

Describe, Adapt and Combine: Empowering CLIP Encoders for Open-set 3D Object Retrieval

Zhichuan Wang¹ Yang Zhou² Zhe Liu³ Rui Yu⁴ Song Bai⁵

Yulong Wang^{1*} Xinwei He^{1*} Xiang Bai⁶

¹Huazhong Agricultural University ²Shenzhen University ³The University of Hong Kong

⁴University of Louisville ⁵ByteDance ⁶Huazhong University of Science and Technology

wzc_65@webmail.hzau.edu.cn, xwhe@hzau.edu.cn

Abstract

Open-set 3D object retrieval (3DOR) is an emerging task aiming to retrieve 3D objects of unseen categories beyond the training set. Existing methods typically utilize all modalities (i.e., voxels, point clouds, multi-view images) and train specific backbones before fusion. However, they still struggle to produce generalized representations due to insufficient 3D training data. Being contrastively pre-trained on web-scale image-text pairs, CLIP inherently produces generalized representations for a wide range of downstream tasks. Building upon it, we present a simple yet effective framework named Describe, Adapt and Combine (DAC) by taking only multi-view images for open-set 3DOR. DAC innovatively synergizes a CLIP model with a multi-modal large language model (MLLM) to learn generalized 3D representations, where the MLLM is used for dual purposes. First, it describes the seen category information to align with CLIP’s training objective for adaptation during training. Second, it provides external hints about unknown objects complementary to visual cues during inference. To improve the synergy, we introduce an Additive-Bias Low-Rank adaptation (AB-LoRA), which alleviates overfitting and further enhances the generalization to unseen categories. With only multi-view images, DAC significantly surpasses prior arts by an average of +10.01% mAP on four open-set 3DOR datasets. Moreover, its generalization is also validated on image-based and cross-dataset setups. Code is available at <https://github.com/wangzhichuan123/DAC>.

1. Introduction

Retrieving 3D objects from a given large 3D repository is an important task in 3D computer vision, whose applications include virtual reality [40], computer-aided design [16], and

*Corresponding author

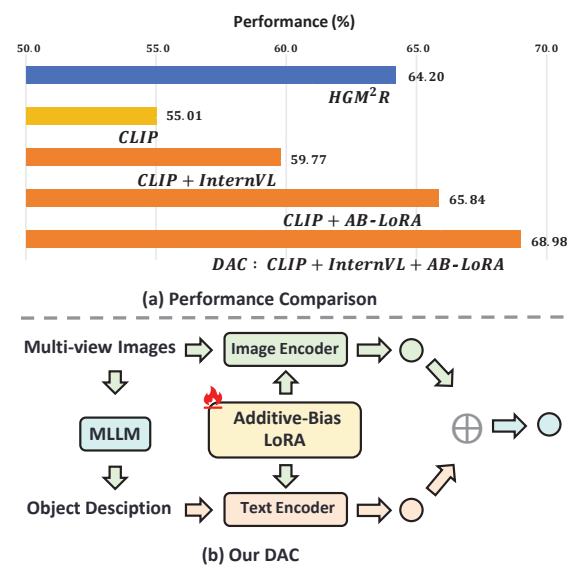


Figure 1. We empower CLIP encoders with an MLLM model (e.g., InternVL [8]) for open-set 3D object retrieval, which significantly outperforms previous state-of-the-art HGM²R [15] (a). Unlike HGM²R, which utilizes all modalities (point cloud, voxel, and view images) and involves test data for training, our method relies solely on multi-view inputs and eliminates test data (b).

3D cultural heritage preservation [31]. Existing 3D object retrieval (3DOR) algorithms [2, 12, 22, 28, 37, 42, 49, 58, 59] usually learn discriminative representations on various 3D data formats (e.g., view images [51], voxels [60], and point clouds [45]). Despite progress, the recent *SHape REtrieval Contest (SHREC) 2022* competition [14] has shown that these methods have great difficulties in open-set scenarios, where the test labels are unseen during training. Recently, HGM²R [15] uses all modalities and incorporates test data for training. Although improvement is achieved, it is complex and impractical in real-world applications.

The challenges of open-set 3DOR are twofold. First, ex-

isting methods typically assume the same domain and category shared by training and testing data, resulting in low performance in open-set scenarios. Second, these methods tend to overfit a handful of known categories due to the *limited* 3D data, leading to poor generalization to unseen ones.

To tackle the issue, we are motivated by the fact that existing pre-trained large vision and language models, such as CLIP, can produce generalized and effective embeddings for a variety of data-scarce downstream tasks [17, 26, 48]. We thus first establish a multi-view CLIP baseline to investigate its potential by plainly aggregating view-wise features from CLIP image encoder. This simple baseline empirically gives promising results, as shown in Figure 1.

But imagine how humans react when faced with a 3D object of an unknown category. Typically, we examine it from various angles to collect visual cues for inference. Yet, in addition to visual analysis, we can articulate descriptions of the object using language, thereby engaging in textual reasoning. For instance, “a horse-like animal but with black and white stripes”, these descriptive sentences provide important attributes and greatly facilitate inference (feature matching) of unknown categories. This fact encourages us to further involve textual descriptions of unknown 3D objects to visual features.

In this paper, therefore, we introduce **Describe, Adapt, and Combine (DAC)**, a novel framework empowering CLIP encoders for open-set 3DOR by combining Multi-Modal Large Language Models (MLLM). Specifically, inputting multi-view images, DAC consists of a three-step process:

1) **Describe:** We utilize a pretrained MLLM model (e.g., InternVL [8]) to acquire textual information. During training, the MLLM is used to derive expressive category descriptions about the objects, which better aligns with the training objective of CLIP in adaptation. During inference time, we only use it to describe the appearance and semantic cues encoded in the multi-view images, thereby obtaining out-of-box textual knowledge for the feature combination.

2) **Adapt:** Paired multi-view images with enriched descriptions of seen categories, we introduce Low-Rank Adaptation (LoRA) in both text and visual encoders for efficient adaptation. This step effectively mitigates the domain gap between view projections and natural images with which CLIP is trained. However, we observed that LoRA always easily overfits the seen categories. We analyze it (see Sec. 3.2) and further introduce an additive bias, which effectively enhances the generalization to unseen categories.

3) **Combine:** After CLIP adaptation, we combine both the multi-view image and textual features for shape retrieval. Thanks to the aligned feature space of CLIP for image and text inputs, we can directly embed the acquired descriptions via CLIP encoder and fuse the visual and textual cues in CLIP space via simple operations like addition.

DAC demonstrates strong performance across four exist-

ing open-set 3DOR benchmarks on five retrieval tasks. In particular, it surpasses state-of-the-art by over +10% mAP on average in open-set 3D object retrieval.

In summary, we make the following contributions:

- We propose a simple yet effective framework named DAC for open-set 3DOR. To our best knowledge, it represents the first attempt at synergizing an existing pretrained vision-language model and a multi-modal large language model to effectively address 3D data scarcity and unknown category challenges in open-set 3DOR.
- To maximize its potential, we introduce a scheme for multi-view image adaptation. It makes use of enriched category descriptions to enhance text-image alignment through contrastive learning. Furthermore, we insert an additive bias to LoRA for better adaptation, effectively improving the generalization to unseen test categories.
- DAC performs remarkably in open-set 3DOR with limited 3D data. In addition, it also effectively extends to various other 3DOR tasks like cross-dataset, single-image-based, and even zero-shot point cloud retrieval.

2. Related Work

3D Object Retrieval. 3D object retrieval is a long-standing task in computer vision [52]. While early research focused on hand-crafted descriptors [5, 6, 30], modern works generally center around learning representations upon 3D data collections. *One group* learns based on raw 3D formats, including voxel grids [41, 53, 60], point clouds [9, 18, 38, 45, 46, 56, 65], or meshes [13, 33]. Although they can leverage 3D geometric information, they often face computation challenges, e.g., cubic computation complexity for voxels and the irregular and unordered issue of point clouds and meshes. *Another group* [10–12, 19–21, 29, 32, 51, 57] focuses on learning to aggregate multi-view features, exhibiting superior potential than 3D methods. Besides, images are 2D grid data, which is more computationally efficient.

Despite great progress, all above methods concentrate on closed-set conditions and do not generalize well to open-set settings involving unseen 3D object categories, as per SHREC’22 [14]. Subsequently, Feng *et al.* [15] proposed jointly learning multi-modal embeddings, including multi-view images, point clouds, and voxels, and utilizing hypergraphs to connect seen and unseen category data. Although it leads to improved generalization, involving multi-modal and test set data brings extra complexity. In this work, we take a *different* route. By leveraging current large vision and language models, we address the generalization challenges in open-set 3DOR, *solely* based on multi-view images.

Transferring Vision-Language Models in 3D. Pre-trained by large-scale image-text pairs, large vision-language models (VLM) (particularly CLIP [47]) can be utilized to produce discriminative representations for 2D depth/view im-

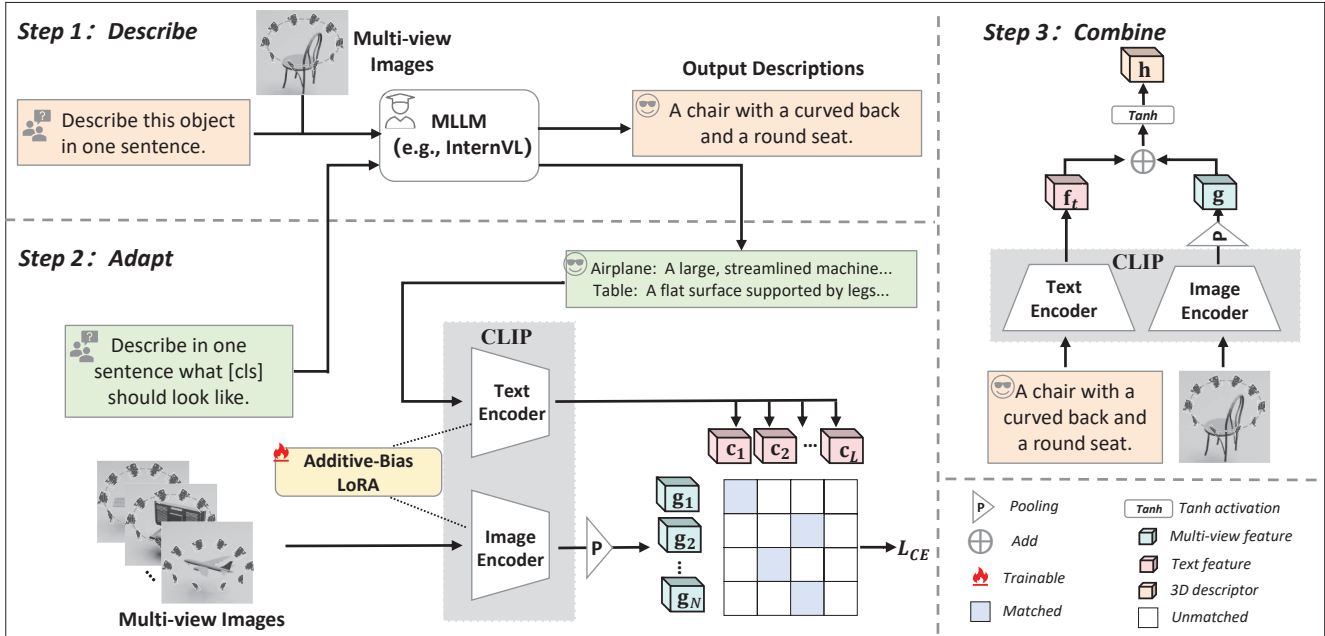


Figure 2. **Overview of DAC.** Given a 3D object, we first project it into multi-view images and utilize a pretrained MLLM (*e.g.*, InternVL) to describe the shape information. Next, we efficiently finetune a pretrained CLIP ViT model with a additive-bias LoRA to adapt to the multi-view projections. Finally, we combine the textual and multi-view embeddings to derive 3D descriptors for retrieval.

ages of 3D objects. One simple way is to use these pretrained models directly in a zero-shot manner. For instance, MV-CLIP [50] directly extracts multi-view features with pretrained CLIP’s visual encoder, while works like ULIP [61] and ULIP-2 [62] incorporate unified embeddings that align language, image, and point data to improve 3D shape analysis across modalities. OpenShape [36] builds on CLIP-based embeddings to jointly learn image, text, and 3D shape features, facilitating robust cross-modal retrieval in open-set scenarios. Some works [26, 64, 70] apply CLIP to process 2D depth maps of point clouds, reaching promising results. Even so, these methods are still restrained in performance by the amount of 3D data.

When a certain amount of downstream training data is available, it is always desirable to finetune the pretrained models to adapt to the downstream tasks for optimized results. Some works, *e.g.*, CoOp [69], CLIP-Adapter [17], and Tip-Adapter [63], learn a minuscule adapter while keeping the pretrained main model frozen. CG3D [23] replaces the handcrafted prompts with learnable ones. In this paper, we propose to use Low-Rank Adaptation (LoRA) [24] but add a simple bias term, enabling efficiently adapts to multi-view projection images of known categories with reduced overfitting risk. In addition, it brings negligible testing latency after merging.

Multi-model Large Language Models. Multi-model Large Language Models (LLMs) [1, 3, 4, 7, 8, 34, 35, 39, 44, 54, 55] possess excellent reasoning abilities on vi-

sual content, driving progress in generic visual-linguistic tasks, such as image captioning, visual question answering (VQA), and visual dialogue. Over the past few years, many commercial and open-source MLLMs have been introduced, *e.g.*, GPT-4 [1], LLaVA series [34, 35, 39] and many others [4, 7, 44, 54, 55]. In particular, InternVL [8] has made significant strides in enhancing the interaction between visual and language components, which is essential for tasks that require detailed object understanding. For open-set 3D object retrieval, the ability of MLLMs to generate high-level semantic descriptions is particularly beneficial. By combining these descriptions with fine-grained visual features, our model can *generalize better* to unseen categories, addressing the key challenge in open-set 3DOR.

3. Method

In open-set 3D object retrieval, our goal is to learn discriminative, more importantly, generalizable embeddings for 3D objects of unseen categories. Formally, let $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_t}$ be a training set consisting of N_t 3D objects. Here \mathbf{x}_i represents i -th 3D object, y_i are the corresponding known labels. Based on it, we aim to train an embedding network $\mathcal{E}(\cdot)$ to produce generalizable embeddings for a retrieval set $\mathcal{D}_{\text{ret}} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^{N_r}$, which consists of N_r 3D objects belonging to unseen categories. The retrieval set \mathcal{D}_{ret} is further split into the query and target sets, denoted by $\mathcal{D}_{\text{query}}$ and $\mathcal{D}_{\text{target}}$. Note that the label \hat{y}_i from \mathcal{D}_{ret} and y_i

from $\mathcal{D}_{\text{train}}$ are drawn from disjoint label sets.

Existing works [14, 15] deal with this task by taking advantage of multi-modal inputs of 3D objects (voxels, point clouds, multi-view images). By contrast, our approach simplifies the paradigm by only utilizing view images. We first project each 3D object \mathbf{x}_i into a set of 2D view images from different viewpoints following the same scheme as HGM²R [15]. We produce M 2D grey-scale images $\{I_{i,m} \in \mathbb{R}^{1 \times H \times W}\}_{m=1}^M$, where H and W represent the height and width of the view image $I_{i,m}$, respectively. These multi-view projections comprehensively capture the object from various angles, ensuring the 3D structure is well-represented. Based solely on multi-view images, we present a simple yet effective framework named **Describe, Adapt, and Combine (DAC)** (Figure 2), which takes advantage of existing large multi-modal models and turn them into a strong 3D embedding learner for open-set 3D object retrieval. DAC offers a new alternative paradigm for open-set 3D representation learning, structured as a three-step process: Describe, Adapt, and Combine. In the following, we describe each part in detail.

3.1. Describe: Describe 3D objects

For this step, we utilize an MLLM to generate text descriptions. Note that any off-the-shelf MLLM can be used (see Appendix I for more choices). Here we mainly use InternVL [8] due to its superior instruction-following ability in open-ended tasks like visual question answering. We leverage the MLLM for *dual purposes*.

To better align with CLIP’s contrastive pretraining objective on image-text pairs, we first use the MLLM to enrich the category with detailed descriptions. For a training set $\mathcal{D}_{\text{train}}$ with L classes $\{l_i\}_{i=1}^L$, we use prompt: “Describe in one sentence what [cls] should look like”, replacing [cls] with each of the L categories, to obtain L sentences $\{t_i\}_{i=1}^L$. Each sentence serves as a description of the category, involving in the Adapt step to fine-tune CLIP to align with the multi-view images. Compared with handcrafted prompts using templates “a photo of [cls]”, using detailed descriptions is conducive to producing more generalized representations reflecting fine-grained aspects of categories.

We further enhance the inference stage by integrating MLLM-generated sentences as external hints. Given multi-view images, we feed them into pretrained InternVL models [8] together with multi-view information-gathering prompt: “There are images of an object from different angles. Describe this object in one sentence.” The prompt instructs the models to generate descriptions s_i to summarize the object’s appearances and semantic features across all views. They are further processed by the adapted CLIP to generate textual embeddings, further improving generalization to unseen categories.

By integrating textual guidance from the MLLM dur-

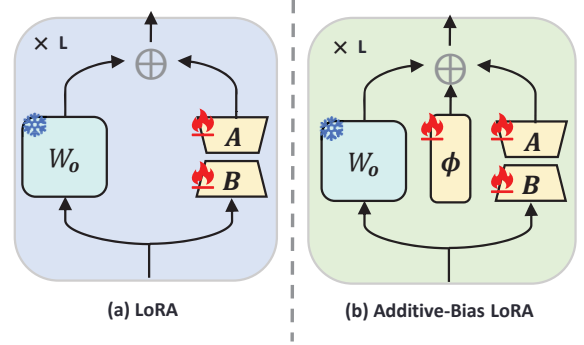


Figure 3. Comparison of LoRA and Additive-Bias LoRA.

ing both training and inference, we effectively combine the strengths of MLLM and CLIP for open-set 3DOR, avoiding somewhat impractical usage of test data for training [15].

3.2. Adapt: Adapt CLIP to Multi-view Images

The next step is to embed the multi-view images and the attached category descriptions for retrieval. To achieve this goal, we build upon the released CLIP model [47] which naturally supports multimodal inputs and embeds them into aligned and discriminative feature space.

Nevertheless, CLIP is primarily pretrained on natural images, which are quite different from multi-view projections. To extract suitable 3D representations, it is essential to adapt to the new distributions. To this end, we first introduce Low-Rank Adaptation (LoRA) [24] with *additive bias* and insert them into the fixed CLIP model in DAC for adaption.

A Revisit to LoRA. Low-Rank Adaptation (LoRA) [24] rewrites the *updates* ($\Delta\mathbf{W}$) to the pretrained weight matrix $\mathbf{W}_o \in \mathbb{R}^{d_1 \times d_2}$ of a linear layer into a product of two small low-rank matrices, denoted by $\mathbf{A} \in \mathbb{R}^{r \times d_2}$ and $\mathbf{B} \in \mathbb{R}^{d_1 \times r}$: $\mathbf{W}_o + \gamma\Delta\mathbf{W} = \mathbf{W}_o + \gamma\mathbf{BA}$. Here \mathbf{A} and \mathbf{B} forms a LoRA module (see Figure 3 (a)), the rank $r \ll \min(d_1, d_2)$, and γ is a scaling factor. Note that we only train \mathbf{A} and \mathbf{B} and keep \mathbf{W}_o frozen. \mathbf{A} and \mathbf{B} can also be viewed as a reparametrization and merged into \mathbf{W}_o readily. Given the input $\mathbf{z} \in \mathbb{R}^{d_2}$, then the output $\mathbf{o} \in \mathbb{R}^{d_1}$ to a linear layer with LoRA is formulated as: $\mathbf{o} = \mathbf{W}_o\mathbf{z} + \gamma\Delta\mathbf{W}\mathbf{z} = \mathbf{W}_o\mathbf{z} + \gamma\mathbf{BA}\mathbf{z}$.

Additive-bias LoRA. For open-set 3DOR, we face a severe risk of overfitting on seen categories due to limited training data, which often causes poor generalization to testing unseen categories. To alleviate it, we rethink LoRA and propose to inject a “bias” inside, which has been widely overlooked but has unreasonable effectiveness for improving unseen category generalization. Given loss \mathcal{L} , the gradient with $\Delta\mathbf{W}$ as a whole is: $\frac{\partial\mathcal{L}}{\partial\Delta\mathbf{W}} = \gamma \left(\frac{\partial\mathcal{L}}{\partial\mathbf{o}}\right) \mathbf{z}^T$. Following the SGD update rule: $\Delta\mathbf{W} \leftarrow \Delta\mathbf{W} - \eta \left(\gamma \left(\frac{\partial\mathcal{L}}{\partial\mathbf{o}}\right) \mathbf{z}^T\right)$, where η is the learning rate. Note that $\Delta\mathbf{W}$ directly accumulates information from \mathbf{z} , which comes from *seen categories*. To

break this tight linkage, we add a bias term $\Phi \in \mathbb{R}^{d_1}$ to prevent the model from being overfitted quickly and maintain generalization to *unseen categories* as much as possible:

$$\mathbf{o} = \mathbf{W}_o \mathbf{z} + \gamma \mathbf{B} \mathbf{A} \mathbf{z} + \Phi. \quad (1)$$

To initialize, we use a standard normal distribution for \mathbf{A} and set \mathbf{B} and Φ to zero, which ensures that the model starts without perturbations to the original pre-trained weights.

In our architecture, we use this LoRA variant in both the textual and visual encoders of CLIP ViT models (see Figure 3 (b)). Following LoRA [24], we limit LoRA modules to the weights of the self-attention modules (*i.e.*, $\mathbf{W}_{\text{query}}$, \mathbf{W}_{key} , $\mathbf{W}_{\text{value}}$).

Training Objective. We train both textual and visual encoders in CLIP with contrastive learning, enabling better alignment between multi-view images and their associated category descriptions, with the limited 3D training data. Given L sentences $\{t_i\}_{i=1}^L$ from the first step, we generate L classification weights by forwarding each into the CLIP textual encoder:

$$\mathbf{c}_i = \mathcal{T}(t_i) \in \mathbb{R}^d, i = 1, \dots, L. \quad (2)$$

Next, we feed the multi-view images $\{I_{k,m}\}_{m=1}^M$ of 3D object \mathbf{x}_k into CLIP visual encoder $\mathcal{V}(\cdot)$:

$$\mathbf{f}_{k,m} = \mathcal{V}(I_{k,m}) \in \mathbb{R}^d, \quad (3)$$

This will give us M view-wise representations $\{\mathbf{f}_{k,m}\}_{m=1}^M$. We further aggregate them into a global compact descriptor for the 3D object via mean-pooling:

$$\mathbf{g}_k = \frac{1}{M} \sum_{m=1}^M \mathbf{f}_{k,m}. \quad (4)$$

Finally, we apply cross-entropy loss to supervise the training process of LoRA, ensuring that the global features $\{\mathbf{g}_k\}$ are aligned with their corresponding text descriptions $\{\mathbf{c}_y\}$ closely in the embedding space:

$$\mathcal{L}_{CE} = -\frac{1}{N_t} \sum_{k=1}^{N_t} \log \frac{\exp(\mathbf{g}_k \cdot \mathbf{c}_y / \tau)}{\sum_{i=1}^L \exp(\mathbf{g}_k \cdot \mathbf{c}_i / \tau)}. \quad (5)$$

where τ is the temperature, and y is the ground-truth category. By efficiently fine-tuning additive bias LoRA with limited 3D training objects, we adapt pre-trained CLIP ViT models to multi-view projections with minimal effort, while demonstrating excellent performance.

3.3. Combine: Combine Textual-Visual Features

After the processes of description and adaptation, we proceed to combine textual and visual embeddings to form discriminative and generalized descriptors for 3D objects. For

this, we utilize the finetuned CLIP model in Sec. 3.2 to perform feature extraction for the two modalities (*i.e.*, descriptive text and multi-view images).

With the generated description s and multi-view images $\{I_m\}$ of 3D objects, we easily derive each modality’s global representations by forwarding them to the adapted CLIP ViT models: $\mathbf{f}_t = \mathcal{T}(s)$, $\mathbf{f}_m = \mathcal{T}(I_m)$. Here we aggregate the multi-view representations into \mathbf{g} with a simple mean-pooling layer. Then, we propose a weighted fusion scheme, which balances visual and textual features with a scalar and then normalizes to enhance discriminativeness:

$$\mathbf{h} = \tanh(\mathbf{g} + \alpha \mathbf{f}_t). \quad (6)$$

where $\alpha \in [0, 1]$ is a weighting factor. The \tanh activation function is applied for normalization after the fusion of \mathbf{g} and \mathbf{f}_t . Finally, $\mathbf{h} \in \mathbb{R}^d$ represents the final descriptor for the 3D object. Without bells and whistles, this simple fusion method gives us surprisingly good performance. For retrieval, we adopt cosine similarity as the metric function.

3.4. Discussions

Why not directly use InternVL for retrieval? While InternVL excels at multi-modal generative tasks, its pretraining is not specifically geared toward producing discriminative features suitable for retrieval. In contrast, CLIP’s encoder is designed to produce more discriminative features, which are crucial for distinguishing between categories in open-set scenarios. Empirical evidence from OS-ESB-core shows that using InternVL’s embeddings led to an inadequate 38.20% mAP, much lower than CLIP’s 53.93% mAP. **Extending to other 3DOR.** DAC indeed represents a general framework and can be easily extended to a variety of 3DOR problems such as cross-domain retrieval. By leveraging semantic descriptions from an MLLM and fine-tuning discriminative models like CLIP, DAC can even handle scenarios where domain shifts present. This simple synergy showcases strong performance with minimal complexity, making it a versatile solution for various 3D retrieval tasks.

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics. We conduct extensive experiments on four public open-set 3DOR datasets [15]: OS-ESB-core, OS-NTU-core, OS-MN40-core, and OS-ABO-core. Each dataset is divided into training, probe, and gallery sets. The training set consists of seen categories for model training, while the probe and gallery sets contain unseen categories for evaluation. Following HGM²R [15], we adopt mean Average Precision (mAP), Normalized Discounted Cumulative Gain (NDCG) and Average Normalized Modified Retrieval Rank (ANMRR) metrics to report the retrieval performance. For both mAP and NDCG, higher values are better; while for ANMRR, lower is better.

Method	Modality	OS-ESB-core			OS-NTU-core			OS-MN40-core			OS-ABO-core		
		mAP↑	NDCG↑	ANMRR↓	mAP↑	NDCG↑	ANMRR↓	mAP↑	NDCG↑	ANMRR↓	mAP↑	NDCG↑	ANMRR↓
Zero-shot Setup													
OpenShape (PB-CLIP B/32) [36]	P.	37.64	18.38	65.57	25.53	15.41	74.51	30.31	46.34	67.01	40.29	46.09	59.42
OpenShape (PB-CLIP L/14) [36]	P.	38.58	18.81	65.10	24.71	15.02	75.18	29.64	44.79	67.64	38.65	45.56	60.90
ULIP-2 (PB-CLIP G/14) [62]	P.	45.14	21.00	59.15	31.50	17.89	68.99	32.76	48.92	65.22	44.26	49.04	55.64
Uni3D (Uni3D-Giant) [68]	P.	44.42	20.92	59.96	32.02	18.04	68.49	33.21	50.51	65.11	45.92	49.79	53.96
MV-CLIP [‡] (CLIP B/32) [50]	I.	46.74	21.72	57.34	47.78	23.69	54.48	52.46	66.24	48.33	52.55	54.95	48.33
MV-CLIP [‡] (CLIP L/14) [50]	I.	49.81	22.75	53.71	<u>57.71</u>	<u>26.46</u>	<u>45.25</u>	63.74	74.85	37.80	<u>63.07</u>	<u>59.18</u>	<u>38.67</u>
Our DAC (CLIP B/32)	I.	56.16	23.58	48.39	55.03	25.70	48.32	55.39	68.08	45.96	60.77	56.28	41.44
Our DAC (CLIP L/14)	I.	56.60	23.94	47.61	61.33	27.53	41.96	<u>59.77</u>	<u>72.08</u>	<u>41.46</u>	65.93	59.04	36.60
Open-set Setup													
TCL [20]	P., I., V.	49.31	21.89	52.68	39.37	21.23	61.00	48.11	63.83	52.30	49.33	53.86	51.05
SDML [25]	P., I., V.	49.59	21.75	52.36	40.16	21.52	60.49	50.75	65.70	50.22	47.44	52.79	52.42
CMCL [28]	P., I., V.	50.01	21.97	53.06	41.08	21.72	59.43	51.38	65.98	49.75	49.83	50.89	50.24
MMSAE [59]	P., I., V.	49.88	22.06	53.69	40.85	21.70	59.99	52.08	66.57	49.00	50.51	53.80	50.49
MCWSA [66]	P., I., V.	49.48	21.34	53.75	39.22	20.69	62.14	48.78	63.85	51.95	45.61	51.05	54.70
PROSER [67]	P., I., V.	48.69	21.13	53.95	39.47	21.24	60.96	49.00	64.54	51.66	50.33	53.27	50.34
InfoNCE [43]	P., I., V.	50.26	21.91	52.63	40.03	21.19	61.09	47.37	63.31	53.02	46.83	52.14	53.50
HGM ² R [15]	P., I., V.	51.74	22.73	51.28	44.88	22.81	56.67	<u>64.20</u>	<u>72.91</u>	<u>38.27</u>	63.39	57.96	37.96
Our DAC (CLIP B/32)	I.	58.70	24.27	45.67	<u>59.21</u>	<u>27.06</u>	<u>44.58</u>	<u>62.40</u>	<u>72.63</u>	<u>39.82</u>	<u>66.10</u>	<u>59.01</u>	<u>36.12</u>
Our DAC (CLIP L/14)	I.	<u>57.80</u>	<u>24.36</u>	<u>47.44</u>	65.83	28.78	37.46	68.98	77.59	33.87	70.74	60.87	32.14

Table 1. **Performance comparisons (%) on open-set 3DOR benchmarks.** **Bold** and underline indicate the best and second best results, respectively. [‡] in gray means we provide *ground-truth* category sets for the view selection process in MV-CLIP, which is impractical, as category information for unseen classes is unknown in open-set scenarios. PB-CLIP refers to PointBERT-CLIP, while P, I., and V. stand for Point Cloud, Multi-view Images, and Voxel, respectively.

Implementation Details. For a fair comparison, we follow the same rendering scheme as HGM²R [15] and project 24 view images of size 256×256 for each 3D object, for all the experiments. We experiment on pretrained CLIP models [47] with ViT-B/32 and ViT-L/14 as the backbone, respectively. Empirically, we set the rank r of AB-LoRA to 8 by default. We also regularize its input by a dropout layer with $p = 0.25$. To train, we utilize Stochastic Gradient Descent (SGD) with a learning rate of 2×10^{-4} , a batch size of 4, and a cosine learning rate scheduler. The model is trained for 30 epochs on two NVIDIA RTX 4090 GPUs. For the MLLM, we primarily use InternVL-4B [8]. For more dataset and implementation details, please refer to [Appendix A](#).

4.2. Intra-dataset Open-set 3DOR

Compared Methods. For comprehensive comparisons, we select twelve representative methods, which can be categorized into four groups: 1) Vision-language-based methods, *i.e.*, Uni3D [68], ULIP-2 [62] and OpenShape [36], which leverage CLIP for 3D learning ¹. We also re-implement a superior zero-shot method MV-CLIP [50], which requires a ground-truth set for view selection. 2) Previous excellent 3DOR methods including TCL [20],

¹We directly evaluate released models from them in a zero-shot manner. Yet, we observe worse results after fine-tuning in the open-set setup.

SDML [25], and CMCL [28]). They have shown strong performance in close-set settings and are now extended to this challenging setup. 3) Auto-encoder-based methods (MMSAE [59], MCWSA [66]), which aim to learn compressed representations in an unsupervised manner. 4) Specially-designed open-set methods (PROSER [67], InfoNCE [43] and HGM²R [15]), which is specifically crafted to handle open-set scenarios.

Result Analysis. Table 1 compares open-set retrieval performance between our DAC and other representative methods. We conduct experiments under two settings: Zero-shot and Open-set. In the zero-shot setup, DAC achieved the best results across the OS-ESB-core, OS-NTU-core, and OS-ABO-core datasets, notably surpassing the previous state-of-the-art by over 6% on the OS-ESB-core dataset. Although MV-CLIP achieved better results on the OS-MN40-core dataset, it cannot be directly applied to open-set tasks. In MV-CLIP’s view selection process, category information must be provided, which is *incompatible* with open-set scenarios where category labels are inherently unknown in advance. In open-set setup, as shown, DAC is better on OS-ESB-core which only has 98 training objects, while much worse on the remaining three larger benchmarks. Among existing methods, HGM²R is the previous state-of-the-art. However, DAC with ViT-B/32 and ViT-L/14 surpass it significantly by over 6.9% and 6.0% in mAP on OS-ESB-

Backbone	InternVL	AB-LoRA	OS-ESB-core	OS-NTU-core	OS-MN40-core	OS-ABO-core
CLIP ViT-B/32	✗	✗	53.93 / 23.00 / 49.70	49.35 / 23.70 / 53.03	49.60 / 65.71 / 50.88	47.64 / 51.55 / 52.47
	✓	✗	56.16 / 23.58 / 48.39	55.03 / 25.70 / 48.32	55.39 / 68.08 / 45.96	60.77 / 56.28 / 41.44
	✗	✓	57.45 / 23.96 / 47.13	54.57 / 25.80 / 49.08	59.35 / 71.89 / 42.72	56.45 / 55.91 / 45.33
	✓	✓	58.70 / 24.27 / 45.67	59.21 / 27.06 / 44.58	62.40 / 72.63 / 39.82	66.10 / 59.01 / 36.12
CLIP ViT-L/14	✗	✗	54.68 / 23.39 / 48.67	57.29 / 26.24 / 45.47	55.01 / 70.55 / 45.72	57.35 / 56.13 / 44.42
	✓	✗	56.60 / 23.94 / 47.61	61.33 / 27.53 / 41.96	59.77 / 72.08 / 41.46	65.93 / 59.04 / 36.60
	✗	✓	56.72 / 23.96 / 48.18	62.77 / 28.03 / 40.34	65.84 / 76.10 / 36.65	62.92 / 58.48 / 38.69
	✓	✓	57.80 / 24.36 / 47.44	65.83 / 28.78 / 37.46	68.98 / 77.59 / 33.87	70.74 / 60.87 / 32.14

Table 2. **Effectiveness of the designed modules.** Values are presented in mAP/NDCG/ANMRR format.

core, respectively. We further observe DAC with both backbones demonstrates remarkable performance on OS-NTU-core, bringing over 14% and 20% higher mAP than HGM²R, respectively. OS-MN40-core and OS-ABO-core have larger training sets. With more 3D objects for training, HGM²R exhibits promising outcomes by connecting seen and unseen categories with hypergraphs. In contrast, DAC with ViT-B/32 shows inferior results on OS-MN40-core and comparable performance on OS-ABO-core. It is noteworthy that, Unlike HGM²R, we do not 1) utilize test data for training and 2) solely make use of multi-view images rather than multi-modality (voxels, point clouds, multi-view images). DAC with stronger ViT-L/14 further enhances the performance, surpassing HGM²R greatly.

4.3. Analyses

In this section, we study the core design choices of DAC. More ablation studies are induced in [Appendix B](#).

Impact of InternVL. We first thoroughly assess the impact of InternVL [8] (see [Appendix I](#) for more choices) on all four datasets with both CLIP ViT-B/32 and CLIP ViT-L/14. The multi-view features extracted by CLIP after AB-LoRA fine-tuning (DAC w.o InternVL) serve as our baselines. As shown in Table 2, incorporating InternVL predictions for text embeddings leads to consistent improvements across all datasets, regardless of backbones. As shown, on the challenging real-world OS-ABO-core dataset, we significantly improve mAP and ANMRR by +9.65% and +9.21%, respectively, when adopting CLIP ViT-B/32 as the backbone. Similar encouraging observations are shown on OS-MN40-core and OS-NTU-core, regardless of the backbones. We also note that on OS-ESB-core, incorporating InternVL results in less pronounced improvements. For instance, when adopting CLIP ViT-B/32, we increase mAP from 57.45 to 58.70. It can be attributed to the nature of OS-ESB-core, which contains high-genus objects [27], such as mechanical parts, posing great difficulties for InternVL in producing semantic descriptions. The above experiments show that incorporating InternVL can offer an effective solution to enhance DAC for open-set 3DOR.

Impact of Adaptation. AB-LoRA enables efficient adap-

tation of CLIP to the multi-view projection image distribution. Table 2 studies the impact of adaptation on all four datasets. As shown, incorporating it yields significant performance gains, particularly on OS-MN40-core, which shows +7.01% mAP improvement using ViT-B/32 and +9.21% using ViT-L/14. DAC incorporates a variant of LoRA for efficient adaptation but introduces negligible testing costs. These excellent results show robustness, generalization, and versatility of our DAC.

AB-LoRA vs LoRA. Open-set 3DOR requires learning robust representations from known categories in the training set to achieve generalization to unknown categories, making it highly susceptible to overfitting on known classes. To mitigate this, we introduce bias into LoRA during training. We conduct experiments on OS-MN40-core to evaluate the impact of introducing bias. As shown in Table 3, compared with LoRA without bias, AB-LoRA increases the performance by +2.55% mAP. This empirical evidence validates the surprising effectiveness of a simple bias in enhancing the model’s generalization to unknown categories.

Method	mAP↑	NDCG↑	ANMRR↓
Without LoRA	55.39	68.08	45.96
LoRA	59.85	70.25	41.75
AB-LoRA	62.40	72.63	39.82

Table 3. **Ablations of Additive Bias LoRA.**

Analysis on Fusion Scheme. We analyze two common parameterless fusion schemes for integrating textual and multi-view features: concatenation (*Concat.*) and element-wise addition (*Add.*). As shown in Table 4, the *Add.* method consistently outperforms the *Concat.* method across all metrics and backbones by a large margin. For the ViT-B/32 backbone, *Add.* improves mAP from 55.30 to 62.40. Likewise, for the ViT-L/14 backbone, *Add.* achieves the best results with a mAP of 68.98, NDCG of 77.59, and ANMRR of 33.87. It implies that element-wise addition can more effectively utilize the two modalities to generate more discriminative representations for retrieval.

Backbone	Fusion Method	mAP \uparrow	NDCG \uparrow	ANMRR \downarrow
ViT-B/32	<i>Concat.</i>	55.30	65.12	46.11
	<i>Add.</i>	62.40	72.63	39.82
ViT-L/14	<i>Concat.</i>	56.74	66.08	45.04
	<i>Add.</i>	68.98	77.59	33.87

Table 4. **Impact of fusion Schemes** on different backbones.

4.4. Cross-dataset Open-set 3DOR

Setup. To evaluate DAC on cross-dataset generalization capacity, We train all the models on OS-MN40-core and evaluate them on OS-ABO-core.

Results. The results are summarized in Table 5. It can be observed that DAC outperforms the other compared methods impressively. Specifically, with the ViT-L/14 backbone, our approach surpasses the previous state-of-the-art by +12.31% mAP. This substantial improvement highlights the effectiveness of our method in cross-dataset scenarios, showcasing its strong ability to process instances of unknown categories from an entirely distinct domain.

Method	OS-MN40-core \rightarrow OS-ABO-core		
	mAP \uparrow	NDCG \uparrow	ANMRR \downarrow
CMCL [28]	53.90	53.79	47.28
MMSAE [59]	52.83	52.80	48.02
MCWSA [66]	49.20	50.99	51.11
PROSER [67]	50.80	52.37	49.73
InfoNCE [43]	51.63	52.75	49.16
HGM ² R [15]	57.55	54.14	45.35
Ours (ViT-B/32)	63.45	57.34	38.48
Ours (ViT-L/14)	69.86	60.13	32.42

Table 5. **Comparisons (%) on cross-dataset retrieval.**

4.5. Single Image based Open-set 3DOR

Setup. Recall that OS-MN40-core contains synthetic 3D objects, whereas OS-ABO-core includes real-world 3D objects, each with an attached real-world image. Here we only utilize one projected image (*Prj.*) or one real image (*Rel.*) depicting each 3D object. We alternate the two datasets for training and retrieval, yielding two setups: *MN40 \rightarrow ABO* and *ABO \rightarrow MN40*. In each setup, we have two subsettings: *Prj. \rightarrow Rel.* (or *Rel. \rightarrow Prj.*) and *Prj. \rightarrow Prj.*

Results. The results are summarized in Table 6. Encouragingly, DAC still has the best generalization ability despite the great domain gap between real and synthetic data: We note that the results under the *Prj. \rightarrow Rel.* setting are better than the *Prj. \rightarrow Prj.* setting. This can be attributed to both CLIP and InternVL being pre-trained on real images, allowing it to generalize well to real-world 3DOR. It highlights DAC’s potential in the challenging real-world 3DOR.

Method	<i>MN40 \rightarrow ABO</i>		<i>ABO \rightarrow MN40</i>	
	<i>Prj. \rightarrow Rel.</i>	<i>Prj. \rightarrow Prj.</i>	<i>Rel. \rightarrow Prj.</i>	<i>Prj. \rightarrow Prj.</i>
CMCL [28]	34.92	45.58	40.41	41.40
MMSAE [59]	35.91	45.24	40.12	41.80
MCWSA [66]	32.63	44.35	38.78	41.14
PROSER [67]	33.65	44.32	39.62	41.14
InfoNCE [43]	33.48	44.22	39.63	41.12
HGM ² R [15]	43.59	50.68	48.93	49.77
Ours (ViT-B/32)	55.59	55.00	49.34	49.71
Ours (ViT-L/14)	63.96	60.12	52.92	56.24

Table 6. **Comparisons (mAP) on single image based open-set 3DOR.** *Prj.* denotes Projected Image, and *Rel.* denotes Real Image. *A \rightarrow B* denotes adopting A for training and B for retrieval.

4.6. Visualization and Discussions on Limitations

Figure 4 presents some retrieval examples. Our method can faithfully retrieve related 3D assets for the query. Nevertheless, DAC has the following possible limitations. 1) Incorporating MLLM at inference increases costs. *Yet, note that DAC already outperforms prior art HGM²R even without it during inference* (see Table 2), e.g., 65.84 v.s. 63.74 mAP on OS-MN40-core. We plan to explore more training strategies like providing region-level descriptions with MLLM to enhance the training and remove the reliance. 2) DAC only fuses global textual and visual embeddings, which fails in some cases when the query and sample have similar global shapes but from hard category pair (e.g., wardrobe and bookshelf) (see *Appendix H*). We plan to adopt more fine-grained discriminative features to address this issue.

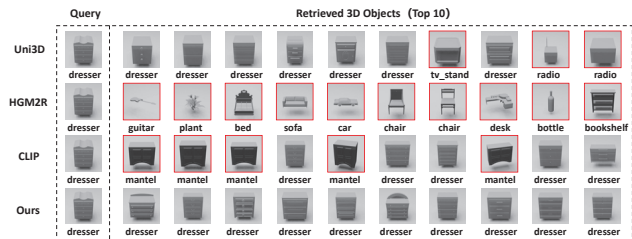


Figure 4. **Retrieval example comparisons** with other methods on OS-MN40-core. Incorrect matches are in red boxes.

5. Conclusion

In this paper, we have presented a simple yet effective framework named DAC for open-set 3D object retrieval. In contrast to previous methods, we make the first attempt to synergize generative and discriminative large VLLMs to derive discriminative and generalized embeddings for open-set 3D object retrieval. We also showed that adapting CLIP with LoRA based on multi-view images can greatly boost performance. Finally, we extended it to other challenging 3D representation learning tasks such as cross-dataset and zero-shot retrieval, further validating its strong generality.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.62302188, 62225603, 62276111 and 62076041); Fundamental Research Funds for the Central Universities (No.2662023XXQD001).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. pages 40–49, 2018. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736, 2022. 3
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3
- [5] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002. 2
- [6] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, pages 223–232, 2003. 2
- [7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1, 2, 3, 4, 6, 7
- [9] Silin Cheng, Xiwu Chen, Xinwei He, Zhe Liu, and Xiang Bai. Pra-net: Point relation-aware network for 3d point cloud analysis. *IEEE TIP*, 30:4436–4448, 2021. 2
- [10] Guoxian Dai, Jin Xie, Yi Fang, et al. Siamese cnn-bilstm architecture for 3d shape representation learning. In *IJCAI*, pages 670–676, 2018. 2
- [11] Carlos Esteves, Yinshuang Xu, Christine Allen-Blanchette, and Kostas Daniilidis. Equivariant multi-view networks. In *ICCV*, pages 1568–1577, 2019.
- [12] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *CVPR*, pages 264–272, 2018. 1, 2
- [13] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: Mesh neural network for 3d shape representation. In *AAAI*, pages 8279–8286, 2019. 2
- [14] Yifan Feng, Yue Gao, Xibin Zhao, Yandong Guo, Nihar Bagewadi, Nhat-Tan Bui, Hieu Dao, Shankar Gangisetty, Ripeng Guan, Xie Han, et al. Shrec’22 track: Open-set 3d object retrieval. *Computers & Graphics*, 107:231–240, 2022. 1, 2, 4
- [15] Yifan Feng, Shuyi Ji, Yu-Shen Liu, Shaoyi Du, Qionghai Dai, and Yue Gao. Hypergraph-based multi-modal representation for open-set 3d object retrieval. *IEEE TPAMI*, 2023. 1, 2, 4, 5, 6, 8
- [16] Ahmed Fradi, Borhen Louhichi, Mohamed Ali Mahjoub, and Benoit Eynard. 3d object retrieval based on similarity calculation in 3d computer aided design systems. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 160–165, 2017. 1
- [17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 2, 3
- [18] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE TPAMI*, 43(12):4338–4364, 2020. 2
- [19] Zhizhong Han, Honglei Lu, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. 3d2seqviews: Aggregating sequential views for 3d global feature learning by cnn with hierarchical attention aggregation. *IEEE TIP*, 28(8):3986–3999, 2019. 2
- [20] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *CVPR*, pages 1945–1954, 2018. 6
- [21] Xinwei He, Tengting Huang, Song Bai, and Xiang Bai. View n-gram network for 3d object retrieval. In *ICCV*, pages 7515–7524, 2019. 2
- [22] Xinwei He, Silin Cheng, Dingkan Liang, Song Bai, Xi Wang, and Yingying Zhu. Latformer: Locality-aware point-view fusion transformer for 3d shape recognition. *PR*, 151: 110413, 2024. 1
- [23] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *ICCV*, pages 2028–2038, 2023. 3
- [24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 4, 5
- [25] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019. 6
- [26] 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 *ICCV*, pages 22157–22167, 2023. 2, 3
- [27] Subramaniam Jayanti, Yagnanarayanan Kalyanaraman, Natraj Iyer, and Karthik Ramani. Developing an engineering shape benchmark for cad models. *Computer-Aided Design*, 38(9):939–953, 2006. 7

- [28] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *CVPR*, pages 3142–3151, 2021. 1, 6, 8
- [29] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *CVPR*, pages 5010–5019, 2018. 2
- [30] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, pages 156–164, 2003. 2
- [31] David Koller, Bernard Frischer, and Greg Humphreys. Research challenges for digital archives of 3d cultural heritage models. *Journal on Computing and Cultural Heritage (JOCCH)*, 2(3):1–17, 2010. 1
- [32] Zhaoqun Li, Cheng Xu, and Biao Leng. Angular triplet-center loss for multi-view 3d shape retrieval. In *AAAI*, pages 8682–8689, 2019. 2
- [33] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *ECCV*, pages 37–54, 2022. 2
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 3
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 3
- [36] 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. *NeurIPS*, 36, 2023. 3, 6
- [37] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. In *AAAI*, pages 8778–8785, 2019. 1
- [38] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *CVPR*, pages 8895–8904, 2019. 2
- [39] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023. 3
- [40] Eleni Mangina. 3d learning objects for augmented/virtual reality educational ecosystems. In *2017 23rd International Conference on virtual system & Multimedia (VSMM)*, pages 1–6, 2017. 1
- [41] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. pages 922–928, 2015. 2
- [42] Weizhi Nie, Qi Liang, An-An Liu, Zhendong Mao, and Yangyang Li. Mmjn: Multi-modal joint networks for 3d shape recognition. In *ACM MM*, pages 908–916, 2019. 1
- [43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6, 8
- [44] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3
- [45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1, 2
- [46] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 2
- [47] 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. pages 8748–8763, 2021. 2, 4, 6
- [48] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, pages 2765–2775, 2023. 2
- [49] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, pages 2437–2446, 2019. 1
- [50] Dan Song, Xinwei Fu, Weizhi Nie, Wenhui Li, and Anan Liu. Mv-clip: Multi-view clip for zero-shot 3d shape recognition. *arXiv preprint arXiv:2311.18402*, 2023. 3, 6
- [51] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, pages 945–953, 2015. 1, 2
- [52] Johan WH Tangelder and Remco C Veltkamp. A survey of content based 3d shape retrieval methods. *Proceedings Shape Modeling Applications, 2004.*, pages 145–156, 2004. 2
- [53] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM TOG*, 36(4):1–11, 2017. 2
- [54] Weiyun Wang, Min Shi, Qingyun Li, Wenhui Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 3
- [55] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 36, 2024. 3
- [56] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM TOG*, 38(5): 1–12, 2019. 2
- [57] Zhichuan Wang, Yang Zhou, Jinhai Xiang, Yulong Wang, and Xinwei He. Teda: Boosting vision-language models for zero-shot 3d object retrieval via testing-time distribution alignment. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 1442–1451, 2025. 2
- [58] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *CVPR*, pages 1850–1859, 2020. 1

- [59] Yiling Wu, Shuhui Wang, and Qingming Huang. Multi-modal semantic autoencoder for cross-modal retrieval. *Neurocomputing*, 331:165–175, 2019. 1, 6, 8
- [60] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 1, 2
- [61] 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 *CVPR*, pages 1179–1189, 2023. 3
- [62] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *CVPR*, pages 27091–27101, 2024. 3, 6
- [63] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3
- [64] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, pages 8552–8562, 2022. 3
- [65] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, pages 5565–5573, 2019. 2
- [66] Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE TMM*, 2022. 6, 8
- [67] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pages 4401–4410, 2021. 6, 8
- [68] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *ICLR*, 2024. 6
- [69] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3
- [70] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, pages 2639–2650, 2023. 3