inspired by the success of masked modeling in vision [20] and language [14], Point-BERT [57] and Point-MAE [29] use transformer-based architectures to predict masked regions of the point cloud. Point-M2AE [61] extends Point-MAE with a multi-scale pyramid design, while PCP-MAE [63] addresses centroid leakage by adding centroid prediction as



(a) Procedural 3D generation (b) Learning 3D representations with masked autoencoding (c) Downstream tasks

Figure 2. **Learning from procedural 3D programs. (a)** Synthetic 3D point clouds are generated by sampling, compositing, and augmenting simple primitives using procedural 3D programs [48]. **(b)** We experiment with multiple state-of-the-art self-supervised learning frameworks for learning 3D representations from synthetic data. Here, we illustrate the pretraining pipeline using Point-MAE [57], naming this variant *Point-MAE-Zero*, where "Zero" emphasizes the absence of any human-made 3D shapes. **(c)** We evaluate the pretrained models across various 3D shape understanding tasks.

an auxiliary objective. Scene-level SSL methods that operate directly on large-scale point clouds have also gained traction [21, 26, 39, 44, 50]. In this work, we adopt Point-MAE, Point-M2AE, PCP-MAE, and MSC as our primary SSL frameworks due to their strong performance on downstream 3D tasks. For ablation studies, we use Point-MAE as our baseline, as it represents a foundational and widely adopted approach in this line of research.

3. Learning from Procedural 3D Programs

We first introduce the procedural 3D programs [48, 52, 53] for generating unlimited number of synthetic 3D shapes using composition of simple primitive shapes (*e.g.*, cylinders) and shape augmentation (Sec. 3.1). We then describe the masked autoencoding scheme [29, 40, 57] for learning 3D representations from synthetic 3D datasets (Sec. 3.2).

3.1. Procedural 3D programs

There is a line of work synthesizing procedural 3D shapes for vision tasks such as novel view synthesis [53], relighting [52], and material estimation [25]. Following recent methods [48] that use procedural 3D shapes for sparse-view reconstruction, we address self-supervised 3D representation learning from purely synthetic datasets. Fig. 2a illustrates our data pipeline:

(1) Randomly sample $\mathbf{K}$ primitive shapes (cubes, spheres, cylinders, cones, tori) and apply affine transformations to combine them;

(2) Apply geometric augmentations (e.g., boolean differences, wireframe conversions) to enrich shape diversity (see [48] for details);

(3) Uniformly sample $\mathbf{N}$ surface points per synthesized shape as inputs for representation learning (Fig. 2b).

We experiment with various shape-generation configurations, such as changing the number of sampled primitives and applying augmentations. By default, each dataset consists of 150K shapes with $\mathbf{N} = 8192$ points each. Sec. 4 further analyzes the effects of dataset size and shape complexity on learned representations.

In order to generate procedural 3D scenes, we follow MegaSynth [22] to procedurally generate 4K synthetic 3D scenes. Specifically, we first generate a floor plan and generate procedural 3D shapes with the above pipeline and place procedural 3D shapes in the scene based on the generated floor plan. We provide details on the generation of 4K procedural 3D scenes and visualizations of the generated shapes in the supplementary material.

3.2. Procedural Pretraining

Pretraining. We adopt Point-MAE [20], Point-M2AE [61], PCP-MAE [63] to train on procedural 3D shapes (Fig. 2b). These methods rely on a masked autoencoding scheme [14, 20, 57], where the input point cloud is split into irregular patches and a large portion of them (60% by default) is randomly masked. A Transformer-based encoder–decoder network then attempts to reconstruct these masked patches, thereby learning 3D representations. The reconstruction loss is computed as the L_2 Chamfer Distance between the predicted point patches P_{pre} and the ground-truth patches P_{gt}:

$$L = \sum_{x \in \{P_{\text{pre}}, P_{\text{gt}}\}} \frac{1}{|x|} \sum_{a \in x} \min_{b \in x'} \|a - b\|_2^2 \qquad (1)$$

where $x' = P_{\text{gt}}$ if $x = P_{\text{pre}}$, and vice versa. For scene-level SSLs, we adopt MSC [44], which combines masked autoencoding and contrasive learning, to train on procedurally generated 3D scenes.

Downstream Probing. We evaluate baselines on several 3D tasks, as summarized in Fig. 2c. For shape classification, we augment the pretrained Transformer encoder with a three-layer MLP classification head. For part segmentation, we aggregate features from the 4th, 8th, and final layers of the encoder, upsample them to all 2048 input points,

Methods	ModelNet40	OBJ-BG	OBJ-ONLY	PB-T50-RS	Avg.
PointNet [30]	89.2	73.3	79.2	68.0	77.4
SpiderCNN [51]	92.4	77.1	79.5	73.7	80.7
PointNet++ [31]	90.7	82.3	84.3	77.9	83.8
DGCNN [41]	92.9	86.1	85.5	78.5	85.8
PointCNN [24]	–	86.1	85.5	78.5	–
PTv1 [64]	93.7	–	–	–	–
PTv2 [43]	94.2	–	–	–	–
OcCo [40]	92.1	84.9	85.5	78.8	85.3
Point-BERT [57]	93.2	87.4	88.1	83.1	88.0
Point-MAE-Scratch [57]	91.4	79.9	80.6	77.2	82.3
Point-MAE-SN [29]	**93.8**	90.0	88.3	85.2	89.3 (+7.0)
Point-MAE-Zero	93.0	**90.4**	**88.6**	**85.5**	89.4 (+7.1)
Point-M2AE-Scratch	92.2	90.0	87.6	85.6	88.9
Point-M2AE-SN [61]	**94.0**	**91.2**	88.8	86.4	90.1 (+1.2)
Point-M2AE-Zero	92.9	90.4	**89.8**	**87.0**	90.0 (+1.1)
PCP-MAE-Scratch	91.5	88.8	88.5	83.8	88.2
PCP-MAE-SN [63]	**94.0**	**95.5**	**94.3**	90.4	93.6 (+5.4)
PCP-MAE-Zero	92.4	94.0	92.3	**90.5**	92.3 (+4.1)

Table 1. **Object Classification.** We evaluate the object classification performance on ModelNet40 and three variants of ScanObjectNN. Classification accuracy (%) is reported (*higher is better*). **Top**: Performance of existing methods with various neural network architectures and pretraining strategies. **Bottom**: Comparison with our baseline methods. The rightmost column shows the average accuracy, with red text indicating the improvement over the corresponding Scratch baseline.

and employ a segmentation head. For masked point cloud reconstruction, we use both the pretrained encoder and decoder with no architectural modifications. For scene-level methods, we use both instance segmentation and semantic segmentation finetuned from the pretrained SSLs. Detailed implementation settings are in the supplementary material.

4. Experiments

We present a comprehensive evaluation of 3D shape representations pretrained with procedural 3D programs across various downstream object-level and scene-level tasks, including object classification, part segmentation, and 3D scene understanding (Sec. 4.1 – 4.4). We further provide an in-depth analysis of model behavior and ablation studies (Sec. 4.5). For each downstream task, we report the performance of relevant existing methods as a reference and focus on comparisons with SSLs pretrained on manually-curated 3D datasets, as well as models trained from scratch. Specifically, for object-level 3D understanding tasks, we evaluate the following three pretraining strategies: (1) *Scratch*: All network parameters are randomly initialized, with no pretraining. (2) *ShapeNet Pretrained (SN)*: Pretrained on 41,952 models in the ShapeNet [8] training split, relying on the officially released weights. (3) *Procedural 3D Programs Pretrained (Zero)*: Pretrained on 150K procedurally generated 3D models, using no human-crafted shapes. For scene-level tasks, we compare SSL models pretrained on ScanNet [11] and procedurally generated 3D scenes.

4.1. Object Classification

Benchmarks. We use ModelNet40 [47] and ScanObjectNN [37] as the benchmarks for the shape classification task. ModelNet40 contains 12,311 clean 3D CAD objects across 40 categories, with 9,843 samples for training and 2,468 for testing. Following Point-MAE, we apply random scaling and translation as data augmentation during training, and a voting strategy during testing [27]. Following prior works [29, 40, 57], we also evaluate the few-shot classification performance on ModelNet40. ScanObjectNN is a more complex real-world 3D dataset, consisting of approximately 15,000 objects across 15 categories, with items scanned from cluttered indoor scenes. We report results on three ScanObjectNN variants: OBJ-BG, OBJ-ONLY, and PB-T50-RS, the latter being the *most challenging* due to its additional noise and occlusions.

Transfer Learning. Table 1 summarizes object classification results across several settings. On ModelNet40, the "-Zero" variants (*e.g.*, Point-MAE-Zero, Point-M2AE-Zero, PCP-MAE-Zero) generally fall slightly behind their ShapeNet-pretrained counterparts ("-SN"), reflecting the larger domain gap between synthetic shapes and the clean 3D models in ModelNet40. By contrast, on ScanObjectNN—which contains real-world scans with broader geometric variability—the "-Zero" models often match or exceed the performance of their "-SN" counterparts. For instance, PCP-MAE-Zero outperforms PCP-MAE-SN on the PB-T50-RS variant, and Point-M2AE-Zero closely matches or exceeds Point-M2AE-SN in several cases. These findings indicate that the diverse geometry in procedurally synthesized data can be advantageous for certain real-world tasks. Meanwhile, all pretrained models (including both "-SN" and "-Zero") surpass their respective from-scratch baselines and outperform existing self-supervised approaches [40, 57] which we highlight with the rightmost column.

Few-shot Classification. We evaluate few-shot classification on ModelNet40 using standard n-way, m-shot protocols, where n denotes the number of randomly selected classes and m the number of examples per class. Each evaluation samples 20 unseen instances from each class. We repeat this procedure 10 times, reporting mean accuracy (%) and standard deviation. Table 2 presents results for $n = \{5, 10\}$ and $m = \{10, 20\}$. Similar to transfer learning experiments, Point-MAE-Zero performs on par or slightly below Point-MAE-SN, likely due to the larger domain gap between procedural shapes and ModelNet40 data. Nonetheless, both methods substantially outperform their scratch-trained counterparts, as reflected by the performance deltas, and also surpass prior approaches such as DGCNN [41] and Transformer-OcCo [40].

Methods	5w/10s	5w/20s	10w/10s	10w/20s	Avg.
DGCNN-rand [41]	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5	27.3
DGCNN-OcCo [41]	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2	88.1
Transformer-OcCo [40]	94.0±3.6	95.9±2.3	89.4±5.1	92.4±4.6	92.9
Point-BERT [57]	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1	93.7
Point-MAE-Scratch [57]	87.8±5.2	93.3±4.3	84.6±5.5	89.4±6.3	88.8
Point-MAE-SN [29]	96.3±2.5	**97.8±1.8**	**92.6±4.1**	95.0±3.0	95.4 (+6.6)
Point-MAE-Zero	**96.6±2.2**	97.6±1.4	91.9±4.2	**95.2±3.0**	95.3 (+6.5)
Point-M2AE-Scratch	87.5±2.6	90.0±5.5	86.4±3.2	89.6±4.3	88.4
Point-M2AE-SN* [29]	93.4±3.1	**96.2±1.5**	91.8±4.5	92.9±3.2	93.6 (+5.2)
Point-M2AE-Zero	**95.4±2.8**	94.4±2.8	**94.3±2.2**	**93.8±3.2**	94.5 (+6.1)
PCP-MAE-Scratch	86.4±2.6	85.0±6.0	88.9±4.1	90.7±4.2	87.8
PCP-MAE-SN [63]	**97.4±2.3**	**99.1±0.8**	93.5±3.7	**95.9±2.7**	96.5 (+8.7)
PCP-MAE-Zero	95.5±3.4	98.6±1.6	**94.2±3.5**	95.6±3.0	96.0 (+8.2)

Table 2. **Few-shot classification on ModelNet40.** We evaluate performance on four n-way, m-shot configurations. For example, 5w/10s denotes a 5-way, 10-shot classification task. The table reports the mean classification accuracy (%) and standard deviation across 10 independent runs for each configuration. **Top**: Results from existing methods for comparison. **Bottom**: Comparison with our baseline methods. Note that results for Point-M2AE-SN are reproduced using publicly available code, as the original configuration was not provided. The final column shows the average accuracy across configurations, with subscripts indicating improvements over the corresponding baseline.

4.2. Part Segmentation

The 3D part segmentation task aims to assign a part label to each point in a shape. We evaluate our methods and baselines on ShapeNetPart [56], which contains 16,881 models across 16 object categories. Consistent with previous works [29, 30, 57], we sample 2,048 points from each shape, resulting in 128 patches in our masked autoencoding pipeline (see Sec. 3).

Table 3 presents the mean Intersection-over-Union (mIoU) across all instances, along with per-category IoU. Across various models, both Point-MAE-Zero and Point-MAE-SN deliver comparable performance, indicating that procedurally generated shapes can learn robust 3D representations without explicit semantic content. Similarly, Point-M2AE-Zero and PCP-MAE-Zero achieve results on par with their ShapeNet-pretrained counterparts, further highlighting the versatility of procedural data in self-supervised representation learning.

In line with our observations in Sec. 4.1, the "-Zero" and "-SN" models surpass scratch-trained baselines and earlier methods that use different architectures [30, 31, 41] or alternative pretraining strategies [40, 57]. Despite lacking high-level semantic cues, these procedurally trained autoencoders still capture sufficient geometric structure to achieve strong segmentation performance.

4.3. Masked Point Cloud Completion

Masked point cloud completion reconstructs missing regions of 3D point clouds as a self-supervised pretext task for learning 3D representations [29] (see Fig. 2 and Sec. 3).

Methods	mIoU_I	aero	bag	cap	car	chair	earphone	guitar	knife
PointNet [30]	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9
PointNet++ [31]	85.1	82.4	79.0	87.7	77.3	90.8	71.8	91.0	85.9
DGCNN [41]	85.2	84.0	83.4	86.7	77.8	90.6	74.7	91.2	87.5
OcCo [40]	85.1	83.3	85.2	88.3	79.9	90.7	74.1	91.9	87.6
Point-BERT [57]	85.6	84.3	84.8	88.0	79.8	91.0	81.7	91.6	87.9
Point-MAE-Scratch [57]	85.1	82.9	**85.4**	87.7	78.8	90.5	**80.8**	91.1	**87.7**
Point-MAE-SN [29]	**86.1**(+1.0)	84.3	85.0	88.3	80.5	91.3	78.5	**92.1**	87.4
Point-MAE-Zero	**86.1**(+1.0)	**85.0**	84.2	**88.9**	**81.5**	**91.6**	76.9	**92.1**	87.6
Point-M2AE-Scratch	84.7	85.1	86.8	88.6	**81.1**	91.5	79.9	**92.1**	87.8
Point-M2AE-SN [61]	**85.0**(+0.3)	84.5	87.2	**89.3**	**81.1**	**91.8**	**80.1**	92.0	**89.2**
Point-M2AE-Zero	84.9(+0.2)	**85.3**	**87.3**	88.7	**81.1**	91.7	79.4	91.9	88.2
PCP-MAE-Scratch	83.8	84.3	83.1	88.7	80.3	91.2	77.1	92.0	**88.1**
PCP-MAE-SN [63]	84.3(+0.5)	**85.0**	84.0	**88.7**	81.0	**91.6**	77.6	91.8	87.6
PCP-MAE-Zero	**84.4**(+0.6)	84.6	**84.3**	88.5	**81.7**	91.5	**81.1**	**92.1**	87.0

Methods	lamp	laptop	motor	mug	pistol	rocket	skateboard	table
PointNet [30]	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
PointNet++ [31]	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
DGCNN [41]	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
OcCo [40]	84.7	95.4	75.5	94.4	84.1	63.1	75.7	80.8
Point-BERT [57]	85.2	95.6	75.6	94.7	84.3	63.4	76.3	81.5
Point-MAE-Scratch [57]	85.3	95.6	73.9	**94.9**	83.5	61.2	74.9	80.6
Point-MAE-SN [29]	**86.1**	**96.1**	75.2	94.6	84.7	63.5	77.1	**82.4**
Point-MAE-Zero	86.0	96.0	**77.8**	94.8	**85.3**	**64.7**	**77.3**	81.4
Point-M2AE-Scratch	85.7	96.0	76.4	**95.4**	**85.5**	63.8	76.3	82.4
Point-M2AE-SN [61]	**86.4**	95.8	**77.7**	95.3	85.2	**65.3**	**77.0**	82.2
Point-M2AE-Zero	85.8	**96.2**	76.6	94.9	84.8	64.4	76.8	**82.5**
PCP-MAE-Scratch	84.9	95.0	76.0	95.0	85.0	63.2	75.4	81.0
PCP-MAE-SN [63]	85.8	**96.4**	76.1	**95.2**	84.8	**64.0**	**77.4**	**81.4**
PCP-MAE-Zero	**86.0**	96.1	**76.6**	94.6	**85.1**	63.6	76.8	80.4

Table 3. **Part Segmentation Results.** We report the mean Intersection over Union over instances (mIoU_I) and the per-category IoU (%) on the ShapeNetPart benchmark. Higher values indicate better performance. The subscripted values in mIoU_I represent performance improvement over the corresponding baseline.

	With Guidance		Without Guidance	
Methods	ShapeNet	Synthetic	ShapeNet	Synthetic
Point-MAE-SN [29]	**0.015**	**0.024**	**0.024**	0.039
Point-MAE-Zero	0.016	**0.024**	0.026	**0.037**
Point-M2AE-SN [61]	**0.002**	**0.005**	**0.007**	0.011
Point-M2AE-Zero	0.003	**0.005**	0.010	**0.009**
PCP-MAE-SN [63]	-	-	**0.016**	**0.028**
PCP-MAE-Zero	-	-	**0.016**	**0.028**

Table 4. **Masked Point Cloud Completion.** The table reports the L_2 Chamfer distance (*lower is better*) between predicted masked points and ground truth on the test set of ShapeNet and procedurally synthesized 3D shapes. *With Guidance*: center points of masked patches are added to mask tokens in the pretrained decoder, guiding masked point prediction during inference. *Without Guidance*: *Without Guidance*: no information from masked patches is available during inference.

During pretraining, points are grouped into patches, with a subset of patches (60% by default) randomly masked. Only visible patches are encoded, while masked patch centers can optionally guide the decoder ("with guidance") or be omitted entirely ("without guidance"). After pretraining, models can reconstruct masked points even without such guidance. We quantitatively compare Point-MAE and Point-M2AE pretrained on ShapeNet ("-SN") and procedural shapes ("-Zero") in both guidance conditions, using the

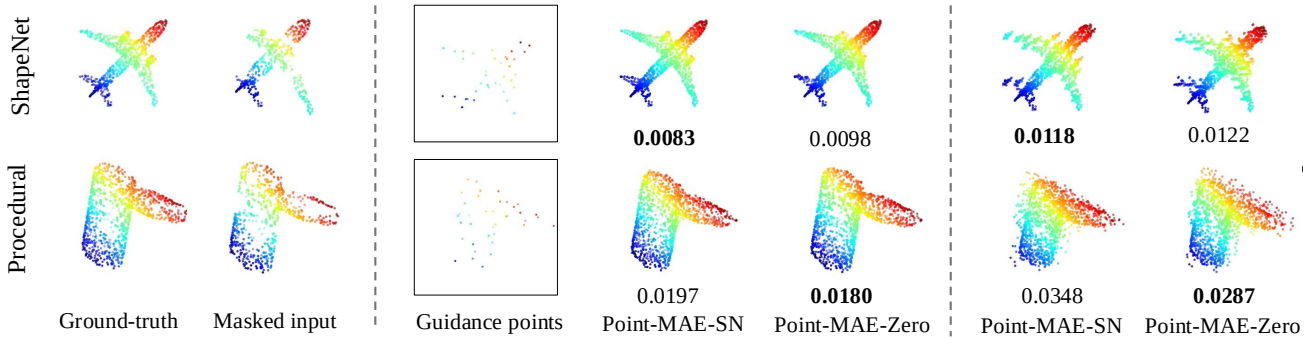


Figure 3. **Masked Point Cloud Completion.** This figure visualizes shape completion results with Point-MAE-SN and Point-MAE-Zero on the ShapeNet test split and procedurally synthesized 3D shapes. **Left**: Ground truth point clouds and masked inputs (60% mask ratio). **Middle**: Completions guided by masked input patch centers [29]. **Right**: Reconstructions without any guidance points. The L_2 Chamfer distance (*lower is better*) between the predicted 3D point clouds and the ground truth is displayed below each reconstruction.

	Semantic Seg.		Instance Segmentation		
Methods	mIoU	mAcc	mAP	AP50	AP25
MSC-scan [44]	73.85	81.80	39.75	60.51	76.49
MSC-Zero (1k)	72.69	80.80	39.03	58.57	75.24
MSC-Zero (2k)	73.86	82.03	40.81	62.28	76.26
MSC-Zero (4k)	**74.34**	**82.16**	**41.47**	**63.12**	**76.54**

Table 5. **Masked Scene Contrast Results.** Performance comparison between MSC-Scan and MSC-Zero with different amounts of pretraining data for semantic and instance segmentation tasks.

ShapeNet test split and 2,000 unseen synthetic shapes (see Tab. 4). All methods perform slightly better on their in-domain data. Removing guidance significantly decreases performance across all methods, highlighting its importance during masked reconstruction. Notably, Point-MAE-Zero and Point-M2AE-Zero closely match or even surpass their SN counterparts in reconstructing synthetic shapes, and remain competitive on ShapeNet shapes despite the lack of semantic training signals. PCP-MAE is a special case since it predicts centers before decoding point cloud and we find PCP-MAE-SN and PCP-MAE-Zero achieve similar performances both on seen and unseen domains.

Fig. 3 further illustrates that procedural-only models (*e.g.*, Point-MAE-Zero) effectively reconstruct familiar ShapeNet objects (*e.g.*, airplane wings, chair legs) without semantic supervision, likely by exploiting geometric symmetries. Similarly, SN-pretrained models generalize effectively to synthetic shapes not encountered during pretraining. Overall, these findings from Fig. 3 and Tab. 4 underscore that masked autoencoding primarily captures geometric rather than semantic information, enabling robust reconstruction across domains.

4.4. Scene-level 3D Understanding Tasks

Given the effectiveness of procedural 3D programs for pretraining self-supervised learning (SSL) models on 3D objects, a natural question arises: *Can procedural 3D programs similarly benefit SSL for 3D scenes?* We adopt Masked Scene Contrast (MSC) [44], a popular SSL method for 3D scenes. We pretrain MSC on ScanNet [11] (1K scenes), commonly used for 3D scene SSL pretraining, and compare it against MSC pretrained on our procedurally generated scenes (denoted MSC-Zero). We conduct experiments with MSC-Zero using varying amounts of data (1K, 2K and 4K procedural scenes). MSC-Zero trained with 4K procedurally generated 3D scenes achieves outperforms MSC pretrained on ScanNet in both 3D semantic and instance segmentation tasks. This demonstrates the effectiveness of procedurally generated data for 3D scene self-supervised learning and reinforces our findings from object-level 3D understanding tasks. Moreover, our results demonstrate that increasing the number of procedural scenes consistently improves performance across semantic and instance segmentation tasks. We discuss the exact procedures of generating such data and implementation details in the supplementary materials.

4.5. Analysis

Complexity of Synthetic 3D shapes. We examine how the geometric complexity of synthetic datasets impacts pretraining and downstream performance. We consider four progressively complex configurations: **(a) Single Primitive**: a single shape with affine transformations; **(b) Multiple Primitives ($\leq$3)**: up to three combined shapes; **(c) Complex Primitives ($\leq$9)**: up to nine combined shapes; **(d) Shape Augmentation**: further modified via boolean differences and wire-frame conversions.

Fig. 4 displays samples from each configuration alongside quantitative comparisons of pretraining performance

Methods	Pre-train Loss	Downstream Accuracy
Scratch	–	77.24
Point-MAE-SN	2.62	85.18
Point-MAE-Zero		
(a) Single Primitive	3.17	83.93
(b) Multiple Primitives	4.10	84.52
(c) Complex Primitives	4.43	84.73
(d) Shape Augmentation	**5.28**	**85.46**

Figure 4. **Impact of 3D Shape Complexity on Performance. Left:** Examples of procedurally generated 3D shapes with increasing complexity, used for pretraining. *Textures are shown for illustration purposes only; in practice, only the surface points are used.* **Right:** Comparison of pretraining masked point reconstruction loss (Eqn. 1) [29] and downstream classification accuracy on the ScanObjectNN dataset [37]. Each row in Point-MAE-Zero represents an incrementally compounded effect of increasing shape complexity and augmentation, with the highest accuracy achieved using shape augmentation.

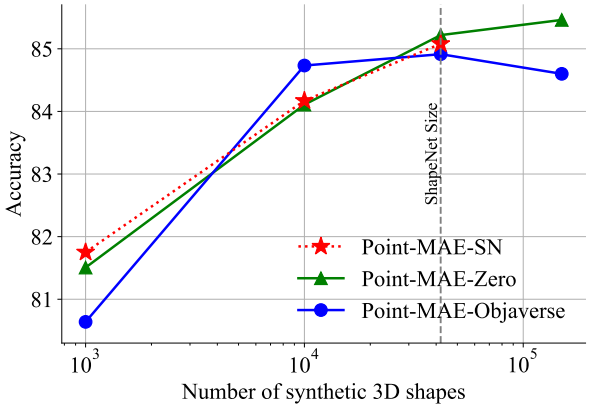


Figure 5. **Impact of pretraining dataset size.** We report the classification accuracy (%) on the PB-T50-RS subset of ScanObjectNN [37] as a function of the pretraining dataset size.

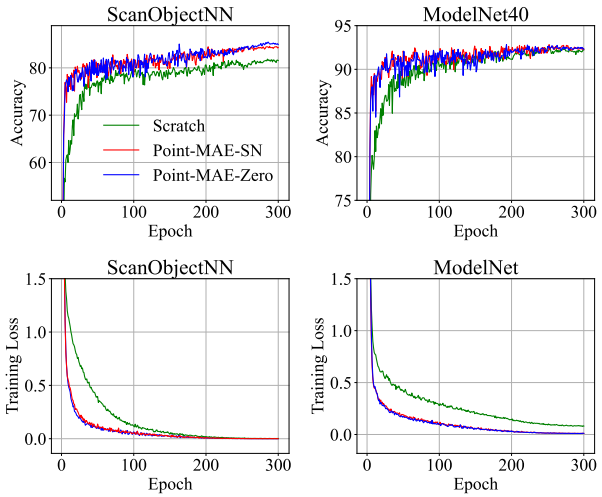


Figure 6. **Learning curves in downstream tasks.** We present validation accuracy (top row) and training curves (bottom row) in object classification tasks on ScanObjectNN (left column) and ModelNet40 (right column).

and downstream classification accuracy on PB-T50-RS, the most challenging variant of ScanObjectNN [37]. As shape complexity increases, the pretraining task becomes more difficult, leading to higher reconstruction losses at the 300th training epoch. However, the downstream classification performance of Point-MAE-Zero improves. This underscores the importance of topological diversity in shapes for effective self-supervised point cloud representation learning.

We observe that the reconstruction loss on our dataset with single primitives (*i.e.*, 3.17) is higher than on ShapeNet (*i.e.*, 2.62) which consists of more diverse 3D shapes. We hypothesize that this is because ShapeNet is relatively smaller than our dataset (50K vs. 150K) and ShapeNet models are coordinate-aligned.

Dataset Size and Comparison with Objaverse. Fig. 5 presents a scaling analysis of Point-MAE-Zero on the PB-T50-RS benchmark using procedurally generated data. When matched for dataset size, Point-MAE-Zero and Point-MAE-SN achieve comparable downstream performance, underscoring the viability of synthetic data as a substitute for curated datasets like ShapeNet. Importantly, we observe that performance further improves with increasing dataset size, despite the lack of semantics in the synthetic data.

To contextualize our results, we compare against pretraining on randomly sampled subsets of Objaverse [12], matched in scale. While models pretrained on Objaverse exhibit strong performance at smaller scales (*e.g.*, 10^3 or 10^4 shapes), their downstream performance plateaus—and eventually declines—as more shapes are added. We pretrain and finetune Point-MAE-Objaverse 5 times on PB-T50-RS and we report the best performance. We hypothesize that this trend is due to many randomly sampled shapes from Objaverse having simple geometry (*e.g.*, avatars, boxes, cups), which makes the pretraining task too easy and limits the ability to learn robust 3D representations. In contrast,

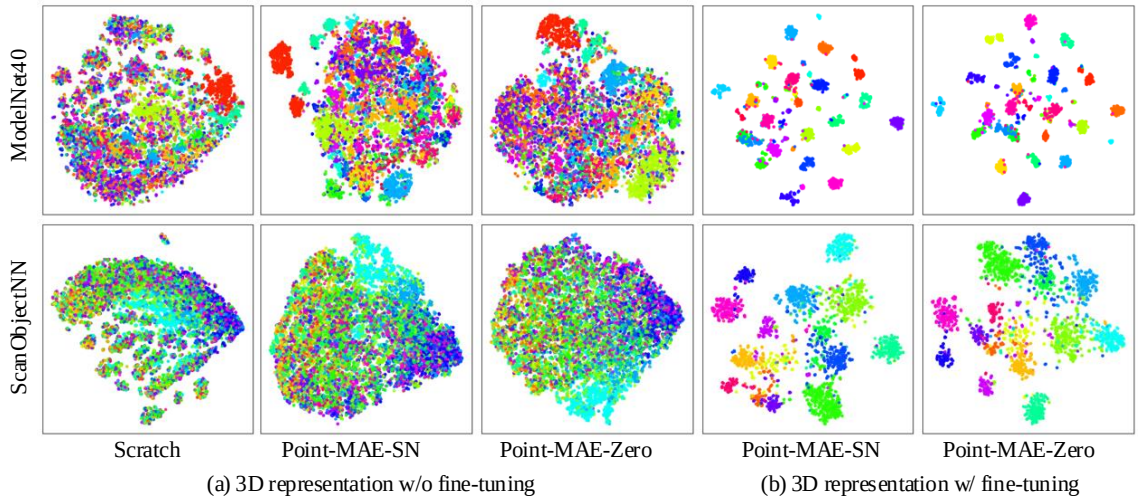


Figure 7. **t-SNE visualization of 3D shape representations. (a)** Shows representations from transformer encoders: Scratch, Point-MAE-SN (ShapeNet), and Point-MAE-Zero (procedural shapes). **(b)** Displays fine-tuned representations for object classification on ModelNet40 (top) and ScanObjectNN (bottom). Each point represents a 3D shape while the color denotes the semantic categories.

our procedurally generated data, enriched through augmentation and shape composition, consistently improves performance as the dataset scales, demonstrating its effectiveness and scalability for point cloud self-supervised learning.

Efficiency of Transfer Learning. Fig. 6 shows the learning curves for training from scratch, Point-MAE-SN, and Point-MAE-Zero on shape classification tasks in the transfer learning setting. Both Point-MAE-SN and Point-MAE-Zero demonstrate faster training convergence and higher accuracy compared to training from scratch, consistently across both ModelNet40 and ScanObjectNN benchmarks.

t-SNE Visualization. Fig. 7 visualizes the distribution of 3D shape representations from Point-MAE-SN and Point-MAE-Zero via t-SNE [38], before and after fine-tuning on specific downstream tasks. It also includes representations from a randomly initialized neural network as a reference.

First, compared to the representations from scratch, both Point-MAE-SN and Point-MAE-Zero demonstrate ***visually improved separation*** between different categories in the latent space. For example, this is evident in the red and light blue clusters on ModelNet40 and the blue and light blue clusters on ScanObjectNN. This highlights the effectiveness of masked auto-encoding for self-supervised 3D learning.

Second, when comparing representations after fine-tuning, both Point-MAE-SN and Point-MAE-Zero show ***much less clear separation*** between categories in the latent space. This raises the question of whether high-level semantic features are truly learned through the masked autoencoding pretraining scheme.

Finally, the t-SNE visualization reveals structural similarities between Point-MAE-Zero and Point-MAE-SN. Most categories ***lack clear separation*** in both models, except for the red and light blue clusters on ModelNet40 and the blue and light blue clusters on ScanObjectNN. This suggests that Point-MAE-Zero and Point-MAE-SN may have learned similar 3D representations, despite differences in the domains of their pretraining datasets. We provide more in-depth analysis in the supplementary material.

5. Discussion

In this work, we propose to learn 3D representations from synthetic data automatically generated using procedural 3D programs. We conduct an comprehensive empirical analysis of existing 3D SSLs and perform extensive comparisons with learning from well-curated, semantically meaningful 3D datasets.

We demonstrate that learning with procedural 3D programs performs comparably to learning from recognizable 3D models, despite the lack of semantic content in synthetic data. Our experiments highlights the importance of geometric complexity and dataset size in synthetic datasets for effective 3D representation learning. Our analysis further reveals that existing 3D SSLs primarily learns geometric structures (*e.g.*, symmetry) rather than high-level semantics.

This work has several limitations. For example, due to limited computational resources, we were unable to further scale up our experiments, such as by increasing the dataset size or conducting more detailed ablation studies on procedural 3D generation. Additionally, our findings may be influenced by potential biases in visualization tools (*e.g.*, t-SNE) or benchmarks (*e.g.*, data distribution and evaluation protocols). Furthermore, in 3D vision, the distinction between geometric structures and semantics remains an open question, as well-stated by Xie *et al*. [48]. This work also does not provide any novel representation learning method. Nevertheless, we hope our findings will inspire further exploration into self-supervised 3D representation learning.

References

[1] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pages 247–263. Springer, 2024. 2

[2] Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. *Advances in Neural Information Processing Systems*, 35:6450–6462, 2022. 2

[3] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34:2556–2569, 2021. 2

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[5] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2

[8] 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. 1, 2, 4

[9] Prakash Chandra Chhipa, Richa Upadhyay, Rajkumar Saini, Lars Lindqvist, Richard Nordenskjold, Seiichi Uchida, and Marcus Liwicki. Depth contrast: Self-supervised pretraining on 3dpm images for mining material classification. In *European Conference on Computer Vision*, pages 212–227. Springer, 2022. 2

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 2

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 4, 6

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 1, 2, 7

[13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[14] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 3

[15] 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?, 2023. 1, 2

[16] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1, 2

[17] Matheus Gadelha, Aruni RoyChowdhury, Gopal Sharma, Evangelos Kalogerakis, Liangliang Cao, Erik Learned-Miller, Rui Wang, and Subhransu Maji. Label-efficient learning on point clouds using approximate convex decompositions. In *European Conference on Computer Vision*, pages 473–491. Springer, 2020. 2

[18] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzhi Li, and Pheng-Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training, 2023. 2

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[20] 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, 2, 3

[21] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts, 2021. 3

[22] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, Georgios Pavlakos, and Hao Tan. Megasynth: Scaling up 3d scene reconstruction with synthesized data. 2024. 1, 2, 3

[23] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 2

[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. 4

[25] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2, 3

[26] Hao Liu, Minglin Chen, Yanni Ma, Haihong Xiao, and Ying He. Point cloud unsupervised pre-training via 3d gaussian splatting, 2024. 3

[27] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8895–8904, 2019. 1, 4

[28] Wufei Ma, Guanning Zeng, Guofeng Zhang, Qihao Liu, Letian Zhang, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *arXiv preprint arXiv:2406.09613*, 2024. 2

[29] 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. 1, 2, 3, 4, 5, 6, 7

[30] 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. 4, 5

[31] 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. 4, 5

[32] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning (ICML)*, 2023. 2

[33] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2

[34] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[35] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018. 2

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[37] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. 1, 4, 7

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

[39] Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, Bohao Peng, Hengshuang Zhao, and Jiaya Jia. Groupcontrast: Semantic-aware self-supervised representation learning for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4917–4928, 2024. 3

[40] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021. 1, 2, 3, 4, 5

[41] 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. 4, 5

[42] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2

[43] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 4

[44] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning, 2023. 1, 2, 3, 6

[45] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025. 1

[46] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1912–1920, 2015. 2

[47] 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. 4

[48] Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. *arXiv preprint arXiv:2406.09371*, 2024. 1, 2, 3, 8

[49] 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. 2

[50] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, 2020. 3

[51] 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. 4

[52] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37 (4):1–13, 2018. 1, 2, 3

[53] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 1, 2, 3

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[55] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1, 2

[56] 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. 1, 5

[57] 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

[58] Xueyang Yu, Xinlei Chen, and Yossi Gandelsman. Learning video representations without natural videos. *arXiv preprint arXiv:2410.24213*, 2024. 2

[59] Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders, 2023. 2

[60] Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders. *arXiv preprint arXiv:2312.10726*, 2023. 2

[61] 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. *arXiv preprint arXiv:2205.14401*, 2022. 1, 2, 3, 4, 5

[62] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders, 2022. 2

[63] Xiangdong Zhang, Shaofeng Zhang, and Junchi Yan. Pcp-mae: Learning to predict centers for point masked autoencoders. *arXiv preprint arXiv:2408.08753*, 2024. 1, 2, 3, 4, 5

[64] 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. 4

[65] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 2