# Negative Prompt Driven Complementary Parallel Representation for Open-World 3D Object Retrieval

**Yang Xu**, **Yifan Feng** and **Yue Gao***
BNRist, THUIBCS, KLISS, BLBCI, School of Software, Tsinghua University, China
{xuyang9610, evanfeng97}@gmail.com, gaoyue@tsinghua.edu.cn

## Abstract

The limited availability of supervised labels (positive information) poses a notable challenge for open-world retrieval. However, negative information is more easily obtained but remains underexploited in current methods. In this paper, we introduce the Negative Prompt Driven Complementary Parallel Representation (NPCP) framework, which navigates the complexities of open-world retrieval through the lens of *Negative Prompts*. Specifically, we employ the Parallel Exclusive Embedding (PEE) module to effectively utilize the prompt information, bilaterally capturing both explicit negative and implicit positive signals. To address the challenges of embedding unification and generalization, our method leverages high-order correlations among objects through the Complementary Structure Tuning (CST) module, by constructing a complementary hypergraph based on bi-directional and cross-category correlations. We have developed four multimodal datasets for open-world 3D object retrieval with negative prompts: NPMN, NPAB, NPNT, and NPES. Extensive experiments and ablation studies on these four benchmarks demonstrate the superiority of our method over current state-of-the-art approaches.

## 1 Introduction

With the gradual increase in 3D data, 3D object retrieval (3DOR) has emerged as a central area of interest within computer vision [Krause *et al.*, 2013; Gao *et al.*, 2012]. The essence of 3DOR lies in establishing the relationship between the query and target datasets through training. Although recent advancements have significantly propelled the development of 3DOR, most existing methods operate under a closed-set assumption. This suggests that all object categories encountered in the testing phase have been previously seen during training [Chen *et al.*, 2022a]. However, in practical open-world applications, training sets often fail to encompass all potential categories, owing to limitations of data volume and insufficient labeling [Vaze *et al.*, 2021]. In open-world scenarios, the abundance of data coupled with incomplete labels makes it challenging to acquire sufficient and precise positive information for 3DOR. Against this backdrop, negative information, being more readily accessible than positive labels, can serve as effective hints or guidance [Wang *et al.*, 2023; Chen *et al.*, 2022b] within retrieval tasks. However, there is still a lack of research on the utilization of negative information for open-set retrieval.

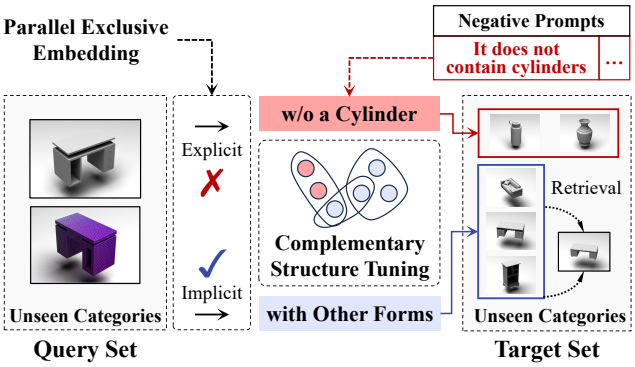



Figure 1: Illustration of negative prompts and proposed NPCP framework for open-world 3D object retrieval. Given unknown 3D objects of unseen categories, our method generates the prompt embedding in both explicit negative and implicit positive directions. Then tuned embeddings are generated via structure-aware tuning for open-world retrieval with unseen categories generalization.

Prompts, as essential components in representation learning [Liu *et al.*, 2023] and generative models [Wu *et al.*, 2022], have the capability to represent information, whether it is positive or negative. Recently, there has been an increasing application of prompt-driven methods in the realm of computer vision tasks [Jia *et al.*, 2022; Bahng *et al.*, 2022; Sung *et al.*, 2022]. However, most current methods for computer vision generate the prompt embeddings only from positive labels unidirectionally, which is a challenging path to implement in an open-world environment. Besides, typical open-world learning approaches are designed to generalize invariant knowledge from seen to unseen categories directly, employing specific approaches such as predicting the Out-of-Distribution (OOD) and In-Distribution (ID) scores [Yu *et*

*Corresponding author: Yue Gao

*al.*, 2020], or structured-aware knowledge learning [Feng *et al.*, 2024]. However, these methods still struggle to overcome the substantial distribution gaps [Zhou, 2022; Parmar *et al.*, 2023]. In this paper, we explore the use of negative prompts to bridge the gap of distributions between seen and unseen categories, aiming to enhance the generalization performance for open-world 3D object retrieval.

Distinct from direct open-set retrieval methods, the prompt-driven open-set retrieval method focuses more on fine-grained similarity in feature representations and demands greater ability to constrain and guide object embeddings using prompt information. This leads to several challenges for prompt-driven open-set retrieval, including: **First, the difficulty of achieving sufficient embeddings from limited prompt information**. Prompts for retrieval serve as cues for coarse categories in the open-world space, rather than simply classifying between OOD and ID [Wang *et al.*, 2023], thus offering broader possibilities. Hence, there is a strong motivation to extract more information from a limited set of prompts. **Second, the difficulty in achieving embedding unification across different spaces**, which involves aligning and fusing the features of prompts with 3D object representations from diverse modalities into a unified space. **Third, the difficulty in generalizing the prompt-based embeddings to unseen categories**, which requires achieving spatial generalization of feature representations under the guidance of prompt information.

Addressing the aforementioned challenges, we explore a method for open-world retrieval tasks through the lens of *Negative Prompts*. As shown in Figure 1, we introduce the Negative Prompt Driven Complementary Parallel Representation framework (NPCP) for open-world 3D object retrieval. On one hand, to tackle the challenge of sufficient prompt embedding, we utilize the Parallel Exclusive Embedding (PEE) to fully leverage the prompt information from both explicit negative and implicit positive directions. On the other hand, to overcome the difficulty in embedding unification and generalization, we construct a complementary hypergraph based on the bi-directional and cross-category correlations. This hypergraph captures the mutually exclusive and complementary structure between negative and positive information. Besides, we adopt the Complementary Structure Tuning (CSL) approach to exploit high-order correlations among objects for category generalization by the complementary hypergraph structure. Our contributions are summarized as follows:

- We explore a method to navigate the complexities of open-world 3D object retrieval through the lens of *Negative Prompts*, and we construct four datasets with multiple negative prompts for benchmarking.
- We propose the NPCP framework for prompt-based open-world 3D object retrieval tasks, including the Parallel Exclusive Embedding (PEE) and the Complementary Structure Tuning (CST) modules, which are designed to fully leverage the prompt information in both positive and negative directions and overcome the distribution deviation of categories.
- We propose a complementary hypergraph structure to capture high-order correlations among objects, guided by the bi-directional and cross-category correlations.
- Extensive experiments are conducted on the four benchmarks for evaluation, demonstrating the superiority of NPCP over current state-of-the-art 3D object retrieval methods.

## 2 Related Work

### 2.1 3D Object Retrieval

Traditional 3D object retrieval methods are based on the close-set assumption, which means the training set and testing set share the same category distribution space. In the retrieval phase, all categories of objects in the testing (query and target) set have been seen in the training phase. Current close-set 3D object retrieval methods can be divided into two categories according to the modality representation of objects: *i.e.*, single-modal retrieval and multi-modal retrieval.

Single-modal 3D object retrieval refers to detecting similar objects within a single modality of 3D data. [Su *et al.*, 2019] and [Wei *et al.*, 2020] construct a view-based graph model to generate the aggregated embeddings for retrieval from multi-view. [He *et al.*, 2018] propose a triplet-center loss to pull objects from the same category closer together and push objects from different categories farther away. HGNN [Feng *et al.*, 2019] proposes a hypergraph-based structure-aware method to capture the high-order correlations among objects and generate better embeddings. As for multi-modal retrieval, existing methods [Nie *et al.*, 2019; Liang *et al.*, 2021; You *et al.*, 2018; You *et al.*, 2019; Dong *et al.*, 2020; Bai *et al.*, 2021] propose the weighted fusion or feature fusion network to generate the aggregation embeddings from different modality-specific basic features. Besides, CMCL [Jing *et al.*, 2021] designs the cross-modal center loss to reduce the difference across different 3D modalities by the common center embeddings.

### 2.2 Open-World Learning

Open-world (open-set) learning aims to do the machine learning research in *open-world* scenarios where important factors are subject to change [Zhou, 2022]. Most existing methods focus on the recognition problem, [Vaze *et al.*, 2021] introduce a benchmark for open-set recognition, named Semantic Shift Benchmark (SSB). [Zhou *et al.*, 2021] proposes a "none-of-above" classifier to detect whether the sample belongs to the seen categories or not. [Chen *et al.*, 2021] introduces an adversarial-based method to minimize the overlap of known distributions and unknown distributions without loss of known classification accuracy. Furthermore, more open-set recognition methods are proposed for 3D object learning [Bendale and Boult, 2016; Joseph *et al.*, 2021; Alliegro *et al.*, 2022; Zhu *et al.*, 2023]. Different from the recognition task, retrieval tasks in the open-world are more practical. While only a few methods [Feng *et al.*, 2024; Liu *et al.*, 2024] address the open-set 3DOR task, they just focus on structure learning networks and overlook the specific conditions in the open-world.

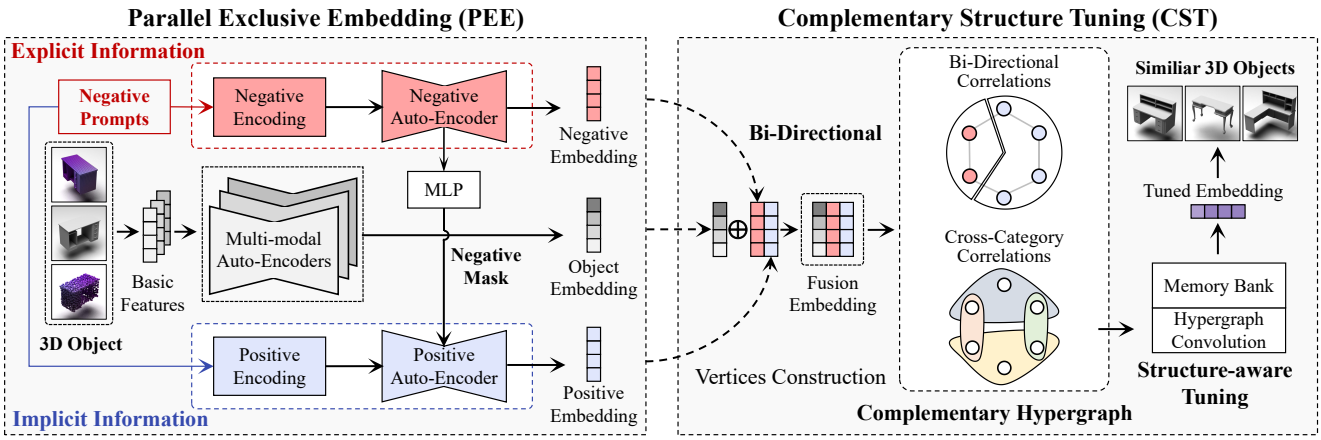



Figure 2: An overview of the proposed Negative Prompt Driven Complementary Parallel Representation (NPCP) framework for open-world 3D object retrieval. Our framework comprises two modules: Parallel Exclusive Embedding and Complementary Structure Tuning, which are used for bi-directional embedding for prompt information and structure-aware category generalization, respectively.

### 2.3 Prompt Learning

Prompt learning aims to justify pre-trained models to downstream tasks [Radford *et al.*, 2021; Li *et al.*, 2022] by *modus operandi* methods, and it has achieved remarkable success in Natural Language Processing (NLP) [Radford *et al.*, 2021; Shin *et al.*, 2020] tasks such as GPT series. Recently, prompt learning has been increasingly applied in computer vision studies [Jia *et al.*, 2022; Bahng *et al.*, 2022; Li *et al.*, 2022; Hegde *et al.*, 2023], which is treated as the task-specific embedding instead of model parameters. In this paper, we focus on extracting sufficient information carried by prompts, and the generalization for open-world scenarios.

## 3 Problem Setup

### 3.1 Open-World Retrieval

Given the query set $\mathcal{D}_q$ of 3D objects, the 3D object retrieval (3DOR) task is to find the matched or similar objects from the target set $\mathcal{D}_t$. The key for the 3DOR task is to find the relationship between the query and the target database through the training set $\mathcal{D}_{trn}$. Each 3D object is represented by $(o_i, y_i)$, the $o_i = \{m_r\}_{r=1}^{M}$ denotes a 3D object represented by $M$ modalities, such as multi-view, point cloud, voxel and others. The $y_i \in \mathcal{Y} = \{c_j\}_{j=1}^{Y}$ indicates the category label associated with the 3D object $o_i$.

In the open-set retrieval assumption, the training set $\mathcal{D}_{trn}$ and testing set $\mathcal{D}_{tes} = \{\mathcal{D}_q, \mathcal{D}_t\}$ are drawn from the different distributions, which means during the retrieval phase, every category of objects present in the query set has not been encountered and learned in the training phase. Specifically, for the training set $\mathcal{D}_{trn} = \{(o_i, y_i)\}_{i=1}^{L}$ and testing set $\mathcal{D}_{tes} = \{(o_i, \hat{y}_i)\}_{i=1}^{R} = \{\mathcal{D}_q, \mathcal{D}_t\}$, the category spaces of the training set and the retrieval set are not the same indicating $y_i \in \mathcal{Y} = \{c_j\}_{j=1}^{Y}$, $\hat{y}_i \in \hat{\mathcal{Y}} = \{\hat{c}_j\}_{j=1}^{\hat{Y}}$, and $\mathcal{Y} \neq \hat{\mathcal{Y}}$.

### 3.2 Negative Prompt

Each 3D object can be classified according to the basic geometric forms $b_i \in \mathcal{B} = \{c_j^*\}_{j=1}^{B}$ it contains, which is inherently independent of the dataset. Face the emergence of the open-world environment, it is much easier to determine the object category from a negative perspective, *i.e.*, geometric forms that the object must not contain or does not belong to. We term this negative cue as *negative prompt*:

$$n_i \in \mathcal{N} = \{\bar{c_j^*} \sim \mathcal{S}_r(\{1, .., B\} \setminus \{c_j^*\})\}, \tag{1}$$

where $n_i$ denotes a negative prompt of object $o_i$, which is in the form of one-hot. $\{1, \cdots, B\} \setminus c_j^*$ denotes the set of category labels whose correct labels are removed. $\mathcal{S}_r(\cdot)$ is proposed to select elements randomly from the set.

Consequently, the negative-prompt driven open-world 3DOR task aims to design a method using the training set $\mathcal{D}_{trn} = \{(o_i, n_i, y_i)\}_{i=1}^{L}$ and then use to search similar objects of the query in the testing set $\mathcal{D}_{ret} = \{(o_i, n_i, \hat{y}_i)\}_{i=1}^{R} = \{\mathcal{D}_q, \mathcal{D}_t\}$. The negative-prompt driven open-world 3DOR aims to minimize the expected risk:

$$f^* = \underset{f \in \mathcal{H}}{argmin} \mathbb{E}_{(D_i, D_j) \sim (\mathcal{D}_q, \mathcal{D}_t)} \Big[ 1_{\{\hat{y}_i \neq \hat{y}_j\}} e^{-\|f(o_i|n_i) - f(o_j|n_j)\|_2} + 1_{\{\hat{y}_i = \hat{y}_j\}} (1 - e^{-\|f(o_i|x_i) - f(o_j|x_j)\|_2}) \Big], \tag{2}$$

where $D_i = (o_i, n_i, \hat{y}_i)$ and $D_j = (o_j, n_j, \hat{y}_j)$ are object instance sampled from the query set $\mathcal{D}_q$ and target set $\mathcal{D}_t$, $1_{\{\cdot\}}$ is the indicator function, which returns 1 if the expression is true and 0 otherwise. $f := o_i|n_i \to z_i$ is the model that maps the 3D object $o_i$ into a fusion embedding $z_i \in \mathbb{R}^d$ under the condition of negative prompt $n_i$. $\mathcal{H}$ is the hypothesis space of map $f(\cdot|\cdot)$. $\|\cdot\|_2$ is the $\mathcal{L}_2$ norm function for distance metric, which could measure the Euclidean distance between two embeddings.

## 4 Methodology

### 4.1 Framework Architecture

As shown in Figure 2, the architecture of NPCP consists of two modules: *Parallel Exclusive Embedding (PEE)* and *Complementary Structure Tuning (CST)*. The framework takes basic features of the different modalities of 3D objects as input.

The PEE module is designed to generate the prompt embeddings bilaterally from both the explicit negative and implicit positive directions. Next, in the CST stage, the complementary hypergraph structure is constructed based on the exclusive and cross-category correlations. Guided by this structure, hypergraph convolution is adopted to leverage the high-order correlations based the bi-directional and cross-category correlations. Finally, the tuned embeddings are distilled and generated by the memory bank for retrieval or other downstream tasks.

## 4.2 Parallel Exclusive Embedding

To fully leverage the information conveyed by negative prompts, the PEE module is designed here. Specifically, the PEE is composed of a multi-modal auto-encoder and two parallel prompt auto-encoders (negative and positive). The multi-modal auto-encoder is employed for integrating features of different modalities to obtain the object embeddings of 3D objects. The two prompt auto-encoders are utilized to get the bi-directional (negative and positive) embeddings parallelly, from the explicit negative and implicit positive information carried by prompts.

**Bi-Directional Embedding**

As shown in Figure 2, the PEE module takes the extracted basic features $\{r_i^k\}_{k=1}^M$ ($r_i \in \mathbb{R}^{N\times d_r}$) of $N$ instances and $M$ modalities. Object embeddings $u_i$ are generated by the multi-modal auto-encoder $\mathcal{A}_m(\{r_i^k\}_{k=1}^M)$, where $u_i \in \mathbb{R}^{N\times d_u}$ and $\mathcal{A}_m(\cdot)$ denotes the fusion function of the multi-modal auto-encoder. Given a negative prompt $n_i$ for the $i$-th 3D object $o_i$, we represent it as the index of form that the object does not contain. The PEE module first encodes it as a negative encoding $e_i^n$ with the same size as $u_i$.

For each object $o_i$, the negative auto-encoder $\mathcal{A}^n$ first compresses negative encoding $e_i^n$ aligned with object embedding $u_i$ into the negative-prompt space $\mathbb{S}^n$, then does the reverse reconstruction while mapping it to a prompt mask for positive auto-encoder $\mathcal{A}^p$ through Multilayer Perceptron (MLP) layers. Specially,

$$\begin{cases} c_i^n = \lambda^n(u_i + e_i^n) \\ x_i^n = \Gamma(c_i^n) \\ \hat{u}_i^n = \omega^n(c_i^n) \end{cases}, \tag{3}$$

where $c_i^n \in \mathbb{R}^{d_u}$ and $\hat{u}_i^n \in \mathbb{R}^{d_u}$ denotes the negative prompt embedding and reconstructed feature, and $x_i^n \in \mathbb{R}^{d_c}$ denotes the negative mask for positive auto-encoder. The encoder and decoder are defined as $\lambda^n := \mathbb{S}_x^n \to \mathbb{S}^n$ and $\omega^n := \mathbb{S}^n \to \mathbb{S}_x^n$, which map the representation between negative-prompt space $\mathbb{S}_x^n$ and negative-mixed space $\mathbb{S}_m^n$. $\Gamma(\cdot)$ denotes the MLP layers for mask generation.

During the embedding for negative information, the positive auto-encoder $\mathcal{A}^p$ also leverages the opposite information using a negative mask. Specifically, $\mathcal{A}^p$ compresses the object embedding $u_i$ to positive-prompt embedding under the guidance of prompt mask $x_i^n$ and does the reverse reconstruction. For better representation,

$$\begin{cases} c_i^p = \lambda^p(u_i \odot x_i^n) \\ \hat{u}_i^p = \omega^p(c_i^p) \end{cases}, \tag{4}$$

where $c_i^p \in \mathbb{R}^{d_c}$ and $\hat{u}_i^p \in \mathbb{R}^{d_u}$ denote the negative prompt embedding and reconstructed feature, the encoder and decoder are defined as $\lambda^p := \mathbb{S}_x^p \to \mathbb{S}^p$ and $\omega^p := \mathbb{S}^p \to \mathbb{S}_x^p$, which map the representation between positive-prompt space $\mathbb{S}_x^p$ and positive-mixed space $\mathbb{S}_x^p$.

Finally, the PEE module generates the object embedding $u_i$, negative embedding $c_i^n$, and positive embedding $c_i^p$ for each object $o_i$.

**Loss Function**

To better construct the prompt embeddings for both negative and positive auto-encoders, we adopt the Dual Binary-Entropy loss $\mathcal{L}_{de}$ for $\mathcal{A}^n$ and $\mathcal{A}^p$.

**Dual Binary-Entropy Loss.** Given a negative prompt $n_i = \{\bar{c_j^*}\}_{j=1}^B$ of the 3D object $o_i$, the object may contain or belong to any other forms other than negative prompt as described in Section 3.2. The implicit positive label can be denoted as $\bar{n_i} = \{1, .., B\} \setminus n_i$, which means all forms except negative prompt. For each 3D object, the explicit negative and implicit positive labels are mutually exclusive and complementary, and together constitute a complete form label set $n_i \cup \bar{n_i} = \{1, .., B\}$. To guide the negative and positive embedding upon prompt information, the dual binary-entropy loss can be defined as follows:

$$\mathcal{L}_{de} = -\sum\nolimits_{k=1}^{B} \Big( n_{i,k} \log(p_{i,k}^n) + \bar{n}_{i,k} \log(p_{i,k}^p) \Big), \tag{5}$$

where $p_{i,k}^n = \frac{e^{\hat{u}_{i,k}^n}}{\sum_{m=1}^A e^{\hat{u}_{i,m}^n}}$ indicates the prediction score of that the 3D object $o_i$ does not contain the $k$-th form (negative label), and $p_{i,k}^p = \frac{e^{\hat{u}_{i,k}^p}}{\sum_{m=1}^A e^{\hat{u}_{i,m}^p}}$ denotes the positive prediction. Both predictions are classified from the reconstruction feature $\hat{u}_i^n$ and $\hat{u}_i^p$. $n_{i,k}$ is the $k$-th value of the one-hot encoded explicit negative label, and $\bar{n}_{i,k}$ is the $k$-th value of multi-hot encoded implicit positive labels, and $B$ is the number of basic forms.

**Joint Optimization.** In the PEE stage, the overall loss function is given:

$$\mathcal{L}_{pee} = \alpha \mathcal{L}_{de} + (1-\alpha)\mathcal{L}_{mm}, \tag{6}$$

where $\alpha$ are the hyper-parameter to trade-off between the loss of object and prompt embeddings, and $\mathcal{L}_{mm}$ denotes the multi-modal fusion loss defined by Homology Loss $\mathcal{L}_{hm}$ and Bi-Reconstruction Loss $\mathcal{L}_{br}$ following [Feng *et al.*, 2024], which can be calculated by $\mathcal{L}_{mm} = \mathcal{L}_{hm} + \mathcal{L}_{br}$.

## 4.3 Complementary Structure Tuning

Although the PEE module generates the negative and positive embeddings from the prompts, the open-world learning paradigm is frequently affected by the distribution gaps across seen and unseen categories. As shown in Figure 2, we proposed the CST module for generalization. Specifically, the complementary hypergraph is constructed to model the high-order correlations among objects in terms of category observability and prompt commonality. After structure construction, the combination of hypergraph convolution and memory bank is adopted for structure-aware smoothing and distilling to get the tuned embedding of each object.

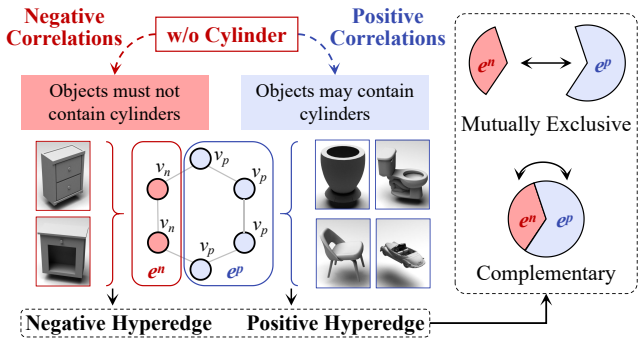



Figure 3: Illustration of the proposed bi-directional hyperedges. For each negative edge $e^n \in \mathcal{E}_n$, there is a corresponding mutually exclusive and complementary edge $e^p \in \mathcal{E}_p$.

### Complementary Hypergraph Generation

To get the most out of potential correlation information from category observability and prompt commonality, the complementary hypergraph structure is designed here. As shown in Figure 3, we use a complementary hypergraph from two perspectives: bi-directional correlations and cross-category correlations.

A hypergraph can be represented as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ are the vertex set and the hyperedge set, respectively. In our complementary hypergraph, each vertex is composed of the object embedding $u_i$ and bi-directional prompt embeddings $c_i^n$ and $c_i^p$ of each object. For better representation, we define the vertex of our complementary hypergraph as:

$$v_i = \delta(c_i^n + c_i^p) + (1-\delta)u_i, \tag{7}$$

where $c_i^n$ and $c_i^p$ denote the negative and positive embedding of the prompt, respectively. $u_i$ denotes the object embedding and $\delta$ is the hyper-parameter for trade-off.

**Bi-Directional Hyperedge.** As discussed in Section 4.2, each negative prompt encapsulates dual information: the object is excluded from this specific form and concurrently included in one of the alternative forms. As shown in Figure 3, bi-directional hyperedges are constructed based on this mutually exclusive information. Specially,

$$\begin{cases} \mathcal{E}_n = \{P_v(n) \mid n \in \mathcal{N}\} \\ \mathcal{E}_p = \{\mathcal{V} \setminus P_v(n) \mid n \in \mathcal{N}\} \end{cases}, \tag{8}$$

where $P_v(n)$ denotes the vertex subset that shares the same negative prompt $n$. For each negative edge $e_j^n \in \mathcal{E}_n$, there is a corresponding mutually exclusive and complementary edge $e_j^p$ in $\mathcal{E}_p$, these two hyperedges have no overlapping vertices and together constitute all the vertices in the hypergraph. $\mathcal{N}$ denote the space of the negative prompts and their implicit positive prompts. In this way, we got $B$ hyperedges for both negative and positive, where $B$ is the number of basic forms.

**Cross-Category Hyperedge.** Followed [Gao *et al.*, 2022], cross-category hyperegdes are constructed through the k-nearest neighbors (KNN) algorithm. For each vertex, we construct a hyperedge $\mathcal{E}_i$ to link it and its $K-1$ neighbor vertices:

$$\mathcal{E}_c = \{M_{\text{KNN}_k}(v) \mid v \in \mathcal{V}\} \tag{9}$$

| | | NPMN | NPAB | NPNT | NPES |
|---|---|---|---|---|---|
| Category | All | 40 | 21 | 67 | 41 |
| | Seen | 8 | 4 | 13 | 17 |
| | Unseen | 32 | 17 | 54 | 24 |
| Number | Training | 2821 | 1082 | 378 | 98 |
| | Testing | 7116 | 4154 | 1155 | 429 |
| | Query | 120 | 63 | 202 | 90 |
| | Target | 6996 | 4091 | 953 | 339 |
| Prompts per Object | | 3 | 3 | 3 | 3 |

Table 1: The statistics of the NPOR datasets.

where $M_{\text{KNN}_k}(v)$ denotes the k-nearest neighbors of vertex $v$.

In this way, we construct cross-category hyperedges with the same number of vertices. By combining these two kinds of hyperedges, we get the final complementary hypergraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_n \cup \mathcal{E}_p \cup \mathcal{E}_c\}$.

### Structure-Aware Tuning

To leverage the potential collaborative information, we utilize the modified hypergraph convolution from [Gao *et al.*, 2022] to smooth the complementary structure:

$$\tilde{\mathbf{V}} = \sigma\left(\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-\frac{1}{2}}\mathbf{V}\mathbf{\Theta}\right), \tag{10}$$

where $\mathbf{H}$ denotes the incidence matrix of the hypergraph $\mathbf{G}$. $\mathbf{D}_v$ and $\mathbf{D}_e$ are the diagonal degree matrices for vertice and hyperedges, respectively.

After getting the tuned embedding $\tilde{v}_i$ of the 3D object $o_i$, we construct a memory bank $\mathcal{M}$ that contains $L$ invariant memory anchors for knowledge distillation. We rebuild the embedding of each object $o_i$ by $z_i = \sum_{j=1}^{L} s'_{ij} h_j, z_i \in \mathbb{R}^{d_c}$, where $s'_{ij}$ denotes the normalization of activation score and calculated by $s_{ij} = \mathcal{D}_m(\tilde{v}_i, h_j)$, $h_j$ denotes the anchor and $D_m(\cdot,\cdot)$ denotes the distance metric function.

### Loss Function

To get better embeddings through hypergraph convolution and distillation, we adopt Memory Reconstruction Loss $\mathcal{L}_{mr}$ and the common Cross-entropy Loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{mr} = \left\|\tilde{v}_i - z_i\right\|_2, \tag{11}$$

$$\mathcal{L}_{ce} = -\sum_{k=1}^{Y}\left(y_{i,k}\log(p_{i,k}) + y_{i,k}\log(\tilde{p}_{i,k})\right), \tag{12}$$

where $\tilde{v}_i$ and $z_i$ denote the structure-aware embedding and memory reconstruction embedding, $\|\cdot\|_2$ is the $\mathcal{L}_2$ norm function. $\tilde{p}_{i,k} = \frac{e^{\tilde{v}_{i,k}}}{\sum_{m=1}^{Y} e^{\tilde{v}_{i,m}}}$ and $p_{i,k} = \frac{e^{z_{i,k}}}{\sum_{m=1}^{Y} e^{z_{i,m}}}$ is the predicted probability score of the 3D object $o_i$ in $k$-th category. $y_{i,k}$ is the $k$-th value of the one-hot encoded ground truth label of $o_i$, and $Y$ is the number of categories.

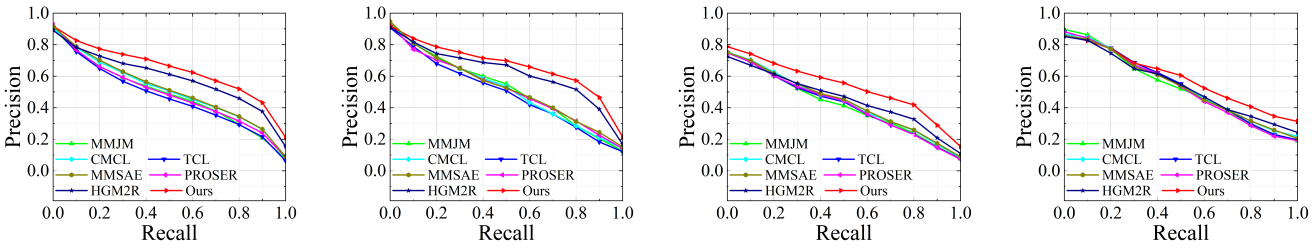
In the Complementary Structure Tuning stage, the overall loss function is given by combining Eq. 11 and Eq. 12:

$$\mathcal{L}_{cst} = \beta\mathcal{L}_{mr} + (1-\beta)\mathcal{L}_{ce}, \tag{13}$$

where $\beta$ is the hyper-parameter for trade-off.

| | NPMN | | | NPAB | | | NPNT | | | NPES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ |
| **MMJN** | 0.4841 | 0.5985 | 0.5206 | 0.5197 | 0.5229 | 0.4861 | 0.4110 | 0.1923 | 0.6093 | 0.5227 | 0.1902 | 0.5120 |
| **TCL** | 0.4674 | 0.5856 | 0.5367 | 0.4897 | 0.5010 | 0.5159 | 0.4130 | 0.1913 | 0.6050 | 0.5353 | 0.1911 | 0.5001 |
| **CMCL** | 0.5108 | 0.6098 | 0.4998 | 0.5115 | 0.5133 | 0.4961 | 0.4281 | 0.1957 | 0.5934 | 0.5349 | 0.1918 | 0.5046 |
| **MMSAE** | 0.5169 | 0.6165 | 0.4930 | 0.5223 | 0.5144 | 0.4900 | 0.4317 | 0.1954 | 0.5895 | 0.5304 | 0.1911 | 0.5229 |
| **PROSER** | 0.4886 | 0.6006 | 0.5199 | 0.5159 | 0.5122 | 0.4968 | 0.4144 | 0.1918 | 0.6036 | 0.5250 | 0.1884 | 0.5013 |
| **HGM$^2$R** | 0.5880 | 0.6504 | 0.4314 | 0.6291 | 0.5516 | 0.3933 | 0.4499 | 0.1984 | 0.5793 | 0.5330 | 0.1944 | 0.5161 |
| **Ours** | **0.6443** | **0.6931** | **0.3827** | **0.6697** | **0.5751** | **0.3500** | **0.5316** | **0.2209** | **0.4995** | **0.5805** | **0.2012** | **0.4701** |

Table 2: Experimental results on the NPMN, NPAB, NPNT, and NPES datasets.




(a) PR Curve on NPMN. (b) PR Curve on NPAB. (c) PR Curve on NPNT. (d) PR Curve on NPES.

Figure 4: The precision-recall curves of the proposed method and compared methods on four datasets, respectively.

# 5 Experiments

## 5.1 Experimental Settings

**NPOR Datasets.** We generate four negative-prompt driven open-world 3D object retrieval (NPOR) datasets, including NPMN, NPAB, NPNT, NPES based on the public datasets ModelNet40 [Wu *et al.*, 2015], ABO [Collins *et al.*, 2022], NTU [Chen *et al.*, 2003], and ESB [Jayanti *et al.*, 2006]. As shown in Table 1, we construct three negative prompts for each object in both the training set and retrieval set according to the basic geometric forms, and we remove some objects that are difficult to categorize under this rule. These datasets consist of seen and unseen categories, each object has three modalities including multi-view, voxel, and point cloud.

**Implemental Details.** For a fair comparison, we fix the random seed as 2022 for all experiments in this paper. The basic features of multi-view, point cloud, and voxel are extracted by MVCNN [Su *et al.*, 2015], PointNet [Qi *et al.*, 2017], and 3DShapeNet [Wu *et al.*, 2015], respectively. We set $\alpha = 0.5$, $\delta = 0.8$, $\beta = 0.9$ for the hyper-parameters of NPCP in Eq. 6, Eq. 7, and Eq. 13, respectively. The two modules are trained separately with 40 epochs on learning rate $lr = 0.1$ and 120 epochs on $lr = 0.001$.

## 5.2 Retrieval Performance

**Compared Methods.** As no methods are specifically designed for the prompt-driven open-world 3DOR, we refine the current state-of-the-art methods of close-set 3DOR (MMJN [Nie *et al.*, 2019], TCL [He *et al.*, 2018], CMCL [Jing *et al.*, 2021], MMSAE [Wu *et al.*, 2019]), and open-world 3D recognition or recognition (PROSER [Zhou *et al.*, 2021], HGM$^2$R [Feng *et al.*, 2024]), then we added a prompt tuning module for each methods following [Wang *et al.*, 2023; Li *et al.*, 2022].

**MMJN [Nie *et al.*, 2019]**: MMJN is a multi-modal joint network that employs weighted fusion to integrate features across multiple modalities for retrieval.

**TCL [He *et al.*, 2018]**: TCL is a method based on metric learning, combining triplet and center loss to get unified fusion embeddings from different modalities.

**CMCL [Jing *et al.*, 2021]**: CMCL designs an adversarial center loss to minimize the distances of features from objects belonging to the same class across all modalities.

**MMSAE [Wu *et al.*, 2019]**: MMSAE is a multi-modal retrieval method using auto-encoders. It trains encoders with a reconstruction loss function to align embeddings from various modalities into a unified latent space.

**PROSER [Zhou *et al.*, 2021]**: PROSER is an open-world recognition method that extends the closed-set classifier to determine if a sample belongs to seen categories or not.

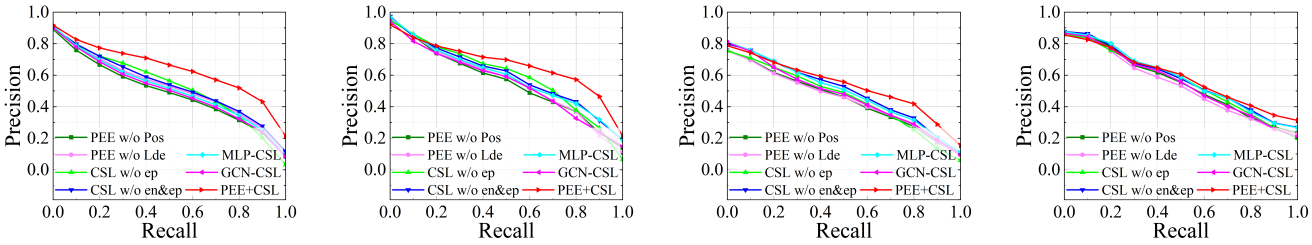
**HGM$^2$R [Feng *et al.*, 2024]**: HGM$^2$R is an open-world 3D multi-modal retrieval method, which retrieves the objects from unseen categories through structure-aware learning.

**Evaluation Metric.** For a fair comparison, we employ the commonly used retrieval metrics, including Mean Average Precision (mAP), Normalized Discounted Cumulative Gain (NDCG), Average Normalized Modified Retrieval Rank (ANMRR), and the Precision-Recall Curve (PR-Curve).

**Comparison Analysis.** As shown in Table 2, we evaluate the prompt-based open-world retrieval results from NPCP framework and other sate-of-the-art methods. Comparison results show that the proposed method outperforms the other methods on all four datasets. In particular, on the NPNT and

| | NPMN | | | NPAB | | | NPNT | | | NPES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ | mAP↑ | NDCG↑ | ANMRR↓ |
| **PEE w/o Pos** | 0.4908 | 0.5927 | 0.5211 | 0.5481 | 0.5300 | 0.4638 | 0.4414 | 0.1985 | 0.5786 | 0.5452 | 0.1938 | 0.4934 |
| **PEE w/o $\mathcal{L}_{de}$** | 0.5257 | 0.6143 | 0.4866 | 0.5596 | 0.5437 | 0.4490 | 0.4407 | 0.1980 | 0.5796 | 0.5255 | 0.1933 | 0.5216 |
| **CST w/o $\mathcal{E}_p$** | 0.5271 | 0.6280 | 0.4762 | 0.5909 | 0.5523 | 0.4128 | 0.4489 | 0.1913 | 0.5606 | 0.5531 | 0.1939 | 0.4893 |
| **CST w/o $\mathcal{E}_n$&$\mathcal{E}_p$** | 0.5389 | 0.6313 | 0.4770 | 0.5957 | 0.5596 | 0.4202 | 0.4951 | 0.2151 | 0.5352 | 0.5681 | 0.2006 | 0.4842 |
| **GCN-based CST** | 0.5159 | 0.6160 | 0.4958 | 0.5855 | 0.5556 | 0.4308 | 0.4889 | 0.2146 | 0.5407 | 0.5715 | 0.2026 | 0.4863 |
| **MLP-based CST** | 0.5034 | 0.6064 | 0.5078 | 0.5508 | 0.5430 | 0.4556 | 0.4614 | 0.2061 | 0.5577 | 0.5427 | 0.1944 | 0.5000 |
| **PEE+CST** | **0.6443** | **0.6931** | **0.3827** | **0.6697** | **0.5751** | **0.3500** | **0.5316** | **0.2209** | **0.4995** | **0.5805** | **0.2012** | **0.4701** |

Table 3: Albation Studies of the Parallel Exclusive Embedding and Complementary Structure Tuning modules on the NPES, NPNT, NPMN, and NPAB datasets. "w/o" denotes "without".




(a) PR Curve on NPMN. (b) PR Curve on NPAB. (c) PR Curve on NPNT. (d) PR Curve on NPES.

Figure 5: The precision-recall curves of the ablation studies on the four datasets, respectively.

NPMN datasets, our method achieves 0.5316/0.6443 mAP with about 18.2%/9.6% improvements compared with the second-best method. We also provide the Precision-Recall (PR) Curve to evaluate the performance of the proposed NPCP framework and other compared methods, as illustrated in Figure 4. The larger area below the curve indicates better performance. From the results, we can observe that our method outperforms all other compared methods. The better performance indicates that by the PEE and CST modules, the proposed method can take full advantage of negative prompts and has the generalized capability for unseen categories. Besides, as shown in Table 2 and Figure 4, our method presents a more significant improvement on the NPNT dataset. Specifically, the NPNT dataset has the maximum of categories and the lowest average number of objects per category. The proposed method can provide sufficient information for open-world retrieval with limited data to the greatest extent.

### 5.3 Ablation Study

We conduct ablation studies to verify the effectiveness of the proposed modules. For the PEE module, we remove the positive auto-encoder (PEE w/o Pos) and dual binary-entropy loss (PEE w/o $\mathcal{L}_{de}$) for comparison. As shown in Table 3 and Figure 5, the performance of PEE degrades after positive auto-encoder and dual binary-entropy loss are removed, respectively. Results show that direct embedding from only one direction of the negative prompts cannot provide sufficient auxiliary information for open-world retrieval. The proposed bi-directional embedding can take full advantage of prompt from both explicit negative and implicit positive directions.

As for the CST module, we compared the hypergraph structure without complementary structure (CST w/o $\mathcal{E}_p$ or CST w/o $\mathcal{E}_n$&$\mathcal{E}_p$), also we replace the hypergraph-based structure learning with MLP and GCN. Table 3 and Figure 5 show that the proposed CST outperforms all other structure learning methods, and the combination of PEE and CST yields the best performance. These results demonstrate the proposed framework can effectively utilize the high-order structure-aware correlations among seen and unseen categories under the open-world setting, which has the generalized capability for unseen categories with the help of prompts.

## 6 Conclusion

In this paper, we introduce the Negative Prompt Driven Complementary Parallel Representation (NPCP) framework, which navigates the complexities of open-world retrieval through the lens of *Negative Prompts*. To tackle the challenge of sufficient prompt embedding, we utilize the Parallel Exclusive Embedding (PEE) to fully leverage the prompt information from both explicit negative and implicit positive directions. Besides, we adopt the Complementary Structure Tuning (CSL) approach to exploit high-order correlations among objects for category generalization by the complementary hypergraph structure. This module constructs a complementary hypergraph based on bi-directional and cross-category correlations. We have developed four multimodal datasets for open-world 3D object retrieval with negative prompts, *i.e.*, NPMN, NPAB, NPNT, and NPES. Extensive experiments and ablation studies on these four benchmarks demonstrate the superiority of our method over current state-of-the-art approaches. We believe that this paper will offer innovative perspectives for the research of open-world retrieval.

## Acknowledgments

This work was supported by the National Natural Science Funds of China (No. 62088102, 62021002), and the Beijing Natural Science Foundation (No. 4222025).

## References

[Alliegro *et al.*, 2022] Antonio Alliegro, Francesco Cappio Borlino, and Tatiana Tommasi. Towards open set 3d learning: A benchmark on object point clouds. *arXiv preprint arXiv:2207.11554*, 2022.

[Bahng *et al.*, 2022] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[Bai *et al.*, 2021] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

[Bendale and Boult, 2016] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2016.

[Chen *et al.*, 2003] 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. Wiley Online Library, 2003.

[Chen *et al.*, 2021] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021.

[Chen *et al.*, 2022a] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Chen *et al.*, 2022b] Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. *Annual Conference on Neural Information Processing Systems*, 35:23908–23922, 2022.

[Collins *et al.*, 2022] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022.

[Dong *et al.*, 2020] Yihe Dong, Will Sawin, and Yoshua Bengio. Hnhn: Hypergraph networks with hyperedge neurons. *arXiv preprint arXiv:2006.12278*, 2020.

[Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI Conference on Artificial Intelligence*, pages 3558–3565, 2019.

[Feng *et al.*, 2024] 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 Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2206–2223, 2024.

[Gao *et al.*, 2012] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing*, 21(9):4290–4303, 2012.

[Gao *et al.*, 2022] Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2022.

[He *et al.*, 2018] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018.

[Hegde *et al.*, 2023] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2028–2038, 2023.

[Jayanti *et al.*, 2006] 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.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[Jing *et al.*, 2021] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021.

[Joseph *et al.*, 2021] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision Workshop*, pages 554–561, 2013.

[Li *et al.*, 2022] Haoran Li, Chun-Mei Feng, Tao Zhou, Yong Xu, and Xiaojun Chang. Prompt-driven efficient open-set semi-supervised learning. *arXiv preprint arXiv:2209.14205*, 2022.

[Liang *et al.*, 2021] Qi Liang, Mengmeng Xiao, and Dan Song. 3d shape recognition based on multi-modal information fusion. *Multimedia Tools and Applications*, 80:16173–16184, 2021.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Liu *et al.*, 2024] 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. *Annual Conference on Neural Information Processing Systems*, 36, 2024.

[Nie *et al.*, 2019] Weizhi Nie, Qi Liang, An-An Liu, Zhendong Mao, and Yangyang Li. Mmjn: Multi-modal joint networks for 3d shape recognition. In *ACM International Conference on Multimedia*, pages 908–916, 2019.

[Parmar *et al.*, 2023] Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023.

[Qi *et al.*, 2017] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Shin *et al.*, 2020] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[Su *et al.*, 2015] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 945–953, 2015.

[Su *et al.*, 2019] Yuting Su, Yuqian Li, Weizhi Nie, Dan Song, and An-An Liu. Joint heterogeneous feature learning and distribution alignment for 2d image-based 3d object retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3765–3776, 2019.

[Sung *et al.*, 2022] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Annual Conference on Neural Information Processing Systems*, 35:12991–13005, 2022.

[Vaze *et al.*, 2021] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*, 2021.

[Wang *et al.*, 2023] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-grained retrieval prompt tuning. In *AAAI Conference on Artificial Intelligence*, pages 2644–2652, 2023.

[Wei *et al.*, 2020] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1850–1859, 2020.

[Wu *et al.*, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[Wu *et al.*, 2019] Yiling Wu, Shuhui Wang, and Qingming Huang. Multi-modal semantic autoencoder for cross-modal retrieval. *Neurocomputing*, 331:165–175, 2019.

[Wu *et al.*, 2022] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando D De la Torre. Generative visual prompt: Unifying distributional control of pre-trained generative models. *Annual Conference on Neural Information Processing Systems*, 35:22422–22437, 2022.

[You *et al.*, 2018] Haoxuan You, Yifan Feng, Rongrong Ji, and Yue Gao. Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition. In *ACM International Conference on Multimedia*, pages 1310–1318, 2018.

[You *et al.*, 2019] Haoxuan You, Yifan Feng, Xibin Zhao, Changqing Zou, Rongrong Ji, and Yue Gao. Pvrnet: Point-view relation neural network for 3d shape recognition. In *AAAI Conference on Artificial Intelligence*, pages 9119–9126, 2019.

[Yu *et al.*, 2020] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *European Conference on Computer Vision*, pages 438–454. Springer, 2020.

[Zhou *et al.*, 2021] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2021.

[Zhou, 2022] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.

[Zhu *et al.*, 2023] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023.