Retrieval-Augmented Primitive Representations for Compositional Zero-Shot Learning

Chenchen Jing[1], **Yukun Li**[2], **Hao Chen**[1*], **Chunhua Shen**[1]

[1] Zhejiang University, China
[2] School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, China
{jingchenchen,haochen.cad,chunhuashen}@zju.edu.cn, liyk@mail.nwpu.edu.cn

Abstract

Compositional zero-shot learning (CZSL) aims to recognize unseen attribute-object compositions by learning from seen compositions. Composing the learned knowledge of seen primitives, *i.e.*, attributes or objects, into novel compositions is critical for CZSL. In this work, we propose to explicitly retrieve knowledge of seen primitives for compositional zero-shot learning. We present a retrieval-augmented method, which augments standard multi-path classification methods with two retrieval modules. Specifically, we construct two databases storing the attribute and object representations of training images, respectively. For an input training/testing image, we use two retrieval modules to retrieve representations of training images with the same attribute and object, respectively. The primitive representations of the input image are augmented by using the retrieved representations, for composition recognition. By referencing semantically similar images, the proposed method is capable of recalling knowledge of seen primitives for compositional generalization. Experiments on three widely-used datasets show the effectiveness of the proposed method.

Introduction

Compositional generalization, understanding unseen combinations composed of seen primitives, is one of the fundamental properties of human intelligence (Fodor and Pylyshyn 1988). Aiming to evaluate such ability of vision models, compositional zero-shot learning (CZSL) (Misra, Gupta, and Hebert 2017; Purushwalkam et al. 2019) requires recognizing unseen attribute-object compositions by learning from seen compositions. Specifically, the training set in CZSL contains images with compositional concepts, such as *wet-sand* and *young-cat*. Given a testing image, the goal is to assign a novel compositional concept, *e.g. wet-cat*, to the image by composing the primitives, wet and cat, learned from the training data, as shown in Figure 1 (a).

To compose the seen primitives into unseen compositions, two challenges must be considered. Firstly, there are semantic entanglements between objects and attributes (Atzmon et al. 2021; Anwaar, Pan, and Kleinsteuber 2022). For an image labeled as *ancient-building*, it is hard to tell which visual features can be captured as a building, and which, as ancient. Disentangling and composing primitives of training samples is thus non-trivial. Secondly, visual concepts usually follow a long-tailed distribution (Salakhutdinov, Torralba, and Tenenbaum 2011; Purushwalkam et al. 2019). Many classes are rare and may not be well-learned for compositional generalization.

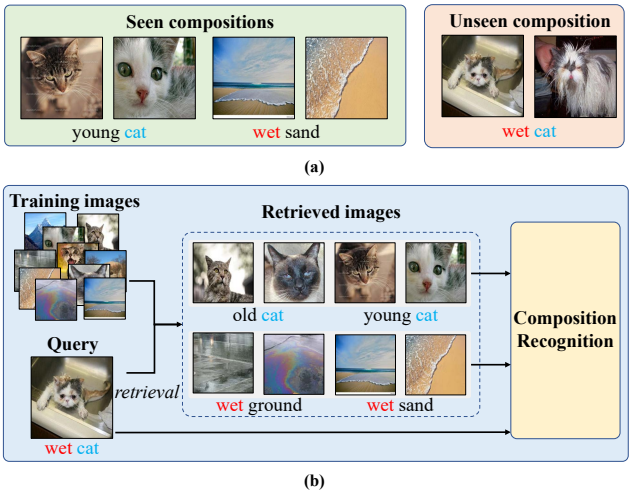

Figure 1: Illustration of primitive retrieval for compositional zero-shot learning. (a) shows two seen compositions in the training set of the MIT-States (Isola, Lim, and Adelson 2015) dataset on the left and an unseen composition in the testing set on the right. (b) shows explicitly retrieving relevant images to identify novel compositions.

Inspired by the ability of humans to perform associative learning to recall relevant concepts in deep memories (Chen et al. 2022), in this work, we propose to improve compositional generalization with retrieval and association. The intuition is that the aforementioned challenges can be alleviated if we construct a knowledge base from the training data. Referencing related knowledge in the test time could provide a strong enhancement signal to help the model recall learned primitives for generalization, as shown in Figure 1 (b).

We introduce a retrieval-augmented method, which enables explicit knowledge retrieval of seen primitives, for

*corresponding author

compositional zero-shot learning. The proposed method augments standard multi-path classification pipelines (Yang et al. 2022; Wang et al. 2023) with retrieval modules. Specifically, we extract attribute representations and object representations of training images, and construct two databases to store the attribute and object representations, respectively. For an input training/testing image, the retrieval modules retrieve representations of training images with the same attribute/object, and use the retrieved representations to augment the attribute/object representations, respectively.

The databases of the proposed method serve as external memories along with the composition recognition pipeline. After each training epoch, the representations of the databases will be updated. Intuitively, throughout the learning process, the model refines the current representations continuously with the retrieved relevant representations. And with the refined representations, the retrieval modules can find more relevant images. Thus the proposed method is supposed to learn more separable primitive representations, benefiting from the retrieval. In addition, the databases explicitly store the knowledge for primitive representation learning, specifically for the tail primitives. We conduct extensive experiments on three widely-used CZSL datasets under two settings. The experimental results show the effectiveness of the proposed method.

The contributions of this paper are summarized as follows:

1. We propose to explicitly retrieve knowledge of seen primitives to recognize unseen compositions for compositional generalization.
2. We present a retrieval-augmented method, which augments multi-path classification methods with two retrieval modules, for compositional zero-shot learning.

Related Work

Compositional Zero-Shot Learning

The task of CZSL aims to recognize unseen attribute-object compositions by learning from seen compositions. Existing methods mainly cast the task of CZSL into a supervised classification task by training one classifier for composition (Misra, Gupta, and Hebert 2017; Naeem et al. 2021), or two classifiers for attribute and object (Li et al. 2020; Purushwalkam et al. 2019), or three classifiers for composition, attribute, and object (Yang et al. 2022; Wang et al. 2023).

To learn disentangled representations for CZSL, Atzmon *et al.* (Atzmon et al. 2021) propose to ensure conditional independence between attribute and object representations via causal inference. Saini *et al.* (Saini, Pham, and Shrivastava 2022) use visually decomposed features to hallucinate representative embeddings of the seen and novel compositions to regularize the model learning. Zhang et al. (Zhang et al. 2022) treated CZSL as a domain generalization task to learn attribute-invariant and object-invariant representations. The aforementioned methods enforce constraints on the model learning, but may not be well-compatible for unseen compositions in testing. By contrast, our method enables to perform retrieval in both the training and testing phase.

With the recent advance in pre-trained vision-language models, CLIP-based CZSL methods (Nayak, Yu, and Bach 2023; Lu et al. 2023a; Huang et al. 2023; Bao et al. 2023) achieved state-of-the-art performance, benefiting from the vision-language alignments learned from large-scale data. CSP (Nayak, Yu, and Bach 2023) first uses the CLIP (Radford et al. 2021) in CZSL. They replace the classes in textual prompts with trainable attributes and object tokens. DFSP (Lu et al. 2023a) uses a cross-modal decomposed fusion module to exploit decomposed language features in image feature learning. Troika (Huang et al. 2023) jointly models the vision-language alignments for the attribute, object, and composition using the CLIP. PLID (Bao et al. 2023) leverages pre-trained large language models to formulate the language-informed class distribution, and enhance the compositionality of the softly prompted class embedding. The aforementioned work mainly focuses on parameter-efficient fine-tuning of the CLIP. By contrast, the proposed method focuses on primitive retrieval and uses the CLIP as the backbone.

Retrieval-Augmented Models

Augmenting traditional models with external memories have recently drawn attention in computer vision (Long et al. 2022; Blattmann et al. 2022; Chen et al. 2022; Rong et al. 2023). RAC (Long et al. 2022) retrieves relevant images and uses textual representations of corresponding labels for the long-tailed image classification task. RePrompt (Rong et al. 2023) retrieves images to learn visual prompts for few-shot image classification. RAC and RePrompt focus on object category classification. In both methods, the retrieved images for an image are determined in an offline manner and remain unchanged across model learning. By contrast, the proposed method aims to recognize both the object category and the attribute category. The associated images and the corresponding representations are constantly changing across different training epochs for the separability of the learned primitive representations.

Method

The proposed method explicitly retrieves seen primitives for CZSL, by building databases containing representations of training images and using retrieve modules to augment the representations of an input image, as shown in Figure 2. In the following, we first formulate the task of CZSL and then introduce the proposed method in detail.

Formulation

Compositional zero-shot learning aims at learning a model from limited compositions of attributes (*e.g.*, yellow, wet) and objects (*e.g.*, flower, ground) to recognize an image from novel compositions. Given an attribute set $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ and an object set $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$, the compositional class set $\mathcal{C} = \mathcal{A} \times \mathcal{O}$ is defined as their Cartesian product. The class set $\mathcal{C}$ can be divided into two disjoint sets, the seen set $\mathcal{C}^s$ and the unseen set $\mathcal{C}^u$, where $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$ and $\mathcal{C}^s \cup \mathcal{C}^u \subset \mathcal{C}$. The training images only contain classes from the $\mathcal{C}^s$, while the testing set contains

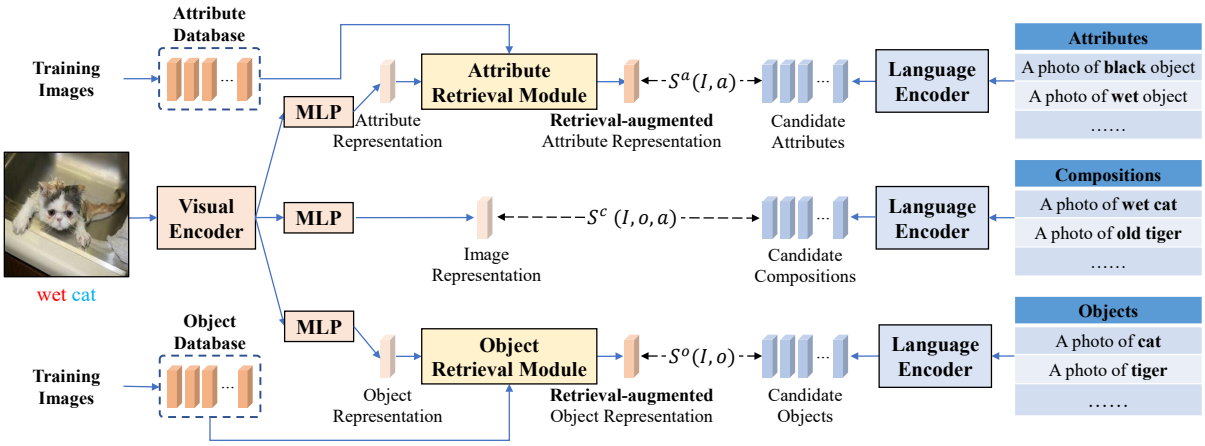

Figure 2: Overview of the proposed method. For an input image, the method uses a visual encoder and three adapters to obtain the image representation, object representation, and attribute representation. For compositions, we use a language encoder to obtain the textual representations of all candidate compositions, attributes, and objects. We build two databases to store attribute representations and object representations of training images, and use two retrieval modules to retrieve representations of images with the same object/attribute to augment the object/attribute representation, respectively. The obtained visual and textual representations are used to compute the compatibility scores for composition recognition.

both seen classes and unseen classes, as the standard generalized zero-shot learning (Pourpanah et al. 2022; Liu et al. 2021).

Given a test image $I \in \mathcal{I}$, the task of CZSL is to predict a class label $c = (a, o)$ from the testing class set. In the closed-world setting, only the known compositions (compositions of the whole dataset) are considered, *i.e.*, $\mathcal{C}^{test} = \mathcal{C}^s \cup \mathcal{C}^u$. That is, the test class set contains all seen classes for the training images and unseen classes of the test set. By contrast, in the challenging open-world setting, the test class set is all possible compositions, *i.e.*, $\mathcal{C}^{test} = \mathcal{C}$.

Formally, CZSL models are required to model a compatibility score function $S : \mathcal{I} \times \mathcal{S} \times \mathcal{O} \rightarrow \mathbb{R}$ between an image I and a candidate composition. To fully characterize the contextuallity of attributes and objects, existing three-path methods (Wang et al. 2023; Huang et al. 2023) usually jointly consider three kinds of compatibility, that is, the attribute compatibility $S^a(I, a)$, the object compatibility $S^o(I, o)$, and the composition compatibility $S^c(I, a, o)$.

Feature Encoding

We use the image encoder and text encoder of the CLIP (Radford et al. 2021) as the visual backbone and textual backbone, respectively. Given an input image I, we use the visual backbone, a vision transformer (ViT) (Dosovitskiy et al. 2021), to obtain the visual representations. The image is split into patches and inputted to the ViT to obtain the representation of the [CLS] token $\boldsymbol{v} \in \mathbb{R}^d$, where d is the dimension of feature embedding of the CLIP. Three multi-layer perceptions (MLPs) are used to transform the representation to the image representation $\boldsymbol{v}^I$, the attribute representation $\boldsymbol{v}^a$, and the object representation $\boldsymbol{v}^o$, respectively.

For the textual inputs, we use the soft prompt (Nayak, Yu, and Bach 2023) to obtain the textual representations for all candidate compositions, attributes, and objects. We create a prompt template like"a photo of *[class]*" for each compatibility scoring sub-task. For the composition compatibility, we feed the text encoder with "a photo of *[attribute] [object]*", as shown in the Figure 2. The "a photo of *[attribute]* object" and "a photo of *[object]*" are used for candidate attributes and candidate objects, respectively. The [attribute] and [object] tokens are trainable and initiated with the corresponding word embeddings extracted by the CLIP. These prompts are fed into the textual backbone to obtain textual representations $\boldsymbol{T}^c \in \mathbb{R}^{N_c \times d}$, $\boldsymbol{T}^a \in \mathbb{R}^{N_a \times d}$, $\boldsymbol{T}^o \in \mathbb{R}^{N_o \times d}$. The N_c, N_a, and N_o are the numbers of candidate compositions, attributes, and objects, respectively.

Database Construction

We construct two databases, $\mathcal{D}^a$ and $\mathcal{D}^o$ with training images, for retrieving images with the same object and the same attribute, respectively. For each database $\mathcal{D}^p, p \in \{\mathrm{a}, \mathrm{o}\}$, we first select a subset of training images, and extract corresponding visual representations.

Taking the attribute database as an example, we choose N_D images for each attribute to avoid attribute-level biases. Considering there are usually multiple compositions associated with the attribute, we sample images from these compositions as evenly as possible by using the Greedy algorithm. For example, suppose that we need to choose 16 images for an attribute *wet*, the training set contains three relevant compositions, *wet ground*, *wet sand*, *wet basement*, *wet well*, and there are 15, 7, 5, 3 images belong to these compositions, respectively. Then we will sample 5, 4, 4, and 3 images for these compositions, respectively.

After the image sampling, we extract corresponding visual representations with the visual backbone to construct the databases. The obtained databases can be represented

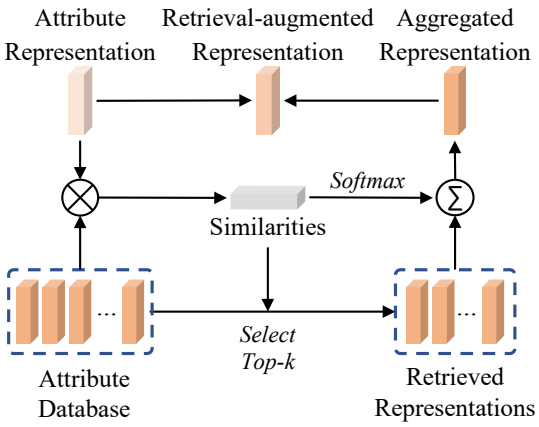

Figure 3: The architecture of the attribute retrieval module. For an attribute representation, the module retrieval relevant representations from the attribute database and aggregate the retrieved representations to obtain retrieval-augmented representation.

as $\mathcal{D}^a = [\boldsymbol{f}_1^a, \boldsymbol{f}_2^a, \cdots, \boldsymbol{f}_{N_I^a}^a]$ and $\mathcal{D}^o = [\boldsymbol{f}_1^o, \boldsymbol{f}_2^o, \cdots, \boldsymbol{f}_{N_I^o}^o]$. Note that we extract visual representations of these images after each epoch to update the databases.

Retrieval Module

In model training/testing, the retrieval modules retrieve images from the databases and use the representations of images to augment the primitive representations of an input image. Specifically, two retrieval modules are introduced to augment the attribute representation and object representation, respectively. Figure 3 shows the architecture of the attribute retrieval module. In the following, we illustrate the retrieval and augmentation process by taking the attribute retrieval module as an example.

For the attribute representation $\boldsymbol{v}^a$ of an input image, the attribute retrieval module computes the similarities between the representation and all representations of the attribute database as

$$s_i^a = \cos(\boldsymbol{v}^a, \boldsymbol{f}_i^a), \tag{1}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity function of two vectors.

The representations in the database are sorted with similarities in a descending manner. We select top-K representations with the highest similarity to augment the attribute representation. These representations are aggregated via weighted average to obtain the aggregated representation $\boldsymbol{u}_a$ as

$$\boldsymbol{u}^a = \sum_{i=1}^{K} \alpha_i^a \boldsymbol{f}_{idx(i)}^a, \quad \alpha_i^a = \frac{\exp(s_i^a)}{\sum_{j=1}^{K} \exp(s_j^a)}, \tag{2}$$

where $idx(i)$ is a function that returns the index of the i-th seclted representation in the attribute database D^a and α_i^a is the weight of the i-th representation.

Then we fuse the aggregated representation with the original attribute representation to obtain the retrieval-augmented attribute representation as

$$\boldsymbol{v}_r^a = \beta \boldsymbol{u}^a + (1-\beta)\boldsymbol{v}^a, \tag{3}$$

where $\beta \in [0,1]$ is a hyper-parameter to balance the two representations. Similarly, the object retrieval module searched relevant representations in D^a with the representation $\boldsymbol{v}^o$, to obtain the retrieval-augmented object representation $\boldsymbol{v}_r^o$.

Optimization

To encourage the retrieval modules to retrieve representations of relevant images, two losses are introduced in model learning. Firstly, considering the entanglement of the attribute and the object in an image, we devise a de-bias loss. We penalize the object representations for predicting the ground truth attribute labels, and attribute representations for predicting the ground truth object labels, and compute the loss as

$$\mathcal{L}_{de} = \cos(\boldsymbol{v}_r^a, \boldsymbol{T}_{gt}^o) + \cos(\boldsymbol{v}_r^o, \boldsymbol{T}_{gt}^a), \tag{4}$$

where $\boldsymbol{T}_{gt}^o$ and $\boldsymbol{T}_{gt}^a$ are the textual representations for the ground-truth object label and ground-truth attribute label, respectively. By reducing the entanglement between the object representations and attribute representations, the retrieval modules can more accurately find representations of images with the same object/attribute. Thus the entanglement will be further reduced. In other words, the de-bias loss and the retrieval module can promote each other.

Secondly, we introduce a retrieval loss to directly enforce the retrieval module to obtain representations of images with the same attribute/object. Specifically, for each representation, we sample the top-M representations with the highest similarities.

$$\begin{aligned}\mathcal{L}_{re} = &\sum_{i}^{M} \left(\sigma\left(s_i^a\right)\right)^{1-l_i^a} \left(1-\sigma\left(s_i^a\right)\right)^{l_i^a} \\ &+ \sum_{i}^{M} \left(\sigma\left(s_i^o\right)\right)^{1-l_i^o} \left(1-\sigma\left(s_i^o\right)\right)^{l_i^o},\end{aligned} \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, l_i^a indicates whether the i-representation has the same attribute/object label with an input image. Note that, the representations in the database are not trainable and the retrieve module is non-parametric. Thus the retrieval loss only optimizes the primitive representations for retrieval.

Apart from the aforementioned two losses, we use the standard cross-entropy loss to encourage the model to explicitly recognize the composition, attribute, and object. The compatibility scores between an image I with the ground-truth composition $c_{gt} = (a_{gt}, o_{gt})$ with the aforementioned representations can be computed as

$$\begin{aligned}\mathcal{S}^a(I, a_{gt}) &= \cos(\boldsymbol{v}_r^a, \boldsymbol{T}_{gt}^a), \\ \mathcal{S}^o(I, o_{gt}) &= \cos(\boldsymbol{v}_r^o, \boldsymbol{T}_{gt}^o), \\ \mathcal{S}^c(I, a_{gt}, o_{gt}) &= \cos(\boldsymbol{v}^I, \boldsymbol{T}_{gt}^c),\end{aligned} \tag{6}$$

	Closed-world	MIT-States				C-GQA				UT-Zappos			
	Model	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
Without CLIP	CompCos (Mancini et al. 2021)	4.5	16.4	25.3	24.6	2.6	12.4	28.1	11.2	28.7	43.1	59.8	62.5
	OADis (Saini, Pham, and Shrivastava 2022)	5.9	18.9	31.1	25.6	-	-	-	-	30.0	44.4	59.5	65.5
	CGE (Naeem et al. 2021)	6.5	21.4	32.8	28.0	4.2	15.5	33.5	16.0	33.5	**60.5**	64.5	71.5
	Co-CGE (Mancini et al. 2022)	6.6	20.0	32.1	28.3	4.1	14.4	33.3	14.9	33.9	48.1	62.3	66.3
	CANet (Wang et al. 2023)	5.4	17.9	29.0	26.2	3.4	14.5	30.0	13.2	33.1	47.3	61.0	66.3
	CAPE (Khan et al. 2023)	6.7	20.4	32.1	28.0	4.6	16.3	33.0	16.4	35.2	49.5	62.3	68.5
With CLIP	CSP (Nayak, Yu, and Bach 2023)	19.4	36.3	46.6	49.9	6.2	20.5	28.8	26.8	33.0	46.6	64.2	66.2
	DFSP (Lu et al. 2023a)	20.6	37.3	46.9	52.0	10.5	27.1	38.2	32.9	36.9	47.2	66.7	71.7
	DRPT (Lu et al. 2023b)	-	-	-	-	6.5	20.5	29.2	28.7	38.5	52.3	64.5	69.4
	Troika (Huang et al. 2023)	22.1	39.3	49.0	53.0	12.4	29.4	41.0	35.7	41.7	54.6	66.8	**73.8**
	PLID (Bao et al. 2023)	22.1	39.0	49.7	52.4	11.0	27.9	41.0	**38.8**	38.7	52.4	67.3	68.8
	Ours	**22.5**	**39.2**	**50.0**	**53.3**	**14.4**	**32.0**	**45.6**	36.0	**44.5**	56.5	**69.4**	72.8

Table 1: The results of the proposed methods and the state-of-the-art on CZSL datasets in the *closed-world* setting.

where $\boldsymbol{v}_r^a$ and $\boldsymbol{v}_r^o$ are retrieval-augmented primitive representations. Thus classification losses are calculated as

$$\begin{aligned}\mathcal{L}^a &= -log\frac{\exp(\mathcal{S}^a(I,a_{gt})/\tau)}{\sum_{k=1}^{|\mathcal{A}|}\exp(\mathcal{S}^a(I,a_k)/\tau)},\\ \mathcal{L}^o &= -log\frac{\exp(\mathcal{S}^o(I,o_{gt})/\tau)}{\sum_{k=1}^{|\mathcal{O}|}\exp(\mathcal{S}^o(I,o_k)/\tau)},\\ \mathcal{L}^c &= -log\frac{\exp(\mathcal{S}^c(I,a_{gt},o_{gt})/\tau)}{\sum_{k=1}^{|\mathcal{C}^s|}\exp(\mathcal{S}^c(I,a_k,o_k)/\tau)},\end{aligned} \tag{7}$$

where $\tau \in \mathbb{R}$ is the pre-defined temperature parameter of CLIP. The overall loss in the model learning is given by

$$\mathcal{L} = \lambda^1\mathcal{L}^s + (1-\lambda^1)(\mathcal{L}^o + \mathcal{L}^c) + \lambda^2\mathcal{L}_{de} + \lambda^3\mathcal{L}_{re}, \tag{8}$$

where λ^1, λ^2, and λ^3 are hyper-parameters to balance the losses.

Inference

During inference, the primitive-level scores and the composition-level scores are combined to complement the composition recognition. The overall compatibility score $\mathcal{S}(I,a,o)$ is calculated as

$$\mathcal{S}(I,a,o) = \lambda^1\mathcal{S}^c(I,a,o) + (1-\lambda^1)(\mathcal{S}^a(I,a) + \mathcal{S}^o(I,o)). \tag{9}$$

The composition with the highest score is predicted. Note that we use the same hyper-parameter λ^1 to balance the scores as in model learning.

Dataset	Attr	Obj	Train	Val		Test	
			Seen	Seen	Unseen	Seen	Unseen
MIT-States	115	245	1262	300	300	400	400
UT-Zappos	16	12	83	15	15	18	18
C-GQA	453	870	6963	1173	1368	1022	1047

Table 2: The statistics of the MIT-States, the UT-Zappos, and the C-GQA.

Experiment

Datasets

We evaluate the proposed method on three CZSL datasets, *i.e.*, MIT-States (Isola, Lim, and Adelson 2015), UT-Zappos (Yu and Grauman 2014), and C-GQA (Naeem et al. 2021). The MIT-States consists of $53,753$ crawled web images labeled with 1962 attribute-object (*e.g.*, *wet-dog*). The dataset contains $1,262$ seen and 300/400 unseen compositions for training and validation/testing, respectively. The UT-Zappos is a fine-grained dataset consisting of 116 kinds of shoe classes composed of 16 attributes (*e.g.*, *rubber*) and 12 objects (*e.g.* sandal. The dataset is split into 83 seen and 15/18 unseen compositions for training and validation/testing. The C-GQA is built based on the GQA dataset (Hudson and Manning 2019) for the visual question answering task (Wu et al. 2017; Jing et al. 2020). The C-GQA dataset contains 453 common attribute classes (*e.g.*, wet and old) and 870 common object classes (*e.g.*, dog and cat), and over $9,000$ composition classes. The dataset is split into $5,592$ seen and $1,040$/923 unseen compositions for training and validation/testing, respectively. The detailed dataset statistics are shown in Table 2.

Metrics

We report the standard metrics of CZSL evaluation protocol in both closed-world and open-world settings, including the best seen accuracy (**Seen**), the best unseen accuracy (**Unseen**), the best harmonic mean (**HM**) between the seen and unseen accuracy, and the area under the curve (**AUC**) of unseen versus seen accuracy. Specifically, the AUC is computed by varying the value of the calibration bias added to unseen compositions, and is thus able to describe the overall performance of a model (Purushwalkam et al. 2019). In the open-world setting, the GloVe (Pennington, Socher, and Manning 2014) is used to obtain the feasibility calibration to filter out infeasible compositions.

Implementation Details

We implement our method based on PyTorch. For the backbone, the CLIP architecture ViT-L/14 is used as previs work (Lu et al. 2023a). A single NVIDIA RTX 3090 GPU is

	Open-world Model	MIT-States AUC	HM	Seen	Unseen	C-GQA AUC	HM	Seen	Unseen	UT-Zappos AUC	HM	Seen	Unseen
Without CLIP	CompCos (Mancini et al. 2021)	0.8	5.8	21.4	7.0	0.43	3.3	26.7	2.2	18.5	34.5	53.3	44.6
	CGE (Naeem et al. 2021)	1.0	6.0	32.4	5.1	0.47	2.9	32.7	1.8	23.1	39.0	61.7	47.7
	KG-SP (Karthik, Mancini, and Akata 2022)	1.3	7.4	28.4	7.5	0.78	4.7	31.5	2.9	26.5	42.3	61.8	52.1
	Co-CGE^{CW} (Mancini et al. 2022)	1.1	6.4	31.1	5.8	0.53	3.4	32.1	2.0	23.1	40.3	62.0	44.3
	Co-CGE^{open} (Mancini et al. 2022)	2.3	10.7	30.3	11.2	0.78	4.8	32.1	3.0	23.3	40.8	61.2	45.8
With CLIP	CSP (Nayak, Yu, and Bach 2023)	5.7	17.4	46.3	15.7	1.20	6.9	28.7	5.2	22.7	38.9	64.1	44.1
	DFSP (Lu et al. 2023a)	6.8	19.3	47.5	18.5	2.40	10.4	38.3	7.2	30.3	44.0	66.8	60.0
	Troika (Huang et al. 2023)	7.2	20.1	48.8	18.7	2.7	10.9	40.8	7.9	33.0	47.8	66.4	61.2
	PILD (Bao et al. 2023)	7.3	20.4	49.1	18.7	2.5	10.6	39.1	7.5	30.8	46.6	67.6	55.5
	Ours	**8.18**	**21.8**	**49.9**	**20.1**	**4.4**	**14.6**	**45.5**	**11.2**	**33.3**	**47.9**	**69.4**	**59.4**

Table 3: The results of the proposed methods and the state-of-the-art on CZSL datasets in the *open-world* setting.

used for training and testing. For the UT-Zappos, the hyper-parameters λ^1, λ^2, and λ^3 in the losses are set as 0.8, 5.0, and 1.0. For the MIT-States, the three hyper-parameters are set as 0.2, 1.0, and 0.1. For the C-GQA, the three hyper-parameters are set as 0.2, 5.0, and 0.1. The number of retrieved images K is set as 32 for UT-Zappos and 16 for both MIT-States and C-GQA. The number of images N_D of each primitive in database construction is set as 128 for the UT-Zappos and 16 for both the MIT-States and the C-GQA, considering the classes in the MIT-States and the C-GQA are much more than classes of the UT-Zappos. The number of selected images M for the retrieval loss is is set as 256 for the UT-Zappos and 512 for both the MIT-States and the C-GQA. The weight of aggregated features β is set as 0.8 for UT-Zappos and 0.5 for both MIT-States and C-GQA. We set the training epochs of each dataset as 20. After each epoch, all the representations of the databases are updated. For the C-GQA, we tune the top 12 layers of the image encoder of CLIP with LoRA (Hu et al. 2021), a lightweight parameter efficient fine-tuning (PEFT) strategy.

Quantitative Results

Main results. We compare our method with various state-of-the-art methods, including both methods without CLIP and CLIP-based methods. The results of all the methods on the test split of MIT-States, UT-Zappos, and C-GQA under the standard closed-world setting are listed in Table 1. We observe that our method outperforms all other methods. The main reason is that benefiting from the primitive retrieval, our method can use representations with relevant images to refine the current primitive representations, and thus learn more disentangled representations progressively for compositional generalization. Besides, the databases serve as external memories explicitly storing the knowledge of the tail primitives, thus our method is able to learn more informative representations for these primitives.

We also evaluate the proposed method in the challenging open-world setting. Table 3 shows the results of all the methods on the three datasets in the open-world setting. The proposed method also outperforms all other methods, which demonstrates the effectiveness of our method for open-world compositional zero-shot learning. We obverse that the performance gap between the troika (Huang et al. 2023) and the proposed method in the open-world setting is larger than that in the closed-world setting. A possible reason is that in the challenging open-world setting, all possible compositions should be considered, which requires disentangled and composable primitive representations. Note that we use identical model weights for the two settings.

	RM	L_{de}	L_{re}	AUC	HM	Seen	Unseen
1				39.4	52.2	64.7	72.6
2	✓			40.6	54.2	66.5	71.5
3		✓		40.4	54.1	68.2	69.2
4	✓	✓		44.0	56.1	69.3	72.6
5	✓		✓	41.0	55.0	66.3	69.7
6	✓	✓	✓	**44.5**	**56.5**	**69.4**	**72.8**

Table 4: Results of different variants of our model on the the Ut-Zappos dataset. RM denotes the retrieval module. L_{de} and L_{re} are the losses.

Ablation studies. To study the effectiveness of several important components of our method, we evaluate different variants of our model by ablating certain components. The results of those models on the test split of the UT-Zappos dataset in the closed-world setting are shown in Table 4.

We firstly ablate the retrieval modules and the de-bias loss and the retrieval loss, and obtain a baseline model. The AUC of this model is much lower than our full model, which demonstrates that these components bring substantial improvements. Then, we add the retrieval module and the de-bias loss on top of the baseline model and obtain the second and the third model, respectively. The comparisons between the two models with the baseline model show that the two components are both beneficial. We further add the retrieval loss for the second model and the fourth model to obtain the fifth model and our full model, respectively. These comparisons demonstrate that the effectiveness of the retrieval loss.

Qualitative Results

Feature distributions. We visualize the feature distributions in Figure 5 to demonstrate the effectiveness of retrieval for primitive representation learning. We select three typical attributes and three typical objects and choose 16 images for each attribute/object of the C-GQA dataset. The object rep-

Figure 4: Qualitative results of the proposed method on the UT-Zappos, the MIT-States and the C-GQA. For each sample, we show an image with the ground-truth composition and the prediction of our method on the left. The retrieved images of our method are shown on the right, where the retrieved images of the attribute retrieval module are shown on the top and these of the object retrieval module are shown on the bottom. The concepts in red/blue denote the ground-truth attribute/object classes.

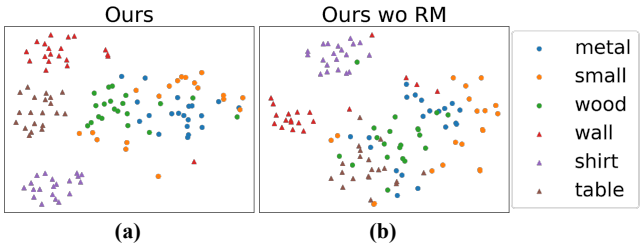

Figure 5: Feature distributions on the C-GQA dataset of our full model (a) and a model without retrieval modules (b).

resentations and the attribute representations of these images of our full model and a model without retrieval modules are shown by using the t-SNE tool (Van der Maaten and Hinton 2008). The circles of different colors denote attribute representations of images with different attributes, and the triangles of different colors denote object representations of images with different objects. It is clearly shown that the primitive representations of our model are more separable than these of the model without retrieval modules.

Qualitative examples. We provide qualitative examples from the UT-Zappos, the MIT-States and the C-GQA in Figure 4. The examples from the three datasets are shown on the upper-left/upper-right/bottom-left, respectively. For each sample, we show the input image with the ground-truth composition and the prediction of our method on the left. The retrieved images of our method are shown on the right, where the images retrieved by the object retrieval module are shown on the top and these of the attribute retrieval module are shown on the bottom. We observe that the retrieve module can relatively accurately find images with the same attribute/object. Benefiting from referencing semantically relevant images, our method can recognize the compositions. We also provide a failure case of the proposed method on the bottom-right. In the example, the "computer" is regarded as an attribute. Thus the model can not find relevant images with the same attribute and fail to figure out the correct attribute label. In this case, performing attribute retrieval by using the object information as condition may be helpful. We leave it as future work.

Conclusion

In this work, we have presented a retrieval-augmented method for compositional zero-shot learning. The proposed method enables explicitly knowledge retrieval of seen primitives for compositional generalization using two retrieval modules. Our method explicitly store attribute and object representations of training images by constructing two databases. By using the retrieval modules, our method obtains representations of relevant images from the databases to enhance the primitive representations of input images. The introduction of the de-bias loss and the retrieval loss further encourage the retrieval modules to retrieve representations of relevant images. Extensive experiments show that our method can learn separable attribute representations and object representations and achieves state-of-the-art performance for compositional zero-shot learning.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2022ZD0118700) and the China Postdoctoral Science Foundation (No. 2023M743003). This work was also partially supported by the National Key R&D Program of China (No.2022ZD0160101).

References

Anwaar, M. U.; Pan, Z.; and Kleinsteuber, M. 2022. On leveraging variational graph embeddings for open world compositional zero-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4645–4654.

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2021. A causal view of compositional zero-shot recognition. *Advances in neural information processing systems (NeurIPS)*.

Bao, W.; Chen, L.; Huang, H.; and Kong, Y. 2023. Prompting Language-Informed Distribution for Compositional Zero-Shot Learning. *arXiv preprint arXiv:2305.14428*.

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35: 15309–15324.

Chen, X.; Li, L.; Zhang, N.; Liang, X.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. *Advances in Neural Information Processing Systems*, 35: 23908–23922.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

Fodor, J. A.; and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2): 3–71.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, S.; Gong, B.; Feng, Y.; Lv, Y.; and Wang, D. 2023. Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning. *arXiv preprint arXiv:2303.15230*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 6700–6709.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections.

Jing, C.; Wu, Y.; Zhang, X.; Yunde, J.; and Wu, Q. 2020. Overcoming Language Priors in VQA via Decomposed Linguistic Representations. In *Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*, 11181–11188.

Karthik, S.; Mancini, M.; and Akata, Z. 2022. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9336–9345.

Khan, M. G. Z. A.; Naeem, M. F.; Van Gool, L.; Pagani, A.; Stricker, D.; and Afzal, M. Z. 2023. Learning Attention Propagation for Compositional Zero-Shot Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3828–3837.

Li, Y.-L.; Xu, Y.; Mao, X.; and Lu, C. 2020. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11316–11325.

Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.; Gu, L.; Zhou, J.; and Harada, T. 2021. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3794–3803.

Long, A.; Yin, W.; Ajanthan, T.; Nguyen, V.; Purkait, P.; Garg, R.; Blair, A.; Shen, C.; and van den Hengel, A. 2022. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 6959–6969.

Lu, X.; Guo, S.; Liu, Z.; and Guo, J. 2023a. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23560–23569.

Lu, X.; Liu, Z.; Guo, S.; Guo, J.; Huo, F.; Bai, S.; and Han, T. 2023b. DRPT: Disentangled and Recurrent Prompt Tuning for Compositional Zero-Shot Learning. *arXiv preprint arXiv:2305.01239*.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2021. Open World Compositional Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5222–5230.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2022. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1792–1801.

Naeem, M. F.; Xian, Y.; Tombari, F.; and Akata, Z. 2021. Learning graph embeddings for compositional zero-shot learning.

Nayak, N. V.; Yu, P.; and Bach, S. 2023. Learning to Compose Soft Prompts for Compositional Zero-Shot Learning. In *International Conference on Learning Representations (ICLR)*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. J. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Purushwalkam, S.; Nickel, M.; Gupta, A.; and Ranzato, M. 2019. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3593–3602.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Rong, J.; Chen, H.; Chen, T.; Ou, L.; Yu, X.; and Liu, Y. 2023. Retrieval-Enhanced Visual Prompt Learning for Few-shot Classification. *arXiv preprint arXiv:2306.02243*.

Saini, N.; Pham, K.; and Shrivastava, A. 2022. Disentangling Visual Embeddings for Attributes and Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13658–13667.

Salakhutdinov, R.; Torralba, A.; and Tenenbaum, J. 2011. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, 1481–1488. IEEE.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, Q.; Liu, L.; Jing, C.; Chen, H.; Liang, G.; Wang, P.; and Shen, C. 2023. Learning Conditional Attributes for Compositional Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11197–11206.

Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; and van den Hengel, A. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163: 21–40.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2022. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*.

Yu, A.; and Grauman, K. 2014. Fine-grained visual comparisons with local learning.

Zhang, T.; Liang, K.; Du, R.; Sun, X.; Ma, Z.; and Guo, J. 2022. Learning invariant visual representations for compositional zero-shot learning. In *European Conference on Computer Vision*, 339–355. Springer.