

Cross-Modal 3D Representation with Multi-View Images and Point Clouds

Ziyang Zhou¹, Pinghui Wang^{1†}, Zi Liang², Haitao Bai¹, Ruofei Zhang³,

¹ MOE KLINNS Lab, Xi'an Jiaotong University

² The Hong Kong Polytechnic University ³ Apple

Abstract

The advancement of 3D understanding and representation is a crucial step for the next phase of autonomous driving, robotics, augmented and virtual reality, 3D gaming and 3D e-commerce products. However, existing 3D semantic representation research has primarily focused on point clouds to perceive 3D objects and scenes, overlooking the rich visual details offered by multi-view images, thereby limiting the potential of 3D semantic representation. This paper introduces OpenView, a novel representation method that integrates both point clouds and multi-view images to form a unified 3D representation. OpenView comprises a unique fusion framework, sequence-independent modeling, a cross-modal fusion encoder, and a progressive hard learning strategy. Our experiments demonstrate that OpenView outperforms the state-of-the-art by 11.5% and 5.5% on the R@1 metric for cross-modal retrieval and the Top-1 metric for zero-shot classification tasks, respectively. Furthermore, we showcase some applications of OpenView: 3D retrieval, 3D captioning and hierarchical data clustering, highlighting its generality in the field of 3D representation learning.

1. Introduction

The core aim of 3D understanding is to extract valuable insights from 3D data, which is a crucial technology for refining autonomous driving, robotics, and augmented/virtual reality. Additionally, 3D games comprise a large number of 3D models, and their development can benefit from automated 3D generation and retrieval. Furthermore, e-commerce product displays are transitioning from 2D images to 3D models, requiring 3D semantic understanding and representation. These real-world applications show the importance of research in 3D understanding. However, previous studies have taken point clouds as the only medium for understanding 3D objects or scenes, overlooking the rich visual details contained in multi-view images, which limits the depth of 3D

understanding. This oversight calls for a paradigm shift in 3D representation learning by fusing multi-modal 3D data.

A crucial element in 3D understanding and representation is the data format used to represent raw 3D data. Common formats include meshes, volumes, point clouds, and multi-view images. Among these, meshes [5] and volumes [34] require particularly high resolution and are less suitable as inputs to neural networks. Point clouds, on the other hand, already have many foundation models [40, 41, 70] and 3D representations methods [30, 64, 65, 72]. Recently, based on single-view or multi-view images, some remarkable 3D generation [31, 32, 49, 56] and 3D reconstruction [13, 67, 69, 73] works have been proposed, suggesting the unexplored potential of multi-view images for 3D understanding.

Existing 3D representation methods can be categorized into two types: (a) those focusing solely on 3D data [17, 44, 55, 70], no multi-modal training, can be utilized for classification and segmentation tasks, but fail to support cross-modal downstream tasks; (b) those unifying the representation of images, text, and 3D point clouds [30, 64, 65, 72], exhibiting commendable performance on cross-modal tasks, such as cross-modal retrieval and 3D captioning. However, these existing methods represent 3D data only through point clouds, overlooking the rich texture and color details provided by multi-view images. This overlook results from previous research [27, 59] being driven by autonomous driving, therefore focusing on the physical information from point clouds, while ignoring the advantage of multi-view images for semantic understanding. We seek to build a more comprehensive framework that leverages precise geometry from point clouds and rich visual details from multi-view images.

This paper proposes a new representation learning method called *OpenView* that combines multi-view images and point clouds to represent 3D visuals. OpenView emphasizes four key factors: (1) a novel fusion framework, (2) sequence-independent modeling, (3) a cross-modal fusion encoder, and (4) a progressive hard learning strategy. The combination of point clouds and multi-view images necessitates a new fusion framework and encoder. Since multi-view images are often unordered, sequence-independent modeling is

[†]Corresponding author. Email: phwang@mail.xjtu.edu.cn, {dakandao, haitao.bai}@stu.xjtu.edu.cn, rfzhang@gmail.com.

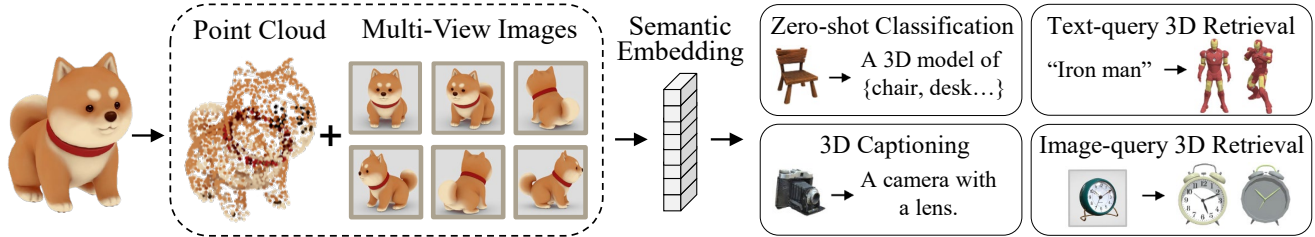


Figure 1. Overview. 3D objects can be rendered and scanned as multi-view images and point clouds. We propose a novel framework that combines point clouds and multi-view images to obtain 3D semantic embeddings, which can be applied to downstream tasks such as zero-shot 3D classification, text-query 3D retrieval and image-query 3D retrieval.

necessary. Furthermore, introducing the image modality can cause the model to focus on a few easy-to-understand images, which becomes a limitation for understanding global information. Therefore, we design a progressive hard training strategy to address this issue.

We first evaluate OpenView on cross-modal retrieval tasks and zero-shot classification tasks. Experiments indicate that on the challenging Objaverse-LVIS [10] dataset, OpenView outperforms the state-of-the-art methods by 11.5% on the R@1 retrieval metric and 5.5% on the Top-1 classification metric. Beyond classification and retrieval, we present demos showcasing OpenView’s generalization ability in understanding 3D visuals. For instance, OpenView can be employed in 3D captioning to generate text descriptions for 3D data, through integrating with a downstream caption model [37]. Furthermore, OpenView can organize 3D data hierarchically and perform clustering. Such hierarchical clustering enables the optimization of retrieval complexity in real-world databases [16, 68].

In summary, OpenView is a 3D semantic representation framework that perceives and understands 3D visuals through multi-view images and point clouds. A new cross-modal fusion method is designed for this purpose, which includes a new fusion framework, a cross-modal encoder, and sequence-independent modeling. Additionally, a new training strategy is developed to facilitate global understanding through progressive hard training. Experiments have shown superior performance compared to state-of-the-art methods, achieving an average improvement of 18.6% and 5.9% in cross-modal retrieval and classification tasks, respectively, across three benchmark datasets. With this work, we make steady steps towards an up-and-coming paradigm shift in 3D understanding.

2. Related Work

2.1. 3D Rendering based on Multi-View Images

3D rendering based on multi-view images has been studied for a long time in the history of computer graphics [4, 22, 46, 47, 50]. The basic approach is to render new views by interpolating pixel values from a given set of multi-view

images. With the rise of deep learning, some studies [3, 8, 39, 57, 58] start to combine deep learning techniques to predict depth information [3, 8, 39] and image features [57, 58] to facilitate pixel matching between different views. Recently, there have emerged many studies focusing on 3D generation [29, 31–33] and 3D reconstruction [13, 20, 25, 73] from multi-view or single-view images. For instance, SparseFusion [73] is based on a view-conditioned latent diffusion model with few views as input and a 3D scene representation as output, which can render novel views. The above studies demonstrate that multi-view images contain rich 3D visual information and reveal the feasibility of 3D representation learning based on multi-view images.

2.2. Single-Modal 3D Representation Learning

Learning 3D Representations from Point Clouds. This line of methods focuses on the self-supervised pretraining of point clouds using pretext tasks, without involving multimodal alignment. These pretext tasks include self-reconstruction [11, 44, 55], mask point modeling [38, 70], and contrastive learning [45, 63, 71]. Further, beyond the self-reconstruction task, Rao et al. [44] performed self-supervised training by mapping global features and different levels of local features to the same feature space. Besides, Yu et al. [70] utilized a discrete variational auto-encoder to generate discrete point tokens, mask local point patches, and predict discrete point tokens during the training. Their downstream tasks include 3D classification, 3D pose estimation [66], etc. However, the absence of multimodal alignment hindered its applicability to cross-modal downstream tasks, such as text-based 3D retrieval and 3D captioning.

Learning 3D Representations from Images. This line of methods [18, 48, 60] learns a compact latent code that encodes the texture, geometry, and pose details of the 3D shape. The typical downstream task is 3D reconstruction using the compact latent code. Both GINA-3D [48] and LRM [18] translate an image into a triplane representation based on a visual transformer. Then, a downstream generator can generate a new view of the 3D shape with the triplane representation as input. Such representations focus on the details of 3D vision rather than semantic understanding.

2.3. Multi-Modal 3D Representation Learning

This line of methods aims to establish unified representations across images, text, and 3D visuals. ULIP [64] initially proposed a multimodal alignment framework based on contrastive learning [21], using point clouds, although it was limited by small-scale datasets. Following ULIP’s framework, with the release of the large-scale dataset Objaverse [10], works such as OpenShape [30], Uni3D [72], and ULIP-2 [65] explored scalable 3D representation pre-training. However, these efforts focused on point clouds as the sole medium for 3D perception, overlooking the potential of multi-view images, which are abundant in visual details. In this work, we aim to combine point clouds and multi-view images to propose a new multi-modal 3D representation learning framework.

3. Method

We propose *OpenView*, an innovative approach for learning a unified representation across text, 2D images, and 3D visuals. Our proposed fusion framework for multi-view images and point clouds is detailed in Section 3.1 and illustrated in Figure 2. The model includes a cross-modal fusion encoder, as described in Section 3.2. The multi-modal alignment training objective is presented in Section 3.3. Finally, we explored a progressive hard training strategy for OpenView in Section 3.4.

3.1. MVI-PC Fusion Framework

Objective. Given a 3D shape or scene in the form of (P, M) as input, where M denotes a set of multi-view images $\{I_i\}_{i=1}^m$ and $P \in \mathbb{R}^{g \times 6}$ denotes the point cloud, the model outputs a 3D representation $e^V \in \mathbb{R}^{d_E}$, aligned with the image-text embedding space of OpenCLIP [7]. Here, I_i denotes the i -th image, m is the total number of multi-view images, d_E is the latent dimension of the model, g is the number of points, and each point has x, y, z coordinates and RGB color.

Components. The fusion model comprises three key components: an image encoder, a point cloud encoder, and a fusion encoder, as depicted in Figure 2. The framework is based on advanced embedding-based modeling, where embeddings are extracted from multi-view images and point clouds, allowing cross-modal fusion based on the attention mechanism. The image encoder independently extracts features $\{e_i^I\}_{i=1}^m \in \mathbb{R}^{d_E}$ from m images of different views. The point cloud encoder extracts features $\{e^P\}_{i=1}^k \in \mathbb{R}^{d_E}$ from the point cloud. The cross-modal fusion encoder takes these features as input and outputs an embedding $e^V \in \mathbb{R}^{d_E}$ as the 3D semantic representation.

Image Encoder. The image encoder utilizes the prevalent transformer-based ViT [12] network. The image is divided into multiple 16×16 patches. Each patch is mapped as

an embedding by a fully connected layer, added with a positional embedding. With an extra [cls] embedding, these embeddings interact with a multi-layer transformer. In this process, each image is extracted into only one semantic embedding, i.e., the [cls] embedding of the last layer is output. The motivation behind this is: if all embeddings are outputted to the subsequent fusion encoder, the GPU memory for inference will be extremely high. The batch size will need to be reduced, which negatively impacts downstream contrastive learning [43].

Point Cloud Encoder. The point cloud encoder uses the network structure of Point-bert [70]. The point cloud encoder divides the point cloud $P \in \mathbb{R}^{g \times 6}$ into k sub-point clouds and then extracts each sub-point cloud as a patch embedding using a mini-PointNet [40]. Later, as in ViT [12], the positional embedding is added. With an additional [cls] embedding, these embeddings are interactively inferred in a multi-layer transformer.

Cross-modal Fusion Encoder. It is necessary to design a cross-modal fusion encoder to fuse the embeddings $\{e_i^P\}_{i=1}^k$ from the point cloud modality and the embeddings $\{e_i^I\}_{i=1}^m$ from the image modality. The cross-modal fusion encoder is designed as a variant of the transformer [54] structure. In particular, two special tokens are set to aggregate the global information. The point-cloud modality utilizes the [cls] token output by the point cloud encoder, while the image modality takes the average feature of the image embeddings $\{e_i^I\}_{i=1}^m$ as the special token. Furthermore, to modeling the intra-modal dependencies and inter-modal correlations between the two modalities, intra-modal and inter-modal attention mechanisms are designed, which will be described in detail in Section 3.2.

Sequence-Independent Modeling. Since multi-view images are typically unordered, the output should ideally remain consistent regardless of the order of the input images. For this reason, we did not add additional positional embedding in the subsequent cross-modal encoder for image embeddings. Furthermore, to allow the cross-modal fusion encoder to distinguish and accommodate the two modalities, before feeding into the fusion encoder, each feature is added with a learnable type embedding $t^I, t^P \in \mathbb{R}^{d_E}$. As a result, OpenView can adapt to any number of multi-view images, and the output remains consistent regardless of the input order of the multi-view images.

Parameter Initialization. The cross-modal fusion encoder utilizes random initialization based on a uniform distribution. The point cloud encoder is initialized with the pre-trained OpenShape [30] model. The image encoder is initialized with the pre-trained OpenCLIP [7]. This strategy yields two critical benefits: (a) it effectively leverages prior knowledge from large-scale pre-trained models, reducing training costs; and (b) it paves the way for aligning the embedding space of 3D data with the text-image embedding

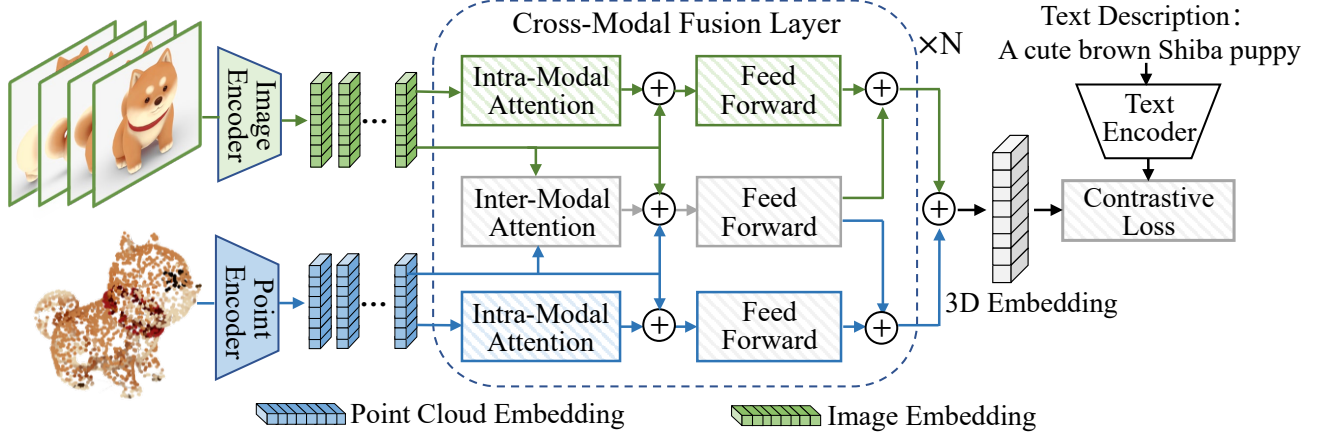


Figure 2. Framework of OpenView. The inputs are multi-view images and point clouds. These inputs are extracted into embeddings by an image encoder and a point cloud encoder. These embeddings are then passed through a cross-modal fusion encoder, which outputs 3D representations. Based on contrastive learning [43], alignment training is performed with the corresponding text embedding.

space of the OpenCLIP model.

3.2. Cross-Modal Fusion Encoder

Motivation. A key challenge in fusing features from multi-view images and point clouds is the significant gap between image features $\{e_i^I\}_{i=1}^m$ and point cloud features $\{e_i^P\}_{i=1}^k$. Our experimental findings indicate that the cosine similarity between image features in the same 3D object is approximately 0.85, while the image features with point cloud features are typically around 0.25. This gap implies that utilizing self-attention [54] alone for inter-modal interaction would be inefficient. Based on this observation, we incorporate cross-attention [52] into the inter-modal attention mechanism. This cross-attention prompts the image features to be updated using the point cloud features and vice versa. Meanwhile, the self-attention is retained to ensure the interaction of global features. The two attention are parallel to facilitate efficient inter-modal interaction.

Intra-modal. We utilize the self-attention [54] to model the intra-modality relationship for image features and point clouds features, respectively. Given a set of features $F = \{e_1; e_2; \dots; e_k\}$, where $e_i \in \mathbb{R}^{1 \times d^E}$, $F \in \mathbb{R}^{k \times d^E}$, the query (Q), key (K) and value (V) are computed as follows:

$$Q = FW^Q, K = FW^K, V = FW^V, \quad (1)$$

where $W^Q, W^K \in \mathbb{R}^{d^E \times d^K}$, $W^V \in \mathbb{R}^{d^E \times d^E}$. The feature F' updated by self-attention is:

$$F' = \text{softmax}\left(\frac{QK^T}{\sqrt{d^K}}\right)V. \quad (2)$$

OpenView also utilizes the multi-head attention mechanism [54], but omits its introduction. Then, the feature $F' = \{e'_1; e'_2; \dots; e'_k\}$ passes through a feed-forward layer and it can be described as:

$$\text{FFN}(e'_i) = \text{ReLU}(e'_i W_1 + b_1) W_2 + b_2, \quad (3)$$

where $W_1 \in \mathbb{R}^{d^E \times d^F}$, $W_2 \in \mathbb{R}^{d^F \times d^E}$, $b_1 \in \mathbb{R}^{1 \times d^F}$, $b_2 \in \mathbb{R}^{1 \times d^E}$, d^E and d^F are latent dimensions.

Inter-modal. We utilize cross-attention [52] and self-attention [54] in parallel for inter-modal modeling. Given a set of image features $F_I = \{e_1^I; e_2^I; \dots; e_m^I\}$ and a set of point cloud features $F_P = \{e_1^P; e_2^P; \dots; e_k^P\}$, the query, key and value are computed as follows:

$$Q_I = F_I W^Q, K_I = F_I W^K, V_I = F_I W^V, \quad (4)$$

$$Q_P = F_P W^Q, K_P = F_P W^K, V_P = F_P W^V, \quad (5)$$

where $W^Q, W^K \in \mathbb{R}^{d^E \times d^K}$, $W^V \in \mathbb{R}^{d^E \times d^E}$, $F_I \in \mathbb{R}^{m \times d^E}$, $F_P \in \mathbb{R}^{k \times d^E}$. The features F'_I and F'_P updated by cross-attention and self attention is computed as:

$$F'_I = \text{softmax}\left(\frac{Q_I K_P^T}{\sqrt{d^K}}\right)V_P + \text{softmax}\left(\frac{Q_I K_I^T}{\sqrt{d^K}}\right)V_I, \quad (6)$$

$$F'_P = \text{softmax}\left(\frac{Q_P K_I^T}{\sqrt{d^K}}\right)V_I + \text{softmax}\left(\frac{Q_P K_P^T}{\sqrt{d^K}}\right)V_P. \quad (7)$$

Then, the features F'_I and F'_P passes through the same feed-forward layer, as described in intra-modal.

Subsequently, the outputs of inter-modal and intra-modal are added together, and normalized to be the input for the next layer of the encoder. In addition, to mitigate the gradient vanishing, residual connections [15] from the input are added to the output of each layer. Finally, the representation output is the normalized addition of the two special token embeddings (detailed in 3.1) in the last layer of the cross-modal fusion encoder.

3.3. Cross-Modal Training Objective

Motivation. This study aims to align the embedding space of 3D data with the text-image embedding space of OpenCLIP [7]. To achieve this, contrastive learning [43] is a prevalent paradigm for cross-modal representation learning. Previous

works [30, 64, 65, 72] have used contrastive learning to set images and text as alignment targets for 3D representations. However, a notable drawback of prior works is the random sampling of one view from the multi-view images as the alignment target. Some views of a 3D shape can be misleading, for example, a wall lamp may appear like a frying pan from above. This can introduce unnecessary noise during alignment training. Therefore, unlike previous studies, we remove the image modality from the alignment targets.

Given a sampled batch of triplets $\{(P_i, M_i, T_i)\}_{i=1}^n$, where P_i and M_i denote a point cloud and multi-view images of 3D data, T_i denotes corresponding text, and n is the number of samples in a batch. With a learnable temperature τ , the contrastive loss is calculated as:

$$Loss = -\frac{1}{2n} \sum_{i=1}^n \left(\log \frac{\exp(e_i^V \cdot e_i^T / \tau)}{\sum_{j=1}^n \exp(e_i^V \cdot e_j^T / \tau)} + \log \frac{\exp(e_i^T \cdot e_i^V / \tau)}{\sum_{j=1}^n \exp(e_i^T \cdot e_j^V / \tau)} \right) \quad (8)$$

where $e_i^V = f^V(P_i, M_i)$ and $e_i^T = f^T(T_i)$ denote normalized representation vectors of 3D data and corresponding text. f^V denotes OpenView’s 3D representation model. f^T denotes the text encoder of OpenCLIP [7] and is frozen in the training.

3.4. Progressive Hard Learning

Motivation. In comparison to prior work, we have introduced the image modality to enhance 3D understanding. However, 3D representation models may focus on a few easy-to-understand views of images, but give up a global 3D understanding. This limits a more comprehensive and deeper understanding. Inspired by curriculum learning [51], we propose a training strategy that employs a more challenging training mode. This mode forces the 3D representation model to learn a more comprehensive correlation between the point cloud and the images.

Specifically, in the hard training mode, partial easy-to-understand multi-view images are removed. The criteria for determining whether an image is easy-to-understand are based on the text description T of the 3D data. Multi-view images are sorted based on cosine similarity between image features and text features. To enable batch training in the hard mode, it is designed to remove a progressive proportion of input images for each training sample, rather than setting a threshold. For each epoch, 10% of the multi-view image is reduced until 50% of the multi-view image remains.

4. Experiments

4.1. Datasets

We utilize the six datasets following [30, 72]. The ensemble training datasets include Objaverse [10], ShapeNetCore

[2], ABO [9], and 3D-Future [14]. The three evaluation benchmarks are Omniobject3D [61], ModelNet40 [62] and Objaverse-LVIS [10]. Each 3D model was sampled into a point cloud of 10,000 colored points and a set of 10 images captured from different viewpoints.

- **Objaverse** dataset [10] comprises 798.8k 3D models. Note that Objaverse-LVIS is the evaluation subset of 46.8k 3D models across 1,156 categories that are removed during the training.
- **ShapeNetCore** dataset [2] comprises 52.5k human-verified CAD 3D models across 55 categories.
- **ABO** dataset [9] comprises 8.0k real-world 3D models derived from Amazon.com products.
- **3D-Future** dataset [14] comprises 16.6k 3D furniture models across 34 categories.
- **ModelNet40** test dataset [62] comprises 2.4k synthetic CAD 3D models across 40 categories.
- **Omniobject3D** test dataset [61] comprises 6.0k real-world 3D models across 190 categories, with text descriptions annotated by human experts.

Textual descriptions in training. Liu et al. [30] have used BLIP [24] and Azure Cognitive Services to generate textual descriptions for the ensemble training datasets. This results in two unique textual descriptions for each 3D model, which are used during the training phase. Specifically, for each training iteration, one of the two descriptions is randomly selected to serve as the target text.

Manual text description in testing. In the cross-modal retrieval task, two test subsets of 1,000 3D models were sampled from Objaverse-LVIS [10] and Omniobject3D [61]. Furthermore, to ensure the quality of the evaluation, we manually annotated text descriptions for the subset of Objaverse-LVIS. In the zero-shot classification task, to be fair, we follow the template-based prompt engineering technique from ULIP [64] and OpenShape [30].

4.2. Implementation Details

Training. Notably, the evaluation subset Objaverse-LVIS has been excluded from the training. The pre-trained image encoder and text encoder from OpenCLIP are frozen in training due to their robust generalisation and large parameters. OpenView is trained on 8 Tesla V100 GPUs (32G) with a batch size of 64, a learning rate of $1e-7$, and 10 epochs, taking 24 hours. The optimizer is AdamW [35] and utilizes the cosine schedule for the learning rate.

Hyperparameter. The latent dimension d^E, d^F are set as 1280 and 4096. The layers of the cross-modal encoder and the point cloud encoder are set to 6 and 12, respectively. The image encoder and the text encoder utilises the ViT-bigG-14 model from OpenCLIP [7].

Metric. For cross-modal retrieval tasks, the metric is recall@ k , i.e., the percentage of ground-truth appearing in

Table 1. Zero-shot classification on Objaverse-LVIS [10], ModelNet40 [62] and Omniobject3D [61]. The best results are **bolded**, and the second-best results are underlined.

Method	Input Type Modality	Objaverse-LVIS [10]			ModelNet40 [62]			Omniobject3D [61]		
		Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
PointCLIP v2 [75]	Point Cloud	4.7	9.5	12.9	63.6	77.9	85.0	7.3	12.0	14.8
ReCon [42]		1.1	2.7	3.7	61.2	73.9	78.1	10.0	16.8	19.6
CLIP2Point [19]		2.7	5.8	7.9	49.5	71.3	81.2	7.5	14.2	18.2
ULIP [64]		6.2	13.6	17.9	60.4	79.0	84.4	7.6	15.1	19.7
OpenShape [30]		39.1	60.8	68.9	85.3	96.2	97.4	33.2	49.6	56.7
Uni3D [72]		<u>47.2</u>	<u>68.8</u>	<u>76.1</u>	<u>86.8</u>	97.3	<u>98.4</u>	37.9	<u>54.9</u>	62.0
PointCLIP v2 [75]	Multi Views	24.9	40.9	47.6	50.7	69.1	74.6	32.8	49.4	57.2
Transformer [54]	Point Cloud	38.2	59.5	67.7	81.6	93.5	95.3	<u>39.2</u>	54.7	63.6
OpenView	+ Multi Views	52.7	74.6	81.1	87.5	<u>97.0</u>	98.5	58.5	73.5	78.6

Table 2. Cross-modal 3D retrieval on Objaverse-LVIS [10] and Omniobject3D [61]. "Text $\rightarrow$ 3D" denotes retrieving the corresponding 3D model given a textual description. "3D $\rightarrow$ Text" denotes retrieving the corresponding textual description given a 3D model.

Method	Input Type Modality	Objaverse-LVIS [10]						Omniobject3D [61]					
		Text $\rightarrow$ 3D			3D $\rightarrow$ Text			Text $\rightarrow$ 3D			3D $\rightarrow$ Text		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
PointCLIP v2 [75]	Point Cloud	15.4	38.1	47.8	4.8	15.0	20.4	3.5	12.2	19.0	1.5	5.4	8.3
ReCon [42]		7.0	20.5	29.9	2.8	11.1	16.5	2.8	12.4	19.3	1.7	4.4	7.5
CLIP2Point [19]		13.7	35.8	45.2	4.6	14.3	19.0	4.2	14.2	20.8	2.2	6.4	9.1
ULIP [64]		5.2	17.0	24.7	3.9	15.0	21.6	2.6	8.3	12.9	1.5	6.4	11.0
OpenShape [30]		<u>56.1</u>	<u>89.0</u>	<u>93.2</u>	<u>54.0</u>	<u>87.9</u>	<u>93.1</u>	11.0	33.3	45.6	9.8	32.2	45.8
Uni3D [72]		44.1	78.2	86.1	48.8	84.3	91.0	8.0	24.0	33.2	10.9	30.2	39.5
PointCLIP v2 [75]	Multi Views	46.9	80.9	88.0	28.3	61.6	73.0	27.3	61.7	75.2	22.0	51.3	61.6
Transformer [54]	Point Cloud	50.0	83.4	90.4	45.7	78.8	86.7	<u>29.9</u>	<u>63.2</u>	<u>78.4</u>	<u>26.6</u>	<u>55.8</u>	<u>65.9</u>
OpenView	+ Multi Views	67.6	92.8	96.4	62.9	91.5	94.6	41.0	77.8	86.1	36.8	71.4	80.2

the top k search results. For zero-shot classification tasks, the metric is top- k accuracy, i.e., the percentage of the correct category appearing in the top k categories.

Baselines. In order to facilitate a comprehensive comparison, we take six previous works [19, 30, 42, 64, 72, 75] as baselines, utilising their open-source models and checkpoints. Furthermore, two new baselines are devised for the comparison of input types: 1) In PointCLIP v2 [75], the projection of the point cloud is replaced with multi-view images. 2) The cross-modal fusion encoder in OpenView was replaced with a base transformer [54] of the same size.

4.3. Zero-Shot Classification

In zero-shot classification task, the evaluation is conducted on three benchmarks: Objaverse-LVIS [10], ModelNet40 [62], Omniobject3D [61]. The zero-shot classification is based on the cosine similarity between the 3D representation and the template-based prompt text representation. For example, the template "a 3D model of {}", which will be filled with category names. The text encoder is used to obtain multiple template representations and perform averaging for classification. To be fair, we use the same template-based prompt as previous work [30, 64, 72].

Table 1 shows the results. It is evident that OpenView outperforms previous work on all three benchmarks. In particular, on the Objaverse-LVIS dataset, which is of high

quality and has a large number of categories, OpenView achieves a new level of classification accuracy, increasing the Top-1, Top-3, and Top-5 accuracy to 52.7%, 74.6%, and 81.1%, respectively. This indicates a more comprehensive understanding of the real-world 3D domain. Furthermore, substituting the cross-modal encoder with a base transformer resulted in an 14.5% reduction in Top-1 accuracy, which reflects the effectiveness of the cross-modal encoder proposed in this paper. On the ModelNet40 dataset, the advantage of OpenView is not significant. We analyze the reason is that the ModelNet40 dataset is a CAD synthetic 3D model with no color and no texture details, and thus the information benefit brought by the multi-view images is not substantial.

4.4. Cross-Modal Retrieval

In cross-modal retrieval task, the 3D models and texts are employed to retrieve each other, thereby validating the broad understanding of 3D domain. It should be noted that to ensure the quality of the evaluation, the textual descriptions in the experiment are annotated by human experts.

Table 2 presents the results. OpenView exhibits a notable advantage over previous work. It is notable that OpenView outperforms the state-of-the-art method by 11.5% and 11.2% on the R@1 metric. This result indicates that OpenView has a superior understanding of the colours and texture details of 3D objects and scenes. Moreover, the new PointCLIP

Table 3. Ablation study. Cross-modal retrieval and zero-shot classification on Objaverse-LVIS [10]. When no extra notes, the experiments in this paper utilize 10 multi-view images. Avg. (Δ) denotes the difference from the average score of the full model.

Variant	Cross-modal Retrieval						Zero-shot Classification			Avg. (Δ)
	Text $\rightarrow$ 3D			3D $\rightarrow$ Text						
	R@1	R@5	R@10	R@1	R@5	R@10	Top-1	Top-3	Top-5	
No multi-view images No point clouds	56.0	88.2	93.0	53.9	86.7	92.1	39.0	60.8	68.9	-7.2
	64.6	90.9	94.0	61.1	88.3	93.6	49.9	70.9	77.6	-1.4
No fusion encoder No sequence-independent modeling	50.0	83.4	90.4	45.7	78.8	86.7	38.2	59.5	67.7	-11.6
	64.7	89.7	93.3	61.7	88.6	92.9	50.0	71.7	77.4	-1.5
No inter-modal attention No intra-modal attention	64.0	89.8	93.7	60.2	87.4	93.1	50.4	72.3	78.2	-1.6
	65.1	90.4	93.9	60.9	89.7	93.5	50.9	72.7	78.4	-0.9
3 multi-view images 6 multi-view images	63.4	88.8	93.5	61.0	87.7	92.7	50.8	72.1	77.8	-1.7
	65.7	91.0	93.9	61.8	89.8	93.7	51.5	72.5	78.6	-0.5
Full Full + progressive hard learning	66.2	91.6	94.3	62.3	90.8	93.9	51.8	73.2	79.3	0.0
	67.6	92.8	96.4	62.9	91.5	94.6	52.7	74.6	81.1	+1.2

v2 [75], which utilises multi-view images, performs better than using point cloud projection. This experiment illustrates the wealth of visual information in multi-view images and underscores the benefits of utilising multi-view images. Similar to the classification task, a reduction in performance was observed when our cross-modal encoder was replaced with a base transformer [54], with an average reduction in R@k of 12.8%. Besides, OpenShape [30] and Uni3D [72] perform poorly on the Omnibject3D [61] dataset, mainly because Omnibject3D does not provide the RGB values of the point cloud, which makes it unable to obtain the colour information for the methods based on the point cloud only.

4.5. Ablation Study

Input of multi-view images and point clouds. The results presented in Table 3 demonstrate that removing the multi-view images or the point clouds data resulted in a notable decline in the model’s performance. Furthermore, removing the multi-view image led to a more pronounced reduction in performance, with $> 10\%$ decreases in both the Top-1 classification and R@1 retrieval metrics. These findings substantiate the visual information benefits that multi-view images provide to the model.

Fusion encoder. The fusion encoder has been replaced with the base transformer [54]. This has resulted in a 13.6% decrease in Top-1 and a 16.2% decrease in R@1. This comparison illustrates the pivotal role of the cross-modal fusion encoder that we developed for OpenView.

Sequence-independent modeling. The sequence-independent modelling introduced in 3.1 is removed and positional embedding [23] is added. This variation results in a reduction in model performance, which is primarily due to the fact that multi-view images are typically unordered.

Inter-modal and intra-modal attention. We remove inter-modal and intra-modal attention respectively to test their effectiveness. The model performance all decreased, demonstrating the necessity of modeling inter-modal relationships and intra-modal dependencies.

Number of multi-view images. In the majority of experiments in this paper, the number of multi-view images is 10. As the number of images decreases, there is a slight decline in the model performance, indicating that OpenView is not sensitive to the number of multi-view images.

Progressive hard learning. To enable the model to focus on global features, we developed progressive hard learning introduced in 3.4. By gradually reducing the number of easy-to-understand images in training, the model’s ability to understand point clouds and ambiguous images is enhanced.

5. Applications

Text-query and image-query 3D Retrieval. Figure 3a and 3b shows the retrieval results. The 3D, text and image representation vectors are extracted using OpenView, the text encoder and image encoder of OpenCLIP [7], respectively. Their similarity is measured using the cosine similarity measure. This application utilises texts and images to retrieve 3D objects in the Objaverse [10] dataset respectively, demonstrating the wide range of understanding of OpenView.

3D captioning. Figure 3c shows the 3D captioning results. The 3D embedding space of OpenView is aligned with the text-image embedding space of the CLIP [43] model, enabling the integration of OpenView with CLIP-based captioning models [37] to implement 3D captioning. In particular, we utilise CLIP ViT-B/32 to construct OpenView to adapt the downstream ClipCap [37] caption model.

Hierarchical Clustering. Figure 3d shows some examples of data clustering results, utilising the t-SNE technique [53]. The semantic embedding extracted by OpenView can be employed to organize the data hierarchically. In this example, furniture and transport are two larger categories. By clustering, the subcategories table, chair, car, and bicycle can be separated, which can assist in optimising the retrieval complexity of real-world databases [16, 68].

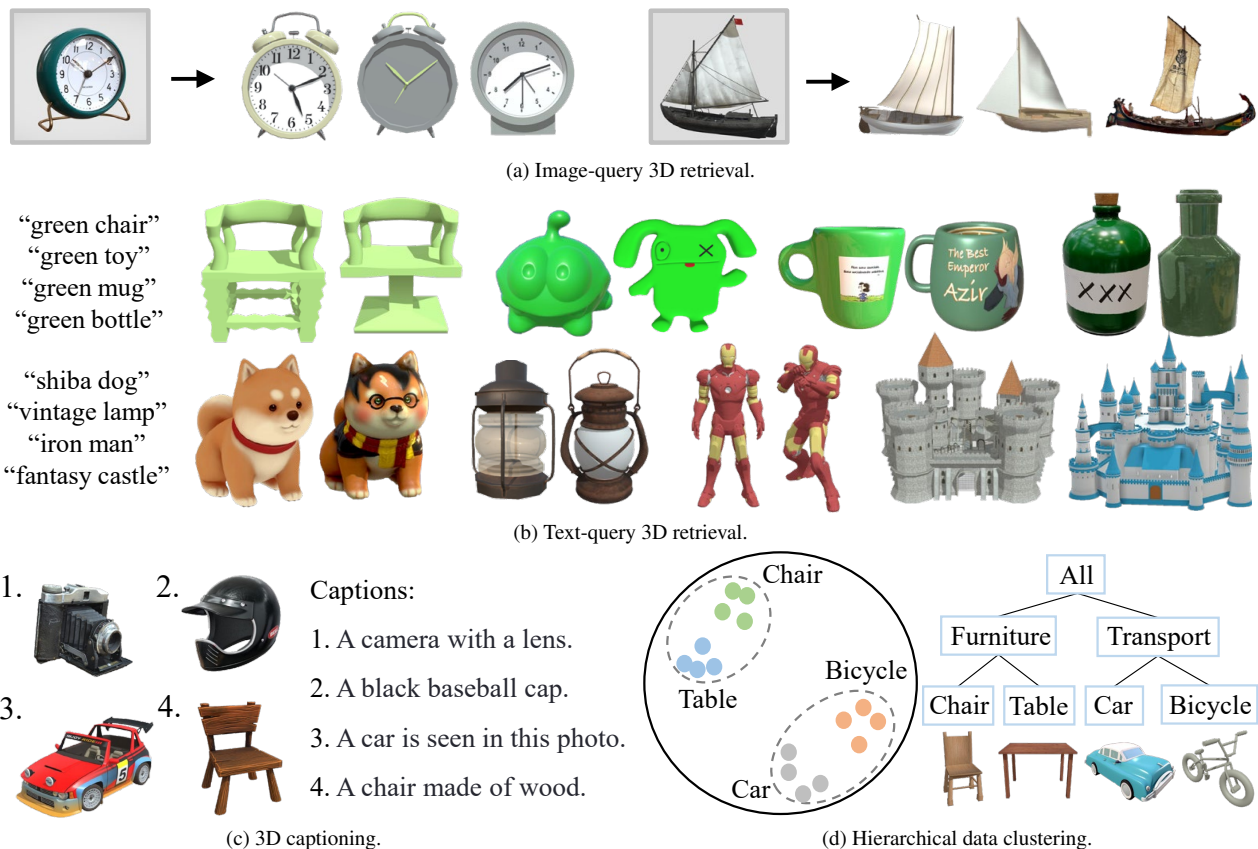


Figure 3. Applications. OpenView can be employed in a wide range of applications. (a) Image-query 3D retrieval. Retrieve 3D objects based on the input images. (b) Text-query 3D retrieval. Retrieve 3D objects based on the input texts. (c) 3D captioning. Generate text descriptions of the input 3D objects. (d) Hierarchical data clustering. Utilize 3D embedding to hierarchically cluster and organize data.

6. Discussion

Does adding multi-view images to 3D representation introduce complexity in data collection? Multi-view images are one of the common data formats for 3D perception. For scenarios such as 3D games, e-commerce products and virtual reality, the 3D models are always visualised for the user, which makes it straightforward to collect multi-view images. Even in the autonomous driving scenarios, Tesla is explicitly based on the technical route of multi-view images [6].

Does the ambiguity of some images affect the understanding of the model? One example of ambiguity is that a wall lamp may appear like a frying pan from above. In order to mitigate the effects of ambiguity, we have developed a progressive hard learning strategy to enhance the understanding of these ambiguous images.

Why not utilize a large language model [28, 36] as the backbone? Instead of stacking large-scale model parameters and training data, our research focuses on exploring a fusion framework for multi-view images and point clouds. Our framework and idea can be leveraged by future work to incorporate 3D modalities into large language models.

Does the addition of multi-view images have an impact on the robustness of the model? In contrast to previous work [30, 64, 65, 72] in the field of 3D semantic representation, our method adds the input of multi-view images. However, neural networks are vulnerable to adversarial attacks [1]. For the image modality, researchers have proposed many adversarial attack [26] and defense [74] methods. Thus, it is necessary to explore the effect of adding multi-view images on model robustness in future work.

7. Conclusion

This work presents OpenView, an innovative framework for learning unified multi-modal 3D representations by fusing point clouds and multi-view images. Experiments indicate that OpenView outperforms the state-of-the-art in text-based retrieval and zero-shot classification tasks. Furthermore, OpenView aligns with the image-text embedding space of CLIP, thereby enabling a wide range of cross-modality applications, such as 3D retrieval, 3D captioning and hierarchical clustering for database management, highlighting its versatility in the field of 3D representation learning.

References

- [1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017.
- [2] 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.
- [3] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013.
- [4] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 423–432. 2023.
- [5] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. *ACM Transactions on Graphics (TOG)*, 28(3):1–12, 2009.
- [6] Bojing Cheng, Hongye Liu, Ying Wang, and Yongchao Liu. Application prospect of high-precision map mode and pure vision mode in automatic driving. In *International Conference on Artificial Intelligence and Security*, pages 86–94. Springer, 2022.
- [7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023.
- [8] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7781–7790, 2019.
- [9] 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 (CVPR)*, pages 21126–21136, 2022.
- [10] 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 (CVPR)*, pages 13142–13153, 2023.
- [11] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 602–618, 2018.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.
- [13] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1536–1546, 2022.
- [14] 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 (IJCV)*, 129:3313–3337, 2021.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [16] Alexandra Henzinger, Emma Dauterman, Henry Corrigan-Gibbs, and Nickolai Zeldovich. Private web search with tiptoe. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 396–416, 2023.
- [17] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christoffer Petersson, and Lennart Svensson. Masked autoencoder for self-supervised pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 350–359, 2023.
- [18] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- [19] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22157–22167, 2023.
- [20] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9778–9787, 2019.
- [21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 18661–18673, 2020.
- [22] Stephane Laveau and Olivier D Faugeras. 3-d scene representation as a collection of images. In *Proceedings of 12th International Conference on Pattern Recognition (ICPR)*, pages 689–691. IEEE, 1994.
- [23] Weixian Lei, Yixiao Ge, Jianfeng Zhang, Dylan Sun, Kun Yi, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. *arXiv preprint arXiv:2308.10185*, 2023.
- [24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022.
- [25] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision—ECCV 2020: 16th European Conference*,

- Glasgow, UK, August 23–28, 2020, *Proceedings, Part XIV 16*, pages 677–693. Springer, 2020.
- [26] Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su, and Hui Xue. Qair: Practical query-efficient black-box attacks for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3330–3339, 2021.
 - [27] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 32(8):3412–3432, 2020.
 - [28] Zi Liang, Pinghui Wang, Ruofei Zhang, Nuo Xu, Shuo Zhang, Lifeng Xing, Haitao Bai, and Ziyang Zhou. Merge: Fast private text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 19884–19892, 2024.
 - [29] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023.
 - [30] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.
 - [31] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.
 - [32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023.
 - [33] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.
 - [34] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
 - [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018.
 - [36] Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, et al. Debate on graph: a flexible and reliable reasoning framework for large language models. *arXiv preprint arXiv:2409.03155*, 2024.
 - [37] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
 - [38] 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 (ECCV)*, pages 604–621. Springer, 2022.
 - [39] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017.
 - [40] 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 (CVPR)*, pages 652–660, 2017.
 - [41] 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 (NeurIPS)*, 30, 2017.
 - [42] 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)*, pages 28223–28243. PMLR, 2023.
 - [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.
 - [44] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5385, 2020.
 - [45] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 626–642. Springer, 2020.
 - [46] Steven M Seitz and Charles R Dyer. Physically-valid view synthesis by image interpolation. In *Proceedings IEEE Workshop on Representation of Visual Scenes (In Conjunction with ICCV’95)*, pages 18–25. IEEE, 1995.
 - [47] Steven M Seitz and Charles R Dyer. View morphing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 21–30, 1996.
 - [48] Bokui Shen, Xinchun Yan, Charles R Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, and Dragomir Anguelov. Gina-3d: Learning to generate implicit neural assets in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4913–4926, 2023.
 - [49] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.
 - [50] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, pages 2–13. SPIE, 2000.
 - [51] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision (IJCV)*, 130(6):1526–1565, 2022.

- [52] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics.
- [53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- [55] 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 (ICCV)*, pages 9782–9792, 2021.
- [56] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.
- [57] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2021.
- [58] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8534–8543, 2021.
- [59] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–1893. IEEE, 2018.
- [60] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9065–9075, 2023.
- [61] 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 (CVPR)*, pages 803–814, 2023.
- [62] 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 (CVPR)*, pages 1912–1920, 2015.
- [63] 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.
- [64] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 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 (CVPR)*, pages 1179–1189, 2023.
- [65] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023.
- [66] Honghong Yang, Hongxi Liu, Yumei Zhang, and Xiaojun Wu. Fmr-gnet: Forward mix-hop spatial-temporal residual graph network for 3d pose estimation. *Chinese Journal of Electronics*, 33(6):1346–1359, 2024.
- [67] Liying Yang, Zhenwei Zhu, Xuxin Lin, Jian Nong, and Yanyan Liang. Long-range grouping transformer for multi-view 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18257–18267, 2023.
- [68] Zhibo Yang, Muhammet Bastan, Xinliang Zhu, Douglas Gray, and Dimitris Samaras. Hierarchical proxy-based loss for deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1859–1868, 2022.
- [69] Wang Yifan, Carl Doersch, Relja Arandjelović, Joao Carreira, and Andrew Zisserman. Input-level inductive biases for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6176–6186, 2022.
- [70] 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 (CVPR)*, pages 19313–19322, 2022.
- [71] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10252–10263, 2021.
- [72] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.
- [73] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12588–12597, 2023.
- [74] Ziyang Zhou, Pinghui Wang, Zi Liang, Ruofei Zhang, and Haitao Bai. Pair: Pre-denoising augmented image retrieval model for defending adversarial patches. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 5771–5779, 2024.

- [75] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2639–2650, 2023.