# ReGen3D: Generalizable Unified Representation Learning for 3D Understanding

Haotian Zhang[1*], Jincen Jiang[1*], Yuhang Li[2], Jianjun Zhang[1], Meili Wang[3✉], and Jian Chang[1✉]

[1] National Centre for Computer Animation, Bournemouth University, Dorset, UK
[2] Shanghai Theatre Academy, Shanghai, China
[3] Northwest A&F University, Yangling, China
{zhangh, jiangj, jzhang, jchang}@bournemouth.ac.uk
yuhangli007@gmail.com wml@nwsuaf.edu.cn

**Abstract.** Existing 3D understanding models are typically built on a fixed geometric representation and often fail when the representation format changes. Unlike domain generalization, which studies distribution shift under a fixed input structure, we focus on representation shift, where identical 3D semantics are expressed through heterogeneous geometric encodings such as points, voxels, meshes, and 3D Gaussian primitives. As 3D representations continue to evolve beyond a predefined set, this challenge is becoming increasingly important. We formalize this problem as Representation Generalization (RG), a new setting for unified 3D learning. RG requires a model to share knowledge across representations during training, remain effective when only a single representation is available at test time, and extend efficiently to unseen or newly emerging representations. To this end, we propose ReGen3D, a unified framework with three modules: Unified Representation Tokenization (URT), which maps heterogeneous representations into a shared token space; Shared Memory Interaction (SMI), which enables cross-representation learning during training while supporting inference from any single available representation; and Efficient Representation Adapter (ERA), which supports lightweight adaptation to unseen representations without modifying the shared backbone. We further establish RepShift-3D for systematic evaluation under the RG setting. Extensive experiments on classification, segmentation, and reconstruction show that ReGen3D achieves strong and robust performance across diverse representation scenarios.

**Keywords:** Representation Generalization · Unified 3D Modeling · 3D Understanding

## 1 Introduction

3D understanding has moved beyond a single geometric representation. Modern 3D data can be expressed as point clouds [37–39, 52], voxels [62], meshes [10,

[*] Equal contributions.
[✉] Corresponding authors.

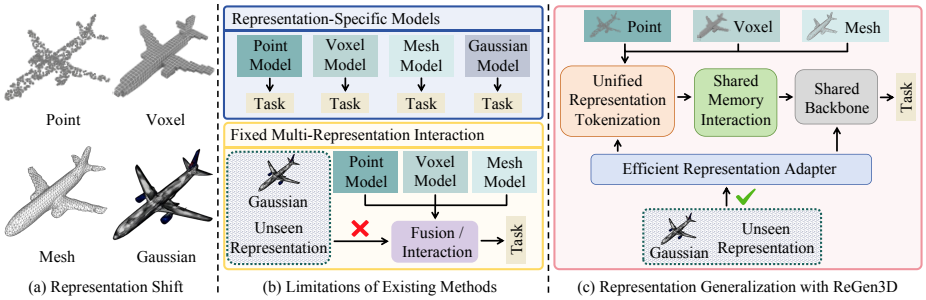


**Fig. 1:** From representation shift to representation generalization. Left: the same 3D object can be represented as points, voxels, meshes, or 3D Gaussian primitives. Middle: existing methods are either representation-specific or tied to fixed multi-representation inference. Right: ReGen3D addresses these limitations with URT, SMI, and ERA.

27], or, more recently, 3D Gaussian primitives [23,28,49], each offering different structural priors and modeling affordances. However, most existing methods are still built on a fixed representation assumption: they are designed, trained, and evaluated on one representation. Once the representation changes, performance often drops substantially [9,20,21,59,61], even when the underlying 3D semantics remain the same. As illustrated in Fig. 1(a), the same 3D object can be encoded by heterogeneous geometric representations, yet existing models often fail to transfer across them. This reveals a practical but largely overlooked challenge in 3D learning, which we refer to as *representation shift*.

Existing paradigms do not fully address this challenge. Domain generalization studies distribution shift under a fixed input structure [1,25,42,61], whereas representation shift arises from the geometric encoding itself. Multimodal learning has explored interaction across language, 2D, and 3D [29, 41, 55], but focuses on semantic complementarity rather than structural variation within 3D. A more relevant direction is joint learning across heterogeneous 3D representations [16, 17]. However, as illustrated in Fig. 1(b), existing methods still suffer from two key limitations: (1) most methods encode each representation with separate representation-specific models, with any subsequent interaction introduced only afterward. Such designs do not reduce representation discrepancy within a unified model, but merely fuse fragmented representation spaces; (2) existing interaction schemes typically assume a fixed representation set and couple inference to the training-time configuration, making it difficult to support single-representation inference or extend to unseen representations.

To address this gap, we introduce Representation Generalization (RG), a new setting for unified 3D learning under representation shift. RG requires a unified model to satisfy three properties: (1) shared learning across heterogeneous representations during training, so that complementary structural cues can be exploited; (2) decoupled inference, where only a single representation may be available at test time; and (3) scalable expansion beyond a predefined representation set, allowing unseen or emerging representations to be integrated

efficiently. These requirements make RG fundamentally different from standard multi-representation fusion: the goal is not simply to process multiple representations jointly, but to learn a unified model that remains robust when representation availability changes and the representation space continues to evolve.

In this paper, we propose **ReGen3D**, a unified framework for **Re**presentation **Gen**eralizable **3D** understanding. ReGen3D resolves representation dependency at three levels. *Unified Representation Tokenization (URT)* maps heterogeneous representations into a shared token space while preserving representation-specific structural cues. *Shared Memory Interaction (SMI)* enables cross-representation knowledge learning during training through a shared latent memory, while allowing inference from any available representation. *Efficient Representation Adapter (ERA)* supports lightweight extension to unseen representations without modifying the shared backbone. We further establish *RepShift-3D*, a benchmark for RG, with aligned heterogeneous representations and unified protocols across classification, segmentation, and reconstruction. Our contributions are three-fold:

- We introduce Representation Generalization (RG), a new setting for 3D understanding that studies robustness under representation shift, where identical semantics are expressed through heterogeneous geometric encodings and representation availability may vary between training and inference.
- We propose ReGen3D, a unified framework that combines URT, SMI, and ERA to support shared learning across multiple representations, decoupled single-representation inference, and extension to unseen representations.
- We establish RepShift-3D, a benchmark for systematic evaluation under the RG setting. Extensive experiments on classification, segmentation, and reconstruction demonstrate strong performance and robust generalization across diverse representation scenarios.

## 2 Related Work

**Representation-Specific 3D Learning.** Most 3D understanding methods are developed under a fixed representation assumption [52], where the architecture is tightly coupled to a specific geometric format [10,23,38,43,44,52]. Point-based methods learn from unordered point sets through permutation-invariant aggregation, hierarchical grouping, graph construction, or attention-based neighborhood modeling [4,15,30,37–39,46,51,54,56,60]. Voxel-based methods exploit regular volumetric grids for convolutional processing [5,8,52], while mesh-based methods leverage surface connectivity and topology [6,10]. More recently, Gaussian-based representations have emerged as efficient primitives for 3D reconstruction and rendering [11,14,23,49,50,53,58]. Despite their success, these methods remain representation-specific, with architectures and inductive biases tied to the native input format. As a result, they do not naturally transfer when the representation changes, even if the underlying 3D semantics remain the same. Our work instead studies robustness under representation shift.

**Multi-Representation 3D Learning.** Several studies explore learning from multiple geometric representations to exploit complementary structural cues

[16, 17, 40]. Common designs either process each representation with a dedicated encoder and introduce interaction via alignment, contrastive learning, or cross-attention, or convert heterogeneous inputs into a shared format before processing [18, 19, 31]. In parallel, transformer-based architectures have become a dominant paradigm for 3D representation learning [24, 35, 47, 60] thanks to tokenized modeling, which offers a flexible interface and suggests mapping heterogeneous inputs into a shared token space while preserving structural priors. Yet most transformer-based 3D models remain representation-specific, *e.g.,* for point clouds [36] or meshes [27], and their tokenization is tightly coupled to the native format. As a result, existing multi-representation schemes still rely on separate pipelines and introduce interaction only after independent encoding; moreover, they typically assume a predefined representation set and couple multiple streams at inference, which does not naturally support single-representation inference or extension to unseen representations.

**Generalization beyond Distribution Shift.** Robustness under distribution shift has been extensively studied in domain adaptation, domain generalization, and test-time adaptation [1,7,32,45,48]. These paradigms improve transfer across datasets, domains, or environments through invariant feature learning, distribution alignment, or online adaptation. Related work on parameter-efficient tuning further supports scalable extension [12, 13, 26] of pretrained models to new domains while preserving learned knowledge. However, most existing generalization frameworks assume a fixed input structure and focus on statistical variation rather than structural variation induced by heterogeneous geometric representations. As a result, they do not explicitly address representation shift, nor do they study how a unified model should support shared learning, decoupled inference, and scalable expansion under changing representation availability. In contrast, our work formalizes RG as a distinct setting for 3D learning across heterogeneous geometric encodings within a unified framework.

## 3 Method

### 3.1 Problem Formulation

We study Representation Generalization (RG) for 3D understanding under representation shift. Let $\mathcal{R} = \{r_1, \ldots, r_K\}$ denote a set of geometric representations, *e.g.,* point clouds, voxels, meshes, and 3D Gaussian primitives. For an object with semantic label $y$, each representation $r_k$ provides a structured observation $x^{(k)} \in \mathcal{X}^{(k)}$. Although $\{x^{(k)}\}_{k=1}^{K}$ describe the same 3D semantics, they differ substantially in geometric encoding and structural organization.

Unlike standard domain generalization, RG does not assume a fixed input structure. Training may access multiple representations of the same sample, whereas inference may rely on only a single available representation. Moreover, the representation set is open-ended, so a practical unified model should support efficient extension to unseen representations. Formally, let $\mathcal{S} \subseteq \mathcal{R}$ denote the set of available representations for a sample. RG seeks a unified predictor $f_\Theta : \{x^{(k)} \mid k \in \mathcal{S}\} \mapsto y$ that satisfies three requirements: (1) shared learning

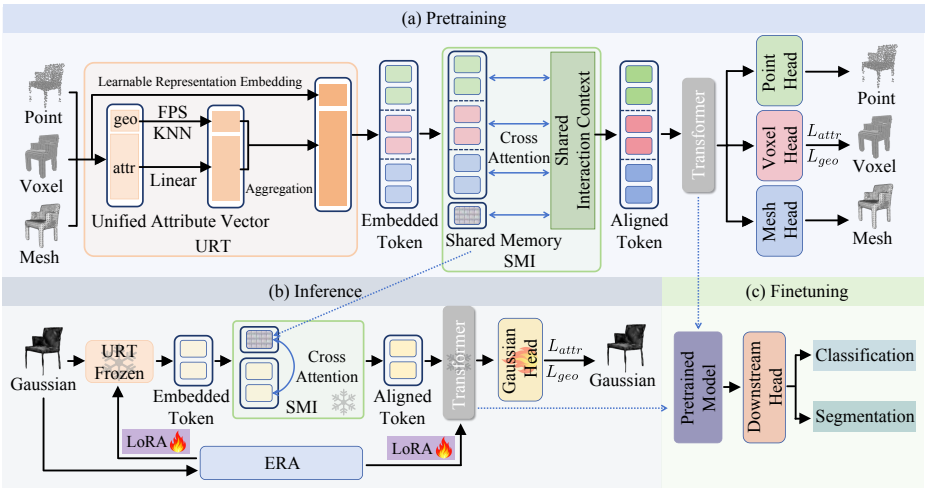


**Fig. 2:** Overview of ReGen3D. Given heterogeneous representations, we first perform reconstruction pretraining with Unified Representation Tokenization (URT) and Shared Memory Interaction (SMI), enabling cross-representation learning in a unified architecture. For an unseen representation, Efficient Representation Adapter (ERA) provides a lightweight adaptation to the same frozen shared encoder. The learned model is further finetuned for downstream classification and segmentation.

across heterogeneous representations during training, (2) decoupled inference under varying representation availability, and (3) scalable expansion to unseen representations via lightweight adaptation while keeping the backbone frozen.

These requirements distinguish RG from representation-specific learning and fixed multi-representation fusion. The goal is to learn a unified model that remains robust when the available representation changes.

### 3.2 Overview of ReGen3D

ReGen3D addresses three representative 3D tasks under the RG setting: reconstruction, classification, and segmentation. As shown in Fig. 2, it learns from heterogeneous geometric inputs in a shared architecture, enables cross-representation interaction during training, and extends efficiently to unseen representations. Specifically, Unified Representation Tokenization (URT) maps heterogeneous inputs into a shared token space while preserving native structural cues. The resulting tokens are processed by a shared encoder with Shared Memory Interaction (SMI), where cross-attention enables information exchange across representations and a shared memory bank accumulates transferable priors. Task-specific heads are then used for reconstruction, classification, and segmentation. To extend the model beyond a closed representation set, Efficient Representation Adapter (ERA) introduces a lightweight low-rank adaptation pathway for unseen representations.

In practice, the shared encoder is pretrained for reconstruction on heterogeneous representations and then finetuned for classification and segmentation. For unseen representations, ERA aligns their inputs to the learned token space without modifying the shared backbone, allowing the same framework to support both pretraining and downstream inference under the RG setting.

### 3.3 Unified Representation Tokenization

ReGen3D begins by mapping heterogeneous 3D representations to a shared token interface. Instead of explicit structural conversion (*e.g.,* collapsing all inputs to points), which may discard representation-specific priors (occupancy, topology, radiance), we design Unified Representation Tokenization (URT) to unify the input space while preserving native structural cues.

For representation $r_k$, each primitive is described by a unified attribute vector

$$a_i^{(k)} = [\, g_i^{(k)};\, s_i^{(k)} \,] \in \mathbb{R}^{d_a}, \tag{1}$$

where $g_i^{(k)} \in \mathbb{R}^3$ is the normalized coordinate, and $s_i^{(k)}$ collects representation-dependent attributes, *e.g.,* voxel size, face normal, curvature or Gaussian opacity, scale, and radiance coefficients. Since different representations expose different attribute types, missing entries are zero-padded so that all primitives share the same attribute dimension $d_a$, yielding $A^{(k)} \in \mathbb{R}^{N_k \times d_a}$. The detailed attribute instantiations are provided in the supplementary.

To produce tokens that can be processed by a shared backbone, URT decouples geometric grouping from attribute projection. We build local neighborhoods purely in $\mathbb{R}^3$ by selecting $G$ anchors via FPS on $\{g_i^{(k)}\}$ and gathering neighbors by KNN, ensuring geometry-consistent patching across encodings. Attributes are then mapped to a unified feature space by a shared linear projection $\phi : \mathbb{R}^{d_a} \to \mathbb{R}^D$, avoiding learning separate representation-dependent feature spaces, and zero-padded entries contribute no informative signal under $\phi$.

For each anchor $c_g$, neighbor attributes are aggregated into a patch token

$$z_g^{(k)} = \rho\Big(\Big\{\phi\Big(a_j^{(k)} - a_{c_g}^{(k)}\Big) \,\Big|\, j \in \mathcal{N}_g(c_g)\Big\}\Big) + e_k, \tag{2}$$

where $\rho(\cdot)$ is a permutation-invariant aggregation operator and $e_k$ is a learnable representation embedding that injects representation identity into the shared token space. Stacking all patch tokens gives

$$Z^{(k)} = [z_1^{(k)}, \ldots, z_G^{(k)}] \in \mathbb{R}^{G \times D}. \tag{3}$$

URT maps heterogeneous representations to the same token dimension $D$ with geometry-consistent patching and shared attribute projection, providing the unified input interface for subsequent cross-representation learning.

### 3.4 Shared Memory Interaction

URT aligns heterogeneous inputs in a shared token space, but RG still requires cross-representation interaction during training. A direct solution is pairwise fusion between representations, but such designs depend on the representation combination and become cumbersome as the number of representations grows. We instead introduce Shared Memory Interaction (SMI), which performs cross-representation interaction through a shared memory space.

Let $M \in \mathbb{R}^{T\times D}$ denote a learnable memory bank with $T$ slots, and let $Z^{(k)} \in \mathbb{R}^{G\times D}$ denote the token sequence of representation $r_k$. For a sample with available representation set $\mathcal{S}$, we build a shared interaction context by concatenating all available representation tokens with the memory bank:

$$\bar{Z}_{\mathcal{S}} = \mathrm{Concat}\big(\{Z^{(j)}\}_{j\in\mathcal{S}}, M\big). \tag{4}$$

Each representation then attends to this shared context:

$$\hat{Z}^{(k)} = Z^{(k)} + \mathrm{Attn}\big(Q = Z^{(k)},\, K = \bar{Z}_{\mathcal{S}},\, V = \bar{Z}_{\mathcal{S}}\big), \tag{5}$$

where $\mathrm{Attn}(\cdot)$ denotes cross-attention [2] followed by a feed-forward block.

SMI serves two purposes. First, it enables each representation to access information beyond its own token stream by attending to other available representations and the shared memory. Second, the memory bank acts as a persistent carrier of transferable priors, absorbing geometric regularities shared across heterogeneous encodings and feeding them back during interaction. Rather than hard-wiring pairwise correspondence, SMI lets different representations communicate through a shared latent space.

A key property of SMI is that its interaction form remains unchanged when representation availability changes. When only a single representation is available at test time, the shared context reduces to

$$\bar{Z}_{\mathcal{S}} = \mathrm{Concat}\big(Z^{(k)}, M\big), \tag{6}$$

and Eq. (5) is applied without architectural modification. The same mechanism also benefits unseen representations once they are mapped into the shared token space. Although their token statistics may differ from seen representations, cross-attention to the shared memory exposes them to priors accumulated from heterogeneous training representations. This helps pull unseen tokens toward a feature space structured by shared geometric semantics, making the encoder less sensitive to representation-specific discrepancies.

Compared with explicit pairwise fusion, SMI keeps the interaction operator independent of any fixed representation combination and scales naturally as the representation set grows. It is the core mechanism that enables training-time knowledge sharing and flexible test-time generalization in ReGen3D.

### 3.5 Efficient Representation Adapter

New representations may emerge after pretraining, but retraining the full network for each new format is inefficient and undermines the idea of a shared

representation-agnostic backbone. To address this issue, we introduce Efficient Representation Adapter (ERA), which extends ReGen3D to unseen representations through parameter-efficient fine-tuning (PEFT) [12, 13].

Suppose a new representation $r_{K+1} \notin \mathcal{R}$ becomes available after pretraining. ERA keeps the shared components of ReGen3D frozen, *i.e.,* the URT embedding layers, the SMI mechanism, and the Transformer backbone, and introduces lightweight low-rank adaptation (LoRA) branches [13] on top of the frozen network. For a learned weight $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, the adapted weight is written as

$$W' = W + BA, \qquad B \in \mathbb{R}^{d_{\text{in}} \times r}, \quad A \in \mathbb{R}^{r \times d_{\text{out}}}, \quad r \ll \min(d_{\text{in}}, d_{\text{out}}), \tag{7}$$

where $W$ remains frozen and only the low-rank parameters $(A, B)$ are optimized.

The adaptation scope depends on the downstream task. We attach LoRA branches to the URT embedding layers and the Transformer. For reconstruction, we additionally retrain the reconstruction head for the unseen representation since the output structure is representation-dependent. For classification and segmentation, we instead attach LoRA branches to the task head, enabling adaptation to the new input structure while preserving the shared backbone.

This design is enabled by URT and SMI in the framework. URT maps heterogeneous inputs into a shared token interface, and SMI organizes interaction in a shared latent space through a fixed memory-guided operator. As a result, adapting an unseen representation does not require re-learning the shared encoder or memory, but only lightweight task-dependent adjustments that align the new representation with the pretrained feature and prediction spaces.

Once adapted, the unseen representation can be processed by the same framework. As shown in Fig. 2(b), its tokens are aligned to the shared token space through the adapted URT embedding, then interact with the fixed shared memory through the same cross-attention mechanism, while ERA compensates for representation-specific discrepancies without disturbing learned priors. This enables ReGen3D to extend beyond the original representation set while preserving shared knowledge acquired from heterogeneous training representations.

### 3.6 Training Objective

ReGen3D is trained in two stages: reconstruction pretraining on heterogeneous representations, followed by finetuning for classification and segmentation.

**Reconstruction pretraining.** The model reconstructs the target representation from encoded tokens. For representation $r_k$, let $\hat{A}^{(k)}$ and $A^{(k)}$ denote predicted and ground-truth attributes. The attribute reconstruction loss is

$$\mathcal{L}_{\text{attr}}^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} \left\| \hat{a}_i^{(k)} - a_i^{(k)} \right\|_2^2, \tag{8}$$

where $N_k$ is the number of primitives in $r_k$, and the geometry loss is

$$\mathcal{L}_{\text{geo}}^{(k)} = \text{CD}\left( \hat{P}^{(k)}, P^{(k)} \right), \tag{9}$$

where $\mathrm{CD}(\cdot,\cdot)$ is the Chamfer Distance between reconstructed and target geometry. The overall reconstruction objective is

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\left(\mathcal{L}_{\mathrm{attr}}^{(k)} + \lambda\mathcal{L}_{\mathrm{geo}}^{(k)}\right), \tag{10}$$

where $\lambda$ balances attribute and geometry reconstruction.

**Classification and segmentation finetuning.** For downstream tasks, the pretrained encoder is finetuned with standard supervised loss:

$$\mathcal{L}_{\mathrm{task}} = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\ell\left(\hat{y}^{(k)}, y\right), \tag{11}$$

where $\ell(\cdot,\cdot)$ is the cross-entropy loss for classification or segmentation.

ReGen3D does not require explicit pairwise alignment loss across representations. Instead, cross-representation consistency is induced by shared supervision on identical semantics, the unified token interface of URT, and the latent interaction of SMI. This keeps the objective simple while enabling shared learning during training and flexible generalization across representations.

## 4 Experiments

### 4.1 RepShift-3D Benchmark

To evaluate Representation Generalization (RG), we establish **RepShift-3D**, a benchmark built on ModelNet40 [52] and ShapeNet-Part [3, 57]. Each object has four representations, *i.e.,* point clouds, voxels, meshes, and 3D Gaussian primitives, and annotated for reconstruction, classification, and segmentation.

For classification and reconstruction, RepShift-3D contains 9,840 training samples and 2,467 test samples. For segmentation, it contains 12,052 training samples, 1,851 validation samples, and 2,841 test samples. All representations are normalized in a unified setting. Coordinates are centered and scaled to a unit sphere, and structural attributes are constructed accordingly. Point clouds are uniformly sampled to 1,024 points. Meshes are processed by manifold [22] reconstruction and resampling, standardized to 1,024 faces by default and 4,096 faces for reconstruction. Voxels are obtained by voxelizing mesh-derived samples. Gaussian representations are derived from ShapeSplat [34] and downsampled to a comparable scale. For segmentation, part labels are defined on point clouds and mapped to corresponding representations. Details are provided in supplementary.

### 4.2 Experimental Setup

**Implementation details.** All models are implemented in PyTorch and trained on two TITAN RTX GPUs. The shared encoder is pretrained for reconstruction and finetuned for classification and segmentation. We use AdamW [33] with learning rate 0.001, cosine decay, and weight decay 0.05. The batch size is 128,

**Table 1:** Reconstruction results on the three training representations under LORO. The upper block reports CD ($\times 10^{-3}$), and the lower block reports $\text{MSE}_{\text{attr}}$. '–' indicates that the representation has no structural attributes.

| Method | Ori. Rep. | VMG | | | PMG | | | PVG | | | PVM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V | M | G | P | M | G | P | V | G | P | V | M |
| *Chamfer Distance (CD)* ($\times 10^{-3}$) | | | | | | | | | | | | | |
| PointNet [38] | Point | 42.5 | 144.2 | 467.3 | 35.2 | 148.5 | 474.1 | 34.2 | 43.4 | 475.8 | 34.9 | 43.3 | 142.4 |
| DGCNN [37] | Point | 31.2 | 138.6 | 40.3 | 28.8 | 135.8 | 38.2 | 30.2 | 35.4 | 48.6 | 35.7 | 43.4 | 148.3 |
| PCT [60] | Point | 47.2 | 192.4 | 38.3 | 47.4 | 207.9 | 40.1 | 57.2 | 58.4 | 54.2 | 50.6 | 47.3 | 186.5 |
| Point-MAE [36] | Point | 34.3 | 41.2 | 52.3 | 33.0 | 42.5 | 54.5 | 25.9 | 26.0 | 36.5 | 26.3 | 26.8 | 48.1 |
| MeshNet [6] | Mesh | 34.6 | 186.3 | 108.4 | 33.0 | 195.1 | 110.2 | 34.7 | 38.1 | 97.4 | 38.6 | 38.2 | 181.2 |
| VoxelNet [62] | Voxel | 53.2 | 186.1 | 48.3 | 51.5 | 182.4 | 47.9 | 56.2 | 54.8 | 46.2 | 53.4 | 52.1 | 190.5 |
| MeshMAE [27] | Mesh | 38.2 | 102.3 | 32.3 | 34.9 | 108.8 | 28.3 | 36.4 | 35.1 | 28.4 | 32.6 | 35.4 | 107.2 |
| ShapeSplat [34] | Gaussian | 43.2 | 44.2 | 46.7 | 44.2 | 43.6 | 43.1 | 47.2 | 45.6 | 42.3 | 43.1 | 41.9 | 42.3 |
| Baseline | Multiple | 35.1 | 41.1 | 55.3 | 34.4 | 42.0 | 53.5 | 27.8 | 26.6 | 37.4 | 28.3 | 26.5 | 49.3 |
| **ReGen3D** | Multiple | **23.0** | **22.1** | **30.7** | **23.4** | **23.6** | **29.2** | **25.1** | **25.2** | **31.2** | **27.8** | **28.0** | **24.3** |
| *Attribute Mean Squared Error* ($\mathbf{MSE_{attr}}$) | | | | | | | | | | | | | |
| PointNet [38] | Point | – | 0.86 | 0.13 | – | 0.87 | 0.13 | – | – | 1.19 | – | – | 0.85 |
| DGCNN [37] | Point | – | 0.29 | 0.51 | – | 0.28 | 0.53 | – | – | 0.51 | – | – | 0.28 |
| PCT [60] | Point | – | 0.26 | 0.56 | – | 0.27 | 0.57 | – | – | 0.58 | – | – | 0.26 |
| Point-MAE [36] | Point | – | 0.19 | 0.58 | – | 0.18 | 0.58 | – | – | 0.70 | – | – | 0.17 |
| VoxelNet [62] | Voxel | – | 0.27 | 0.64 | – | 0.26 | 0.67 | – | – | 0.56 | – | – | 0.21 |
| MeshNet [6] | Mesh | – | 0.28 | 0.51 | – | 0.26 | 0.52 | – | – | 0.53 | – | – | 0.29 |
| MeshMAE [27] | Mesh | – | 0.13 | 0.57 | – | 0.14 | 0.56 | – | – | 0.58 | – | – | 0.14 |
| ShapeSplat [34] | Gaussian | – | 0.06 | 0.54 | – | 0.07 | 0.51 | – | – | 0.43 | – | – | 0.06 |
| Baseline | Multiple | – | 0.21 | 0.62 | – | 0.19 | 0.59 | – | – | 0.73 | – | – | 0.19 |
| **ReGen3D** | Multiple | – | **0.04** | **0.40** | – | **0.04** | **0.42** | – | – | **0.38** | – | – | **0.05** |

the mask ratio 0.6, and training lasts 300 epochs. Each representation is sampled to 1,024 primitives and partitioned into $G = 64$ patches of 16 primitives.

**LORO protocol.** We adopt a leave-one-representation-out (LORO) protocol. Given four representations, *i.e.,* Point, Voxel, Mesh, and Gaussian, we hold out one and train on the remaining three, yielding VMG (Point held out), PMG (Voxel held out), PVG (Mesh held out), and PVM (Gaussian held out). For each split, we report results on the three training representations to evaluate unified representation learning, and on the held-out representation to evaluate unseen representation generalization.

**Compared methods.** Most existing 3D methods are designed for a single representation without modeling interaction across heterogeneous formats. For fair comparison, we re-instantiate representative methods in our setting and train them under the same LORO splits, including PointNet [38], DGCNN [37], VoxelNet [62], MeshNet [6], PCT [60], Point-MAE [36], MeshMAE [27], and ShapeSplat [34]. We also construct a naive interaction baseline by adding cross-attention before the shared backbone; it will degenerate to self-attention when only one representation is available at test time. This provides a direct comparison against simple multi-representation interaction.

**Metrics.** For reconstruction, we use Chamfer Distance (CD) [52] and attribute Mean Squared Error ($\text{MSE}_{\text{attr}}$) when applicable. For classification, we report accuracy. For segmentation, we report mean Intersection over Union (mIoU).

**Table 2:** Reconstruction results on the held-out representation under LORO. CD is scaled by $10^{-3}$; $\text{MSE}_{\text{attr}}$ is reported when applicable.

| Method | Ori. Rep. | Point | | Voxel | | Mesh | | Gaussian | |
|---|---|---|---|---|---|---|---|---|---|
| | | CD | $\text{MSE}_{\text{attr}}$ | CD | $\text{MSE}_{\text{attr}}$ | CD | $\text{MSE}_{\text{attr}}$ | CD | $\text{MSE}_{\text{attr}}$ |
| PointNet [38] | Point | 58.4 | – | 61.2 | – | 206.2 | 0.93 | 491.4 | 1.57 |
| DGCNN [37] | Point | 48.3 | – | 53.2 | – | 226.2 | 0.31 | 58.7 | 0.85 |
| PCT [60] | Point | 76.2 | – | 80.4 | – | 294.5 | 0.29 | 62.4 | 0.88 |
| Point-MAE [36] | Point | 52.4 | – | 55.3 | – | 59.2 | 0.24 | 63.1 | 0.84 |
| VoxelNet [62] | Voxel | 74.1 | – | 72.4 | – | 238.2 | 0.34 | 59.8 | 0.85 |
| MeshNet [6] | Mesh | 48.2 | – | 50.4 | – | 212.4 | 0.15 | 137.1 | 0.84 |
| MeshMAE [27] | Mesh | 64.2 | – | 71.7 | – | 121.3 | 0.25 | 48.8 | 0.85 |
| ShapeSplat [34] | Gaussian | 54.5 | – | 55.7 | – | 52.4 | 0.08 | 52.2 | 0.82 |
| Baseline | Multiple | 55.6 | – | 57.2 | – | 59.1 | 0.22 | 63.4 | 0.86 |
| **ReGen3D** | Multiple | **34.5** | – | **33.1** | – | **32.1** | **0.06** | **32.2** | **0.73** |

### 4.3 Pretraining: Reconstruction

**Unified representation learning.** We first evaluate reconstruction on the three training representations under each leave-one-representation-out split. As shown in Table 1, ReGen3D achieves the strongest and most stable performance across all splits, consistently yielding the lowest CD on the training representations. Under VMG, for example, it reduces CD to 23.0 on Voxel and 22.1 on Mesh, compared with 34.4/41.2 for Point-MAE [36] and 38.2/102.3 for MeshMAE [27]. The same trend holds under other splits and in attribute reconstruction, where ReGen3D achieves the lowest $\text{MSE}_{\text{attr}}$ on both Mesh and Gaussian. These results show that URT and SMI enable the shared encoder to accumulate transferable geometric priors across representations, improving single-representation reconstruction after joint multi-representation pretraining.

**Unseen representation generalization.** We next evaluate reconstruction on the held-out representation under the same LORO splits. As shown in Table 2, baseline performance drops substantially once the test representation is excluded from training. The degradation is especially severe for Mesh and Gaussian, where both geometry and structural attributes must be recovered consistently. On unseen Mesh, for example, the best baseline CD is 52.4 from ShapeSplat, whereas ReGen3D reduces it to 32.1; on unseen Gaussian, it further achieves 32.2, compared with 48.8 from MeshMAE and 52.2 from ShapeSplat. The same trend holds in attribute reconstruction, where ReGen3D achieves 0.06 on unseen Mesh and 0.73 on unseen Gaussian. These gains are obtained without changing the shared backbone, showing that priors learned through URT and SMI transfer beyond the training representation set, while ERA provides an efficient adaptation path for the held-out format.

### 4.4 Finetuning: Downstream Tasks

**Unified representation learning.** We then turn to downstream classification and segmentation on the three training representations under each LORO split. Table 3 shows that ReGen3D achieves the best performance across all splits. For

**Table 3:** Downstream results on the three training representations under LORO. The upper and lower blocks report classification and segmentation results, respectively.

| Method | Ori. Rep. | VMG | | | PMG | | | PVG | | | PVM | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | V | M | G | P | M | G | P | V | G | P | V | M |
| *Classification Accuracy (%)* | | | | | | | | | | | | | |
| PointNet [38] | Point | 76.2 | 74.9 | 80.2 | 75.4 | 76.0 | 79.5 | 76.9 | 77.1 | 77.8 | 76.2 | 76.9 | 75.8 |
| DGCNN [37] | Point | 82.1 | 70.2 | 81.4 | 81.5 | 73.7 | 82.9 | 83.6 | 83.8 | 85.0 | 80.5 | 79.7 | 68.9 |
| PCT [60] | Point | 58.6 | 46.2 | 42.1 | 60.3 | 49.3 | 35.9 | 78.9 | 79.0 | 70.3 | 78.1 | 77.8 | 58.6 |
| Point-MAE [36] | Point | 46.7 | 71.4 | 86.2 | 43.3 | 68.4 | 84.3 | 86.3 | 86.0 | 87.8 | 76.5 | 75.3 | 52.4 |
| VoxelNet [62] | Voxel | 42.3 | 45.2 | 48.1 | 40.2 | 44.7 | 47.6 | 44.5 | 43.3 | 47.4 | 31.9 | 31.0 | 34.4 |
| MeshNet [6] | Mesh | 64.2 | 79.4 | 80.2 | 65.3 | 79.0 | 81.6 | 76.1 | 74.9 | 78.9 | 75.4 | 75.6 | 75.3 |
| MeshMAE [27] | Mesh | 72.1 | 71.2 | 66.8 | 54.8 | 70.9 | 66.2 | 53.9 | 74.2 | 66.1 | 52.6 | 74.4 | 72.4 |
| ShapeSplat [34] | Gaussian | 82.4 | 79.2 | 87.1 | 81.2 | 80.9 | 86.9 | 83.8 | 83.6 | 88.2 | 82.1 | 81.7 | 80.7 |
| Baseline | Multiple | 45.2 | 70.1 | 81.3 | 43.1 | 62.2 | 80.0 | 81.4 | 84.3 | 83.2 | 71.5 | 74.9 | 51.2 |
| **ReGen3D** | Multiple | **84.6** | **83.2** | **90.1** | **85.6** | **82.1** | **89.8** | **86.5** | **86.8** | **88.5** | **83.4** | **83.4** | **78.6** |
| *Segmentation mIoU (%)* | | | | | | | | | | | | | |
| PointNet [38] | Point | 43.5 | 46.5 | 49.1 | 50.4 | 41.5 | 43.6 | 50.4 | 49.5 | 42.5 | 47.3 | 49.4 | 48.2 |
| DGCNN [37] | Point | 48.8 | 45.8 | 47.8 | 47.6 | 43.1 | 42.6 | 41.2 | 40.3 | 41.4 | 42.6 | 46.7 | 50.1 |
| PCT [60] | Point | 44.6 | 46.3 | 50.6 | 44.7 | 44.9 | 50.0 | 44.3 | 45.9 | 50.0 | 44.6 | 45.8 | 49.6 |
| Point-MAE [36] | Point | 45.4 | 47.2 | 41.7 | 47.0 | 45.3 | 50.7 | 48.2 | 46.8 | 50.7 | 46.4 | 46.6 | 50.1 |
| VoxelNet [62] | Voxel | 40.3 | 44.2 | 46.7 | 46.3 | 41.2 | 47.0 | 42.3 | 43.7 | 47.0 | 52.2 | 41.3 | 45.5 |
| MeshNet [6] | Mesh | 32.9 | 39.7 | 33.7 | 32.0 | 31.4 | 33.3 | 32.1 | 49.9 | 33.3 | 32.7 | 33.7 | 34.7 |
| MeshMAE [27] | Mesh | 45.5 | 51.8 | 49.7 | 47.1 | 43.5 | 48.6 | 51.4 | 46.5 | 48.6 | 45.9 | 46.7 | 48.0 |
| ShapeSplat [34] | Gaussian | 54.3 | 56.4 | 46.7 | 59.7 | 57.9 | 56.4 | 55.8 | 54.1 | 51.4 | 55.5 | 55.8 | 54.0 |
| Baseline | Multiple | 43.4 | 46.1 | 40.2 | 46.3 | 44.1 | 48.3 | 46.9 | 46.2 | 50.1 | 46.2 | 46.1 | 48.9 |
| **ReGen3D** | Multiple | **77.8** | **65.1** | **58.2** | **77.5** | **64.9** | **58.9** | **78.9** | **80.0** | **57.1** | **79.0** | **78.4** | **63.8** |

**Table 4:** Downstream results on the held-out representation under LORO. Cls. and Seg. denote classification accuracy (%) and segmentation mIoU (%), respectively.

| Method | Ori. Rep. | Point | | Voxel | | Mesh | | Gaussian | |
|---|---|---|---|---|---|---|---|---|---|
| | | Cls. | Seg. | Cls. | Seg. | Cls. | Seg. | Cls. | Seg. |
| PointNet [38] | Point | 43.3 | 39.7 | 48.1 | 40.1 | 26.2 | 39.0 | 24.0 | 44.6 |
| DGCNN [37] | Point | 45.1 | 46.1 | 48.3 | 38.6 | 46.3 | 35.9 | 35.2 | 39.7 |
| PCT [60] | Point | 55.7 | 37.5 | 61.2 | 37.5 | 38.0 | 40.6 | 43.9 | 42.6 |
| Point-MAE [36] | Point | 42.6 | 24.9 | 36.8 | 24.6 | 30.6 | 26.9 | 25.2 | 27.7 |
| VoxelNet [62] | Voxel | 24.1 | 37.1 | 26.5 | 37.4 | 25.7 | 39.6 | 22.8 | 38.8 |
| MeshNet [6] | Mesh | 44.1 | 26.0 | 47.7 | 25.2 | 27.2 | 32.2 | 26.4 | 21.2 |
| MeshMAE [27] | Mesh | 32.2 | 42.5 | 30.9 | 42.1 | 28.3 | 26.5 | 38.5 | 44.2 |
| ShapeSplat [34] | Gaussian | 61.2 | 46.7 | 58.8 | 47.4 | 29.9 | 51.5 | 45.2 | 50.1 |
| Baseline | Multiple | 42.1 | 25.5 | 36.2 | 22.1 | 29.4 | 25.1 | 23.4 | 26.7 |
| **ReGen3D** | Multiple | **78.4** | **75.2** | **77.6** | **76.6** | **67.5** | **62.0** | **85.1** | **56.6** |

classification, it reaches 84.6/83.2/90.1 under VMG and 83.4/83.4/78.6 under PVM. The advantage is larger for segmentation, where it achieves 77.5/64.9/58.9 under PMG and 78.9/80.0/57.1 under PVG. These results show that cross-representation interaction during finetuning improves single-representation inference across heterogeneous training representations.

**Unseen representation generalization.** We further examine downstream performance on the held-out representation under LORO. Table 4 shows that all baselines degrade markedly once the test representation is excluded from train-

**Table 5:** Ablation under the LORO protocol with Point, Gaussian, and Mesh for training and Voxel held out. The left block reports unified learning, and the right block reports unseen generalization. Reconstruction is measured by CD ($\times 10^{-3}$); classification and segmentation are reported in accuracy (%) and mIoU (%), respectively.

| URT | SMI | ERA | Unified Representation Learning | | | | | | | | | Unseen Representation Generalization | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Reconstruction | | | Classification | | | Segmentation | | | Reconstruction | Classification | Segmentation |
| | | | P | M | G | P | M | G | P | M | G | V | V | V |
| | | | 34.4 | 42.0 | 53.5 | 43.1 | 62.2 | 80.0 | 46.3 | 44.1 | 48.3 | 57.2 | 36.2 | 22.1 |
| ✓ | | | 27.2 | 37.1 | 38.6 | 70.9 | 76.9 | 85.2 | 65.4 | 53.1 | 51.2 | 55.7 | 51.8 | 52.4 |
| ✓ | ✓ | | 23.4 | 23.6 | 29.2 | 85.6 | 82.1 | 89.8 | 77.5 | 64.9 | 58.9 | 50.6 | 72.6 | 58.6 |
| ✓ | ✓ | ✓ | 23.4 | 23.6 | 29.2 | 85.6 | 82.1 | 89.8 | 77.5 | 64.9 | 58.9 | 33.1 | 77.6 | 76.6 |

**Table 6:** Robustness evaluation under the VMG setting on real-scan ScanObjectNN data and geometrically perturbed RepShift-3D point clouds.

| Settings | Real Scan | Noise | Rotation | Shift | Scale |
|---|---|---|---|---|---|
| Condition | ScanObjectNN | 10% | $\pm 20°$ | $\pm 0.1$ | $\pm 10\%$ |
| Results | 40.1 CD / 70.6 Acc. | 75.3 Acc. | 78.2 Acc. | 79.9 Acc. | 78.3 Acc. |

ing, indicating limited transfer across representations. By contrast, ReGen3D remains stronger on both classification and segmentation. For classification, it improves the best baseline from 45.2 to 85.1 on unseen Gaussian and from 46.3 to 67.5 on unseen Mesh. For segmentation, it achieves 75.2/76.6/62.0/56.6 on unseen Point, Voxel, Mesh, and Gaussian, respectively. These results show that the finetuned shared encoder retains representation-transferable semantic structure, which can be extended to held-out formats with lightweight adaptation.

### 4.5 Ablation Studies

**Key Modules.** We ablate URT, SMI, and ERA under the same LORO protocol, using Point, Gaussian, and Mesh for training while holding out Voxel. Table 5 reports results on both training and held-out representations. URT already brings clear gains over the baseline, especially for classification and segmentation, showing that a unified token interface reduces structural discrepancy at the input level. Its improvement on held-out reconstruction remains limited, indicating that token alignment alone is insufficient for stable transfer across representations. Adding SMI further improves all three tasks, particularly reconstruction and segmentation, confirming the importance of shared-memory interaction for accumulating transferable priors across heterogeneous representations. ERA then yields additional gains only on the held-out Voxel representation while leaving the training representations unchanged, consistent with its role as an efficient adaptation module for unseen formats. These progressive gains validate the roles of URT, SMI, and ERA in reducing representation discrepancy, strengthening shared learning, and extending the model to unseen representations.

**Robustness evaluation.** We further evaluate ReGen3D on real scans and geometric perturbations under the VMG setting. Since ScanObjectNN contains

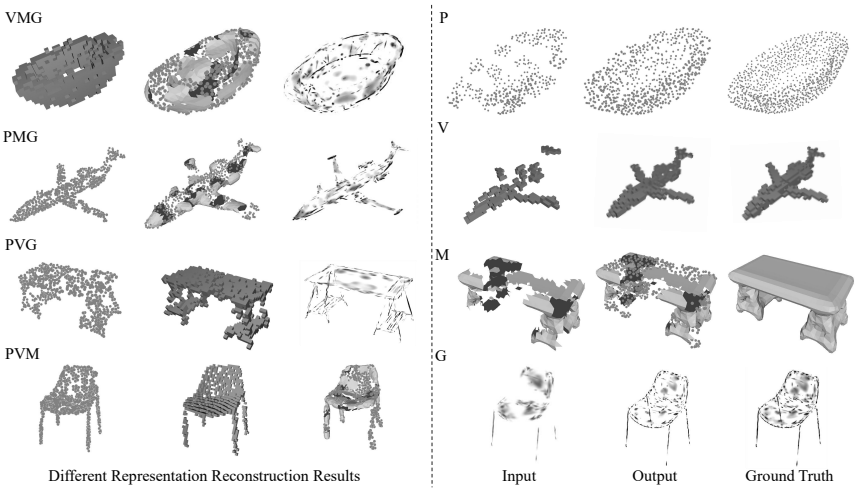


**Fig. 3:** Qualitative reconstruction results. Left: unified learning results on the three training representations. Right: RG results on the held-out unseen representation. We follow MeshMAE [27] and ShapeSplat [34] for mesh and Gaussian visualization.

only point clouds without paired mesh, voxel, or Gaussian counterparts, it cannot form a fully aligned four-representation benchmark. We therefore use it as a joint domain and representation shifts stress test: ReGen3D is pretrained on the voxel, mesh, and Gaussian representations of RepShift-3D, with point clouds held out, and ERA adapts the unseen ScanObjectNN point inputs while keeping the shared backbone frozen. We also apply controlled noise, rotation, translation, and scaling to RepShift-3D point clouds to simulate real-world acquisition and preprocessing variations. As shown in Table 6, ReGen3D adapts effectively to real-scan inputs and remains robust across these controlled perturbations.

### 4.6 Visualization Results

**Visualization for representation generalization.** Fig. 3 shows qualitative reconstruction results under the LORO setting. ReGen3D preserves complete geometry and faithful structural details across different held-out representations, which are consistent with the quantitative gains in Table 2.

**t-SNE visualization.** We further visualize features under the same LORO setting with Gaussian as the unseen representation. As shown in Fig. 4(a), features remain clearly separated after URT, indicating that unified tokenization preserves representation-specific structure. After introducing SMI, the seen representations become more aligned in Fig. 4(b), showing reduced representation discrepancy and shared learning. After applying ERA, the unseen Gaussian features in Fig. 4(c) align to the shared feature space. This progressive transition illustrates the roles of URT, SMI, and ERA in preserving native structure, enabling shared learning, and aligning unseen representations.

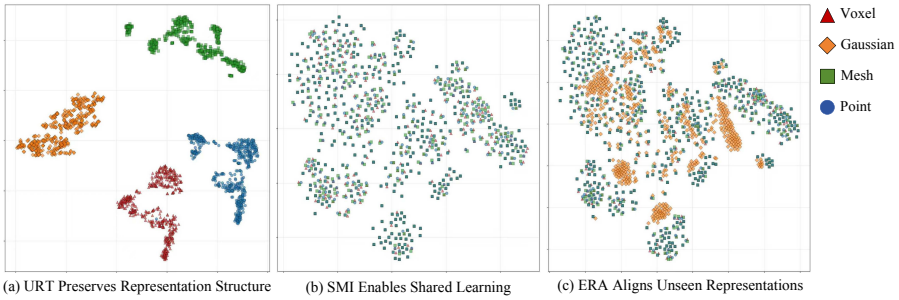


**Fig. 4:** t-SNE visualization under the LORO setting with Gaussian held out.

### 4.7 Limitations

Despite strong robustness to representation shift, ReGen3D still has several limitations. First, URT uses geometry-guided FPS and KNN grouping to construct unified tokens, which reduces representation discrepancy but does not explicitly preserve all fine-grained native structures, such as mesh connectivity. Second, ERA keeps the shared backbone frozen but still requires lightweight optimization to align a new representation with the learned space. Third, appearance attributes of 3D Gaussians, such as color and radiance, remain more challenging to reconstruct than geometry, motivating richer appearance-aware modeling.

## 5 Conclusion

We introduced Representation Generalization (RG), a new setting for 3D understanding under representation shift, where the semantics are expressed through heterogeneous geometric encodings and representation availability may change between training and inference. To address this problem, we proposed ReGen3D, a unified framework built on Unified Representation Tokenization (URT), Shared Memory Interaction (SMI), and Efficient Representation Adapter (ERA), which together enable shared learning, decoupled inference, and efficient extension to unseen representations. We further established RepShift-3D for systematic evaluation under the RG setting. Extensive experiments on reconstruction, classification, and segmentation demonstrate strong and robust performance across diverse representation scenarios. We hope this work can encourage future research on scalable unified 3D learning beyond fixed representation assumptions.

## Acknowledgements

Haotian Zhang is supported by the China Scholarship Council (Grant Number 202506300056) and the Research and Development Fund of Bournemouth University. Meili Wang is supported by the National Key Research and Development Program of China (Grant Number 2022ZD04014).

## References


1. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3D model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chen, G., Wang, M., Yang, Y., Yu, K., Yuan, L., Yue, Y.: Pointgpt: Auto-regressively generative pre-training from point clouds. Advances in Neural Information Processing Systems **36**, 29667–29679 (2023)
5. Choy, C., Gwak, J., Savarese, S.: 4D spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
6. Feng, Y., Feng, Y., You, H., Zhao, X., Gao, Y.: Meshnet: Mesh neural network for 3D shape representation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8279–8286 (2019)
7. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
8. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9224–9232 (2018)
9. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
10. Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., Cohen-Or, D.: Meshcnn: A network with an edge. ACM Transactions on Graphics **38**(4), 1–12 (2019)
11. He, B., Chen, Y., Lu, G., Wang, Q., Gu, Q., Xie, R., Song, L., Zhang, W.: H3D-DGS: Exploring heterogeneous 3D motion representation for deformable 3D gaussian splatting. arXiv preprint arXiv:2408.13036 (2024)
12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of the International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)
14. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3D gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 634–644 (2024)
15. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11108–11117 (2020)
16. Hu, Z., Bai, X., Shang, J., Zhang, R., Dong, J., Wang, X., Sun, G., Fu, H., Tai, C.L.: Voxel-mesh network for geodesic-aware 3d semantic segmentation of indoor scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

17. Hunyuan3D, T., Zhang, B., Guo, C., Liu, H., Yan, H., Shi, H., Huang, J., Yu, J., Li, K., Wang, P., et al.: Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. arXiv preprint arXiv:2509.21245 (2025)
18. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
19. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)
20. Jiang, J., Zhou, Q., Li, Y., Lu, X., Wang, M., Ma, L., Chang, J., Zhang, J.J.: Dgpic: Domain generalized point-in-context learning for point cloud understanding. In: Proceedings of the European Conference on Computer Vision. pp. 455–474. Springer (2024)
21. Jiang, J., Zhou, Q., Li, Y., Zhao, X., Wang, M., Ma, L., Chang, J., Zhang, J., Lu, X., et al.: Pcotta: Continual test-time adaptation for multi-task point cloud understanding. Advances in Neural Information Processing Systems **37**, 96229–96253 (2024)
22. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing. vol. 7 (2006)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 139–1 (2023)
24. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J.: Stratified transformer for 3D point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8500–8509 (2022)
25. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
26. Liang, D., Feng, T., Zhou, X., Zhang, Y., Zou, Z., Bai, X.: Parameter-efficient fine-tuning in spectral domain for point cloud learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
27. Liang, Y., Zhao, S., Yu, B., Zhang, J., He, F.: Meshmae: Masked autoencoders for 3D mesh data analysis. In: Proceedings of the European Conference on Computer Vision. pp. 37–54. Springer (2022)
28. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4D reconstruction with dynamic 3D gaussian particle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21136–21145 (2024)
29. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3D shape representation towards open-world understanding. Advances in Neural Information Processing Systems **36**, 44860–44879 (2023)
30. Liu, Y., Fan, B., Xiang, S., Pan, C.: Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8895–8904 (2019)
31. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel CNN for efficient 3D deep learning. Advances in Neural Information Processing Systems **32** (2019)
32. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: Proceedings of the International Conference on Machine Learning. pp. 97–105. PMLR (2015)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

34. Ma, Q., Li, Y., Ren, B., Sebe, N., Konukoglu, E., Gevers, T., Van Gool, L., Paudel, D.P.: Shapesplat: A large-scale dataset of gaussian splats and their self-supervised pretraining. arXiv preprint arXiv:2408.10906 (2024)
35. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual MLP framework. arXiv preprint arXiv:2202.07123 (2022)
36. Pang, Y., Tay, E.H.F., Yuan, L., Chen, Z.: Masked autoencoders for 3D point cloud self-supervised learning. World Scientific Annual Review of Artificial Intelligence **1**, 2440001 (2023)
37. Phan, A.V., Le Nguyen, M., Nguyen, Y.L.H., Bui, L.T.: Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks **108**, 533–543 (2018)
38. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
39. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
40. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. In: European Conference on Computer Vision. pp. 214–238. Springer (2024)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763 (2021)
42. Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: Proceedings of the International Conference on Machine Learning. pp. 18347–18377. PMLR (2022)
43. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3D representations at high resolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3577–3586 (2017)
44. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 945–953 (2015)
45. Sun, B., Saenko, K.: Deep CORAL: Correlation alignment for deep domain adaptation. In: Proceedings of the European Conference on Computer Vision. pp. 443–450. Springer (2016)
46. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: KPConv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420 (2019)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
48. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: TENT: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
49. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4D gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)

50. Wu, Q., Esturo, J.M., Mirzaei, A., Moenne-Loccoz, N., Gojcic, Z.: 3dgut: Enabling distorted cameras and secondary rays in gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26036–26046 (2025)
51. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9621–9630 (2019)
52. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1912–1920 (2015)
53. Xiang, P., Wen, X., Liu, Y.S., Cao, Y.P., Wan, P., Zheng, W., Han, Z.: Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5499–5509 (2021)
54. Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Y.: Spidercnn: Deep learning on point sets with parameterized convolutional filters. In: Proceedings of the European Conference on Computer Vision. pp. 87–102 (2018)
55. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1179–1189 (2023)
56. Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S.: Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5589–5598 (2020)
57. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3D shape collections. ACM Transactions on Graphics **35**(6), 1–12 (2016)
58. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: Diverse point cloud completion with geometry-aware transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12498–12507 (2021)
59. Zhang, L., Tao, Y., Lin, J., Zhang, F., Fallon, M.: Visual localization in 3D maps: Comparing point cloud, mesh, and NeRF representations. arXiv preprint arXiv:2408.11966 (2024)
60. Zhao, H., Jiang, L., Jia, J., Torr, P.H.S., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268 (2021)
61. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4396–4415 (2022)
62. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2018)