Exploring Transferable Homogenous Groups for Compositional Zero-Shot Learning

Zhijie Rao , **Jincai Guo*** , **Miaoge Li** , **Yang Chen** and **Mengzhu Wang**
Department of COMP/LSGI, The Hong Kong Polytechnic University, Hong Kong SAR
zhijie96.rao@connect.polyu.hk, jc-jingcai.guo@polyu.edu.hk

Abstract

Conditional dependency presents one of the trickiest problems in Compositional Zero-Shot Learning, leading to significant property variations of the same state (object) across different objects (states). To address this problem, existing approaches often adopt either *all-to-one* or *one-to-one* representation paradigms. However, these extremes create an imbalance in the seesaw between transferability and discriminability, favoring one at the expense of the other. Comparatively, humans are adept at analogizing and reasoning in a hierarchical clustering manner, intuitively grouping categories with similar properties to form cohesive concepts. Motivated by this, we propose Homogeneous Group Representation Learning (HGRL), a new perspective that formulates state (object) representation learning as multiple homogeneous subgroup representation learning. HGRL seeks to achieve a balance between semantic transferability and discriminability by adaptively discovering and aggregating categories with shared properties, learning distributed group centers that retain group-specific discriminative features. Our method integrates three core components designed to simultaneously enhance both the visual and prompt representation capabilities of the model. Extensive experiments on three benchmark datasets validate the effectiveness of our method. Code is available at: https://github.com/zjrao/HGRL.

1 Introduction

Deep neural networks have achieved impressive results in object recognition tasks, yet their capabilities remain underexplored in recognizing abstract concepts and performing compositional reasoning. Humans, on the other hand, possess a remarkable ability to decompose and reorganize underlying visual knowledge to recognize new concepts. For instance, after recognizing a *red car* and a *green truck*, one can easily identify a *green car* and a *red truck*, even without having seen these specific visual samples before. To exploit the compositional recognition potential of models, Compositional Zero-Shot Learning (CZSL) [Chen and Grauman, 2014][Misra *et al.*, 2017] has been proposed, aiming to generalize to unseen compositions by learning from seen state-object sample pairs.

Conditional dependency [Wei *et al.*, 2019][Ge *et al.*, 2022] is one of the most intractable problems in CZSL, which reflects the intrinsic connection between states and objects, often referred to as domain bias. The dependency results in the same state (or object) exhibiting vastly different visual features when paired with different objects (or states), making it difficult to generalize semantic information learned from seen compositions to unseen ones. To tackle this problem, existing approaches typically follow two fundamental paradigms, the *all-to-one* and *one-to-one* paradigms. The former focuses on learning a versatile, domain-invariant prototype, whereas the latter explicitly models the dependencies between objects and states to learn instance-specific representations.

Despite the promising results, it remains a challenge to strike a balance between transferability and discriminability [Chen *et al.*, 2019]. The *all-to-one* paradigm overly emphasizes transferability, aiming for a pure and uniform central representation. This focus forces the model to sacrifice instance-specific semantic information. In contrast, the *one-to-one* paradigm prioritizes discriminability, leading the model to develop a bias toward the seen domain. Comparatively, humans are adept at analogizing and reasoning via hierarchical clustering to form group-specific conceptual cognition (Fig. 1 a). Such concepts maximize transferability among group members while preserving distinct and recognizable features. Thus, when recognizing an *old dog*, we would easily associate it with a *bear* or a *cat* to reduce the complexity of inference. Fortunately, despite the lack of guidance by hierarchical labels in CZSL, we find that homogeneous semantic structures are naturally maintained in the deep feature space (Fig. 1 b-c).

Motivated by this, we propose Homogeneous Group Representation Learning (HGRL). Analogous to the idea of hierarchical clustering, we cluster categories with similar properties together and call it a homogeneous group. The objective of HGRL is to identify unseen compositions by leveraging the high transferability among homogeneous group members. Our method integrates three critical components with CLIP as the backbone. Group-Aware Visual Representation

*Corresponding author: *Jingcai Guo*.

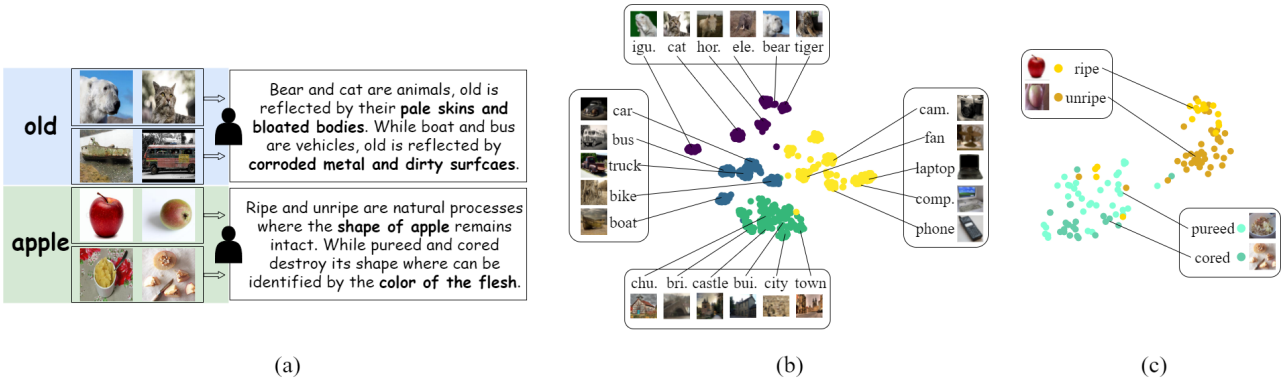

Figure 1: (a) *Motivation*. Humans utilize higher-order knowledge structure to perform hierarchical clustering, enabling effective analogies and reasoning. (b) *Visualization of the state—old*. The semantic structure of homogeneous groups is well maintained in deep feature space, e.g., animals are naturally clustered together. (c) *Visualizaiton of the object—apple*. Similarly, apples in various states cluster in groups due to visual variation.

(GAVR) aims to perceive and capture homogeneous groups and to learn group-specific representations to preserve semantic richness and completeness. To mimic the human knowledge system, we introduce the text co-occurrence probability graph to guide the hierarchical clustering process. Decoupled Group Prompt (DGP) is a simple yet effective group-based prompt representation. Traditional prompts provide a centralized prototype for each category. However, finding a common prototype representation is extremely challenging. To this end, DGP equips each group with an exclusive context prompt to learn distributed prototype representations. Group-Aware Pair Enhancement (GAPE), which combines state and object features into pairs to facilitate joint recognition. To further improve the robustness and generalization of the model, we utilize the compatibility of homogeneous group members to enhance the state (object) features to obtain diverse pairs. We conduct experiments on three mainstream datasets, and the results demonstrate that our method achieves state-of-the-art performance.

Our contributions are summarized below:

- We analyze that the limitation of existing paradigms dealing with conditional dependency lies in the zero-sum game between transferability and discriminability. Inspired by the hierarchical clustering mechanism, we provide a new view to tackle this problem.
- We present a new framework, named HGRL, which integrates three essential components to enhance both visual and prompt representations. HGRL explores latent homogeneous groups and maintains group-specific semantic information with multi-expert networks and distributed prompts.
- We conduct experiments on three major benchmark datasets, and the results show that the proposed method achieves state-of-the-art performance.

2 Related Work

2.1 Compositional Zero-Shot Learning

The goal of Compositional Zero-Shot Learning (CZSL) is to recognize unseen state-object pairs by learning and reorganizing seen compositions. The primary challenge is the distributional bias between seen and unseen compositions, which is caused by the dependency relationship between objects and states. Some studies adopt the *all-to-one* paradigm, i.e., learning a versatile semantic representation for a single object or state. The dominant technique is decoupling representation, which is based on the idea that objects and states can exist and be represented independently, so that the two need to be divorced. To this end, prototype learning [Ruis *et al.*, 2021][Lu *et al.*, 2023] and attention-based decoupling techniques [Zhang *et al.*, 2022][Li *et al.*, 2023][Hao *et al.*, 2023] have been introduced. In addition, some studies are inspired by causal representations to uncouple the potential connection between objects and states through decorrelation techniques [Atzmon *et al.*, 2020]. Nan et al. [Nan *et al.*, 2019] introduce contrastive learning to compress the intra-class semantic space to purify the representations. This paradigm is prone to sacrifice rich semantic features or learn pseudo-related features [Rao *et al.*, 2024b].

Another prominent paradigm is *one-to-one* representation. The objective is explicitly modeling the dependency relations between states and objects. For example, CoT [Kim *et al.*, 2023] uses object visual features as induced information to generate state visual features, while CANet [Wang *et al.*, 2023] and Troika [Huang *et al.*, 2024] leverage visual features to correct text embeddings. Other notable approaches include constructing conditional reconstruction networks [Nagarajan and Grauman, 2018][Li *et al.*, 2020] and conditional classifiers [Liu *et al.*, 2023][Huo *et al.*, 2024]. This paradigm tends to bias the model in favor of the seen domain.

Contribution. The *all-to-one* and *one-to-one* paradigms explore solutions to the issue of conditional dependency in almost opposing forms, yet both reach an extreme of neglecting discriminability or transferability. We provide a new perspective on it—find a balance between the two thus maximizing the generalization ability of the model.

2.2 Hierarchical Clustering

Hierarchical clustering is a classic clustering idea that considers both merging and splitting, drawing inspiration from

tree-like human knowledge system [Silla and Freitas, 2011]. It plays a crucial role in various structured data tasks, such as text classification [Chalkidis *et al.*, 2020], speech classification [Dekel *et al.*, 2004], and protein function prediction [Otero *et al.*, 2009]. Recently, hierarchical clustering has shown promising potential in vision tasks. For example, B-CNN [Zhu and Bain, 2017] combines the inter-layer outputs of convolutional networks with tree labels, significantly improving classification efficiency. Detclipv3 [Yao *et al.*, 2024] introduces hierarchical labels for open-vocabulary object detection, enabling the learning of multi-granularity semantic information. Additionally, some studies [Novack *et al.*, 2023][Rao *et al.*, 2024a] explore the potential of hierarchical clustering in zero-shot inference tasks.

Contribution. To the best of our knowledge, we are the first to apply the idea of hierarchical clustering to CZSL to solve the problem of imbalance between discriminability and transferability. Meanwhile, we introduce the text co-occurrence probability graph to cope with the problem that no hierarchical labels are available in CZSL.

3 Methodology

3.1 Preliminary

Suppose we have n_s states $\mathcal{S} = \{s_1, s_2, ..., s_{n_s}\}$ and n_o objects $\mathcal{O} = \{o_1, o_2, ..., o_{n_o}\}$, then there are $n_s \times n_o$ compositions defined as $\mathcal{C} = \{c_1 = (s_1, o_1), c_2 = (s_1, o_2), ..., C_{n_s \times n_o} = (s_{n_s}, o_{n_o})\}$, i.e., $\mathcal{C} = \mathcal{S} \times \mathcal{O}$. During the training phase, only partial compositions are available which are denoted as seen compositions $\mathcal{C}^{se}$. The goal of CZSL is to train a generalizable model thereby identifying unseen compositions $\mathcal{C}^{use}$. Note that $\mathcal{C}^{se} \bigcap \mathcal{C}^{use} = \emptyset$. Generally, $|\mathcal{C}^{se}| + |\mathcal{C}^{use}| < |\mathcal{C}|$ since some compositions in $\mathcal{C}$ do not exist in reality. So there are two settings in CZSL, closed-world CZSL where the label space is $\mathcal{C}^{se} \bigcup \mathcal{C}^{use}$ and open-world CZSL where the label space is $\mathcal{C}$. Formally, the training set is defined as $\mathcal{T} = \{(x_i, c_i) | x_i \in \mathcal{X}, c_i \in \mathcal{C}^{se}\}$, where $\mathcal{X}$ denotes the image space. An overview of our approach is presented in Fig. 2.

3.2 Group-Aware Visual Representation

The representations learned by paradigm *all-to-one* lose substantial key semantic information, while by paradigm *one-to-one* tend to favor the seen domain. To solve this problem, we present Group-Aware Visual Representation (GAVR) to extract the state and object features, whose motivation is learning group-specific representations. The advantages is twofold: 1) Maximizing semantic transferability among group memebers; 2) Preserving semantic integrity. Since the two branches are symmetric, for ease of illustration, we next introduce the state branch only.

The immediate problem is how to extract group-specific features. Inspired by mixture-of-expert network [Jacobs *et al.*, 1991], we introduce multiple expert sub-networks to learn the features of different groups. Given an instance $(x, c) \in \mathcal{T}$, we obtain its semantic features via the image encoder Φ_I. The feature $\Phi_I(x)$ at this point are a mixture of state and object semantic information. To get its state features, we use a sub-network Φ_S to map it to the state semantic space, which is denoted as $\Phi_S(\Phi_I(x))$. Previous methods would have used the feature to train directly, leading to a loss of group-specific information. We access a multi-expert network to extract group-specific semantic information. It consists of a route network Φ_{RS} and k_s expert sub-networks $\Phi_E^S = \{\Phi_{E_1}^S, ..., \Phi_{E_{k_s}}^S\}$. First, the feature $\Phi_S(\Phi_I(x))$ is routed through the route network to determine the belonging of the group, the confidence for expert i is denoted as:

$$p_r^{S_i} = \frac{\exp(\Phi_{RS}(\Phi_S(\Phi_I(x)))_i)}{\sum_1^{k_s} \exp(\Phi_{RS}(\Phi_S(\Phi_I(x))))}. \tag{1}$$

We then assign features to the corresponding expert networks based on the route scores. To balance sparsity and smoothness, we use a top filtering mechanism, i.e., selecting K experts with the largest confidence to participate in the next process. The obtained feature can be expressed as a weighted sum of K experts:

$$\mathbf{x}_g = \sum^{K} p_r^{S_i} \cdot \Phi_{E_i}^S(\Phi_S(\Phi_I(x))). \tag{2}$$

Finally, we fuse the group-specific feature with the original feature:

$$\mathbf{x}_s = \Phi_S(\Phi_I(x)) + \beta \cdot \mathbf{x}_g, \tag{3}$$

where β is a learnable parameter.

How to Capture Homogenous Groups. Another problem is the lack of supervision of hierarchical labels makes the capture of homogeneous groups uncontrollable, despite the natural heterogeneity exploration capability of route network. To this end, we utilize a word embedding model, GloVe [Pennington *et al.*, 2014], to evaluate compatibility between categories. The core idea of GloVe is to learn co-occurrence relationships between words from a large-scale corpus. We argue that the probability of co-occurrence between words largely matches the probability that they belong to the same superclass in reality, and thus can be used as a substitute for hierarchical labels. For state branch, we need to evaluate the heterogeneity of different objects with the same state. For any two object words, we compute cosine similarity to assess the co-occurrence probability of them, which is denoted as $M_{ij}^o = \cos(\mathbf{v}_i^o, \mathbf{v}_j^o)$, where $\mathbf{v}$ denotes the word embedding extracted by GloVe.

Now the most straightforward approach is to assign hard group labels by clustering through similarity threshold filtering, which involves sensitive threshold effect and is prone to introduce noise. Inspired by graph representation learning [Scarselli *et al.*, 2008], we encourage the model to adaptively discover homogeneous groups via local clustering. Specifically, we aggregate the feature of one sample with its nearest compatible samples. For example, for an *old dog*, we fuse it with *old cat*, *old tiger*, *etc.* to form a tiny homogeneous group. The model iteratively learns and summarizes the relationships of the tiny homogeneous groups to extend to the whole homogeneous group. Given a batch of instances $\{x_i, c_i = (s_i, o_i)\}^B \in \mathcal{T}$, where B denotes the batch size. We can get their relation map $A \in \mathbb{R}^{B \times B}$, and we have:

$$A_{ij} = M_{ij}^o \cdot \mathbb{I}[M_{ij}^o \geq \zeta] \cdot \mathbb{I}[s_i = s_j], \tag{4}$$

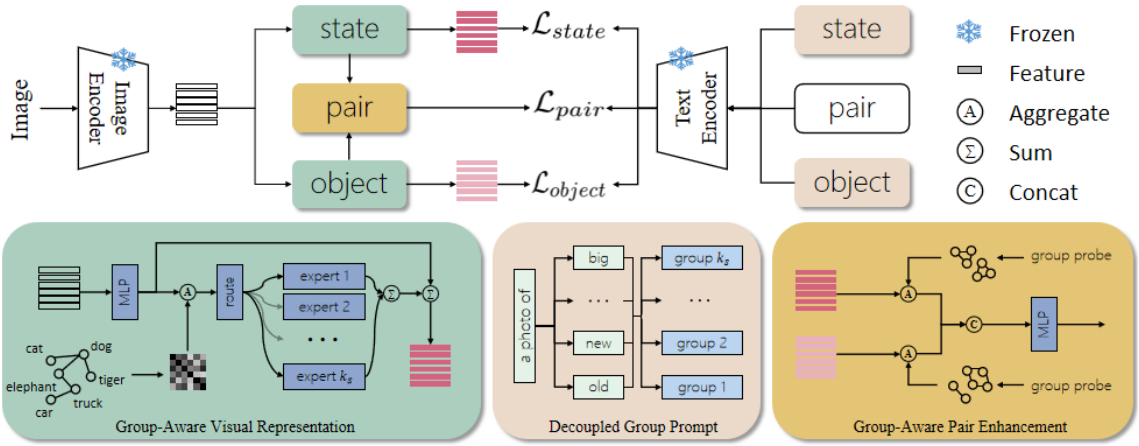

Figure 2: Overview of the proposed method, which comprises three main components to enhance both visual and prompt representations. Note that the two branches of state and object are symmetric, so just one branch is presented. *State (object) visual enhancement:* GAVR deeply explores latent homogeneous groups and utilizes multi-expert networks to extract group-specific representations to maintain semantic integrity and transferability. *Pair visual enhancement:* GAPE integrates state and object features for joint recognition and introduces group-aware feature augmentation to improve pair diversity. *Text prompt enhancement:* Unlike traditional category prompts, DGP additionally learns a customized contextual prompt for each group.

where $\mathbb{I}[\cdot]$ denotes indicator function and ζ is a constant which is fixed to 0.5. After normalization for A, we can get the aggregated features which denoted as

$$\mathbf{x}^s_{agg} = A \otimes \Phi_S(\Phi_I(x)), A = \frac{\exp(A+\Lambda)}{\sum_{j=1}^{B}\exp(A_{ij}+\Lambda)} \quad (5)$$

where $\otimes$ denotes matrix multiplication and Λ denotes the unit diagonal matrix. Next the aggregated features are fed into route and expert networks for training.

3.3 Decoupled Group Prompt

Textual prompts are important components in visual language models (VLMs) [Radford *et al.*, 2021]. In the classification tasks, they act as primitive representations of categories thus providing optimized targets for visual features. The common practice of designing prompts in the fixed form of *a photo of [class]*, which has been proven to perform inferior to learnable prompts [Zhou *et al.*, 2022]. The central idea of learnable prompts is to learn the task-specific context. To this end, previous CZSL approaches equip each branch with a separate contextual prompt, which are represented as:

$$\mathbf{P}^S_i = [\mathbf{p}^s_1, ..., \mathbf{p}^s_m, \mathbf{w}^s_i], \quad (6)$$

$$\mathbf{P}^O_i = [\mathbf{p}^o_1, ..., \mathbf{p}^o_m, \mathbf{w}^o_i], \quad (7)$$

$$\mathbf{P}^C_i = [\mathbf{p}^c_1, ..., \mathbf{p}^c_m, \mathbf{w}^s_i, \mathbf{w}^o_i], \quad (8)$$

where $\mathbf{P}^S_i, \mathbf{P}^O_i, \mathbf{P}^C_i$ denote the state, object and pair prompts, respectively. And $\{\mathbf{p}^s_1, ..., \mathbf{p}^s_m\}$, $\{\mathbf{p}^o_1, ..., \mathbf{p}^o_m\}$, $\{\mathbf{p}^c_1, ..., \mathbf{p}^c_m\}$ denote their contextual prompts and $\mathbf{w}^s_i$, $\mathbf{w}^o_i$ denote the state and object words. For state and object branches, the learned prompts capture the global context yet ignore the discrepancies between heterogeneous groups. To this end, we add the corresponding prompt for each group, which are represented as:

$$\mathbf{P}^{S_j}_i = [\mathbf{p}^s_1, ..., \mathbf{p}^s_m, \mathbf{g}^s_j, \mathbf{w}^s_i], j = 1, 2, ..., k_s, \quad (9)$$

$$\mathbf{P}^{O_j}_i = [\mathbf{p}^o_1, ..., \mathbf{p}^o_m, \mathbf{g}^o_j, \mathbf{w}^o_i], j = 1, 2, ..., k_o, \quad (10)$$

where k_s, k_o represent the groups of state and object, $\mathbf{g}^s$, $\mathbf{g}^o$ represent the group prompts. Then these prompts are fed into the text encoder Φ_T to obtain representations, which are denoted as:

$$\mathbf{t}^{S_j}_i = \Phi_T(\mathbf{P}^{S_j}_i), \mathbf{t}^{O_j}_i = \Phi_T(\mathbf{P}^{O_j}_i). \quad (11)$$

Training Objectives. We adopt InfoNCE loss function to train the network. At the same time, we use the confidence scores in Eq. (1) as soft group labels for samples. For the state branch, the loss formula is:

$$\mathcal{L}_{state} = -\frac{1}{B}\sum^{B}\log\frac{\sum_{j=1}^{k_s} p_r^{S_j}\cdot\exp(\mathbf{x}_s\cdot\mathbf{t}^{S_j}_i/\tau)}{\sum_{i=1}^{n_s}\sum_{j=1}^{k_s}\exp(\mathbf{x}_s\cdot\mathbf{t}^{S_j}_i/\tau)}, \quad (12)$$

where τ denotes the temperature coefficient. Since the object and state branches are symmetric, similarly, we can get the loss formula for the object branch as:

$$\mathcal{L}_{object} = -\frac{1}{B}\sum^{B}\log\frac{\sum_{j=1}^{k_o} p_r^{O_j}\cdot\exp(\mathbf{x}_o\cdot\mathbf{t}^{O_j}_i/\tau)}{\sum_{i=1}^{n_o}\sum_{j=1}^{k_o}\exp(\mathbf{x}_o\cdot\mathbf{t}^{O_j}_i/\tau)}. \quad (13)$$

3.4 Group-Aware Pair Enhancement

To improve the compositional recognition ability of the model, we introduce a pair branch. It utilizes the outputs of state and object branches as inputs to learn semantic reorganization and interaction. The pair branch consists of a learnable network Φ_P. We concat the state features $\mathbf{x}_s$ and object features $\mathbf{x}_o$ as inputs, and the outputs are features with the same dimension as $\Phi_I(x)$. To enhance the performance of Φ_P, we introduce group-aware feature enhancement to augment the input space. Although previous approaches have used similar enhancement techniques [Panda and Mukherjee, 2024][Jing *et al.*, 2024], the difference is that we argue that

not all enhancements are meaningful due to the heterogeneity of groups. For example, it is inappropriate to enhance an *old dog* with *old* semantics extracted from an *old car*. Therefore, we suggest that compatibility between instances should first be checked.

To realize this, we employ the group confidence scores in Eq. (1) to detect homogeneous groups. For one instance, the group confidence scores of its state features can be denoted as $p_r^S(i) = [p_r^{S_1}(i), p_r^{S_2}(i), ..., p_r^{S_{k_s}}(i)]$. The compatibility of two instances with respect to the state can then be quantized as $sim(i,j)$. Similar to Eq. (4), we can obtain an adjacency matrix $\hat{A}^s$:

$$\hat{A^s}_{ij} = sim(i,j) \cdot \mathbb{I}[sim(i,j) \geq \zeta] \cdot \mathbb{I}[s_i = s_j], \tag{14}$$

where $sim(i,j) = \cos(p_r^S(i), p_r^S(j))$. Then the enhanced state features are formulated as:

$$\hat{\mathbf{x}}_s = \hat{A^s} \otimes \mathbf{x}_s, \hat{A^s} = \frac{\exp(\hat{A^s} + \Lambda)}{\sum_{j=1}^{B} \exp(\hat{A^s}_{ij} + \Lambda)}. \tag{15}$$

Similarly, we can get the enhanced object features, denoted as $\hat{\mathbf{x}}_o$. Then we concat them to form the pair features, which are denoted as $\bar{\mathbf{x}}_p = \hat{\mathbf{x}}_s \oplus \hat{\mathbf{x}}_o$, where $\oplus$ means concat. Then we can get the output of pair branch and pair prompt representations according to Eq. (8):

$$\mathbf{x}_p = \Phi_P(\bar{\mathbf{x}}_p), \mathbf{t}_i^C = \Phi_T(\mathbf{P}_i^C). \tag{16}$$

The training loss is formulated as:

$$\mathcal{L}_{pair} = -\frac{1}{B} \sum^{B} \log \frac{\exp(\mathbf{x}_p \cdot \mathbf{t}_i^C / \tau)}{\sum_{i=1}^{|\mathcal{C}^{se}|} \exp(\mathbf{x}_p \cdot \mathbf{t}_i^C / \tau)}. \tag{17}$$

3.5 Summarize

Total Training Loss. The basic loss that applies CLIP to the CZSL task is denoted as:

$$\mathcal{L}_{base} = -\frac{1}{B} \sum^{B} \log \frac{\exp(\Phi_I(x) \cdot \mathbf{t}_i^C / \tau)}{\sum_{i=1}^{|\mathcal{C}^{se}|} \exp(\Phi_I(x) \cdot \mathbf{t}_i^C / \tau)}. \tag{18}$$

Then the total training loss of the propose method is denoted as:

$$\mathcal{L}_{HGRL} = \mathcal{L}_{base} + \lambda \cdot (\mathcal{L}_{state} + \mathcal{L}_{object} + \mathcal{L}_{pair}), \tag{19}$$

where λ is a hyper-parameter.

Inference. We integrate the outputs of backbone, state, object, and pair branches for joint prediction. The output of backbone is represented as:

$$p_{base}(c|\Phi_I(x)) = \frac{\exp(\Phi_I(x) \cdot \mathbf{t}_i^C / \tau)}{\sum_{i=1}^{|\mathcal{C}^{se}|} \exp(\Phi_I(x) \cdot \mathbf{t}_i^C / \tau)}. \tag{20}$$

Similarly, the output of pair branch is denoted as $p_{pair}(c|\mathbf{x}_p)$. For output of state branch, it is denoted as:

$$p_{state}(s|\mathbf{x}_s) = \frac{\sum_{j=1}^{k_s} \cdot \exp(\mathbf{x}_s \cdot \mathbf{t}_i^{S_j} / \tau)}{\sum_{i=1}^{n_s} \sum_{j=1}^{k_s} \exp(\mathbf{x}_s \cdot \mathbf{t}_i^{S_j} / \tau)}. \tag{21}$$

Similarly, we can get the output of object branch, denoted as $p_{object}(o|\mathbf{x}_o)$. The final prediction is:

$$\arg\max_{c \in \mathcal{C}^{tar}} p_{base} + p_{pair} + p_{state} \times p_{object}, \tag{22}$$

where $\mathcal{C}^{tar}$ is $\mathcal{C}^{se} \bigcup \mathcal{C}^{use}$ for closed-world setting or $\mathcal{C}$ for open-world setting.

3.6 Theoretical Insights

We provide theoretical insights for our approach from the perspective of domain adaptation [Ben-David *et al.*, 2006] since one state (object) on different objects (states) can be viewed as different domains [Zhang *et al.*, 2022].

Definition 1 *Given a seen domain, $\mathcal{S}$, and an unseen domain, $\mathcal{U}$. We have,*

$$\epsilon_U(h) \leq \epsilon_S(h) + d(\mathcal{S}, \mathcal{U}) + \gamma, \tag{23}$$

where $\epsilon_U(h), \epsilon_S(h)$ is the expected risks for unseen and seen domains and h is the model. $d(,)$ is the distributional divergence and γ is a constant denotes the minimum of the risk sum of h over $\mathcal{S}$ and $\mathcal{U}$.

Def. 1 reveals a loose upper bound which is determined by the distributional divergence between $\mathcal{S}$ and $\mathcal{U}$. Let $\mathcal{S}^* \in \mathcal{S}$ denotes the homogeneous group for $\mathcal{U}$ and we have $\mathcal{S}^\sharp = \mathcal{S} - \mathcal{S}^*$. Then $d(\mathcal{S}, \mathcal{U})$ could be decomposed into:

$$d(\mathcal{S}^*, \mathcal{U} | p(\mathbf{g}_u = \mathbf{g}_{s^*})) + d(\mathcal{S}^\sharp, \mathcal{U} | p(\mathbf{g}_u \neq \mathbf{g}_{s^*})), \tag{24}$$

where $\mathbf{g}$ denotes group label. Obviously $d(\mathcal{S}^*, \mathcal{U}) \leq d(\mathcal{S}, \mathcal{U})$ and $d(\mathcal{S}^*, \mathcal{U}) \leq d(\mathcal{S}^\sharp, \mathcal{U})$. To this end, our method first creates conditions for separating homogeneous group $\mathcal{S}^*$ by constructing a multi-expert network as well as distributed group prompts. Then, the text co-occurrence probability graph is utilized to guide the clustering of homogeneous categories thereby increasing the probability $p(\mathbf{g}_u = \mathbf{g}_{s^*})$.

4 Experiments

4.1 Experiment Settings

Datasets and Evaluation Metrics. We perform experiments on three commonly used datasets including MIT-States [Isola *et al.*, 2015], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem *et al.*, 2021]. MIT-States has 115 states and 245 objects. The number of seen compositions is 1262 and unseen compositions is 400. UT-Zappos is a small footwear dataset with 16 states and 12 objects. There are 83 seen compositions for training and 18 unseen compositions for testing. C-GQA is a challenging dataset containing 413 states and 674 objects. There are 5592 seen compositions and 923 unseen combinations. We follow the evaluation protocol of previous work[Huang *et al.*, 2024][Jing *et al.*, 2024]. Accuracy curve is plotted by adjusting the calibration bias for the scores of unseen pairs. The area under the curve called **AUC** is recorded to measure model performance. Simultaneously the best seen accuracy **S** and the best unseen accuracy **U** are reported. We also report the best harmonic mean **HM** between seen accuracy and unseen accuracy.

Implementation Details. For a fair comparison, we adopt the pre-trained CLIP [Radford *et al.*, 2021] ViT-L/14 model as the backbone. The MLP in GAVR is one-layer Multi-head Attention and the expert is two-layer fully-connected networks. The MLP in GAPE is two-layer fully-connected networks. The learning rate is $5e^{-4}$ for UT-Zappos and $5e^{-5}$ for MIT-States and C-GQA. The batch size is 180 for UT-Zappos and 32 for MIT-States and C-GQA. We use Adam optimizer to train the model. The group number of state k_s and object k_o are set to 3 for UT-Zappos and 5 for MIT-States and C-GQA. The hyper-parameter λ is set to 1.0 for UT-Zappos and 0.1 for MIT-States and C-GQA.

	METHOD	MIT-States S	MIT-States U	MIT-States HM	MIT-States AUC	UT-Zappos S	UT-Zappos U	UT-Zappos HM	UT-Zappos AUC	C-GQA S	C-GQA U	C-GQA HM	C-GQA AUC
Closed-World	CLIP	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
	CoOp	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
	CSP	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
	PromptCompVL	48.5	47.2	35.3	18.3	64.4	64.0	46.1	32.2	-	-	-	-
	DFSP	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
	PLID	49.7	52.4	39.2	22.5	67.3	68.8	52.4	38.7	41.0	38.8	27.9	11.0
	Troika	49.0	53.0	39.3	22.1	66.8	73.8	54.6	41.7	41.0	35.7	29.4	12.4
	CDS-CZSL	50.3	52.9	39.2	22.4	63.9	74.8	52.7	39.5	38.3	34.2	28.1	11.1
	Retri-Aug	50.0	53.3	39.2	22.5	69.4	72.8	56.5	44.5	45.6	36.0	32.0	14.4
	HGRL(Ours)	51.8	53.0	**40.2**	**23.3**	73.2	73.6	**59.0**	**46.8**	44.6	41.9	**34.6**	**16.3**
Open-World	CLIP	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.3
	CoOp	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.7
	CSP	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.2
	PromptCompVL	48.5	16.0	17.7	6.1	64.6	44.0	37.1	21.6	-	-	-	-
	DFSP	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3	38.3	7.2	10.4	2.4
	PLID	49.1	18.7	20.4	7.3	67.6	55.5	46.6	30.8	39.1	7.5	10.6	2.5
	Troika	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33.0	40.8	7.9	10.9	2.7
	CDS-CZSL	49.4	21.8	22.1	8.5	64.7	61.3	48.2	32.3	37.6	8.2	11.6	2.7
	Retri-Aug	49.9	20.1	21.8	8.2	69.4	59.4	47.9	33.3	45.5	11.2	14.6	4.4
	HGRL(Ours)	51.4	21.1	**22.3**	**8.7**	73.2	59.6	**51.0**	**37.5**	44.6	11.8	**15.2**	**4.5**

Table 1: Main results on three datasets.

4.2 Main Results

For a fair comparison, we compare our methods with the state-of-the-art methods that employ the same backbone, including CLIP [Radford *et al.*, 2021], CoOp[Zhou *et al.*, 2022], CSP [Nayak *et al.*, 2022], PromptCompVL [Xu *et al.*, 2022], DFSP [Lu *et al.*, 2023], PLID [Bao *et al.*, 2025], Troika [Huang *et al.*, 2024], CDS-CZSL [Li *et al.*, 2024] and Retri-Aug [Jing *et al.*, 2024]. The main results are presented in Table 1.

We compare the results under both closed-world and open-world settings. AUC and HM are the two most important metrics. From the table, we can observe that our method beats the other methods by a significant margin. Specifically, in both AUC and HM metrics, we achieve the best results on all three datasets. In particular, we lead the second place by 2.3% and 2.5% in AUC and HM in the closed-world setting on UT-Zappos, and by 4.2% and 2.8% in AUC and HM in the open-world setting. In the closed-world setting on the challenging dataset CGQ we lead the second place by 1.9% and 2.6% in AUC and HM. For MIT-States, which has relatively more noise. Our method leads the second place by 0.8% and 0.9% in AUC and HM in the closed-world setting. These results illustrate that the proposed method effectively improves the generalization of the model.

4.3 Ablation Study

We conduct extensive empirical analysis to assess the impact of the individual modules as well as the hyper-parameters.

Effect of key modules. A series of ablation experiments are conducted to investigate the role of each module. The results of the experiments are shown in Table 2. From the table, it can be observed that removing any of the modules degrades the performance of the model, indicating the rationality of the design of the proposed method. Among them, GAVR and DGP have the greatest impact on the model's performance, indicating that they contribute greatly to the optimization of the visual and prompt representations, respectively. In addition, the results show that the introduction of the text co-occurrence probability graph (TCPG) improves the performance of the model, demonstrating that it indirectly facilitates the representational ability of GAVR, guiding it to discover and aggregate homogeneous groups.

		UT-Zappos HM	UT-Zappos AUC	C-GQA HM	C-GQA AUC
CW	Full	59.0	46.8	34.6	16.3
	w/o GAVR	52.1	39.4	28.9	11.7
	w/o TCPG	55.5	43.1	32.1	14.4
	w/o DGP	52.8	40.4	31.3	13.8
	w/o GAPE	55.5	43.4	33.3	15.1
OW	Full	51.0	37.5	15.2	4.5
	w/o GAVR	47.0	33.0	12.5	3.2
	w/o TCPG	50.1	36.4	13.8	3.9
	w/o DGP	48.5	35.0	13.6	3.7
	w/o GAPE	49.6	36.0	13.5	3.9

Table 2: Ablation studies. *CW* and *OW* mean closed-world and open-world. *TCPG* means text co-occurrence probability graph. w/o means remove.

Effect of group number k_s and k_o. Due to the lack of supervision by hierarchical labels, the potential number of

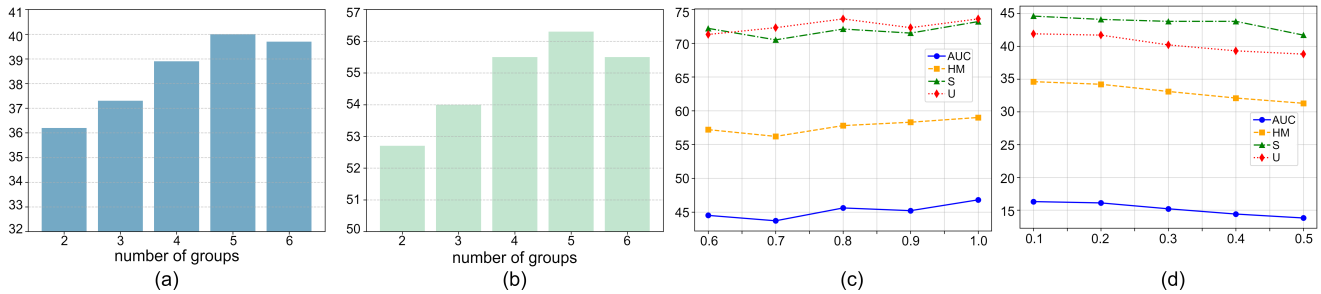

Figure 3: (a) The effect of group number for state accuracy. (b) The effect of group number for object accuracy. (c) The sensitivity of λ on UT-Zappos. (d) The sensitivity of λ on C-GQA.

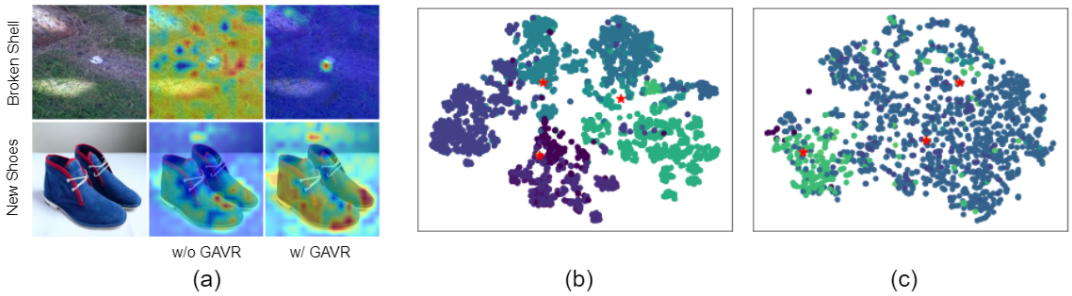

Figure 4: (a) Attention visualization for GAVR. (b-c) T-SNE analysis for DGP. Red pentagrams indicate group prompt representations. Dots indicate image representations. (b) Shows the state whose label is *Synthetic* in UT-Zappos. Different colored dots indicate different objects. (c) Shows the object whose label is *Shoes.Sneakers.and.Athletic.Shoes* in UT-Zappos. Different colored dots indicate different states.

groups is agnostic. We investigate the relationship between the setting of the group number and the model performance. The results are shown in Fig. 3 (a) and (b), which demonstrates the effect of the setting of the group number on the state accuracy and object accuracy on MIT-States. The results show that when the number of groups is set too small, e.g., 2, a substantial performance decay occurs, indicating that the weak expert panel has difficulty in handling the heterogeneous group problem. And when the number of groups gradually increases, the performance does not increase linearly, which may be caused by the saturation of the number of groups.

Effect of hyper-parameter λ. We investigate the effect of the coefficient λ in Eq. 19 on the performance of the model. The results are shown in Fig. 3 (c) and (d). It can be observed that the sensitivity of the model to λ is within a reasonable range. We suggest setting a larger value on a small dataset such as UT-Zappos to promote fast model fitting, and a smaller value on a large dataset such as C-GQA to avoid unstable training.

4.4 Visualization Analysis

To deeply investigate the role of the key components, we perform a series of visualization analyses.

Attention visualization for GAVR. To further explore the effect of GAVR, we perform a visual comparison. As can be seen in Fig. 4 (a), single network (w/o GAVR) is driven to learn semantic information shared between heterogeneous groups thus easily capture pseudo-related features. In contrast, GAVR has multiple expert sub-networks that can adaptively process data from different groups, and thus the captured features are more accurate (top row). Meanwhile, the expert networks complements the group-specific semantic knowledge, so the extracted features are richer and more discriminative (bottom row).

T-SNE Analysis for DGP. To study the context learned by group prompts, we map them to the visual space. Fig. 4 (b) and (c) show the feature distribution of the same state across different objects and the same object across different states, respectively. It can be observed that the significant group heterogeneity in the same state (object). Meanwhile, The proposed group prompts capture different contexts. These group prompts provide distributed centers so that visual features do not have to be compressed into a single prototype thereby maintaining group-specific semantic information.

5 Conclusion

In this paper, we explore the drawback of existing approaches to deal with conditional dependencies, i.e., the zero-sum game of transferability and discriminability. To solve this problem, we provide a new perspective, inspired by the human notion of hierarchical clustering, that formulates state (object) representation learning into homogeneous group representation learning. To this end, we propose a new framework, HGRL, which aims to adaptively discover homogeneous groups and extract group-specific semantic information. Our method consists of three key components to facilitate both visual and prompt representations. Experiments demonstrate the effectiveness of the proposed method.

Acknowledgements

This research was supported by funding from the Hong Kong RGC General Research Fund (152211/23E, 15216424/24E, and 152115/25E), the PolyU Internal Fund (P0056171), and the Huawei Gifted Fund.

References

[Atzmon *et al.*, 2020] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33:1462–1473, 2020.

[Bao *et al.*, 2025] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. In *European Conference on Computer Vision*, pages 107–123. Springer, 2025.

[Ben-David *et al.*, 2006] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[Chalkidis *et al.*, 2020] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, 2020.

[Chen and Grauman, 2014] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 200–207, 2014.

[Chen *et al.*, 2019] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019.

[Dekel *et al.*, 2004] Ofer Dekel, Joseph Keshet, and Yoram Singer. An online algorithm for hierarchical phoneme classification. In *International workshop on machine learning for multimodal interaction*, pages 146–158. Springer, 2004.

[Ge *et al.*, 2022] Jiannan Ge, Hongtao Xie, Shaobo Min, Pandeng Li, and Yongdong Zhang. Dual part discovery network for zero-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3244–3252, 2022.

[Hao *et al.*, 2023] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Learning attention as disentangler for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2023.

[Huang *et al.*, 2024] Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24005–24014, 2024.

[Huo *et al.*, 2024] Fushuo Huo, Wenchao Xu, Song Guo, Jingcai Guo, Haozhao Wang, Ziming Liu, and Xiaocheng Lu. Procc: Progressive cross-primitive compatibility for open-world compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12689–12697, 2024.

[Isola *et al.*, 2015] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.

[Jacobs *et al.*, 1991] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[Jing *et al.*, 2024] Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2652–2660, 2024.

[Kim *et al.*, 2023] Hanjae Kim, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Hierarchical visual primitive experts for compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5675–5685, 2023.

[Li *et al.*, 2020] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11316–11325, 2020.

[Li *et al.*, 2023] Yun Li, Zhe Liu, Saurav Jha, and Lina Yao. Distilled reverse attention network for open-world compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1782–1791, 2023.

[Li *et al.*, 2024] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17037–17046, 2024.

[Liu *et al.*, 2023] Zhe Liu, Yun Li, Lina Yao, Xiaojun Chang, Wei Fang, Xiaojun Wu, and Abdulmotaleb El Saddik. Simple primitives with feasibility-and contextuality-dependence for open-world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Lu *et al.*, 2023] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023.

[Misra *et al.*, 2017] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017.

[Naeem *et al.*, 2021] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021.

[Nagarajan and Grauman, 2018] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018.

[Nan *et al.*, 2019] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019.

[Nayak *et al.*, 2022] Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2022.

[Novack *et al.*, 2023] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023.

[Otero *et al.*, 2009] Fernando EB Otero, Alex A Freitas, and Colin G Johnson. A hierarchical classification ant colony algorithm for predicting gene ontology terms. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 7th European Conference, EvoBIO 2009 Tübingen, Germany, April 15-17, 2009 Proceedings 7*, pages 68–79. Springer, 2009.

[Panda and Mukherjee, 2024] Aditya Panda and Dipti Prasad Mukherjee. Compositional zero-shot learning using multi-branch graph convolution and cross-layer knowledge sharing. *Pattern Recognition*, 145:109916, 2024.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[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.

[Rao *et al.*, 2024a] Zhijie Rao, Jingcai Guo, Xiaocheng Lu, Jingming Liang, Jie Zhang, Haozhao Wang, Kang Wei, and Xiaofeng Cao. Dual expert distillation network for generalized zero-shot learning. *arXiv preprint arXiv:2404.16348*, 2024.

[Rao *et al.*, 2024b] Zhijie Rao, Jingcai Guo, Luyao Tang, Yue Huang, Xinghao Ding, and Song Guo. Srcd: Semantic reasoning with compound domains for single-domain generalized object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Ruis *et al.*, 2021] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34:10641–10653, 2021.

[Scarselli *et al.*, 2008] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[Silla and Freitas, 2011] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72, 2011.

[Wang *et al.*, 2023] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2023.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3741–3749, 2019.

[Xu *et al.*, 2022] Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077*, 2022.

[Yao *et al.*, 2024] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27391–27401, 2024.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014.

[Zhang *et al.*, 2022] Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. In *European Conference on Computer Vision*, pages 339–355. Springer, 2022.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[Zhu and Bain, 2017] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017.